Utilização de Notas Escolares para Predição da Nota ENEM em Ciências Humanas

Juvenal Antônio Cordeiro Filho[1], Seiji Isotani[2], Bruno Elias Penteado[3]

Resumo

O presente trabalho explora a hipótese de predição da nota ENEM em ciências humanas a partir de dados pedagógicos oriundos de notas de avaliações escolares de estudantes de ensino médio de um colégio particular em São Paulo. A partir da análise, foi possível a predição da nota ENEM de ciências humanas no decurso do 3º ano do ensino médio com uma acurácia de 60,2%. O modelo apontou, ainda, conteúdos relevantes fortemente correlacionados com o desempenho ENEM já desde o 1º ano do ensino médio.

Palavras-chave: *ENEM, ensino médio, notas escolares, nota ENEM, ciências humanas*

Abstract

In the actual paper is explored the hypothesis of ENEM human sciences grade prediction based on high school test grades from students in a private school in São Paulo. The analysis allowed a prediction model correlating ENEM human sciences grade and high school grades from the 3rd high school year with an accuracy of 60.2%. Besides this, the model found important academic contents strongly correlated to students ENEM performance since the 1st high school year.

Keywords: *ENEM, high school, scholar grades, ENEM grade, human sciences*

1. Introdução

O Exame Nacional do Ensino Médio – ENEM – é o maior exame de admissão universitária do país e um dos maiores do mundo [MEC, 2015, *online*]. Sua importância,

[1] Pós-Graduando em Computação Aplicada à Educação, USP, juvenalfilho@usp.br.

[2] professor titular do Instituto de Ciências Matemáticas e da Computação, da Universidade de São Paulo (ICMC/USP), sisotani@usp.br.

[3] doutor em Ciências da Computação pelo do Instituto de Ciências Matemáticas e da Computação, da Universidade de São Paulo (ICMC/USP), brunopenteado@usp.br.

Cite as: Filho, J. & Isotani, S. & Penteado, B. (2020). Utilização de Notas Escolares para Predição da Nota ENEM em Ciências Humanas. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

contudo, vai além de ser o portal de ingresso para quase todas as universidades federais do país, além de outras tantas vagas em universidades estaduais e no ensino superior privado via ProUni e FIES.

Em seu desenho original com 63 questões, o exame foi concebido para ser um indicador da qualidade do Ensino Médio no Brasil, e, mesmo com os muitos desvios que sofreu desde sua primeira proposta[4], segue oferecendo uma série de dados importantes para a educação brasileira[5].

Justamente por sua importância, o ENEM atraiu a atenção de escolas e cursinhos pré-vestibulares na estruturação de suas grades curriculares – ao menos até a emergência da nova Base Nacional Comum Curricular – BNCC – e no posicionamento de suas respectivas estratégias comerciais[6]. Esta última parte deste processo foi desconstruída pelo próprio INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – responsável pelo exame. Em 2017, o Instituto publicou nota informando que não divulgaria mais o ranking de escolas[7].

Para além das questões políticas e econômicas, no contexto da discussão posta, assumem importância dois elementos que aqui interessam: em primeiro lugar, as relações entre ENEM e a grade curricular de ensino médio das escolas, de um lado, e os dados disponíveis, sobre os alunos nas escolas e sobre o ENEM, de outro. No que tange o primeiro, as informações oriundas da proposta do ENEM, especificamente habilidades e competências, orientarão modelos de currículo de Ensino Médio, uma vez que estão baseadas nos Parâmetros Curriculares Nacionais – PCNs – e aparecem como alternativa à falta de currículos oficiais em muitos estados e municípios [MOREIRA JÚNIOR, ARAÚJO, online; STADLER, HOUSSEIN, 2017, online]. Já sobre o segundo aspecto, os dados, interessa particularmente ao presente trabalho a informação de que os dados de acompanhamento de alunos, bem como os microdados do ENEM são muito volumosos, carecendo de serem minerados para que suas informações possam ser convertidas em respostas às questões de aprendizagem envolvendo alunos/as de Ensino Médio e, assim, apoiarem a tomada de decisão pedagógica.

No que se refere aos dados ENEM, alguns bons trabalhos têm aparecido nos últimos anos sobre o tema, parte deles citados nas próximas seções do presente artigo.

---

[4] A discussão sobre esses “desvios” e seu impacto técnico ultrapassa os limites do nosso artigo aqui. O apontamento de seus problemas, contudo, podem ser encontrados em Machado [apud. BARROS, 2014, p. 1073]. Disponível em: https://www.scielo.br/pdf/ensaio/v22n85/v22n85a09.pdf

[5] O dados do ENEM fazem parte de um conjunto gigantesco de informações do governo federal sobre Educação Básica e Ensino Superior. As informações são públicas e abertas, podendo ser acessadas em: http://portal.inep.gov.br/web/guest/microdados

[6] A fim de melhorarem seu posicionamento no “ranking de escolas”, divulgado pelo INEP, muitas instituições acabaram por criar turmas com seus alunos de melhores resultados, abrindo CNPJs separados para que estas “escolas” figurassem em melhores posições. O fato foi largamente exposto por veículos de comunicação entre 2010 e 2017. Cf., por exemplo: http://g1.globo.com/educacao/noticia/2015/08/metade-no-top-20-do-enem-recebe-maioria-dos-alunos-no-ano-da-prova.html

[7] A este respeito, cf.: http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/nota-de-esclarecimento-encerramento-do-enem-por-escola/21206

Ocorre, contudo, que, ao procurar por trabalhos que tratassem da mineração de dados relativos à relação entre dados ENEM e, aqueles oriundos das grades de notas do ensino médio, evidenciou-se ausência de material nas bases de pesquisa escolhidas [Revista Brasileira de Informática na Educação; Scholar Google; Scielo, Sociedade Brasileira de Computação].

Outro dado importante chamou a atenção, das publicações encontradas, nenhuma fazia referência à área de ciências humanas, sobre a qual versa a presente pesquisa.

Nesse sentido, há um vácuo importante no que diz respeito a trabalhos em Mineração de Dados que correlacionem dados pedagógicos ENEM-escola e, em especial, na área de ciências humanas. De tal maneira, há, por consequência, uma impossibilidade de se encontrarem modelos de classificação e associação fiáveis no apoio à decisão pedagógica por parte de professores e gestores. É neste vácuo que o presente trabalho procura justificar sua existência: a oferta de informação de qualidade ao professor dentro de sala e ao gestor pedagógico fora de sala no apoio à tomada de decisão é um espaço que precisa ser ocupado. Trata-se, portanto, de um estudo de natureza exploratória.

A fim de se aprofundar nas questões apontadas, o trabalho busca investigar o seguinte: é possível predizer a nota de ciências humanas no ENEM de um dado aluno ainda no curso do ensino médio, através de algoritmo de regressão, com uma confiança mínima de 60%?

Para proceder esta investigação, foi escolhida amostra de alunos oriundos do ensino médio de um tradicional colégio do Tucuruvi, bairro situado na zona norte da cidade de São Paulo, e que prestaram o ENEM. O colégio em questão orienta o currículo do ensino médio para a prova do ENEM.

Para a pesquisa, foram coletados dados pedagógicos quantitativos – notas de avaliações – disponíveis em dois bancos de dados da instituição, tendo como recorte os egressos dos anos de 2017, 2018 e 2019.

De outro lado, foram coletados dados referentes ao desempenho dos referidos alunos na área de Ciências Humanas do ENEM, entre os mesmos anos, 2017 a 2019. Tais informações foram captadas via questionário no google forms enviado individualmente a cada um.

A pesquisa objetivou, então, de maneira geral, avaliar a procedência de modelo preditivo de nota ENEM, a partir de dados pedagógicos em relação à área de ciências humanas, no contexto apresentado, como componente para a tomada de decisão pedagógica por parte de professores e gestores. Outrossim, de maneira específica, buscou encontrar regras adequadas para predição da nota no exame, com as notas dos alunos em avaliações de desempenho durante o itinerário do ensino médio. Além disso, procurou, ainda, avaliar a importância da associação de habilidades e competências com as notas dos alunos para a tomada de decisão pedagógica.

O trabalho está estruturado em quatro seções, além da introdução, a saber: 2 – fundamentação teórica, abordando a Mineração de Dados e o que há de relevante ao presente artigo; 3 - metodologia e procedimentos utilizados na pesquisa; 4 – apresentação e discussão dos resultados da pesquisa; 5 – conclusões.

Uma observação final importante sobre a escrita do trabalho. Foi escolhida uma abordagem que pudesse, de um lado, ser acessível aos pares do autor na área educacional.

Assim, há, no texto, o cuidado de fornecer certas explicações sobre conceitos e processos de programação e estatística, que podem soar desnecessárias aos colegas das áreas de computação. O mesmo, na face oposta, vale para explicações metodológicas e epistemológicas do lado pedagógico. Conta-se, portanto, com a empatia epistêmica de ambos os lados.

2. Fundamentação teórica

2.1. Mineração de Dados Educacionais - MDE

A Mineração de Dados Educacionais – MDE – pode ser considerada uma área de estudo relativamente nova nos espectros da computação e da educação [BISPO Jr., 2019, p. 1541]. Como área interdisciplinar, exige adaptações em ambas as direções. Como área emergente (Ibid.), requer, nesta via de mão dupla, a adaptação de técnicas de Mineração de Dados – MD – ao contexto educacional, de um lado [SILVA, NUNES, 2015, p 1113]; e, de outro, o desenvolvimento de recursos, no escopo pedagógico, em auxílio ao trabalho de professores e professoras [ROMERO et. al., 2008].

Conforme Bispo Jr. [op. cit.], a área de MDE surge possivelmente com o trabalho de Anjewierden et. al. (2007). Şahin e Yurdugül [2020, p. 123], contudo, apontam para um artigo pioneiro na área, em 1995. De toda forma, é sabido que a área ganhou impulso na passagem da primeira para a segunda década de 2000. Nesses anos de existência, apareceram trabalhos importantes mundo afora, com um potencial riquíssimo de contribuição para a melhoria da qualidade da educação.

Exemplo disso, o trabalho de San Pedro et. al. [2013], que conseguiu, a partir de dados oriundos da interação de crianças de sexto ano (12 anos de idade) com software educacional, predizer, com uma acurácia (taxa de acerto) de 68,7%, se estas mesmas crianças iriam ou não ingressar na universidade anos depois.

No caso do Brasil, uma primeira referência importante sobre o tema aparece em Baker et. al. [2011], que abre a discussão sobre o tema no Brasil, num contexto de expansão do Ensino a Distância – EAD – da criação da Universidade Aberta do Brasil e da massiva expansão do ensino superior. O trabalho aponta possibilidades para a pesquisa na área no Brasil.

2.2. Trabalhos relacionados

Após esta primeira publicação de 2011, sobre Ensino Médio e ENEM, respectivamente, aparecem alguns trabalhos que apresentam dados importantes utilizando técnicas de MDE para classificação de dados e predição de resultados.

Mais genericamente, sobre ensino médio, Silva e Nunes [2015], já citadas nas páginas anteriores, aplicam o algoritmo J48 a dados de alunos em uma escola particular de ensino médio, na região de Campina Grande, na Paraíba, a fim de encontrar alunos com risco de reprovação e agir precocemente. Esse algoritmo, J48 possibilita a classificação de dados e a criação de árvores de decisão, que nada mais são do que uma estrutura que, por meio de uma regra divide sucessivamente um grupo grande de registros (dados) em conjuntos menores, possibilitando a análise. No caso do trabalho relatado, as autoras conseguiram evidenciar algumas correlações importantes. Soube-se, por exemplo, que os alunos bolsistas têm índice zero de reprovação. Também, que alunos de outras cidades têm

menores taxas de reprovação do que aqueles de Campina Grande. Com esses dados, foi possível desenvolver ferramentas para atuação pedagógica específica com os grupos de alunos que têm maior propensão a reprovação. Este trabalho é um bom exemplo sobre a mineração de grandes quantidades de dados no ensino médio.

Sobre especificamente o ENEM, há alguns trabalhos que podem ser destacados com a utilização das técnicas de MDE.

Furtado [2014] propõe uma nova utilização para o algoritmo SKATER – que faz agrupamento espacial – para agrupar municípios no estado do Rio de Janeiro com notas semelhantes em Matemática no ENEM de 2011. A escolha por Matemática foi arbitrária, embora o autor deixe claro nos objetivos da pesquisa a importância de avaliar as pessoas que desejam ingressar em universidades públicas. O trabalho demonstrou que a proposta de agrupamento geoespacial do autor produz resultados de melhor qualidade em relação à abordagem tradicional.

Stearns et. al. [2017] analisam a possibilidade de prever a performance de estudantes no ENEM somente a partir das suas informações socioeconômicas. Para tal, os autores escolheram modelos de regressão baseados em árvore de decisão combinados através de técnicas de boosting – algoritmos que basicamente fazem combinações entre classificadores. A escolha, segundo apontam, se deve ao fato de que “um modelo de Árvore de decisão quando utilizado sozinho é considerado um algoritmo preditivo ‘fraco’ (weak learner)” [p. 2523]. Pela alta variância, os autores escolheram a nota de Matemática, e, em relação aos dados, utilizaram os métodos AdaBoost e Gradient Boosting, tendo sido encontrados os melhores resultados com esse último. Como resultado da pesquisa, os autores apontam que existe um viés dos dados do questionário socioeconômico do ENEM sobre sua nota, sendo possível a predição de nota com valores de métricas adequadas[8].

Simon e Cazella [2017], na mesma direção, empreendem uma análise dos chamados microdados do ENEM (citados na introdução do presente trabalho) de 2015. Os autores procuram gerar um modelo preditivo que indique o desempenho médio na área de ciências da natureza e suas tecnologias a partir de fatores socioeconômicos em todo o território nacional. A escolha da área de ciências da natureza foi feita, segundo os autores, a partir da evidência dos baixos índices de laboratórios de ciências nas escolas, a partir do Censo Escolar da Educação Básica, de 2016. Para a análise, os autores propõem a construção de um modelo preditivo utilizando técnica de árvore de decisão através do algoritmo J48. Os dados foram coletados da base do ENEM por escola, a qual inclui as instituições em que pelos 10 alunos estiveram em fase de conclusão do ensino médio regular, e, no mínimo, 50% prestaram o exame no ano de 2015. Os autores encontraram, com uma acurácia de 77,02% correlações entre dados socioeconômicos e nota. De acordo com suas conclusões, as notas mais altas são evidenciadas entre: a) escolas privadas – apenas no nível socioeconômico muito alto; b) federais – nos níveis muito alto, alto e médio alto; c) estaduais – apenas no nível muito alto; d) municipais – apenas no nível médio alto.

Essa correlação entre perfil socioeconômico e desempenho foi evidenciado, também, em um interessante trabalho feito por Leonardo Jorge Sales a pedido do jornal “O Estado de

[8] Métricas utilizadas: MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) e R^2

São Paulo" ("Estadão"), em 2018 [9] . A intenção inicial do trabalho foi produzir uma aplicação web em que, através do fornecimento de algumas informações, o candidato poderia ter uma previsão de nota. Para chegar a esta "calculadora", o autor analisou dados de 1.330.294 alunos, incluindo a nota final do ENEM 2017 e mais 199 variáveis explicativas. Destas, 168 oriundas do questionário socioeconômico, e mais 31 oriundas do Censo Escolar de 2017. O autor selecionou desse conjunto as 30 variáveis mais fortemente correlacionadas com a nota do candidato na prova. Para modelagem, o autor utilizou o algoritmo DecisionTreeRegressor, que basicamente produz um modelo de árvore de decisão, como já visto noutros trabalhos acima. Ao final da análise, relata ter encontrado uma acurácia média de 85,87% correlacionando variáveis que impactam positivamente ou negativamente a nota com a predição do possível resultado no exame. O erro médio na previsão ficou na casa de 59,85 pontos (para mais ou para menos). Das variáveis que impactam positivamente a nota, destacam-se cinco: ter estudado em uma escola privada; a renda per capita familiar; o nível de utilização de equipamentos multimídia na escola; o número de funcionários (relativo à quantidade de alunos) da escola; se a escola possui parque infantil. De outro lado, as variáveis que que impactam mais negativamente a nota são: ter estudado em escola pública estadual ou municipal; não haver computador no domicílio; não haver carro no domicílio; falta de acesso à internet no domicílio; falta de telefone fixo no domicílio.

Há, ainda, um trabalho interessante de ser destacado, explorando um viés mais intrínseco ao conteúdo do próprio ENEM. É levado a cabo por Lima et. al. (2019), que buscam uma análise de conteúdo em relação ao ENEM e ao Exame Nacional de Desempenho dos Estudantes – ENADE – este último voltado ao ensino superior brasileiro. A análise dos autores se assenta sobre uma metodologia proposta por Lima et. al. [2018], a qual primeiramente classifica as questões do exame em domínios de conhecimento (por exemplo, Português, Matemática, Química, etc) e em análises que esses domínios possibilitam. Após isso, um software separa os resultados dos estudantes por temas, de modo a produzir relatórios que oferecem um conhecimento mais apurado tanto da estrutura do teste quanto do desempenho de cada estudante em determinado domínio do conhecimento. Segundo os autores, a aplicação da metodologia possui a vantagem de se poder automatizar praticamente todo o processo, exceto o download dos microdados do ENEM. Finalmente, apontam essa metodologia permite, com os dados obtidos, aplicações em Data Mining, entre outras de estatística e análise de dados.

2.3 – Questões epistêmico-metodológicas intrínsecas

Como observado, os trabalhos acima, ainda que escolhidos por sua relevância e proximidade em relação ao tema aqui proposto, de um lado, e sua atualidade, por outro, apresentam análises distintas da que se relata aqui. A exposição deles se divide, basicamente, em três frentes: a) análise de dados de perfil socioeconômico de estudantes durante o curso do ensino médio e sua relação com o desempenho no curso; b) análise de perfil socioeconômico dos candidatos do ENEM e sua relação com o desempenho no exame (e aqui está a maioria dos trabalhos); c) análise das informações do próprio exame

[9] O texto, embora possa ser considerado um exemplo de literatura cinza, apresenta e discute dados com bastante assertividade metodológica, sendo sua inclusão no presente trabalho previamente discutida entre o autor e orientador, os quais, em comum acordo, consideraram a referida pesquisa relevante para o que é proposto aqui.

e as possibilidades de investigação que ela oferece (o último estudo relatado, o qual, diga-se, é exploratório).

Nenhum dos trabalhos relatados, contudo – e nenhum trabalho encontrado durante a presente pesquisa – trata da correlação entre as notas de alunos nas avaliações escolares e as notas obtidas por estes mesmos alunos no ENEM. Ao que parece, há, então, um vácuo neste ponto.

Essa inexistência, contudo, não é sem razão. Há diferenças importantes entre os métodos aplicados por professores(as) em avaliações escolares e aqueles empregados pelo INEP na elaboração do ENEM. Isso porque as avaliações escolares distribuem pontos de acordo com questões, enquanto o ENEM os distribui a itens (a famosa TRI – Teoria de Resposta ao Item). Explicando melhor, nas escolas, as avaliações podem ser classificadas de acordo com a função no dia a dia da sala de aula [WEBER, 2007, p. 30]. Há, assim, avaliações para acompanhamento (avaliação formativa), identificação do momento de construção de conhecimento do aluno (avaliação diagnóstica), acompanhamento mais próximo mediado por diálogo aluno-professor (avaliação mediadora), avaliação e transformação da realidade escolar (avaliação emancipadora), superação de modelos tradicionais de avaliar (avaliação dialógica), ou mesmo identificação de conceitos e informações aprendidos de forma mais tradicional (avaliação somativa). Em resumo, a avaliação, no âmbito escolar, serve ao propósito de acompanhar, verificar aprendizado, e permitir condições de desenvolvimento e transformação. No ENEM, por outro lado, a avaliação está baseada em 30 competências e 120 habilidades distribuídas entre as 4 grandes áreas do conhecimento (ciências humanas, ciências da natureza, matemática e linguagens e códigos), sustentadas por 5 eixos cognitivos comuns a todas (Dominar linguagens, compreender fenômenos, enfrentar situações-problema, construir argumentação, elaborar propostas). A avaliação é mensurada através de pesos distintos por questão, a partir de critérios que incluem, entre outros, a articulação entre habilidades e competências (mesmo entre questões de disciplinas distintas) e um sistema inteligente que “entende” quando o candidato “chuta” uma resposta [INEP, online; Id., online]. Em resumo, a avaliação, no âmbito da prova do ENEM, serve ao propósito de diagnosticar níveis de proficiência em habilidades e competências, além de classificar candidatos em relação a estes mesmos níveis.

Fora esta questão da natureza das avaliações, há, ainda, problemas em relação à coleta de dados que vale a pena relatar. Ao se compararem os dados de acompanhamento escolar com os dados do ENEM surge um primeiro problema importante: os resultados do exame são divulgados sempre no ano seguinte à sua realização. Assim, ao se compararem dados de histórico, por exemplo, de alunos concluintes de ensino médio, com sua nota ENEM, será preciso procurar por estes mesmos alunos já egressos, o que obriga o pesquisador a encontrá-los um a um, o que dificulta o processo e aumenta o tempo e a dificuldade da pesquisa. Este ponto poderia ser contornado com o fornecimento do número de inscrição no ENEM pelos alunos, o qual eles recebem no momento em que a efetuam, ainda no 3º ano do ensino médio (esta foi uma tentativa, inclusive, do presente trabalho). Ocorre, contudo, que os números de inscrição divulgados pelo INEP em seus dados públicos são uma mera máscara que nada têm que ver com os números reais – estratégia para proteger a identidade e os dados dos inscritos. Deste modo, sem poder contar com nenhuma correspondência, o trabalho de coleta para uma pesquisa como a que este trabalho relata precisa ser feito em duas etapas: uma, dentro da escola, recolhendo os dados fornecidos

pela instituição; e outra, fora da escola, buscando-se os agora ex-alunos que prestaram o exame para a coleta de informações e resultados no exame. Uma estratégia para diminuir este trabalho é conhecer, ainda na instituição, os alunos que irão prestar o exame para filtrá-los do total. No caso da pesquisa aqui apresentada, por exemplo, de 72 concluintes do ensino médio em 2019, na escola investigada, apenas 29 prestaram o ENEM.

Um último ponto, ainda, decorrente do anterior, parece relevante de ser colocado e se refere à praticidade da pesquisa. Ora, com os dados socioeconômicos em mãos, é possível, como mostraram os trabalhos correlatos, conhecer bastante a realidade dos alunos de determinada escola e, com estes dados, estabelecer estratégias de ação que possam mitigar tais problemas. Neste aspecto, poderia, inicialmente, fazer pouco sentido a elaboração de uma pesquisa mais trabalhosa, envolvendo, inclusive, dados de coleta manual para, ao final, chegar-se a resultados semelhantes aos demais.

Os problemas relatados nos parágrafos acima, um de natureza metodológica, outro relacionado aos procedimentos de pesquisa, e um último aparentado com o objeto de pesquisa, aparecem como motivadores de três dentre as questões que este trabalho levanta e, na medida das possibilidades metodológicas e práticas, procura responder: a) mesmo com esta diferença metodológica entre as notas de avaliação escolar e as notas das avaliações ENEM, é possível encontrar correlações fiáveis entre ambas?; b) os limites de ordem prática para coleta e tratamento de dados podem ser contornados? Se sim, sob quais estratégias?; c) O levantamento e a análise de dados sobre o ENEM fora da base de dados públicos e abertos, vale, em seus resultados, o esforço levado a cabo em sua coleta e processamento?

A busca de respostas a estas questões está implícita no itinerário metodológico que as seções abaixo percorrem.

3. Metodologia da pesquisa

3.1. Materiais e métodos

Como já citado, a presente pesquisa buscou um modelo preditivo que pudesse correlacionar as notas escolares de alunos do ensino médio com seu respectivo desempenho no ENEM ao final do curso. Para tal, foram escolhidos como amostra os dados de egressos do ensino médio, entre 2017 e 2019, que prestaram o ENEM, de um tradicional Colégio do Tucuruvi, bairro situado na zona norte da cidade de São Paulo.. A escolha do período se deveu à escola pesquisada ter assumido, entre 2014 e 2015, uma orientação curricular voltada para o ENEM, sendo os dados dos egressos de 2017 os primeiros a relatarem esta orientação, sendo estendida a análise até os egressos de 2019, de modo a diversificar a amostra e permitir sua ampliação. A escolha da instituição, por sua vez, se deveu ao fato de ser uma escola voltada para o ENEM, com currículo elaborado a partir de conteúdos, habilidades e competências relacionados ao exame. Além destes, foram adotados também os seguintes critérios: a) a amostra só compreenderia dados de estudantes que iniciaram e terminaram o ensino médio na instituição; b) a amostra só compreenderia dados de alunos que concluíram o ensino médio sem nenhuma reprovação; c) a amostra incluiria, preferencialmente, dados de alunos que tivessem prestado prestado o ENEM na terceira série do ensino médio e não em anos subsequentes.

Em relação às notas detalhadas de alunos do colégio, elas estão disponíveis em duas bases de dados administradas por empresas terceirizadas, contratadas pela instituição, uma com informações entre 1986 e 2017; outra, de 2018 até a atualidade. Os dados compreendem todas as notas de todas as avaliações escritas, bem como de trabalhos, tarefas de casa, dentre outras, conforme exemplo descrito na figura 3.1.

1º Trimestre											2º Trimestre									3º Trimestre												
Inst			AM				AT		Falta	Nº	Inst				Mensal	Trimest		Falta	Nº	Inst				mens	tri		Falta	Nº	Média Parcial	Média Final	Faltas	Resultado
For	vis	Total	met	Unif	Ava	Total					form	Sem	ApJLK	Total						form	vis	atv	Total									
85	87	86	0	57	57	38	70	65	-	1	85	80	83	83	79	88	83	1	1	87	100	95	94	90	70	85	2	1	81	81	3	APROVADO
78	43	61	80	92	92	88	75	75	2	2	75	65	85	75	76	80	77	5	2	83	100	65	83	99	75	86	7	2	81	81	14	APROVADO
81	87	84	65	50	50	55	80	73	2	3	80	70	83	78	45	92	72	1	3	80	100	95	92	100	77	90	2	3	81	81	5	APROVADO
85	87	86	0	68	68	45	75	69	1	4	85	75	85	82	85	94	87	3	4	88	100	95	94	100	100	98	3	4	90	90	7	APROVADO
80	57	69	0	70	70	47	75	64	-	5	85	95	83	88	75	92	85	-	5	87	100	90	92	100	90	94	1	5	86	86	1	APROVADO
85	86	86	0	87	87	58	75	73	-	6	80	75	85	80	65	90	78	1	6	83	100	90	91	100	100	97	-	6	87	87	1	APROVADO
85	100	93	80	52	52	61	70	75	-	7	80	75	83	79	81	88	83	1	7	87	100	85	91	100	45	79	-	7	80	80	1	APROVADO
88	87	88	0	70	70	47	75	70	-	8	85	75	85	82	65	85	77	-	8	87	100	95	94	96	65	85	2	8	80	80	2	APROVADO
80	71	76	75	69	69	71	70	72	-	9	80	80	83	81	56	88	75	3	9	84	100	95	93	100	55	83	2	9	79	79	5	APROVADO
85	86	86	0	40	40	27	85	66	-	10	80	85	85	83	82	95	87	1	10	85	100	67	84	100	88	91	4	10	86	86	5	APROVADO

Figura 3.1: Exemplo de notas de alunos durante um ano letivo

Fonte: O autor

As notas no ensino médio do colégio, independente da disciplina, são divididas em três conjuntos: instrumentos (que aparecem como “inst” na figura 3.1); avaliação mensal (“AM”, na figura 3.1); avaliação trimestral (“AT”, na figura 3.1). O primeiro conjunto inclui a parte de avaliação diversificada (trabalhos em grupo, notas por participação e comportamento, etc); o segundo e o terceiro, notas de avaliações escritas (únicas ou subdivididas em avaliações menores), com calendário comum a todas as turmas e horários específicos, elaborados pela coordenação pedagógica do colégio. Os valores vão de 0 a 100 em cada conjunto, e a nota (N) dos alunos ao final do trimestre é calculada por média simples entre os três conjuntos, sendo necessário um mínimo de 65 pontos para aprovação. O cálculo pode ser expresso pela seguinte fórmula:

$$=(\quad+\quad+\quad)\div 3$$

O ano escolar é dividido, por sua vez, em três trimestres, de janeiro a novembro, sendo a nota anual (NA) calculada por meio de média ponderada entre as médias de cada trimestre no decorrer do ano (N1, N2, N3), com pesos 1, 2 e 3, respectivamente. O cálculo pode ser expresso pela seguinte fórmula:

$$NA=\frac{}{}$$

Para a análise aqui empreendida, foram utilizadas as notas “AM” e “AT” de cada trimestre, em cada uma das disciplinas que compõem a área de ciências humanas: História, Geografia, Filosofia e Sociologia. A escolha de tais notas foi feita por duas razões: primeiro, por serem avaliações escritas e formadas, em boa parte, por testes – como o ENEM o é; segundo, por estas avaliações privilegiarem, embora não somente, questões ENEM de anos anteriores. O total de variáveis (notas de avaliações) encontradas a partir desta escolha pode ser visualizado na tabela 3.2, abaixo:

Tabela 3.2: número de avaliações realizadas por estudantes durante o ensino médio

Fonte: o autor

Disciplina	Ano Escolar						
	1º Ano		2º Ano		3º Ano		Subtotais
	AM	AT	AM	AT	AM	AT	
Filosofia	3	3	3	3	3	3	18
História	3	3	3	3	3	3	18
Geografia	3	3	3	3	3	3	18
Sociologia	3	3	3	3	3	3	18
						Total geral	72

A partir da relação das notas, ficou evidente um problema. Para cada aluno, as notas foram dispostas horizontalmente, num total de 72 colunas. Com o acréscimo da nota ENEM em ciências humanas, chegar-se-ia a 73 colunas. O número total de egressos nas bases de dados do colégio, incluindo os que se inscreveram no ENEM e os que não, compreendia informações de 238 alunos. Estatisticamente, esta construção não era possível[10] . Foi adotado, então, como solução, o cálculo de uma média simples entre todas as notas mensais e trimestrais de cada ano para cada disciplina. Não foi possível a adoção de média ponderada, já que, para isso, eram necessários de dados de habilidades e competências por questão em cada avaliação, bem como nas questões do ENEM. Os dados referentes a habilidades e competências nas questões das avaliações na escola, ao menos, inexistem. A solução adotada, contudo, permitiu reduzir o número de variáveis relativas a notas de 73 para 13, conforme figura 3.3.

[10] Há uma conhecida regra estatística que atende pelo nome de “1 in 10 rule”, ou “1 para 10”. A grosso modo, ela aponta que deve haver uma proporção de 1 para 10 na relação entre colunas e linhas em uma planilha de dados (para cada coluna, 10 linhas, em média). Para mais informações a respeito, cf., por exemplo: https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/206-30.pdf

Conforme exposto, ainda, na figura 3.3, foram adotados nomes específicos para as médias de notas de cada ano em cada disciplina, valor formado pelas três primeiras letras da disciplina seguida pela letra “A” e por um valor numérico entre 1 e 3. A letra “A” representa a palavra “Ano Escolar” e o valor numérico, o ano de referência. Assim, os valores para “FIL-A1”, por exemplo, se referem à média simples entre todos os valores “AM” e “AT” dos alunos amostrados durante o 1º ano do ensino médio. Tais valores são expressos através de números inteiros entre 0 e 100. Na última coluna, aparece a sigla “ENEM-CH”, que se refere à média ENEM em ciências humanas. Os valores nessa coluna são expressos através de números reais (inteiros ou decimais) e não têm intervalo definido.

A	B	C	D	E	F	G	H	I	J	K	L	M
FIL-A1	GEO - A1	HIS-A1	SOC-A1	FIL-A2	GEO-A2	HIS-A2	SOC-A2	FIL-A3	GEO-A3	HIS-A3	SOC-A3	ENEM-CH
80	87	90	93	80	84	84	93	86	81	99	92	634.8
67	53	62	70	76	65	65	80	65	61	64	84	509.6
86	78	71	89	82	75	75	88	75	72	90	96	501.4
78	66	56	75	63	61	61	76	75	64	74	87	602
75	78	75	79	76	75	75	92	80	65	75	98	589.1
76	85	79	94	84	79	79	95	72	79	88	91	605
66	77	67	75	73	79	79	85	74	75	85	94	612.1
77	80	75	79	76	77	77	84	73	77	98	97	585.5
71	68	66	86	75	74	74	75	80	68	72	87	587.2
73	91	79	90	74	73	73	88	83	80	94	93	659
81	65	68	84	76	65	65	86	80	80	92	97	608.3
81	67	63	84	78	68	68	96	81	61	89	100	578.6
82	85	78	96	86	77	77	95	83	82	96	100	619.5
80	69	53	83	65	62	62	68	61	41	39	75	530.5
81	88	87	90	80	78	78	94	83	85	95	95	646.2
79	87	90	92	76	79	79	93	76	82	100	98	645.4
71	80	70	62	73	78	78	84	76	85	99	89	655.3
77	72	77	76	72	77	77	78	72	77	97	93	630.4
82	87	81	73	82	78	78	91	80	76	97	94	620
75	78	71	82	74	67	67	83	76	68	74	94	590.9
73	64	70	66	75	65	65	72	64	68	53	78	525.5
85	86	89	88	84	87	87	99	78	84	100	94	665.8

Figura 3.3: Reprodução de tabela final para análise

Fonte: o autor

Para obtenção dos dados referentes à média ENEM de ciências humanas para cada aluno, foi elaborado um questionário via Google Forms, o qual foi enviado para todos/as aqueles/as constantes nas bases de dados do colégio com quem fosse possível fazer contato. Não houve discriminação a priori entre alunos que prestaram ou não prestaram o ENEM, dado não haver esta informação entre os dados disponíveis. No processo de localização dos egressos, houve mobilização por parte de professores e funcionários. Além disso, muitos dos concluintes possuem familiares estudando na instituição, o que também ajudou no processo. A partir disso, foram recebidas, no total, 118 respostas ao questionário no Forms.

Ocorreu, contudo, uma falha na elaboração do questionário. Ao pedir informações sobre o ENEM, o questionário dispunha somente de pergunta sobre o número de inscrição, sendo esperado, a partir disso, encontrar as demais informações no banco de microdados do ENEM. Como anteriormente dito, não há como localizar qualquer informação no banco de microdados a partir do número de inscrição que o candidato possui. Assim, o questionário precisou ser refeito e as pessoas que já haviam respondido precisaram ser

contatadas novamente. Com isso, houve perda de informação, sendo que, ao final, restou uma amostra com dados de 67 inscritos. O número é ainda considerável e valida a pesquisa, embora mais modesto que o total inicial.

A partir das médias de notas escolares e das notas de ciências humanas recebidas, foi possível formatar, finalmente, a planilha com as informações a serem mineradas por algoritmo.

Para o processamento dos dados, foi escolhido o algoritmo SimpleLinearRegression no WEKA – Waikato Environment for Knowledge Analysis, programa bastante popular que inclui, entre outras, a facilidade de não requerer conhecimentos em linguagens de programação para ser operado.. Inicialmente, o projeto da pesquisa estava orientado para o uso do classificador J48, já citado na seção 2 e bastante popular para aplicações em educação [SIMON, CAZELLA, 2017, op. cit].

A apuração do dados durante a coleta e pré-processamento, contudo, revelou que não seria possível um modelo de classificação com árvore de decisão utilizando-se este algoritmo. A diferença entre este modelo e a regressão linear, aqui efetivada, se dá em relação ao tipo de dados da variável que se deseja prever. No caso da classificação, de um lado, seriam valores nominais (por exemplo, “satisfatório”, “avançado”); já a regressão, de outro, busca prever valores numéricos inteiros ou reais (se será “580,7” ou “615”, por exemplo). Como o trabalho aqui procura prever a nota ENEM (um número real, portanto), fica claro que esteja incluso no segundo caso.

4. Apresentação e discussão dos resultados

Os dados foram processados no WEKA utilizando-se o algoritmo SimpleLinearRegression, o qual retornou resultados conforme a figura 4.1.

```
=== Classifier model (full training set) ===

Linear regression on HIS-A3

2.2 * HIS-A3 + 427.1

Predicting 0 if attribute value is missing.


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.602
Mean absolute error                     31.6801
Root mean squared error                 39.6781
Relative absolute error                 77.4918 %
Root relative squared error             79.85   %
Total Number of Instances               67
```

Figura 4.1: resumo de resultado de regressão linear

Fonte: o autor

Os dados acima podem ser, resumidamente, expressos de forma gráfica através da figura 4.2 abaixo:

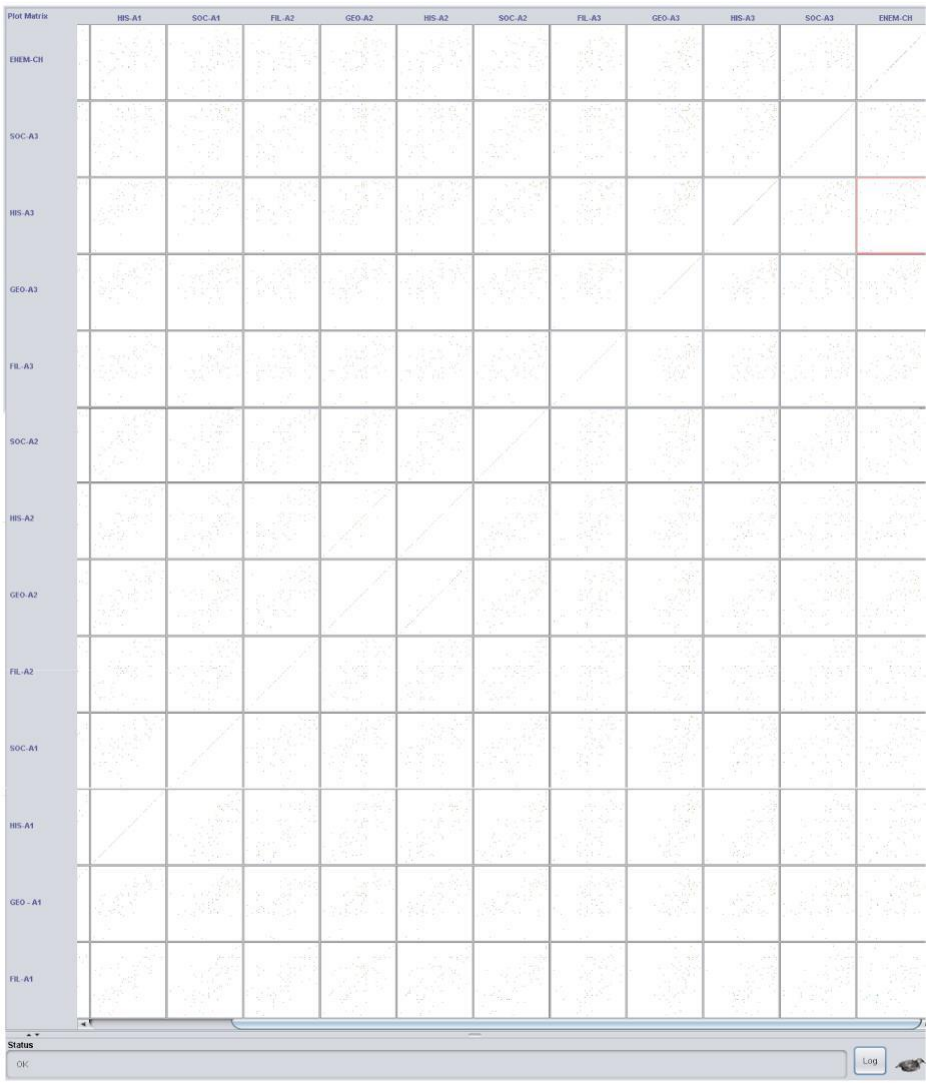

Figura 4.2: Ilustraçao gráfica de setor de regressão linear

Fonte: o autor

Os dados expressos na figura 4.1 e ilustrados na figura 4.2 apontam que foi verificada uma correspondência entre as notas escolares e o desempenho ENEM em ciências humanas. A regressão aponta o valor “HIS-A3” para predição da nota no exame. Isto quer dizer que a nota de História do 3º ano do ensino médio tem uma correlação muito forte com a nota de ciências humanas ENEM para os alunos/candidatos amostrados. Como também apontado no resultado, a acurácia (taxa de acerto da predição) esteve em 0.602, o que significa que a previsão acertou o resultado em 60,2% dos casos. A variável MAE (*Mean Absolute Error*), que mede a margem de erro, para mais ou para menos, esteve em 31.6801, ou seja, pouco mais de 31,6 em número da nota ENEM.

A observação de dados no gráfico permitiu, ainda, uma percepção importante: a partir da linha estatística resultante da regressão, é observável uma correlação forte entre os dados “HIS-A3”, “GEO-A3”, “HIS-A2, “GEO-A2” “HIS-A2” e “GEO-A1”, de modo que

coube interrogar qual seria o tamanho da correlação entre essas variáveis e a variável “ENEM-CH”. Para investigar essa questão, os dados totais da amostra pesquisada foram processados novamente no WEKA utilizando-se o algoritmo CorrelationAtributeEval, que basicamente produz um ranking relativo à força de correlação entre variáveis. O resultado de ranqueamento aparece abaixo na figura 4.3

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 13 ENEM-CH):
        Correlation Ranking Filter
Ranked attributes:
 0.602  11 HIS-A3
 0.595  10 GEO-A3
 0.574   7 HIS-A2
 0.541   6 GEO-A2
 0.507   2 GEO - A1
 0.5     3 HIS-A1
 0.486   8 SOC-A2
 0.401   9 FIL-A3
 0.393  12 SOC-A3
 0.356   4 SOC-A1
 0.299   5 FIL-A2
 0.266   1 FIL-A1

Selected attributes: 11,10,7,6,2,3,8,9,12,4,5,1 : 12
```

Figura 4.3: Ranking de correlação

Fonte: o autor

Este ranking é também bastante valioso pois informa, desde o 1º ano do ensino médio, quais conteúdos devem ser olhados com mais atenção justamente por sua forte correlação com o ENEM. Vale ressaltar “conteúdos” e não “disciplinas”, uma vez que, na grade de ciências humanas, os conteúdos são revisitados em diferentes disciplinas, reforçando e ampliando o trabalho com certas habilidades e certas competências. Por exemplo, em História, no 2º ano do ensino médio (“HIS-A2”), são discutidas, no seio da modernidade, teorias políticas liberais; estas mesmas teorias serão aprofundadas em filosofia no ano seguinte (“FIL-A3). De igual modo, em Sociologia, no 2º ano do ensino médio (“SOC-A2) são trabalhadas questões relativas à sociedade contemporânea, dita informacional; de igual modo, as especificações desta sociedade, em termos de trabalho, riqueza e outros serão aprofundadas em Geografia, no ano seguinte (“GEO-A3”). Assim, sendo, quando se indica o valor “HIS-A3” como o mais fortemente correlacionado ao desempenho em ciências humanas no ENEM, indicam-se, por consequência, os conteúdos estudados no período a que o valor se refere (“História e sociedade contemporânea”; “Brasil pós regime miliar”; etc).

5. Conclusões

Desta forma, respondem-se, inicialmente, as perguntas-problema indicadas ao final da seção 2. Em primeiro lugar, foi verificada, ainda que de forma exploratória (inicial), uma correlação entre notas escolares e a nota ENEM (no âmbito de ciências humanas, ao menos), mesmo que a metodologia para obtenção de ambas seja distinta. Em segundo, ficou claro que, mesmo com as ditas dificuldades de ordem prática, o trabalho é possível, havendo, como relatado, estratégias, e mesmo percalços, que podem servir de base para trabalhos futuros.

Por último, vale ressaltar a diferença entre dados socioeconômicos e dados pedagógicos, em resposta à terceira pergunta-problema. Os primeiros informam questões extrínsecas à sala de aula que impactam no desempenho dos estudantes, como foi observado em outros trabalhos (“risco de reprovação”; “nota ENEM alta”, por exemplo). Estes dados, importantes, sem dúvida, para a escola, possuem uma granularidade menor, no que tangem os aspectos pedagógicos. Os segundos apontam para questões intrínsecas ao ambiente de sala de aula e ao seu planejamento, esbarrando diretamente em conteúdos, estratégias de aprendizagem e seu planejamento. Neste caso, ao se responderem a quais conteúdos prestar mais atenção, obtêm-se, certamente, respostas com uma granularidade bem maior, o que salvaguarda o modelo de pesquisa aqui empreendido, mesmo com o esforço extra exigido na sua condução.

De todo modo, embora os progressos acima relatados, há evidentes limites na pesquisa que podem, e, espera-se, sejam tratados em trabalhos posteriores. O mais relevante se refere à amostra. Com amostras maiores, podem-se aproveitar mais variáveis de notas em lugar de uma média anual simples, por exemplo. Esse cenário pode melhorar sensivelmente a precisão na correlação de dados e na predição. De igual modo, o acompanhamento pedagógico da elaboração e condução de avaliações escolares, tendo-se bases de habilidades e competências relativas às questões de cada avaliação pode ser um interessante passo na ponderação de médias anuais quando for o caso de necessária simplificação.

Espera-se aqui, finalmente, que o trabalho ora apresentado seja o primeiro de muitos na direção da investigação de pontes pedagógicas entre dados de notas dentro da escola e o ENEM.

6. Referências

Anjewierden, A., Kolloffel, B., Hulshof, C. [2007] “Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes” In: International Workshop on Applying Data Mining in e-Learning, http://hal.cirad.fr/EIAH/hal-00190067

Baker, R., Isotani, S., & Carvalho, A. [2011] “Mineração de Dados Educacionais: Oportunidades para o Brasil” In: Revista Brasileira de Informática na Educação, 19(02), 03, http://dx.doi.org/10.5753/rbie.2011.19.02.03

Bispo Jr., E. [2019] “Questões Epistemológicas em Mineração de Dados Educacionais”, In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), 30(1), 1541, http://dx.doi.org/10.5753/cbie.sbie.2019.1541

Brasil, “Matriz de Referencia ENEM”, http://download.inep.gov.br/download/enem/matriz_referencia.pdf

Brasil, “Nota técnica:Teoria de Resposta ao Item”,

http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf

De Oliveira, T. S. [2016] “O ENEM: breves considerações sobre importância avaliativa e reforma educacional” In: Educação Por Escrito, 7(2), 275-285, https://doi.org/10.15448/2179-8435.2016.2.23995

Furtado, V. M. [2014] “Agrupamento de Conjunto de Instâncias: uma aplicação ao ENEM”, [Dissertação], https://www.cos.ufrj.br/uploadfile/1417018604.pdf

Laisa, J., & Nunes, I. [2015] “Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio” In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), 26(1), 1112, http://dx.doi.org/10.5753/cbie.sbie.2015.1112

Lima, P., Rosa, E., Ambrósio, A., and Oliveira, J. [2019] “Applying Content Analysis to Brazilian National Exams”, In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 8(1), 139, http://dx.doi.org/10.5753/cbie.wcbie.2019.139, November.

Moreira Júnior, R. L., Araújo, F. M. L. [2014] “O Enem Como Proposta De Restruturação [sic] Curricular Do Ensino Médio: Uma Reflexão Teórica Sobre O Currículo De História”, In: Didática e Prática de Ensino na relação com a Escola, http://www.uece.br/endipe2014/ebooks/livro1/347-%20O%20ENEM%20COMO%20PROPOSTA%20DE%20RESTRUTURA%C3%87%C3%83O%20CURRICULAR%20DO%20ENSINO%20M%C3%89DIO%20UMA%20REFLEX%C3%83O%20TE%C3%93RICA%20SOBRE%20O%20CURR%C3%8DCULO%20DE%20HIST%C3%93RIA..pdf

Romero, C., Ventura S., Espejo, P. G. and Hervás C. [2008] “ Data Mining Algorithms to Classify Students”, In:International Conference on Educational Data Mining, 1, 08, https://www.researchgate.net/publication/221570435_Data_Mining_Algorithms_to_Classify_Students, November.

Şahin, M., & Yurdugül, H. [2020] “Educational data mining and learning analytics: past, present and future”, In: Bartın University Journal of Faculty of Education, 9(1), 121-131, https://www.researchgate.net/publication/339140442_Educational_Data_Mining_and_Learning_Analytics_Past_Present_and_Future

Sales, L. J. “Diz-me Quem És e Calcularei Tua Nota no ENEM”, https://leosalesblog.wordpress.com/2018/10/31/diz-me-quem-es-e-calcularei-tua-nota-no-enem/

San Pedro, S., Baker, R., Bowers, A. J., Heffernan, N. T. [2013] “Predicting College Enrollment From Student Interaction With an Intelligent Tutoring System in Middle School”, In: Educational Data Mining (EDM) Conference, http://www.columbia.edu/~rsb2162/EDM2013_SBBH.pdf

Simon, A., & Cazella, S. [2017] “Mineração de Dados Educacionais nos Resultados do ENEM de 2015”, In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 6(1), 754, http://dx.doi.org/10.5753/cbie.wcbie.2017.754

Stadler, João Paulo, & Hussein, Fabiana Roberta Gonçalves e Silva. [2017] “O perfil das questões de ciências naturais do novo Enem: interdisciplinaridade ou contextualização?”, In: Ciência & Educação (Bauru), 23(2), 391-402, https://doi.org/10.1590/1516-731320170020007

Stearns, B., Rangel, F., Firmino, F., Rangel, F., & Oliveira, J. [2017] “Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos” In: Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC, https://sol.sbc.org.br/index.php/ctic/article/view/3244

Weber, S. S. F. [2008] “Avaliação da Aprendizagem Escolar: práticas em novas perspectivas”, [Dissertação], https://repositorio.ufsm.br/bitstream/handle/1/6785/SONIAWEBER.pdf?sequence=1&isAllowed=y.