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

Juvenal Cordeiro Filho, ICMC-USP, jcconsultoriaacademica@gmail.com (0000-0002-2250-8206)
Bruno Elias Penteado, ICMC-USP, brunopenteado@alumni.usp.br (0000-0002-8366-5512)
Ig Ibert Bittencourt, UFAL, ig.ibert@ic.ufal.br (0000-0001-5676-2280)
Seiji Isotani, ICMC-USP, sisotani@icmc.usp.br (0000-0003-1574-0784)

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 correlação (r de Pearson) de 60,2% e RMSE de 39,7 pontos na escala ENEM. 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.

The use of school grades for the ENEM's score prediction in Human Sciences

Abstract. This paper explores the hypothesis of the prediction of ENEM human sciences grade prediction based on high school test grades from students in a private school in São Paulo. We collected test grades from human sciences disciplines over all high school period from 67 students and correlated them with their performance in the ENEM exam. The analysis allowed a prediction model correlating ENEM human sciences grade and high school grades from the 3rd high school year with a correlation (Pearson's r) of 60.2%. and RMSE of 39.7 points in ENEM's scale. Besides, the model found important academic contents strongly correlated to students ENEM performance since the 1st high school year.

Keywords: ENEM, high school, school 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, 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[1], segue oferecendo uma série de dados importantes para a educação brasileira[2].

[1] 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

[2] Os dados do ENEM fazem parte de um conjunto grande 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

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[3]. Esta utilização comercial foi desconstruída pelo próprio Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) - responsável pelo exame. Em 2017, o Instituto publicou nota informando que não divulgaria mais o ranking de escolas[4].

Para além das questões políticas e econômicas, no contexto da discussão posta, importam dois elementos que aqui interessam: primeiro, 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 primeiro, as informações oriundas da proposta do ENEM, especificamente habilidades e competências, orientarão modelos de currículo de Ensino Médio. Isso considerando que tais habilidades e competências 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]. Sobre o segundo, 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. Assim, é preciso que sejam minerados para que suas informações possam ser convertidas em respostas às questões de aprendizagem envolvendo alunos/as de Ensino Médio para apoiarem a tomada de decisão pedagógica.

Sobre os dados ENEM, alguns trabalhos têm aparecido nos últimos anos com o tema, parte deles citados nas próximas seções do presente artigo. 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 escassez de material nas bases de pesquisa escolhidas. Foram levantados dados em: Revista Brasileira de Informática na Educação; Scholar Google; Scielo, Sociedade Brasileira de Computação. Os filtros escolhidos prezaram, especialmente, pela recência – trabalhos publicados na última década, exceto os que tratam da perspectiva histórica – e pela associação entre o uso de ferramentas computacionais, o ENEM e/ou o ensino médio. Outro dado importante chamou a atenção, das publicações encontradas, nenhuma fazia referência à área de ciências humanas (CH), sobre a qual versa a presente pesquisa.

Nesse sentido, há uma lacuna 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.

---

[3] 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

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

A fim de se aprofundar nas questões apontadas, o trabalho busca investigar: qual o grau de precisão das notas ENEM em CH baseado apenas no desempenho dos alunos nas disciplinas de CH durante todo seu Ensino Médio? É possível, a partir de dados prévios, estimar a nota de estudantes no ENEM em CH utilizando Regressão Linear? A estimativa prévia dos resultados no ENEM para estudantes de ensino médio pode auxiliar na tomada de decisão pedagógica por professores e gestores? [5] Para proceder esta investigação, foi escolhida amostra de alunos oriundos do ensino médio de um colégio em um bairro situado na zona norte da cidade de São Paulo, e que prestaram o ENEM.

A pesquisa objetivou, então, de maneira geral, avaliar a procedência de um modelo preditivo de nota ENEM, a partir de dados pedagógicos em relação à área de ciências humanas no contexto apresentado. 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.

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

2. Fundamentação teórica

2.1. Sobre a área e trabalhos relacionados

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].

Trabalhos recentes em diferentes níveis de ensino, especialmente na América Latina, têm demonstrado, de toda forma, a relevância desta área, seja na análise de engajamento e resultados de estudantes [QUEIROGA et. al., 2021], seja em relação a níveis de desistência [FERNANDEZ et. al., 2021], seja em revisões sistemáticas que, em última análise, trazem contribuições para o desenho de políticas públicas [CECHINEL, et. al., 2020].[6]

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 (UAB) e da massiva expansão do ensino superior. O trabalho aponta possibilidades para a pesquisa na área no Brasil, como a mineração dos volumes de dados oriundos de plataformas EAD – àquela época recentes em território nacional. Tal procedimento foi apontado pelos autores com o intuito de melhoria na qualidade educação brasileira, o que

[5] O trabalho poderia ampliar estas questões na direção de pensar a aplicação deste modelo preditivo em apoio à constituição de políticas públicas mais eficientes em relação à educação. Este propósito, contudo, ultrapassaria os limites da pesquisa aqui exposta, a qual busca apoiar o trabalho de professores e gestores, a partir dos dados oriundos de suas próprias disciplinas, no contexto das instituições em que estão inseridos.

[6] A versão original do presente artigo, apresentada como Trabalho de Conclusão, apresentada no curso de especialização em Computação Aplicada à Educação, em 2020, traz outros elementos relevantes sobre este aspecto. A esse respeito, Cf. https://www.researchgate.net/publication/354539324_Utilizacao_de_Notas_Escolares_para_Predicao_da_Nota_ENEM_em_Ciencias_Humanas/stats.

inclui: identificação de perfis de alunos, identificação de alunos com alto risco de evasão e/ou com dificuldades de aprendizagem.

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] 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 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, 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.

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* – basicamente, algoritmos que fazem combinações entre classificadores. A escolha, 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, 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[7].

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, pela evidência dos baixos índices de laboratórios de ciências nas escolas, a partir do Censo Escolar da Educação Básica de 2016. Para a análise, propõem a construção de um modelo preditivo utilizando 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;

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

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 trabalho feito por Leonardo Jorge Sales a pedido do jornal "O Estado de São Paulo" ("Estadão"), em 2018[8]. A intenção inicial do trabalho foi produzir uma aplicação web em que, pelo 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, utilizou o algoritmo DecisionTreeRegressor, que também produz um modelo de árvore de decisão. 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 relevante a 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 – voltado ao ensino superior brasileiro. A análise 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, etc) e em análises que esses domínios possibilitam. Após isso, um software separa os resultados dos estudantes por temas, produzindo 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 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.

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

[8] 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 orientadores, os quais, em comum acordo, consideraram a referida pesquisa relevante para o que é proposto aqui.

Nenhum dos trabalhos relatados, contudo – e nenhum trabalho encontrado durante a presente pesquisa – trata da relação entre as notas de alunos nas avaliações escolares e seu desempenho no ENEM. Tal 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 escritas, normalmente (e no caso da amostra utilizada nesse trabalho), distribuem pontos de acordo com questões (por média simples ou ponderada), enquanto o ENEM os distribui a itens (por meio da TRI – Teoria de Resposta ao Item). [WEBER, 2007, p. 30; INEP, online].

Além disso, há, ainda, problemas na coleta de dados: os resultados do exame são divulgados sempre no ano seguinte à sua realização. com alunos já egressos, o que obriga o pesquisador a encontrá-los um a um, fora da escola. Isso 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 (tentativa, inclusive, do presente trabalho). Porém, os números de inscrição divulgados pelo INEP em seus dados públicos são uma mera máscara que nada têm que ver com os números reais (para proteger a identidade e os dados dos inscritos). Assim, a coleta precisa ser feito em duas etapas: uma, dentro da escola, nos bancos de dados da instituição; e outra, fora da escola, procurando os agora ex-alunos que prestaram o ENEM para a coleta de seus 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. Na pesquisa aqui apresentada, por exemplo, de 72 concluintes de 2019, na escola investigada, apenas 29 prestaram o ENEM.

Um último ponto 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 uma pesquisa mais trabalhosa, envolvendo, inclusive, dados de coleta manual para, ao final, chegar-se a resultados semelhantes aos demais.

Os problemas relatados acima, um de natureza metodológica, outro relacionado aos procedimentos de pesquisa, e um último aparentado com o objeto de pesquisa, aparecem como motivadores de 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? Estas questões importam não apenas em viabilidade prática, mas em termos de metodologia de coleta e processamento de dados. 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 coletados como amostra os dados de 67 dos 238 egressos do ensino médio, entre 2017 e 2019, de um colégio do Tucuruvi, bairro situado na zona norte da cidade de São Paulo. A quantidade expressa o número de egressos localizados que responderam o formulário de pesquisa e cujos dados contemplavam os critérios abaixo relatados.

A escolha da instituição e do período se deveu à escola pesquisada ter assumido, entre 2014 e 2015, uma orientação curricular voltada para o ENEM (currículo baseado

em temas ENEM, habilidades e competências), sendo os dados dos egressos de 2017 os primeiros a relatarem esta orientação, com extensão até 2019 para diversificação. Além destes, foi adotado, também, que a amostra: a) compreenderia dados de estudantes que iniciaram e terminaram o ensino médio na instituição; b) só compreenderia dados de alunos que concluíram o ensino médio sem reprovação; c) incluiria, preferencialmente, dados de alunos que tivessem prestado o ENEM na terceira série do ensino médio e não em anos subsequentes.

Sobre as notas detalhadas de alunos do colégio, elas estão disponíveis em duas bases de dados administradas por empresas terceirizadas, por período: 1986-2017; 2018-presente. Os dados compreendem todas as notas de todas as avaliações escritas, bem como de trabalhos, tarefas de casa, dentre outras. No ensino médio do colégio, as avaliações são divididas em três conjuntos: instrumentos ("inst"); avaliação mensal ("AM"); avaliação trimestral ("AT"). "Inst" inclui a parte de avaliação diversificada (trabalhos em grupo, notas por participação e comportamento, etc); "AM" e "AT", 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 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.

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. A escolha de tais notas foi feita por serem avaliações escritas e formadas, em boa parte, por testes – como o ENEM; e por estas avaliações privilegiarem, embora não somente, questões ENEM de anos anteriores. Há assim, 72 variáveis: 2 avaliações ("AM" e "AT) x 3 trimestres x 3 anos x 4 disciplinas. Para cada aluno, então, as notas foram dispostas horizontalmente, em 72 colunas. Acrescida a nota ENEM em ciências humanas, chegar-se-ia a 73 colunas. O número total de egressos nas bases de dados do colégio, inscritos ou não no ENEM era de 238. Estatisticamente, esta construção não era recomendada[9]. Adotou-se, então, uma média simples entre as "AM" e "AT" de cada disciplina anualmente. A média ponderada por habilidade/competência, como no ENEM, não foi possível, por inexistirem tais dados na instituição. Foram adotados nomes específicos para as médias de notas de cada ano em cada disciplina ("FIL-A1", por exemplo, se referem à média simples entre todos os valores "AM" e "AT" durante o 1º ano do ensino médio em Filosofia), numa escala de 0 a 100. Na última coluna, aparece a sigla "ENEM-CH", que se refere ao escore do ENEM em ciências humanas. Os valores nessa coluna são expressos através de números reais e não têm intervalo definido.

Para obtenção dos dados da média ENEM em CH para cada aluno, foi elaborado um questionário via Google Forms, enviado para todos/as aqueles/as egressos/as quem fosse possível fazer contato, para posterior seleção dos que tivessem prestado o ENEM e coleta de suas médias. Estes dados foram processados pelo algoritmo *SimpleLinearRegression*, do software WEKA (*Waikato Environment for Knowledge*

[9] 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

Analysis). Este algoritmo busca estimar os parâmetros para um hiperplano que minimize a distância dos pontos de dados coletados. Com isso, espera-se haver uma generalização para novos pontos de dados, baseados nesses parâmetros.

4. Apresentação e discussão dos resultados

Os dados foram processados no WEKA, com 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

Os dados da Figura 4.1 demonstram uma correspondência moderada 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, o coeficiente de correlação de Pearson ficou em 0,602. A métrica 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.

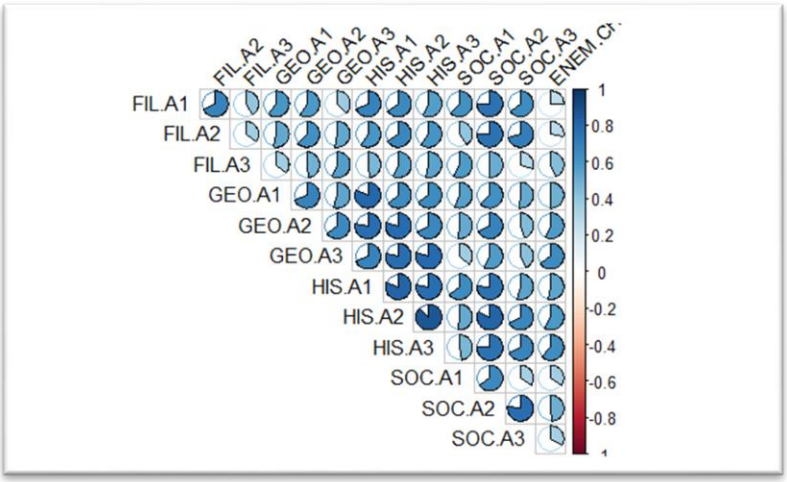
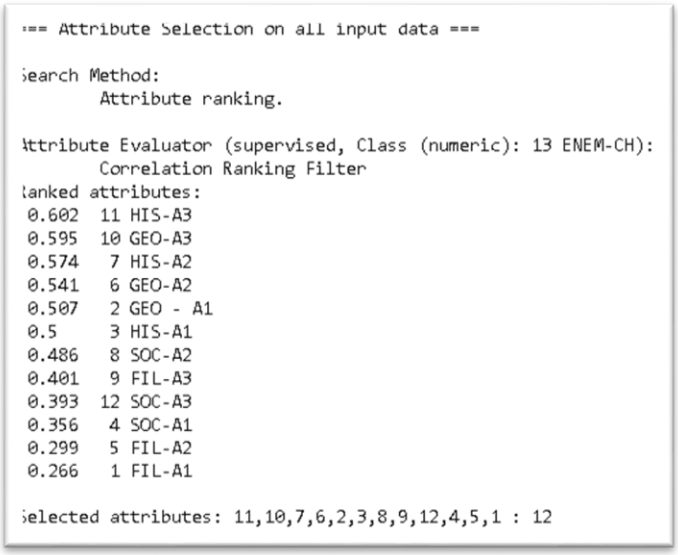


Figura 4.2: Correlação entre as notas das disciplinas a cada ano.

Complementarmente, a observação dos dados na Figura 4.2 permitiu uma percepção importante: é notável 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 nas Figuras 4.2 e 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

Este ranking é também bastante relevante 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. 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 militar"; "Marxismo e Liberalismo" etc).

5. Conclusões

Desta forma, respondem-se, inicialmente, as perguntas-problema indicadas ao final da seção 2. Em primeiro lugar, contribuição prática desta pesquisa, foi verificada, ainda que de forma exploratória, uma correlação entre notas escolares e a nota ENEM (no âmbito de ciências humanas, ao menos), mesmo que a metodologia para aferição 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. 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. Esta é, indubitavelmente, uma contribuição teórica do presente trabalho.

De todo modo, embora os progressos acima relatados, há limites na pesquisa que podem, e, espera-se, sejam tratados em trabalhos posteriores. O mais relevante se refere à amostra. Amostras maiores e de outros contextos escolares, permitem aproveitar mais variáveis de notas em lugar de uma média anual simples, por exemplo. Isso pode melhorar a precisão na correlação de dados e na predição. Também, o acompanhamento pedagógico da elaboração e condução de avaliações escolares, com base em habilidades e competências relativas às questões de cada avaliação pode ser um interessante passo na ponderação de médias 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. 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. Disponível em: <http://hal.cirad.fr/EIAH/hal-00190067>.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19(02), 03, 2011. Disponível em <http://dx.doi.org/10.5753/rbie.2011.19.02.03>. Acesso em 29/08/2021.

BISPO JR., Esdras. Questões Epistemológicas em Mineração de Dados Educacionais. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - SBIE, [S.l.], nov. 2019. P. 1541 Disponível em: <https://br-ie.org/pub/index.php/sbie/article/view/8887>. Acesso em: 29/08/2021.

BRASIL. Matriz de Referência ENEM. Disponível em: <http://download.inep.gov.br/download/enem/matriz_referencia.pdf>. Acesso em 29/08/2021

BRASIL. Nota técnica: Teoria de Resposta ao Item. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf>. Acesso em 20/08/2021.

CECHINEL, C.; et. al. Mapping Learning Analytics initiatives in Latin America. British Journal of Educational Technology, 51(4), Julho, 2020. Disponível em: < https://doi.org/10.1111/bjet.12941>. Acesso em: 12/12/2021.

FERNANDEZ, J. P. S.; et. al. Curricular Analytics to Characterize Educational Trajectories in High-Failure Rate Courses That Lead to Late Dropout. Applied Sciences, 2021, 11(4), 1436. Disponível em: <https://doi.org/10.3390/app11041436>. Acesso em: 11/12/2021.

FURTADO, V. M. Agrupamento de Conjunto de Instâncias: uma aplicação ao ENEM. Rio de Janeiro: UFRJ, 2014. Dissertação de mestrado. 95p.

LAISA, Jéssica; NUNES, Isabel. Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - SBIE), [S.l.], out. 2015. p. 1112. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/5430>. Acesso em: 29/08/2021.

OLIVEIRA, T. S. O ENEM: breves considerações sobre importância avaliativa e reforma educacional. In: Educação Por Escrito, [S.l.]7(2), 2015. p. 275-285. Disponível em: < https://doi.org/10.15448/2179-8435.2016.2.23995>. Acesso em 29/08/2021.

QUEIROGA E. M.; et. al. Using Virtual Learning Environment Data for the Development of Institutional Educational Policies. Applied Sciences. 2021; 11(15):6811. Disponível em: <https://doi.org/10.3390/app11156811>. Acesso em: 11/12/2021.

ROMERO, C.; VENTURA S.; ESPEJO, P. G.; HERVÁS C. Data Mining Algorithms to Classify Students. In: IST INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, [S.l.], 1, 2008. p. 08-17.

ŞAHIN, M.; YURDUGÜL, H. Educational data mining and learning analytics: past, present and future. Bartın University Journal of Faculty of Education, Bartin, 2020, 9(1), 121-131.

SALES, L. J. Diz-me Quem És e Calcularei Tua Nota no ENEM. Disponível em: <https://leosalesblog.wordpress.com/2018/10/31/diz-me-quem-es-e-calcularei-tua-nota-no-enem/>. Acesso em 29/08/2021.

SAN PEDRO, S., BAKER, R., BOWERS, A. J., HEFFERNAN, N. T. Predicting College Enrollment From Student Interaction With an Intelligent Tutoring System in Middle School. In: EDUCATIONAL DATA MINING CONFERENCE, 2013. Disponível em: <http://www.columbia.edu/~rsb2162/EDM2013_SBBH.pdf>. Acesso em 29/08/2021.

SIMON, Augusto; CAZELLA, Sílvio. Mineração de Dados Educacionais nos Resultados do ENEM de 2015. ANAIS DOS WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, [S.l.], out. 2017. p. 754. Disponível em: <http://www.br-ie.org/pub/index.php/wcbie/article/view/7461>. Acesso em: 29/08/2021.

STADLER, J. P.; HUSSEIN, F. B. G. S. O perfil das questões de ciências naturais do novo Enem: interdisciplinaridade ou contextualização? Ciência & Educação, Bauru, 23(2). p. 391-402. Acesso em 29/08/2021. Disponível em: <https://www.scielo.br/j/ciedu/a/yX7KS7nc5s4THFs3fXW8cJk/?lang=pt#>.

STEARNS, Bernardo; RANGEL, Flavio; FIRMINO, Fabrício; RANGEL, Fabio;

OLIVEIRA, Jonice. Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos. *In*: CONCURSO DE TRABALHOS DE INICIAÇÃO CIENTÍFICA DA SBC (CTIC-SBC), 36, 2017, São Paulo. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2017.

WEBER, S. S. F. Avaliação da Aprendizagem Escolar: práticas em novas perspectivas. Santa Maria: UFSM, 2007. 185p. Dissertação de mestrado.