IV Congresso sobre Tecnologias na Educação (Ctrl+E 2019)
Recife, Pernambuco – Brasil
28 a 30 de agosto de 2019

Mineração dos Perfis Acadêmico e Socioeconômico de Estudantes do 3º Ano do Ensino Médio da Rede Pública de Pernambuco, com Base nos Questionários da Avaliação Nacional da Educação Básica (ANEB)

Hugo Augusto Vasconcelos Medeiros[1], Katarina Tatiana Marques Santiago[1]

[1]Instituto de Gestão Pública de Pernambuco
50040-090, 1377 – Recife – PE – Brasil

hugo.vasconcelos@seplag.pe.gov.br, katarina.santiago@seplag.pe.gov.br

Abstract. *This research aims to understand the school profiles of students and the relationship between these profiles and the achievement of the results of public education policy in the State of Pernambuco. For this purpose, machine learning techniques were used to quantify the responses of the contextual questionnaire, which is applied to the students along with the mathematics and Portuguese language tests by the National Assessment of Basic Education (ANEB), more impacting on the results of these tests; and then to estimate the relationships between the student profiles and these results from the observation of the variations of the test scores according to the variation in a previously selected variable.*

Resumo. *Esta pesquisa objetiva entender os perfis escolares dos estudantes e a relação entre esses perfis e o alcance dos resultados da política pública de educação no Estado de Pernambuco. Para isso foram utilizadas técnicas de aprendizagem de máquina para minerar quais respostas do questionário contextual, que é aplicado aos estudantes juntamente às provas de matemática e de língua portuguesa pela Avaliação Nacional da Educação Básica (ANEB), mais impactaram nos resultados dessas provas; e, posteriormente, estimar quais as relações entre os perfis estudantis e esses resultados a partir da observação das variações das nas notas dos testes de acordo com a variação em determinada variável selecionada anteriormente.*

1. Introdução

Dentro do Mapa da Estratégia do Governo do Estado de Pernambuco, um dos objetivos estratégicos é o Pacto pela Educação (PPE). Essa iniciativa pública visa a "elevar o nível de escolaridade, a qualidade da educação pública e promover ações de incentivo a cultura" (PERNAMBUCO, s/d). Neste sentido, o foco da política pública é o Ensino Médio, em virtude de esse nível ser uma responsabilidade direta dos Estados (BRASIL, 1996); e utilizam-se como medidas de alcance do objetivo o Índice de Desenvolvimento da Educação Básica de Pernambuco (IDEPE) e o Índice de Desenvolvimento da Educação Básica (IDEB).

Os dois índices têm a mesma fórmula de medição, a mesma população-alvo e utilizam a mesma metodologia (teoria de resposta ao item). Contudo, diferenças como população de referência (o IDEB exclui escolas técnicas do cálculo), tipologia (até 2015, o IDEB era censitário), periodicidade (o IDEB é bianual), além de serem distintas as matrizes de referência das habilidades testadas na avaliação, fazem com que seja interessante para um Estado o acompanhamento dos dois índices. Assim, por exemplo, com o IDEPE o Estado poder comparar-se interna e anualmente; e, com o IDEB, nacionalmente, percebendo onde se coloca em relação a outras unidades federativas.

Como dito, os dois índices utilizam a mesma fórmula, qual seja a média das proficiências em língua portuguesa e em matemática, multiplicada pelo fluxo escolar. O fluxo escolar, por sua vez, é a média harmônica das aprovações em cada ano de uma dada etapa da Educação Básica. Com esta fórmula, pretende-se observar quanto tempo determinada escola ou sistema escolar levou para construir um dado nível de proficiência com seus alunos.

Dentro do escopo do monitoramento e da avaliação da política pública estadual de educação, uma das questões que se coloca é o entendimento dos perfis escolares dos estudantes e a relação entre esses perfis e o alcance dos resultados, tendo em vista servir como *feedback* e insumo para reorientar o sistema educacional, de forma geral, e a prática pedagógica, de forma mais específica.

Contudo, em virtude da complexidade do sistema educacional (JACOBSON, 2018), se colocam algumas dificuldades para identificação de padrões de aprendizagem e *feedback* corretivo para a política pública:

- *Gama de Prazos* (idem, 337): a disponibilidade de dados para traçar o perfil dos estudantes é difusa, pois ao longo do ano, numa periodicidade mensal, por exemplo, estão quase sempre indisponíveis, uma vez que os sistemas educacionais tendem a não coletá-los; porém, anual ou bianualmente, tendem a tornarem-se disponíveis todos de uma vez só, a partir das respostas dos questionários contextuais das avaliações externas – questionários esses que, usualmente, possuem um número elevado de respostas qualitativas, dificultando a análise, por questões como seleção de variáveis e a estimativa das relações e impactos delas sobre o resultado;
- *Dinâmica Ambiental Complexa*, fazendo com que os estudantes "aprendem em uma variedade de contextos: em ambientes de aprendizagem formais, com professores em escolas e universidades, e em ambientes de aprendizagem informais, como museus de ciência, mídia em massa, publicações impressas e, cada vez mais, fontes on-line mediadas pela internet" (idibem).

Assim, para considerar essas duas dificuldades de análise e manter o sentido do objetivo (identificar padrões que permitam melhorar a política pública), este artigo objetiva contribuir com o estudo acerca do perfil escolar dos estudantes, utilizando técnicas de aprendizagem de máquina para minerar quais respostas mais impactaram nos resultados dos estudantes nas provas de matemática e de língua portuguesa da Avaliação Nacional da Educação Básica (ANEB) 2017; e, posteriormente, estimar quais as relações entre os perfis estudantis e os resultados, observando as variações nas notas dos testes de acordo com a variação em determinada variável selecionada anteriormente.

2. Aprendizagem de Máquina e os questionários da ANEB.

Nesta seção, primeiro são descritos os dados do questionário do ANEB 2017, e, em seguida, apresenta-se uma discussão sobre a aprendizagem de máquina, e a escolha do algoritmo de florestas aleatórias, tomando por base as características dos dados disponíveis.

Para Penteado, Bittencourt e Isotani (2019), cumpre destacar que esse tipo de dado é um dataset educacional aberto de nível micro, posto que focado no estudante, o qual permite, dentre outras ações, avaliar iniciativas públicas.

De acordo com o INEP (2018), a ANEB 2017 teve dois instrumentos: os testes de língua portuguesa e matemática, e os questionários contextuais, respondidos por alunos, professores e diretores (esses respondem sobre si e sobre a escola). No caso do questionário dos alunos, os jovens respondem sobre "ambiente e nível socioeconômico familiar, hábitos de estudo e de leitura, motivação, trajetória escolar, entre outros aspectos" (idem, p. 8).

Ao todo, o questionário dos alunos possui 60 questões, que vão do sexo, raça, idade, até perguntas sobre o dever de casa das disciplinas e hábitos de estudo, passando por questões de perfil socioeconômico, a exemplo de "Na sua casa tem televisão a cores?", ou " Na sua casa tem banheiro?". Logo, as questões permitem perceber com maior acuidade a dinâmica ambiental complexa à qual os estudantes estão submetidos durante a construção de sua aprendizagem.

As respostas às perguntas são binárias ou fatoriais. Mesmo quando poderia ser apresentada uma contagem (como na pergunta de banheiros), existe um valor limítrofe, do tipo "Sim, quatro ou mais". Assim, temos uma base de dados cuja variável de interesse – a proficiência dos estudantes nos testes – é contínua, e as variáveis independentes são todas categóricas. Em adição, o número de casos, 43.167 no total (após a remoção dos casos sem resposta), torna difícil a mineração do perfil dos estudantes por métodos tradicionais, sobretudo ao nível do indivíduo.

Neste sentido, entendemos mineração de dados de forma semelhante à de Rigo e Cazella (2014), para quem a mineração pode ser definida como a extração de conhecimento implícito através de métodos específicos, tendo em vista identificar "padrões válidos, novos, potencialmente úteis e compreensíveis" (idem, p. 136).

Em virtude de oferecerem soluções para resolver problemas como esse, na última década, os algoritmos de Aprendizagem de Máquina (AM) voltaram a ganhar espaço, sobretudo, em virtude de sua capacidade de lidarem com muitos dados; de lidarem com dados de tipos diversos, inclusive categóricos; e por possuírem menos exigências quanto à distribuição dos dados – em comparação com os métodos estatístico/computacionais tradicionais de avaliação, como as regressões, por exemplo.

Além disso, a utilização desses modelos tem-se mostrado eficiente e eficaz, em virtude de as máquinas apresentarem, usualmente, melhor desempenho do que os operadores humanos (no tocante a tempo e a consumo de recursos) em tarefas como seleção e classificação.

Neste sentido, Rand (2015, p. 52) resume bem as possibilidades de AM: "aprendizagem automática permite a pesquisadores de sistemas complexos inferir

modelos em nível do indivíduo a partir de grandes conjuntos de dados, que podem ser usados para avaliar como uma nova política vai afetar as decisões destes indivíduos".

Mas o que é AM? Em resumo, pode-se descrever a aprendizagem de máquina como um campo multidisciplinar de estudo que se debruça sobre como as máquinas resolvem tarefas de aprendizagem, sendo capazes de melhorar seu desempenho conforme treinam e ganham experiência na resolução do problema proposto (MITCHELL, 1997).

As máquinas podem aprender de duas formas, de acordo com a disponibilidade de informação a respeito da classificação dos dados, ou seja, a existência ou não de uma classe. Quadros de dados que possuem uma variável classe permitem à máquina uma aprendizagem supervisionada, uma vez que ela é capaz de identificar se acertou ou não; enquanto quadros de dados sem classe permitem aprendizagem não supervisionada, uma vez que a máquina consegue apenas agrupar instâncias (casos) de acordo com os atributos.

Para que a máquina possa resolver um problema na perspectiva de AM, é necessário estruturá-lo com uma tarefa, um indicador de desempenho e uma fonte de experiência de aprendizagem. Para este artigo, a tarefa é estruturar o perfil dos estudantes do 3º Ano do Ensino Médio da Rede Estadual, percebendo quais variáveis são mais importantes na construção do resultado em matemática e em língua portuguesa; o indicador de desempenho é o nível de erro do modelo, ou seja, o quão distantes estão os resultados reais dos testes e as predições elaboradas pela máquina; e a fonte de experiência são os dados dos questionários contextuais da ANEB 2017, divididos em atributos, que são as variáveis explicativas, ou melhor, as respostas das questões contextuais, com as características do fenômeno sobre o qual se deseja aprender, e classe, a variável de interesse, sobre a qual se deseja aprender, que este caso são os resultados dos testes de proficiência.

Além da tarefa, do desempenho e da experiência, é necessário informar à máquina uma função de aprendizagem, que permite interpretar os dados e aprender a resolver a tarefa. Neste artigo, é utiliza a técnica de Florestas Aleatórias: uma função de aprendizagem supervisionada que toma por base a construção aleatória de árvores de decisão, para aumentar a precisão da estimativa das variáveis importantes e a capacidade de predição do modelo.

Essa função de aprendizagem usa a ideia de árvore da teoria dos grafos. Os grafos são um método de representação, gráfica ou escrita, de uma situação-problema através de vértices (pontos ou nós) e arestas (caminhos) (SCHEINERMAN, 2003). As árvores, por sua vez, são grafos conexos, ou seja, todos os pontos possuem conexões com outros, e não cíclicos, uma vez que não há caminho com retorno ao ponto inicial (idem).

As árvores de decisão, por sua vez, são árvores cuja decisão de divisão ou não se dá com base em testes de atributo, e que, para o caso em questão, são relevantes, pois, ao estruturar de maneira simbólica um problema, ganham informações sobre ele. De maneira geral, a decisão de dividir ou não um nó – ou seja, de criar ou não arestas – envolve o nível de ganho de informação que aquela divisão trará sobre o problema. Logo, as árvores começam com um nó inicial, chamado raiz, no qual estão contidos todos os dados disponíveis. A raiz pode ser então dividida em ramos (arestas) de acordo com um teste de atributo, relacionado ao ganho de informação proveniente da partição.

No fim, as arestas recebem novos nós, que trazem os dados divididos de acordo com as regras aprendidas sobre os atributos. Esse processo é repetido até que sejam formadas as folhas, que são nós terminais, nós de classificação, nos quais não há mais testes de atributo, uma vez que não se pode mais conseguir ganhos de informação, ou o ganho de informação é feito a um custo muito alto, como criar folhas com um total muito baixo de casos, ou tornar a árvore muito complexa (MITCHELL, 1997).

Por exemplo, considere a famosa árvore (idem) que representa o caminho da decisão de jogar ou não uma partida de tênis, começando pelas características climáticas (variável mais importante) e passando pelas características de umidade – caso o tempo esteja ensolarado – e de vento – caso o tempo esteja chuvoso. Assim, pode-se prever se haverá ou não uma partida perguntando-se "como está o tempo?" Se "nublado", haverá jogo. Se "ensolarado", faz-se a pergunta "Como está a umidade?" Se alta, tende a não haver jogo; se normal, tendência de haver o jogo. Caso a resposta à primeira pergunta seja "chuvoso", pergunta-se sobre o vento. Caso esteja forte, tendência é não haver jogo; e caso esteja fraco, tendência é haver.

Percebe-se que, embora a função de aprendizagem possa ser complexa, por considerar tanto o ganho de informação quanto o custo do ganho, a representação gráfica do problema é relativamente simples, facilitando a avaliação e a tomada de decisão.

Além disso, segundo Mitchell (1997), esta função de aprendizagem é capaz de se acomodar a uma classificação categórica, sem a necessidade de transformar em binárias as variáveis de interesse, bem como permite a utilização de atributos categóricos, e é robusto o bastante para produzir classificações acuradas mesmo quando há algum nível de erro nos dados.

Finalmente, como dito anteriormente, não será utilizada a técnica originária de árvores de decisão, e, sim, a técnica de florestas aleatórias. Observa-se que as florestas aleatórias são uma técnica de aprendizagem conjunta, ou seja, uma forma de aprendizagem em que a máquina é submetida a vários modelos e aprende considerando esses modelos de maneira conjunta.

Em resumo, esse processo de aprendizagem conjunta pode ocorrer de duas formas: incremental (*boosting*) ou paralela (*bagging*). No primeiro, os modelos são acrescentados progressivamente, fazendo com que cada geração seja mais precisa que a anterior; no segundo, os modelos não são acrescentados, mas, sim, elaborados paralelamente, com amostras diferentes dos dados, fazendo com que a máquina aprenda através da "votação" da classificação ente os modelos (ou seja, a moda) ou da média, em casos de regressão.

De acordo com Zhou (2012), a força dos métodos paralelos vem da independência das aprendizagens, o que tende a reduzir a variância e o viés da aprendizagem, moldando-se melhor a aprendizagens instáveis; ou seja, quando as variáveis são não lineares, como as variáveis qualitativas do questionário ANEB, e/ou quando a amostragem tende a alterar bastante os resultados – o que também é de se esperar, no caso específico, em virtude de estarmos trabalhando com respostas individuais de 43.167 estudantes a 60 questões diferentes contextuais diferentes.

As florestas aleatórias enquadram-se, decerto, em modelos paralelos, uma vez que a aprendizagem se dá por meio da criação de árvores de decisão que selecionam

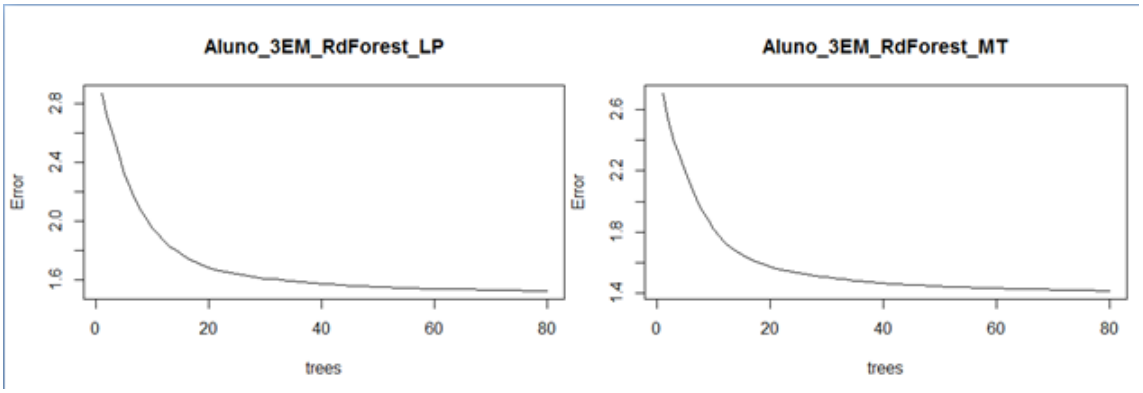
amostras probabilísticas dos atributos a cada geração (BREIMAN, 2001). Assim, ao criar uma floresta com novos atributos a cada rodada, a técnica permite que cada árvore desenvolva uma aprendizagem especializada nos atributos que foram selecionados aleatoriamente.

Em relação à implantação da solução, foi utilizada linguagem de programação R, em seu ambiente de desenvolvimento integrado R Studio, com o uso do pacote *randomForest*, cujo algoritmo tem por base o modelo de florestas aleatórias sugerido por Leo Breiman (idem).

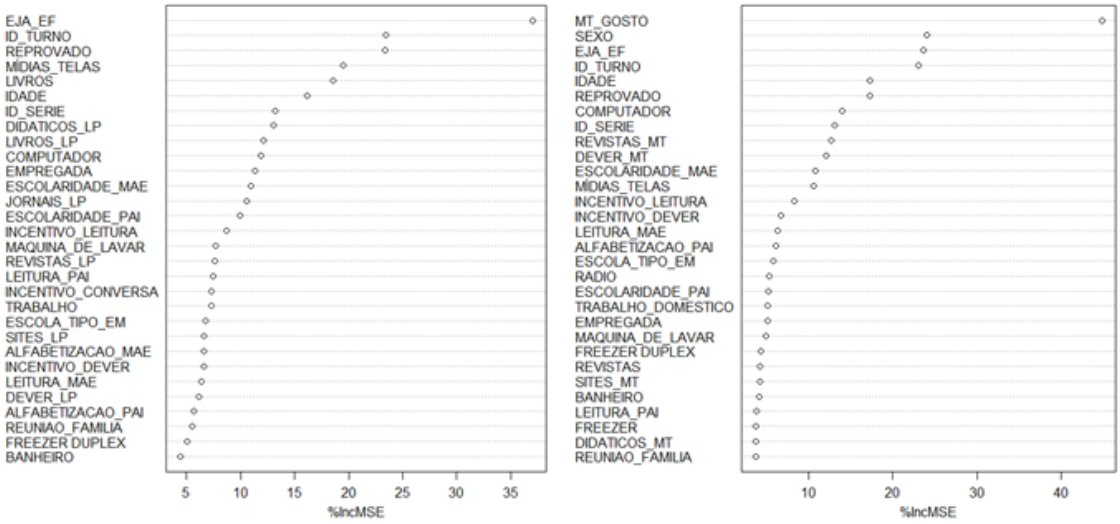
Como dito anteriormente, foram utilizados apenas casos em que todas as respostas do questionário estavam disponíveis. Além disso, foram usadas as proficiências padronizadas nos testes de língua portuguesa e matemática. Finalmente, os dados foram divididos de forma aleatória em treino (70%) e teste (30%) tendo em vista diminuir a possibilidade de superajuste na aprendizagem de máquina.

3. Mineração dos perfis dos estudantes.

Na mineração do perfil, foram criadas 80 árvores para língua portuguesa e 80 árvores para matemática; e, em ambos os casos, foram selecionadas aleatoriamente 17 variáveis em cada árvore. Para língua portuguesa, a média dos resíduos quadráticos foi 1.52, e a porcentagem explicada da variância foi de 25.52%; para matemática, respectivamente, 1.41 e 26.83%. Pelos gráficos observou-se que a saturação da aprendizagem deu-se por volta de 20 árvores, e estacionou com 40, não havendo melhorias marginais relevantes daí em diante, conforme mostram os gráficos abaixo:



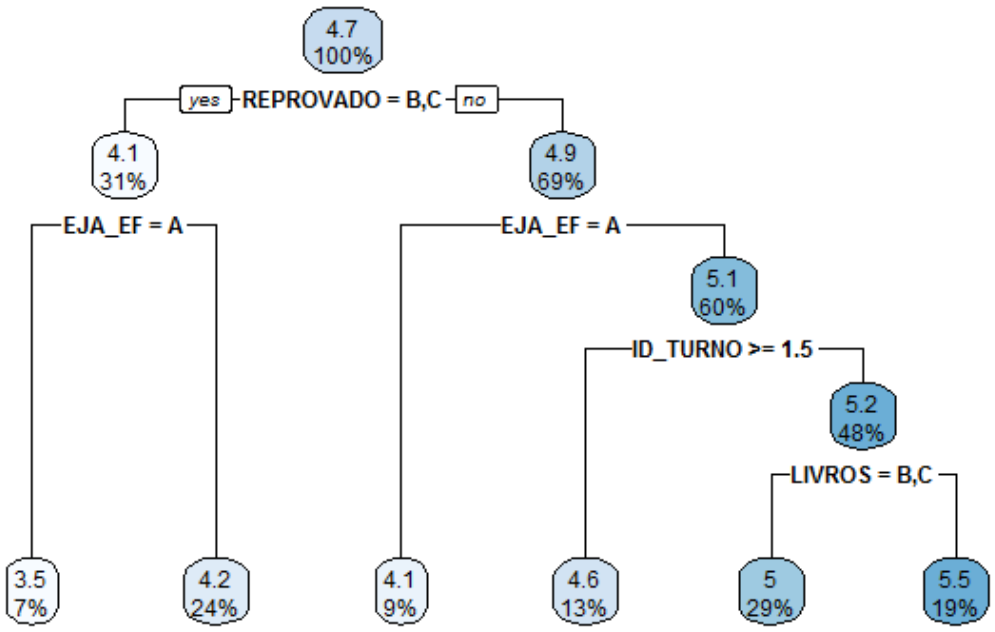
Em relação às questões que mais intensamente se relacionam com os resultados nos testes de língua portuguesa e matemática, foi utilizada a estatística %IncMSE, a qual indica qual a variação no Erro Quadrático Médio quando a variável é retirada do modelo – o que é uma forma de identificar quais as variáveis mais importantes.



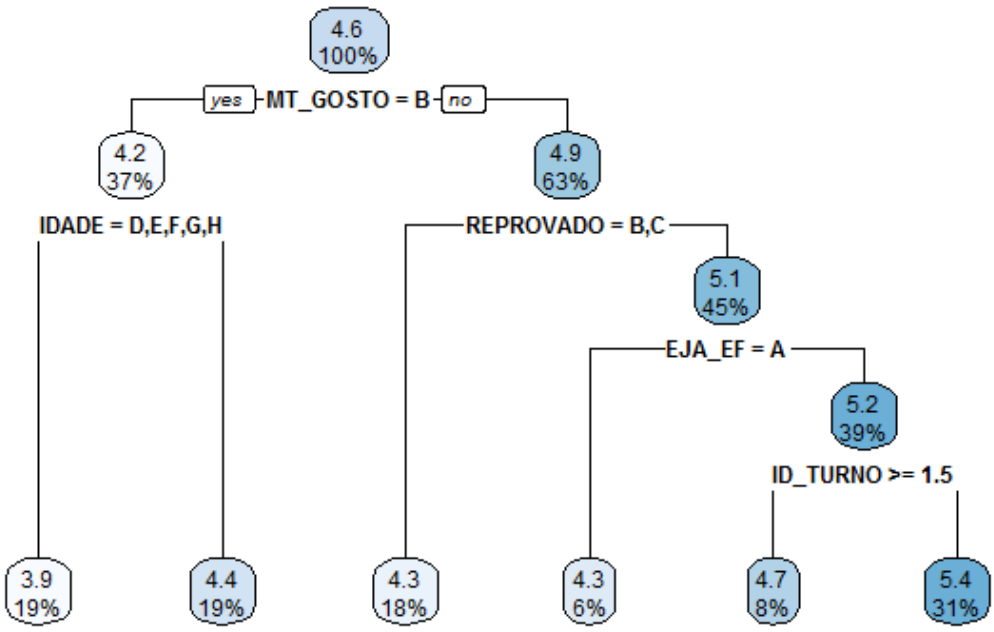
Pelos gráficos de importância das variáveis, observa-se que os perfis dos resultados são consideravelmente diferentes, uma vez que, para língua portuguesa, a variável mais importante é se o(a) estudante cursou o ensino fundamental no EJA (>35% de variação no MSE quando a variável é retirada); e, para matemática, se o(a) estudante gosta ou não da disciplina (~45% de variação no MSE).

Porém, algumas variáveis aparecem nas duas listas, como a própria EJA_EF (que é a terceira na lista de MT); ID_TURNO; IDADE, ID_SERIE, variável que muda de acordo com se o ensino é integral ou não; e REPROVADO, variável que indica se o (a) estudante possui ou não reprovação.

Em virtude dessas diferenças, e para aumentar a compreensão sobre os perfis dos (as) estudantes, foram criadas árvores isoladas, apenas com as variáveis mais importantes (ou seja, %IncMSE >= 15%). Ficando assim as listas de variáveis usadas na construção das árvores finais: LP ~ EJA_EF + ID_TURNO + REPROVADO + MÍDIAS_TELAS + LIVROS + IDADE + ID_SERIE + DIDATICOS_LP; e, MT ~ MT_GOSTO + SEXO + EJA_EF + ID_TURNO + IDADE + REPROVADO).



Como é possível notar, estudantes com reprovação B (Sim, uma vez) ou C (Sim, mais de uma vez), apresentam, na média, notas menores em língua portuguesa que aqueles que nunca tiveram reprovação. O próximo split, que aparece nos dois lados da árvore, mostra as diferenças entre aqueles que fizeram o ensino fundamental na Educação de Jovens e Adultos (A = Sim) e aqueles que não. Em seguida, os turnos vespertino e noturno apresentam notas piores que o diurno. Finalmente, estudantes que leem livros com pouca frequência ou que não leem (LIVROS = B, C, respectivamente), tendem a apresentar notas menores do que aqueles que leem com frequência. Em resumo, estudantes que já apresentaram reprovação e que fizeram EF no EJA tendem as menores notas (3.5); enquanto estudantes que não reprovaram, não fizeram EF no EJA, estudam pela manhã e leem livros tendem as maiores notas (5.5).



Aqui, temos que o gosto por matemática encaminha para árvores de valores mais altos (uma vez que MT_GOSTO = B representa ausência de gosto pela disciplina). As idades mais altas que 17 anos (D representa 18 anos, E, 19 anos...) tendem a notas mais baixas. Do outro lado da árvore, estudantes que possuem reprovações ou que concluíram o EF no EJA também possuem notas mais baixas, enquanto, mais uma vez, estudantes que estudaram no turno diurno tendem às maiores notas[1].

4. Conclusão

De maneira geral, observando-se as repetições nos perfis, é possível identificar um espaço de atuação para políticas públicas mais tradicionais, conectadas ao combate à distorção e à reprovação, evitando, especialmente, que as reprovações e distorções sejam tão frequentes ao ponto de estudante necessitar concluir o fundamental com a EJA; e também espaço para atuação com políticas públicas focalizadas em padrões individuais, usando, por exemplo, economia comportamental, para incentivar os hábitos de leitura, no caso de língua portuguesa, e o gosto por estudar matemática, no caso dessa disciplina.

Além disso, a escolha por florestas aleatórias mostrou-se acertada, já que a técnica foi robusta para lidar com o alto número de casos e variáveis, sem necessidade de agregar em níveis superiores, ou de criar novas variáveis. Contudo, para uma melhor compreensão do sistema, talvez fosse necessário utilizar técnicas específicas para modelos complexos, como a modelagem baseada em agentes. Isso por que uma das coisas que não fica clara, por exemplo, é como os estudantes que não gostam de estudar

[1] Destacam-se, nos modelos finais das árvores, a ausência de algumas variáveis importantes, como sexo para MT, por exemplo. Possivelmente, a ausência dessas variáveis no modelo simbólico final pode estar relacionada à carga fatorial dessas variáveis sendo captadas por outras – como, no exemplo citado, o gosto por estudar matemática.

se relacionam com as atividades, os outros estudantes, e como escolhem, por exemplo, fazer ou não o dever de casa. Evidentemente, isso passa não somente pela modelagem, mas pela disponibilidade dos dados, que em virtude da gama de prazos, se encontram em temporalidade muito diferente daquela do contato dos estudantes com seu ambiente.

Então, para prosseguimento dos estudos, cumpre, uma vez que o mapeamento encontra-se estruturado, encaminhar para o passo de formulação de soluções e de formulação de políticas públicas para os problemas identificados, fazendo com que o uso de inteligência artificial possa refletir na melhoria tanto do fazer pedagógico, quando das políticas públicas educacionais.

Referências

Breiman, L. (2001) Random Forests. In *Machine Learning*, 45, p. 5-32.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2018). *Microdados da Aneb e da Anresc 2017*. Recuperado em 31 de outubro de 2018, de http://portal.inep.gov.br/basica-levantamentosacessar.

Jacobson, M. (2015) "A educação como sistema complexo: implicações para a pesquisa educacional e políticas", Modelagem de sistemas complexos para políticas públicas, Bernardo Furtado, Patrícia Sakowski and Marina Tóvolli, Brasília, Brasil, IPEA, p. 335-350.

Mitchell, T. (1997) Machine Learning, McGraw-Hill Science/Engineering/Math.

Penteado, B.E., Bittencourt, I. I, Isotani, S. (2019) Análise exploratória sobre a abertura de dados educacionais no Brasil: como melhorar o ecossistema de dados na Web? Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 27(1), 175-195. DOI: 10.5753/RBIE.2019.27.01.175.

Pernambuco. (2018) Lei nº 16.379, de 6 de junho de 2018 - Altera a Lei nº 12.985, de 2 de janeiro de 2006, que dispõe sobre o Sistema Estadual de Informática de Governo - SEIG. Recife, Brasil, Assembleia Legislativa de Pernambuco.

Pernambuco. (s/d) Mapa da Estratégia 2015-2018. https://www.seplag.pe.gov.br/servicos-da-seplag/65-mapa-da-estrategia.

Rand, W. (2015) "Sistemas Complexos: conceitos, literatura, possibilidades e limitações", Modelagem de sistemas complexos para políticas públicas, Bernardo Furtado, Patrícia Sakowski and Marina Tóvolli, Brasília, Brasil, IPEA, p. 43-63.

Rigo, J.S., Cazella, S. C. (2014). Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 22(1), 132-146.

Scheinerman, E. (2003) Matemática Discreta: Uma Introdução, São Paulo, Pioneira Thomson Learning.

Zhou, Z-H. (2012) Ensemble Methods: Foundations and Algortithms, Nova Iorque, CRC Press.