Usando Mineração de Dados para Identificar Fatores mais Importantes do Enem dos Últimos 22 Anos

Jacinto José Franco[1], Fernanda Luzia de Almeida Miranda[1], Davi Stiegler[1], Felipe Rodrigues Dantas[1] Jacques Duilio Brancher[3] Tiago C. Nogueira[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT)
Barra do Garças, MT – Brazil

[2]Instituto Federal de Educação, Ciência e Tecnologia Baiano (IFBaiano)
Guanambi, BA – Brasil

[3]Departamento de Ciências da Computação
Universidade Estadual de Londrina (UEL) – Londrina, PR – Brasil

{jacinto.franco,fernanda.miranda}@bag.ifmt.edu.br, davistiegler@gmail.com,

felipe29082002@gmail.com, jacques@uel.br, tiago.nogueira@ifbaiano.edu.br

Abstract. *The national high school exam (Enem) is a test required by most Brazilian universities to select students. In this exam, several characteristics about the candidates are collected. Some of these characteristics do not contribute significantly to the performance prediction, thus representing an excessively large dataset, requiring exponential computational resources to identify them. To solve this problem, this work applies algorithms for the selection and classification of attributes, identifying twenty main characteristics that contribute to the high or low performance of students in this exam, in the last twenty-two years (1998-2019).*

Resumo. *O Exame Nacional do Ensino Médio (Enem) é uma prova requerida pela maioria das universidades brasileiras para seleção de estudantes. Neste exame, são coletadas várias características sobre os candidatos. Algumas dessas características não contribuem de forma significativa para a predição de desempenho, representando assim, um conjunto demasiadamente grande de dados, o que requer recursos computacionais exponenciais para identificá-las. Para solucionar tal problema, este trabalho aplica algoritmos de seleção e de classificação de atributos, identificando as vinte principais características que contribuem para o desempenho alto ou baixo dos estudantes neste exame, nos últimos vinte e dois anos (1998-2019).*

1. Introdução

No Brasil, os estudantes de nível médio que almejam ingressar em algum curso de nível superior precisam realizar testes para avaliar seus conhecimentos. Essa avaliação pode ocorrer de duas formas, via vestibular ou pelo Exame Nacional do Ensino Médio, o Enem. Este é aplicado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) [Lima et al. 2019].

Nos anos iniciais de aplicação do Enem, o teste contou com um número pequeno de inscritos; contudo, esse número foi crescendo ao longo dos anos ao longo dos anos à

DOI: 10.5753/cbie.sbie.2020.1112

medida que as instituições de ensino superior brasileiras foram adotando-o como requisito para o ingresso em seus cursos. Atualmente, essa prova é largamente utilizada por grande parte das instituições públicas e privadas do país, pois provê uma métrica confiável para avaliar os conhecimentos do público estudantil.

O resultado de cada edição do Enem realizada nos últimos anos está disponível via microdados anonimizados pelo Inep[1], o que viabiliza análises quanto a relevância de cada fator. Uma dificuldade que muitos pesquisadores podem ter ao analisar o Enem diz respeito ao número de fatores coletados pelo Inep em cada prova, que pode chegar a 327 variáveis coletadas em um único ano. Para resolver esse tipo de problema, em mineração de dados, vários métodos foram desenvolvidos com o intuito de reduzir a dimensionalidade, permitindo dizer quais fatores são mais importantes e eliminar os que são redundantes ou irrelevantes [Stańczyk and Jain 2015].

Considerando a mencionada constatação, este trabalho fez uso de métodos de seleção de atributos e de classificação para identificar os fatores que mais contribuem para o desempenho do estudante como sendo baixo ou alto. Nos datasets avaliados foram considerados todos os fatores como entrada dos algoritmos de seleção de atributos, com exceção dos que se referiam às notas, às informações do gabarito e à presença dos candidatos.. Ao final, escolheram-se somente os 10 melhores atributos retornados por cada algoritmo de acordo com o *ranking*. As 10 características de cada ano foram aglutinadas em um *ranking* global após a conclusão dos experimentos de cada ano.

Para mensurar a qualidade dos fatores retornados pelos métodos de seleção, este trabalho fez uso de algoritmos de classificação e da média das dez execuções, estabelecendo-se a melhor combinação de fatores cuja taxa de acertos foi maximizada.

Com a finalidade de avaliar os dados entre 1998 e 1999, este trabalho aplicou uma abordagem exploratória para identificar algoritmos que possuem melhores desempenhos em relação às informações contidas nos datasets do Enem. A partir dos resultados obtidos, os quatros melhores métodos de seleção e os dois melhores algoritmos de classificação foram selecionados para evidenciar os fatores mais relevantes no contexto educacional para a performance dos estudantes.

Quanto à sua estrutura, além desta introdução (seção 1), este artigo apresenta: a seção 2, que traz a revisão da literatura, com intuito de identificar os principais trabalhos sobre mineração de dados educacionais no Enem; a seção 3, na qual se realiza o delineamento experimental demonstrando as fases de pré-processamento, seleção e classificação das melhores características aplicadas no Enem; a seção 4, que apresenta os resultados e discussões; e a seção 5, que traz as conclusões deste estudo.

2. Revisão da Literatura

Para a identificação de possíveis lacunas na literatura, buscou-se por trabalhos que aplicassem a mineração de dados educacionais no conjunto de dados do Enem. Essa procura ocorreu de maneira exploratória, utilizando-se as engines de buscas *Google Scholar*, *ISI Web of Science*, *ACM Digital Library*, *IEEE Digital Library*, *Scielo*, *Science Direct*, *Scopus* e *Springer*.

[1] `http://inep.gov.br/microdados`

Dessa forma, definiu-se uma string de busca genérica a fim de detectar o maior número de estudos possíveis sobre a temática, elencando-se, como a principal *string*, a palavra-chave 'Enem'. Todos os resultados advindos das engines acima elencadas foram considerados, desde que a quantidade de entradas fosse inferior a 1.000 registros. Não obstante, por ser considerado uma *string* de busca genérica, que gera muitos registros no *Google Scholar* passíveis de descarte por estarem fora do escopo deste trabalho, acrescentou-se ao protocolo mais uma palavra-chave. Assim, utilizando-se dos conectores lógicos, a *string* de busca final foi ('Enem' *and* 'mineração de dados'), limitando as buscas em 500 registros. Ao final da aplicação deste protocolo, obteve-se o quantitativo de 9 estudos sobre mineração de dados e Enem, em todas as *engines* utilizadas nesta pesquisa.

Para a tarefa de filtrar os registros do Enem foram considerados apenas artigos publicados em periódicos e eventos, sendo sua seleção feita pelo título, resumo, pelas palavras chaves e conclusões. Por essa filtragem, constatou-se que somente 9 artigos de fato usavam algum método da mineração de dados aplicado aos datasets do Enem. Os estudos listados na Tabela 1 estão organizados de acordo com a classe de algoritmo utilizado.

Tabela 1. Listagem de trabalhos organizados por pela classe de algoritmos.

Classe de algoritmos	Trabalhos
Seleção de atributos	[Abreu et al. 2018] [Gomes et al. 2017] [Costa et al. 2017]
Classificação	[Alves et al. 2018] [Braga and Drummond 2016] [Costa et al. 2017] [Adeodato et al. 2014] [Abreu et al. 2018] [Stearns et al. 2017] [Gomes et al. 2017]
Regressão	[Alves et al. 2018]
Agrupamento	[Ideas 2019] [Albertini and Backes 2017]
Regras de associação	[Alves et al. 2018] [Braga and Drummond 2016] [Gomes et al. 2017] [Adeodato et al. 2014]

Observa-se na Tabela 1 que a maior parte dos artigos focou-se em um único *dataset*, com exceção de [Ideas 2019] que explorou os datasets de 2009 à 2015 e [Albertini and Backes 2017] usou os dados de 2010 e 2011, os demais artigos focaram-se em apenas um ano do Enem. Destes trabalhos, é possível notar que nenhum dos autores fez uso de métodos da mineração de dados para prover uma visão geral acerca dos fatores de todos os *datasets* do Enem, desde 1998. Em decorrência dessa constatação, este trabalho utilizou métodos de seleção de atributos para identificar os 10 fatores mais importantes de todos os anos do Enem, tendo sua validação provida por meio de algoritmos de classificação.

A partir do que se verificou na literatura, observou-se que muito do que se coletou em todos os anos do Enem ainda não foi analisado utilizando métodos de mineração de dados. Sendo assim, o presente trabalho preenche essa lacuna com um desenho experimental replicado para todos os anos, provendo uma visão de alto nível acerca dos 20 fatores considerados mais importantes na classificação.

3. Delineamento experimental

Nesta seção são apresentados os três passos usados para obter as 10 melhores características do Enem, que consistem no pré-processamento, na seleção de 10 atributos mais relevantes para cada dataset desde 1998 e na validação pela classificação.

3.1. Pré-processamento

O primeiro passo efetuado consistiu em calcular as médias aritméticas de todas as notas para cada prova. O limite de 600 pontos determina se o aluno tem desempenho baixo ou alto, sendo considerado alto quando o aluno alcança a média cujo valor é igual ou superior a 600 pontos e baixo se o valor for inferior ao mencionado limite.

Essa abordagem foi aplicada somente para o ano de 2018; para os demais, considerou-se uma constante, que é 1,1306926609053397, multiplicada pela média para estabelecer o limite de notas altas. É válido esclarecer que se chegou ao valor da constante por meio da divisão de 600 pela média obtida em 2018. Isso se fez necessário para normalizar entre os anos o conceito de aluno com desempenho alto ou baixo, pois a média de cada ano foi distinta, em virtude, dentre outros motivos, de possíveis variações quanto ao grau de dificuldade das provas entre as diferentes edições do exame.

Saliente-se, ainda, que as informações de todos os estudantes foram consideradas no experimento, com exceção às que se referiam aos alunos que não preencheram os questionários, que eram treineiros, que zeraram na(s) prova(s) e aos que faltaram às etapas do exame ou a alguma delas.

Após a definição do filtro inicial e do perfil de aluno com alto desempenho, o próximo passo consistiu em converter o dataset de cada ano em números inteiros com o comando apply da biblioteca Pandas em todas as colunas dos datasets. Cada nível das características coletadas tem um valor inteiro único, mesmo os valores nulos. Para estes, o mesmo resultado pode ser obtido usando a *SimpleImputer*[2] disponível no *Scikit-learn*.

Um problema observado quando se utilizou de classificação foi o desbalanceamento das classes. Nessa situação, ao se aplicar os algoritmos de classificação, percebeu-se que enquanto houve grande índice de acerto na apresentação da classe de alunos com baixo desempenho, o mesmo não ocorreu em relação à categoria de alunos com alto desempenho, o que invalidou o resultado. Uma abordagem muito comum em ciência de dados é usar algum método sintético para balancear as classes. Essa tarefa foi executada usando o SMOTE (*Synthetic Minority Oversampling Technique*) implementado na biblioteca imblearn[3].

Com o pré-processamento concluído, o *dataset* ficou pronto para a próxima etapa, a seleção de atributos, podendo, na sequência, ser realizada a execução dos classificadores. Um aspecto relevante a ser observado sobre os datasets em questão referiu-se ao número de fatores coletados. Nota-se que esse número variou entre 100 e 327, conforme a Figura 1, e com apenas 10 características foi possível obter taxas de acerto de 69.2% a 89%, o que pode ser verificado na Figura 2 da seção 4.

3.2. Seleção dos atributos mais relevantes

Inicialmente o experimento partiu de uma abordagem exploratória no sentido de testar vários algoritmos de seleção de atributos e de classificação, não sendo viável executar em datasets muito grandes, devido ao tempo requerido para isso. Portanto, escolheram-se os dados de 1999 para apontar os algoritmos mais eficientes em identificar os fatores que propiciam melhor resultado na classificação.

[2] https://scikit-learn.org/stable/modules/impute.html
[3] https://imbalanced-learn.readthedocs.io/en/stable/index.html

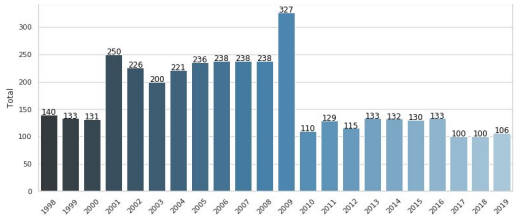

Figura 1. Total de características consideradas de todos os anos do Enem.

Considerando haver na literatura muitos algoritmos de seleção de atributos e que a execução de todos iria requerer muito tempo, neste trabalho foram escolhidos alguns algoritmos disponíveis no software Weka na versão 3.9.4, Scikit-learn na versão 0.23.1 e na biblioteca MLxtend[4] na versão 0.17.2. Outra abordagem usada para evidenciar os atributos mais significativos consistiu em executar algoritmos de classificação e extrair o feature_importances_ após o treinamento do modelo. Somado aos algoritmos, também foi testado os atributos cujo fator de correlação e Predictive Power Score (PPS) é maior. O fator de correlação foi obtido pela biblioteca pandas na versão 1.0.1 e o PPS pela biblioteca ppscore 0.0.2.

Foram considerados inicialmente os seguintes algoritmos de seleção de atributos:

- **Weka** - InfoGainAttributeEval, SymmetricalUncertAttributeSetEval e GainRatioAttributeEval
- **Scikit-learn** - SelectKBest, RFE, GenericUnivariateFeatures e PCA.
- **MLxtend** - SequentialFeatureSelector (SFS) combinado com o classificador LinearRegression do Scikit-learn.

Os classificadores abaixo foram usados para extrair as características mais importantes providas pelo atributo feature_importances_:

- XGBoost provido pela biblioteca xgboost na versão 1.0.2.
- LinearSVC e ExtraTreesClassifier disponíveis no Scikit-learn.

3.3. Classificação

No experimento, os algoritmos de classificação serviram para dizer o quanto é possível chegar em termos de classificação e definir a melhor combinação de fatores. Quanto ao número de classificadores, o experimento não explorou tantas opções devido à quantidade de dados que precisaram ser processados.

Os seguintes classificadores foram considerados: *XGBoost*, *LightGBM*, redes neurais e árvores de decisão implementados nas classes *MLPClassifier* e *DecisionTreeClassifier* do *Scikit-learn*. Nos algoritmos *XGBoost* e *LightGBM* não foi aplicada nenhuma personalização, no *DecisionTreeClassifier* alterou-se somente o parâmetro n_jobs com valor -1 para permitir o uso pleno de todos os núcleos. Como entrada, os algoritmos tiveram 30% do dataset para teste e 70% para treinamento e cada classificador foi repetido 10 vezes com conjuntos de teste e de treinamento distintos gerados pelo train_test_split do *Scikit-learn*, alterando-se apenas o random_state de cada execução.

[4] `http://rasbt.github.io/mlxtend/`

4. Resultados e discussão

Nesta seção são apresentados todos os resultados obtidos experimentalmente combinando fatores, algoritmos de seleção e de classificação que proporcionam a melhor taxa de acerto. Inicialmente, parte-se de uma abordagem exploratória quanto ao número de algoritmos testados e, depois, passa-se a discorrer sobre a escolha dos quatro melhores algoritmos para a seleção de atributos e dos dois melhores para a classificação.

4.1. Dataset de 1998 e 1999

Os dois primeiros anos do Enem foram testados com o intuito de identificar os algoritmos de seleção e classificações que seriam mais confiáveis quanto à qualidade dos resultados possíveis de serem obtidos nos demais anos.

Tabela 2. 4 melhores resultados combinando todos os algoritmos no dataset de 1999.

Ranking	Método	Algoritmo de classificação	Quantidade de características	Taxa de acerto
1	XGBClassifier	LightGBM	10	76.31567
2	ExtraTreesClassifier	XGBClassifier	10	76.31279
3	SFS	XGBClassifier	10	76.24993
4	ExtraTreesClassifier	NeuralNet	10	76.24670

Tabela 3. Duas melhores combinações com o dataset de 1998.

Ranking	Método de seleção	Algoritmo de classificação	Quantidade de características	Taxa de acerto
1	SFS	LightGBM	10	80.46193
2	PCA	XGBoost	10	80.36647

A partir da execução, repetida por 10 vezes, de todos os algoritmos com 10 características, foram escolhidos para a tarefa de classificação os algoritmos *XGBoost* e *LightGBM*; já para a tarefa de seleção, os *XGBoost*, *ExtraTreesClassifier*, *PCA* e *SFS*. A partir do ano 2000 somente esses algoritmos foram usados, pois executar todos os algoritmos disponíveis geraria um alto custo computacional.

4.2. Resultado combinado de todos os anos

Conforme a Figura 2, os resultados obtidos experimentalmente indicaram que, no período de 1998 a 2019 a taxa de acerto possível variou entre 69.2% e 89.0% com 10 características.

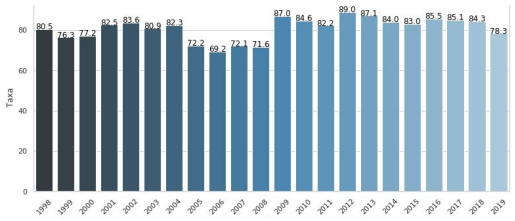

Figura 2. Taxa de acerto do experimento dos últimos 22 anos.

Em razão do grande número de características gerado pela busca efetuada nos 22 anos de exame, para elencar as mais relevantes em cada ano, optou-se por usar o atributo feature_importances_ obtido a partir de uma execução em cada dataset com o

XGBoost para indicar os 20 atributos considerados mais importantes. A partir da média do valor feature_importances_, permitiu-se indicar os atributos que em média foram mais relevantes para a atividade de classificação.

Tabela 4. Ranking dos melhores atributos dos últimos 22 anos do Enem.

Ranking	Descrição	Anos	Score Médio
1	Língua Estrangeira	2016, 2018, 2019, 2011, 2012, 2013, 2014, 2010	0,302309875
2	Grau de importância quanto aos motivos que levaram a participar do ENEM Para Conseguir uma bolsa de estudos (ProUni, outras)	2010	0,280232
3	O quanto você se interessa e acompanha: a politica internacional	1999, 2000,	0,248994
4	Se indicou indígena, qual(is) língua(s) você domina	2005, 2006, 2007, 2008	0,2179125
5	Indique os motivos que levaram você a participar do ENEM: Conseguir uma bolsa de estudos (ProUni, outras).	2015,2016, 2012, 2013, 2014	0,1878978
6	Indique o que levou você a participar do ENEM: Conseguir uma bolsa de estudos (ProUni, outras)	2011	0,16978
7	Você tem em sua casa? Microcomputador	1999, 2000, 2002, 2003, 2004, 2006, 2007, 2009, 2011, 2012, 2013, 2014	0,160423333333333
8	Fez curso de língua estrangeira	2009, 2000, 2001, 2002, 2003, 2004, 2013, 2010	0,155758625
9	Indique os cursos que você frequenta ou frequentou: Curso superior	2013, 2014, 2011, 2010	0,153418
10	Em que tipo escola cursou ou está cursando o ensino médio (2º grau)	2000, 2001, 2002, 2003, 2004, 2005, 2008, 2015, 2016, 2017, 2018, 1998	0,144833846153846
11	Você considera que conhece suficientemente a atividade de trabalho Que você escolheu.	1999	0,129628
12	O quanto você se interessa pela política dos outros países	2001, 2002, 2003, 2004, 2005, 2006	0,1238565
13	Qual é a renda mensal de sua família? (Some a sua renda com a dos Seus familiares.)	2015, 2016, 2017, 2018, 2019, 2005, 2006, 2007, 2008, 2012, 2010	0,1196753
14	Sexo	2011, 2013, 2014, 2017, 2003, 2004, 2005, 2006, 2007, 2008	0,11832
15	Tem Acesso à Internet e quantos	2003, 2004, 2009, 2011	0,11703875
16	Em que tipo de escola cursou o ensino fundamental (1ª grau)	2004, 2006, 2007, 2008, 2010, 2012	0,10934
17	Durante o Ensino Fundamental, você abandonou os estudos e/ou Foi reprovado?	2015,2016	0,103279
18	Em que turno cursou ou esta cursando o ensino médio (2º grau)	2000, 2001, 2002, 2003, 2005, 1998, 2005	0,093806428571429
19	Ao concluir o ensino médio (2º grau) o candidato do ENEM pretende fazer curso(s) profissionalizante(s) e me preparar para o trabalho.	1999	0,087494
20	Você tem em sua casa? Banheiro	2012	0,08709

A Tabela 4 elencou os fatores mais importantes para os classificadores. O resultado seria diferente se fosse utilizado o fator de correlação como referência. Embora a correlação dos fatores com o resultado não seja o foco principal deste trabalho, é importante dizer que a renda familiar se tornou o fator de maior correlação com a definição do desempenho como baixo ou alto somente a partir de 2011. Na Tabela 5, considerou-se o maior fator de correlação que mais se aproxima de 1 positivo ou negativo.

Quanto ao algoritmo que permitiu a melhor classificação, o *XGBoost* destacou-se em quase todos os anos. Em média, foi possível obter a taxa de 80,85% de acerto, conforme a Tabela 6, o que proporcionou maior segurança para dizer que os fatores elencados na Tabela 4 poderiam ser considerados relevantes. O melhor resultado foi o obtido com o *XGBoost* para o ano de 2012, tendo assertividade de 89%.

Tabela 5. Maior fator de correlação de todos os anos do Enem.

Descrição	Fator	Ano
Tipo de escola que cursou o ensino médio (2ºgrau)	0,422846794455483	1998
O quanto você se interessa e acompanha: a politica internacional	-0,32451507473847	1999
	-0,346604306461826	2000
Em que tipo escola cursou ou está cursando o ensino médio (2º grau)	0,380182101996118	2001
Tem Microcomputador e quantos	-0,290258678805769	2002
Está frequentando um curso profissionalizante	0,256403650140639	2003
Até quando a mãe estudou	0,28581360289315	2004
Se indicou indígena, qual(is) língua(s) você domina	0,249277195498918	2005
	0,216173206398373	2006
	0,219766219644129	2007
	0,257849686576848	2008
Até quando sua mãe estudou	0,21839460163376 1	2009
Grau de importância quanto aos motivos que levaram a participar do ENEM para Conseguir uma bolsa de estudos (ProUni, outras)	-0,329040009597119	2010
Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	0,397946320171181	2011
	0,410388728724133	2012
	0,39844400990799	2013
	0,401047691589809	2014
	0,248671127151669	2015
	0,404529502005125	2016
	0,407309507632146	2017
	0,410839404783366	2018
	0,439258036363207	2019

Observa-se que a melhor combinação de algoritmos obtida no experimento foi o do *ExtraTreesClassifier* com o *XGBoost* para a classificação, em 8 anos, seguido do *XGBoost* para a classificação e seleção, em 7 anos, e do *SFS* com o *XGBoost*, em 4 anos. Entre os quatro métodos escolhidos para a seleção de atributos, o *PCA* é o seletor de atributos menos eficiente.

Utilizando o método de classificação supracitado, foi possível constatar que o que é perguntado aos candidatos no questionário de cada Enem influencia diretamente o resultado da classificação. Partindo dessa premissa, o questionário do exame de 2012 foi considerado o melhor estruturado de todas as edições, pois, no referido ano,foi possível obter a taxa de acerto de 88.9774% executando o *XGBoost*.

Segundo a Tabela 6, corroborada pela Figura 3, o fator de classificação considerado mais importante para 2012 foi o tipo de escola que o estudante cursou o ensino fundamental. Uma possível justificativa para isso seria o fato de, na educação escolar brasileira, esse nível de ensino representar $3/4$ dos estudos anteriores ao ingresso da maioria dos candidatos em algum curso de nível superior. Infelizmente tal informação não pôde ser coletada nas últimas edições do Enem, em virtude de seus respectivos questionários não contemplarem mais essa pergunta

Tabela 6. Melhores características para a classificação de 2012.

Importância	Fator	Descrição
0.231240	Q32	Em que tipo de escola você cursou o Ensino Fundamental?
0.191640	TP_LINGUA	Tipo de Língua Estrangeira
0.104269	Q3	Qual é a renda mensal de sua família? (Some a sua renda com a Dos seus familiares.)
0.087090	Q21	Você tem em sua casa? Banheiro
0.082882	Q38	Caso você ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades? Bolsa de estudos da empresa onde trabalho.
0.076745	Q10	Você tem em sua casa? Microcomputador
0.064528	Q30	Quantos anos você levou para concluir o Ensino Fundamental?
0.061394	Q16	Você tem em sua casa? Telefone celular
0.050122	Q28	Indique os motivos que levaram você a participar do ENEM: Conseguir uma bolsa de estudos (ProUni, outras)
0.050088	Q4	Quantas pessoas moram em sua casa (incluindo você)?

Em 2012, o desempenho do aluno foi considerado alto, se obteve média maior ou igual a 570.30 pontos. A partir das porcentagens da Tabela 7, notou-se que somente 7.71% dos alunos que cursaram o ensino fundamental na rede pública conseguiram atingir notas acima de 570.30 pontos, ao passo que as alcançaram 45.92% dos que cursaram

esse nível escolar na rede particular de ensino. Essa constatação provavelmente se aplica aos demais anos, pois o Brasil não avançou muito em termos de qualidade do ensino fundamental, que, conforme se averiguou neste estudo, foi o nível que fez toda a diferença para classificar o desempenho dos estudantes em alto ou baixo.

Tabela 7. Porcentagem de alunos cuja nota é considerada é baixo ou alto para 2012.

Descrição	Conceito	Quantidade	Porcentagem
Somente em escola pública	Alto	336.782	7,71%
	Baixo	4.029.581	**92,29%**
Maior parte em escola pública	Alto	78.548	17,75%
	Baixo	364.073	**82,25%**
Somente em escola particular	Alto	338.140	45,92%
	Baixo	398.168	**54,08%**
Maior parte em escola particular	Alto	68.270	28,17%
	Baixo	174.068	**71,83%**
Somente em escola indígena	Alto	99	6,86%
	Baixo	1.345	**93,14%**
Maior parte em escola indígena	Alto	55	7,97%
	Baixo	635	**92,03%**
Somente em escola situada em comunidade quilombola.	Alto	53	8,33%
	Baixo	583	**91,67%**
Maior parte em escola situada em comunidade Quilombola	Alto	74	11,13%
	Baixo	591	**88,87%**

5. Conclusões

A partir dos dados evidenciados nesta pesquisa, foi possível inferir que a redução da dimensionalidade do conjunto de características simplificou o estudo, pois diminuiu seu escopo, permitindo que se focasse no que é de fato relevante para avaliar o desempenho da maior parte dos estudantes brasileiros no ENEM.

Cumpre ressaltar que a consideração de um maior número de atributos não pressupõe necessariamente uma melhor classificação, pois alguns deles são redundantes e/ou irrelevantes para os resultados, conforme verificou-se em alguns classificadores pela variável feature_importance e pelo ranking fornecido pelos algoritmos de seleção de atributos.

Foi possível perceber que alguns atributos considerados importantes nos experimentos foram removidos do questionário do Enem dos últimos anos, a exemplo dos dados relacionados ao ensino fundamental do estudante. Do ponto de vista educacional, observou-se que algumas informações retiradas eram mais relevantes que as que permaneceram no questionário.

Notou-se, ainda, que embora sejam computacionalmente menos onerosos, os métodos fiters selecionam características menos eficientes para classificar os alunos quanto aos seus desempenhos no referido Exame.

Pelos resultados obtidos, mesmo que este estudo não tenha explorado muitos fatores para todos os anos e tenha executado os esembles sem qualquer otimização, mostrou-se mais eficiente que os trabalhos analisados na literatura.

Verificou-se que a importância e a correlação dos atributos entre os anos analisados foram distintas. Após análise das Tabelas 4, 5 e 6, percebeu-se também que os fatores nelas descritos foram os que permitiram classificar melhor os alunos brasileiros de nível médio. Além disso, constatou-se a necessidade de uma reformulação no questionário a ser preenchido pelos estudantes, a fim de melhorar a acurácia do modelo de aprendizagem de máquina, pois, em 2019, foi possível obter apenas 78.3% de acerto contra 88.9774% de 2012.

Diante de todo o exposto, o presente estudo pode servir de base para um trabalho futuro, pois provê subsídios para a realização de análises estatísticas acerca dos fatores evidenciados experimentalmente, bem como fornece meios para averiguar quais interações os fatores podem ter entre si, permitindo, assim, uma melhor compreensão dos atributos e arranjos que mais influenciam o desempenho dos estudantes no ENEM.

Referências

Abreu, D., Bittencourt, I., Paiva, R., and Dermeval, D. (2018). Pedagogical recommendation to improve the quality of writing: A case study in a public school. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pages 75–76. IEEE.

Adeodato, P. J., Santos Filho, M. M., and Rodrigues, R. L. (2014). Predição de desempenho de escolas privadas usando o enem como indicador de qualidade escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 25, page 891.

Albertini, M. K. and Backes, A. R. (2017). Visualization of clusters in an educational data set based on convex-hull shape preservation algorithm. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(02):1750004.

Alves, R. D., Cechinel, C., and Queiroga, E. (2018). Predição do desempenho de matemática e suas tecnologias do enem utilizando técnicas de mineração de dados. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 7, page 469.

Braga, L. C. and Drummond, I. N. (2016). Uma abordagem de mineração descritiva aplicada a dados abertos governamentais empregando a ferramenta r. *Anais do Computer on the Beach*, pages 051–060.

Costa, J. A. R., Reis, A. L., SOUZA, D. C., CRISTINO, K. G., Aureliano, M. M., Soares, S. R., Santos, T. E., and SILVA, Y. V. (2017). Técnicas de mineração de dados aplicadas em dados do enem 2015. In *9ª Jornada Científica e Tecnológica e 6º Simpósio da Pós-Graduação do IFSULDEMINAS*.

Gomes, T., Gouveia, R., and Batista, M. (2017). Dados educacionais abertos: associações em dados dos inscritos do exame nacional do ensino médio. In *Anais do Workshop de Informática na Escola*, volume 23, page 895.

Ideas, N. (2019). Desempenho das Instituições Brasileiras no ENEM: uma Abordagem Usando Mineração de Dados. pages 106–113.

Lima, P. d. S. N., Ambrósio, A. P. L., Ferreira, D. J., and Brancher, J. D. (2019). Análise de dados do enade e enem: uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 24(1):89–107.

Stańczyk, U. and Jain, L. C. (2015). *Feature selection for data and pattern recognition*. Springer.

Stearns, B., Rangel, F., Firmino, F., Rangel, F., and 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*. SBC.