Estudo sobre as relações socioeconômicas e o desempenho dos alunos na prova do ENADE 2018.

Caique Augusto Cardoso de Moraes, Maíra Matos Araújo, Jander Almeida Silva, Lucas Souza Nogueira Santos, Ruan Nilton, Yasmim Thasla Santos Ferreira, Marcio Rene Brandão Soussa.

Centro Universitário Senai Cimatec

Resumo: O Exame Nacional de Desempenho dos Estudantes (ENADE) foi proposto pelo INEP (Instituto Nacional de Pesquisas Educacionais Anísio Teixeira) em 2004 com o foco de avaliar o desempenho dos graduandos das Instituições de Ensino Superior (IES) brasileiras, por esse motivo o presente trabalho tem como objetivo compreender as possíveis correlações entre fatores socioeconômicos, educacionais e o desempenho dos alunos, por meio da aplicação de técnicas de mineração de dados no conjunto de dados do ENADE em 2018. O trabalho teve como meta o desenvolvimento de uma ferramenta de painel que inclui análise de indicadores e pode gerar relatórios com o intuito de auxiliar na preparação de trabalhos futuros relacionados a exames de profissionais que atuam na área de ensino superior. Para atingir o objetivo, desenvolveu-se uma metodologia baseada na aquisição dos dados, revisão da literatura e levantamento de hipóteses paralelamente a aplicação de técnicas de Data Mining na base de dados escolhida e desenvolvimento da aplicação. Os resultados foram similares aos de outros autores e indicaram a importância do ensino superior público .

1. Introdução

O Exame Nacional de Desempenho de Estudantes (ENADE) foi proposto pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) em 2004, compondo o Sistema Nacional de Avaliação da Educação Superior (SINAES) BRITO, Márcia Regina (2018)[1]. Tem como função avaliar o desempenho dos concluintes dos cursos de graduação nas Instituições de Ensino Superior (IES) do Brasil de acordo com o Inep (2020) [2].

A aplicação da avaliação é considerada complementar e abrangente, visto que considera os aspectos de ensino, vivência e construção do processo de aprendizado durante a graduação. Considera-se que combinando os resultados do exame com os indicadores de qualidade e com os resultados alcançados na avaliação externa por meio das visitas *in loco* (em que são aplicados questionários e entrevistas) reflete-se sobre a forma de avaliar a qualidade de cursos de GRIBOSKI, CLAUDIA (2012) [3].

Sendo assim, a cada aplicação do ENADE, o volume de informações coletadas sobre as instituições, alunos e conteúdos programáticos dos cursos, forma uma grande base de dados que é utilizado como material de estudo para autores como NOGUEIRA, ANDRINO (2015) [4], que desenvolveu a pesquisa Mineração de Dados para análise da relação entre características socioeconômicas de concluintes do ensino superior e o desempenho desses estudantes no ENADE e LIMA, PRISCILA (2019) [5] com a Análise de dados do Enade e Enem: uma revisão sistemática da literatura.

O Enade de 2018, campo de estudo deste artigo, registrou a inscrição de quase 1,2 milhão de estudantes. Sendo possível, utilizar a massiva quantidade de informações coletadas nas aplicações da prova para desenhar perfis de alunos, habilidades, competências correlatas e compreender padrões de desempenho. Da mesma forma, é possível confrontar ou confirmar hipóteses sobre como os planos de ensino e oportunidades oferecidas pelas IES são fatores influenciadores no desempenho que os estudantes obtêm na prova.

Portanto, o presente trabalho tem como objetivo compreender através da aplicação de técnicas de

mineração de dados no *dataset* do ENADE de 2018 possíveis correlações entre fatores socioeconômicos, educacionais e o desempenho dos alunos, desenvolvendo uma ferramenta dashboard que contenha análises de indicadores e possibilite a geração de relatórios, visando auxiliar a confecção de trabalhos futuros relacionados a prova pelos profissionais que atuam na área de educação das instituições de ensino superior.

2. Metodologia

2.1 Obtenção de Dados

Os dados foram obtidos a partir do site do Governo Federal (2020) [6] que contém dados do exame realizado no ano de 2018, possibilitando assim uma grande gama de informações que possibilitou uma análise exploratória dos dados direcionando-se nas características que impactam diretamente na nota no aluno, como fatores socioeconômicos citados no artigo de MEDEIROS FILHO (2020) [7].

O dataset do Enade 2018 é dividido em 137 colunas, possui 548127 tuplas contendo informações referentes ao curso, nota geral e específica, tipos de presença, percepção da prova e questionário do estudante, onde se encontram as questões que abrangem os dados socioeconômicos dos alunos.

2.2 Data Mining

A figura 1 apresenta o processo de construção do modelo, dividido em quatro passos principais para a obtenção de resultados.

Figura 1. Etapas para a construção do modelo

No pré-processamento para o *data mining* foram retirados alunos ausentes, eliminados e com resultados desconsiderados pelo aplicador da prova para que tivéssemos somente alunos presentes e com resultados válidos. .Também foram retiradas tuplas com informações nulas ou com Not a Number(NaN). Reduzindo o *dataset* para apenas 72.6% em relação ao original com um total de 397928 tuplas.

Os campos de interesse selecionadas do dataset foram sexo, idade, unidade federativa e região do curso, modalidade do curso, renda, etnia, tipo de bolsa de estudo, tipo de ensino medio, membro da familia com ensino superior , situação trabalhista,horas de estudo e a nota geral do aluno. Campos referentes a área do questionário do aluno realizado pelo ENADE contendo dados socioeconômicos fornecidos pelo próprio aluno. As demais colunas foram descartadas por conter informação redundante ou de pouco peso para o modelo.

Após a seleção das partes de interesse do dataset, foram arbitradas 4 classes: Notas < 25 (A), 25 <= Notas < 50 (B), 50 <= Notas < 75 (C), Notas >=75 (D) para a realização da classificação e identificação do perfil do aluno em relação ao seu desempenho. Todo o restante do *dataset* foi

transformado para formato *One Hot Encoder*, formato utilizado pela biblioteca do *Sklearn*. No final obtemos um dataset com 4 classes desbalanceadas contendo as informações descritas na tabela 1.

Na construção do modelo foi utilizado o algoritmo otimizado de Classification and Regression Tree (CART) da biblioteca *Sklearn* . Foram implementadas várias versões da árvore, variando a profundidade, atributos selecionados e divisão dos dados de treino e teste com o intuito de atingir a melhor acurácia. Também foram obtidos os valores da importância do campo, valor que denomina o peso do campo para o modelo.

Tabela 1. Variáveis da Mineração

Nome da variável	Descrição	Categorias
MODALIDADE	Código da Modalidade de Ensino	Educação presencial ou a distância
UF	Unidade Federativa do Curso	Acre (AC) , Alagoas (AL), Amapá (AP), Amazonas (AM), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Mato Grosso (MT), Mato Grosso do Sul (MS), Minas Gerais (MG), Pará (PA), Paraíba (PB), Paraná (PR), Pernambuco (PE), Piauí (PI), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Rondônia (RO), Roraima (RR), Santa Catarina (SC), São Paulo (SP), Tocantins (TO)
REGIÃO	Região de funcionamento do curso	Região Norte (NO), Região Nordeste (NE), Região Sudeste (SE), Região Sul (SUL), Região Centro-Oeste (CO)
SEXO	Sexo	Masculino(M) ou Feminino(F)
IDADE	Idade do inscrito em 25/11/2018	Valores entre 4 e 94
NOTA_GERAL	Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%) (valor de 0 a 100)	Min = 0 Max = 93,7 (Máximo para o ano de 2018)
ETNIA	Qual é a sua cor ou raça?	Branca.
		Preta.
		Amarela.
		Parda.
		Indígena.
		Não quero declarar.
RENDA	Qual a renda total de sua família, incluindo seus rendimentos?	Até 1,5 salário mínimo (até R$ 1.431,00).
		De 1,5 a 3 salários mínimos (R$ 1.431,01 a R$ 2.862,00).
		De 3 a 4,5 salários mínimos (R$ 2.862,01 a R$ 4.293,00).
		De 4,5 a 6 salários mínimos (R$ 4.293,01 a R$ 5.724,00).
		De 6 a 10 salários mínimos (R$ 5.724,01 a R$ 9.540,00).
		De 10 a 30 salários mínimos (R$ 9.540,01 a R$ 28.620,00).
		Acima de 30 salários mínimos (mais de R$ 28.620,00).
SITUACAO_TRABALHO	Qual alternativa a seguir descreve sua situação de trabalho (exceto estágio ou bolsas)?	Não estou trabalhando.
		Trabalho eventualmente.
		Trabalho até 20 horas semanais.
		Trabalho de 21 a 39 horas semanais.
		Trabalho 40 horas semanais ou mais.
TP_BOLSA	Que tipo de bolsa de estudos ou financiamento do curso você	Nenhum, pois meu curso é gratuito.
		Nenhum, embora meu curso não seja gratuito.

	recebeu para custear todas ou a maior parte das mensalidades? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.	ProUni integral.
		ProUni parcial, apenas.
		FIES, apenas.
		ProUni Parcial e FIES.
		Bolsa oferecida por governo estadual, distrital ou municipal.
		Bolsa oferecida pela própria instituição.
		Bolsa oferecida por outra entidade (empresa, ONG, outra).
		Financiamento oferecido pela própria instituição.
		Financiamento bancário.
TP_ENSINO_MEDIO	Em que tipo de escola você cursou o ensino médio?	Todo em escola pública.
		Todo em escola privada (particular).
		Todo no exterior.
		A maior parte em escola pública.
		A maior parte em escola privada (particular).
		Parte no Brasil e parte no exterior.
SUPERIOR_FAMILIA	Alguém em sua família concluiu um curso superior?	Sim.
		Não.
HORAS_ESTUDO	Quantas horas por semana, aproximadamente, você dedica aos estudos, excetuando as horas de aula?	Nenhuma, apenas assisto às aulas.
		De uma a três.
		De quatro a sete.
		De oito a doze.
		Mais de doze.

2.3 Revisão da literatura e definição de hipóteses

Em paralelo ao desenvolvimento da etapa de Data Mining, ocorreu a revisão do estado da arte para fundamentação teórica necessária para o levantamento de hipóteses envolvendo os atributos da base de dados utilizada, tais hipóteses foram consideradas na aplicação das técnicas de mineração de dados visando o confronto e confirmação das mesmas a partir dos resultados obtidos.

Foram usadas as bases eletrônicas SCIELO e Google Scholar para seleção de sete trabalhos correlatos que embasaram o entendimento das especificidades de aplicação do ENADE e utilização das bases de dados geradas para pesquisa científica. Entre as palavras chaves de pesquisa, destacam-se mineração de dados e ENADE.

Um dos pontos avaliados na literatura diz respeito à relação entre a entrada na IES por políticas de inclusão e o resultado que os alunos obtêm no desenvolvimento no curso e prova do ENADE. Do ponto de vista da avaliação de uma política pública, acredita-se que esta é a medida mais correta para o resto da sociedade: os alunos que se beneficiaram de uma política de inclusão não terminam no ensino superior como profissionais de menor qualidade WAINER, Jacques (2018)[8]

A partir do entendimento de diferentes abordagens científicas, foram definidas 10 hipóteses de possíveis relações entre os atributos do dataset:

Tabela 2. Hipóteses e atributos relacionados

Hipótese	Atributos relacionados
1. Relação entre a avaliação dos alunos sobre a instituição e a categoria de organização acadêmica da IES	CO_ORGACAD, QE_I43, QE_I44, QE_I45

2. Relação entre modalidade de ensino e nota bruta na prova	CO_MODALIDADE, NT_GER
3.Relação de indicação da presença dos estudantes por município/estado	CO_MUNIC_CURSO, CO_UF_CURSO, TP_PRES
4. Relação entre a nota na parte de formação geral e a avaliação sobre os planos de ensino das instituições	NT_FG, QE_I38
5. Relação entre renda familiar, descrição da situação financeira do aluno e nota bruta da prova	QE_I08, QE_I09, NT_GER
6. Relação entre alunos com ingresso no curso de graduação por políticas de ação afirmativa ou inclusão social e oportunidade de programa curricular no exterior	QE_I15, QE_I14
7. Relação entre categoria da escola no ensino médio e modalidade de acesso à graduação	QE_I17, QE_I15
8. Relação entre motivações para escolhas de cursos de graduação, primeiros membros da família a possuir uma graduação e renda familiar	QE_I08, QE_I21, QE_I25
9. Relação entre horas de dedicação aos estudos e situação de trabalho do aluno	QE_I23, QE_I10
10. Relação entre capacidade de argumentação e nota bruta na parte discursiva de formação geral	QE_I33, NT_DIS_FG

2.3 Prototipação da Aplicação e Banco de dados

Partiu-se então para prototipação da aplicação que, visando os princípios de Interação Humano Máquina, entendeu-se quais os indicadores de relevância e objetivos de comunicação a serem implementados no *dashboard* de visualização. O objetivo principal da aplicação foi exibir de forma visual os dados processados e as árvores de regressão geradas pelo processo de consulta e mineração de dados, respectivamente. A construção da aplicação se deu pelas tecnologias NodeJS com Typescript para o *backend*, NextJS com Typescript para o *frontend* e para o banco de dados foi utilizado a tecnologia MongoDB, um banco de dados não relacional.

Após aquisição dos dados, o *dataset* passou pela fase do pré-processamento, sendo necessário para prevenção de erros no dashboard. O *dataset* foi separado em duas bases, a primeira contendo alunos apenas presentes e com resultados válidos e a segunda com alunos presentes ou ausentes com notas válidas e inválidas, sendo tratados individualmente e filtrados em 22 colunas a fim de melhorar o desempenho e otimizar o processo da geração de gráficos e posteriormente relatórios.

A escolha do banco não-relacional, no qual teve como princípios a flexibilidade da organização e armazenamento dos dados, como o “documentos embutidos” onde temos todas as referências englobadas dentro de um único documento, e mesmo não sendo um Data Warehouse propriamente dito, o modelo desenvolvido e utilizado para a aplicação possui algumas características desse tipo de organização de banco de dados, tendo em vista que os dados são não-voláteis e o banco é orientado por assunto.

3. Resultados e discussão

3.1. Árvore de Classificação

Foram geradas duas árvores, a primeira árvore para as classes A e D e a segunda referente às classes B e C para extrair o maior número de perfis e aumentar a acurácia para classes pouco representativas. A árvore 1 tem a acurácia de 0.77 e a árvore 2 possui acurácia de 0.60 possuindo 4 níveis de profundidade. Ambas as árvores possuíram um desempenho maior após a diminuição dos parâmetros

passados para o modelo,a retirada de parâmetros referentes a localização do estudante possuíam uma importância que tende a zero.

Também foram medidos valores referentes ao *Recall,* a frequência em que o modelo encontrou uma determinada classe; e a Precisão, que é a razão entre o número de *True Positive* pela soma dos *False Positive com* o *True Positive ;*e o *F1-Score,* que representa a relação entre o *Recall* e a Precisão. Essas métricas de avaliação do modelo estão expressas na tabela 2. A precisão de cada classe é influenciada tanto pela diversidade de perfis de alunos quanto pela quantidade de informações. Pode -se observar também que o modelo classifica muitos falsos positivos para classes com número muito grande de dados.As variáveis com maior impacto nos modelos foram o faculdade pública (0.24), bolsa prouni (0,19) e estudo na a escola privada(0.42). A presença desses campos marcam os nós principais das árvores.

Tabela 2. Reporte de Classificação

	Árvore 1		**Árvore 2**	
	Nota < 25 (A)	**Nota > 75 (D)**	**25 <= Nota < 50(C)**	**50<= Nota < 75(B)**
Precisão	0.82	0.69	0.65	0.57
F1-Score	0.73	0.78	0.65	0.56
Recall	0.74	0.77	0.66	0.55

Os perfis gerados tiveram algumas semelhanças, A classe de alunos A teve como principal perfil pessoas que não estudam em faculdades públicas entretanto estudaram em escolas não particulares e não participam do programa de bolsas PROUNI, além de não possuírem uma renda entre 10 e 30 salários mínimos, a quantidade de alunos que se enquadram nesse perfil são de 21773 alunos. Para a classe D, existiram inúmeros perfis, entretanto, os dois perfis que mais se destacaram estudam em escola particular, contudo o primeiro perfil, com 3784 alunos, não estudava de 1 hora há 3 horas por semana e não possuíam uma renda de entre 10 e 30 salários mínimos; o segundo perfil, com 2890 alunos, estudaram em escola pública, não trabalham mais de 40 horas semanais e não possuem uma renda menor que 1.5 salários mínimos.Os alunos da classe B tiveram um perfil similar a classe A,com 127707 alunos, entretanto, os alunos da classe C frequentaram uma escola particular e estudam em uma faculdade pública, não trabalham mais de 40 horas semanais e não possuem uma renda inferior a 1.5 salários mínimos.

Figura 2. Árvore 1:nós à esquerda

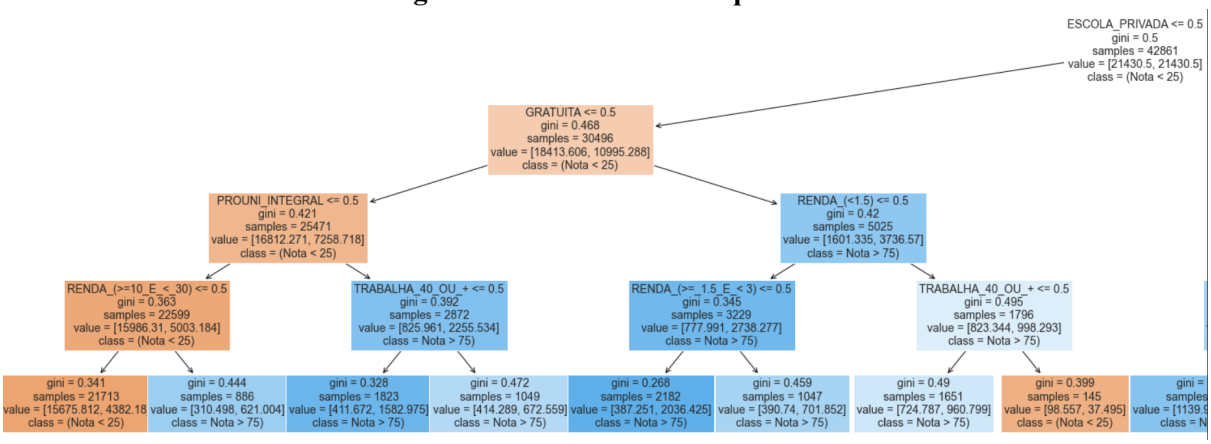


Figura 3. Árvore 1:nós à direita

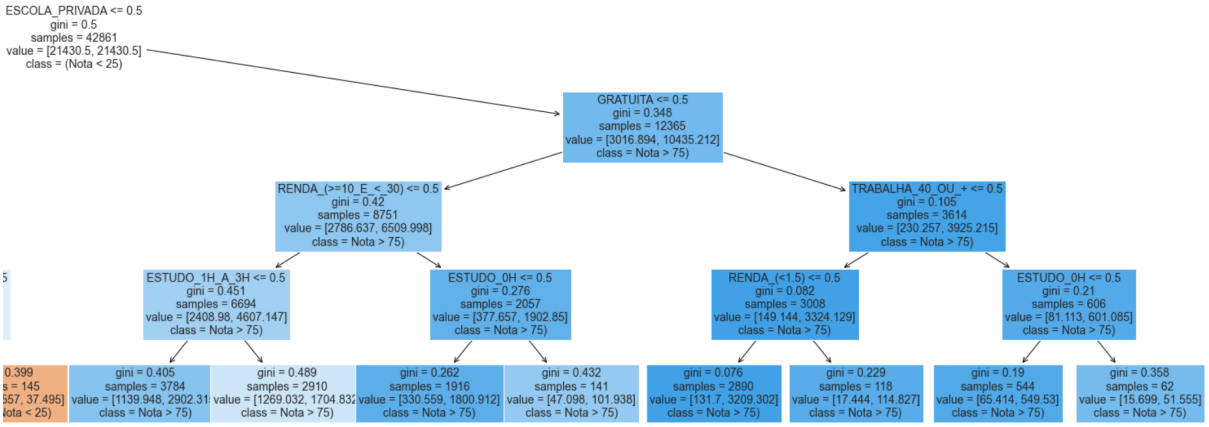


Figura 4. Árvore 2:nós à esquerda

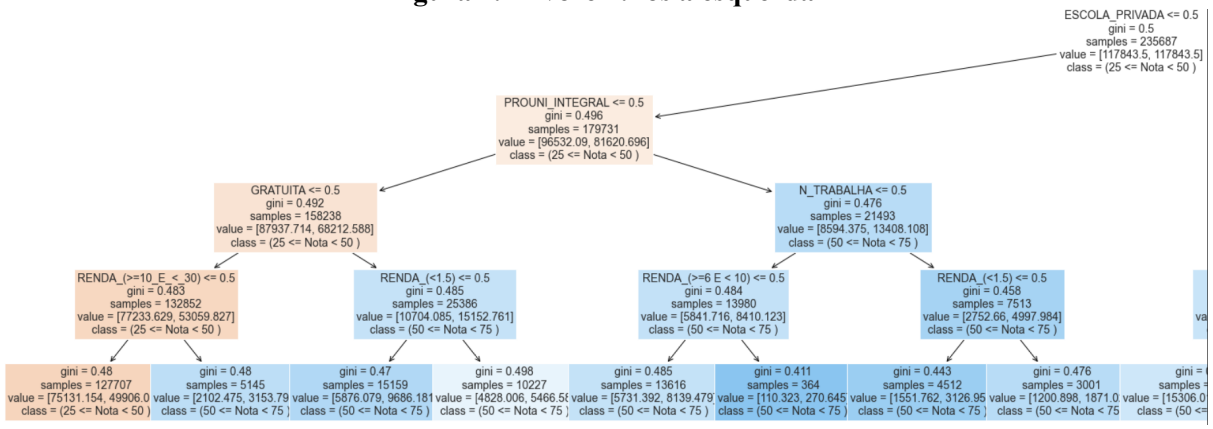
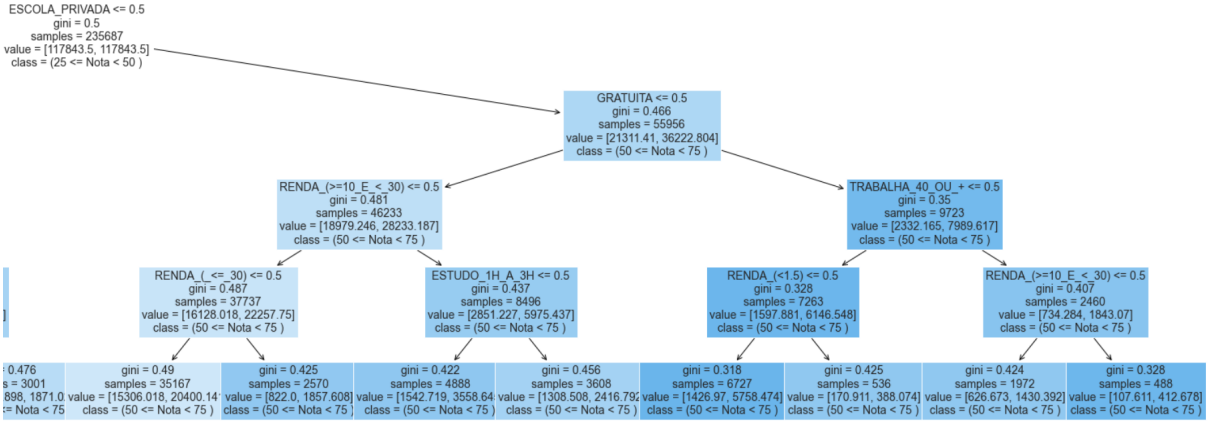


Figura 4. Árvore 2:nós à direita



4. Conclusão

Os resultados obtidos neste artigo se assimilam a outros trabalhos, como o de NOGUEIRA, ADRIANO *et al*, onde alunos que cursaram o ensino médio em uma escola particular e/ou estudam em uma faculdade pública possuem um desempenho melhor no ENADE 2012. Contudo, existem vários pontos de melhoria para trabalhos futuros, o número de classes, por exemplo, poderia deixar de ser arbitrado e ser definido por um algoritmo de agrupamento para que se possa definir os perfis dos grupos gerados e alcançar perfis mais concretos. Outro ponto de melhora seria uma análise sobre a percepção da prova, e como isso impacta em seu desempenho.

A justificativa central desse projeto é tornar os dados da devolutiva do ENADE acessíveis a qualquer público. Servindo de auxílio para órgãos governamentais e profissionais da educação no processo de identificação de pontos de melhorias na aplicação da prova, com o objetivo de melhorar o desempenho dos alunos, resultando em uma melhoria na educação do país.

Segundo o presidente do INEP durante uma coletiva em 2019, destaca: "Se mergulharmos nos estudos das devolutivas, podemos obter avanços importantes. É importante nos aprofundarmos nesses números com o objetivo de aprimorar a educação superior no Brasil". Destacando a relevância dos resultados do Enade para o aperfeiçoamento da educação superior

A aplicação desenvolvida, deverá servir como ferramenta de auxílio nacional na preparação e estudo posterior a aplicação das provas e avaliações do ENADE, visto que, propõe-se a disponibilizar uma análise de dados coerente a partir da técnica de data mining implementada neste trabalho. Busca-se influenciar positivamente nos métodos avaliativos da qualidade das Instituições de Ensino Superior do Brasil, e dessa forma, ser um vetor de melhoria para a educação brasileira.

5. Referências:

[1] BRITO, Márcia Regina F. de. O SINAES e o ENADE: da concepção à implantação. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 13, n. 3, p. 841-850, 2008.

[2] Inep. Governo Federal, 2018. Exame Nacional de Desempenho dos Estudantes (Enade). Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em: 02 de Dezembro de 2020.

[3] GRIBOSKI, Claudia Maffini. O Enade como indutor da qualidade da educação superior. Estudos em avaliação educacional, v. 23, n. 53, p. 178-195, 2012.

[4] ANDRINO NOGUEIRA, EDUARDO DIMAS; TSUNODA, DENISE FUKUMI. MINERAÇÃO DE DADOS PARA ANÁLISE DA RELAÇÃO ENTRE AS CARACTERÍSTICAS SOCIOECONÔMICAS DE CONCLUINTES DO ENSINO SUPERIOR E O DESEMPENHO DESSES ESTUDANTES NO ENADE 2012. Revista Percurso, v. 15, n. 1, 2015.

[5] Inep. Governo Federal, 2020. Microdados do Exame Nacional de Desempenho dos Estudantes. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade. Acesso em: 02 de Dezembro de 2020.

[6] DE MEDEIROS FILHO, Antonio Evanildo Cardoso; ROSEIRA, Ítalo Breno Rocha; PONTES JR, Jose Airton Freitas. Perfil socioeconômico e desempenho de estudantes de licenciatura em educação física no ENADE/BRASIL. Tendencias pedagógicas, n. 35, p. 90-101, 2020.

[7] LIMA, Priscila da Silva Neves et al. Análise de dados do Enade e Enem: uma revisão sistemática da literatura. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 24, n. 1, p. 89-107, 2019.

[8] WAINER, Jacques; MELGUIZO, Tatiana. Políticas de inclusão no ensino superior: avaliação do desempenho dos alunos baseado no Enade de 2012 a 2014. Educação e Pesquisa, v. 44, 2018.

[9] Berry, M.,& Linoff, S. (2000). Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: Wiley