#tear
Revista de Educação, Ciência e Tecnologia

LEARNING ANALYTICS EM EXAME DE DESEMPENHO DE ALUNOS DE UMA INSTITUIÇÃO DE ENSINO SUPERIOR

Learning Analytics Applied in Performance Exams for Higher Education Institution Students

Romão Matheus Martines de Jesus[1]

Eliana Alves Moreira[2]

Resumo: O *Learning Analytics* tornou-se uma ferramenta poderosa no aprimoramento dos processos de aprendizagem e ensino, pois por meio dessa análise é possível entender diversos problemas, tais como os altos índices de repetência e evasão escolar, ou até mesmo construir planos customizados para cada aluno. Este trabalho apresenta um estudo sobre o desempenho de alunos de um curso de graduação brasileiro em relação ao Enade, exame nacional que mede a qualidade do ensino superior, destacando fatores que podem ter influenciado sua performance. Um modelo de Regressão Linear foi desenvolvido para prever o desempenho dos alunos e também a nota do Conceito Enade em aplicações futuras. Resultados mostraram que a pandemia de COVID-19 pode ter influenciado a formação dos estudantes, fator destacado também pelo alto número de ausências no exame em 2021, o primeiro realizado durante o período pandêmico. Essas informações podem sinalizar pontos de atenção que devem ser levados em consideração frente aos alunos/disciplinas ao longo da graduação.

Palavras-chave*: Learning Analytics. Data Analytics*. Regressão Linear.

Abstract: Learning Analytics became a powerful tool that enhances the learning and teaching processes. By means of the analysis it is possible to understand problems such as high failure and dropout rates, or even build customized plans for each student. This study analyzes students' performance at a specific Brazilian graduation program in order to understand which factors might impact the final student performance in Enade, a national exam that measures the quality of graduation programs in Brazil. A Linear Regression model was developed to predict students' performance as well as the Enade Concept score. The results showed the COVID-19 pandemic influenced the students' performance, highlighting the high absence rate at Enade's 2021 exam application, the first one applied during the pandemic. That information can flag the attention points of the students' development throughout the graduation program.

Keywords: Learning Analytics. Data Analytics. Linear Regression.

[1] Bacharel em Engenharia de Computação pela UFSCAR. Estudante de Especialização em Ciência de Dados no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Campus Campinas. ORCID: https://orcid.org/0009-0004-2893-5614. E-mail: r.martines@aluno.ifsp.edu.br.

[2] Doutora em Ciência da Computação pela UNICAMP. Professora no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Campus Campinas. ORCID: https://orcid.org/0000-0002-0042-639X. E-mail: eliana.moreira@ifsp.edu.br.

1 Introdução

Nos últimos anos, a área de *Learning Analytics* (LA) tem se destacado cada vez mais na educação, permitindo a análise de grandes volumes de dados relacionados ao processo de ensino e aprendizagem. Com base nestes conceitos, é possível extrair informações tanto sobre o desempenho dos alunos quanto sobre o processo de ensino.

Segundo Siemens e Long (2011), através da análise de dados sobre a performance dos alunos é possível entender melhor seus desafios e necessidades, o que pode levar a melhorias significativas na qualidade da educação. A importância da *LA* pode ser percebida ao se considerar os desafios enfrentados pela educação, uma vez que com a crescente demanda por ensino de qualidade, torna-se cada vez mais necessário utilizar essas tecnologias para melhorias no processo, como aponta Kovanović (2015).

Nesse contexto, este trabalho tem como objetivo aplicar a técnica de *LA* em microdados referentes a duas edições do Exame Nacional de Desempenho de Estudantes (Enade), realizadas ao longo de dois anos, 2017 e 2021, referentes a um curso de ensino superior. O Enade é um exame de avaliação do desempenho dos estudantes referente a seus respectivos cursos superiores. Com base no desempenho dos candidatos, é possível calcular métricas que refletem a formação dos estudantes de uma determinada instituição, bem como traçar comparações do desempenho entre Instituições de Ensino Superior (IES) dentro de regiões e estados do Brasil.

Analisar os dados do Enade dentro da ótica da *Learning Analytics* permite detalhar a situação dos alunos, destacando os principais influenciadores do desempenho final. O estudo apresentado no presente trabalho visa analisar a performance dos alunos com o intuito de buscar possíveis pontos de atenção que possam afetar diretamente o desempenho desses. As análises estatísticas podem ajudar a destacar variáveis que influenciam direta ou indiretamente no desempenho das turmas, auxiliando em processos de melhoria que podem ser aplicados de forma preventiva. Como parte do estudo, foi desenvolvido um modelo de predição em que será possível estimar a nota de conceito do Enade referente ao curso, bem como o desempenho geral da turma. Os discentes do curso em questão realizam a cada semestre uma simulação do Enade que, além de ser parte das notas finais dos alunos, os prepara para o exame oficial. Logo, o resultado do estudo poderia ser utilizado como parâmetro comparativo com os resultados dos simulados semestrais, servindo como base em uma abordagem colaborativa entre professores, alunos e gestores da universidade.

Este trabalho está organizado da seguinte forma: a seção 2 contém uma revisão do estado da arte, trabalhos relacionados e uma breve fundamentação teórica; na seção 3 é descrita a metodologia utilizada no estudo e como foi realizada a coleta dos dados; a seção 4 apresenta a análise do desempenho dos alunos com base nos dados coletados; a seção 5 apresenta a criação de um modelo de predição para projetar a nota de conceito do Enade para os próximos anos com base no desempenho dos anos anteriores; e na seção 6 realiza-se uma breve discussão e conclusão.

2 Fundamentação teórica e revisão de literatura

O uso de técnicas de análise de dados possibilita fornecer informações que podem ser utilizadas para melhorar a aprendizagem e o ensino, bem como identificar problemas específicos na educação. A *Learning Analytics* é uma área interdisciplinar que consiste em medição, coleta, análise e relato de dados sobre alunos e seus contextos, numa tentativa de

entender e otimizar o processo de aprendizagem nos ambientes nos quais ela ocorre, como elencam Siemens e Long (2011).

De acordo com Foster e Siddle (2020), a análise de dados pode ajudar a identificar a evasão escolar, baixo desempenho acadêmico e dificuldades de aprendizagem em disciplinas específicas e, ainda, ajudar na personalização do ensino, fornecendo um ambiente de aprendizagem mais eficaz. O estudo de Kovanović (2015) sugere que a análise de dados pode ser utilizada para fornecer *feedback* personalizado e melhorar a experiência dos alunos em ambientes de aprendizagem *online*. Selwyn e Facer (2013) destacam a importância da análise de dados para avaliar a eficácia de políticas educacionais e práticas de ensino. Ainda, a análise de dados pode ajudar a prever o desempenho dos alunos em avaliações futuras, como proposto por Brandt *et al.* (2019).

Dentro desse contexto, o uso de técnicas de *LA* pode trazer *insights* para a tomada de decisão educacional. O estudo de Ferreira e Andrade (2013) mostra em uma abordagem mais ampla sobre como foi o processo de implementação de *LA* nos cursos da Universidade Católica Portuguesa de Porto (UCP), relatando o processo de integração e a recepção do sistema pelos alunos. Foi implementado na instituição um *Learning Content Management System (LCMS)*, sistema cujo intuito é coletar informações sobre como os alunos usam as plataformas de ensino, a fim de se obter um estudo de como a *LA* poderia auxiliar nas tomadas de decisões no ensino superior, dado que até aquele momento os estudos estavam focados nos ensinos primários.

Já no trabalho realizado por Morais et al. (2017), o modelo de extração de métricas consegue entender alunos que estão em processo de evasão escolar com base em frequências nas aulas presenciais, acesso à plataforma de ensino e desempenho acadêmico, atingindo uma média correlação dessas variáveis no processo de desmotivação do aluno.

Outros estudos trazem abordagens mais observadoras do movimento de adoção da *LA* no ensino superior brasileiro. Como já apontaram Ferreira e Andrade (2013), a adoção de *LA* como um processo de melhoria e gerenciamento do ensino ainda é muito nova no país, e nos estudos de Scheneider (2022) e Luz et al. (2021) temos, respectivamente, um panorama da maturidade da *LA* no ensino superior brasileiro e uma revisão sistemática da literatura referente ao tema. Embora recente, a inserção das técnicas de *LA* já apresenta melhorias, como mostra o estudo de Portal (2016), que traz uma observação de documentos de frequência em cursos EaD e com isso desenvolve um modelo que consegue predizer o processo de evasão de um aluno e agir de forma a evitar que isso realmente venha a ocorrer.

Semelhante ao que Morais et al. (2017) e Ferreira e Andrade (2013) realizaram, a ideia deste estudo é que, através de um modelo que consiga predizer o desempenho dos alunos numa aplicação do Enade, seja possível que este processo passe a ser introduzido como ferramenta de melhoria da formação de estudantes de um curso superior. Este modelo também pode destacar pontos de atenção nos quais os alunos apresentam dificuldades e, com base nisso, ações de melhoria poderiam ser tomadas de forma mais assertiva, de maneira semelhante ao relatado por Portal (2016) em seu estudo.

No entanto, a área de *LA* ainda apresenta algumas limitações. A falta de dados ou qualidade inadequada dos dados disponíveis pode afetar a precisão das análises e previsões, como apontado por Romero e Ventura (2013). A privacidade e a segurança dos dados dos alunos também são questões importantes a serem consideradas, como discutem Siemens e Long (2011). Logo, é relevante desenvolver estratégias eficazes para coletar e analisar dados, especialmente em um contexto em que a educação faz cada vez mais uso e está cada vez mais dependente de tecnologias digitais e plataformas de ensino *online*.

2.1 Enade - Exame nacional de desempenho dos estudantes

O Enade é uma avaliação do Governo Federal, aplicada por meio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, também conhecido como Inep, órgão federal responsável pelas evidências educacionais que atua em avaliações e exames educacionais, pesquisas estatísticas, indicadores educacionais e gestão do conhecimento e estudos (Brasil, 2024). O Enade avalia o desempenho dos estudantes de cursos de graduação e é aplicado próximo à conclusão de um determinado curso. O ciclo avaliativo do Enade define as áreas de conhecimento que terão os cursos avaliados a cada ano e as áreas têm seus estudantes concluintes avaliados de três em três anos. Como concluinte, entende-se o estudante que cumpriu 80% ou mais da carga horária exigida para formação.

A presença no Enade é uma exigência para a formação do estudante. Analisando os materiais disponibilizados por meio do exame, é possível ter acesso à nota do conceito do curso e a outras notas que dizem respeito ao desempenho dos alunos. Em relação às notas, essas são divididas em dois segmentos: (a) Formação Geral, que diz respeito a temas mais abrangentes e universais; (b) Formação Específica, que, com base na definição da grade curricular, foca em assuntos relacionados diretamente a temas do curso em questão.

O conceito do Enade é calculado através da seguinte fórmula (Brasil, 2022):

$$NC_C = 0.25NP_{FG} + 0.75N_{CE} \qquad \text{(I)}$$

Em que *NC* é a nota dos concluintes do Enade, *FG* a nota em Formação Geral e *CE* a nota em Componente Específico. *NP* diz respeito à nota padronizada, que através de uma interpolação linear faz com que todas as notas fiquem dentro de uma escala entre 0 e 5. Ao final, a depender do valor de *NC* temos as notas do conceito Enade atribuídas conforme a Tabela 1, onde 5 é a maior nota possível.

Tabela 1 – Parâmetros de Conversão do *NC* conceito Enade

Conceito Enade (Faixa)	NC_C **(Valor contínuo)**
1	$0 \leq NC_C < 0.945$
2	$0.945 \leq NC_C < 1.945$
3	$1.945 \leq NC_C < 2.945$
4	$2.945 \leq NC_C < 3.945$
5	$3.945 \leq NC_C \leq 5.0$

Fonte: Brasil, 2022.

Os dados disponibilizados pelo relatório do Enade fazem um comparativo de um determinado curso em relação às notas médias de outros três setores:

a) UF, que diz respeito à média geral de todos os cursos correspondentes naquela Unidade Federativa;

b) região, que diz respeito à média geral de todos os cursos correspondentes naquela região;

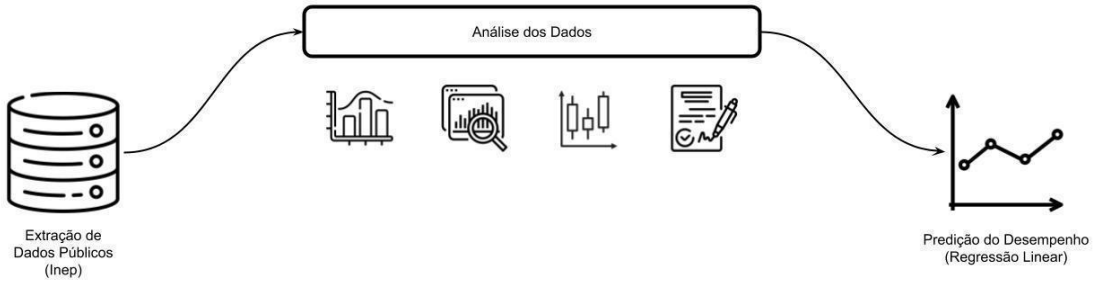
c) Brasil, que diz respeito à média geral de todos os cursos correspondentes no país.

A comparação de resultados realizada neste estudo se baseou nas provas do Enade aplicadas nos anos de 2017 e 2021. É necessário ressaltar que o Enade 2020 foi adiado para 2021 em decorrência da pandemia de COVID-19. Nos dados referentes ao Enade de 2021 é também encontrada uma seção com resultados de um questionário utilizado para conhecer a percepção dos alunos em relação à prova. Essas notas dizem respeito ao nível de dificuldade que os alunos sentiram ao responder as questões do exame e à correspondência entre o conteúdo abordado e o que foi coberto pelo plano pedagógico do curso, entre outros. Estas notas serão utilizadas para validar e justificar o desempenho dos alunos com base nos resultados que o modelo de predição irá fornecer.

3 Metodologia

Este trabalho tem como objetivo entender e analisar o desempenho dos alunos em aplicações do Enade por meio de um estudo qualiquantitativo no qual se determina a projeção de desempenho na próxima aplicação do exame. Para além das notas dos componentes da prova, existem as questões sobre percepção, que emitem dados sobre a dificuldade da prova, bem como a visão dos alunos em relação ao preparo do curso para suas respectivas formações. A Figura 1 ilustra as fases de desenvolvimento deste estudo.

Figura 1 – Fases de desenvolvimento do estudo



Fonte: Elaborada pelos autores, 2024.

Para esse estudo de caso, dois momentos do Enade foram analisados: 2017 e 2021. O curso de graduação objeto de estudo teve sua fundação em 2013 e, portanto, até o momento foram essas as aplicações do exame relacionadas a ele. Os dados utilizados foram extraídos de uma base pública do Inep, como está descrito detalhadamente na seção 3.1 e na seção 2.1, que falam sobre o conjunto de dados utilizado e sobre o funcionamento e propósito do Enade, respectivamente.

A primeira análise feita no estudo conta com um comparativo entre o desempenho ao longo das aplicações do exame. Devido à Lei Nº 13.709, de 14 de agosto de 2018, conhecida como Lei Geral de Proteção de Dados (LGPD) (Brasil, 2018), todas as métricas relacionadas a desempenho são tratadas como um conjunto, sendo ele uma turma de alunos. Nesse comparativo, as possíveis evoluções dos conjuntos são estudadas e observadas ao longo do tempo. Em uma próxima análise, as questões sobre percepção auxiliam no entendimento da coerência ou não dos resultados atingidos, bem como destacam possíveis variáveis que afetam diretamente o desempenho final dos alunos. Essas análises estatísticas são apresentadas na seção 4.

Por fim, um modelo de predição é desenvolvido para que, com base no desempenho de uma turma de estudantes em questão, seja possível projetar a nota do conceito Enade e o desempenho das turmas subsequentes. O propósito desse modelo, apresentado na seção 5, é projetar o andamento do curso.

3.1 Coleta de dados

A base de dados utilizada para este estudo é fornecida pelo Inep e consiste em microdados do Enade, que são os dados referentes ao desempenho de todos os alunos que realizaram o exame em um determinado ano, coletados através da aplicação da prova.

Neste trabalho, a meta é analisar o desempenho dos alunos de um curso de graduação cujas atividades iniciaram-se em 2013. Até o momento, as aplicações do Enade que tiveram participantes do curso datam de 2017 e 2021. A próxima aplicação ocorrerá no ano de 2024.

Os microdados são fornecidos em um conjunto de arquivos (Brasil, 2023a) que são separados por temáticas. Ao todo, são 42 arquivos referentes ao ano de 2017 e 43 referentes ao ano de 2021, uma vez que esse foi o primeiro ano a incluir um questionário sobre a percepção dos alunos em relação à prova, com ênfase no estudo durante o período da pandemia de COVID-19. Para o estudo apresentado neste trabalho, foram utilizados três arquivos:

a) arquivo nº 43, referente ao questionário de pandemia;
b) arquivo nº 1, que identifica o curso;
c) arquivo nº 3, referente às notas dos estudantes em todos os componentes do exame.

Devido à LGPD, os microdados do Enade são fornecidos de forma anonimizada. Sendo assim, por respeito à privacidade de cada estudante, e pela impossibilidade de identificação de cada aluno, não há como desenvolver uma análise multivariada dos estudantes.

4 Resultados

Através de uma manipulação do arquivo nº 1, foi possível descobrir o código de identificação do curso. A partir desse código foram extraídas as informações referentes aos alunos do curso em questão nos anos de 2017 e 2021. Com base na aplicação do filtro no arquivo nº 3 (de notas), descobriu-se quantos alunos realizaram a prova em cada um dos anos e, com base nos registros vazios, foi também possível descobrir o número de ausências em cada ano, como é mostrado na Tabela 2.

Tabela 2 – Quantidade de alunos e taxa de ausência em cada aplicação do Enade.

Ano	Número de Candidatos	Número de Ausências
2017	29	1
2021	59	12

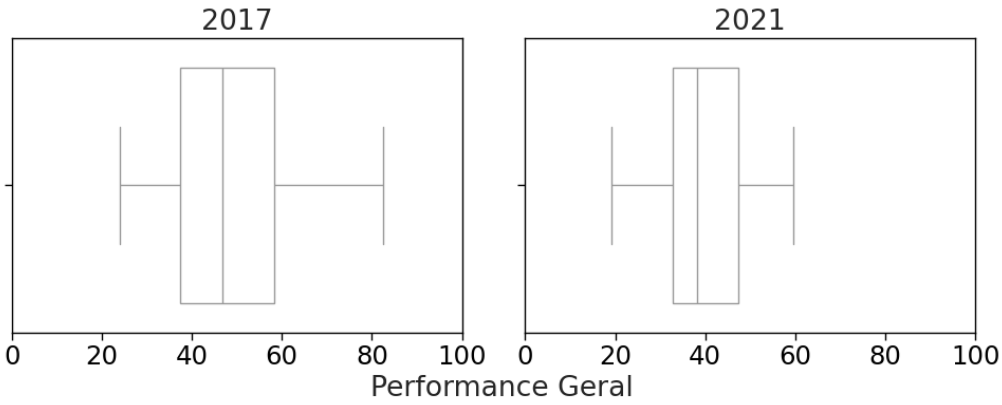
Fonte: Elaborada pelos autores, 2023.

Um dado para se analisar é o número significativo de ausências em 2021, uma vez que cerca de 20% dos alunos se ausentaram do exame em 2021. Segundo o § 5º do artigo 5º da Lei nº 10.861 de 2004 destacada no último Edital do Enade publicado (Brasil, 2023b), o Exame é um componente curricular obrigatório e, assim sendo, o aluno que se ausenta da aplicação do exame fica em caráter de irregularidade com sua respectiva IES. Isso pode acarretar inclusive o impedimento de sua colação de grau e, portanto, a ausência significativa de alunos no ano de 2021 é um problema importante, dadas suas consequências. A hipótese levantada é de que,

devido a pandemia, muitos candidatos se ausentaram da prova, que foi realizada presencialmente. É importante pontuar que, por ser o Enade obrigatório, o aluno que não realiza o exame precisa apresentar uma justificativa para a ausência.

Ao se analisar as notas de cada turma nos anos de aplicação do Enade, há uma leve diminuição no desempenho de 2021 quando comparado a 2017: enquanto a média da nota geral dos alunos em 2017 foi de aproximadamente 48.5 pontos, em 2021 foi de 39.3 pontos. Quando é feita a observação da distribuição de notas nos dois anos, por meio dos *boxplots* da Figura 2, percebe-se um deslocamento da média em 2021 para um valor menor em relação ao ano de 2017. Ainda, percebe-se que as notas do ano de 2021 estão distribuídas num intervalo menor, o que se traduz em uma maior homogeneidade entre as notas dos alunos.

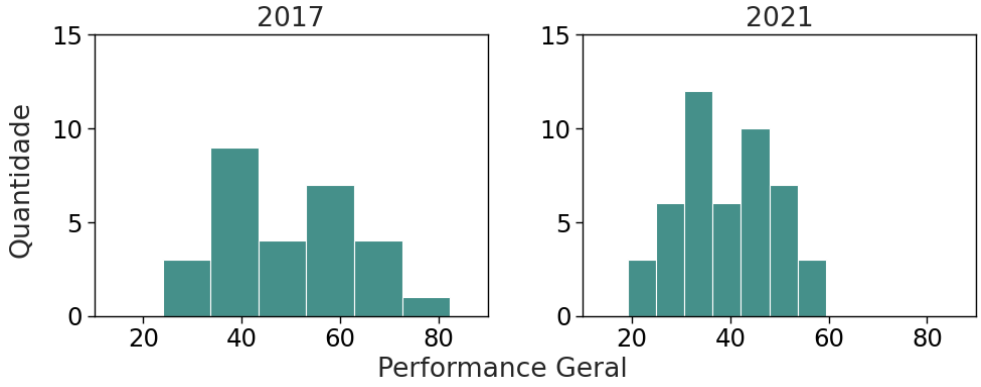
Figura 2 – Performance Geral



Fonte: Elaborada pelos autores, 2023.

No histograma da Figura 3, é possível notar que o comportamento da distribuição de notas é semelhante. Apesar do deslocamento da média, a quantidade de alunos que atingem as faixas de notas mais altas é proporcionalmente semelhante nos últimos dois anos, assim como a proporção de alunos que atingem as faixas de notas mais baixas.

Figura 3 – Distribuição das Notas

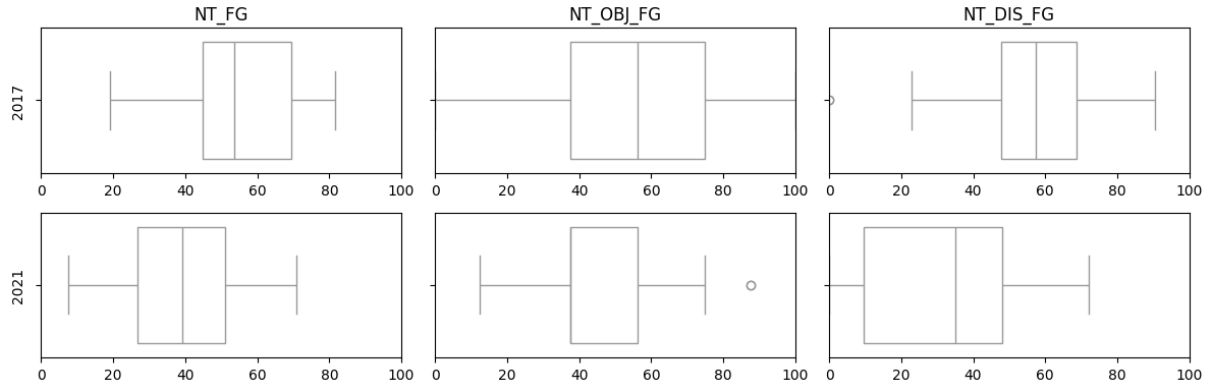


Fonte: Elaborada pelos autores, 2023.

Na Figura 4, consegue-se observar o desempenho dos alunos no componente de Formação Geral (FG). Os *boxplots* da primeira coluna representam as notas gerais do componente, ao passo que as colunas com prefixo NT_OBJ e NT_DIS representam as notas nas questões objetivas e nas questões dissertativas, respectivamente. Analisando os

componentes ano a ano, percebe-se que a queda mais significativa de desempenho está atrelada às questões de formação geral, principalmente nas questões discursivas. Quando nos atentamos aos valores numéricos, a mediana das notas objetivas de formação geral em 2021 equivale a nota do primeiro quartil do ano de 2017: 37.5. A queda é ainda maior nas questões dissertativas, sendo a mediana de 57.25 em 2017 e de 35.0 em 2021.

Figura 4 – Desempenho dos alunos no componente de Formação Geral

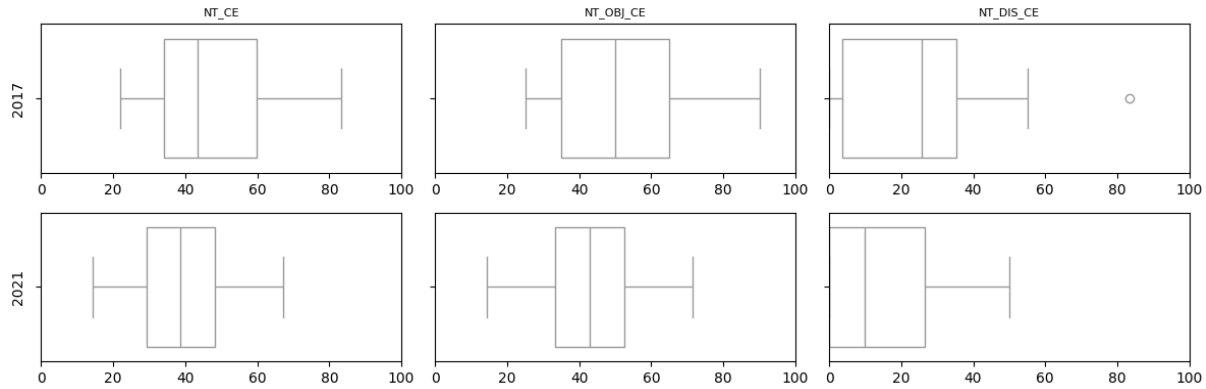


Fonte: Elaborada pelos autores, 2023.

No caso do componente de Conhecimentos Específicos (CE), ilustrado na Figura 5, a queda de desempenho é mais sutil do que nos componentes de formação geral. Ainda assim, é possível observar uma diferença significativa nas medianas das notas das questões objetivas (50.0 em 2017 e 42.9 em 2021) e principalmente das questões dissertativas, que são as que apresentam a maior queda de desempenho dos componentes: 25.85 em 2017 e 10.0 em 2021.

Na medida em que a relativa constância no desempenho dos componentes específicos possa significar uma consistência na formação do estudante, a queda brusca de desempenho no que tange às questões dissertativas pode ser um argumento contrário. Para uma análise mais minuciosa, seria interessante entender outros fatores que possam influenciar direta ou indiretamente o desempenho dos estudantes.

Figura 5 – Desempenho dos alunos no componente de Conhecimentos Específicos



Fonte: Elaborada pelos autores, 2023.

Tabela 3 - Questionário de Percepção da Pandemia.

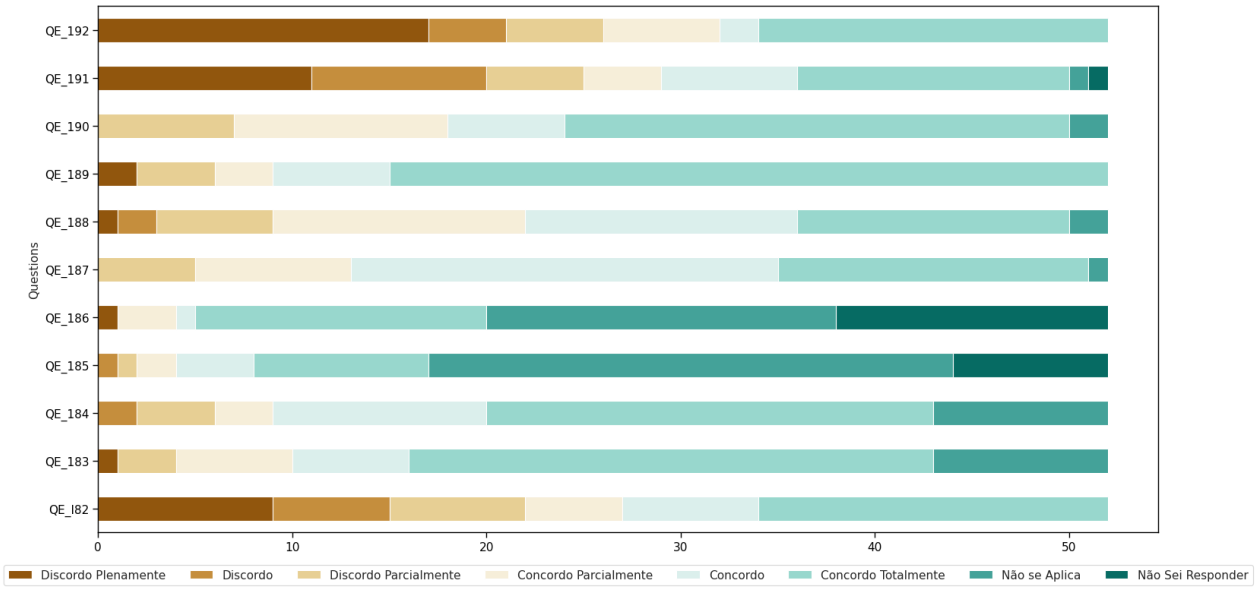
Cód. da Questão	Questão Apresentada
QE_I82	Com o início da pandemia, sua instituição passou rapidamente a ofertar aulas não presenciais.
QE_I83	Sua instituição ofereceu suporte para os estudantes superarem dificuldades tecnológicas de acesso às atividades não presenciais.
QE_I84	As referências bibliográficas (livros, artigos, textos) necessárias às aulas continuaram acessíveis após o início da pandemia.
QE_I85	As atividades de pesquisa e/ou extensão que você participava antes do início da pandemia continuaram sendo ofertadas.
QE_I86	As atividades de estágio supervisionado puderam ser realizadas ao longo da pandemia
QE_I87	Os professores demonstraram domínio dos recursos tecnológicos que passaram a ser utilizados nas aulas não presenciais.
QE_I88	A didática dos seus professores foi adequada para as aulas não presenciais.
QE_I89	Os recursos tecnológicos e o acesso à internet que você possuía no início da pandemia eram adequados para acompanhar as aulas não presenciais.
QE_I90	Durante a pandemia, você desenvolveu a capacidade de aprender por meio do ensino não presencial.
QE_I91	A implementação de aulas não presenciais e uso de tecnologias digitais decorrentes da pandemia prejudicaram seu processo formativo.
QE_I92	As dificuldades geradas pela pandemia para a continuidade dos estudos levaram você a pensar em trancar ou desistir do curso.

Fonte: Brasil, 2022.

Outro ponto de observação é o fato de que o Enade de 2021 foi aplicado dentro de um contexto pandêmico. Como destacam Menezes e Francisco (2020), muitas ferramentas de ensino tiveram de ser revisitadas para contemplar o isolamento social e ainda assim manter o conteúdo programático. Além disso, fatores como a própria estrutura domiciliar de cada aluno podem adicionar camadas e ruídos no processo de aprendizagem e consequentemente no desempenho final. Os questionários de percepção da prova (Tabela 3) podem ajudar a criar uma relação entre as notas e a situação em que os estudantes se encontravam, além de fornecer mais ferramentas e embasamento na identificação de aspectos que influenciam diretamente o desempenho dos alunos.

Na Figura 6, há a distribuição de respostas dos alunos a este questionário sobre percepção da pandemia. As questões que mais tiveram divergência de respostas foram questões relacionadas à percepção pessoal do aluno, como por exemplo as questões QE_I91 e QE_I92, o que mostra que nem todos os alunos se adaptaram bem ao modelo remoto. A questão QE_I82 também apresentou bastante divergência, uma vez que o conceito de “rapidamente” pode ser subjetivo.

Figura 6 - Respostas ao Questionário de Percepção da Pandemia.



Fonte: Elaborada pelos autores, 2023.

Neste estudo, um modelo de predição foi desenvolvido para que, com base no desempenho de uma turma de estudantes, fosse possível projetar a nota de desempenho geral dos alunos e, então, calcular a nota do conceito Enade para sua próxima aplicação. A próxima seção apresenta o modelo de regressão linear que foi utilizado para a predição das notas de desempenho dos alunos.

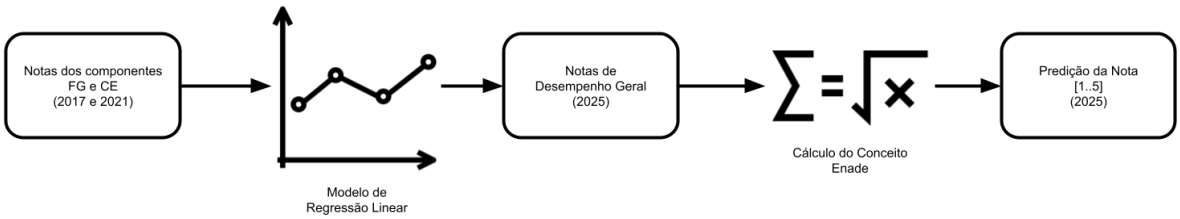
5 Modelo de predição

Um dos objetivos deste estudo é a criação de um modelo de predição que seja capaz de projetar a nota de conceito do Enade em sua próxima aplicação. Com o uso dos microdados foi possível fazer um *framework* que permitisse predizer essa nota com base no desempenho dos alunos nas aplicações do Enade nos anos anteriores.

Os microdados foram utilizados como forma de estudo para treinamento de um modelo que pudesse prever as notas futuras. Para que a abordagem fosse análoga ao cenário desejado, considerou-se que os dados coletados referentes aos anos de 2017 e 2021 foram tratados como notas de semestres consecutivos.

Tratando-se de um cálculo de conceito, a abordagem utilizada foi uma regressão linear. O treinamento do modelo consistiu em utilizar as notas de Formação Geral (NT_FG) e de Conhecimentos Específicos (NT_CE) para a predição da nota geral (NT_GER). Para se adequar ao objetivo final do modelo, o treinamento e a validação aconteceram da seguinte forma: com base nos dados fornecidos pelo Inep em 2017 e 2021, por meio das notas obtidas pelos candidatos desenvolveu-se um modelo de Regressão Linear, cujo treinamento tinha como objetivo prever as notas gerais de cada um dos anos, conforme o diagrama da Figura 7.

Figura 7 – Diagrama da estrutura do modelo de predição.



Fonte: Elaborado pelos autores, 2023.

A validação foi feita com as notas reais e, ao final, o modelo obteve as métricas exibidas na Tabela 4.

Tabela 4 - Métricas obtidas no modelo de Regressão Linear.

Métrica	**Valor**
MSE	0.004434
MAE	0.049896
RMSE	0.066585
R^2	0.999962

Fonte: Elaborado pelos autores, 2023.

Como as métricas apresentam resultados ótimos num pequeno conjunto de dados, os valores podem indicar *overfit*. Sendo assim, foram realizados estudos de validação cruzada a fim de se entender se o modelo era capaz de interpretar novos dados ou se estava apenas reproduzindo o comportamento dos dados iniciais. Utilizando a técnica de validação cruzada *k-folds* e uma validação de cinco *folds* (isto é, cinco conjuntos de dados), ao final obteve-se os seguintes resultados: [0.00078233, 0.0006637, 0.00112541, 0.00104607, 0.00069054]. Estes resultados se referem a *Mean Squared Error* (*MSE*)[3], por meio da qual se pode observar que o resultado da validação cruzada do modelo treinado apresenta uma diferença muito pequena entre o valor predito e o valor real, significando que o modelo é capaz de interpretar novos dados.

Mesmo apresentando resultados promissores, o *overfitting* do modelo não pode ser descartado uma vez que o conjunto de dados é pequeno. Em contrapartida, uma vez que esteja disponível um grande conjunto de dados, o modelo pode ser treinado e aperfeiçoado a fim de funcionar como um indicador de desempenho dos alunos do curso.

Após a fase de treinamento do modelo, foi então criado uma predição de quais seriam as possíveis notas dos alunos na próxima aplicação do exame. A predição gerou dados referentes às Notas de Formação Geral (NT_FG) e de Conhecimentos Específicos (CE). A partir das notas preditas foi criada uma terceira coluna, que é referente a Nota Geral (NT_GER) e basicamente consiste na seguinte equação:

$$NT\_GER = 0.25 * NT\_FG + 0.75 * NT\_CE \quad \text{(II)}$$

[3] *Mean Squared Error (MSE)* é uma métrica que consiste em calcular a distância média entre o valor predito e o valor real, numa escala quadrática. Sendo assim, quanto menor for nosso MSE, melhor nosso modelo se comporta, significando que a diferença entre o valor predito e o valor real é muito pequena.

Com as novas notas é possível então calcular o novo conceito Enade utilizando a equação (I). Com as notas preditas, o conceito Enade projetado para a próxima aplicação é de três. Comparado com o valor quatro relativo aos anos de 2017 e 2021, observou-se uma queda de desempenho. Logo, supõe-se que essa situação pode ter sido influenciada pelos resultados referentes ao ano de 2021, que, possivelmente, foram afetados pelo cenário de pandemia, com representativa quantidade de ausências no exame, o que também pode ter sido um indicador da queda no desempenho geral dos alunos.

6 Conclusão

Este trabalho apresentou um modelo de predição utilizando regressão linear para projetar a nota de conceito do Enade. Para isso, foi aplicada a técnica de *LA* em microdados do exame de desempenho de anos anteriores. O modelo de regressão desenvolvido pode ser uma ferramenta útil ao projetar o desempenho da turma na próxima aplicação do exame, e com isso possíveis ações podem ser tomadas visando a melhoria do desempenho geral.

É importante lembrar que o *LA* tem suas limitações e restrições, como a dificuldade de coletar dados precisos e confiáveis e a necessidade de garantir a privacidade dos alunos. Outro problema é a falta de clareza em relação à definição do objeto de estudo em *LA*: como ressaltado por Johnson *et al*. (2016), esse pode variar desde o desempenho individual do aluno até a análise de políticas educacionais em nível macro. Essa falta de clareza pode tornar difícil a identificação dos limites e alcances da área, e levar a interpretações equivocadas dos resultados obtidos.

A falta de generalização dos resultados também se apresenta como um obstáculo. Leitner *et al*. (2019) mencionam que muitas destas análises são baseadas em dados de uma única instituição de ensino, dificultando assim a generalização. Além disso, mesmo quando os resultados são generalizáveis, pode ser difícil determinar a causalidade das relações encontradas, devido a fatores externos que não foram considerados.

No cenário estudado, a falta de possibilidade de cruzar os dados para entender outros fatores se tornou um limitante significativo. Devido a anonimização dos dados fornecidos pelo Inep, análises que poderiam identificar características relacionadas a gênero, faixa etária, ou até mesmo condição social ficaram impossibilitadas, uma vez que não há como identificar a qual aluno exatamente pertence cada resposta ou nota. Em um cenário em que os dados fornecidos sejam atrelados à identificação do aluno, análises mais profundas poderiam ser feitas a fim de, por exemplo, entender se há algum obstáculo de aprendizagem relacionado a grupos específicos por meio de medidas de correlação entre as variáveis analisadas. Com estes dados em mãos, seria possível montar programas que atendessem grupos ou turmas específicas, visto que o modelo de análise é capaz de destacar pontos importantes do desempenho dos alunos.

Outro número a ser observado foi a significativa queda de desempenho dos alunos em relação à aplicação do Enade em um ano de pandemia. Talvez com um volume maior de dados históricos, isto é, com os resultados das próximas aplicações, seja possível haver um melhor entendimento de como a pandemia pode ter ou não afetado o processo de aprendizagem dos alunos. Com base nesses resultados e traçando paralelos com o estudo de Menezes e Francisco (2020) poderiam ser implementadas ações que estudam os efeitos da pandemia e do ensino a distância a longo prazo em cenários como o apresentado neste trabalho.

Em trabalhos futuros pretende-se analisar os dados dos simulados realizados semestralmente com alunos do curso cujos dados de Enade serviram para compor este estudo, para que seja possível comparar o desempenho da turma com o resultado previsto. Esse tipo de aplicação pode trazer melhorias como as observadas no estudo de Morais *et al*. (2017), podendo ainda ser parte de um *framework* do estudo da qualidade do curso. Portanto, o modelo pode ser usado como ferramenta de apoio, para que ações de possíveis melhorias no processo de aprendizagem dos alunos sejam tomadas de forma preditiva. Todas essas ações seriam embasadas em métricas que podem ser adicionadas para tornar o estudo cada vez mais amplo e completo. O modelo poderia ainda ser introduzido em outros cursos da instituição que também façam a aplicação periódica dos simulados, ou ainda servir de base para analisar o desempenho dos alunos através de outras avaliações e exames periódicos.

Referências

BRANDT, J. Z.; TEJEDO-ROMERO, F.; ARAUJO, J. F. F. E. Fatores influenciadores do desempenho acadêmico na graduação em administração pública. **Educação e Pesquisa**, São Paulo, v. 46, e202500, 2020. DOI 10.1590/S1678-4634202046202500. Disponível em https://www.scielo.br/j/ep/a/RF8cFBPnKjNqYPJkLjZVpHg/?lang=pt&format=html. Acesso em 5 maio 2024.

BRASIL. **Descrição da Metodologia de Cálculo do Conceito Enade**. Ministério da Educação, Inep. Brasil. 2022. Disponível em: https://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2019/nota_tecnica_n_7_2022_CGCQES_DAES_metodologia_calculo_conceito_enade_2021.pdf. Acesso: em 15 de nov. de 2023.

BRASIL. **Exame Nacional de Desempenho dos Estudantes (Enade).** Ministério da Educação, Inep, 2024. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em 5 de maio de 2024.

BRASIL. **Microdados do Enade.** Ministério da Educação, Inep, 2023a. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade. Acesso em: 15 maio 2024.

BRASIL. **Edital Nº 37, de 25 de maio de 2023**. Exame Nacional de Desempenho dos Estudantes (ENADE), 2023b. Diário Oficial da União. Disponível em: https://www.in.gov.br/en/web/dou/-/edital-n-37-de-25-de-maio-de-2023-486214440. Acesso em: 15 de nov. de 2023.

BRASIL. **Dicionário de Arquivos e Variáveis.** Ministério da Educação, Inep. 2022. Disponível em:https://download.inep.gov.br/microdados/microdados_enade_2021.zip. Acesso em: 3 de jul. de 2024.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 15 de nov. de 2023.

FERREIRA, Sérgio André; ANDRADE, António. Desenhar e implementar um sistema de *learning analytics* no ensino superior. **Gestão e Desenvolvimento**, n. 21, 123-146. 2013. DOI

10.7559/gestaoedesenvolvimento.2013.244. Disponível em: https://journals.ucp.pt/index.php/gestaoedesenvolvimento/article/view/244. Acesso em: 15 maio 2024.

FOSTER, E.; SIDDLE, R. The effectiveness of learning analytics for identifying at-risk students in higher education. **Assessment & Evaluation in Higher Education**, v. 45, n. 6, p. 842-854, 2020. DOI: 10.1080/02602938.2019.1682118. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/02602938.2019.1682118. Acesso em: 15 maio 2024.

JOHNSON, L. et al. **NMC horizon report: 2016 higher education edition**. 13. ed. Austin, Texas: The New Media Consortium, 2016. 56 p.

KOVANOVIĆ, V. et al. What public media reveals about MOOCs: A systematic analysis of news reports. **British Journal of Educational Technology**, v. 46, n. 3, p. 510-527, 2015. DOI: 10.1111/bjet.12277. Disponível em: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12277. Acesso em: 15 maio 2024.

LEITNER, P.; KHALIL, M.; EBNER, M. Learning analytics in higher education—a literature review. *In*: PEÑA-AYALA, A. (ed). **Learning analytics:** Fundaments, applications, and trends: A view of the current state of the art to enhance E-learning. v. 94, 2017. p. 1-23.

LUZ, J. W. P.; REHFELDT, M. J. H.; SCHORR, M. C. Revisão Sistemática da Literatura sobre o uso de *Learning Analytics* no ensino de programação. **Revista Novas Tecnologias na Educação**, v. 19, n. 2, p. 203-212, 2021. DOI: 10.22456/1679-1916.121207. Disponível em: https://seer.ufrgs.br/renote/article/view/121207. Acesso em: 15 maio 2024.

MENEZES, S. K. O.; FRANCISCO, D. J. Educação em tempos de pandemia: aspectos afetivos e sociais no processo de ensino e aprendizagem. **Revista Brasileira de Informática na Educação**, v. 28, p. 985-1012, 2020. DOI: 10.5753/rbie.2020.28.0.985. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/v28p985. Acesso em: 15 maio 2024.

MORAIS, C.; ALVES, P.; MIRANDA, L. *Learning analytics* na obtenção de indicadores de desempenho no ensino superior. *In*: **12th Iberian Conference on Information Systems and Technologies (CISTI)**. Lisboa: IEEE, 2017. p. 1-6. Disponível em: https://bibliotecadigital.ipb.pt/handle/10198/22550. Acesso em: 15 maio 2024.

PORTAL, C. **Estratégias para minimizar a evasão na educação a distância: o uso de um sistema de mineração de dados educacionais e *learning analytics***. 2016. Dissertação (Mestrado em Educação, Desenvolvimento e Tecnologias) - Universidade do Vale do Rio dos Sinos. Disponível em: https://www.abed.org.br/congresso2015/anais/pdf/BD_317.pdf. Acesso em: 15 maio 2024.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)**, v. 40, n. 6, p. 601-618, 2010. DOI: 10.1109/TSMCC.2010.2053532. Disponível em: https://ieeexplore.ieee.org/abstract/document/5524021. Acesso em: 15 maio 2024.

SCHENEIDER, T. F. *et al*. Análise do nível de maturidade na adoção de *learning analytics* em instituições de ensino superior brasileiras. *In*: **Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil**. SBC, 2022.

SELWYN, N.; FACER, K. (ed.). **The politics of education and technology:** Conflicts, controversies, and connections. New York: Palgrave Macmillan, 2013. 250 p.

SIEMENS, G.; LONG, P. Penetrating the fog: Analytics in learning and education. **EDUCAUSE Review**, v. 46, n. 5, p. 30-32, 2011. Disponível em: https://eric.ed.gov/?id=EJ950794. Acesso em: 15 maio 2024.

Recebido em março de 2024

Aprovado em junho de 2024