ANÁLISES EXPLORATÓRIA E MULTIVARIADA COMO FERRAMENTAS PARA DIAGNOSTICAR O CURSO DE ENGENHARIA DE PRODUÇÃO NO BRASIL

Josicleyton Azevedo dos Santos
Universidade Federal Rural do Semi-Árido (Ufersa)
Av. Francisco Mota, 572. Bairro Costa e Silva Mossoró-RN, CEP: 59.625-900
josicleyton.academico@gmail.com

Miriam Karla Rocha
Universidade Federal Rural do Semi-Árido (Ufersa)
Av. Francisco Mota, 572. Bairro Costa e Silva Mossoró-RN, CEP: 59.625-900
miriam.rocha@ufersa.edu.br

RESUMO

Este trabalho aborda o uso das técnicas estatísticas de análises exploratória e multivariada de dados, dentro do contexto educacional. Tem como objetivo uma melhor percepção em torno das características que compõem o atual ensino em Engenharia de Produção no Brasil. Quanto à metodologia, este estudo se caracteriza como uma pesquisa exploratória, através da coleta, tratamento, modelagem e análise dos dados presentes no Exame Nacional de Desempenho dos Estudantes (Enade) do ano de 2019, do curso de Engenharia de Produção, aplicando as ferramentas de análise multivariada de dados, por meio da linguagem de programação Python. Os resultados levaram a conclusões pertinentes do atual cenário de ensino de Engenharia de Produção, de acordo com a categoria administrativa, organização acadêmica, modalidade de ensino e região, bem como as características mais relevantes de ensino. Pretende-se assim auxiliar a todos os interessados na qualidade, melhoria e gestão do ensino superior, especialmente, em Engenharia de Produção.

PALAVRAS CHAVE: Ensino superior; Regressão Múltipla; Clusterização.

EST&MP – Estatística e Modelos Probabilísticos; AdP&ED – PO na Administração Pública e Educação

ABSTRACT

This study approaches statistical techniques of exploratory and multivariate data analysis within the educational context. Its objective is to raise the characteristics that make up the current teaching in Industrial Engineering in Brazil. As for the methodology, this study is characterized as exploratory research, through the collection, treatment, modeling, and analysis of data present in the Exame Nacional de Desempenho dos Estudantes (Enade) of the year 2019 of the Production Engineering course, applying the tools exploratory and multivariate analysis of data, through the Python programming language. The results led to relevant conclusions regarding the current scenario of Industrial Engineering teaching, according to the administrative category, academic organization, teaching modality, region, and the most relevant teaching characteristics. It is thus intended to help all those interested in the quality, improvement, and management of higher education, especially in Industrial Engineering.

KEYWORDS. Undergraduate school; Multiple Regression; Clustering.
EST&MP – Statistics and Probabilistic Models; AdP&ED – PO in Public Administration and Education

1. Introdução

Atualmente, devido ao enorme avanço tecnológico, há uma grande quantidade de dados disponíveis, dos mais diversos tipos e fontes, desde os mais simples, aos extremamente complexos e confusos. Uma pesquisa recente revelou que até 2025, cerca de 150 trilhões de gigabytes de dados necessitarão de análise, ao passo que 95% das empresas encontram algum tipo de dificuldade em gerenciá-los [KULKARNI, 2019]. Desta forma, pode-se observar que nem sempre este acesso a dados em massa é transformado em significado, ou seja, em informação e posteriormente, o que é mais importante, em conhecimento [PICHETTI et al., 2020].

Para tanto, segundo Crespo (2002), a estatística tem proporcionado técnicas para "coleta, organização, descrição, resumo, análise e interpretação de dados", advindos de experimentos ou observações de estudos. Dentro de suas diversas áreas, as análises exploratória e multivariada de dados, aliadas a programação, destacam-se pela possibilidade de extrair informações úteis de um grande volume de dados [FÁVERO; BELFIORE, 2015].

Além do mais, um dos ramos de aplicações mais recorrentes em estudos de dados é a educação, principalmente, a mineração de dados educacionais, em que a partir da extração e obtenção de dados brutos, é possível adquirir e fornecer elementos importantes que auxiliem na análise e tomada de ações ou resolução de problemas no âmbito educacional [BORGES, 2017].

Neste contexto, o Exame Nacional de Desempenho dos Estudantes (Enade) é uma ferramenta de avaliação da qualidade dos cursos e instituições de ensino superior no Brasil e, de acordo com Canan e Eloy (2016), "tem impactado na gestão dos cursos que, em busca de resultados positivos, acaba gerando estratégias de gestão que contemplem as atividades vinculadas ao preparo dos alunos para a prova", proporcionado um estímulo a uma melhoria contínua no ensino.

Entre os cursos mais relevantes avaliados no Enade, evidencia-se a Engenharia de Produção (EP), já que das engenharias ela é vista como a mais sistêmica, envolvendo diversas áreas de formação [CARDOSO, 2007]. Outro fator considerável, segundo os dados disponíveis no Cadastro Nacional de Cursos e Instituições de Educação Superior [MEC, s.d.], é que até o final de 2021 houveram 1128 cursos de EP em atividade em todo o Brasil, ao passo que até o final de 2005 (ano que o curso passou a ser avaliado pelo Enade) eram apenas 251 cursos, o que representa um aumento de aproximadamente 350 % de cursos.

Os dados brutos disponibilizados nos resultados das edições do Enade [INEP, s.d.], uma vez analisados, poderiam possibilitar a visualização, avaliação e um possível diagnóstico da qualidade de ensino dos cursos e instituições, bem como identificar quais os fatores são mais determinantes neste objetivo. Alguns estudos já utilizaram os dados do Enade no curso de EP [STURM, 2015; DE ASSUNÇÃO E FILHO, 2019; RIBEIRO et al., 2019; NASCIMENTO, 2020. MORAES et al., 2021], no entanto, as técnicas de análises exploratória e multivariada de dados foram pouco utilizadas e, ao mesmo tempo, foram exploradas às edições mais antigas do Enade e pouco se encontra da edição de 2019, a mais atual sobre EP até a presente data, que contém algumas variáveis e indicadores a mais em relação às outras edições.

Desde modo, o presente trabalho tem como objetivo compreender o desenvolvimento e as características que circundam o cenário do ensino superior em Engenharia de Produção no Brasil, por meio da aplicação de técnicas estatísticas de análises exploratória e multivariada dos dados deste curso no Exame Nacional de Desempenho dos Estudantes de 2019, com o auxílio de linguagem de programação. Pretende-se assim auxiliar instituições, docentes, discentes, administradores, pesquisadores e demais interessados no conhecimento, melhoria, qualidade e gestão dos sistemas educacionais na formação dos profissionais do futuro.

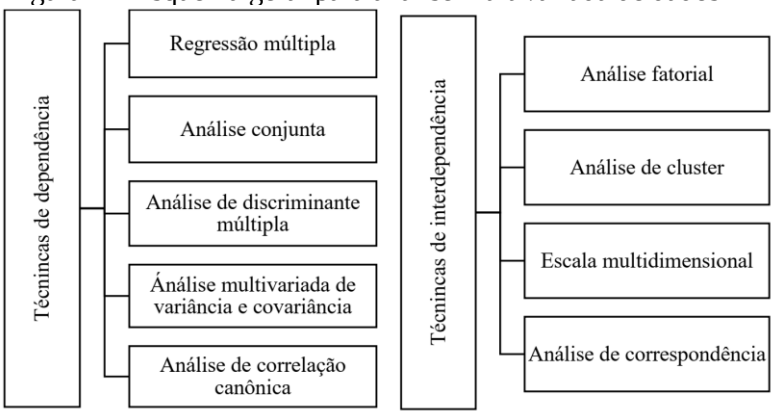
2. Análise multivariada de dados

A Análise Multivariada de Dados (AMD), segundo Loesh e Hoeltgebaum [2012], trata-se de qualquer técnica estatística para análise simultânea de mais de duas variáveis presentes nas

observações. É interessante notar que a AMD também se utiliza de conceitos e técnicas da AED, especialmente, no trabalho de identificar possíveis relações entre as variáveis.

De acordo com Hair Jr. et al. [2009], na prática, a AMD se utiliza de dois tipos de técnicas de análise dependendo do relacionamento entre os dados e sua natureza: dependência e interdependência. No primeiro grupo de técnicas, há algum tipo relação de dependência ou independência entre as variáveis, de maneira que é possível descrevê-las ou prevê-las. Já no segundo, não se estabelece relacionamentos entre as variáveis, o que permite que elas sejam analisadas concomitantemente. O esquema representado na Figura 1 apresenta as principais técnicas [FERREIRA et al., 2021].

Figura 1 – Esquema geral para análise multivariada de dados



Fonte: Baseada em Ferreira et al. (2021)

O presente trabalho irá abordar a regressão múltipla, como técnica de dependência e, finalmente, a análise de cluster, como técnica de interdependência.

2.1. Regressão múltipla

Segundo Barbetta et al. [2010], a regressão múltipla, também conhecida por regressão linear múltipla, possibilita predizer uma variável dependente e identificar a relação com as suas variáveis independentes. Desta forma, o valor esperado da variável dependente é dado de acordo com a Equação (1):

$$E\{Y\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j \quad (1)$$

Em que Y é a variável dependente e x_1, x_2 até x_j são as variáveis independentes; j é o número de variáveis independentes; β_1, β_2 até β_j são os parâmetros estimados pelo modelo que corresponde aos coeficientes de suas respectivas variáveis aleatórias; e, por fim, α é o parâmetro estimado pelo modelo que equivale a intercessão entre os planos das variáveis.

É importante mencionar o R-quadrado ajustado, que corresponde ao coeficiente de determinação, uma medida que vai de 0 a 1, e mostra o quanto a equação está ajustada aos dados de interesse bem como o quão "forte" é a relação linear entre as variáveis dependente e independentes [TRIOLA, 2017].

Ainda é importante definir testes para a verificação de significância do modelo. Uma forma prática é calcular a razão F na análise de variância (ANOVA), representa por f, que equivale a razão entre o modelo e seu erro. Na prática, estabelecido um intervalo de confiança, caso o valor-p associado à f for extremamente pequeno, as variáveis independentes em questão são significativas para a variância de Y [BARBETTA et al., 2010].

Também é necessário realizar testes estatísticos individuais para cada um dos coeficientes β_1, β_2 a β_j da equação. Novamente, umas das formas de se analisar se dá pela observação dos valores-p de cada coeficiente: caso estes forem inferiores ao nível de significância, os coeficientes são significativos para a explicação do modelo. Caso contrário, os que estiverem acima, devem ser desconsiderados do modelo [BARBETTA et al., 2010].

2.2. Análise de cluster

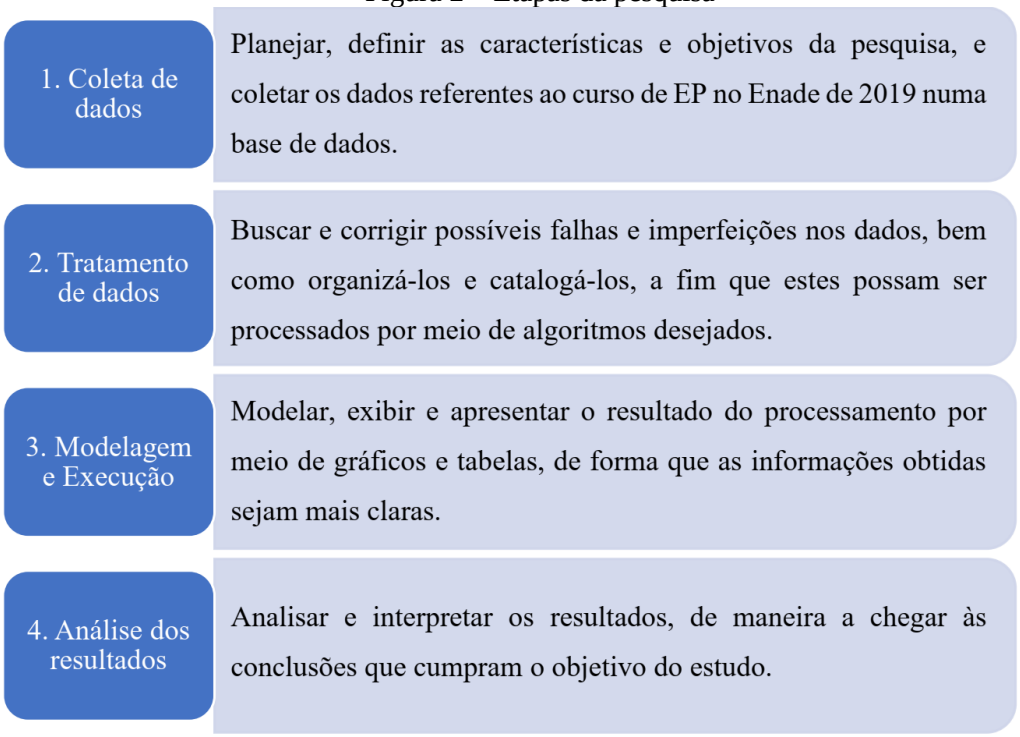
De acordo com Fávero e Belfiore [2015], a análise de cluster, também conhecida por análise de agrupamentos ou conglomerados, busca-se características próximas entre os dados e variáveis das observações em estudo, permitindo a criação de grupos, chamados de clusters, que são semelhantes internamente e dissimilares entre si. Na prática, existem inúmeras técnicas de clusterização. As mais conhecidas se classificam em dois métodos: os hierárquicos e os não hierárquicos. Cada método tem a sua medida de distância ou de semelhança que será utilizada na maneira de aglomerar os grupos. Com a definição dos grupos se torna possível analisá-los e comparar os resultados [LOESCH; HOELTGEBAUM, 2012].

Acontece que alguns métodos que utilizam medidas de distância, como a euclidiana, costumam funcionar, na maioria dos casos, apenas para variáveis quantitativas. Quando se lida com variáveis numéricas e categóricas, normalmente, outras abordagens devem ser consideradas [HUNT; JORGENSEN, 2011]. Nestes casos, uma das possibilidades é a utilização do algoritmo K-medoides (K-medoids, em inglês), em que é possível a aplicação com variáveis heterogêneas [DE OLIVEIRA et al., 2020], com menor vulnerabilidade à dados discrepantes ou outliers e possibilitando o uso de outras medidas de distância [DE ASSIS; DE SOUZA, 2011].

3. Método de pesquisa

O presente trabalho se caracteriza como uma pesquisa aplicada, exploratória, documental e, ao mesmo tempo, um estudo de caso [PRODANOV; FREITAS, 2013], em que se busca compreender as características, relações, condições e padrões referentes aos cursos e instituições que ofertam a graduação em EP, a partir da análise de dados presentes na edição de 2019 do Enade disponibilizados na página do Inep: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade/resultados (INEP, s.d.), aplicando as ferramentas e técnicas estáticas da AED e AMD. As etapas da pesquisa, podem ser visualizadas na Figura 2.

Figura 2 – Etapas da pesquisa



As variáveis que foram consideradas para a análise foram divididas em quantitativas e qualitativas. Dentre as quantitativas, encontram-se os quatro indicadores: o Conceito Enade; o Conceito Preliminar do Curso (CPC); o Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD); e, por fim, o Índice Geral de Cursos Avaliados da Instituição (IGC). Estes indicadores podem ser classificados em faixas, de 1 a 5 (MEC, 2016); no entanto, os cursos que não possuem o indicador específico são identificados no lugar do indicador com a sigla “SC”.

Já as variáveis qualitativas ou categóricas foram agrupadas em: (i) categoria administrativa: especiais (ESP), privadas com fins lucrativos (PCFL), públicas estaduais (PE), públicas federais (PF), públicas municipais (PM) e privadas sem fins lucrativos (PSFL); (ii) Organização acadêmica: Centros Federais de Educação Tecnológica ou apenas Centros Federais (CEFET), Centros Universitários (CU), Faculdades (FA), Institutos Federais de Educação Tecnológica ou apenas Institutos Federais (IF) e Universidades (UN); (iii) Modalidade de Ensino: presencial ou à distância; (iv) região: Centro-Oeste (CO), Nordeste (NE), Norte (NO), Sudeste (SE) e Sul (SU).

Ressalta-se ainda, que na parte de processamento, especificamente na AMD, foram considerados apenas os cursos que possuem todos os indicadores ao mesmo tempo, descartando os que faltavam algum ou mais indicadores. Ainda foram desconsiderados os cursos especiais quanto à categoria administrativa, já que de todos os analisados, esta classificação correspondia a apenas um curso, podendo gerar assim alguma incongruência no modelo. Desta forma, na AMD foram analisados 654 cursos, enquanto que na AMD foram apenas 607. Além disso, as variáveis quantitativas foram normalizadas pelo Min-Max e as qualitativas foram transformadas em variáveis dummy.

Por fim, todas as análises, tratamentos, manipulações e resultados foram feitos utilizando a linguagem de programação Python, devido a sua portabilidade, código livre, simplicidade com robustez e a facilidade do aprendizado [BANIN, 2018]. O ambiente virtual de trabalho foi o Google Colab, que escreve e executa códigos em Python na nuvem, sem nenhuma configuração necessária.

4. Resultados

A seguir são apresentados os resultados da AMD do curso de EP do Enade de 2019, aplicando as técnicas de regressão múltipla e análise de cluster.

4.1 Regressão múltipla

Foram analisados 607 cursos, desconsiderando os dados do tipo "SC" ou especiais quanto à categoria administrativa. O resultado da regressão múltipla, considerando como variável dependente o Conceito Enade e as demais variáveis analisadas como independentes, e um nível de significância de 2,5 % pode ser visto na Figura 3.

A Figura 3 revela algumas informações importantes. A primeira delas é o R-quadrado ajustado que neste caso corresponde à 0,723, o que significa que aproximadamente 72 % da variação do Conceito Enade é explicada pela variação das demais variáveis, o que equivale a um valor considerável para a análise.

Em relação aos testes de significâncias, antes de tudo, o valor-p associado à f é de $2{,}49 \cdot 10^{-155}$, um valor extremamente pequeno, o que implica que as variáveis analisadas são significativas na explicação da variância do Conceito Enade.

Porém, ao analisar os valores-p associados aos os testes estatísticos individuais para cada um dos coeficientes, nota-se que apenas as variáveis numéricas CPC, IDD e IGC, e as variáveis categóricas "privadas com fins lucrativos", "privadas sem fins lucrativos", "pública federal", "presencial", "Norte" e "Sudeste" estão abaixo do nível de significância. Em outras palavras, somente estas variáveis são significativas na explicação do modelo e devem ser consideradas.

Figura 3 – Resultado da regressão múltipla considerando todas as variáveis

Dep. Variable:	ConceitoENADE	R-squared:	0.731
Model:	OLS	Adj. R-squared:	0.723
Method:	Least Squares	F-statistic:	94.22
Date:	Wed, 15 Jun 2022	Prob (F-statistic):	2.49e-155
Time:	16:13:36	Log-Likelihood:	-393.95
No. Observations:	607	AIC:	823.9
Df Residuals:	589	BIC:	903.3
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.2839	0.087	-3.255	0.001	-0.455	-0.113
CPC	0.7145	0.064	11.176	0.000	0.589	0.840
IDD	0.1548	0.033	4.627	0.000	0.089	0.221
IGC	0.1906	0.064	2.959	0.003	0.064	0.317
CentroFederal	-0.1493	0.185	-0.806	0.420	-0.513	0.214
CentroUnivesitário	-0.0487	0.073	-0.669	0.504	-0.192	0.094
Faculdade	0.0181	0.071	0.253	0.800	-0.122	0.158
InstitutoFederal	-0.1157	0.168	-0.688	0.491	-0.446	0.214
Universidade	0.0117	0.065	0.181	0.857	-0.116	0.139
Especial	-0.2862	0.402	-0.712	0.476	-1.075	0.503
PrivadaComFinsLucrativos	-0.4024	0.086	-4.689	0.000	-0.571	-0.234
PrivadaSemFinsLucrativos	-0.3591	0.085	-4.209	0.000	-0.527	-0.192
PúblicaEstadual	0.2068	0.115	1.793	0.074	-0.020	0.433
PúblicaFederal	0.5730	0.103	5.582	0.000	0.371	0.775
PúblicaMunicipal	-0.0161	0.180	-0.089	0.929	-0.369	0.337
Presencial	-0.1512	0.065	-2.339	0.020	-0.278	-0.024
ÀDistância	-0.1328	0.082	-1.624	0.105	-0.293	0.028
CentroOeste	-0.0902	0.074	-1.214	0.225	-0.236	0.056
Nordeste	-0.0752	0.051	-1.480	0.139	-0.175	0.025
Norte	-0.3304	0.086	-3.822	0.000	-0.500	-0.161
Sudeste	0.1384	0.038	3.609	0.000	0.063	0.214
Sul	0.0735	0.050	1.466	0.143	-0.025	0.172

Ao aplicar novamente a regressão apenas com estas variáveis, mesmo assim, notou-se que a categoria presencial apresentou um valor-p de 0,894, ou seja, acima do nível de significância, como mostra a Figura 4.

Figura 4 – Resultado da segunda regressão múltipla

Dep. Variable:	ConceitoENADE	R-squared:	0.726
Model:	OLS	Adj. R-squared:	0.722
Method:	Least Squares	F-statistic:	175.6
Date:	Wed, 15 Jun 2022	Prob (F-statistic):	2.38e-161
Time:	16:13:36	Log-Likelihood:	-399.90
No. Observations:	607	AIC:	819.8
Df Residuals:	597	BIC:	863.9
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3102	0.201	-1.545	0.123	-0.704	0.084
CPC	0.7091	0.063	11.232	0.000	0.585	0.833
IDD	0.1529	0.033	4.592	0.000	0.088	0.218
IGC	0.2156	0.061	3.535	0.000	0.096	0.335
PrivadaComFinsLucrativos	-0.5716	0.090	-6.364	0.000	-0.748	-0.395
PrivadaSemFinsLucrativos	-0.5190	0.090	-5.772	0.000	-0.696	-0.342
PúblicaFederal	0.3782	0.098	3.857	0.000	0.186	0.571
Presencial	-0.0156	0.117	-0.134	0.894	-0.245	0.214
Norte	-0.3177	0.105	-3.020	0.003	-0.524	-0.111
Sudeste	0.1352	0.040	3.393	0.001	0.057	0.213

Desta forma, uma terceira regressão foi feita, desconsiderando a variável presencial. O resultado pode ser visto na Figura 5.

Figura 5 – Resultado da terceira regressão múltipla

Dep. Variable:	ConceitoENADE	R-squared:	0.726
Model:	OLS	Adj. R-squared:	0.722
Method:	Least Squares	F-statistic:	197.9
Date:	Wed, 15 Jun 2022	Prob (F-statistic):	1.65e-162
Time:	16:13:37	Log-Likelihood:	-399.91
No. Observations:	607	AIC:	817.8
Df Residuals:	598	BIC:	857.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3252	0.166	-1.956	0.051	-0.652	0.001
CPC	0.7090	0.063	11.241	0.000	0.585	0.833
IDD	0.1526	0.033	4.596	0.000	0.087	0.218
IGC	0.2159	0.061	3.545	0.000	0.096	0.336
PrivadaComFinsLucrativos	-0.5713	0.090	-6.368	0.000	-0.748	-0.395
PrivadaSemFinsLucrativos	-0.5192	0.090	-5.779	0.000	-0.696	-0.343
PúblicaFederal	0.3782	0.098	3.860	0.000	0.186	0.571
Norte	-0.3181	0.105	-3.028	0.003	-0.524	-0.112
Sudeste	0.1354	0.040	3.401	0.001	0.057	0.214

As informações contidas na Figura 5 confirmam a significância do modelo e de cada uma das variáveis consideradas. O R-quadrado ajustado corresponde a $0{,}722$ e o valor-p associado à f é cerca de $1{,}65 \cdot 10^{-162}$, novamente próximo a zero. E os valores-p dos testes individuais de cada coeficiente são todos abaixo do nível de significância.

Assim, o valor esperado do Conceito Enade de acordo com a Equação (2) e o modelo de regressão realizado é:

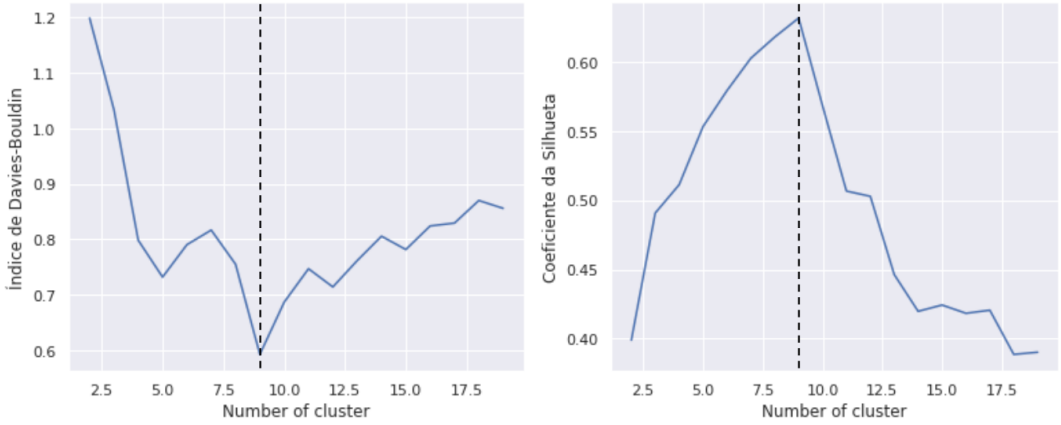
$$E\{ConceitoEnade\} = -0{,}3252 + 0{,}7090 \cdot CPC + 0{,}1526 \cdot IDD + 0{,}2159 \cdot IGC - 0{,}5713 \cdot PrivadaComFinsLucrativos - 0{,}5192 \cdot PrivadaSemFinsLucrativos + 0{,}3782 \cdot PúblicaFederal - 0{,}3181 \cdot Norte + 0{,}1354 \cdot Sudeste \quad (2)$$

Pode-se estimar o valor do Conceito Enade, desde que o resultado, os demais indicadores sejam expressos no formato de normalização Min-Max, e as variáveis categóricas como variáveis *dummy*.

4.1 Análise de cluster

Considerando apenas as variáveis significativas do modelo de regressão, antes de aplicar o algoritmo do *K*-medoides, foi necessário definir o número de *clusters* pelo índice de Davies-Bouldin e o coeficiente de Silhueta, com o auxílio do Figura 6.

Figura 6 – Resultados para o índice de Davies-Bouldin e o coeficiente da Silhueta



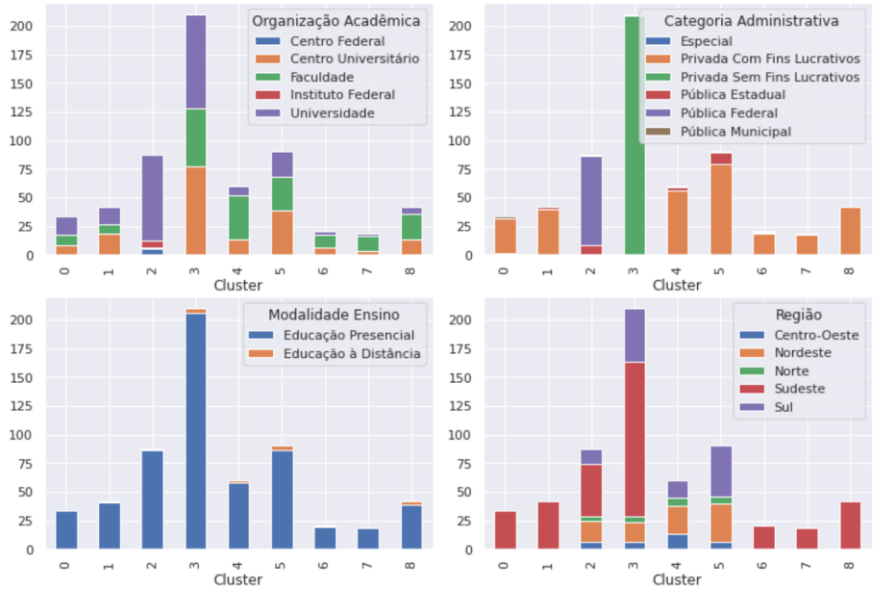
Pelo Figura 6 observar-se que o número ideal de *clusters* para o caso em estudo é 9. Desta forma, foi aplicado o método do *K*-medoides para o agrupamento dos dados em 9 grupos. Vale ressaltar que na linguagem de programação utilizada, a numeração destes *clusters* foi de 0 a 8, e das 607 observações foram elencadas de 0 a 606. Para a análise destes grupos de cursos de EP no Enade de 2019, antes de tudo, convém identificar a quantidade de cursos de cada *cluster* bem como as principais medidas descritivas estatísticas do Conceito Enade, como é mostrado na Tabela 1.

Tabela 1 – Quantidade de cursos e medidas descritivas dos clusters no Conceito Enade

Cluster	Quantidade de cursos (%)	Medidas descritivas do Conceito Enade				
		Média	D. Padrão	Mediana	Máximo	Mínimo
0	34 (5,60 %)	2,75	0,65	2,72	4,05	1,58
1	42 (6,92 %)	2,30	0,45	2,36	3,32	1,43
2	88 (14,50 %)	3,61	0,79	3,66	5,00	1,79
3	210 (34,60 %)	2,12	0,73	2,08	4,18	0,41
4	60 (9,88 %)	1,38	0,40	1,41	2,39	0,50
5	91 (14,99 %)	2,25	0,55	2,27	3,61	1,25
6	21 (3,46 %)	1,39	0,46	1,36	2,12	0,21
7	19 (3,13 %)	1,90	0,41	1,82	2,64	1,16
8	42 (6,92 %)	1,75	0,31	1,79	2,39	1,09

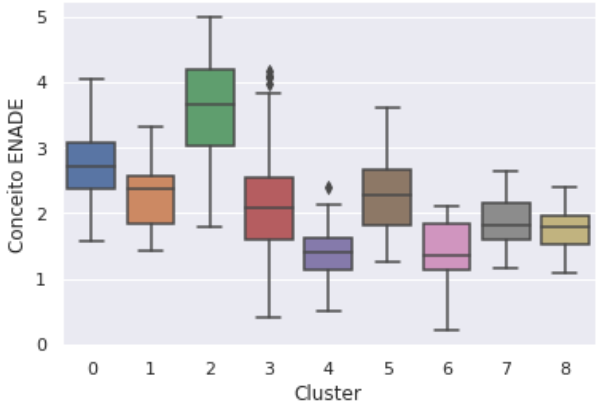
Outro detalhe importante é a visualização da proporção de *clusters* em relação a cada variável categórica original do problema mostrado na Figura 7.

Figura 7 – Quantidade de cursos por clusters em cada variável categórica



Por fim, pode-se observar a concentração ou dispersão dos dados de cada grupo em relação ao Conceito Enade por meio do *boxplot*, Figura 8.

Figura 8 – Boxplot do Conceito Enade de acordo com o cluster



Apresentado este resultado, é possível descrever as características de cada *cluster*. Vale lembrar novamente que todos os dados são na sua maioria de cursos presenciais e, portanto, o mesmo vale para os grupos descritos.

- *Cluster* 0: formado por 34 cursos, com média e mediana aproximadas no Conceito Enade de 2,75 e 2,72, respectivamente, e distribuídos entre Universidades, Centros Federais e Faculdades, de sua maioria privadas com fins lucrativos da região Sudeste.
- *Cluster* 1: formado por 42 cursos, com média e mediana aproximadas no Conceito Enade de 2,30 e 2,36, respectivamente, e distribuídos entre Universidades, Centros Federais e Faculdades, de sua maioria privadas com fins lucrativos da região Sudeste.
- *Cluster* 2: formado por 88 cursos, com média e mediana aproximadas no Conceito Enade de 3,61 e 3,66, respectivamente, e na sua maioria de Universidades, públicas federais das regiões Sudeste e Nordeste.

- *Cluster* 3: formado por 210 cursos, com média e mediana aproximadas no Conceito Enade de 2,12 e 2,08, respectivamente, e distribuídos entre Universidades, Centros Federais e Faculdades, na sua maioria privadas sem fins lucrativos das regiões Sudeste e Sul.
- *Cluster* 4: formado por 60 cursos, com média e mediana aproximadas no Conceito Enade de 1,38 e 1,41, respectivamente, e na sua maioria de Faculdades, privadas com fins lucrativos das regiões Nordeste e Sul.
- *Cluster* 5: formado por 91 cursos, com média e mediana aproximadas no Conceito Enade de 2,25 e 2,27, respectivamente, e distribuídos entre Universidades, Centros Federais e Faculdades, na sua maioria privadas com fins lucrativos das regiões Sul e Nordeste.
- *Cluster* 6: formado por 21 cursos, com média e mediana aproximadas no Conceito Enade de 1,39 e 1,36, respectivamente, e distribuídos entre Universidades, Centros Federais e Faculdades, na sua maioria privadas com fins lucrativos da região Sudeste.
- *Cluster* 7: formado por 19 cursos, com média e mediana aproximadas no Conceito Enade de 1,90 e 1,82, respectivamente, e na sua maioria de Faculdades, privadas com fins lucrativos da região Sudeste.
- *Cluster* 8: formado por 42 cursos, com média e mediana aproximadas no Conceito Enade de 1,75 e 1,79, respectivamente, e distribuídos entre Universidades, Centros Federais e Faculdades, na sua maioria privadas com fins lucrativos da região Sudeste.

Como em todos os grupos considerados, a média e a mediana estão relativamente próximas, pode-se observar que uma congruência nos dados.

Dos 9 *clusters* analisados se destaca o grupo 2, já que este possui a maior média do conceito Enade de 3,61, sendo o único agrupamento com média acima de 3, que corresponde ao mínimo satisfatório e, portanto, em tese, representaria os cursos com as melhores condições de qualidade em seu ensino.

Já os *clusters* 4 e 6 possuem as menores médias, correspondentes a 1,38 e 1,39, respectivamente, e teoricamente, retratam os cursos com as menos favoráveis circunstâncias de qualidade.

5. Considerações Finais

Diante do que foi exposto ao longo deste trabalho, é possível perceber o quanto as ferramentas de análises exploratória e multivariada de dados apresentadas, aliadas com a programação, são pertinentes para se obter informações valiosas, antes não identificadas apenas com dados brutos.

Os resultados obtidos a partir do Enade de 2019 permitiram chegar em conclusões importantes acerca das realidades que circundam a qualidade e desempenho no ensino dos cursos e instituições de Engenharia de Produção.

Por um lado, a Análise Exploratória de Dados permitiu a visualização da distribuição dos indicadores de qualidade, além das especificidades quanto as categorias administrativas, organizações acadêmicas, modalidades de ensino e regiões. Pode-se revelar aspectos relevantes como o acesso ao ensino, condições socioeconômicas e as capacidades do ensino público ou privado, presencial ou à distância.

Por outro lado, a Análise Multivariada de Dados, por meio dos modelos de regressão e de clusterização, permitiu observar quais os fatores são mais importantes na determinação do Conceito Enade, além de identificar a formação de grupos de acordo com esses fatores, bem como suas características.

Sugere-se ainda, para trabalhos futuros, de acordo com as próximas edições do Enade uma comparação dos cursos e instituições de Engenharia de Produção antes e depois da pandemia do Covid-19, ou ainda comparar com algum outro método, como o Stepwise, próprio para seleção de variáveis.

Desta forma, espera-se que este trabalho sirva na cooperação e benefício da comunidade acadêmica e científica e, sobretudo, facilite numa melhor organização e incentivo no aprendizado e gestão de todos aqueles que buscam a qualidade do ensino na formação dos futuros profissionais que atuarão na sociedade.

Referências

BANIN, Sérgio Luiz. **Python 3: Conceitos e Aplicações: Uma Abordagem Didática**. 1. ed. São Paulo: Érica, 2018.

BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antonio Cezar. **Estatística para Cursos de Engenharia e Informática**. 3. ed. São Paulo: Atlas, 2010.

BORGES, Vanessa Araújo. **Definição de um Modelo de Referência de Dados Educacionais para a Descoberta de Conhecimento**. Tese de Doutorado. Universidade de São Paulo, 2017.

CANAN, Silvia Regina; ELOY, Vanessa Taís. **Políticas de Avaliação em Larga Escala: O Enade Interfere na Gestão dos Cursos?**. Práxis Educativa, v. 11, n. 3, p. 621-640, 2016.

CARDOSO, Patrícia Alcântara. **Interdisciplinaridade na Prática: A Experiência da Aplicação do Software Preactor como Ferramenta de Integração Curricular no Curso de Graduação em Engenharia de Produção**. Revista Gestão Industrial, v.03, n.01, p.103-110, 2007.

CRESPO, Antônio Arnot. **Estatística Fácil**. 18. ed. São Paulo: Saraiva, 2002.

DE ASSIS, Elaine Cristina; DE SOUZA, Rafaela M. C. R. **A K-Medoids Clustering Algorithm For Mixed Feature-Type Symbolic Data**. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, p. 527–531, 2011.

DE ASSUNÇÃO, Marina Heil; FILHO, Roberto Ivo da Rocha Lima. **Ensino de Engenharia de Produção no Brasil: O Crescimento a Partir dos Anos 2000 e a Qualidade dos Cursos**. XXXIX Encontro Nacional de Engenharia de Produção. Santos, 2019.

DE OLIVEIRA, Fernanda Robes; KLEINA, Mariana; MARQUES, Marcos Augusto Mendes; GAYER, Jessika Alvares Coppi Arruda; TAMACHIRO, Thiago Shoji Obi. **Clusterização de Clientes: Um Modelo Utilizando Variáveis Categóricas e Numéricas**. X Congresso Brasileiro de Engenharia de Produção: ConBRepro, Abepro, 2020.

FÁVERO, Luis Paulo; BELFIORE, Patrícia. **Análise de Dados, Técnicas Multivariadas Exploratórias com SPSS ® e STATA ®**. 1. ed. Rio de Janeiro: Elsevier, 2015.

FERREIRA, Rafael Gastão Coimbra; DE MIRANDA, Leandro Botelho Alves; PINTO, Rafael Albuquerque; PESSUTTO, Lucas Rafael Costella; PEREIRA, Mariana Araújo; DE ANDRADE, Ana Luiza Cerchiari; MARQUES, Leonardo Torres. **Preparação e Análise Exploratória de Dados**. Porto Alegre: SAGAH, 2021.

HAIR JR., Joseph F.; BLACK, William C.; BABIN, Barry J.; ANDERSON, Rolph E.; TATHAM, Ronald L. **Análise Multivariada de Dados**. 6. ed. Porto Alegre: Bookman, 2009.

HUNT, Lynette Anne; JORGENSEN, Murray A. **Clustering Mixed Data**. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 1, n. 4, p. 352–361, 2011.

INEP. Instituto Nacional de Pesquisas Educacionais Anísio Teixeira. **Exame Nacional de Desempenho dos Estudantes: Resultados**. s.d. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade/resultados>. Acesso em: 07 de dezembro de 2021.

KULKARNI, Rohit. **Big Data Goes Big**. Forbes. 2019. Disponível em: https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/. Acesso em: 20 de fevereiro de 2022.

LOESCH, Claudio; HOELTGEBAUM, Marianne. **Métodos Estatísticos Multivariados**. 1 ed. São Paulo: Saraiva, 2012.

MEC. Ministério da Educação. **Cadastro Nacional de Cursos e Instituições de Educação Superior Cadastro e-MEC**. s.d. Disponível em: <https://emec.mec.gov.br/>. Acesso em: 07 de dezembro de 2021.

MEC. Ministério da Educação. **Portaria Normativa nº 23, de 20 de dezembro de 2016**. Altera dispositivos da Portaria Normativa MEC nº 40, de 12 de dezembro de 2007, republicada em 2010, para efeitos imediatos nos processos de cálculo e divulgação dos Indicadores de Qualidade da Educação Superior, a partir da edição de 2015. Brasília, 2016. Disponível em: < https://abmes.org.br/arquivos/legislacoes/portaria-normativa-n-23-2016-12-20.pdf>. Acesso em: 07 de dezembro de 2021.

MORAES, Renato de Oliveira; KISTE, Gerson; PEREIRA, Priscila Maria Santiago. **Cursos de Engenharia de Produção no Brasil – Uma Tipologia Baseada nos Dados do INEP**. XLI Encontro Nacional de Engenharia de Produção. Foz do Iguaçu, 2021.

NASCIMENTO, Francisco Eduardo de Lima. **Desempenho dos Estudantes dos Cursos de Engenharia de Produção no ENADE dos Anos 2014 e 2017: Uma Análise Estatística do Comportamento da Nota Média Geral**. XL Encontro Nacional de Engenharia de Produção. Foz do Iguaçu, 2020.

PICHETTI, Roni Francisco Vida; CORTES, Ednilson da Silva; PAIXÃO, Vanessa Stangherlin Machado. **Banco de Dados**. Porto Alegre: SAGAH, 2020.

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani Cesar. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Novo Hamburgo: Feevale, 2013.

RIBEIRO, Isabelly Santos; DOS ANJOS, Evandro Souza; BARBOSA, Ava Santana; SANTANNA, Angelo Marcio Oliveira. **Avaliação dos Cursos de Engenharia de Produção das Universidades Federais do Nordeste Baseada no Método Multicritério Promethee**. XXXIX Encontro Nacional de Engenharia de Produção. Santos, 2019.

TRIOLA, Mario F. **Introdução à Estatística**. 12. ed. Rio de Janeiro: LTC, 2017.