



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
CENTRO DE ENGENHARIAS
DEPARTAMENTO DE ENGENHARIA E CIÊNCIAS AMBIENTAIS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

JOSICLEYTON AZEVEDO DOS SANTOS

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

MOSSORÓ

2022

JOSICLEYTON AZEVEDO DOS SANTOS

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

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia de Produção.

Orientadora: Profa. Dra. Miriam Karla Rocha

MOSSORÓ

2022

JOSICLEYTON AZEVEDO DOS SANTOS

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

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia de Produção.

Defendida em: 14 / 06 / 2 022.

BANCA EXAMINADORA

**Miriam
Karla Rocha**

Assinado digitalmente
por Miriam Karla Rocha
Data: 2022-06-20 11:
49:49

Profa. Dra. Miriam Karla Rocha (UFERSA)
Presidente

**ANDRE LUIZ SENA DA
ROCHA:04986338455**

Assinado de forma digital por
ANDRE LUIZ SENA DA
ROCHA:04986338455
Dados: 2022.06.22 12:54:00 -03'00'

Prof. Me. André Luiz Sena da Rocha (UFERSA)
Membro Examinador

**TADILA LARISSA
GOMES MOURA**

Assinado de forma digital por
TADILA LARISSA GOMES
MOURA
Dados: 2022.06.22 14:40:41
-03'00'

Administradora Tádila Larissa Gomes Moura
Membro Examinador

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

D237a Dos Santos, Josicleyton Azevedo.
Análises Exploratória e Multivariada Como
Ferramentas Para Diagnosticar o Curso de
Engenharia de Produção no Brasil / Josicleyton
Azevedo Dos Santos. - 2022.
45 f. : il.

Orientadora: Miriam Karla Rocha.
Monografia (graduação) - Universidade Federal
Rural do Semi-árido, Curso de Engenharia de
Produção, 2022.

1. Engenharia de Produção.. 2. Análise
Exploratória. 3. Análise Multivariada. 4.
Regressão Múltipla. 5. Clusterização. I. Rocha,
Miriam Karla, orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade
com AACR2 e os dados fornecidos pelo(a) autor(a).
Biblioteca Campus Mossoró / Setor de Informação e Referência
Bibliotecária: Keina Cristina Santos Sousa e Silva
CRB: 15/120

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

AGRADECIMENTOS

Antes de tudo, à Deus que conduziu a minha história até aqui. À Ele dedico este trabalho e toda a minha ocupação profissional. Agradeço, igualmente, à Sua Santa Mãe, a Virgem Maria, cuja proteção confio toda a minha vida.

Aos meus pais Josivaldo e Socorro, por acreditarem em mim e me apoiarem em todos os momentos, mas, acima de tudo, por me concederem o dom da vida.

À minha irmã Josivânia, pela presença e amparo em qualquer situação.

Aos meus demais familiares, tios e tias, primos e primas e, de modo especial, aos meus avôs e avós paternos e maternos, que já se encontram na eternidade junto de Deus.

A todos os docentes e demais profissionais da UFERSA, que me acompanharam ao longo desta jornada, especialmente, a minha orientadora a Profa. Dra. Miriam Karla, pela atenção, tempo e apoio prestados.

Enfim, agradeço à todas as pessoas, amigos, colegas, conhecidos, de fora ou de dentro da Universidade que, de alguma maneira, contribuíram para mais uma etapa da minha vida.

A todos vocês dedico este trabalho.

Cumpra o pequeno dever de cada momento;
faz o que deves e está no que fazes.

Josemaria Escrivá.

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 das análises exploratória e 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: Engenharia de Produção; Análise Exploratória; Análise Multivariada; Regressão Múltipla; Clusterização.

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: Industrial Engineering; Exploratory Analysis; Multivariate Analysis; Multiple Regression; Clustering.

LISTA DE FIGURAS

Figura 1 – Esquema geral para análise exploratória de dados.....	20
Figura 2 – Esquema geral para análise multivariada de dados.....	21
Figura 3 – Etapas da pesquisa.....	24
Figura 4 – Resultado da regressão múltipla considerando todas as variáveis	34
Figura 5 – Resultado da segunda regressão múltipla	35
Figura 6 – Resultado da terceira regressão múltipla.....	36
Figura 7 – Resultado parcial da clusterização	37

LISTA DE GRÁFICOS

Gráfico 1 – Quantidade de cursos por categoria administrativa por faixa de indicador	26
Gráfico 2 – Boxplots dos indicadores por categoria administrativa.....	27
Gráfico 3 – Quantidade de cursos por organização acadêmica por faixa de indicador.....	28
Gráfico 4 – Boxplots dos indicadores por organização acadêmica	29
Gráfico 5 – Quantidade de cursos por modalidade de ensino por faixa de indicador	30
Gráfico 6 – Boxplots dos indicadores por modalidade de ensino	31
Gráfico 7 – Quantidade cursos por região por faixa de indicador.....	32
Gráfico 8 – Boxplots dos indicadores por região	33
Gráfico 9 – Resultados para o índice de Davies-Bouldin e o coeficiente da Silhueta	37
Gráfico 10 – Quantidade de cursos por clusters em cada variável categórica	38
Gráfico 11 – Boxplot do Conceito Enade de acordo com o cluster	39

LISTA DE TABELAS

Tabela 1 – Parâmetros de conversão dos indicadores discretos e contínuos do Enade.....	17
Tabela 2 – Quantidade de cursos e medidas descritivas dos clusters no Conceito Enade	38

LISTA DE ABREVIATURAS E SIGLAS

AED	Análise Exploratória de Dados
AMD	Análise Multivariada de Dados
ANOVA	Análise de Variância
CEFET	Centros Federais de Educação Tecnológica ou Centros Federais
CO	Centro-Oeste
CPC	Conceito Preliminar do Curso
CU	Centros Universitários
Enade	Exame Nacional de Desempenho dos Estudantes
EP	Engenharia de Produção
ESP	Especiais
FA	Faculdades
IDD	Indicador de Diferença entre os Desempenhos Observado e Esperado
IF	Institutos Federais de Educação Tecnológica ou Institutos Federais
IGC	Índice Geral de Cursos Avaliados da Instituição
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
NE	Nordeste
NO	Norte
PMA	Particionamento sobre os Medoides
PCFL	Privadas com fins lucrativos
PE	Públicas estaduais
PF	Públicas federais
PM	Públicas municipais
PSFL	Privadas sem fins lucrativos
SC	Sem Conceito ou Indicador
SE	Sudeste
Sinaes	Sistema Nacional de Avaliação da Educação Superior
SU	Sul
UN	Universidades

SUMÁRIO

1 INTRODUÇÃO	14
2 FUNDAMENTAÇÃO TEÓRICA.....	16
2.1 EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES	16
2.1.1 Indicadores de Qualidade do Enade.....	16
2.1.2 A Engenharia de Produção e o Enade.....	17
2.2 ANÁLISE EXPLORATÓRIA DE DADOS.....	18
2.2.1 Banco de dados e variáveis	18
2.2.2 Processo de análise exploratória de dados	19
2.3 ANÁLISE MULTIVARIADA DE DADOS	20
2.3.1 Regressão múltipla	21
2.3.2 Análise de cluster.....	22
2.3.2.1 Método K-Medoides	23
3 MÉTODOS.....	24
4 RESULTADOS	26
4.1 ANÁLISE EXPLORATÓRIA DE DADOS DE ENGENHARIA DE PRODUÇÃO NO ENADE 2019.....	26
4.1.1 Categoria administrativa.....	26
4.1.2 Organização acadêmica	28
4.1.3 Modalidade de ensino.....	29
4.1.4 Região	31
4.2 ANÁLISE MULTIVARIADA DE DADOS DE ENGENHARIA DE PRODUÇÃO NO ENADE 2019.....	34
4.2.1 Regressão múltipla	34
4.2.2 Análise de cluster.....	36
5 CONCLUSÃO.....	41
REFERÊNCIAS	42

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 FUNDAMENTAÇÃO TEÓRICA

Antes de apresentar as análises, métodos e resultados do presente trabalho se faz necessário abordar alguns conceitos importantes para a referenciação e aplicação destes posteriormente.

2.1 EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES

De acordo com a Lei nº 10.861, de 14 de abril de 2004, o Sistema Nacional de Avaliação da Educação Superior (Sinaes) foi instituído como uma forma de priorizar a qualidade da educação superior no Brasil, através da análise e avaliação de ensino, pesquisa, extensão, docência, além de outros aspectos, inclusive sociais, em relação aos cursos, instituições e estudantes (BRASIL, 2004).

Dentro dos processos avaliativos do Sinaes, destaca-se o Exame Nacional de Desempenho de Estudantes (Enade), realizado desde 2004 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), órgão do Ministério da Educação (MEC), com o objetivo de aferir o desempenho dos estudantes a partir das diretrizes curriculares aliadas com as habilidades e competências de cada curso (BRASIL, 2004).

A sua realização é baseada em um ciclo avaliativo, em que instituições e cursos superiores, periodicamente (geralmente, a cada três anos), são avaliados de acordo com os componentes de formação geral e específicos de cada área de conhecimento e eixo tecnológico (BRASIL, 2017). Ressalta-se ainda a obrigatoriedade da inscrição para os ingressantes e concluintes dos cursos de bacharelado e superiores de tecnologia participantes das áreas de avaliação do exame em cada edição (INEP, s.d.).

2.1.1 Indicadores de qualidade do Enade

Uma das principais vantagens do Enade é a divulgação e acompanhamento dos indicadores de qualidade, que se tornaram instrumentos de avaliação e desempenho da qualidade do ensino superior no Brasil, nos respectivos cursos e instituições examinadas.

Nos últimos anos, entre os resultados fornecidos, evidenciam-se 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).

O principal indicador, o Conceito Enade, corresponde a Nota dos Concluintes no Enade (INEP, 2020a), dada pela Equação (1) como:

$$NC_c = 0,25 \cdot NP_{FG_c} + 0,75 \cdot NP_{CE_c} \quad (1)$$

Em que:

- NC_c é a nota dos concluintes no Enade do curso de graduação c ;
- NP_{FG_c} é a nota padronizada de Formação Geral do exame do curso de graduação c ;
- NP_{CE_c} é a nota padronizada do Componente Específico do exame do curso de graduação c .

As especificações, bem como a base de cálculo e outros detalhes de cada um destes indicadores podem ser encontrados nas suas respectivas normas técnicas (INEP, 2020a, 2020b, 2020c, 2020d).

Além disso, é importante ressaltar que os indicadores são expressos em escala contínua e em cinco níveis ou faixas, de 1 a 5, sendo 3 considerado como satisfatório mínimo, e 5 corresponde ao melhor (MEC, 2016), como mostra a Tabela 1.

Tabela 1 – Parâmetros de conversão dos indicadores discretos e contínuos do Enade

Faixas dos indicadores na forma discreta	Faixas dos indicadores na forma contínua
1	$0,000 \leq \text{indicador} < 0,945$
2	$0,945 \leq \text{indicador} < 1,945$
3	$1,945 \leq \text{indicador} < 2,945$
4	$2,945 \leq \text{indicador} < 3,945$
5	$3,945 \leq \text{indicador} < 5,000$

Fonte: INEP (2019)

2.1.2 A Engenharia de Produção e o Enade

De acordo com Batalha (2013), a Engenharia de Produção (EP) lida com o projeto, desenvolvimento e concretização de sistemas integrados para a produção de bens e serviços, como a administração de pessoas, materiais, informações, recursos etc. envolvidos, de forma

econômica, ética e cultural, tendo como base os conhecimentos e habilidades específicas da profissão.

A participação de cursos de EP no Enade, seja “plenos” ou com ênfases em outras áreas (FURLANETTO, 2006), até o presente momento foram nos anos de 2005, 2008, 2011, 2014, 2017 e 2019. Há de se observar que esta participação foi acompanhando a evolução e crescimento do curso ao longo dos anos, já que ao analisar os resultados do Enade disponíveis na página do Inep (INEP, s.d.), observa-se um aumento de aproximadamente 320 % de cursos participantes dos anos de 2005 a 2019, passando de 156 a 654, respectivamente.

2.2 ANÁLISE EXPLORATÓRIA DE DADOS

A seguir são apresentados alguns conceitos e ferramentas importantes a respeito da Análise Exploratória de Dados para o presente trabalho.

2.2.1 Banco de dados e variáveis

Ao longo dos anos, o ser humano se defrontou com a necessidade de registrar as informações mais importantes ao seu redor, de maneira a suprir um objetivo específico. O avanço tecnológico permitiu que estas informações fossem guardadas com uma maior usabilidade, eficiência e segurança por meio de bancos de dados, em que é possível extrair algum significado, ou ainda retratar características da realidade (ALVES, 2014).

Pelo estabelecimento de alguma relação entre os dados, torna-se possível tratá-los, compartilhá-los ou fazer aplicações com programas, de acordo o interesse desejado (PICHETTI et al., 2020). Pode-se entender este conjunto de dados como uma informação referente a uma determinada variável, ou seja, uma característica ou atributo.

As variáveis, por sua vez, podem expressar qualidades (variável qualitativa), em que importa a ordem (ordinal) ou não (nominal). ou ainda podem representar quantidades (variável quantitativa), advindas de um processo de contagem (discreta) ou de uma medição (contínua) de dados (VIEIRA, 2019).

Por exemplo, a escolaridade ou classe social de um indivíduo, poderiam ser classificadas como variáveis qualitativas ordinais; já a sua profissão, sexo ou religião, representariam variáveis qualitativas nominais. Por outro lado, a idade, o número de alunos ou o número de cursos, por expressarem quantidades inteiras, seriam variáveis quantitativas discretas; enquanto

que a renda, o peso ou a altura, na maioria das vezes, são expressas por números reais não inteiros, portanto, expressariam variáveis quantitativas contínuas.

Há de se observar ainda que as variáveis podem ser chamadas de categóricas, já que, em alguns casos, sua escala de mensuração corresponde a um conjunto de categorias (AGRESTI; FINLAY, 2012).

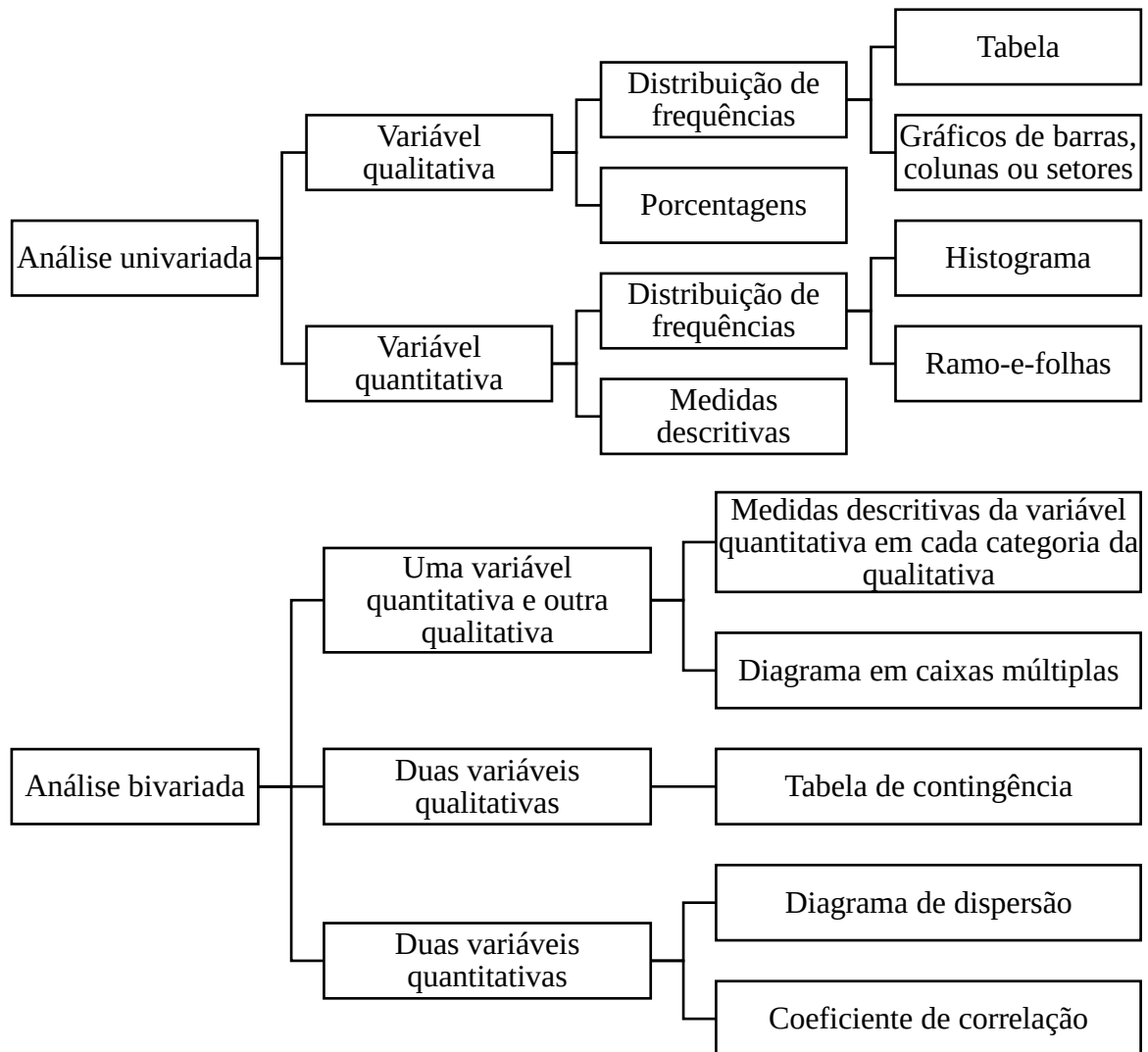
2.2.2 Processo de análise exploratória de dados

Dentro da estatística, a Análise Exploratória de Dados (AED), de acordo com Ferreira et al. (2021), busca averiguar os dados para uma possível aplicação estatística, de forma a compreendê-los melhor, bem como identificar as relações entre as suas variáveis. Na prática, busca-se organizar, resumir e aplicar cálculos importantes nos dados para posteriormente apresentá-los por meio de uma visualização e interpretação adequadas

Dentre as inúmeras técnicas utilizadas pela AED, segundo Barbetta et al. (2010), destaca-se a utilização de recursos como tabelas, gráficos e medidas, a fim que estes revelem informações que estavam implícitas na forma bruta dos dados, mas que agora se tornam evidentes e claras.

Há de se atentar que quando se trata de uma enorme quantidade de dados, torna-se mais conveniente explorar as associações entre as variáveis, dependendo de suas naturezas (BARBETTA et al., 2010), como é possível observar na Figura 1.

Figura 1 – Esquema geral para análise exploratória de dados



Fonte: Adaptado de Barbetta et al. (2010)

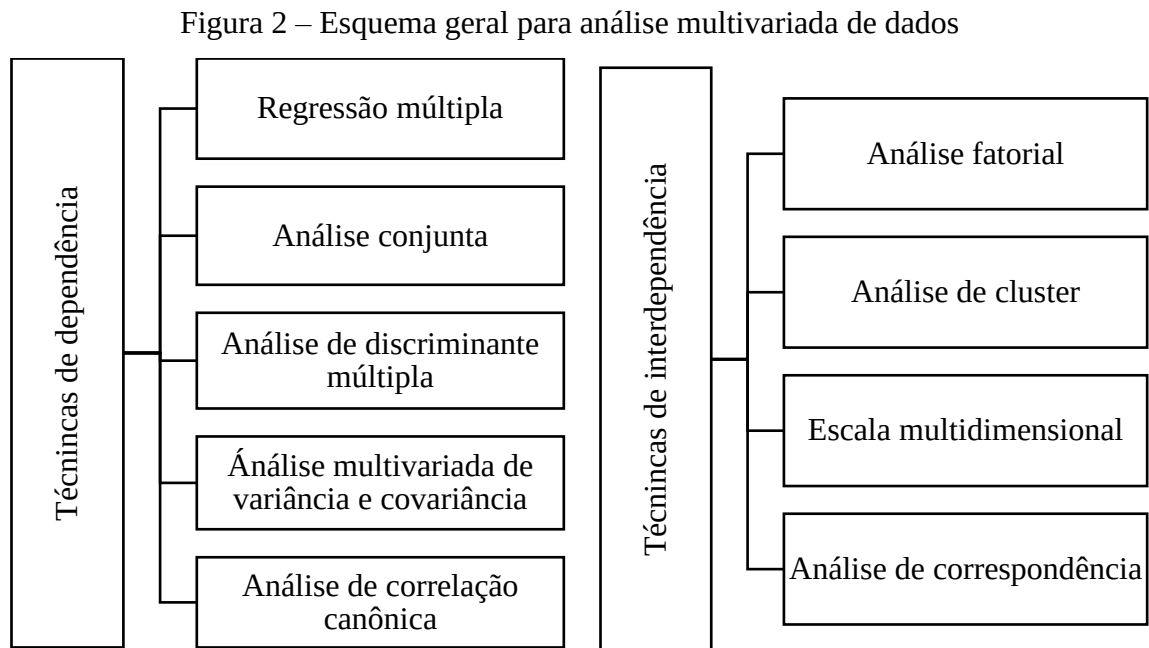
2.3 ANÁLISE MULTIVARIADA DE DADOS

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 2 apresenta as principais técnicas (FERREIRA et al., 2021) considerando a classificação descrita no tópico 2.3.



Fonte: Autoria própria, baseado 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.

2.3.1 Regressão múltipla

Segundo Barbetta et al. (2010), a regressão múltipla, também conhecida por regressão linear múltipla, possibilita prever 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 (2):

$$E\{Y\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j \quad (2)$$

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).

2.3.2 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).

2.3.2.1 Método K-Medoides

O método não hierárquico *K-medoides* consiste na escolha do objeto ou medoide central em um conjunto de dados, de forma que cada elemento é selecionado de acordo com sua similaridade com algum medoide (PALMA, 2018). Para tanto, costuma-se utilizar a técnica de particionamento sobre os medoides (em inglês, *partitioning around medoids*, ou a sigla PAM), que pode ser resumida nos seguintes passos (REYNOLDS et al., 2004):

1. Seleção de k medoides aleatórios;
2. Cálculo das distâncias entre os elementos;
3. Atribuição de cada elemento ao medoide mais próximo;
4. Recálculo das posições dos k medoides;
5. Repetição das etapas 2 e 3 até que os medoides se tornem fixos.

Antes de tudo, é útil aplicar alguma técnica de normalização de variáveis para aumentar a exatidão do algoritmo, colocando os dados numa mesma escala de grandeza (FERREIRA et al., 2021). Além do mais, convém que os dados categóricos se tornem binários ou variáveis *dummy*, isto é, assumam apenas os valores 0 ou 1. Sendo assim, pode-se aplicar a normalização Min-Max para as variáveis numéricas estarem entre 0 e 1, e assim, todos os dados, tanto numéricos quanto categóricos, encontrarem-se numa mesma escala (DE OLIVEIRA et al., 2020).

Quanto à medida de semelhança, pode-se utilizar a distância de Gower, que permite analisar ao mesmo tempo dados quantitativos e qualitativos (ROCHA et al., 2010).

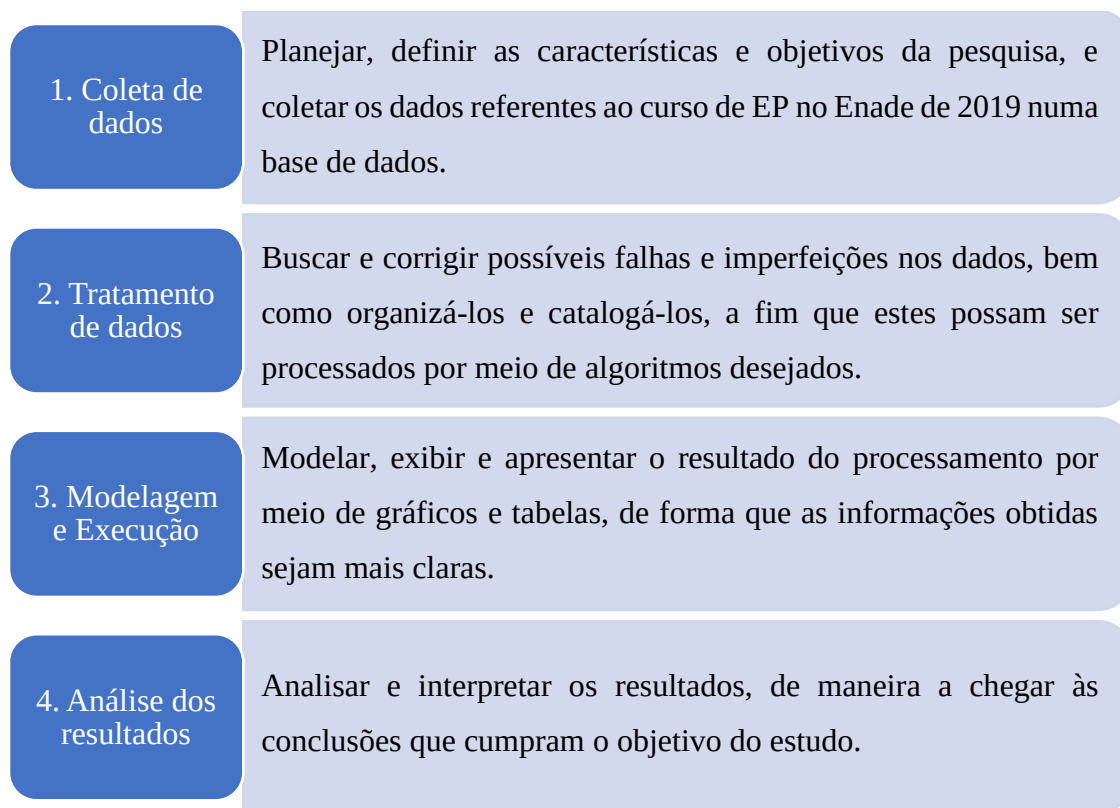
Por fim, a escolha do número de *clusters* se costuma validar por alguns índices, como o de Davies-Bouldin, em que quanto menor o valor, melhor o agrupamento entre os grupos; e o coeficiente da Silhueta, em que o maior valor significa uma escolha mais adequada (DA SILVA et al., 2021).

3 MÉTODOS

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 3.

Figura 3 – Etapas da pesquisa



Fonte: Autoria própria (2022)

As variáveis que foram consideradas para a análise foram divididas em quantitativas e qualitativas. Dentre as quantitativas, encontram-se os indicadores, a saber: Conceito Enade, CPC, IDD e IGC. Recorda-se que estes podem ser classificados em faixas, de 1 a 5; 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:

- Categoria administrativa: especiais (ESP), privadas com fins lucrativos (PCFL), públicas estaduais (PE), públicas federais (PF), públicas municipais (PM) e privadas sem fins lucrativos (PSFL);
- 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);
- Modalidade de Ensino: presencial ou à distância;
- 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

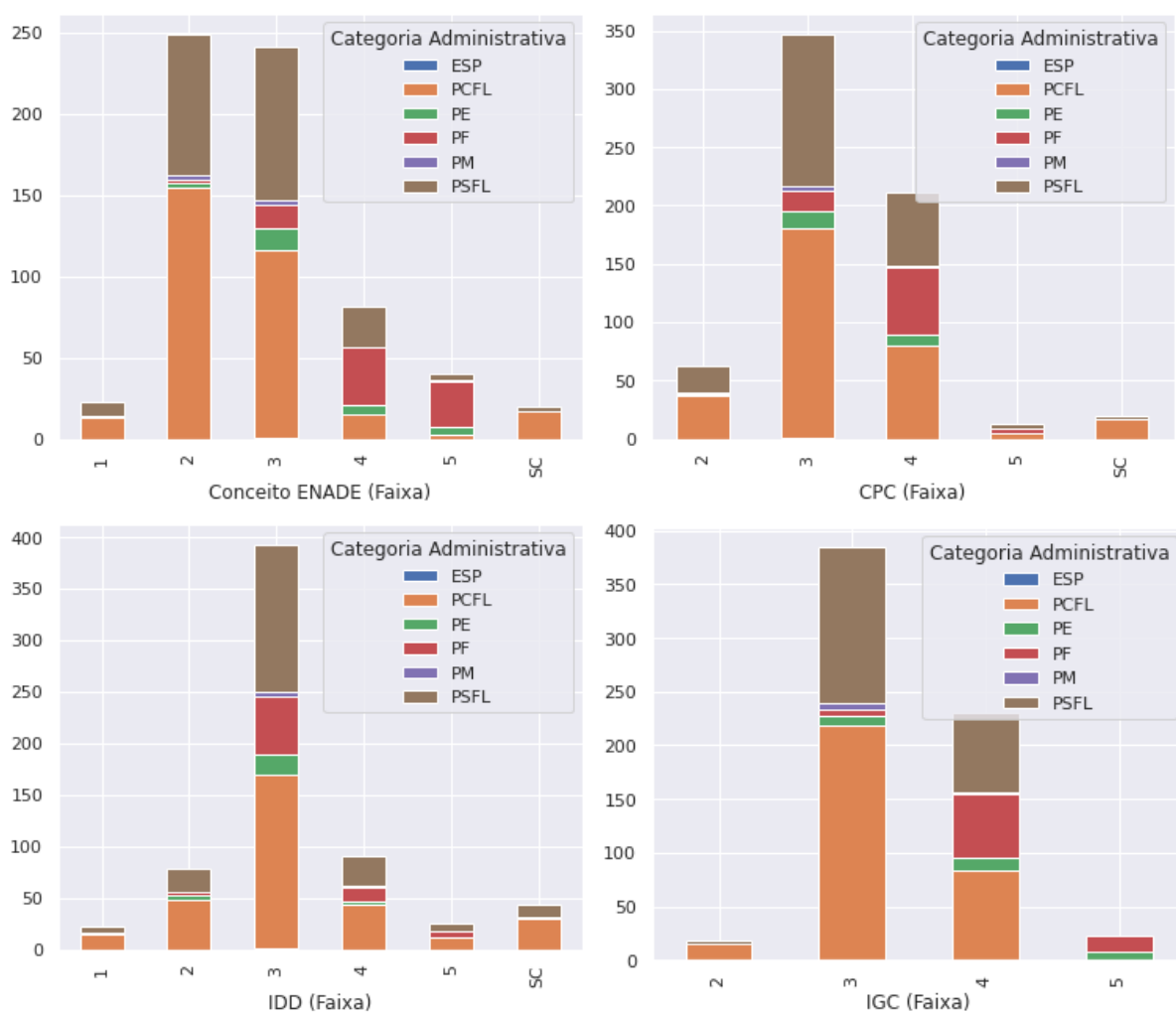
4.1 ANÁLISE EXPLORATÓRIA DE DADOS DE ENGENHARIA DE PRODUÇÃO NO ENADE 2019

A seguir são apresentados os resultados da AED dos 654 cursos de EP do Enade de 2019, relacionando as variáveis categóricas com cada indicador.

4.1.1 Categoria administrativa

Quanto à categoria administrativa, o Gráfico 1 apresenta os dados absolutos de cada indicador por faixa, por meio de gráficos de barras.

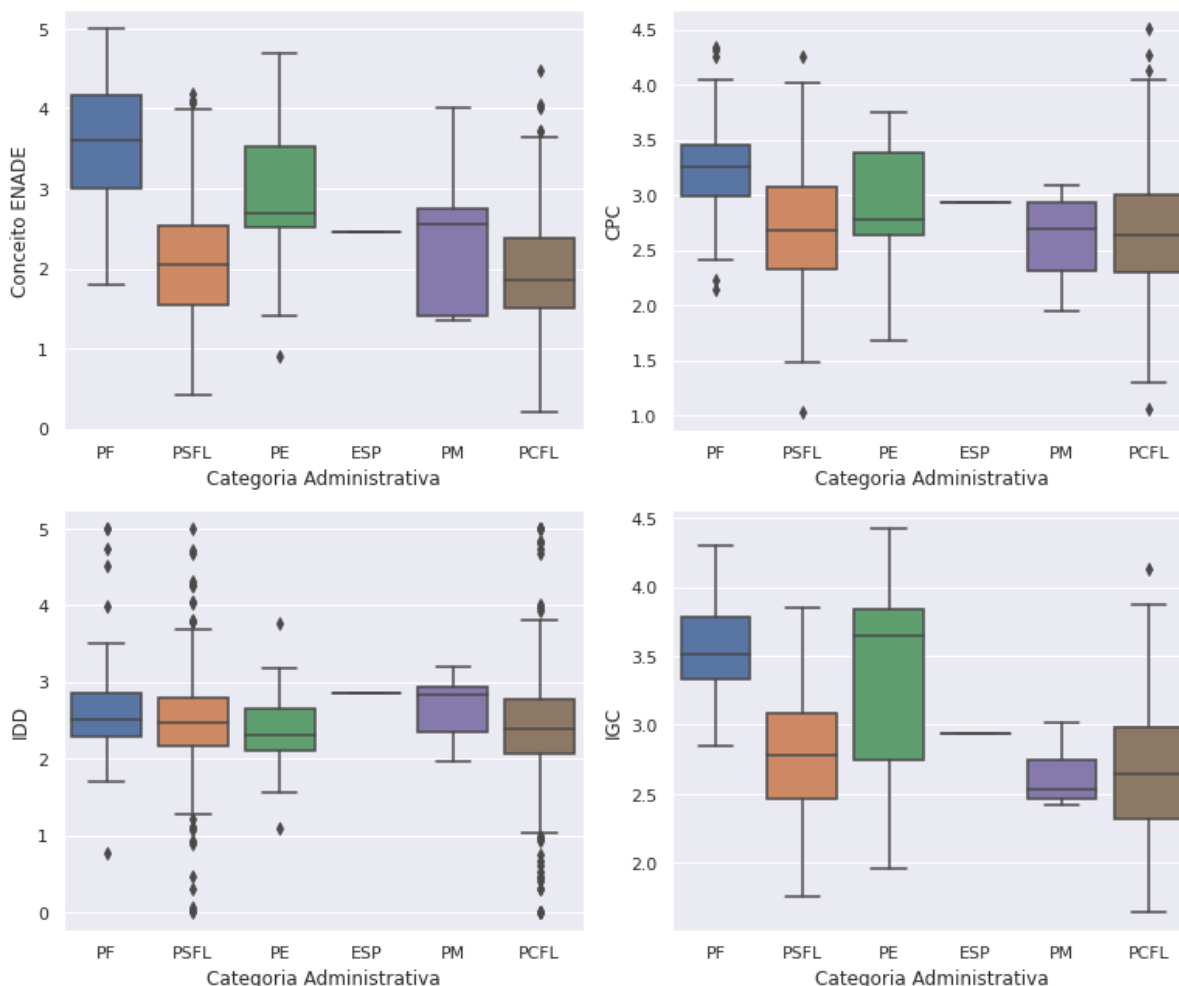
Gráfico 1 – Quantidade de cursos por categoria administrativa por faixa de indicador



Fonte: Autoria própria (2022)

Já o Gráfico 2 representa a variabilidade destes dados por meio de *boxplot*.

Gráfico 2 – Boxplots dos indicadores por categoria administrativa



Fonte: Autoria própria (2022)

Diante dos resultados apresentados, é possível perceber que a maioria dos cursos são de instituições privadas com ou sem fins lucrativos seguidas por públicas federais. Isto reflete sobre a falta de cursos ofertados na área pública, o que restringe muitas vezes o acesso a graduação de EP como de outros cursos à uma parcela da população com uma melhor condição socioeconômica.

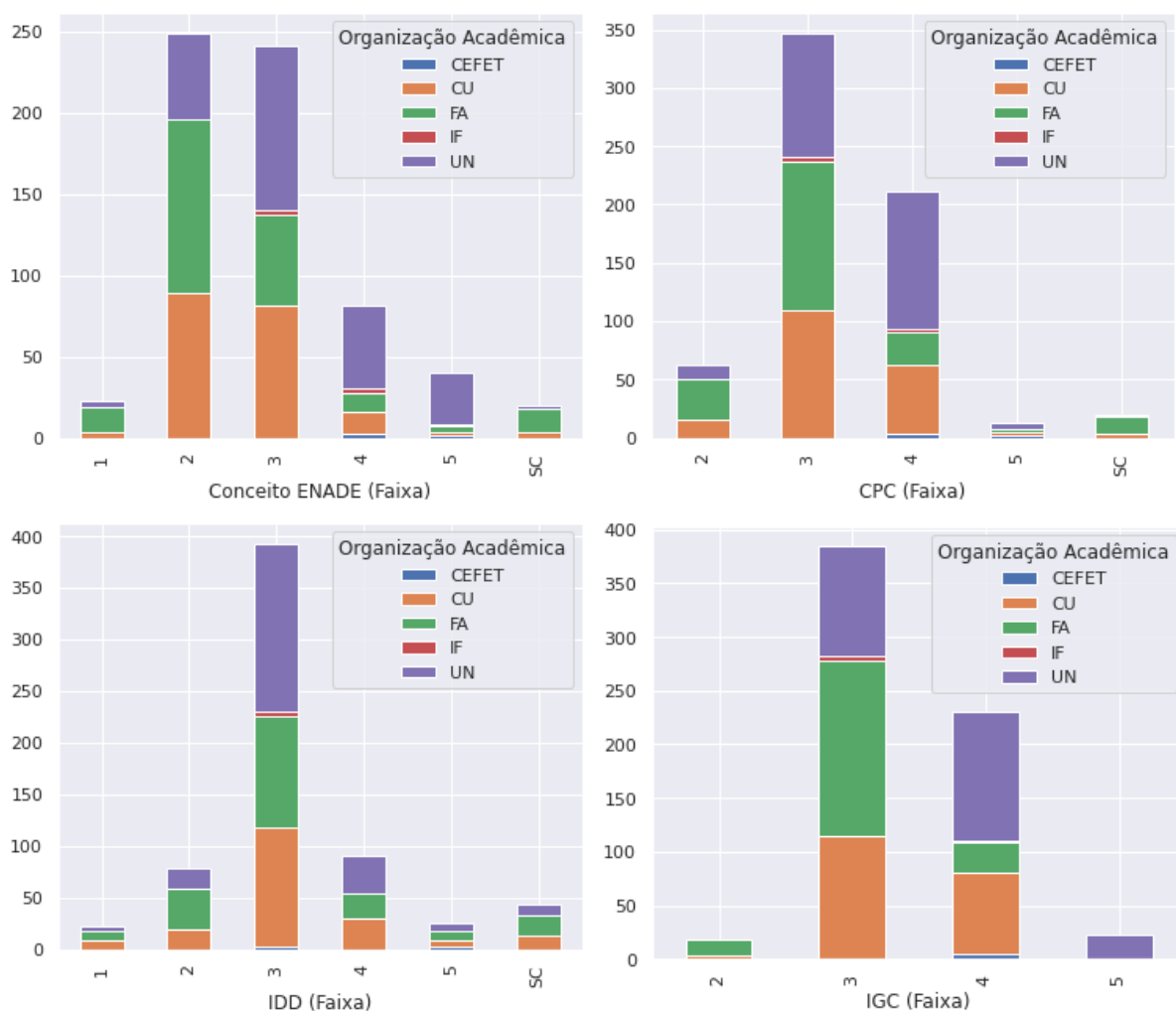
Já em relação a variabilidade dos dados, as instituições privadas apresentaram mais dados discrepantes quanto às demais e, apesar de corresponderem a maior quantidade, seus indicadores estão concentrados em valores inferiores às outras instituições, principalmente no Conceito Enade e no IGC. Já as instituições públicas possuem, em geral, um melhor desempenho. Tal fator é importante na análise da qualidade de ensino por categoria administrativa.

Ainda é válido notar que independente da variável analisada – categoria administrativa, organização acadêmica, modalidade de ensino ou região –, foi possível observar que a maioria dos cursos apresentaram indicadores com faixa 3, o que corresponde ao mínimo esperado, como demonstram os Gráficos 1, 3, 5 e 7. Mesmo assim, quanto ao Conceito Enade, pode-se perceber uma grande quantidade de dados na faixa 2, o que é um fator preocupante.

4.1.2 Organização acadêmica

Quanto à organização acadêmica, o Gráfico 3 caracteriza os dados absolutos de cada indicador por faixa, por meio de gráficos de barras.

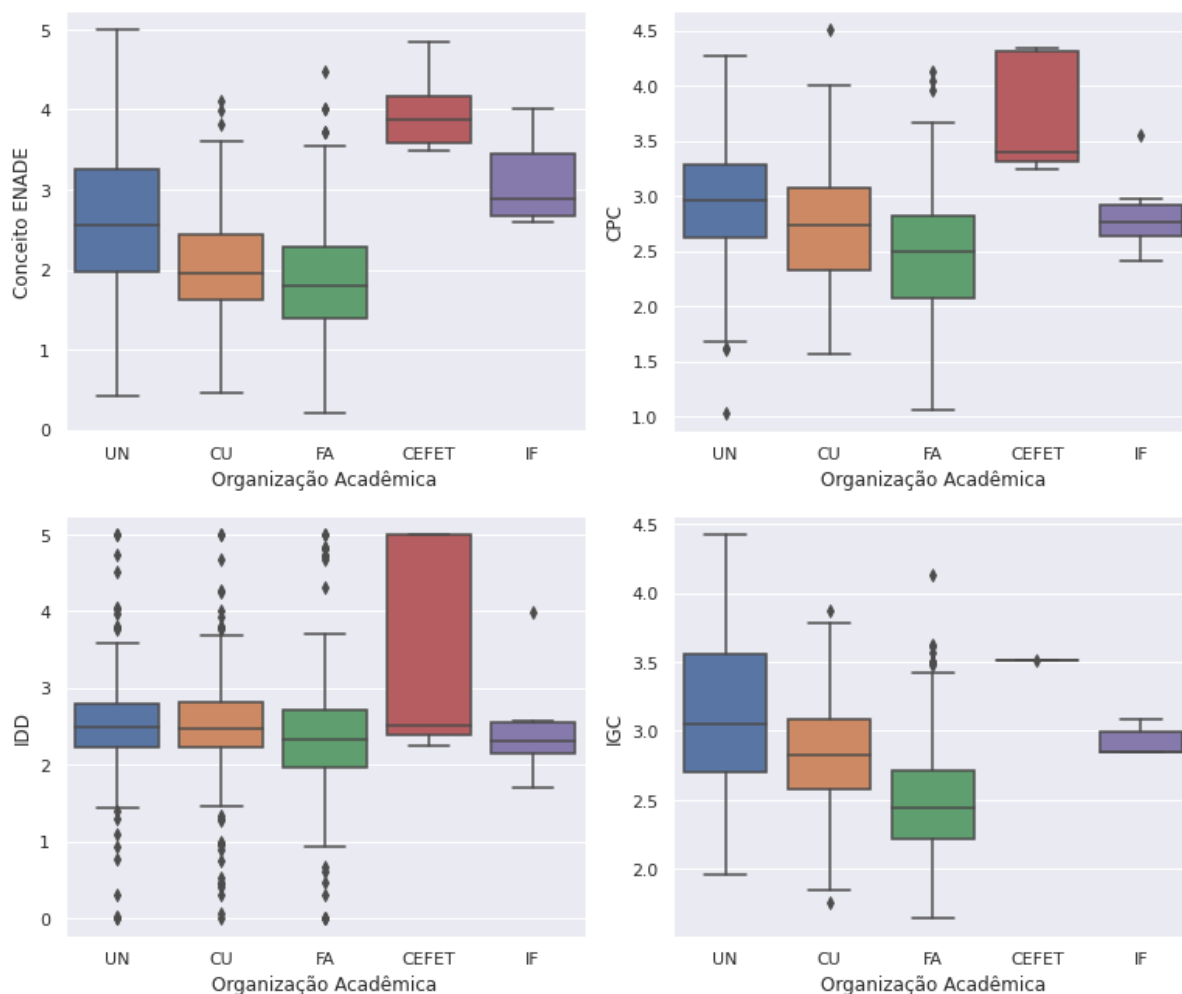
Gráfico 3 – Quantidade de cursos por organização acadêmica por faixa de indicador



Fonte: Autoria própria (2022)

Já o Gráfico 4 configura a variabilidade destes dados por meio de *boxplots*.

Gráfico 4 – Boxplots dos indicadores por organização acadêmica



Fonte: Autoria própria (2022)

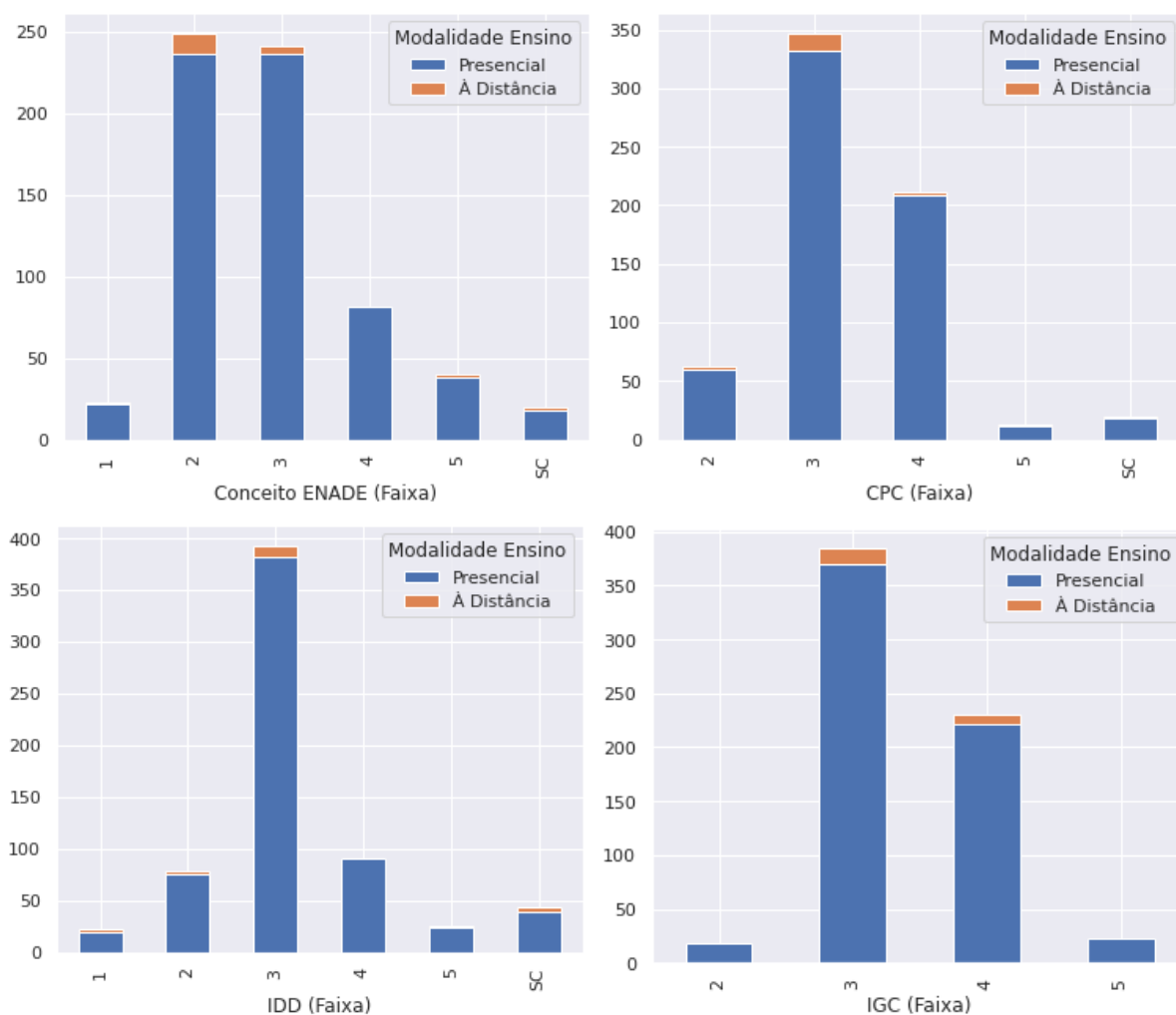
Diante dos resultados revelados, é possível averiguar que a maioria dos cursos se concentram em Universidades, Faculdades e Centros Universitários.

Em relação a variabilidade dos dados, quando se considera todos os indicadores ao mesmo tempo, as Universidades, Centros e Institutos Federais possuem um melhor desempenho, enquanto que as Faculdades possuem um desempenho inferior, principalmente, em relação ao Conceito Enade e ao IGC.

4.1.3 Modalidade de ensino

Quanto à modalidade de ensino, o Gráfico 5 retrata os dados absolutos de cada indicador por faixa, por meio de gráficos de barras.

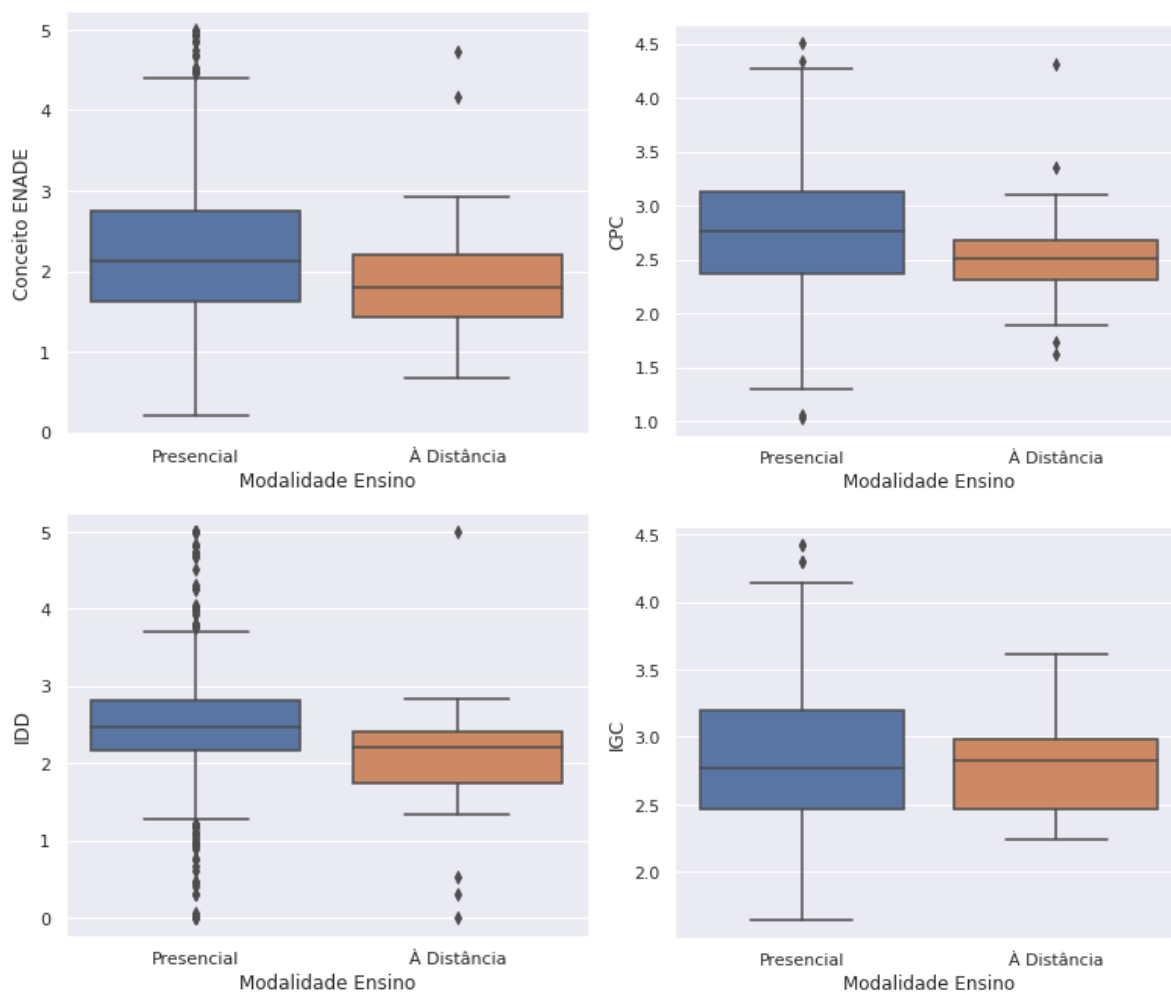
Gráfico 5 – Quantidade de cursos por modalidade de ensino por faixa de indicador



Fonte: Autoria própria (2022)

Já o Gráfico 6 demonstra a variabilidade destes dados através de *boxplots*.

Gráfico 6 – Boxplots dos indicadores por modalidade de ensino



Fonte: Autoria própria (2022)

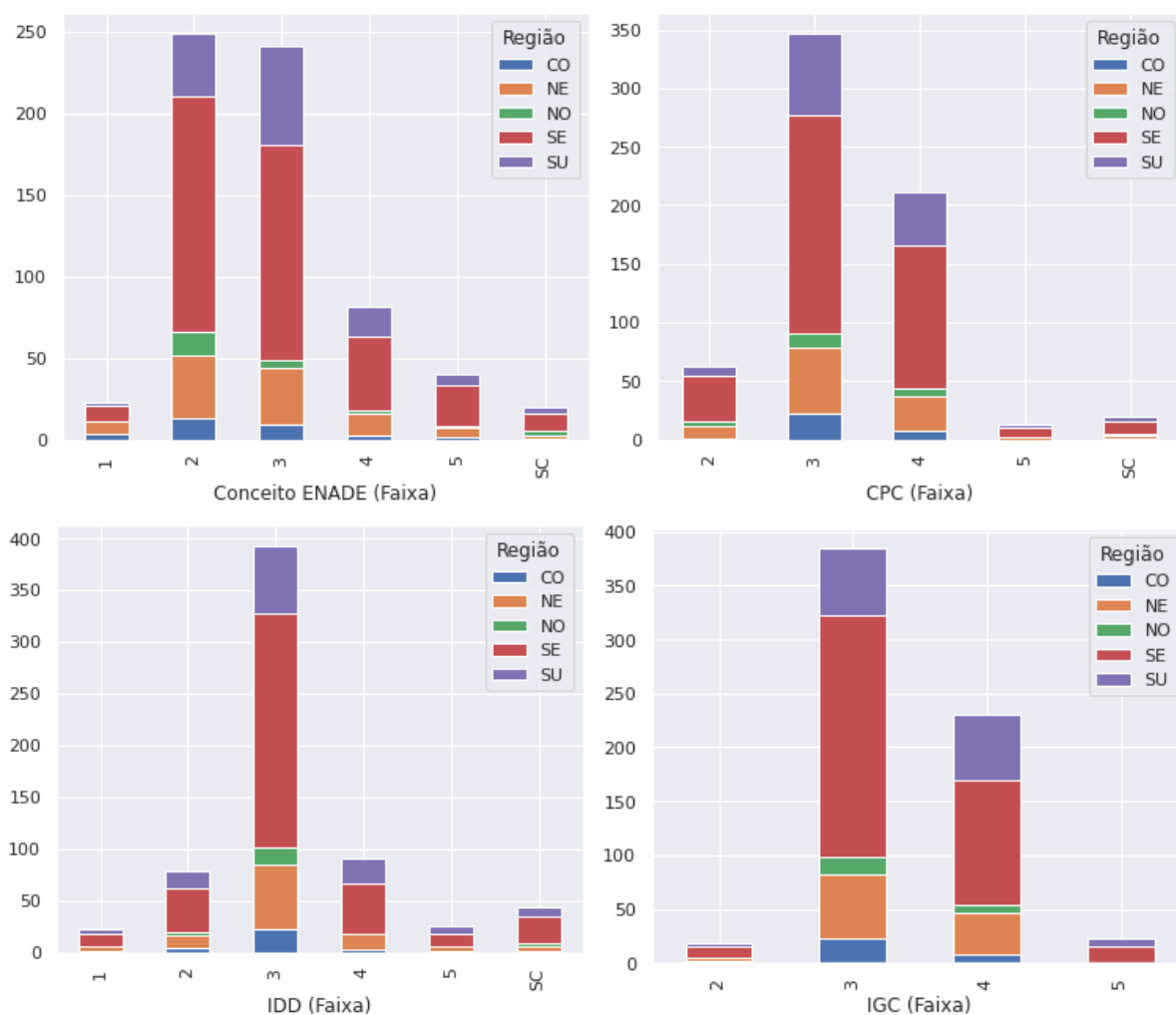
Diante dos resultados apresentados, fica evidente notar que na sua esmagadora maioria os cursos correspondem ao ensino presencial. Em outras palavras, o ensino à distância ainda é pouco valorizado no âmbito da EP.

Quanto à variabilidade dos dados, já que os cursos presenciais refletem à maioria, estes apresentam mais dados discrepantes em relação ao ensino à distância. Mesmo assim, ambos possuem valor médio aproximado dos indicadores, o que induz a percepção de que a qualidade de ensino de EP é semelhante tanto no presencial quanto à distância.

4.1.4 Região

Quanto à região, o Gráfico 7 exprime os dados absolutos de cada indicador por faixa, por meio de gráfico de barras.

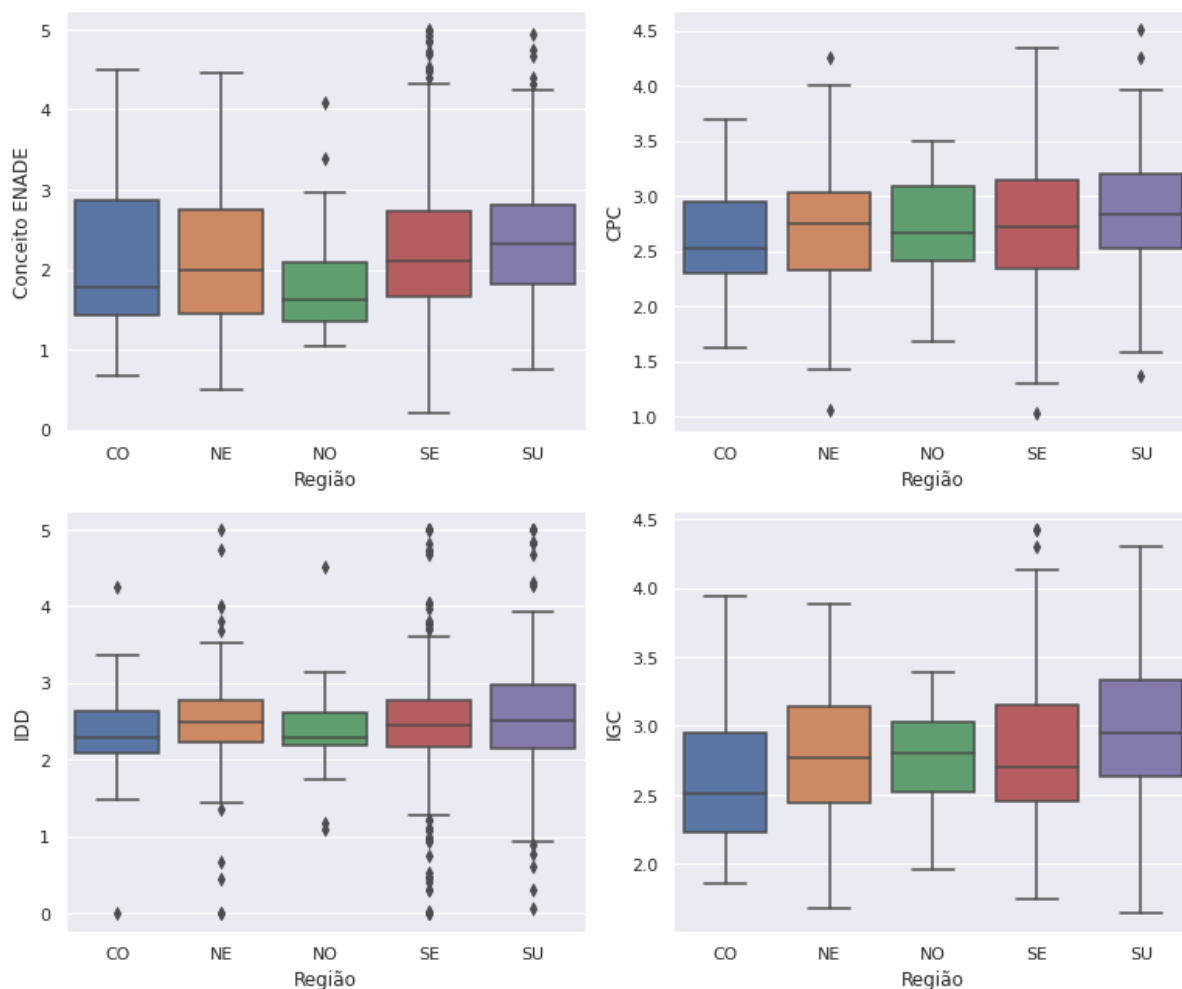
Gráfico 7 – Quantidade cursos por região por faixa de indicador



Fonte: Autoria própria (2022)

Já o Gráfico 8 traduz a variabilidade destes dados através de *boxplots*.

Gráfico 8 – Boxplots dos indicadores por região



Fonte: Autoria própria (2022)

Diante dos resultados apresentados, fica evidente que a maioria dos cursos se encontram na região Sudeste do país, enquanto que região Norte é a que possui menos cursos. Este contraste reflete mais uma vez o quanto as condições socioeconômicas influenciam na oferta de cursos de EP no Brasil.

Quanto à variabilidade dos dados, as regiões possuem um valor médio aproximado dos indicadores, o que induz a conclusão que a qualidade de ensino de EP não depende tanto da região. Mesmo assim, a região Sul se destaca um pouco a mais que as outras.

Por fim, é importante mencionar que a presença de valores discrepantes como os mostrados nos gráficos 2, 4, 6 e 8, podem superestimar ou subestimar a média de cada uma das variáveis categóricas sobre os indicadores.

4.2 ANÁLISE MULTIVARIADA DE DADOS DE ENGENHARIA DE PRODUÇÃO NO ENADE 2019

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*.

4.2.1 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 4.

Figura 4 – 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

Fonte: autoria própria (2022)

A Figura 4 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.

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 5.

Figura 5 – 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

Fonte: autoria própria (2022)

Desta forma, uma terceira regressão foi feita, desconsiderando a variável presencial. O resultado pode ser visto na Figura 6.

Figura 6 – 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

Fonte: autoria própria (2022)

As informações contidas na Figura 6 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:

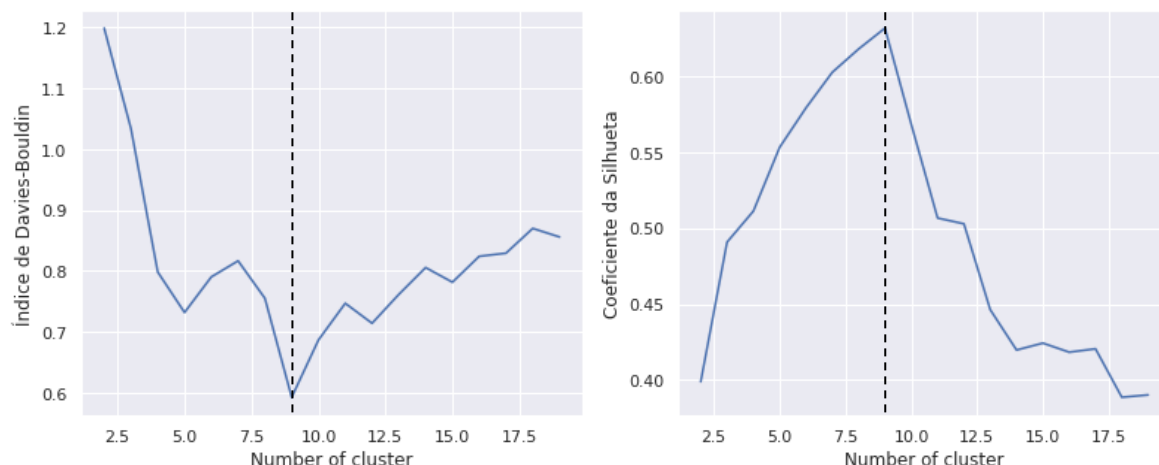
$$\begin{aligned}
 E\{\text{ConceitoEnade}\} = & -0,3252 + 0,7090 \cdot \text{CPC} + 0,1526 \cdot \text{IDD} + 0,2159 \cdot \text{IGC} \\
 & -0,5713 \cdot \text{PrivadaComFinsLucrativos} - 0,5192 \cdot \text{PrivadaSemFinsLucrativos} \\
 & + 0,3782 \cdot \text{PúblicaFederal} - 0,3181 \cdot \text{Norte} + 0,1354 \cdot \text{Sudeste}
 \end{aligned}$$

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*.

4.2.2 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 Gráfico 9.

Gráfico 9 – Resultados para o índice de Davies-Bouldin e o coeficiente da Silhueta



Fonte: autoria própria (2022)

Pelo Gráfico 9 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. Parte do resultado pode ser visualizado na Figura 7.

Figura 7 – Resultado parcial da clusterização

	ConceitoENADE	CPC	IDD	IGC	PrivadaComFinsLucrativos	PrivadaSemFinsLucrativos	PúblicaFederal	Norte	Sudeste	Cluster
0	0.894616	0.683530	0.47346	0.826204	0.0	0.0	1.0	0.0	0.0	2
1	0.633765	0.572866	0.42402	0.619780	0.0	0.0	1.0	0.0	0.0	2
2	0.496035	0.522851	0.45726	0.504926	0.0	0.0	1.0	1.0	0.0	2
3	0.365609	0.430296	0.48676	0.504926	0.0	0.0	1.0	1.0	0.0	2
4	0.649624	0.547859	0.44760	0.473451	0.0	0.0	1.0	0.0	0.0	2
...	...	...	...	...	...	...	...	...	...	...
602	0.412563	0.661972	0.75248	0.397134	0.0	0.0	0.0	0.0	0.0	5
603	0.609349	0.530325	0.61044	0.587650	0.0	0.0	1.0	0.0	0.0	2
604	0.476210	0.545272	0.43632	0.555864	0.0	0.0	1.0	0.0	0.0	2
605	0.276711	0.605346	0.46578	0.592052	1.0	0.0	0.0	0.0	0.0	5
606	0.406093	0.647025	0.59122	0.595571	1.0	0.0	0.0	0.0	1.0	1

Fonte: autoria própria (2022)

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 2.

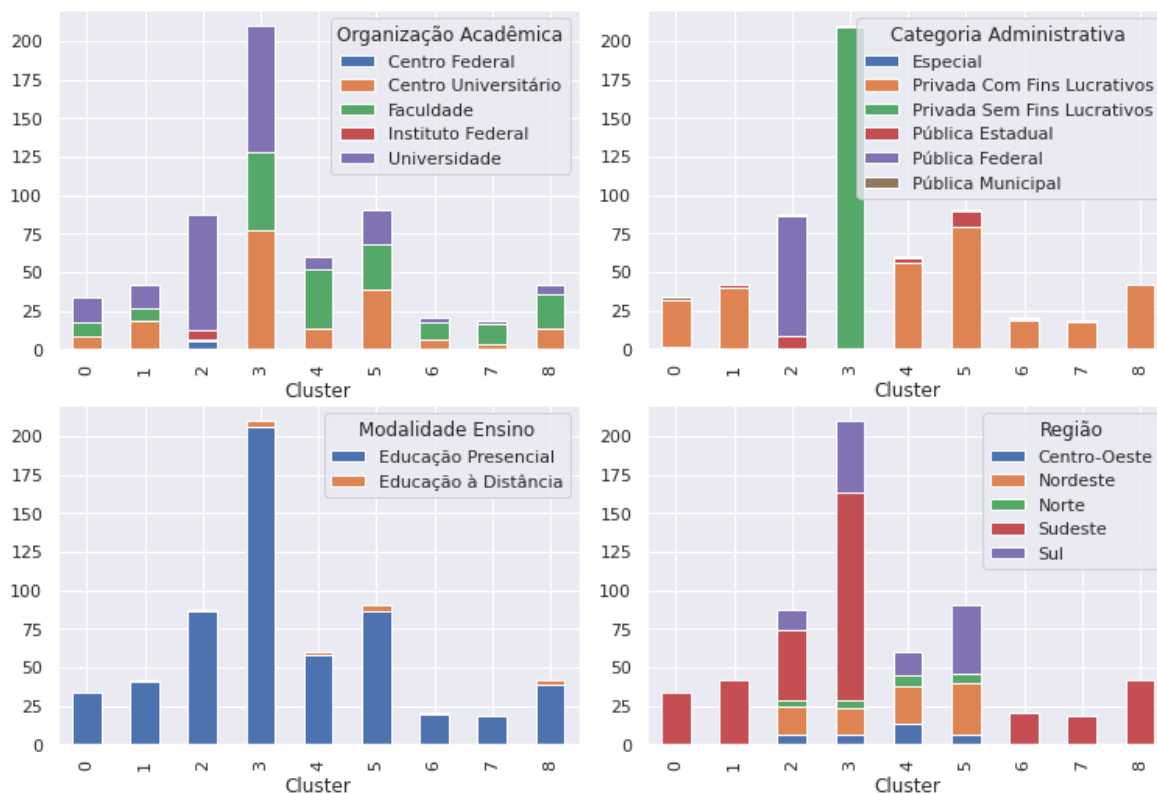
Tabela 2 – Quantidade de cursos e medidas descritivas dos clusters no Conceito Enade

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

Fonte: autoria própria (2022)

Outro detalhe importante é a visualização da proporção de *clusters* em relação a cada variável categórica original do problema mostrado no Gráfico 10.

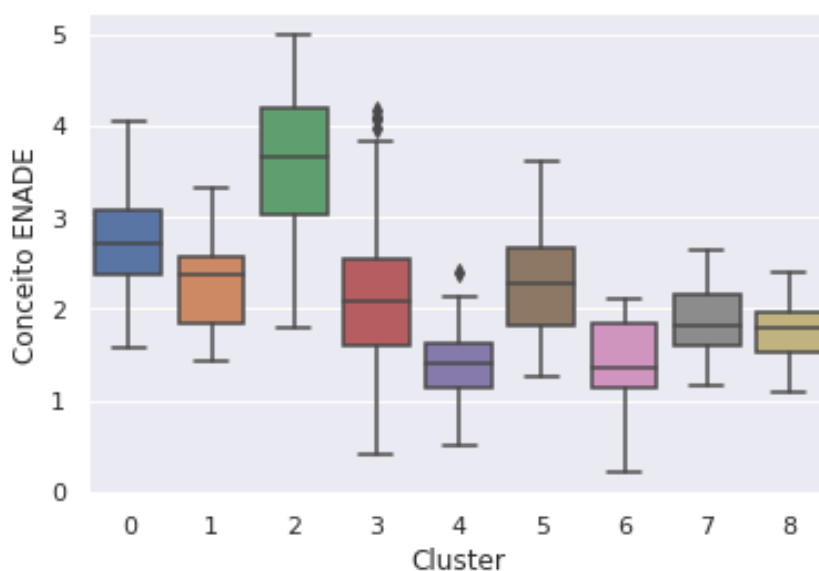
Gráfico 10 – Quantidade de cursos por clusters em cada variável categórica



Fonte: autoria própria (2022)

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

Gráfico 11 – Boxplot do Conceito Enade de acordo com o cluster



Fonte: autoria própria (2022)

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 CONCLUSÃO

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

AGRESTI, Alan; FINLAY, Barbara. **Métodos Estatísticos para as Ciências Sociais**. 4 ed. Porto Alegre: Penso Editora, 2012.

ALVES, William Pereira. **Banco de Dados**. 1. ed. São Paulo: Érica, 2014.

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.

BATALHA, Mario. **Introdução à Engenharia de Produção**. Elsevier Brasil, 2013.

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.

BRASIL. Ministério da Educação. Gabinete do Ministro. **Portaria Normativa nº 23, de 21 de dezembro 2017**. Dispõe sobre os fluxos dos processos de credenciamento e credenciamento de instituições de educação superior e de autorização, reconhecimento e renovação de reconhecimento de cursos superiores, bem como seus aditamentos. Brasília, 2017. Disponível em: <<https://abmes.org.br/arquivos/legislacoes/Republicada-Port-Normativa-023-2017-12-21.pdf>>. Acesso em: 07 de dezembro de 2021.

BRASIL. Presidência da República Casa Civil. Subchefia para Assuntos Jurídicos. **Lei n. 10.861, de 14 de abril de 2004**. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES e dá outras providências. Brasília, 2004. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/110.861.htm>. Acesso em: 07 de dezembro de 2021.

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.

DA SILVA, Ingrid L. A. MELLO, Rafael Ferreira; MIRANDA, Pércles B. C.; NASCIMENTO, André C. A.; MALDONADO, Isabel W. S.; COELHO FILHO, José L. M. **Assessment of Text Clustering Approaches For Legal Documents**. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. Porto Alegre: Sociedade Brasileira de Computação, p. 37-48, 2021.

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.

FURLANETTO, Egidio Luiz; NETO, Henri Geraldo Malzac; NEVES, Cleiber Pereira. **Engenharia de Produção no Brasil: Reflexões Acerca da Atualização dos Currículos dos Cursos de Graduação**. Revista Gestão Industrial, v. 2, n. 4, 2006.

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.

INEP. Instituto Nacional de Pesquisas Educacionais Anísio Teixeira. **Nota Técnica nº 5/2020/CGCQES/DAES**. Diretoria de Avaliação da Educação Superior, 2020a.

INEP. Instituto Nacional de Pesquisas Educacionais Anísio Teixeira. **Nota Técnica nº 34/2020/CGCQES/DAES**. Diretoria de Avaliação da Educação Superior, 2020b.

INEP. Instituto Nacional de Pesquisas Educacionais Anísio Teixeira. **Nota Técnica nº 58/2020/CGCQES/DAES**. Diretoria de Avaliação da Educação Superior, 2020c.

INEP. Instituto Nacional de Pesquisas Educacionais Anísio Teixeira. **Nota Técnica nº 59/2020/CGCQES/DAES**. Diretoria de Avaliação da Educação Superior, 2020d.

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.

MONTGOMERY, Douglas C.; RUNGER, George C. **Estatística Aplicada e Probabilidade para Engenheiros**. 7. ed. Rio de Janeiro: LTC, 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.

PALMA, Luann Farias. **Agrupamento de Dados: K-Médias**. Universidade Federal do Recôncavo da Bahia Centro de Ciências Exatas e Tecnológicas, 2018.

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.

REYNOLDS, Alan P; GRAEME, Richards; VIC, J. Rayward-Smith. **The Application of K-Medoids and PAM to the Clustering of Rules**. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, 2004. p. 173-178.

ROCHA, Mariella Camargo; GONÇALVES, Leandro Simões Azeredo; RODRIGUES, Rosana; DA SILVA, Paula Renata Alves; DO CARMO, Margarida Goréte Ferreira; ABBOUD, Antonio Carlos de Souza. **Uso do Algoritmo de Gower na Determinação da Divergência Genética entre Acessos de Tomateiro do Grupo Cereja**. Acta Scientiarum. Agronomy, v. 32, n. 3, p. 423-431, 2010.

SYAKUR, M. A.; KHOTIMAH, B. K.; ROCHMAN, E. M. S.; SATOTO, B. D. **Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster**. IOP Conf. Series: Materials Science and Engineering, v. 336, p. 012017, 2018.

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

VIEIRA, Sonia. **Fundamentos de Estatística**. 6. ed. São Paulo: Atlas, 2019.