MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Curso de Graduação em Engenharia de Produção

Caracterização do perfil de alunos dos cursos do Icea: uma análise guiada por dados e indicadores de sucesso

André Abner Rocha Ferreira

João Monlevade, MG
2023

André Abner Rocha Ferreira

Caracterização do perfil de alunos dos cursos do Icea: uma análise guiada por dados e indicadores de sucesso

Trabalho de conclusão de curso apresentado ao curso de graduação em Engenharia de Produção do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Bacharel em Engenharia de Produção.

Orientador: Prof. Dr. Thiago Augusto de Oliveira Silva

Coorientadora: Profa. Dra. Helen de Cássia Sousa da Costa Lima

João Monlevade, MG

2023

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

F383c Ferreira, Andre Abner Rocha.
Caracterização do perfil de alunos dos cursos do Icea [manuscrito]: uma análise guiada por dados e indicadores de sucesso. / Andre Abner Rocha Ferreira. - 2023.
82 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Thiago Augusto de Oliveira Silva.
Coorientadora: Profa. Dra. Helen de Cássia Sousa da Costa Lima.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Produção .

1. Ensino superior - Rendimento escolar. 2. Evasão universitária. 3. Frequência universitária. 4. Indicadores educacionais - Sucesso. 5. Mineração de dados (Computação). I. Silva, Thiago Augusto de Oliveira. II. Lima, Helen de Cássia Sousa da Costa. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 004.62:378

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE ENGENHARIA DE PRODUCAO - ICEA

FOLHA DE APROVAÇÃO

André Abner Rocha Ferreira

Caracterização do perfil de alunos dos cursos do Icea: uma análise guiada por dados e indicadores de sucesso

Monografia apresentada ao Curso de Graduação em Engenharia de Produção da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Produção.

Aprovada em 18 de setembro de 2023

Membros da banca

Prof. Dr. Thiago Augusto de Oliveira Silva - Orientador (Deenp/Ufop)
Prof[a]. Dr[a]. Helen de Cássia Sousa da Costa Lima - Coorientadora (Decsi/Ufop)
Prof. Dr. Felipe Nunes Ribeiro (Decsi/Ufop)
Prof. Dr. Paganini Barcellos de Oliveira (Deenp/Ufop)

Thiago Augusto de Oliveira Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 18/09/2023

Documento assinado eletronicamente por **Thiago Augusto de Oliveira Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 19/09/2023, às 14:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0591759** e o código CRC **F6E47A48**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0817 - www.ufop.br

Este trabalho é dedicado à todos aqueles que contribuíram para a minha formação como Engenheiro de Produção.

Agradecimentos

Quero agradecer aos meus pais, Flávio e Ione, que me suportam com amor e tanto me apoiaram durante essa longa jornada na vida universitária. Sou grato pelo suporte nos momentos em que eu mais quis desistir, pelo apoio nas decisões difíceis e pelo cuidado. Ao meu irmão, Nathan, que é muito mais do que um melhor amigo. Ao meu irmão, Gabriel, que faz tanta falta. Eu não sei se existem palavras pra descrever o quanto eu sou grato por vocês dois na minha vida, eu amo vocês. À Letícia que, com tanto amor, me acolhe, encoraja e me motiva a ser melhor, dividindo comigo as conquistas e os revéses da vida.

Aos meus professores da graduação pelo empenho e pela dedicação que colocam no trabalho como Educadores. Em especial, agradeço aos meus orientadores, Thiago e Helen: vocês foram peças chave para o meu desenvolvimento como aluno e como profissional.

Agradeço aos muitos amigos que fiz ao longo da vida e que carrego no peito como família. Seria injusto citar um a um, mas saibam que eu sou muito grato pelo carinho, pelo apoio, pelas palavras de afirmação e por todo amor que recebo de todos os lados.

Reconheço o papel do meu psicólogo durante o período acadêmico. A psicoterapia foi fundamental para que eu me entendesse como humano, como aluno e como profissional.

Por último, agradeço aos companheiros de Clubpetro e Hashdex pela paciência e oportunidade de aprender e me desenvolver como profissional.

"Gota a gota de suor serão recompensadas"

Resumo

Este trabalho apresenta uma caracterização do perfil de alunos dos cursos do Instituto de Ciências Exatas e Aplicadas (Icea), uma unidade acadêmica da Universidade Federal de Ouro Preto (Ufop), a partir de uma análise guiada por dados e indicadores de sucesso. O objetivo é identificar os fatores que influenciam o desempenho e a evasão dos estudantes, bem como as características comuns entre os diplomados, os evadidos e os matriculados de cada curso. Para isso, foram utilizados dados sociais e acadêmicos fornecidos pela unidade, referentes aos ingressantes entre 2011 e 2022. Esses dados foram tratados e analisados por meio de técnicas estatísticas e de mineração de dados, como análise exploratória, redução de dimensão e clusterização. Os resultados mostram que existem diferenças significativas entre os grupos de alunos, formados pela clusterização, tanto em termos de variáveis pré-universidade quanto de indicadores acadêmicos. Além disso, foram identificados alguns dos fatores que mais contribuem para a evasão e o sucesso dos alunos, bem como levantadas sugestões possíveis ações para melhorar a qualidade do ensino e da aprendizagem no Icea.

Palavras-chaves: Evasão. Clusterização. Desempenho acadêmico. Indicadores

Abstract

This paper presents a characterization of the profile of students taking courses at Icea, an academic unit of Ufop, based on an analysis guided by data and success indicators. The objective is to identify the factors that influence student performance and dropout, as well as the common characteristics between graduates, dropouts and those enrolled in each course. For this, social and academic data provided by the unit, referring to freshmen between 2011 and 2022, were used. These data were processed and analyzed using statistical and data mining techniques, such as exploratory analysis, dimension reduction and clustering. The results show that there are significant differences between the groups of students, formed by clustering, both in terms of pre-university variables and academic indicators. In addition, some of the factors that most contribute to student dropout and success were identified, as well as suggestions for possible actions to improve the quality of teaching and learning at Icea.

Keywords: Academic dropout. Clustering. Academic performance. Key Indicators

Lista de ilustrações

Lista de tabelas

Lista de quadros

Lista de Algoritmos

Lista de abreviaturas e siglas

Decea Departamento de Ciências Exatas e Aplicadas

DM Mineração de Dados

Enem Exame Nacional do Ensino Médio

Icea Instituto de Ciências Exatas e Aplicadas

IDE *Integrated Development Environment*

KDD *Knowledge-discovery in Databases*

MDE Mineração de Dados Educacionais

ML Aprendizado de Máquina

PCA Análise de componentes principais

Ufop Universidade Federal de Ouro Preto

Sumário

1 Introdução

Com o avanço da tecnologia, os dados, que antes eram perdidos ou acumulados em pastas empoeiradas, são tratados hoje como o novo petróleo. Na era do *Big Data*, expressão em inglês que, ao pé da letra, significa grandes dados, dados existem em grande variedade e assumem volumes que podem chegar à petabytes, que é equivalente à 1×10^9 megabytes. Isso tudo, aliado à velocidade de processamento dos computadores e ao avanço das tecnologias de informação fez com que esses dados pudessem ser analisados e transformados em informações relevantes para quem os obtém, processo conhecido como a Mineração de Dados (DM).

Bitencourt, Silva e Xavier (2021) defendem que o DM aplicado a dados educacionais de cursos de ensino superior pode ajudar as instituições de ensino a entender melhor o comportamento acadêmico dos alunos. Essa compreensão pode ser usada para desenvolver ações práticas de combate à evasão. Isso pode ser feito analisando dados como frequência às aulas, notas, histórico escolar e fatores socioeconômicos. Ao identificar esses alunos, as instituições podem oferecer apoio e orientação personalizados para ajudá-los a permanecer na universidade.

A literatura indica caminhos para entender fenômenos como a evasão escolar, a inadimplência, avaliação de sinistro e casos de quebra de contrato. O que todos esses fenômenos têm em comum é a possibilidade de análise a partir de algoritmos, que são obtidos com o Aprendizado de Máquina (ML) e a DM conforme descrito por Goldschmidt, Passos e Bezerra (2015).

O Icea, situado no Campus da Ufop da cidade de João Monlevade, oferece, até o presente momento, os cursos de Engenharia de Computação, Engenharia de Produção, Engenharia Elétrica e Sistemas de Informação. A unidade acadêmica sofre com uma alta taxa de evasão e retenção de alunos, um problema que é presente em todo o país. Tanto a desistência quanto a retenção dos alunos na universidade pública são problemas que precisam ser compreendidos e intervenções precisam ser feitas, afinal, além de ser um investimento público, cujo retorno não é materializado, são vagas que poderiam ser ocupadas por outras pessoas que eventualmente alcançariam a diplomação. Sabendo disso, o presente trabalho visa entender, a partir de técnicas de DM e dos dados acadêmicos e sociais dos alunos do Icea, o comportamento, curso a curso, dos alunos evadidos, retidos e diplomados e, consequentemente, auxiliar a instituição no processo de tomadsa de decisão e desenvolvimento de políticas para tratar o problema da evasão e da retenção de alunos no Icea

A evasão e retenção escolar é um problema latente nas instituições de ensino superior, visto que provoca graves consequências sociais, acadêmicas e econômicas (BAGGI; LOPES, 2011). Um levantamento da Secretaria de Modalidades Especializadas de Educação (Semesp) indica que mais de 300 mil alunos trancaram seus cursos em universidades públicas no ano de 2020, onde a pandemia do COVID-19 fez com que a maioria das instituições adotassem o modelo de ensino à distância. Além disso, a pesquisa da Semesp mostra que a taxa de evasão nas universidades, que de 2015 até 2019 era de, em média, 18,5% subiu para 21,7% no ano em que a pandemia começou no Brasil (SEMESP, 2022).

Não há um entendimento claro das razões da evasão e retenção de alunos no contexto do Icea. Existem trabalhos feitos por alunos do próprio Icea que abordam o tema, como Gonçalves (2022), Paranhos (2021) e Rodrigues (2022). A partir dos estudos anteriores identificou-se a necessidade e oportunidade de analisar os dados de forma estratificada, isto é, dividindo, curso a curso, como este trabalho objetiva.

1.1 Objetivo geral

O objetivo geral do trabalho é analisar, por meio de caracterização e agrupamento das informações pré-universidade e do histórico de desempenho dos estudantes do Icea, os casos de sucesso e insucesso acadêmico em cada curso.

Para cumprimento do objetivo geral, é necessário atender os seguintes objetivos específicos:

1. coletar os dados disponibilizados pela Seção de Ensino do Icea;
2. realizar a limpeza e o pré-processamento dos dados;
3. fazer uma análise exploratória e descritiva dos dados;
4. fazer uma clusterização das bases de cada curso ofertado no Icea;
5. entender, a partir da clusterização, os perfis dos alunos desistentes e dos retidos e caracterizar os perfis encontrados em cada curso.

1.2 Organização do Trabalho

O restante deste trabalho está estruturado do seguinte modo. No Capítulo 2 são apresentados os trabalhos relacionados e a fundamentação teórica em que este trabalho é embasado. No Capítulo 3 serão abordados os processos utilizados para o desenvolvimento do trabalho e a modelagem do problema. No Capítulo 4 serão discutidos os resultados da clusterização, curso a curso. Nos Capítulos 5 e 6 serão discutidas as considerações finais e os trabalhos futuros.

2 Revisão da Literatura

Este capítulo tem como objetivo apresentar a fundamentação teórica em que o presente trabalho é embasado, conceituando os termos que são pertinentes para a pesquisa. Além disso, o capítulo apresenta trabalhos relacionados ao tema.

2.1 Evasão

A evasão no contexto do ensino superior é um fenômeno social e econômico complexo, definido como interrupção no ciclo de estudos (GAIOSO, 2005). Este fenômeno tem preocupado as instituições de ensino superior, haja vista que a alta taxa de desistência dos alunos pode ser um indicador ruim para a instituição e também porque a saída desses alunos provoca graves consequências sociais, acadêmicas e econômicas (BAGGI; LOPES, 2011). Já Vargas e Lima (2004) faz um levantamento de definições de evasão de acordo com diferentes autores e a amplitude desses conceitos, que é mostrado na Tabela.

Quadro 2.1 – Definição de evasão e amplitude do conceito

Trabalhos	**Definição**	**Amplitude do conceito**
Utiyama e Borba (2003) Apud Vargas e Lima (2004)	Evasão é entendida como a saída definitiva do aluno de seu curso de origem sem concluí-lo	Ampla. Não foi estabelecido nenhum critério de tempo no curso para a saída do aluno
Maia e Meirelles(2005) Apud Vargas e Lima (2004)	Evasão consiste em alunos que não completam cursos ou programas de estudo podendo ser considerada como evasão aqueles alunos que se matriculam e desistem antes mesmo de iniciar o curso.	Especifica que mesmo os alunos que nunca começaram o curso devem ser considerados no cálculo das taxas de evasão
Abbad; Carvalho e Zerbini(2005) Apud Vargas e Lima (2004)	Evasão refere-se à desistência definitiva do aluno em qualquer etapa do curso.	Não deixa claro se evasão se aplicaria apenas aos alunos que chegaram a iniciar o curso ou se abrangeria também àqueles que apenas se matricularam e nunca iniciaram o curso

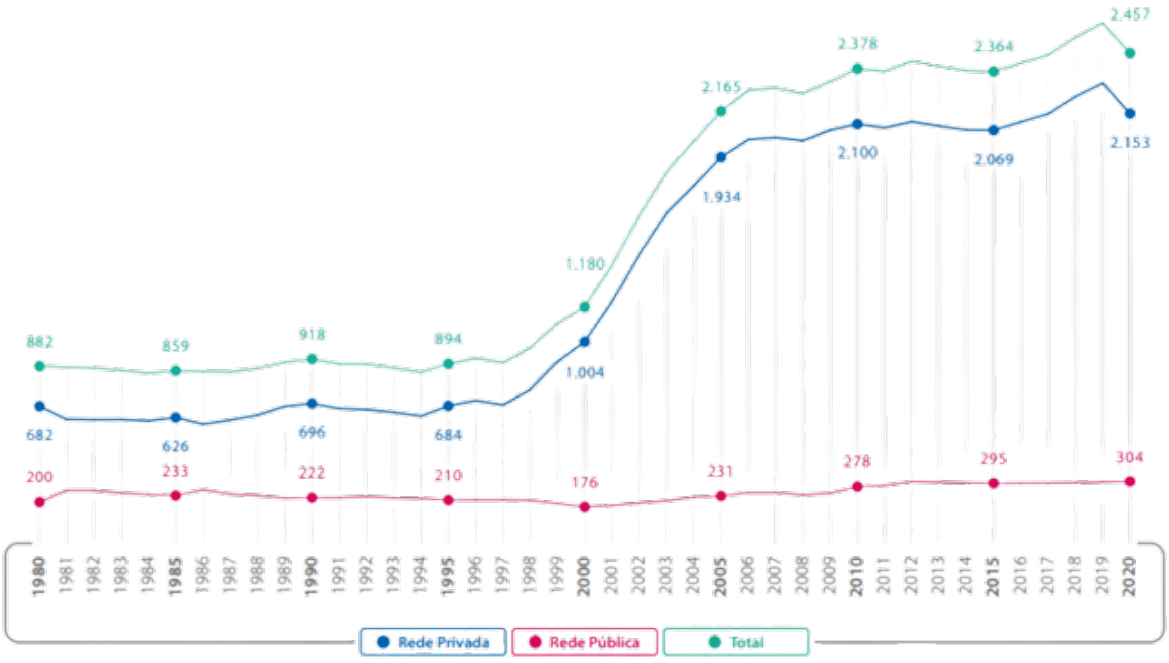
Fonte: Vargas e Lima (2004)

O Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (INEP), que é responsável pelo levantamento dos dados do sistema educacional, conceitua a evasão conforme trecho a seguir.

> Evasão: saída antecipada, antes da conclusão do ano, série ou ciclo, por desistência (independentemente do motivo), representando, portanto, condição terminativa de insucesso em relação ao objetivo de promover o aluno a uma condição superior a de ingresso, no que diz respeito à ampliação do conhecimento, ao desenvolvimento cognitivo, de habilidades e de competências almejadas para o respectivo nível de ensino. Obviamente, a interrupção do programa em decorrência de falecimento do discente não pode ser atribuída como insucesso, dado que, de forma geral, se trata de caso fortuito e não se pode presumir uma intencionalidade do indivíduo em interromper o curso, cessá-lo ou uma incapacidade do indivíduo de manter-se no programa educacional (INEP, 2017, p.9) .

Com a democratização do acesso ao ensino superior e a pressão do mercado por profissionais qualificados, a procura por um diploma aumentou significativamente, o que fez com que a oferta de cursos aumentasse de forma proporcional. Um levantamento da Semesp mostra um aumento de quase 110% de instituições de ensino superior de 2000 a 2020, sendo que, das 1277 novas instituições, apenas 128 são da rede pública, como mostrado na Figura 1.

Figura 1 – Evolução do número de instituições de ensino superior no Brasil



Fonte: (SEMESP, 2022)

Com o aumento da oferta de instituições e, consequentemente, de cursos, há também uma movimentação maior dos alunos na busca por outros cursos e/ou faculdades/centros de ensino/universidades. As razões pelas quais isso acontece não são facilmente mensuráveis, no entanto, evidências empíricas sugerem que fatores internos e externos influenciam a decisão do estudante no que diz respeito à evadir ou se manter no curso, como por exemplo causas demográficas, acadêmicas, pessoais e familiares (BITENCOURT; SILVA; XAVIER, 2021).

2.2 Trabalhos relacionados

Como visto anteriormente, a evasão escolar é tema de diversos estudos que objetivam entender, prever ou intervir na saída prematura ou na retenção de estudantes dentro das instituições de ensino no país. O trabalho de Bitencourt, Silva e Xavier (2021) sugere o uso de técnicas de Mineração de Dados Educacionais (MDE) por ML, partindo da ideia de Goldschmidt, Passos e Bezerra (2015) de que dados educacionais brutos podem ser tanto transformados em informações relevantes para as pesquisas sobre a educação quanto para a prática do processo educacional. Essas informações podem ser usadas para identificar os dados estudantis por classificação ou clusterização. Ainda segundo Bitencourt, Silva e Xavier (2021) os processos de classificação desses dados podem ser feitos através de algoritmos de ML que têm como objetivo desenvolver programas de computador que podem aprender regras de decisão a partir de treinamentos.

No Icea também existem pesquisas direcionadas ao entendimento, de alguma forma, sobre a evasão universitária. Gonçalves (2022), por exemplo, buscou desenvolver um modelo de inteligência artificial para identificar o padrão curricular dos alunos do Icea. A autora conseguiu desenvolver um modelo computacional capaz de prever com uma acurácia de 87,90% a evasão dos alunos contidos na base de dados da unidade, elencando as variáveis mais relevantes para o modelo e, consequentemente, para o sucesso ou insucesso escolar dos alunos.

Rodrigues (2022) fez, por sua vez, uma pesquisa voltada para os cursos da área de computação da unidade, objetivando uma análise de caracterização quantitativa e predição da evasão dos alunos a partir de técnicas de DM. Apesar da limitação aos cursos de computação, os resultados da caracterização do aluno egresso são muito semelhantes aos resultados de Gonçalves (2022).

Paranhos (2021), por sua vez, desenvolveu um *dashboard* para análise dos dados educacionais dos alunos do Icea que é de grande valia para a universidade, visto que as informações podem ser vistas por aluno, departamento e disciplina, além de dados gerais sobre colação, evasão, admissão e demográficos, o que possibilita perspectivas diferentes sobre os problemas enfrentados pelos alunos e pela própria universidade.

Caldeira (2021) caracteriza a evasão de discentes usando a metodologia *Knowledge-discovery in Databases* (KDD). Isso possibilitou uma melhor compreensão do problema da desistência dos alunos, já que os cinco grupos obtidos através da clusterização apresentaram comportamentos distintos entre si.

Essas são apenas algumas das pesquisas que têm metodologias semelhantes ao presente trabalho, o que evidencia a dificuldade da instituição em lidar com alunos egressos e tomar providências adequadas a respeito do problema. De diferente forma, o presente trabalho trata a evasão, curso a curso, o que ainda não havia sido feito em outros trabalhos.

2.3 Python

A linguagem de programação Python é muito utilizada no universo da mineração de dados por ser considerada uma linguagem de alto nível e possuir diversas bibliotecas para análise e processamento dos dados. Mueller e Massaron (2019) defendem que o uso da linguagem para a Ciência de Dados é também sobre desempenho, já que é uma linguagem que facilita o uso de multiprocessamento em grandes conjuntos de dados e que, além disso, possui diversos *Integrated Development Environment* (IDE) especializados e que facilitam a visualização e os cálculos necessários para o desenvolvimento do projeto, e por isso, será a linguagem de programação utilizada para o desenvolvimento da parte programática do presente trabalho.

2.4 Clusterização - *k-means*

O algoritmo de clusterização *k-means*, ou k-médias é, segundo Jain, Murty e Flynn (1999), o mais popular dos algoritmos para clusterização pela facilidade de implementação e sua baixa ordem de complexidade $O(n)$, que escala linearmente com o número n de padrões.

O *k-means* é um método de aprendizado não supervisionado que agrupa dados de acordo com suas características, formando k clusters (ou grupos) com centroides (ou pontos centrais) definidos. Jain, Murty e Flynn (1999) apresenta o passo a passo conforme o algoritmo 1

Algoritmo 1: Algoritmo k-means

Entrada: Dados, k

Saída: Agrupamentos

```
1  início
2      Escolha k centros de cluster aleatoriamente ;
3      para cada ponto em Dados faça
4          Calcule a distância euclidiana do ponto para cada centro de cluster;
5          Atribua o ponto ao centro de cluster mais próximo;
6      fim
7      Atualize a posição dos centros de cluster, calculando a média aritmética de todos os
         pontos em cada cluster;
8      repita
9          Recalcule os centros de cluster usando as associações de cluster atuais;
10     até que o critério de convergência for alcançado;
11 fim
```

Considerando um conjunto de dados X contendo n pontos de dados multidimensionais e k agrupamentos para dividir a base. A distância euclidiana é selecionada como o índice de similaridade para dividir a base em k agrupamentos. Os alvos de agrupamento minimizam a soma dos quadrados dos desvios intra-agrupamentos tal qual mostrado na Equação (2.1).

$$d = \sum_{k=1}^{k} \sum_{i=1}^{n} ||(x_i - u_k)||^2 \tag{2.1}$$

onde k representa os k centros de cluster, u_k representa o centro k, e x_i representa o ponto i no conjunto de dados. A solução para o centróide uk é mostrado na Equação 2.2.

$$\frac{\partial d}{\partial u_k} = \sum_{i=1}^{n} 2(x_i - u_k) \tag{2.2}$$

2.5 Análise de componentes principais

A PCA, do inglês *Principal Component Analysis*, é um método estatístico que permite reduzir a dimensão de um conjunto de dados, mantendo a maior parte da informação contida neles (JOLLIFFE, 2002). O PCA transforma as variáveis originais em novas variáveis chamadas componentes principais, que são combinações lineares das variáveis originais. As componentes principais são ordenados de forma decrescente de acordo com a sua variância, que mede a quantidade de informação que cada componente captura. A mínima variância explicada é um critério que define o número mínimo de componentes principais que devem ser mantidos para preservar uma certa percentagem da variância total dos dados. Por exemplo, para manter 95% da variância total, deve-se escolher o número de componentes principais, tal que a soma das suas variâncias seja igual ou superior a 0,95 vezes a variância total (HAIR, 2009).

Segundo Hongyu, Sandanielo e Junior (2016), essa técnica transforma, de forma linear, um conjunto original de variáveis em um conjunto menor de variáveis não correlacionadas que contém a maior parte das informações do conjunto de dados original. O PCA é associado à ideia de redução de massa de dados, com menor perda possível de informação.

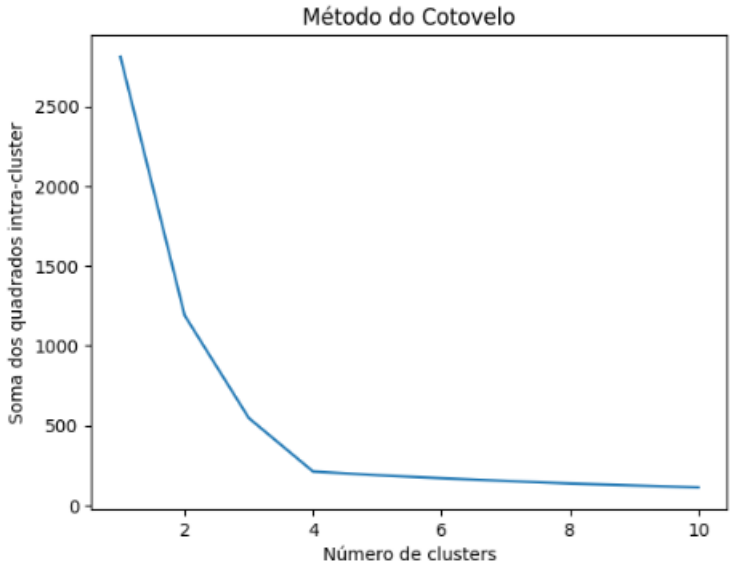
Além disso, Fan et al. (2018) defende o uso da análise de componente principal para a clusterização no contexto do *big data* por ser uma ferramenta poderosa que pode otimizar a execução computacional de problemas e, mesmo com diminuição no número de componentes em um conjunto de dados, fornece fortes garantias estatísticas.

2.6 Método do Cotovelo

A ideia principal por trás do Método do Cotovelo é identificar o ponto em que o aumento do número de clusters não resulta em um ganho significativo na explicação da variância. Este ponto é visualmente representado como um "cotovelo"em um gráfico que mostra a variância explicada em função do número de clusters (KETCHEN; SHOOK, 1996). No eixo *y* do gráfico do Método do Cotovelo mostra a soma das distâncias quadráticas entre os pontos do clusters e o centróide do mesmo cluster.

O método do Cotovelo é intuitivo e fácil de implementar, mas também tem suas limitações. Por exemplo, nem sempre é possível identificar claramente um "cotovelo"no gráfico, especialmente em conjuntos de dados complexos ou de alta dimensão (BAI; LIANG; CAO, 2020). Além disso, o Método do Cotovelo não leva em consideração a estrutura dos clusters, o que pode resultar em uma subestimação ou superestimação do número ótimo de clusters (MILLIGAN; COOPER, 1985). A Figura 2 mostra como seria um gráfico do método do Cotovelo onde o número ótimo de clusters é 4.

Figura 2 – Método do Cotovelo



Fonte: Elaborado pelo autor

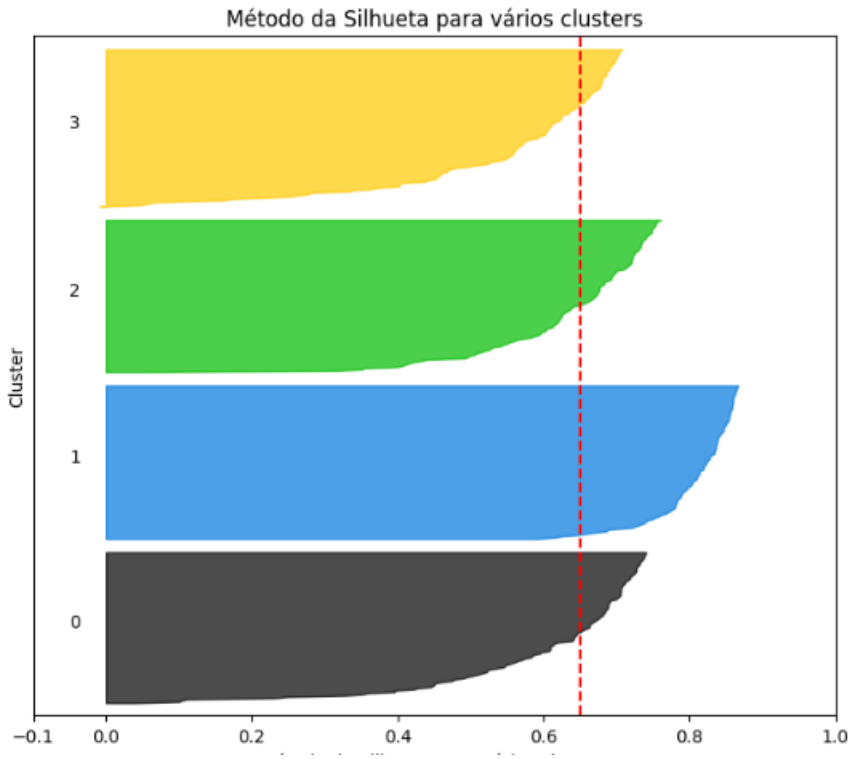
.

Apesar dessas limitações, o Método do Cotovelo é uma ferramenta valiosa na análise de clusterização, especialmente quando usado em conjunto com outras técnicas, como a análise de silhueta ou a validação cruzada (ROUSSEEUW, 1987).

2.7 Método da Silhueta

O método da Silhueta, proposto por Rousseeuw (1987) é uma técnica usada para interpretar e validar a consistência dentro de clusters de dados. Ele fornece uma representação gráfica de quão bem cada objeto foi classificado. O valor da silhueta é uma medida de quão semelhante um objeto é ao seu próprio cluster (coesão) em comparação com outros clusters (separação). A silhueta varia de -1 a +1, onde um valor alto indica que o objeto está bem combinado com seu próprio cluster e mal combinado com os clusters vizinhos. Se a maioria dos objetos tiver um valor próximo à média, então a configuração do cluster é apropriada. Se muitos pontos tiverem um valor baixo ou negativo, então a configuração dos clusters será ruim. A Figura 3 mostra o comportamento dos clusters no método da Silhueta

Figura 3 – Método da Silhueta



Fonte: Elaborado pelo autor

.

O gráfico do método da Silhueta, mostrado na Figura 3 é baseado no conceito de silhueta, que é uma medida da coesão interna e da separação externa de um cluster. A coesão interna mede o quanto os pontos de dados de um cluster são semelhantes entre si. A separação externa mede o quanto os clusters são diferentes entre si. Para cada ponto de dados, a silhueta é calculada como a diferença entre a distância média do ponto ao seu cluster e a distância média do ponto ao cluster mais próximo. O valor da silhueta varia de -1 a 1, sendo:

- +1: o ponto está bem ajustado ao seu cluster e mal ajustado aos outros clusters.
- 0: o ponto está na fronteira entre dois clusters.

- -1: o ponto está bem ajustado aos outros clusters e mal ajustado ao seu cluster.

A linha vertical vermelha é o coeficiente médio da silhueta, que é a média dos valores de silhueta de todos os pontos de dados. Um valor alto do coeficiente médio da silhueta indica que os clusters são bem ajustados e bem separados. O coeficiente médio da silhueta é usado como referência da qualidade dos agrupamentos, ou seja, se no gráfico a silhueta do cluster não tocar a linha que representa o coeficiente médio da silhueta, este cluster está mal ajustado.

3 Metodologia

Este capítulo tem como objetivo definir a metodologia que será utilizada nesta pesquisa, desde a obtenção dos dados até a implementação dos modelos. O presente trabalho tem natureza quantitativa empírica descritiva e usará a metodologia CRISP-DM nas etapas de mineração de dados.

3.1 Classificação da Pesquisa

Antes da coleta de dados, é necessário determinar o caminho a ser traçado para obtenção do que foi proposto na pesquisa. A metodologia de pesquisa, que é o caminho determinado pelo pesquisador para o desenvolvimento do estudo, precisa ser estabelecida de forma adequada, de acordo com o assunto abordado. A metodologia usada neste trabalho pode ser classificada como quantitativa, já que há uma mensuração de variáveis de pesquisa, que segundo Cauchick-Miguel et al. (2018) é a característica mais marcante da abordagem quantitativa. Além disso, a metodologia quantitativa pode ser segmentada em metodologias quantitativas por pesquisa axiomática e por pesquisa empírica, que será a abordagem adotada neste trabalho pois:

> (...) preocupa-se com testes em processos reais, com a validade dos modelos científicos obtidos pela pesquisa teórica quantitativa, e com a utilidade e o desempenho das soluções resultantes (CAUCHICK-MIGUEL et al., 2018, p. 176).

Por fim, ainda é possível separar a pesquisa empírica em três classificações, sendo elas: empírica quantitativa, empírica descritiva e empírica normativa. O presente trabalho é classificado como uma pesquisa empírica descritiva, afinal, busca desenvolver um modelo descritivo para o entendimento de processos reais.

3.2 CRISP-DM

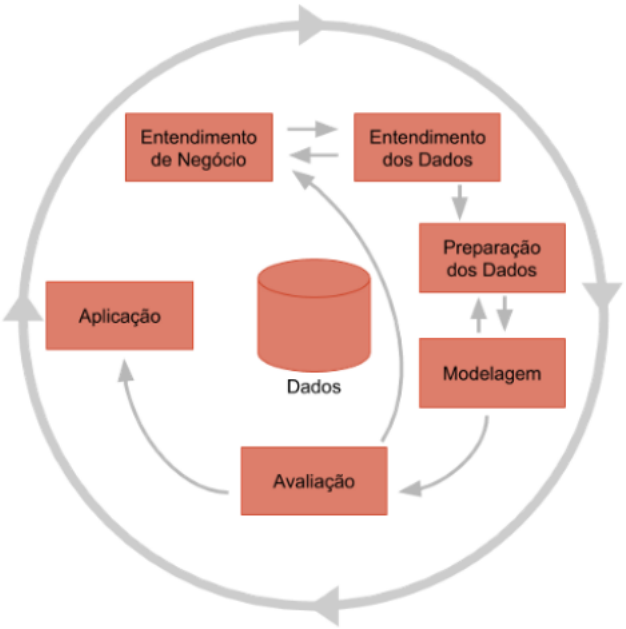
A sigla CRISP-DM significa *CRoss-Industry Standard Process for Data Mining* ou, em português, Padrão Intersetorial de Processo para Mineração de Dados é uma metodologia de mineração de dados que, segundo Shearer (2000), segue seis etapas:

1. Entendimento do negócio;
2. Entendimento dos dados;
3. Preparação dos dados;

4. Modelagem;
5. Avaliação; e
6. Aplicação.

Essa metodologia para a mineração de dados pode ser vista no esquema da Figura 4. As setas indicam a sequência do processo, indicando que algumas etapas podem ser feitas de forma concomitante, como o entendimento do problema e dos dados. Ademais, a etapa de avaliação é importante para entender os próximos passos do projeto, já que, resultados ruins podem indicar uma má interpretação nas primeiras etapas. Por último, o círculo em volta do esquema indica a natureza cíclica do CRISP-DM, afinal, é esperado que a aplicação da mineração de dados traga resultados, sendo um deles, o desenvolvimento de novos projetos.

Figura 4 – Fluxograma da Metodologia

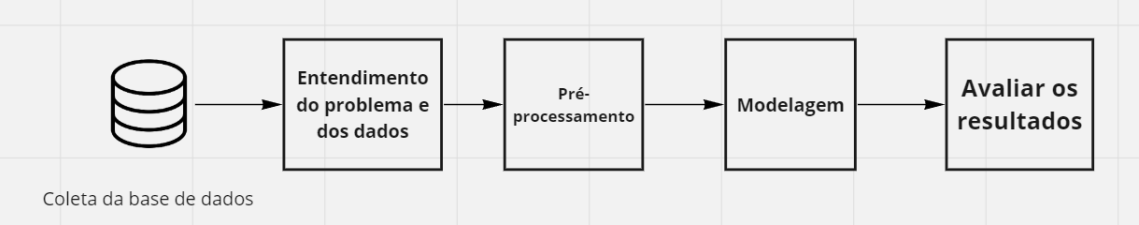


Fonte: Adaptado de Shearer (2000).

3.2.1 Fluxograma da metodologia

A Figura 5 como será aplicada a metodologia CRISP-DM no presente trabalho. Na etapa de entendimento do problema e dos dados, a base de dados da Seção de Ensino será separada em quatro bases de dados, uma para cada curso ofertado no Icea. O pré-processamento será feito em seguida, com a limpeza e simplificação das bases já separadas. A etapa de modelagem é a etapa de clusterização, que será discutida nos resultados.

Figura 5 – Fases do CRISP-DM



Fonte: Elaborado pelo autor.

3.3 Coleta de dados

A Seção de Ensino do Icea possui uma base de dados muito diversa que coleta informações dos alunos ingressantes período a período. As fontes de dados apresentadas podem se resumir a duas, que são os dados gerais, que são as informações de todos os alunos coletadas para a obtenção da matrícula e a base de dados de notas, em que uma linha é uma disciplina onde o aluno foi matriculado. Como alguns desses dados são sensíveis, as informações que não devem ser divulgadas foram anonimizadas antes de serem disponibilizadas para o trabalho, em atendimento ao disposto na Lei Geral de Proteção de Dados Pessoais (LGPD), (BRASIL, 2018).

Foram cinco bases de dados fornecidas pela Seção de Ensino da Ufop, sendo elas: Uma base de dados contendo informações de todos os alunos que já passaram pelo Icea desde 2001, com os seguintes campos no Quadro 3.1.

Quadro 3.1 – Banco dos dados dos alunos

Nome das colunas	**Tipo de dado**
ano de admissão, ano de diplomação, ano de nascimento	Datetime
mask, sexo, código do modo de admissão, descrição modo de admissão, origem, turno, código da situação aluno, descrição situação aluno, código do curso, código do curso admissão, curso, código do currículo, código do habilitação, código da ênfase, código do polo, polo, modalidade concorrência, modalidade concorrência homologada, aluno computado censo, participou política afirmativa, usou política afirmativa	Categórica
pontuação no vestibular, carga horária do curso, carga horária cursada, ano de nascimento, semestre de admissão, semestre de diplomação	Numérica

Fonte: Elaborado pelo autor.

Quatro bases de dados, uma para cada curso ofertado no Icea, com informações do desempenho dos alunos nas disciplinas ofertadas pelos departamentos da instituição, com os campos descritos no Quadro 3.2.

Quadro 3.2 – Banco de Notas

Nome das colunas	**Tipo de dado**
ano, ano de nascimento	Datetime
mask, cor da pele, sexo, código do curso, caráter, situação, descrição do modo de admissão, tipo de escola, code, mask, código do departamento, código da disciplina,	Categórica
média final, exame especial, aula dada, faltas, ano de nascimento, semestre	Numérica

Fonte: Elaborado pelo autor.

Essas bases de dados foram coletadas no começo do primeiro semestre letivo de 2023, portanto, a base de alunos contém dados sociais de todos os alunos matriculados até 2023 e a base de notas contém as informações acadêmicas dos alunos que cursaram disciplinas até o fim do segundo semestre letivo de 2022.

3.4 Pré-processamento

O pré-processamento de bases de dados é uma etapa fundamental para garantir a qualidade e consistência dos dados utilizados em análises e estudos. Nesse sentido, são realizadas diversas operações, tais como limpeza, transformação, integração e redução de dados, visando preparar as bases para uma análise mais eficiente.

As bases de notas dos cursos foram agregadas para simplificar o pré-processamento, já que as colunas são as mesmas e, assim, os cálculos são feitos apenas uma vez na base inteira. Para a base de dados de alunos da universidade, as seguintes operações foram realizadas:

- mudança de nome de colunas para facilitar o tratamento em Python;
- remoção de colunas desnecessárias para a análise;
- exclusão dos alunos que ingressaram antes da padronização do vestibular através do Exame Nacional do Ensino Médio (Enem);
- concatenação com a base de notas a partir do número de matrícula mascarado (mask).

Já para as bases de dados de notas, as seguintes etapas foram realizadas:

- mudança de nome de colunas para facilitar o tratamento em Python;
- remoção de colunas desnecessárias para a análise;
- exclusão dos alunos que ingressaram antes da padronização do vestibular através do Enem;

- padronização de formato de dados para a consistência na manipulação;
- criação de um Id de semestre, para facilitar os cálculos de comparação entre diferentes semestres;.
- correção dos períodos letivos especiais (como, por exemplo, o semestre 18.3) para normalizar o número de semestres no ano;
- remoção de disciplinas que foram aproveitadas de outras universidades em caso de transferência.

Por fim, a base de alunos ficou com as colunas apresentadas no Quadro 3.3 e a base de notas ficou com as colunas descritas no Quadro 3.4.

Quadro 3.3 – Banco de Alunos pós limpeza

Nome das colunas	**Tipo de dado**
ano de admissão, ano de diplomação, ano de nascimento	Datetime
mask, sexo, código do modo de admissão, descrição modo de admissão, origem, turno, código da situação aluno, descrição situação aluno, código do curso, código do curso admissão, curso, código do currículo, código do habilitação, código da ênfase, código do polo, polo, modalidade concorrência, modalidade concorrência homologada, aluno computado censo, participou política afirmativa, usou política afirmativa, cor da pele, tipo de escola	Categórica
pontuação no vestibular, carga horária do curso, carga horária cursada, ano de nascimento, semestre de admissão, semestre de diplomação, porcentagem cursada, idade que entrou na Ufop	Numérica

Fonte: Elaborado pelo autor.

Quadro 3.4 – Banco de Notas pós limpeza

Nome das colunas	**Tipo de dado**
ano, ano de nascimento	Datetime
mask, cor da pele, sexo, código do curso, caráter, situação, descrição do modo de admissão, tipo de escola, code, mask, código do departamento, código da disciplina,	Categórica
média final, exame especial, aula dada, faltas, ano de nascimento, semestre	Numérica

Fonte: Elaborado pelo autor.

3.5 Análises descritivas das bases

Esta Seção contém informações observadas através da análise descritiva das bases de dados fornecidas pelo Icea.

3.5.1 Análise descritiva dos atributos sociais

Algumas observações presentes na base de dados do aluno que não cursaram nenhuma disciplina, ou seja, se matricularam na universidade e saíram antes mesmo de concluir o primeiro semestre, isso pode ser um problema caso o aluno evada depois do prazo limite para chamadas de vagas remanescentes. Separando a base dos alunos que não cursaram nada e a base dos alunos que cursaram (Cursou / Não cursou), é possível comparar informações quantitativas, como por exemplo, as notas do vestibular e a idade em que o aluno entrou na universidade. Além disso, é possível descrever o perfil dos alunos a partir de variáveis categóricas fornecidas nas bases de dados de alunos, por exemplo, sexo, o curso em que o aluno se matriculou, a modalidade de concorrência (para alunos que ingressaram através do vestibular), dentre outros. A diferença no total de observações das colunas se dá pelos alunos que ingressaram na universidade por transferência externa, obtenção de novo título ou alguma outra modalidade que não o Enem.

Tabela 1 – Estatística descritiva dos alunos que não cursaram nenhuma disciplina

Estatística	**Pontuação Enem**	**Idade que entrou**
Número de observações	251	268
Média	640.85	20.73
Desvio padrão	66.33	4.03
Mín	393.60	18
25%	602.55	18.75
50%	652.20	19.5
75%	689.95	21
Máx	764.80	52

Fonte: Elaborado pelo autor.

Tabela 2 – Estatística descritiva dos alunos que cursaram pelo menos uma disciplina

Estatística	**Pontuação Enem**	**Idade que entrou**
Número de observações	3816	4115
Média	645.60	21.46
Desvio padrão	47.38	4.64
Mín	436.40	16
25%	617.47	19
50%	646.75	20
75%	677.52	22
Máx	774.60	62

Fonte: Elaborado pelo autor.

É difícil diferenciar os alunos que não cursaram nenhuma disciplina dos alunos que cursaram. As métricas, tanto de idade quanto de pontuação no vestibular, parecem se comportar de forma quase semelhante e, portanto, faz se necessária uma abordagem diferente.

A Tabela 3 mostra os tipos de cotas e a participação de cada um deles nas bases de dados. Vale ressaltar que as modalidades PAA foram extintas e deram lugar às modalidades L_n, mas englobam o mesmo grupo de pessoas.

Tabela 3 – Descrição de cotas por tabela

Tipo de cota	**Modalidade de Concorrência**	**% na base Cursou**	**% na base Não Cursou**
Ampla Concorrência	AC	54.41%	52.58%
Ensino médio em escola pública sem comprovação de renda	L5, PAA3	7.17%	7.17%
Ensino médio em escola pública com comprovação de renda	L1, PAA4	8.09%	11.15%
Cota racial sem comprovação de renda	L6, PAA2	10.60%	12,74%
Cota racial com comprovação de renda	L2, PAA1	10.72%	15.14%
Política de Ação Afirmativa (até 2012)	PAA	8.80%	0.39%
Candidatos com deficiência	L9, L10, L13, L14	0.16%	0.79%

Fonte: Elaborado pelo autor.

Nota-se uma dominância da Ampla Concorrência nos dois cortes da base de dados e uma presença maior, em proporção, de alunos autodeclarados negros (pretos ou pardos) ou indígenas, que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas na porção de alunos que não cursou nada e evadiu.

A ausência da categoria PAA, que a partir de 2012 foi dividida em várias classes, só evidenciou o efeito da evasão de alunos que não chegaram a cursar alguma disciplina e usaram cotas. Portanto, apesar das políticas de cotas e das bolsas fornecidas pela Ufop, fica evidente a dificuldade de alunos que estão em um grupo de vulnerabilidade social em cursar o Ensino Superior.

3.5.2 Análise descritiva da situação dos alunos na base

A base de alunos fornecida pela Ufop conta com a categoria em que o aluno se encontra no processo de formação, são:

- Afastado: suspensão da matrícula e das obrigações acadêmicas, que pode ser requeria apenas uma vez pelo aluno, em face de situações especiais, devidamente comprovadas, ao Colegiado de Curso, por um prazo de até 4 (quatro) anos.
- Diplomado;
- Evadido;
- Matriculado;
- Mobilidade;
- Trancado: o trancamento é a suspensão, durante o semestre letivo, da matrícula em uma ou mais disciplinas. Quando abranger todas as disciplinas, recebe o nome de trancamento total de matrícula, e neste caso, o período correspondente não será contado no tempo de permanência do aluno nesta Universidade.

A Tabela 4 descreve a presença de cada categoria na base de dados.

Tabela 4 – Descrição da situação dos alunos na base de alunos

Descrição situação aluno	**Quantidade absoluta**	**Porcentagem**
Evadido	2512	54.96%
Matriculado	1268	27.74%
Diplomado	736	16.10%
Trancado	34	0.74%
Afastado	10	0.22%
Mobilidade	4	0.08%

Fonte: Elaborado pelo autor.

Dentre os diplomados, é possível entender o tempo médio em semestres para a diplomação na universidade. É importante salientar que os cursos de engenharia têm 10 semestres e o curso de Sistemas de Informação tem 8 semestres e que há alunos que se graduaram em poucos semestres pois entraram na Ufop através de transferência ou reingressaram através do Enem, aproveitando as disciplinas que cursaram anteriormente. A Tabela 5 mostra o comportamento dos alunos diplomados no que diz respeito ao número de semestres feitos até a diplomação.

Tabela 5 – Estatística descritiva do tempo para a diplomação no Icea

Estatística	**Semestres até a diplomação**
Número de observações	736.00
Média	11.22
Desvio padrão	2.51
Mín	2.00
25%	10.00
50%	11.00
75%	13.00
Máx	19.00

Fonte: Elaborado pelo autor.

3.6 Construção dos atributos acadêmicos

A partir do pré-processamento, é possível incluir atributos que podem metrificar a jornada de todos os alunos no Icea. Isso é importante para agrupar as informações acadêmicas dos estudantes de forma a não enviesar o algoritmo da clusterização.

3.6.1 Atributos

Com o objetivo de ter apenas um aluno por linha no conjunto de dados a ser clusterizado, foi criada uma tabela intermediária, exemplificada no Quadro 3.5 contendo apenas as informações: aluno, código da disciplina, maior nota ao cursar a disciplina e quantidade de vezes que cursou a disciplina. O Quadro 3.5 é um exemplo genérico do que seria uma linha dessa base de dados, que possui as colunas: Mask, Código da Disciplina (**CD**), Nota Final (**NF**), Quantidade Cursada (**QC**) e Horas Cursadas (**HC**). Na representação, v_k^{ij} representa o valor do indicador (coluna) k ($k \in \{nf, qc, hc\}$) para o estudante i na disciplina j.

Quadro 3.5 – Estrutura da base antes de rotacionar a tabela

Mask	**CD**	**NF**	**QC**	**HC**
...	...	...	...	...
$mask_i$	COD_{ij}	v_{nf}^{ij}	v_{qc}^{ij}	v_{hc}^{ij}
$mask_i$	COD_{ij+1}	v_{nf}^{ij+1}	v_{qc}^{ij+1}	v_{hc}^{ij+1}
$mask_i$	...	...	...	...
...	...	...	...	...

Fonte: Elaborado pelo autor.

A tabela representada no Quadro 3.5 foi rotacionada para que haja apenas um aluno por linha na base para a clusterização. É importante salientar que, se a base contiver 100 disciplinas diferentes, a nova tabela possuirá 100*3 colunas para cada mask. O quadro 3.6 é um exemplo de como a tabela do Quadro 3.5 fica depois de rotacionada

Quadro 3.6 – Estrutura da base depois da rotação

Mask	$\cdots$	NF_j	QC_j	HC_j	NF_{j+1}	QC_{j+1}	HC_{j+1}	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$mask_i$	$\cdots$	v_{nf}^{ij}	v_{qc}^{ij}	v_{hc}^{ij}	v_{nf}^{ij+1}	v_{qc}^{ij+1}	v_{hc}^{ij+1}	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Fonte: Elaborado pelo autor.

O resultado é uma base que contém uma linha por aluno e colunas referentes ao desempenho dele em todas as disciplinas na base. Para disciplinas que o aluno não cursou, as informações de nota, quantidade cursada e horas cursadas serão 0.

3.6.2 Estatísticas de atributos acadêmicos

Os atributos mencionados na Seção 3.6 foram criados para entender a jornada de cada aluno durante o período acadêmico, no entanto, ainda há uma diferença muito grande entre alunos de mesmo curso. Por isso, algumas informações foram usadas para calcular atributos (*features* acadêmicos - FAs) que agrupassem as informações do aluno independentemente de quais disciplinas ele cursou no curso.

1. AP: total de disciplinas em que o aluno foi aprovado no histórico;
2. RPN: total de disciplinas em que o aluno foi reprovado por nota no histórico;
3. RPF: total de disciplinas em que o aluno foi reprovado por falta no histórico;
4. RNF: total de disciplinas em que o aluno foi reprovado por nota e falta no histórico;
5. TR: total de disciplinas trancadas no histórico;
6. CA: total de disciplinas canceladas no histórico;
7. APD: taxa de aprovação nas matérias do Departamento de Ciências Exatas e Aplicadas (Decea);

$$APD = \frac{\text{\# disciplinas aprovadas no DECEA}}{\text{\# disciplinas cursadas no DECEA}}$$

8. SEM: máximo número de semestres cursados;
9. CH/P: distorção carga horária período;

$$CH/P = \frac{\text{carga horária completa}}{\text{300*número de semestres cursado}}$$

Os atributos taxa de aprovação nas matérias do Decea e distorção carga horária período têm valor ótimo igual a 1, que significa, respectivamente, que o aluno não reprovou em nenhuma disciplina do Decea e que não está com períodos atrasados no curso.

Unindo esses atributos à base rotacionada, é possível agrupar os alunos levando em consideração as disciplinas cursadas, mas também, o desempenho geral no curso. Com isso é possível obter uma base com as informações descritas no Quadro 3.7.

Quadro 3.7 – Tabela final para a clusterização

index	**QC**	**NF**	AP	RPN	RPF	RNF	TRA	CA	APD	SEM	CH/P
Mask											

O Quadro 3.7 é um exemplo de como ficou o conjunto de dados usado na clusterização. Como a clusterização será feita curso a curso, cada curso teve um conjunto de dados diferente e com diferentes dimensões. A Tabela 6 mostra a dimensão das bases de dados de cada curso.

Tabela 6 – Dimensão das bases por curso

Curso	**Linhas**	**Colunas**
Engenharia Elétrica	1026	218
Engenharia de Computação	1016	252
Engenharia de Produção	1065	282
Sistemas de Informação	1010	358

4 Resultados

As análises do presente trabalho serão feitas curso a curso, e, como mostrado na Seção 3.3 as bases de dados de alunos e de notas foram separadas e analisadas individualmente. Portanto, este capítulo trará os resultados por curso.

4.1 Engenharia Elétrica

Nas Seções a seguir, serão tratados e discutidos os dados do curso de Engenharia Elétrica.

4.1.1 Atributos pré universidade

A Tabela 7 apresenta as estatísticas descritivas dos atributos Idade em que entrou na Universidade e nota no Enem.

Tabela 7 – Idade e Notas de Enem do curso

Estatística	**Pontuação Enem**	**Idade que entrou**
Média	647.96	21.66
Desvio padrão	50.19	4.98
Mínimo	479.30	17.00
25%	622.40	19.00
50%	649.10	20.00
75%	679.40	22.00
Máximo	760.60	52.00

Fonte: Elaborado pelo autor.

A Tabela 8 apresenta a frequência de valores para os atributos categóricos sexo e uso de ações afirmativas. É possível traçar um perfil esperado do aluno egresso no curso por meio dos atributos descritos nas Tabelas 7 e 8.

Tabela 8 – Sexo e uso de cotas no curso

Sexo	**Usou política afirmativa**	**Quantidade**
M	Não	292
M	Sim	265
F	Não	64
F	Sim	64

Fonte: Elaborado pelo autor.

4.1.2 Atributos Acadêmicos

Como mencionado no Capítulo 3.6, todos os alunos tiveram informações agregadas para a clusterização. Essas informações são usadas para entender a média e a moda dos atributos acadêmicos dos alunos deste curso. Na Seção 3.6.2 estão os significados de cada atributo.

Tabela 9 – Estatística descritiva dos atributos do curso

Estatística	*AP*	*RPN*	*APD*	*SEM*	CH/P
Média	19.34	5.97	0.42	6.39	0.31
Desvio padrão	20.83	6.95	0.37	4.79	0.31
Mín	0.00	0.00	0.00	1.00	0.00
25%	1.00	0.00	0.00	2.00	0.00
50%	10.00	4.00	0.41	5.50	0.23
75%	36.00	9.00	0.76	10.00	0.57
Máx	66.00	40.00	1.00	21.00	1.00

Fonte: Elaborado pelo autor.

Unindo as Tabelas 8 e 9 tem-se o perfil do aluno do curso, segundo as bases de dados disponibilizadas pela Universidade.

Portanto, um perfil esperado do aluno que entra no curso de Engenharia Elétrica no Icea se dá por:

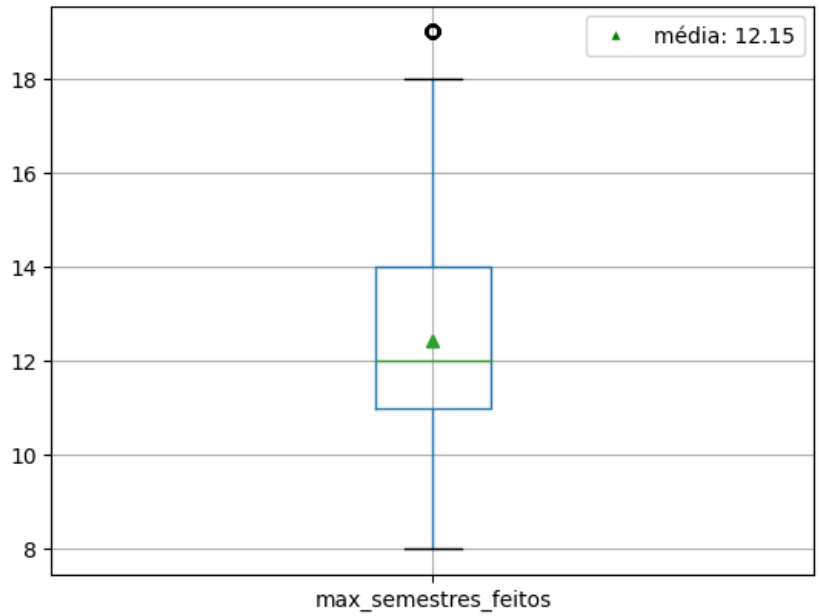
- sexo: masculino
- pontuação no Enem: 647.96
- idade que entrou: 22
- usou política afirmativa: não
- aprovações (*AP*): 19.34
- reprovações (*RPN*): 5.97
- taxa de aprovação no Decea (*APD*): 0.42
- semestres feitos (*SEM*): 6.39
- distorção carga horária período (CH/P): 0.31

Como esperado, há mais homens que mulheres no curso, replicando a realidade do Icea. Também é visível o quanto os atributos taxa de aprovação no Decea e a distorção carga horária período estão longe de um cenário idealizado.

4.1.3 Tempo para a diplomação

O curso de Engenharia Elétrica tem cinco anos de duração, ou seja, dez semestres. A Figura 6 mostra o comportamento da base de alunos do curso que foram diplomados até a coleta de dados, excluindo os alunos que se formaram em menos de oito períodos, já que são, provavelmente, alunos que reingressaram através do Enem ou alunos egressos por transferência.

Figura 6 – *Box plot*: Número de semestres feitos pelos alunos diplomados



Fonte: Elaborado pelo autor.

É notável que menos de 25% dos alunos se formam dentro do tempo esperado de diplomação no curso de Engenharia Elétrica. Vê-se também como a média se comporta e como os quartis se distanciam no gráfico.

4.1.4 Situação do aluno

A Tabela 10 mostra a participação de cada categoria de situação do aluno na base de alunos do curso de Engenharia Elétrica.

Tabela 10 – Situação dos Alunos na Base do curso

Descrição situação aluno	**Total**	**%**
Evadido	575	56.04
Matriculado	259	25.24
Diplomado	183	17.83
Trancado	8	0.78
Afastado	1	0.10

Fonte: Elaborado pelo autor.

Na Tabela 10 já fica evidente a alta taxa de evasão da Engenharia Elétrica no Icea. Essa porcentagem poderia ser ainda maior se os alunos matriculados não fossem levados em consideração. Ainda assim, mesmo com 54.05% dos alunos evadidos e apenas 17.8% dos alunos diplomados até a coleta da base de dados, o curso não é o líder em evasão no Icea.

4.1.5 Análise de Componente Principal

A base de dados com os atributos dos alunos da Engenharia Elétrica continha 1026 linhas e 217 colunas além do mask, e, para explicar 80% da variância total, foram necessárias 23 componentes. Usando apenas essas três componentes, a nova base explica 53.10% da variância da base de dados que foi usada para o PCA.

4.1.5.1 Scores features x componentes

No mapa de calor mostrado na Figura 7, vemos a relação das componentes 0, 1 e 2 com as colunas originais da base de dados de atributos. Por serem muitas colunas, o peso da componente (score) vai de, aproximadamente, -0.2 a 0.2. Os extremos são os que contém maior variabilidade de dados e, por isso, contribuem mais para explicar a variância da base de dados.

Na primeira componente, os valores mais escuros no mapa de calor se dão pelas notas de disciplinas específicas do curso, e o maior valor é a nota final na disciplina CEA570.

Na segunda componente, há uma presença maior de valores negativos, mas que também são referentes às notas dos alunos nas disciplinas específicas do curso de Engenharia Elétrica. O maior valor se dá pela nota final em CEA597 e o menor valor se dá pela nota final na disciplina CEA743.

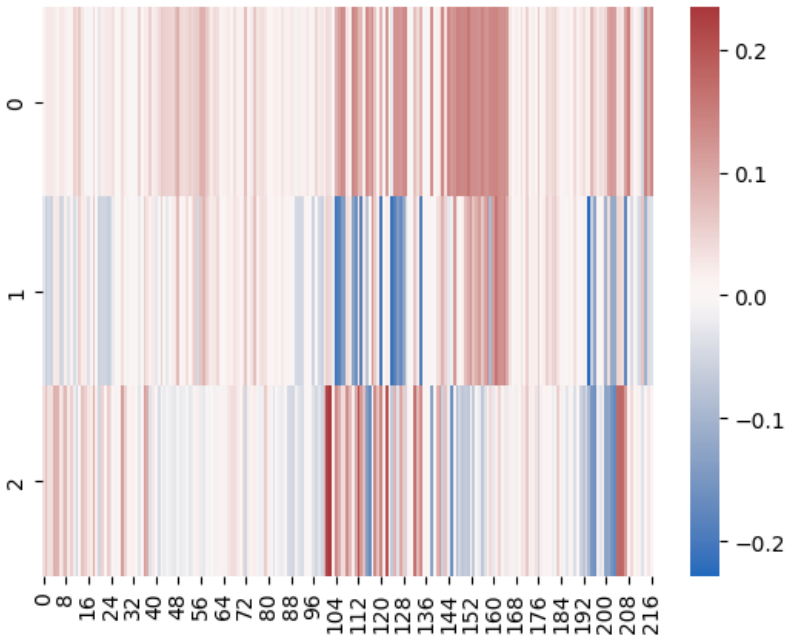
Por último, a terceira componente tem um o maior valor dentre os mostrados no mapa de calor, que é a coluna quantidade cursada EAD702, e o menor valor desta componente é a coluna que mostra a nota final na disciplina EAD344.

É importante salientar que essas disciplinas são as que mais contribuem para a variabilidade da base de dados, mas isso não significa que sejam as disciplinas fundamentais para o sucesso ou o insucesso do aluno ao longo do curso.

4.1.6 Clusterização

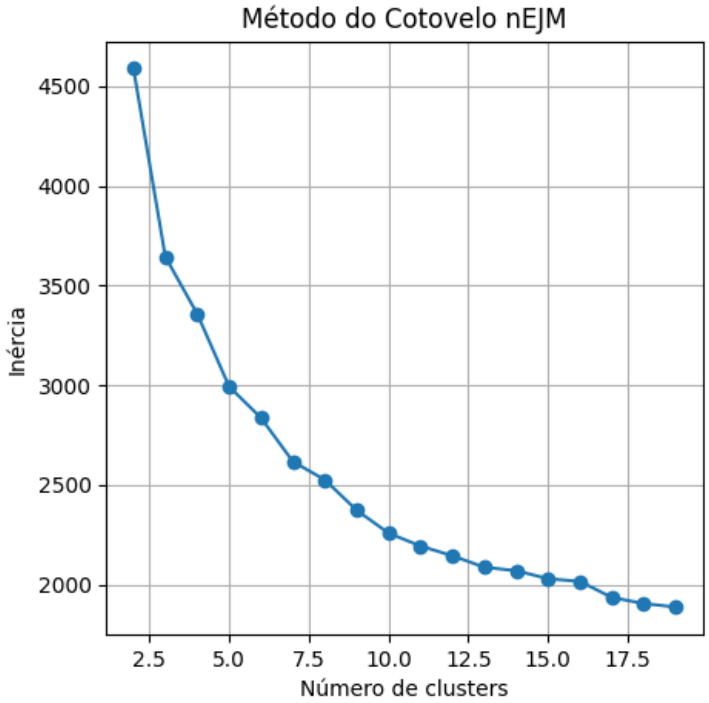
A base com os valores transformados após a aplicação do PCA possui 23 colunas que foram utilizadas para a clusterização. Utilizamos o método da silhueta e o método do cotovelo para definir qual o melhor número possível de clusters para essa base. O método do cotovelo, que consiste em calcular a soma das distâncias quadráticas dos dados intra-clusters, é mostrado na Figura 8

Figura 7 – Mapa de calor: Explicando os componentes da ACP



Fonte: Elaborado pelo autor.

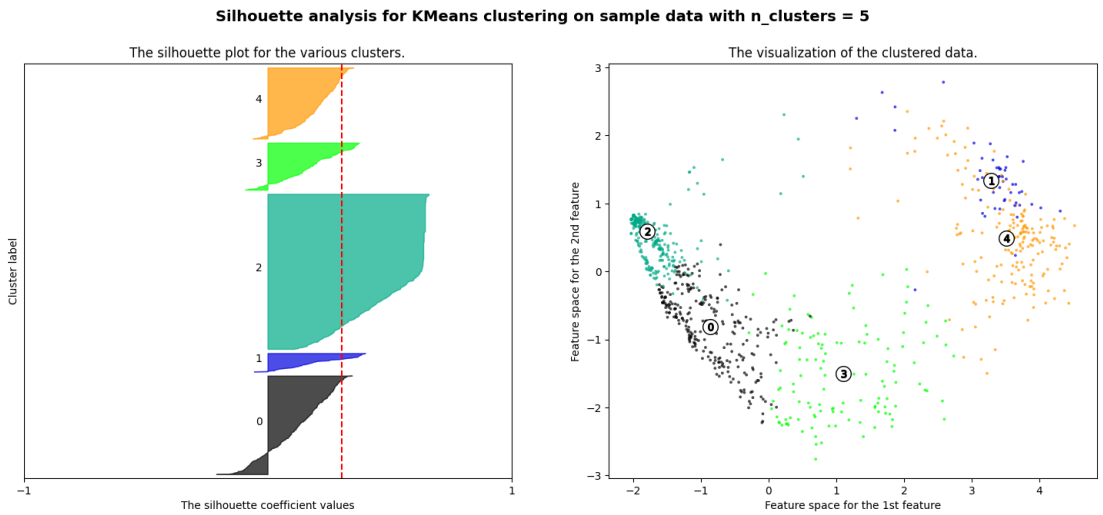
Figura 8 – Método do cotovelo



Fonte: Elaborado pelo autor.

No gráfico mostrado na Figura 8, é possível ver a redução das distâncias com o aumento do número de clusters. Com isso, é visível o ganho de 4 para 5 clusters e o ganho de 6 para 7 clusters e por isso, serão os valores testados no método da silhueta.

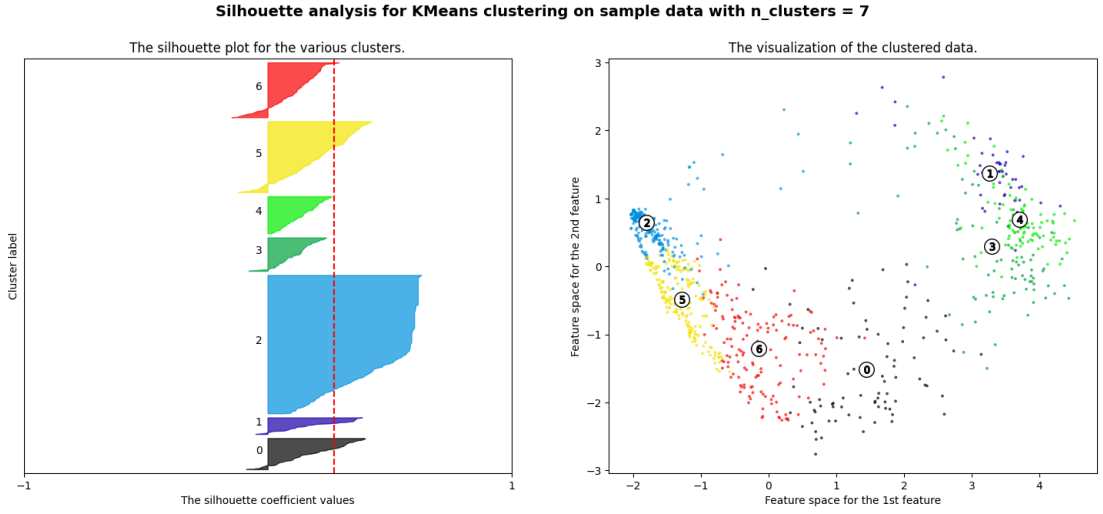
Figura 9 – Método da Silhueta para 5 clusters e seu *scatterplot*



Fonte: Elaborado pelo autor.

Com 5 clusters, todos os valores de coeficiente da silhueta tocam a média de coeficiente, traçada em vermelho na Figura 9. Como discutido na Seção 2.7, a proximidade do coeficiente da silhueta dos clusters com a média de coeficiente é importante para entender a qualidade dos agrupamentos. O posicionamento dos Clusters considerando as duas primeiras componentes é apresentado no *scatterplot* à direita.

Figura 10 – Método da Silhueta para 7 clusters e seu *scatterplot*



Fonte: Elaborado pelo autor.

Por outro lado, com 7 clusters, ver Figura 10, os valores de coeficiente da silhueta do Cluster 3 não tocam a linha média de coeficiente da silhueta.

Portanto, o melhor número de clusters para essa base de dados é 5. Aplicando esse número de clusters ao algoritmo k-means, temos a divisão da base, mostrada na Tabela 11.

Tabela 11 – Numero de observações por cluster

Cluster	Quantidade de alunos
0	369
1	176
2	142
3	248
4	64

Fonte: Elaborado pelo autor.

4.1.7 Análise estatística dos clusters

Esta Seção trata dos atributos que não foram usados para a clusterização a fim de relacionar com os atributos usados e os resultados obtidos.

A Tabela 12 mostra a distribuição étnica nos Clusters da Engenharia Elétrica. É visível a participação dominante de alunos brancos e pardos no curso, o que também acontece no Icea.

Tabela 12 – Etnia por cluster

Cluster	Amarela %	Branca %	Indígena %	Não Declarado %	Parda %	Preta %
0	0.51	35.86	0.00	5.56	45.20	12.88
1	2.27	35.23	0.57	2.27	53.41	6.25
2	2.11	30.28	0.00	2.82	49.30	15.49
3	0.40	40.32	0.81	0.81	44.35	13.31
4	0.00	45.31	0.00	3.12	45.31	6.25

Fonte: Elaborado pelo autor.

A Tabela 13, por sua vez, mostra a porcentagem de alunos de cada Cluster que usou política afirmativa, independente de qual categoria.

Tabela 13 – Uso de política afirmativa por Cluster

Cluster	Não %	Sim %
0	55.31	44.69
1	54.46	45.54
2	48.04	51.96
3	48.63	51.37
4	46.15	53.85

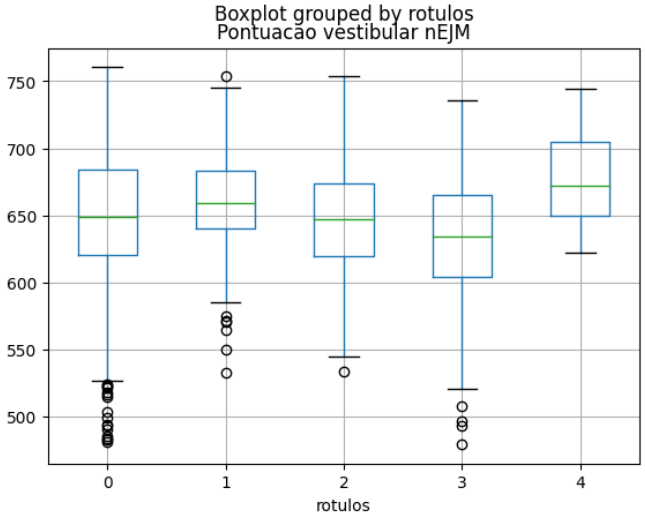
Fonte: Elaborado pelo autor.

Os Clusters 2, 3 e 4 têm uma participação um pouco maior de alunos que entraram na Universidade através do uso de cotas.

4.1.7.1 Nota no Enem

A partir das Tabelas 12 e 13, é possível obter uma relação de cor de pele e modalidades de concorrência para cada cluster. Essas informações podem ser relevantes para entender como os alunos dos Clusters se saíram no Enem. O box plot da Figura 11 mostra como os Clusters se comportam nesse quesito.

Figura 11 – Box plot: Nota no Enem por Cluster



Fonte: Elaborado pelo autor.

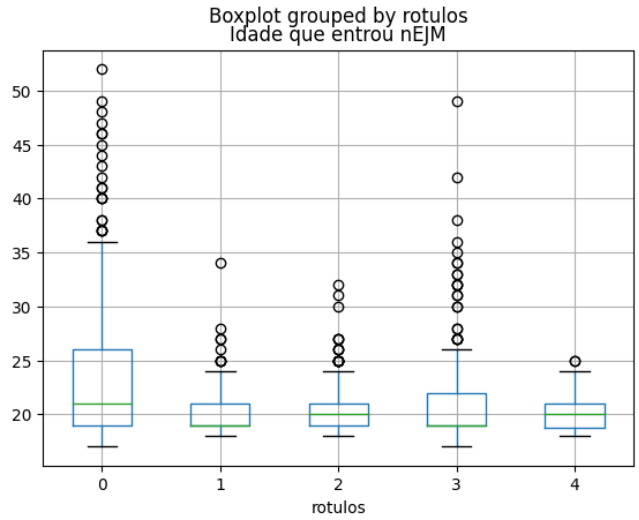
No box plot mostrado na Figura 11 é possível ver a dispersão maior nos Clusters 0 e 3, sendo o Cluster 3 com a menor média nas notas no Enem. Olhando a Tabela 13, o Cluster 3 tem mais da metade de alunos que usaram políticas afirmativas. Por outro lado, o Cluster 4, que é o Cluster com maior porcentagem de alunos que usaram políticas auto afirmativas, tem uma dispersão muito menor que os outros Clusters e a maior média de notas no Enem.

4.1.7.2 Idade que entrou

O box plot da Figura 12 mostra a distribuição da idade dos alunos quando ingressaram na universidade. Esse é um cálculo aproximado, já que a base de dados fornece apenas o ano em que o aluno entrou na universidade e seu ano de nascimento.

De novo, há uma diferença visível dos Clusters 0 e 3 para o resto dos clusters. Ambos têm uma dispersão maior no que diz respeito à idade em que o aluno entrou na universidade, apesar da média não ser tão diferente da média dos outros clusters. Essa informação pode justificar a alta dispersão nas notas no Enem e, consequentemente, no desempenho acadêmico dos clusters. Há uma relação, então, da idade em que o aluno ingressa na universidade com o desempenho acadêmico dele, já que até o momento de coleta dessa base, o aluno mais velho a se formar no curso entrou no curso com 34 anos em 2018 e se formou em 2022, com quase 40 anos.

Figura 12 – Box plot: Idade por Cluster

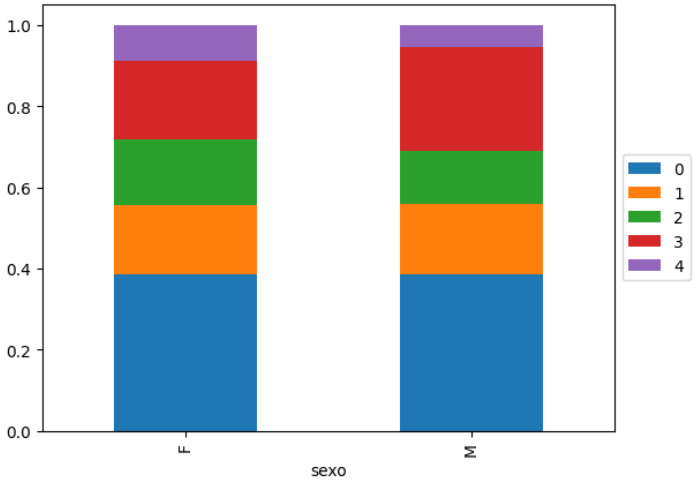


Fonte: Elaborado pelo autor.

4.1.7.3 Sexo

É sabido que há um número muito maior de homens do que de mulheres, não só no Icea, mas em cursos de tecnologia em geral, por isso, para uma análise mais clara, o gráfico da Figura 13 é a porcentagem dos Clusters em cada sexo.

Figura 13 – Gráfico de barras empilhadas: Porcentagem dos Clusters por sexo



Fonte: Elaborado pelo autor.

Os Clusters 0 e 1 têm proporções similares de homens e mulheres em relação ao todo dessas populações, mas há uma diferença nos outros clusters, chamando a atenção para a baixa participação de mulheres no Cluster 3 comparando com os homens e, em contrapartida, a alta participação das alunas no Cluster 4 comparando com os alunos.

Com essas informações a Tabela 14 traça um perfil social de cada cluster, usando as médias das variáveis numéricas e a moda das variáveis categóricas.

Tabela 14 – Perfil social por cluster

Cluster	**Sexo**	**Cor da Pele**	**Nota no Enem**	**Desvio padrão**	**Idade em que entrou**	**Desvio padrão**	**Usou política afirmativa ?**
0	Masc.	Parda (22.6%)	648.02	57.43	23.43	6.45	Não
1	Masc.	Parda (26.7%)	660.32	40.15	20.16	2.36	Não
2	Masc.	Parda (24.6%)	645.58	44.33	20.56	2.69	Sim
3	Masc.	Parda (22.16%)	633.72	46.30	20.96	4.37	Sim
4	Masc.	Parda / Branca (22.66%)	676.76	33.30	19.92	1.78	Sim

Fonte: Elaborado pelo autor.

Como esperado, há uma presença muito maior de homens e de pessoas pardas nos clusters, que é apenas um reflexo do curso como um todo. Nos clusters, não há uma relação aparente entre o uso de política afirmativa e a nota do Enem.

4.1.8 Estatísticas de atributos acadêmicos

A Tabela 15 agrega, por cluster, alguns atributos que foram usados para a clusterização.

Tabela 15 – Informações acadêmicas médias por cluster

Cluster	AP	RPN	RPF	APD	SEM	CH/P
0	1.54	1.89	2.24	0.06	2.09	0.02
1	52.15	9.22	1.27	0.86	12.47	0.75
2	27.96	11.39	3.63	0.71	10.04	0.46
3	10.71	6.43	2.77	0.41	5.39	0.23
4	53.56	8.55	0.92	0.82	12.00	0.79

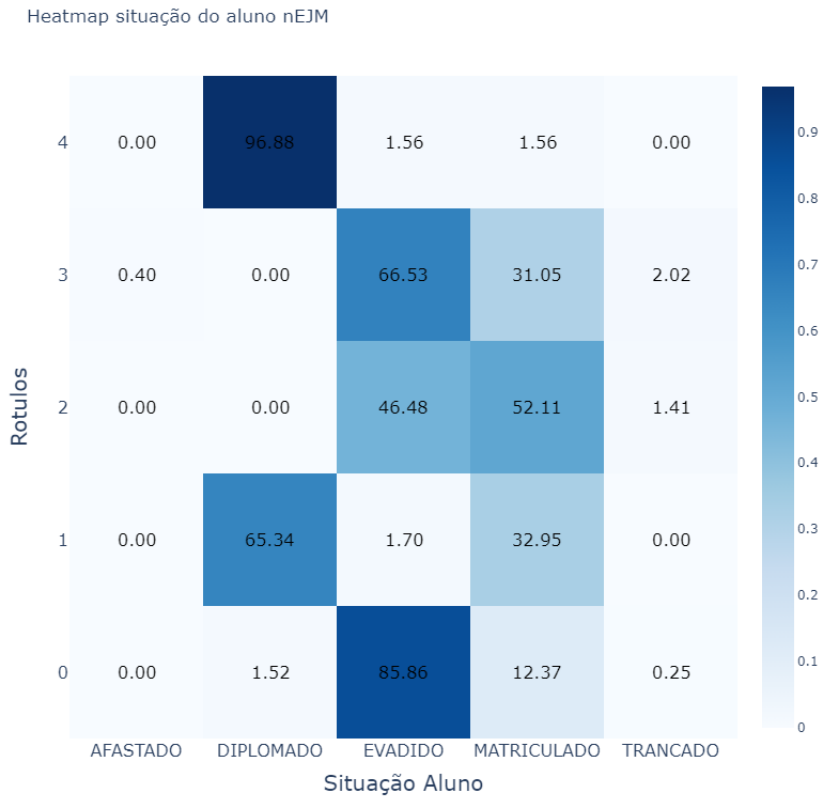
Fonte: Elaborado pelo autor.

A Tabela 15 mostra que os atributos escolhidos para a clusterização foram suficientes para separar os perfis de alunos no curso de Engenharia Elétrica. Com essas estatísticas traçamos um perfil acadêmico médio de cada Cluster na Engenharia Elétrica. Os Clusters 1 e 4 são Clusters com maior sucesso acadêmico e com o maior número de diplomados pelo alto número médio de aprovações e de semestres feitos. Por outro lado, os Clusters 0 e 3 tem um média muito baixa de disciplinas aprovadas frente ao número médio de semestres feitos, e isso se torna preocupante visto que são os Clusters mais populosos. O Cluster 2, por sua vez, é um Cluster médio, provavelmente populado por alunos que se encaixam no fenômeno de retenção. Esses alunos têm uma taxa de aprovação razoável nas disciplinas do Decea, mas mostram uma distorção de carga horária/período muito mais alta que a dos Clusters 0 e 3.

4.1.8.1 Estatística da situação do aluno

No mapa de calor na Figura 14 vemos as situações de aluno por Cluster e, assim, podemos comparar com as Tabelas 14 e 15.

Figura 14 – Mapa de calor: Clusters x Situação de aluno



Fonte: Elaborado pelo autor.

O gráfico da Tabela 14 é um mapa de calor que relaciona a situação dos alunos com o Cluster em que eles estão localizados. Cada célula é a porcentagem de alunos na situação descrita presente no Cluster e a soma das linhas tem resultado 100%.

Apenas pelo mapa de calor fica evidente a separação entre os alunos diplomados e os alunos evadidos. Em Clusters onde há uma quantidade significativa de diplomados, há uma quantidade muito pequena de evadidos e o contrário também é verdade. Os Clusters 0 e 3, como esperado, são Clusters que contém a maior porcentagem da evasão, sendo que, o Cluster 3 ainda tem um número significativo de alunos matriculados. Os alunos do Cluster 0, como é possível ver na Tabela 15, são alunos que têm dificuldade de adaptação à Universidade e isso fica evidente pela baixa média na taxa de aprovação em matérias do Decea e pela taxa de distorção de carga horária média muito próxima de zero. Esses alunos são, segundo a literatura, os alunos mais propensos a evadirem.

Portanto, a partir da clusterização, vemos que os alunos matriculados do Cluster 2 têm mais semelhanças com os alunos evadidos do curso de Engenharia Elétrica do que com os alunos diplomados, mas que são resilientes às intempéries do caminho, como vimos na Tabela 15, esses alunos têm uma distorção média de carga horária baixa e a maior quantidade média de reprovações e trancamentos dentre os clusters, mas ainda assim, fizeram em média, quase metade das disciplinas do curso. Por outro lado, o Cluster 1 tem uma quantidade significativa dos alunos diplomados, isso indica uma proximidade dos alunos do Cluster 1 que ainda estão matriculados com a diplomação.

O Cluster 4, em particular, tem um comportamento muito peculiar por ser composto, quase inteiramente, por alunos diplomados. Esse Cluster tem uma quantidade média de reprovações baixa, frente aos Clusters 1 e 2, mesmo tendo quase a mesma média de períodos feitos. Esse Cluster se caracteriza por ter a melhor média de notas, com 7,41, que é o que difere o Cluster 4 do Cluster 1.

4.1.9 Análise da situação do aluno em cada Cluster

4.1.9.1 Evadidos

Os alunos evadidos do curso de engenharia elétrica no Icea têm um perfil diverso e representam 56% da base clusterizada do curso. Alguns indicadores pré-Ufop, como a nota do Enem, a idade de ingresso e o acesso por meio de políticas afirmativas, permitem comparar as semelhanças e diferenças entre esses alunos. Contudo, os indicadores acadêmicos apontam para um padrão comum de dificuldade de adaptação na universidade, com uma média elevada de reprovações por período, uma taxa de distorção carga horária/período baixa. O posicionamento desses alunos no cluster depende da quantidade de disciplinas que ele cursou e, consequentemente, do número de períodos feitos.

4.1.9.2 Diplomados

São 138 os alunos diplomados no curso de Engenharia Elétrica no Icea até o momento de coleta da base original. Esses alunos não mostraram muitas divergências no que diz respeito às estatísticas acadêmicas. Há uma proximidade nos indicadores criados para a clusterização, como mostra a Tabela 15 e o que difere os diplomados são as notas nas disciplinas e indicadores que foram usados para a agregação dos alunos.

O Cluster 4 é um cluster modelo que deveria ser explorado pelo curso, buscando entender a jornada dos estudantes dentro e fora da universidade.

4.1.9.3 Matriculados

Os alunos ainda matriculados no curso representam 25.24% da base que foi clusterizada e foram bem distribuídos entre os primeiros clusters. No Cluster 0, que agrupa os estudantes com baixo rendimento e baixa carga horária, há também os alunos recém-matriculados e que, portanto, não devem ser comparados da mesma forma que os alunos matriculados em outros clusters. Dentre os alunos admitidos a partir de 2021, 66% deles são do Cluster 0.

Por outro lado, os matriculados no Cluster 1 e 4 são estudantes com um alto potencial de diplomação, já que se aproximam dos alunos já formados através das disciplinas cursadas e dos indicadores criados para a clusterização.

Por último, os alunos matriculados dos Clusters 2 e 3 precisam ser tratados com cautela pelo curso, já que são os Clusters em que muitos dos alunos evadem ou ficam retidos por muito tempo na universidade.

4.1.10 Perfil dos Clusters

A Tabela 16 mostra o pefil de cada Cluster.

Tabela 16 – Informações médias por cluster ordenado por evasão

Cluster	**Sexo**	**Cor da Pele**	**Enem**	**Idade**	**PAA**	*AP*	*RP*	*APD*	*SEM*	CH/P
0	M (80%)	Parda (22.6%)	648.02	23.43	Não	1.54	1.89	0.06	2.09	0.02
3	M (84%)	Parda (22.2%)	633.72	20.96	Sim	10.71	6.43	0.41	5.39	0.23
2	M (77%)	Parda (24.6%)	645.58	20.56	Sim	27.96	11.39	0.71	10.04	0.46
1	M (80%)	Parda (26.7%)	660.32	20.16	Não	52.15	9.22	0.86	12.47	0.75
4	M (72%)	Parda / Branca (22.7%)	676.76	19.92	Sim	53.56	8.55	0.82	12.00	0.79

Fonte: Elaborado pelo autor.

4.1.11 Resultado do curso

Como visto na Figura 14, a clusterização separou bem os alunos diplomados dos evadidos e, com isso, separou também os alunos matriculados de forma a evidenciar a proximidade deles com a evasão ou a diplomação.

O curso de Engenharia Elétrica é, proporcionalmente, o segundo curso que mais forma alunos no Icea, mas isso não significa que os resultados sejam positivos.

Como visto na Tabela 10, há um número alto de alunos evadidos e, calculando uma taxa de alunos formados por alunos evadidos o resultado é de 0.30. Isso quer dizer que para cada 3 alunos diplomados no curso, 10 alunos desistiram.

Além da alta taxa de evasão, muitos alunos do curso contribuem com o fenômeno da retenção, tendo até nos alunos diplomados uma taxa de distorção carga horária/período distante de 1.

Por fim, é importante salientar o desempenho do Cluster 4. Os resultados deste Cluster superam a expectativa do perfil esperado do curso em muito e merecem a atenção do colegiado.

4.2 Engenharia de Computação

Nas Seções a seguir, serão tratados e discutidos os dados do curso de Engenharia de Computação.

4.2.1 Atributos pré universidade

A Tabela 17 apresenta as estatísticas descritivas dos atributos Idade em que entrou na Universidade e nota no Enem.

Tabela 17 – Idade e Notas no Enem do curso

Estatística	**Pontuação Enem**	**Idade que entrou**
Média	646.03	21.22
Desvio padrão	47.51	4.76
Mín.	476.20	17.00
25%	618.72	19.00
50%	645.40	20.00
75%	680.25	22.00
Máx.	768.40	62.00

Fonte: Elaborado pelo autor.

A Tabela 18 apresenta a frequência de valores para os atributos categóricos sexo e uso de ações afirmativas. É possível traçar um perfil esperado do aluno egresso no curso por meio dos atributos descritos nas Tabelas 17 e 18.

Tabela 18 – Sexo e uso de cotas no curso

Sexo	**Usou política afirmativa**	**Quantidade**
M	Não	286
M	Sim	283
F	Não	60
F	Sim	60

Fonte: Elaborado pelo autor.

4.2.2 Atributos Acadêmicos

Como mencionado na Seção 3.6.1, todos os alunos tiveram informações agregadas para a clusterização. Essas informações serão usadas para entender a média dos atributos acadêmicos dos alunos deste curso. Na Seção 3.6.2 estão os significados de cada atributo.

Tabela 19 – Estatística descritiva dos atributos do curso

Estatística	AP	RPN	APD	SEM	CH/P
Média	14.09	5.19	0.31	5.17	0.22
Desvio padrão	18.72	6.22	0.34	4.36	0.29
Mín	0.00	0.00	0.00	1.00	0.00
25%	0.00	0.00	0.00	1.00	0.00
50%	6.00	3.00	0.20	4.00	0.08
75%	18.00	8.00	0.56	8.00	0.38
Máx	65.00	49.00	1.00	22.00	1.11

Fonte: Elaborado pelo autor.

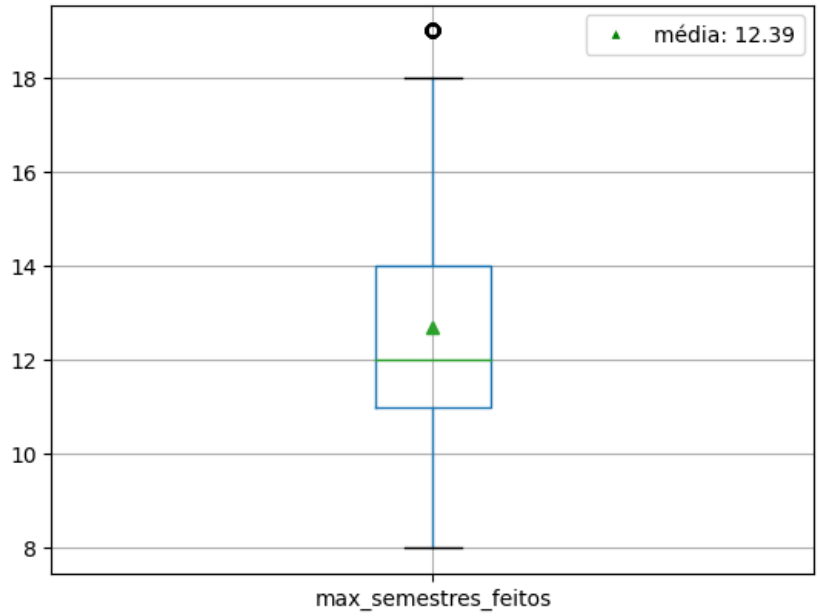
Com base nas informações contidas nas Tabelas 17, 18, e 19, tem-se o perfil esperado do aluno do curso, segundo as bases de dados disponibilizadas pela Universidade.

- sexo: masculino
- pontuação no Enem: 646.03
- idade que entrou: 21.22
- usou política afirmativa: Não
- aprovações (AP): 14.09
- reprovações (RPN): 5.19
- taxa de aprovação no Decea (APD): 0.31
- semestres feitos (SEM): 5.17
- distorção carga horária/período (CH/P): 0.22

4.2.3 Tempo para a diplomação

O curso de Engenharia de Computação tem cinco anos de duração, ou seja, dez semestres. O *box plot* na Figura 15 mostra o comportamento da base de alunos do curso que foram diplomados até a coleta de dados, excluindo os alunos que se formaram em menos de oito períodos, já que são, provavelmente, alunos que reingressaram através do Enem ou alunos egressos por transferência.

Figura 15 – *Box plot*: Número de semestres feitos pelos alunos diplomados



Fonte: Elaborado pelo autor.

Menos de 25% dos alunos se formam dentro do tempo esperado de diplomação no curso de Engenharia de Computação. É possível ver também como a média se comporta e como os quartis se distanciam no gráfico.

4.2.4 Situação do aluno

Por último, a Tabela 20 apresenta a participação de cada categoria de situação do aluno na base de alunos de Engenharia de Computação.

Tabela 20 – Situação do aluno

Situação aluno	**Total**	**%**
Evadido	665	65.45
Matriculado	230	22.64
Diplomado	115	11.32
Trancado	5	0.49
Afastado	1	0.10

Fonte: Elaborado pelo autor.

4.2.5 Análise de Componente Principal

A base de dados com os atributos dos alunos da Engenharia de Computação continha 1016 linhas e 252 colunas além do mask, e, para explicar 80% da variância total, foram necessárias 26 componentes. Usando apenas quatro componentes, a nova base explica 55.44% da variância da base de dados que foi usada para o PCA.

4.2.5.1 Scores features x componentes

No mapa de calor da Figura 16, é possível ver a relação das componentes 0, 1, 2 e 3 com as colunas originais da base de atributos. Por serem muitas colunas, os *scores* vão de, aproximadamente, -0.3 a 0.2. Os extremos são os que contém maior variabilidade de dados e, por isso, contribuem mais para explicar a variância da base de dados.

Na primeira componente, os valores mais escuros no mapa de calor se dão pelas notas de disciplinas específicas do curso, e o maior valor é a nota final na disciplina ENP493, que tem como pré-requisito 1800 horas.

Na segunda componente, há um valor negativo que chama a atenção, mas que também é referente à nota final na disciplina CSI148, que é uma disciplina de quarto período. O maior valor se dá pela nota final em CEA160, uma disciplina de primeiro período. Isso foge do padrão que foi observado nos outros cursos.

A terceira componente tem o maior valor dentre os mostrados no mapa de calor, que é a nota final na disciplina CEA148, que curiosamente, é a mesma disciplina de menor valor da segunda componente, mas que é dada pelo departamento de Engenharia Elétrica.

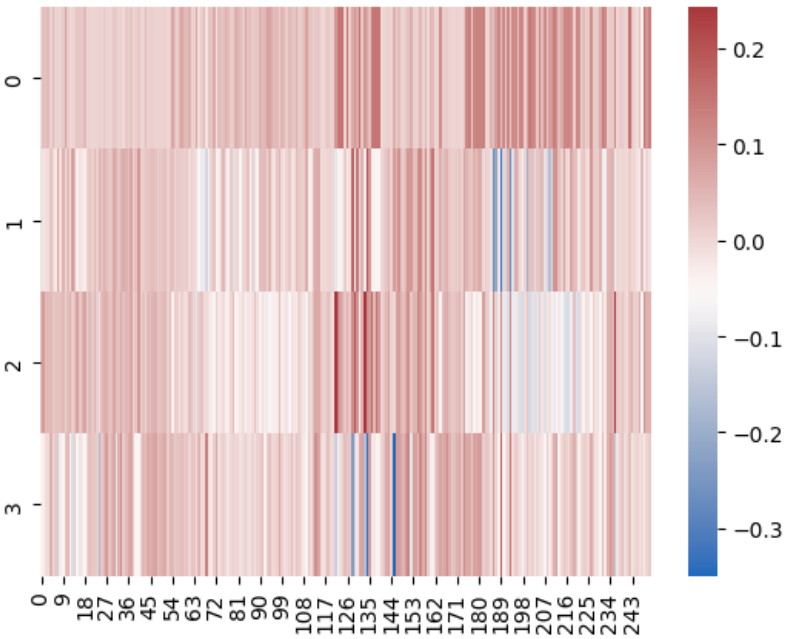
Por último, a quarta componente tem o menor valor negativo dentre os mostrados no mapa de calor, que é a nota final na disciplina CEA422, que também é uma disciplina oferecida pelo departamento de Engenharia Elétrica.

É importante salientar que essas disciplinas são as que mais contribuem para a variabilidade da base de dados, mas isso não significa que sejam as disciplinas fundamentais para o sucesso ou o insucesso do aluno ao longo do curso.

4.2.6 Clusterização

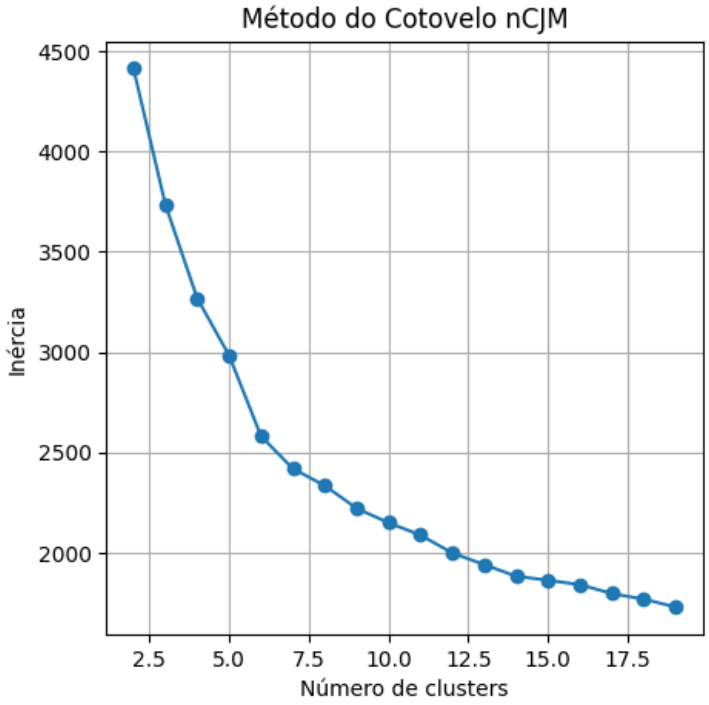
A base com os valores transformados após a aplicação do PCA possui 26 colunas que foram utilizadas para a clusterização. Utilizamos o método da silhueta e o método do cotovelo para definir qual o melhor número possível de clusters para essa base. O método do cotovelo, que consiste em calcular a soma das distâncias quadráticas dos dados intra-clusters, é mostrado na Figura 17

Figura 16 – Mapa de calor: Explicando as componentes da PCA



Fonte: Elaborado pelo autor.

Figura 17 – Método do Cotovelo



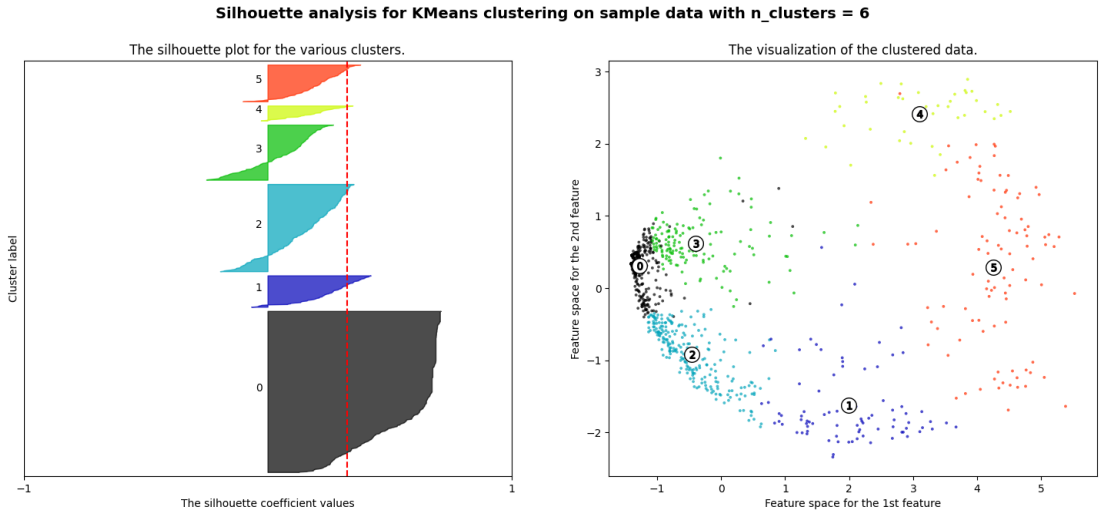
Fonte: Elaborado pelo autor.

Com esse gráfico, vê-se a redução das distâncias com o aumento do número de clusters. É visível o ganho de 4 para 5 clusters e, ainda mais, o ganho de 5 para 6 clusters e por isso, serão os valores testados no método da silhueta.

Com 6 clusters, todos os valores de coeficiente da silhueta tocam a média de coeficiente, traçada em vermelho na Figura 18. Como discutido na Seção 2.7, a proximidade do coeficiente da silhueta dos clusters com a média de coeficiente é importante para entender a qualidade dos agrupamentos.

O posicionamento dos Clusters considerando as duas primeiras componentes é apresentado no *scatterplot* à direita das Figuras 18 e 19.

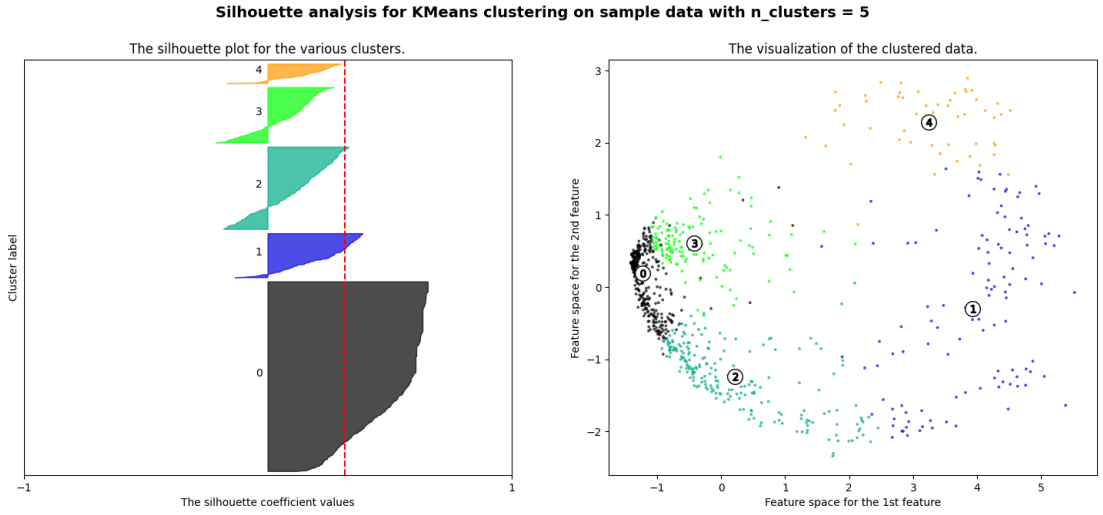
Figura 18 – Método da silhueta para 6 clusters



Fonte: Elaborado pelo autor.

Por outro lado, com 5 clusters, os valores de coeficiente da silhueta do cluster 3 e 4 não tocam a linha média de coeficiente da silhueta.

Figura 19 – Método da silhueta para 5 clusters



Fonte: Elaborado pelo autor.

A partir da análise dos métodos, foi definida a quantidade de 6 clusters como ótima para o conjunto de dados. Aplicando esse número de clusters ao algoritmo k-means, temos a divisão da base, mostrada na Tabela 21

Tabela 21 – Tamanho de cada Cluster

Cluster	Observações
0	416
1	89
2	97
3	225
4	40
5	149

4.2.7 Análise estatística dos Clusters

A Tabela 22 mostra a distribuição étnica nos clusters da Engenharia de Computação. É visível a participação dominante de alunos brancos e pardos no curso, o que também acontece nos demais cursos do Icea.

Tabela 22 – Etnia por cluster

Cluster	Amarela %	Branca %	Indígena %	Não Declarado %	Parda %	Preta %
0	0.72	39.42	0.48	7.21	44.47	7.69
1	2.25	46.07	0.00	3.37	40.45	7.87
2	1.03	45.36	0.00	2.06	37.11	14.43
3	0.89	48.44	0.44	0.44	40.89	8.89
4	2.50	42.50	0.00	5.00	42.50	7.50
5	1.34	40.27	0.00	3.36	42.95	12.08

Fonte: Elaborado pelo autor.

A Tabela 23, por sua vez, mostra a porcentagem de alunos de cada cluster que usou política afirmativa.

Tabela 23 – Uso de Política Afirmativa por cluster

Cluster	Não%	Sim%
0	50.98	49.02
1	53.12	46.88
2	48.33	51.67
3	48.57	51.43
4	58.82	41.18
5	47.76	52.24

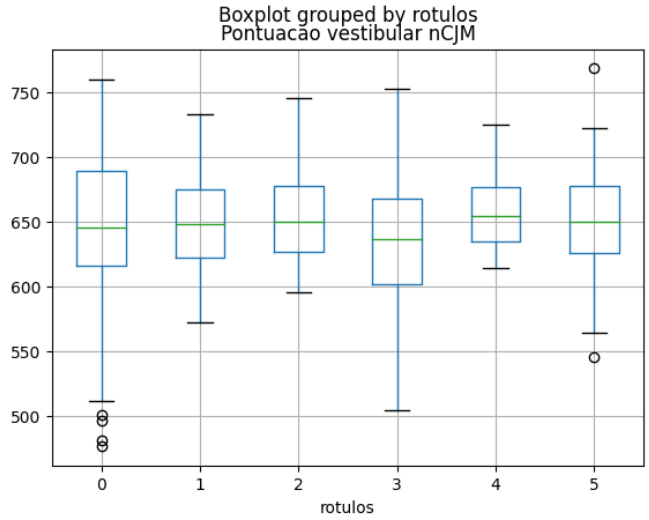
Fonte: Elaborado pelo autor.

Os Clusters 2, 3 e 5 têm uma participação um pouco maior de alunos que entraram na Universidade através do uso de cotas.

4.2.7.1 Nota no Enem

A partir das Tabelas 22 e 23, é possível obter uma relação de cor de pele e uso de políticas afirmativas para cada cluster. Essas informações podem ser relevantes para entender como os alunos dos clusters se saíram no vestibular. O *box plot* apresentado na Figura 20 mostra como os clusters se comportam nesse quesito.

Figura 20 – *Box plot*: Nota no Enem por Cluster



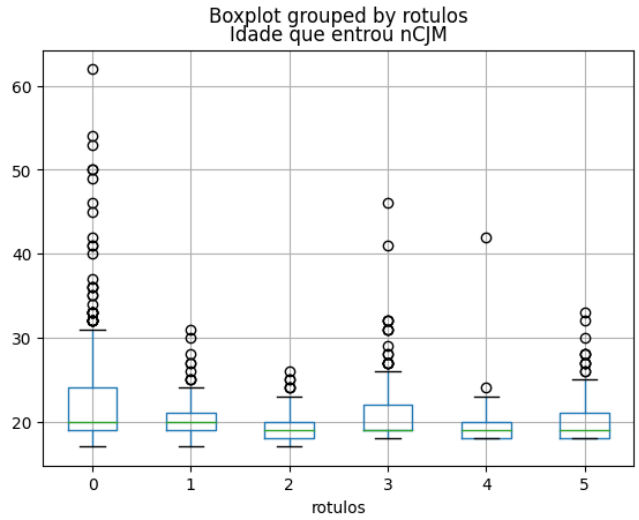
Fonte: Elaborado pelo autor.

No *box plot* apresentado na Figura 20 fica evidente uma alta dispersão de notas do vestibular nos Clusters 0 e 3, sendo o Cluster 3 o com menor média entre os clusters. Por outro lado, nos Clusters 1, 2 e 4 há uma dispersão bem menor, sendo o Cluster 3 com a maior média de nota dos clusters.

4.2.7.2 Idade que entrou

O *box plot* da Figura 21 mostra a distribuição da idade dos alunos quando ingressaram na universidade.

Figura 21 – *Box plot*: Idade que entrou por Cluster



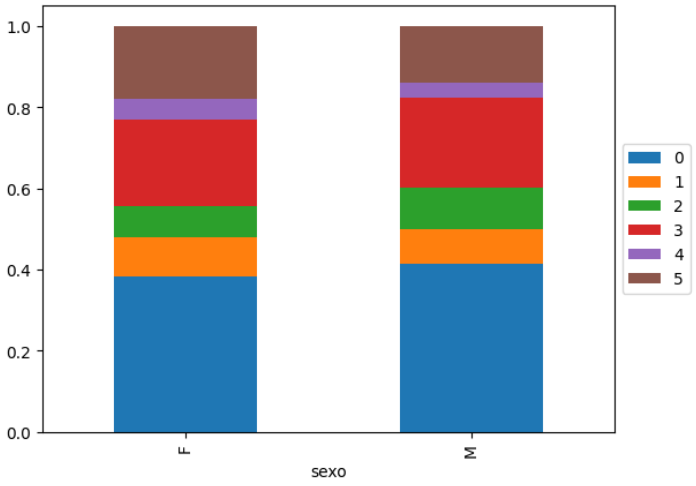
Fonte: Elaborado pelo autor.

O *box plot* da Figura 21 mostra uma alta dispersão nas idades em que o aluno entrou nos clusters 0 e 3, que têm também, alta dispersão nas notas de vestibular. Isso é esperado pois são os clusters com mais estudantes. Por outro lado os Clusters 1, 2 e 4, por sua vez, têm uma dispersão mais baixa de idade e também nas notas de vestibular. Portanto, existe uma relação entre a idade do aluno e a nota dele no vestibular nesses Clusters.

4.2.7.3 Sexo

Novamente, pela discrepância entre o número de homens e de mulheres o gráfico apresentado na Figura 22 é a porcentagem dos clusters em cada sexo.

Figura 22 – Gráfico de Barras Empilhadas: Porcentagem dos clusters por sexo



Fonte: Elaborado pelo autor.

A participação do Cluster 1 é maior no sexo masculino enquanto a participação do Cluster 5 é maior no sexo feminino. Os outros clusters não tem tanta diferença de participação entre os sexos.

Com essas informações, foi traçado um perfil social esperado de cada cluster, usando as médias das variáveis numéricas e a moda das variáveis categóricas.

Tabela 24 – Perfil esperado de cada cluster

Cluster	**Sexo**	**Cor da Pele**	**Nota do vestibular**	**Desvio padrão**	**Idade em que entrou**	**Desvio padrão**	**Usou política afirmativa ?**
0	Masc.	Parda (44.47%)	645.59	53.62	23.47	6.20	Não
1	Masc.	Branca (46.07%)	650.01	34.23	20.42	2.73	Não
2	Masc.	Branca (45.36%)	653.63	36.10	19.62	1.84	Sim
3	Masc.	Branca (48.44%)	634.67	52.26	20.71	6.65	Sim
4	Masc.	Parda / Branca (42.50%)	658.90	30.60	20.00	3.95	Não
5	Masc.	Parda (42.95%)	652.64	36.51	20.38	2.93	Sim

Fonte: Elaborado pelo autor

Como esperado, há uma presença muito maior de homens. No curso de Engenharia de Computação no Icea há uma dominância de alunos brancos e pardos, e isso se repete nos clusters. Não há uma diferença muito grande nas médias de notas de vestibular, apesar do desvio padrão ser muito diferente nos Clusters 0 e 3, como discutido na Seção 4.2.7.1.

4.2.8 Estatísticas de atributos acadêmicos

A Tabela 25 agrega, por Cluster, algumas das informações que foram usadas para a clusterização.

Tabela 25 – Análise dos atributos acadêmicos por cluster

Cluster	*AP*	*RPN*	*RPF*	*APD*	*SEM*	*CH/P*
0	0.87	1.60	2.92	0.03	1.82	0.01
1	31.91	8.98	3.46	0.74	9.38	0.54
2	54.82	7.73	1.80	0.84	12.54	0.78
3	9.68	6.44	3.91	0.36	5.13	0.19
4	54.02	13.22	1.93	0.73	13.80	0.72
5	9.79	7.30	2.70	0.33	4.99	0.19

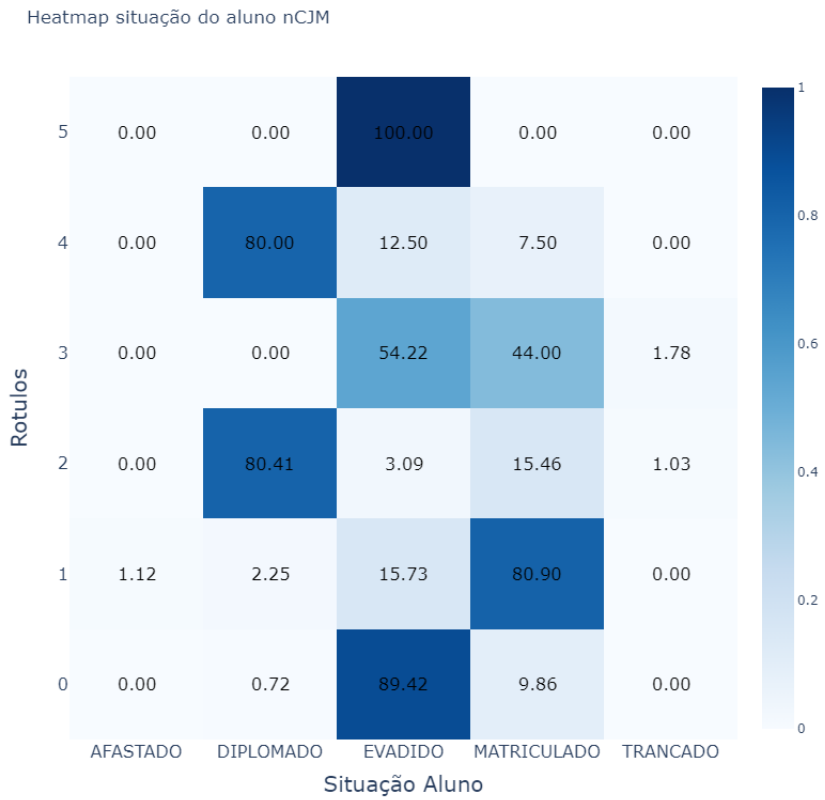
Fonte: Elaborado pelo autor

Os Clusters 2 e 4 são os Clusters com maior sucesso acadêmico e os Clusters 0, 3 e 5 por sua vez, os que têm pior desempenho. Por fim, o Cluster 1 se comporta de forma diferente dos outros clusters, tendo um número razoável de aprovações, mas também tendo um número grande de reprovações por nota.

4.2.8.1 Estatística da situação do aluno

No mapa de calor mostrado na Figura 23 vemos as situações de aluno por Cluster e, assim, podemos comparar com as Tabelas 24 e 25.

Figura 23 – Mapa de calor: Cluster x Situação de aluno



Fonte: Elaborado pelo autor.

O gráfico da Tabela 23 é um mapa de calor que relaciona a situação dos alunos com o Cluster em que eles estão localizados. Cada célula é a porcentagem de alunos na situação descrita presente no Cluster e a soma das linhas tem resultado 100%.

Apenas pelo mapa de calor da Figura 23 fica clara a separação entre os alunos diplomados e os alunos evadidos. Em clusters onde há uma quantidade significativa de diplomados, há uma quantidade muito pequena de evadidos e o contrário também é verdade. Como esperado, os Clusters 0, 3 e 5 são os clusters de menos sucesso acadêmico, sendo o Cluster 5 composto apenas por evadidos e o Cluster 0 sendo quase 90% composto por evadidos, o que é esperado dado o baixíssimo número de aprovações. O Cluster 3, por sua vez, conta tanto com alunos evadidos quanto com matriculados, isso, somado à distorção de carga horária/período, ao baixo número de aprovações e a alta taxa de reprovações indica a proximidade dos alunos matriculados dos alunos evadidos.

Os Clusters 2 e 4, por sua vez, têm uma taxa de diplomação de 80% e, apesar de ter também alunos evadidos, mostra um sucesso muito grande em relação aos outros clusters. Ainda assim, os clusters têm um número de reprovações médias muito alto e, como consequência, um número de semestres médios bem mais alto do que o número esperado de semestres para a diplomação.

Por fim, o Cluster 1 é composto, em sua maioria, por alunos que ainda estão matriculados. O número de semestres médios no Cluster 1 mostra que os alunos deste cluster são alunos que contribuem com o fenômeno da retenção no curso e no instituto.

4.2.9 Análise da situação do aluno em cada Cluster

4.2.9.1 Evadidos

Os alunos evadidos do curso de Engenharia de Computação no Icea têm um perfil social e acadêmico diverso e representam mais de 65% da base do curso. Os indicadores pré-Ufop não têm tanto impacto no desempenho dos alunos evadidos como os indicadores acadêmicos. Esses sim, apontam uma alta taxa de reprovação e uma dificuldade dos alunos de evoluir nas disciplinas do curso. O posicionamento desses alunos nos agrupamentos depende da quantidade de disciplinas que ele cursou e, consequentemente, do número de períodos feitos.

4.2.9.2 Diplomados

São apenas 115 alunos diplomados no curso de Engenharia de Computação no Icea até o momento de coleta da base de dados. Esses alunos mostraram algumas divergências na clusterização, principalmente pelo número de reprovações e o número de semestres feitos.

Os Clusters 2 e 4 são os clusters com mais sucesso na universidade e merecem ser explorados pelo colegiado do curso, buscando entender a jornada dos estudantes destes Clusters dentro e fora da Universidade.

4.2.9.3 Matriculados

Os alunos ainda matriculados no curso representam 22.64% da base clusterizada e foram distribuídos entre todos os clusters, exceto o Cluster 5. Os alunos matriculados que estão nos Clusters 2 e 4 são estudantes com um alto potencial de diplomação, já que se aproximam dos alunos já formados através das disciplinas cursadas e dos indicadores criados para a clusterização.

Por outro lado, os alunos matriculados que se encontram nos Clusters 0 e 3 merecem uma atenção especial do colegiado do curso, já que têm alta taxa de insucesso.

Por último, os alunos matriculados do Cluster 1 são alunos que também precisam ser observados, já que há uma taxa razoável de evasão e um número de reprovações médio muito alto.

4.2.10 Perfil dos Clusters

A Tabela 26 mostra o pefil de cada Cluster.

Tabela 26 – Informações médias por cluster ordenado por evasão

Cluster	Sexo	Cor da Pele	Enem	Idade	PAA	*AP*	*RPN*	*APD*	*SEM*	*CH*/*P*
5	M (79%)	Parda (42.95%)	652.64	20.38	Sim	9.79	7.3	0.33	4.99	0.19
0	M (84%)	Parda (44.47%)	645.59	23.47	Não	0.87	1.6	0.03	1.82	0.01
3	M (84%)	Branca (48.44%)	634.67	20.71	Sim	9.68	6.44	0.36	5.13	0.19
1	M (81%)	Branca (46.07%)	650.01	20.42	Não	31.91	8.98	0.74	9.38	0.54
4	M (78%)	Parda / Branca (22.66%)	676.76	19.92	Sim	54.02	13.22	0.73	13.8	0.72
2	M (87%)	Branca (45.36%)	653.63	19.62	Sim	54.82	7.73	0.84	12.54	0.78

Fonte: Elaborado pelo autor.

4.2.11 Resultado do curso

Como visto na Figura 23, a clusterização separou bem os alunos diplomados dos evadidos e, com isso, separou também os alunos matriculados de forma a evidenciar a proximidade deles com a evasão ou a diplomação.

O curso de Engenharia de Computação tem a maior taxa de evasão no Icea. Tendo formado apenas 115 alunos, o curso tem 10 diplomados para cada 100 ingressantes.

Com as taxas de distorção carga horária/período média muito distante de 1 em todos os clusters e o número médio de semestres para a diplomação mais alto entre os cursos de engenharia, o curso de Engenharia de Computação tem o pior desempenho frente aos atributos criados na Seção 3.6.

Nos outros cursos, o Cluster com mais evadidos é sempre o cluster com poucas disciplinas, taxa de aprovação no Decea e distorção de carga horária período perto de 0. No entanto, o Cluster 5 da Engenharia de Computação, que tem a maior taxa de evasão se comporta diferente. Os alunos deste Cluster têm indicadores um pouco melhores que o do Cluster 0 e mais tempo de curso e, ainda assim, evadem mais.

4.3 Engenharia de Produção

4.3.1 Atributos pré universidade

A Tabela 27 apresenta as estatísticas descritivas dos atributos Idade em que entrou na Universidade e nota no Enem.

Tabela 27 – Atributos pré universidade dos alunos no curso

Estatística	**Pontuação Enem**	**Idade que entrou**
Média	652.68	21.03
Desvio padrão	46.28	3.91
Mín	475.70	16.00
25%	626.10	19.00
50%	654.00	20.00
75%	685.20	22.00
Máx	774.60	48.00

Fonte: Elaborado pelo autor.

Por meio dos dados apresentados na Tabela 27 e dos atibutos categóricos sexo e uso de política afirmativa, disponíveis na Tabela 28, traçamos um perfil esperado do aluno egresso no curso .

Tabela 28 – Uso de Política afirmativa no curso

Sexo	**Usou política afirmativa**	**Quantidade**
F	Sim	202
M	Sim	195
M	Não	148
F	Não	139

Fonte: Elaborado pelo autor.

4.3.2 Atributos Acadêmicos

Como mencionado na Seção 3.6, todos os alunos tiveram informações agregadas para a clusterização. Essas informações foram usadas para entender a média dos atributos acadêmicos dos alunos deste curso.

Tabela 29 – Atributos acadêmicos no curso

Estatística	*AP*	*RPN*	*APD*	*SEM*	*CH/P*
Média	21.79	4.08	0.43	6.36	0.36
Desvio padrão	21.64	4.92	0.37	4.51	0.36
Mín	0.00	0.00	0.00	1.00	0.00
25%	1.00	0.00	0.00	2.00	0.00
50%	12.00	2.00	0.45	6.00	0.28
75%	47.00	6.00	0.76	10.00	0.72
Máx	67.00	27.00	1.00	18.00	1.08

Fonte: Elaborado pelo autor.

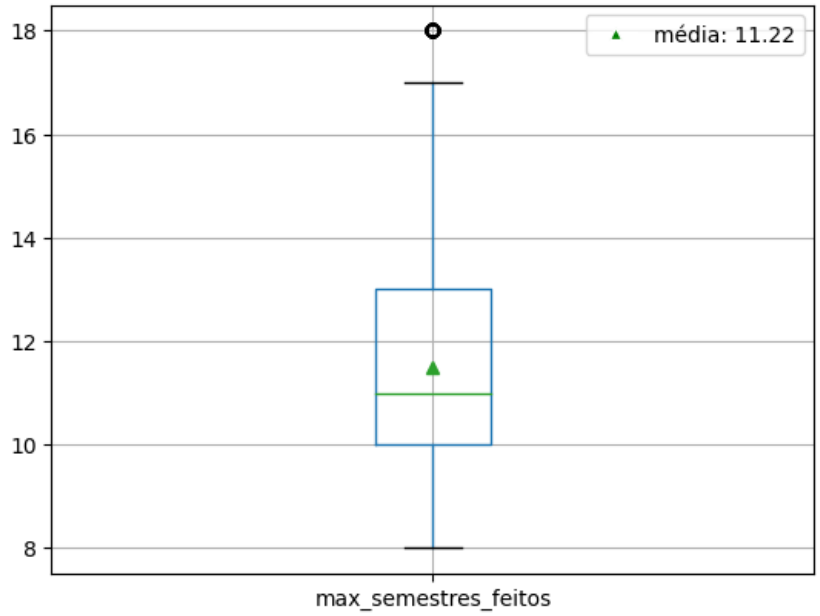
Unindo as Tabelas 27, 28 e 29 temos o perfil esperado do aluno do curso, segundo as bases de dados disponibilizadas pela Universidade. Portanto, um perfil esperado do aluno que entra no curso de Engenharia de Produção se dá por.

- sexo: masculino.
- pontuação no Enem: 652.68
- idade que entrou: 21.03
- usou política afirmativa: sim
- aprovações: 21.79
- reprovações: 4.08
- taxa de aprovação no Decea: 0.43
- semestres feitos: 6.36
- distorção carga horária/ período: 0.36

4.3.3 Tempo para a diplomação

O curso de Engenharia de Produção tem cinco anos de duração, ou seja, dez semestres. O *box plot* da Figura 24 mostra o comportamento da base de alunos do curso que foram diplomados até a coleta de dados, excluindo os alunos que se formaram em menos de oito períodos, já que são, provavelmente, alunos que reingressaram através do Enem ou alunos egressos por transferência.

Figura 24 – *Box plot*: Número de semestres feitos pelos alunos diplomados



Fonte: Elaborado pelo autor.

No curso de Engenharia de Produção, 25% dos alunos se formam dentro do tempo esperado de diplomação e a média de tempo de formatura é de 11.22 semestres. É possível ver também como a mediana se comporta e como os quartis se distanciam no gráfico apresentado na Figura 24.

4.3.4 Situação do aluno

Por último, é importante descrever a participação de cada categoria de situação do aluno na base de alunos da Engenharia de Produção. A Tabela 30 apresenta essas informações.

Tabela 30 – Descrição da situação dos alunos da base

Descrição situação aluno	**Total**	*%*
Evadido	485	45.54
Diplomado	307	28.83
Matriculado	258	24.23
Trancado	8	0.75
Mobilidade	4	0.38
Afastado	3	0.28

Fonte: Elaborado pelo autor.

4.3.5 Análise de Componente Principal

A tabela com os atributos dos alunos da Engenharia de Produção continha 1065 linhas e 281 colunas além do *mask*, e, para explicar 80% da variância total, foram necessárias 27 componentes. Usando apenas três componentes, a nova base explica 53.34% da variância da base de dados que foi usada para o PCA.

4.3.5.1 Scores features x componentes

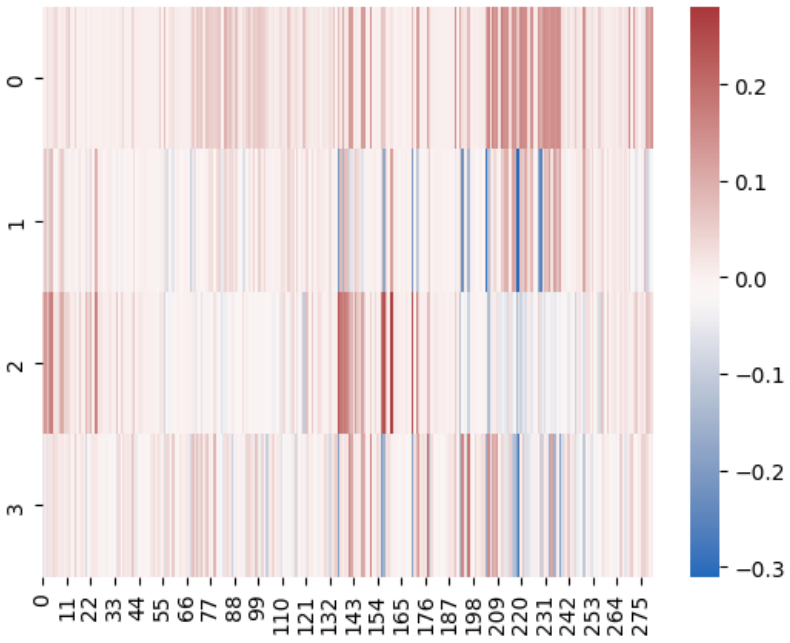
No mapa de calor mostrado na Figura 25, vemos a relação das componentes 0, 1 e 2 com as colunas originais da base de atributos. Por serem muitas colunas, os *scores* vão de, aproximadamente, -0.3 a 0.2. Os extremos são os que contém maior variabilidade de dados e, por isso, contribuem mais para explicar a variância da base de dados.

Na primeira componente, os valores mais escuros no mapa de calor se dão pelas notas de disciplinas específicas do curso, e o maior valor é a nota final na disciplina ENP018.

Na segunda componente, alguns valores negativos chamam a atenção, mas que também é referente à nota final de uma disciplina, mais especificamente, a disciplina ENP029. O maior valor na segunda componente é referente à nota final na disciplina ENP160.

A terceira componente tem o maior valor dentre os mostrados no mapa de calor, que é a nota final na disciplina CEA034, que curiosamente, é uma disciplina que não é obrigatória para o curso de Engenharia de Produção.

Figura 25 – Mapa de calor: Explicando as componentes do PCA



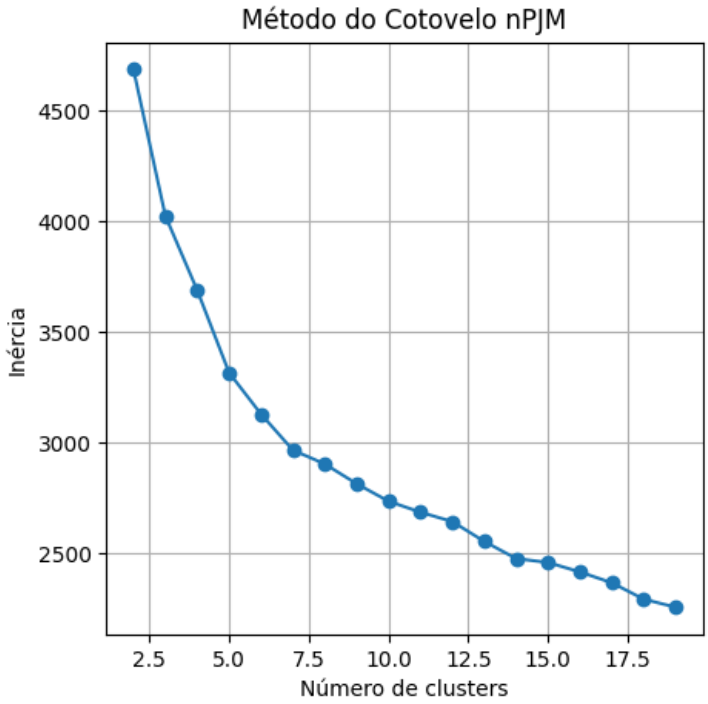
Fonte: Elaborado pelo autor.

É importante salientar que essas disciplinas são as que mais contribuem para a variabilidade da base de dados, mas isso não significa que sejam as disciplinas fundamentais para o sucesso ou o insucesso do aluno ao longo do curso.

4.3.6 Clusterização

A base com os valores transformados após a aplicação do PCA possui 27 colunas que foram utilizadas para a clusterização. Utilizamos o método da silhueta e o método do cotovelo para definir qual o melhor número possível de clusters para essa base. O método do cotovelo, que consiste em calcular a soma das distâncias quadráticas dos dados intra-clusters, é mostrado na figura 26.

Figura 26 – Método do cotovelo

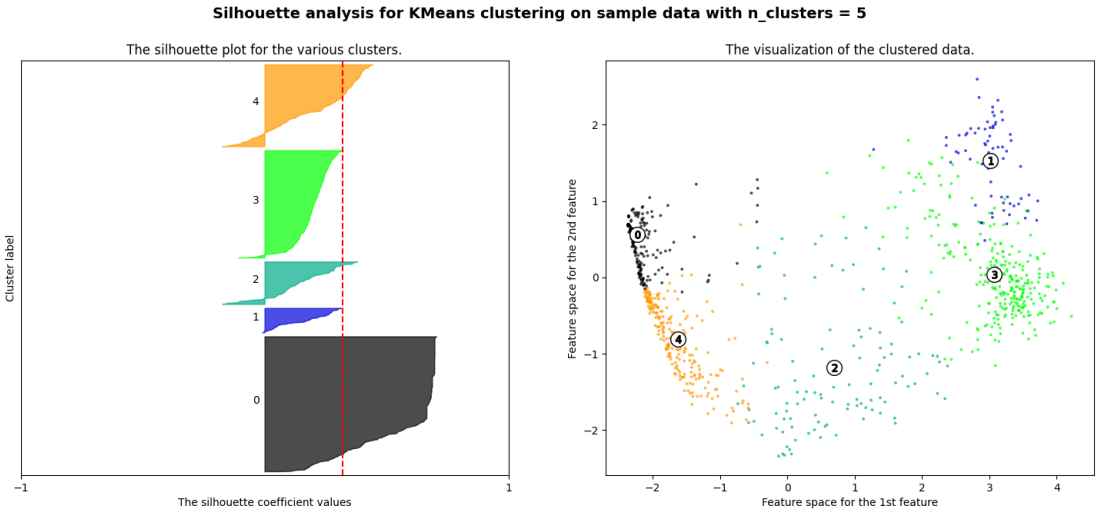


Fonte: Elaborado pelo autor.

Com o gráfico mostrado na Figura 26, vê-se a redução das distâncias com o aumento do número de clusters. É visível o ganho de 3 para 4 clusters e, ainda mais, o ganho de 4 para 5 clusters e por isso, serão os valores testados no método da silhueta. Como discutido na Seção 2.7, a proximidade do coeficiente da silhueta dos clusters com a média de coeficiente é importante para entender a qualidade dos agrupamentos.

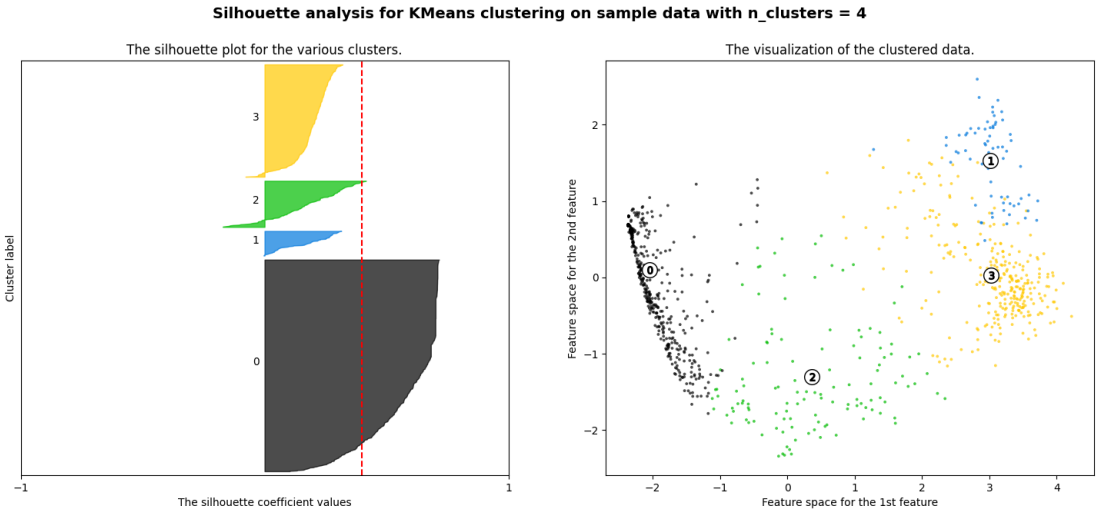
Com 5 clusters, todos os valores de coeficiente da silhueta tocam a média de coeficiente, traçada em vermelho na Figura 27. O posicionamento dos Clusters considerando as duas primeiras componentes é apresentado no *scatterplot* à direita das Figuras 27 e 28.

Figura 27 – Método da silhueta para 5 clusters



Fonte: Elaborado pelo autor.

Figura 28 – Método da silhueta para 4 clusters



Fonte: Elaborado pelo autor.

Por outro lado, com 4 clusters, os valores de coeficiente da silhueta dos Clusters 1 e 3 não tocam a linha média de coeficiente da silhueta.

Portanto, o melhor número de clusters para esse conjunto de dados é 5. Aplicando esse número de clusters ao algoritmo k-means, temos a divisão da base, mostrada na Tabela 31.

4.3.7 Análise estatística dos Clusters

A Tabela 32 mostra a distribuição étnica nos clusters da Engenharia da Produção. É visível a participação dominante de alunos brancos e pardos no curso, o que também acontece nos demais cursos do Icea.

Tabela 31 – Número de observações por cluster no curso

Cluster	Quantidade de observações
0	369
1	310
2	225
3	44
4	117

Fonte: Elaborado pelo autor.

Tabela 32 – Participação étnica por cluster

Cluster	Amarela %	Branca %	Não declarado %	Parda %	Preta%
0	1.63	45.53	5.42	39.30	8.13
1	0.32	45.48	1.29	42.90	10.00
2	1.78	43.11	0.89	46.67	7.56
3	2.27	45.45	11.36	34.09	6.82
4	0.85	47.86	1.71	43.59	5.98

Fonte: Elaborado pelo autor.

A Tabela 33, por outro lado, mostra a porcentagem de alunos de cada cluster que usou política afirmativa. O curso de Engenharia de Produção têm dados faltantes no que diz respeito ao uso de políticas afirmativas, por isso, os clusters na Tabela 33 têm dimensões diferentes das dimensões da Tabela 31.

Tabela 33 – Uso de política afirmativa por cluster

Cluster	Total	Não %	Sim %
0	260	49.62	50.38
1	168	38.10	61.90
2	159	37.11	62.89
3	9	22.22	77.78
4	88	37.50	62.50

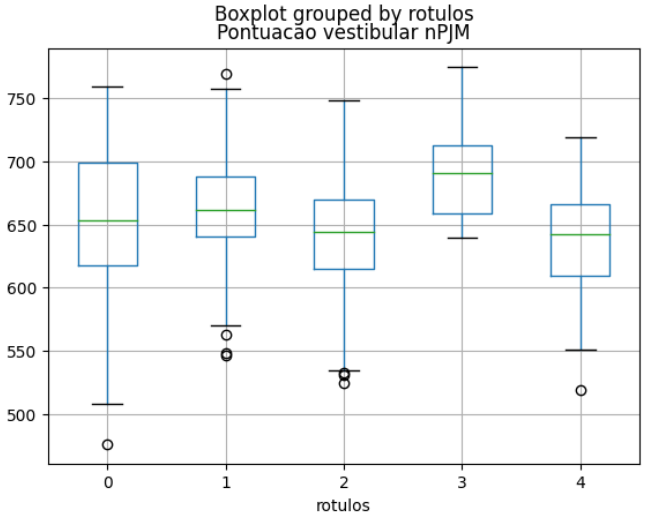
Fonte: Elaborado pelo autor.

Todos os Clusters do curso de Engenharia de Produção têm mais alunos que usaram alguma política afirmativa, no entanto, não é possível afirmar se os alunos que não têm essa informação na base de dados do Icea usaram ou não cotas, enviesando a estatística.

4.3.7.1 Nota no Enem

A partir das Tabelas 32 e 33, temos uma relação de cor de pele e uso de políticas afirmativas para cada cluster. Essas informações podem ser relevantes para entender como os alunos dos clusters se saíram no Enem. O *box plot* apresentado na Figura 29 mostra como os clusters se comportam nesse quesito.

Figura 29 – *Box plot*: Nota no Enem por Cluster



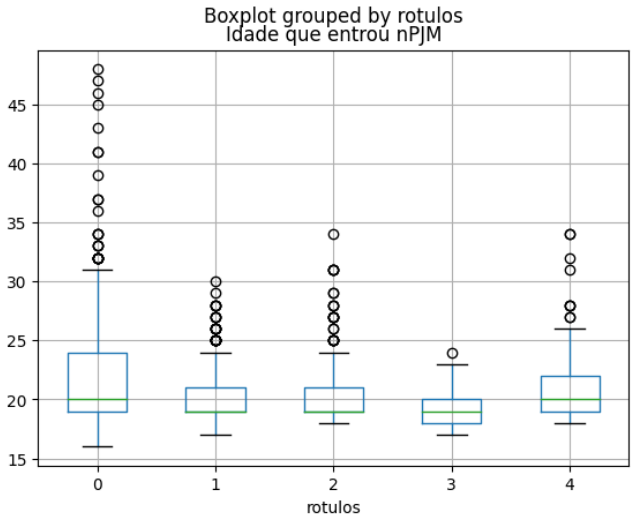
Fonte: Elaborado pelo autor.

O gráfico apresentado na Figura 29 mostra uma alta dispersão de notas no Enem no Cluster 0, apesar da menor média de nota ser a do Cluster 4. O curso de Engenharia de Produção tinha, até 2022, uma nota mínima para aprovação, o que explica a baixa dispersão entre os clusters.

4.3.7.2 Idade que entrou

O *box plot* da Figura 30 mostra a distribuição da idade dos alunos quando ingressaram na Universidade.

Figura 30 – *Box plot*: Idade que entrou por Cluster



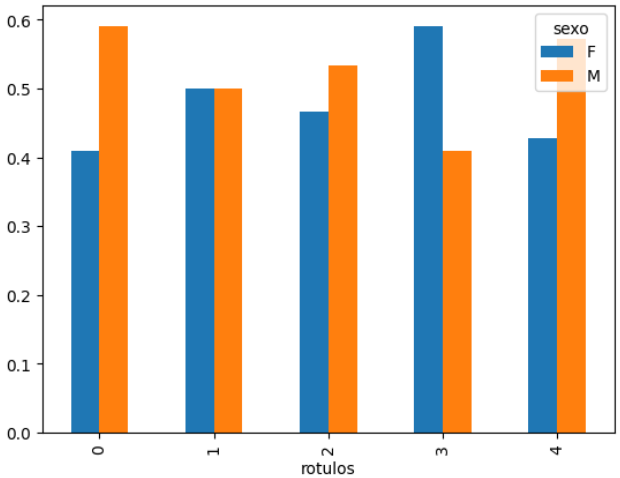
Fonte: Elaborado pelo autor.

O *box plot* da Figura 30 mostra uma alta dispersão de idades no Cluster 0 e o Cluster 3 com uma dispersão mais baixa em relação aos outros clusters. A relação da dispersão na nota do Enem e a dispersão de idade em que o aluno entrou na Universidade também é verdadeira no curso de Engenharia de Produção.

4.3.7.3 Sexo

Nas outras análises de sexo por cluster, foi necessário o uso do gráfico de barra empilhado para melhor entendimento dos dados, no entanto, pela primeira vez, um curso tem mais mulheres que homens em algum cluster. Isso é mostrado na Figura 31.

Figura 31 – Gráfico de barras: Sexo por Cluster



Fonte: Elaborado pelo autor.

Os Clusters 0, 2 e 4 têm maior participação de homens, enquanto o Cluster 3 tem maior participação de mulheres. O Cluster 1 tem 50% de participação de cada um.

Com essas informações traçamos um perfil esperado de cada cluster, usando as médias das variáveis numéricas e a moda das variáveis categóricas, mostrado na Tabela 34.

Tabela 34 – Perfil social esperado por cluster

Cluster	**Sexo**	**Cor da Pele**	**Nota do Enem**	**Desvio padrão**	**Idade em que entrou**	**Desvio padrão**	**Usou política afirmativa ?**
0	Masc.	Branca (45.53%)	652.08	53.39	22.18	5.24	Sim
1	Masc./ Fem.	Branca (45.48%)	662.03	36.34	20.28	2.48	Sim
2	Masc.	Parda (46.67%)	640.72	45.71	20.44	3.00	Sim
3	Fem.	Branca (45.45%)	688.58	34.27	19.30	1.37	Sim
4	Masc.	Branca (47.86%)	638.64	38.89	21.15	3.32	Sim

Fonte: Elaborado pelo autor.

Diferentemente dos outros cursos, há uma participação muito maior de mulheres. No entanto, o padrão de cor de pele branca ou parda permanece presente em todos os clusters.

4.3.8 Estatísticas de atributos acadêmicos

A Tabela 35 agrega, por cluster, algumas das informações que foram usadas para a clusterização.

Tabela 35 – Estatística dos atributos acadêmicos por cluster

Cluster	AP	RPN	RPF	APD	SEM	CH/P
0	1.08	1.01	2.11	0.04	1.87	0.02
1	49.13	5.92	1.38	0.79	11.03	0.79
2	8.88	5.00	2.64	0.37	4.98	0.18
3	49.55	5.66	1.20	0.78	11.30	0.80
4	29.09	6.50	2.82	0.71	8.96	0.53

Fonte: Elaborado pelo autor.

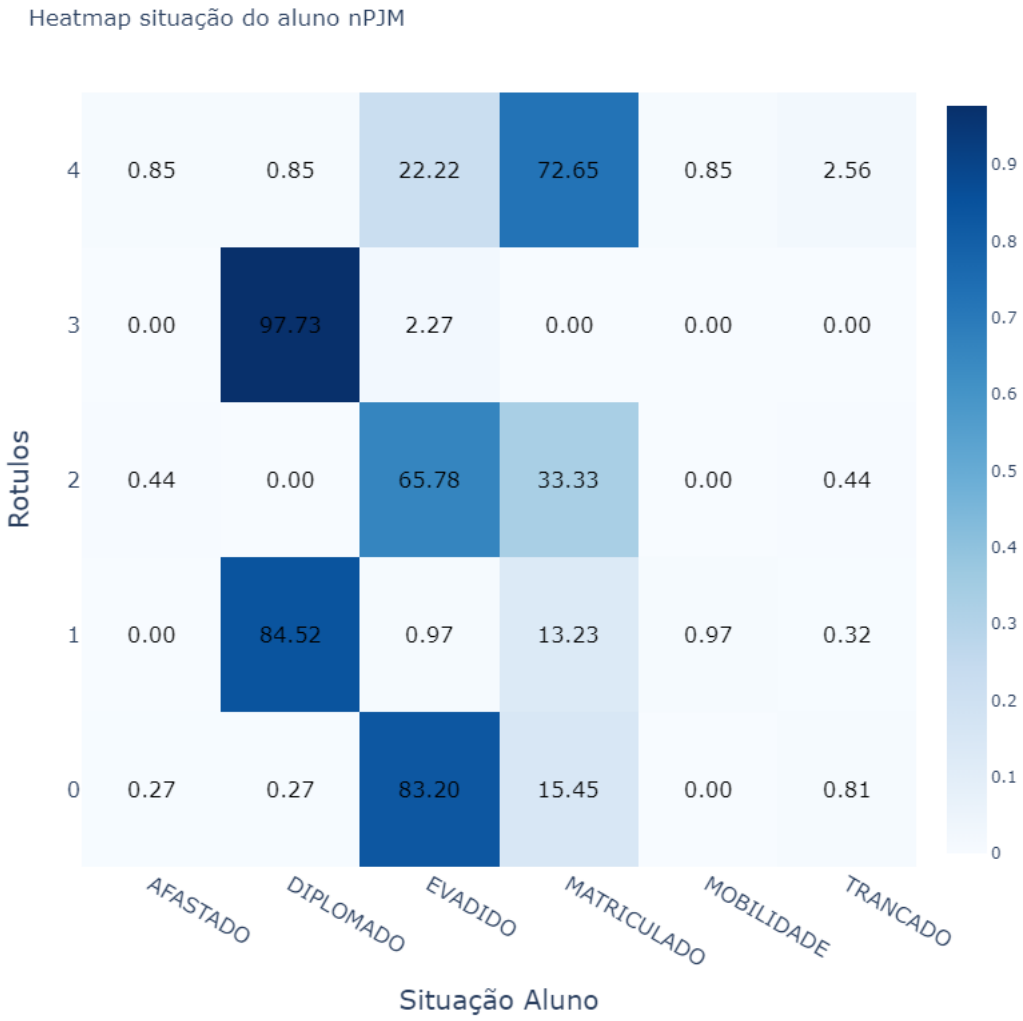
Com essas estatísticas traçamos um perfil acadêmico médio de cada cluster na Engenharia de Produção. Além disso, fica evidente que os atributos escolhidos para a clusterização foram suficientes para separar os perfis de alunos do curso.

Os Clusters 1 e 3 são os Clusters com maior sucesso acadêmico e os Clusters 0 e 2, os que têm pior desempenho. Por fim, o Cluster 4 se comporta de forma diferente dos outros clusters, tendo um número razoável de aprovações, mas também tendo um número grande de reprovações.

4.3.8.1 Estatística da situação do aluno

No mapa de calor da Figura 32 podemos ver as situações de aluno por cluster e comparar com as Tabelas 34 e 35.

Figura 32 – Mapa de calor: Situação do aluno por Cluster



Fonte: Elaborado pelo autor.

O mapa de calor da Figura 32 relaciona a situação dos alunos com o cluster em que eles estão localizados. Cada célula é a porcentagem de alunos na situação descrita presente no cluster e a soma das linhas tem resultado 100%.

Apenas pela Figura 32 fica clara a separação entre os alunos diplomados e os alunos evadidos. Em clusters onde há uma quantidade significativa de diplomados, há uma quantidade muito pequena de evadidos e o contrário também é verdade. Como esperado, dado o baixo número de aprovações e de períodos, os Clusters 0 e 2 são os clusters de menos sucesso acadêmico. Os Clusters 1 e 3, por sua vez, são clusters com alta taxa de diplomação, sendo que o Cluster 3 é composto praticamente por diplomados.

O Cluster 4 tem uma alta taxa de alunos matriculados e pelas médias dos indicadores de semestres, e a taxa de distorção de carga horária / período, é o Cluster que concentra os alunos retidos.

4.3.9 Análise da situação do aluno em cada Cluster

4.3.9.1 Evadidos

Os alunos evadidos do curso de Engenharia de Produção no Icea têm um perfil social e acadêmico diverso e representam 45.54% da base do curso. As notas do Enem parecem ter uma relação com o desempenho acadêmico dos alunos, sendo os clusters 0 e 2 dois dos clusters com menor nota média e pior desempenho em disciplinas do Decea.

Os indicadores acadêmicos, como o número de reprovações e a taxa de distorção carga horária / período evidenciam a dificuldade de adaptação dos alunos evadidos na Universidade. Mais uma vez, o posicionamento desses alunos no cluster depende da quantidade de disciplinas que ele cursou e, consequentemente, do número de períodos feitos. Com isso, conclui-se que os primeiros períodos são determinantes para a evasão ou não dos alunos do curso.

4.3.9.2 Diplomados

São 307 alunos diplomados, sendo 166 mulheres, no curso de Engenharia da Produção no Icea, até o momento da coleta da base de dados. Esse é o único curso que tem mais mulheres diplomadas do que homens.

O Cluster 3, que é o único cluster de todo o trabalho com mais mulheres que homens, é o Cluster com mais sucesso na Universidade e merece ser explorado pelo colegiado, buscando entender a jornada dos estudantes deste Cluster dentro e fora da Ufop.

4.3.9.3 Matriculados

Ao todo, os alunos ainda matriculados no curso representam 24.23% da base clusterizada e foram distribuídos entre todos os clusters, exceto pelo Cluster 3.

Os alunos matriculados que se encontram nos Clusters 0 e 2 são estudantes com um alto potencial de evasão e devem ser observados. Os alunos do Cluster 4 também precisam de atenção, já que é um cluster que tem alta taxa de reprovações e aparenta ser o cluster com os alunos que são retidos. Por último, os alunos do Cluster 1 por sua vez, têm alto potencial de diplomação.

4.3.10 Perfil dos Clusters

A Tabela 36 mostra o pefil de cada Cluster.

Tabela 36 – Informações médias por cluster ordenado por evasão

Cluster	**Sexo**	**Cor da Pele**	**Enem**	**Idade**	**PAA**	*AP*	*RPN*	*APD*	*SEM*	*CH/P*
0	M (59%)	Branca (45.53%)	652.08	22.18	Sim	1.08	1.01	0.04	1.87	0.02
2	M (53%)	Parda (46.67%)	640.72	20.44	Sim	8.88	5	0.37	4.98	0.18
4	M (57%)	Branca (47.86%)	638.64	21.15	Sim	29.09	6.5	0.71	8.96	0.53
1	M/F (50%)	Branca (45.48%)	662.03	20.28	Sim	49.13	5.92	0.79	11.03	0.79
3	F (59%)	Branca (45.45%)	688.58	19.3	Sim	49.55	5.66	0.78	11.3	0.8

Fonte: Elaborado pelo autor.

4.3.11 Resultado do curso

Como visto na Figura 32, a clusterização separou bem os alunos diplomados dos evadidos e, com isso, separou também os alunos matriculados de forma a evidenciar a proximidade deles com a evasão ou a diplomação.

Quando comparado com os outros cursos, o curso de Engenharia de Produção tem o melhor desempenho do Icea. Com uma taxa de diplomação mais alta e, consequentemente, uma taxa de evasão mais baixa que os outros cursos, o curso forma, em média, 58 alunos para cada 100 evadidos. Ainda assim, o curso enfrenta um pouco dos mesmos problemas que ecoam pelo Instituto.

Apesar de ter o menor número médio de semestres para a diplomação entre os cursos de engenharia, os alunos da Engenharia de Produção também enfrentam dificuldades de adaptação nos primeiros períodos, fator é determinante para a evasão ou não do aluno no curso.

Vale ressaltar também a relevância do sexo feminino nos Clusters 1 e 3, que tem os melhores indicadores no curso e, no Icea, ficam atrás apenas do Cluster 4 da Engenharia Elétrica.

4.4 Sistemas de Informação

Nas Seções a seguir, serão tratados e discutidos os dados do curso de Sistemas de Informação.

4.4.1 Atributos pré-universidade

A Tabela 37 apresenta as estatísticas descritivas dos atributos Idade em que entrou na Universidade e nota no Enem.

Tabela 37 – Atributos pré universidade do curso

Estatística	**Pontuação Enem**	**Idade que entrou**
Média	21.97	635.44
Desvio padrão	4.80	43.85
Mín	17.00	436.40
25%	19.00	606.02
50%	20.00	636.40
75%	24.00	667.33
Máx	55.00	763.50

Fonte: Elaborado pelo autor.

A Tabela 38 apresenta a frequência de valores para os atributos categóricos sexo e uso de ações afirmativas. É possível traçar um perfil esperado do aluno egresso no curso por meio dos atributos descritos nas Tabelas 37 e 38.

Tabela 38 – Uso de políticas afirmativas no curso

Sexo	**Usou política afirmativa**	**Quantidade**
M	Não	498
M	Sim	281
F	Não	148
F	Sim	83

Fonte: Elaborado pelo autor.

4.4.2 Atributos Acadêmicos

Como mencionado na Seção 3.6, todos os alunos tiveram informações agregadas para a clusterização. Essas informações serão usadas para entender a média dos atributos acadêmicos dos alunos deste curso.

Tabela 39 – Atributos acadêmicos do curso

Estatística	*AP*	*RPN*	*APD*	*SEM*	*CH/P*
Média	12.79	4.04	0.32	4.93	0.23
Desvio padrão	15.52	4.90	0.38	3.86	0.29
Mín	0.00	0.00	0.00	1.00	0.00
25%	0.00	0.00	0.00	1.00	0.00
50%	6.00	2.00	0.05	4.00	0.12
75%	20.00	6.00	0.67	8.00	0.40
Máx	57.00	29.00	1.00	20.00	1.17

Fonte: Elaborado pelo autor.

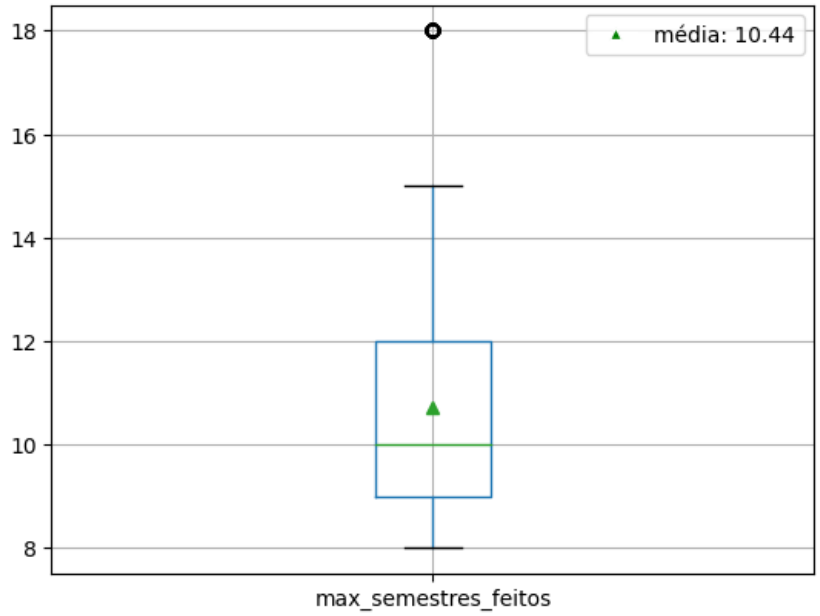
Unindo as Tabelas 37, 38 e 39 determinamos o perfil esperado do aluno do curso, segundo as bases de dados disponibilizadas pela Universidade. Portanto, um perfil esperado do aluno que entra no curso de Sistemas de Informação se dá por:

- sexo: masculino
- pontuação no Enem: 635.44
- idade que entrou: 21.97
- usou política afirmativa: não
- aprovações: 12.79
- reprovações: 4.04
- taxa de aprovação no Decea: 0.32
- semestres feitos: 4.93
- distorção carga horária/ período: 0.23

4.4.3 Tempo para a diplomação

O curso de Sistemas de Informação tem quatro anos de duração, ou seja, oito semestres. O *box plot* apresentado na Figura 33 mostra o comportamento da base de alunos do curso que foram diplomados até a coleta de dados, excluindo os alunos que se formaram em menos de seis períodos, já que são, provavelmente, alunos que reingressaram através do Enem ou alunos egressos por transferência.

Figura 33 – *Box plot*: Número de semestres feitos pelos alunos diplomados



Fonte: Elaborado pelo autor.

No curso de Sistemas de Informação, menos de 25% dos alunos se formam dentro do tempo esperado de diplomação e a média de tempo de formatura é de 10.22 semestres. Vê-se também como a mediana se comporta e como os quartis se distanciam no gráfico, apresentado na Figura 33.

4.4.4 Situação do aluno

Por último, a Tabela 40 mostra a participação de cada categoria de situação do aluno na base de alunos de Sistemas de Informação.

Tabela 40 – Descrição da situação dos alunos da base

Descrição situação aluno	**Total**	**%**
Evadido	610	60.40
Matriculado	253	25.05
Diplomado	131	12.97
Trancado	13	1.29
Afastado	3	0.30

Fonte: Elaborado pelo autor.

4.4.5 Análise de Componente Principal

A base de dados com os atributos dos alunos de Sistemas de Informação continha 1010 linhas e 357 colunas além do mask, e, para explicar 80% da variância total, foram necessárias 38 componentes. No curso de Sistemas de Informação há um número substancialmente maior de colunas, em relação aos cursos de Engenharia do Icea. Usando apenas três componentes, a nova base explica 39.82% da variância da base de dados que foi usada para o PCA.

4.4.5.1 Score features x componentes

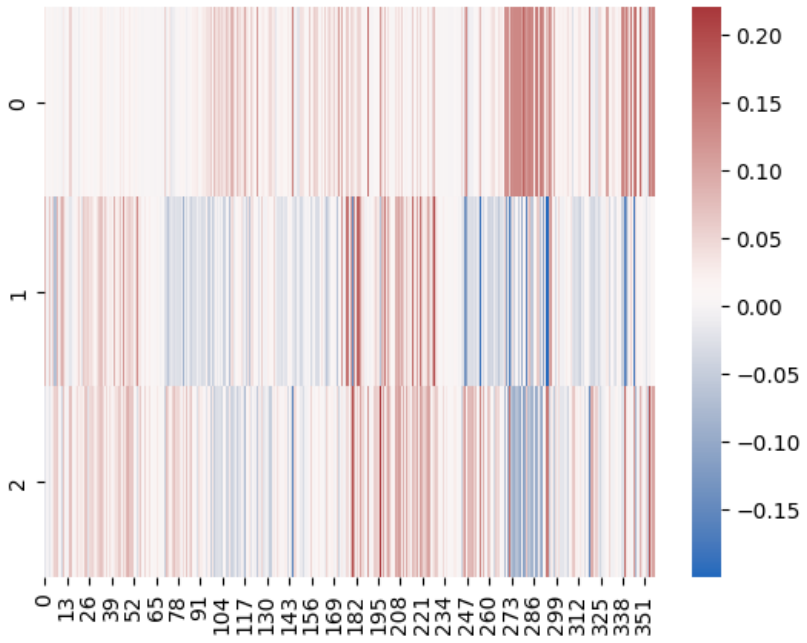
No mapa de calor apresentado na Figura 34, vemos a relação das componentes 0, 1 e 2 com as colunas originais da base de atributos. Por serem muitas colunas, os scores vão de, aproximadamente, -0.2 a 0.2. Os extremos são os que contém maior variabilidade de dados e, por isso, contribuem mais para explicar a variância da base de dados.

Na primeira componente, os valores mais escuros no mapa de calor se dão pelas notas de disciplinas específicas do curso, e o maior valor é a nota final na disciplina CSI443.

Na segunda componente, alguns valores negativos chamam a atenção, mas que também é referente à nota final de uma disciplina, mais especificamente, a disciplina CSI489. O maior valor na segunda componente é referente à nota final na disciplina CEA488.

A terceira componente tem maior valor na nota final na disciplina CEA422. Já o menor valor é dado pela quantidade de vezes em que a disciplina CSI734 foi cursada.

Figura 34 – Mapa de calor: Explicando as componentes da ACP



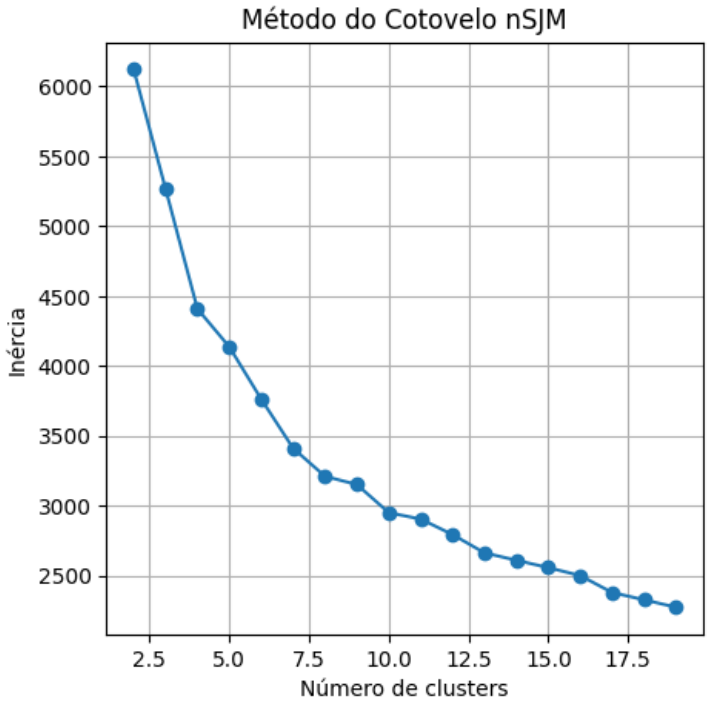
Fonte: Elaborado pelo autor.

É importante salientar que essas disciplinas são as que mais contribuem para a variabilidade da base de dados, mas isso não significa que sejam as disciplinas fundamentais para o sucesso ou o insucesso do aluno ao longo do curso.

4.4.6 Clusterização

Com o PCA aplicado, o número de colunas da base de dados foi reduzido para 1065 linhas e 27 colunas, e a partir disso, usando os métodos da silhueta e do cotovelo, define-se qual o melhor número possível de clusters para essa base. O método do cotovelo, que consiste em calcular a soma das distâncias quadráticas dos dados intra-clusters tem resultado mostrado na Figura 35.

Figura 35 – Método do cotovelo

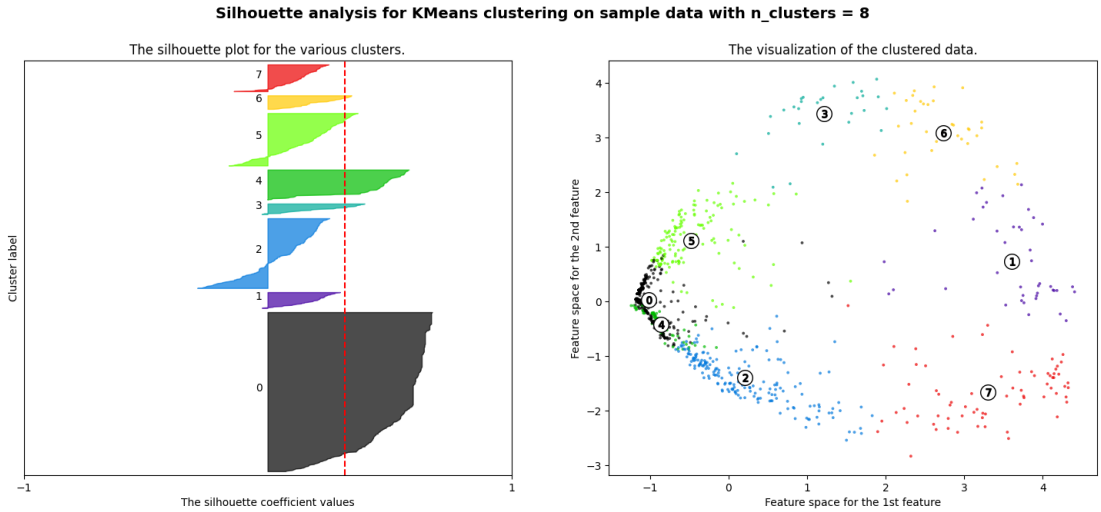


Fonte: Elaborado pelo autor.

Com o gráfico da Figura 35, fica visível a redução das distâncias com o aumento do número de clusters. Há um ganho de 2 para 3 clusters e um ganho de 7 para 8 clusters e por isso, serão os valores testados no método da silhueta. Como discutido na Seção 2.7, a proximidade do coeficiente da silhueta dos clusters com a média de coeficiente é importante para entender a qualidade dos agrupamentos.

Com 8 clusters, quase todos os valores de coeficiente da silhueta tocam a média de coeficiente, traçada em vermelho no gráfico à esquerda da Figura 36. O posicionamento dos Clusters considerando as duas primeiras componentes é apresentado no *scatterplot* à direita das Figuras 36 e 37.

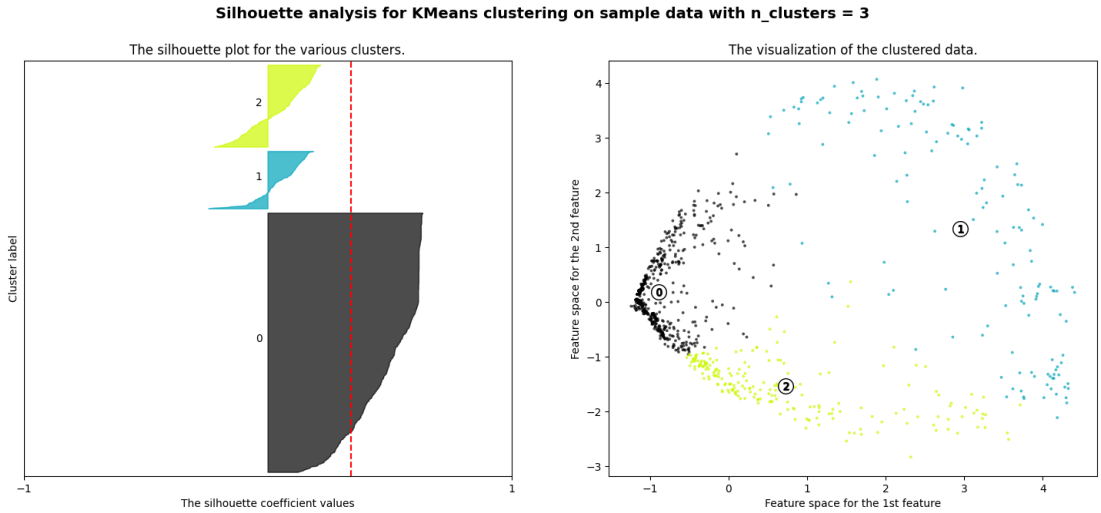
Figura 36 – Método da silhueta para 8 clusters



Fonte: Elaborado pelo autor.

Por outro lado, com 3 clusters, os valores de coeficiente da silhueta dos Clusters 1 e 2 não tocam a linha média de coeficiente da silhueta por uma distância considerável.

Figura 37 – Método da silhueta para 3 clusters



Fonte: Elaborado pelo autor.

Portanto, o melhor número de clusters para esse conjunto de dados é 8. Aplicando esse número de clusters ao algoritmo k-means, temos a divisão da base, mostrada na Tabela 41.

Tabela 41 – Observações por cluster no curso

Cluster	Número de observações
0	80.00
1	137.00
2	55.00
3	419.00
4	26.00
5	73.00
6	176.00
7	44.00

Fonte: Elaborado pelo autor.

4.4.7 Análise estatística dos clusters

A Tabela 42 mostra a distribuição étnica nos clusters do curso de Sistemas de Informação. É visível a participação dominante de alunos brancos e pardos no curso, o que também acontece no instituto.

Tabela 42 – Participação étnica por cluster

Cluster	Amarela %	Branca %	Não declarado %	Parda %	Preta %
0	2.50	55.00	1.25	35.00	6.25
1	3.65	29.20	2.19	48.18	16.79
2	0.00	43.64	0.00	50.91	5.45
3	1.19	38.42	4.53	45.82	10.02
4	0.00	53.85	0.00	34.62	11.54
5	0.00	31.51	2.74	53.42	12.33
6	1.14	44.89	1.14	39.77	13.07
7	0.00	34.09	0.00	56.82	9.09

Fonte: Elaborado pelo autor.

A Tabela 43, por sua vez, mostra a porcentagem de alunos de cada cluster que usou política afirmativa.

Tabela 43 – Uso de política afirmativa por cluster

Cluster	Não %	Sim %
0	51.25	48.75
1	73.72	26.28
2	63.64	36.36
3	67.30	32.70
4	69.23	30.77
5	61.64	38.36
6	51.70	48.30
7	75.00	25.00

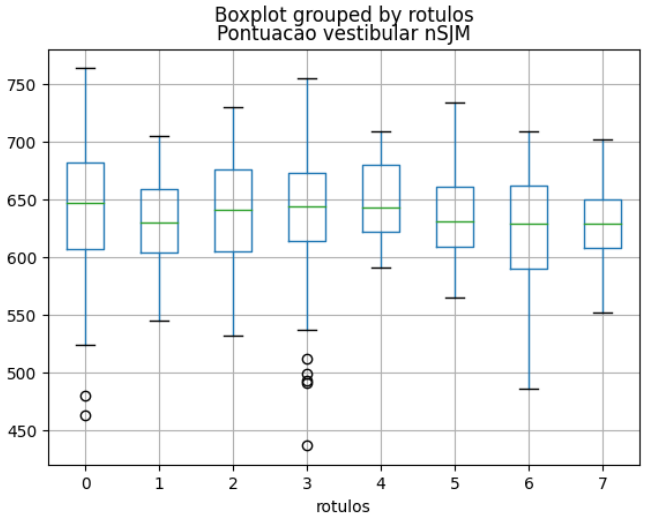
Fonte: Elaborado pelo autor.

Mais uma vez, a presença de alunos brancos e pardos é dominante. A participação de alunos que usaram algum tipo de política afirmativa é muito mais baixa nos Clusters 1, 2, 3, 4 e 5. Além disso, nenhum dos Clusters teve maior participação de alunos que usaram cotas.

4.4.7.1 Nota no Enem

A partir das Tabelas 42 e 43, temos uma relação de cor de pele e uso de políticas afirmativas para cada Cluster. Essas informações podem ser relevantes para entender como os alunos dos Clusters se saíram no Enem. O *box plot* apresentado na Figura 38 mostra como os Clusters se comportam nesse quesito.

Figura 38 – *Box plot*: Nota no Enem por Cluster



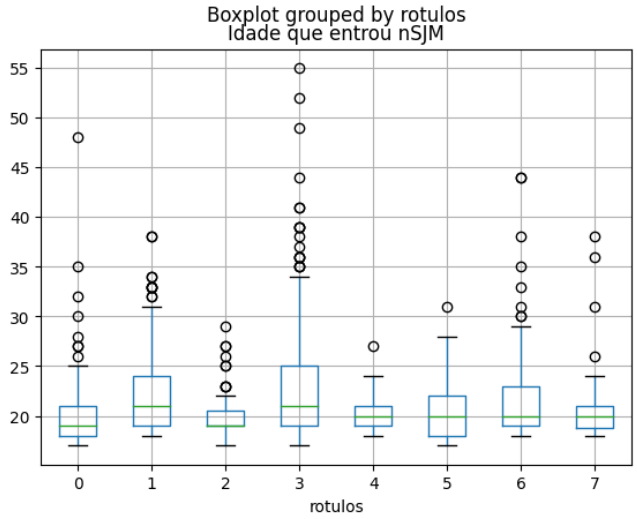
Fonte: Elaborado pelo autor.

O gráfico apresentado na Figura 38 mostra uma alta dispersão de notas no Enem nos Cluster 0, 3 e 6. O Cluster 4, que tem poucos alunos, chama a atenção pela menor dispersão de notas entre todos os clusters. Os outros clusters não chamam atenção neste gráfico.

4.4.7.2 Idade que entrou

O *box plot* da Figura 39 mostra a distribuição da idade dos alunos quando ingressaram na universidade.

Figura 39 – *Box plot*: Idade que entrou por Cluster



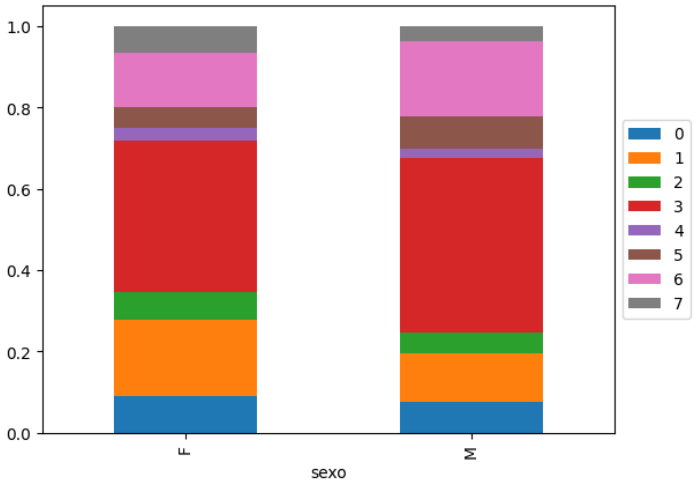
Fonte: Elaborado pelo autor.

O box plot mostra uma alta dispersão de idade nos clusters 0, 3 e 6, que são, justamente, os clusters mencionados com a maior variabilidade de notas no Enem. Essa relação de idade e nota no Enem se repete mais uma vez ao ver, principalmente, como os Clusters 3, 4 e 5 se comportam nas Figuras 38 e 39.

4.4.7.3 Sexo

Mais uma vez, o sexo masculino é dominante no curso e então, para uma análise mais clara, o gráfico apresentado na Figura 40 é a porcentagem dos clusters em cada sexo.

Figura 40 – Porcentagem de sexo por cluster



Fonte: Elaborado pelo autor.

A participação dos Clusters 3, 5 e 6 é maior no sexo masculino enquanto a participação dos Clusters 1 e 7 é maior no sexo feminino. Os outros Clusters não tem tanta diferença de participação entre os sexos. Por fim, a Tabela 44 mostra o perfil social esperado de cada Cluster no curso de Sistemas de Informação.

Tabela 44 – Perfil esperado de cada Cluster

Cluster	**Sexo**	**Cor da Pele**	**Nota no Enem**	**Desvio padrão**	**Idade em que entrou**	**Desvio padrão**	**Usou política afirmativa ?**
0	Masc.	Branca (55.00%)	640.37	57.70	20.73	4.53	Não
1	Masc.	Parda (48.18%)	629.90	32.61	22.20	4.57	Não
2	Masc.	Parda (50.91%)	638.52	46.92	20.27	2.56	Não
3	Masc.	Parda (45.82%)	640.58	44.07	22.87	5.52	Não
4	Masc.	Branca (53.85%)	647.12	34.40	20.31	2.00	Não
5	Masc.	Parda (53.42%)	634.18	38.76	20.71	2.96	Não
6	Masc.	Branca (44.89%)	624.92	46.62	21.78	4.33	Não
7	Masc.	Parda (56.82%)	628.36	33.29	20.98	4.32	Não

Fonte: Elaborado pelo autor.

Alunos brancos e pardos que não usaram cotas são maioria no curso de Sistemas de Informação, replicando o que acontece no instituto. Apesar das diferenças de dispersão nas notas no Enem, a média dos clusters converge para a faixa dos 635 pontos.

4.4.8 Estatísticas de atributos acadêmicos

A Tabela 45 agrega, por cluster, algumas das informações que foram usadas para a clusterização.

Tabela 45 – Estatística dos atributos acadêmicos por cluster

Cluster	*AP*	*RP*	*RPN*	*APD*	*SEM*	CH/P
0	8.12	1.1	0.69	0.71	2.99	0.19
1	8.15	6.34	3.42	0.26	5.14	0.16
2	37.29	7.33	1.8	0.69	9.58	0.65
3	1.05	1.43	2.44	0.04	1.93	0.01
4	40.35	5.46	0.96	0.79	10.23	0.73
5	42.82	6.63	1.1	0.75	10.93	0.72
6	14.91	6.38	3.35	0.4	6.5	0.31
7	42.18	8.5	1.73	0.73	11.25	0.73

Fonte: Elaborado pelo autor.

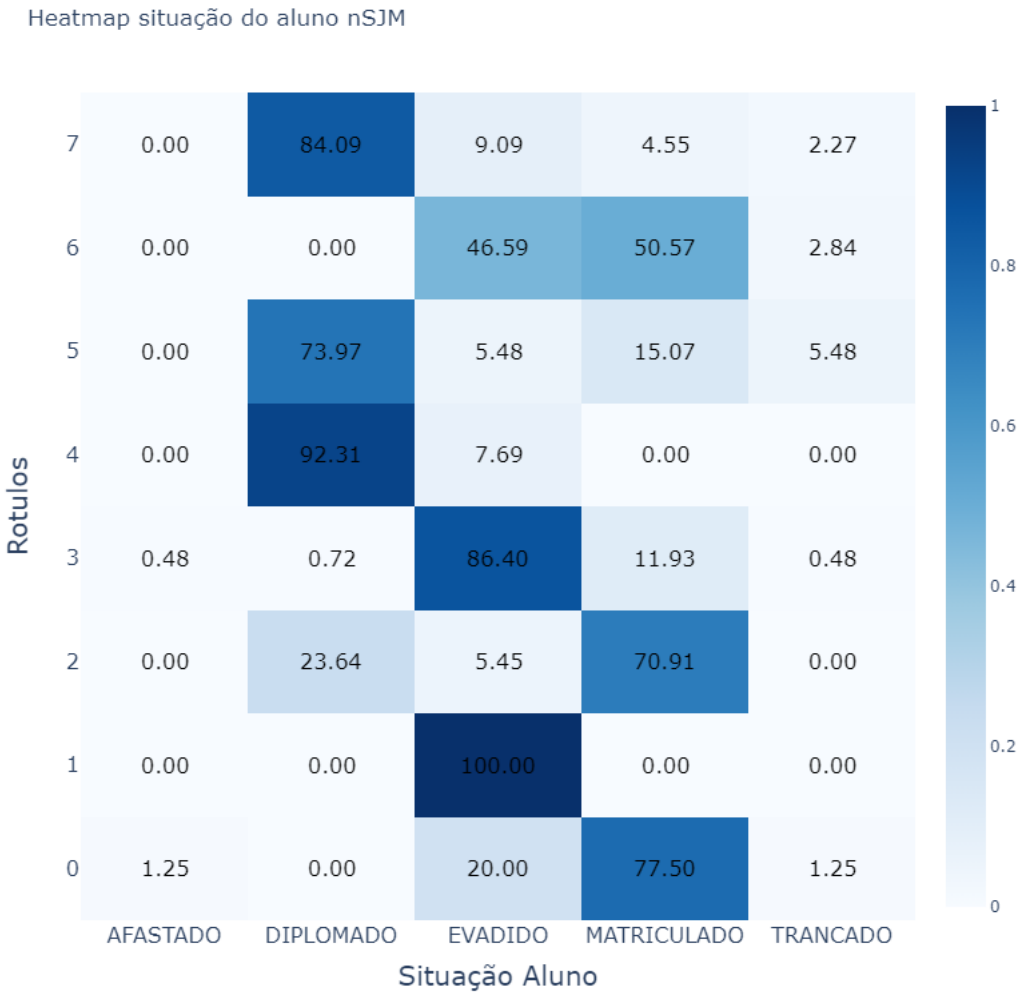
Com essas estatísticas traçamos um perfil acadêmico esperado de cada Cluster no curso de Sistemas de Informação. Além disso, fica evidente que os atributos escolhidos para a clusterização foram suficientes para separar os perfis de alunos do curso.

Os Clusters 2, 4, 5 e 7 têm o maior sucesso acadêmico e os Clusters 0, 1, 3 e 6 são os que têm pior desempenho, chamando a atenção para o Cluster 3 que tem os piores indicadores, não só no curso, mas em todo o Icea.

4.4.8.1 Estatística da situação do aluno

No mapa de calor da Figura 41 podemos ver as situações de aluno por cluster e comparar com as Tabelas 44 e 45.

Figura 41 – Mapa de calor: Situação do aluno por Cluster



Fonte: Elaborado pelo autor.

O mapa de calor da Figura 41 relaciona a situação dos alunos com o cluster em que eles estão localizados. Cada célula é a porcentagem de alunos na situação descrita presente no cluster e a soma das linhas tem resultado 100%.

Apenas pelo mapa de calor fica evidente a separação entre os alunos diplomados e os alunos evadidos. Em clusters onde há uma quantidade significativa de diplomados, há uma quantidade muito pequena de evadidos e o contrário também é verdade. Por serem muitos clusters, eles serão analisados um a um.

- Cluster 0: Uma grande quantidade de alunos matriculados. Esses alunos ainda estão no começo da jornada universitária, por isso, foram agrupados com alunos evadidos que cursaram poucas matérias. É difícil afirmar se os matriculados deste Cluster se aproximam de fato da evasão.

- Cluster 1: Esse Cluster chama a atenção, não só pelos indicadores acadêmicos ruins, mas também por ser um cluster com 100% dos alunos evadidos. Esse Cluster deve ser analisado pelo colegiado para entender o perfil e a jornada acadêmica dos alunos contidos neste agrupamento.
- Cluster 2: É um Cluster com muitos alunos matriculados e alguns já diplomados. Por ter uma distorção de carga horária/período alta, indica que os alunos deste Cluster são parte do fenômeno de retenção, mas que têm caminhado à diplomação
- Cluster 3: O Cluster com mais alunos têm indicadores ruins e uma taxa de desistência alta. Os alunos deste Cluster que ainda estão matriculados devem ser acompanhados mais de perto pelo colegiado.
- Cluster 4: Alta taxa de diplomação e nenhum aluno matriculado. Por ser o menor Cluster, tende a ter indicadores médios que convergem.
- Cluster 5: O Cluster 5 tem uma taxa de diplomação alta, mas chama a atenção pela alta quantidade de alunos que trancaram o curso perto da diplomação.
- Cluster 6: O Cluster 6 tem uma alta taxa de evasão e muitos alunos matriculados. Os indicadores deste agrupamento são ruins e os alunos matriculados devem ser acompanhados pelo colegiado.
- Cluster 7: Alta taxa de diplomação e bons indicadores, apesar do número médio de semestres ser maior que o número médio de semestres para ser diplomado no curso.

4.4.9 Análise da situação do aluno em cada Cluster

4.4.9.1 Evadidos

Os alunos evadidos do curso de sistemas de informação representam 60.40% da base clusterizada. Os indicadores acadêmicos dos alunos que evadiram são muito ruins se comparados com os alunos diplomados do curso ou de alunos evadidos de outros cursos. O alto número de reprovações e a taxa de distorção carga horária / período deixam clara a dificuldade de adaptação dos alunos evadidos do curso. Desta vez, o posicionamento desses alunos no cluster depende menos da quantidade de disciplinas que ele cursou e do número de períodos feitos.

No curso de Sistemas de Informação, a adaptação e o desempenho do aluno nos primeiros períodos são determinantes para a evasão do curso.

4.4.9.2 Diplomados

São 131 alunos diplomados pelo curso de Sistemas de Informação até o momento de coleta da base, isso representa quase 13% da base. Os alunos diplomados têm bons indicadores, chamando a atenção para o Cluster 4, que apesar dos poucos alunos, tem indicadores bons quando comparados com os outros clusters com alta taxa de diplomação.

4.4.9.3 Matriculados

Ao todo, os alunos ainda matriculados no curso representam 25.05% da base clusterizada e foram distribuídos entre todos os clusters, exceto pelos clusters 1 e 4. Os alunos matriculados que se encontram nos clusters 3 e 6 são estudantes com um alto potencial de evasão e devem ser observados pelo curso . Os alunos do Cluster 0, como mencionado anteriormente, são alunos que ainda estão no começo da jornada acadêmica e por isso é difícil classificá-los usando a clusterização.

Os alunos matriculados dos Clusters 2 são alunos com potencial de diplomação mas que ainda têm alguns períodos pela frente. Por último, os clusters 5 e 7 agregam os alunos com maior tempo de Universidade e que estão mais próximos da diplomação que os alunos do Cluster 2.

4.4.10 Perfil dos Clusters

A Tabela 46 mostra o pefil de cada Cluster.

Tabela 46 – Perfil esperado por cluster ordenado por evasão

Cluster	**Sexo**	**Cor da Pele**	**Enem**	**Idade**	**PAA**	*AP*	*RP*	*APD*	*SEM*	CH/P
1	M (69%)	Parda (48.18%)	629.9	22.2	Não	8.15	6.34	0.26	5.14	0.16
3	M (79%)	Parda (45.82%)	640.58	22.87	Não	1.05	1.43	0.04	1.93	0.01
6	M (82%)	Branca (44.89%)	624.92	21.78	Não	14.91	6.38	0.4	6.5	0.31
0	M (74%)	Branca (55.00%)	640.37	20.73	Não	8.12	1.1	0.71	2.99	0.19
7	M (66%)	Parda (56.82%)	628.36	20.98	Não	42.18	8.5	0.73	11.25	0.73
4	M (73%)	Branca (53.85%)	647.12	20.31	Não	40.35	5.46	0.79	10.23	0.73
5	M (84%)	Parda (53.42%)	634.18	20.71	Não	42.82	6.63	0.75	10.93	0.72
2	M (71%)	Parda (50.91%)	638.52	20.27	Não	37.29	7.33	0.69	9.58	0.65

Fonte: Elaborado pelo autor.

4.4.11 Resultado do curso

Como visto na Figura 41, a clusterização separou bem os alunos diplomados dos evadidos e, com isso, separou também os alunos matriculados de forma a evidenciar a proximidade deles com a evasão ou a diplomação.

O curso de Sistemas de Informação tem um desempenho um pouco melhor que o curso de Engenharia de Computação, mas ainda assim, sofre com o fenômeno da evasão. Com 57.88% dos alunos da base sendo evadidos, o curso tem a segunda maior taxa de evasão no Icea e, consequentemente, é o segundo curso que, proporcionalmente, menos forma alunos.

Com a taxa de distorção carga horária/período média muito distante de 1 e os e o número médio de semestres para a diplomação alto, o curso de Sistemas de Informação tem o segundo pior desempenho frente aos atributos criados na Seção 3.6.

5 Discussão de resultados

5.1 Comparação global entre o encontrado em cada curso

Comparando os resultados da clusterização de cada curso do Icea é possível entender algumas das razões pelas quais quase 55% dos alunos matriculados de 2011 pra cá evadem do curso em algum momento. Há evidência suficiente para afirmar que os primeiros períodos dos cursos são determinantes para a evasão ou não do aluno. Isso se prova pela relação entre o número de aprovações e reprovações médias e a média de semestres feitos nos clusters onde há um grande número de alunos evadidos. Essa estatística não é única e puramente a razão do alto índice de evasões no instituto, afinal, há indicadores que não podem ser medidos, como a adaptação à cidade, renda familiar do aluno, distância do lar, e vários outros fatores que a Universidade não tem o controle. No entanto, sobre o que a Ufop tem controle, fica evidente a necessidade de um nivelamento dos alunos nos primeiros semestres, já que o desempenho dos alunos no início da graduação é fundamental para a diplomação.

5.2 Comparação entre cursos

A Tabela 47 mostra, além da distorção de carga horária / período (CH/P), as taxas de evasão, diplomação e a taxa de diplomados por evadidos que são dadas por:

- TX_1: taxa de diplomação

$$TX_1 = \frac{\text{\# total de alunos diplomados}}{\text{\# total de alunos na base}}$$

- TX_2: taxa de evasão

$$TX_2 = \frac{\text{\# total de alunos evadidos}}{\text{\# total de alunos na base}}$$

- TX_3: taxa de diplomados por evadidos

$$TX_3 = \frac{\text{\# total de alunos diplomados}}{\text{\# total de alunos evadidos}}$$

Os números apresentados na Tabela 47 evidenciam a dificuldade em diplomar alunos em todos os cursos do Icea. Além disso, fica evidente a relação entre as taxas TX_1, TX_2e TX_3 com o atributo CH/P, que foi calculado para o Icea usando a média ponderada dos valores de CH/P dos cursos e suas respectivas contribuições para a base de dados.

Tabela 47 – Taxas de evasão e diplomação no Icea

	TX_1	TX_2	TX_3	CH/P
Icea	0.295	0.550	0.162	0,28
Engenharia Elétrica	0.297	0.551	0.164	0.31
Engenharia de Computação	0.161	0.623	0.100	0.22
Engenharia de Produção	0.583	0.452	0.264	0.36
Sistemas de Informação	0.203	0.579	0.117	0.23

Fonte: Elaborado pelo autor.

As bases de dados usadas para a clusterização se comportaram de forma semelhante entre os cursos de engenharia na quantidade de colunas, de componentes do PCA, no número de clusters e também no comportamento dos clusters. Por outro lado, o curso de Sistemas de Informação teve comportamento diferente, com mais colunas e, consequentemente, mais componentes do PCA e do número de clusters.

Os desempenhos dos clusters com maior taxa de evasão em cada curso foram semelhantes. Com excessão ao curso de Engenharia de Computação, os alunos que tiveram menor taxa de distorção carga horária período, menor taxa de aprovação no Decea e menor número de semestres feitos são os alunos com maior possibilidade de insucesso nos cursos.

Por último, foi observado que há uma relação entre nota no Enem, idade do aluno quando ingressou e a evasão em todos os cursos. Apesar de observados, os fatores dessa relação não foram evidenciados através do entendimento das bases, mas é relevante o suficiente para ser discutido em um trabalho futuro.

6 Considerações Finais

Este trabalho analisou dados acadêmicos e sociais dos alunos do Icea. Com do uso de técnicas de DM foi possível entender, de forma orientada a dados, o perfil dos alunos evadidos, diplomados e retidos de cada curso do Icea através da clusterização. Além disso, o uso da metodologia CRISP-DM foi importante para o entendimento das etapas do trabalho.

A análise das matrizes curriculares possibilitou observações mais refinadas sobre a situação dos alunos nos cursos, já que, os atributos foram construídos usando informações da matriz. O desenvolvimento destes atributos acadêmicos foi fundamental para que a clusterização tivesse um resultado satisfatório na segmentação dos alunos.

A clusterização com atributos acadêmicos, apesar de ser uma técnica de análise não supervisionada, conseguiu segregar os alunos de acordo com a proximidade deles com a evasão, diplomação ou retenção. Os resultados alcançados mostram-se promissores em relacionar os atributos usados para a clusterização com a situação do aluno. Com esses atributos, a Universidade pode desenvolver políticas para acompanhar os atributos e tomar medidas preventivas para reduzir as taxas de evasão e retenção dos alunos ainda matriculados.

Por último, o presente trabalho conseguiu mostrar, através dos dados, as taxas de diplomação e evasão dos cursos e relacioná-las com os atributos médios dos cursos. Isso foi importante para evidenciar as relações de desempenho acadêmico com a evasão e retenção.

Como trabalhos futuros, sugere o uso de técnicas supervisionadas de análise dos atributos dos alunos e dos cursos a fim de identificar e estratificar a situação do aluno evadido por causa de evasão. Outra sugestão é compreender, junto aos alunos evadidos, os critérios que contribuíram para a evasão para além do desempenho acadêmico. Por fim, um trabalho que perpetue as análises feitas no presente trabalho para acompanhamento de alunos que estão em clusters com alta taxa de evasão.

Referências

BAGGI, C. A. d. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 16, n. 02, p. 355–374, 2011.

BAI, L.; LIANG, J.; CAO, F. A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. **Information Fusion**, Elsevier, v. 61, p. 36–47, 2020.

BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. d. C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 30, p. 669–694, 2021.

BRASIL. **Lei Nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD)**. Brasília, DF, 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

CALDEIRA, D. M. Caracterização do problema de evasão de discentes nos cursos do icea mediante técnicas de mineração de dados. 2021.

CAUCHICK-MIGUEL, P. A.; FLEURY, A.; MELLO, C. H. P.; NAKANO, D. N. **Metodologia de pesquisa em engenharia de produção e gestão de operações**. 3. ed. Rio de Janeiro: Elsevier, 2018. 244 p. ISBN 9788535291346. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788595153561>. Acesso em: 8 abr. 2022.

FAN, J.; SUN, Q.; ZHOU, W.-X.; ZHU, Z. Principal component analysis for big data. **arXiv preprint arXiv:1801.01602**, 2018.

GAIOSO, N. P. d. L. O fenômeno da evasão escolar na educação superior no brasil. **Brasília, DF: Universidade Católica de Brasília**, p. 20, 2005.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining**. [S.l.]: Elsevier Brasil, 2015.

GONÇALVES, A. P. F. Definição de um modelo de inteligência artificial para a identificação do padrão curricular dos alunos do icea. 2022.

HAIR, J. F. Multivariate data analysis. 2009.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and science**, v. 5, n. 1, p. 83–90, 2016.

INEP. **Metodologia de Cálculo dos indicadores de fluxo da educação superior**. [S.l.]: INEP, Brasília, 2017.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys (CSUR)**, Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

JOLLIFFE, I. T. **Principal component analysis for special types of data**. [S.l.]: Springer, 2002.

KETCHEN, D. J.; SHOOK, C. L. The application of cluster analysis in strategic management research: an analysis and critique. **Strategic management journal**, Wiley Online Library, v. 17, n. 6, p. 441–458, 1996.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, Springer, v. 50, p. 159–179, 1985.

MUELLER, J. P.; MASSARON, L. **Aprendizado de máquina para leigos**. [S.l.]: Alta Books Editora, 2019.

PARANHOS, H. P. Desenvolvimento de um dashboard para análise e visualização dos dados educacionais dos discentes do instituto de ciências exatas e aplicadas da ufop. 2021.

RODRIGUES, E. H. A. Análise de caracterização quantitativa e predição da evasão escolar nos cursos da área de computação do icea por meio de técnicas de data science. 2022.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987.

SEMESP. Mapa do ensino superior no brasil. **SEMESP, São Paulo**, 2022.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000.

VARGAS, M. R. M.; LIMA, S. M. V. Barreiras à implantação de programas de educação e treinamento a distância. In: **CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA**. [S.l.: s.n.], 2004. v. 11.