CORRELAÇÃO ENTRE PERFIL E DESEMPENHO DE ESTUDANTES DE GRADUAÇÃO E AS TAXAS DE CONCLUSÃO DE CURSO E EVASÃO

André Luis Maciel Leme[1]

RESUMO

As instituições de ensino buscam compreender o universo de estudantes e seus egressos através de diversos recursos e ferramentas. As estatísticas ajudam consideravelmente a entender o cenário quantitativo dos resultados obtidos quanto a estudantes formados, evadidos e outras classificações inerentes ao contexto. Observa-se um constante estudo e preocupação com as taxas de evasão, porém neste estudo, objetiva-se olhar para o estudante concluinte, buscando coletar dados que sinalizem os fatores que os levam ao sucesso. Tais fatores podem contribuir para focalizar os esforços no que dá certo ao contrário de olhar para o que não deu certo, neste contexto, a evasão. O trabalho fez uso dos dados de estudantes de um curso de graduação (tecnólogo) de uma instituição pública federal e de um campus específico. Os dados foram saneados e tratados para que pudessem ser submetidos a duas técnicas de aprendizado de máquina não supervisionado, *k-means* e *Self Organizing Maps*. Os atributos selecionados para o processamento mostraram resultados de agrupamento satisfatórios, sobretudo em *Self Organizing Maps* onde um gráfico com 400 quadrantes (dimensões 20x20) apresentou agrupamentos de estudantes com características bastante similares denotando a correlação proposta no estudo. Também foi possível discorrer sobre a diferença entre as duas técnicas a partir de amostragens apresentadas. A conclusão foi que características específicas contribuem para o sucesso na conclusão do curso por um grupo de estudantes com atributos similares.

Palavras-chave: Algoritmos de Clusterização. Evasão. Análise de Dados. Aprendizado de Máquina.

1 INTRODUÇÃO

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), criado no final da década de 1930, tem como missão, segundo (INEP, 2020), subsidiar a formulação de políticas educacionais dos diferentes níveis de governo com intuito de contribuir para o desenvolvimento econômico e social do país.

Abrange diversas áreas da educação no país, entre elas o SINAES que tem com um dos instrumentos o ENADE que avalia o desempenho de concluintes dos cursos de graduação no que diz respeito ao seu rendimento quanto aos conteúdos programáticos, habilidade e competências adquiridas ao longo de sua formação.

A partir dos resultados do ENADE o INEP apresenta indicadores de qualidade da educação superior. Vale observar que esses indicadores são divulgados para cursos que tiveram pelo menos dois estudantes concluintes participantes da prova.

A menção ao ENADE neste trabalho remete as críticas que são feitas sobre a efetividade de tal instrumento para avaliação de um curso de graduação. Segundo (EDUCAÇÃO, 2019) e (KUZUYABU, 2019) a Organização para Cooperação e Desenvolvimento Econômico (OCDE), em seu último relatório, “Repensando a Garantia da Qualidade do Ensino Superior”, apresenta algumas críticas que vão desde a sua concepção até a avaliação dos resultados. São

[1] Estudante de pós-graduação latu sensu do SENAC São Paulo, andre.m.leme@gmail.com

críticas construtivas tanto que sugerem aprimoramentos que, segundo eles, tornariam os indicadores m ais relevantes e eficazes.

O curso objeto deste projeto teve nota 5 no ENADE, portanto, evidenciando a efetividade do ensino e qualidade do curso, sob os conceitos ENADE. A reflexão a ser considerada é a amostragem de estudantes que participam de tal avaliação

Neste contexto, os indicadores do ENADE não são suficientes, porque tratam-se de concluintes. Aqueles que entram nas estatísticas de sucesso na evolução do curso. Este fato deve ser levado em consideração neste estudo para buscar entendimento do que é relevante e pertinente para esse sucesso ser alcançado e buscar replicar ou expandir os "sintomas" para alcançar maior índice de sucesso.

Em contra ponto é necessário vislumbrar os indicadores e sintomas de insucesso. A oportunidade de visualizar o que deu e o que deu errado dentro de um mesmo contexto pode ajudar a rever conceitos e necessidades de aprimoramento e ajustes.

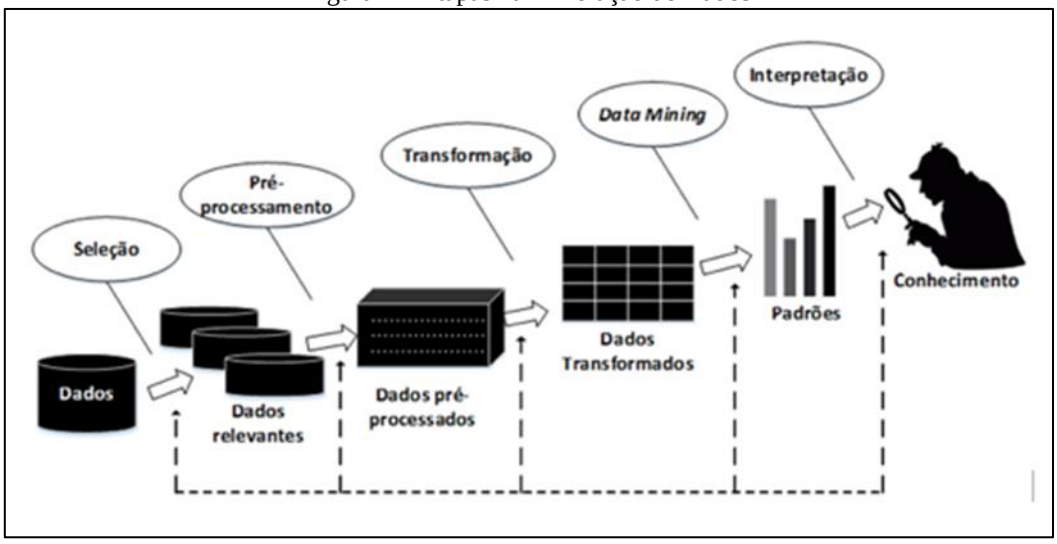
E educação 3.0 fomentada por diversos autores, trata de endereçar o aparato digital atualmente vigente a nova geração de estudantes. (LENGEL, 2012) menciona a evolução dos ambientes de trabalho sinalizando a mudança comportamental e o impacto nas relações humanas com o advento de novas formas se realizar as tarefas. Esses aspectos devem ser levados em consideração ao analisar o entorno dos estudantes, a princípio com base em seus dados do registro escolar e a posteori buscar informações complementares que possa corroborar com a análise de sucesso e insucesso na realização do curso de graduação.

Uma possível confrontação da avaliação do ENADE com os estudantes daquele período letivo ou até daquela turma pode apresentar indicadores significativos.

O uso dos registros escolares deve ser executado respeitando diretrizes de anonimização sem que se possa identificar quaisquer um dos indivíduos participantes das informações. A LGPD (BRASIL, 2018) entrou em vigor em 2021 e traz regulamentações e esclarecimento sobre terminologias usuais na área de segurança de dados, mas que agora precisam receber mais atenção. Segundo a lei, em seu artigo 5º item III, dado anonimizados é aquele em que o titular não pode ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento.

Para a análise dos dados serão aplicados conceitos e técnicas de mineração de dados que compreendem etapas que vão desde a coleta até a produção de conhecimento sobre as informações como ilustra a Figura 1.

Figura 1 – Etapas na Mineração de Dados



Fonte: Adaptado de Fayyad et al. (1996)

Segundo (SHARDA, DELEN e TURBAN, 2018), a mineração de dados é o caminho para desenvolver inteligência a partir de um conjunto de dados. O primeiro passo é ter uma visão, não necessariamente precisa, do que se espera encontrar com a amostragem de dados. Neste trabalho existe um objetivo geral, porém os resultados na aplicação dos algoritmos podem suscitar outras abordagens.

A seleção dos dados relevantes e seu tratamento, identificado na Figura 1 como pré-processamento envolve diversos. Conforme (CUESTA e KUMAR, 2016), construir soluções analíticas de dados do mundo real requer dados precisos. Depois da coleta é preciso limpar, normalizar e transformar dados brutos em um formato padrão que possam ser processados por algoritmos específicos.

Os dados transformados ou tratados para serem processados serão então submetidos a algoritmos de aprendizado de máquina do tipo *clustering* ou para agrupamento de dados. (SHARDA, DELEN e TURBAN, 2018) descrevem estes tipos de algoritmos com aqueles que particionam em segmentos (ou agrupamentos naturais), uma coleção de coisas (objetos, eventos) em um conjunto de dados estruturados cujos membros compartilham características semelhantes. Esta é a proposta com os dados dos registros escolares, encontrar similaridades entre os estudantes bem-sucedidos em um grupo e evadidos em outro e depois buscar correlações ou contra pontos entre um e outro.

Os algoritmos de *clusterização* mais comumente utilizados, segundo (SHARDA, DELEN e TURBAN, 2018), incluem *k-means* (de estatísticas) e mapas auto organizáveis (de aprendizado de máquina), que é uma arquitetura de rede neural exclusiva desenvolvida por (KOHONEN, 2001).

Por fim será adotada a análise de cluster para dissertar sobre os resultados.

> A Análise de Clusters é um procedimento da Estatística Multivariada que tenta agrupar um conjunto de dados em grupos homogéneos, chamados clusters; os dados podem ser objetos ou variáveis. Ou seja, cada observação pertencente a um determinado cluster é idêntica a todas as outras pertencentes a esse cluster e é diferente das observações pertencentes aos outros clusters (FARIA, 2014).

2 REFERENCIAL TEÓRICO

Os três pilares da formação acadêmica se completam enquanto consolidadores dos conhecimentos adquiridos na academia. Ensino, pesquisa e extensão estão presentes nas organizações didáticas de todas os cursos e instituições buscando conduzir o estudante ao processo investigativo e ao compartilhamento e aquisição de conhecimento além do âmbito da sala de aula.

A pesquisa acadêmica voltada para um resultado prático que contribui para o aprimoramento de algum produto, serviço ou instituição tem resposta e aceitação mais rápida.

Uma busca por trabalhos similares ao que se pretende realizar, pode ajudar a organizar o direcionamento do texto. Evasão são temas mais frequentes nas pesquisas acadêmicas sobre cursos e instituições de ensino. (HOED, 2016) argumenta que o problema de evasão é conhecido e acrescenta frequentemente se notícia que o número de estudantes formados é bem menor do que aqueles matriculados.

Em trabalho de teor similar, (MACHADO e CAVALCANTI, 2010) argumentam que quando ocorre a evasão de um estudante, as repercussões sociais adquirem uma extensão preocupante porque além do significado social que a educação possui, ela é considerada um dos setores mais importantes para o desenvolvimento de uma nação.

O INEP fornece ampla variação de dados para análise, mas são oriundos de avaliações específicas como o ENADE, não se pode obter dados individualizados dos estudantes de um

curso todo porque os participantes do exame são pré-selecionados conforme o momento em que estão no curso. A amostragem para a proposta desta pesquisa é muito pequena.

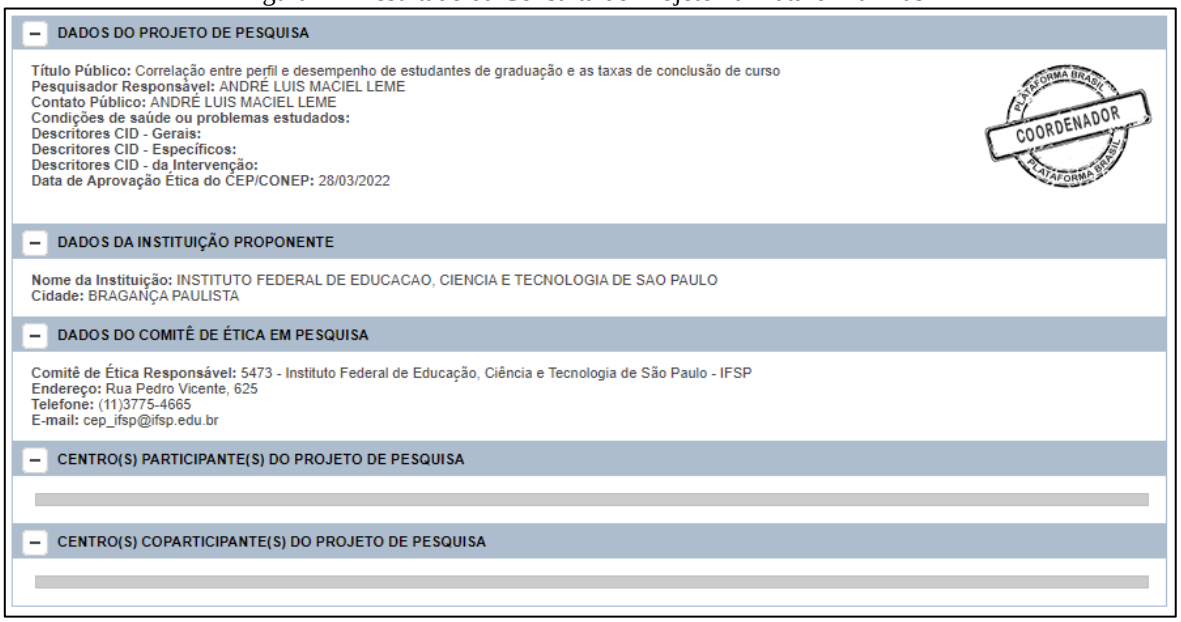
O uso de técnicas estatísticas gerais e algoritmos de aprendizado de máquina serão as ferramentas fundamentais para a obtenção dos resultados e sua análise. (CUESTA e KUMAR, 2016) e (SHARDA, DELEN e TURBAN, 2018) fornecem suporte para a utilização das técnicas.

3 METODOLOGIA

Os dados foram extraídos do Sistema Unificado de Administração Pública (SUAP), usado para apoio a gestão educacional da instituição de ensino foco do estudo.

O projeto foi submetido ao comitê de ética do Instituto Federal de São Paulo através da Plataforma Brasil, recebendo aprovação em 28/03/2022. O projeto pode ser localizado pesquisando-se na Plataforma Brasil[2] pelo nome deste pesquisador ou pelo título do trabalho. A Figura 2 apresenta uma captura da tela com o resultado da pesquisa e dados do projeto..

Figura 2 – Resultado da Consulta do Projeto na Plataforma Brasil



Fonte: Autoria Própria

Foram extraídos dados dos estudantes através das funcionalidades oferecidas pelo sistema mencionado. Os parâmetros de extração abrangeram o período de 2009 a 2021, de onde se obteve informações diversas sobre o perfil dos estudantes como ano de ingresso, data de nascimento, escola de origem, tempo de conclusão do curso (para os formados), entre outros dados.

Embora todos os estudantes tenham sido considerados para a extração, o tratamento dos dados para análise limitou aos estudantes com situação “evasão” e “formado”, totalizando 695 registros de dados. As demais situações não entraram no escopo desta pesquisa.

O Quadro 1 apresenta o dicionário de dados do *dataset* obtido e a relevância e significado dos dados.

[2] https://plataformabrasil.saude.gov.br/

Quadro 1 – Atributos do *dataset*, aqueles assinalados com X são aqueles utilizados nos algoritmos.

Atributo	**Considerações**	
matricula	Apenas referencial para identificar o indivíduo e seus dados	
nome	idem acima	
ano_conclusao_ensino_anterior	Usado para os cálculos mencionados	
ano_ingresso	Usado para os cálculos mencionados	
diferenca_ingresso_ensino_anterior	Cálculo efetuado entre ano_ingresso e ano_conclusao_ensino_anterior	X
bairro	não utilizado - sem relevância para o estudo	
cidade	não utilizado - sem relevância para o estudo	
grade	O curso teve uma grade até 2012 e uma nova a partir de 2013, sinaliza em qual das duas se situa o indivíduo (2009 ou 2013) - não relevante para o estudo	
idade_ingresso	Cálculo efetuado entre ano_ingresso e data_nascimento	X
data_conclusao_curso	Usado para os cálculos mencionados	
tempo_conclusao	Cálculo efetuado entre o ano_conclusao e ano_ingresso	X
data_nascimento	Usado para os cálculos mencionados	
idade_conclusao	Cálculo efetuado entre o ano_conclusao e data_nascimento	
deficiencia	não utilizado - sem relevância para o estudo	
estado	não utilizado - sem relevância para o estudo	
estado_civil	não utilizado - maioria não informado	
forma_ingresso	Categoria conforme legislação (SISU, Cotas, Vestibular, etc.)	X
frequencia_periodo	não utilizado - relevante apenas na data de corte da extração dos dados	
instituicao_ensino_anterior	Nome da escola anterior	
tipo_escola	Classificação feita para este trabalho (EE, Técnica, Privada, Pública)	X
municipio_residencia	não utilizado - sem relevância para o estudo	
naturalidade	não utilizado - sem relevância para o estudo	
necessidade_especial	não utilizado - sem relevância para o estudo	
nivel_ensino_anterior	não utilizado - maioria ensino médio - sem relevância	
percentual_progresso	Percentual de presença no curso - relevante para observar "evasão"	
percentual_progresso_faixa	Cálculo sobre o percentual_progresso, dividido em faixas com intervalos de 25%	X
sexo	não utilizado - sem relevância para o estudo	
situacao_curso	não utilizado - sem relevância para o estudo	
tipo_escola_origem	não utilizado - sem relevância para o estudo	
turno	não utilizado - sem relevância para o estudo	
rendimento_boletim	não utilizado - sem relevância para o estudo	

Fonte: Autoria Própria

Os dados foram tratados por um programa em *python* com biblioteca *pandas*[3] e *sklearn*[4]. Os dados faltantes (sem preenchimento) foram complementados usando a regra "*most_frequent*" da classe *SimpleImputer*[5] que consiste em preencher o dado com o conteúdo mais frequente no dataset.

Os atributos categóricos forma_ingresso, tipo_escola e percentual_progresso_faixa foram codificados com *LabelEncoder*[6] que consiste em atribuir uma numeração sequencial aos dados após classificá-los alfabeticamente.

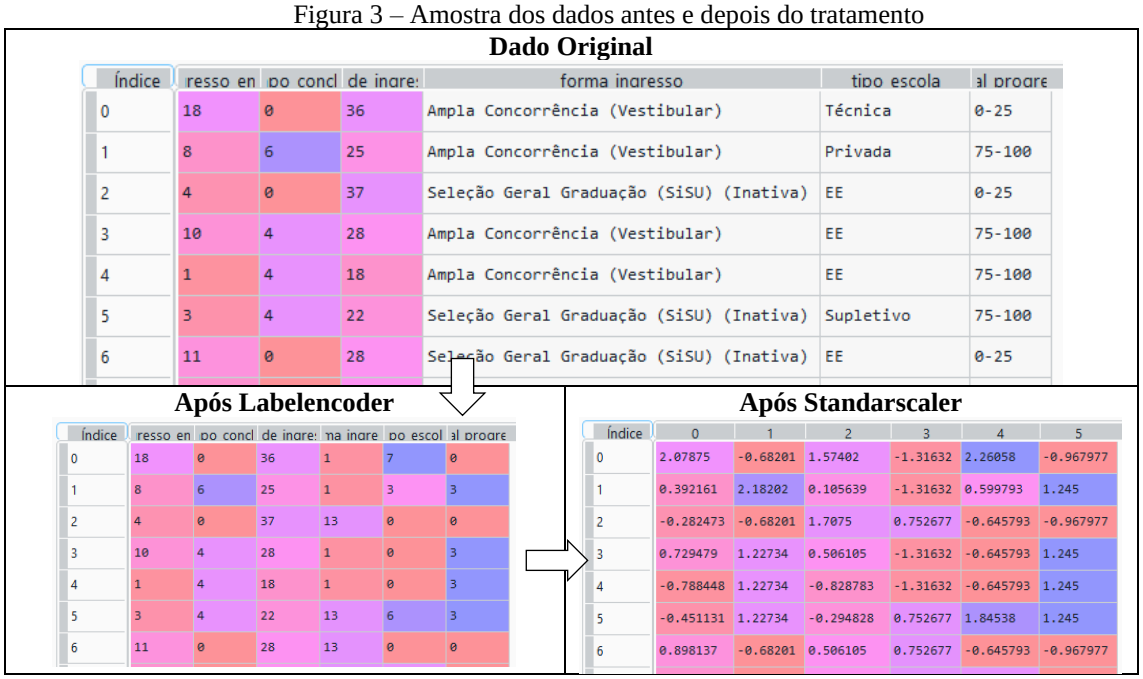
Depois deste processo de classificação todos os dados foram normalizados usando StandarScaler[7] que equaliza os dados conforme a fórmula:

$$valor\ normalizado = \frac{valor - média}{desvio\ padrão} \quad (1)$$

A Figura 3 apresenta uma amostra de como os dados foram transformados mostrando como estavam originalmente e como ficaram após aplicação do tratamento de dados necessários para a execução dos algoritmos.

[3] https://pandas.pydata.org/
[4] https://scikit-learn.org/stable/index.html
[5] https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Figura 3 – Amostra dos dados antes e depois do tratamento



Fonte: Autoria Própria

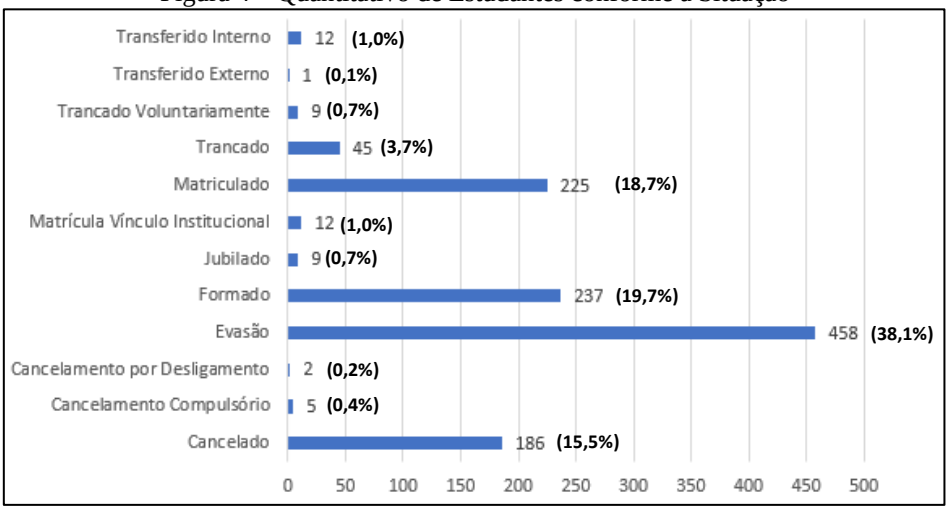
Os resultados da análise dos dados serão discutidos no próximo tópico.

4 ANÁLISE DOS DADOS

4.1 Cenário de Dados

A extração de dados gerou 1201 registros distribuídos conforme a situação mostrada no gráfico da Figura 4.

Figura 4 – Quantitativo de Estudantes conforme a Situação

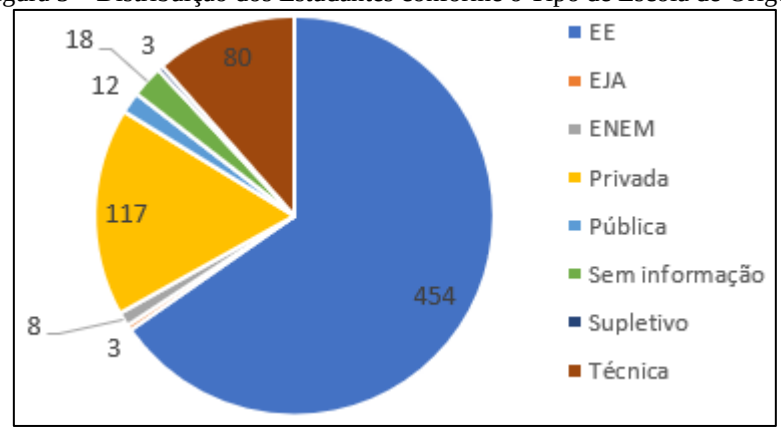


Fonte: Autoria Própria

Embora o percentual de “Cancelados” seja significativo, os motivos para enquadrar os indivíduos nessa situação são diversificados, alguns poderiam ser computados como evasão outros por abandono do curso em seu início, mas não existe registro efetivo, portanto esse subconjunto de dados foi desconsiderado nesta pesquisa.

Dentro do universo de 695 estudantes (237 Formados e 458 evadidos), substrato de dados escolhido para esta análise, a Figura 5 mostra que 65,3% (454) dos estudantes são oriundos de escolas estaduais, seguido pelos 16,8% (117), egressos de escolas privadas.

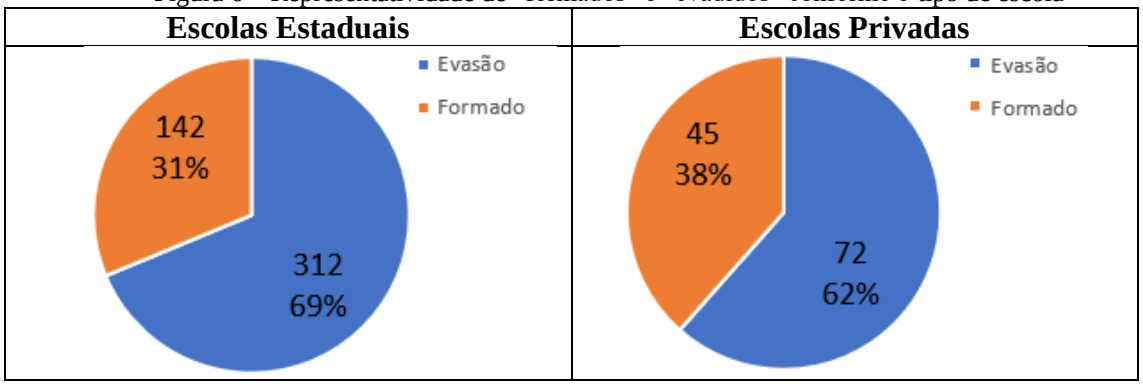
Figura 5 – Distribuição dos Estudantes conforme o Tipo de Escola de Origem



Fonte: Autoria Própria

Neste outro cenário (Figura 6) é apresentado a representatividade dos "Formados" e "Evadidos" conforme o tipo de escola, destacando as de maior percentual.

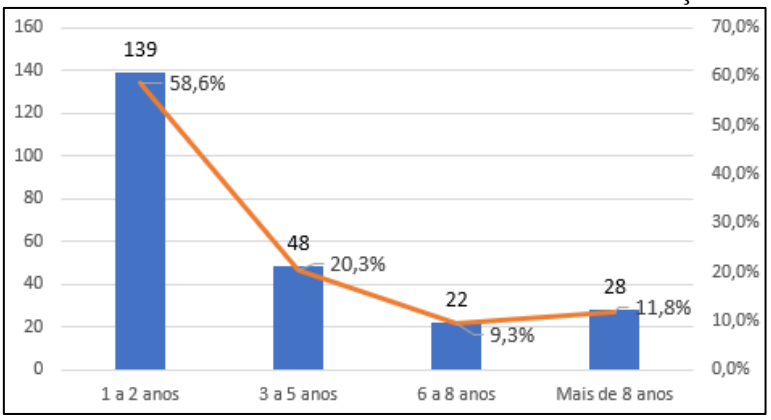
Figura 6 – Representatividade de "formados" e "evadidos" conforme o tipo de escola



Fonte: Autoria Própria

Outro cenário significativo observado entre os formados, foi a concentração de indivíduos (58,6% dos 237 que concluíram o curso) cujo interstício entre a última formação e o ingresso no curso de graduação em questão ficou na faixa de 1 a 2 anos.

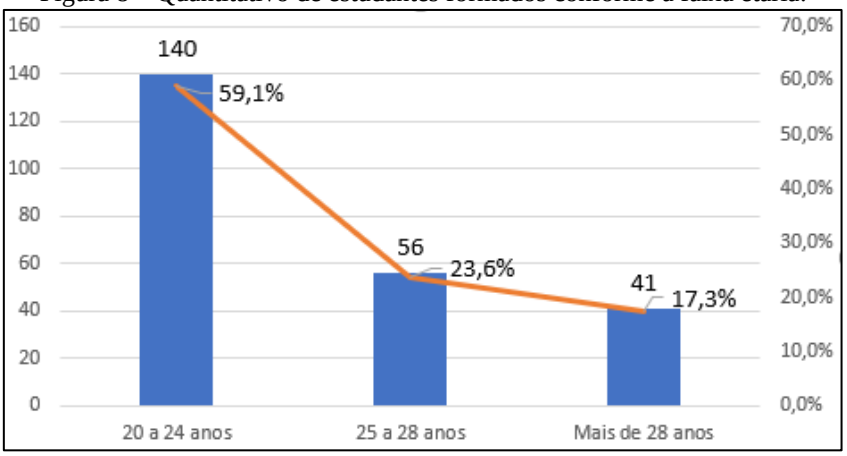
Figura 7 – Quantitativo de estudantes conforme o interstício entre a última formação e o ingresso no curso.



Fonte: Autoria Própria

Finalizando esta abordagem estatística simples, outro destaque é a faixa etária entre os formados, conforme demonstrado no gráfico da Figura 8. Nota-se que 59,1% (140 indivíduos) estão na faixa etária de 20 a 25 anos. Vale constar que apenas 3 dos 140 estão com 20 anos, os demais se distribuem pelas demais idades.

Figura 8 – Quantitativo de estudantes formados conforme a faixa etária.



Fonte: Autoria Própria

4.2 Algoritmos de Agrupamento

K-means

O algoritmo *K-means*[8] faz uso da similaridade entre os dados a partir de cálculos de distâncias, neste caso foi usada a distância euclidiana[9]. O algoritmo se inicializa posicionando pontos aleatórios na quantidade de agrupamentos desejado, estes pontos são denominados centroides. Os centroides representam os clusters onde acontecerá o agrupamento dos dados, representam o "k" (quantidade de centroides ou *clusters*) da denominação do método. Em seguida é calculada a distância euclidiana entre os atributos do *dataset* e os centroides. Assim os dados são atribuídos (classificados) como pertencentes aquele cluster (grupo). A cada iteração uma média, por isso o termo "*means*", das distâncias dos dados até o centroide é calculada e o centroide é reposicionado conforme o resultado, e novamente os dados são reposicionados conforme a proximidade com o centroide.

O número de iterações fica a critério do analista de dados. Um número maior de iterações agrupa os dados mais efetivamente, porém conforme o universo de informações os cálculos de proximidade podem se estabilizar e nenhum ajuste mais é feito no agrupamento.

Para este trabalho foi usada biblioteca *sklearn* do *python* para aplicar o algoritmo, assim como tratar os dados como mencionado anteriormente. Os parâmetros mais importantes a mencionar foi o uso de 5 clusters e 5000 iterações (Figura 9).

Figura 9 - Trecho do código fonte python onde o "k-means" é acionado.

```
97    from sklearn.cluster import KMeans
98    kmeans = KMeans(n_clusters=5,
99                    init='k-means++',
100                   algorithm='auto',
101                   max_iter=5000,
102                   n_init=100,
103                   random_state=0)
```

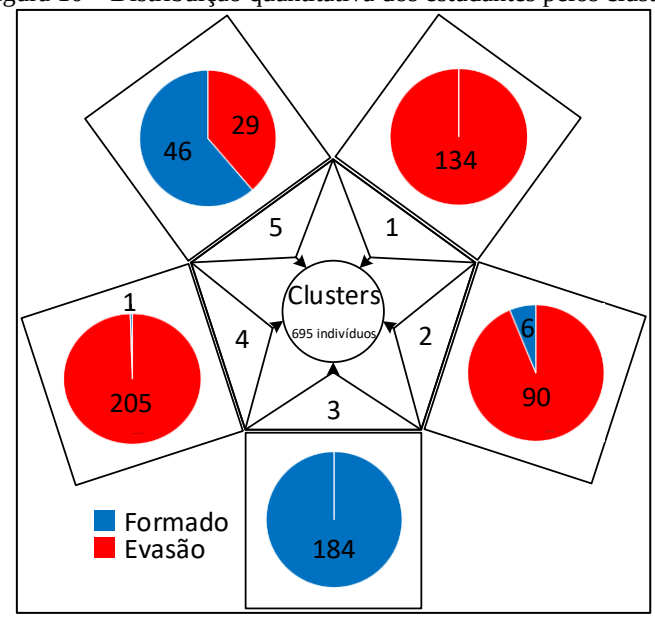
Fonte: Autoria Própria

O resultado está representado no gráfico da Figura 10, construído manualmente para este fim, os gráficos produzidos pelo programa não se mostraram elucidativos.

[8] referncia k-means

[9] Não é objeto deste trabalho se aprofundar nas técnicas matemáticas e de programação. Para aprofundamento no assunto pode ser acessado o link xxxxxxx.

Figura 10 – Distribuição quantitativa dos estudantes pelos clusters



Fonte: Autoria Própria

Depois da execução do algoritmo coube uma análise mais detalhada dos agrupamentos observando a similaridade encontrada entre os registros, naturalmente os dados confirmam o resultado do algoritmo ao proceder o agrupamento dos estudantes.

Com auxílio de uma planilha eletrônica e arquivos de apoio gerados nos programas em *python*, foi possível identificar cada indivíduo no grupo de modo a explicitar os dados para análise mais apurada.

Como a quantidade de registros é grande, optou-se por escolher uma amostragem de alguns *clusters* para discorrer sobre os resultados.

A Figura 11 apresenta um substrato de 10% dos registros do cluster número 3 onde agrupou-se 77,6% dos formados.

Figura 11 – Amostragem de 10% (18) dos registros do cluster 3

	diferenca_ingresso_ensino_anterior	idade_ingresso	tempo_conclusa	forma_ingresso	tipo_escola
1	1	19	3	Ampla Concorrência (Vestibular)	EE
2	2	19	2	SiSU L1 (SGC L1) - Candidatos com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo que tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	EE
3	1	19	4	Ampla Concorrência (Vestibular)	EE
4	3	20	4	Matrícula Direta (Inativa)	EE
5	1	19	4	Seleção Geral Graduação (SiSU) (Inativa)	EE
6	6	23	4	Seleção Geral Graduação (SiSU) (Inativa)	EE
7	3	21	3	Seleção Geral Graduação (SiSU) (Inativa)	EE
8	4	23	2	Matrícula Direta (Inativa)	EE
9	4	21	5	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
10	3	20	6	Ampla Concorrência (Vestibular)	Privada
11	1	19	4	Ampla Concorrência (Vestibular)	EE
12	4	21	6	Ampla Concorrência (Vestibular)	Privada
13	1	18	3	Ampla Concorrência (Vestibular)	EE
14	5	24	4	Seleção Geral Graduação (SiSU) (Inativa)	EE
15	3	20	6	Ampla Concorrência (Vestibular)	Privada
16	11	29	5	Ampla Concorrência (Vestibular)	EE
17	1	18	3	Ampla Concorrência (Vestibular)	Privada
18	1	19	4	Ampla Concorrência (Vestibular)	EE

Fonte: Autoria Própria

Pode-se observar que a amostragem traz conteúdos diversificados entre os dados não direcionando ou indicando claramente a similaridade entre os registros, exceto que todos apresentam a situação “Formado” e uma certa coerência em relação a idade de ingresso e o tempo de conclusão. O atribuo sobre percentual de progresso (frequência) foi omitido porque todos têm 100% de progresso por conta da dita situação.

A Figura 12 apresenta um substrato de 10% do cluster 5, onde se agrupam, segundo o algoritmo, 19,4% dos formados (46) e 6,3% dos evadidos.

Figura 12 – Amostragem de 10% (8) dos registros do cluster 5

	diferenca_ingresso_ensino_anterior	idade_ingresso	tempo_conclusa	forma_ingresso	tipo_escola	percentual_progresso_faixa
1	3	21	3	Seleção Geral Graduação (SiSU) (Inativa)	Técnica	75-100
2	1	18	3	Seleção Geral Graduação (SiSU) (Inativa)	Técnica	75-100
3	1	19	5	Seleção Geral Graduação (SiSU) (Inativa)	Técnica	75-100
4	15	32	5	Seleção Geral Graduação (SiSU) (Inativa)	Técnica	75-100
5	0	22	0	Transferência Facultativa	Técnica	25-50
6	1	18	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica	75-100
7	2	20	0	Processo Seletivo Simplificado (Inativa)	Técnica	25-50
8	4	22	0	Seleção Geral Graduação (SiSU) (Inativa)	Técnica	25-50

Fonte: Autoria Própria

A amostra extraída aleatoriamente mostra um pouco mais de similaridade do que a amostra anterior (Figura 11). Dentre os 8 registros, aqueles com percentual de progresso na faixa “25-50” são aqueles com situação “Evasão”.

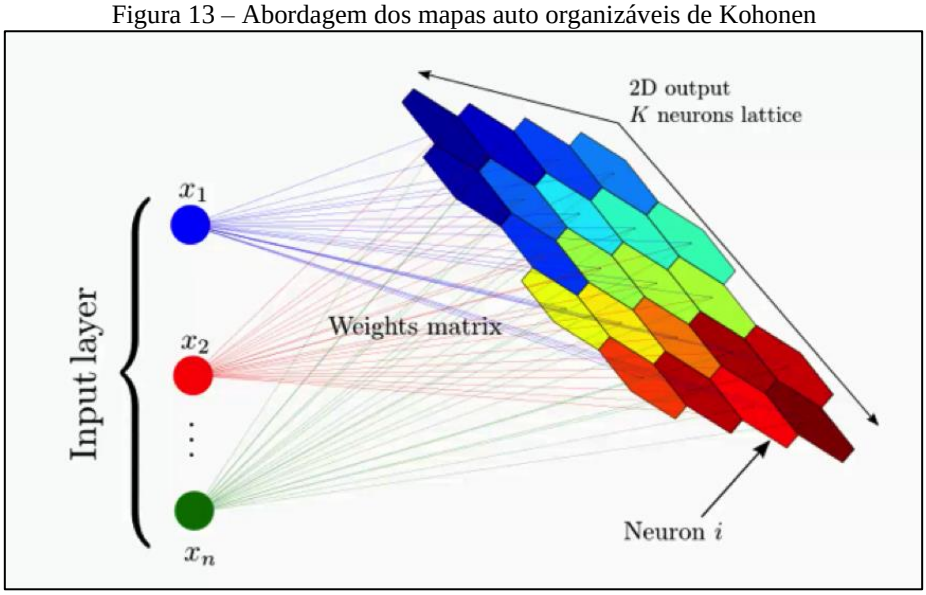
Mapas Auto organizáveis de Kohonen (KOHONEN, 2001)

Esse algoritmo é baseado em redes neurais artificiais sem as camadas ocultas que caracterizam as redes neurais multicamadas e as tornam eficientes.

Redes neurais artificias, como outros algoritmos de aprendizado de máquina supervisionados, precisam de uma massa de dados considerável e de um resultado esperado/ conhecido. Os processos de cálculos dessas redes buscam aproximar os resultados encontrados com os resultados esperados disponíveis na massa de dados. Esse processo todo é chamado de treinamento da rede.

No caso dos mapas auto organizáveis (Figura 13) não se faz necessário existir um resultado conhecido/esperado. O objetivo é agrupar os dados em células, tantas quantas se desejar, usando cálculos de distância entre os registros com alguma semelhança ao *k-means*, com outra abordagem.

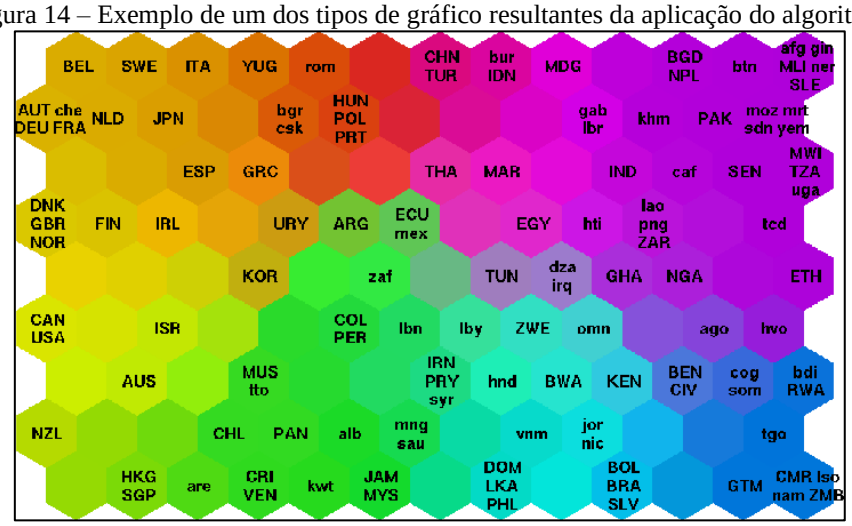
Figura 13 – Abordagem dos mapas auto organizáveis de Kohonen



Fonte: Disponível em https://visimaps.blogspot.com/2019/09/self-organizing-map-architecture.html, acessado em 09/06/2022

O resultado do processamento será um gráfico (Figura 14), onde os dados ficam distribuídos conforme proximidade que tem. Vale salientar que não é foco deste trabalho o aprofundamento na técnica descrita, mas sim a análise dos resultados[10].

Figura 14 – Exemplo de um dos tipos de gráfico resultantes da aplicação do algoritmo.



Fonte: Disponível em https://electricarchaeology.ca/2015/05/05/a-quick-note-on-visualizing-topic-models-as-self-organizing-map/, acesso em 09/06/2022

Neste trabalho foi construído um programa, adaptado de diversos códigos públicos divulgados em artigos e tutoriais. O código construído faz uso de lógica direta sem a inclusão de bibliotecas especificas para *self organizing maps*. Essa abordagem permitiu maior flexibilidade na customização dos parâmetros e captura de dados relevantes para apoio a interpretação do gráfico resultante (Figura 16).

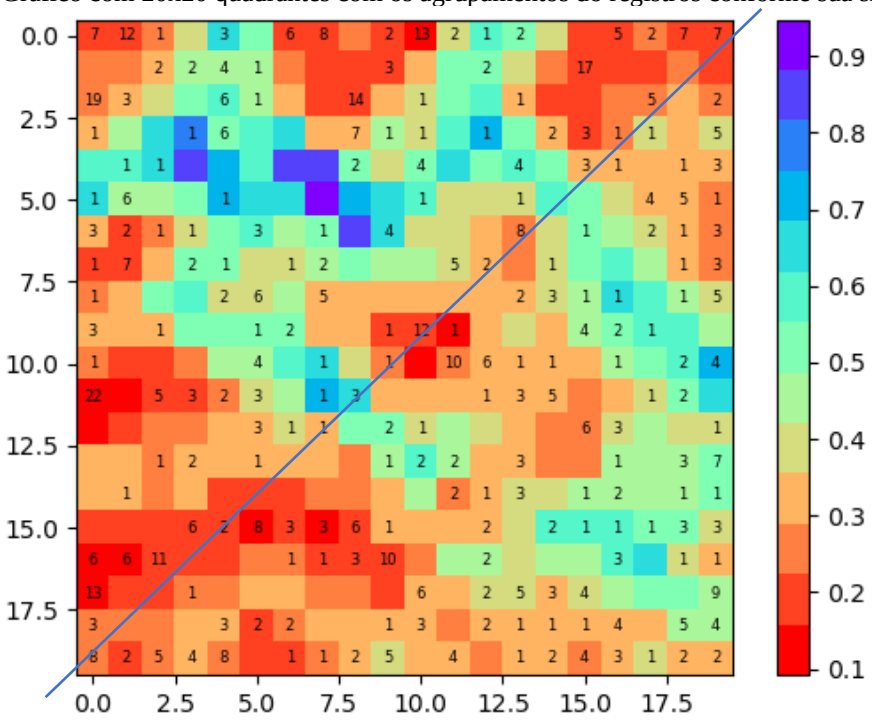
Figura 15 – Trecho do código fonte em python

```
160    # 3. Construção da Matriz de Distâncias
161    u_matrix = np.zeros(shape=(Rows,Cols), dtype=np.float64)
162    for i in range(Rows):
163      for j in range(Cols):
164        v = pesos[i][j]  # a vector
165        sum_dists = 0.0; ct = 0
166
167        if i-1 >= 0:    # above
168          sum_dists += distancia_euclidiana(v, pesos[i-1][j]); ct += 1
169        if i+1 <= Rows-1:   # below
170          sum_dists += distancia_euclidiana(v, pesos[i+1][j]); ct += 1
171        if j-1 >= 0:   # left
172          sum_dists += distancia_euclidiana(v, pesos[i][j-1]); ct += 1
173        if j+1 <= Cols-1:   # right
174          sum_dists += distancia_euclidiana(v, pesos[i][j+1]); ct += 1
175
176        u_matrix[i][j] = sum_dists / ct
```

Fonte: Autoria Própria

[10] Mais informações sobre mapas auto organizáveis pode ser adquiridas em https://towardsdatascience.com/kohonen-self-organizing-maps-a29040d688da, acesso em 09/06/2022

Figura 16 – Grafíco com 20x20 quadrantes com os agrupamentos do registros conforme sua similaridade



Fonte: Autoria Própria

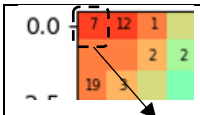
A barra de intensidade a direita indica que quanto maior o valor, mais distante os dados estão entre si. Assim, quanto mais se aproxima da cor avermelhada, mais próximo o registro de dados está.

Os números dentro dos quadrantes representam a quantidade de estudantes que ali foram agrupados ou classificados. Usando a referência de linhas e colunas (sempre começando com zero), observa-se que na linha 16, coluna 9, o quadrante apresenta 10 registros agrupados, eles têm bastante semelhança com aqueles que estão no quadrante **L**16**C**2 (**L**inha, **C**oluna) segundo a interpretação do mapa auto organizável conforme a notação de cores.

Dados agrupados nos quadrantes mais acima, L2C8 por exemplo, apresentam a mesma tonalidade, porém estão distantes daqueles mencionados no parágrafo anterior. Isso acontece por uma limitação na construção do gráfico, em verdade existe uma linha limítrofe, colocada manualmente, que separa diagonalmente os "formados" (lado superior esquerdo) e os "Evadidos" lado inferior esquerdo. Não se trata de uma separação exata e linear como desenhado, a linha é simbólica para apoiar o esclarecimento.

A Figura 17 apresenta os dados dos registros dos 7 estudantes agrupados no quadrante L0C0. Observa-se que todos possuem o mesmo tempo de conclusão do curso, o mesmo tipo de escola de origem, mesma forma de ingresso e pouca diferença de idade.

Figura 17 – Dados dos estudantes alocados no quadrante L0C0

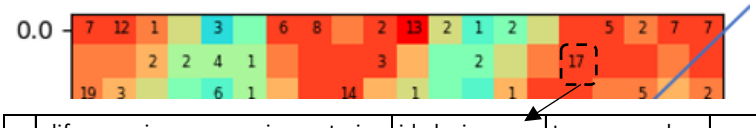


	diferenca_ingresso_ensino_anterior	idade_ingresso	tempo_conclusa	forma_ingresso	tipo_escola
1	2	20	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
2	1	18	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
3	2	20	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
4	1	18	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
5	2	20	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
6	1	19	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica
7	1	19	4	Seleção Geral Graduação (SiSU) (Inativa)	Técnica

Fonte: Autoria Própria

A Figura 18 mostra os dados dos 17 estudantes no quadrante L1C15, eles apresentam bastante similaridade, mas tem forma de ingresso e tipo de escola diferentes, por isso estão distantes daqueles mostrados na Figura 17.

Figura 18 – Estudantes agrupados no quadrante L1C15.

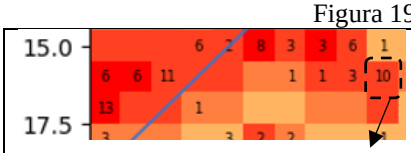


	diferenca_ingresso_ensino_anterior	idade_ingresso	tempo_conclusa	forma_ingresso	tipo_escola
1	1	18	4	Ampla Concorrência (Vestibular)	EE
2	1	18	4	Ampla Concorrência (Vestibular)	EE
3	2	19	4	Ampla Concorrência (Vestibular)	EE
4	1	18	4	Ampla Concorrência (Vestibular)	EE
5	1	18	4	Ampla Concorrência (Vestibular)	EE
6	1	18	4	Ampla Concorrência (Vestibular)	EE
7	1	19	4	Ampla Concorrência (Vestibular)	EE
8	1	19	4	Ampla Concorrência (Vestibular)	EE
9	1	18	4	Ampla Concorrência (Vestibular)	EE
10	2	19	4	Ampla Concorrência (Vestibular)	EE
11	1	18	4	Ampla Concorrência (Vestibular)	EE
12	2	19	4	Ampla Concorrência (Vestibular)	EE
13	3	20	4	Ampla Concorrência (Vestibular)	EE
14	1	19	4	Ampla Concorrência (Vestibular)	EE
15	1	19	4	Ampla Concorrência (Vestibular)	EE
16	3	20	4	Ampla Concorrência (Vestibular)	EE
17	2	19	4	Ampla Concorrência (Vestibular)	EE

Fonte: Autoria Própria

A Figura 19 apresenta os estudantes na parte inferior esquerda. Trata-se de um conjunto de registros com a situação “evasão” que foram adequadamente agrupados nesse quadrante. O algoritmo se comportou como esperado como pode ser observado na similaridade entre os dados.

Figura 19 - Estudantes agrupados no quadrante L16C9.

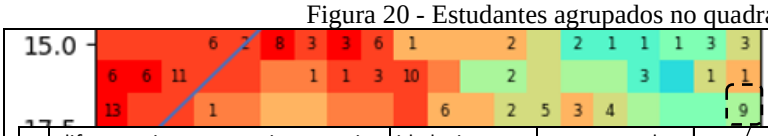


	diferenca_ingresso_ensino_anterior	idade_ingresso	tempo_conclusa	forma_ingresso	tipo_escola
1	11	28	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
2	8	29	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
3	10	28	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
4	10	28	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
5	8	29	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
6	12	29	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
7	11	29	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
8	12	29	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
9	11	28	0	Seleção Geral Graduação (SiSU) (Inativa)	EE
10	10	28	0	Seleção Geral Graduação (SiSU) (Inativa)	EE

Fonte: Autoria Própria

Cabe examinar agrupamentos atípicos com o do quadrante L17C19 (Figura 20) que apresenta 9 estudantes, sendo 2 com situação “formado” e 7 com situação “evasão”. O agrupamento ocorreu por conta da faixa etária e interstício entre a última formação e o ingresso no curso.

Figura 20 - Estudantes agrupados no quadrante L17C19.



	diferenca_ingresso_ensino_anterior	idade_ingresso	tempo_conclusa	forma_ingresso	tipo_escola
1	27	46	6	Ampla Concorrência (Vestibular)	EE
2	26	46	5	Ampla Concorrência (Vestibular)	EE
3	25	47	0	Ampla Concorrência (Vestibular)	Privada
4	22	40	0	Ampla Concorrência (Vestibular)	EE
5	24	45	0	Exame de Seleção (Inativa)	Privada
6	18	49	0	Processo Seletivo Simplificado (Inativa)	EE
7	26	43	0	Ampla Concorrência (Vestibular)	EE
8	24	45	0	Exame de Seleção (Inativa)	Privada
9	21	49	0	Ampla Concorrência (Vestibular)	EE

Fonte: Autoria Própria

Cabe salientar que o dado “situacao_curso”, cujo domínio é “Formado” e “Evasão” não foi utilizado no processamento do algoritmo.

5 CONSIDERAÇÕES FINAIS

A aplicação dos dois algoritmos e uma análise apurada e individualizada de cada agrupamento mostrou a similaridade entre o conteúdo dos seis atributos escolhidos e a correlação com a situação correspondente.

Cabe salientar que o dado “situacao_curso”, cujo domínio é “Formado” e “Evasão” não foi utilizado no processamento do algoritmo, assim a lógica de agrupamento não foi influenciada por esses dados.

Essa argumentação é importante porque direciona as considerações finais para o objetivo principal e título do trabalho.

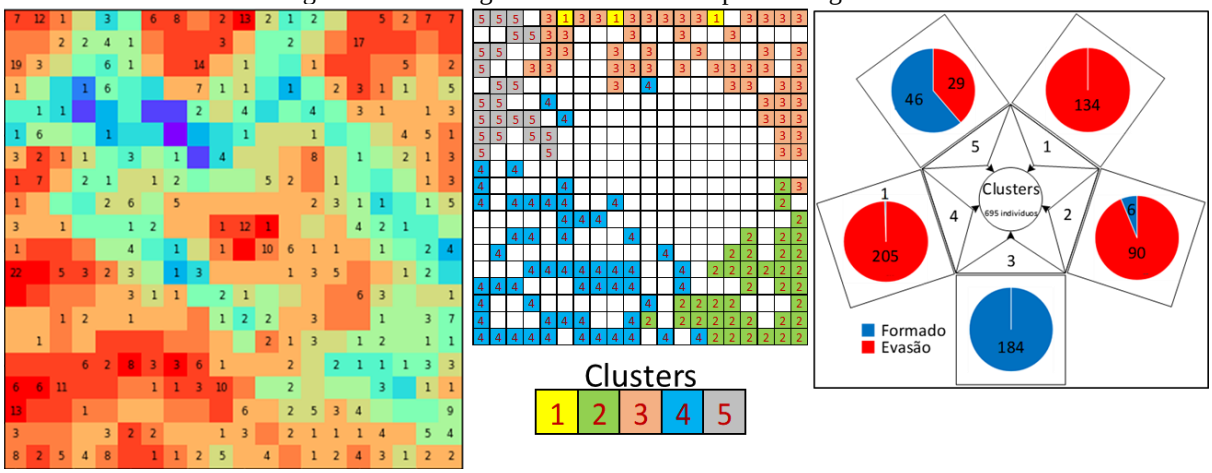
A comparação entre as duas técnicas mostra um melhor eficiência e coerência nos resultados obtidos com o algoritmo de mapas auto organizáveis, embora deve ser considerado que as diferenças de amplitude dos grupos criados em uma e outra técnica, ou seja, em *k-means* forame 5 grupos (clusters) de registros e em mapas auto organizáveis foram 400 grupos (gráfico com 20x20 quadrantes).

A abordagem de um e outro é diferente, embora façam uso do cálculo de distâncias entre os dados dos registros.

O algoritmo *k-means* tem suas limitações por conta do número de clusters e a inicialização dos centroides e ainda pode ter desempenho prejudicado quando existem atributos categóricos codificados (AHMED e RAIHAN SERAJ, 2020) de alguma forma como é o caso do *dataset* utilizado neste trabalho.

A Figura 21, apresenta uma comparação, por amostragem, entre os registros agrupados nos clusters e sua posição nos quadrantes do mapa auto organizável.

Figura 21 – Visão geral dos clusters e mapa auto organizável.



Fonte: Autoria Própria

Alguns fatores significativos foram observados, os quais podem significar a correlação objetivada na pesquisa. A Tabela 1 apresenta algumas características do universo de dados que contribuíram para os agrupamentos resultantes do processamento dos algoritmos.

Tabela 1 – Resumo percentual de algumas características dos dados considerando os 237 estudantes formados.

Característica	Qtde	Pct
1 ano desde a última formação	110	46,4%
18 a 23 anos de idade	189	79,7%
3 a 4 anos para conclusão do curso	146	61,6%
Seleção Geral (SISU e Vestibular)	214	90,3%

(A)

Tipo de Escola	Qtde	Pct
Escola Estadual	142	59,9%
Privada	45	19,0%
Técnica	47	19,8%
Outros	3	1,3%

(B)

Fonte: Autoria Própria

É perceptível que a faixa etária e o tempo de interstício entre a entrada no curso e a última formação estão entre as principais características dos estudantes que concluíram o curso.

Um estudo mais abrangente com estudantes do mesmo curso em todos os campi resultaria em uma análise mais rica, ficando com sugestão para trabalhos futuros assim como o treinamento de uma rede neural para verificar se esses atributos representam dados suficientes para alcançar um baixo nível de erro, permitindo o uso em dados estatísticos preditivos a partir dos estudantes ingressantes.

Este autor acredita ainda que outros dados relacionados ao perfil dos estudantes podem ter inferência no sucesso ou insucesso na conclusão do curso. Alguns deles existem no dataset, porém sem preenchimento adequado, como forma de deslocamento ao campus e estado civil, portanto não foram usados no estudo.

O estudo pode avançar analisando a performance do estudante ao longo dos semestres em relação a notas nas disciplinas e frequência, fatores que poderiam indicar em que momento ocorre a evasão.

Enfim, entende-se que o presente estudo contribui para uma visão mais ampla do histórico do curso e seu universo de estudantes, podendo ser fomento para fóruns de direcionamento de iniciativas que possam contribuir para aumentar as estatísticas de estudantes que concluem o curso.

6 REFERÊNCIAS

AHMED, M.; RAIHAN SERAJ, S. M. S. The k-means algoritm; A comprehensive survey and preformance evalution. **Electronics**, Basel, 9, n. 8, 12 Agosto 2020. 1295. Disponivel em: <https://www.mdpi.com/2079-9292/9/8/1295/htm>. Acesso em: 9 Junho 2022.

BRASIL. L13709. **Lei 13709**, 14 Agosto 2018. Disponivel em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm>. Acesso em: 29 Novembro 2020.

CERVO, A. L.; BERVIAN, P. A.; SILVA, R. D. **Metodologia Científica**. 6ª. ed. São Paulo: Pearson, 2010.

CUESTA, H.; KUMAR, D. S. **Practical Data Analysis:** A practical guide to obtaining, transforming, exploring, and. 2ª. ed. Birmingham: Packt, 2016.

EDUCAÇÃO, R. D. S. D. D. Por que a OCDE "reprova" o modelo do Enade. **Desafios da Educação**, 20 Fevereiro 2019. Disponivel em: <https://desafiosdaeducacao.grupoa.com.br/ocde-reprova-enade/>. Acesso em: 29 Novembro 2020.

FARIA, S. M. S. M. L. **Educational Data Mining e Learning Analytics na melhoria do ensino on-line**. Universidade Aberta de Portugal. [S.l.], p. 138. 2014.

HOED, R. M. **Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação**. Universidade de Brasília. Brasília, p. 188. 2016.

INEP. Inep. **Instituto Nacional de Estudos e Pesquisas Educionais Anísio Teixeira**, 2020. Disponivel em: <https://www.gov.br/inep/pt-br>. Acesso em: 29 Novembro 2020.

KOHONEN, T. **Sef-Organizing Maps**. 3ª. ed. Berlin: Springler-Verlag, 2001.

KUZUYABU, M. Relatório aponta falhas no Enade, no CPC e nos indicadores de referência. **Revista Ensino Superior**, 27 Fevereiro 2019. Disponivel em: <https://revistaensinosuperior.com.br/enade-cpc/>. Acesso em: 29 Novembro 2020.

LENGEL, J. **Education 3.0:** seven steps to better schools. New York: Columbia University, 2012.

LJUBLJANA, U. O. Orange. **Orange**, 2020. Disponivel em: <https://orange.biolab.si/>. Acesso em: 29 Novembro 2020.

MACHADO, R. C.; CAVALCANTI, E. L. D. **Desempenho acadêmico e sucesso/insucesso escolar dos estudantes do curso de química:** relações possíveis. XV Encontro Nacional de Ensino de Química (XV ENEQ). Brasília: [s.n.]. 2010. p. 10.

MEDEIROS, C. A. D. **Estatística Aplicada à Educação**. Brasília: Universidade de Brasília, 2009.

PRODANOV, C. C.; FREITAS, E. C. D. **Metodologia do Trabalho Científico:** Métodos e Técnicas da Pesquisa e do trabalho Acadêmico. 2ª. ed. Novo Hamburgo: Feevale, 2013.

SANTANA, R. C. G. Ciclo de Vida dos Dados. **Informação & Informação**, p. 116-142, 2016.

SHARDA, R.; DELEN, D.; TURBAN, E. **Business Intelligence, Analytics and Data Science**. 4ª. ed. Harlow: Pearson, 2018.