

# Análise de Desempenho do Poscomp usando técnicas de agrupamento

Jean Carlos de Carvalho Costa<sup>1</sup>, Adonney Allan de Oliveira Veras<sup>2</sup>

<sup>1</sup>Instituto de Ciências Exatas e Naturais – Universidade Federal do Pará (UFPA)  
Caixa Postal 66.075.110 – Belém – PA – Brasil

<sup>2</sup>Programa de Pós-Graduação de Computação Aplicada (PPCA) – Universidade Federal  
do Pará (UFPA) – Tucuruí – PA – Brasil

jeancc.costa@gmail.com, allanveras@ufpa.br

**Abstract.** *To evaluate and analyze the performance of computer science students, we will use data science and mining techniques. These techniques will contribute to solving the research problem, for which the KDD process will be used to discover useful and non-trivial information presented in the POSCOMP database. This exam aims to evaluate the knowledge in the field of computer science of candidates seeking to enter postgraduate programs in computer science in Brazil. Based on this information, we will conduct the research.*

**Resumo.** *Para avaliar e analisar o desempenho dos discentes dos cursos de computação, utilizaremos técnicas de ciência e mineração de dados. Essas técnicas contribuirão na solução do problema de pesquisa, para o qual será utilizado o processo KDD, a fim de descobrir informações úteis e não triviais apresentadas no banco de dados do POSCOMP. Esse exame tem como objetivo avaliar os conhecimentos na área de computação dos candidatos que buscam ingressar em programas de pós-graduação em computação no Brasil. Com base nessas informações, realizaremos a pesquisa.*

## 1. Introdução

Avaliar o desempenho dos alunos das instituições de ensino tornou-se primordial para o desenvolvimento da educação de qualidade. É fundamental buscar novas metodologias de ensino e aprendizagem para os alunos visando o crescimento acadêmico [Hui et al. 2020]. Nesse contexto, analisar o desempenho dos alunos ao longo do curso é uma maneira de buscar estratégias para evitar problemas futuros tanto para os alunos quanto para as instituições. Assim, a utilização de tecnologias para solucionar esses problemas é uma prática comum.

Com isso, a aplicação da ciência de dados na área educacional, utilizando técnicas de mineração de dados para extrair informações úteis e não triviais dos candidatos que prestam provas durante a vida acadêmica, tornou-se uma forma eficaz de levantar informações que possam beneficiar a instituição de ensino, prevendo até mesmo a evasão de alunos e a probabilidade de reprovação [Lauría et al. 2013]. A mineração de dados aborda conceitos e ideias para classificação e agrupamento, permitindo observar o progresso dos alunos com modelos de previsão que utilizam técnicas de mineração de dados educacionais para explorar e melhorar o desempenho dos alunos [de Castro and Ferrar 2016].

O objetivo desta pesquisa é analisar o desempenho dos candidatos e discentes que realizaram a prova POSCOMP por meio de técnicas de ciência e mineração de dados, combinadas com técnicas de Machine Learning e algoritmos de agrupamento, para avaliar as possibilidades de ingresso na pós-graduação dos cursos de computação. Com este propósito, serão avaliados o desempenho dos candidatos em relação aos assuntos abordados na prova e identificados os destaques nas áreas, bem como as lacunas que precisam ser melhoradas.

Observar o desempenho dos discentes dos cursos de computação ao longo das atividades acadêmicas é difícil para a gestão, coordenação e professores na busca por melhorias no processo de ensino. Um dos problemas abordados pelas instituições de ensino é a previsão do desempenho dos discentes, de modo que aqueles em risco possam ser identificados o mais cedo possível e medidas de intervenção possam ser aplicadas, reduzindo assim os índices de reaprovação em exames de avaliação dos cursos ou de ingresso em programas de pós-graduação na área de computação [Senthil and Lin 2017]. Dessa forma, a pesquisa poderá contribuir para que gestores, coordenadores e professores de universidades do estado do Pará que possuem cursos de Computação melhorem o processo de ensino e aprendizagem dos alunos, verificando os resultados obtidos a partir do desempenho acadêmico dos alunos.

Para os discentes que desejam se dedicar ao mestrado e doutorado nas áreas de computação, é necessário realizar um exame oferecido pela Sociedade Brasileira de Computação (SBC) desde 2002, aplicado em todas as regiões do Brasil e que também possui parceria com a Sociedade Peruana de Computação desde 2006, tendo passado a ser realizado no Peru. O Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) busca testar os conhecimentos na área de computação, e, segundo a SBC, um dos seus objetivos é avaliar os conhecimentos dos candidatos e, com a nota obtida, permitir a submissão aos programas de pós-graduação em computação oferecidos no país [de Computação 2022].

A prova aborda as áreas de Matemática, Fundamentos da Computação e Tecnologia da Computação. A cada ano, a prova é composta por 70 questões de múltipla escolha (a, b, c, d, e), e os candidatos precisam obter uma nota suficiente para participar dos cursos de pós-graduação em computação.

O exame tem relevância nacional, pois parte dos Programas de Pós-Graduação no país utilizam de alguma forma o resultado em seu processo seletivo [de Computação 2022]. Outro fator relevante é que este exame permite que o candidato a um programa de pós-graduação possa realizá-lo com o mínimo de deslocamento, não sendo necessário se locomover para fazer a prova na instituição do programa em questão [Moura et al. 2012]. Esse fato aumenta consideravelmente a acessibilidade dos candidatos aos melhores programas de pós-graduação disponíveis no país.

O trabalho está organizado da seguinte forma: na seção 1, é apresentada uma breve contextualização do tema, a apresentação do problema e a justificativa; na seção 3, são mostradas as ferramentas utilizadas para o desenvolvimento do trabalho, bem como o processo de preparação dos dados; na seção 4, são apresentados os resultados das aplicações dos algoritmos de agrupamento; e na seção 5, são apresentadas as considerações do trabalho.

## 2. Referencial Teórico

Com a realização do Mapeamento Sistemático da Literatura (MSL), percebeu-se que não há trabalhos que abordam análises nos microdados do POSCOMP com aplicações de técnicas de ciências e mineração de dados, acredita-se que motivo seja dos dados não serem públicos. Porém, foram buscados estudos que se relacionam com o tema, estudos esses que abordam sobre ENADE, dados públicos disponíveis em sites do governo e banco de dados das próprias instituições de ensino.

Os autores [Fernando Raguro et al. 2022] em seu trabalho buscaram avaliar o desempenho dos alunos para que levasse a melhora dos cursos, com isso utilizaram técnicas de mineração de dados educacionais, na qual abordam possuir métodos para extrair informações úteis dos desempenhos dos alunos e prever resultados futuros utilizando técnicas de aprendizado de máquina, precisamente técnicas de árvore de decisão. Para a preparação dos dados utilizaram o processo KDD, para extrair conhecimentos e encontrar padrões úteis nos conjuntos de dados educacionais.

Para [Amazona and Hernandez 2019] usaram abordagem de mineração de dados educacional para modelar os desempenhos dos alunos ao aplicar modelos de classificação, tais como: *Naïve Bayes*, *Decision Tree* e *Deep Learning in Neural Network*. E relata que o classificador *Deep Learning* superar os outros ao obter uma precisão geral da previsão de 95% de acurácia. O trabalho foi desenvolvido utilizando dados do curso de bacharelado em tecnologia da informação realizado em um período de seis semestres.

[Silva Guerra et al. 2018] relatam em seu estudo que os cursos da área de Computação, a evasão de discentes são maiores. Então, com este problema, eles buscam aplicar técnicas de mineração de dados, pois ele vem ganhando espaço na área educacional. Eles aplicaram técnicas de árvore de decisão e utilizada três diferentes opções de execução: *Use Training Set*, *Supplied Test Set* e *Cross-validation*. Concluíram que o *Use Training Set* aponta a melhor taxa de acerto na tarefa de classificação, porém no resultado aponta que o teste não é realista para realizar o treinamento. E de fato, o que demonstrou boa taxa de acertos para classificação foi utilizando o algoritmo J4.8 e a *Cross-validation*.

[Ahmed et al. 2020] utiliza as técnicas de mineração de dados educacionais para analisar e prever o desempenho acadêmico do aluno, para propor uma intervenção na melhoria do desempenho. O objetivo da pesquisa dos autores é calcular o desempenho acadêmico dos discentes de graduação usando técnicas de mineração de dados, precisamente, algoritmos de classificação, para o registro de 800 alunos do curso de Ciência da Computação. Para avaliar o desempenho, foi utilizado quatro métodos de seleção de característica: algoritmos genéticos, razão de ganho, relevos e ganho de informação, e para algoritmos de classificação: *K-Nearest Neighbor*, *Naïve Bayes*, *Bagging*, *Random Forest* e *J48 Decision Tree*. Diante disso, os resultados experimentais do trabalhos mostraram que o método de algoritmos genéticos fornece a melhor precisão de 91,37% com o classificador KNN.

[Carrillo and Parraga-Alava 2018] relatam em seu trabalho que as instituições de ensino superior possuem sucesso ao avaliar o desempenho dos alunos utilizando três classificadores: C5.0, *Random Forest* e *CART* em que são aplicados em dados com 1086 instâncias, informações acadêmicas do curso de Ciência da Computação da *Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix Lopez* (ESPM MFL) entre

2009 a 2017, utilizado algoritmos de árvores de decisões como abordagem preliminar. No trabalho conclui-se que o algoritmo CART foi considerado o melhor baseado em desempenho.

[Gunawan et al. 2019] utilizam técnicas de mineração de dados para prever a conclusão do aluno na graduação, caso ele não termine no tempo hábil, possibilita à instituição realizar o planejamento, acompanhamento e orientações para o aluno. Então, o objetivo do trabalho é prever o tempo de graduação dos alunos usando a árvore de decisão com os algoritmos C4.5 e descobrir quais atributos influenciam a previsão.

A Mineração de dados educacionais é uma área que está crescendo cada vez mais e está se tornando um campo de pesquisa interdisciplinar [Islam et al. 2019]. O principal objetivo é entender o processo de aprendizagem dos alunos e identificar a maneira pela qual eles podem aprender para melhorar os resultados educacionais. Desta forma, os pesquisadores [Islam et al. 2019] usaram quatro modelos de classificação: *Support Vector Machine (SVM)*, *Logistic Regression (LR)*, *Decision Tree* e *Random Forest (RF)* para prever o desempenho dos alunos.

[Jain et al. 2018] enfatizam sobre as instituições educacionais, sobre a importância da análise, previsão e geração de resultados precisos dos alunos no decorrer das atividades acadêmicas. Para realizar a pesquisa, utilizaram a mineração de dados para descoberta de conhecimentos. Assim, para classificação usaram o perceptron multicamada e a árvore de decisão, configurando-as para trabalharem juntas e prever as notas dos alunos. Com isso, concluíram ser útil para os professores, alunos, coordenadores e pais para que possam tomar medidas preventivas.

[Nabil et al. 2021] analisam os dados dos alunos em ambientes educacionais para prever o desempenho. O objetivo do trabalho é explorar a eficiência do aprendizado profundo com mineração de dados educacionais. Os dados coletados são de 4 anos e de universidade pública para desenvolver o modelo preditivo para prever o desempenho acadêmico dos alunos. Os métodos usados foram rede neural profunda, árvore de decisão, floresta aleatória, *gradient boosting*, *logistic regression*, *support vector classifier*, and *K-nearest neighbor*. O resultado demonstra que a rede neural profunda apresenta o melhor algoritmo, com precisão de 89%.

[Arun et al. 2021] usam WEKA para realizar análise com vários modelos de aprendizado de máquina para problemas de classificação e regressão. No trabalho aplicam técnicas de mineração de dados educacionais coletado conjunto de dados da *BMS College of Engineering*. Os autores realizaram testes e análises em vários algoritmos para previsão do desempenho dos alunos, analisando por assunto e na previsão de GPA. Os resultados dos testes e análises mostraram que o *Random Forest* apresentou melhores resultados usando a análise de regressão.

Com isso, está pesquisa busca analisar o desempenho dos candidatos que prestaram a prova do POSCOMP dos anos de 2016 a 2019 utilizando técnicas de ciência e mineração de dados, o que possibilita averiguar os desempenhos nôs conteúdo da área da Computação e assim prever o sucesso para o ingresso em Programas de Pós-Graduação em Computação no Brasil.

### **3. Metodologia**

Os dados foram disponibilizados pela Sociedade Brasileira de Computação (SBC) contendo as informações dos candidatos que prestaram o Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP).

Neste estudo, busca-se aplicar técnicas de Machine Learning, precisamente algoritmos de agrupamento, para buscar padrões nos candidatos que realizaram a prova POSCOMP, avaliando-os por área de conhecimento, como Matemática, Fundamentos da Computação e Tecnologia da Computação, além de identificar a origem dos candidatos que demonstram bom desempenho. A clusterização é um método robusto de análise para Machine Learning e projetos de Data Science. Ele avalia os dados e tenta encontrar padrões sem a base da supervisão, como é o caso de outros algoritmos. Desse modo, torna-se poderoso para extrair insights gerais que ajudam as empresas na tomada de decisões.

Os dados disponibilizados pela SBC contêm informações sensíveis, e por isso foram excluídas para manter o anonimato dos candidatos que realizaram o exame nos anos anteriores. As informações disponibilizadas são: homologação, notas do POSCOMP, gabarito e exportação de respostas.

As ferramentas utilizadas para o desenvolvimento do estudo foram as bibliotecas Pandas, NumPy, Jupyter Notebook e Scikit-learn. Com elas foi possível realizar a análise exploratória dos dados e a aplicação dos algoritmos de agrupamento das técnicas de aprendizado de máquina não supervisionado. Para a visualização dos dados, foram utilizadas as bibliotecas Seaborn e Matplotlib, a fim de facilitar a interpretação dos dados.

A metodologia utilizada para o processo de preparação dos dados e o entendimento do negócio possibilita boas práticas para executar um projeto. De modo simplificado, o processo KDD é uma metodologia abrangente de mineração de dados e um modelo de processo que fornece para qualquer usuário de DM, seja ele iniciante ou especialista, um modelo completo para realização de um projeto de DM. O projeto é dividido em seis fases: entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação [Shearer 2000].

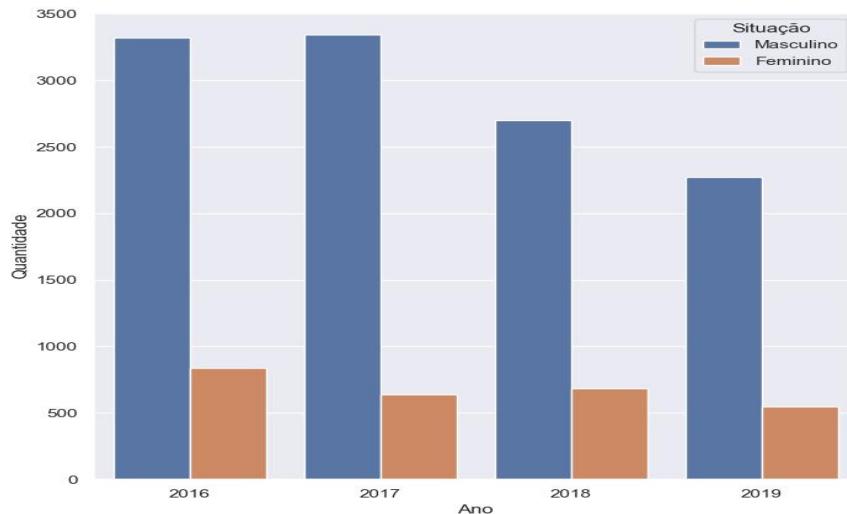
Para a preparação dos dados, foram selecionados os atributos **sexo, estado, região, ano, disciplinas e conteúdos**. Esses atributos permitem mostrar a quantidade de participantes por sexo, estado e nível de escolaridade (mestrado/doutorado ou autoavaliação), bem como por ano e disciplinas/áreas de conhecimento específicas.

Durante a limpeza dos dados, identificou-se que apenas dois candidatos haviam sido registrados incorretamente, portanto, foram removidos sem prejuízo do conjunto de dados de homologação dos candidatos. Na base de dados de notas, foram eliminados os valores ausentes para executar os algoritmos k-means, DBSCAN e Agrupamento Hierárquico. Dessa forma, não houve perda significativa para a aplicação de técnicas de aprendizado de máquina não supervisionado.

### **4. Resultados e Experimentos**

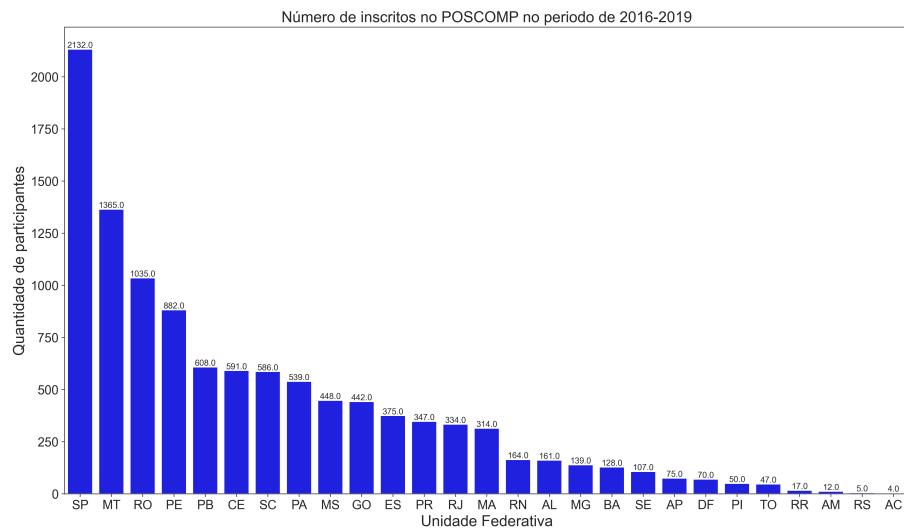
Na base de dados dos candidatos foram calculados a média, desvio padrão, entre outros, utilizando os recursos das bibliotecas Pandas, NumPy e algoritmos de agrupamento.

Na figura 1, observa-se a participação dos candidatos interessados em ingressar no mestrado ou doutorado, e uma pequena parcela de candidatos que buscam avaliar apenas seus conhecimentos nas áreas de computação.



**Figura 1. Quantidade de participantes por cargo em relação ao sexo**

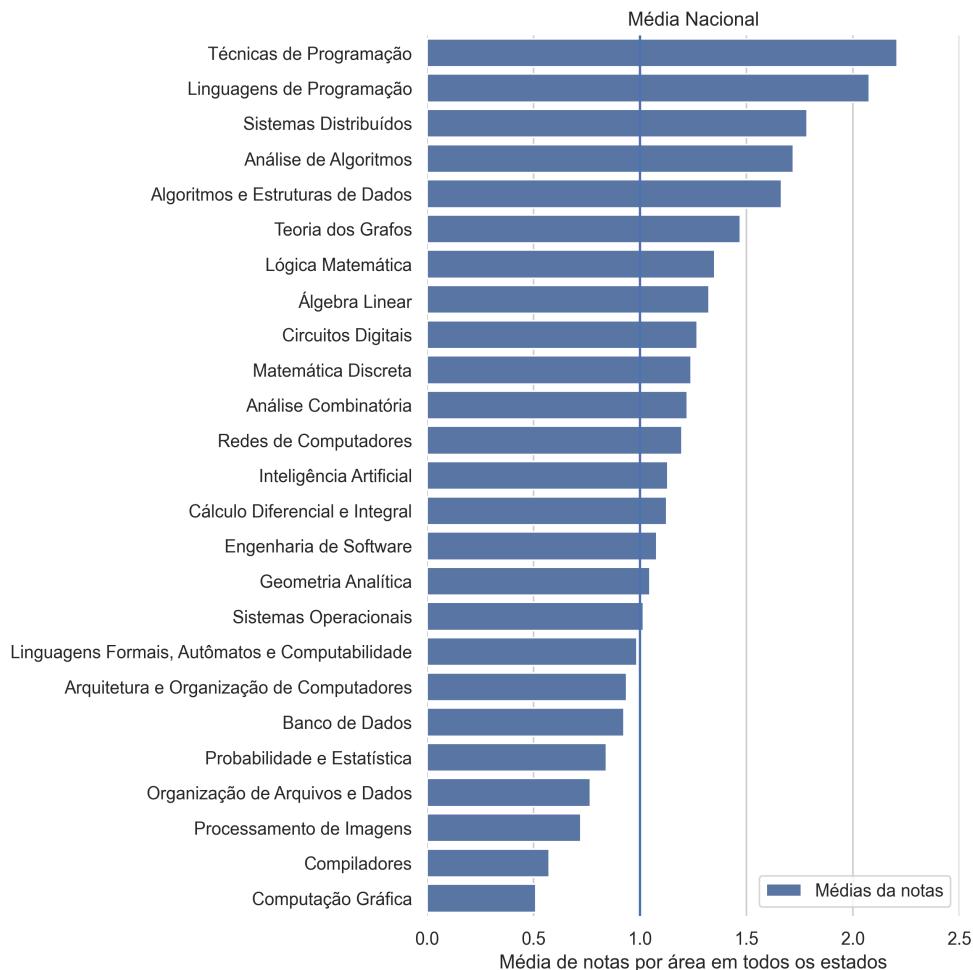
A figura 2 mostra o número de participantes por estados do Brasil. Os estados de Sergipe, Mato Grosso e Rondônia estão com maior número de participantes no POS-COMP. Já os estados de Acre, Rio Grande do Sul são os estados com menor número de participantes. De acordo com as informações fornecidas na plataforma Sucupira, no estado de Acre, o PPGCC da UFAC teve início em 2017.



**Figura 2. Total de inscritos nos anos de 2016 a 2019**

Na figura 3, são apresentados os conteúdos abordados no POSCOMP, juntamente com a média das notas obtidas pelos candidatos. Observa-se que os estudos sobre algoritmos de programação tiveram o maior número de acertos pelos candidatos. Isso ocorre

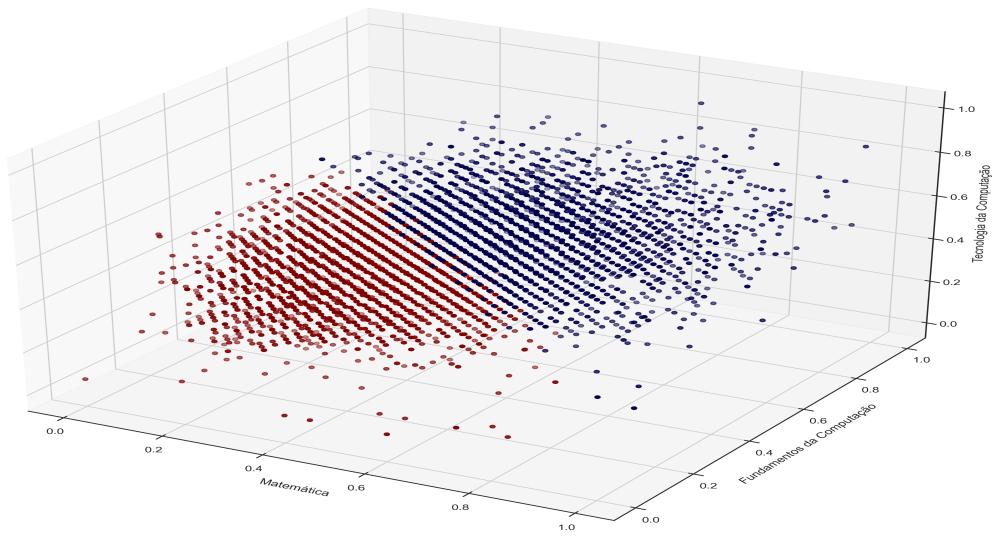
devido a diversos fatores, como o grande número de vagas para programadores, além de incentivos na área. Além disso, é importante ressaltar que há um crescente interesse no ensino de programação desde o ensino básico, o que ajuda os alunos a aprimorar e despertar o raciocínio lógico.



**Figura 3. Média dos candidatos nos conteúdos**

Então, foram aplicados algoritmos de agrupamento para buscar padrões dos candidatos. No primeiro momento, foi aplicado na base de dados que contém as somas das notas em cada eixo temático. No entanto, com a visualização dos dados, foi identificado que não há padrões entre os dados, como demonstra o gráfico seguinte.

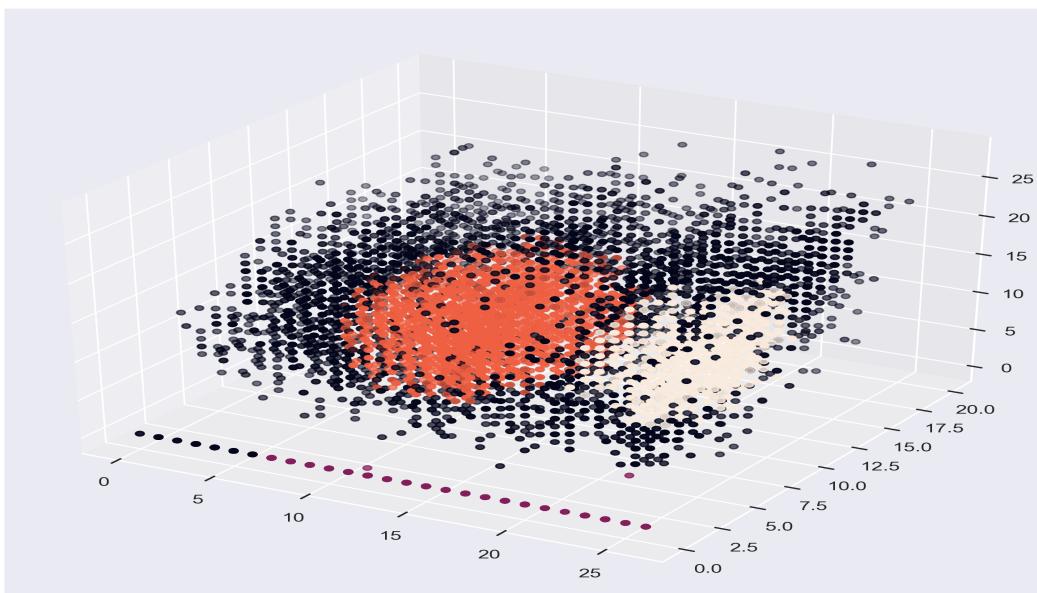
Na aplicação do algoritmo K-means, foi utilizado 2 clusters para gerar os grupos, como identificado na figura. O K-means é um dos algoritmos mais aplicados em estudos para identificação de padrões dentro de bases de dados.



**Figura 4. Aplicação do algoritmos K-means**

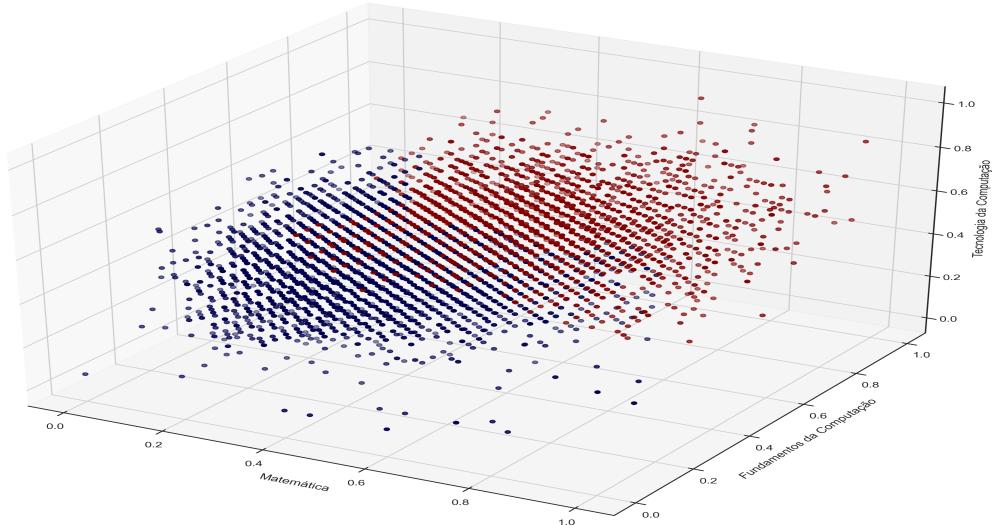
Com a identificação dos grupos podemos analisar quais candidatos se enquadram nos eixos de estudos do POSCOMP.

Também foi utilizado o algoritmo DBSCAN para identificação dos grupos. O DBSCAN não precisa definir o número de cluster como no K-means, mas é preciso definir alguns parâmetros para encontrar o número de cluster, o que torna complicado e sendo necessário uma aplicação para encontrar os parâmetros ideias de acordo com a dados a serem trabalhos. Com o DBSCAN, foram identificados *outliers*, pois o algoritmo demonstra eles como -1 ao definir os números de cluster.



**Figura 5. Aplicação do algoritmos DBSCAN**

Com a utilização do agrupamento hierárquico, percebe-se que os resultados são parecidos com os algoritmos anteriores, porém é apresentado o número de 2 cluster na aplicação. Na figura 6 é mostrado o resultado da aplicação do algoritmo.



**Figura 6. Aplicação do algoritmos de agrupamento hierárquico**

De acordo com os algoritmos *K-means*, DBSCAN e agrupamento hierárquico são apresentados resultados que chegam próximos em relação aos grupos. Os algoritmos *K-means* e agrupamento hierárquico demonstram resultados semelhantes, já o algoritmo DBSCAN mostra diferença em relação aos outros, percebendo que alguns dados estão próximos ou dentro dos outros grupos. Na tabela 1 é mostrado o desempenho dos algoritmos desenvolvido neste trabalho.

**Tabela 1. Desempenho dos algoritmos**

Algoritmos	<b>K-means</b>	<b>Agrupamento Hierárquico</b>	<b>DBSCAN</b>
<b>Coeficiente de Silhueta</b>	0.33	0.30	-0.4
<b>Índice de Davies-Bouldin</b>	1.14	1.22	4.73
<b>Índice de Calinski-Harabasz</b>	7051	5761	—

## 5. Considerações finais

Neste estudo, utilizamos técnicas de agrupamento para identificar padrões entre os candidatos que realizaram o POSCOMP nos anos de 2016 a 2019. Foi possível realizar uma análise exploratória dos dados, compreendendo a distribuição dos candidatos por estado, sexo e região, bem como suas notas médias por disciplina e conteúdo. Além disso, o estudo possibilitou uma avaliação da aplicação de algoritmos de agrupamento e técnicas de aprendizado de máquina na análise de dados do POSCOMP.

Durante a realização deste trabalho, foram identificadas algumas dificuldades, como a definição dos dados a serem trabalhados e a escolha dos hiperparâmetros dos

algoritmos de agrupamento, especialmente do K-means. No entanto, essas dificuldades foram superadas com a utilização de bibliotecas e ferramentas disponíveis para a análise de dados.

Como trabalhos futuros, sugerimos a realização de estudos utilizando outros algoritmos de agrupamento e técnicas de aprendizado de máquina para a identificação de padrões nos dados do POSCOMP. Além disso, é possível explorar outras técnicas de ajuste de hiperparâmetros para obter resultados mais precisos e robustos. Acreditamos que este estudo possa contribuir para a compreensão da distribuição dos candidatos e suas notas no POSCOMP, auxiliando na identificação de áreas de melhoria para a preparação dos candidatos.

## Referências

- Ahmed, M. R., Tahid, S. T. I., Mitu, N. A., Kundu, P., and Yeasmin, S. (2020). A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6.
- Amazona, M. V. and Hernandez, A. A. (2019). Modelling student performance using data mining techniques: Inputs for academic program development. In *Proceedings of the 2019 5th International Conference on Computing and Data Engineering, ICCDE' 19*, page 36–40, New York, NY, USA. Association for Computing Machinery.
- Arun, D. K., Namratha, V., Ramyashree, B. V., Jain, Y. P., and Roy Choudhury, A. (2021). Student academic performance prediction using educational data mining. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–9.
- Carrillo, J. M. and Parraga-Alava, J. (2018). How predicting the academic success of students of the espam mfl?: A preliminary decision trees based study. In *2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6.
- de Castro, L. N. and Ferrar, D. G. (2016). *Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações*. Saraiva, São Paulo, first edition.
- de Computação, S. B. (2022). Exame nacional para ingresso na pós-graduação em computação (poscomp).
- Fernando Raguro, M., Carpio Lagman, A., P. Abad, L., and S. Ong, P. L. (2022). Extraction of lms student engagement and behavioral patterns in online education using decision tree and k-means algorithm. In *2022 4th Asia Pacific Information Technology Conference, APIT 2022*, page 138–143, New York, NY, USA. Association for Computing Machinery.
- Gunawan, Hanes, and Catherine (2019). Information systems students' study performance prediction using data mining approach. In *2019 Fourth International Conference on Informatics and Computing (ICIC)*, pages 1–8.
- Hui, H., Ming-jie, T., Qing-tao, Z., and Xiao-liang, Z. (2020). Application of student achievement analysis based on apriori algorithm. In *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pages 19–22.

- Islam, R., Sazid, M. T., Mahmud, S. R., Ferdous, C. N., Reza, R., and Hossain, S. A. (2019). Parametric study of student learning in it using data mining to improve academic performance. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 286–290.
- Jain, A., shah, K., Chaturvedi, P., and Tambe, A. (2018). Prediction and analysis of student performance using hybrid model of multilayer perceptron and random forest. In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, pages 1–7.
- Lauría, E. J. M., Moody, E. W., Jayaprakash, S. M., Jonnalagadda, N., and Baron, J. D. (2013). Open academic analytics initiative: Initial research findings. LAK ’13, page 150–154, New York, NY, USA. Association for Computing Machinery.
- Moura, N., Gordiano, R. S., Silva, R. K. J., and Santos, S. S. (2012). Poscomp e a importância da pós-graduação para aprimoramento profissional. *Centro Federal de Educac,ao Tecnol ~ ogica de Minas-Gerais*.
- Nabil, A., Seyam, M., and Abou-Elfetouh, A. (2021). Prediction of students’ academic performance based on courses’ grades using deep neural networks. *IEEE Access*, 9:140731–140746.
- Senthil, S. and Lin, W. M. (2017). Applying classification techniques to predict students’ academic results. In *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pages 1–6.
- Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.
- Silva Guerra, M., Asseiss Neto, H., and Azevedo Oliveira, S. (2018). A case study of applying the classification task for students’ performance prediction. *IEEE Latin America Transactions*, 16(1):172–177.