# identificação de perfis socioeconômicos de universitários prestadores do enade utilizando k-means

Guilherme Zafra Garcia (e-mail); João Paulo Dutra Kreling (jpdk.97@gmail.com); Marlon Alves Bomfim (marhlon@hotmail.com.br)

***Resumo:*** *Devido à necessidade de identificar e ser assertivos ao criar medidas públicas e uma melhor relação com os estudantes, os governos têm disponibilizado bases de dados. Ter maior confiabilidade e precisão nas informações recebidas é vital na tomada de decisões, e a utilização de técnicas de aprendizado de máquina (do inglês, Machine Learning - ML) tem auxiliado para a criação de campanhas e políticas que melhor se adequam ao povo. Com este foco, este trabalho teve como objetivo a aplicação de técnicas de ML de agrupamento para identificar grupos de características socioeconômicas dos universitários, que realizaram o ENADE. Os dados foram pré-processados e foi utilizado a técnica* -*means. O método buscou estabelecer uma relação entre os valores da base de dados para identificar a melhor distância e identificar a quais clusters pertencem. Os resultados foram satisfatórios, por permitirem identificar os perfis dos estudantes, assim como a relação com os cursos escolhidos.*

***Palavras-chave:*** *ENADE; Ensino Superior; Mineração de Dados; Agrupamento;* -*means.*

## Introdução

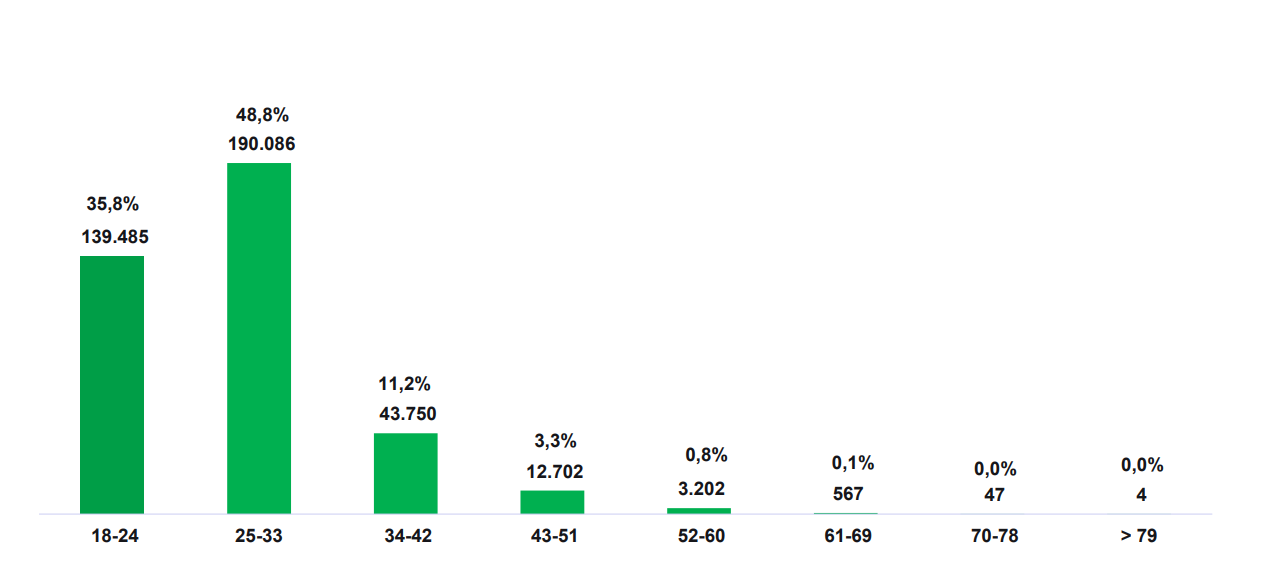
Segundo Bourdieu & Passeron (1970), a escola é o principal meio de legitimar e/ou perpetuar a diferença de classes socioeconômicas. Os resultados de vários testes de habilidade usados ​​para monitorar as políticas públicas fortalecem esta tese.

O Exame Nacional de Desempenho dos Estudantes (ENADE) avalia o desempenho dos egressos dos cursos de graduação de acordo com os conteúdos programáticos especificados no guia do curso, aprofunda o desenvolvimento de competências e habilidades necessárias à formação geral e profissional e atualiza o nível de relacionamento do aluno com o Brasil e as realidades do mundo (Ministério da Educação - MEC, 2020).

Com aplicação realizada desde 2004, o ENADE faz parte do Sistema Nacional de Avaliação da Educação Superior (SINAES), que inclui também avaliação de cursos de graduação e avaliação institucional. Juntos, eles formam um tripé de avaliação para entender a qualidade dos cursos e instituições de ensino superior brasileiros. Os resultados do ENADE, juntamente com as respostas aos questionários dos alunos, são a base para o cálculo dos indicadores de qualidade do ensino superior. Devem se inscrever calouros e graduados qualificados de cursos de graduação e técnicos avançados relacionados à área de avaliação da edição (Ministério da Educação - MEC, 2020).

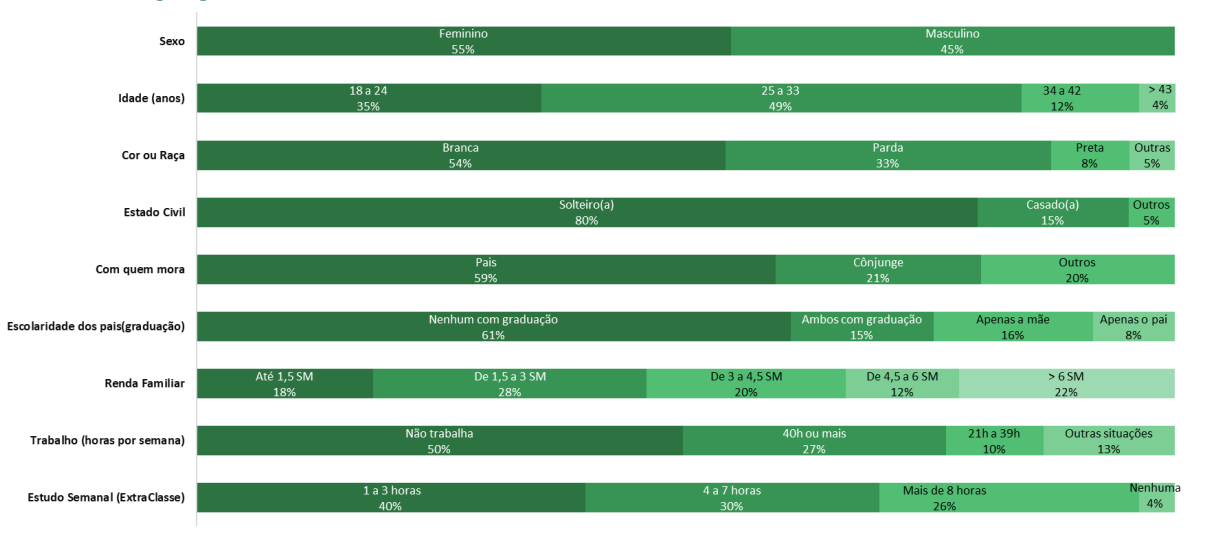
Segundo o INEP, na edição de 2019, o Exame apresentou 435.469 concluintes inscritos no exame, sendo 333.894 (77%) estudantes de Instituições de Ensino Privadas e 101.575 (23%) de Instituições Públicas. Com base nos dados apresentados pelo INEP, pode ser observado que a maioria dos inscritos ainda são jovens, conforme pode-se verificar na Figura 1:

Figura 1 - Faixa etária ENADE 2019

Fonte: INEP (2020).

Entrando em maiores detalhes socioeconômicos, observa-se a seguinte estratificação ilustrada na Figura 2:

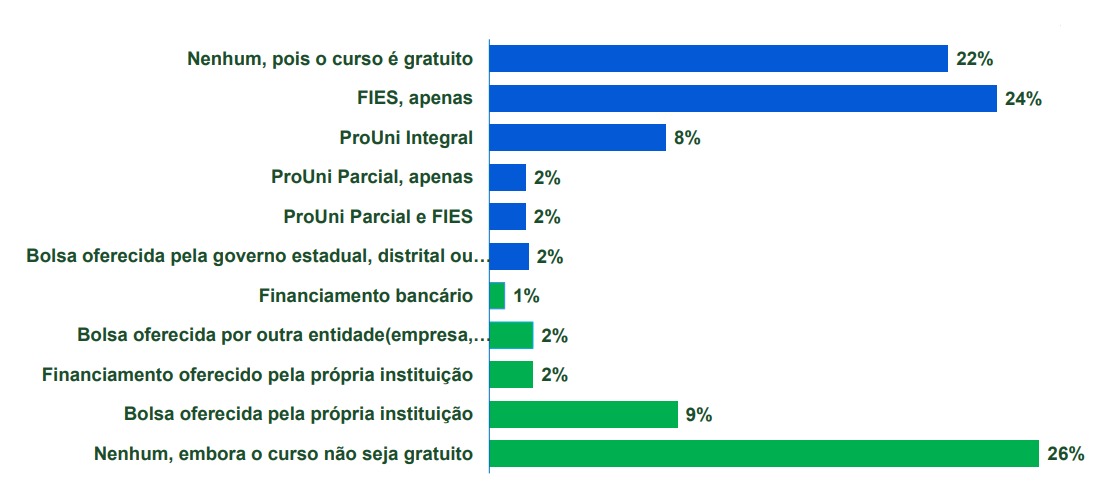
Figura 2 - Estratificação socioeconômica



Fonte: INEP (2020).

Conforme citado anteriormente a respeito do monitoramento de políticas públicas por parte do governo para melhorias e adequações, 60% dos inscritos são beneficiados por financiamento ou subsídios públicos conforme Figura 3:

Figura 3 - Estratificação por beneficiários do Estado



Fonte: INEP (2020).

Além das pontuações do ENADE (conceito ENADE), que visam avaliar diretamente o desempenho dos alunos, existem outras métricas utilizadas para formar indicadores de qualidade para cursos de ensino superior, por meio do conceito preliminar de curso (CPC), e instituições de ensino superior, que são avaliadas pelo índice geral de cursos (IGC).

O conceito ENADE é um indicador de qualidade que avalia o curso com base no desempenho dos alunos no ENADE. Apenas são calculados e divulgados anualmente os cursos em que pelo menos dois alunos concluintes fazem o exame (Ministério da Educação - MEC, 2020).

Segundo o Ministério da Educação (2020), no CPC, é calculado o conceito dos cursos avaliados, atribuindo notas em escala de 1 a 5. Os seguintes fatores são considerados no cálculo:

* Conceito ENADE (desempenho do aluno na prova ENADE);
* Indicador da diferença entre desempenho observado e esperado (IDD);
* Professor: (informações sobre mestrado, doutorado e sistemas de trabalho do censo avançado))
* Alunos: visão sobre o processo de formação (informações do Questionário do Aluno do ENADE) (Ministério da Educação - MEC, 2020).

O IGC é um indicador expresso em conceitos, com pontuação variável de um a cinco pontos. O índice é resultado da média ponderada do (CPC), indicador de avaliação de cursos de graduação, e obedece a um ciclo de três anos, em combinação com o resultado do ENADE, que mede o desempenho dos estudantes (Ministério da Educação - MEC, 2020).

Com a crescente acessibilidade a dispositivos mais tecnológicos, para Alpaydin (2016), a maioria das tarefas que eram executadas manualmente no passado agora são realizadas automaticamente por máquinas. Outrora, os dados tinham um comportamento passivo, apenas manipulados para processamento e devolvidos ao utilizador. Hoje, o programador começa a atingir esta função mais passiva, pois o próprio algoritmo aprende por conta própria, definindo o passo seguinte, se tornando cada vez mais independente do programador. De acordo com Alpaydin (2016), dados coletados previamente podem ser usados ​​para encontrar padrões de comportamento e, apoiado nestes padrões, é possível desenvolver uma previsão de um próximo comportamento futuro, ou seja, antever o que vai acontecer e ajudar a compreender o sistema como um todo. Isso é chamado de mineração de dados (data mining), o qual grandes quantidades de dados são processadas ​​para elaborar modelos mais simples, mas valiosos.

Na mineração de dados, existem algumas técnicas de aprendizado de máquina (*Machine Learning*). Segundo Mahesh (2018), essas técnicas são utilizadas para orientar as máquinas a processar dados de maneira mais eficiente por meio de algoritmos, devido não ser possível interpretar e extrair o conteúdo necessário do banco de dados. Além disso, a procura por esse recurso segue ampliando devido ao grande número de novos conjuntos de dados disponíveis.

De acordo com Ayodele (2010), o aprendizado de máquina envolve arquitetar algoritmos que possibilitem que os computadores pratiquem, assimilem e aprendam, o qual o comportamento de aprendizagem se resume a encontrar estatísticas, ordenamentos e outros padrões nos dados em estudo. Portanto, o tipo de algoritmo usado depende de fatores como o contexto de problema a ser resolvido, o número de variáveis, dentre outros (AYODELE, 2010).

O foco do presente artigo é a utilização de técnicas de clusterização com o algoritmo de agrupamento de dados k-means para realizar a identificação do perfil dos alunos avaliados no ENADE, buscando compreender as características e individualidades socioeconômicas dos inscritos.

O seguinte trabalho visa a demonstração prática dos vários perfis socioeconômicos e suas particularidades relacionadas aos inscritos do Exame Nacional de Desempenho dos Estudantes (ENADE).

Nesta seção foram apresentados o desenvolvimento em relação sobre os perfis socioeconômicos dos alunos, métricas de avaliação do ensino superior e os principais estudos sobre *Machine Learning* e *Data Mining*. Na Seção 2 foi feita uma descrição dos passos executados durante o desenvolvimento da pesquisa e a descrição do conjunto de dados e as ferramentas utilizadas. Na Seção 3 é discorrido sobre os resultados obtidos e uma análise para interpretação dos mesmos. Na Seção 4 conclui-se este artigo com uma discussão sobre os resultados empíricos e suas implicações para trabalhos futuros.

## Materiais e Métodos

Os dados foram pré-processados e analisados utilizando a linguagem de programação Python, versão 3.7.12, a partir da IDE online Google Colab. As etapas de implementação das fases executadas estão descritas na sequência e representadas no fluxograma da Figura 4:

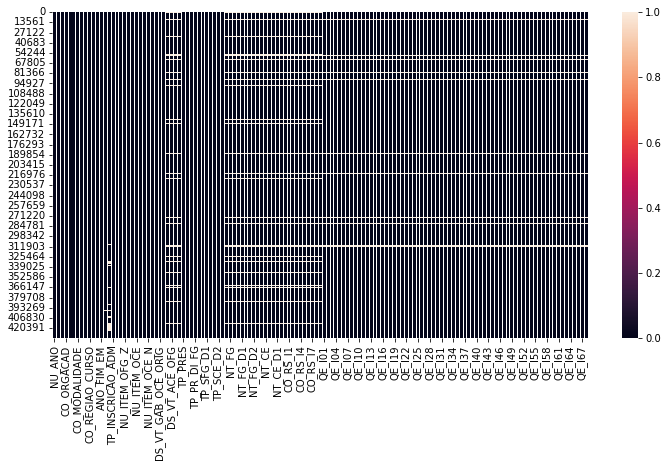
Figura 4 - Etapas do processamento dos dados



Fonte: os autores (2021).

A base de dados bruta foi importada no programa e o filtro inicial foi realizado. Para esta primeira etapa, houve a análise de dados faltantes nas mais de 430 mil instâncias e mais de 137 atributos. A Figura 5 ilustra a ocorrência de dados faltantes, representada pela cor azul clara, enquanto que existe determinado dado onde há o azul escuro.

Figura 5 - *Heatmap* dos dados faltantes



Fonte: os autores (2021).

Percebe-se na Figura 5 que há variáveis com inúmeros dados faltantes, precisamente 3.064.184 valores faltantes, correspondendo a 5,15% de toda a base. Para o pré-processamento, primeiramente foi realizada a filtragem dos atributos, as variáveis foram selecionadas por ter relação com perguntas socioeconômicas, e informações do curso, mantendo-se um conjunto com as quantidades iniciais de instâncias, com 23 atributos e apenas 4,95% dos dados faltantes. Os dados dos alunos cujo exame ENADE não foram prestados, ou haviam informações faltantes, foram removidos, resultando em 387.351 instâncias.

Na sequência foi realizado a alteração do nome das colunas, para facilitação da identificação dos atributos dentro da base de dados, e substituição das instâncias caracterizadas como “*strings*” para valores “*integer 64*”. Foi então analisado o agrupamento dos dados com base na Idade, Sexo, Raça, Modalidade, e dificuldades ao longo do ensino superior.

Para realizar a tarefa de agrupamento, foi aplicado a técnica -*means*. O número de *clusters* de maneira empírica, foi considerado 4 para execução das atividades de agrupamento devido a quantidade de instâncias identificadas nesta base de dados. A avaliação do modelo foi feita a partir da plotagem gráfica, e estabelecendo a relação dos dados entre eles.

## Resultados e Discussão

Após a coleta dos dados e a seleção dos mesmos, chegou-se a uma base de dados com um total de 387342 alunos inscritos no ensino superior que prestaram o ENADE, caracterizados por 23 atributos. Este conjunto de dados foi submetido a análise do número de grupos ideal, parâmetro demandado pelo algoritmo -*means*.

Como a base de dados é relacionada ao exame aplicado em 2019, foram avaliados 49 cursos, das áreas de ciências biológicas e exatas, e a constituição das informações dos cursos foram mantidas com os seus códigos junto ao MEC durante a realização da avaliação dos *clusters*.

**3.1 Análise dos *cluster* entre as relações de atributos**

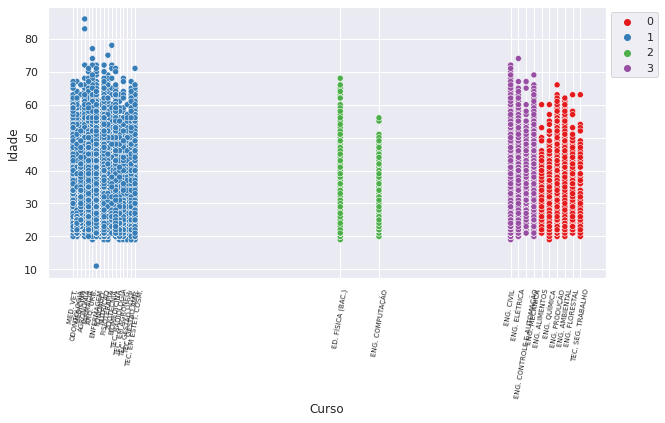
Os *clusters* analisados permitiram compreender que a avaliação de quais características é mais forte quando se trata a qual curso está matriculado e a idade do universitário, porém não deixa de traçar alguns paralelos importantes dentro de cada *cluster*.

No cluster 00, é formado por cursos da área exatas, entre eles estão cursos como Engenharia de Produção e Tecnologia em segurança do trabalho, tenho uma média de idade de 26,71 anos, em que 52,32% dos estudantes é do sexo masculino, com a faixa de renda entre 3 a 4,5 salários mínimos como predominante. Para o *cluster 01*, há uma predominância dos cursos de ciências biológicas (como medicina, enfermagem etc), cuja média de idade é de 27,06 anos, frequentado por 73,97% das pessoas do sexo feminino, com a maior concentração da faixa de renda entre 1,5 a 3 salários mínimos.

Enquanto no *cluster 02*, é composto pelos cursos de bacharelado em educação física e engenharia de computação, e a idade dos universitários é de 27,08 anos em média, com presença do público masculino de 67,25% nestes cursos, cuja renda principal varia entre 1,5 e 3 salários mínimos. E o *cluster* 03, compreende demais engenharias, como civil, elétrica entre outros, com faixa de idade média dos universitários de 27,28 anos, a participação do público masculino corresponde a três quartos dos estudantes, cuja renda está distribuída principalmente entre 3 e 4,5 salários mínimos.

Com base de utilização dos quatro *clusters* para a avaliação dos dados, a primeira relação avaliada foi com base no curso dos universitários e a distribuição da idade no momento da realização do Enade, conforme ilustra a Figura 6:

Figura 6 - *Clusters* de Curso x Idade

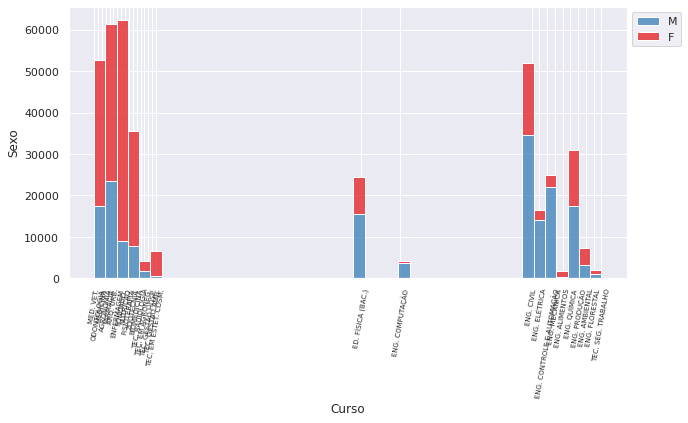


Fonte: os autores (2021).

Ao avaliar o curso de engenharia de computação, que é uma área com crescente em relação ao mercado de trabalho, é a que possui uma menor distribuição na faixa etária dos alunos, porém é agrupada junto ao curso de bacharelado de Educação Física.

Outro ponto importante, é conhecer a distribuição por sexo dentro dos cursos, se baseando na quantidade de *clusters* estabelecidos, conforme ilustrado pela Figura 7:

Figura 7 - Relação curso sexo

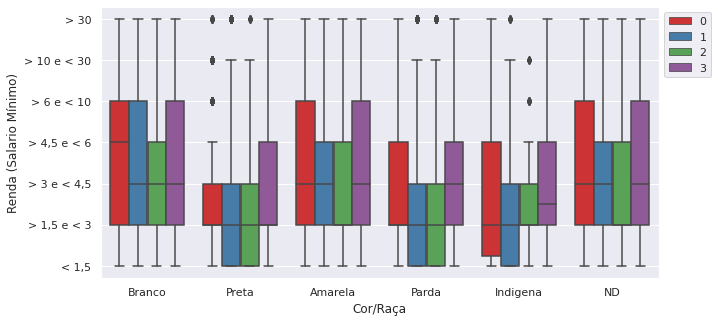


Fonte: os autores (2021).

Percebe-se que na Figura 7 que há muito mais mulheres nos cursos envolvendo a biologia, enquanto para as engenharias, ocorre o inverso com a predominância dos homens.

A partir do momento que se tem os perfis definidos para os *clusters* envolvendo curso, sexo e idade, cabe avaliar como é a relação da renda e a raça autodeclarada do estudante, permitindo a melhor compreensão de quem compõe a classe discente para cada curso ofertado nas universidades. A Figura 8 mostra esta relação entre os *clusters*, renda e cor/raça.

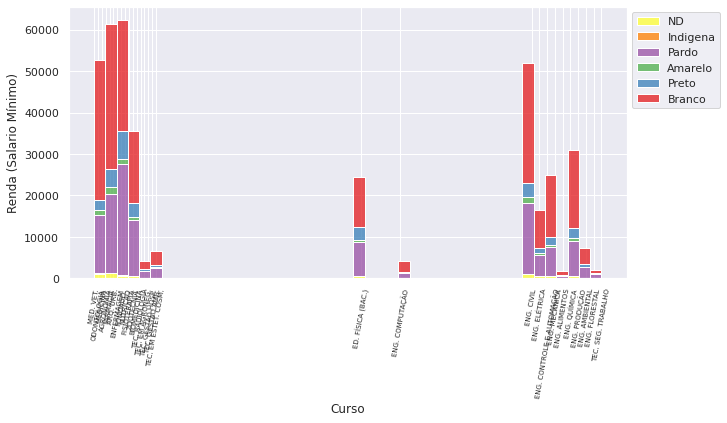
Figura 8 - Relação Raça/Cor autodeclarada X Renda



Fonte: os autores (2021).

A partir da figura 8, começa a ser feito um mapeamento do perfil dos estudantes do ensino superior, porém para a afirmação acima feita, é necessário complementar com a Figura 9:

Figura 9 - Relação Raça/Cor autodeclarada X Renda

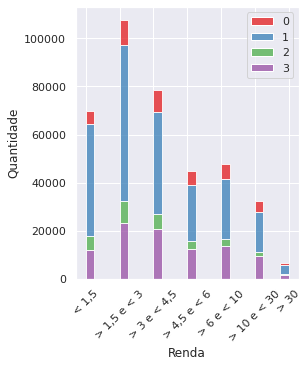
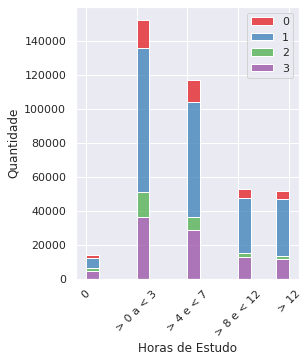


Fonte: os autores (2021).

O ponto de destaque com a análise de ambas as figuras, constata-se que a população autodeclarada negra, parda e indígena, são as que menos estão inseridas no meio acadêmico, quanto aos autodeclarados brancos, assim como na análise de renda, são os que possuem uma média menor em seus rendimentos financeiros.

Como mapeado a renda, tem-se que entender o comportamento das horas de estudos e a forma de ingresso do universitário. Este agrupamento de informações permite compreender quem são os alunos que mais necessitam de auxílios da universidade para continuar os estudos, sejam eles financeiros, reforços disciplinares, entre outros.

Figura 10 - Relação Quantidade de horas de estudo e renda.

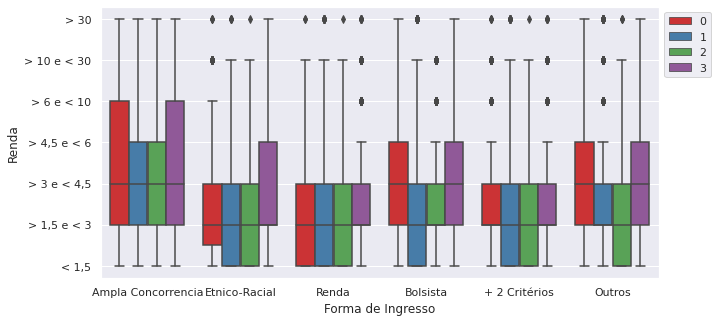


Fonte: os autores (2021).

A distribuição das instâncias obtidas monstra que a maior concentração dos universitários estuda entre 0 e 7 horas por semana e possuem renda que variam entre menor do que 1,5 salários mínimos, e 4,5 salários mínimos.

Destaca-se que, com exceção dos universitários que ingressaram no ensino superior, todas as outras políticas afirmativas apresentam média de renda entre 1,5 a 3 salários mínimos, demonstrando que todos os *clusters* neste critério definem que rendas maiores que 10 salários mínimos são *outliers*, conforme ilustra a Figura 11:

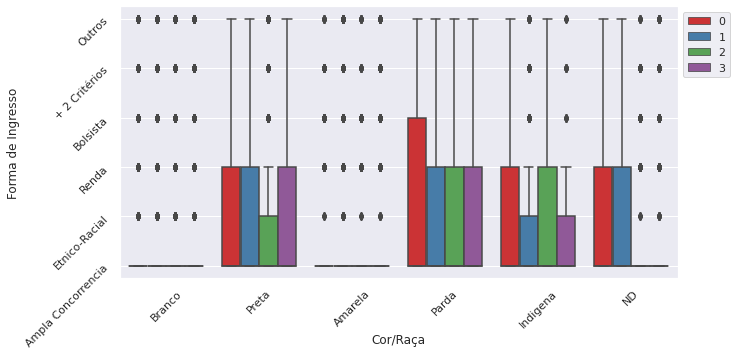
Figura 11 - *Boxplot* dos *clusters* em relação a forma de ingresso X Renda.



Fonte: dos autores (2021).

Outro demonstrativo ao comparar os *clusters*, é a utilização de políticas públicas para ingresso no ensino superior, cabe também destacar um pequeno detalhe que, para os universitários autodeclarados brancos há outliers informando que utilizaram cotas étnico-raciais. Aqui volta-se à discussão sobre a diferença entre os ingressos, por ficar evidente que boa parte do acesso à universidade de Pretos, Pardos e Indígenas, se dá pelas políticas de inclusão, conforme a Figura 12:

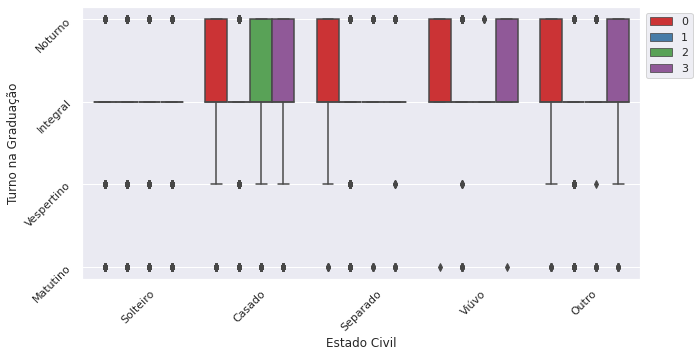
Figura 12 - *Boxplot* dos *clusters* em relação a forma de ingresso X Cor/Raça autodeclarada.



Fonte: os autores (2021).

Também é importante compreender em qual período o universitário realiza seus estudos, e a relação com o seu estado civil, para compreender muitas vezes o que leva a escolher o curso superior, a Figura 13 mostra esta relação, que permite compreender que, em sua grande maioria os cursos integrais possuem pessoas solteiras, enquanto os demais, são mais diversificados e o noturno ganha espaço.

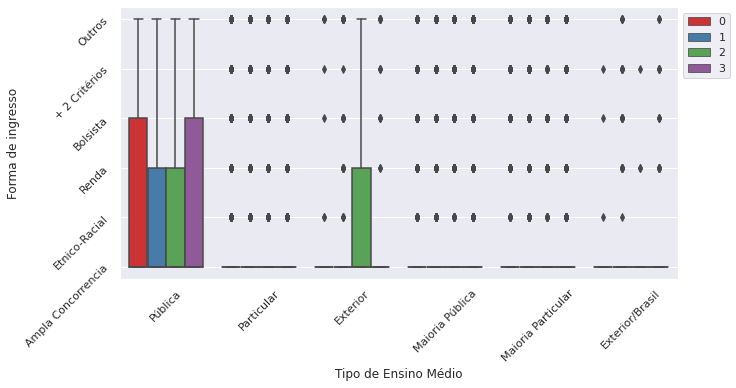
Figura 13 - *Boxplot* dos *clusters* em relação ao Estado Civil X Turno da Graduação.



Fonte: os autores (2021).

Por último, a análise feita compreende a relação de ingresso com o tipo de ensino médio cursado, pois a origem é um dos pontos que têm maior impacto nas dificuldades de cursar e acompanhar os conteúdos quando o aluno ingressa no ensino superior. Conforme a Figura 14 ilustra, em sua maioria, os alunos que utilizaram políticas afirmativas para entrar no ensino superior, vieram de escolas públicas.

Figura 14 - *Boxplot* dos *clusters* em relação a forma de ingresso X Tipo de Ensino Médio.



Fonte: os autores (2021).

Cabe ressaltar que, em todos os *clusters*, nota-se que a maioria dos alunos se auto declararam brancos, cuja renda varia entre 1,5 a 6 salários mínimos, enquanto pretos, pardos e indígenas, possuem a menor faixa de renda, e são ingressos a partir de políticas públicas.

## Conclusões

O artigo aplicou o algoritmo de agrupamento de dados -*means* para realizar a identificação do perfil dos alunos avaliados no ENADE, buscando compreender as características de quem cursa o ensino superior no Brasil.

Com a definição de quatro *clusters* para a execução do trabalho, pode-se identificar os perfis dos universitários, ingressos em cursos das áreas das ciências biológicas e das ciências exatas, tais como a faixa etária, sexo, renda, origem, raça/cor autodeclarada e tempo de estudo.

Portanto, os resultados do trabalho são considerados satisfatórios e dentro do esperado, desde o cálculo da melhor quantidade de grupos até o agrupamento em si, o qual foi validado pelo especialista de domínio.

As dificuldades encontradas para a realização do trabalho se deram por conta do tamanho da base de dados que consta com mais de 430 mil instâncias e 137 atributos, e que mesmo quando reduzidas para os dados necessários, ainda permanece com aproximadamente 380 mil instâncias e 23 atributos, e em alguns momentos, o processamento dos dados teve que ser interrompido ou não finalizado, por estar levando mais de 15 horas para a sua execução.

Outro ponto, foi identificar os melhores gráficos para relacionar duas variáveis com as quantidades de *clusters*. Devido ao tamanho da base de dados, muitas análises foram deixadas de ser executadas para priorizar as principais citadas neste trabalho.

As características obtidas dos *clusters*, permite sugerir às universidades e governo, melhorias nas políticas de condução e integração dos universitários no meio acadêmico, e políticas públicas de integração e permanência dos mesmos.

Como sugestão de trabalhos futuros, pode-se citar a redução de dimensionalidade, como o PCA, ou de seleção de instâncias, para a melhoria do conjunto de dados, para que as etapas de avaliação de hiper parâmetros possam ser executadas. Além disso, recomendação da utilização de outros atributos, para identificação de outras características que os *clusters* possam ter.

### Referências Bibliográficas

AYODELE, T. O. **TYPES OF MACHINE LEARNING ALGORITMS**. New Advances in Machine Learning. InTech. 2010. 10.5772/9385.

ALPAYDIN, E. **MACHINE LEARNING:** THE NEW AI. The MIT Press Essential Knowledge Series. Massachusetts Institute of Technology. 2016.

BOURDIEU, P.; PASSERON, J. C. **La reproduction: éléments d’une théorie du système d’enseignement**. Paris: Minuit, 1970. (Collection le sens commun)

Conceito Enade. **Ministério da Educação**, 2020. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/conceito-enade>. Acesso em 02/12/2021

## Exame Nacional de Desempenho dos Estudantes (Enade). Ministério da Educação, 2020. Disponível em:<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade> . Acesso em 02/12/2021

MINISTÉRIO DA EDUCAÇÃO. **Inep completa 82 anos de serviços prestados à educação**. 2020. Disponível em: <http://portal.mec.gov.br/component/tags/tag/indice-geral-de-cursos>. Acesso em 02/12/2021.

MINISTÉRIO DA EDUCAÇÃO. **Inep divulga indicadores que avaliam cursos e instituições**. 2020. Disponível em: <http://portal.mec.gov.br/component/tags/tag/32911>. Acesso em 02/12/2021.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), **Resultados Enade 2019**. Conceito Enade e IDD 2019. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 2020. Disponível emhttps://download.inep.gov.br/educacao\_superior/enade/apresentacao/2020/Apresentacao\_Resultados\_Enade\_Conceito\_Enade\_IDD\_2019.pdf. Acesso em 07/12/2021.

MAHESH, B. **MACHINE LEARNING ALGORITHMS - A REVIEW**. International Journal of Science and Research, v.9, i.1, p. 381-386. 2018. 10.21275/ART20203995.