UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

LUIS OTÁVIO OLIVEIRA CAPELARI

PREDIÇÃO DO DESEMPENHO NO ENADE DOS DISCENTES DE COMPUTAÇÃO

CAMPO MOURÃO

2022

LUIS OTÁVIO OLIVEIRA CAPELARI

PREDIÇÃO DO DESEMPENHO NO ENADE DOS DISCENTES DE COMPUTAÇÃO

Performance prediction in ENADE of computing students

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. André Luis Schwerz

CAMPO MOURÃO

2022

LUIS OTÁVIO OLIVEIRA CAPELARI

PREDIÇÃO DO DESEMPENHO NO ENADE DOS DISCENTES DE COMPUTAÇÃO

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 15/junho/2022

André Luis Schwerz
Doutorado
Universidade Tecnológica Federal do Paraná

Diego Bertolini Gonçalves
Doutorado
Universidade Tecnológica Federal do Paraná

Juliano Henrique Foleis
Doutorado
Universidade Tecnológica Federal do Paraná

CAMPO MOURÃO

2022

AGRADECIMENTOS

Agradeço ao meu orientador Prof. Dr. André Luis Schwerz, por ter me ajudado durante muito tempo, ter compartilhado sabedoria e experiencia, pela paciência comigo, pelo bom relacionamento que foi estabelecido, além de ter me ajudado a crescer na área de Ciência de Dados.

Agradeço aos professores Dr. Diego Bertolini Gonçalves e Dr. Juliano Henrique Foleis, pelas suas colaborações, com conselhos, dicas e mentorias, além de disporem tempo para me auxiliarem. Além deles, agradeço também ao Prof. Dr. Marco Aurélio Graciotto Silva, por ter sido a ponte entre mim e o Prof. Dr. André Luis Schwerz, fazendo com que esse trabalho se iniciasse.

Por fim, agradeço aos meus amigos e familiares que me apoiaram e estiveram comigo durante toda minha jornada acadêmica, por terem sidos importantes e marcantes nessa trajetória, além de me darem forças durante vários momentos.

RESUMO

A avaliação da educação superior é importante para garantir sua qualidade, além de ser do interesse de discentes, docentes, governos e da sociedade. O Sistema Nacional de Avaliação da Educação Superior (SINAES) é uma rede de avaliação educacional com vários métodos de avaliação, sendo um deles o Exame Nacional de Desempenho de Estudantes (ENADE), que é organizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). O ENADE avalia os alunos de cursos de graduação que estão no último ano do curso ou completaram no mínimo 80% da carga horaria do curso. Esse exame possui um ciclo avaliativo de três anos, avaliando um conjunto de cursos diferentes em cada um desses anos, sendo os cursos da área de computação avaliados no Ano II. O resultado das avaliações e dos questionários complementares são disponibilizados como dados abertos pelo INEP, servindo como fonte rica para análises e a descoberta de novos conhecimentos. Embora os dados estejam disponíveis, ainda são restritas as pesquisas cientificas que exploram suas relações e evoluções, principalmente, no que diz respeito a identificação de fatores que auxiliam a tomada de decisão dos gestores das Instituições de Ensino Superior. Tendo isso em mente, o objetivo do presente trabalho é utilizar o método de Descoberta de Conhecimento em Base de Dados junto a técnicas de classificação para predizer o desempenho dos discentes de computação participantes do ENADE a partir de fatores socioeconômicos, buscando descobrir qual o melhor modelo de classificação e qual o melhor ponto de corte para gerar classes do desempenho dos participantes. Nesse trabalho foram avaliados cinco modelos de classificação: Árvore de Decisão, K-vizinhos mais próximos, Máquina de Vetores de Suporte, Floresta Aleatória e Regressão Logística; além de serem testadas cinco formas de classificar os participantes de acordo com o desempenho. Como resultado, foi descoberto que a melhor combinação para a classificação de desempenho dos participantes é a utilização do modelo Máquina de Vetores de Suporte adotando a mediana como forma de separar as classes de participantes. O resultado obtido contribui no entendimento de como os dados do ENADE podem ser utilizados para predizer o desempenho dos participantes, na definição de modelos de Aprendizado de Máquina adequados para esse contexto, além da descoberta de como a forma de divisão de desempenho dos participantes impacta no desempenho dos modelos de classificação e de qual delas é a mais indicada.

Palavras-chave: predição; desempenho; enade; classificação; aprendizado de máquina.

ABSTRACT

The evaluation of higher education is important to ensure its quality, in addition to being of interest to students, teachers, governments and society. The National Higher Education Assessment System (SINAES) is an educational assessment network with several assessment methods, one of which is the National Student Performance Examination (ENADE), which is organized by the National Institute for Educational Studies and Research Anísio Teixeira (INEP). ENADE evaluates undergraduate course students who are in the final year of the course or who have completed at least 80% of the course workload. This exam has an evaluation cycle of three years, evaluating a set of different courses in each of these years, with courses in the computing area being evaluated in Year II. The results of assessments and complementary questionnaires are made available as open data by INEP, serving as a rich source for analysis and the discovery of new knowledge. Although the data are available, scientific research that explores their relationships and evolutions is still restricted, especially with regard to the identification of factors that help decision-making by managers of Higher Education Institutions. With this in mind, the objective of the present work is to use the Discovery of Knowledge in Database method together with classification techniques to predict the performance of computing students participating in ENADE based on socioeconomic factors, seeking to discover the best model. of classification and what is the best cut-off point to generate classes of the participants' performance. In this work, five classification models were evaluated: Decision Tree, K-nearest neighbors, Support Vector Machine, Random Forest and Logistic Regression; in addition to testing five ways of classifying participants according to performance. As a result, it was found that the best combination for classifying participants' performance is to use the Support Vector Machine model, adopting the median as a way of separating the classes of participants. The result obtained contributes to the understanding of how ENADE data can be used to predict the performance of participants, in the definition of Machine Learning models suitable for this context, in addition to the discovery of how the way participants' performance is divided impacts the performance of classification models and which one is the most suitable.

Keywords: prediction; enade; performance; classification; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Abreviaturas

CEFET	Centro Federal de Educação Tecnológica
CPC	Conceito Preliminar de Curso
DT	Árvore de Decisão
ENADE	Exame Nacional de Desempenho de Estudantes
ENEM	Exame Nacional do Ensino Médio
IDD	Indicador de Diferença entre os Desempenhos Observado e Esperado
IES	Instituições de Ensino Superior
IFECT	Instituto Federal de Educação, Ciência e Tecnologia
IGC	Índice Geral de Cursos
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	Descoberta de Conhecimento em Base de Dados
KNN	K-Vizinhos Mais Próximos
LR	Regressão Logística
MDE	Mineração de Dados Educacionais
MEC	Ministério da Educação
RF	Floresta Aleatória
SINAES	Sistema Nacional de Avaliação da Educação Superior
SVM	Máquina de Vetores de Suporte

SUMÁRIO

1 INTRODUÇÃO

A avaliação do ensino superior é um tema amplo e complexo de interesse de diversos grupos como discentes, docentes, governos, instituições e também da sociedade (ALVARES; CAMPOS; GOMES, 2015). As avaliações educacionais medem índices de aprendizado dos discentes e têm como um dos principais objetivos garantir a qualidade do ensino. Os dados gerados por essas avaliações permitem observar o desempenho dos estudantes, as debilidades e potencialidades das instituições (BRITO, 2015), além de possibilitar o cálculo de indicadores de qualidade que podem auxiliar a proporcionar melhorias nos processos de ensino e de aprendizagem (LIMA *et al.*, 2019).

Em 2004, o Sistema Nacional de Avaliação da Educação Superior (SINAES) foi instituído como uma rede de avaliação educacional composta de vários métodos. Entre eles, encontra-se o Exame Nacional de Desempenho de Estudantes (ENADE), um exame que avalia o rendimento dos concluintes de cursos de graduação em relação aos conteúdos programáticos dos cursos, o desenvolvimento de competências e habilidades necessárias para a formação geral e profissional, e o nível de atualização dos estudantes com relação à realidade brasileira e mundial. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é um órgão que promove estudos, pesquisas e avaliações a respeito do sistema educacional brasileiro (ARAÚJO *et al.*, 2019), além de ser o responsável por realizar as avaliações. Com base nas informações geradas pelo SINAES, as Instituições de Ensino Superior (IES) podem tomar decisões quanto a seus cursos (VERHINE; DANTAS; SOARES, 2006; LIMA *et al.*, 2019).

O ENADE foi proposto como um exame obrigatório aos discentes das IES, com um ciclo avaliativo de três anos, em que a cada ano desse ciclo são avaliadas algumas áreas de ensino, sendo o Ano II responsável por avaliar as áreas de Ciências Exatas, Licenciaturas e afins, na qual estão incluídos os cursos da área de computação. Esse exame é aplicável a todos os alunos que estejam no último ano dos cursos avaliados ou que tenham concluído mais que 80% da carga horária do curso. A prova aplicada baseia-se no conteúdo programático de cada curso. Além de ser submetido a questões de formação geral e questões de conhecimento específico, o aluno tem que responder a dois questionários: um sobre sua percepção da prova e o outro a respeito de suas informações socioeconômicas (VERHINE; DANTAS; SOARES, 2006; LIMA *et al.*, 2019).

Desde sua concepção, o SINAES propôs que os procedimentos, dados e resultados de suas avaliações deveriam ser públicos e abertos, podendo ser utilizados por qualquer pessoa (ISOTANI; BITTENCOURT, 2015). Com este intuito, o Ministério da Educação (MEC) disponibiliza no portal do INEP dados abertos de domínio público referentes a cada ano de execução das avaliações. Os chamados microdados oferecem de forma anonimizada informações a respeito do desempenho de cada participante, percepção da prova, questionário socioeconômico, entre outras informações (VERHINE; DANTAS; SOARES, 2006; LIMA *et al.*, 2019). Os dados gerados

por essas avaliações oferecem uma grande quantidade de informações relevantes a gestores, pesquisadores, educadores e comunidade (CRISPIM NETO, 2020).

Em relação ao ENADE, o interesse das IES está no Conceito ENADE, que é um indicador da qualidade de ensino medido com base no desempenho dos alunos participantes da prova (VERHINE; DANTAS; SOARES, 2006; LIMA *et al.*, 2019). O Conceito ENADE é uma nota que varia de 1 a 5, sendo que cursos com conceitos 1 ou 2 estarão impossibilitados de abrir novas turmas, podendo também terem de encerrar suas atividades. Portanto, é de interesse das IES que seus cursos tenham um conceito acima de 3 (NUNES, 2018). O resultado no ENADE é somado a outros indicadores de desempenho das IES referentes à infraestrutura, titulação dos docentes e ao projeto pedagógico, obtendo assim a nota final da avaliação do MEC de um curso (SILVA *et al.*, 2015). Esses indicadores são o Conceito Preliminar de Curso (CPC), que avalia separadamente cada curso de uma instituição e é calculado levando em conta quatro dimensões: o Conceito ENADE, o Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD), o perfil dos professores, e a percepção dos estudantes sobre seu processo formativo com base no questionário do estudante do ENADE; e o Índice Geral de Cursos (IGC), que avalia a instituição como um todo e sua nota considera três pontos: média do CPC dos últimos três anos, a média dos conceitos de avaliação dos programas de pós-graduação *stricto sensu*, e a distribuição dos estudantes entre os diferentes níveis de ensino, graduação ou pós-graduação *stricto sensu*; e (ALVARES; CAMPOS; GOMES, 2015).

Para as IES públicas, o desempenho acadêmico impacta na distribuição de verbas orçamentarias. Por outro lado, para as IES privadas, ter uma melhor avaliação pode gerar uma maior procura por parte dos discentes, consequentemente, acarretando um aumento na receita da IES (FERREIRA *et al.*, 2015). Outros benefícios advindos de uma boa avaliação incluem uma maior facilidade na captação de recursos para financiamento de pesquisas, aumento da procura por parte da indústria para criar parcerias, além de dar mais destaque para o corpo docente (BRITO, 2015).

Obter o conhecimento dos fatores que influenciam o processo de ensino-aprendizagem é relevante para auxiliar a definição de políticas educacionais e de estratégias no processo de ensino-aprendizagem para que, além de melhorarem o desempenho dos participantes no ENADE, possam contribuir para a melhoria da qualidade do ensino dos cursos superiores e no desenvolvimento social brasileiro (ARAÚJO *et al.*, 2013; LEMOS; MIRANDA, 2015; ARAUJO, 2017). As descobertas sobre a influência dos fatores também podem contribuir para melhorar a relação do discente com áreas mais críticas de aprendizado, além de que estratégias administrativas e pedagógicas podem ser implementadas para melhorar a relação do professor com o conteúdo ministrado (LIMA *et al.*, 2021).

A melhor forma de se descobrir os fatores relacionados ao desempenho dos discentes e, além disso, conseguir predizer o desempenho dos mesmos, é através da análise e extração de informações com base nos dados fornecidos pelo MEC. No entanto, os microdados disponibilizados pelo MEC oferecem um enorme volume de dados. Segundo Fayyad, Piatetsky-Shapiro

e Smyth (1996), é inviável analisar um grande volume de dados utilizando métodos tradicionais como planilhas e relatórios informativos operacionais. Para viabilizar a exploração desses dados se faz necessária a utilização de técnicas específicas e eficientes, como a Mineração de Dados.

A Mineração de Dados é uma área que envolve Banco de Dados, Estatística e Aprendizado de Máquina (SILVA, 2015). Ela tem a função de extrair ou minerar informações e conhecimentos, que geralmente não são facilmente percebidas por humanos, em grande volumes de dados. Com ela é possível encontrar padrões e relações entre os dados. A Mineração de Dados compõe um processo maior, denominado Descoberta de Conhecimento em Base de Dados (KDD) (VISTA; FIGUEIRÓ; CHICON, 2017), que tem como objetivo converter dados brutos em informações úteis (ARAÚJO *et al.*, 2019).

A aplicação de técnicas de Mineração de Dados no campo da educação ocasionou no surgimento de uma nova área de investigação denominada Mineração de Dados Educacionais (MDE). Essa área tem como foco o desenvolvimento de métodos para explorar dados obtidos em ambientes educacionais (CARVALHO, 2011), como, por exemplo, plataformas de ambiente virtual de aprendizagem e dados de avaliações institucionais (CRISPIM NETO, 2020). O uso de MDE tem se mostrado vantajoso, contribuindo na tomada de decisão de docentes, no controle de evasão de alunos em IES, na detecção de fatores que influenciam a aprendizagem, na predição do desempenho dos alunos e também em pesquisas aplicadas ao Exame Nacional do Ensino Médio (ENEM) e ENADE (ARAÚJO *et al.*, 2019). Os resultados da MDE são de interesse de estudantes, educadores, administradores e pesquisadores, pois ajudam a melhorar o ensino e aprendizado, a entender melhor as estruturas educacionais, além de fornecer ferramentas para tomada de decisões, organização e coordenação das instituições (CRISPIM NETO, 2020). Com a MDE é possível, por exemplo, minerar dados para verificar a relação entre uma abordagem pedagógica e o aprendizado do aluno, permitindo compreender se a abordagem está ajudando o aluno (CARVALHO, 2011). Segundo (ROMERO; VENTURA, 2010), um dos tipos de estudo realizado pela MDE é a educação *offline* que, dentre várias coisas, analisa dados de desempenho dos alunos.

Levando isso em consideração, o presente trabalho tem como objetivo avaliar algoritmos de Aprendizado de Máquina e opções de pontos de corte para discretização dos dados, no intuito de predizer o desempenho de participantes do ENADE. Mais especificamente, foram analisados os discentes dos cursos das áreas de computação, participantes das edições de 2008, 2011, 2014 e 2017 do ENADE. Para alcançar o objetivo, foi utilizado o método KDD em conjunto com a técnica MDE. Com a realização dos processos do KDD foi descoberto que a melhor opção para predição do desempenho dos participantes do ENADE é utilizar a Máquina de Vetores de Suporte tendo como ponto de corte a mediana da Nota Geral. Além disso, com os resultados desse trabalho espera-se que as informações descobertas possam ser utilizadas pelas instituições para melhorar a qualidade de ensino dos cursos da área de computação no Brasil, promover políticas de ensino e ajudar em pontos fracos dos discentes, para alcançarem um bom desempenho no ENADE.

Para ajudar no entendimento do tema e apresentar o que já foi realizado nessa área, o Capítulo 2 traz alguns trabalhos relacionados. Ao Capítulo 3 temos a explicação da metodologia adotada nesse trabalho. No Capítulo 4 temos a descrição e análise do conjunto de dados utilizado no presente trabalho. O Capítulo 5 apresenta os resultados obtidos nos experimentos. E, por fim, o Capítulo 6 traz a conclusão desse trabalho.

2 TRABALHOS RELACIONADOS

Este capitulo sintetiza os vários trabalhos que já foram desenvolvidos buscando entender quais fatores estão relacionados ao desempenho dos discentes ou tentando predizer seu desempenho no ENADE. Na Seção 2.1, os trabalhos que buscam os fatores que influenciam o desempenho dos alunos são apresentados. E na Seção 2.2, são apresentados trabalhos que buscam predizer o desempenho dos participantes no ENADE.

2.1 Descoberta de fatores que influenciam o desempenho

Em seu trabalho, Silva, Hoed e Saraiva (2019) utilizaram-se da técnica de mineração de regras de associação, com base no algoritmo Apriori, e o método Cross Industry Standard Process for Data Mining (CRISP-DM), para descobrir quais fatores estão relacionados ao desempenho dos estudantes da área de computação. Para realizar isso, o questionário socioeconômico disponibilizado pelos microdados do ENADE de 2017 foi avaliado. Foram analisadas as variáveis de idade, sexo, nota bruta na prova, nota bruta de Formação Geral, nota bruta no Componente Específico e mais outras 76 variáveis do questionário socioeconômico. Como resultado, interessantes regras foram apresentadas, como a de que 88% dos alunos que tiveram nota alta na parte de Componente Específico e que disse que seu curso disponibilizou monitores ou tutores obtiveram uma nota bruta geral alta. Os autores também conseguiram determinar as variáveis que se relacionam com o desempenho do aluno, como o fato da IES ter uma boa estrutura e boas políticas de acompanhamento.

Moreira (2010) fez um estudo sobre quais fatores institucionais influenciam o rendimento de alunos de Biologia, Engenharia Civil, História e Pedagogia, com base nos microdados do ENADE de 2005. Foi utilizado o método de Regressão Múltipla para verificar os efeitos institucionais sobre o desempenho dos estudantes, tendo como variável resposta a nota geral no exame, e uma árvore de classificação para caracterizar as IES em relação a fatores institucionais, sendo a variável resposta a categoria administrativa da IES. O objetivo era entender quais políticas e estratégias de gestão da IES levam a um bom desempenho do estudante. De acordo com o modelo de regressão, as variáveis explicativas são: tipo de organização acadêmica da IES; região geoeconômica onde se situa o curso; grupos etários; escolaridade paterna; faixa de renda familiar; tipo de atividade extracurricular; qualidade dos fatores institucionais; titulação docente aproximada, e nível de exigência do curso. As variáveis explicativas descobertas na árvore de decisão foram: organização acadêmica da IES, índice de qualidade de fatores institucionais por IES, pós-graduação: mestrado/doutorado, nível de exigência do curso e região geoeconômica em que se situa a IES.

Araujo (2017) se dedicou a verificar se os fatores como renda familiar, atividade remunerada, nível de escolaridade dos pais, forma de ingresso à universidade e o tipo de escola do Ensino Médio impactam o desempenho de alunos de Ciências Contábeis no ENADE, com

a intenção de incentivar melhorias nas esferas discente, docente e institucional. Foram analisados 44.370 participantes presentes nos microdados do ENADE de 2012. Em seu trabalho foi utilizada uma Regressão via Bootstrap para determinar a relação entre as variáveis estudadas e o desempenho acadêmico no ENADE. Com isso, foi descoberto que as variáveis renda familiar, nível de escolaridade da mãe e forma de ingresso à Universidade podem explicar o desempenho de um aluno no ENADE.

Com o intuito de verificar se há relação de fatores socioeconômicos, de trajetória acadêmica, e de mecanismos de ingresso com o desempenho acadêmico, Machado (2019) analisou as variáveis gênero, estado civil, etnia, renda mensal familiar, situação de trabalho, mecanismo de ingresso, escolaridade do pai, escolaridade da mãe, e tipo de escola que cursou o ensino médio. O estudo observou os cursos de Ciências Contábeis das universidades públicas federais da região Nordeste. Foram analisados 2.774 alunos, participantes das edições de 2012 e 2015 do ENADE. Os dados utilizados vieram de microdados, relatórios e sinopses estatísticas disponíveis no portal do INEP. Para alcançar o objetivo, foi realizada uma análise descritiva e a aplicação de regressão linear múltipla. Os resultados indicaram que a renda familiar é o principal fator de impacto no desempenho do discente.

Para verificar a relação entre o desempenho do estudante de Nutrição no ENADE e fatores socioeconômicos, a trajetória acadêmica e o perfil da instituição, Rocha, Leles e Queiroz (2018) analisaram 23.746 estudantes de Nutrição presentes nos microdados do ENADE de 2004, 2007, 2010 e 2013, porém em seu experimento utilizando uma regressão linear simples e uma regressão múltipla foram selecionados apenas 16.983 participantes. Com seu trabalho foi descoberto que o principal fator associado ao desempenho é a categoria acadêmica da IES. Além disso, foi descoberto que alunos que usaram políticas afirmativas obtiveram notas maiores do que aqueles que não usaram.

Já Brito (2015) buscou entender como as características do corpo docente de uma IES influenciam o desempenho dos concluintes de seus cursos de graduação em Administração no ENADE. Para realizar o estudo, foram utilizados os dados disponibilizados pelos microdados do ENADE de 2012 e pelo Censo de Educação Superior de 2012. Foram criados três modelos de regressão múltipla, um modelo para todas IES, um para as IES públicas e um para as IES privadas. Ao final do trabalho, concluiu-se que o nível de escolaridade dos docentes, a quantidade de docentes e o volume de cursos ofertados pela IES influenciam o desempenho do discente no ENADE.

2.2 Predição do desempenho

O trabalho de Araújo *et al.* (2019) se propôs a criar uma ferramenta para análise exploratória dos microdados do ENADE de 2017, além de um modelo classificatório para predizer o desempenho de um estudante. Como algoritmo de classificação foram utilizados o Classification and Regression Trees (CART), o C4.5 e o Random Forest, tendo o algoritmo CART se

destacado com a melhor acurácia e sensibilidade, e a segunda melhor especificidade. Nesse trabalho foram analisadas 40 variáveis a respeito de 170.875 participantes. Após a aplicação do algoritmo CART, foram descobertas algumas regras de associação, como: alunos de IES públicas tendem a ter um alto desempenho; alunos acima dos 30 anos com renda familiar abaixo de 4,5 salários mínimos tendem a ter um baixo desempenho; e alunos que não são de IES da categoria Pública e que a renda familiar é superior a 4,5 salários mínimos tendem a ter um desempenho alto.

No trabalho de Silva *et al.* (2015), com base nos microdados do ENADE de 2012, foram analisados quais fatores podem impactar no desempenho dos estudantes de Administração no ENADE. Foram utilizados os métodos de análise fatorial para agrupar algumas variáveis e de regressão múltipla junto ao método de Anderson-Rubin para predizer as notas a partir de 47 variáveis independentes. Com isso, três modelos de regressão foram criados para predição das notas de (i) formação geral, (ii) formação específica e (iii) geral. Os resultados indicaram que as variáveis independentes são pouco eficientes para explicar as variáveis dependentes.

Rosa *et al.* (2021) aplicaram o método KDD em conjunto com o Estudo Longitudinal Transversal Repetido, um estudo que observa dados ao longo do tempo, cada vez utilizando uma amostra diferente. Foram observados os participantes do curso de Bacharelado em Ciência da Computação presentes nos microdados do ENADE das edições de 2008, 2011, 2014 e 2017. Os participantes foram divididos em duas classes (alto e baixo), utilizando como limiar de classificação 60% da maior nota, realizando isso a cada ano. Foram comparados os algoritmos Árvore de Decisão, Random Forest e Suporte Vector Machine (SVM). Em seus resultados, a Árvore de Decisão apresentou o melhor resultado com a acurácia média de 71,84%.

O trabalho realizado por Rezende *et al.* (2022) teve como objetivo determinar os aspectos socioeconômicos que mais influenciam no desempenho no ENADE de estudantes do curso Sistema de Informações. Para alcançar o objetivo, foi utilizado o modelo de Árvore de Decisão com base nos dados do ENADE de 2017. Como forma de criar classes de desempenho dos participantes, suas notas foram separas em quarto classes, de acordo com os valores de quartis da nota. Em seus resultados a Árvore de Decisão obteve 30% de acurácia, por isso foi adotado uma segunda maneira de encontrar as variáveis mais relacionadas com o desempenho, o método de coeficiente de correlação. Assim foram destacadas que as variáveis mais relevantes foram a renda familiar, a participação em intercambio e bolsa acadêmica, o tipo de financiamento de mensalidade e o tipo de escola no Ensino Médio.

2.3 Considerações finais

No melhor do nosso conhecimento, não há nenhum estudo que se utilize dos microdados do ENADE para analisar todos cursos da área da computação, com o objetivo de comparar diferentes modelos de classificação e analisar diferentes formas de categorização do desempenho no ENADE. A importância da avaliação dessas diferentes formas de categorização se

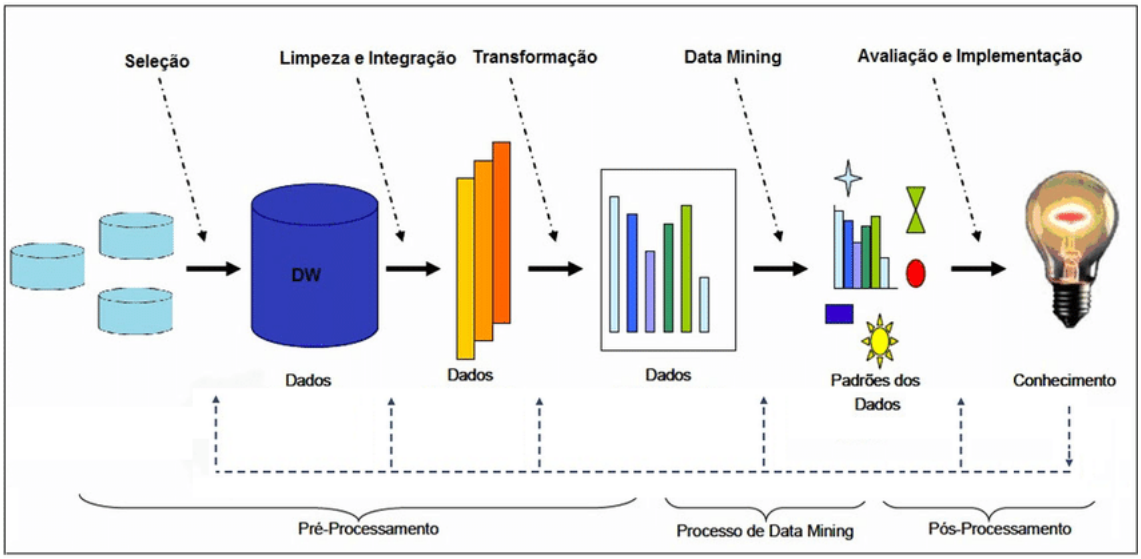
deve ao fato dessa separação não ser algo natural, sendo algo artificial, ou seja, não há uma definição oficial sobre como definir o que é um bom ou mau desempenho de um participante, se fazendo necessário testar diversas opções de como realizar essa divisão.

3 METODOLOGIA

O método adotado nesse trabalho consistiu em empregar o método KDD utilizando dados educacionais provenientes do ENADE. Além da Mineração de Dados, o KDD conta com duas outras grandes fases: pré-processamento, que inclui a seleção, limpeza e transformação dos dados; e o pós-processamento, que faz a interpretação e avaliação do conhecimento obtido (CALIL *et al.*, 2008), conforme apresentado na Figura 1. De acordo com (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), as etapas do KDD são definidas da seguinte forma:

- Seleção de Dados – compreende em descobrir e conhecer o domínio de aplicação, e então selecionar um subconjunto de dados de interesse para se trabalhar;
- Limpeza: consiste em analisar os dados para corrigir inconsistências, buscando por ruídos, tratando valores ausentes ou duplicados, e garantindo a padronização dos dados;
- Transformação – com base no conhecimento do domínio são realizados métodos de normalização, redução e discretização dos dados, visando encontrar atributos úteis para as fases seguintes;
- Mineração de Dados – é a etapa em que se aplica algum algoritmo para gerar um modelo que represente um conjunto de dados, procurando por padrões e relações; e
- Interpretação e Avaliação – serve para apresentar os conhecimentos descobertos e averiguá-los, verificando se os mesmos auxiliam na resolução do problema inicial.

Figura 1 – Etapas do KDD.



Fonte: (SANTOS; MENEZES; HORA, 2014, p. 7).

Neste trabalho foram utilizados os microdados dos concluintes de computação disponibilizados no portal do INEP referentes aos exames de 2008, 2011, 2014 e 2017. Os dados do

exame realizado em 2005 foram descartados devido à falta de compatibilidade entre os dados, como, por exemplo, não ter informação se o curso é EaD ou presencial, ter um enquadramento da área dos cursos divergente dos demais anos, entre outros fatores. Os cursos de computação também realizam o exame em 2021; entretanto, os resultados não foram divulgados até o momento da coleta dos dados deste trabalho.

As operações que foram realizadas em cada uma das fases do KDD são descritas nas subseções seguintes.

3.1 Seleção de Dados

Para facilitar o entendimento, além de que, nas diferentes edições do ENADE as variáveis podem ter nomenclaturas diferentes, as variáveis selecionadas serão apresentadas através de suas descrições. Nessa etapa foram selecionadas vinte e nove variáveis presentes nos microdados, referentes a cada participante:

1. ano de realização do exame;
2. área de enquadramento do curso no ENADE;
3. UF de funcionamento do curso;
4. região de funcionamento do curso;
5. categoria administrativa da IES (Pública ou Privada);
6. categoria da organização acadêmica da IES (Universidade, Faculdade, Centro Universitário, Centro Federal de Educação Tecnológica (CEFET) ou Instituto Federal de Educação, Ciência e Tecnologia (IFECT));
7. turno do curso;
8. nota geral no exame;
9. ano de conclusão do Ensino Médio;
10. ano de início da graduação;
11. tipo de escola que o participante cursou o Ensino Médio (Pública ou Privada);
12. modalidade de Ensino Médio;
13. idade do participante na data de realização do exame;
14. sexo do participante;
15. cor ou raça do participante;

16. estado civil;
17. nível de escolaridade do pai do participante;
18. nível de escolaridade da mãe do participante;
19. renda familiar do participante;
20. tipo de moradia;
21. quantidade de pessoas em sua moradia;
22. política de ação afirmativa ou inclusão social utilizada para ingresso no curso;
23. situação de trabalho do participante;
24. quantidade de livros lidos no ano de realização do exame;
25. horas dedicadas à estudos, além das horas de aula;
26. opinião sobre as condições das salas de aulas;
27. opinião sobre as condições das salas de aulas práticas;
28. opinião sobre o plano de ensino do curso;
29. avaliação pessoal da contribuição do curso para a formação.

Além disso, outras quatro variáveis que foram derivadas das variáveis selecionadas:

30. indicativo dicotômico de utilização de política de ação afirmativa ou inclusão social;
31. idade do participante no início da graduação;
32. quantidade de anos que o participante esteve ocioso entre a conclusão do Ensino Médio até o início da graduação;
33. quantidade de anos que o participante está na graduação.

A partir desses dados foram selecionados somente os participantes presentes que obtiveram resultado válido e que cursavam algum dos cursos da área da computação citados na Tabela 1.

Assim, após a etapa de seleção de dados restaram 170.389 participantes.

Tabela 1 – Cursos analisados e nomenclatura adotada no trabalho.

Sigla	Área de Enquadramento	Grau
ADS	Análise e Desenvolvimento de Sistemas	Tecnologia
BCC	Ciência da Computação	Bacharelado
EC	Engenharia da Computação	Bacharelado
GTI	Gestão da Tecnologia da Informação	Tecnologia
LCC	Ciência da Computação	Licenciatura
RC	Redes de Computadores	Tecnologia
SI	Sistemas de Informação	Bacharelado

Fonte: Autoria própria (2022).

3.2 Limpeza

Para remover qualquer dado que pudesse enviesar o modelo de classificação e para haver compatibilidade entre as edições analisadas foram tomadas as seguintes medidas:

- Remoção de participantes com informações faltantes;
- Remoção de participantes que reponderam ter cursado o Ensino Médio metade em escola privada e metade em escola pública, já que essa alternativa só existia em 2008 e 2011;
- Remoção de participantes que não conseguiram concluir a prova;
- Remoção de participantes ingressantes, presentes somente em 2008, tendo em mente que o foco desse trabalho são os participantes concluintes;
- Remoção de participantes que estudaram o Ensino Médio parcial ou inteiramente no exterior, que só constavam nas edições de 2014 e 2017.

Após realizados esses procedimentos de limpeza, sobraram 121.330 participantes.

3.3 Transformação

Levando em consideração que foram utilizados dados de várias edições do ENADE, essas edições continham algumas diferenças nas questões. Para padronizar os dados de todas edições, os seguintes procedimentos foram realizados nas variáveis apresentadas na etapa de seleção:

- Algumas variáveis tiveram suas respostas transformadas, como, por exemplo, nas edições de 2008 até 2014 havia uma coluna para cada tipo de turno de curso, enquanto

em 2017 existe apenas uma coluna com alternativas: "diurno", "noturno", "matutino"e "integral". Nos anos de 2008 até 2014, o turno foi transformado para uma única coluna contendo o turno do curso do participante;

- Algumas variáveis tiveram suas repostas convertidas para serem humanamente compreensíveis, como, por exemplo, a UF de residência da IES, que são originalmente representadas por números (códigos) e foram transformadas para suas siglas (em letras), como exemplo temos o Paraná que era representado pelo código 41 e foi transformado para "PR";

- Algumas variáveis tiveram suas alternativas de respostas agrupadas, como, por exemplo, o tipo de categoria administrativa da IES, onde originalmente haviam alternativas como "Pessoa Jurídica de Direito Público - Federal", "Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Civil", "Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Fundação", entre outras, foram resumidas apenas para "IES privada"e "IES pública".

Além dos itens citados acima, três procedimentos adicionais de transformação dos dados foram realizados: normalização, discretização e codificação.

3.3.1 Normalização

Este trabalho considera o resultado de diferentes exames para diferentes cursos ao longo de quatro edições do ENADE. Naturalmente, as notas sofrem uma variação de acordo com o nível do exame e o desempenho dos estudantes. Desta forma, para transformar os valores na mesma ordem de grandeza, optou-se pela normalização das notas dos estudantes.

A normalização foi realizada de maneira separada para os participantes de cada curso em cada edição analisada, seguindo o método MinMax, colocando os valores do atributo Nota Geral no intervalo entre 0 e 1. A Equação 1 exibe a normalização MinMax, tal que x_i' é o valor normalizado de x_i com base nos valores mínimo e máximo do conjunto X. No caso da nota no ENADE, a nota de cada participante era normalizada com base na nota máxima e mínima dos participantes de seu curso na mesma edição.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

3.3.2 Discretização

Para realizar a classificação, foi necessário a criação de classes com base nas notas dos participantes. Como não existe uma definição oficial da separação das classes, foram criadas

cinco opções de cortes para gerar as classes, realizadas com base nas notas normalizadas. As cinco opções de corte foram:

- 60% da nota mais alta (por ano e por curso);
- Quartil 60 da nota (por ano e por curso);
- Nota 0,6;
- 10% acima da média da nota (por ano e por curso);
- Mediana da nota (por ano e por curso).

Essas várias alternativas de opções de corte foram necessárias porque, já que a criação das classes de desempenho é algo artificial, não se tinha conhecimento sobre qual seria a melhor maneira de realizar a divisão das classes, por isso foi preciso testar as opções, visando descobrir qual delas era a melhor.

3.3.3 Codificação

Para que fosse possível utilizar os modelos de classificação, além de evitar um aprendizado enviesado dos modelos, os dados passaram por um processo de codificação, transformando as variáveis categóricas em uma representação binaria com números. As variáveis que possuíam duas opções de respostas simplesmente tiveram suas respostas convertidas para números, utilizando o método *Label Encoder*. As demais variáveis seguiram o processo chamado *OneHot Encoding*, que consiste em transformar cada alternativa de uma coluna em uma nova coluna.

O resultado da aplicação das etapas de seleção, limpeza e transformação produziram um conjunto de dados que será usado como entrada para a tarefa de mineração. Esse conjunto de dados é apresentado por meio de uma análise descritiva no Capítulo 4.

3.4 Mineração de dados

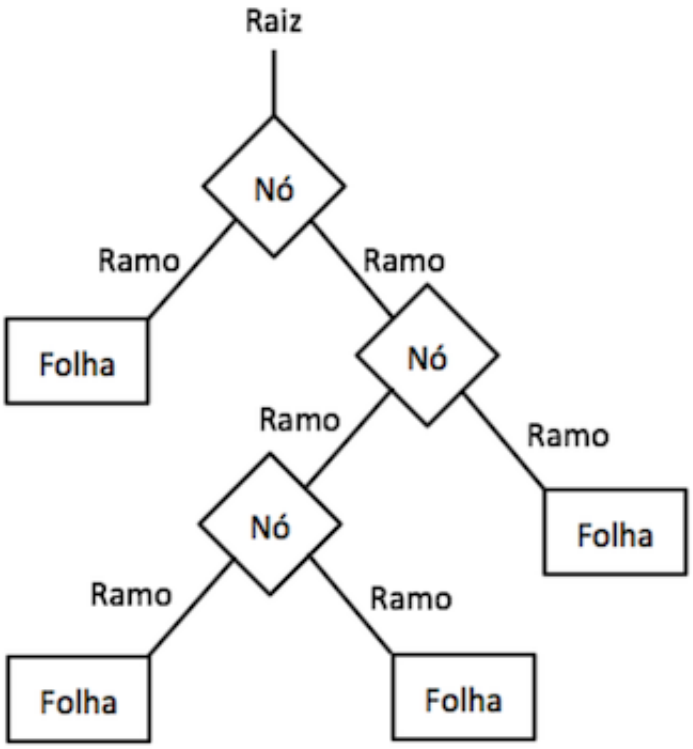
Como já dito anteriormente, a Mineração de Dados é uma área que envolve Banco de Dados, Estatística e Aprendizado de Máquina. Sendo assim, para executar a etapa de Mineração de Dados é necessário que hajam modelos de Aprendizado de Máquina e também formas de avaliar o desempenho desses modelos. Levando isso em conta, nas subseções seguintes são descritos os modelos, as métricas de desempenho e o procedimento para a execução desses modelos.

3.4.1 Classificadores

Como classificadores foram testados cinco dos mais conhecidos e comumente utilizados modelos de aprendizado supervisionado: Árvore de Decisão (DT), K-Vizinhos Mais Próximos (KNN), Máquina de Vetores de Suporte (SVM), Floresta Aleatória (RF) e Regressão Logística (LR).

A **Árvore de Decisão** é um modelo que utiliza uma estrutura de árvore para separar os dados em subgrupos. Nela são criados nós de forma binária ou múltipla, a partir de um nó raiz, separando os dados de acordo com os grupos (AMORIM, 2021). Os nós da árvore são compostos por condições vinculadas aos atributos dos dados e possíveis decisões aparecem nas ramificações desses nós. A classificação começa do nó raiz partindo para os nós folhas, que representam alguma classe (VALADARES *et al.*, 2021). Seu funcionamento pode ser explicado como uma divisão de problemas complexos em problemas mais simples, de maneira recursiva (TERAMACHI, 2020). O processo recursivo da árvore só acaba quando todos os dados de um nó pertencem a uma única classe ou quando não há mais recursos para expandir a árvore (AMORIM, 2021). Este modelo é pouco sensível a ruídos, como pontos fora da curva (*outliers*) e valores faltantes (BINUESA, 2020). A Figura 2 exemplifica uma Árvore de Decisão.

Figura 2 – Árvore de decisão.

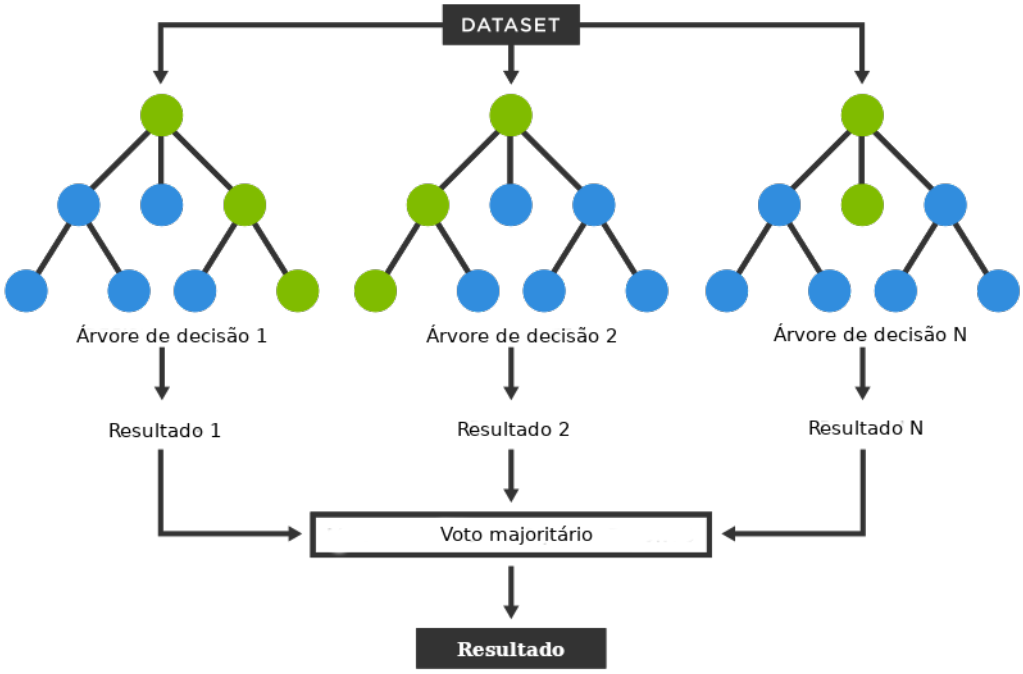


Fonte: Adaptado de Sakurai (2022).

A **Floresta Aleatória** é um método que utiliza uma coleção de árvores de decisão, em que cada árvore gera um voto para a classe mais popular dos dados de entrada (AMORIM, 2021). É um método do tipo *ensemble*, que oferece maior resistência ao *overfiting* comparado

a uma única arvore de decisão. Nele são geradas várias árvores, em que cada árvore recebe apenas uma amostra das instâncias e cada nó considera somente um subconjunto de atributos para realizar a divisão da árvore (SOUZA; FERNANDES; DUTRA, 2020). Esse modelo usa o conceito de *bagging*, que consiste em gerar uma amostragem aleatória para substituir as combinações de entrada dos dados de treinamento. O método de *bagging* ajuda a aumentar a acurácia dos resultados, além de fornecer estimativas dos erros de generalização de forma contínua das árvores de decisão (AMORIM, 2021). Cada árvore é construída com variáveis aleatórias, gerando árvores totalmente diferentes, ajudando a gerar menos erro de generalização e redução no tempo de processamento (BINUESA, 2020). Após todas as árvores votarem, a classe é determinada com base na maioria dos votos (SOUZA; FERNANDES; DUTRA, 2020). A Figura 3 demonstra o funcionamento de uma Floresta Aleatória.

Figura 3 – Floresta aleatória.

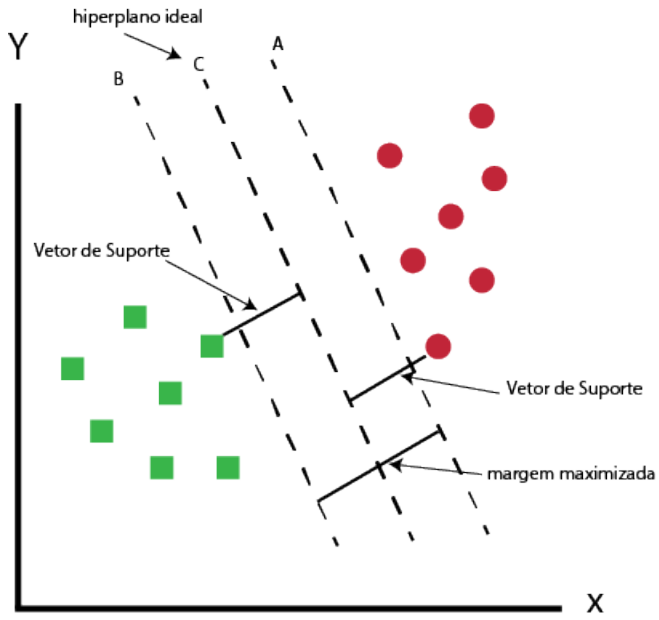


Fonte: Adaptado de TIBCO (2022).

A **Máquina de Vetores de Suporte** é um classificador linear que faz mapeamento não-linear do espaço de entrada em um espaço de alta dimensão. Este mapeamento é realizado utilizando as funções *kernel*. Trabalhar com espaço de alta dimensão é uma vantagem, pois permite flexibilizar a aplicação para vários cenários. Outro ponto de destaque é que o SVM consegue lidar com dados não linearmente separáveis (SOUZA; FERNANDES; DUTRA, 2020). Esse modelo usa o conceito de hiperplano que é uma generalização para mais de três dimensões (SOUZA; FERNANDES; DUTRA, 2020). O hiperplano perfeito é definido como aquele

que apresenta a máxima margem de separação das classes (TERAMACHI, 2020). A Figura 4 apresenta o conceito da SVM.

Figura 4 – SVM.

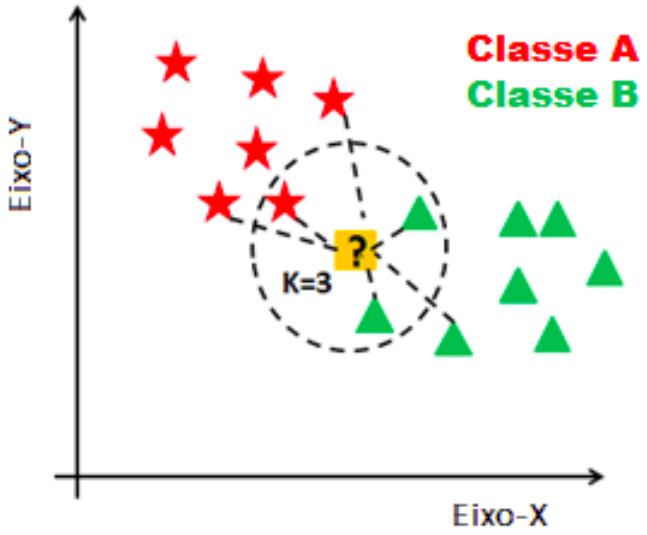


Fonte: Adaptado de Tussevana (2022).

O **K-Vizinhos Mais Próximos** é um modelo simples e eficaz. Sua funcionalidade consiste em analisar uma quantidade k de vizinhos mais próximos para cada instancia a ser predita (SOUZA; FERNANDES; DUTRA, 2020). Cada instancia representa um ponto em um espaço que é definido pelos atributos dos dados de entrada. Com a chegada de uma nova instância, é calculada sua distância para todas instâncias no espaço. Após descobertos os k vizinhos mais próximos à nova instância é classificada com base na classe majoritária dos k vizinhos (LIMA, 2013). Os vizinhos mais próximos são descobertos com base em um cálculo de distância da nova instância para as demais já presentes (SOUZA; FERNANDES; DUTRA, 2020). O KNN trabalha com base na hipótese de que dados similares tendem a estar próximos e dados diferentes tendem a estar distantes entre si (GALDINO, 2020). A figura exemplifica o funcionamento do KNN.

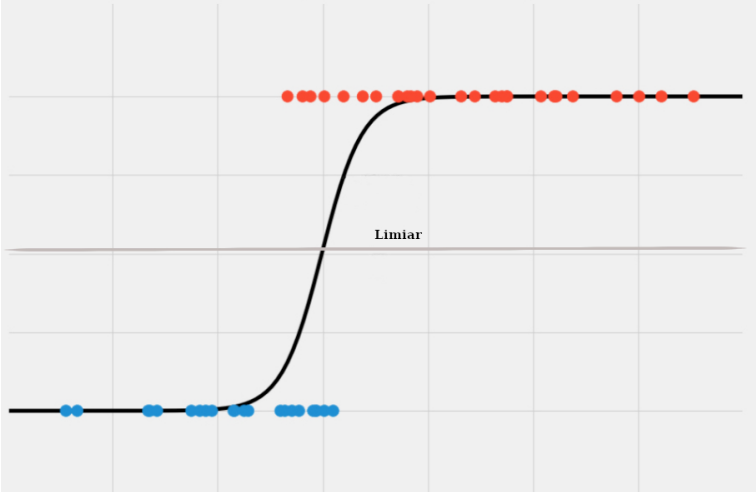
A **Regressão Logística** é um modelo estatístico que usa uma função logística para classificar uma variável alvo binária, retornando a probabilidade de uma certa instância pertencer às possíveis classes (VALADARES *et al.*, 2021). Os modelos de regressão têm o proposito de entender a relação das variáveis independentes com a variável dependente. A regressão logística é aplicável para classificação em que a variável dependente pode ser dicotômica ou multinomial e as variáveis independentes podem ser qualitativas ou quantitativas (SILVA, 2022). Na Figura 6 vemos como o modelo LR faz a classificação.

Figura 5 – KNN.



Fonte: Adaptado de Remigio (2020).

Figura 6 – Regressão logística.



Fonte: Adaptado de StatsTest (2022).

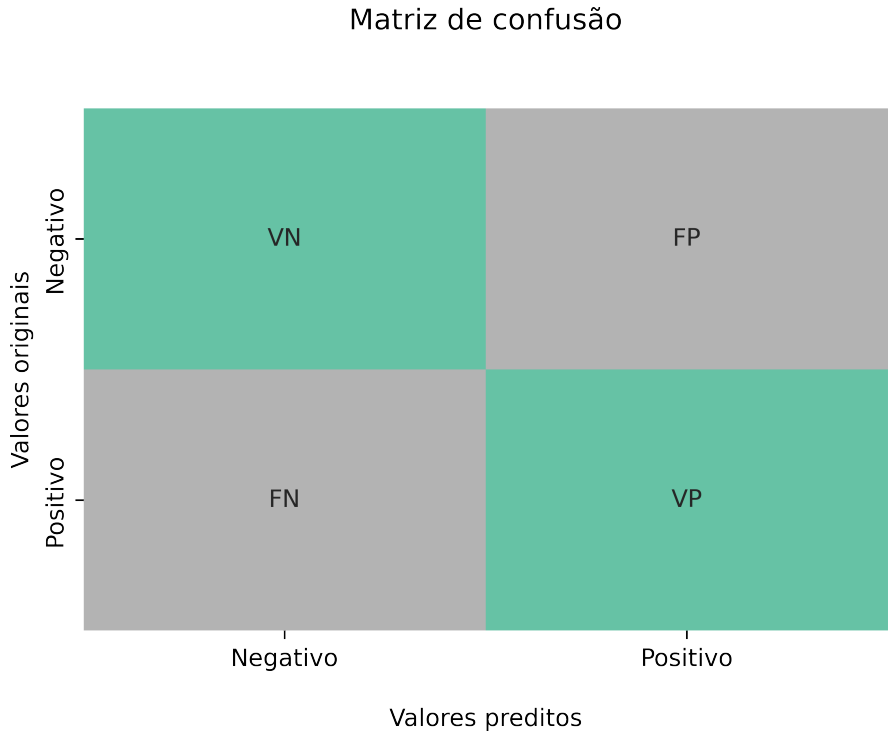
3.4.2 Métricas de desempenho

Após a predição do modelo, gera-se uma estrutura que descreve os acertos e erros do modelo denominada matriz de confusão. A matriz de confusão é uma tabela que apresenta as frequências de cada classe predita, e com base nela são geradas várias métricas de desempenho. A Figura 7 demonstra a visualização de uma matriz de confusão. Em um caso de classificação binária, a matriz contém quatro quadrantes com medidas da classificação (SILVA, 2022):

- Verdadeiro Negativo (VN) - classificação correta de instâncias da classe negativa;
- Falso Positivo (FP) - classificação incorreta de instâncias da classe negativa;

- Falso Negativo (FN) - classificação incorreta de instâncias da classe positiva;
- Verdadeiro Positivo (VP) - classificação correta de instâncias da classe positiva.

Figura 7 – Matriz de confusão.



Fonte: Autoria própria (2022).

A acurácia é a métrica de desempenho mais utilizada por representar a porcentagem de acertos no total de predições, sendo representada pela Equação 2. Entretanto, ela não é recomendada para casos de dados desbalanceamentos, pois pode-se imputar um viés de preferência às classes majoritárias, desconsiderando o impacto das classes minoritárias (AMORIM, 2021), sendo passível que um modelo classifique corretamente todas instancias da classe majoritária, mas acertando pouco na classe minoritária, e mesmo assim obtenha uma alta acurácia (VALADARES *et al.*, 2021).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2}$$

A sensibilidade (*recall*) descrita na Equação 3 verifica o percentual de instâncias da classe positiva preditos corretamente (SOUZA; FERNANDES; DUTRA, 2020), ou seja, de todas instâncias originalmente positivas, quantas foram classificadas corretamente como positivas (VALADARES *et al.*, 2021).

$$Sensibilidade = \frac{VP}{VP + FN} \tag{3}$$

Especificidade (Equação 4) representa a taxa de acertos da classe negativa, ou seja, de todas instâncias originalmente negativas, quantas foram corretamente classificadas como negativas (TERAMACHI, 2020).

$$Especificidade = \frac{VN}{VN + FP} \tag{4}$$

A precisão (*precision*) descrita na Equação 5 calcula a proporção de previsões positivas corretas, ou seja, de todas instâncias que foram classificadas como positivas quantas realmente eram positivas(VALADARES *et al.*, 2021).

$$Precisão = \frac{VP}{VP + FP} \tag{5}$$

F1-score é uma métrica calculada como a média harmônica entre sensibilidade e precisão como descrita na Equação 6. Para que seu valor seja alto é necessário que ambas métricas (sensibilidade e precisão) sejam altas e próximas, pois um desequilíbrio entre elas ocasiona em uma queda no valor do F1-score(VALADARES *et al.*, 2021).

$$F1\text{–}score = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade} \tag{6}$$

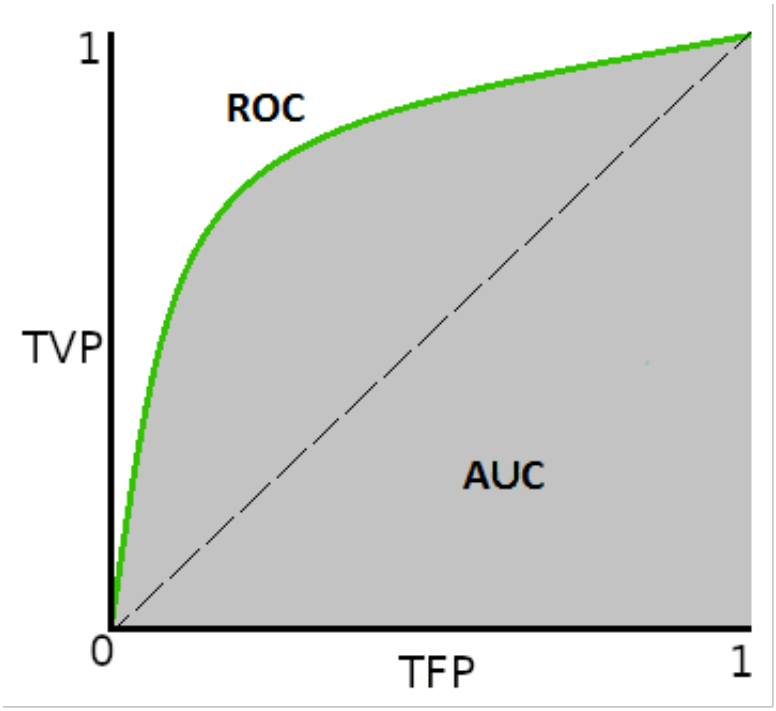
A curva ROC é uma representação gráfica que utiliza a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1 - especificidade) (AMORIM, 2021). ROC é uma curva de probabilidade enquanto AUC representa o grau de separabilidade, representando o quanto um modelo é capaz de diferenciar as classes (VALADARES *et al.*, 2021). AUC consiste na área abaixo da curva ROC (AMORIM, 2021). Quanto maior o valor da AUC, melhor é o modelo em classificar negativo como negativo e positivo como positivo (VALADARES *et al.*, 2021). Os gráficos da curva ROC possuem o eixo Y dado pela sensibilidade e o eixo X pela taxa de falsos positivos, como é apresentado na Figura 8. Nesses gráficos o ponto (0,1) representa a classificação perfeita, com todas instâncias sendo classificadas corretamente, enquanto o ponto(1,0) significa que todas instancias foram classificadas erroneamente (TERAMACHI, 2020).

Todas métricas retornam um valor de 0 a 1, sendo que quanto maior seu valor, melhor foi o modelo.

3.4.3 Experimento

A execução dos testes foi realizada em uma máquina com sistema operacional Ubuntu 18.04, 32 GB de memória RAM, processador Intel Core i7 de 4° geração, placa de vídeo NVIDIA Titan X Pascal, utilizando o Python 3.6.9 com as bibliotecas pandas 1.1.5 e scikit-learn 0.24.2.

Figura 8 – Gráfico ROC AUC.



Fonte: Adaptado de (NARKHEDE, 2018, p. 2).

Na etapa de mineração de dados foram testados os modelos DT, KNN, SVM, RF e LR. Além da execução dos modelos, as etapas de normalização e discretização também foram realizadas utilizando a técnica de validação cruzada K-fold, com 4 vias (folds).

No método K-fold os dados são divididos em k subconjuntos de mesmo tamanho. Seu processo realiza k iterações onde, a cada iteração, um dos subconjuntos será usado para teste e os outros k-1 subconjuntos para treino, como demonstrado na Figura 9. Sendo assim, ao final das k iterações, todos dados foram utilizados tanto para treino quanto para teste (CUNHA, 2019). Esse método visa avaliar a capacidade de generalização de um modelo, verificando se o mesmo possui sub-ajuste ou sobre-ajuste aos dados treino (SIMIONATO, 2022).

Em ambas etapas, normalização e discretização, o procedimento foi realizado primeiramente nos dados de treino e, com os parâmetros gerados com base nos dados de treino, os dados de teste eram transformados.

Como poderia haver nos dados de teste notas menores do que o mínimo ou maiores que o máximo do treino, após a normalização dos dados de teste, os mesmos foram ajustados, forçando eles a ficar no intervalo entre 0 e 1.

Na execução dos modelos, a cada iteração do K-fold, os modelos eram treinados com dados de treino, fazendo a busca por hiperparâmetros, que retornava os melhores hiperparâmetros para cada modelo de acordo com a métrica ROC AUC.

Hiperpârametros são configurações utilizadas na criação do modelo, como, por exemplo, a profundidade de uma Árvore de Decisão, a quantidade k de vizinhos no KNN, e o núcleo (*kernel*) do SVM. A otimização dos hiperparâmetros tem a intenção de melhorar o desempenho

Figura 9 – Validação cruzada com 4 folds.

Fonte: Adaptado de (SCHNEIDER, 2018, p. 32).

do modelo. A Tabela 2 apresenta os hiperparâmetros e os valores que foram buscados para cada modelo de classificação a cada iteração da validação cruzada.

Tabela 2 – Hiperparâmetros buscados.

Modelo	Hiperparâmetro	Valores
DT	criterion	gini, entropy
	ccp_alpha	0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0.7; 0,8; 0,9; 1
KNN	n_neighbors	3; 5; 7; 9
	weights	uniform, distance
SVM	C	1; 10; 100
	gamma	auto; scale; 0,1; 0,01
RF	criterion	gini, entropy
	ccp_alpha	0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0.7; 0,8; 0,9; 1
LR	tol	$1^{-5}; 1^{-4}; 1^{-3}$
	C	1; 10; 100

Fonte: Autoria própria (2022).

Após o treinamento foi feito a predição dos dados de teste e computadas as métricas de avaliação F1-score e ROC AUC. Ao fim das iterações das 4 vias, foram calculadas as médias e desvio padrão de cada métrica, além de ser gerada a matriz de confusão de cada modelo.

No Capítulo 4 é apresentada uma análise descritiva e a visualização dos dados utilizados. E no Capítulo 5 são relatados os resultados dos experimentos realizados.

4 ANÁLISE E VISUALIZAÇÃO DOS DADOS

Neste capítulo serão apresentadas a descrição e visualização do conjunto de dados, analisando a cada seção, uma variável, porém apresentando apenas algumas variáveis, aquelas que apresentaram informações inesperadas, ou seja, padrões e conhecimentos novos e/ou incomuns. Os dados analisados contam com 121.330 participantes, sendo 10.578 da edição de 2008, 31.589 de 2011, 38.599 de 2014 e 40.564 de 2017, e também já passaram pela etapa de pré-processamento.

4.1 Curso

Nesses dados temos 31,86% dos participantes sendo do curso SI, 24,05% do curso ADS, 23,20% do curso BCC, 9,02% de RC, 6,47% de EC, 2,88% de LCC e 2,54% de GTI. Com a Tabela 3 vemos a quantidade de participantes por curso em cada uma das edições analisadas.

Tabela 3 – Quantidade de participantes por curso em cada edição.

Sigla	2008	2011	2014	2017
ADS	1.988	6.849	11.057	9.282
BCC	2.581	9.298	8.210	8.054
EC	210	1.863	2.521	3.255
GTI	-	-	-	3.079
LCC	-	665	1.773	1.052
RC	1.553	3.168	3.837	2.380
SI	4.246	9.746	13.166	11.497
Total	10.578	31.589	38.599	40.564

Fonte: Autoria própria (2022).

4.2 Região

Começando pelo fator região, a análise do *dataset* nos informa que o mesmo é composto principalmente por alunos que estudam na região Sudeste, são 62.338 participantes de IES do Sudeste em contraste com os 6.712 participantes de IES do Norte.

Os dados também mostram que SI é o curso com mais participantes em quase todas regiões, com exceção à região Sul, onde o curso ADS possui a maior quantidade de participantes, como mostra a tabela 4.

Com base na Figura 10, observa-se que boa parte dos cursos têm uma maior presença de alunos nas regiões Sudeste, Sul e Nordeste, com ressalvas para o curso LCC, que possui

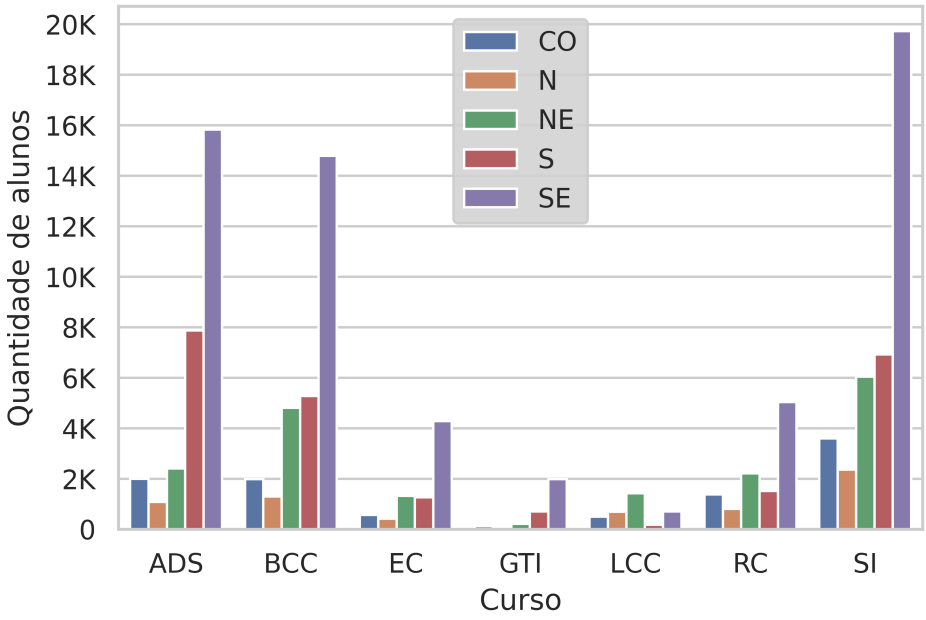
Tabela 4 – Quantidade de alunos por curso em cada região.

Sigla	S	SE	CO	N	NE
ADS	7.869	15.826	1.993	1.085	2.403
BCC	5.272	14.781	1.986	1.295	4.809
EC	1.263	4.282	567	413	1.324
GTI	699	1.980	118	66	216
LCC	173	708	497	689	1.423
RC	1.520	5.034	1.375	801	2.208
SI	6.923	19.727	3.600	2.363	6.042
Total	23.719	62.338	10.118	6.712	18.425

Fonte: Autoria própria (2022).

maior presença nas regiões Sudeste, Norte e Nordeste, e também para o curso RC, com maior presença nas regiões Sudeste, Nordeste e Centro-Oeste.

Figura 10 – Quantidade de alunos por região.



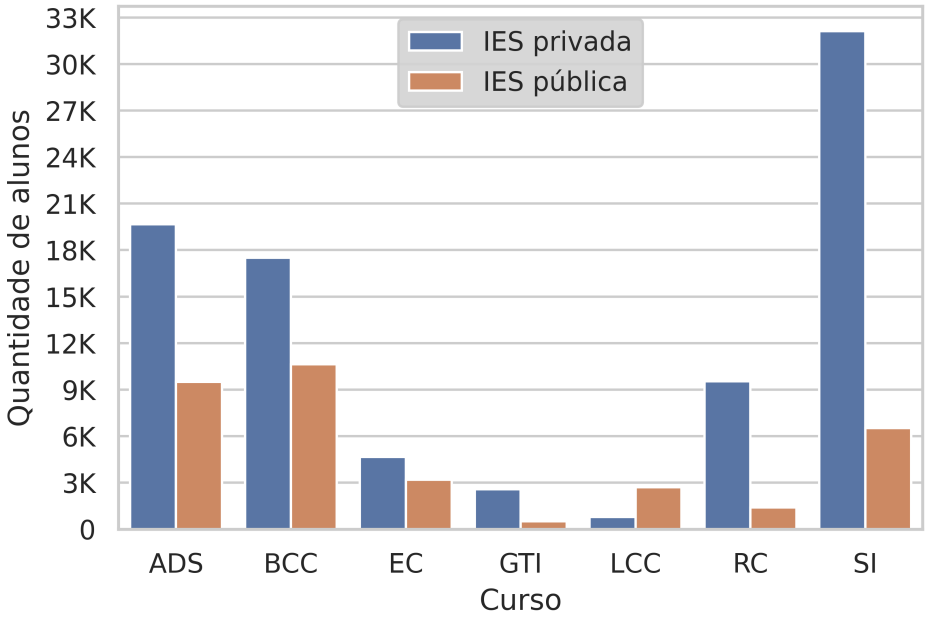
Fonte: Autoria própria (2022).

4.3 Instituições de Ensino Superior

No quesito IES, são 86.849 participantes de IES privadas e 34.481 participantes de IES públicas. Também foi descoberto que nas IES privadas há mais alunos do curso SI, enquanto, em IES públicas, há mais alunos do curso BCC.

Analisando a Figura 11, nota-se que praticamente todos os cursos têm mais alunos em IES privadas do que em IES públicas, exceto pelo curso LCC, onde ocorre o oposto.

Figura 11 – Quantidade de alunos por categoria de IES.



Fonte: Autoria própria (2022).

4.4 Turno

De acordo com os dados, a maioria dos participantes faz curso no turno noturno, sendo representado por 73,54% dos participantes.

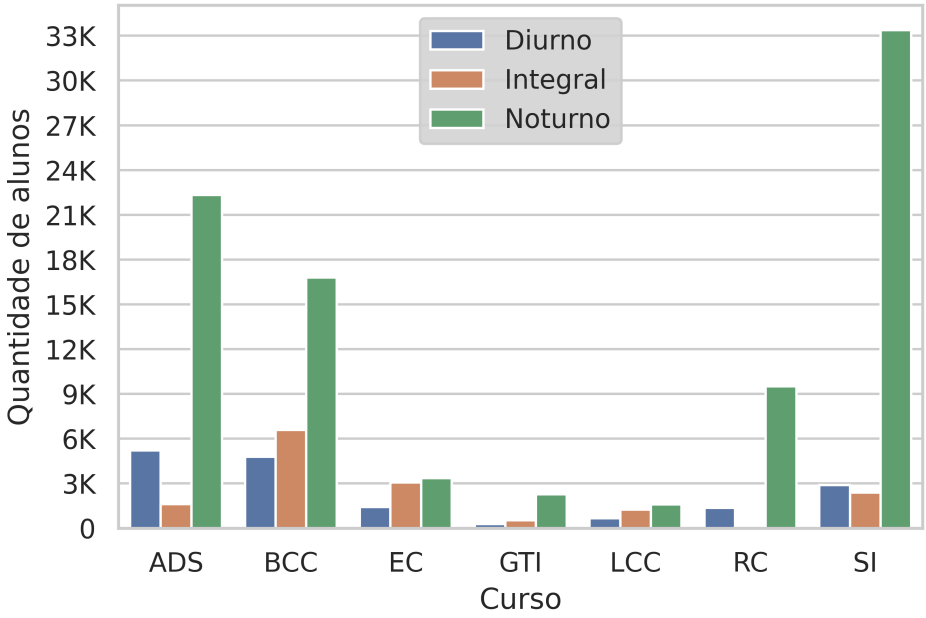
A Figura 12 mostra que todos os cursos têm predominância de alunos no turno noturno. Também é possivel constatar que os cursos EC e LCC possuem um maior equilibrio de participantes por turno. E outra observação feita, é que os cursos GTI e RC praticamente não possuem alunos nos cursos de turno integral.

4.5 Ensino Médio

Constatou-se que a maioria dos participantes estudou o ensino médio em escolas públicas. São 81.936 alunos, contra 39.394 estudantes de escolas privadas. Também foi descoberto que independente do tipo de escola, a maioria dos alunos acaba indo estudar em IES privadas.

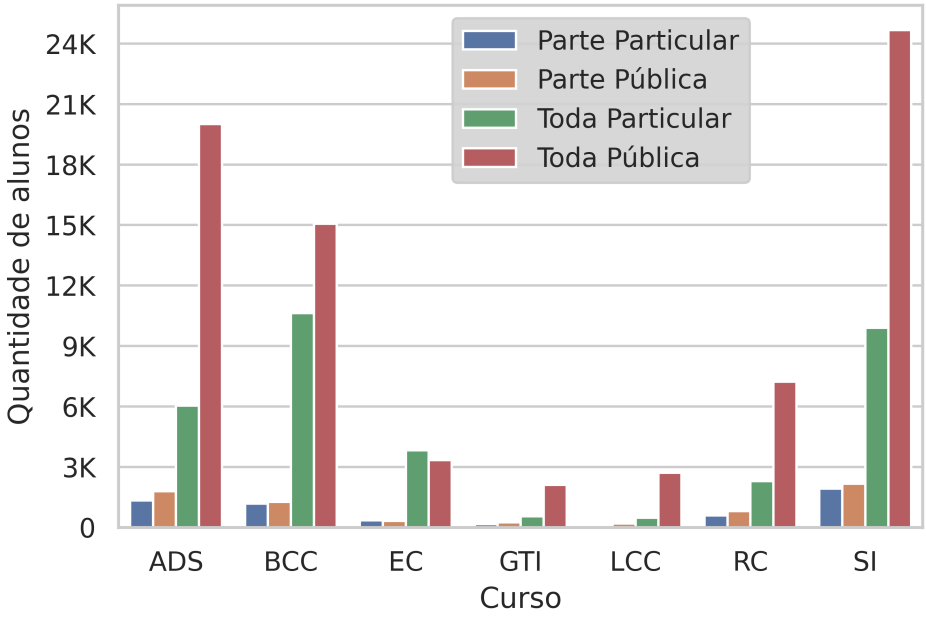
A Figura 13 mostra que, em praticamente todos os cursos, os participantes são em sua maioria advindos de escolas públicas, com exceção ao curso EC, onde além de haver um certo equilibrio, há mais participantes que cursaram o Ensino Médio em escolas particulares.

Figura 12 – Quantidade de alunos por turno de curso.



Fonte: Autoria própria (2022).

Figura 13 – Quantidade de alunos por tipo de Ensino Médio.



Fonte: Autoria própria (2022).

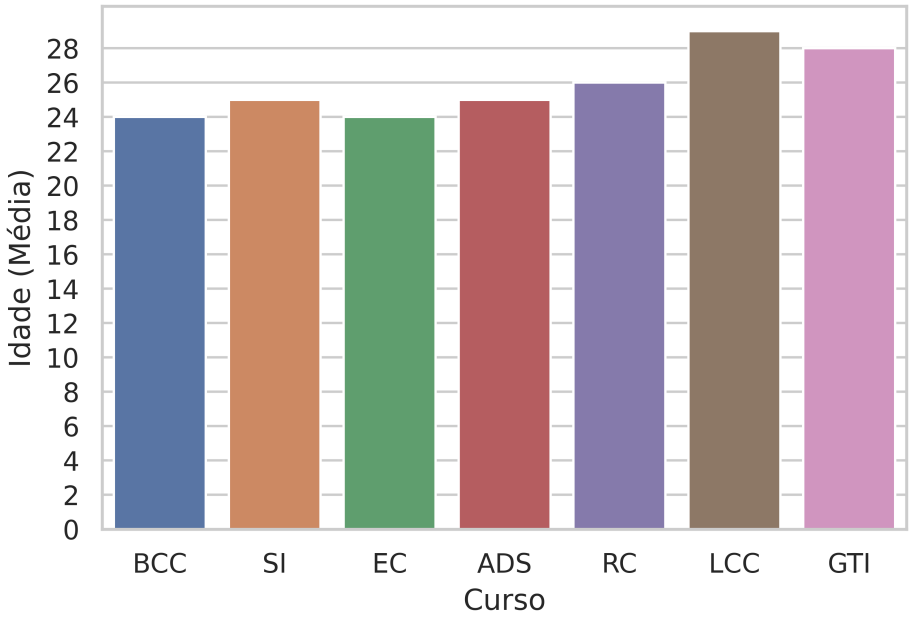
4.6 Idade

A análise dos dados permitiu descobrir que a maioria dos alunos, possuía 22 anos na data de realização do exame, que em média tinham 26 anos e que 52,34% deles tinham entre 22 e 25 anos. Com as variáveis geradas, descobriu-se que a maioria dos participantes começou

seu curso aos 18 anos e também que ingressou em uma IES no ano seguinte de sua formação no Ensino Médio.

A Figura 14 apresenta a idade média em cada área da computação. Analisando o gráfico, percebe-se que, em boa parte dos cursos, os alunos têm em média entre 24 e 26 anos, mas em muitos cursos a média é mais próxima dos 24 anos. Os cursos de LCC e GTI são aqueles que apresentam os alunos mais velhos, com uma média acima dos 27 anos, contribuindo assim para que a média geral (aproximadamente 26 anos) seja superior aos demais cursos.

Figura 14 – Idade média dos alunos por curso.



Fonte: Autoria própria (2022).

4.7 Sexo

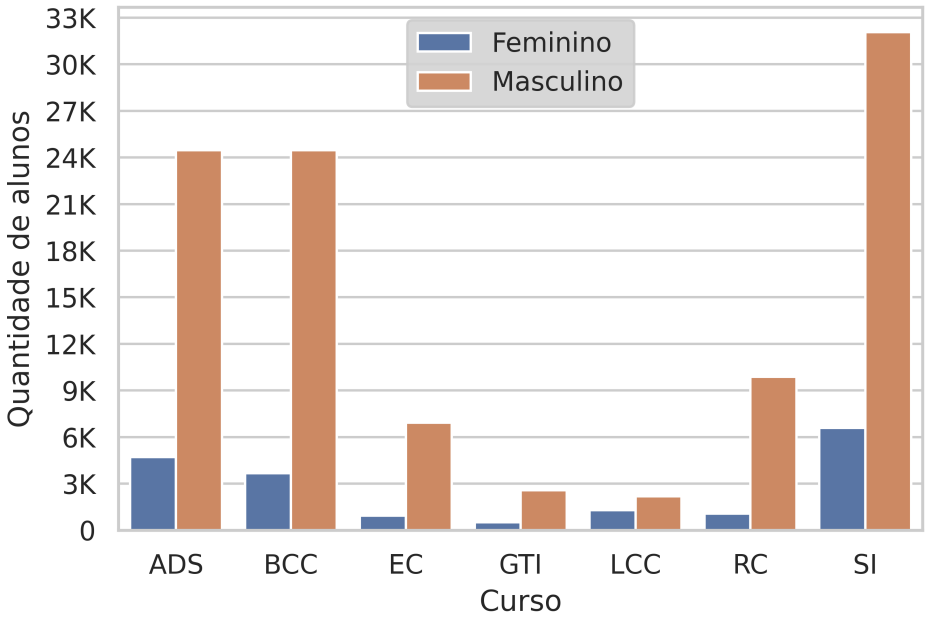
A análise dos dados também evidência a discrepância entre os gêneros, mostrando que o sexo masculino é predominante na área de computação, sendo representado por 84,53% dos participantes. Essa é uma característica de todos cursos analisados, como é observável de acordo com a Figura 15, porém o curso LCC é o que detêm a menor desproporção.

4.8 Cor e raça

A extração de informações também apresentou que a cor/raça branca é disparada a maioria entre os participantes, sendo mais que o dobro de participantes pardos. A Tabela 5 apresenta a quantidade de participantes por cor/raça.

Ao interpretar a Figura 16, observa-se que a cor branca é predominante na maioria dos cursos analisados, porém esse padrão não acontece no curso de LCC. E com base nos

Figura 15 – Quantidade de alunos por sexo.



Fonte: Autoria própria (2022).

Tabela 5 – Quantidade de participantes de acordo com sua cor/raça.

Cor/raça	Quantidade de participantes	Proporção em relação ao todo (%)
Branca	74.697	61,56
Parda	33.935	27,97
Preta	8.842	7,29
Amarela	3.154	2,60
Indígena	702	0,58
Total	121.330	100

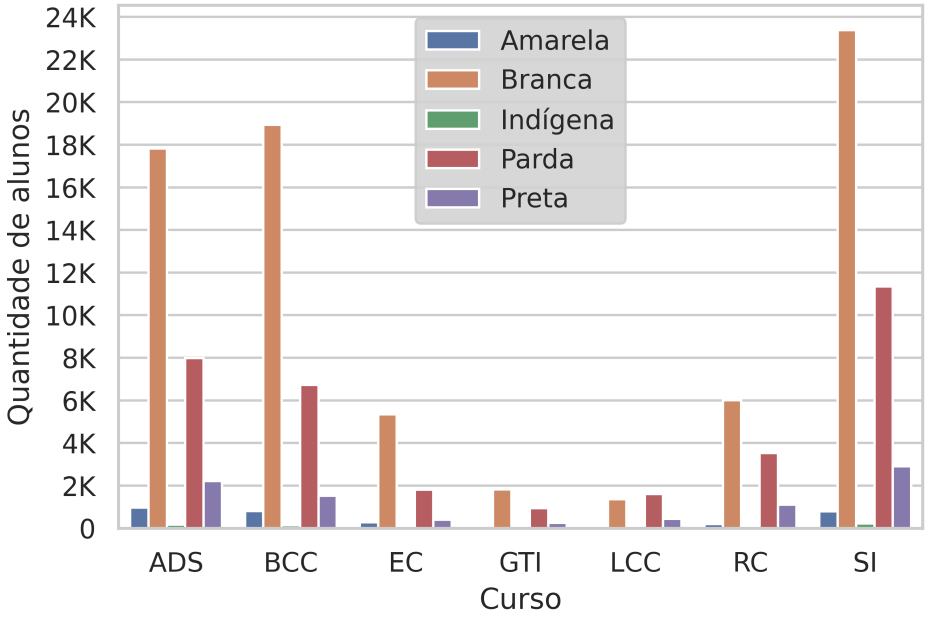
Fonte: Autoria própria (2022).

dados, foi identificado que em quase todos cursos a cor branca representa entre 54% e 67% dos participantes, com exceção a LCC, onde a cor branca representa 39,16% enquanto a cor parda representa 45,95% dos participantes.

4.9 Nível de escolaridade dos pais

No questionário socioeconômico uma das questões é a respeito o nível de escolaridade de seu pai e de sua mãe, variando desde nenhuma escolaridade até a pós-graduação. Nas subseções seguintes serão apresentados as análises referentes a cada uma dessas questões.

Figura 16 – Quantidade de alunos por cor/raça.



Fonte: Autoria própria (2022).

4.9.1 Nível de escolaridade do pai

Dentre os participantes analisados, a maioria deles respondeu que seu pai estudou até o Ensino Fundamental, são 45.747 participantes cujo o pai estudou até o Ensino Fundamental, 44.030 cujo o pai estudou até o Ensino Médio, e apenas 27.375 dos pais concluíram o Ensino Superior, o restante dos pais não possui estudos.

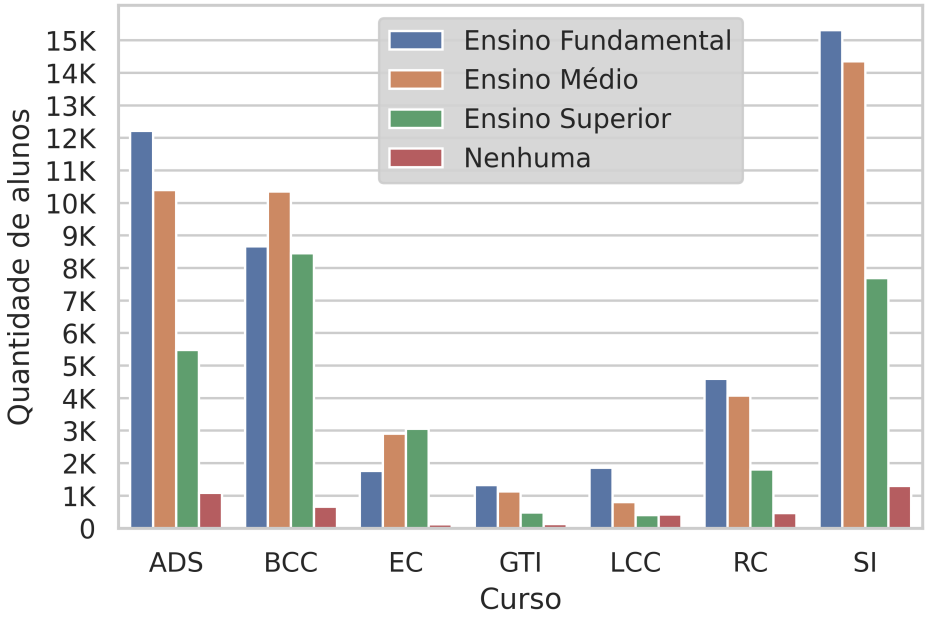
A Figura 17 ilustra o padrão que, em quase todos os cursos, os alunos são geralmente filhos de pais que estudaram até o Ensino Fundamental ou Ensino Médio. A exceção é o curso EC em que a maioria dos participantes tem o pai que estudou até o Ensino Superior.

4.9.2 Nível de escolaridade da mãe

Dando continuidade, também se descobriu que a maioria dos participantes possui mãe que estudou até o Ensino Médio. São 45.807 mães que estudaram até o Ensino Médio, 41.089 que estudaram até o Ensino Fundamental e 31.969 que estudaram até o Ensino Superior, as demais não possuem estudos.

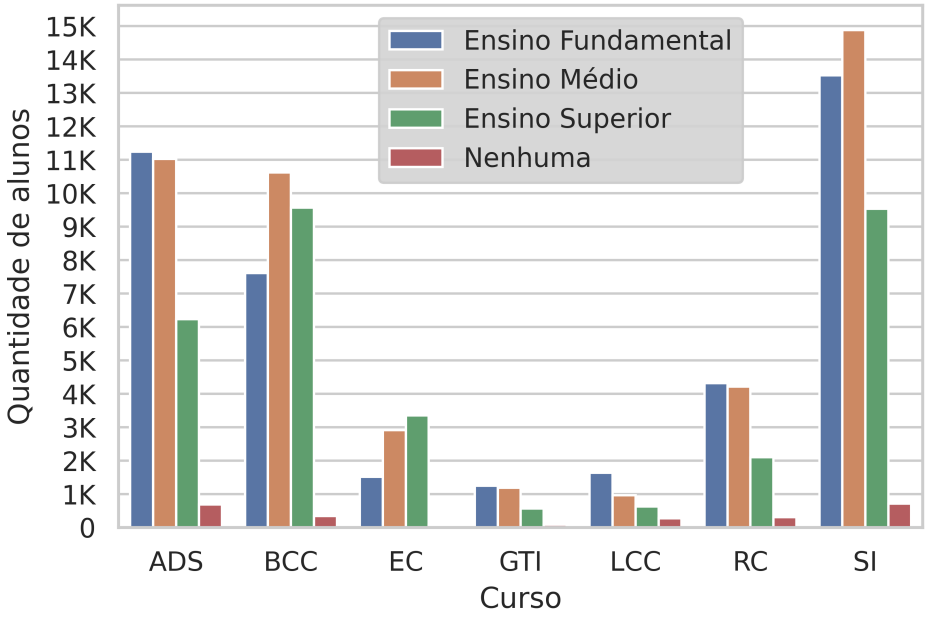
Novamente o padrão dos dados é representado em forma de gráfico, pela Figura 18 observa-se que as mães dos participantes na maioria dos cursos estudaram até Ensino Fundamental ou Ensino Médio, com uma ressalva para os cursos de BCC e EC, na qual a maioria das mães estudaram até Ensino Médio ou Ensino Superior.

Figura 17 – Quantidade de alunos de acordo com a escolaridade do pai.



Fonte: Autoria própria (2022).

Figura 18 – Quantidade de alunos de acordo com a escolaridade da mãe.



Fonte: Autoria própria (2022).

4.10 Renda familiar

Com a análise dos dados foi descoberto que a maioria dos participantes tem uma renda familiar entre 3 e 10 salários mínimos. A Tabela 6 apresenta a quantidade de participantes de acordo com sua renda familiar (em salários mínimos).

Tabela 6 – Quantidade de participantes de acordo com sua renda familiar.

Renda familiar	Quantidade de participantes	Proporção em relação ao todo (%)
Menos que 3	36.779	30,31
De 3 a 10	67.419	55,57
De 10 a 30	15.425	12,71
Mais que 30	1.707	1,41
Total	121.330	100

Fonte: Autoria própria (2022).

Também foi descoberto que participantes que tem a renda inferior a 10 salários mínimos tendem a cursar SI, enquanto participantes com renda superior a 10 salários mínimos tendem a cursar BCC.

Outra informação relevante foi a descoberta da relação entre a escolaridade dos pais e a renda familiar. A Tabela 7 indica que uma maior renda familiar está relacionada com uma maior escolaridade dos pais.

Tabela 7 – Escolaridade dos pais mais comum de acordo com a renda familiar do participante.

Renda familiar	Escolaridade do pai	Escolaridade da mãe
Menos que 3	Ensino Fundamental	Ensino Fundamental
De 3 a 10	Ensino Médio	Ensino Médio
De 10 a 30	Ensino Superior	Ensino Superior
Mais que 30	Ensino Superior	Ensino Superior

Fonte: Autoria própria (2022).

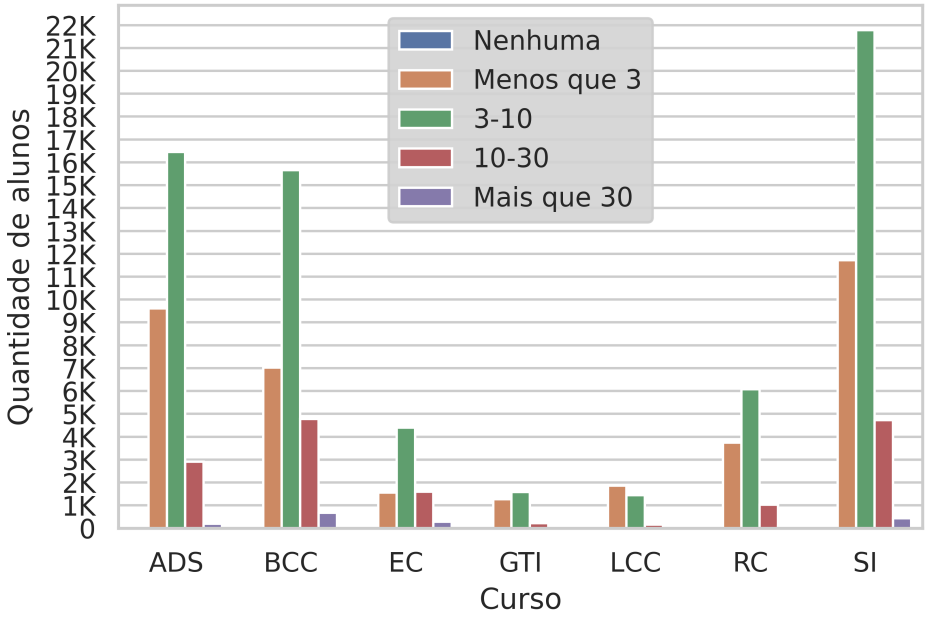
Pela análise da Figura 19, percebe-se que a maioria dos participantes tem renda familiar de 3 a 10 salários mínimos, com uma exceção para concluintes de LCC, onde a maioria deles possuem uma renda menor do que 3 salários mínimos.

4.11 Emprego

Outra questão do questionário socioeconômico diz respeito ao fato do participante estar ou não exercendo alguma forma de trabalho. Entre as possíveis respostas havia: não trabalho, trabalho eventualmente, trabalho de até 20 horas semanais, trabalho de 21 a 39 horas semanais e trabalho com 40 horas semanais ou mais.

De acordo com os dados, 58,40% dos participantes trabalham 40 horas semanais ou mais. Os alunos que trabalham estão principalmente concentrados em IES privadas, enquanto os que não trabalham estão distribuídos de forma balanceada pelas IES. Outra descoberta

Figura 19 – Quantidade de alunos por renda familiar (em salários mínimos).

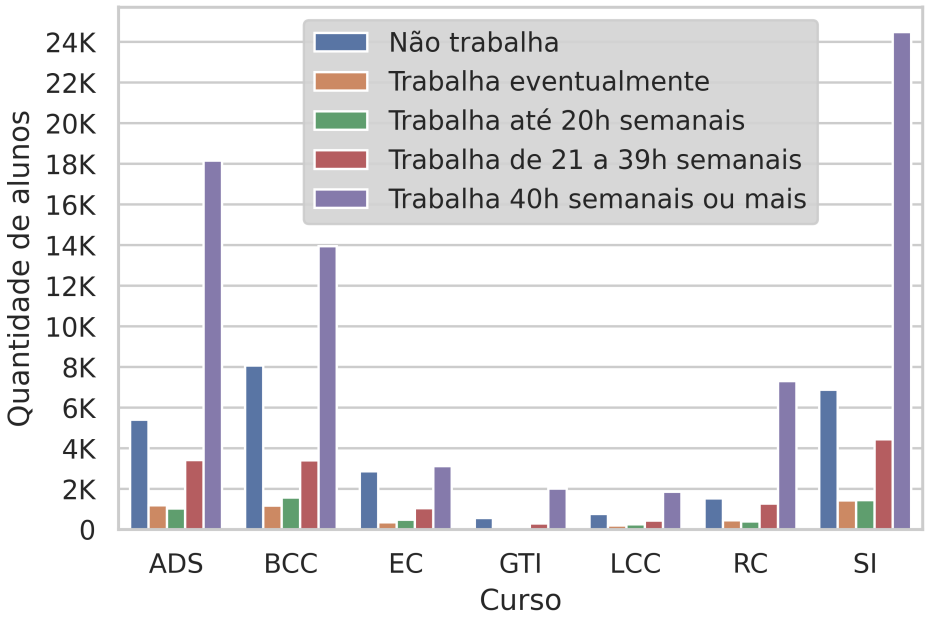


Fonte: Autoria própria (2022).

interessante foi a de que independente de sua condição de trabalho, os participantes possuem principalmente a renda familiar menor que 3 salários mínimos ou de 3 a 10 salários mínimos.

A Figura 20 mostra que em todos cursos a maioria dos participantes trabalha 40 horas semanais ou mais, porém, a segunda maior parcela dos participantes é daqueles que não trabalham.

Figura 20 – Quantidade de alunos trabalhando.



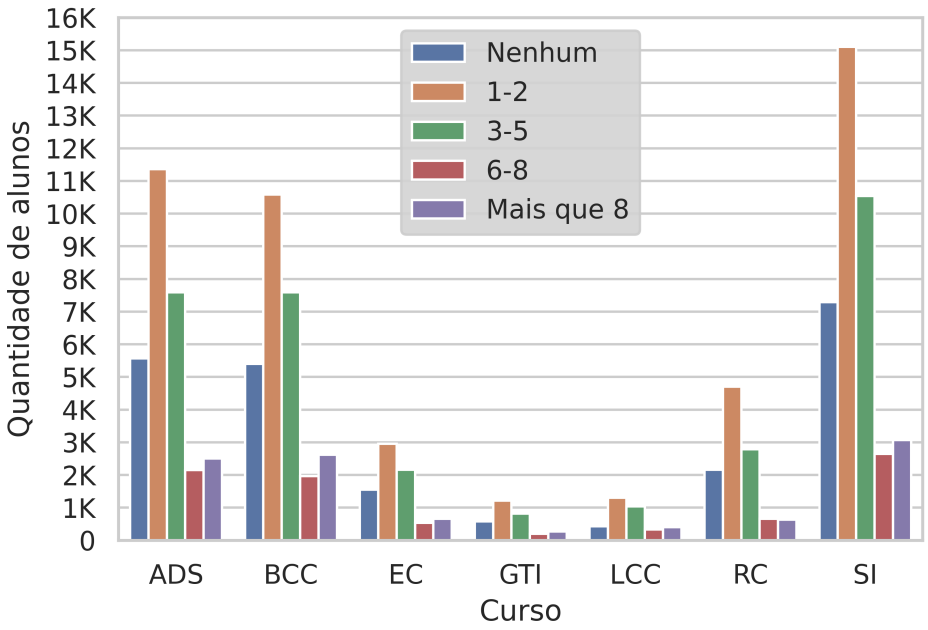
Fonte: Autoria própria (2022).

4.12 Leitura de livros

A análise dos dados também possibilitou descobrir que a maioria dos participantes, 38,91% deles, leu 1 ou 2 livros no último ano, porém 18,92% deles não leram livros.

A Figura 21 apresenta que, em todos os cursos, a maioria dos alunos lê entre 0 e 5 livros, principalmente de 1 a 2 livros, no entanto, são muitos os participantes que não leem livros.

Figura 21 – Quantidade de alunos por quantidade de livros.



Fonte: Autoria própria (2022).

4.13 Horas de estudo

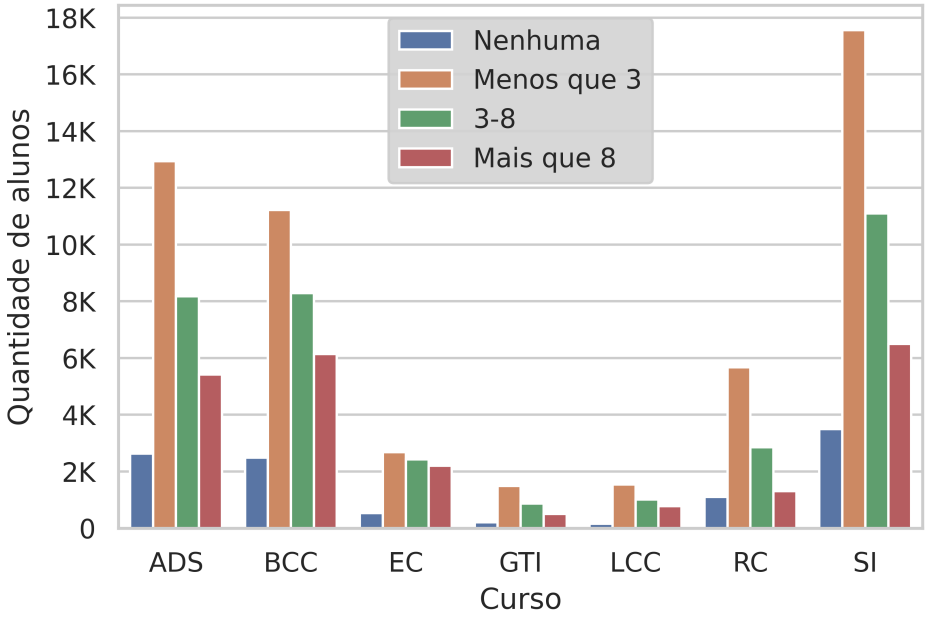
No presente conjunto de dados a maioria dos participantes relatou estudar menos de 3 horas por semanas, além das horas de aula. 53.115 disseram estudar menos de 3 horas, 34.732 estudaram de 3 a 8 horas, 22.866 estudaram mais de 8 horas e 10.617 reportaram não estudar fora do horário de aula.

Visualizando a Figura 22 é possível detectar que, em todos os cursos, a maioria dos alunos estuda menos do que 3 horas por semana, mas que também há uma boa quantia de alunos que estudam de 3 a 8 horas semanais, e que os participantes que não estudam além das horas de aula são a minoria em todos cursos.

4.14 Desempenho

O ENADE é composto por dois tipos de questões:

Figura 22 – Quantidade de alunos por horas de estudo.



Fonte: Autoria própria (2022).

- 10 questões de Formação Geral, elaboradas com base na matriz de referência do Exame Nacional do Ensino Médio (ENEM), que para os cursos da área da computação tratam assuntos de democracia, ética e cidadania (SILVA *et al.*, 2015). Essas 10 questões são compostas por 2 discursivas e 8 objetivas (BRITO, 2015);
- 30 questões de Componente Específico, que avaliam os conhecimentos obtidos durante a graduação, compostas por 3 questões discursivas e 27 objetivas (BRITO, 2015).

Com as respostas dos participantes são calculadas três notas: Nota na Formação Geral, Nota no Componente Específico e Nota Geral. A Nota na Formação Geral é a média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral. A Nota no Componente Específico é a média ponderada da parte objetiva (85%) e discursiva (15%) no componente específico. E a Nota Geral é a média ponderada da formação geral (25%) e componente específico (75%). Ambas notas tem um valor de 0 (zero) a 100. Nesse trabalho o foco está na Nota Geral.

4.14.1 Nota Geral

Em um contexto geral, os participantes tiveram em média uma nota geral de 40,23 pontos, no entanto, mais da metade dos alunos ficaram abaixo da média. Além disso, a nota mais comum foi de 39,4 pontos.

A Tabela 8 mostra os valores associados a nota geral dos participantes de cada curso.

Tabela 8 – Descrição da Nota Geral de cada curso.

Curso	Média	Desvio Padrão	Mediana
ADS	40,47	13,57	39,70
BCC	39,55	14,18	38,50
EC	43,45	14,56	42,70
GTI	45,46	13,38	45,60
LCC	43,59	15,07	43,20
RC	38,53	13,01	37,50
SI	39,65	13,40	38,80

Fonte: Autoria própria (2022).

Já a Tabela 9 apresenta a evolução da nota geral média de cada curso ao longo dos anos. Sendo possível observar que o curso ADS obteve a maior consistência na média ao longo dos anos.

Tabela 9 – Nota Geral média de cada curso por ano.

Curso	2008	2011	2014	2017	Desvio padrão
ADS	42,08	40,47	40,43	40,19	0,86
BCC	36,37	32,97	44,41	43,22	5,48
EC	40,91	36,94	46,0	45,37	4,23
GTI	-	-	-	45,46	0,00
LCC	-	32,55	49,50	40,61	8,47
RC	38,28	37,53	42,45	33,72	3,57
SI	35,45	30,24	43,60	44,66	6,86

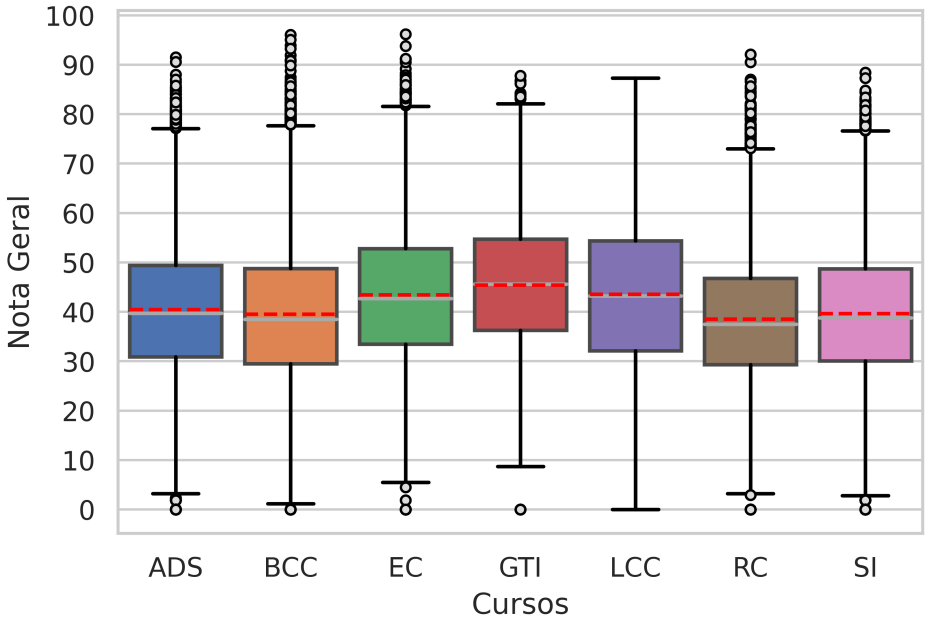
Fonte: Autoria própria (2022).

No gráfico do tipo *boxplot* que será apresentado a partir daqui, os círculos cinzas representam os *outliers*, a linha cinza no meio da caixa indica a mediana, e a linha tracejada vermelha indica a média.

Com a Figura 23 temos uma noção maior a respeito a nota geral dos alunos, obtendo algumas informações além das apresentadas na Tabela 8.

Pela Figura 23 é mais fácil perceber que o curso GTI tem a maior nota geral média, porém isso se deve ao fato de esse curso não estar presente em outras edições, não havendo assim variação na média ao longo das edições. Vemos também que todos os cursos estão com uma média próxima aos 40 pontos e que a média de cada curso vária pouco em relação à mediana. Em praticamente todos os cursos 75% dos participantes atingiram mais de 30 pontos, com exceção ao curso BCC, porém menos de 25% passaram de 50 pontos, com exceção aos cursos EC, GTI e LCC.

Figura 23 – Nota Geral por curso.



Fonte: Autoria própria (2022).

4.15 Considerações Finais

Com base nas informações descobertas, é possível dizer que o perfil do discente da área de computação, participante do ENADE, é de maneira geral uma pessoa que: nasceu na região Sudeste; estuda em IES privada; faz curso no período noturno; fez o Ensino Médio em escola pública; tem entre 22 e 25 anos; é do sexo masculino; tem cor de pele branca; com o pai tendo estudado até o Ensino Fundamental e a mãe estudado até o Ensino Médio; possuí uma renda familiar entre 3 e 10 salários mínimos; já tem um emprego; leu 1 ou 2 livros no último ano; estuda menos de 3 horas semanais, além das horas de aula; e obteve uma Nota Geral de 39,4 pontos. Resultados parciais relacionados à descrição do perfil socioeconômico do discente da área da computação, semelhantes às descobertas de perfil desse trabalho, foram detalhados no trabalho de Capelari e Schwerz (2021).

5 RESULTADOS

Neste capitulo serão apresentados os resultados desse trabalho, abordando os resultados parciais de cada etapa do KDD, começando pelas etapas de seleção e limpeza, passando para a etapa de transformação e por último os resultados da execução da mineração de dados.

Inicialmente, considerando todo conjunto de dados, haviam 1.857.112 participantes, e após realizadas as etapas de seleção e limpeza restaram 121.330 participantes, essa queda abrupta foi devida principalmente ao fato de haver outras áreas de cursos nos dados coletados. A Tabela 10 descreve a distribuição dos participantes ao longo das edições, além da redução no volume dos dados ao longo dos procedimentos realizados. Em relação aos dados apresentados, é importante observar que: (i) houve um aumento no número de alunos submetidos ao exame ao longo dos anos; (ii) em 2008, os alunos ingressantes também eram registrados, embora não realizassem nenhuma prova.

Tabela 10 – Quantidade de participantes e redução no volume de dados após as etapas de seleção e limpeza dos dados.

Edição	Quantidade de participantes	Após seleção	Após limpeza	Redução total
2017	537.436	40.592	38.599	92,82%
2014	481.720	40.888	40.564	91,58%
2011	376.180	33.927	31.589	91,60%
2008	461.776	54.982	10.578	97,71%
Total	**1.857.112**	**170.389**	**121.330**	**93,47%**

Fonte: Autoria própria (2022).

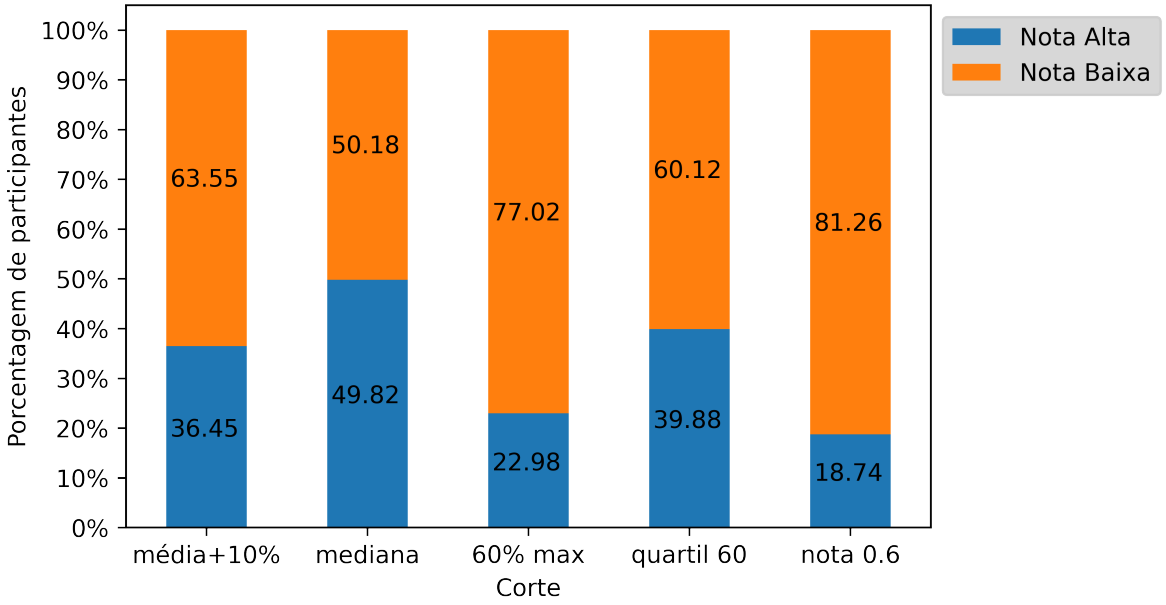
Após feita a seleção e limpeza, os dados foram transformados, realizando a padronização, normalização, discretização e codificação dos mesmos, conforme descrito no Capítulo 3.

A Figura 24 apresenta os dados após a etapa de transformação, com ela pode-se observar as distribuições dos participantes em cada classe de acordo com os pontos de corte, ressaltando o impacto da escolha do limiar no balanceamento dos dados. Analisando o gráfico é possível observar que praticamente todas opções de pontos de corte tornaram os dados desbalanceados, sendo que somente o a divisão com base na mediana tornou dados balanceados.

Depois do pré-processamento dos dados chegou a vez de realizar os experimentos de mineração de dados, testando diversas combinações entre algoritmos e opções de limiar de desempenho.

Para apresentar os resultados da etapa de mineração de dados, foram geradas duas tabelas apresentando os resultados dos algoritmos de aprendizado de máquina de acordo com cada alternativa de ponto de corte. Primeiramente, na Tabela 11 é apresentando a métrica F1-score e, em seguida, na Tabela 12 é a métrica ROC AUC.

Figura 24 – Proporção das classes de acordo com o limiar de corte.



Fonte: Autoria própria.

Tabela 11 – Resultados dos experimentos de acordo com a métrica F1-score.

Classificador	media +10%	mediana	60% nota máxima	quartil 60	nota 0,6
DT	43,76% ± 0,07	54,88% ± 0,32	34,47% ± 0,31	46,9% ± 0,59	30,06% ± 0,44
KNN	40,74% ± 0,23	58,23% ± 0,38	24,29% ± 0,41	45,46% ± 0,16	17,66% ± 0,24
SVM	56,59% ± 0,43	62,34% ± 0,35	49,90% ± 0,36	58,26% ± 0,32	45,39% ± 0,51
RF	37,85% ± 0,65	60,71% ± 0,39	15,75% ± 0,59	43,94% ± 0,63	8,66% ± 0,46
LR	55,44% ± 0,28	61,72% ± 0,32	46,65% ± 0,17	57,35% ± 0,37	42,37% ± 0,19

Fonte: Autoria própria (2022).

Analisando inicialmente a Tabela 11 temos que o SVM demonstrou ser o melhor modelo, já que em todas opções de corte obteve o melhor valor de F1-score. Também destaca-se que o melhor ponto de corte foi a mediana, pois, para todos modelos, resultou no melhor desempenho.

Na Tabela 12 podemos observar que novamente o modelo SVM se saiu melhor, obtendo o melhor desempenho pelo ROC AUC em todos pontos de cortes. Porém, não é possível afirmar qual foi o melhor ponto de corte, pois não há um ponto de corte que seja melhor para todos ou para a maioria dos modelos.

Sendo assim, definimos que o melhor desempenho é a combinação do melhor modelo com o melhor ponto de corte, levando em consideração as métricas F1-score e ROC AUC,

Tabela 12 – Resultados dos experimentos de acordo com a métrica ROC AUC.

Classificador	media +10%	mediana	60% nota máxima	quartil 60	nota 0,6
DT	55,46% ± 0,08	55,12% ± 0,28	57,39% ± 0,24	55,64% ± 0,37	56,74% ± 0,25
KNN	57,38% ± 0,12	57,83% ± 0,25	55,28% ± 0,14	57,84% ± 0,07	53,89% ± 0,08
SVM	64,50% ± 0,23	63,48% ± 0,17	68,48% ± 0,15	64,19% ± 0,12	69,04% ± 0,47
RF	58,93% ± 0,20	62,23% ± 0,11	53,71% ± 0,15	60,31% ± 0,24	51,98% ± 0,12
LR	63,32% ± 0,10	62,38% ± 0,20	65,55% ± 0,18	63,12% ± 0,14	66,21% ± 0,21

Fonte: Autoria própria (2022).

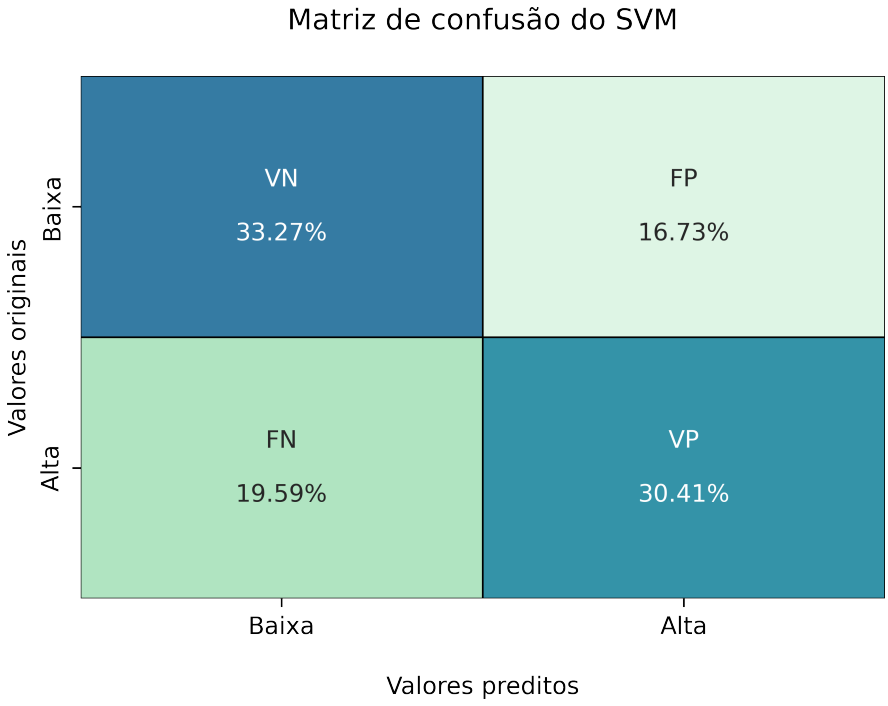
sendo necessário que ambas sejam altas. Essa combinação de métricas foi utilizada porque a métrica ROC AUC indica qual o melhor modelo para predizer corretamente todas classes, enquanto um alto de valor de F1-score indica que há poucos erros relacionados a predição da classe positiva. Portanto, foi estabelecida como melhor combinação para predizer o desempenho dos participantes o modelo SVM utilizando o ponto de corte na mediana da Nota Geral.

A Figura 25 nos mostra como ficou a matriz de confusão do nosso melhor resultado, apresentando em cada quadrante a proporção de predições em relação ao total de dados utilizados, 121.330 participantes. A partir dela, pode-se concluir que o melhor modelo conseguiu acertar e errar as predições de maneira equilibrada, sem ser tendenciosa para alguma das classes e obteve mais acertos do que erros para todas classes.

Por mais que hajam outros trabalhos relacionados a predição do desempenho de participantes do ENADE, e que até possam ter obtidos melhores resultados, há trabalhos que não analisaram vários cursos, como é o caso do trabalho de Rosa *et al.* (2021) que analisou apenas os cursos de BCC, e trabalhos que não avaliaram mais de um modelo de Aprendizado de Máquina, diferentemente desse trabalho. Além disso, os trabalhos não comentavam sobre o desbalanceamento que pode acontecer devido à escolha da forma de separação das classes, nem consideraram a avaliação de métricas como o F1-score e a ROC AUC para casos de dados desbalanceados.

Existem outras técnicas que também podem ser aplicadas nesse estudo, como utilizar um *ensemble* de classificadores, que combina várias instâncias de modelos, visando um melhor desempenho, também seria possível realizar a avaliação de modelos isoladamente para cada edição, ou até mesmo, utilizar dados de uma edição para treinar os modelos e testá-los com os dados das demais edições. Outro método que pode ser aplicado nesse estudo é a variação da quantidade de classes de desempenho, no entanto, com o trabalho de Rezende *et al.* (2022) foi visto que uma alta quantidade de classes resulta em um baixo desempenho.

Figura 25 – Matriz de confusão do SVM com o ponto de corte na mediana.



Fonte: Autoria própria.

6 CONCLUSÃO

Este trabalho consistiu na exploração os microdados do ENADE das edições de 2008, 2011, 2014 e 2017 disponibilizados pelo MEC, com o objetivo de descobrir qual é o melhor modelo de classificação e qual a melhor forma de discretização do desempenho dos discentes de cursos da área da computação participantes do ENADE.

A predição do desempenho dos participantes é importante, pois contribui para o entendimento do potencial dos dados disponibilizados pelo MEC, além de servir de apoio para docentes e Instituições de Ensino Superior na tomada de decisão em relação aos discentes, visando uma melhora no processo de ensino e aprendizagem, que ocasione em melhores desempenhos no ENADE, além de servir como base na escolha de medidas para melhorar a nota de avaliação do MEC de um curso.

Para alcançar os objetivos foi adotado o método KDD, seguindo cada uma de suas etapas: seleção, limpeza, transformação e mineração de dados; e utilizadas como ferramentas a linguagem Python e suas bibliotecas. Foram analisados cinco algoritmos de aprendizagem de máquina mais comumente utilizados: DT, KNN, SVM, RF e LR, em combinação com cinco estratégias de classificação do desempenho dos concluintes: 10% acima da média, 60% da nota máxima, mediana, quartil 60 e nota 0.6.

Com a realização dos experimentos desse trabalho, foi identificado que o algoritmo SVM, em combinação com a mediana como limiar de discretização, apresenta o melhor resultado para a predição do desempenho no ENADE de discentes da computação, obtendo 62,34% e 63,48% de acerto segundo as métricas F1-score e ROC AUC, respectivamente. O uso das métricas F1-score e ROC AUC se fez necessário porque na maioria dos cenários avaliados os dados estavam desbalanceados, e utilizar uma métrica mais conhecida, como a acurácia, é inapropriado para esses casos. Também foi possível notar que o modelo SVM teve um desempenho superior aos outros modelos, independente da métrica analisada e independente da forma de discretização adotada. Além do que, foi percebido que diferentes formas de discretização ocasionam em diferentes desbalanceamentos dos dados e que, de acordo com a métrica F1-score, esse desbalanceamento afeta o desempenho do modelo, podendo gerar uma preferência em predizer a classe majoritária.

A predição do desempenho dos participantes do ENADE é uma tarefa considerada difícil, principalmente em decorrência da artificialidade da criação das classes, pois não há uma forma ideal para realizar a categorização do desempenho dos participantes. Para resolver esse problema, foram analisadas várias formas de discretização, na intenção de descobrir qual a mais adequada para essa situação.

Ao longo do processo do KDD foram encontradas várias dificuldades, como encontrar nos questionários do ENADE perguntas se repetiam em todas edições, algo que para ser resolvido exigiu atenção, analisando em todas edições quais eram as questões presentes e as possíveis alternativas de resposta, verificando se havia compatibilidade entre as diferentes edi-

ções. Outra dificuldade foi realizar a padronização das respostas dessas perguntas, pois, como haviam edições com respostas diferentes de outras edições, era necessário achar uma forma única de representar essas respostas, e para contornar esse problema foi preciso agrupar algumas alternativas de respostas em uma única alternativa.

Existem diversos experimentos que ainda podem complementar os resultados desse trabalho, sendo sugeridos para trabalhos futuros. Como, por exemplo, os dados do ENADE podem ser explorados considerando todos cursos presentes nas edições. Outra opção é analisar mais formas de discretizar os dados, com diferentes pontos de cortes e até mesmo gerar mais classes de desempenho. Também seria possível utilizar um *ensemble* de modelos, combinando diferentes modelos, como, por exemplo, combinar KNN, SVM e LR. Além dessas opções, também é cabível treinar os modelos somente com os dados de uma das edições e predizer utilizando as demais edições, ou gerar um modelo para cada edição e analisar separadamente o desempenho a cada edição.

REFERÊNCIAS

ALVARES, R. V.; CAMPOS, N. d. S.; GOMES, V. B. Adoção de data discovery para apoio ao processo de análise de dados do enade. *In*: **CONGRESSO INTERNACIONAL DE INFORMÁTICA EDUCATIVA**. [*S.l.*: *s.n.*], 2015. v. 20, p. 480–485.

AMORIM, F. S. **Previsão de indícios de fraude em fundos de pensão utilizando modelos de aprendizado de máquina supervisionados e técnicas de balanceamento de dados**. [*S.l.*]: Programa de Pós-Graduação em Administração da Universidade de Brasília, 2021.

ARAÚJO, E. A. T. *et al.* Desempenho acadêmico de discentes do curso de ciências contábeis: Uma análise dos seus fatores determinantes em uma ies privada. **Contabilidade Vista & Revista**, v. 24, n. 1, p. 60–83, 2013.

ARAÚJO, R. A. *et al.* **Análise dos microdados do Enade: Proposta de uma ferramenta de exploração utilizando mineração de dados**. [*S.l.*]: Universidade Federal de Goiás, 2019.

ARAUJO, R. A. E. F. **Variáveis socioeconômicas e desempenho acadêmico**. [*S.l.*]: Universidade Federal de Uberlândia, 2017.

BINUESA, F. **Previsão de mortalidade após cirurgia cardíaca congênita utilizando aprendizagem de máquina**. [*S.l.*]: Centro Universitário FEI, São Bernardo do Campo, 2020.

BRITO, T. F. d. **Corpo Docente: fatores determinantes do desempenho discente no ENADE**. 2015. Tese (Doutorado) — Universidade de São Paulo, 2015.

CALIL, L. A. d. A. *et al.* **Mineração de dados e pós-processamento em padrões descobertos**. 2008.

CAPELARI, L. O. O.; SCHWERZ, A. L. O perfil socioeconômico dos concluintes de computação do sul do brasil. **Anais do Computer on the Beach**, v. 12, p. 133–140, 2021. Acessado em 06 jun. 2022. Disponível em: https://periodicos.univali.br/index.php/acotb/article/view/17392/9884.

CARVALHO, R. B. e Seiji Isotani e A. Mineração de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011. ISSN 2317-6121. Disponível em: http://br-ie.org/pub/index.php/rbie/article/view/1301.

CRISPIM NETO, W. R. **O uso de mineração de dados educacionais sob o ENADE como apoio ao processo de tomada de decisão de gestores do ensino superior**. 2020.

CUNHA, J. P. Z. **Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos**. 2019. Tese (Doutorado) — Universidade de São Paulo, 2019.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, Mar. 1996. Acessado em 06 jun. 2022. Disponível em: https://ojs.aaai.org/index.php/aimagazine/article/view/1230.

FERREIRA, M. A. *et al.* **Determinantes do desempenho discente no ENADE em cursos de Ciências Contábeis**. [*S.l.*]: Universidade Federal de Uberlândia, 2015.

GALDINO, M. V. **Modelos probabilísticos e não probabilísticos de classificação binária para pacientes com ou sem demência como auxílio na prática clínica em geriatria.** [*S.l.*]: Universidade Estadual Paulista (UNESP), 2020.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados: Em busca da Web do Conhecimento**. [*S.l.*]: Novatec Editora, 2015.

LEMOS, K. C. S.; MIRANDA, G. J. Alto e baixo desempenho no enade: que variáveis explicam? **REVISTA AMBIENTE CONTÁBIL-Universidade Federal do Rio Grande do Norte-ISSN 2176-9036**, v. 7, n. 2, p. 101–118, 2015.

LIMA, M. J. A. **Classificação Automática de Emails**. 2013.

LIMA, P. d. S. N. *et al.* Análise de dados do enade e enem: uma revisão sistemática da literatura. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, SciELO Brasil, v. 24, p. 89–107, 2019.

LIMA, P. d. S. N. *et al.* Análise de conteúdo das provas do enade para os alunos do curso de bacharelado em ciência da computação. **Revista Brasileira de Informática na Educação**, v. 29, p. 385–413, 2021.

MACHADO, P. M. O. **ENADE: Uma análise dos fatores determinantes do rendimento dos discentes de Ciências Contábeis das Universidades Federais do Nordeste.** [*S.l.*]: Universidade Federal do Maranhão, 2019.

MOREIRA, A. M. d. A. **Fatores institucionais e desempenho acadêmico no enade: um estudo sobre os cursos de biologia, engenharia civil, história e pedagogia**. 2010.

NARKHEDE, S. **Understanding AUC - ROC Curve**. 2018. Acessado em 06 jun. 2022. Disponível em: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

NUNES, R. T. M. Cálculo preditivo de classificação das notas do enade utilizando redes neurais artificiais. **Revista de Tecnologia Aplicada**, v. 7, n. 2, 2018.

REMIGIO, M. **Aprendizagem Baseada em Instâncias — KNN**. 2020. Acessado em 06 jun. 2022. Disponível em: https://medium.com/@msremigio/aprendizagem-baseada-em-inst\%C3\%A2ncias-knn-7e2c6f0778bc.

REZENDE, C. C. d. S. *et al.* O impacto de aspectos socioeconômicos no desempenho de estudantes de sistemas de informação no enade. **Revista Brasileira de Informática na Educação**, v. 30, p. 157–181, maio 2022. Disponível em: https://sol.sbc.org.br/journals/index.php/rbie/article/view/2093.

ROCHA, A. L. d. P.; LELES, C. R.; QUEIROZ, M. G. Fatores associados ao desempenho acadêmico de estudantes de nutrição no enade. **Revista brasileira de Estudos pedagógicos**, SciELO Brasil, v. 99, p. 74–94, 2018.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 40, n. 6, p. 601–618, 2010.

ROSA, E. R. *et al.* Estudo exploratório através de análises longitudinais aplicado à ciência da computação a partir da base de dados do enade. **Revista Brasileira de Informática na Educação**, v. 29, p. 1463–1486, dez. 2021. Disponível em: https://sol.sbc.org.br/journals/index.php/rbie/article/view/2073.

SAKURAI, R. **Decision Tree: Aprendendo a classificar flores do tipo Iris**. 2022. Acessado em 06 jun. 2022. Disponível em: https://www.sakurai.dev.br/classificacao-iris/.

SANTOS, A. C. d. S. G. d.; MENEZES, T. d. P.; HORA, H. R. M. da. Análise do perfil de aluno e egresso de cursos técnicos por meio de data mining: estudo de caso no instituto federal

fluminense. **#Tear: Revista de Educação, Ciência e Tecnologia**, v. 3, n. 1, jun. 2014. Acessado em 06 jun. 2022. Disponível em: https://periodicos.ifrs.edu.br/index.php/tear/article/view/1828.

SCHNEIDER, C. F. **Machine learning aplicado na previsão de resultados de partidas de futebol: um estudo de caso para comparação de diferentes classificadores**. 2018. 32 p.

SILVA, A. F.; HOED, R. M.; SARAIVA, P. F. Análise do desempenho dos alunos de cursos superiores em computação no enade–uma abordagem usando mineração de dados. **WWW/INTERNET 2019**, p. 207, 2019.

SILVA, C. *et al.* Fatores determinantes para o desempenho dos alunos de administração no enade. INPEAU/UFSC, 2015. Acessado em 06 jun. 2022.

SILVA, L. A. S. e L. Fundamentos de mineração de dados educacionais. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, v. 3, n. 1, p. 568, 2015. ISSN 2316-8889. Disponível em: http://br-ie.org/pub/index.php/wcbie/article/view/3281.

SILVA, V. d. O. **Detecção de fraudes na utilização de cartões usando a técnica de regressão logística: uma aplicação com dados desbalanceados**. [*S.l.*]: Universidade Estadual Paulista (Unesp), 2022.

SIMIONATO, D. **Estudo e comparação das técnicas de validação cruzada desenvolvidas para séries temporais**. [*S.l.*]: Universidade Federal de São Carlos, 2022.

SOUZA, J. P. L. de; FERNANDES, D. Y. de S.; DUTRA, J. F. **Predição precoce de problemas de desempenho de estudantes em modalidade de educação online: um estudo de caso no ensino médio integrado**. 2020.

STATSTEST. **Simple Logistic Regression**. 2022. Acessado em 06 jun. 2022. Disponível em: https://www.statstest.com/simple-logistic-regression/.

TERAMACHI, A. G. **Aprendizado de máquina para classificação da desocupação de leitos pós-cirúrgicos: um estudo sobre pacientes cardiopatas congênitos**. [*S.l.*]: Centro Universitário FEI, São Bernardo do Campo, 2020.

TIBCO. **O que é uma floresta aleatória?** 2022. Acessado em 06 jun. 2022. Disponível em: https://www.tibco.com/pt-br/reference-center/what-is-a-random-forest.

TUSSEVANA, M. **Máquina de vetores de suporte (SVM)**. 2022. Acessado em 06 jun. 2022. Disponível em: https://aprenderdatascience.com/maquina-de-vetores-de-suporte-svm/.

VALADARES, J. A. *et al.* Identificação de perfis de comportamento de usuários no ethereum utilizando técnicas de aprendizado de máquina. *In*: SBC. **Anais do IV Workshop em Blockchain: Teoria, Tecnologias e Aplicações**. 2021. p. 60–73. Disponível em: https://sol.sbc.org.br/index.php/wblockchain/article/view/17129.

VERHINE, R. E.; DANTAS, L. M. V.; SOARES, J. F. Do provão ao enade: uma análise comparativa dos exames nacionais utilizados no ensino superior brasileiro. **Ensaio: avaliação e políticas públicas em Educação**, SciELO Brasil, v. 14, p. 291–310, 2006.

VISTA, N. P. B.; FIGUEIRÓ, M. F.; CHICON, P. M. M. Técnicas de mineração de dados aplicadas aos microdados do enade para avaliar o desempenho dos acadêmicos do curso de ciencia da computação no rio grande do sul utilizando o software r. **I Seminário de Pesquisa Científica e Tecnológica**, v. 1, n. 1, 2017.