UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
Sistemas De Informação
Luiz Araujo de Souza

APLICAÇÃO DE CIÊNCIA DE DADOS NO EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES

Diamantina
2021

Luiz Araujo de Souza

APLICAÇÃO DE CIÊNCIA DE DADOS NO EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES

Trabalho de conclusão de curso apresentado ao curso de Sistemas de Informação, como parte dos requisitos exigidos para a obtenção do título de bacharel em Sistemas de Informação.

Orientador: Alessandro Vivas Andrade

Diamantina
2021

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

FOLHA DE APROVAÇÃO

Luiz Araujo de Souza

APLICAÇÃO DE CIÊNCIA DE DADOS NO EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES

Trabalho de Conclusão de Curso apresentado ao Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisitos parcial para conclusão do curso.

Orientador: Prof. Dr. Alessandro Vivas Andrade

Data de aprovação: 16/09/2021

Profª. Dra. Claudia Beatriz Berti

Faculdade de Ciências Exatas - UFVJM

Prof. MSc. Rafael Santin

Faculdade de Ciências Exatas - UFVJM

Documento assinado eletronicamente por **Alessandro Vivas Andrade**, **Servidor**, em 17/09/2021, às 14:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Claudia Beatriz Berti**, **Servidor**, em 17/09/2021, às 16:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Santin**, **Servidor**, em 20/09/2021, às 07:47, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufvjm.edu.br/sei/controlador_externo.php?

acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0462397** e o código CRC **0E275FBE**.

Referência: Processo nº 23086.008168/2021-07 SEI nº 0462397

AGRADECIMENTOS

Agradeço primeiramente a Deus, por abençoar minha jornada durante o curso, perante situações de altos e baixos. Agradeço também a minha família por todo apoio de sempre e principalmente durante a graduação. Aos meus ex-colegas de trabalho da DEAD, onde tive umas das melhores experiências da minha vida. Aos amigos de curso, principalmente do grupo "Os Doidão de Jamantina", onde compartilhamos conhecimento e ótimos momentos desde o início do curso. Por último, mas não menos importante, agradeço imensamente aos professores por todo o conhecimento passado e por mim adquirido.

RESUMO

Ao se pensar em formas de medir a qualidade do ensino superior brasileiro, o ENADE é um dos principais métodos de avaliação com esse intuito. O processo do ENADE coleta dados da instituição, pessoais dos alunos e os seus resultados individuais na avaliação. Foi pensando nesses dados disponibilizados que este trabalho propõe identificar propriedades pertencentes ao aluno e a instituição de ensino que podem impactar de alguma forma no resultado dos estudantes nessa avaliação. Para a execução dessa pesquisa, foram realizadas análises exploratórias sobre algumas das propriedades relacionadas à renda do estudante, procurando identificar se a situação financeira pode ou não impactar na nota média dos alunos. Em outra análise sobre alguns grupos de estudantes, foi utilizado Aprendizagem de Máquina com um algoritmo de classificação, para identificar possíveis propriedades pertencentes ao aluno e a instituição que podem predizer a nota na avaliação do ENADE.

Palavras-chave: Ciência de Dados; ENADE; Análise exploratória; Analise preditiva; Aprendizagem de Máquina; Estatística; Árvore de Decisão

ABSTRACT

When thinking about ways to measure the quality of Brazilian higher education, ENADE is one of the main assessment methods for this purpose. The ENADE process collects data from the institution, students' personal data and also their individual results in the assessment. It was considering these available data that this work proposes to identify properties that can somehow impact the final result of students in this assessment. To carry out this research, exploratory analyzes were carried out on some of the properties related to student income, seeking to identify whether the financial situation may or may not impact the average grade of students. In another analysis of some groups of students, Machine Learning with a classification algorithm was used to identify possible properties belonging to the student and the institution that can predict the grade in the ENADE assessment.

Keywords: Data Science; ENADE; Exploratory analysis; Predictive analysis; Machine learning; Statistic; Decision Tree

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As Instituições de Ensino Superior (IES) procuram apresentar uma melhor qualidade no ensino motivado por vários fatores, como a oportunidade de o estudante escolher dentre os cursos oferecidos tomando base o conceito do ENADE; às instituições privadas podem receber incentivos através de programas do Governo Federal como PROUNI e FIES; já as instituições públicas tem a oportunidade de aumentar a oferta de recursos para atividades acadêmicas e auxílios estudantis.

Para determinar a qualidade do ensino superior brasileiro, o Ministério da Educação (MEC) utiliza de três tipos de avaliações, como a Avaliação de Cursos de Graduação, Avaliação Institucional e o Exame Nacional de Desempenho dos Estudantes (ENADE), em conjunto, esses instrumentos de avaliações apresentam um panorama geral e classificam as instituições e seus cursos em níveis de qualidade de ensino.

Este trabalho tem o objetivo analisar os dados do ENADE 2018, que tem por base mensurar o nível de qualidade dos cursos das IES dispostos em uma classificação de 1 a 5. As classificações de 1 a 2 correspondem a uma nota insatisfatória, 3 é satisfatória e as notas 4 e 5 são satisfatórias com alto nível de qualidade.

Assim como nas demais áreas afetadas pela análise e interpretação de grandes volumes de dados variados, muitas das abordagens educacionais atuais geram cada vez mais dados e também demandam análises precisas com o intuito de melhorar o planejamento das áreas da educação (SILVA *et al.*, 2017). Isso se justifica porque cada vez mais o governo e as instituições de ensino estão automatizando seus processos, e isso é uma ótima oportunidade de aplicar a Ciência de Dados sobre os registros dos sistemas

De acordo com (SILVA *et al.*, 2017), ”A análise de dados educacionais, representa uma área de pesquisa emergente da informática, para o desenvolvimento de métodos que exploram dados oriundos de ambientes educacionais com a finalidade de entender melhor os estudantes e os cenários em que eles aprendem”.

A utilização de Ciência de Dados é potencialmente fundamental para identificar particularidades que estão relacionadas ao desempenho do aluno na avaliação do ENADE, e assim calcular tendências de resultados de acordo com essas particularidades retiradas na base de dados. Com base nisso, os cursos das Instituições de Ensino Superior podem redirecionar seus esforços para que seus alunos tenham um ensino com cada vez mais qualidade.

Obter informações da base de dados do ENADE com milhares de registros e realizar análises sobre esses dados de forma eficaz e eficiente, é perfeitamente possível ao utilizar Ciências de Dados, pois visa a extração de conhecimento, detecção de padrões e a capacidade preditiva em futuros cenários.

1.1 Objetivos Geral

Esse trabalho tem o objetivo de analisar a base de dados do ENADE procurando identificar e prever padrões de propriedades presentes nesta base de dados que podem influenciar no resultado do aluno na prova do ENADE.

1.1.1 Objetivos Específicos

- Utilizar de técnicas e ferramentas de Ciências de Dados para realizar análises sobre os dados disponibilizados pela plataforma do ENADE.
- Argumentar com base em análise de dados, se existe a relação do meio social e suas características que podem impactam o resultado final para alguns grupos de estudantes no ENADE 2018.
- Com análises preditivas, pretende-se conhecer quais propriedades do meio do estudante que possibilitam classificação de seu resultado na avaliação.

2 FUNDAMENTAÇÃO TEÓRICA

De acordo com INEP/MEC (2021) o Exame Nacional de Desempenho dos Estudantes (ENADE) é uma das formas de avaliações pertencentes ao Sistema Nacional de Avaliação da Educação Superior (SINAES), juntamente com os dados do Censo, Conceito Preliminar de Curso (CPC) e Índice Geral de Cursos Avaliados da Instituição (IGC).

De acordo o INEP/MEC (2021), o ENADE tem o objetivo de medir o desempenho dos discentes dos cursos superiores em relação aos conteúdos programáticos nas ementas. Os alunos são avaliados trienalmente de acordo com suas habilidades necessárias sobre formação geral e profissional (conteúdo específico), ligados à realidade de profissionais do Brasil e do mundo.

Segundo Dias, Porto e Nunes (2016), são inscritos no exame estudantes concluintes do ano de avaliação do curso. No histórico escolar do estudante fica registrada a situação de regularidade em relação ao exame do ENADE.

INEP/MEC (2021) deixa claro que de acordo com a Lei (nº. 10.861/2004), para o estudante apto a realizar o exame, concluir seu curso e obter o diploma, é imprescindível que realize a prova. A situação de irregularidade do estudante junto ao ENADE irá ocorrer quando o estudante: 1) Não realizar a prova, e não apresentar atestado ou dispensa oficial; 2) Não preencher o Questionário do Estudante; 3) ter a sua participação na prova desconsiderada por ação indevida.

INEP/MEC (2021) descreveu as categorias de cursos pertencentes a cada ano de aplicação:

- **Ano I:** Cursos de bacharelado nas áreas de conhecimento de Ciências Agrárias, Ciências da Saúde e áreas afins; Cursos de bacharelado nas áreas de conhecimento de Engenharias e Arquitetura e Urbanismo; Cursos Superiores de Tecnologia nas áreas de Ambiente e Saúde, Produção Alimentícia, Recursos Naturais, Militar e Segurança.
- **Ano II:** Cursos de bacharelado nas áreas de conhecimento de Ciências Biológicas; Ciências Exatas e da Terra; Linguística, Letras e Artes e áreas afins; Cursos de licenciatura nas áreas de conhecimento de Ciências da Saúde; Ciências Humanas; Ciências Biológicas; Ciências Exatas e da Terra; Linguística, Letras e Artes; Cursos de bacharelado nas áreas de conhecimento de Ciências Humanas e Ciências da Saúde, com cursos avaliados no âmbito das licenciaturas; Cursos Superiores de Tecnologia nas áreas de Controle e Processos Industriais, Informação e Comunicação, Infraestrutura e Produção Industrial.
- **Ano III:** Cursos de bacharelado nas Áreas de Conhecimento Ciências Sociais Aplicadas e áreas afins; Cursos de bacharelado nas Áreas de Conhecimento Ciências Humanas e áreas afins que não tenham cursos também avaliados no âmbito das licenciaturas; Cursos Superiores de Tecnologia nas áreas de Gestão e Negócios, Apoio Escolar, Hospitalidade e Lazer, Produção Cultural e Design.

O INEP juntamente com o SINAES [1] relacionaram os instrumentos para a avaliação do ENADE, são eles:

- **Avaliação individual** Possui quatro horas de duração, dividida em 40 questões, sendo 10 questões de formação geral (8 questões de múltipla escolha e 2 discursivas), e 30 questões da parte de formação específica da área (27 questões de múltipla escolha e 3 discursivas). A parte de componente geral possui 25% do peso da prova final e componente específico possui 75% da nota final.
- **Questionário do estudante:** destinado a levantar informações que permitam caracterizar o perfil dos estudantes e o contexto de seus processos formativos, relevantes para a compreensão dos resultados no ENADE;
- **Questionário de percepção da avaliação:** destinado a levantar informações que permitam aferir a percepção dos estudantes em relação à prova, auxiliando, também, na compreensão dos resultados dos estudantes no ENADE;
- **Questionário do coordenador de curso:** destinado a levantar informações que permitam caracterizar o perfil do coordenador de curso e o contexto dos processos formativos, auxiliando, também, na compreensão dos resultados dos estudantes no ENADE.

De acordo com INEP/MEC (2021), após a aplicação do exame, O Ministério da Educação disponibiliza a conjunto de dados gerados pelas avaliações e questionários de estudantes de todo o Brasil, com intuito de tornar público essas informações. Os arquivos podem ser encontrados no portal do INEP, onde estão disponíveis para consulta e download todos os dados gerados pelo ENADE de cada ano de aplicação.

2.1 ENADE 2018

No ENADE referente ao ano de 2018, foco das análises desse trabalho, foram avaliados os cursos que conferem diploma de bacharel nas áreas de Administração, Administração Pública, Ciências Contábeis, Ciências Econômicas, Comunicação Social, Jornalismo, Comunicação Social, Publicidade e Propaganda, Design, Direito, Psicologia, Relações Internacionais, Secretariado Executivo, Serviço Social, Teologia e Turismo.

Também foram avaliados os cursos que conferem diploma de tecnólogo nas áreas de Tecnologia em Comércio Exterior, Tecnologia em Design de Interiores, Tecnologia em Design de Moda, Tecnologia em Design Gráfico, Tecnologia em Gastronomia, Tecnologia em Gestão Comercial, Tecnologia em Gestão da Qualidade, Tecnologia em Gestão de Recursos Humanos, Tecnologia em Gestão Financeira, Tecnologia em Gestão Pública, Tecnologia em Logística, Tecnologia em Marketing e Tecnologia em Processos Gerenciais.

[1] Sistema Nacional de Avaliação da Educação Superior (SINAES)

3 CIÊNCIA DE DADOS

Durante a última década, a computação percorreu um caminho novo com o surgimento da Ciência de Dados e o amadurecimento da Aprendizagem de Máquina. É o que relata (CLEAR; PARRISH, 2020), complementando que a Ciência de Dados combina métodos computacionais e estatística para identificar tendências nos dados existentes e gerar novos conhecimentos.

Morettin e Singer (2020), cita que o termo Ciência de Dados é empregado costumeiramente como um termo novo, diferentemente de métodos de análises de dados que os estatísticos lidam a bastante tempo. Esse termo novo foi apresentado por Jeff Wu em 1980, numa palestra na Universidade de Michigan, EUA. Onde Jeff Wu julgou necessário os rótulos, Ciência de Dados Estatísticos ou Ciência de Dados, no lugar de simplesmente estatística, para dar maior notoriedade e visibilidade ao trabalho dos estatísticos desta área.

Sharma (2019) define Ciência de Dados como a utilização de estatística, computação, algoritmos e Aprendizagem de Máquina com o objetivo de descobrir e prever padrões a partir dos dados brutos.

Segundo Curty e Cervantes (2016), hoje está claro que as organizações que conseguem analisar e entender os seus dados, possuem uma importante vantagem competitiva sobre organizações que não aproveitam dessa tendência. Para se adaptar a esse cenário, é preciso que as organizações invistam em Ciência de Dados para lidar com o enorme volume e variedades de dados, com o intuito de traçar as melhores decisões estratégicas, planejar objetivos com maior exatidão e segurança.

É importante deixar claro que a Ciência de Dados pode ser aplicada em diversas áreas, contextos e organizações diferentes com objetivos diversos. Alguns exemplos são a utilização em mecanismos de busca na Internet, sistemas de recomendação, comparadores de preços, análises de fraude e risco, comparação sobre metodologias de ensino e recomendações de tratamento de saúde.

Para entender Ciência de Dados, é necessário diferenciá-la de análise de dados, pois, são termos que podem gerar certa confusão, às vezes sendo interpretados como sinônimos.

Como pode-se observar a figura 1, onde Sharma (2019) descreve Analista de Dados como um profissional capaz de coletar, analisar e interpretar os dados. Já o Cientista de Dados além de realizar essa análise exploratória, também utiliza-se de algoritmos de Aprendizagem de Máquina para fazer previsões longínquas com mais exatidão,

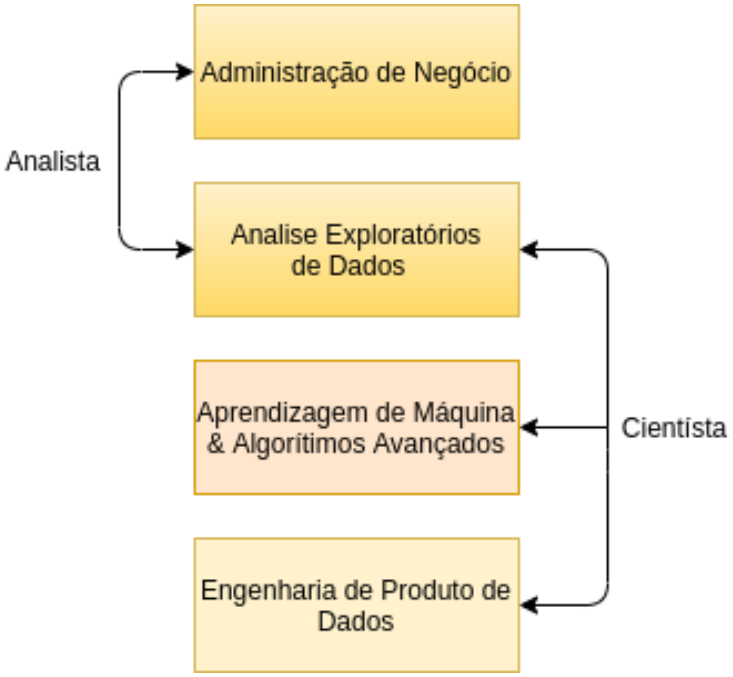

Figura 1 – Análise de dados contra Ciência de Dados
(SHARMA, 2019)

Tão importante quanto a própria análise, Wing (2019) relata que os Cientistas de Dados devem utilizar-se de uma metodologia na aplicação da Ciência de Dados, iniciando pela extração dos dados, remoção de registros inválidos ou vazios, processamento, análises e pôr fim a elaboração de relatórios e gráficos sobre os resultados. É importante o cientista de dados possuir um nível de destreza e percepção para maximizar os retornos em cada fase do processo.

Curty e Cervantes (2016) Argumenta que existem organizações que pretendem armazenar grande quantidade de conteúdo digital para futuras análises, surgindo assim a era da Big Data, essa volumosa quantidade de dados carecem de formas inovadoras e econômicas para serem armazenadas, organizadas e processadas. Como resultado, a análise sobre esses dados auxiliará organizações nas tomadas de decisões dos níveis mais operacionais aos mais estratégicos.

3.1 Big Data no contexto de Ciência de Dados

Loukides (2012) cita que grandes organizações acabam gerando quantidades absurdas de dados constantemente, dentre elas podem se destacar as empresas de tecnologia, telecomunicações e indústrias. Como a demanda por capacidade de armazenamento continua a expandir, o que é considerado grande hoje, certamente será o médio de amanhã e o pequeno da próxima semana.

Chunarkar-Patil e Bhosale (2018) afirma que Big Data é entendido como o grande volume de dados armazenados que variam de gigabytes à petabytes de dados, tornando as técnicas tradicionais de armazenamento, organização e processamento não serem suficientemente ideais. Em tese, os analistas de Big Data necessitam de ferramentas especiais e inovadoras para

armazenar dados de fontes distintas, organizar e transformar em informações que possibilitam uma análise em um curto espaço de tempo.

De acordo com Song e Zhu (2015), a Ciência de Dados está intimamente ligada com Big Data, pois, com a necessidade de analisar esse problema emergente sobre grandes volumes de dados, se fez necessário um novo estudo multidisciplinar, hoje conhecido como Ciência dos Dados, que representa um aglomerado de técnicas e ferramentas, como a infraestrutura de big data, o ciclo de vida analítico de big data, habilidades de gerenciamento de dados e entendimento sobre comportamento organizacional, são conhecimentos necessários para resolver os desafios impostos pela Big Data.

Como pode ser observado pela figura 2, é possível entender como originou a construção da disciplina de Ciência de Dados, a partir da Big Data. Onde os avanços da tecnologia, computação e exploração de dados, aumentou abruptamente a quantidade de dados disponíveis, surgindo o termo Big Data. Tornando necessário a elaboração de infraestrutura de coleta, armazenamento, análises e visualização de dados.

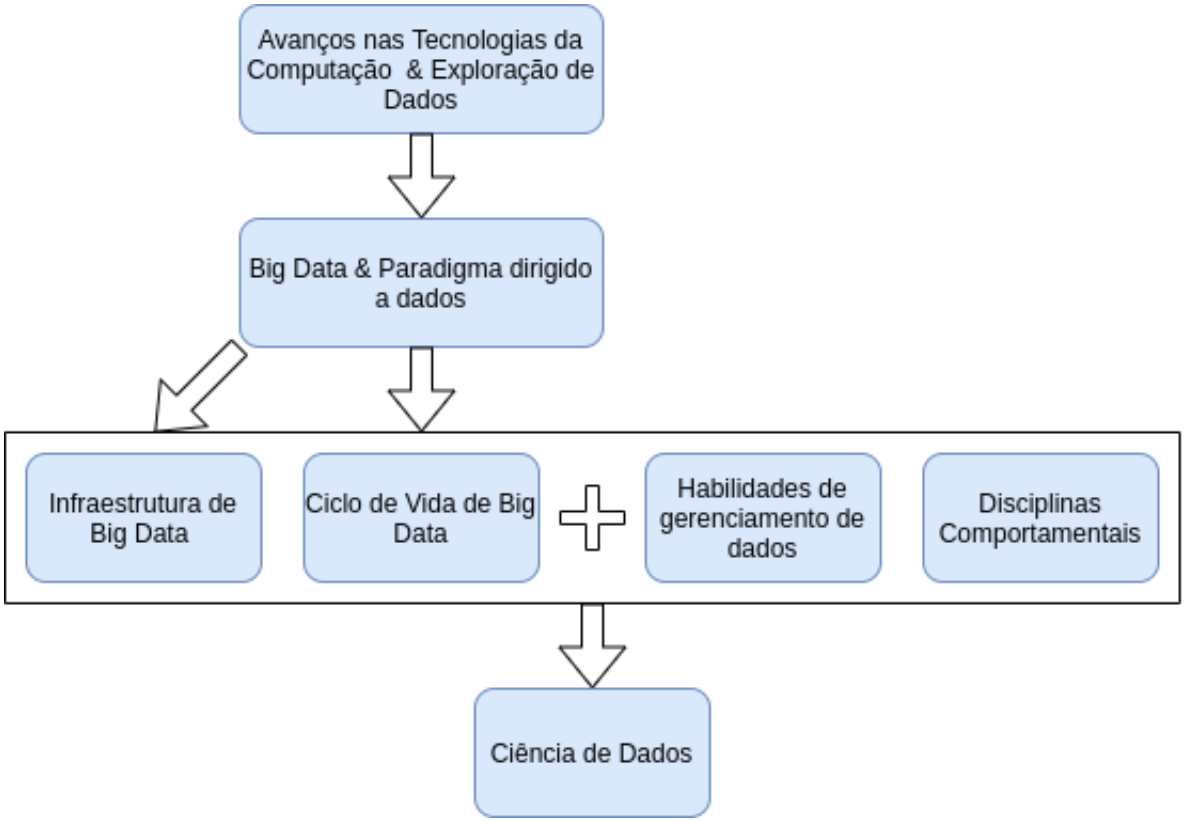

Figura 2 – O Surgimento da Ciência de Dados à partir da BigData

(SONG; ZHU, 2015)

Ciência de Dados se beneficia com o conjunto de técnicas e ferramentas pertencentes as áreas da Estatística, Aprendizagem de Máquina, Mineração de Dados e Computação. Como discutido nesse capítulo, a Ciência de Dados é a combinação de todas essas áreas, com o intuito de aumentar o domínio das análises e predições nos dados.

3.2 Estatística

A maioria dos estatísticos trabalham com dados coletados com perguntas em mente. De acordo com Hand (1998), com a ajuda de subdisciplinas emergentes, como: design de experimentos e design de pesquisa, a coleta de dados ficou mais eficiente, de modo a responder às questões apresentadas.

Enquadrar as perguntas estatisticamente nos permite aproveitar os recursos de dados para extrair conhecimento e melhores respostas, é o que argumenta Dyk *et al.* (2015), complementando que a estatística possibilita aos pesquisadores distinguir entre causalidade e correlação e, assim, identificar intervenções que irão causar mudanças nos resultados. Também lhes permite estabelecer métodos de previsão, estimativa e quantificar seu grau de certeza e incerteza.

Hand (1998) relata que a estatística, ensinada na maioria dos conteúdos estatísticos, pode ser descrita como análise em conjuntos de dados pequenos e limpos, que permitem respostas diretas através de análises intensivas desse conjunto de dados. Essas características não estão erradas, mas a estatística aplicada no contexto de Ciência de Dados, não mais faz análises em pequenos conjuntos de dados, mas sim em enormes conjuntos vindo de fontes diferentes e muitas das vezes não tratadas.

3.3 Computação

Conforme Dijkstra (1972) A computação mudou radicalmente como humanos resolvem problemas, ela transformou o mundo mais do que qualquer outra invenção dos últimos cem anos, e passou a permear quase todos os empreendimentos. A computação pode incluir uma variedade de interpretações, a partir da construção e programação de sistemas de hardware e software para uma ampla gama de finalidades, como processamento, criação de estudos científicos; Aprendizagem de Máquina e meios de comunicação Clear e Parrish (2020)

A computação pode ser definida como processo de implementar programas de computadores, sistematicamente utilizando modelos matemáticos com vários conjuntos de instruções, com o intuito de permitir que o sistema execute uma determinada tarefa, resolva problemas Balanskat e Engelhardt (2015). Compreendendo essa definição, fica claro que a Ciência de Dados está ancorada no conceito de computação, tendo em vista que esta ciência se utiliza de linguagens de programação e Aprendizagem de Máquina para resolver os problemas impostos e possibilitar a realização de análises sobre os dados.

3.4 Aprendizagem de Máquina

De acordo com (CARVALHO, 2015) no contexto de Ciências de Dados, a Aprendizagem de Máquina surge a partir da dificuldade dos humanos em analisar os dados, pois, faltavam especialistas, o custo era elevado e existia subjetividade nas informações. As técnicas de análise de dados permitiam apenas consultas simples, como: ”Quantos produtos foram vendidos hoje?”, mas Não conseguem responder perguntas do tipo: ”Qual novo filme eu gostaria de assistir?”.

De acordo com Morettin e Singer (2020), a Aprendizagem de Máquina antes de 1980 era entendida como um sistema no qual existe uma entrada, uma regra de negócios basadas em cálculos, para no final obter uma resposta. A partir da década de 1990, começou a surgir

problemas mais complexos, onde simples regras de negócios não conseguem mais resolver, logo se fez necessário melhorar a tecnologia com o passar dos anos, surgindo assim a Aprendizagem de Máquina como conhecemos hoje, com capacidade de reconhecimento de imagens, voz, escrita, comportamental, etc.

Existem muitos problemas para os quais não temos uma certa teoria ou algoritmo que resolva, mas podemos ter muitos dados relacionados a esse problema, assim necessitando identificar as informações que os dados podem transmitir. Existe a possibilidade de não ser possível conseguir identificar isso completamente, podendo não perto de uma boa solução. É o que Alpaydin (2014) explica, mas salienta que ao construir modelos de Aprendizagem de Máquina, podemos obter um boa e útil aproximação de respostas sobre este problema com base nas informações contidas no conjunto de dados.

(LOPES, 2018), cita alguns dos algoritmos de Aprendizagem de Máquina que costumeiramente são aplicados a bases de dados, como a Regressão Linear, Regressão Lógica, Árvore de Decisão, Máquinas de Vetor de Suporte, Baías ingênuas, K-ésimo Vizinho mais Próximo, K-means, Floresta Aleatória, Algoritmos de redução de Dimensionalidade, Gradient Boost e AdaBoost.

A figura 3 separa por áreas os tipos de algoritmos. Alpaydin (2014) cita que ao utilizar algoritmos de Aprendizagem de Máquina, podemos detectar certos padrões ou regularidades, que auxiliará o cientista de dados a compreender o processo, ou utilizar esses padrões para fazer previsões, assim os acontecimentos no futuro, mesmo que próximo, não será muito diferente das amostra que foram coletadas na passado ou presente.

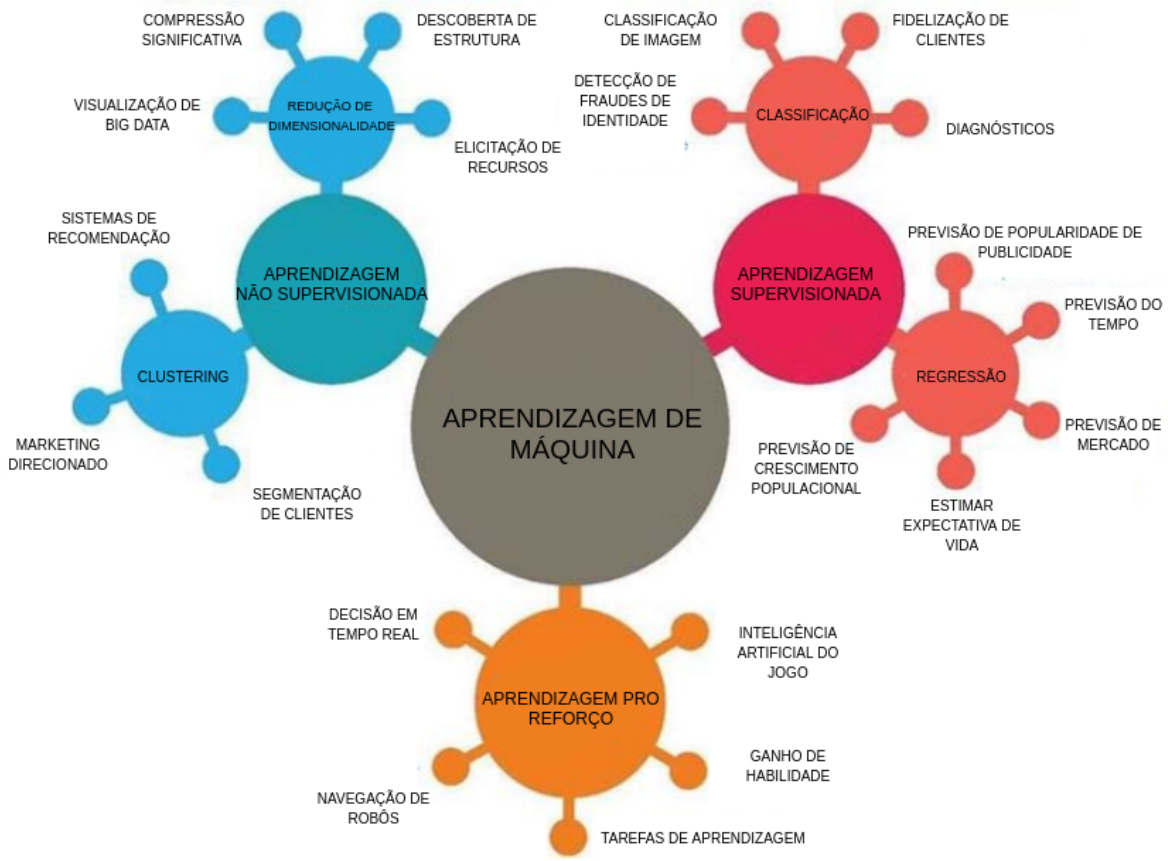

Figura 3 – Áreas de Aprendizagem de Máquina

(LOPES, 2018)

3.5 Mineração de Dados

Alpaydin (2014) cita que a aplicação de métodos de Aprendizagem de Máquina em grandes bancos de dados é chamada de mineração de dados. Fazendo analogia a mineração de matéria-prima que é explorada em minas, e que quando processada apresenta pequena quantidade de material muito precioso. Da mesma forma, na mineração de dados, um grande volume de dados é processado para obter um modelo simples, porém, valioso, com alta precisão preditiva.

Hand (1998) relata que podemos definir mineração de dados como o processo de análise de grandes bancos de dados com o objetivo de encontrar relações insuspeitadas que sejam de interesse ou valor para os proprietários do banco de dados. Com essa característica, a mineração de dados é uma das principais ferramentas da Ciência de Dados, em outras palavras, ela consegue obter a partir de dados brutos, conhecimento que antes não eram claros.

3.6 Etapas do Processo de Ciência de Dados

De acordo com Wing (2019), existe a ideia de ciclo de vida da Ciência de Dados, esse ciclo pode ser dividido em: geração dos dados, criação da coleção, análises e pôr fim a visualização dos resultados. Sharma (2019) reforça essa ideia, pois, segundo ele, é muito importante que as fases do ciclo de vida de Ciência de Dados sejam seguidas para garantir a

qualidade no projeto, pois, um erro comum é adiantar-se na coleta e análise de dados, sem compreender os requisitos, ou mesmo, erroneamente enquadrar o problema do negócio.

Song e Zhu (2015) salienta que o cientista de dados após criar métodos de pesquisa, elaborar perguntas, reunir os dados, organizar hipóteses, construir modelos, visualizar a saída, interpretar os resultados e validá-los, é importante que repita esse ciclo, até que os resultados estejam satisfatórios. De acordo com Stodden (2020), essa metodologia em ciclo de vida não é imutável, e pode ser adaptada ao contexto, isso dependerá do tipo de conjunto de dados.

3.6.1 Geração dos Dados

Cientistas de dados podem ser levados a lidar com uma grande variedade de dados estruturados, semi-estruturados e não estruturados. Dessa forma, o especialista pode gerá-los de diferentes formas. Sobre a diferença na obtenção destes três tipos de estrutura de dados, podemos dizer que.

> Dados estruturados é a maneira mais fácil de organizá-los e analisá-los, porque geralmente estão contidos em linhas e colunas; seguido pelos dados semi-estruturados, que podem ser encontrados em base de dados relacionais como SQL; já os dados não estruturados, são mais complicados de lidar, pois, surgem de vídeos, músicas ou textos. Hoje em dia, apenas 20% dos dados gerados são estruturados, sendo o restante muitas vezes descartados das análises, pois, a falta de conhecimentos técnicos e tecnologia tornou o processo mais difícil, porém, com a proliferação recente de Aprendizagem de Máquina e técnicas mais sofisticadas de análise de dados, dados semi-estruturados e não-estruturados estão ficando mais comuns de estarem presentes em análises (MARR, 2019).

3.6.2 Criação da Coleção dos Dados

Após ter os dados gerados, é preciso armazená-los em uma coleção para facilitar o processo de limpeza e análise. Neste ponto, os dados de diferentes fontes podem ser mesclados em uma única coleção, que costumeiramente é criada em formato de planilha, isto é, os dados são exibidos em linhas e categorizados em colunas.

De acordo com Wing (2019) Nem todos os dados gerados são coletados, porque não precisamos ou não queremos, ou por razões práticas, porque os dados fluem mais rapidamente do que podemos processar.

3.6.3 Limpeza dos Dados

A limpeza de dados, trata da remoção de erros e inconsistências dos dados, de modo a melhorar a sua qualidade. Os problemas de qualidade de dados são motivados geralmente por erros de ortografia durante a entrada de dados, informações ausentes ou outros dados inválidos Rahm e Do (2000).

Quando se fala no processo de limpeza de dados, a tarefa de tratar esses estão faltantes é quase cotidiana na jornada dos cientistas de dados, pois, é um trabalho que precisa ser feito em grandes partes das bases de dados que serão analisadas, para evitar análises estatísticas erradas,

resultando em intervalos de confiança prejudicados e estimativas tendenciosas. De acordo com Soley-Bori (2013) em uma pesquisa, os entrevistados podem não querer revelar algumas informações, uma pergunta pode ser inaplicável ou o participante do estudo simplesmente esqueceu de responder, assim é inevitável que os pesquisadores ao realizarem pesquisas empíricas terão que decidir como os dados ausentes serão tratados.

De acordo com Acaps (2016), é importante implementar estratégias de limpeza de dados para a prevenção de erros. No entanto, as estratégias de prevenção de erros podem reduzir, mas não eliminar alguns erros, eles serão detectados muitas das vezes de forma acidental ao explorar, analisar ou gerar relatórios e visualizações. Dessa forma, Acaps (2016) complementa que esse processo de limpeza de dados envolve ciclos repetidos de triagem, diagnóstico, tratamento e documentação desse processo. Quando detectado erro no dado, deve ser implementada uma atualização na coleta para corrigir e reduzir erros futuros.

3.6.3.1 Formas de lidar com Dados Faltantes

De acordo com (SOLEY-BORI, 2013) Existem diferentes suposições sobre as causas de dados estarem faltantes em bases de dados:

- **Totalmente ausente ao acaso:** o fato de dados estarem ausentes é completamente aleatório. Não contém relação entre quaisquer outros valores no conjunto de dados.
- **Faltando aleatoriamente:** refere-se a dados ausentes condicionalmente em aleatório, porque a ausência é condicional a outra variável. Significa que há uma relação sistemática entre os valores ausentes e outros dados observados. Por exemplo, se as mulheres são menos propensas a dizerem seu peso do que os homens, o peso é um dado faltante aleatório.

3.6.4 Análise Exploratória e Preditiva

Após ter concluído as etapas das subseções anteriores, já é possível realizar as análises, essas que podemos dividir em análise exploratória e análise preditiva, cada uma com suas peculiaridades e objetivos diferentes, mas juntas obtém resultados mais interessantes e concretos.

Analise Exploratória É originária da estatística, essa ciência possui várias formas de medir alguma população ou amostras de dados, como as medidas de tendência central, medidas de variabilidade e de posição, dentre elas, podemos destacar a média aritmética simples, mediana, moda, desvio padrão, percentis, etc.

No contexto de Ciência de Dados, é a análise inicial realizada no conjunto de dados. Utilizamos essa análise para entender como os dados se comportam após aplicação de algumas das técnicas estatísticas. Segundo Perkel (2021) também é possível na análise exploratória organizar os dados em conjuntos, descrever seus comportamentos, resumir as importantes características de cada conjunto observado e compará-las entre dois ou mais conjuntos. Ao se concluir um estudo, os resultados podem ser exibidos em formato de tabela, gráfico e de forma escrita, dependendo da necessidade.

Com uma boa análise exploratória, o cientista de dados pode organizar bons materiais para formular hipóteses e auxiliar na aplicabilidade de alguma técnica da análise preditiva.

Cruz (2015) relata que podemos utilizar também da análise exploratória para encontrar anomalias nos dados, isto é, aqueles registros que se divergem muito da tendência geral, que pode ser até mesmo o resultado do registro ou cadastramento incorreto de valores.

Analise Preditiva é definida por Cruz (2015) como a capacidade de predizer o que pode acontecer com base nos dados obtidos no presente e passado, e de que forma podemos influenciar, antecipar, retardar ou prolongar acontecimentos futuros. No contexto de Big Data, não é possível fazer essas estimativas e predições sem a utilização de algorítimos de Aprendizagem de Máquina, com conhecimentos estatísticos.

Podemos utilizar de vários tipos de algoritmos de Aprendizagem de Máquinas para podermos realizar análises preditivas. Alguns algoritmos são caracterizados como de classificação, agrupamento, regressão, redução de dados e métodos de séries temporais, dentre eles se destacam as Árvores de classificação, agrupamento de k-médias, regressão linear e análise fatorial.

3.6.5 Visualização dos Resultados

Após os processos de análises, o cientista de dados pode expor suas conclusões de forma mais atrativa ao humano com intenção de facilitar a compreensão, já que quem estará consumindo esses resultados pode estar mais preocupado com o que significam e representam, logo é essencial expor essas conclusões de forma compreensível.

Segundo Wing (2019), É muito importante ao apresentar um gráfico, contextualizá-lo com o motivo daquele estudo e legendas que facilitam a compreensão dos resultados. Isso porque dependendo do gráfico, apenas a imagem não é suficiente para alguém interpretá-lo da melhor forma.

3.7 Linguagens Utilizadas

O que potencializou as pesquisas sobre análise de dados, foi a possibilidade de utilização de algoritmos para automatizar, diminuir o tempo de execução das tarefas e ter um poder preditivo maior.

Nesta sessão, foram relacionadas três linguagens de programação diferentes que possibilitam o processo de Ciência de Dados. Foi relacionado duas linguagens comuns ao contexto (Python e R) e uma linguagem relativamente nova (Julia), mas com grande potencial. Vale salientar que a melhor linguagem é aquela que melhor se adapta à base de dados e aos cientistas que a utilizará.

3.7.1 Python

Python é uma linguagem de programação de código aberto, é robusta e de fácil aprendizagem, pois, sua estrutura de dados é escrita em alto nível e segue uma abordagem simples, porém, competente Python (2021). Em outras palavras, codificar utilizando Python aproxima-se da escrita de instruções em linguagem natural.

Unindo a sintaxe descomplicada e seu ambiente interpretado, é considerada uma linguagem ideal para scripts e rápido desenvolvimento de aplicativos em grande parte das áreas. Outra vantagem da linguagem é que sua comunidade de desenvolvedores é enorme. Além de

algoritmos nativos, existem outros milhares de algoritmos desenvolvidas por terceiros disponíveis para uso.

Como é uma linguagem robusta e adaptativa, é muito fácil de encontrar Python em mineração de dados, desenvolvimento da Web, sistemas embarcados, computação científica, Aprendizagem de Máquina e muito mais.

Especificamente sobre a utilização da linguagem para a Ciência de Dados, (HEBBAR, 2019) cita que Python é a principal linguagem de programação para aplicação desta ciência, ela é usada diariamente pelos cientistas de dados, justificado pela grande quantidade de ferramentas e bibliotecas que auxiliam a Ciência de Dados.

Podemos citar algumas bibliotecas que são fortemente utilizadas em problemas de Ciência de Dados, como Pandas, NumPy, SciPy, Scikit-Learn e Matplotlib. Cada uma dessas tem suas peculiaridades, quando usadas em conjunto, fornecem uma gama de ferramentas poderosas para realizar análises e expor os resultados.

3.7.2 R

R foi escrita por estatísticos para ser um ambiente para computação estatística Malik (2020). A linguagem oferece uma variedade de algoritmos que são constantemente usados para inteligência artificial, classificação de dados, clusterização e modelagem linear. Ela integra bem algumas linguagens de programação como C++, Java e SQL.

Conforme R-Project (2021), a linguagem R foi desenvolvida para tratar de computação estatística, ela fornece uma grande variedade de funcionalidades como modelagem linear e não linear, testes estatísticos clássicos, classificações e técnicas gráficas. R é uma linguagem científica de scripts, então desenvolver funcionalidades como sistemas web, não é indicado.

De acordo com MRAN (2021) com o passar dos anos a comunidade da linguagem foi aumentando e atualmente é extremamente ativa e extensa, existem milhares de algorítimos de terceiros implementados, assim aprimorando as funcionalidades do R.

Os cientistas de dados que utilizam R têm em suas mãos grande quantidade de bibliotecas de algorítimos como Prophet, Plotly, Janitor, Caret, Mlr, Lubridate, Ggpolot2, Dplyr, Forcats e Dplyr. Essas bibliotecas oferecem um ótimo suporte para aplicação da Ciência de Dados, fazendo a linguagem R ser uma das preferidas dos profissionais.

3.7.3 Julia

De acordo com Julialang (2021), a linguagem foi projetada pensando na computação científica e numérica, pois, existe a forte demanda por linguagens de grande poder computacional, logo uma linguagem dinâmica e flexível como Julia é apropriada. Em relação a desempenho, ela se compara às linguagens estaticamente tipadas.

Como um dos objetivos da criação da linguagem, ela permite que cientistas e pesquisadores usem a computaçãomais rapidamente, descomplicada e eficaz, a linguagem se baseia nesta ideia pelo fato de facilitar o trabalho de cientistas, para não se preocuparem em aprender uma linguagem de programação difícil. Outra vantagem de utilizá-la, é porque

quando escrita, tem um desempenho aproximado das linguagens como C, possibilitando realizar pesquisas que necessitam de linguagens com alto desempenho.

Inicialmente a linguagem foi criada com o intuito de resolver as deficiências do Python e R, que atualmente são muito usadas para computação científica e processamento de dados Yegulalp (2020).

Segundo Julialang (2021), Julia tem melhor desempenho que Python em técnicas como computação científica, Aprendizagem de Máquina, mineração de dados, álgebra linear em grande escala, computação distribuída e paralela.

Como é uma linguagem recente, ela não possui tantas bibliotecas se comparado com linguagens mais consolidadas, porém, com o pouco que já tem, ela é bastante forte e eficiente, mas em casos específicos Yegulalp (2020) relata que Julia pode chamar bibliotecas Python, C e Fortran, interagindo diretamente com essas linguagens externas.

4 METODOLOGIA

Como fonte de análise deste trabalho, usaremos os conjuntos de dados disponibilizados pelo ENADE 2018, pela grande quantidade de registros e propriedades presentes neste conjunto, será possível analisar aspectos que podem impactar no desempenho do aluno.

Para estruturar o trabalho em etapas, foi adaptado um ciclo de vida em seis etapas para o processo de Ciência de Dados.

1. Pesquisa e entendimento de como funciona o ENADE e como interpretar seus dados, pois, é importante que antes de iniciar um projeto de Ciência de Dados, é preciso entender o problema a ser resolvido.
2. Pesquisas sobre Ciência de Dados, mais especificamente sobre suas aplicações e ferramentas de apoio em análises.
3. Geração dos dados e importação da base de dados em formato reconhecido pela linguagem de programação.
4. Limpeza dos dados e eliminação de registros com dados imprecisos ou ausentes, para evitar erros nos algoritmos e resultados incorretos.
5. Análise exploratória e análise preditiva, primeiro para identificar o que os dados podem nos dizer além da modelagem formal, em seguida utilização de aprendizagem de máquina para realizar projeções sobre resultados dos alunos na avaliação.
6. Visualização dos resultados utilizando elementos visuais como gráficos e relatórios com as informações obtidas após análises. A visualização de dados é uma forma fácil de observar e entender exceções, tendências e padrões nos dados.

Todo esse processo é cíclico, isso quer dizer que após feita a análise e obtido os resultados de uma linha de pesquisa, é possível retornar a etapa inicial para analisar outras linhas de pesquisa.

Inicialmente foi escolhido a utilização da ferramenta Jupyter Notebook com a linguagem de programação Python e suas bibliotecas de algoritmos que auxiliam o processo de Ciência de Dados, essas bibliotecas incluem subprogramas com funções matemáticas, estatísticas e de aprendizagem de máquina.

Na etapa de limpeza da base de dados, optou-se por escolher a abordagem de eliminar os registros que estão com dados faltantes, evitando erros durante as análises. Esses dados se referem a alunos que não responderam ao questionário do estudante ou alunos que não compareceram ao dia da avaliação, ou que compareceram, mas tiveram a participação na prova desconsiderada.

Esses dados ausentes impactam diretamente o lançamento das informações nas propriedades da base, já que grande parte delas estão relacionadas. Logo é preciso que o aluno conclua todo o processo do ENADE para que seu registro fique completamente preenchido.

Complementando a etapa de limpeza e adaptação da base de dados, foi necessário padronizar os dados de algumas propriedades em formato numérico, para possibilitar a utilização

de aprendizagem de máquina e funções pertencentes as bibliotecas de algoritmos do Python, de modo a evitar problemas de execução das análises.

Foi realizado análises exploratórias e análises preditivas de acordo com os objetivos da pesquisa. Os resultados foram apresentados após as análises.

5 DESCRIÇÃO SOBRE A BASE DE DADOS

A base dedados estudada possui 548.127 linhas e 137 colunas/propriedades de dados brutos, isto é, antes de realizar a limpeza e a filtragem dos dados que serão utilizados para a pesquisa.

Como pode ser observado no anexo A, as propriedades da base de dados são organizadas em 9 categorias, são elas:

- Parte 1 - Informações sobre a instituição de ensino superior e do curso.
- Parte 2 - Informações referentes ao estudante.
- Parte 3 - número de questões da parte objetiva, questões válidas e inválidas.
- Parte 4 - Gabaritos do estudante, gabaritos originais e gabaritos finais.
- Parte 5 - índices que apontam a presença ou não do aluno durante todo o processo do ENADE.
- Parte 6 - Situação em relação à realização ou não das questões discursivas
- Parte 7 - Notas nas provas de formação geral e componente específico
- Parte 8 - respostas do aluno no Questionário de Percepção das Provas.
- Parte 9 - respostas sobre o questionário do estudante, com perguntas voltadas na vivência do estudante durante a graduação, desde contextos pessoais a contextos estudantis.

Pela grande quantidade de registros em linhas e de propriedades em colunas, a base de dados possui abundância de informações úteis e relevantes para análises.

Uma dessas fontes de informação é o questionário do estudante, que possui 68 perguntas sobre diversos aspectos da vida pessoal e da graduação dos estudantes. Este é um questionário bem completo e somente dele pode realizar boas análises. A parte relacionada às informações pessoais do estudante pode ser de grande importância para auxiliar e propor análises. Também existe a parte relacionada à percepção da prova, neste caso é avaliado pelo próprio estudante os blocos da avaliação e o tempo de realização da prova.

5.0.1 Processo de Importação dos Dados

Para importar os dados podem ser utilizado a combinação de duas funções. A primeira é uma função nativa da linguagem de programação Python *open()*, que tem a finalidade de carregar a base de dados para um formato reconhecido por essa linguagem de programação; a segunda função utilizada, pertence à biblioteca *Pandas*, é a *read_csv()*, ela tem a finalidade de transformar a base de dados em um formato denominado *DataFrame*, pois, muitas ferramentas de análise operam utilizando esse formato. Sua estrutura assemelha-se muito a estrutura de uma planilha.

```
1 import pandas as pd
2
3 name_path_csv = "DADOS_ENADE_2018/xaa"
4 arquivo = open(name_path_csv)
5 df = pd.read_csv(arquivo, delimiter=";")
6
7 #Trocar a virgula para ponto, como separador de numeros decim is
```

```
df['NT_GER'] = [str(val).replace(',', '.') for val in df['NT_GER']]

#cria a coluna Notas convertendo para numerico
df['Notas']=pd.to_numeric(df.NT_GER,errors='coerce')
```

5.1 Percentual de Valores Faltantes

Para descobrir a porcentagem exata de valores faltantes nas propriedades podemos somar todos os valores faltantes, dividir pelo número de linhas da base de dados, e no fim, multiplicar por 100. Essa estratégia pode ser implementada da seguinte forma.

```
pd.set_option('display.max_columns', None) # or 1000
pd.set_option('display.max_rows', None) # or 1000
pd.set_option('display.max_colwidth', -1) # or
percentual =round(100*(df.isnull().sum()/len(df)),2)
print(percentual)
```

Como é possível ser observado na figura 4, existem algumas propriedades que precisaram ser tratadas devido aos seus dados ausentes. Para isso deve ser estudado estratégias que melhor se adaptam a base de dados para tratar a falta de dados.

Foi identificado um ponto positivo ao gerar esse percentual. A base de dados não possui grande quantidade de dados faltantes por propriedade, pois, a propriedade com o maior percentual possui 26.11% dos dados ausentes. Neste sentido, foi relacionado na tabela 1 as propriedades com a maior porcentagem de dados ausentes. Essa porcentagem variou de 20,31 à 26,11

NU_ANO	0.00
CO_IES	0.00
CO_CATEGAD	0.00
CO_ORGACAD	0.00
CO_GRUPO	0.00
CO_CURSO	0.00
CO_MODALIDADE	0.00
CO_MUNIC_CURSO	0.00
CO_UF_CURSO	0.00
CO_REGIAO_CURSO	0.00
NU_IDADE	0.00
TP_SEXO	0.00
ANO_FIM_EM	0.00
ANO_IN_GRAD	0.00
CO_TURNO_GRADUACAO	0.00
TP_INSCRICAO_ADM	0.00
TP_INSCRICAO	0.00
NU_ITEM_OFG	0.00
NU_ITEM_OFG_Z	0.00
NU_ITEM_OFG_X	0.00
NU_ITEM_OFG_N	0.00
NU_ITEM_OCE	0.00
NU_ITEM_OCE_Z	0.00
NU_ITEM_OCE_X	0.00
NU_ITEM_OCE_N	0.00
DS_VT_GAB_OFG_ORIG	0.00
DS_VT_GAB_OFG_FIN	0.00
DS_VT_GAB_OCE_ORIG	0.00
DS_VT_GAB_OCE_FIN	0.00
DS_VT_ESC_OFG	15.67
DS_VT_ACE_OFG	15.58
DS_VT_ESC_OCE	15.67
DS_VT_ACE_OCE	15.58
TP_PRES	0.00
TP_PR_GER	0.00
TP_PR_OB_FG	0.00
TP_PR_DI_FG	0.00
TP_PR_OB_CE	0.00
TP_PR_DI_CE	0.00
TP_SFG_D1	0.00
TP_SFG_D2	0.00
TP_SCE_D1	0.00
TP_SCE_D2	0.00
TP_SCE_D3	0.00
NT_GER	0.00
NT_FG	15.67
NT_OBJ_FG	15.67
NT_DIS_FG	15.67
NT_FG_D1	15.67
NT_FG_D1_PT	15.67
NT_FG_D1_CT	15.67
NT_FG_D2	15.67
NT_FG_D2_PT	15.67
NT_FG_D2_CT	15.67
NT_CE	15.67
NT_OBJ_CE	15.67
NT_DIS_CE	15.67
NT_CE_D1	15.67
NT_CE_D2	15.67
NT_CE_D3	15.67
CO_RS_I1	15.64
CO_RS_I2	15.64
CO_RS_I3	15.64
CO_RS_I4	15.64
CO_RS_I5	15.64
CO_RS_I6	15.64
CO_RS_I7	15.64
CO_RS_I8	15.64
CO_RS_I9	15.64
QE_I01	10.96
QE_I01	10.96
QE_I02	10.97
QE_I03	10.98
QE_I04	11.02
QE_I05	11.11
QE_I06	11.27
QE_I07	11.39
QE_I08	11.42
QE_I09	11.42
QE_I10	11.42
QE_I11	11.42
QE_I12	11.42
QE_I13	11.42
QE_I14	11.42
QE_I15	11.42
QE_I16	11.42
QE_I17	11.42
QE_I18	11.42
QE_I19	11.42
QE_I20	11.42
QE_I21	11.42
QE_I22	11.42
QE_I23	11.42
QE_I24	11.42
QE_I25	11.42
QE_I26	13.85
QE_I27	11.62
QE_I28	12.51
QE_I29	15.27
QE_I30	20.31
QE_I31	26.11
QE_I32	24.14
QE_I33	16.60
QE_I34	12.74
QE_I35	12.50
QE_I35	12.50
QE_I36	12.42
QE_I37	12.52
QE_I38	12.41
QE_I39	12.32
QE_I40	12.44
QE_I41	12.37
QE_I42	12.49
QE_I43	13.06
QE_I44	13.58
QE_I45	13.86
QE_I46	14.68
QE_I47	14.46
QE_I48	15.71
QE_I49	17.09
QE_I50	18.85
QE_I51	19.07
QE_I52	18.71
QE_I53	17.63
QE_I54	17.60
QE_I55	20.70
QE_I56	24.06
QE_I57	25.92
QE_I58	25.43
QE_I59	20.61
QE_I60	15.69
QE_I61	13.65
QE_I62	13.89
QE_I63	14.27
QE_I64	14.59
QE_I65	15.70
QE_I66	16.31
QE_I67	17.34
QE_I68	17.70
Notas	15.67

Figura 4 – Percentual de dados ausentes

Tabela 1 – Propriedades com mais dados ausentes

Propriedade	Assunto abordado
QE_I30	Se o curso propiciou experiências de aprendizagem inovadoras.
QE_I31	Se o curso contribuiu para o desenvolvimento da consciência ética para o exercício profissional.
QE_I32	Se no curso teve oportunidade de aprender a trabalhar em equipe.
QE_I55	Se as avaliações de aprendizagem realizadas durante o curso foram compatíveis com os conteúdos ou temas trabalhados pelos professores.
QE_I56	Se os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.
QE_I57	Se os professores demonstraram domínio dos conteúdos abordados nas disciplinas.
QE_I58	Se os professores utilizaram tecnologias da informação e comunicação (TIC's) como estratégia de ensino (projetor, multimídia, laboratório de informática, ambiente virtual de aprendizagem).
QE_I59	Se a instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico.

5.2 Limpeza da Base de Dados

Optou-se por manter na base de dados somente os casos dos alunos que realizaram a avaliação e tiveram os resultados válidos em todos os blocos, isso inclui as questões discursivas e objetivas. A justificativa dessa ação se baseou em alunos que em muitas das vezes deixaram de responder a algum bloco, reduzindo consideravelmente sua nota na prova e prejudicando a média do todo.

```
1 # 555: codigo que representa que determinado bloco da avalia  o foi
    respondido de forma valida
2 df = df.query("TP_SFG_D1 == 555 & TP_SFG_D2 == 555 & TP_SCE_D1 == 555 &
    TP_SCE_D2 == 555 & TP_SCE_D3 == 555 & TP_PRES == 555 & TP_PR_GER ==
    555 & TP_PR_OB_FG == 555 & TP_PR_DI_FG == 555 & TP_PR_OB_CE == 555 &
    TP_PR_DI_CE == 555")
```

Após essa limpeza, restaram os registros que realmente interessa estar presentes nas análises. Como comparativo, antes do processo de limpeza, a base de dados possuía 548.127 registros, após o processo esse número passou para 278.402.

6 RESULTADOS

Podemos dividir os resultados desse trabalho em duas sessões, a primeira relacionada aos resultados das análises exploratórias e a segunda aos resultados das análises preditivas.

6.1 Análise Exploratória dos Dados

Como o foco dessa pesquisa está em volta da nota dos alunos na avaliação do ENADE, para conhecer um pouco mais sobre essa propriedade, é possível destacar que a nota pode variar de 0 a 100, e seu cálculo é a média ponderada da nota da formação geral (25%) e de componente específico (75%).

É viável a realização de algumas medidas estatísticas relacionadas às notas dos alunos, com essas medidas capacita conhecer alguns detalhes sobre essa propriedade.

- Maior nota: 93.7
- Menor nota: 1
- Média: 45.23
- Desvio Padrão: 14.18
- Variância: 201.18
- Moda: 41.8
- 1º quartil: 34.8
- 2º quartil ou mediana: 44.7
- 3º quartil: 55.2

Observa-se que a média e mediana das notas dos estudantes estão em 45. Considerando que as notas podem chegar até 100, 45% de acerto em uma prova não pode ser considerado ótimo ou perfeito.

O desvio padrão está relativamente alto, isso quer dizer que as notas tendem a se estabelecerem entre 31 e 58.

Se observamos os quartis, podemos confirmar que, em geral, as notas estão relativamente baixas. De acordo com o 3º quartil existe a probabilidade de 75% das notas não serem maior que 55.20 pontos.

pode-se analisar o gráfico de barras na figura 5 e o gráfico gaussiano equivalente na figura 6, que os alunos estão normalmente distribuídos aproximadamente entre as notas 25 a 50. De acordo com Gordon (2006), ao se agrupar grande quantidade de registros, eles tendem a seguir um caminho natural relacionado a determinado parâmetro. Neste caso as notas.

Código criado para a geração desses dois gráficos das imagens.

```

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm

qnt_por_classe = df['classe_notas2'].value_counts()
print(qnt_por_classe.sort_index())
```

```
qnt_por_classe = qnt_por_classe.sort_index().array

classes = ['0-25','25-50','50-75','75-100']

fig, ax=plt.subplots()
ax.bar(x=classes, height=qnt_por_classe, alpha=0.5)
plt.xlabel('Classes de notas')
plt.ylabel('Quantidade de alunos')
plt.show()

x_axis = np.arange(0, 5, 0.001)
plt.xlabel('Classes de notas')
plt.ylabel('Quantidade de alunos')
plt.plot(x_axis, norm.pdf(x_axis,2.32,0.69))
plt.show()
```

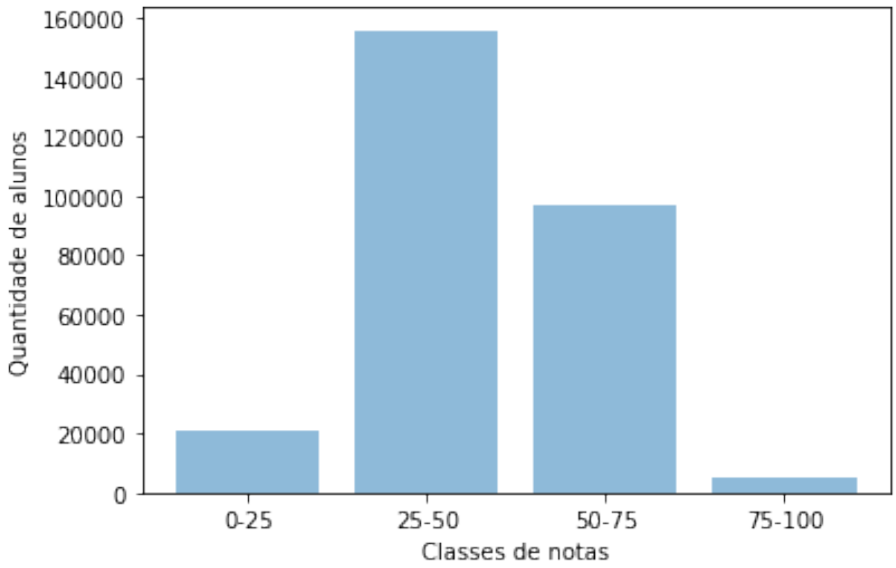

Figura 5 – Quantidade de alunos por grupo de renda familiar

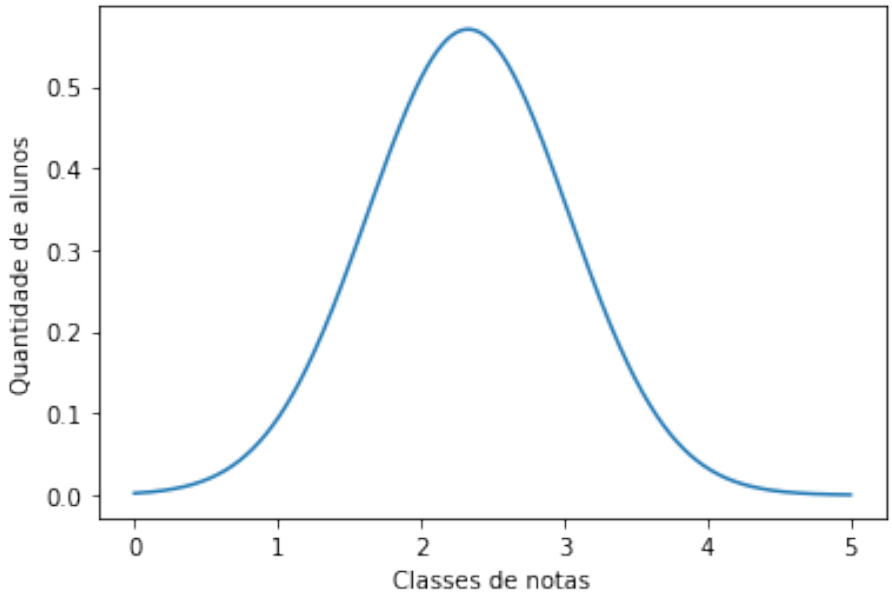

Figura 6 – Curva Normal: quantidade de alunos por grupo de renda familiar

6.1.1 Correlação Entre Propriedades em Relação à Nota

Segundo (ROCHA, 2011), a correlação linear é uma forma de medir o grau de associação entre duas variáveis contínuas, determinando se essas são independentes ou variam juntas. Porém, é importante destacar que esse cálculo determina o nível de correlação e isso não significa que uma variável é a causa raiz da outra.

O resultado da correlação pode variar de -1 à 1.

- Quando correlação é positiva às duas variáveis crescem ou decrescem juntas.
- Quando correlação é negativa uma variável cresce enquanto a outra decresce, ou vice-versa.
- Quando não tem correlação o crescimento das variáveis ou o decrescimento não estão correlacionados

Foi aplicado esse cálculo estatístico na base de dados para determinar o grau de correlação dentre a nota do aluno com as demais propriedades. Na imagem 7 são exibidos os resultados obtidos.

Observa-se que de todas as propriedades do conjunto de dados, nenhuma tem grau de correlação alto. Isso quer significa que possuem maior correlação são: QE_I08 = 0.21; QE_I04 = 0.17; CO_GRUPO = 0.16; QE_I05 = 0.16 e CO_MODALIDADE = -0.14. Logo é demonstrado que essas propriedades de forma individuais não possuem grau de correlação suficientemente alto para poder impactar consideravelmente nas notas.

O algoritmo utilizado para realizar as correlações.

```
#Colunas de notas ou que nao fazem sentido serem analisadas
columns_disctart = df_alunos_presentes[['DS_VT_ACE_OCE','DS_VT_ACE_OFG',
    'DS_VT_ESC_OCE','DS_VT_ESC_OFG','DS_VT_GAB_OCE_FIN',
'DS_VT_GAB_OCE_ORIG', 'DS_VT_GAB_OFG_FIN', 'DS_VT_GAB_OFG_ORIG','NT_CE',
    'NT_CE_D2','NT_CE_D3','NT_DIS_CE',
```

```
'NT_DIS_FG','NT_FG','NT_FG_D1','NT_FG_D1_CT','NT_FG_D1_PT','NT_FG_D2','NT_FG_D2_CT','NT_FG_D2_PT','NT_GER','NT_CE_D1',
'NT_OBJ_CE','NT_OBJ_FG','NU_ITEM_OCE','NU_ITEM_OCE_N','NU_ITEM_OCE_X','NU_ITEM_OCE_Z','NU_ITEM_OFG','NU_ITEM_OFG_N',
'NU_ITEM_OFG_X','NU_ITEM_OFG_Z','Notas','TP_INSCRICAO','TP_INSCRICAO_ADM','TP_PRES','TP_PR_DI_CE','TP_PR_DI_FG',
'TP_PR_GER','TP_PR_OB_CE','TP_PR_OB_FG','TP_SFG_D2','TP_SCE_D1','TP_SCE_D2','NU_ANO','TP_SCE_D3','TP_SFG_D1']]

columns_disctart.columns

columns = list(set(columns) - set(columns_disctart.columns))

dict_correlacao = {}

for nome_coluna in columns:

df_classificar = df_alunos_presentes[[nome_coluna,'Notas']]
df_classificar = df_classificar.dropna(subset=[nome_coluna,'Notas']) # Removendo as linhas cuja alguma celula pertencente a essas coluna est  vazia

eh_inteiro = df_alunos_presentes[nome_coluna].iloc[0]
eh_decimal = df_alunos_presentes[nome_coluna].iloc[0]
#Se a coluna for numerica, n o precisa normalizar
if type(eh_inteiro) != np.int64 and type(eh_decimal) != np.float64:

df_classificar[nome_coluna] = df_classificar[nome_coluna].apply(lambda x: x.replace('A','1').replace('B', '2').replace('C', '3').replace('D', '4').replace('E', '5').replace('G', '6').replace('F', '7').replace('H', '8').replace('I', '9').replace('J', '10').replace('K', '11').replace('E', '5').replace('G', '6').replace('F', '7').replace('H', '8').replace('I', '9').replace('J', '10').replace('K', '11').replace('L', '12').replace('M', '13').replace('.', '12').replace('*', '13'))
df_classificar[nome_coluna] = df_classificar[nome_coluna].astype(int)

correlation = df_classificar.corr()

dict_correlacao[nome_coluna] = round(correlation['Notas'][0],2)

sorted_x = sorted(dict_correlacao.items(), key=lambda kv: kv[1],reverse=True)
sorted_x
```

6.1.2 Análises Relacionadas a Renda

Analisando a propriedade QE_I08 que possui a maior correlação dentre todas as outras propriedades da base de dados em relação à nota. É uma das perguntas do questionário

```
('QE_I08', 0.21),
('QE_I04', 0.17),
('CO_GRUPO', 0.16),
('QE_I05', 0.16),
('QE_I23', 0.13),
('QE_I25', 0.1),
('QE_I14', 0.08),
('TP_SEXO', 0.07),
('CO_ORGACAD', 0.07),
('QE_I17', 0.07),
('QE_I22', 0.07),
('CO_RS_I8', 0.07),
('QE_I15', 0.05),
('ANO_FIM_EM', 0.04),
('QE_I13', 0.04),
('QE_I12', 0.03),
('ANO_IN_GRAD', 0.03),
('QE_I16', 0.02),
('CO_RS_I9', 0.02),
('QE_I45', 0.02),
('QE_I66', 0.02),
('QE_I54', 0.01),

('QE_I36', -0.01),
('QE_I57', -0.01),
('QE_I39', -0.02),
('QE_I46', -0.02),
('CO_UF_CURSO', -0.02),
('QE_I59', -0.02),
('QE_I32', -0.02),
('QE_I47', -0.02),
('QE_I28', -0.02),
('CO_REGIAO_CURSO', -0.02),
('CO_MUNIC_CURSO', -0.02),
('QE_I64', -0.03),
('QE_I65', -0.03),
('QE_I63', -0.03),
('QE_I41', -0.03),
('CO_IES', -0.03),
('QE_I03', -0.03),
('QE_I19', -0.03),
('QE_I18', -0.04),
('QE_I29', -0.04),
('QE_I62', -0.04),
('QE_I68', -0.04),
('CO_RS_I2', -0.04),
('QE_I43', 0.01),
('QE_I33', 0.01),
('QE_I58', -0.0),
('CO_CURSO', 0.0),
('QE_I49', 0.0),
('QE_I55', -0.0),
('CO_RS_I7', 0.0),
('QE_I20', 0.0),
('QE_I35', -0.0),
('QE_I50', -0.0),
('QE_I51', -0.0),
('QE_I06', -0.0),
('QE_I34', -0.0),
('QE_I27', -0.01),
('CO_RS_I3', -0.01),
('QE_I44', -0.01),
('CO_CATEGAD', -0.01),
('QE_I56', -0.01),
('QE_I53', -0.01),
('QE_I31', -0.01),
('QE_I52', -0.01),
('QE_I67', -0.01),

('QE_I42', -0.04),
('QE_I61', -0.04),
('QE_I09', -0.05),
('QE_I60', -0.05),
('QE_I40', -0.05),
('CO_RS_I6', -0.06),
('QE_I37', -0.06),
('QE_I02', -0.06),
('QE_I48', -0.06),
('CO_RS_I5', -0.06),
('QE_I01', -0.06),
('CO_RS_I4', -0.07),
('QE_I38', -0.07),
('QE_I07', -0.07),
('QE_I30', -0.07),
('CO_RS_I1', -0.08),
('NU_IDADE', -0.09),
('QE_I10', -0.09),
('QE_I21', -0.1),
('QE_I11', -0.11),
('QE_I24', -0.11),
('QE_I26', -0.13),
('CO_TURNO_GRADUACAO', -0.13),
('CO_MODALIDADE', -0.14)]
```

Figura 7 – Grau de Correlação Entre Nota e demais Propriedades

do estudante que diz a respeito da renda familiar total do aluno. Pretende-se analisar o quão significante é o impacto da renda familiar, mesmo que as análises indicaram baixa correlação com a nota do aluno no exame do ENADE. Essa propriedade está melhor detalhada na tabela 2.

Tabela 2 – Propriedade QE_I08: Opções de escolha

Opções	Descrição
A	Até 1,5 salário mínimo (até R$ 1.431,00).
B	De 1,5 a 3 salários mínimos (R1.431,01aR 2.862,00).
C	De 3 a 4,5 salários mínimos (R$ 2.862,01 a R$ 4.293,00).
D	De 4,5 a 6 salários mínimos (R$ 4.293,01 a R$ 5.724,00).
E	De 6 a 10 salários mínimos (R$ 5.724,01 a R$ 9.540,00).
F	De 10 a 30 salários mínimos (R$ 9.540,01 a R$ 28.620,00).
G	Acima de 30 salários mínimos (mais de R$ 28.620,00).

Foi elaborado um gráfico de pizza na imagem 8 onde contém a divisão entre as opções de resposta da propriedade QE_108 e a quantidade de alunos pertencentes a cada uma das opções disponibilizadas.

É possível observar que grande parte dos estudantes se encontram em famílias que recebem até 4,5 salários mínimos. Quando se trata dos alunos nos quais a renda familiar ultrapassa 10 salários mínimos, a quantidade de registros cai expressivamente.

Código para obter o gráfico de Pizza da imagem 8:

```
def qtd_alunos_por_categoria(var_media,var_analisada):
#alunos = df[df[var_media] &gt;= 0]
alunos = df
plt.figure(figsize=(10,6))
alunos[var_analisada].value_counts().plot.pie(title=var_analisada)
print(alunos[var_analisada].value_counts())

var_analisada = 'QE_I08'
opcoes_escolhidas=df.loc[:,[var_analisada,"Notas"]]

qtd_alunos_por_categoria('Notas',var_analisada)
```

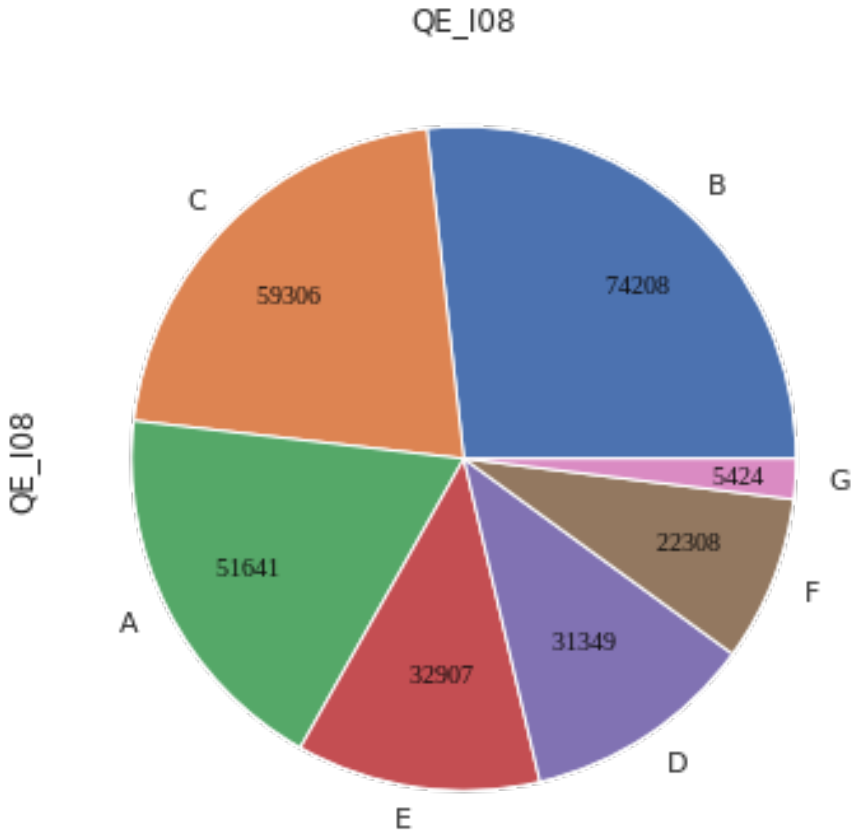

Figura 8 – Quantidade de alunos separados por renda familiar

Foi construído um algoritmo que obtém a nota média dos alunos pertencentes a cada opção de escolha da propriedade QE_I08. O resultado desse algoritmo foi disponibilizado em um gráfico do tipo BoxPlot na imagem 9. Neste gráfico foram ordenadas da esquerda para direita os grupos de famílias dos estudantes que possuem maior nota média.

Apesar do gráfico BoxPlot representar uma média da nota constantemente maior quando o aluno pertence à famílias com maior renda, o desvio padrão dessas notas é alto o suficiente para indicar que essa correlação entre a propriedade QE_I08 e a nota não tem significância estatística.

Código para obter o gráfico de Boxplot da imagem 9:

```
def notas_medias_na_variavel(var_analisada,opcoes_escolhidas):
# Determine the order of boxes
order = opcoes_escolhidas.groupby(by=[var_analisada])["Notas"].mean().
    iloc[::-1].index

# change the graphic size
fig, ax = plt.subplots()
fig.set_size_inches(10,6)

bplot=sns.boxplot(x=var_analisada,y='Notas', data=opcoes_escolhidas,
    order=order, ax=ax)
```

```
bplot.axes.set_title(str(var_analisada)+" versus Nota ",fontsize=16)
bplot.set_xlabel("Op oes",fontsize=14)
bplot.set_ylabel("Notas",fontsize=14)

var_analisada = 'QE_I08'
Media_e_desvio(var_analisada,opcoes_escolhidas)
notas_medias_na_variavel(var_analisada,opcoes_escolhidas)
```

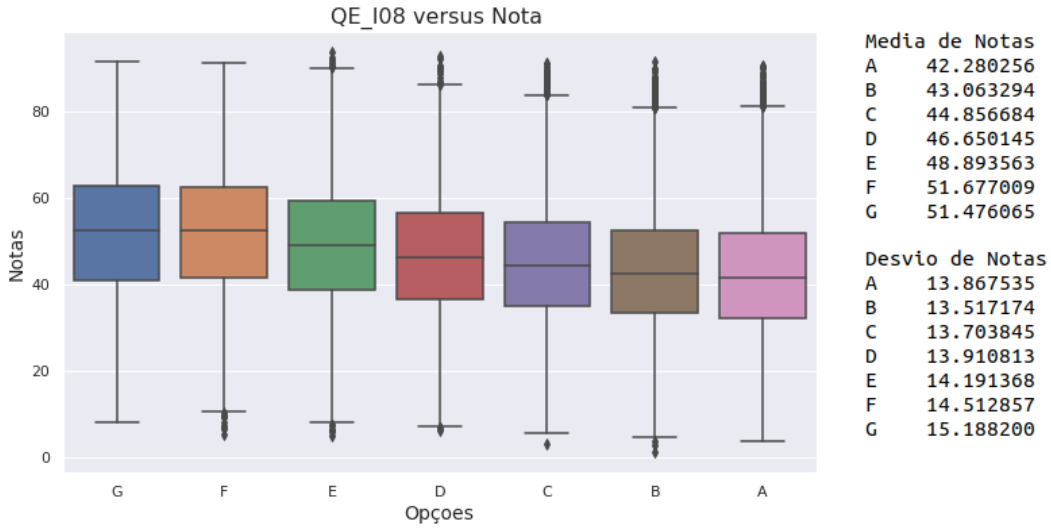

Figura 9 – Nota média dos alunos separados por renda familiar

Aprofundando um pouco mais os estudos exploratórios neste tema social, foi relacionada a propriedade QE_I08 com a QE_I09, essa segunda propriedade diz a respeito da situação financeira individual do estudante. Foi melhor detalhada na tabela 3.

Tabela 3 – Propriedade QE_I09: Opções de escolha

Opções	Descrição
A	Não tenho renda e meus gastos são financiados por programas governamentais
B	Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas
C	Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos
D	Tenho renda e não preciso de ajuda para financiar meus gastos
E	Tenho renda e contribuo com o sustento da família
F	Sou o principal responsável pelo sustento da família

O objetivo dessa relação entre QE_I08 e QE_I09 é identificar se existem variações de respostas entre essas duas propriedades que podem gerar impactos positivos ou negativos na

nota do ENADE. Em outras palavras, queremos saber se um determinado padrão de respostas entre QE_I08 e QE_I09 pode indicar um aumento ou redução da nota média do aluno.

Para relacionar as duas propriedades, foi considerado que cada uma delas fosse tratada como conjunto, sendo suas alternativas os itens pertencentes a esse conjunto. Logo o resultado seria: QE_I08 = {A,B,C,D,E,F,G} e QE_I09 = {A,B,C,D,E,F}.

Foi utilizado uma técnica de conjuntos chamada produto cartesiano que é a multiplicação entre pares ordenados envolvendo conjuntos distintos. Ex: QE_I08 x QE_I09 = (A,A),(A,B),(A,C),(A,D),(A,E),(A,F),(B,A)... .

Algoritmo utilizado para filtrar os dados, separar em conjuntos, fazer o produto cartesiano e expor os resultados de forma que a média da nota fique ordenada da mais baixa para a mais alta, está disponível logo abaixo.

```
1
2  # Opcoes de escolha de cada quest o
3  QE_I08 = ['A','B','C', 'D', 'E', 'F', 'G']
4  QE_I09 = ['A','B','C', 'D', 'E', 'F']
5  array_aux = []
6
7  for i in QE_I08:
8  for j in QE_I09:
9  dados = df[(df['QE_I08'] == i) & (df['QE_I09'] == j)]
10 grade = dados['Notas'].mean()
11 array_aux.append(f'Media das notas: {grade:.2f} - QE_I08 = '+str(i)+' e
       QE_I09 = '+str(j) +". Numero de registros: " + str(len(dados)))
12
13 # Ordenacao dos resultados
14 ordenado = sorted(array_aux)
15 cont = 0
16
17 for i in ordenado:
18 cont+=1;
19 print(str(cont) + ") " + i)
```

1) Media das notas: 37.80 - QE_I08 = G e QE_I09 = A. Numero de registros: 10
2) Media das notas: 39.07 - QE_I08 = A e QE_I09 = E. Numero de registros: 5381
3) Media das notas: 40.06 - QE_I08 = A e QE_I09 = F. Numero de registros: 4607
4) Media das notas: 41.27 - QE_I08 = B e QE_I09 = E. Numero de registros: 17188
5) Media das notas: 41.72 - QE_I08 = A e QE_I09 = D. Numero de registros: 3111
6) Media das notas: 41.98 - QE_I08 = B e QE_I09 = F. Numero de registros: 6669
7) Media das notas: 42.02 - QE_I08 = A e QE_I09 = B. Numero de registros: 14346
8) Media das notas: 42.32 - QE_I08 = B e QE_I09 = D. Numero de registros: 7547
9) Media das notas: 42.42 - QE_I08 = A e QE_I09 = C. Numero de registros: 10801
10) Media das notas: 43.07 - QE_I08 = C e QE_I09 = E. Numero de registros: 14813
11) Media das notas: 43.60 - QE_I08 = C e QE_I09 = D. Numero de registros: 7741
12) Media das notas: 43.95 - QE_I08 = B e QE_I09 = C. Numero de registros: 21368
13) Media das notas: 44.13 - QE_I08 = B e QE_I09 = B. Numero de registros: 16365
14) Media das notas: 44.50 - QE_I08 = B e QE_I09 = A. Numero de registros: 5071
15) Media das notas: 44.64 - QE_I08 = A e QE_I09 = A. Numero de registros: 13395
16) Media das notas: 44.92 - QE_I08 = D e QE_I09 = E. Numero de registros: 6711
17) Media das notas: 44.99 - QE_I08 = C e QE_I09 = F. Numero de registros: 4998
18) Media das notas: 45.58 - QE_I08 = D e QE_I09 = D. Numero de registros: 4737
19) Media das notas: 45.84 - QE_I08 = C e QE_I09 = C. Numero de registros: 17966
20) Media das notas: 45.98 - QE_I08 = C e QE_I09 = A. Numero de registros: 2049
21) Media das notas: 46.10 - QE_I08 = E e QE_I09 = A. Numero de registros: 348ca
22) Media das notas: 46.18 - QE_I08 = C e QE_I09 = B. Numero de registros: 11739
23) Media das notas: 46.38 - QE_I08 = D e QE_I09 = A. Numero de registros: 651
24) Media das notas: 47.09 - QE_I08 = D e QE_I09 = F. Numero de registros: 2750
25) Media das notas: 47.14 - QE_I08 = E e QE_I09 = E. Numero de registros: 6025
26) Media das notas: 47.23 - QE_I08 = G e QE_I09 = D. Numero de registros: 698
27) Media das notas: 47.27 - QE_I08 = E e QE_I09 = D. Numero de registros: 5305
28) Media das notas: 47.55 - QE_I08 = D e QE_I09 = C. Numero de registros: 9752
29) Media das notas: 47.66 - QE_I08 = D e QE_I09 = B. Numero de registros: 6748
30) Media das notas: 47.77 - QE_I08 = F e QE_I09 = A. Numero de registros: 95
31) Media das notas: 49.46 - QE_I08 = E e QE_I09 = B. Numero de registros: 7706
32) Media das notas: 49.60 - QE_I08 = E e QE_I09 = F. Numero de registros: 2833
33) Media das notas: 50.00 - QE_I08 = F e QE_I09 = D. Numero de registros: 3327
34) Media das notas: 50.18 - QE_I08 = E e QE_I09 = C. Numero de registros: 10690
35) Media das notas: 50.61 - QE_I08 = F e QE_I09 = E. Numero de registros: 2713
36) Media das notas: 51.78 - QE_I08 = F e QE_I09 = B. Numero de registros: 6224
37) Media das notas: 51.88 - QE_I08 = G e QE_I09 = E. Numero de registros: 332
38) Media das notas: 51.89 - QE_I08 = G e QE_I09 = C. Numero de registros: 2471
39) Media das notas: 52.09 - QE_I08 = G e QE_I09 = B. Numero de registros: 1691
40) Media das notas: 52.18 - QE_I08 = F e QE_I09 = C. Numero de registros: 8238
41) Media das notas: 54.03 - QE_I08 = F e QE_I09 = F. Numero de registros: 1711
42) Media das notas: 55.56 - QE_I08 = G e QE_I09 = F. Numero de registros: 222

Figura 10 – Resultado da associação entre QE_I08 e QE_I09

Ao relacionar as propriedades QE_I08 e QE_I09 detalhadas nas tabelas 2 e 3. Destacou-se que alunos com renda familiar inferior a 4,5 salários mínimos tendem a ter uma nota média levemente maior quando não exercem nenhuma atividade remunerada.

Propriedade QE_I08 = A

- 2) Media das notas: 39.07 - QE_I08 = A e QE_I09 = E
- 3) Media das notas: 40.06 - QE_I08 = A e QE_I09 = F
- 5) Media das notas: 41.72 - QE_I08 = A e QE_I09 = D
- 7) Media das notas: 42.02 - QE_I08 = A e QE_I09 = B
- 9) Media das notas: 42.42 - QE_I08 = A e QE_I09 = C

- 15) Media das notas: 42.64 - QE_I08 = A e QE_I09 = A
  Propriedade QE_I08 = B
- 4) Media das notas: 41.27 - QE_I08 = B e QE_I09 = E
- 6) Media das notas: 41.98 - QE_I08 = B e QE_I09 = F
- 8) Media das notas: 42.32 - QE_I08 = B e QE_I09 = D
- 12) Media das notas: 43.95 - QE_I08 = B e QE_I09 = C
- 13) Media das notas: 44.13 - QE_I08 = B e QE_I09 = B
- 14) Media das notas: 44.50 - QE_I08 = B e QE_I09 = A
  Propriedade QE_I08 = C
- 10) Media das notas: 43.07 - QE_I08 = C e QE_I09 = E
- 11) Media das notas: 43.60 - QE_I08 = C e QE_I09 = D
- 17) Media das notas: 44.99 - QE_I08 = C e QE_I09 = F
- 19) Media das notas: 45.84 - QE_I08 = C e QE_I09 = C
- 20) Media das notas: 45.98 - QE_I08 = C e QE_I09 = A
- 22) Media das notas: 46.18 - QE_I08 = C e QE_I09 = B

Também é possível observar que ao não exercer atividade remunerada, a nota média do aluno tende a ser levemente maior que a nota média quando somente observado o grupo de renda familiar (QE_I08):

- QE_I08 = A: nota média é 42.28
- QE_I08 = A e QE_I09 = A: nota média é 44.64
- QE_I08 = B: nota média é 43.06
- QE_I08 = B e QE_I09 = A: nota média é 44.50
- QE_I08 = C: nota média é 44.85
- QE_I08 = C e QE_I09 = B: nota média é 46.18

6.2 Análise Preditiva Com Árvore de Decisão.

O algoritmo Árvore de Decisão é classificado como um algorítimo de aprendizagem de máquina supervisionado, isso é, dado uma pergunta ou questionamento, o algoritmo é treinado e ajustado para que retorne uma resposta categorizada mais próximo do real. É possível observar na imagem 11, que esse algoritmo é um conjunto de nós em camadas, onde os nós no interior avaliam os dados de acordo com um teste lógico, já os nós das extremidades inferiores são chamados de nós folhas, neles contém a classe predita (MARQUES, 2016).

Para utilizar o algorítimo Árvore de Decisão, (ALPAYDIN, 2014) afirma que a base de dados precisa ser dividida em um conjunto de treino e um conjunto de teste. No treino, o algoritmo compreende os dados e a resposta que eles produzem, e em qual classe ela pertence, após isso o algoritmo é testado com o segundo conjunto. Uma estratégia comum é separar 80% para treino e 20% para teste. Ao final do teste, obtém-se a acurácia do algoritmo, isto é, seu nível de confiança de predição, de acordo com as propriedades analisadas.

Para compreender melhor, podemos citar um exemplo simples. Queremos saber a faixa etária de uma pessoa, se alimentarmos o algoritmo com propriedades como escolaridade,

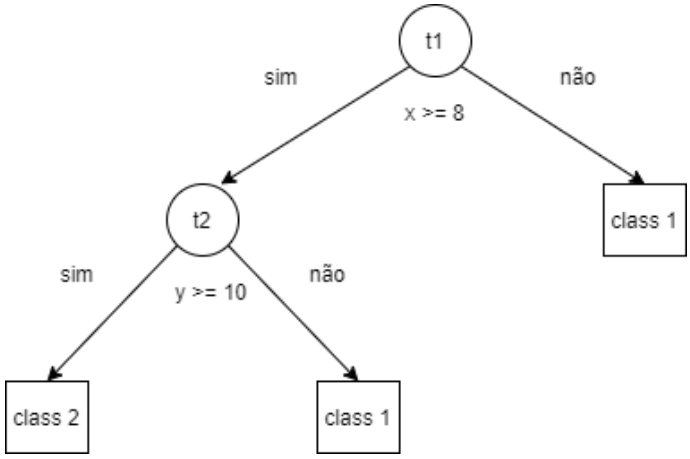

Figura 11 – Exemplo de Árvore de Decisão

estado civil, número de filhos, nome e emprego, dependendo do conteúdo desses dados, o algoritmo poderá predizer qual afaixa etária de uma determinada pessoa. Quanto melhor é o conjunto de propriedades utilizadas no algoritmo, maior é a acurácia.

Para utilizar desse algoritmo, precisamos categorizar a propriedade de observação, isto é, separá-la em classes. Cada classe deve ter aproximadamente o mesmo número de registros que as demais, para que o algoritmo não seja tendencioso em seus cálculos. No contexto desse trabalho, a propriedade escolhida é a nota total do aluno. Algoritmo utilizado para a categorização:

```
from sklearn.utils import shuffle

df_ordenado_por_nota = df_alunos_presentes.query(f"NOME_COLUNA == VALOR"
    ).sort_values("Notas")
print(len(df_ordenado_por_nota))

numero_classes = NUMERO_DE_CLASSES_DESEJADA
valores_em_cada_classe = len(df_ordenado_por_nota) / numero_classes

lista = []
classe_atual = 1
cont = 0
for i in df_ordenado_por_nota.itertuples():

if(cont < valores_em_cada_classe):
lista.append(classe_atual)
cont+=1
else:
cont = 1
classe_atual+=1
lista.append(classe_atual)

df_ordenado_por_nota['classe_notas2'] = lista
```

```
23 print(f"Valores em cada classe:")
24 print(df_ordenado_por_nota['classe_notas2'].value_counts())
25 print("\n")
26
27 for i in range(1,numero_classes+1):
28 classe1 = df_ordenado_por_nota.query(f"classe_notas2 == {i}")
29 print(f"Classe {i}: Notas at  {classe1['Notas'].max()}")
30
31 print("\n------------------")
32
33 df_para_analise = shuffle(df_ordenado_por_nota)
```

Foi optado por separar as notas em três classes. Porque em testes com um número maior de classes, a acurácia do algoritmo reduzia, e apenas duas não é uma quantidade recomentada, visto a variabilidade das notas serem relativamente altas (de 0 à 100).

Para aplicar a Árvore de Decisão é preciso selecionar algumas propriedades(colunas do Dataset) e testar qual a acurácia em relação às notas que essas propriedades resultam. Com o objetivo de melhorar os resultados e automatizar os processos de análise utilizando com Árvore de Decisão, foi implementado um algoritmo complementar em duas etapas.

Algoritimo complementar: 1ª parte:

Foi traçado a estratégia de selecionar as opções das propriedades que melhor se destacaram em um teste de acurácia. Essas opções obtidas devem ser passadas no restante do algoritmo. Exemplo: ao ser analisada a propriedade **CO_GRUPO(Área de enquadramento do curso no Enade)**, foi calculado que alunos pertencentes a opção **ADMINISTRAÇÃO PÚBLICA** são interessantes para serem analisado, logo o grupo dos alunos que cursam ADMINISTRAÇÃO PÚBLICA deverá passar para etapa seguinte do algorítimo.

Um exemplo gráfico do funcionamento dessa parte do algoritmo é demonstrado na figura 12, onde é indicado de verde as propriedades e suas a opções que melhor se destacaram; de amarelo estão indicadas as opções que são regularmente boas de serem analisadas e de vermelho, as opções que tendem a não apresentar bons resultados de acurácia.

Algorítimo complementar: 2ª Parte: Já com os grupos selecionados na 1ª parte, o funcionamento da 2ª parte do algoritmo pode ser observado na figura 13. Essa parte foi implementada para automatizar o processo manual de seleção de propriedades que devem ser testadas na Árvore de Decisão. Sabe-se que é possível realizar essa seleção com menor ou maior quantidade de propriedades, e que esse processo pode ser realizado de forma manual, porém, no contexto deste trabalho, a base de dados possui 91 propriedades aptas a serem selecionadas e testadas, logo é cansativo, improdutivo e pouco eficaz selecionar de propriedades, para no final das análises, poder observar com quais conjuntos de propriedades a acurácia tende a melhorar.

Essa automatização também tem o objetivo filtrar quais propriedades melhoram a acurácia eficientemente e com elevada redução do tempo de análise. Ao final obtém-se o melhor conjunto de propriedades que elevam a possibilidade de predizer a nota do aluno em determinados grupos.

Propriedades	Opções		
P1	Op1	Op2	Op3
P2	Op1	Op2	Op3
P3	Op1	Op2	Op3
P4	Op1	Op2	Op3
P5	Op1	Op2	Op3
P6	Op1	Op2	Op3

Figura 12 – Exemplo de Seleção de Grupo a Ser Analisado (1ª Parte Algorítimo Complementar)

1ª Iteração	
Associação de Propriedades	Acurácia
P1	40,20%
P2	35,82%
P4	**50,50%**
P5	49,30%
P6	31,01%

2ª Iteração	
Associação de Propriedades	Acurácia
P4, P1	43,80%
P4, P2	36,82%
P4, P3	**55,70%**
P4, P5	53,01%
P4, P6	33,50%

3ª Iteração	
Associação de Propriedades	Acurácia
P4, P3, P1	47,70%
P4, P3, P2	44,20%
P4, P3, P5	**59,50%**
P4, P3, P6	37,90%

4ª Iteração	
Associação de Propriedades	Acurácia
P4, P3, P5, P1	52,70%
P4, P3, P5, P2	46,20%
P4, P3, P5, P6	38,00%

Figura 13 – Exemplo de Classificação das Melhores Propriedades (2ª e 3ª Parte Algoritmo Complementar)

6.2.1 Grupos Analisados

Dentre todas as análises preditivas realizadas, foram selecionados alguns dos resultados mais relevantes. Fica destacado que sem a utilização do algoritmo complementar exemplificado nas imagens 12 e 13 não seria possível realizar esse tipo de análise, porque as informações estavam ocultas pela grande quantidade de dados existente na base.

O significado de cada uma das propriedades utilizadas para se obter os resultados das análises preditivas estão relacionados na tabela 4.

Propriedade: **CO_GRUPO (Código da Área de enquadramento do curso no ENADE)**

- Opção de resposta: **ADMINISTRAÇÃO PÚBLICA**
- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 38.4;
  Classe 2: 38.5 à 49.5;
  Classe 3: 49.5 à 100
- Colunas analisadas: **'CO_IES','QE_I68','CO_RS_I9'**
- Profundidade máxima da árvore: **6**
- Acurácia: **59,38%**

Propriedade: **CO_GRUPO (Código da Área de enquadramento do curso no ENADE)**

- Opção de resposta: **SERVIÇO SOCIAL**
- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 34.7;
  Classe 2: 34.8 à 49.8;
  Classe 3: 49.9 à 100
- Colunas analisadas: **'CO_IES','QE_I11','CO_MODALIDADE','QE_I54', 'CO_RS_I6','QE_I08','CO_CATEGAD','QE_I21','QE_I19'**
- Profundidade máxima da árvore: **7**
- Acurácia: **57,91%**

Propriedade: **QE_I16 (Em que (UF) Unidade da Federação concluiu o ensino médio)**

- Opção de resposta: **Pará**
- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 35.3;
  Classe 2: 35.3 à 48.3;
  Classe 3: 48.4 à 100
- Colunas analisadas: **'CO_CURSO','QE_I26','QE_I17'**
- Profundidade máxima da árvore: **13**
- Acurácia: **56,50%**

Propriedade: **QE_I16 (Em que (UF) Unidade da Federação concluiu o ensino médio)**

- Opção de resposta: **Rio Grande do Norte**
- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 38.5;
  Classe 2: 38.6 à 52.4;
  Classe 3: 52.5 à 100
- Colunas analisadas:**CO_CURSO','QE_I14','QE_I03'**
- Profundidade máxima da árvore: **20**
- Acurácia: **55,50%**

Propriedade: **QE_I16 (Em que (UF) Unidade da Federação concluiu o ensino médio)**

- Opção de resposta: **Espírito Santo**

- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 40.4;
  Classe 2: 40.5 à 54;
  Classe 3: 54.1 à 100
- Colunas analisadas:**'CO_CURSO','QE_I67','QE_I02','QE_I14'**
- Profundidade máxima da árvore: **10**
- Acurácia: **52,90%**

Propriedade: **CO_TURNO_GRADUCAO (Código do turno de graduação)**

- Opção de resposta: **Integral**
- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 43.2;
  Classe 2: 43.3 à 57.3;
  Classe 3: 57.4 à 100
- Colunas analisadas: **'CO_CURSO','CO_CATEGAD','CO_IES','CO_GRUPO'**
- Profundidade máxima da árvore: **15**
- Acurácia: **51,71%**

Propriedade: **CO_GRUPO (Código da Área de enquadramento do curso no ENADE)**

- Opção de resposta: **RELAÇÕES INTERNACIONAIS**
- Classe de notas (todas com a mesma quantidade de registros):
  Classe 1: Notas até 44.7;
  Classe 2: 44.8 à 55.6;
  Classe 3: 55.7 à 100
- Colunas analisadas: **'CO_IES','CO_ORGACAD','CO_TURNO_GRADUACAO', 'CO_UF_CURSO','CO_RS_I9'**
- Profundidade máxima da árvore: **7**
- Acurácia: **51.42%**

Tabela 4 – Significado das Propriedades dos Resultados da Análises Preditivas

Propriedades	Significado
CO_CURSO	Código do curso no ENADE
CO_CATEGAD	Código da categoria administrativa da IES
CO_IES	Código da IES (e-MEC)
CO_GRUPO	Código da Área de enquadramento do curso no ENADE
QE_I11	A bolsa de estudos ou financiamento do curso que recebeu para custear as mensalidades
CO_MODALIDADE	Código da Modalidade de Ensino
QE_I54	Se os estudantes participaram de avaliações periódicas do curso (disciplinas, atuação dos professores, infraestrutura)
CO_RS_I6	Se as informações/instruções fornecidas para a resolução das questões foram suficientes para resolvê-las
QE_I08	Renda total da família, incluindo os próprios rendimentos
QE_I21	Se alguém da família concluiu um curso superior
QE_I19	Quem mais incentivou a cursar a graduação
QE_I68	Se a instituição dispôs de refeitório, cantina e banheiros em condições adequadas que atenderam as necessidades dos seus usuários
CO_RS_I9	O tempo gasto para concluir a prova
CO_TURNO_GRADUACAO	Código do turno de graduação
CO_UF_CURSO	Código da UF de funcionamento do curso
QE_I14	Se durante o curso de graduação, participou de programas e ou atividades curriculares no exterior
QE_I03	A nacionalidade
QE_I67	Se a instituição promoveu atividades de cultura, de lazer e interação social
QE_I02	A cor ou raça
QE_I26	A principal razão para ter escolhido a instituição de educação superior
QE_I17	Em que tipo de escola cursou o ensino médio

Tomando um caso como exemplo, ao obter algum aluno que cursa Administração Pública e utilizar os dados desse aluno no algoritmo com as propriedades que tiveram o melhor resultado de acurácia, teremos 59.38% de chance de acertar em qual classe de notas ele pertencerá.

Comparando as estimativas sem e com análise preditiva, podemos concluir que ao obter um conjunto de dados classificados em 3 classes com quantidade de registros equivalente, sem nenhum tipo de cálculo preditivo, é possível afirmar que um aluno tem 33% de chance

aproximadamente de estar em qualquer uma dessas classes. Agora se compararmos esse mesmo aluno, porém, depois que algumas de suas propriedades serem passadas pelo algoritmo de predição, a assertividade aumenta para 59.38%, no s casos de estudantes de Administração Pública, isso é 26,08% de diferença.

Não é destacado que os resultados de acurácias dessas propriedades sejam um máximo local, porém, para encontrar o conjunto de propriedades que obtenham o máximo global ou resultados de acurácias maiores. É necessário a utilização de outras técnicas de aplicação de Aprendizagem de Máquina.

O termo algoritmo preditivo se justifica, pois, existe a possibilidade de analisar as propriedades que têm a maior acurácia em alunos que ainda não realizaram a prova do ENADE, isso quer dizer que podemos fazer uma estimativa de qual seria seu possível resultado quando no futuro este realizar o exame. Com base nessa ideia, instituições do ensino superior podem estimular as propriedades que tendem a melhorar o resultado do aluno no ENADE.

Analisando puramente as propriedades que geram maior acurácia, pode-se observar que dentro de cada conjunto de propriedades de todos os resultados, eles possuem itens relacionadas a instituição, curso, cidade, e não somente dados do aluno. Logo, de acordo com essa informação, é possível indicar que a instituição e o curso pode ser um fator importante para determinar a nota do aluno no ENADE.

7 CONCLUSÃO

Foi observado que Ciência de Dados aperfeiçoa as técnicas de análise de dados, e não por acaso, está revolucionando as áreas que necessitam interpretar informações presentes em conjuntos de dados. Dentre estas áreas, foi destacado a possibilidade de incorporar a área da educação, por gerar grande quantidade de dados progressivamente, neste sentido, existe importantes benefícios do uso de Ciência de Dados na educação, pois possibilita auxiliar profissionais a tomar decisões mais assertivas em metodologias e abordagens de ensino.

No ENADE 2018, foco desse estudo, e pertence aos principais instrumentos para determinar que um curso e sua instituição possui bom grau de qualidade de ensino, foi aplicado a Ciência de Dados para investigar características que podem impactar na nota dos estudantes no exame.

Como um dos resultados, foi possível observar que 75% de todos os alunos que realizaram o ENADE 2018 não obtiveram resultados maiores que 55.2, logo é possível considerar esse resultado como relativamente baixo, pois o limite da nota é 100. É necessário outras análises e pesquisas sobre esse estudo para obter os motivos causadores deste desempenho.

A aplicação do algoritmo de aprendizagem de máquina Árvore de Decisão revelou que um importante fator que impacta na nota média dos alunos no ENADE é a própria instituição e o curso, pois essas informações esteve presente em todos os resultados gerados pelo algoritmo.

Ao se atentar no porquê os cursos e as instituições geram impactos no poder de predição de nota do aluno no ENADE. Vale para pesquisas futuras identificar o fator que difere um curso do outro, ou uma instituição da outra, e como isso pode estar contribuindo ou prejudicando o resultado do estudante neste exame.

Como resultado das análises preditivas sobre a nota na avaliação em grupos de alunos, se torna possível classificar notas de alunos pertencentes a esses grupos que ainda não realizaram o exame do ENADE, com o intuito de predizer essa nota com base nas propriedades analisadas que resultaram em boas acurácias. Isso é possível pois em muitas das vezes, essas propriedades estão relacionadas ao próprio aluno ou a instituição. Por exemplo, com o resultado de 56.5% de acurácia, é possível classificar alunos que concluíram o ensino médio no estado do Pará em classes de notas de acordo com as propriedades 'CO_CURSO','QE_I26','QE_I17', e determinar com 56.5% de certeza qual será o resultado na avaliação do ENADE para estudantes que ainda não foram inscritos neste exame. Dessa forma instituições do ensino superior podem estimular as propriedades que tendem a melhorar o resultado final do aluno no ENADE.

REFERÊNCIAS

ACAPS. **Data Cleaning**. 2016. ⟨https://www.acaps.org/sites/acaps/files/resources/files/acaps_technical_brief_data_cleaning_april_2016_0.pdf⟩. (Accessed on 08/14/2020).

ALPAYDIN, E. **Intruduction to Machine Learning**. 2014. (Accessed on 10/10/2020).

BALANSKAT, A.; ENGELHARDT, K. **Computingour future**. 2015. ⟨http://www.eun.org/documents/411753/817341/Computing+our+future_final_2015.pdf/d3780a64-1081-4488-8549-6033200e3c03/⟩. (Accessed on 10/10/2020).

CARVALHO, A. P. de Leon F de. **Aprendizado de Máquina**. 2015. ⟨http://www2.ic.uff.br/~kdmile/MachineLearning_Andre.pdf/⟩. (Accessed on 10/10/2020).

CHUNARKAR-PATIL, P.; BHOSALE, A. **Big Data Analytics**. 2018. ⟨https://medcraveonline.com/OAJS/OAJS-02-00095.pdf⟩. (Accessed on 14/09/2021).

CLEAR, A.; PARRISH, A. **Computing Curricula 2020 (CC2020) Paradigms for Future Computing Curricula**. 2020. 27 p. ⟨https://drive.google.com/file/d/1LPbxATWYSQlFJEB0ejDlsj1P_tVH0TLm/view/⟩. (Accessed on 10/10/2020).

CRUZ, A. **Ciência dos Dados e a Análise Preditiva**. 2015. ⟨https://riuni.unisul.br/bitstream/handle/12345/4878/JOSIEL_FERREIRA_SOARES.pdf?sequence=1&isAllowed=y⟩. (Acessado em 19/07/2021).

CURTY, R. G.; CERVANTES, B. M. N. Data science: Ciencia orientada a dados. 2016.

DIAS, J. da S.; PORTO, C. de M.; NUNES, A. K. F. **Formação Geral e Conhecimento Específico Na Prova do Enade**. 2016. ⟨https://eventos.set.edu.br/enfope/article/view/2056⟩. (Acessado em 24/10/2020).

DIJKSTRA, E. **Chapter 1. Computing**. 1972. (Accessed on 10/10/2020).

DYK, D. V.; FUENTES, M.; JORDAN, M. I.; NEWTON, M.; RAY, B. K.; LANG, D. T.; WICKHAM, H. **ASA Statement on the Role of Statistics in Data Science**. 2015. ⟨https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/⟩. (Accessed on 10/10/2020).

GORDON, S. **The Normal Distribution**. 2006. ⟨https://www.sydney.edu.au/content/dam/students/documents/mathematics-learning-centre/normal-distribution.pdf⟩. (Acessado em 21/08/2021).

HAND, D. J. **Data Mining: Statistics and More?** 1998. ⟨https://storm.cis.fordham.edu/~gweiss/selected-papers/data-mining-and-statistics-hand.pdf/⟩. (Accessed on 10/10/2020).

HEBBAR, P. Why should you learn python for data science? **Analytics India Magazine**, 2019. Disponível em: ⟨https://analyticsindiamag.com/why-should-you-learn-python-for-data-science/⟩.

INEP/MEC. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**. 2021. ⟨https://www.gov.br/inep/pt-br⟩. (Accessed on 08/18/2020).

JULIALANG. **Julia Documentation The Julia Language**. 2021. ⟨https://docs.julialang.org/en/v1/⟩. (Accessed on 01/08/2020).

LOPES, H. F. **Statistics or Data Science? What about MachineLearning? Predictability or Modeling? Big Data?** 2018. ⟨http://hedibert.org/wp-content/uploads/2018/12/statistics-datascience.pdf/⟩. (Accessed on 10/10/2020).

LOUKIDES, M. **What Is Data Science?** 2012. ⟨http://www.gmsl.it/wp-content/uploads/2014/09/What_Is_Data_Science_.pdf⟩. (Accessed on 10/07/2020).

MALIK, F. **R - Statistical Programming Language**. 2020. ⟨https://towardsdatascience.com/r-statistical-programming-language-6adc8c0a6e3d⟩. (Accessed on 10/08/2020).

MARQUES, D. **A Decision Tree Learner for Cost-Sensitive Binary Classification**. 2016. ⟨https://www.maxwell.vrac.puc-rio.br/28239/28239.PDF⟩. (Acessado em 19/07/2021).

MARR, B. **What's The Difference Between Structured, Semi-Structured And Unstructured Data?** 2019. ⟨https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/#75fe7a192b4d⟩. (Accessed on 07/27/2020).

MORETTIN, P. A.; SINGER, J. M. Introdução à ciência de dados: Fundamentos e aplicações. 2020.

MRAN. **What is R?** 2021. ⟨https://mran.microsoft.com/documents/what-is-r⟩. (Accessed on 01/08/2021).

PERKEL, J. **Julia: Come For The Syntax, Stay for the Speed**. 2021. ⟨https://www.nature.com/articles/d41586-019-02310-3⟩. (Acessado em 19/07/2021).

PYTHON. **Python 3.9.0 documentation**. 2021. ⟨https://docs.python.org⟩. (Accessed on 10/02/2021).

R-PROJECT. **The R Project for Statistical Computing**. 2021. ⟨https://www.r-project.org/⟩. (Accessed on 01/08/2021).

RAHM, E.; DO, H. H. **Data Cleaning: Problems and Current Approaches**. 2000. ⟨http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.8661&rep=rep1&type=pdf⟩. (Accessed on 08/04/2020).

ROCHA, A. **Correlação E Regressão: Modelos Probabilísticos Para a Computação**. 2011. ⟨https://docplayer.com.br/73364988-Correlacao-e-regressao-modelos-probabilisticos-para-a-computacao-professora-andrea-rocha-uni
html⟩. (Acessado em 21/08/2021).

SHARMA, H. What is data science? a beginner's guide to data science. 2019. Disponível em: ⟨https://www.edureka.co/blog/what-is-data-science/⟩.

SILVA, L. A.; SILVEIRA, I. F.; SILVA, L.; RAMOS, J. L. C.; RODRIGUES, R. L. Ciência de dados educacionais: definições e convergências entre as áreas de pesquisa. 2017.

SOLEY-BORI, M. **Dealing with missing data: Key assumptions and methods for applied analysis**. 2013. ⟨http://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf⟩. (Accessed on 08/04/2020).

SONG, Y.; ZHU, Y. **Big Data and Data Science: What Should we Teach**. 2015. ⟨https://www.researchgate.net/publication/282692841_Big_data_and_data_science_What_should_we_teach⟩. (Accessed on 10/10/2020).

STODDEN, V. **The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science**. 2020. ⟨https://cacm.acm.org/magazines/2020/7/245700-the-data-science-life-cycle/fulltext⟩. (Accessed on 07/27/2020).

WING, J. M. **The Data Life Cycle · Harvard Data Science Review**. 2019. ⟨https://hdsr.mitpress.mit.edu/pub/577rq08d/release/3⟩. (Accessed on 07/27/2020).

YEGULALP, S. **Julia vs. Python: Which is Best for Data Science?** 2020. ⟨https://www.infoworld.com/article/3241107/julia-vs-python-which-is-best-for-data-science.html⟩. (Accessed on 10/08/2020).

ANEXO A – DICIONÁRIO DE PROPRIEDADES DO ENADE 2018

PARTE 1 - INFORMAÇÕES DA INSTITUIÇÃO DE ENSINO SUPERIOR E DO CURSO		
PROPRIEDADE	**DESCRIÇÃO**	**CATEGORIAS/OPÇÕES**
NU_ANO	Ano de realização do exame	2018
CO_IES	Código da IES (e-MEC)	Entre 1 e 23410
CO_CATEGAD	Código da categoria administrativa da IES	118=Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Civil
		120=Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Associação de Utilidade Pública
		121=Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Fundação
		10005=Privada com fins lucrativos
		10006=Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Mercantil ou Comercia
		10007=Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Associação de Utilidade Pública
		10008=Privada sem fins lucrativos
		10009=Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Sociedade
		17634=Fundação Pública de Direito Privado Municipal
		93=Pessoa Jurídica de Direito Público - Federal
		115=Pessoa Jurídica de Direito Público - Estadual
		116=Pessoa Jurídica de Direito Público - Municipal
		10001=Pessoa Jurídica de Direito Público - Estadual
		10002=Pessoa Jurídica de Direito Público - Federal
		10003=Pessoa Jurídica de Direito Público - Municipal
CO_ORGACAD	Código da organização acadêmica da IES	10019=Centro Federal de Educação Tecnológica
		10020=Centro Universitário
		10022=Faculdade
		10026=Instituto Federal de Educação, Ciência e Tecnologia
		10028=Universidade
		1=ADMINISTRAÇÃO
		2=DIREITO
		13=CIÊNCIAS ECONÔMICAS
		18=PSICOLOGIA
		22=CIÊNCIAS CONTÁBEIS
		26=DESIGN
		29=TURISMO
		38=SERVIÇO SOCIAL
		67=SECRETARIADO EXECUTIVO
		81=RELAÇÕES INTERNACIONAIS

CO_GRUPO	Código da Área de enquadramento do curso no Enade	83=TECNOLOGIA EM DESIGN DE MODA
		84=TECNOLOGIA EM MARKETING
		85=TECNOLOGIA EM PROCESSOS GERENCIAIS
		86=TECNOLOGIA EM GESTÃO DE RECURSOS HUMANOS
		87=TECNOLOGIA EM GESTÃO FINANCEIRA
		88=TECNOLOGIA EM GASTRONOMIA
		93=TECNOLOGIA EM GESTÃO COMERCIAL
		94=TECNOLOGIA EM LOGÍSTICA
		100=ADMINISTRAÇÃO PÚBLICA
		101=TEOLOGIA
		102=TECNOLOGIA EM COMÉRCIO EXTERIOR
		103=TECNOLOGIA EM DESIGN DE INTERIORES
		104=TECNOLOGIA EM DESIGN GRÁFICO
		105=TECNOLOGIA EM GESTÃO DA QUALIDADE
		106=TECNOLOGIA EM GESTÃO PÚBLICA
		803=COMUNICAÇÃO SOCIAL - JORNALISMO
		804=COMUNICAÇÃO SOCIAL - PUBLICIDADE E PROPAGANDA
CO_CURSO	Código do curso no Enade	Entre 1 e 5001389
CO_MODALIDADE	Código da Modalidade de Ensino	1=Educação Presencial
		2=Educação a Distância
CO_MUNIC_CURSO	Código do município de funcionamento do curso	Ir para Planilha MUNICÍPIOS
CO_UF_CURSO	Código da UF de funcionamento do curso	Conforme código do IBGE
		11 = Rondônia (RO)
		12 = Acre (AC)
		13 = Amazonas (AM)
		14 = Roraima (RR)
		15 = Pará (PA)
		16 = Amapa (AP)
		17 = Tocantins (TO)
		21 = Maranhão (MA)
		22= Piauí (PI)
		23 = Ceará (CE)
		24 = Rio Grande do Norte (RN)
		25 = Paraíba (PB)
		26 = Pernambuco (PE)
		27 = Alagoas (AL)
		28 = Sergipe (SE)
		29 = Bahia (BA)
		31 = Minas gerais (MG)
		32 = Espírito Santo (ES)
		33 = Rio de Janeiro (RJ)

		35 = São Paulo (SP)
		41 = Paraná (PR)
		42 = Santa Catarina (SC)
		43 = Rio Grande do Sul (RS)
		50 = Mato Grosso do Sul (MS)
		51 = Mato Grosso (MT)
		52 = Goiás (GO)
		53 = Distrito federal (DF)
CO_REGIÃO_CURSO	Código da região de funcionamento do curso	1 = Região Norte (NO)
		2 = Região Nordeste (NE)
		3 = Região Sudeste (SE)
		4 = Região Sul (SUL)
		5 = Região Centro-Oeste (CO)
PARTE 2 - INFORMAÇÕES DO ESTUDANTE		
NU_IDADE	Idade do inscrito em 25/11/2018	valores entre 4 e 94
TP_SEXO	Sexo	M = Masculino
		F = Feminino
ANO_FIM_EM	Ano de conclusão do Ensino Médio	AAAA (valores entre 0 e 2686)
ANO_IN_GRAD	Ano de início da graduação	AAAA (valores entre 1973 e 2099)
CO_TURNO_GRADUCAO	Código do turno de graduação	1 = Matutino
		2 = Vespertino
		3 = Integral
		4 = Noturno
TP_INSCRICAO_ADM	Forma pela qual foi realizada a inscrição	0 = Tradicional
		1 = Judicial
		2 = Administrativa
TP_INSCRICAO	Tipo de inscrição	0 = Inscrito
		1 = Não inscrito
AVALIAÇÃO - FORMAÇÃO GERAL E COMPONENTE ESPECÍFICO PARTE 3 - NÚMERO DE ITENS DA PARTE OBJETIVA		
NU_ITEM_OFG	Número de itens da parte objetiva da Formação Geral	8 itens
NU_ITEM_OFG_Z	Número de itens da parte objetiva da Formação Geral que foram excluídos devido a anulação	Min=0 Max=0
NU_ITEM_OFG_X	Número de itens da parte objetiva da Formação Geral que foram excluídos devido ao coeficiente ponto-bisserial menor que 0,20	Min=0 Max=0
NU_ITEM_OFG_N	Número de itens da parte objetiva da Formação Geral que não se aplicam ao grupo de curso	Min=0 Max=0
NU_ITEM_OCE	Número de itens da parte objetiva de Componente Específico	27 itens
NU_ITEM_OCE_Z	Número de itens da parte objetiva de Componente Específico que foram excluídos devido a anulação	Min=0 Max=2 (Áreas 22, 67 e 104)
NU_ITEM_OCE_X	Número de itens da parte objetiva de Componente Específico que foram excluídos devido ao coeficiente ponto-bisserial menor que 0,20	Min=0 Max=9 (Área 22)

NU_ITEM_OCE_N	Número de itens da parte objetiva de Componente Específico que não se aplicam ao grupo de curso	Min=0 Max=0
PARTE 4 - VETORES		
DS_VT_GAB_OFG_ORIG	Vetor que representa o gabarito original de Formação Geral	1 letra por item (intervalo de A a E)
		Z = questão excluída devido à anulação
DS_VT_GAB_OFG_FIN	Vetor que representa o gabarito final de Formação Geral	1 letra por item (intervalo de A a E)
		Z = questão excluída devido à anulação
		X = questão excluída devido ao coeficiente ponto-bisserial < 0,20
DS_VT_GAB_OCE_ORIG	Vetor que representa o gabarito original de Componente Específico	1 letra por item (intervalo de A a E)
		Z = questão excluída devido à anulação
DS_VT_GAB_OCE_FIN	Vetor que representa o gabarito final de Componente Específico	1 letra por item (intervalo de A a E)
		Z = questão excluída devido à anulação
		X = questão excluída devido ao coeficiente ponto-bisserial < 0,20
DS_VT_ESC_OFG	Vetor que representa a escolha de resposta da parte objetiva da Formação Geral	1 letra por item (intervalo de A a E), "."=em branco, "*"=múltiplo
DS_VT_ACE_OFG	Vetor que representa os acertos da parte objetiva na Formação Geral	0 = Errado
		1 = Certo
		8 = Anulada pela comissão
		9 = Anulada pelo índice de discriminação (correlação pontobisserial < 0,20)
DS_VT_ESC_OCE	Vetor que representa a escolha de resposta da parte objetiva do Componente Específico	1 letra por item, "."=em branco, "*"=múltiplo
DS_VT_ACE_OCE	Vetor que representa os acertos da parte objetiva do Componente Específico	0 = Errado
		1 = Certo
		8 = Anulada pela comissão
		9 = Anulada pelo índice de discriminação
PARTE 5 - TIPOS DE PRESENÇA		
TP_PRES	Tipo de presença no Enade	222 = Ausente
		334 = Eliiminado por participacao indevida
		444 = Ausente devido a dupla graduação
		555 = Presente com resultado válido
		556 = Presente com resultado desconsiderado pela Aplicadora
TP_PR_GER	Tipo de presença na prova	222 = Prova não realizada devido a ausência do estudante
		333 = Participação com prova em branco
		555 = Participação com respostas na prova
		556 = Participação com resultado desconsiderado pela Aplicadora
		222 = Prova não realizada devido a ausência do estudante

TP_PR_OB_FG	Tipo de presença na parte objetiva na formação geral	333 = Participação com prova em branco
		555 = Participação com respostas na prova
		556 = Participação com resultado desconsiderado pela Aplicadora
TP_PR_DI_FG	Tipo de presença na parte discursiva na formação geral	222 = Prova não realizada devido a ausência do estudante
		333 = Participação com prova em branco
		555 = Participação com respostas na prova
		556 = Participação com resultado desconsiderado pela Aplicadora
TP_PR_OB_CE	Tipo de presença na parte objetiva no componente específico	222 = Prova não realizada devido a ausência do estudante
		333 = Participação com prova em branco
		555 = Participação com respostas na prova
		556 = Participação com resultado desconsiderado pela Aplicadora
TP_PR_DI_CE	Tipo de presença na parte discursiva no componente específico	222 = Prova não realizada devido a ausência do estudante
		333 = Participação com prova em branco
		555 = Participação com respostas na prova
		556 = Participação com resultado desconsiderado pela Aplicadora
PARTE 6 - TIPOS DE SITUAÇÃO DAS QUESTÕES DA PARTE DISCURSIVA (*) resultado considerados para o cálculo da nota do estudante; (**) resultado desconsiderado para o cálculo da nota do estudante.		
TP_SFG_D1	Tipo de situação da questão 1 da parte discursiva da formação geral	222 = Não se aplica (estudante ausente)**
		333 = Questão em branco (estudante presente) *
		335 = Questão zerada por motivo de resposta nula*
		336 = Questão zerada por motivo de resposta divergente com a temática *
		555 = Questão com resultado válido *
		556 = Questão com resultado desconsiderado por problemas administrativos **
TP_SFG_D2	Tipo de situação da questão 2 da parte discursiva da formação geral	222 = Não se aplica (estudante ausente)**
		333 = Questão em branco (estudante presente) *
		335 = Questão zerada por motivo de resposta nula*
		336 = Questão zerada por motivo de resposta divergente com a temática *
		555 = Questão com resultado válido *
		556 = Questão com resultado desconsiderado por problemas administrativos **

TP_SCE_D1	Tipo de situação da questão 1 da parte discursiva do componente específico	222 = Não se aplica (estudante ausente)**
		333 = Questão em branco (estudante presente) *
		335 = Questão zerada por motivo de resposta nula*
		336 = Questão zerada por motivo de resposta divergente com a temática *
		555 = Questão com resultado válido *
		556 = Questão com resultado desconsiderado por problemas administrativos **
TP_SCE_D2	Tipo de situação da questão 2 da parte discursiva do componente específico	222 = Não se aplica (estudante ausente)**
		333 = Questão em branco (estudante presente) *
		335 = Questão zerada por motivo de resposta nula*
		336 = Questão zerada por motivo de resposta divergente com a temática *
		555 = Questão com resultado válido *
		556 = Questão com resultado desconsiderado por problemas administrativos **
TP_SCE_D3	Tipo de situação da questão 3 da parte discursiva do componente específico	222 = Não se aplica (estudante ausente)**
		333 = Questão em branco (estudante presente) *
		335 = Questão zerada por motivo de resposta nula*
		336 = Questão zerada por motivo de resposta divergente com a temática *
		555 = Questão com resultado válido *
		556 = Questão com resultado desconsiderado por problemas administrativos **
PARTE 7 - NOTAS NA FORMAÇÃO GERAL E COMPONENTE ESPECÍFICO		
NT_GER	Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%) (valor de 0 a 100)	Min = 0 Max = 93,7
NT_FG	Nota bruta na formação geral - Média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral (valor de 0 a 100)	Min = 0 Max = 99,2
NT_OBJ_FG	Nota bruta na parte objetiva da formação geral (valor de 0 a 100)	Min = 0 Max = 87,5
NT_DIS_FG	Nota bruta na parte discursiva da formação geral (valor de 0 a 100)	Min = 0 Max = 98,0
NT_FG_D1	Nota da questão 1 da parte discursiva da formação geral - Média ponderada da parte de Língua Portuguesa (20%) e Conteúdo (80%) da Questão 1 da parte discursiva (valor de 0 a 100)	Min = 0 Max = 100,0

NT_FG_D1_PT	Nota de **Língua Portuguesa** da **questão 1** da parte discursiva da formação geral (valor de 0 a 100)	Min = 0 Max = 100,0
NT_FG_D1_CT	Nota de **Conteúdo** da **questão 1** da parte discursiva da formação geral (valor de 0 a 100)	Min = 0 Max = 100,0
NT_FG_D2	Nota da questão 2 da parte discursiva na formação geral - Média ponderada da parte de Língua Portuguesa (20%) e Conteúdo (80%) da Questão 2 da parte discursiva (valor de 0 a 100)	Min = 0 Max = 99,0
NT_FG_D2_PT	Nota de **Língua Portuguesa** da **questão 2** da parte discursiva da formação geral (valor de 0 a 100)	Min = 0 Max = 100,0
NT_FG_D2_CT	Nota de **Conteúdo** da **questão 2** da parte discursiva da formação geral (valor de 0 a 100)	Min = 0 Max = 100,0
NT_CE	Nota bruta no componente específico - Média ponderada da parte objetiva (85%) e discursiva (15%) no componente específico (valor de 0 a 100)	Min = 0 Max = 97,5
NT_OBJ_CE	Nota bruta na parte objetiva do componente específico (valor de 0 a 100)	Min = 0 Max = 96,3
NT_DIS_CE	Nota bruta na parte discursiva do componente específico (valor de 0 a 100)	Min = 0 Max = 98,3
NT_CE_D1	Nota da questão 1 da parte discursiva do componente específico (valor de 0 a 100)	Min = 0 Max = 100,0
NT_CE_D2	Nota da questão 2 da parte discursiva do componente específico (valor de 0 a 100)	Min = 0 Max = 100,0
NT_CE_D3	Nota da questão 3 da parte discursiva do componente específico (valor de 0 a 100)	Min = 0 Max = 100,0
QUESTIONÁRIOS PARTE 8 - QUESTIONÁRIO DE PERCEPÇÃO DA PROVA		
CO_RS_I1	1 - Qual o grau de dificuldade desta prova na parte de Formação Geral?	A = Muito fácil.
		B = Fácil.
		C = Médio.
		D = Difícil.
		E = Muito difícil.
		. = Sem resposta
		* = Resposta anulada
CO_RS_I2	2 - Qual o grau de dificuldade desta prova na parte do Componente Específico?	A = Muito fácil.
		B = Fácil.
		C = Médio.
		D = Difícil.
		E = Muito difícil.

		. = Sem resposta
		* = Resposta anulada
CO_RS_I3	3 - Considerando a extensão da prova, em relação ao tempo total, você considera que a prova foi:	A = Muito longa.
		B = Longa.
		C = Adequada.
		D = Curta.
		E = Muito curta.
		. = Sem resposta
		* = Resposta anulada
CO_RS_I4	4 - Os enunciados das questões da prova na parte de Formação Geral estavam claros e objetivos?	A = Sim, todos.
		B = Sim, a maioria.
		C = Apenas cerca da metade.
		D = Poucos.
		E = Não, nenhum.
		. = Sem resposta
		* = Resposta anulada
CO_RS_I5	5 - Os enunciados das questões na parte do Componente Específico estavam claros e objetivos?	A = Sim, todos.
		B = Sim, a maioria.
		C = Apenas cerca da metade.
		D = Poucos se apresentam.
		E = Não, nenhum.
		. = Sem resposta
		* = Resposta anulada
CO_RS_I6	6 - As informações/instruções fornecidas para a resolução das questões foram suficientes para resolvê-las?	A = Sim, até excessivas.
		B = Sim, em todas elas.
		C = Sim, na maioria delas.
		D = Sim, somente em algumas.
		E = Não, em nenhuma delas.
		. = Sem resposta
		* = Resposta anulada
CO_RS_I7	7 - Você se deparou com alguma dificuldade ao responder à prova. Qual?	A = Desconhecimento do conteúdo.
		B = Forma diferente de abordagem do conteúdo.
		C = Espaço insuficiente para responder às questões.
		D = Falta de motivação para fazer a prova.
		E = Não tive qualquer tipo de dificuldade para responder à prova.
		. = Sem resposta
		* = Resposta anulada
CO_RS_I8	8 - Considerando apenas as questões objetivas da prova, você percebeu que:	A = Não estudou ainda a maioria desses conteúdos.
		B = Estudou alguns desses conteúdos, mas não os aprendeu.
		C = Estudou a maioria desses conteúdos, mas não os aprendeu.
		D = Estudou e aprendeu muitos desses conteúdos.
		E = Estudou e aprendeu todos esses conteúdos.
		. = Sem resposta
		* = Resposta anulada
		A = Menos de uma hora.

CO_RS_I9	9 - Qual foi o tempo gasto por você para concluir a prova?	B = Entre uma e duas horas.
		C = Entre duas e três horas.
		D = Entre três e quatro horas.
		E = Quatro horas e não consegui terminar.
		. = Sem resposta
		* = Resposta anulada
PARTE 9 - QUESTIONÁRIO DO ESTUDANTE		
QE_I01	Qual o seu estado civil?	A = Solteiro(a).
		B = Casado(a).
		C = Separado(a) judicialmente/divorciado(a).
		D = Viúvo(a).
		E = Outro.
QE_I02	Qual é a sua cor ou raça?	A = Branca.
		B = Preta.
		C = Amarela.
		D = Parda.
		E = Indígena.
		F = Não quero declarar.
QE_I03	Qual a sua nacionalidade?	A = Brasileira.
		B = Brasileira naturalizada.
		C = Estrangeira.
QE_I04	Até que etapa de escolarização seu pai concluiu?	A = Nenhuma.
		B = Ensino Fundamental: 1º ao 5º ano (1ª a 4ª série).
		C = Ensino Fundamental: 6º ao 9º ano (5ª a 8ª série).
		D = Ensino Médio.
		E = Ensino Superior - Graduação.
		F = Pós-graduação.
QE_I05	Até que etapa de escolarização sua mãe concluiu?	A = Nenhuma.
		B = Ensino Fundamental: 1º ao 5º ano (1ª a 4ª série).
		C = Ensino Fundamental: 6º ao 9º ano (5ª a 8ª série).
		D = Ensino médio.
		E = Ensino Superior - Graduação.
		F = Pós-graduação.
QE_I06	Onde e com quem você mora atualmente?	A = Em casa ou apartamento, sozinho.
		B = Em casa ou apartamento, com pais e/ou parentes.
		C = Em casa ou apartamento, com cônjuge e/ou filhos.
		D = Em casa ou apartamento, com outras pessoas (incluindo república).
		E = Em alojamento universitário da própria instituição.
		F = Em outros tipos de habitação individual ou coletiva (hotel, hospedaria, pensão ou outro).
	Quantas pessoas da sua família moram com você? Considere seus	A = Nenhuma.
		B = Uma.
		C = Duas.

QE_I07	moram com você? Considere seus pais, irmãos, cônjuge, filhos e outros parentes que moram na mesma casa com você.	D = Três.
		E = Quatro.
		F = Cinco.
		G = Seis.
		H = Sete ou mais.
QE_I08	Qual a renda total de sua família, incluindo seus rendimentos?	A = Até 1,5 salário mínimo (até R$ 1.431,00).
		B = De 1,5 a 3 salários mínimos (R$ 1.431,01 a R$ 2.862,00).
		C = De 3 a 4,5 salários mínimos (R$ 2.862,01 a R$ 4.293,00).
		D = De 4,5 a 6 salários mínimos (R$ 4.293,01 a R$ 5.724,00).
		E = De 6 a 10 salários mínimos (R$ 5.724,01 a R$ 9.540,00).
		F = De 10 a 30 salários mínimos (R$ 9.540,01 a R$ 28.620,00).
		G = Acima de 30 salários mínimos (mais de R$ 28.620,00).
QE_I09	Qual alternativa a seguir melhor descreve sua situação financeira (incluindo bolsas)?	A = Não tenho renda e meus gastos são financiados por programas governamentais.
		B = Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas.
		C = Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos.
		D = Tenho renda e não preciso de ajuda para financiar meus gastos.
		E = Tenho renda e contribuo com o sustento da família.
		F = Sou o principal responsável pelo sustento da família.
QE_I10	Qual alternativa a seguir melhor descreve sua situação de trabalho (exceto estágio ou bolsas)?	A = Não estou trabalhando.
		B = Trabalho eventualmente.
		C = Trabalho até 20 horas semanais.
		D = Trabalho de 21 a 39 horas semanais.
		E = Trabalho 40 horas semanais ou mais.
QE_I11	Que tipo de bolsa de estudos ou financiamento do curso você recebeu para custear todas ou a maior parte das mensalidades? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.	A = Nenhum, pois meu curso é gratuito.
		B = Nenhum, embora meu curso não seja gratuito.
		C = ProUni integral.
		D = ProUni parcial, apenas.
		E = FIES, apenas.
		F = ProUni Parcial e FIES.
		G = Bolsa oferecida por governo estadual, distrital ou municipal.
		H = Bolsa oferecida pela própria instituição.
		I = Bolsa oferecida por outra entidade (empresa, ONG, outra).

		J = Financiamento oferecido pela própria instituição.
		K = Financiamento bancário.
QE_I12	Ao longo da sua trajetória acadêmica, você recebeu algum tipo de bolsa de permanência? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.	A = Nenhum.
		B = Auxílio moradia.
		C = Auxílio alimentação.
		D = Auxílio moradia e alimentação.
		E = Auxílio Permanência.
		F = Outro tipo de auxílio.
QE_I13	Ao longo da sua trajetória acadêmica, você recebeu algum tipo de bolsa acadêmica? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.	A = Nenhum.
		B = Bolsa de iniciação científica.
		C = Bolsa de extensão.
		D = Bolsa de monitoria/tutoria.
		E = Bolsa PET.
		F = Outro tipo de bolsa acadêmica.
QE_I14	Durante o curso de graduação, você participou de programas e ou atividades curriculares no exterior?	A = Não participei.
		B = Sim, Programa Ciência sem Fronteiras.
		C = Sim, programa de intercâmbio financiado pelo Governo Federal (Marca; Brafitec; PLI; outro).
		D = Sim, programa de intercâmbio financiado pelo Governo Estadual.
		E = Sim, programa de intercâmbio da minha instituição.
		F = Sim, outro intercâmbio não institucional.
QE_I15	Seu ingresso no curso de graduação se deu por meio de políticas de ação afirmativa ou inclusão social?	A = Não.
		B = Sim, por critério étnico-racial.
		C = Sim, por critério de renda.
		D = Sim, por ter estudado em escola pública ou particular com bolsa de estudos.
		E = Sim, por sistema que combina dois ou mais critérios anteriores.
		F = Sim, por sistema diferente dos anteriores.
QE_I16	Em que unidade da Federação você concluiu o ensino médio?	Conforme código do IBGE
		11 = Rondônia (RO)
		12 = Acre (AC)
		13 = Amazonas (AM)
		14 = Roraima (RR)
		15 = Pará (PA)
		16 = Amapa (AP)
		17 = Tocantins (TO)
		21 = Maranhão (MA)
		22= Piauí (PI)
		23 = Ceará (CE)
		24 = Rio Grande do Norte (RN)
		25 = Paraíba (PB)
		26 = Pernambuco (PE)
		27 = Alagoas (AL)
		28 = Sergipe (SE)
		29 = Bahia (BA)
		31 = Minas gerais (MG)

		32 = Espírito Santo (ES)
		33 = Rio de Janeiro (RJ)
		35 = São Paulo (SP)
		41 = Paraná (PR)
		42 = Santa Catarina (SC)
		43 = Rio Grande do Sul (RS)
		50 = Mato Grosso do Sul (MS)
		51 = Mato Grosso (MT)
		52 = Goiás (GO)
		53 = Distrito federal (DF)
		99 = Não se aplica
QE_I17	Em que tipo de escola você cursou o ensino médio?	A = Todo em escola pública.
		B = Todo em escola privada (particular).
		C = Todo no exterior.
		D = A maior parte em escola pública.
		E = A maior parte em escola privada (particular).
		F = Parte no Brasil e parte no exterior.
QE_I18	Qual modalidade de ensino médio você concluiu?	A = Ensino médio tradicional.
		B = Profissionalizante técnico (eletrônica, contabilidade, agrícola, outro).
		C = Profissionalizante magistério (Curso Normal).
		D = Educação de Jovens e Adultos (EJA) e/ou Supletivo.
		E = Outro modalidade.
QE_I19	Quem mais lhe incentivou a cursar a graduação?	A = Ninguém.
		B = Pais.
		C = Outros membros da família que não os pais.
		D = Professores.
		E = Líder ou representante religioso.
		F = Colegas/Amigos.
		G = Outras pessoas.
QE_I20	Algum dos grupos abaixo foi determinante para você enfrentar dificuldades durante seu curso superior e conclui-lo?	A = Não tive dificuldade.
		B = Não recebi apoio para enfrentar dificuldades.
		C = Pais.
		D = Avós.
		E = Irmãos, primos ou tios.
		F = Líder ou representante religioso.
		G = Colegas de curso ou amigos.
		H = Professores do curso.
		I = Profissionais do serviço de apoio ao estudante da IES.
		J = Colegas de trabalho.
		K = Outro grupo.
QE_I21	Alguém em sua família concluiu um curso superior?	A = Sim.
		B = Não.
QE_I22	Excetuando-se os livros indicados na bibliografia do seu curso,	A = Nenhum.
		B = Um ou dois.
		C = De três a cinco.

	quantos livros você leu neste ano?	D = De seis a oito.
		E = Mais de oito.
QE_I23	Quantas horas por semana, aproximadamente, você dedicou aos estudos, excetuando as horas de aula?	A = Nenhuma, apenas assisto às aulas.
		B = De uma a três.
		C = De quatro a sete.
		D = De oito a doze.
		E = Mais de doze.
QE_I24	Você teve oportunidade de aprendizado de idioma estrangeiro na Instituição?	A = Sim, somente na modalidade presencial.
		B = Sim, somente na modalidade semipresencial.
		C = Sim, parte na modalidade presencial e parte na modalidade semipresencial.
		D = Sim, na modalidade a distância.
		E = Não.
QE_I25	Qual o principal motivo para você ter escolhido este curso?	A = Inserção no mercado de trabalho.
		B = Influência familiar.
		C = Valorização profissional.
		D = Prestígio Social.
		E = Vocação.
		F = Oferecido na modalidade a distância.
		G = Baixa concorrência para ingresso.
		H = Outro motivo.
QE_I26	Qual a principal razão para você ter escolhido a sua instituição de educação superior?	A = Gratuidade.
		B = Preço da mensalidade.
		C = Proximidade da minha residência.
		D = Proximidade do meu trabalho.
		E = Facilidade de acesso.
		F = Qualidade/reputação.
		G = Foi a única onde tive aprovação.
		H = Possibilidade de ter bolsa de estudo.
		I = Outro motivo.
As variáveis QE_I27 a QE_I68 foram marcadas conforme o grau de concordância do estudante para cada assertiva, segundo a escala que varia de 1 (discordância total) a 6 (concordância total). A resposta 7 foi marcada qunado o estudante julgou não ter elementospara avaliar a assertiva (Não sei responder) e reposta 8 quando considerou a não pertinente ao seu curso (Não se aplica).		
QE_I27	As disciplinas cursadas contribuíram para sua formação integral, como cidadão e profissional.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
QE_I28	Os conteúdos abordados nas disciplinas do curso favoreceram sua atuação em estágios ou em atividades de iniciação profissional.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.

	profissional.	7 = Não se aplica.
		8 = Não sei responder.
QE_I29	As metodologias de ensino utilizadas no curso desafiaram você a aprofundar conhecimentos e desenvolver competências reflexivas e críticas.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
QE_I30	O curso propiciou experiências de aprendizagem inovadoras.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
QE_I31	O curso contribuiu para o desenvolvimento da sua consciência ética para o exercício profissional.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
QE_I32	No curso você teve oportunidade de aprender a trabalhar em equipe.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
QE_I33	O curso possibilitou aumentar sua capacidade de reflexão e argumentação.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
QE_I34	O curso promoveu o desenvolvimento da sua capacidade de pensar criticamente, analisar e refletir sobre soluções para problemas da sociedade.	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.
		4 = Concordo parcialmente.
		5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.
	O curso contribuiu para você	1 = Discordo totalmente.
		2 = Discordo.
		3 = Discordo parcialmente.

QE_I35	ampliar sua capacidade de comunicação nas formas oral e escrita.	4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I36	O curso contribuiu para o desenvolvimento da sua capacidade de aprender e atualizar-se permanentemente.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I37	As relações professor-aluno ao longo do curso estimularam você a estudar e aprender.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I38	Os planos de ensino apresentados pelos professores contribuíram para o desenvolvimento das atividades acadêmicas e para seus estudos.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I39	As referências bibliográficas indicadas pelos professores nos planos de ensino contribuíram para seus estudos e aprendizagens.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I40	Foram oferecidas oportunidades para os estudantes superarem dificuldades relacionados ao processo de formação.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I41	A coordenação do curso esteve disponível para orientação acadêmica dos estudantes.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.

QE_I42	O curso exigiu de você organização e dedicação frequente aos estudos.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I43	Foram oferecidas oportunidades para os estudantes participarem de programas, projetos ou atividades de extensão universitária.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I44	Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I45	O curso ofereceu condições para os estudantes participarem de eventos internos e/ou externos à instituição.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I46	A instituição ofereceu oportunidades para os estudantes atuarem como representantes em órgãos colegiados.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I47	O curso favoreceu a articulação do conhecimento teórico com atividades práticas.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I48	As atividades práticas foram suficientes para relacionar os conteúdos do curso com a prática, contribuindo para sua formação	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo.

	contribuindo para sua formação profissional.	6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I49	O curso propiciou acesso a conhecimentos atualizados e/ou contemporâneos em sua área de formação.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I50	O estágio supervisionado proporcionou experiências diversificadas para a sua formação.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I51	As atividades realizadas durante seu trabalho de conclusão de curso contribuíram para qualificar sua formação profissional.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I52	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios no país.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I54	Os estudantes participaram de avaliações periódicas do curso (disciplinas, atuação dos professores, infraestrutura).	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
		1 = Discordo totalmente. 2 = Discordo.

QE_I55	As avaliações de aprendizagem realizadas durante o curso foram compatíveis com os conteúdos ou temas trabalhados pelos professores.	3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I56	Os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I57	Os professores demonstraram domínio dos conteúdos abordados nas disciplinas.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I58	Os professores utilizaram tecnologias da informação e comunicação (TIC's) como estratégia de ensino (projetor, multimídia, laboratório de informática, ambiente virtual de aprendizagem).	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I59	A instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I60	O curso disponibilizou monitores ou tutores para auxiliar os estudantes.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I61	As condições de infraestrutura das salas de aula foram adequadas.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica.

		8 = Não sei responder.
QE_I62	Os equipamentos e materiais disponíveis para as aulas práticas foram adequados para a quantidade de estudantes.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I63	Os ambientes e equipamentos destinados às aulas práticas foram adequados ao curso.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I64	A biblioteca dispôs das referências bibliográficas que os estudantes necessitaram.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I65	A instituição contou com biblioteca virtual ou conferiu acesso a obras disponíveis em acervos virtuais.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I66	As atividades acadêmicas desenvolvidas dentro e fora da sala de aula possibilitaram reflexão, convivência e respeito à diversidade.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I67	A instituição promoveu atividades de cultura, de lazer e interação social.	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente. 5 = Concordo. 6 = Concordo totalmente. 7 = Não se aplica. 8 = Não sei responder.
QE_I68	A instituição dispôs de refeitório, cantina e banheiros em condições	1 = Discordo totalmente. 2 = Discordo. 3 = Discordo parcialmente. 4 = Concordo parcialmente.

QE_I68	adequadas que atenderam as necessidades dos seus usuários.	5 = Concordo.
		6 = Concordo totalmente.
		7 = Não se aplica.
		8 = Não sei responder.