Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Descoberta de conhecimento através da análise e mineração em dados do Enem

Klark Gable Souza Porto

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Jan Mendonça Corrêa

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Descoberta de conhecimento através da análise e mineração em dados do Enem

Klark Gable Souza Porto

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Jan Mendonça Corrêa (Orientador)
CIC/UnB

Prof.a Dr.a Letícia Gonçalves Nunes Coelho
IFD/UnB

Prof. Dr. Pedro Antonio Dourado Rezende
CIC/UnB

Prof. Dr. José Edil Guimarães de Medeiros
Coordenador do Curso de Engenharia da Computação

Brasília, 9 de Julho de 2019

# Dedicatória

Dedico este trabalho ao meu avô Manoel Felipe de Sousa (In Memorian), exemplo de vida, e especialmente à minha mãe que lutou sempre para que eu alcançasse meus objetivos e não deixou de me apoiar nessa etapa da vida.

# Agradecimentos

Sou grato a todos os professores que contribuíram com a minha trajetória acadêmica, especialmente ao Prof. Dr. Jan Mendonça Correa pela orientação. Agradeço à minha esposa, que jamais negou apoio, incentivo e carinho. Por fim, quero agradecer família, amigos e colegas de trabalho que sempre estiveram ao meu lado.

# Resumo

O Exame Nacional do Ensino Médio (Enem) pode ser usado como parâmetro de avaliação de desempenho e servir como indicador de qualidade do ensino básico no Brasil. Esse trabalho tem como objetivo fazer uma análise sobre os microdados do Enem 2017 e descobrir padrões de desempenho, médias de nota, questões com maior acerto e erro, estatísticas por regiões do país, entre outras.

O Enem é um requisito de entrada para algumas Instituições de Ensino Superior e é feito por milhões de candidatos anualmente. Além dos indicativos de desempenho em diversas áreas de conhecimento, esse exame coleta informações importantes dos inscritos, como idade, gênero, raça, estado de residência, escola onde concluiu o ensino médio e também dados socioeconômicos.

Por meio do processo de mineração de dados e análise estatística, esse trabalho tenta descobrir os fatores que têm impacto no desempenho do Enem. Os dados foram retirados do portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Foi possível retirar informações acerca do nível de conhecimento dos estudantes através da avaliação de desempenho de questão por questão em partes da prova do Enem, além de realizar comparativos sobre a distribuição de inscritos, sobre notas entre as regiões do Brasil, e sobre as notas separadas por gênero, raça/cor, renda e tipo de escola.

**Palavras-chave:** Mineração de Dados, Informação, Análise Estatística, Microdados do Enem

# Abstract

The National High School Examination (Enem) can be used as a parameter of performance evaluation and serve as an indicator of the quality of basic education in Brazil. The purpose of this work is to analyze the Enem 2017 microdata and discover patterns of performance, grade average, questions with greater accuracy or error, among others.

Enem is a requirement for entry into some Higher Education Institutions and is taken by millions of candidates. In addition to performance indicators in several areas of knowledge, this exam collects important information from enrollees, such as age, gender, race, state of residence, high school, and socioeconomic data.

Through the process of data mining and statistical analysis, this paper attempts to discover the factors that have an impact on Enem's performance. The data were taken from the portal of the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

It was possible to withdraw information about students' level of knowledge through the evaluation of performance of question-by- question in parts of the Enem test, in addition to comparing the distribution of enrollees, grades between Brazilian regions, and grades separated by gender, race/color, income and type of school.

**Keywords:** Data Mining, Information, Statistics Analysis, Enem microdata

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

**ANA** Avaliação Nacional da Alfabetização.

**BNCC** Base Nacional Comum Curricular.

**CBIE** Congresso Brasileiro de Informática na Educação.

**CRISP-DM** Cross-Industry Standard Process for Data Mining.

**Enem** Exame Nacional do Ensino Médio.

**Fies** Fundo de Financiamento Estudantil.

**IES** Instituições de Ensino Superior.

**INEP** Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

**KDD** *Knowledge Discovery in Databases.*

**Ideb** Índice de Desenvolvimento da Educação Básica.

**MEC** Ministério da Educação.

**OCDE** Organização para a Cooperação e Desenvolvimento Econômico.

**PNE** Plano Nacional de Educação.

**Prouni** Programa Universidade para Todos.

**Saeb** Sistema de Avaliação da Educação Básica.

**SGBD** Sistema Gerenciador de Banco de Dados.

**Sisu** Sistema de Seleção Unificada.

**SVM** *Support Vector Machine.*

**Weka** *Waikato Environment for Knowledge Analysis.*

# Capítulo 1

# Introdução

A Constituição da República Federativa do Brasil de 1988 enuncia que "a educação é um direito de todos e dever do Estado e da família". Portanto, ela deve promovida e incentivada com a participação da sociedade, levando em conta o desenvolvimento da pessoa, assim como seu preparo para o exercício da cidadania e sua qualificação para o trabalho [1]. O processo de ensino começa na educação infantil, segue para o ensino fundamental até chegar ao nível médio, que é o caminho para ingresso em uma Instituições de Ensino Superior (IES). O artigo 206 da Constituição fala sobre a incumbência da União em garantir a qualidade do ensino em todo o território nacional e uma das iniciativas foi começar a promover o Exame Nacional do Ensino Médio (Enem).

O Enem foi criado em 1998 [2] como avaliação de desempenho dos estudantes de escolas públicas e particulares do Ensino Médio. Porém, desde 2009, o Enem agregou outra função e tornou-se também uma avaliação que seleciona estudantes de todo o país para Instituições Federais de ensino superior e para programas de assistência do Governo Federal, como o Sistema de Seleção Unificada (Sisu), o Programa Universidade para Todos (Prouni) e o Programa do Fundo de Financiamento Estudantil (Fies). Esses programas visam pessoas que querem ingressar em universidades públicas ou que precisam de auxílio do governo para pagar a mensalidade da universidade particular [3].

A prova do Enem tem duração de dois dias e acontece todo ano, normalmente entre o final de outubro e as primeiras semanas de novembro. O conteúdo avaliado no exame envolve todas as matérias dos três anos do ensino médio, organizadas em quatro áreas do conhecimento [4]:

- Matemática e suas Tecnologias: Cálculos, Teoria dos conjuntos, etc.
- Ciências da Natureza e suas Tecnologias: Química, Física e Biologia.
- Ciências Humanas e suas Tecnologias: Geografia, História, Filosofia, Sociologia e Conhecimentos Gerais.

- Linguagens, Códigos e suas Tecnologias: Língua Portuguesa (Gramática e Interpretação de Texto), Língua Estrangeira Moderna, Literatura, Artes, Educação Física e Tecnologias da Informação.

O exame também é composto por uma redação em Língua Portuguesa, na qual os participantes devem elaborar um texto dissertativo-argumentativo sobre o tema proposto. Tal tema só é revelado na hora da prova e normalmente envolve um assunto importante ou polêmico da atualidade [4].

O portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) [5], órgão vinculado ao Ministério da Educação (MEC) [6], disponibiliza o acesso aos dados das provas do Enem, organizados em forma de microdados para cada ano. Os dados mais antigos disponíveis no portal são do Enem 1998. Os microdados possibilitam a realização de análises, com auxílio de técnicas de mineração de dados, de forma a obter padrões de desempenho de estudantes de todo o Brasil. Este trabalho pretende analisar os dados do Enem aplicado no ano de 2017.

## 1.1 Justificativa

O Enem é uma exame com a participação de milhões de candidatos por ano e é utilizada para avaliar a qualidade do ensino médio no país. Como essa fase escolar é a porta de entrada para os próximos níveis de ensino, a ocorrência de deficiência estudantil aqui acarreta consequências para o desenvolvimento de uma futura geração de alunos do ensino superior, técnico ou tecnólogo. O que deriva em prejuízo na qualificação desses estudantes no mercado de trabalho.

Nesse sentido, surge a necessidade de obtenção de estatísticas relacionadas ao ensino médio e de estudos capazes de fornecer informações para as Instituições de Ensino, que podem usar tal conhecimento para definir suas metas, avaliar seus alunos e fazer possíveis melhorias. Assim, a mineração de dados é essencial para tentar descobrir padrões e informações úteis, além dos dados já apresentados nas bases do governo.

## 1.2 Problema

Já existe na literatura estudos sobre o Enem e os dados do INEP, como relatado por Rodrigues [7] em 2014. O problema proposto é como utilizar ferramentas de mineração de dados para estudar os microdados do Enem aplicado em 2017.

## 1.3 Hipótese

A hipótese é que, por meio da mineração de dados, é possível identificar padrões relacionados ao desempenho dos alunos, além de comparar os resultados das prova e realizar comparativos entre os alunos em diversas características diferentes.

## 1.4 Objetivos

O objetivo principal deste trabalho é analisar os microdados do Enem aplicado em 2017, buscando fatores que influenciam no desempenho dos alunos, de forma a obter informações sobre a qualidade do ensino médio no Brasil e estatísticas de desempenho nas diferentes áreas de conhecimento.

O objetivo secundário é criar um sistema de análise de microdados do Enem que possa ser reutilizado, bastando alterar os dados de entrada, o que servirá como base para estudos futuros.

## 1.5 Descrição dos Capítulos

O Capítulo 2 apresenta uma visão geral da situação do ensino no Brasil e dos princípios relacionados ao Enem, além de mostrar trabalhos realizados com análise de dados e pesquisas críticas com tema similar ao deste trabalho. O Capítulo 3 aborda os conceitos básicos de banco de dados e de técnicas de mineração de dados, a importância dos sistemas de informação, além das ferramentas utilizadas neste trabalho. Os Capítulos 4 a 5 descrevem a metodologia adotada para a análise e mineração dos dados do Enem, apresentam o processo e análise estatística realizados, e por fim, mostram os resultados obtidos. As conclusões e sugestões para trabalhos futuros são apresentadas no Capítulo 6.

# Capítulo 2

# O Ensino no Brasil

Neste capítulo será feita uma discussão sobre o cenário do ensino no Brasil e sobre o desempenho geral dos estudantes, com foco no Enem. Na Seção 2.1 é mostrado o cenário da educação básica no Brasil e na Seção 2.2, o cenário da educação em nível superior. A Seção 2.3 trata dos métodos de avaliação de desempenho de alunos, como os tipos de exames. Finalmente, na Seção 2.4, é feito um levantamento dos estudos sobre dados do Enem ao longo dos anos que aborda também a influência de trabalhos anteriores na melhora ou busca de soluções de problemas no ensino básico e superior.

## 2.1 O Cenário da Educação Básica

No portal do INEP encontra-se o Censo Escolar [8], o maior levantamento de dados estatísticos da educação brasileira, que é realizado anualmente com a colaboração de todas as instituições públicas e privadas com oferta de ensino infantil, fundamental e médio. O censo abrange os diferentes níveis da educação básica e profissional (ensino regular, educação especial, educação de jovens e adultos e educação profissional). Faz-se a coleta de dados das escolas em caráter declaratório em duas etapas. A primeira consiste em preencher a "Matrícula Inicial", que se coleta informações sobre os estabelecimentos de ensino, turmas, alunos e profissionais escolares em sala de aula. A segunda etapa trata do preenchimento de informações sobre a situação do aluno e considera os dados sobre o movimento e rendimento escolar ao final do ano letivo. Para se compreender a situação educacional utiliza-se um conjunto amplo de indicadores para monitorar o desenvolvimento da educação brasileira, como o Índice de Desenvolvimento da Educação Básica (Ideb) [9], as taxas de rendimento e de fluxo escolar, a distorção "idade-série", entre outros, que servem de referência para o que denominam de metas do Plano Nacional da Educação (PNE) [10], que podem ser acompanhadas no Observatório do PNE [11]. Esses indicadores são calculados com base nos dados do Censo Escolar. Após a coleta da 1ª etapa

do Censo Escolar, que seria a Matrícula Inicial, os dados informados são consolidados e publicados no Diário Oficial da União.

Pelos dados informados no Boletim n°15 do mês de setembro de 2018 [12], em 2017, de todas as escolas do país, 4.040 não preencheram o Censo Escolar. Dentre essas instituições, encontra-se aquelas que em anos anteriores responderam ao Censo, porém em 2017 não o fizeram. O maior número de escolas faltantes ficou na região Nordeste, com 1.163 escolas.

Os resultados do Censo mostrados na Sinopse estatística encontrados no site do INEP dividem os dados referentes a estabelecimento, matrícula, função docente, movimento e rendimento escolar, para as diferentes modalidades de ensino brasileiras e a tabela em formato de planilha pode ser acessada para análise.

Sendo assim, com os dados obtidos, o INEP disponibiliza o resumo técnico [13] que em um só indicador (Ideb), mostra os resultados de dois conceitos importantes para a qualidade da educação: o fluxo escolar (progressão dos anos na educação básica) e as médias de desempenho nas avaliações. Esse indicador obedece a uma fórmula bastante simples relacionando as notas de provas de língua portuguesa e matemática e é padronizado em uma escala de zero a dez. Depois, a média dessas duas notas é multiplicada pela média (harmônica) das taxas de aprovação das séries da etapa (anos iniciais, anos finais e ensino médio), que, em percentual, varia de 0 (zero) a 100 (cem).

Com relação ao ensino médio que é o foco da pesquisa, em 2017 pela primeira vez o INEP passou a calcular Ideb para escolas de ensino médio. Nas tabelas de resultados nota-se que depois de três edições consecutivas sem alteração, o Ideb do ensino médio avançou 0.1 ponto em 2017. Apesar do crescimento observado, o país está distante da meta que foi projetada. Nesse cenário, cinco estados tiveram redução no valor do Ideb. O registro de destaque vai para o Espírito Santo, estado com o melhor desempenho no país. Com relação às escolas nas regiões Norte e Nordeste, cerca de 30% das escolas estão em limite inferior. Os estados com as maiores proporções de escolas com Ideb igual ou superior a 4,2 são: Espírito Santo (60,3%), Goiás (56,9%), São Paulo (48,0%) e Pernambuco (46,1%). Tratando de escolas de rede privada, ela participa com 12,2% da matrícula no ensino médio e alcançou em 2017 um desempenho 2,3 pontos superior ao obtido pela rede estadual, ou seja, Ideb igual a 5,8 contra 3,5 da rede estadual. Em função da estabilidade do desempenho no Ideb, a rede privada não alcançou nenhuma meta proposta para o ano de 2017, sendo registrada, ainda, uma queda de desempenho nas escolas de Roraima. Os maiores resultados foram obtidos pelas escolas privadas de Minas Gerais (6,3), Espírito Santo (6,1), Santa Catarina (6,0) e Distrito Federal(6,0).

## 2.2 O Cenário da Educação Superior

É importante analisar o cenário da educação superior, pois o foco do Enem é avaliar os alunos vindos do ensino médio e servir de acesso para as instituições de ensino superior. Portanto, a relação de alunos do nível superior está diretamente ligada ao desempenho deles no Enem, se considerar que muitos utilizam a nota do Enem para ingressar nas Universidades.

O censo da Educação Superior que é realizado anualmente também se encontra no site do INEP. Esse censo é o instrumento de pesquisa considerado o mais completo do Brasil sobre as Instituições de Educação Superior (IES) e a coleta de dados deles tem o objetivo de oferecer informações sobre as tendências do setor. Para a pesquisa é importante saber sobre a situação do ensino superior, pois o Enem é a porta de entrada, e pode demonstrar alguma fraqueza do ensino médio que vem prejudicar o desempenho dos estudantes no ensino superior.

Pelos dados divulgados pelo INEP em 2018, no Resumo Técnico: Censo da Educação Superior 2015 [14], verifica-se que há uma predominância de Instituições de Ensino Superior (IES) privadas no país (2.069), seguidas pelas estaduais (120), federais (107) e municipais (68). Foi notado que houve uma pequena diminuição do total de IES nos últimos anos.

O INEP divide os resultados do censo em subseções, uma para falar sobre as IES, outra para os cursos de graduação, outra referente aos alunos e uma última com relação aos docentes. A seção sobre os alunos fica organizada em Matrículas, Ingressantes e Concluintes que correspondem às estatísticas de condições que os alunos tem vínculo, considerando o período de 2013 a 2015.

Em 2015, a graduação alcançou o total de 8.027.297 matrículas, sendo 75,7% em categoria privada. Foi possível ver que a variação percentual do crescimento do número de matrículas, de 2015 em relação a 2014 foi menor do que de 2014 com relação a 2013.

A Tabela 2.1 mostra a distribuição do número de cursos de graduação, separando em áreas gerais do conhecimento e também pela categoria administrativa das IES. Verifica-se que se manteve o padrão de maior oferta de cursos públicos na área de Educação (38.7%) e, na categoria privada, predominam os cursos na área de Ciências Sociais, Negócios e Direito (36.6%).

O censo concluiu também que o graduando em 2015, é predominantemente do sexo feminino, representando 55,2% dos alunos ingressantes, e contando os que conseguem concluir a graduação, 64,1% são do sexo feminino. Nos cursos presenciais predominam o bacharelado e o turno noturno. No caso dos cursos a distância, maioria de matrículas e ingressantes são nas licenciaturas. Para os concluintes, predominam vínculos a cursos tecnólogos. Na mesma direção de resultados de edições anteriores do Censo Brasil (2012,

Tabela 2.1: Número de cursos de graduação, por categoria administrativa, segundo a área geral do conhecimento (OCDE) - Brasil - 2015. Fonte: INEP.

| **Área Geral do Conhecimento** | **Total** | **Pública** | **Privada** |
|---|---|---|---|
| Total | 33.501 | 10.769 | 22.732 |
| Agricultura e Veterinária | 959 | 578 | 381 |
| Ciências Sociais, Negócios e Direito | 9.935 | 1.608 | 8.327 |
| Ciências, Matemática e computação | 3.292 | 1.193 | 2.099 |
| Educação | 7.626 | 4.165 | 3.461 |
| Engenharia, Produção e Construção | 4.937 | 1.543 | 3.394 |
| Humanidade e Artes | 1.568 | 605 | 963 |
| Saúde e Bem-Estar Social | 4.029 | 814 | 3.215 |
| Serviços | 1.155 | 263 | 892 |

2013, 2014, 2015), pode-se dizer que a modalidade a distância abriga alunos cujos ingressos se dão, em média, mais tardiamente. Também para esta modalidade, verifica-se uma maior amplitude da distribuição das idades atendidas nas diferentes condições de matrícula, ingressante e concluinte.

Com relação às funções docentes, também avaliadas pelo censo, nota-se que nas Universidades, há predominância de doutores (51,6%), seguido da participação de mestres (32,3%). Nos centros universitários, são maioria os mestres (52,0%), seguidos do agrupamento de até especialistas (26,1%). Esse cenário assemelha-se grandemente ao das faculdades, com 46,9% de mestres e 36,6% do agrupamento de até especialistas. Finalmente, nos Institutos Federais e nos Centros Federais de Educação Tecnológica, verifica-se predominância de mestres (48,6%), seguida de doutores (30,4%). Nas IES públicas, o doutorado é o grau de formação mais expressivo (57,9%) do total de funções docentes em exercício das IES públicas. Nas IES privadas, por sua vez, o mestrado aparece com maior participação (48,2%) notadamente nas faculdades e, secundariamente, nas universidades.

## 2.3 A Necessidade de Avaliar Desempenho

Para buscar soluções relacionadas ao desempenho dos alunos no ensino médio, é importante entender quais são os tipos de avaliações de aprendizagem e como elas se aplicam no país. É um tema muito discutido entre educadores e ainda não chegaram em um consenso do que seria a melhor forma de avaliação de aprendizado. Geralmente o ato de

avaliar está associado com realização de provas, atribuição de notas, e tem ligação com aprovação e reprovação de alunos.

A avaliação, no livro de Luckesi [15], é dividida em três tipos:

- Avaliação Somativa: Tem como objetivo de classificar o aluno em aprovado ou reprovado com relação ao conteúdo visto até o final da unidade, semestre ou ano.
- Avaliação Formativa: Vem com o intuito de informar o aluno e professor sobre o resultado da aprendizagem para propor reformulações no ensino.
- Avaliação Diagnóstica: Trata-se de uma etapa do processo educacional para medir os conhecimentos aprendidos ou não pelos alunos, servindo como uma sondagem para alcançar os objetivos propostos.

Sendo assim, em outro livro de Luckesi [16] focado em avaliação da aprendizagem escolar, ele faz algumas observações sobre a prática educativa brasileira e denota que a maior característica dela é ser focada no que ele chama de "pedagogia do exame", ou seja, resumidamente aplicação de provas como forma de avaliar os alunos. Isso é visto claramente no terceiro ano do segundo grau, onde todas as atividades docentes e discentes estão voltadas para o treinamento em resolução de provas, ainda mais que soma com a pressão de preparar os estudantes para o ingresso nas Universidades. Luckesi ainda diz que o nosso sistema social se conforma com as notas obtidas nos exames e que o sistema de ensino "aparentemente" se importa apenas com os resultados gerais. Contudo, ele conclui que é necessário aprender a avaliar, ou seja, aprender os conceitos teóricos sobre avaliação através de estudos em livros e artigos relacionados para aprender a prática da avaliação e traduzir para o cotidiano somando com a própria experiência dos professores, cotidiano em sala de aula, podendo arriscar com experimentação e investigação.

Assim como o professor avalia seus alunos, o País também tem suas formas de avaliação de desempenho para acompanhar o desenvolvimento do ensino brasileiro e faz isso por meio dos exames padronizados aplicados pelo INEP. No sítio do projeto "Todos pela Educação" [17], fundado em 2006 no Museu do Ipiranga, há uma descrição das principais avaliações de educação básica no Brasil. Essas informações também se encontram no portal do MEC [18], porém estão desatualizadas:

- Educação infantil: Para o ensino infantil foi criada a avaliação chamada Saeb (Sistema de Avaliação da Educação Básica) que entra em vigor em 2019. Era uma avaliação já prevista nas Diretrizes Curriculares Nacionais para Educação Infantil e no PNE que atrasou na implantação.
- Anos Iniciais do Ensino Fundamental: Aqui se aplica a ANA (Avaliação Nacional da Alfabetização). Esse exame está passando por uma transição devido a criação

da BNCC (Base Nacional Comum Curricular) que fez uma antecipação do fim do ciclo em uma parte do ensino fundamental. A mudança altera o ano avaliado e nome da prova que possivelmente se chamará "Saeb dos Anos Iniciais do Ensino Fundamental", mantendo o foco da avaliação.

- Anos Finais do Ensino Fundamental: Há um exame padronizado aplicado desde 2005 aos estudantes do 5º e 9º anos do Ensino Fundamental da rede pública. São questões de língua portuguesa e matemática, além de um questionário socioeconômico. Esta era a avaliação conhecida como prova Brasil e receberá novo nome em 2019 também, passando para Saeb dos Anos Finais do Ensino Fundamental. Alunos do 9º ano passarão a responder questões de ciências humanas e ciências da natureza.

- Ensino Médio: Alunos do ensino médio também prestam o Saeb. Até 2015, a prova só avaliava um pequeno grupo de jovens que serviam de representantes do país. Em 2017 passou a ser obrigada a avaliar todos. Aqui se destaca a outra avaliação que é o Enem. Como referenciado em outras partes desse trabalho, o primeiro programa a incorporar os resultados do Enem foi o Prouni. O caráter vestibular foi intensificado apenas em 2010, quando a nota passou a ser utilizada como único critério de seleção via Sisu. Em 2014 integraram o Fies para assegurar custeio em instituições privadas. O Enem é opcional, precisa de inscrição, diferentemente da prova do Saeb.

## 2.4 Estudos Sobre Microdados do Enem

Há diversos trabalhos publicados que utilizaram mineração em dados educacionais abertos. No panorama nacional, segundo Rodrigues [7], passou de 20 trabalhos relatados nos principais periódicos e conferências de 2006 a 2010 para mais de 10 trabalhos publicados apenas nos anais do Congresso Brasileiro de Informática na Educação (CBIE) [19] em 2015. Alguns trabalhos merecem ser analisados por tratarem diretamente do assunto Enem.

Em uma dissertação de mestrado defendida por Jorge Luiz [20], ele propôs uma ferramenta para a obtenção e análise de dados do ENEM, usando os dados de 2012. A ideia final dele era capacitar outras pessoas para obter os dados e desenvolver suas próprias ferramentas utilizando o programa *Excel*. A ferramenta serve para conhecer, analisar, tomar decisões e criar estratégias para o aprimoramento do projeto pedagógico do Estabelecimento de Ensino. Ele filtrou os dados, separou em disciplinas e fez comparativos de notas, apresentando os gráficos e ao fim ele disponibiliza um aplicativo feito em Excel que ao inserir o código da escola, os gráficos são gerados automaticamente para o interessado.

No trabalho feito por Tancicleide Simões [21] em 2015, aplicou-se técnicas e métodos de mineração de dados a fim de descobrir padrões e regras de associação em dados estatísticos do Enem nos anos de 2013 e 2014 no âmbito da região Nordeste. Utilizaram os algoritmos *Apriori* e *J48* para análise de desempenho e outras ferramentas de mineração de dados como *IBM Intelligent Miner7* e *Oracle Data Mining, o SQL Server Analysis Services* e também *Rapid Miner, SAS Enterprise Miner e a Weka* que possuem interface gráfica com o usuário. No trabalho dela, os dados foram organizados em quatro bases distintas considerando a região Nordeste e o estado de Pernambuco nos anos de 2013 e 2014. Conseguiu gerar 30 regras com o algoritmo *Apriori* e algumas regras especiais forma destacadas:

- Inscritos que realizaram o ensino fundamental somente em escola pública apresentam forte tendência de terem realizado o ensino médio apenas em escolas públicas (grau de confiança de 96%);
- Inscritos que cursaram o ensino médio apenas em escolas públicas apresentam forte tendência de serem oriundos de famílias de classe E, ou seja com renda familiar de até dois salários mínimos (grau de confiança de 94%);
- Inscritos que indicaram ter realizado seus estudos na modalidade regular têm forte tendência de residirem na zona urbana (grau de confiança de 83%);
- Inscritos oriundos do interior têm forte tendência de pertenceram a famílias de classe E (grau de confiança de 83%).

Com relação às notas, focadas apenas no cenário de Pernambuco e observando apenas notas não nulas de Matemática (304.955), notou-se que aproximadamente 70%(215.062) das notas estão na faixa entre 300 e 500 pontos. 30% (90.999) de todos participantes obtiveram notas entre 300 e 400 pontos, o que é abaixo da nota média nacional (473,5) e até mesmo da média regional (442,7). 40% (124.063) de todos participantes obtiveram notas entre 400 e 500 pontos. O trabalho concluiu que as análises de desempenho utilizando o algoritmo *Apriori* endossam uma forte relação entre a renda da família e o desempenho dos estudantes no exame, sobretudo, os provenientes de escolas públicas.

O trabalho do Rafael Alves [22], foi direcionado às notas das redações e pegou a base de dados do Enem 2016. A classe analisada foi a nota da redação, categorizada como: baixo, médio, alto e nulo. Os modelos finais foram treinados e testados por meio dos algoritmos: Naive Bayes e J48. Esses algoritmos foram utilizados através do pacote de *software* Weka (Waikato Environment for Knowledge Analysis). Ele separou os dados para o estado de Santa Catarina e para a cidade de Araranguá. Ao longo do trabalho ele tentou gerar modelos para predição do desempenho da redação do Enem, e também gerar árvores de

decisão, buscando auxiliar na melhora do desempenho dos alunos na prova de redação. Ele concluiu que apesar do crescimento na realização de pesquisas na área de mineração de dados educacionais, ainda é uma área que necessita ser mais explorada e estudada. Uma prova disso é que não se encontrou trabalhos semelhantes ao dele, apesar de como relatado, as redações do Enem serem decisivas para o desempenho final do candidato e serem uma das maiores preocupações de alunos e professores.

Em um pequeno artigo apresentado na 9ª Jornada Científica e Tecnológica do Instituto Federal do Sul de Minas (6° Simpósio da Pós-Graduação), foi feita a utilização de técnicas de mineração sobre dados do Enem 2015 [23] e concluíram que algumas variáveis, tais como, nível socioeconômico e taxa de permanência dos alunos influenciam no resultado das escolas. Para avaliar com relação ao fator socioeconômico, utiliza-se a técnica com indução de regras nos dados. Eles utilizaram os dados do INEP e aplicaram as regras de associação e classificação para os experimentos, usando algoritmos como o *Apriori* e a metodologia de teste *Cross Validation*. Dessa forma, o algoritmo *Apriori* detectaria regras das instituições que correspondem à maioria dos dados. Portanto, eles decidiram analisá-los de forma isolada por tipos de escolas.

Concluíram ao fim que o Support Vector Machine (SVM) possui melhor taxa de acerto, mas tem o segundo pior tempo de criação do modelo, demorando 65,55 segundos. O algoritmo *J48* (Árvore de Decisão) teve um bom resultado interessante, pois alcançou a segunda melhor taxa de acerto (77,21%) e o tempo de criação do modelo é de 0,76 segundos.

# Capítulo 3

# Sistemas de Informação e Mineração de Dados

Neste capítulo, serão apresentados os principais conceitos utilizados nesse trabalho, relacionados à análise de informações, às tecnologias para a geração de conhecimento usando dados e à importância desses tópicos para processos decisórios e em um ambiente organizacional. Além disso, será abordada a ferramenta principal desse trabalho através do conhecimento básico sobre mineração de dados como apoio para análise e descoberta de informação. Por fim, serão vistas as etapas, as técnicas e alguns dos algoritmos utilizados em mineração de dados.

Na Seção 3.1 serão apresentados os conceitos de dado, informação e conhecimento. Na Seção 3.2, é apresentado o conceito de sistema de informação e seus benefícios. Na Seção 3.3, listam-se algumas características desejáveis para o gerenciamento de informações. Na Seção 3.4, o conceito geral de mineração de dados é descrito junto com as etapas do processo de extração de conhecimento, e a arquitetura de um sistema de mineração de dados é apresentada. Na seção 3.5, é apresentado o modelo CRISP para mineração de dados. A Seção 3.6, mostra algumas das técnicas de mineração de dados. Por fim, na Seção 3.7, a linguagem *Python*, bibliotecas e alguns dos algoritmos empregados para análise de dados e descoberta de informações são apresentados.

## 3.1 Dado, Informação e Conhecimento

"Dado", "Informação" e "Conhecimento" são conceitos da Ciência da informação, área que se preocupa com a análise, coleta, classificação e manipulação de dados de forma geral. Essa área estuda a informação desde a sua gênese até o momento da transformação em conhecimento. Alguns autores enxergam esses conceitos como sinônimos. Porém,

cada uma tem sua característica distinta na participação no processo de manipulação de informação.

Claude Elwood Shannon, que é considerado o "pai da teoria da informação", em seu artigo de 1948 [24] fez um estudo para descobrir a melhor forma de codificar a informação que um remetente precisa transmitir. Shannon fundou a teoria matemática da informação e propôs uma medida para a incerteza sobre espaços desordenados. Para ele, informação está ligado ao que é transmitido por um canal de comunicação de fonte ao destino, onde a incerteza do que não é antecipável em relação ao que pode ser esperado pelo receptor determina essa medida.

Para Siqueira [25], dado é a menor unidade, ou unidade primária que compõe a informação. A informação pode assumir diversos valores que dependem da análise e avaliação humana para ser interpretada.

Para Russo [26], os dados são sinais que não foram processados, correlacionados, integrados, avaliados, ou interpretados de qualquer forma, e, por sua vez, representam matéria-prima a ser utilizada na produção de informações. A informação seria como dados contextualizados, que visam fornecer uma solução para determinada situação de decisão.

## 3.2 Sistemas de Informação Aplicado à Análise de Informações

Há uma necessidade natural de acesso rápido à informação, tanto em consultas como acompanhamento de algum processo ou atividade. Assim, os sistemas de informação surgem para proporcionar velocidade e eficiência na gestão de qualquer que seja o meio informacional.

Para Laudon e Laudon [27], um sistema de informação pode ser definido tecnicamente como um conjunto de componentes inter-relacionados que coletam, processam, armazenam e distribuem informações destinadas a apoiar a tomada de decisões, a coordenação e o controle de uma organização. Além disso, auxiliam os gerentes e trabalhadores a analisar problemas, visualizar assuntos complexos e criar novos produtos. Esse autor faz um esclarecimento sobre as dimensões dos sistemas de informação, que são as organizações (sistemas de informação são parte integrante das organizações), pessoas (um sistema é tão bom quanto as pessoas que o formam) e a tecnologia (*hardware* e *software* atuais).

O'Brien [28] define três papéis para um sistema de informação, que envolvem as operações, decisão gerencial e a vantagem estratégica:

- Operações envolvem a produção de diversidade de produtos de informação para uso interno e externo e trata do monitoramento e controle de processos físicos;
- Decisão gerencial é mais voltada para fornecer informações e apoio a tomada de decisão com relatórios e análise de dados por exemplo.
- Vantagem estratégica está ligada com o sucesso a longo prazo.

Para ele, o sistema se compõe de entrada (captação e reunião de elementos que entram no sistema), processamento (converte para produto) e saída (destino final). O principal objetivo dos sistemas de informação vem no sentido de criar e distribuir conhecimento de uma maneira que possa resolver os problemas das organizações.

## 3.3 Características da Informação

Stair e Reynolds [29] descrevem algumas características das informações que as tornam valiosas em uma organização. Elas tem que ser de fácil acesso aos usuários autorizados e que tenham exatidão em seu conteúdo para eliminar imprecisões. Precisam ser completas, econômicas e flexíveis para poder proporcionar o uso em diversos propósitos. Além disso, as informações tem que trazer confiança, segurança e simplicidade agregada para garantir o não acesso de usuários não autorizados. Por último, eles dizem que a informação precisa ser entregue em tempo hábil e serem verificáveis, ou seja, tem que ser permitida uma análise sobre ela para assegurar que estão corretas.

## 3.4 Mineração de Dados e o Processo de Extração de Conhecimento

A facilidade de armazenamento em sistemas de informação tem contribuído para que a quantidade de dados só aumente, o que começa a trazer mais importância ainda para a mineração de dados. Os dados são produzidos por qualquer pessoa e qualquer meio que tenha acesso computacional, em seus diversos tipos de aparelhos. A quantidade de dados cresce em ritmo mais acelerado do que a evolução do entendimento sobre eles. Assim, surge a preocupação em identificar padrões nos dados armazenados, e propor ferramentas de automatização para isso. Para isso, foram desenvolvidas várias ferramentas de gerência de dados, que apresentam uma metodologia de pesquisa e análise, utilizando das áreas de computação e estatística para tratamento dos dados. Sendo assim, os sistemas conseguem isso por meio dos *softwares* específicos para otimizar captura,verificação, filtragem e organização dos dados.

Um artigo sobre mineração de dados de Côrtes et al. [30] comenta que o conceito de mineração está ligado aos princípios da "Análise Exploratória de Dados (*Exploratory Data Analysis* - EDA)" e as organizações vem se tornando muito eficientes em armazenar grande quantidade de dados, porém poucas se preocupam em usá-los para descoberta de conhecimento e solução de problemas. Nesse mesmo artigo, são apresentadas diversas definições encontradas na literatura sobre mineração de dados, e em seguida a própria visão dos autores: "Mineração de dados é um processo altamente cooperativo entre homens e máquinas, que visa a exploração de grandes bancos de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamento entre variáveis, conhecimentos esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressividade estatística". Aqui estão algumas definições citadas por Côrtes et al.:

- Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens (Sholom M. Weis, Nitim Indurkhya–1999).

- Mineração de dados é a exploração e análise de dados, por meios automáticos ou semi-automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes (Michael J. A. Berry; Gordon Linoff–1997).

- Mineração de dados, em poucas palavras, é a análise de dados indutiva (Jesus Mena–1999).

- Mineração de dados é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, frequentemente desconhecidos, a partir de grande quantidade de dados armazenada em bancos de dados (Bhavani Thuraisingham–1999).

- Mineração de dados, de forma simples, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados (Jiawei Han, Micheline Kamber–2001).

Além desses, outros autores como Goldschmidt e Passos [31] definem Mineração de Dados como uma das etapas da Descoberta de Conhecimento em Bases de Dados, enquanto Groth [32] afirma que se trata do processo de encontrar tendências e padrões nos dados.

Han e Kamber [33] mostram que os sistemas constituem de vários componentes integrados e a arquitetura organizada principalmente por:

- Base de dados: Banco de dados, *Data Warehouse*, ou outros repositórios de informação. Aqui, técnicas de limpeza e integração de dados podem ser aplicadas.
- Servidores de banco de dados ou *Data Warehouse*: É responsável por trazer os dados relevantes, de acordo com a requisição do usuário.
- Base de conhecimento: Domínio de conhecimento usado para guiar a pesquisa, ou avaliar o interesse dos padrões resultantes. Pode incluir hierarquias de conceitos usados para organizar atributos ou valores de atributos em diferentes níveis de abstração.
- *Engine* de mineração de dados: É essencial para o sistema de mineração de dados e consiste de um conjunto de módulos funcionais para realização de tarefas tais como caracterização, associação, classificação, análise de cluster, evolução e análise de desvio.
- Módulo de avaliação de padrões: Esse componente emprega, tipicamente, medidas de interesse e interage com os módulos de mineração de dados para concentrar a busca em padrões interessantes. Pode ser integrado ao módulo de mineração de dados, levando em conta as especificidades dele.
- Interface de usuário: Esse módulo realiza a comunicação entre os usuários e o sistema de mineração de dados, permitindo que o usuário interaja com o sistema. Ele tem que especificar uma consulta ou tarefa, realizando a exploração da mineração de dados. A interface de usuário permite também, a consulta em esquemas de banco de dados, *data warehouse* ou estruturas de dados, avaliar padrões de mineração e visualizar os padrões em formatos diversos.

Por haver similaridade entre os conceitos de mineração de dados e descoberta de conhecimento em bases de dados (KDD - Knowledge Discovery in Database), vários autores abordam o assunto. Fayyad et al. [34], por exemplo dividem o processo de KDD em cinco etapas:

1. Seleção: Diz respeito à análise da disponibilidade e relevância dos dados. Ocorre a seleção do conjunto ou subconjunto de variáveis ou amostras de dados, onde o processo de descoberta será executado;
2. Pré-processamento: Fase de filtragem e limpeza dos dados, visando a remoção de dados inconsistentes, redundantes, valores faltantes ou extremos, que possam interferir nos resultados ou que sejam irrelevantes durante a mineração;

3. Transformação: Formatação adequada dos dados, para que os algoritmos possam ser aplicados corretamente;

4. Mineração de Dados: É a fase de mineração propriamente dita e consiste na exploração e análise dos dados. São aplicadas as técnicas de mineração de dados (classificação, sumarização, regressão, clusterização, dentre outras) com a finalidade de detecção dos padrões de comportamento dos dados e geração de novas descobertas;

5. Interpretação e Avaliação: Fase em que se alcança as informações desejadas. Os resultados obtidos são analisados, sintetizados, avaliados e organizados para apresentação e publicação.

## 3.5 Modelo CRISP-DM

O modelo CRISP-DM (*Cross-Industry Standard Process for Data Mining*) se refere a um método de organização industrial aplicado à mineração de dados. Há dois guias práticos que serviram de base aqui, um de autoria de Rüdiger Wirth et al. [35] e o outro de autoria da IBM Corporation [36].

O CRISP-DM pode ser visto como uma metodologia, onde é dividido em fases típicas de um projeto, relacionando tarefas e ordem de execução ou como um modelo de processo, onde há um ciclo de fases dependentes sem sequência estrita, ou seja, pode ocorrer movimentação entre as fases de acordo com o necessário. Há uma distinção também entre o modelo de referência (visão geral rápida de fases, tarefas e suas saídas e descreve o que fazer em um projeto de mineração de dados) e o guia de usuário (fornece formas mais detalhadas para cada fase e cada tarefa em uma fase e descreve como realizar uma mineração de dados no projeto).

O modelo de referência será visto a seguir, com mais detalhes. Nele, o ciclo de vida de um processo de mineração de dados é dividido em seis fases:

1. **Compreensão de negócios** - Fase inicial concentrada na compreensão dos objetivos e requisitos do projeto de uma perspectiva de negócios, para em seguida, converter esse conhecimento em uma definição de problema de mineração de dados e um plano básico projetado para alcançar os objetivos.

2. **Compreensão de dados** - Começa com a coleta de dados inicial e prossegue com atividades relacionadas aos dados, como identificar problemas na qualidade de dados, descobrir os primeiras intuições sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

3. **Preparação de dados** - Abrange todas as atividades necessárias para construir o conjunto de dados final, dados que serão aplicados na ferramenta de mineração, a partir dos dados iniciais. As tarefas incluem seleção de tabela, registro e atributo, bem como transformação e limpeza de dados para ferramentas de modelagem.

4. **Modelagem** - Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados para valores. Existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas possuem requisitos sobre a forma de dados. Assim, pode ser necessário voltar à fase de preparação de dados.

5. **Avaliação** - Neste estágio do projeto, já foi construído um modelo que parece ter qualidade a partir de uma análise de dados em perspectiva. Antes de prosseguir para a implantação final do modelo, é importante avaliar as etapas executadas para criá-lo, para ter certeza de que o modelo atinge adequadamente os objetivos de negócios. Um objetivo chave é determinar se existe algum problema comercial importante que não tenha sido suficientemente considerado. No final deste fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.

6. **Implantação** - A criação do modelo geralmente não é o final do projeto. O modelo deve ser apresentado de forma intuitiva e de fácil utilização. Dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar uma mineração de dados repetível. Em muitos casos, é o cliente, não o analista de dados, quem executa a implantação dos passos. No entanto, mesmo que o analista execute o esforço de implantação, é importante que o cliente compreenda quais ações precisam ser realizadas para realmente fazer uso dos modelos criados.

Nesse trabalho, essas etapas foram seguidas com algumas alterações na parte de modelagem, avaliação e implantação. Isso é desenvolvido e mostrado nos Capítulos 4 e 5.

## 3.6 Técnicas de Mineração de Dados

Como dito na Seção 3.4, a mineração de dados surge para descobrir padrões nos dados e obter conhecimento disso. Portanto, existem várias técnicas interessantes para especificar os tipos de padrões a serem encontrados.

Côrtes et al. [30] listam as formas de tratamento das funcionalidades ou técnicas de mineração:

- Descoberta de conhecimento e Predição;

- Classificação, Estimação, Predição, Afinidade em grupos, Agrupamentos e Descrição;
- Classificação, Detecção de sequência, Análise de dependência de dados e Análise de desvio;
- Previsão, Identificação, Classificação e Otimização;
- Descrição e Predição;
- Predição, Classificação, Agrupamento, Segmentação, Associação, Visualização e Otimização;
- Classificação, Estimação, Segmentação e Descrição;
- Predição, Detecção de desvio, Segmentação, Agrupamento, Análise de ligações e Regras de associação, Sumarização e Visualização e Garimpagem em textos.

A Figura 3.1 mostra, em camadas, as interações entre funcionalidades, técnicas e algoritmos, para esclarecer a interatividade entre elas e a mineração de dados.

É necessário definir as principais técnicas de mineração de dados e decidir qual usar para obter os melhores resultados. Côrtes et al. [30], classificam a funcionalidade em mineração de dados como Análise Descritiva e Análise de Prognóstico, como demonstrado na Figura 3.2. A análise descritiva busca descrever fatos relevantes desconhecidos dos usuários para validar o processo de mineração e seus resultados. A análise de prognóstico busca inferir resultados a partir dos padrões encontrados na análise descritiva.

Uma das vantagens de se focar as funcionalidades da mineração de dados dessa forma, segundo Côrtes et al. [30], diz respeito as facilidades que podem ser obtidas quando surge a necessidade de uma nova análise de dados. Neste caso, basta identificar que tipo de resultado se deseja chegar e imediatamente partir para identificação de que técnica aplicar.

A seguir, serão descritas as principais técnicas ou funcionalidades aplicadas em diversas etapas do processo de mineração de dados.

### 3.6.1 Classificação

A classificação serve para dividir os itens em categorias ou classes de destino para prever o que pode ocorrer dentro de uma classe. Isso seria útil por exemplo para identificar em uma escola, qual a turma mais indicada para certo tipo de aluno.

Han e Kamber [33], definem a classificação como um processo de duas etapas. Na primeira etapa (fase de treino), cria-se um classificador que descreve um conjunto de classes e conceitos pré-determinados. O algoritmo usado constrói esse classificador por meio

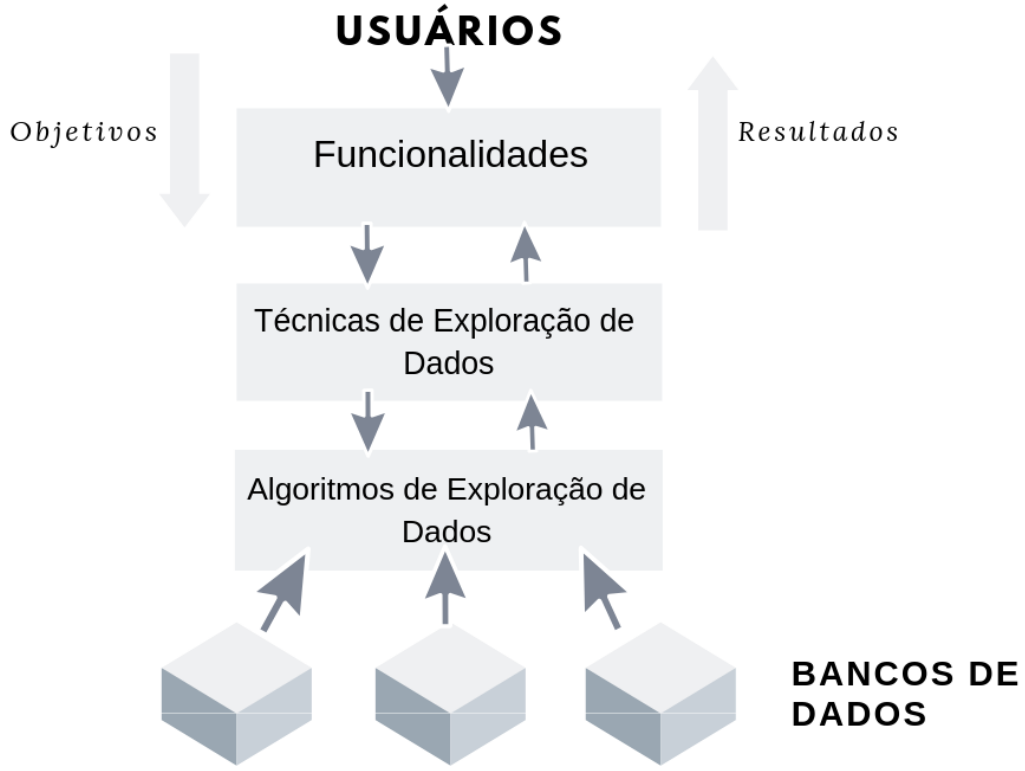



Figura 3.1: Interatividade entre as funcionalidades e técnicas da mineração de dados. Imagem baseada na de Côrtes et al. (Fonte: [30]).

do aprendizado baseado no conjunto de treinamento. A representação desse classificador é feita por meio de uma tupla (gerada pelo banco de dados e seus rótulos de classe associados) que conterá atributos e pertencerá a uma classe.

Na segunda etapa, é feita a aplicação do classificador sobre o conjunto de dados teste, e criada as tuplas de teste e os rótulos das classes. Então, a classificação ocorre efetivamente aqui, onde é necessário estimar a acurácia da classificação. Nessa fase, é verificada a porcentagem de tuplas do conjunto de testes que foram classificadas corretamente pelo classificador, que será a medida da acurácia do mesmo.

### Árvores de Decisão

As árvores de decisão são usadas para categorizar ou fazer previsão de dados. Gonçalves [37] define árvore de decisão como uma composição formada por um conjunto de regras de classificação. Cada caminho da raiz até uma folha representa uma destas regras. A árvore de decisão deve ser definida de forma que, para cada observação da base de dados,


Figura 3.2: Funcionalidades em mineração de dados. Imagem baseada na de Côrtes et al. (Fonte: [30]).

haja um e apenas um caminho da raiz até a folha. Han and Kamber [33], dizem que a indução por árvores de decisão é a aprendizagem de árvores de decisão a partir de classes rotuladas nas tuplas de treinamento. Cada nó interno (não-folha) da árvore indica um teste de um atributo, cada ramo representa o resultado de um teste e cada nó da folha possui um rótulo de classe.

### Classificação por Regras

As regras tem a ver com buscar padrões mais fortes de associação entre os itens. Para isso, utiliza-se o conceito de frequência e categorias. Han and Kamber [33], definem a classificação baseada em regras como o modelo representado por um conjunto de regras "se-então", o qual estuda o relacionamento entre itens de dados que ocorrem com frequência, proporcionando a obtenção de regras relevantes em um conjunto de itens aplicados a outros itens e identificando grupos que apresentam uma coocorrência entre si.

No artigo de Cassio e Joao [38], a classificação por regras consiste basicamente de dois passos: Primeiro, os dados de treinamento são analisados para se obtenha os itens mais frequentes. Em seguida, estes itens são usados para a geração das regras. Alguns estudos demostraram que esta técnica tem apresentado mais assertividade do que algoritmos tradicionais, como o C4.5. Alguns exemplos de algoritmos de classificação são: CBA (Classification-Based Association) e CMAR (Classification based on Multiple Association Rules). Uma nova abordagem chamada de CARM (Classification Association Rule Mining) também entra nessa classificação.

### 3.6.2 Clusterização

Clusterização significa dividir uma base de dados em agrupamentos (clusters), de modo que elementos similares fiquem no mesmo agrupamento. A técnica de clusterização trata de agrupar os dados em grupos bem definidos ao invés de apenas usar características particulares. Os critérios usados para definir as semelhanças e dividir em grupos varia pela definição do tipo do algoritmo a ser usado.

Groth [32] define clusterização como o processo de dividir um conjunto de dados em grupos distintos. Han e Kamber [33] tratam da clusterização como o processo de identificar regiões densas e esparsas no espaço do objeto, de forma automatizada, para descobrir padrões globais de distribuição e correlações que sejam interessantes entre os atributos dos dados. Para medir a qualidade dos clusters encontrados, em uma amostra por exemplo, usa-se as distâncias intra-cluster e extra-cluster. A intra-cluster diz respeito a quão distantes estão os elementos de um mesmo grupo. A extra-cluster diz respeito a quão distantes estão os clusters um dos outros. Os clusters são melhores quando seus elementos são mais próximos.

### 3.6.3 Predição

A predição é o processo de predizer algum valor, resultado ou comportamento futuro, baseado em valores já identificados anteriormente.

Han e Kamber [33] veem a predição como a aplicação de modelos de análise com a finalidade de determinar possíveis valores futuros para um determinado atributo. Para eles, há algumas distinções entre o processo de classificação e predição. Na classificação, o classificador consiste em prever rótulos categóricos e na predição, prever valores ordenados.

Segundo Cassio e Joao [38], A tarefa de predição é parecida com as tarefas de classificação e estimação, porém ela tem o objetivo de descobrir o valor futuro de um determinado atributo, como por exemplo predizer o valor de uma ação três meses adiante ou também predizer o vencedor de um campeonato baseando nas comparações estatísticas dos times.

Sendo assim, alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

### 3.6.4 Regressão

Pela definição de Goldschmidt e Passos [31], a regressão seria a busca por funções, lineares ou não, que mapeiem os registros de um banco de dados em valores reais.

Segundo Han e Kamber [33], a regressão tem como saída um valor numérico. É possível determinar qual a função que representa a regressão por meio de equações estatísticas.

### 3.6.5 Detecção de Desvios e Anomalias

Um problema comum que acontece na fase de pré-processamento de dados é a ocorrência de anomalias, que são desvios ou *Outliers* (valores extremos) na base de dados. A detecção de anomalias vem para identificar dados que não se enquadram nos padrões desejados para a mineração.

Um conceito muito usado na detecção de anomalias é o de "distância". Goldschmidt e Passos [31] relatam que, são especificados limiares de tolerância e sempre que a distância entre o registro em análise e o padrão médio representativo da base de dados excede um desses limiares, considera-se que tal registro é uma anomalia. Para Han e Kamber [33], os modelos de comportamento normal, são aqueles usados para comparar e identificar novos padrões que se afastariam dos perfis pré-definidos. Assim, ruídos e exceções podem ser identificados mais precisamente.

As anomalias podem ser retiradas do conjunto de dados usando técnicas de remoção com testes estatísticos que assumem um modelo de distribuição ou probabilidade para os dados. Caso não seja feita a remoção, os resultados podem ser comprometidos e causar distorções. Porém, ao fazer a remoção, pode ocasionar em um resultado tendencioso, já que as anomalias podem ser dados relevantes.

## 3.7 Ferramentas para Mineração de Dados

Cada ferramenta possui suas especificações técnicas e algoritmos próprios para organização dos dados. Um sistema de mineração deve ter recursos que facilitem a pesquisa em dados, portabilidade, capacidade de buscar dados em outros idiomas, sistema de classificação intuitivo, interface amigável, entre outras características que o tornem eficientes. Isso faz com que surjam diversas plataformas. Assim, os profissionais escolhem, de acordo com cada problema, a melhor ferramenta a ser usada.

No artigo de Rodrigo Santana [39], ele fala sobre as ferramentas e bibliotecas mais usadas nessa área:

- Linguagem Python e suas Bibliotecas;
- Anaconda: Bibliotecas embutidas;
- Linguagem R;
- RapidMiner;
- Weka.

A seguir, serão mostradas com mais detalhes algumas dessas ferramentas. As informações foram retiradas do mesmo artigo [39] e das documentações de cada *software*.

### 3.7.1 Python

Segundo o artigo [39], Python foi considerada a linguagem de programação mais usada pela comunidade de profissionais de Mineração de dados e *Big data*. A linguagem oferece uma sintaxe simples e objetiva, permitindo ao programador, focar no problema a ser resolvido sem se preocupar tanto com detalhes de implementações. Exige que o código fonte seja corretamente endentado e assim contribui para leitura e compreensão de forma mais clara e organizada. Além da grande comunidade no mundo inteiro, Python possui um vasto e variado conjunto de bibliotecas para se trabalhar com diversas áreas, desde computação científica, redes, segurança e análise de dados. A documentação do Python se encontra em [40] com descrição também dos módulos e bibliotecas. As ferramentas e bibliotecas Python mais usadas são: *Jupyter, Matplotlib, Pandas, Scikit-Learn*.

#### *Jupyter*

Aplicação baseada em protocolo cliente-servidor que permite a edição e execução de *notebooks* via *browser*. Os notebooks são documentos editáveis contendo células de códigos e elementos visuais como imagens, *links*, equações. A principal vantagem na utilização de *notebooks* é para a descrição de análises e seus resultados de forma dinâmica e interativa. Documentação disponível em [41].

#### *Matplotlib*

Biblioteca Python 2D utilizada na visualização e plotagem de gráficos. É muito utilizada para gerar os diversos tipos de gráficos possíveis como histogramas, gráficos de barras, gráficos de pizza. Documentação disponível em [42].

#### *Pandas*

Pandas é uma das bibliotecas mais utilizadas para análise de dados. Fornece ferramentas para manipulação de estruturas de dados de forma simples. Operações complexas que trabalham com matrizes e vetores podem ser facilmente realizadas em ótima performance. Documentação disponível em [43].

#### *Scikit-Learn*

Biblioteca Python para trabalhar com "Aprendizado de Máquina". Contém diversos algoritmos implementados, métodos de análise e processamento de dados e métricas de avaliação. Essa é uma biblioteca extremamente útil para o cientista de dados. Guia disponível em [44].

### 3.7.2 Anaconda

Anaconda é uma plataforma *open source* para *Data Science*. Contém centenas de pacotes embutidos e as principais bibliotecas Python e R já vem disponíveis. O usuário não precisa se preocupar em ficar instalando biblioteca por biblioteca e procurando uma IDE para trabalhar, pois no menu do Anaconda se encontra as principais ferramentas. Documentação disponível em [45].

### 3.7.3 Linguagem R

R é uma linguagem de programação poderosa e que tem um espaço bem considerável com relação a *Data Science*. Ficou famosa pela sua facilidade em fazer análise de dados, processar instruções estatísticas e modelos gráficos. Documentação disponível em [46].

#### *R-Studio*

A linguagem R possui diversas ferramentas para ajudar o desenvolvedor. A ferramenta que mais se destaca é o R-studio. Esta é uma IDE *open source* para R muito útil pois facilita bastante o desenvolvimento, incluindo editor de código e ferramentas de depuração e visualização.

### 3.7.4 RapidMiner

RapidMiner é uma plataforma para trabalhar com Data Science de forma simples e visual. De acordo com o artigo [39], as ferramentas oferecidas fornecem uma interface

gráfica com objetos e processos que simplificam as diversas tarefas necessárias para trabalhar com mineração de dados. O diferencial do RapidMiner é a facilidade e velocidade para criar modelos preditivos já que não é necessário o trabalho de codificação e transformação dos dados, tornando simples o processo de validação e ajuste do modelo. Os três produtos oferecidos são o RapidMiner Studio (para workflows), RapidMiner Server (utilizado para gerenciar modelos) e RapidMiner Radoop (compilar e executar workflows). Documentação disponível em [47].

### 3.7.5 Weka

A documentação do Weka está disponível em [48]. É um projeto *Open Source* que foi criado como um sistema de *Machine Learning* pela universidade de Waikato na Nova Zelândia. O projeto tem o objetivo de disseminar técnicas de "Aprendizado de Máquina" através da disponibilização do *software* para utilização de pesquisadores e alunos para resolver problemas reais da indústria, além de contribuir com a ciência.

O grande diferencial do Weka é a sua interface gráfica (GUI – Graphical User Interface) que torna as tarefas de mineração de dados extremamente fáceis e rápidas. Através da interface é possível consultar dados em sistemas de bancos de dados, executar métodos de processamento de dados, executar e configurar parâmetros dos algoritmos e visualizar os resultados através de gráficos.

O Weka tem funcionalidades para manipulação de bases de dados, interface para visualização de dados, e ainda disponibiliza diversos algoritmos de *machine learning* e *Data Mining*. Os usuários não precisam saber programar as ferramentas de mineração para fazer um trabalho.

### 3.7.6 Ferramentas escolhidas

A principal ferramenta de desenvolvimento utilizada foi a plataforma *Open Source Anaconda*, pois contem as principais bibliotecas Python já inclusas, centenas de pacotes embutidos, como dito anteriormente. Nesse trabalho vão ser usadas as ferramentas e bibliotecas da linguagem Python, principalmente a biblioteca *Pandas* para análise de dados e *Matplotlib* para plotar os gráficos, usando a aplicação *Jupyter* para programação. A decisão em usar essas ferramentas se deu pelo fato de serem gratuitas e *open source* e por serem fáceis de usar. Soma-se a isso a reusabilidade de código. Nos capítulos posteriores será visto com mais detalhes como essas ferramentas foram usadas para se chegar aos resultados e conclusões desse trabalho.

# Capítulo 4

# Preparação dos Dados

Neste capítulo, será abordada a parte de preparação da análise sobre dados do Enem 2017. O objeto da pesquisa é delimitado na Seção 4.1. O processo de de aquisição dos dados é visto na Seção 4.2. Logo em seguida, na Seção 4.3, são mostradas as técnicas e procedimentos realizados sobre os dados a fim de deixá-los no formato ideal para análise. Finalmente, na Seção 4.4, as operações e transformações feitas nos dados são apresentadas.

## 4.1 Delimitação do Objeto de Pesquisa

Como foi descrito no Capítulo 1, os dados do Enem vão ser analisados, pois esta é a avaliação de ensino médio de maior importância atualmente. Com os dados, especificamente do Enem 2017 (o mais recente que o INEP disponibilizou dados), na tentativa de traçar perfis de estudantes, e nível de desempenho geral no Brasil em suas diversas regiões.

O processo de mineração de dados desse trabalho se baseia no modelo CRISP-DM visto na Seção 3.5 e usa as ferramentas citadas na Subseção 3.7.6. Foram seguidas as seguintes etapas:

1. Compreensão de negócios - fase inicial demonstrada no capítulo 1;
2. Aquisição dos dados (Compreensão de dados);
3. Pré-Processamento dos dados (Preparação dos dados);
4. Transformação dos dados (Modelagem);
5. Descoberta de padrões (Avaliação);
6. Resultados (Implantação).

A forma como foram seguidas e aplicadas cada uma dessas etapas nesse trabalho vão ser demonstradas nas seções seguintes e no Capítulo 5.

## 4.2 Aquisição de Dados

Os dados utilizados para essa pesquisa, como dito anteriormente, foram retirados na seção de microdados no portal do INEP, indo na parte relacionada ao Enem e selecionando o mais atual, que nesse caso foram os microdados do Enem 2017 [49]. Ao clicar em baixar, vem um arquivo compactado com nome "*microdados_enem2017*", dividido nas seguintes pastas e contendo os arquivos:

- DADOS: Pasta com os arquivos principais
  - ITENS_PROVA_2017.csv
  - MICRODADOS_ENEM_2017.csv – Esse é o arquivo de dados principal que vai ser usado para mineração
- DICIONÁRIO: O dicionário é um arquivo que contém a definição e caracterização de cada coluna dos arquivos de itens e microdados
  - Dicionário_Microdados_Enem_2017.xlsx
- INPUTS: Arquivos de entrada de sistema com as diretivas de código para automatização de análise de dados, processamento e entre outras coisas.
  - INPUT_ SPSS_MICRODADOS_ENEM_2017.sps
  - INPUT_R_ITENS_PROVA_2017.R
  - INPUT_R_MICRODADOS_ENEM_2017.R
  - INPUT_SAS_ITENS_PROVA_2017.sas
  - INPUT_SAS_MICRODADOS_ENEM_2017.sas
  - INPUT_SPSS_ITENS_PROVA_2017.sps
- LEIA-ME e DOCUMENTOS TÉCNICOS: Parte de documentação, relatório, manual e referências.
  - edital_enem_2017.pdf
  - Leia_Me_Enem_2017.docx
  - manual_de_redacao_do_enem_2017.pdf

  - matriz_referencia_enem.pdf

- PROVAS E GABARITOS: Arquivos de provas e gabaritos divididos por tipos e cores de prova.
  - GABARITOS – Arquivos de gabaritos de Prova 1 e 2, Dias 1 e 2
  - PROVAS – Arquivos de provas, Provas 1 e 2, Dias 1 e 2

A descrição de como são divididos os dados, a descrição das colunas, número de colunas, quantidade de inscritos, entres outras informações, fica no arquivo de dicionário, como mencionado acima. A parte do dicionário que especifica as colunas dos microdados do Enem é divida em cinco colunas (Nome da Variável, Descrição, Variáveis Categóricas, Tamanho e Tipo) que divide todas a colunas.

Os microdados deixam em sigilo os dados individuais dos candidatos, portanto não daria pra identificar nome da pessoa, endereço, RG, etc. São dados mascarados, mas que mostram diversas informações essenciais, como idade, sexo, local de moradia, se o inscrito é treineiro ou candidato oficial, escola onde estudou, respostas que deu na pesquisa socioeconômica, respostas na prova, tipo do caderno, nota final, gabaritos, entre várias outras informações que possibilitam diversas análises. No caso dessa pesquisa, o foco foi nas questões das provas.

O formato do arquivo de microdados já é ideal para trabalhar com a biblioteca do *Pandas* (escolhida para processamento e análise dos dados), pois vem em formato "csv", o que o torna simples de importar e trabalhar sobre o *DataFrame* que a biblioteca gera para fazer as alterações, pré-processamento, transformação dos dados, e descoberta dos resultados. A estratégia usada foi utilizar amostras menores dos dados, de 1000 linhas inicialmente, pois o arquivo de microdados é muito extenso e para criar um único DataFrame no Pandas iria demandar um tempo elevado de processamento.

## 4.3 Pré-Processamento dos Dados

Para conseguir iniciar a busca de padrões e resultados sobre os dados, é necessário fazer um pré-processamento deles com intuito de remover dados que possam ser redundantes e desnecessários. Serve também para resumir os dados, e delimitar o conjunto de dados principal que será o foco do trabalho.

Após a aquisição dos dados, é feito a importação do arquivo dos microdados através da aplicação *Jupyter* (*Notebook*). Foi usada a versão 5.6.0, dentro do conjunto de trabalho do *Anaconda Navigator* e usando as bibliotecas do Pandas e entre outras que serão citadas posteriormente. A instalação é bem simples, basta entrar no site principal da

iniciativa [50], seguir os passos de download para sistema operacional que usa e seguir os passos de instalação. No caso desse trabalho, foi usado a versão para Windows 64 bits. Como já dito anteriormente, o *Anaconda* disponibiliza várias ferramentas de análise de dados no mesmo conjunto de trabalho, facilitando o tratamento dos dados.

Após iniciar o *Jupyter*, inclui-se a biblioteca Pandas para começar trabalhar sobre os dados. Para isso, basta invocar "import pandas as pd" por exemplo, com "pd" sendo a variável usada. Após isso, é preciso criar o *DataFrame* para trabalho, usando a função *read_csv* e atribuindo a uma variável, "data" por exemplo. Essa função precisa receber o caminho do arquivo de dados, ou seja, a árvore hierárquica de pastas, até chegar ao arquivo. Essa função também tem o argumento para informar a norma de padronização para o *DataFrame* e como será feita a separação dos dados na tabela. Para entender melhor, leia a parte inicial do código gerado nesse trabalho, que se encontra nos Apêndices. No caso em questão, foram passado dois argumentos, um para usar padrão "ISO-8859-1", com "encoding = "ISO-8859-1" e outro para ler o arquivo CSV (que no caso é separado por vírgulas) no *DataFrame* criado. Assim, usa-se "sep =';' " no último argumento senão ele não consegue criar um *DataFrame* padronizado.

O *DataFrame* criado funciona como se fosse a própria planilha de dados, mas vem definida para ser usada pela biblioteca do Pandas. É organizado por linhas e colunas, e entende os dados separando por tipos. O Pandas tem o método ".dtypes" que faz a verificação dos tipos de dados que compõem as colunas. Para acessar a lista de colunas, usa-se o método ".columns" que mostra uma lista na forma de índices. O método ".describe()", verifica resumidamente a disposição estatística dos dados numéricos.

Os microdados do Enem tem 137 colunas e milhões de linhas (cada uma representa um inscrito no exame) na amostra completa. No caso, após as transformações de dados que foram feitas para esse trabalho, aumentaram-se as colunas no *DataFrame* como será visto nas Seções 5.1.7 e 5.1.8. Com o método *describe* dá pra verificar essa informação sempre que necessário. O método ".head()" serve para mostrar sucintamente as primeiras linhas dos dados, e por padrão do Pandas [43], ele mostra as primeiras 5 linhas. Para alterar isso, basta colocar nos parênteses o argumento que se quer com o número de linhas que se deseja mostrar, como por exemplo, ".head(n=10)". Existe também o método ".tail()" que faz a mesma coisa, porém ao invés de apresentar as primeiras 5 linhas, apresenta as 5 últimas.

Como o arquivo é muito grande, foi preciso fazer uma amostra menor dele para facilitar a análise dos dados e para que os algoritmos rodassem adequadamente. Então, os testes iniciais foram feitos em apenas 1000 linhas. Uma vez completados os testes, se aplicou o algoritmo a uma entrada de 10000 linhas.

```
In [11]: dados.head()
```

Out[11]:

| | NU_INSCRICAO | NU_ANO | CO_MUNICIPIO_RESIDENCIA | NO_MUNICIPIO_RESIDENCIA | CO_UF_RESIDENCIA | SG_UF_RESIDENCIA | NU_IDADE | TP_SEXO | TP_E |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 170003336736 | 2017 | 3503208 | Araraquara | 35 | SP | 29 | F | |
| 1 | 170003333545 | 2017 | 5002902 | Cassilândia | 50 | MS | 22 | F | |
| 2 | 170001663644 | 2017 | 3550308 | São Paulo | 35 | SP | 38 | F | |
| 3 | 170001663645 | 2017 | 4209300 | Lages | 42 | SC | 35 | F | |
| 4 | 170001663646 | 2017 | 2704302 | Maceió | 27 | AL | 40 | M | |

5 rows × 137 columns

```
In [17]: dados.describe()
```

Out[17]:

| | NU_INSCRICAO | NU_ANO | CO_MUNICIPIO_RESIDENCIA | CO_UF_RESIDENCIA | NU_IDADE | TP_ESTADO_CIVIL | TP_COR_RACA | TP_NACIONALIDADE | CO_ |
|---|---|---|---|---|---|---|---|---|---|
| count | 9.990000e+02 | 999.0 | 9.990000e+02 | 999.000000 | 999.000000 | 963.000000 | 999.000000 | 999.000000 | |
| mean | 1.700025e+11 | 2017.0 | 3.303199e+06 | 32.864865 | 25.315315 | 0.199377 | 2.024024 | 1.026026 | |
| std | 9.331097e+05 | 0.0 | 9.359837e+05 | 9.350809 | 7.896352 | 0.466909 | 0.991660 | 0.177161 | |
| min | 1.700017e+11 | 2017.0 | 1.100122e+06 | 11.000000 | 14.000000 | 0.000000 | 0.000000 | 0.000000 | |
| 25% | 1.700017e+11 | 2017.0 | 2.703328e+06 | 27.000000 | 19.000000 | 0.000000 | 1.000000 | 1.000000 | |
| 50% | 1.700017e+11 | 2017.0 | 3.204906e+06 | 32.000000 | 23.000000 | 0.000000 | 2.000000 | 1.000000 | |
| 75% | 1.700033e+11 | 2017.0 | 3.550308e+06 | 35.000000 | 29.000000 | 0.000000 | 3.000000 | 1.000000 | |

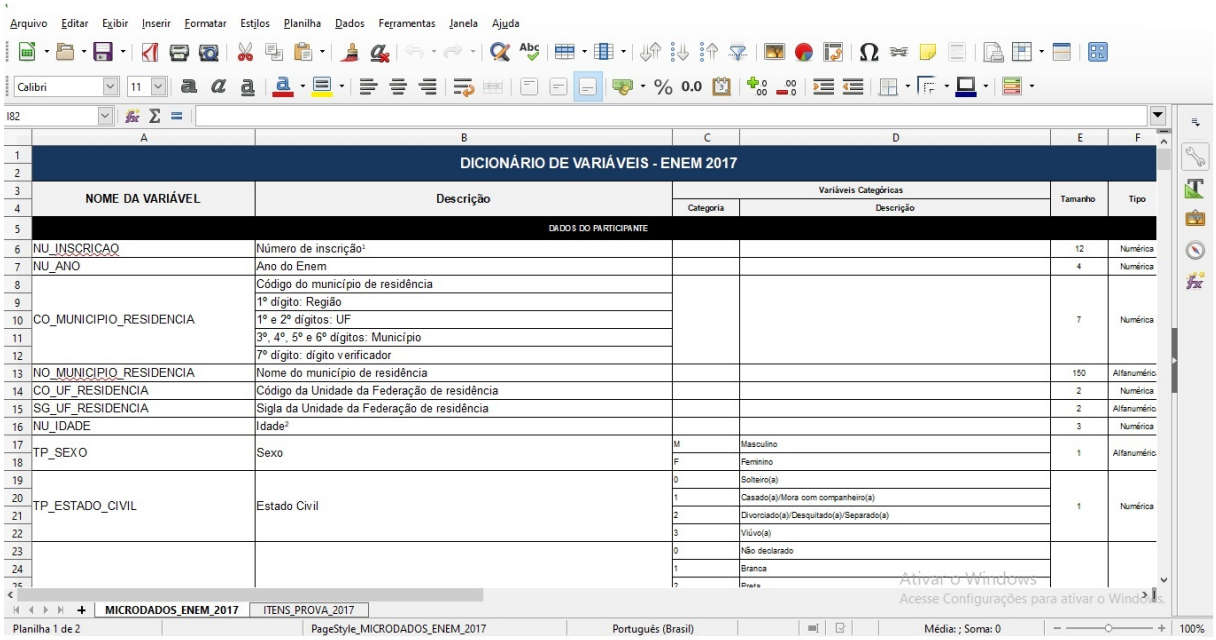


Figura 4.1: Métodos head e e describe no Pandas, exemplo aplicado nas primeiras 1000 linhas dos microdados do Enem 2017.

No Pandas, as informações do *DataFrame* usando os métodos vistos acima aparecem na forma da Figura 4.1

O arquivo de Dicionário de Variáveis do Enem, já citado anteriormente, na aba "MICRODADOS_ENEM_2017", contém a descrição de cada coluna, informando o campo, tipo e a descrição da variável, entre outras informações que também podem ser vistas no Pandas com os métodos visto acima. A Figura 4.2 mostra uma parte de como é organizado o arquivo de Dicionário na aba "MICRODADOS_ENEM_2017". No subarquivo "ITENS_PROVA_2017" são descritos os atributos relacionados às provas, como item, gabarito, identificador da prova, língua escolhida, código do item, entre outros. A Figura 4.3 mostra uma parte de como é organizado o arquivo de Dicionário na aba "ITENS_PROVA_2017".

Não foram usadas todas as colunas presentes na tabela de microdados para esse trabalho, pois o foco foi em avaliar desempenho para descobrir problemas. Então, as colunas com dados muito específicos sobre os estudantes e escolas em que estudaram, como por exemplo, código e sigla da unidade de federação de residência e códigos da escola, do município da escola, dependência administrativa dela, entre outras, foram retiradas, pois não são úteis para o tipo de análise feita aqui. Dados sobre os locais de aplicação de prova, e algumas informações sobre a redação, que apesar de importantes, não serviram para o foco de análise desse trabalho.



DICIONÁRIO DE VARIÁVEIS - ENEM 2017

| NOME DA VARIÁVEL | Descrição | Variáveis Categóricas | | Tamanho | Tipo |
|---|---|---|---|---|---|
| | | Categoria | Descrição | | |
| DADOS DO PARTICIPANTE | | | | | |
| NU_INSCRICAO | Número de inscrição[1] | | | 12 | Numérica |
| NU_ANO | Ano do Enem | | | 4 | Numérica |
| CO_MUNICIPIO_RESIDENCIA | Código do município de residência | | | 7 | Numérica |
| | 1º dígito: Região | | | | |
| | 1º e 2º dígitos: UF | | | | |
| | 3º, 4º, 5º e 6º dígitos: Município | | | | |
| | 7º dígito: dígito verificador | | | | |
| NO_MUNICIPIO_RESIDENCIA | Nome do município de residência | | | 150 | Alfanumérico |
| CO_UF_RESIDENCIA | Código da Unidade da Federação de residência | | | 2 | Numérica |
| SG_UF_RESIDENCIA | Sigla da Unidade da Federação de residência | | | 2 | Alfanumérico |
| NU_IDADE | Idade[2] | | | 3 | Numérica |
| TP_SEXO | Sexo | M | Masculino | 1 | Alfanumérico |
| | | F | Feminino | | |
| TP_ESTADO_CIVIL | Estado Civil | 0 | Solteiro(a) | 1 | Numérica |
| | | 1 | Casado(a)/Mora com companheiro(a) | | |
| | | 2 | Divorciado(a)/Desquitado(a)/Separado(a) | | |
| | | 3 | Viúvo(a) | | |
| | | 0 | Não declarado | | |
| | | 1 | Branca | | |

Figura 4.2: Fragmento do arquivo de Dicionário dos microdados do Enem 2017, aberto em um leitor de planilhas, aba 1.

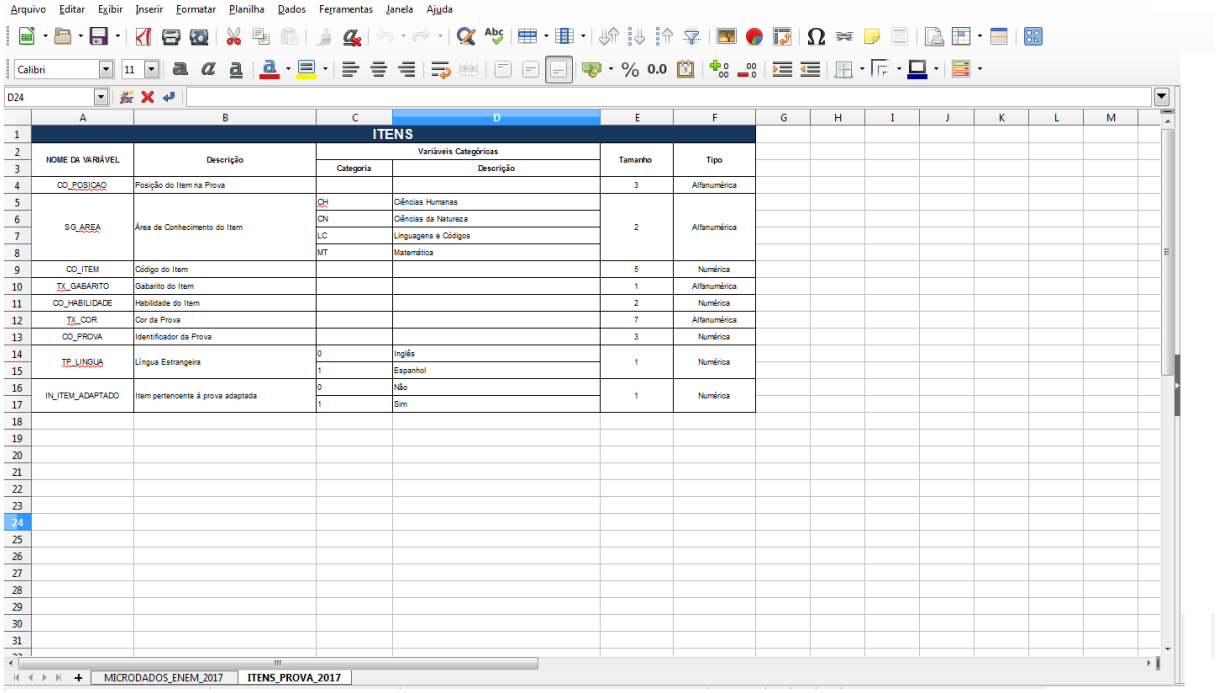
Os dados também contêm vários indicadores de alunos que pediram atendimento especial, como de baixa visão, surdez, deficiência mental, autismo, entre outros, e também para alunos que pediram atendimentos específicos, como, gestantes, lactantes, idosos, cadeirantes, etc. Essas informações foram consideradas, porém o foco do trabalho foi voltado a inscritos em condições regulares.

Por fim, o Enem solicita aos inscritos responder o questionário socioeconômico, onde foram feitas 27 perguntas sobre poder aquisitivo do candidato e da família com quem ele reside. Esses dados foram considerados no trabalho com o intuito de verificar o impacto de variáveis socioeconômicas no desempenho dos candidatos.

## 4.4 Transformação dos Dados

As colunas mais importantes dos dados do Enem para esse trabalho estão relacionadas à parte de respostas dos alunos às questões das provas, os gabaritos correspondentes, os tipos de provas, cadernos e resultados na prova. Ou seja, dados que conseguem descrever por si só o desempenho dos candidatos. Contudo, ainda foi preciso aplicar transformações sobre algumas colunas desses dados para facilitar a busca de resultados e padrões.

Algumas colunas merecem destaque, pois sobre elas foram feitas as transformações que determinaram os resultados desse trabalho. No geral seriam todas as colunas de



ITENS

| NOME DA VARIÁVEL | Descrição | Variáveis Categóricas | | Tamanho | Tipo |
|---|---|---|---|---|---|
| | | Categoria | Descrição | | |
| CO_POSICAO | Posição do Item na Prova | | | 3 | Alfanumérica |
| SG_AREA | Área de Conhecimento do Item | CH | Ciências Humanas | 2 | Alfanumérica |
| | | CN | Ciências da Natureza | | |
| | | LC | Linguagens e Códigos | | |
| | | MT | Matemática | | |
| CO_ITEM | Código do Item | | | 5 | Numérica |
| TX_GABARITO | Gabarito do Item | | | 1 | Alfanumérica |
| CO_HABILIDADE | Habilidade do Item | | | 2 | Numérica |
| TX_COR | Cor da Prova | | | 7 | Alfanumérica |
| CO_PROVA | Identificador da Prova | | | 3 | Numérica |
| TP_LINGUA | Língua Estrangeira | 0 | Inglês | 1 | Numérica |
| | | 1 | Espanhol | | |
| IN_ITEM_ADAPTADO | Item pertencente à prova adaptada | 0 | Não | 1 | Numérica |
| | | 1 | Sim | | |

Figura 4.3: Fragmento do arquivo de Dicionário dos microdados do Enem 2017, aberto em um leitor de planilhas, aba 2.

respostas e gabaritos. As colunas de respostas apresentam um vetor com as respostas dos candidatos em cada parte da prova do Enem:

- TX_RESPOSTAS_CN: Vetor com as respostas da parte objetiva da prova de Ciências da Natureza;
- TX_RESPOSTAS_CH: Vetor com as respostas da parte objetiva da prova de Ciências Humanas;
- TX_RESPOSTAS_LC: Vetor com as respostas da parte objetiva da prova de Linguagens e Códigos;
- TX_RESPOSTAS_MT: Vetor com as respostas da parte objetiva da prova de Matemática.

Já as colunas de gabaritos apresentam um vetor com os gabaritos de cada parte da prova do Enem:

- TX_GABARITO_CN: Vetor com o gabarito da parte objetiva da prova de Ciências da Natureza;

- TX_GABARITO_CH: Vetor com o gabarito da parte objetiva da prova de Ciências Humanas;
- TX_GABARITO_LC: Vetor com o gabarito da parte objetiva da prova de Linguagens e Códigos;
- TX_GABARITO_MT: Vetor com o gabarito da parte objetiva da prova de Matemática.

O problema com essas colunas é que apresentam um vetor completo com as respostas do aluno para todos os itens e apenas informa a nota final do candidato naquela prova em questão. Então, para saber a resposta questão por questão foi preciso varrer o vetor e consultar a posição desejada. Outro problema está no fato de o Enem dividir a prova em tipos de cadernos (caderno azul, amarelo, cinza, entre outros para cada tipo de prova), ou seja, além da preocupação em separar a análise nos tipos principais das provas (matemática, ciências naturais, ciências humanas e linguagens), surge a necessidade de filtrar pelos tipos de cadernos. A diferença entre os cadernos é só na ordem que estão dispostas as questões, o conteúdo é o mesmo. Para os tipos de cadernos, os dados do Enem tem as colunas de código, então para saber qual tipo de prova o candidato fez, basta verificar o campo de código (ex: CO_PROVA_CN==391.0), que no caso, o Enem dividiu em:

- Prova de Ciências da Natureza
  - Caderno Azul(CO_PROVA_CN==391.0)
  - Caderno Amarelo(CO_PROVA_CN==392.0)
  - Caderno Cinza(CO_PROVA_CN==393.0)
  - Caderno Rosa(CO_PROVA_CN==394.0)
- Prova de Ciências Humanas
  - Caderno Azul(CO_PROVA_CH==395.0)
  - Caderno Amarelo(CO_PROVA_CH==396.0)
  - Caderno Branco(CO_PROVA_CH==397.0)
  - Caderno Rosa(CO_PROVA_CH==398.0)
- Prova de Linguagens e Códigos
  - Caderno Azul(CO_PROVA_LC==399.0)
  - Caderno Amarelo(CO_PROVA_LC==400.0)
  - Caderno Rosa(CO_PROVA_LC==401.0)

  - Caderno Branco(CO_PROVA_LC==402.0)

- Prova de Matemática
  - Caderno Azul(CO_PROVA_MT==403.0)
  - Caderno Amarelo(CO_PROVA_MT==404.0)
  - Caderno Rosa(CO_PROVA_MT==405.0)
  - Caderno Cinza(CO_PROVA_MT==406.0)

Então, foram propostas duas transformações principais:

1. Extrair as respostas de cada questão varrendo os vetores "RESPOSTAS" e extrair também os gabaritos de cada questão varrendo os vetores "GABARITOS" e fazer comparação item a item para a partir daí criar uma coluna que diz se o aluno acertou ou não aquele item. Para isso, representar com "0 (zero)" se o candidato errou o item e representar com "1 (um)" se o candidato acertou o item. No fim, cria-se uma coluna por respostas, que no caso são nomeadas como por exemplo: "TX_Result_MT_questaoX", onde "X" representa o número do item. Por exemplo, se o aluno acertou a primeira questão da prova de matemática, então cria-se uma coluna TX_Result_MT_questao1 cujo o conteúdo é 1. Se o aluno errou, então cria-se a coluna TX_Result_MT_questao1 cujo o conteúdo é 0. Um exemplo disso é mostrado na Tabela 4.1. Da mesma forma se faz para as outras provas.

2. Criar colunas com as respostas, de forma similar aos acertos, apenas separando item a item. Por exemplo, se nas respostas da prova de Ciências da Natureza aparecer "ABCAB..", cria-se as colunas "TX_CN_questao1, TX_CN_questao2, TX_CN_questao3, TX_CN_questao4, TX_CN_questao5...." cada uma com sua resposta separada para permitir saber como os candidatos do Enem respondem as perguntas. Isso permite descobrir algum padrão de respostas dos candidatos em determinadas questões das provas. Um exemplo disso é mostrado na Tabela 4.2. Da mesma forma se faz para as outras provas.

Os Apêndices contêm as partes de código referentes a essas transformações. Nesse ponto, houve uma certa dificuldade em simplificar funções para minimizar o trabalho de

Tabela 4.1: Exemplo de Colunas criadas na Primeira transformação.

| **TX_Result_MT _questao1** | **TX_Result_MT _questao2** | **TX_Result_MT _questao3** |
|---|---|---|
| 0 | 0 | 1 |

Tabela 4.2: Exemplo de Colunas criadas na Segunda transformação.

| **TX_CN_questao1** | **TX_CN_questao2** | **TX_CN_questao3** |
|---|---|---|
| A | B | C |

criar por exemplo uma variável para cada resposta (uma média de 45 a 50 questões por tipo de prova necessitou de 45 a 50 variáveis por tipo de prova). Isso porque era preciso muitas vezes, preencher cada coluna baseando nos dados das outras colunas. Olhando o código dá pra notar o grande número de variáveis que foi preciso criar para preencher as diversas colunas de respostas após as transformações. Após o processo de transformação gera-se os arquivos de saída (.csv) com as novas colunas inclusas. Isso é feito no Pandas, convertendo o *DataFrame* em um arquivo desejado, com a função ".to_csv", onde csv representa a extensão do arquivo que se deseja na saída (pode ser csv, xls, etc). Essa função recebe o *DataFrame* no parâmetro de entrada e na saída gera o arquivo com a extensão informada e com o nome do arquivo que se desejar na saída no parâmetro entre parênteses, por exemplo, pd.to_csv('transformação1.csv') irá criar um arquivo csv de nome "transformação1" baseado no *DataFrame* que está associado ao Pandas na variável "pd".

Um ponto final a se destacar aqui seria com relação aos dados nulos encontrados na tabela do Enem. Alguns campos vem em branco quando alunos deixam a prova sem marcação, ou faltaram alguma das provas ou se ausentaram totalmente nos dias de prova. O Enem mantém o registro desse candidato e campos de inscrição, estado onde vive, respostas na prova socioeconômica, etc. Na pesquisa, os campos nulos foram desconsiderados e para isso se fez uma limpeza, deixando o *DataFrame* de trabalho sem dados nulos. No Pandas isso é possível com a função ".dropna".

# Capítulo 5

# Resultados Sobre o Enem 2017

Nesse capítulo, que contém apenas a Seção 5.1, serão feitas a mineração de dados e as análises de algumas colunas geradas sobre os dados. Tais colunas mostram os resultados do desempenho em diversas questões das provas e as diferenças de resultados de região para região do país, levando em consideração diferentes atributos no conjunto de dados.

## 5.1 Mineração e Análise dos Dados

A parte de preparação, processamento e transformação dos dados serviram para melhor extrair informações úteis para o trabalho. Aqui se inicia a etapa de mineração de dados propriamente dita.

Deve ser lembrado que a mineração foi feita sobre uma amostra menor dos microdados. Foram feitos vários testes, que após funcionar foram aplicados à amostra maior. Nesse caso, o tamanho das amostras foi decidido por convenção, escolhendo-se uma amostra para maximizar o tempo de processamento do algoritmo completo e ainda assim ser suficiente para fazer uma análise nos dados. Primeiro, foram feitos os testes em uma amostra de 1000 linhas dos dados, para depois aplicar em uma amostra maior com 10000 linhas. Os resultados apresentados a seguir são com relação ao segundo *DataFrame* com a amostra de 10000 linhas.

A partir disso, foi possível fazer uma análise gráfica e estatística dos dados (o que lembra um pouco o que foi feito no trabalho de Jorge [20]). Nessa parte do trabalho alguns gráficos foram gerados usando as funções da biblioteca *matplotilib* junto ao Pandas na plataforma de desenvolvimento do Anaconda, nas células do *Jupyter*.

É importante destacar a função describe(), que foi utilizada várias vezes para gerar as informações estatísticas das colunas dos microdados. Essa função [51] resume a tendência central, a dispersão e a forma da distribuição dos dados da coluna, incluindo valor máximo, mínimo e os percentis padrão(25%, 50% e 75%). Percentis se referem aos 99 valores que

separam uma série em 100 partes iguais. Para os dados numéricos, o índice do resultado vai incluir:

- ".count" - Número de contagem de observações não nulas;
- ".mean" - Média dos valores;
- ".std" - Desvio padrão das observações;
- "min" - Mínimo dos valores no objeto;
- "max" - Máximo dos valores no objeto;
- Percentis
  - Percentis inferiores
  - 50%
  - Percentis superiores

Por padrão, o percentil inferior é 25 e o percentil superior é 75. O 50 percentil é o mesmo que a mediana. Estes são os chamados "quartis" na análise estatística. Quartil são os valores que dividem um conjunto de dados ordenados em quatro partes iguais:

1. Primeiro quartil (1/4) - Quartil inferior = Valor aos 25% da amostra ordenada. Valor que deixa 25% dos elementos à sua esquerda e 75% à sua direita;
2. Segundo quartil (1/2) - Mediana = Valor até ao qual se encontra 50% da amostra ordenada, ou seja, divide o conjunto de dados ao meio considerando a ordenação. 50% dos elementos estão à sua esquerda e 50% dos elementos estão à sua direita;
3. Terceiro quartil (3/4) - Quartil superior = Valor a partir do qual se encontram 25% dos valores mais elevados: valor aos 75% da amostra ordenada. É o valor que deixa 75% dos elementos à sua esquerda e 25% à sua direita.

Em se tratando de dados do Enem, que são obtidos após uma avaliação, é natural surgir questionamentos sobre o desempenho dos inscritos de forma geral, desempenho por região, por escolas, etc. Por isso, na parte de mineração foram estipuladas algumas ideias do que seria interessante tirar de informação dos dados para compor o resultado do trabalho:

- Distribuição de inscritos no Enem por gênero, cor/raça, renda e tipo de escola na amostra;

- Estatísticas de notas da amostra;
- Distribuição das notas dos inscritos da amostra;
- Estatísticas das notas por gênero, cor/raça, renda e tipo de escola;
- Média das notas por regiões brasileiras na amostra;
- Tirar alguns dados sobre desempenho dentro do Distrito Federal especificamente, que é a região onde se encontra a Universidade de Brasília;
- Porcentagem de acerto por questão em cada prova;
- Distribuição de marcação de alternativas nas questões entre os inscritos.
- Com as principais questões do Enem, comparar as estatísticas por idade, estado civil, cor/raça e tipo de escola;
- Comparar os inscritos com maiores notas por gênero, cor/raça, renda e tipo de escola;

### 5.1.1 Distribuição de Inscritos no Enem em Diferentes Classes da Amostra

Para se ter uma ideia geral de como foi a relação de inscritos no Enem 2017, será dividida em classes e verificada a porcentagem de inscritos em cada uma na amostra de 10000 inscritos.

#### Gênero

Na amostra, foi verificado que 41,76% dos inscritos são do sexo masculino e 58,23% do sexo feminino.

#### Raça

Na amostra, foi verificado que 38,32% se declararam brancos, 13,69% pretos, 43,37% pardos, 2,32% amarelos e 0,39% indígenas e 1,9% não declararam cor ou raça. Nota-se a maioria de inscritos pardos e minoria de inscritos indígenas.

**Renda**

Na amostra, considerando renda total da família, foi verificado que 2,2% declararam que não têm renda nenhuma, 18,66% que têm renda até 937,00 reais, 26,04% que têm renda até 1405,50 reais, 13,52% que têm renda até 2342,50 reais, 10,6% que têm renda até 2811,00 reais, 6,29% que têm renda até 3748,00 reais, 6,81% que têm renda até 3748,00 reais, 4,72% que têm renda até 4685,00 reais, 3,14% que têm renda até 5622,00 reais, 1,95% que têm renda até 6559,00 reais, 1,03% que têm renda até 7496,00 reais, 0,87% que têm renda de até 8433,00 reais, 0,69% que têm renda até 9370,00 reais, 1,15% que têm renda até 11244,00 reais, 0,65% que têm renda até 14055,00 reais, 0,66% que têm renda até 18740,00 reais e 1,01% que têm renda superior a 18740,00 reais.

**Tipo de Escola**

85,35% dos inscritos não responderam em que escola estudaram ou estudam no ensino médio. Entre os que responderam, 81,51% são de escola pública, 17,14% são de escola privada e 1,29% são de escola do exterior.

### 5.1.2 Estatísticas de Notas da Amostra

Para obter essas estatísticas o Pandas possui as funções de máximo e mínimo, bastando então passar a coluna de referência. Uma forma melhor de se fazer isso seria usando a função "describe()" também.

Para gerar gráficos e tabelas, fez mais sentido criar variáveis e armazenar os valores de média, máximo e mínimo e depois plotar passando essas informações como entradas. As notas das diferentes provas do Enem valem 1000 pontos. É importante lembrar que a amostra aqui foi de 10000 inscritos, portanto quando aparece dados sobre notas máximas ou mínimas ou médias, é com relação a essa amostra. Para uma amostra de 6 milhões de candidatos (total de inscritos do Enem), nível de confiança de 95% e margem de erro de 1%, o tamanho de amostra mínima é de 9589 candidatos [52]. Assim, foi escolhido uma amostra de 10000 inscritos.

A Tabela 5.1 mostra as estatísticas gerais de notas em cada parte da prova.

### 5.1.3 Distribuição das Notas dos Inscritos na Amostra

Aqui foram calculadas as médias de notas dos inscritos nas 4 partes da prova do Enem, Ciências da Natureza, Ciências Humanas, Linguagens e Códigos e Matemática. Para isso, bastou executar a função de média no Pandas (.mean), referenciando as colunas que tratam cada seção da prova e salvando em uma variável. Uma outra forma de se obter esses

Tabela 5.1: Estatísticas gerais de notas em cada parte da prova do Enem 2017.

| | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|
| **Média** | 531,54 | 551,33 | 535,87 | 547,37 |
| **Máxima** | 805,70 | 841,70 | 724,90 | 943,60 |
| **Mediana** | 530,30 | 560,20 | 541,10 | 534,40 |
| **Primeiro quartil** | 473,50 | 494,70 | 498,20 | 4620 |
| **Segundo quartil** | 530,30 | 560,20 | 541,10 | 534,40 |
| **Terceiro quartil** | 585,90 | 610,90 | 579,30 | 617,90 |

dados seria com a função “describe()”, explicada na seção 5.1. Lembrando que foi necessário a eliminação de linhas nulas e para isso usou-se o comando “.dropna(inplace=True)”.

### Ciências da Natureza

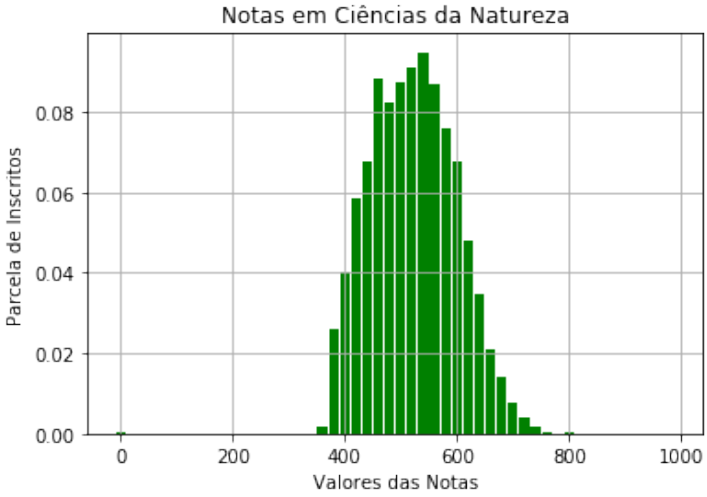



Figura 5.1: Gráfico de distribuição de notas de Ciências da Natureza por inscritos.

A coluna correspondente que foi referenciada aqui é a “NU_NOTA_CN”. Ela contém os dados sobre as notas dos inscritos na prova de Ciências da Natureza. A média nacional encontrada na amostra foi de 531,54 pontos. A Figura 5.1 mostra a distribuição de notas de Ciências da Natureza no país, referenciando a quantidade de inscritos (frequência) pelas notas tiradas na prova. Nota-se que alguns inscritos, apesar de terem feito a prova, ficaram com nota zero. Nos microdados, quando o aluno não compareceu à prova, o dado sobre a nota dele não aparece como zero e sim como dado nulo. Por

isso, acontece de aparecer poucas notas zeros, que são de inscritos que realmente fizeram a prova, porém tiraram zero. As notas aqui ficaram distribuídas entre 400 e 800 pontos, onde apenas um inscrito conseguiu nota acima de 800 pontos nessa amostra.

### Ciências Humanas

A coluna correspondente que foi referenciada aqui é a "NU_NOTA_CH". Ela contém os dados sobre as notas dos inscritos na prova de Ciências Humanas. A média nacional encontrada na amostra foi de 551,33 pontos. A Figura 5.2 mostra a distribuição de notas de Ciências Humanas no país, referenciando a quantidade de inscritos (frequência) pelas notas tiradas na prova. Ocorreu aqui de mais inscritos tirarem zero na prova e novamente apenas um aluno ficou com nota aceima de 800. Comparando com Ciências da Natureza, aqui a distribuição ficou com mais notas abaixo de 400 pontos e mais notas acima de 700 pontos.

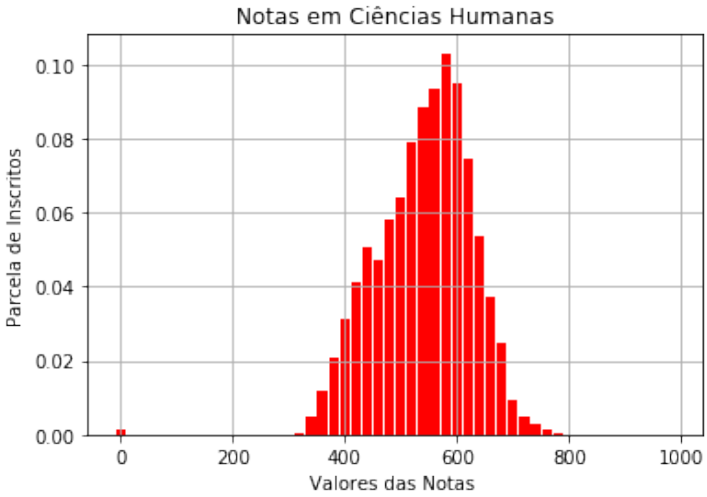



Figura 5.2: Gráfico de distribuição de notas de Ciências Humanas por inscritos.

### Linguagens e Códigos

A coluna correspondente que foi referenciada aqui é a "NU_NOTA_LC". Ela contém os dados sobre as notas dos inscritos na prova de Linguagens e Códigos. A média nacional encontrada na amostra foi de 535,87 pontos. A Figura 5.3 mostra a distribuição de notas de Linguagens e Códigos no país, referenciando a quantidade de inscritos (frequência) pelas notas tiradas na prova. Ocorreram casos de nota zero, e deu pra ver que teve maior quantidade de notas abaixo de 400 pontos se comparar com Ciências da Natureza e Humanas e também as notas ficaram mais distantes dos 800 pontos.

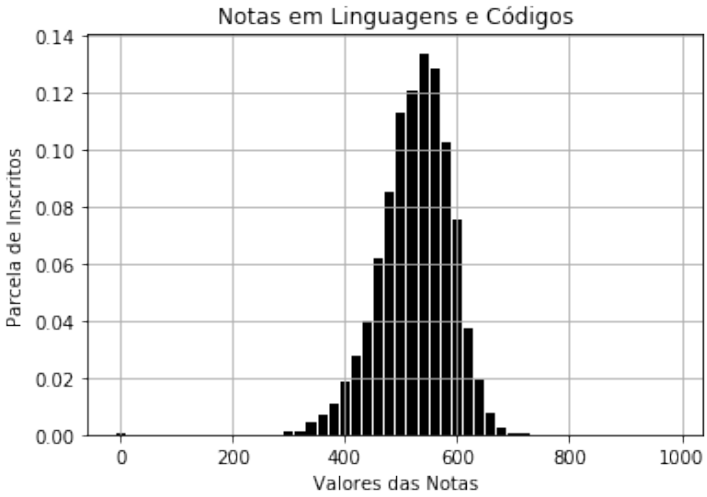



Figura 5.3: Gráfico de distribuição de notas de Linguagens e Códigos por inscritos.

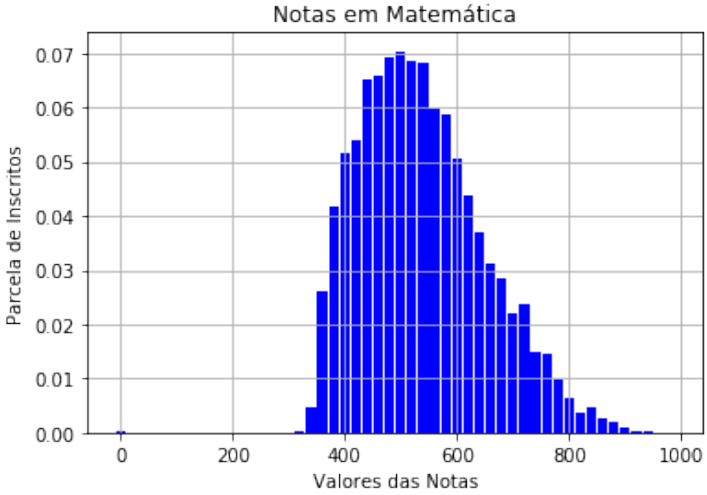



Figura 5.4: Gráfico de distribuição de notas de Matemática por inscritos.

### Matemática

A coluna correspondente que foi referenciada aqui é a "NU_NOTA_MT". Ela contém os dados sobre as notas dos inscritos na prova de Matemática. A média nacional encontrada na amostra foi de 547,37 pontos. A Figura 5.4 mostra a distribuição de notas de Matemática no país, referenciando a quantidade de inscritos (frequência) pelas notas tiradas na prova. Nessa área, aconteceu algo que não era esperado: muitas notas acima de 800 pontos.

### 5.1.4 Estatísticas das Notas por Gênero, Cor/Raça, Renda e Tipo de escola

#### Gênero

A Tabela 5.2 mostra as estatísticas de notas separadas por gênero. De forma geral, não foi possível notar grandes diferenças de desempenho entre os dois gêneros.

Tabela 5.2: Estatísticas de notas por gênero.

| | | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|---|
| Masculino | Média | 543.84 | 565,56 | 537,37 | 573,16 |
| | Mediana | 545,80 | 577,70 | 544,44 | 565,60 |
| | Máxima | 805,70 | 841,70 | 722,00 | 943,60 |
| | 1° quartil | 484,40 | 515,45 | 498,90 | 483,25 |
| | 3° quartil | 601,95 | 623,80 | 582,35 | 655,10 |
| Feminino | Média | 522,77 | 541,1 | 534,79 | 528,99 |
| | Mediana | 520,55 | 549,10 | 538,75 | 516,10 |
| | Máxima | 759,10 | 781,70 | 724,90 | 930,00 |
| | 1° quartil | 466,17 | 484,00 | 497,60 | 452,07 |
| | 3° quartil | 574,12 | 600,20 | 577,07 | 591,05 |

#### Cor/Raça

A Tabela 5.3 mostra as estatísticas de notas separadas por cor/raça.

#### Renda

O Enem dividiu a pesquisa socioeconômica em 17 categorias. Como são muitas, será apresentado apenas o desempenho médio das notas nas provas para cada categoria. Com relação à renda zero e à renda máxima, serão mostradas todas as estatísticas. A Tabela 5.4 mostra as estatísticas gerais de notas da menor e maior faixas de renda. A Tabela 5.5 mostra as médias de notas em todas as faixas de renda.

Percebe-se que o desempenho dos inscritos de renda zero é inferior aos outros e que conforme cresce a renda, o desempenho melhora. Porém, não se torna regra, pois as médias de notas dos inscritos na faixa de renda entre 11244,01 e 14055,00 reais são as maiores do Enem, ultrapassando até a média dos inscritos com a maior renda familiar (acima de 18740,00 reais).

Tabela 5.3: Estatísticas de notas por cor/raça.

| | | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|---|
| Branca | Média | 544,27 | 565,91 | 548,67 | 569,25 |
| | Mediana | 545,80 | 574,90 | 553,30 | 560,75 |
| | Máxima | 805,70 | 781,70 | 724,90 | 943,60 |
| | 1° quartil | 484,27 | 514,80 | 514,15 | 481,67 |
| | 3° quartil | 601,32 | 622,30 | 590,40 | 646,00 |
| Preta | Média | 518,60 | 539,80 | 527,87 | 519,86 |
| | Mediana | 516,00 | 548,45 | 532,55 | 505,90 |
| | Máxima | 750,20 | 748,90 | 694,50 | 919,50 |
| | 1° quartil | 468,45 | 484,32 | 491,90 | 442,90 |
| | 3° quartil | 565,6 | 599,67 | 570,70 | 580,15 |
| Parda | Média | 523,28 | 514,10 | 526,45 | 534,77 |
| | Mediana | 522,10 | 551,15 | 530,75 | 521,35 |
| | Máxima | 755,10 | 841,70 | 722,00 | 909,90 |
| | 1° quartil | 466,40 | 481,00 | 488,52 | 454,60 |
| | 3° quartil | 576,60 | 601,47 | 569,17 | 601,50 |
| Amarela | Média | 526,43 | 545,23 | 532,63 | 553,36 |
| | Mediana | 523,30 | 551,90 | 542,60 | 542,80 |
| | Máxima | 708,50 | 722,00 | 672,00 | 892,50 |
| | 1° quartil | 474,90 | 487,15 | 491,05 | 465,00 |
| | 3° quartil | 579,20 | 603,85 | 576,65 | 628,80 |
| Indígena | Média | 491,70 | 517,20 | 506,93 | 513,50 |
| | Mediana | 465,70 | 536,25 | 505,25 | 493,30 |
| | Máxima | 684,90 | 690,80 | 658,40 | 764,40 |
| | 1° quartil | 441,17 | 450,40 | 470,67 | 445,27 |
| | 3° quartil | 546,62 | 588,97 | 547,92 | 580,57 |

Tabela 5.4: Estatísticas de notas nas duas principais faixas de renda.

| | | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|---|
| Renda zero R$0,00 | Média | 502,57 | 511,07 | 510,95 | 513,34 |
| | Mediana | 497,95 | 505,00 | 511,30 | 500,70 |
| | Máxima | 713,40 | 707,10 | 647,70 | 834,60 |
| | 1° quartil | 453,22 | 437,85 | 468,10 | 442,50 |
| | 3° quartil | 551,22 | 590,60 | 561,95 | 559,30 |
| Renda superior a R$18 740,00 | Média | 609,46 | 638,35 | 594,79 | 688,33 |
| | Mediana | 634,70 | 647,20 | 601,00 | 730,30 |
| | Máxima | 805,70 | 841,70 | 703,80 | 930,00 |
| | 1° quartil | 572,40 | 600,50 | 568,30 | 617,75 |
| | 3° quartil | 682,97 | 685,00 | 632,40 | 774,82 |

Tabela 5.5: Médias de notas em cada faixa de renda

| | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|
| Renda zero R$0,00 | 502,57 | 511,07 | 510,95 | 513,34 |
| Renda até R$937,00 | 503,54 | 516,46 | 508,75 | 504,24 |
| R$937,01 até R$1405,50 | 511,34 | 531,83 | 520,22 | 514,09 |
| R$1405,50 até R$1874,00 | 523,310 | 544,68 | 531,79 | 535,57 |
| R$1874,01 até R$2342,50 | 530,95 | 553,75 | 540,24 | 546,82 |
| R$2342,51 até R$2811,00 | 542,82 | 569,63 | 548,72 | 570,90 |
| R$2811,01 até R$3748,00 | 545,10 | 569,44 | 553,44 | 580,55 |

**Tabela 5.5 – continuação**

| | Ciências da Natureza | Ciências Humanas | Linguagens e Códigos | Matemática |
|---|---|---|---|---|
| R$3748,01 até R$4685,00 | 570,71 | 589,22 | 564,59 | 596,07 |
| R$4685,01 até R$5622,00 | 582,59 | 608,7 | 574,74 | 615,4 |
| R$5622,01 até R$6559,00 | 583,51 | 604,73 | 575,99 | 613,22 |
| R$6559,01 até R$7496,00 | 589,34 | 618,53 | 585,49 | 635,61 |
| R$7496,01 até R$8433,00 | 605,21 | 634,65 | 589,59 | 667,42 |
| R$8433,01 até R$9370,00 | 594,48 | 626,1 | 595,92 | 635,46 |
| R$9370,01 até R$11 244,00 | 602,18 | 621,91 | 584,95 | 651,81 |
| R$11 244,01 até R$14 055,00 | 631,22 | 649,23 | 608,83 | 707,88 |
| R$14 055,01 até R$18 740,00 | 609,12 | 614,89 | 589,75 | 669,56 |
| Renda superior a R$18 740,00 | 609,46 | 638,35 | 594,79 | 688,33 |

**Tipo de Escola**

A Tabela 5.6 mostra as estatísticas de notas com relação a cada tipo de escola. Nota-se que inscritos de escola de exterior tiveram desempenho melhor. Entre as escolas do Brasil, inscritos de escolas privadas tiveram médias maiores do que os de escolas públicas.

### 5.1.5 Média das Notas por Regiões Brasileiras na Amostra

Com relação a média de notas por regiões brasileiras, precisa-se calcular as médias como na seção anterior, porém agora com os filtros separando as regiões. As colunas

Tabela 5.6: Estatísticas de notas por tipo de Escola.

| | | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|---|
| Escola Pública | Média | 512,07 | 527,75 | 518,48 | 525,29 |
| | Mediana | 507,55 | 534,4 | 523,80 | 513,20 |
| | Máxima | 715,4 | 761,80 | 675,40 | 892,80 |
| | 1° quartil | 462,40 | 468,60 | 479,90 | 450,42 |
| | 3° quartil | 561,12 | 589,40 | 561,00 | 588,45 |
| Escola Privada | Média | 576,68 | 587,34 | 560,93 | 622,25 |
| | Mediana | 584,20 | 599,40 | 568,85 | 625,70 |
| | Máxima | 735,50 | 723,90 | 703,80 | 930,00 |
| | 1° quartil | 528,80 | 551,22 | 525,00 | 534,30 |
| | 3° quartil | 622,90 | 643,42 | 601,75 | 704,40 |
| Escola de Exterior | Média | 611,21 | 629,33 | 590,15 | 660,02 |
| | Mediana | 624,90 | 638,60 | 591,50 | 635,10 |
| | Máxima | 755,10 | 735,30 | 665,40 | 911,33 |
| | 1° quartil | 580,45 | 608,15 | 561,95 | 582,55 |
| | 3° quartil | 654,15 | 651,05 | 611,85 | 744,90 |

relacionadas a elas são: "NO_MUNICIPIO_RESIDENCIA" que filtra pelo município onde o inscrito mora e "SG_UF_RESIDENCIA", que filtra pelo estado onde o inscrito habita. Para simplificar foram escolhidos alguns estados de cada região geográfica do país de forma sortida. Os selecionados foram: São Paulo, Rio de Janeiro, Goiás, Minas Gerais, Mato Grosso, Amazonas, Rondônia, Rio Grande do Sul, Paraná, Pará, Maranhão, Bahia, Ceará e Sergipe. Lembrando aqui que o Distrito Federal foi deixado para a seção seguinte pra ser estudado mais a fundo por ser a região onde esse trabalho foi executado. Serão apresentadas as médias de notas tiradas nas diversas regiões do Brasil.

A Tabela 5.7 mostra as médias de notas em cada um desses estados. Nota-se que nenhum estado brasileiro desses analisados ficaram com média inferior a 500 pontos e nenhum chegou a 600 pontos na amostra analisada.

### 5.1.6 Desempenho do Distrito Federal na Amostra

Essa parte foi feita como a seção anterior, mudando apenas o filtro para Distrito Federal, e com a ideia de investigar mais a fundo com relação às regiões administrativas. Foi feita a filtragem nos microdados e tirado uma amostra de 10000 inscritos do DF. Contudo, ao filtrar os dados do DF pelas regiões administrativas, nota-se que o Inep

Tabela 5.7: Médias de notas em cada Estado.

| | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|
| São Paulo | 530,97 | 558,55 | 544,88 | 552,66 |
| Rio de Janeiro | 542,98 | 566,42 | 545,64 | 567,03 |
| Minas Gerais | 546,52 | 566,9 | 548,35 | 567,56 |
| Goiás | 536,95 | 557,95 | 542,1 | 553,98 |
| Mato Grosso | 526,07 | 534,84 | 521,85 | 536,10 |
| Amazonas | 529,02 | 534,89 | 519,57 | 526,25 |
| Rondônia | 511,34 | 542,76 | 523,64 | 508,49 |
| Rio Grande do Sul | 531,51 | 555,31 | 545,86 | 546,49 |
| Paraná | 535,01 | 556,10 | 541,69 | 550,30 |
| Pará | 510,54 | 525,46 | 506,17 | 506,78 |
| Maranhão | 520,19 | 532,52 | 522,26 | 528,96 |
| Bahia | 518,78 | 533,83 | 520,95 | 522,95 |
| Ceará | 530,53 | 549,36 | 529,80 | 553,66 |
| Sergipe | 543,68 | 563,62 | 538,41 | 562,02 |

não fez a separação e considerou todos os cadastros dos inscritos como se estivessem no município de Brasília, impossibilitando diferenciar o desempenho dos inscritos nas diversas regiões do DF.

As Figuras 5.5, 5.6, 5.7 e 5.8 apresentam os gráficos de distribuição de notas por quantidade de inscritos com relação ao DF.

A Tabela 5.8 mostra as estatísticas de notas do DF em cada prova. Nota-se que ele acompanha quase o mesmo padrão dos outros estados, com médias acima de 500 e abaixo de 600 pontos, e com distribuição de notas medianas.

### 5.1.7 Porcentagem de Acerto por Questão em Cada Prova

Até aqui foram vistas informações gerais dos dados de notas dos inscritos, médias e até uma tentativa de descobrir a distribuição das notas nas regiões brasileiras. Porém, não deu para tirar muitas conclusões sobre o que precisa melhorar no conhecimento dos

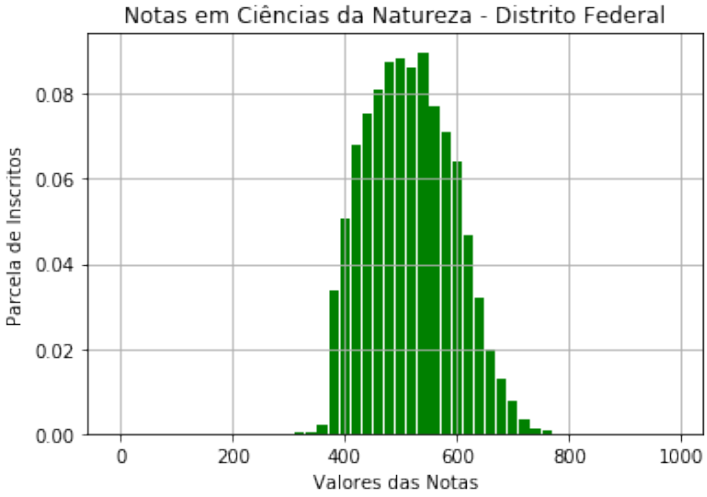



Figura 5.5: Gráfico de distribuição de notas de Ciências da Natureza por inscritos no Distrito Federal.

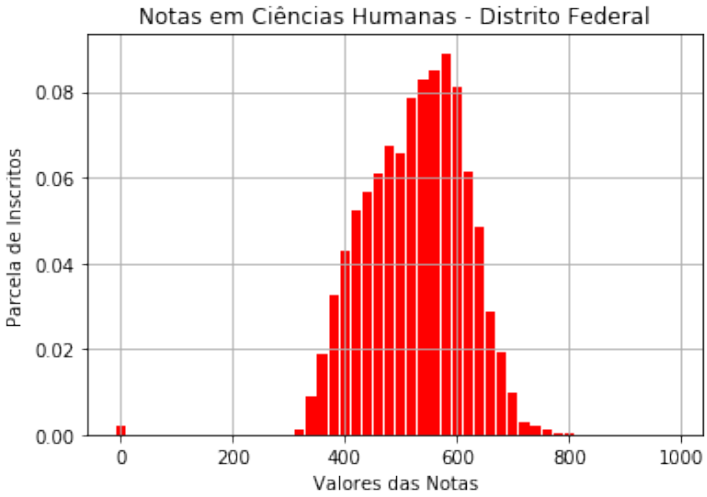



Figura 5.6: Gráfico de distribuição de notas de Ciências Humanas por inscritos no Distrito Federal.

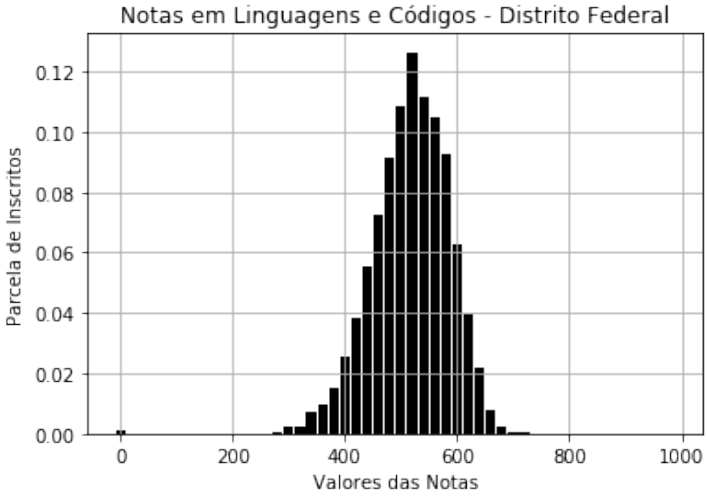



Figura 5.7: Gráfico de distribuição de notas de Linguagens e Códigos por inscritos no Distrito Federal.

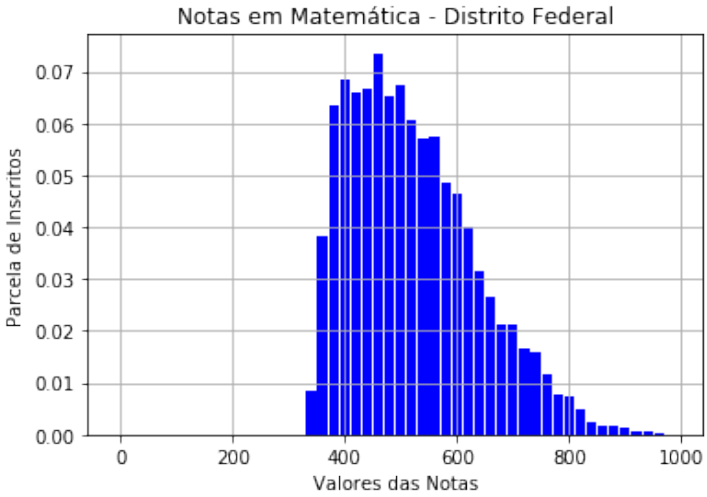



Figura 5.8: Gráfico de distribuição de notas de Matemática por inscritos no Distrito Federal.

Tabela 5.8: Estatísticas de notas do Distrito Federal.

| | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|
| Média | 525,87 | 536,25 | 527,18 | 531,69 |
| Mediana | 522,90 | 543,00 | 531,30 | 515,15 |
| Máxima | 771,50 | 819,70 | 730,90 | 974,70 |
| 1° quartil | 465,62 | 471,70 | 484,70 | 441,52 |
| 3° quartil | 582,50 | 601,10 | 575,20 | 603,87 |

inscritos e na elaboração da prova do Enem. Por isso, foi necessário buscar outro método de análise, separando questão por questão das provas.

Como dito anteriormente, a primeira transformação dos dados foi feita para tornar possível abstrair a porcentagem de acertos por questão em cada área de conhecimento. Na transformação foram criadas as colunas para cada questão da prova, inclusive separando por parte da prova e cor do caderno, contendo um bit de acerto ou erro (1 e 0 respectivamente) em cada linha. Dessa forma, para achar a porcentagem de acerto na questão, bastou aplicar a função de média do Pandas para cada coluna. Isso deu um trabalho maior porque é extensa a quantidade de colunas, porque além de 45 questões por prova, ainda tem a separação por cor de prova, onde é preciso seguir a ordem de questões de cada caderno e consultar corretamente o gabarito para cada cor. O objetivo aqui foi investigar quais questões apresentaram maior ou menor índice de erro e tentar descobrir os motivos disso para poder entender como os alunos respondem a prova.

Nesse caso, como são 45 questões por prova e 4 cadernos diferentes para cada, foi melhor organizar os dados em tabelas para mostrar aqui. No código feito no Pandas, foram aplicadas as funções para cada caderno, porém para não ficar muito extenso e repetitivo, a análise aqui foi sobre o caderno azul apenas. A Tabela 5.9 mostra a porcentagem de acerto em cada questão de cada parte da prova do Enem, com relação ao caderno azul. As colunas separam as provas (Ciências da Natureza, Ciências Humanas, Linguagens e códigos e Matemática) e cada linha é uma questão numerada (Q1 a Q45), além das últimas duas linhas que mostram as questões que tiveram maior e menor porcentagem de acertos. É preciso lembrar que o Enem 2017 teve uma reaplicação de provas em alguns estados que apresentaram queda de energia no dia da prova. Então, para a pesquisa foram desconsiderados os inscritos que fizeram reaplicação.

Tabela 5.9: Porcentagem de acerto por questão em cada parte da prova do Enem

| | **Ciências da Natureza** | **Ciências Humanas** | **Linguagens e Códigos** | **Matemática** |
|---|---|---|---|---|
| Q1 | 18,30 | 34,98 | 29,41 | 17,04 |
| Q2 | 30,60 | 61,22 | 27,52 | 15,61 |
| Q3 | 30,83 | 24,62 | 16,71 | 29,11 |
| Q4 | 44,90 | 59,10 | 31,75 | 45,19 |
| Q5 | 14,64 | 56,99 | 26,18 | 11,89 |
| Q6 | 37,41 | 67,01 | 76,10 | 15,16 |
| Q7 | 31,29 | 44,51 | 47,91 | 24,42 |
| Q8 | 28,26 | 22,72 | 45,73 | 10,12 |
| Q9 | 56,00 | 40,89 | 61,72 | 34,78 |
| Q10 | 54,46 | 51,97 | 33,59 | 26,54 |
| Q11 | 64,18 | 50,52 | 74,37 | 28,08 |
| Q12 | 11,72 | 28,85 | 28,02 | 30,54 |
| Q13 | 18,13 | 68,91 | 26,35 | 34,38 |
| Q14 | 33,00 | 37,99 | 52,86 | 34,61 |
| Q15 | 44,73 | 48,91 | 51,97 | 46,10 |
| Q16 | 8,00 | 58,38 | 43,06 | 25,68 |
| Q17 | 41,87 | 38,49 | 58,49 | 38,50 |
| Q18 | 12,52 | 32,14 | 31,64 | 19,90 |
| Q19 | 40,21 | 38,83 | 73,03 | 69,16 |
| Q20 | 46,10 | 41,33 | 21,55 | 57,83 |
| Q21 | 27,97 | 24,40 | 72,03 | 22,82 |

**Tabela 5.9 – continuação**

| | Ciências da Natureza | Ciências Humanas | Linguagens e Códigos | Matemática |
|---|---|---|---|---|
| Q22 | 24,42 | 48,13 | 47,13 | 45,19 |
| Q23 | 15,96 | 17,93 | 64,84 | 21,51 |
| Q24 | 16,99 | 34,65 | 39,72 | 31,12 |
| Q25 | 51,25 | 23,11 | 17,71 | 6,40 |
| Q26 | 18,42 | 53,59 | 44,12 | 30,77 |
| Q27 | 19,10 | 48,02 | 78,27 | 13,10 |
| Q28 | 52,74 | 35,93 | 62,11 | 30,14 |
| Q29 | 14,30 | 43,17 | 17,32 | 49,02 |
| Q30 | 25,17 | 18,71 | 37,21 | 16,87 |
| Q31 | 19,67 | 44,45 | 59,94 | 19,73 |
| Q32 | 27,05 | 35,15 | 56,04 | 16,76 |
| Q33 | 25,62 | 57,21 | 14,42 | 31,00 |
| Q34 | 24,02 | 25,90 | 86,29 | 17,90 |
| Q35 | 34,49 | 39,77 | 29,02 | 38,78 |
| Q36 | 36,61 | 27,57 | 24,17 | 30,37 |
| Q37 | 22,99 | 26,62 | 63,11 | 12,98 |
| Q38 | 24,65 | 25,68 | 39,38 | 29,17 |
| Q39 | 10,69 | 52,92 | 20,00 | 27,80 |
| Q40 | 32,55 | 25,40 | 52,70 | 24,08 |
| Q41 | 9,43 | 53,09 | 36,04 | 26,65 |
| Q42 | 58,86 | 41,72 | 53,20 | 29,23 |

**Tabela 5.9 – continuação**

| | Ciências da Natureza | Ciências Humanas | Linguagens e Códigos | Matemática |
|---|---|---|---|---|
| Q43 | 16,53 | 36,15 | 63,84 | 57,03 |
| Q44 | 20,53 | 17,15 | 47,29 | 27,80 |
| Q45 | 31,23 | 11,97 | 29,58 | 17,84 |
| Média das Taxas | 29,50 | 39,48 | 44,74 | 28,63 |
| Maior Acerto | 64,18 | 68,91 | 86,29 | 69,16 |
| Menor Acerto | 8,00 | 11,97 | 14,42 | 6,40 |

Agora será feita a análise de cada parte da prova do Enem, utilizando os dados mostrados na Tabela 5.9. Através das provas e gabaritos disponibilizados pelo Inep, junto à pasta que tem os microdados do Enem, será possível consultar as questões de maior destaque.

Como são dois dias de prova na avaliação do Enem, no primeiro aplica-se a prova de Ciências Humanas e Linguagens e códigos, e no segundo a de Ciências da Natureza e Matemática. Porém a primeira questão da prova de Ciências da Natureza vem numerada como questão 91, pois seria a sequência das primeiras duas provas de 45 questões. No código do Pandas para ficar mais simples, foi considerado a separação de prova em prova, considerando cada prova com as 45 questões de 1 a 45. Então, quando for citado aqui, questão 1 de uma prova, para consultar deve-se olhar para o enunciado da primeira questão daquela prova e não para o número da questão no caderno de prova necessariamente.

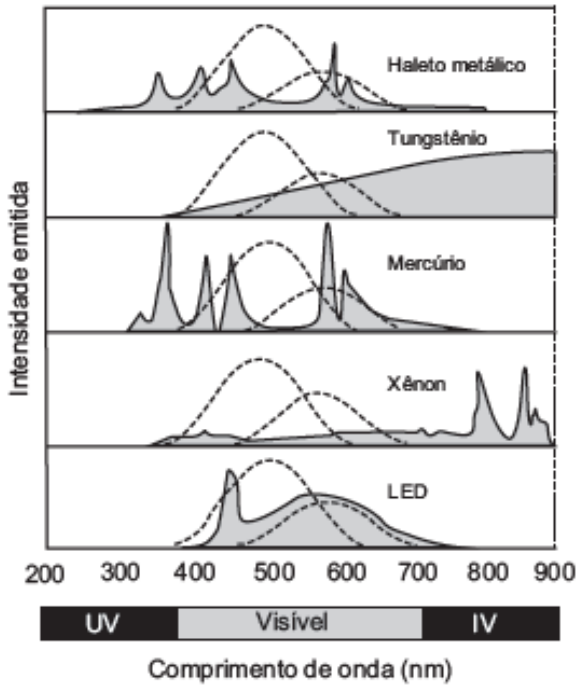
### Ciências da Natureza

A questão com maior porcentagem de acerto foi a questão 11 (questão 101 da prova geral) com 64,18%. Essa questão fala sobre a emissão de radiação eletromagnética por diversos tipos de lâmpada e pede para marcar qual a melhor lâmpada para certa aplicação de um arquiteto, como mostrado na Figura 5.9. A alternativa correta é a letra "E". A questão se torna bem intuitiva por mostrar um desenho esquemático de simples análise. Pode ser por isso que ela teve maior quantidade de acertos.

As questões que tiveram uma média de acertos na faixa de 50% foram as 10, 25, 28 e 42 (questões 100, 115, 118 e 132 respectivamente da prova geral). Questão 10 questiona sobre patologias. Questão 25 questiona sobre o funcionamento de uma centrífuga. Questão 28

A figura mostra como é a emissão de radiação eletromagnética para cinco tipos de lâmpada: haleto metálico, tungstênio, mercúrio, xênon e LED (diodo emissor de luz). As áreas marcadas em cinza são proporcionais à intensidade da energia liberada pela lâmpada. As linhas pontilhadas mostram a sensibilidade do olho humano aos diferentes comprimentos de onda. UV e IV são as regiões do ultravioleta e do infravermelho, respectivamente.

Um arquiteto deseja iluminar uma sala usando uma lâmpada que produza boa iluminação, mas que não aqueça o ambiente.

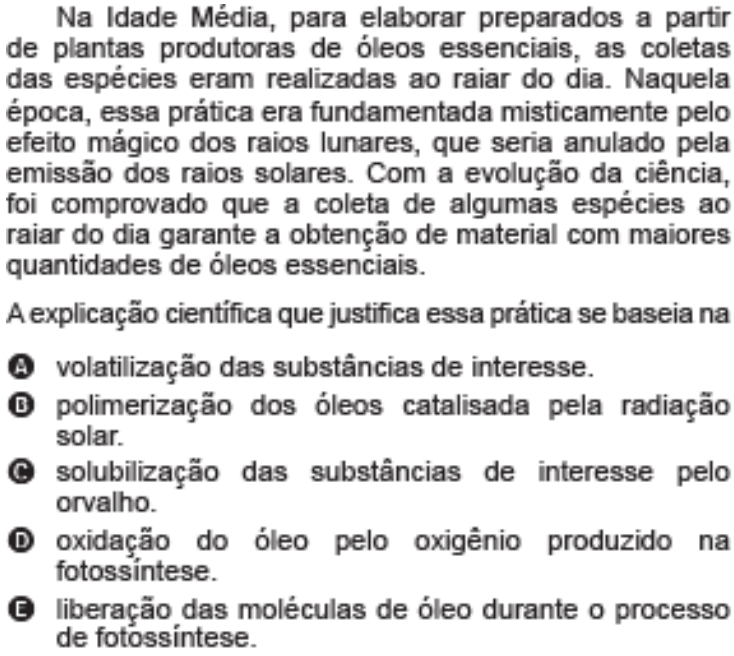



Disponível em: http://zeiss-campus.magnet.fsu.edu. Acesso em: 8 maio 2017 (adaptado).

Qual tipo de lâmpada melhor atende ao desejo do arquiteto?

A Haleto metálico.
B Tungstênio.
C Mercúrio.
D Xênon.
E LED.

Figura 5.9: Fragmento da décima primeira questão da prova de Ciências Naturais do Enem 2017.



Na Idade Média, para elaborar preparados a partir de plantas produtoras de óleos essenciais, as coletas das espécies eram realizadas ao raiar do dia. Naquela época, essa prática era fundamentada misticamente pelo efeito mágico dos raios lunares, que seria anulado pela emissão dos raios solares. Com a evolução da ciência, foi comprovado que a coleta de algumas espécies ao raiar do dia garante a obtenção de material com maiores quantidades de óleos essenciais.

A explicação científica que justifica essa prática se baseia na

A volatilização das substâncias de interesse.
B polimerização dos óleos catalisada pela radiação solar.
C solubilização das substâncias de interesse pelo orvalho.
D oxidação do óleo pelo oxigênio produzido na fotossíntese.
E liberação das moléculas de óleo durante o processo de fotossíntese.

Figura 5.10: Fragmento da décima sexta questão da prova de Ciências Naturais do Enem 2017.

questiona sobre "piso concregrama". E a questão 42 fala sobre usar sal para conservação de alimentos.

Já a questão com menor porcentagem de acerto foi a questão 16 (questão 106 da prova geral) com 8%. Essa questão traz uma curiosidade histórica da época medieval que usavam técnicas em plantas produtoras de óleos essenciais. Na época, acreditavam em algo mágico, mas depois com a evolução da ciência o efeito foi explicado. Assim, se pergunta qual a explicação científica relata o fenômeno, como mostra a Figura 5.10. A alternativa correta é a letra "A". Na seção seguinte, será visto mais a fundo como os inscritos reagiram a essa questão.

### Ciências Humanas

A questão com maior porcentagem de acerto foi a questão 13 (questão 58 da prova geral) com 68,91%. A questão fala sobre sociedade multiétnicas e o texto de referência é bem explicativo e parece sugerir a resposta, como mostrado na Figura 5.11. A alternativa correta é a letra "B". Por isso pode ter tido maior número de acertos.

As questões que tiveram uma média de acertos na faixa de 50% foram as 10, 11, 26, 39 e 41 (questões 55, 56, 71, 84 e 86 respectivamente da prova geral). Questão 10 é sobre empobrecimento do solo em colheitas. A questão 11 fala sobre problemas no comércio de soja no início do século 21. A questão 26 pergunta sobre a usina de Belo Monte.

Muitos países se caracterizam por terem populações multiétnicas. Com frequência, evoluíram desse modo ao longo de séculos. Outras sociedades se tornaram multiétnicas mais rapidamente, como resultado de políticas incentivando a migração, ou por conta de legados coloniais e imperiais.

GIDDENS, A. **Sociologia**. Porto Alegre: Penso, 2012 (adaptado).

Do ponto de vista do funcionamento das democracias contemporâneas, o modelo de sociedade descrito demanda, simultaneamente,

- **A** defesa do patriotismo e rejeição ao hibridismo.
- **B** universalização de direitos e respeito à diversidade.
- **C** segregação do território e estímulo ao autogoverno.
- **D** políticas de compensação e homogeneização do idioma.
- **E** padronização da cultura e repressão aos particularismos.

Figura 5.11: Fragmento da décima terceira questão da prova de Ciências Humanas do Enem 2017.

**Figura 1**

**Recorte fotográfico de Maria Bonita, década de 1930.**

ABRAÃO, B. Disponível em: www.brasilcult.pro.br. Acesso em: 18 maio 2013.

**Figura 2**

**Traje de coleção de Zuzu Angel.**

Disponível em: www.zuzuangel.com.br. Acesso em: 18 maio 2013.

Elaborada em 1969, a releitura contida na Figura 2 revela aspectos de uma trajetória e obra dedicadas à

- **A** valorização de uma representação tradicional da mulher.
- **B** descaracterização de referências do folclore nordestino.
- **C** fusão de elementos brasileiros à moda da Europa.
- **D** massificação do consumo de uma arte local.
- **E** criação de uma estética de resistência.

Figura 5.12: Fragmento da quadragésima quinta questão da prova de Ciências Humanas do Enem 2017.

A questão 39 questiona sobre o conceito de democracia. E a questão 41 cita o último terremoto ocorrido no Chile e pergunta sobre o efeito dele. São perguntas voltadas para meio ambiente.

Já a questão com menor porcentagem de acerto foi a questão 45 (questão 90 da prova geral) com 11,97%. Questão muito interessante sobre releitura de uma foto de "Maria Bonita", que leva o estudante a pensar sobre como é a imagem da mulher aos olhos da sociedade, como mostrado na Figura 5.12. A alternativa correta é a letra "E". Talvez a subjetividade do enunciado e alternativas diversificadas tenha levado ao grande número de erros.

**Linguagens e Códigos**


Época, n. 698, 3 out. 2011 (adaptado).

Os textos publicitários são produzidos para cumprir determinadas funções comunicativas. Os objetivos desse cartaz estão voltados para a conscientização dos brasileiros sobre a necessidade de

A) as crianças frequentarem a escola regularmente.
B) a formação leitora começar na infância.
C) a alfabetização acontecer na idade certa.
D) a literatura ter o seu mercado consumidor ampliado.
E) as escolas desenvolverem campanhas a favor da leitura.

Figura 5.13: Fragmento da trigésima quarta questão da prova de Linguagens e Códigos do Enem 2017.

A questão com maior porcentagem de acerto foi a questão 34 (questão 34 da prova geral) com 86,29%. Essa questão traz um texto publicitário de incentivo à leitura desde criança para analisar e pergunta sobre o objetivo de textos nesse formato, como mostrado

João/Zero (Wagner Moura) é um cientista genial, mas infeliz porque há 20 anos atrás foi humilhado publicamente durante uma festa e perdeu Helena (Alinne Moraes), uma antiga e eterna paixão. Certo dia, uma experiência com um de seus inventos permite que ele faça uma viagem no tempo, retornando para aquela época e podendo interferir no seu destino. Mas quando ele retorna, descobre que sua vida mudou totalmente e agora precisa encontrar um jeito de mudar essa história, nem que para isso tenha que voltar novamente ao passado. Será que ele conseguirá acertar as coisas?

Disponível em: http://adorocinema.com. Acesso em: 4 out. 2011.

Qual aspecto da organização gramatical atualiza os eventos apresentados na resenha, contribuindo para despertar o interesse do leitor pelo filme?

A) O emprego do verbo *haver*, em vez de *ter*, em "há 20 anos atrás foi humilhado".
B) A descrição dos fatos com verbos no presente do indicativo, como "retorna" e "descobre".
C) A repetição do emprego da conjunção "mas" para contrapor ideias.
D) A finalização do texto com a frase de efeito "Será que ele conseguirá acertar as coisas?".
E) O uso do pronome de terceira pessoa "ele" ao longo do texto para fazer referência ao protagonista "João/Zero".

Figura 5.14: Fragmento da trigésima terceira questão da prova de Linguagens e Códigos do Enem 2017.

na Figura 5.13. A alternativa correta é a letra "B". Questão fala de um assunto importante, onde a grande maioria acertou. Nota-se que foi a questão mais acertada em toda a prova do Enem, considerando a amostra de 10000 inscritos.

As questões que tiveram uma média de acertos na faixa de 50% foram as 14, 15, 40 e 42 (questões 14, 15, 40 e 42 respectivamente da prova geral). Questão 14 é de interpretação de texto, onde o texto fala sobre propaganda. Questão 15, uma poesia que traz uma palavra diferente e pergunta qual sentido dela no verso. Questão 40 mostra um fragmento de narração, com um história de divisão social dentro de uma família com serviçal e questiona sobre como o autor abordou isso no texto. A questão 42 traz parte da música "Fim de semana no parque" da banda "Racionais MCs" e pergunta sobre a realidade social criticada na letra. Interessante notar aqui, que todas essas questões foram de interpretação de texto.

Já a questão com menor porcentagem de acerto foi a questão 33 (questão 33 da prova geral) com 14,42%. Questão apresenta um texto tratando de uma sinopse de um filme de Wagner Moura e pergunta qual foi o elemento ou elementos gramaticais usados para chamar atenção para as pessoas verem o filme, como mostrado na Figura 5.14. A alternativa correta é a letra "B". As alternativas apresentam assertivas que soam ser corretas em

primeira análise e avalia o candidato com relação à interpretação do texto e conhecimento gramatical. Isso pode ser o motivo para grande maioria ter errado a questão.

### Matemática

A questão com maior porcentagem de acerto foi a questão 19 (questão 154 da prova geral) com 69,16%. Questão com enunciado bem explicativo, figura com boa legenda e representação simples e fácil de entender, como mostrado na Figura 5.15. A alternativa correta é a letra "A". Esperava-se ainda mais acertos do que 69,16% apenas, o que mostra falta de atenção nos inscritos que erraram esse tipo de questão.

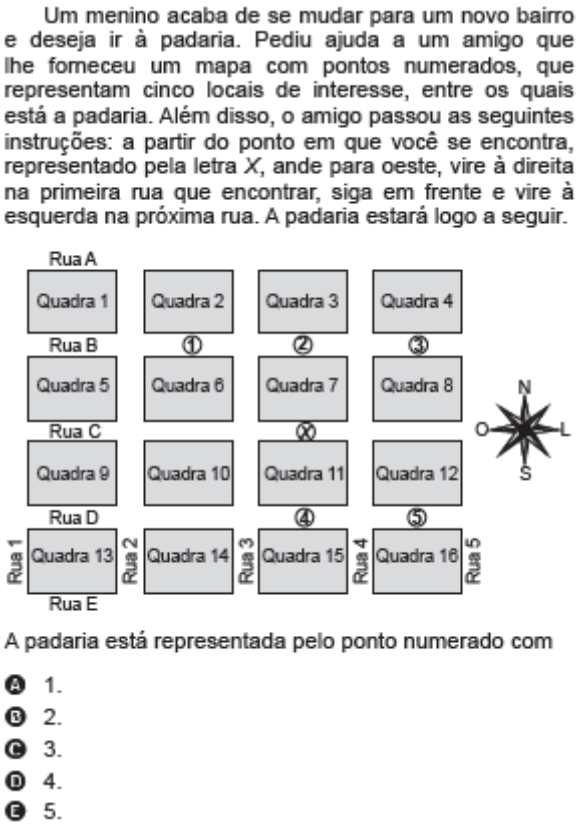



Figura 5.15: Fragmento da décima nona questão da prova de Matemática do Enem 2017.

As questões que tiveram uma média na faixa de 50% de acerto foram as 20, 29 e 43 (155, 164 e 178 respectivamente da prova geral). A questão 20 era de aplicação de média aritmética sobre as notas. Bastava fazer a conta e marcar a alternativa correta. A questão 29 usa um problema de trânsito e percorrimento de distância com carros para o estudante raciocinar e calcular uma razão. A questão 43 seria de raciocínio lógico sobre combinação de partidas de futebol, mas a combinação das alternativas facilitou a exclusão, tornando a questão mais fácil.

Já a questão com menor porcentagem de acerto foi a questão 25 (questão 160 da prova geral) com 6,4%. Essa questão colocou um assunto de copa do mundo de 2014, onde a logomarca da copa seria uma taça onde mãos se unem dando sentido de união em um só ritmo. A questão é sobre quantidade de combinações possíveis de cores que se poderia usar para pintar a taça seguindo um regra, como mostrado na Figura 5.16. A alternativa correta é a letra "E". Esse tipo de questão parece gerar muitas dúvidas nos alunos por não ter um padrão modelado a se seguir. Isso pode explicar a grande quantidade de erros.

O comitê organizador da Copa do Mundo 2014 criou a logomarca da Copa, composta de uma figura plana e o *slogan* "Juntos num só ritmo", com mãos que se unem formando a taça Fifa. Considere que o comitê organizador resolvesse utilizar todas as cores da bandeira nacional (verde, amarelo, azul e branco) para colorir a logomarca, de forma que regiões vizinhas tenham cores diferentes.

JUNTOS NUM SÓ RITMO

Disponível em: www.pt.fifa.com. Acesso em: 19 nov. 2013 (adaptado).

De quantas maneiras diferentes o comitê organizador da Copa poderia pintar a logomarca com as cores citadas?

A 15
B 30
C 108
D 360
E 972

Figura 5.16: Fragmento da vigésima quinta questão da prova de Matemática do Enem 2017.

### 5.1.8 Distribuição de Marcação de Alternativas nas Principais Questões

Como dito anteriormente, a segunda transformação dos dados foi feita para tornar possível calcular a porcentagem de marcação de alternativas em cada questão. Na transformação foram criadas as colunas para cada questão da prova, separadas por parte da

prova e cor do caderno, e nelas informam qual foi a alternativa marcada pelo inscrito. Dessa maneira, fica fácil saber quais alternativas cada questão teve maior marcação ou menor marcação. O objetivo disso foi tentar descobrir um padrão de raciocínio dos inscritos para determinados tipos de questões.

Para isso, no Pandas foram utilizados duas funções: "describe()" e ".value_counts()". A primeira já foi explicada e descreve a coluna em si com dados estatísticos. A segunda retorna a quantidade de vezes que cada dado apareceu na coluna, ideal para o objetivo nessa parte do trabalho, que seria descobrir a quantidade marcações de alternativas na questão em análise.

Nesse ponto, usando a biblioteca do matplotlib foi possível gerar gráficos em barra que mostram a distribuição das marcações de alternativas em cada questão, usando a função ".value_counts().plot.bar()". No código feito no Pandas, foram aplicadas as funções para todas as questões, porém para não ficar muito extenso e repetitivo, a análise aqui foi sobre as questões vistas na seção 4.5.5, caderno azul que tiveram maior,menor e taxa média de acertos.

A seguir, serão apresentados gráficos em barra, onde um eixo determina a alternativa e no outro a quantidade de inscritos que marcaram cada uma delas. O asterisco significa dupla marcação e o ponto é quando o aluno deixou em branco. Os dados usados para os resultados foram de 10000 inscritos, onde 1795 desses fizeram as provas de Ciências Humanas e Linguagens e Códigos, e 1748 fizeram as provas de Ciências Naturais e Matemática, todos pelo caderno azul. Por isso nos gráficos de barras iremos notar menor quantidade de inscritos assinalando as alternativas.

**Ciências da Natureza**

- Questão 11 - Maior porcentagem de acertos. A Figura 5.17 mostra a distribuição de marcação das alternativas. Volte na Figura 5.9 para consultar o que dizia cada alternativa. Lembrando que a resposta correta é a letra "E".

- Questão 16 - Menor porcentagem de acertos. A Figura 5.18 mostra a distribuição de marcação das alternativas. Volte na Figura 5.10 para consultar o que dizia cada alternativa. Lembrando que a resposta correta é a letra "A". Dá pra ver que a maioria assinalou a alternativa E, mas a opinião dos inscritos variou bem com relação a todas as alternativas.

**Ciências Humanas**

- Questão 13 - Maior porcentagem de acertos. A Figura 5.19 mostra a distribuição de marcação das alternativas. Volte na Figura 5.11 para consultar o que dizia cada

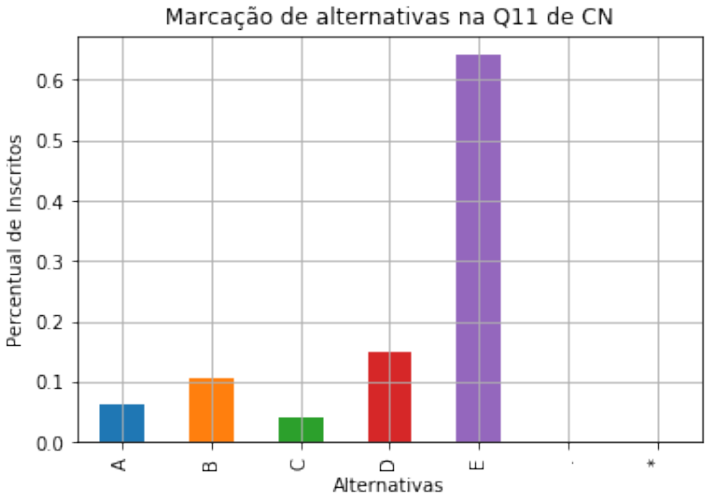



Figura 5.17: Gráfico de distribuição de marcação de alternativas na questão 11 em Ciências da Natureza por quantidade de inscritos.

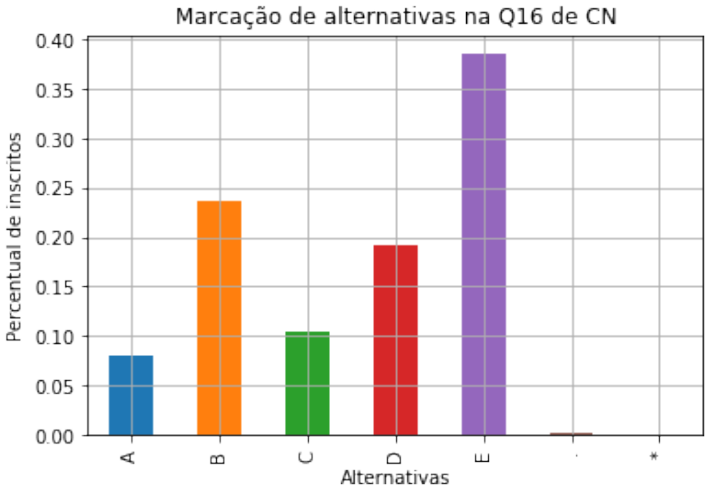



Figura 5.18: Gráfico de distribuição de marcação de alternativas na questão 16 em Ciências da Natureza por quantidade de inscritos.

alternativa. Lembrando que a resposta correta é a letra “B”.

- Questão 45 - Menor porcentagem de acertos. A Figura 5.20 mostra a distribuição de marcação das alternativas. Volte na Figura 5.12 para consultar o que dizia cada alternativa. Lembrando que a resposta correta é a letra “E”. Nota-se que a maioria marcou alternativa “A” e “C”.

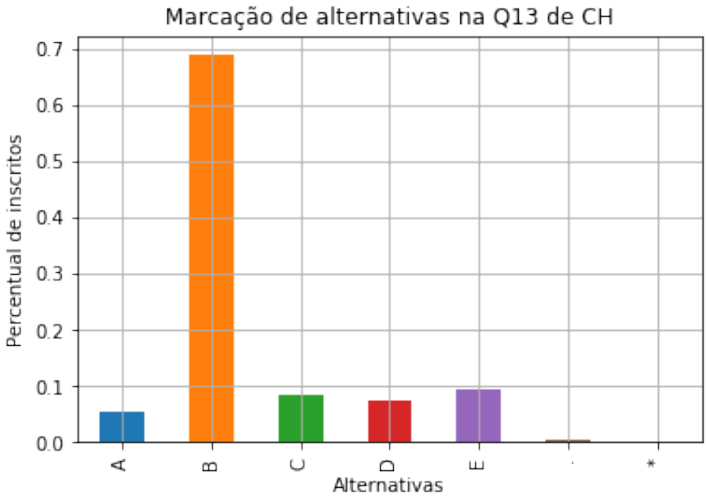



Figura 5.19: Gráfico de distribuição de marcação de alternativas na questão 13 em Ciências Humanas por quantidade de inscritos.

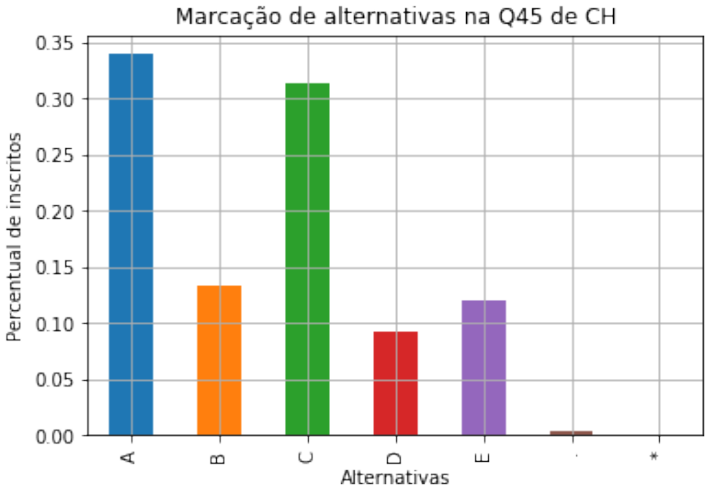



Figura 5.20: Gráfico de distribuição de marcação de alternativas na questão 45 em Ciências Humanas por quantidade de inscritos.

### Linguagens e Códigos

- Questão 34 - Maior porcentagem de acertos. A Figura 5.21 mostra a distribuição de marcação das alternativas. Volte na Figura 5.21 para consultar o que dizia cada alternativa. Lembrando que a resposta correta é a letra “B”.

- Questão 33 - Menor porcentagem de acertos. A Figura 5.22 mostra a distribuição de marcação das alternativas. Volte na Figura 5.22 para consultar o que dizia cada

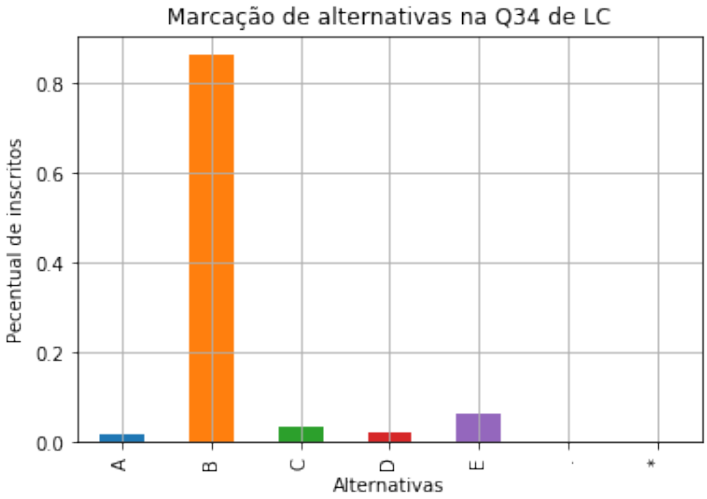



Figura 5.21: Gráfico de distribuição de marcação de alternativas na questão 34 em Linguagens e Códigos por quantidade de inscritos.

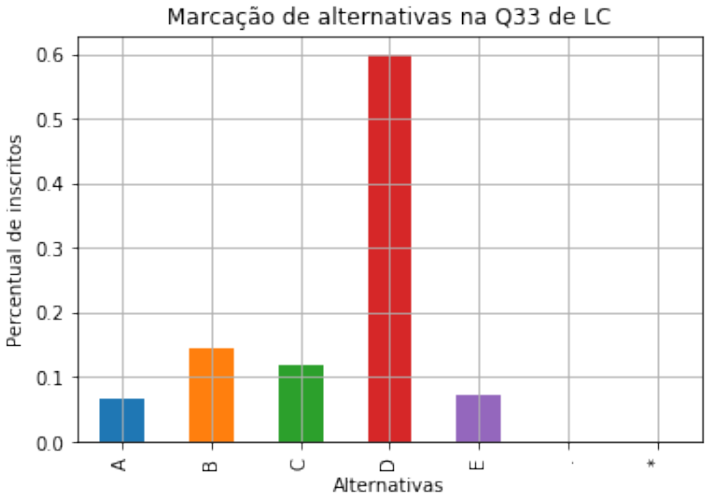



Figura 5.22: Gráfico de distribuição de marcação de alternativas na questão 33 em Linguagens e Códigos por quantidade de inscritos.

alternativa. Lembrando que a resposta correta é a letra "B". Dá pra notar que a maioria marcou a alternativa "D".

### Matemática

- Questão 19 - Maior porcentagem de acertos. A Figura 5.23 mostra a distribuição de marcação das alternativas. Volte na Figura 5.15 para consultar o que dizia cada alternativa. Lembrando que a resposta correta é a letra "A".

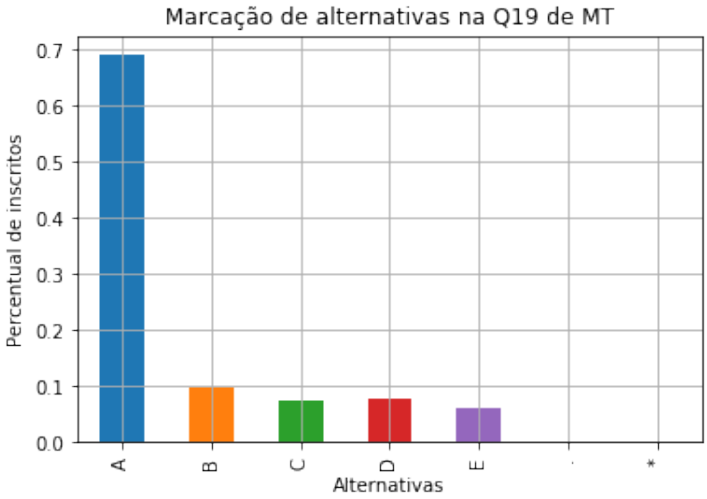



Figura 5.23: Gráfico de distribuição de marcação de alternativas na questão 19 em Matemática por quantidade de inscritos.

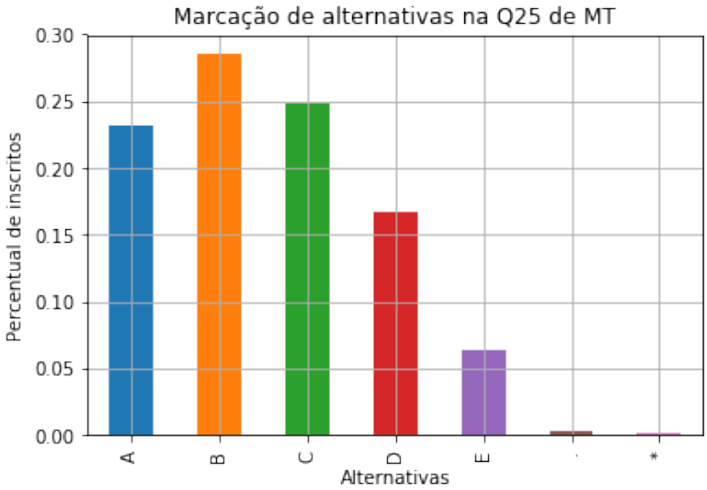



Figura 5.24: Gráfico de distribuição de marcação de alternativas na questão 25 em Matemática por quantidade de inscritos.

- Questão 25 - Menor porcentagem de acertos. A Figura 5.24 mostra a distribuição de marcação das alternativas. Volte na Figura 5.16 para consultar o que dizia cada alternativa. Lembrando que a resposta correta é a letra “E”. Aqui, deu pra ver que os estudantes tiveram uma distribuição bem abrangente entre as alternativas, com a maior parte marcando letra “B”. Poucas pessoas marcaram letra “E”, por ser um valor muito alto.

Tabela 5.10: Comparativo entre as principais questões de Ciências da Natureza.

| | | **Maior acerto - Q11(%)** | **Menor acerto - Q16(%)** |
|---|---|---|---|
| Gênero | Masculino | 67,57 | 9,94 |
| | Feminino | 61,95 | 6,73 |
| Cor/Raça | Branco | 68,94 | 6,36 |
| | Preto | 55,55 | 9,12 |
| | Pardo | 63,12 | 8,70 |
| | Amarelo | 65,71 | 11,42 |
| | Indígena | 20,00 | 0,00 |
| Renda | Renda zero R$0,00 | 52,17 | 13,04 |
| | Renda até R$937,00 | 57,23 | 7,07 |
| | R$1874,01 até R$2342,50 | 68,12 | 10,00 |
| | R$3748,01 até R$4685,00 | 70,42 | 8,45 |
| | R$9370,01 até R$11 244,00 | 78,12 | 21,87 |
| | Renda superior a R$18 740,00 | 84,21 | 5,26 |
| Escola | Pública | 58,11 | 7,54 |
| | Privada | 75,00 | 8,33 |
| | Exterior | 100,00 | 0,00 |

### 5.1.9 Comparação Usando as Principais Questões Apresentadas

Nessa parte, serão apresentadas as médias de acertos nas principais questões vistas comparando entre gênero, raça, renda e tipo de escola. Lembrando sempre que as estatísticas foram tiradas da amostra de 10000 inscritos.

#### Ciências da Natureza

A Tabela 5.10 mostra o comparativo entre as questões 11 (maior taxa de acerto) e 16 (menor taxa de acerto) na prova de Ciências da Natureza.

#### Ciências Humanas

A Tabela 5.11 mostra o comparativo entre as questões 13 (maior taxa de acerto) e 45 (menor taxa de acerto) na prova de Ciências Humanas.

Tabela 5.11: Comparativo entre as principais questões de Ciências Humanas.

| | | **Maior acerto - Q13(%)** | **Menor acerto - Q45(%)** |
|---|---|---|---|
| Gênero | Masculino | 72,19 | 12,41 |
| | Feminino | 66,60 | 11,66 |
| Cor/Raça | Branco | 71,25 | 11,70 |
| | Preto | 63,63 | 15,70 |
| | Pardo | 67,95 | 11,63 |
| | Amarelo | 75,00 | 5,00 |
| | Indígena | 58,33 | 8,33 |
| Renda | Renda zero R$0,00 | 61,29 | 19,35 |
| | Renda até R$937,00 | 58,43 | 9,01 |
| | R$1874,01 até R$2342,50 | 70,58 | 14,11 |
| | R$3748,01 até R$4685,00 | 76,74 | 8,13 |
| | R$9370,01 até R$11 244,00 | 90,00 | 15,00 |
| | Renda superior a R$18 740,00 | 88,23 | 35,29 |
| Escola | Pública | 62,72 | 11,11 |
| | Privada | 74,62 | 7,46 |
| | Exterior | 83,33 | 0,00 |

**Linguagens e Códigos**

A Tabela 5.12 mostra o comparativo entre as questões 34 (maior taxa de acerto) e 33 (menor taxa de acerto) na prova de Linguagens e Códigos.

**Matemática**

A Tabela 5.13 mostra o comparativo entre as questões 19 (maior taxa de acerto) e 25 (menor taxa de acerto) na prova de Matemática.

### 5.1.10 Inscritos com Maiores Notas (1% melhores)

Nessa parte será feita a análise sobre os inscritos que tiveram as maiores notas em cada parte da prova, mais especificamente sobre o 1% mais bem colocados. Com isso, descobrir algum padrão entre eles, gênero dominante, cor/raça, renda, tipo de escola, etc.

Tabela 5.12: Comparativo entre as principais questões de Linguagens e Códigos.

| | | **Maior acerto - Q34(%)** | **Menor acerto - Q33(%)** |
|---|---|---|---|
| Gênero | Masculino | 85,56 | 14,17 |
| | Feminino | 86,81 | 14,61 |
| Cor/Raça | Branco | 89,18 | 13,77 |
| | Preto | 86,77 | 14,87 |
| | Pardo | 83,60 | 14,39 |
| | Amarelo | 82,50 | 20,00 |
| | Indígena | 91,66 | 8,33 |
| Renda | Renda zero R$0,00 | 70,96 | 12,90 |
| | Renda até R$937,00 | 78,48 | 16,86 |
| | R$1874,01 até R$2342,50 | 86,47 | 11,76 |
| | R$3748,01 até R$4685,00 | 94,18 | 10,46 |
| | R$9370,01 até R$11 244,00 | 95,00 | 15,00 |
| | Renda superior a R$18 740,00 | 94,11 | 11,76 |
| Escola | Pública | 81,72 | 14,33 |
| | Privada | 92,53 | 16,41 |
| | Exterior | 100,00 | 0,00 |

No pandas, foi utilizado a função ".quantile(.99)" que retorna o número de linhas que estão no 1% maiores. Com essa parte de inscritos, aplica-se os filtros para as classes que deseja informação e utilizando a função ".value_counts()", temos a distribuição quantitativa de inscritos onde, por exemplo, pode-se determinar quantos inscritos são do sexo masculino e feminino nos 1% melhores. Lembrando que alguns inscritos não fizeram ou estiveram ausentes em algumas das provas, então não necessariamente dentro dos 10000 inscritos há 100 inscritos analisados no 1% melhores.

### Ciências Naturais

Na amostra de 10000 inscritos, apenas 7200 foram válidos. Assim, os 1% melhores foi composto apenas de 72 inscritos se destacando como melhores na prova de Ciências da Natureza. A seguir, serão apresentadas as estatísticas relacionadas aos 1% melhores e quantidades separadas por gênero, cor/raça, renda e tipo de escola:

Tabela 5.13: Comparativo entre as principais questões de Matemática.

| | | **Maior acerto - Q19(%)** | **Menor acerto - Q25(%)** |
|---|---|---|---|
| Gênero | Masculino | 73,91 | 5,18 |
| | Feminino | 66,03 | 7,21 |
| Cor/Raça | Branco | 75,93 | 8,22 |
| | Preto | 60,71 | 3,17 |
| | Pardo | 66,45 | 6,01 |
| | Amarelo | 68,57 | 2,85 |
| | Indígena | 80,00 | 2,00 |
| Renda | Renda zero R$0,00 | 60,86 | 10,86 |
| | Renda até R$937,00 | 59,69 | 7,07 |
| | R$1874,01 até R$2342,50 | 67,50 | 5,62 |
| | R$3748,01 até R$4685,00 | 76,05 | 5,63 |
| | R$9370,01 até R$11 244,00 | 78,12 | 15,62 |
| | Renda superior a R$18 740,00 | 89,47 | 15,78 |
| Escola | Pública | 67,92 | 5,66 |
| | Privada | 81,25 | 20,83 |
| | Exterior | 100,00 | 0,00 |

- A média ficou em 727,54 pontos, acima da média geral que foi de 531,54 pontos;
- Com relação ao gênero, 44 são do sexo Masculino, 28 são do sexo Feminino.
- Com relação à cor/raça, 47 são da cor/raça branca, 5 são pretos, 18 são pardos, 1 amarelo e nenhum indígena. O restante foi de quem não quis declarar cor ou raça.
- Com relação às rendas, houve maior quantidade de inscritos da faixa de renda superior a 18740 reais que foi de 11 inscritos. Tiveram 9 inscritos na faixa de renda de 11.244,01 até 14.055 reais, 6 inscritos na faixa de renda de 6.559,01 até 7.496 reais, 6 inscritos na faixa de 3.748,01 até 4.685 reais, 6 inscritos na faixa de 4.685,01 até 5.622 reais, 5 inscritos na faixa de 2.342,51 até 2.811 reais, 5 inscritos na faixa de 2.811,01 até 3.748 reais. Nas outras faixas de rendas tiveram de 4 inscritos para baixo, e com renda zero apareceu 1 inscrito entre as melhores notas;

- Com relação ao tipo de escola, 62 desses inscritos não responderam a essa pergunta, 7 são de escolas privadas, 4 são de escolas de exterior e 1 de escola pública;
- Analisando pelas principais questões da prova:
  - Questão 11 (Maior porcentagem de acerto na amostra de 10 mil inscritos) - 52 inscritos acertaram e 20 erraram. Ou seja, 72,22% de taxa de acerto no 1% melhores, contra 64,18% na amostra geral;
  - Questão 16 (Menor porcentagem de acerto na amostra de 10 mil inscritos) - 51 inscritos acertaram, 21 erraram. Ou seja, 70,83% de taxa de acerto no 1% melhores, contra 8% na amostra geral.

**Ciências Humanas**

Na amostra de 10000 inscritos, apenas 7400 foram válidos. Assim, os 1% melhores foi composto apenas de 74 inscritos se destacando como melhores na prova de Ciências Humanas. A seguir, serão apresentadas as estatísticas relacionadas aos 1% melhores e as quantidades separadas por gênero, cor/raça, renda e tipo de escola:

- A média ficou em 743,33 pontos, acima da média geral que foi de 551,33 pontos;
- Com relação ao gênero, 44 são do sexo Masculino, 30 são do sexo Feminino.
- Com relação à cor/raça, 42 são da cor/raça branca, 3 são pretos, 24 são pardos, 1 amarelo e nenhum indígena. O restante foi de quem não quis declarar cor ou raça.
- Com relação às rendas, houve 7 inscritos da faixa de renda superior a 18740 reais, 7 inscritos na faixa de renda de 2.811,01 até 3.748 reais, 6 inscritos na faixa de renda de 6.559,01 até 7.496 , 6 inscritos na faixa de 3.748,01 até 4.685 reais, 6 inscritos na faixa de 4.685,01 até 5.622 reais, 5 inscritos na faixa de 11.244,01 até 14.055 reais, 5 inscritos na faixa de 9.370,01 até 11.244 reais, 5 inscritos na faixa de 1.874,01 até 2.342,50 reais, 5 inscritos na faixa de 2.342,51 até 2.811 reais, 5 inscritos na faixa de 5.622,01 até 6.559 reais. Nas outras faixas de rendas tiveram de 4 inscritos para baixo, e com renda zero não apareceu nenhum inscrito entre as melhores notas;
- Com relação ao tipo de escola, 70 desses inscritos não responderam a essa pergunta, 2 são de escolas privadas, 1 de escola de exterior e 1 de escola pública;
- Analisando pelas principais questões da prova:
  - Questão 13 (Maior porcentagem de acerto na amostra de 10 mil inscritos) - 57 inscritos acertaram, 17 erraram. Ou seja, 77,02% de taxa de acerto no 1% melhores, contra 68,91% na amostra geral;

  - Questão 45 (Menor porcentagem de acerto na amostra de 10mil inscritos) - 54 inscritos acertaram, 20 erraram. Ou seja, 72,97% de taxa de acerto no 1% melhores, contra 11,97% na amostra geral.

**Linguagens e Códigos**

Na amostra de 10000 inscritos, apenas 7400 foram válidos. Assim, os 1% melhores foi composto apenas de 74 inscritos se destacando como melhores na prova de Linguagens e Códigos. A seguir, serão apresentadas as estatísticas relacionadas aos 1% melhores e as quantidades separadas por gênero, cor/raça, renda e tipo de escola:

- A média ficou em 677,38 pontos, acima da média geral que foi de 535,87 pontos;
- Com relação ao gênero, 42 são do sexo Feminino, 32 são do sexo Masculino.
- Com relação à cor/raça, 48 são da cor/raça branca, 4 são pretos, 18 são pardos, 1 amarelo e nenhum indígena. O restante foi de quem não quis declarar cor ou raça.
- Com relação às rendas, houve 9 inscritos da faixa de renda de 4.685,01 até 5.622 reais, 7 inscritos na faixa de renda de 5.622,01 até 6.559 reais, 6 inscritos na faixa de renda de 6.559,01 até 7.496 , 6 inscritos na faixa de renda superior a 18740 reais, 6 inscritos na faixa de 2.811,01 até 3.748 reais, 6 inscritos na faixa de 2.342,51 até 2.811 reais, 5 inscritos na faixa de 11.244,01 até 14.055 reais, 5 inscritos na faixa de 6.559,01 até 7.496 reais, 5 inscritos na faixa de 7.496,01 até 8.433 reais, 5 inscritos na faixa de 9.370,01 até 11.244 reais. Nas outras faixas de rendas tiveram de 4 inscritos para baixo, e com renda zero não apareceu nenhum inscrito entre as melhores notas;
- Com relação ao tipo de escola, 66 desses inscritos não responderam a essa pergunta, 4 são de escola privada, 3 são de escolas públicas e 1 de escola de exterior;
- Analisando pelas principais questões da prova:
  - Questão 34 (Maior porcentagem de acerto na amostra de 10 mil inscritos) - 72 inscritos acertaram, 2 erraram. Ou seja, 97,29% de taxa de acerto no 1% melhores, contra 86,29% na amostra geral;
  - Questão 33 (Menor porcentagem de acerto na amostra de 10 mil inscritos) - 43 inscritos acertaram, 31 erraram. Ou seja, 58,10% de taxa de acerto no 1% melhores, contra 14,42% na amostra geral.

**Matemática**

Na amostra de 10000 inscritos, apenas 7100 foram válidos. Assim, os 1% melhores foi composto apenas de 71 inscritos se destacando como melhores na prova de Matemática. A seguir, serão apresentadas as estatísticas relacionadas aos 1% melhores e as quantidades separadas por gênero, cor/raça, renda e tipo de escola:

- A média ficou em 870,46 pontos, acima da média geral que foi de 547,37 pontos;
- Com relação ao gênero, 50 são do sexo Masculino, 21 são do sexo Feminino.
- Com relação à cor/raça, 41 são da cor/raça branca, 4 são pretos, 22 são pardos, 2 são amarelos e nenhum indígena. O restante foi de quem não quis declarar cor ou raça.
- Com relação às rendas, houve maior quantidade de inscritos da faixa de renda superior a 18740 reais que foi de 10 inscritos. Tiveram 9 inscritos na faixa de renda de 9.370,01 até 11.244 reais, 8 inscritos na faixa de renda de 4.685,01 até 5.622 reais, 7 inscritos na faixa de 7.496,01 até 8.433 reais, 6 inscritos na faixa de 3.748,01 até 4.685 reais, 5 inscritos na faixa de 2.342,51 até 2.811 reais. Nas outras faixas de rendas tiveram de 4 inscritos para baixo, e com renda zero não apareceu nenhum inscrito entre as melhores notas;
- Com relação ao tipo de escola, 61 desses inscritos não responderam a essa pergunta, 6 são de escolas privadas, 3 são de escolas de exterior e 2 de escolas públicas;
- Analisando pelas principais questões da prova:
  - Questão 19 (Maior porcentagem de acerto na amostra de 10 mil inscritos) - 59 inscritos acertaram, 12 erraram. Ou seja, 83,09% de taxa de acerto no 1% melhores, contra 69,16% na amostra geral;
  - Questão 25 (Menor porcentagem de acerto na amostra de 10 mil inscritos) - 42 acertaram, 29 erraram. Ou seja, 59,15% de taxa de acerto no 1% melhores, contra 6,4% na amostra geral.

# Capítulo 6

# Conclusão

Os dados do Enem podem revelar informações úteis para Instituições de Ensino aplicarem em seus contextos com o objetivo de melhorar a qualidade do ensino médio no Brasil, uma vez que o Enem é uma avaliação feita por milhões de estudantes com características distintas: diversas regiões do país, diferentes classes sociais, gêneros, cor/raça, vindos de diversas escolas. Sendo assim, esse trabalho vem para contribuir com mais informações educacionais e contribuir com pesquisas na área por meio de uma abordagem com mineração de dados, onde foi possível identificar padrões de desempenho na avaliação e realizar comparativos entre os diferentes inscritos.

Foi mostrado de forma geral como está a educação brasileira, tanto no ensino básico como no ensino superior, focando na avaliação do Enem como porta de entrada nas IES. Para a mineração de dados, foi usado um conjunto de ferramentas que utilizam bibliotecas baseadas em Python: a biblioteca do Pandas que consegue simplificar todo o processo de mineração de dados.

Os objetivos propostos foram alcançados: alguns fatores que influenciam no desempenho dos alunos foram elucidados e foi gerado um sistema de análise para microdados do Enem. Os *scripts* gerados utilizando a biblioteca Pandas funcionam para o modelo de dados do Enem disponibilizado pelo INEP. Para utilizá-los, basta mudar os dados de entrada e fazer as alterações necessárias para a aplicação em amostras maiores.

A etapa de preparação dos dados foi fundamental para a realização do trabalho, além do aprendizado de técnicas de mineração de dados utilizando a biblioteca do Pandas. A biblioteca do Pandas foi muito eficaz, pois dispensou a utilização de um sistema gerenciador de banco de dados (SGBD) externo, já que ela organizou os dados em um *DataFrame* permitindo a sua fácil utilização.

O banco de dados do Enem tem milhões de registros, o que torna inviável fazer o processamento completo com as máquinas disponíveis. Assim, foi necessário dividir os dados e utilizar apenas uma amostra significativa para conseguir os resultados. No caso, eles

foram separados por classes (gênero, raça, renda e tipo de escola), por regiões brasileiras e por diferentes questões para obter melhores resultados.

Após decidir as ferramentas para mineração, foi feito o planejamento com base no modelo CRISP-DM, onde foram seguidos os passos de preparação dos dados (aquisição, pré-processamento e transformação dos dados). Por fim, foi feita a mineração dos dados, descoberta de padrões e resultados finais apresentados e comentados.

Nesse trabalho foram mostradas quais foram as questões com maior índice de acerto e maior índice de erro por prova, assim como as taxas de acerto para os diferentes perfis.

## 6.1 Contribuições

É possível estender o algoritmo gerado no final para qualquer outra pesquisa em cima de dados do Enem, desde que siga o formato dos dados tratados aqui. Assim, mais pesquisas podem surgir nessa área e unir resultados em busca de melhorar a educação brasileira. Pelos comentários feitos nas rotinas e funções, é possível que qualquer pessoa com conhecimento na área possa entender e fazer as mudanças necessárias, além de criar novas funções dependendo do problema.

## 6.2 Trabalhos Futuros

Algumas possibilidades de trabalhos futuros derivadas desse trabalho são mostradas a seguir.

- Realizar um comparativo dos resultados deste trabalho com os obtidos em trabalhos anteriores sobre o Enem (microdados antigos);
- Usar novas técnicas de mineração de dados com outro algoritmo ou ferramenta e comparar com os resultados usando Pandas;
- Otimizar código gerado nesse trabalho, em busca de criar uma ferramenta mais genérica que aceite mais tipos de dados de entrada usando menos linhas de código;
- Fazer um estudo mais aprofundado sobre o Distrito Federal ou qualquer outra região de interesse, ou também sobre em alunos de alguma escola específica;
- Incluir uma análise sobre as notas de redação, que foi desconsiderada nesse trabalho.

# Referências


[1] Brasil: *Constituição da República Federativa do Brasil*, 1988. http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. 1

[2] Ministério da Educação: *ENEM - Apresentação*. http://portal.mec.gov.br/enem-sp-2094708791, Visitado em Janeiro de 2019. 1

[3] Descomplica: *O que é o enem - Tudo Sobre Enem*, 2018. https://descomplica.com.br/tudo-sobre-enem/enem/o-que-e-o-enem/, Visitado em Outubro de 2018. 1

[4] Enem Virtual: *Tudo sobre o enem*, 2018. https://www.enemvirtual.com.br/tudo-sobre-o-enem/, Visitado em Outubro de 2018. 1, 2

[5] Inep - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. http://www.inep.gov.br/. 2

[6] MEC - Ministério da Educação. https://www.mec.gov.br/. 2

[7] João Carlos Sedraz Silva, Rodrigo Lins Rodrigues; Jorge Luis Cavalcanti Ramos; e Alex Sandro Gomes: *A literatura brasileira sobre mineração de dados educacionais*. 3º Congresso Brasileiro de Informática na Educação (CBIE 2014), 2014, ISSN (WCBIE 2014). 2, 9

[8] Inep - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira: *Censo Escolar - Inep*. http://portal.inep.gov.br/web/guest/censo-escolar, Visitado em Outubro de 2018. 4

[9] Ideb - Índice de Desenvolvimento da Educação Básica. http://portal.inep.gov.br/ideb, Visitado em Outubro de 2018. 4

[10] PNE - Plano Nacional da Educação. http://pne.mec.gov.br/, Visitado em Outubro de 2018. 4

[11] Observatório PNE - Plano Nacional da Educação. http://www.observatoriodopne.org.br, Visitado em Outubro de 2018. 4

[12] INEP: *Censo Escolar - Boletim n°°15 09/2018*, Setembro 2018. http://download.inep.gov.br/educacao_basica/educacenso/documentos/2018/boletim_censo_escolar_n15_set2018.pdf, Visitado em Dezembro de 2018. 5

[13] INEP: *Resumo Técnico - Resultados do índice de desenvolvimento da educação básica*, 2018. http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/ResumoTecnico_Ideb_2005-2017.pdf, Visitado em Dezembro de 2018. 5

[14] INEP: *Resumo Técnico - Censo da Educação Superior 2015.* Relatório Técnico, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, 2018. http://download.inep.gov.br/educacao_superior/censo_superior/resumo_tecnico/resumo_tecnico_censo_da_educacao_superior_2015.pdf. 6

[15] Luckesi, Cipriano Carlos: *Maneiras de Avaliar a Aprendizagem.* Pátio. São Paulo, ano 3. nº 12, 2000. 8

[16] Luckesi, Cipriano Carlos: *Avaliação da Aprendizagem Escolar - Estudos e Proposições.* 22ª edição, Cortez Editora, 2014. 8

[17] Todos pela Educação: *Quais são as avaliações brasileiras e por que elas são importantes?*, julho 2018. https://www.todospelaeducacao.org.br/conteudo/uais-sao-as-avaliacoes-brasileiras-e-porque-elas-sao-importantes/, Visitado em Dezembro de 2018. 8

[18] Ministério da Educação: *Avaliações de aprendizagem.* http://portal.mec.gov.br/secretaria-de-educacao-basica/190-secretarias-112877938/setec-1749372213/18843-avaliacoes-da-aprendizagem, Visitado em Janeiro de 2019. 8

[19] CBIE - Congresso Brasileiro de Informática na Educação. http://www.br-ie.org/pub/index.php/wcbie/article/view/6401/4450, Visitado em Outubro de 2018. 9

[20] de Frias, Jorge Luiz Dias: *Uma ferramenta para a obtenção e análise de dados do enem*, 2015. 9, 37

[21] Gomes, Tancicleide Carina Simões: *Descoberta de Conhecimento Utilizando Mineração de Dados Educacionais Abertos*, 2015. 10

[22] Alves, Rafael Damiani: *Predição do desempenho da redação do Enem utilizando técnicas de mineração de dados*, 2018. 10

[23] Juliete A. R. Costa, André L. Reis, Daniel C. L. Souza Kaessa G. S. Cristino Marcelo M.Aureliano Salles R. Soares Thiago E. Santos e Yasmin V. S. Silva: *Técnicas de mineração de dados aplicados em dados do Enem 2015.* ABMES Cadernos, páginas 1–4, 2017, ISSN ISSN 2319–0124. 11

[24] Shannon, Claude Elwood: *A Mathematical Theory of Communication.* The Bell System Technical Journal, XXVII(3), 1948. 13

[25] Siqueira, M. C.: *Gestão Estratégica da Informação.* Brasport, Rio de Janeiro, 2005. 13

[26] Russo, M.: *Fundamentos de Biblioteconomia e Ciência da Informação. Biblioteconomia e gestão de unidades de informação.* Série Didáticos. e-papers, 2010. 13

[27] Laudon, Kenneth C. e Jane Price Laudon: *Sistemas de Informação Gerenciais*. Editora Pearson, 9a edição edição, 2011. 13

[28] O'Brien, James A.: *Sistemas de Informacao e as Decisões Gerenciais na Era da Internet*. Editora Saraiva, 9a edição edição, 2001, ISBN 85-02-03276-3. 13

[29] Stair, Ralph M.: *Princípios de Sistemas de Informação*. tradução da 9ª edição, CENGAGE Learning, 2010. 14

[30] Sérgio da Costa Côrtes, Rosa Maria Porcaro e Sérgio Lifschitz: *Mineração de Dados — Funcionalidades, Técnicas e Abordagens*, 2002. 15, 18, 19, 20, 21

[31] Passos, Emmanuel e Ronaldo Goldshmidt: *Data Mining: Um Guia Prático*. Elsevier, 2005. 15, 23

[32] Groth, Robert: *Data Mining: Building Competitive Advantage*. Prentice Hall, 1999, ISBN 0-13-086271-1. 15, 22

[33] Jiawei Han e Micheline Kamber: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001. 15, 19, 21, 22, 23

[34] Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From Data Mining to Knowledge Discovery in Databases*. Artificial Intelligence Magazine, páginas 36–54, 1996. 16

[35] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR) Thomas Khabaza (SPSS) Thomas Reinartz (DaimlerChrysler) Colin Shearer (SPSS) e Rüdiger Wirth (DaimlerChrysler): *Crisp-dm 1.0: Step-by-step data mining guide*, 2000. `https://www.the-modeling-agency.com/crisp-dm.pdf`. 17

[36] IBM Corporation: *Ibm spss modeler crisp-dm guide*, 2011. `https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf`. 17

[37] Gonsalves, Elisa Pereira: *Conversas Sobre Iniciação à Pesquisa Científica*. Alínea, 4a edição edição, 2005, ISBN 85-7516-002-8. 20

[38] Camilo, Cássio Oliveira e João Carlos da Silva: *Mineração de dados: Conceitos, tarefas, métodos e ferramentas*. RT-INF_001-09, páginas 12–15, 2009. 22

[39] Ferreira, Rodrigo Santana: *10 ferramentas e bibliotecas para trabalhar com data mining e Big Data - Parte 01 e Parte 02*, Junho 2017. `https://imasters.com.br/data/10-ferramentas-e-bibliotecas-para-trabalhar-com-data-mining-e-big-data-parte-01`, Visitado em Janeiro de 2019. 24, 25

[40] Python Software Foundation: *Our Documentation*, 2019. `https://www.python.org/doc/`, Visitado em Janeiro de 2019. 24

[41] Jupyter: *The Jupyter Notebook*, 2019. `https://jupyter-notebook.readthedocs.io/en/stable/`, Visitado em Janeiro de 2019. 24

[42] Matplotlib: *Matplotlib Version 3.1.0 Overview*, 2019. `https://matplotlib.org/3.1.0/contents.html`, Visitado em Janeiro de 2019. 24

[43] Pandas: *pandas 0.21.1 documentation*, 2019. https://pandas.pydata.org/pandas-docs/version/0.21/index.html, Visitado em Janeiro de 2019. 25, 30

[44] Scikit Learn: *User Guide*, 2019. https://scikit-learn.org/stable/user_guide.html, Visitado em Janeiro de 2019. 25

[45] Anaconda: *Anaconda Documentation*, 2019. https://docs.anaconda.com/, Visitado em Janeiro de 2019. 25

[46] R Project: *R - Documentation*, 2019. https://www.r-project.org/other-docs.html, Visitado em Janeiro de 2019. 25

[47] RapidMiner: *rapidminer Documentation*, 2019. https://docs.rapidminer.com/, Visitado em Janeiro de 2019. 26

[48] The University of Waikato: *Documentation*, 2019. https://www.cs.waikato.ac.nz/ml/weka/documentation.html, Visitado em Janeiro de 2019. 26

[49] INEP: *Microdados - Enem*, 2019. http://inep.gov.br/microdados, Visitado em Outubro de 2018. 28

[50] Anaconda: *Anaconda Distribution*, 2019. https://www.anaconda.com/distribution/, Visitado em Janeiro de 2019. 30

[51] Pandas: *pandas.dataframe.describe*, 2019. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html, Visitado em Janeiro de 2019. 37

[52] Conroy, Ronán: *Sample size: A rough guide*. https://pdfs.semanticscholar.org/4781/878153e13322c028c7d8970e7f52fbaa102a.pdf. 40

# Apêndice A

# Códigos em Python - Células do Jupyter

O repositório com todos os códigos utilizados no trabalho e as amostras está disponível em: `https://github.com/klarkgable/ProjetoFinalEngComp`.

## A.1 Transformação dos dados 1 - Colunas com Zero ou Um de Acerto e Erro e Funções Gerais dos Resultados

São os *scripts* usados para realizar a primeira transformação dos dados que criou uma coluna para cada questão das provas e preenchendo com zero ou um caso o inscrito acertasse ou errasse a questão, que serviu para os resultados da Subseção 5.1.7.

Além disso, tem todas as funções para tirar os resultados gerais mostrados nas Subseções 5.1.1, 5.1.2, 5.1.3, 5.1.4, 5.1.5, 5.1.9 e 5.1.10.

O código Python está com nome "Enem01Geral.py" e pode ser acessado em : `https://github.com/klarkgable/ProjetoFinalEngComp/blob/master/Enem01Geral.py`.

## A.2 Transformação dos dados 2 - Colunas com alternativas marcadas em cada questão da prova do Enem

São os *scripts* usados para realizar a segunda transformação dos dados que criou uma coluna para cada questão preenchendo com a alternativa marcada pelo inscrito naquela

questão. Isso permitiu demonstrar a distribuição de marcação das alternativas visto na Subseção 5.1.8.

O código Python está com nome "EnemAB.py" e pode ser acessado em: `https://github.com/klarkgable/ProjetoFinalEngComp/blob/master/EnemAB.py`.

## A.3 Funções para a Amostra do Distrito Federal

São os *scripts* usados para tirar os resultados com relação ao Distrito Federal com uma amostra de 10000 inscritos dessa região, como mostrado na Subseção 5.1.6.

O código Python está com o nome "EnemDF.py" e pode ser acessado em: `https://github.com/klarkgable/ProjetoFinalEngComp/blob/master/EnemDF.py`.