



UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

**Análise das características determinantes para o bom
desempenho dos concluintes de Engenharias no ENADE.**

Eric Laurino Gondran

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientadora: Profa. Dra. Héliida Salles Santos
Co-orientador: Prof. Dr. Eduardo Nunes Borges

Rio Grande, 2024

Ficha Catalográfica

G637a Gondran, Eric Laurino.

Análise das características determinantes para o bom desempenho dos concluintes de Engenharias no ENADE / Eric Laurino Gondran. – 2024.

75 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em computação, Rio Grande/RS, 2024.

Orientador: Dra. Helida Salles Santos.

Coorientador: Dr. Eduardo Nunes Borges.

1. ENADE 2. Seleção de atributos 3. Mineração de dados educacionais 4. Ciências de dados I. Santos, Helida Salles II. Borges, Eduardo Nunes III. Título.

CDU 004



Dissertação de Mestrado

Análise das características determinantes para o bom desempenho dos concluintes de Engenharias no ENADE.

Eric Laurino Gondran

Banca examinadora:

Documento assinado digitalmente
 **GIANCARLO LUCCA**
Data: 06/03/2024 14:29:34-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Giancarlo Lucca

Documento assinado digitalmente
 **LUIS OTONI MEIRELES RIBEIRO**
Data: 06/03/2024 14:36:34-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Luis Otoni Meireles Ribeiro

Documento assinado digitalmente
 **HELIDA SALLES SANTOS**
Data: 06/03/2024 14:41:15-0300
Verifique em <https://validar.iti.gov.br>

Prof^a. Dr^a. Héli da Salles Santos

AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer a Deus, pela minha vida e por me ajudar a superar todos os obstáculos encontrados ao longo do caminho.

Agradeço à minha mãe Marione Laurino e meu pai Alvair Gondran, por todo o amor, dedicação e esforço envolvido desde criança, para que essa conquista fosse possível. Em especial a minha esposa Fernanda Kucharski, por estar presente ao meu lado nos momentos mais difíceis e me incentivar à seguir em busca dos meus sonhos e conquistas.

Gostaria de agradecer aos meus professores orientadores, Hélida Santos e Eduardo Borges, por acreditarem no meu potencial, por estarem dispostos a me ajudar, colaborando com ideias e sugestões que me proporcionaram chegar neste momento. Também gostaria de agradecer à Universidade Federal do Rio Grande e ao Programa de Pós-Graduação em Computação, pela estrutura oferecida.

*Sonhos determinam o que você quer.
Ação determina o que você conquista.*
— ALDO NOVAK

RESUMO

GONDRAN, Eric Laurino. **Análise das características determinantes para o bom desempenho dos concluintes de Engenharias no ENADE..** 2024. 75 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Nas últimas duas décadas, o Brasil fez investimentos na expansão do acesso ao nível superior de ensino, mas este é um fenômeno observável em todo o mundo e não somente no país. Todos estes investimentos fizeram-se necessários, pois a carência de profissionais, mais especificamente nas Engenharias, mostrava-se evidente. No entanto, atualmente esse cenário inverteu-se devido ao fato do fluxo de recém-formados ter sido mais elevado do que o crescimento da demanda. Tendo em vista que para avaliar a qualidade do sistema educacional brasileiro nos diferentes cursos superiores, o Exame Nacional do Desempenho dos Estudantes (ENADE) é aplicado anualmente, este trabalho aplica diferentes técnicas de seleção de características nas bases de dados do ENADE ao longo dos anos. O objetivo principal é analisar como os aspectos pessoais e socioeconômicos são variáveis significativas na explicação da performance dos concluintes de graduação de engenharias ao longo dos anos. As técnicas de seleção de características utilizadas incluem as seguintes medidas: ganho de informação, taxa de ganho, incerteza simétrica, correlação de Pearson e alívio F. Os dados utilizados dizem respeito ao ciclo avaliativo do Ano II, sendo eles: 2005, 2008, 2011, 2014 e 2017. Pode-se concluir, dentre outros resultados, que quatorze atributos foram considerados importantes ao longo de todos os anos analisados. Com destaque para o atributo Idade, que foi importante nos cinco anos analisados. Outros atributos também considerados importantes foram: Situação de Trabalho, Bolsa Acadêmica e Escolarização da Mãe, que se destacaram em quatro anos dentre aqueles do ciclo avaliativo. Por fim, percebeu-se que alguns atributos perderam importância, enquanto outros ganharam relevância com o passar dos anos.

Palavras-chave: ENADE, Seleção de Atributos, Mineração de Dados Educacionais, Ciências de Dados.

ABSTRACT

GONDRAN, Eric Laurino. **Analysis of the determining characteristics for the performance of Engineering Graduates in ENADE exam.** 2024. 75 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Over the last two decades, Brazil has invested in expanding access to higher education. A phenomenon also observed throughout the world, besides Brazil. All these investments were necessary, as the lack of professionals, more specifically in the field of Engineering, was evident. However, currently, this scenario has reversed, because the flow of recent graduates has been higher than the growth in demand. To assess the quality of the Brazilian educational system, for the different higher education courses, the National Student Performance Examination (or the known acronym in Portuguese, ENADE) is applied annually. Taking this into account, this study applies different feature selection techniques, namely: information gain, gain rate, symmetric uncertainty, Pearson correlation, and F relief, in ENADE databases over the years. We aim to analyze how personal and socioeconomic aspects are significant variables in explaining the performance of engineering graduates over the years considering Year II of the assessment cycle, regarding the following years: 2005, 2008, 2011, 2014, and 2017. Among other results, one can conclude that fourteen attributes were considered important throughout the years. We highlight the Age attribute, which was relevant in all of the five years analyzed. Other attributes regarded as important were: Work Status, Scholarship, and Mother's Schooling, which stood out in four years among those of the assessment cycle. Finally, we noticed that some attributes lost importance, while others gained relevance over the years.

Keywords: ENADE, Feature Selection, Educational Data Mining, Data Science.

LISTA DE FIGURAS

1	Etapas do processo de DCBD	19
2	Metodologia do Trabalho.	39
3	Resultado da Correlação.	46
4	Resultado do Ganho de Informação.	47
5	Resultado da Taxa de Ganho.	48
6	Resultado da Incerteza Simétrica.	49
7	Resultado do Alívio F.	50
8	Evolução do perfil de estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos, parte 1/3.	53
9	Evolução do perfil de estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos, parte 2/3.	54
10	Evolução do perfil de estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos, parte 3/3.	55
11	Relação entre Idade e Desempenho.	57
12	Relação da Situação de Trabalho e Desempenho.	58
13	Relação entre Bolsa Acadêmica e Desempenho.	59
14	Relação entre Escolarização da Mãe e Desempenho.	60
15	Relação entre Família com Curso Superior e Desempenho.	60
16	Relação entre Escolarização do Pai e Desempenho.	61
17	Relação entre Bolsa/Financiamento e Desempenho.	61
18	Relação entre Escola Ensino Médio e Desempenho.	62
19	Relação entre Estudo Semanal e Desempenho.	62
20	Relação da Renda Familiar e Desempenho.	63
21	Relação entre Escolha da Instituição e Desempenho.	63
22	Relação entre Onde E Com Quem Mora e Desempenho. (ENADE 2005)	64
23	Relação entre Estado Civil e Desempenho. (ENADE 2008)	65
24	Relação entre Bolsa Permanência e Desempenho. (ENADE 2011)	65

LISTA DE TABELAS

1	Tabela comparativa de trabalhos relacionados, parte 1/2.	37
2	Tabela comparativa de trabalhos relacionados, parte 2/2.	38
3	Relação do número de participantes e atributos considerados neste estudo.	40
4	Relação dos atributos considerados neste estudo, conforme ENADE 2005, 2008, 2011, 2014 e 2017.	41
5	Relação do número de participantes após cada etapa de pré-processamento.	43
6	Relação da discretização das notas e anos.	43
7	Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2005 para concluintes de Engenharia.	51
8	Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2008 para concluintes de Engenharia.	51
9	Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2011 para concluintes de Engenharia.	52
10	Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2014 para concluintes de Engenharia.	53
11	Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2017 para concluintes de Engenharia.	54
12	Resultados obtidos com a Correlação de Spearman.	56
13	Resultados obtidos com a Correlação de Kendall.	56

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AVA	Ambiente Virtual de Aprendizagem
CNCST	Catálogo Nacional de Cursos Superiores de Tecnologia
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CK	Correlação de Kendall
CS	Correlação de Spearman
CGK	Correlação de Goodman e Kruskal
CDS	Correlação D de Somers
CRISP-DM	<i>Cross-Industry Standard Process for Data Mining</i>
Daes	Diretoria de Avaliação de Educação Superior
DCBD	Descoberta de Conhecimento em Banco de Dados
EaD	Ensino a Distância
ENADE	Exame Nacional do Desempenho dos Estudantes
ENEM	Exame Nacional do Ensino Médio
ENC	Exame Nacional de Cursos
EWD	<i>Equal Width Discretization</i>
FIES	Fundo de Financiamento ao Estudante do Ensino Superior
GI	Ganho de Informação
GPL	<i>General Public License</i>
IES	Instituições de Ensino Superior
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	<i>Knowledge Discovery in Database</i>
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MEC	Ministério da Educação
PDA	Plano de Dados Abertos

ProUni	Programa Universidade para Todos
SA	Seleção Atributos
SAEB	Sistema Nacional de Avaliação da Educação Básica

SUMÁRIO

1	Introdução	14
1.1	Objetivos	16
1.2	Organização do trabalho	17
2	Fundamentação Teórica	18
2.1	Descoberta de conhecimento em banco de dados	18
2.2	Seleção de atributos	20
2.2.1	Entropia	23
2.2.2	Ganho de Informação	23
2.2.3	Taxa de Ganho	23
2.2.4	Incerteza Simétrica	24
2.2.5	Correlação de Pearson	24
2.2.6	Alívio F	25
2.3	Discretização	26
2.4	Correlação de Classificação	27
2.4.1	Correlação de Spearman	28
2.4.2	Correlação de Kendall	29
2.4.3	Correlação de Goodman e Kruskal	30
2.4.4	Correlação D de Somers	31
3	Trabalhos Relacionados	33
4	Metodologia	39
4.1	Seleção	39
4.2	Pré-processamento	42
4.3	Transformação	43
4.4	Mineração de dados	43
4.5	Interpretação e Avaliação	44
5	Resultados	45
5.1	Resultados obtidos em cada algoritmo de Seleção de Atributos	45
5.2	Resultados obtidos em cada ano do ENADE	48
5.3	Resultados obtidos de cada Atributo ao Longo dos Anos	52
5.4	Resultados obtidos da Correlação de Classificação	54
5.5	Resultados da Análise das Características mais Importantes	56
6	Conclusão	66

Referências	68
--------------------	-----------

ANEXOS

A Ano e área de aplicação do ENADE	75
---	-----------

1 INTRODUÇÃO

Nas últimas décadas, a taxa de matrícula mundial no ensino superior aumentou em 10% a cada dez anos [5]. Antes da pandemia de Covid-19, quatro em cada dez jovens frequentavam pelo menos um ano de curso de ensino superior [5]. Investimentos na expansão do acesso ao ensino superior constituem um fenômeno observável em todo o mundo [69]. Em 2020, observando países vizinhos na América do Sul, a parcela da população com idade entre 24 e 35 anos que concluiu uma graduação era de 34% no Chile e 40% na Argentina. No entanto, no Brasil era apenas 21%, contrastando com a referência mundial que é a Coreia do Sul com uma parcela de concluintes de 70% [5].

No Brasil, existe uma autarquia federal vinculada ao Ministério da Educação (MEC) que é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) [64] para averiguar a qualidade do sistema educacional brasileiro através de avaliações. Há três avaliações, tendo início com a prova do Sistema Nacional de Avaliação da Educação Básica (SAEB), depois prossegue com o Exame Nacional do Ensino Médio (ENEM), que avalia alunos no fim do Ensino Médio e, por fim, o Exame Nacional do Desempenho dos Estudantes (ENADE), que avalia os concluintes dos cursos de graduação.

Ao realizar avaliações periódicas em todos os níveis educacionais, o Inep detém informações importantes do percurso escolar dos alunos. Como consequência desse processo de avaliação, as escolas e Instituições de Ensino Superior (IES), são cada vez mais cobradas pelo desempenho de seus alunos e formandos, seja pelo governo ou pela sociedade [29]. As avaliações periódicas permitem observar o desempenho dos estudantes, as deficiências e potencialidades das instituições, além de possibilitar o cálculo de indicadores de qualidade que podem proporcionar melhorias no processo de ensino-aprendizagem [75].

A primeira tentativa empreendida pelo governo federal para avaliar cursos de graduação e os processos de ensino e de aprendizagem dos estudantes em IES brasileiras foi o denominado Exame Nacional de Cursos (ENC), conhecido como “Provão”, realizado entre os anos de 1996 e 2003 [13], e substituído pelo ENADE em 2004. Atualmente, o Inep elabora e aplica anualmente o ENADE e a sua concepção e operacionalização está a cargo da Diretoria de Avaliação de Educação Superior (Daes). O objetivo principal

do ENADE é acompanhar e avaliar o rendimento dos alunos concluintes de cursos de graduação, levando em consideração o conteúdo previsto nas diretrizes curriculares do respectivo curso em que o aluno está matriculado, além de avaliar o desenvolvimento de competências e habilidades necessárias para a formação do profissional no mercado de trabalho.

O ciclo avaliativo do ENADE determina as áreas de avaliação e os cursos a elas vinculados [34]. As áreas de conhecimento para os cursos de bacharelado e licenciatura derivam da tabela de áreas do conhecimento divulgada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Já os eixos tecnológicos são baseados no Catálogo Nacional de Cursos Superiores de Tecnologia (CNCST), do MEC. A distribuição das áreas de conhecimento e seu respectivo ano de aplicação podem ser vistos no anexo A. Através do exame são gerados os microdados do ENADE, que são o menor nível de granularidade de dados coletados. Eles atendem à demanda por informações específicas ao disponibilizar as provas, os gabaritos, as informações sobre os itens, as notas e o questionário do estudante.

Na área das ciências exatas, mais especificamente nos cursos de engenharias, é essencial que este tipo de avaliação seja aplicada. Isto porque, no Brasil, a carência de profissionais na área de Engenharia mostrava-se evidente, desde os primeiros anos do século 21 [19]. O número de engenheiros formados tinha um crescimento bastante lento, caracterizado por uma crescente taxa de evasão [19]. Atualmente esse cenário inverteu-se, devido ao fato do fluxo de recém-formados ter sido mais elevado que o crescimento da demanda [47]. O que se observa no mercado de trabalho é uma falta de engenheiros experientes, aliada à falta de qualidade dos engenheiros formados, ainda que hajam poucos dados confiáveis em relação a estes profissionais [66].

Avaliar as medidas de desempenho dos estudantes ao longo do tempo é essencial para entender o avanço da educação brasileira. No entanto, ainda mais importante, é detectar as razões pelas quais determinadas estratégias ou políticas públicas não foram efetivas. Assim, novas políticas podem ser propostas com base nos dados e não na vontade ou nas percepções dos governantes futuros.

Nesse contexto, o presente trabalho considera a crescente expansão do ensino superior das últimas décadas, destacando que conhecer o perfil dos alunos é um quesito fundamental, pois ajuda na tomada de decisão de diferentes políticas. Frente ao panorama atual, esse trabalho ganha ainda mais relevância, pois com o excesso de formandos nas mais diversas engenharias, pode-se compreender características importantes relacionadas aos estudantes.

Para extrair conhecimento deste conjunto de dados, utilizamos a metodologia chamada de Descoberta de Conhecimento em Banco de Dados (DCBD) [18], que se trata de um processo de extração de informações dividido em algumas etapas. Um dos passos da DCBD é a Mineração de Dados (MD) [72], que tem como objetivo analisar registros,

extraindo conteúdo útil por meio de algoritmos que criam um modelo com base nos dados.

A aplicação das técnicas de mineração de dados no campo da educação tem apresentado resultados promissores, dando origem a uma nova área de investigação científica denominada Mineração de Dados Educacionais (MDE) [2]. Tais técnicas têm sido amplamente exploradas para identificação de padrões e comportamentos de alunos através de um Ambiente Virtual de Aprendizagem (AVA) e exames nacionais de avaliação dos alunos.

As principais áreas de aplicação de MDE são: modelagem do estudante, modelagem do domínio, suporte pedagógico, e descoberta científica [3]. A MDE é um método que busca padrões e regras a partir de dados educacionais para gerar conhecimentos úteis e compreensíveis que subsidiem a tomada de decisão aos diversos níveis de gestores educacionais [12]. A MDE pode contribuir para identificação de fatores associados ao desempenho dos alunos, permitindo às instituições, ou mesmo ao próprio sistema educacional, a trabalhar de forma preventiva. Possibilita ainda a atuação em grupos, onde se pode identificar algum fator de risco de baixo desempenho, garantindo a busca constante pela qualidade na educação [64].

1.1 Objetivos

O objetivo geral deste trabalho é explorar os dados do ENADE, e analisar como os aspectos pessoais e socioeconômicos são variáveis significativas na explicação da performance dos concluintes de graduação de engenharias, ao longo dos anos do ciclo avaliativo, sendo eles: 2005, 2008, 2011, 2014 e 2017. Ao analisar os dados em um determinado período de tempo, acompanhando os perfis dos alunos, podemos compreender como eles vem se portando ao longo do tempo. Assim, medidas preventivas podem ser tomadas para melhorar o ensino dos cursos de engenharias de forma geral.

Para atingir o objetivo geral do trabalho, é preciso atingir alguns objetivos específicos, sendo eles:

- Revisar a literatura, para que seja identificado o Estado da Arte no tema proposto.
- Identificar nos artigos da revisão os tipos de análises e estratégias utilizadas.
- Avaliar a estrutura e distribuição dos dados das provas do ENADE.
- Analisar os resultados de cada algoritmo de Seleção de Atributos ao longo dos anos de estudo.
- Analisar o perfil dos alunos concluintes dos cursos de engenharias, para cada ano estudado.
- Analisar cada atributo ao longo dos anos de estudo.

- Analisar os resultados de cada algoritmo de Correlação de Classificação nos anos de estudo.
- Analisar as características mais importantes dos alunos concluintes dos cursos de engenharias, para todos os anos.

1.2 Organização do trabalho

Este trabalho está organizado da seguinte maneira. Este capítulo, introduz o assunto do trabalho juntamente com o problema a ser investigado e objetivos a serem alcançados. No Capítulo 2, apresentamos a fundamentação teórica dividido em seções. Na Seção 2.1 é apresentada a definição de descoberta de conhecimento, na Seção 2.2 e Seção 2.3 estão, respectivamente, a definição de Seleção de Atributos e a descrição dos algoritmos utilizados no estudo. Já na Seção 2.4 apresentamos a definição de Correlação de classificação e a descrição dos algoritmos utilizados. Em seguida, o Capítulo 3 apresenta os trabalhos que mostram diferentes abordagens de aplicação dos dados do ENADE. A metodologia utilizada é apresentada no Capítulo 4. No Capítulo 5, os resultados são apresentados. Por fim, discutimos as considerações finais deste trabalho no Capítulo 6.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo apresentamos a fundamentação teórica necessária para o melhor entendimento deste trabalho.

2.1 Descoberta de conhecimento em banco de dados

Com o crescente avanço da coleta e armazenamento de dados, por empresas e organizações, surgiu-se uma necessidade de extrair informações importantes e conhecimento útil. A partir deste problema, foi desenvolvida uma metodologia que combina estratégias tradicionais para análise de dados com algoritmos para extração de conhecimento, que podem ser aplicados em bases com milhares de instâncias. Este processo é chamado de Descoberta de Conhecimento em Banco de Dados (DCBD), em inglês *Knowledge Discovery in Database (KDD)* [18].

O termo surgiu em 1989, para enfatizar que o produto final do processo de descoberta em banco de dados era o “conhecimento” [18]. Embora DCBD e Mineração de Dados (em inglês - *Data Mining*) sejam frequentemente entendidos como sinônimos, é importante frisar que, segundo [18], a DCBD compreende todas as etapas para a descoberta do conhecimento a partir da existência de dados. A mineração de dados é apenas e tão somente uma das etapas do processo. O processo de DCBD consiste em cinco passos apresentados na Figura 1:

1. **Seleção:** esta etapa consiste em selecionar um conjunto ou subconjunto de dados que farão parte da análise. As fontes de dados podem ser variadas (planilhas, sistemas gerenciais, *data warehouses*) e possuir dados com diferentes formatos (estruturados, semiestruturados e não-estruturados).
2. **Pré-processamento:** esta etapa consiste em fazer a verificação da qualidade dos dados armazenados, pois um dos principais obstáculos para MD são dados de má qualidade. Quando os dados são precários, o produto de qualquer tarefa de MD também é precário. Essa etapa talvez seja a mais trabalhosa de todo o processo, levando a ocupar 70-80% do tempo de DCBD [72]. Alguns dos passos envolvidos nesta etapa são:

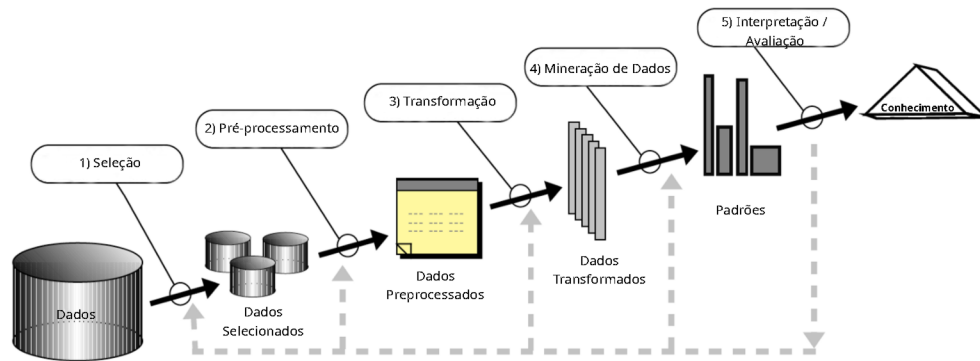


Figura 1: Etapas do processo de DCBD

Fonte: Adaptado de [23].

- **Integração dos dados:** é feita a integração dos diferentes bancos de dados, se existir mais de um, além de remover inconsistências nos nomes ou em valores de atributos de diferentes origens.
- **Limpeza dos dados:** consiste em detectar e corrigir erros nos dados, como a existência de dados ausentes, ruidosos, duplicados, redundantes, incompletos, inconsistentes e anomalias (*outliers*).
- **Seleção de dados:** consiste em escolher atributos e instâncias, que são considerados importantes para a tarefa de mineração. Detalharemos alguns métodos de seleção de atributos, com exemplos, na Seção 2.2.

3. **Transformação:** esta etapa consiste em aplicar técnicas de transformação como:

- normalização, que é um conjunto de regras que visa, principalmente, a organização de um projeto de banco de dados para reduzir a redundância, aumentar a integridade e o desempenho.
- agregação, que é essencialmente dados sumarizados, ou seja, dados agrupados.
- criação ou derivação de novos atributos.
- redução e sintetização dos dados.
- conversão de dados nominais para numéricos.
- discretização, que é transformar atributos contínuos em categóricos, seja porque o algoritmo de MD não trabalha com atributos contínuos ou para melhorar a compreensão do conhecimento descoberto.

Aqui os dados ficam disponíveis e agrupados em um mesmo local para a aplicação dos modelos.

4. **Mineração de Dados:** esta etapa consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo verificar uma hipótese, descobrir novos padrões de forma autônoma e extrair informação útil. Além disso, a descoberta pode ser dividida em: preditiva e descritiva. Na preditiva, o objetivo principal dessa mineração é dizer algo sobre os resultados futuros, não sobre o comportamento atual. Ele usa as funções de aprendizagem supervisionada que são usadas para prever o valor alvo. Os métodos incluídos neste tipo de categoria de mineração são chamados de classificação, análise de série temporal e regressão. Já na descritiva é usado basicamente para produzir correlação, tabulação cruzada, frequência, etc. Essa abordagem é usada para determinar as semelhanças nos dados e para encontrar os padrões existentes. Essa análise enfatiza o resumo e a transformação dos dados em informações significativas para relatórios e monitoramento. Esses modelos (preditivo e descritivo) geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.
5. **Interpretação e Avaliação:** esta é a última etapa, e consiste em avaliar o desempenho do modelo, aplicando testes em cima de dados que não foram utilizados na fase de treinamento ou mineração. A validação pode ser feita de diversas formas, algumas delas são: utilizar medidas estatísticas, ou passar pela avaliação de especialistas. Os resultados podem ser apresentados em tabelas, gráficos e relatórios, para a melhor compreensão e apresentação da informação obtida.

2.2 Seleção de atributos

Aprendizado de Máquina (AM), ou em inglês *Machine Learning*, é baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana [54]. Muitos fatores afetam o sucesso do AM em uma determinada tarefa. A representação e a qualidade dos dados de exemplo estão em primeiro lugar. Teoricamente, mais recursos deveria resultar em maior poder de discriminação. No entanto, experiências práticas com algoritmos de mineração de dados mostram que nem sempre é esse o caso, pois a quantidade de atributos não significa aumento de qualidade [31]. No conjunto de dados selecionados é comum haver atributos irrelevantes para o objetivo pretendido. Assim, é importante selecionar um subconjunto que melhor se adéque ao objetivo.

A Seleção Atributos (SA), em inglês *Feature Selection*, é o processo de identificar e remover tanto atributos irrelevantes, quanto informações redundantes [53]. Isso reduz a dimensionalidade dos dados, e melhora o desempenho dos classificadores [6, 15], já que atributos que não agregam valor à classificação ou que deterioram os resultados são eliminados. Também contribui para um melhor entendimento e análise dos resultados obtidos, além de permitir que os algoritmos de aprendizagem operem de forma mais rápida e efi-

caz. Assim, o principal objetivo da SA é identificar o conjunto de atributos que melhor representa, para um determinado contexto, a informação útil contida nos dados [73].

De acordo com a maneira utilizada para descobrir quais os atributos são os mais importantes para a mineração, as técnicas de SA podem ser agrupadas em três abordagens principais[57, 71, 73, 74]:

- **Método empacotamento (*Wrapper methods*)**: esta abordagem usa um modelo preditivo para pontuar subconjuntos de recursos. Cada novo subconjunto é usado para treinar um modelo, que é testado em um conjunto de validação. Contar o número de erros cometidos naquele conjunto de validação (a taxa de erro do modelo) dá a pontuação para aquele subconjunto. Como os métodos de empacotamento treinam um novo modelo para cada subconjunto, o custo computacional desse método é muito alto, pois o algoritmo de classificação deve ser executado várias vezes, para cada subconjunto candidato a ser avaliado. No entanto, o método geralmente fornece o conjunto de recursos de melhor desempenho para esse tipo específico de modelo ou problema típico.
- **Método embutido (*Embedded methods*)**: nesta abordagem, executa-se a seleção de recursos como parte da execução do algoritmo de aprendizado. Esses métodos tendem a ser mais eficientes computacionalmente que os métodos de empacotamento, porque eles simultaneamente integram a modelagem com a seleção de recursos. Essa abordagem é comumente utilizada pelos algoritmos de classificação que induzem árvores de decisão.
- **Método filtro (*Filter methods*)**: nessa abordagem, o processo de SA ocorre antes da construção de modelos propriamente dita. São utilizadas medidas para selecionar um grupo de atributos considerados relevantes, segundo algum critério, enquanto os outros são descartados. Os filtros geralmente são muito mais rápidos e funcionam independentemente do algoritmo de aprendizado. Isso significa que os recursos selecionados podem ser repassados para qualquer algoritmo de aprendizado. Os métodos de filtro podem ser categorizados pela medida de filtragem que empregam, ou seja, informações, distância, dependência, consistência, similaridade e medidas estatísticas. Este é a abordagem utilizada neste trabalho de dissertação de mestrado.

Independente da forma como as técnicas de atributos são categorizadas, seja como filtro, empacotamento ou embutida, seu funcionamento depende da utilização de dois algoritmos em conjunto [16, 58, 73]. O primeiro tipo estabelece uma forma de como os atributos de maior influência sobre o atributo alvo serão pesquisados e encontrados, chamados de métodos de pesquisa. O segundo tipo são medidas ou algoritmos que estabelecem uma relação entre os atributos que foram encontrados pelo algoritmo de método de

pesquisa, aplicando uma análise sobre eles. Esse tipo é chamado de algoritmo de seleção de atributos. Dentre os principais métodos de pesquisa pode-se citar:

- **Ranker**: Esse algoritmo analisa os atributos individualmente, e cria um ranking em ordem crescente de influência dos atributos sobre o atributo alvo.
- **Best-First**: Inicialmente contém um conjunto vazio de atributos, sendo preenchido a medida que a adição de um atributo melhora o processo preditivo. Esse método guarda um histórico contendo os estados do subconjunto de atributos, possibilitando voltar atrás em decisões tomadas.
- **Genetic-Search**: Esse algoritmo baseia-se nos princípios naturais da biologia, utilizando conceitos de mutação a cruzamento de atributos. Mutação consiste em adicionar e remover atributos de subconjuntos que são utilizados para análise. Cruzamento consiste em agrupar pares de atributos distintos, também para o processo de análise. O objetivo é encontrar um subconjunto de atributos que evoluem ao longo do tempo, gerando melhores previsões.
- **Greedy-Stepwise**: Analisa subconjuntos de atributos, buscando características que melhorem o processo de predição. Encerra o processo de análise apenas quando não se tem aumento no desempenho do modelo de predição. A busca ocorre percorrendo-se o conjunto de atributos, adicionando ou removendo atributos a outro subconjunto. A inserção de atributos no subconjunto pode ocorrer de duas formas: i) os atributos são adicionados a um subconjunto de forma que sua capacidade preditiva aumente, e dá-se o nome de pesquisa para frente (do inglês, *forward*); ii) o subconjunto contém inicialmente todos os atributos, sendo que esses são removidos com a finalidade de melhorar a capacidade preditiva, e nomeia-se de pesquisa para trás (do inglês, *backward*).

Medidas de relevância podem ser definidas como medidas de avaliação que serão utilizadas para o cálculo da relevância dos atributos. Elas são capazes de medir relações de diferentes naturezas, sejam elas lineares ou não-lineares [35], e aplicam-se a algoritmos de SA que utilizam a abordagem filtro. Essas medidas podem ser divididas em duas categorias: medidas que avaliam atributos individualmente e medidas de avaliação de subconjuntos de atributos [48, 59, 73]. As medidas e os algoritmos para avaliação de atributos são explicados nas Subseções 2.2.1 a 2.2.6, e estes mesmos foram utilizados no presente trabalho. Salienta-se que para que um método de pesquisa seja utilizado em conjunto com um algoritmo de seleção de atributos, é necessário haver compatibilidade entre ambos.

Para exemplificar o funcionamento dos algoritmos de SA, recomendamos a leitura do trabalho de Gondran (2021) [23], que disponibiliza uma base de dados fictícia, utilizada como exemplo em todos os métodos considerados neste trabalho.

2.2.1 Entropia

Entropia [67] é definida como uma forma de medir o grau médio de incerteza a respeito de fontes de informação. Consequentemente, permite a quantificação da informação presente que flui no sistema. Em termos simples, o conceito de entropia associa-se à ideia de que, quanto mais incerto é o resultado de um experimento aleatório, maior é a informação obtida ao observar a sua ocorrência.

Também pode ser definida como a quantidade de incerteza que há em uma mensagem, a qual diminui a medida que os símbolos são transmitidos. Ou seja, a medida que a mensagem vai sendo conhecida, tem-se então a informação, que pode ser vista como uma redução da incerteza. A medida de entropia é dada pela Equação (1).

$$E(A) = - \sum_{a \in A} P(a) \log_2 P(a) \quad (1)$$

Sendo A o atributo que vai ser calculado, a é o valor deste atributo e P é a frequência relativa dos valores.

2.2.2 Ganho de Informação

O ganho de informação [32] é definido como a quantidade de informação obtida sobre uma variável aleatória ou sinal da observação de outra variável aleatória. Ele mede a significância do atributo em relação à classe alvo, ou seja, mede a redução da incerteza (entropia) em função da divisão. Como desvantagem tende a preferir divisões que resultam em grande número de partições, cada uma delas sendo pequena, mas pura. O ganho de informação (GI) é calculado conforme a Equação (2).

$$GI(A) = E(C) - \sum_{a \in A} \frac{n_a}{n} E(a) \quad (2)$$

Onde E significa a medida de entropia, C é a classe, A é o atributo a ser avaliado, n_a é o número de instâncias da categoria pertencente ao atributo, n é o número total de instâncias e a é o valor do atributo.

2.2.3 Taxa de Ganho

A taxa de ganho [37] foi desenvolvida para solucionar o problema do ganho de informação. Ela é a razão do ganho de informação pela entropia do atributo, que nada mais é do que o ganho de informação relativo (ponderado) como critério de avaliação. Ajusta o GI pela entropia do particionamento, fazendo com que o particionamento de alta entropia (grande número de pequenas partições) seja penalizado. É dada pela Eq. (3).

$$TG(A) = \frac{GI(A)}{E(A)} \quad (3)$$

Onde GI é ganho de informação e E é a entropia.

2.2.4 Incerteza Simétrica

A incerteza simétrica [76] é uma medida de correlação não linear que foi desenvolvida com o mesmo propósito da taxa de ganho. Ou seja, é uma tentativa de normalizar o ganho de informação de um atributo em relação à classe. É calculada pela Equação (4).

$$IS(A) = 2 \cdot \frac{GI(A)}{E(A) + E(C)} \quad (4)$$

Sendo GI o ganho de informação e E , a entropia.

2.2.5 Correlação de Pearson

A Correlação de Pearson [31], também conhecida como coeficiente linear, mede o grau da correlação (e a direção dessa correlação - se positiva ou negativa) entre duas variáveis de escala métrica. De forma mais simples, é um grau de relação entre duas variáveis quantitativas e exprime o grau de correlação através de valores situados entre -1 e 1.

Quando o coeficiente de correlação aproxima-se de 1, tem-se um aumento no valor de uma variável quando a outra também aumenta, ou seja, há uma relação linear positiva. Quando ele se aproxima de -1, também é possível dizer que as variáveis são correlacionadas. Mas, nesse caso, quando o valor de uma variável aumenta, o da outra diminui, chamado de correlação negativa ou inversa. Um coeficiente de correlação próximo de zero indica que não há relação entre as duas variáveis. A correlação é dada pela Eq. (5) ou pela Eq. (6).

$$P(A) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5)$$

Onde A é o atributo a ser calculado, x_i e y_i são os valores medidos da variável e da classe, respectivamente. Já $\bar{x}$ e $\bar{y}$ são, respectivamente, as médias aritméticas da variável e da classe.

$$P(A) = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}, \quad (6)$$

sendo Cov a covariância entre os dois atributos X, Y e Var é a variância de cada atributo.

Para calcular a correlação de variáveis qualitativas, realiza-se a binarização delas, em que cada valor de cada atributo é substituído por 0 e, os demais, por 1. Os mesmos são

considerados valor por valor, tratando cada valor como um indicador. Uma correlação geral para um atributo nominal é obtida por meio de uma média ponderada. Salienta-se que para atributos binários, o cálculo valor por valor não se faz necessário. Correlações negativas devem ser multiplicadas por -1, na correlação de atributos qualitativos [31].

2.2.6 Alívio F

A ideia principal do método Alívio F [43, 44] é um algoritmo que ordena os atributos baseados em sua capacidade de diferenciar instâncias de classes distintas. Para tanto, dada uma instância, o algoritmo original chamado de Alívio procura seus dois vizinhos mais próximos: um da mesma classe (chamado de acerto mais próximo ou *hit*) e o outro de uma classe diferente (chamado de erro mais próximo ou *miss*). O algoritmo original de Alívio seleciona aleatoriamente n instâncias de treinamento, onde n é o parâmetro definido pelo usuário. Embora o algoritmo de Alívio original possa lidar com atributos discretos e contínuos, ele não pode lidar com dados incompletos e está limitado a problemas de apenas duas classes.

Ao longo dos anos, foi desenvolvida uma extensão do Alívio, chamada Alívio F, que melhora o algoritmo original ao estimar probabilidades de forma mais confiável e o estende para lidar com conjuntos de dados incompletos e de várias classes, mantendo a complexidade [74]. Em primeiro lugar, o Alívio F depende de um parâmetro de usuário k (número de vizinhos) que especifica o uso de k acertos mais próximos da mesma classe e k erros mais próximos de outra classe, para a pontuação de atualização para cada instância de destino (ao invés de um único acerto e erro). Ao encontrar um valor ausente, o algoritmo define a função *diff* igual à classe condicional, calcula a probabilidade de que duas instâncias tenham valores diferentes para o determinado recurso. Durante a pontuação em multiclasse, o Alívio F encontra os erros mais próximos de cada diferente classe e calcula a média da atualização de peso com base na probabilidade anterior de cada classe. O Alívio F é calculado pela Equação (7):

$$W(A) = W(A) - \frac{diff(A, R_i, H)}{n} + \frac{diff(A, R_i, M)}{n} \quad (7)$$

onde A é o atributo, $W(A)$ é um vetor com a pontuação de cada atributo, R_i é a instância alvo, H é a instância mais próxima da mesma classe, M é a instância mais próxima da outra classe e n é o número de instâncias aleatórias selecionadas para fazer parte do cálculo.

Os pesos são estimativas da qualidade dos atributos. O fundamento lógico da fórmula para atualizar os pesos é que um bom atributo deve ter o mesmo valor para instâncias da mesma classe e deve diferenciar entre instâncias de classes diferentes.

A função *diff* na Equação (7) calcula a diferença no valor do recurso A entre duas

instâncias I_1 e I_2 , onde $I_1 = R_i$ e I_2 é H ou M , ao realizar atualizações de peso em W . Para recursos discretos (por exemplo, categóricos ou nominais), $diff$ é definida como:

$$diff(A, I_1, I_2) = \begin{cases} 0, & \text{se o valor de } (A, I_1) = (A, I_2), \\ 1, & \text{caso contrário.} \end{cases}$$

Para atributos discretos, a diferença é 1 (os valores são diferentes) ou 0 (os valores são iguais). Já para recursos contínuos (por exemplo, ordinal ou numérico), $diff$ é definida como:

$$diff(A, I_1, I_2) = \frac{|\text{valor}(A, I_1) - \text{valor}(A, I_2)|}{\max(A) - \min(A)}$$

Os valores máximo e mínimo de A são determinados em todo o conjunto de instâncias. Essa normalização garante que as atualizações de peso fiquem entre 0 e 1 para recursos discretos e contínuos. A função $diff$ também é usada para calcular a distância entre as instâncias ao encontrar os vizinhos mais próximos. A distância total é simplesmente a soma das distâncias de $diff$ sobre todos os atributos (ou seja, distância de Manhattan) [45, 62].

2.3 Discretização

Na matemática, a discretização [36] é o processo de transferência de funções contínuas, modelos, variáveis e equações em contrapartes discretas. Na DCBD, a discretização pode ser vista, simplesmente, como a transformação de atributos numéricos em atributos categóricos. Isto é feito seja porque o algoritmo de MD não trabalha com atributos contínuos, ou para melhorar a compreensão do conhecimento descoberto.

Existem variadas técnicas de discretização, subdividas em discretização supervisionada e não supervisionada. As técnicas supervisionadas examinam a classe da instância quando a discretização é realizada. Quanto às técnicas não supervisionadas, a classe é ignorada [16].

Neste trabalho é utilizada a técnica não supervisionada, chamada de *Equal Width Discretization* (EWD) [61]. Neste método, os valores são divididos em N intervalos de mesmo tamanho. O primeiro passo é achar o máximo e o mínimo do atributo a ser discretizado, e então a largura de cada intervalo será dada pela Equação (8).

$$LI = \frac{Max - Min}{N}, \quad (8)$$

onde LI é largura de cada intervalo, Max e Min são, respectivamente, o valor máximo e mínimo do atributo, e N é o número de categorias a serem discretizadas.

No próximo passo achamos os valores limites da categoria, a partir do pseudo-algoritmo 1 dado por:

Algoritmo 1: Para calcular limites da discretização (EWD)

```

para ( $i \leftarrow 1$  até  $i < N$ ) faça
  |  $L[i-1] = \text{Min} + LI \cdot i$ ;
  |  $i \leftarrow i + 1$ 
fim

```

Onde L é um vetor que armazena os valores limites da categoria e i é o contador.

2.4 Correlação de Classificação

Correlação refere-se à associação entre os valores observados de duas variáveis, ou seja, essa medida descreve por meio de um único número a associação (ou dependência) entre duas variáveis [55]. Existem situações em que o relacionamento entre as duas variáveis não é linear, ou uma delas não é contínua, ou o número de pares das medidas é muito pequeno [14, 50]. Em outras palavras, os dados não tem uma distribuição Gaussiana. Então, para cada uma dessas situações há a necessidade de uma medida adequada de associação entre as variáveis, sendo os métodos de correlação de classificação os mais apropriados a serem utilizados [51]. A importância de conhecer os diferentes métodos e suas possíveis aplicações é exatamente pelo cuidado em não utilizar uma técnica inadequada [30].

A correlação de classificação [46] (ou, em inglês, *Rank Correlation*) é qualquer uma das várias estatísticas que medem uma associação ordinal. Ou seja, a relação entre classificações de diferentes variáveis ordinais ou diferentes classificações da mesma variável, em que uma “classificação” é a atribuição dos rótulos de ordenação “primeiro”, “segundo”, “terceiro”, etc. a diferentes observações de uma determinada variável. Dados ordinais são aqueles que possuem valores de rótulo e possuem um relacionamento de ordem ou classificação. Um coeficiente de correlação de classificação mede o grau de similaridade entre duas classificações e pode ser usado para avaliar a significância da relação entre elas [40, 21]. Como exemplo, em uma tabela com a variável renda (baixa, média e alta) e nível educacional (ensino fundamental, ensino médio, ensino superior) sendo a outra variável, uma correlação de classificação mede a relação entre as variáveis renda e nível educacional.

Um coeficiente de correlação de classificação situa-se no intervalo $[-1, 1]$ e assume os seguintes valores [21]:

- 1 se a concordância entre as duas classificações for perfeita (as duas classificações são iguais).
- 0 se as classificações forem completamente independentes.
- -1 se a discordância entre as duas classificações for perfeita (uma classificação é o inverso da outra).

A correlação de classificação pode ser calculada para variáveis de valor real. Primeiramente, convertem-se os valores de cada variável em dados de classificação. Nesta fase os valores são ordenados e atribuídos a um valor de classificação inteiro. Assim, os coeficientes de correlação de classificação podem ser calculados para quantificar a associação entre as duas variáveis classificadas.

Como nenhuma distribuição para os valores é assumida, os métodos de correlação de classificação são referidos como correlação livre de distribuição ou correlação não paramétrica.

2.4.1 Correlação de Spearman

A correlação de Spearman é uma medida não paramétrica de correlação de classificação (dependência estatística entre as classificações de duas variáveis) [42, 68]. A Correlação de Spearman (CS) é nomeada em homenagem a Charles Spearman. Também pode ser chamada de coeficiente de correlação de Spearman, denotado pela letra grega minúscula rho (ρ).

É uma medida de associação que exige que ambas as variáveis se apresentem em escala de mensuração pelo menos ordinal [30]. É Utilizada quando não existe normalidade e/ou não existe relação linear. O coeficiente de Spearman é apropriado para variáveis ordinais contínuas e discretas e quando as amostras são pequenas, ou seja, é conveniente para número de pares menor que 30 e quando os dados já estão ordenados [50].

Esse método estatístico quantifica o grau em que as variáveis classificadas são associadas por uma função monotônica, significando uma relação crescente ou decrescente. Mede a força e direção da associação entre duas variáveis classificadas. Este coeficiente baseia-se nos postos das observações dentro de cada variável e sobre as diferenças entre os postos observados, nas variáveis X e Y [10].

A intuição para a correlação de classificação de Spearman é que ela calcula uma correlação de Pearson (por exemplo, uma medida paramétrica de correlação) usando os valores de classificação em vez dos valores reais. A correlação de Pearson, discutida na Subseção 2.2.5, é o cálculo da covariância (ou diferença esperada de observações da média) entre as duas variáveis normalizadas pela variância ou dispersão de ambas as variáveis. O coeficiente de correlação de Spearman é definido na Eq. (9):

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}. \quad (9)$$

onde ρ é o coeficiente de Spearman a ser calculado, d_i é $x_i - y_i$, ou seja, a diferença de postos das variáveis X e Y e n é o número de observações.

O coeficiente de Spearman, pode ter valor entre $+1$ e -1 , sendo: i) $+1$ uma associação de classificação perfeita; ii) 0 não há uma associação de classificação; e iii) -1 uma

associação de classificação negativa perfeita. Por exemplo, uma correlação de Spearman igual a -1 significa que o maior valor para a variável X está associado ao menor valor para a variável Y , e assim por diante.

2.4.2 Correlação de Kendall

A Correlação de Kendall (CK) é uma medida não-paramétrica que avalia a associação entre duas variáveis, fornecendo o grau de similaridade entre dois conjuntos de classificações dadas a um mesmo conjunto de objetos [38, 41]. A correlação de classificação de Kendall é nomeada em homenagem a Maurice Kendall. Também é chamado de coeficiente de correlação de Kendall, é frequentemente referido pela letra grega minúscula tau (τ).

É uma medida de correlação utilizada para dados ordinais, como no caso do Coeficiente de Correlação de Spearman. Ambas as variáveis devem ser medidas no mínimo em nível ordinal, de forma que seja possível atribuir postos a cada uma das variáveis [50]. É um método adequado quando amostras têm tamanhos reduzidos, pois o método é mais preciso. Pode ser estendido a correlações parciais, quando o efeito de uma terceira variável, que age sobre X e Y , é retirado antes de determinar se tais variáveis estão relacionadas.

A intuição para o teste é que ele calcula uma pontuação normalizada para o número de classificações correspondentes ou concordantes entre as duas amostras. A CK entre duas variáveis será alta quando as observações tiverem uma classificação semelhante (ou idêntica para uma correlação de valor igual a $+1$), ou seja, o rótulo da posição relativa das observações dentro da variável: 1^a , 2^a , 3^a posições, etc., entre as duas variáveis. Será baixa quando as observações tiverem uma classificação diferente (ou totalmente diferente para uma correlação igual a -1) entre as duas variáveis.

Os valores de τ variam entre -1 e $+1$, no qual valores próximos aos extremos indicam forte correlação. O sinal indica a direção da relação entre as variáveis e os valores próximos a 0 indicam correlações fracas ou inexistentes.

Existe três variantes do tau de Kendall [39]: Tau-a (τ_a) para casos sem empates, Tau-b (τ_b) nos casos com empates, e Tau-c (τ_c) entre uma variável ordinal e outra nominal. Podem ser definidos pelas respectivas Equações (10), (11), e (12).

$$\tau_a = \frac{P - Q}{N}, \quad (10)$$

onde τ_a é o coeficiente de Kendall a ser calculado, P é o número de pares concordantes, Q o número de pares discordantes, e N é o número total de comparações entre os pares.

$$\tau_b = \frac{P - Q}{\sqrt{N - T_x} * \sqrt{N - T_e}}, \quad (11)$$

onde τ_b é o coeficiente de Kendall, T_x é o número de empates apenas em X e T_e é o

número de empates apenas em Y . Se ocorrer um empate para o mesmo par em X e Y , ele não será adicionado a T_x ou T_e .

$$\tau_c = \frac{2 * (P - Q)}{\left(\frac{n^2 * (m-1)}{m}\right)}, \quad (12)$$

onde τ_c é o coeficiente de Kendall, n é o número total de postos e m é o número de valores exclusivos em X ou Y , o que for menor.

Um par pode ser definido como concordante ou discordante. Considere (x_1, y_1) , (x_2, y_2) , $\dots$, (x_n, y_n) um conjunto de observações das variáveis conjuntas de X e Y , respectivamente, tal que todos os valores de x_i e y_i sejam únicos.

- Qualquer par de observações (x_i, y_i) e (x_j, y_j) , em que $i \neq j$, é **concordante** se ambos os elementos concordam um com o outro, isto é, se $(x_i > x_j \text{ e } y_i > y_j)$ ou $(x_i < x_j \text{ e } y_i < y_j)$.
- Qualquer par de observações (x_i, y_i) e (x_j, y_j) , em que $i \neq j$, é **discordante** se os elementos discordam entre si, isto é, se $(x_i > x_j \text{ e } y_i < y_j)$ ou $(x_i < x_j \text{ e } y_i > y_j)$.
- Se $x_i = x_j$ ou $y_i = y_j$, o par não é nem concordante, nem discordante.

Segundo Siegel (2006) [68], em uma comparação entre Correlação de Spearman e Correlação de Kendall, os valores numéricos não são iguais quando calculados para os mesmos pares de postos, e não são comparáveis numericamente. Contudo, pelo fato de utilizarem a mesma quantidade de informação contida nos dados, ambos têm o mesmo poder de detectar a existência de associação.

2.4.3 Correlação de Goodman e Kruskal

A Correlação de Goodman e Kruskal (CGK) é uma medida de correlação de classificação, ou seja, a semelhança das ordenações dos dados quando classificados por cada uma das quantidades [24]. É uma medida não paramétrica da força e direção da associação que existem entre duas variáveis medidas em uma escala ordinal [25]. Os valores variam de -1 (associação negativa ou inversão perfeita) a $+1$ (associação positiva ou concordância perfeita). Um valor igual a 0 indica ausência de associação. Recebeu o nome de Leo Goodman e William Kruskal, que a propuseram em uma série de artigos. Também é chamada de Gama de Goodman e Kruskal.

O coeficiente gama trata as variáveis simetricamente, ou seja, não precisa levantar hipóteses sobre quais podem ser variáveis dependentes e quais podem ser independentes [26, 27]. Pode ser definido pela Equação (13).

$$\gamma = \frac{P - Q}{P + Q}, \quad (13)$$

onde γ é o coeficiente de Goodman e Kruskal a ser calculado, P é o número de pares concordantes e Q o número de pares discordantes.

Os pares são concordantes ou discordantes pela mesma definição de Kendall, vista na Subseção 2.4.2. Quando não há empates (ou seja, $T_x = T_e = 0$), Kendall τ_a e τ_b são iguais aos de Goodman e Kruskal γ , ou seja, $\tau_a = \tau_b = \gamma$.

O gama pode ser calculado para variáveis ordinais que são variáveis contínuas ou variáveis discretas. Embora existam outros coeficientes que podem calcular relações para esses tipos de variáveis, como o D de Somer (discutido na Subseção 2.4.4) ou o Tau de Kendall, o gama de Goodman e Kruskal é geralmente preferido quando os dados têm muitas classificações empatadas. Também é particularmente útil quando seus dados apresentam valores discrepantes, pois eles não afetam muito os resultados.

2.4.4 Correlação D de Somers

É uma medida de associação ordinal entre duas variáveis aleatórias X e Y possivelmente dependentes. A Correlação D de Somers (CDS) também assume valores entre: -1 quando todos os pares de variáveis discordam, 1 quando todos os pares de variáveis concordam e 0 quando não tem relação entre os pares. Recebeu este nome devido a Robert H. Somers, que propôs a medida em 1962, e também é conhecida pela letra grega δ (delta) [70]. A correlação D de Somers desempenha um papel central nas estatísticas de classificação e é o parâmetro por trás de muitos métodos não paramétricos. Adicionalmente é usada como medida de qualidade de escolha binária ou regressão ordinal (por exemplo, regressões logísticas) e modelos de pontuação de crédito [56].

A correlação D de Somers é uma extensão assimétrica de Gama que difere apenas na inclusão do número de pares não relacionados à variável independente [8]. Teoricamente, testes de hipóteses baseados no teste de Kendall e D de Somers devem ser idênticos. Podem ser definidas pelas Equações (14) e (15), respectivamente.

$$\delta = \frac{P - Q}{P + Q + T} \quad (14)$$

onde δ é a correlação D de Somers, P é o número de pares concordantes, Q é o número de pares discordantes e T é o número de pares empatados. Os pares são concordantes ou discordantes pela mesma definição de Kendall da Subseção 2.4.2.

O cálculo da correlação D de Somers em função do tau de Kendall é dado por:

$$\delta = \frac{\tau(X, Y)}{\tau(X, X)}. \quad (15)$$

Observe que essa definição é assimétrica, pois em geral, $\delta(Y|X) \neq \delta(X|Y)$.

Na correlação D de Somers, a previsão da variável dependente melhora com base no conhecimento de um valor da variável independente. Portanto, é importante definir qual

variável é a independente e qual é a dependente ao executar este teste, pois obtém-se dois resultados diferentes para (X, Y) e (Y, X) .

3 TRABALHOS RELACIONADOS

As possibilidades de aplicação de técnicas de MDE no Brasil são apresentadas no trabalho de Baker et al. em 2011 [2]. O estudo demonstra como essa área de pesquisa pode contribuir com a melhor compreensão dos processos de ensino e aprendizagem e para a motivação dos alunos. O trabalho de Gottardo et al. em 2012 [28] propõe a definição de um conjunto de atributos amplos e generalizáveis a serem utilizados para realizar inferências relativas ao desempenho de estudantes. Experimentos realizados demonstraram índices de 76% de acerto na previsão de desempenho. Em outro estudo [63], evidenciam-se as possibilidades de utilização de mineração de dados para a extração de informações relevantes sobre estudantes no contexto educacional utilizando um AVA. No restante deste Capítulo, apresenta-se uma série de trabalhos especificamente sobre o ENADE.

O trabalho realizado por Faria em 2017 [17] tem como objetivo principal a identificação dos fatores determinantes do desempenho dos discentes dos cursos de Administração. A autora faz uso dos microdados do ENADE 2012 do Distrito Federal. A pesquisa tem características de análise predominantemente quantitativa, baseada em uso de técnica estatística descritiva e multivariada. O método de regressão múltipla foi utilizado para verificar se as características próprias dos estudantes, como aspectos pessoais e socioeconômicos, da instituição, e da organização didático-pedagógica, seriam variáveis significativas na explicação do resultado da prova. Os resultados apontaram que os fatores relacionados ao discente que apresentaram relevância na predição de desempenho foram, dentre outros: renda familiar, gênero, escolaridade da mãe e escolaridade do pai.

Em busca de algum padrão ou regra que determine ou influencie diretamente na nota dos alunos, o trabalho [11] aplica técnicas de visualização, análise e mineração de dados, no Enade de 2014, com um foco diferenciado para o curso de Bacharelado em Ciência da Computação e para os alunos da Universidade de Brasília. Foi usado o algoritmo de classificação de dados JRip [60, 65]. Foram feitas diversas análises sobre os dados como: correlações entre colunas, classificação da nota com relação aos atributos e análise das respostas dos estudantes nos itens do exame (incluindo tabelas e gráficos). O objetivo dessas análises foi entender melhor o perfil dos estudantes do ensino superior no Brasil.

O algoritmo JRip encontrou diversas regras como, por exemplo, a seguinte regra: alunos que não estudam no turno vespertino e que gastaram menos de uma hora para concluir a prova obtiveram uma nota menor ou igual a mediana. Outros atributos como renda, idade, e etnia, também foram encontrados em regras semelhantes (relacionando-os com as notas).

A utilização de técnicas de descoberta de conhecimento para desenvolver uma ferramenta de exploração dos dados do exame ENADE foi proposta por [1]. Além de avaliar a estrutura e distribuição dos dados das provas, também propuseram a criação de um modelo classificatório capaz de prever o desempenho do estudante. Para a elaboração de regras, foram utilizados diversos algoritmos de classificação de árvore de decisão, no qual se destacou o CART [7]. Como resultado final, elaborou-se uma ferramenta de análise de dados, que permite comparar cursos e instituições de ensino superior e proporciona uma melhor visualização dessas informações com a finalidade de auxiliar na tomada de decisão. Por fim, foi distribuído um formulário online para que professores, alunos e coordenadores pudessem avaliar e validar o sistema desenvolvido.

A metodologia CRISP-DM (acrônimo em inglês para *Cross-Industry Standard Process for Data Mining*) [9] é aplicada no trabalho [4] para construir modelos que identifiquem os fatores que podem explicar o sucesso do aluno no exame, bem como antecipar suas chances de se classificar no grupo dos mais bem colocados. Na preparação dos dados para modelagem, foi selecionada uma amostra balanceada, no qual se dividiu em duas categorias, sendo elas: estudantes com nota superior a 60 e estudantes com nota inferior a 60. Foi utilizada a técnica de Regressão Logística Binária, tomando-se variáveis independentes extraídas dos microdados, referentes ao exame de 2017. O perfil encontrado para o estudante classificado como de melhor performance, por ordem de importância, foi: discente oriundo de IES Pública Federal, com renda familiar superior a 6 salários mínimos, cuja IES está localizada no Sul ou Sudeste, com idade até 24 anos, morando sozinho, estudando em turno integral, com professores com domínio do conteúdo, possuindo familiar com curso superior, matriculado na modalidade de ensino presencial e com oportunidades em projetos de iniciação científica.

Na tentativa de relacionar os atributos dos alunos com a nota obtida no ENADE, e entender melhor os alunos, cursos e universidades, o trabalho [52] aplica técnicas de mineração de dados e os algoritmos de clusterização K-means [72]. O foco foi o questionário sócio-econômico de cada aluno e a tentativa de categorizar os grupos. Identificou-se que a maioria dos alunos de Ciência da Computação são homens (87%) e perceberam-se indícios de que não há uma relação entre renda da família do estudante e seu desempenho no exame. Novamente os resultados dos autores direcionam a uma conclusão diferente do senso comum sobre o rendimento dos alunos conforme a região que se estuda.

Tendo em consideração que tanto cursos na modalidade presencial quanto os oferecidos à distância (EaD) são considerados na avaliação do ENADE, o estudo inicial realizado

por Gondran et al. em 2022 [22] aplica diferentes técnicas de seleção de características na base de dados do ENADE 2018, com o objetivo de encontrar quais características pessoais e socioeconômicas são determinantes para o desempenho do discente, matriculado na modalidade EaD ou presencial. Pode-se concluir, dentre os resultados, que a escolarização do pai influenciou diretamente no desempenho, assim como quanto maior a renda, melhor constatou-se o desempenho. Também constatou-se que estudantes de cor branca apresentaram melhor desempenho frente a estudantes das cores preta e parda. Por fim, notou-se diferenças nos perfis das duas modalidades, sendo alguns atributos importantes somente para uma das modalidades.

No trabalho [64] foi feito um Estudo Longitudinal Transversal Repetido, buscando analisar os dados ao longo do tempo para verificar seu comportamento nos anos estudados, em particular, os dados do curso de ciência da computação. Foram aplicadas técnicas de mineração de dados nos dados do ENADE, analisando os anos de 2008, 2011, 2014 e 2017. A contribuição deste trabalho foi mostrar que dentre os algoritmos utilizados nos experimentos, a árvore de decisão foi a abordagem que melhor classificou os dados em relação ao desempenho dos alunos, possibilitando identificar que algumas características socioeconômicas como: renda familiar, escolaridade do pai, situação de trabalho do discente em conjunto com a categoria administrativa e o turno de graduação, impactavam no desempenho acadêmico.

Utilizando a técnica de agrupamento de dados conhecida como K-means, nos microdados do ENADE com foco no concluintes dos cursos da área computação das edições de 2008, 2011, 2014 e 2017, os autores em [75] propõe identificar padrões que possam caracterizar o desempenho dos estudantes no exame. Como resultado, observou-se que há uma significativa diferença entre a média do desempenho dos participantes dos agrupamentos, o que permitiu classificar os grupos como alto, médio e baixo desempenho. Questões relevantes também foram observadas, como: o tipo do ensino médio, pois estudantes oriundos de escola particular apresentaram melhor desempenho, e quanto maior a escolarização do pai e da mãe, maior foi o desempenho do concluinte. A categoria administrativa da IES, também influenciou o desempenho, com destaque para a instituição pública. O curso de engenharia de computação apresentou maior proporção de concluintes com alto desempenho, frente aos outros cursos da área. Por fim, os resultados mostraram uma evolução dos grupos ao longo das edições.

Um estudo de caso sobre o perfil socioeconômico dos concluintes de graduação do município de Araçatuba/SP, baseado nos microdados de 2016 a 2018, foi feito no trabalho [20]. A escolha do estudo de caso dá-se, principalmente, pelo interesse de um dos autores em entender as características das IES do município onde atua como docente em uma das instituições. Além de estatística descritiva, técnicas de mineração de dados foram utilizadas para identificar regras de associação, sendo utilizado o algoritmo FP-Growth [33]. Como resultado, foram extraídas algumas regras de associação como: alunos brancos de

IES privadas que escolheram o curso visando inserção no mercado de trabalho, apresentaram baixas notas no ENADE; enquanto alunos autodeclarados pretos, pardos ou indígenas que escolheram o curso pelo mesmo motivo apresentaram notas melhores. Ainda revelou-se que aproximadamente um quarto dos estudantes brancos de IES privadas de Araçatuba possuem pais que concluíram até o ensino médio. Por fim, os autores afirmam, com um alto grau de convicção, que estudantes pretos, pardos ou indígenas de IES públicas, que moram com cônjuge e/ou filhos, que possuem renda, mas recebem ajuda para financiar os gastos, tendem a apresentar baixo desempenho no ENADE.

Em [13] é apresentado um estudo documental que tem por objetivo analisar o perfil das licenciandas em Matemática, levando em consideração os dados do ENADE de Matemática, no período de 2005 a 2017. A partir da análise de conteúdo, foi criado um inventário composto por cinco elementos textuais, que foram tratados e analisados mediante duas categorias. Sendo assim, buscou-se nas edições que apresentavam os dados de Matemática (2005, 2008, 2011, 2014 e 2017) evidenciar, qual o perfil de mulher que estava inserindo-se no referido curso, tendo em conta a quantidade, idade, cor/raça/etnia, escolaridade dos pais (mãe e pai) e tipo de curso concluído no Ensino Médio. Os dados revelaram que há uma predominância de mulheres brancas nos cursos, com idades médias de 18 a 24 anos, com poucos recursos financeiros e, por fim, com pais escolarizados até os anos iniciais do Ensino Fundamental.

O trabalho realizado por [5] constitui uma pesquisa de caráter documental, realizado a partir dos microdados do ENADE entre 2009 e 2017, onde busca discutir em que medida a educação superior brasileira vem efetivamente reduzindo os níveis de desigualdade educacional em termos de conclusão por jovens estudantes oriundos de contextos socioeconômicos e sociais desfavorecidos. Esta pesquisa ateu-se aos dados do questionário socioeconômico, no qual, foram selecionadas sete questões. A partir disso foram identificadas, para cada uma das questões, as respostas oferecidas pelo questionário que indicaram proveniência de contextos socioeconômicos e sociais desfavorecidos. Como resultado, os autores argumentam que, com o desenvolvimento de programas como Prouni, Fies e Cotas, observou-se avanços em termos de igualdade na educação superior brasileira, porém, tratava-se do início de um processo. Ainda percebe-se a persistência de heranças elitistas em cursos que proporcionam maior status social e melhores remunerações.

Ainda pode-se destacar o trabalho [49], que propõe fazer uma revisão sistemática, tendo como objetivo identificar o uso que vem sendo feito dos dados coletados pelo Inep, para além das análises feitas por ele próprio. A pesquisa foi guiada por uma questão de pesquisa principal: que tipos de análise têm sido realizadas com base nos dados do Enem e do Enade e quais são os seus objetivos?. Como resultado o tipo de análise mais utilizada é a estatística descritiva, mostrando a necessidade de se investir em pesquisas que abordem outros tipos de análises, como a mineração de dados. Por fim, os autores destacam que existe, portanto, um vasto campo de pesquisa ainda inexplorado e com

potencial para gerar informação útil, para compreender melhor o perfil dos alunos e o sistema educacional brasileiro.

Tabela 1: Tabela comparativa de trabalhos relacionados, parte 1/2.

Trabalho	Objetivo	Ano
[17]	Identificar os fatores determinantes do desempenho dos discentes, com foco nos cursos de Administração.	2012
[11]	Entender melhor o perfil dos estudantes do ensino superior no Brasil, com foco curso de Ciência da Computação e para os alunos da Universidade de Brasília.	2014
[1]	Criar um modelo classificatório capaz de prever o desempenho do estudante, para comparar cursos e instituições de ensino superior.	2017
[4]	Construir modelos que identifiquem os fatores que podem explicar o sucesso do aluno no exame, bem como antecipar suas chances de estarem no grupo dos mais bem colocados.	2017
[52]	Relacionar os atributos dos alunos com a nota obtida no ENADE, com foco no curso de Ciência da Computação.	2014
[22]	Encontrar quais características pessoais e socioeconômicas são determinantes para o desempenho do discente, matriculado na modalidade EaD ou presencial.	2018
[64]	Analisar os dados ao longo do tempo para verificar seu comportamento nos anos estudados, em particular, os dados do curso de ciência da computação.	2008 / 2011 / 2014 / 2017
[75]	Identificar padrões que possam caracterizar o desempenho dos estudantes nos cursos da área computação.	2008 / 2011 / 2014 / 2017
[20]	Analisar o perfil socioeconômico dos concluintes de graduação do município de Araçatuba/SP.	2016 / 2017 / 2018
[13]	Analisar o perfil das licenciaturas em Matemática, levando em consideração os dados do ENADE.	2005 / 2008 / 2011 / 2014 / 2017

Na Tabela 1 e Tabela 2 podemos ver uma comparação de alguns dos trabalhos apresentados, no qual temos na primeira coluna, de ambas as Tabelas, a identificação do trabalho. Depois o objetivo e o ano da base de dados do ENADE utilizado no trabalho na Tabela 1. Na Tabela 2, encontram-se o algoritmo utilizado e os atributos considerados importantes nestes mesmos trabalhos. Nestas Tabelas, observamos que a maioria dos estudos busca identificar os fatores determinantes para o desempenho dos alunos de uma instituição, curso ou área específica. Notamos que normalmente esse estudo é feito em um único ano de análise e que os algoritmos utilizados tendem a ser os de árvore de decisão, regressão ou agrupamento. Por fim, percebemos que os três atributos que mais apareceram nos trabalhos e considerados importantes foram: renda familiar, escolarização do pai e categoria

Tabela 2: Tabela comparativa de trabalhos relacionados, parte 2/2.

Trabalho	Algoritmo	Atributos
[17]	Análise fatorial exploratória e regressão linear múltipla	Renda familiar, gênero, escolaridade da mãe e do pai.
[11]	JRip	Turno, renda familiar, idade, e cor.
[1]	J48, CART e Random Forest	Categoria administrativa da IES, renda familiar, bolsa financiamento, tipo escola ensino médio.
[4]	Regressão Logística Binária	Categoria administrativa da IES, renda familiar, região da IES, idade, pessoas que moram com você, turno, professores com domínio do conteúdo, familiar com curso superior, modalidade de ensino e bolsa acadêmica.
[52]	K-means	Renda da familiar, gênero e região.
[22]	SA	Escolarização do pai, renda familiar e cor.
[64]	Árvore de decisão	Renda familiar, escolaridade do pai, situação de trabalho, categoria administrativa da IES e o turno.
[75]	K-means	Tipo escola ensino médio, escolarização do pai e da mãe e categoria administrativa da IES.
[20]	FP-Growth	Cor, escolarização dos pais. categoria administrativa da IES, pessoas que moram com você, renda familiar, situação financeira.
[13]	Estudo documental	Cor, idade, situação financeira e escolarização do pai.

administrativa da IES.

A partir do levantamento bibliográfico realizado, não foi possível encontrar trabalhos que utilizassem técnicas de seleção de atributos nos microdados de várias edições do ENADE, de forma a entender a evolução do perfil de estudantes de Engenharia. Assim, o presente trabalho tem o intuito de aplicar diferentes técnicas de seleção de atributos, para apontar quais aspectos pessoais e socioeconômicos são características significativas na explicação do resultado do ENADE, não de um único ano ou curso, mas focando nos cursos de engenharias, ao longo de todos os anos (2005, 2008, 2011, 2014 e 2017) de aplicação do exame e disponíveis até o momento de escrita deste trabalho.

4 METODOLOGIA

Neste Capítulo será apresentada a metodologia utilizada para o presente estudo. Ela foi baseada no processo de Descoberta de Conhecimento em Banco de Dados (DCBD), composta por cinco etapas: seleção, pré-processamento, transformação, mineração de dados e análise dos resultados, apresentadas em detalhes na Seção 2.1. Na Figura 2, podemos visualizar a metodologia do trabalho no formato de um fluxograma.

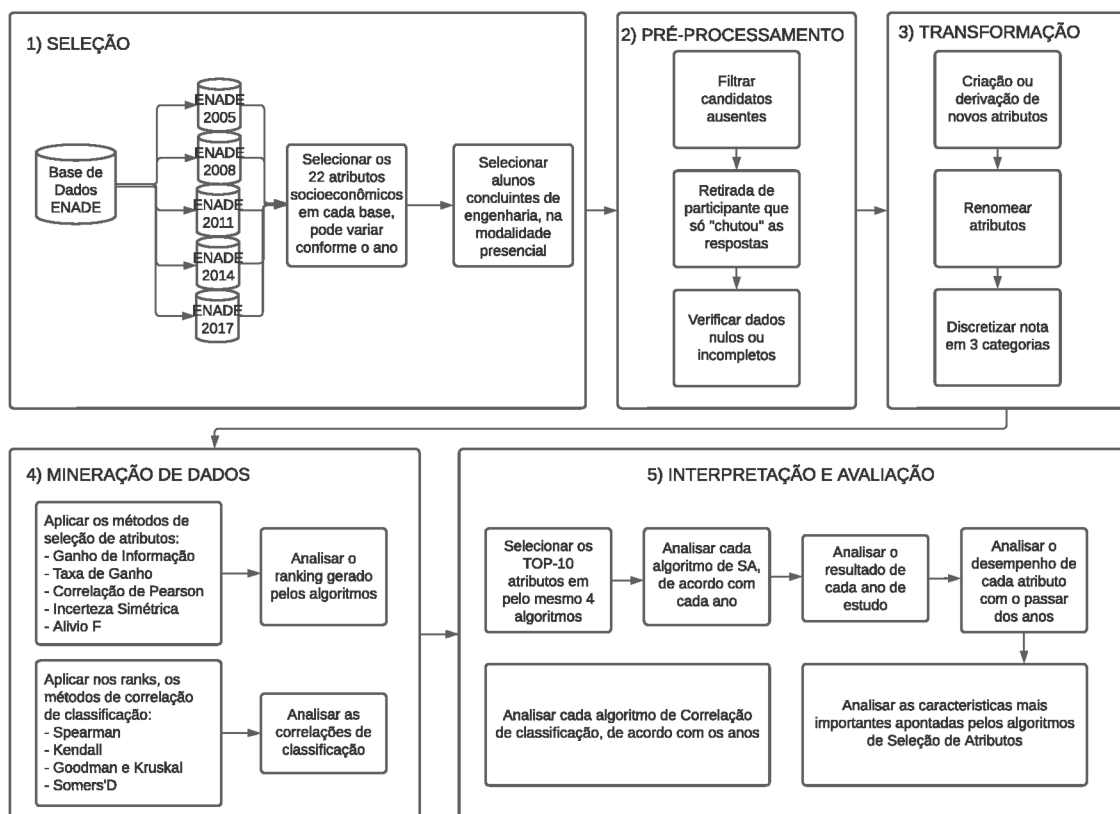


Figura 2: Metodologia do Trabalho.

4.1 Seleção

As bases de dados utilizadas são os microdados do ENADE, que fornecem informações do exame, desde o ano de 2004 até o ano de 2018. Estas bases são disponi-

bilizadas pelo Inep. O portal do Inep disponibiliza o Plano de Dados Abertos (PDA) para consulta de informações estatísticas que ajudam a construir e a interpretar o panorama da educação brasileira¹.

Dentre as bases de dados disponíveis, foram escolhidos os microdados do ENADE dos seguintes anos: 2005, 2008, 2011, 2014 e 2017. Estas bases foram escolhidas por serem as bases com os anos de avaliação dos alunos de engenharia. Os cursos de engenharia que fazem parte destas bases (totalizando 37) são Engenharia: Geológica, Cartográfica, Civil, de Agrimensura, de Recursos Hídricos, Sanitária, da Computação, de Comunicações, de Controle e Automação, de Redes de Comunicação, de Telecomunicações, Elétrica, Eletrônica, Eletrotécnica, Industrial, Mecatrônica, Aeroespacial, Aeronáutica, Automotiva, Mecânica, Naval, Bioquímica, de Alimentos, de Biotecnologia, Química, Têxtil, de Materiais, Física, Metalúrgica, de Fundição, de Produção, Ambiental, de Minas, de Petróleo, Agrícola, Florestal e de Pesca. Alguns cursos possuem mais de uma ênfase.

Os dados obtidos são compostos de três partes:

- (i) Leia-me (contendo um dicionário de variáveis, o manual do usuário e o questionário do estudante)
- (ii) Inputs (*scripts* para a leitura dos arquivos utilizando os softwares SAS² e R³)
- (iii) a base de dados contendo os dados brutos

Na Tabela 3, temos a composição de cada ano, com o número total de participantes (exemplos) e o número de atributos (variáveis), que podem ser do tipo numérico ou caractere. Os atributos são divididos em: informações da instituição e curso, informações do estudante, número de itens na parte objetiva, vetores (com gabaritos, escolhas e acertos dos participantes), tipos de presença, tipos de situações das questões da parte objetiva e discursiva, notas na formação geral e no componente específico, questionário de percepção da prova e questionário do estudante.

Tabela 3: Relação do número de participantes e atributos considerados neste estudo.

Ano	2005	2008	2011	2014	2017
Número de Atributos	210	198	115	154	150
Participantes (Total)	323.338	461.776	376.180	481.720	537.436
Participante Engenharia	20.763	28.145	47.808	91.635	149.300

Em cada ano (2005, 2008, 2011, 2014 e 2017), nas bases de dados originais, foram selecionados os alunos concluintes dos cursos de engenharia, na modalidade presencial. Em cada ano, temos, respectivamente, 20.763, 28.145, 47.808, 91.635 e 149.300 instâncias,

¹Disponível em <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados>

²Disponível em <https://oestatistico.com.br/sas-o-poder-do-conhecimento/>

³Disponível em <https://www.r-project.org/>

como visto na Tabela 3. Somente os concluintes presencias foram escolhidos porque, de acordo com [22], há diferenças nos perfis dos alunos na modalidade da Educação a Distância (EaD) e dos alunos presencias. Além disso, o quantitativo de alunos EaD representava uma pequena minoria. Após analisar todos os atributos de cada base de dados, decidimos focar nos aspectos pessoais e socioeconômicos, totalizando 22 atributos (incluindo a nota do exame) que atendiam este requisito. Neste ponto, tentamos manter os 22 atributos, mas, o número de atributos em cada base de dados pode variar conforme o ano, sendo eles apresentados na Tabela 4. Destacamos que é na base de dados filtrada com estas características que os algoritmos de seleção de atributos vistos na Subseção 2.2 foram aplicados em cada ano do exame.

Tabela 4: Relação dos atributos considerados neste estudo, conforme ENADE 2005, 2008, 2011, 2014 e 2017.

ID	Características
1	idade
2	sexo
3	estado civil
4	cor
5	escolarização pai
6	escolarização mãe
7	onde e com quem mora
8	pessoas que moram com você
9	renda familiar
10	situação financeira
11	situação de trabalho
12	bolsa/financiamento
13	bolsa permanência ^{†,‡}
14	bolsa acadêmica
15	ação afirmativa ou inclusão social [†]
16	tipo escola ensino médio
17	modalidade ensino médio
18	família com curso superior
19	estudo semanal
20	escolha curso ^{†,‡,*}
21	escolha instituição ^{†,‡,*}
22	nota geral

[†] ausente no ENADE 2005

[‡] ausente no ENADE 2008

^{*} ausente no ENADE 2011

Na Tabela 4, além dos atributos de semântica equivalente ao nome da característica, o atributo situação financeira diz respeito à origem e ao destino da renda do indivíduo, enquanto o 11 refere-se a estar trabalhando e quantas horas semanais trabalha. Os atributos 12, 13 e 14 caracterizam os tipos de bolsa ou financiamento. O atributo tipo escola ensino médio categoriza os estudantes em: pública ou privada, no Brasil ou no exterior. O atributo estudo semanal refere-se a quantidade de horas de estudo extraclasse. Os atributos 20 e 21 dizem respeito aos motivos da escolha pelo curso e instituição, respectivamente.

Família com curso superior indica se alguém na família concluiu uma graduação. Por fim, a nota geral é o atributo que refere-se ao desempenho no ENADE. Observamos que o número de atributos em cada base de dados variou ligeiramente porque alguns atributos não estavam presentes em todos os anos considerados.

4.2 Pré-processamento

Na primeira operação de pré-processamento, foi aplicado um filtro para selecionar apenas os participantes que realmente tinham realizado a prova. Os critérios para a remoção de instâncias incluem:

1. Participante Ausente,
2. Participante com prova em branco na parte objetiva na formação geral,
3. Participante com prova em branco na parte discursiva na formação geral,
4. Participante com prova em branco na parte objetiva no componente específico,
5. Participante com prova em branco na parte discursiva no componente específico,
6. Participante com resultado desconsiderado pela Aplicadora,
7. Participante com resultado desconsiderado pelo Inep,
8. Participante com resultado desconsiderado por inscrição indevida.

A segunda etapa consistiu em remover aqueles participantes que preencheram questões da parte discursiva com qualquer resposta, sendo considerado que este participante tentou adivinhar as respostas. Assim, estes foram retirados por considerá-los com perfis não condizentes ao seu desempenho. Para eliminá-los, consideramos aqueles que em todas as questões da parte discursiva estavam em um dos seguintes critérios:

1. Questão em branco,
2. Questão zerada por motivo de resposta nula,
3. Questão zerada por motivo de resposta divergente da temática.

A terceira etapa consistiu em verificar dados nulos ou incompletos. Esses casos incluem notas da prova em branco, a parte do questionário em branco ou aqueles que não quiseram responder o questionário. Por fim, destacamos na Tabela 5 o número de participantes engenheiros e concluintes, retirando-se todos os casos citados acima no pré-processamento, em cada ano de análise e em cada etapa de pré-processamento. Salienta-se que os algoritmos de SA são aplicados no número de engenheiros da 3ª etapa.

Tabela 5: Relação do número de participantes após cada etapa de pré-processamento.

Ano	2005	2008	2011	2014	2017
1ª Etapa	12.718	16.332	32.089	66.178	113.109
2ª Etapa	12.718	16.332	32.089	66.167	113.056
3ª Etapa	8.996	12.281	30.400	66.081	112.285

4.3 Transformação

A primeira operação realizada foi a criação ou derivação de novos atributos, por exemplo: se a família tinha concluído um curso superior e se o aluno percebia alguma bolsa acadêmica. Isto foi necessário pois uma questão poderia ser dividida em várias questões em outros anos. A próxima operação realizada foi a padronização na nomenclatura dos atributos em todos os períodos. Nesta etapa, os atributos possuíam nomes referentes a numeração do questionário do estudante (QE_I01, QE_I02 e QE_In), mas esses atributos foram renomeados conforme a Tabela 4. Os valores nominais dos atributos (A, B, C, D, E, etc.) também foram renomeados para nomes de compreensão mais fácil, por exemplo, o estado civil foi renomeado para: Solteiro, Casado, Divorciado, Viúvo, e Outro. A renomeação de valores nominais variou conforme o ano do exame.

Por fim, as notas obtidas pelos candidatos também foram categorizadas. Os valores numéricos foram discretizados em três categorias de frequências distintas (desempenho: baixo, médio e alto), para que fosse mantida a distribuição original. Além disso, para manter essa distribuição optou-se pela discretização individual de cada ano. Na ??, pode-se ver o resultado da discretização para cada ano:

Tabela 6: Relação da discretização das notas e anos.

Desempenho	Baixo	Médio	Alto
2005	nota $\leq 36,70$	$36,70 < \text{nota} \leq 64,40$	nota $> 64,40$
2008	nota $\leq 29,26$	$29,26 < \text{nota} \leq 58,53$	nota $> 58,53$
2011	nota $\leq 30,26$	$30,26 < \text{nota} \leq 60,53$	nota $> 60,53$
2014	nota $\leq 32,46$	$32,46 < \text{nota} \leq 64,93$	nota $> 64,93$
2017	nota $\leq 32,46$	$32,46 < \text{nota} \leq 64,43$	nota $> 64,43$

4.4 Mineração de dados

Nesta etapa são aplicados os cinco métodos distintos para seleção de atributos, em cada ano do exame. Para a seleção de características são consideradas: as bases de dados do ENADE conforme o ano, e consideramos os n atributos (variando de acordo com o ano) pré-selecionados na etapa de seleção descrita na Seção 4.1.

Os algoritmos de SA aplicados, apresentados em detalhes na Seção 2.2, são:

1. Ganho de Informação,

2. Taxa de Ganho,
3. Incerteza Simétrica,
4. Correlação de Pearson,
5. Alívio F.

Para cada ano selecionado do ENADE, escolhemos as características mais relevantes conforme o ranking gerado pelos algoritmos aplicados. Nos rankings gerados para cada ano, aplicamos os algoritmos de Correlação de Classificação, apresentados em detalhes na Seção 2.4, sendo eles:

1. Correlação de Spearman,
2. Correlação de Kendall,
3. Correlação de Goodman e Kruska,
4. Correlação D de Somers.

4.5 Interpretação e Avaliação

Nesta última etapa, após a aplicação dos cinco métodos de SA e análise do ranking gerados pelo algoritmos, são considerados os top-10 atributos mais frequentes em pelo menos quatro dos algoritmos de seleção de características, para cada ano do ENADE selecionado. Também é feita a análise de cada algoritmo de SA, de acordo com cada ano de aplicação. Na sequência, é feita a análise de cada ano de estudo. E em seguida, analisamos o desempenho de cada atributo com o passar dos anos. Ainda, analisamos os resultados dos algoritmos de Correlação de Classificação e, por fim, analisamos as características mais importantes apontadas pelos métodos SA. Todos esses resultados serão vistos no Capítulo 5.

5 RESULTADOS

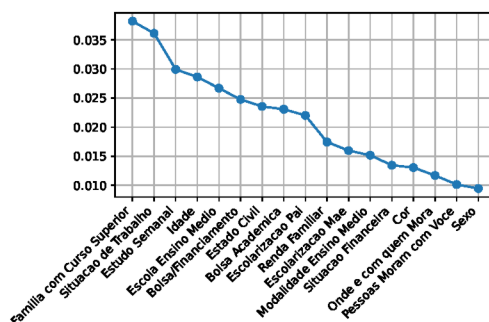
Esse Capítulo apresenta os resultados do estudo quando os diferentes algoritmos de SA são aplicados. Para um melhor entendimento do estudo e melhor apresentação, dividiu-se a seção em cinco partes distintas. Na primeira, é apresentado o resultado de cada algoritmos de SA, ao longo dos anos de análise. Na segunda parte, discutimos os resultados apresentados dos diferentes algoritmos de SA, ao longo dos anos. Na terceira, os resultados de cada atributo (posição no ranking em cada ano) são analisados conforme anos selecionados do ENADE. Na quarta parte, os resultados obtidos de cada algoritmo de correlação de classificação são apresentados e discutidos. Por fim, apresentamos os resultados da análise das características mais importantes apontadas pelos algoritmos de Seleção de Atributos.

5.1 Resultados obtidos em cada algoritmo de Seleção de Atributos

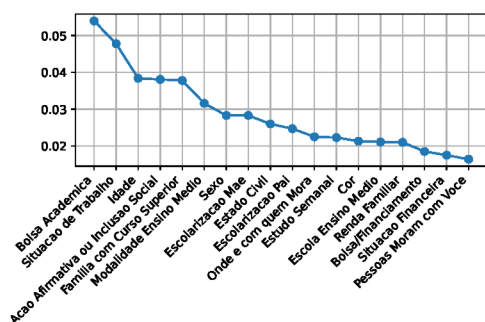
Esta Seção apresenta os resultados obtidos quando cada algoritmo de SA é aplicado nos diferentes anos considerados do ENADE.

A Figura 3 apresenta o resultado do método de seleção de atributos Correlação. Para cada atributo preditor é exibido o valor retornado pelo método. Todos os atributos são ordenados resultando num ranking de variáveis. Os anos são divididos em ordem conforme os itens (a), (b), (c), (d), e (e) da Figura 3. Analisando a área sobre o gráfico do ENADE 2005 (a) e o decaimento da curva, é possível perceber que há quase uma linearidade nos resultados, indicando que os atributos vão perdendo importância aos poucos e quase de forma igualitária. Na Figura 3-(b), percebemos que os dois primeiros atributos sobressaem-se frente aos outros atributos. Na Figura 3-(c), notamos uma maior correlação nos três primeiros atributos, enquanto na Figura 3-(d), observamos uma linearidade nos quatro primeiros atributos. Por fim, na Figura 3-(e), o atributo Idade é o que apresenta a maior correlação, pois os outros atributos ficaram mais abaixo e com menor variação.

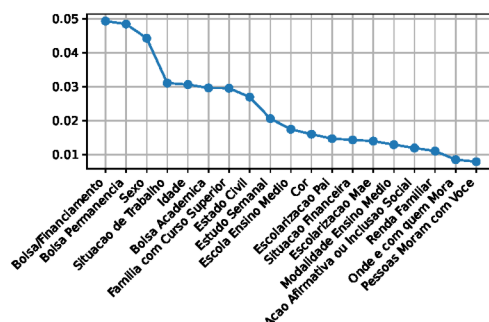
Na Figura 4, são mostrados os resultados do método de seleção de atributos Ganho de Informação ao longo dos anos de estudo do ENADE. A mesma abordagem vista na Figura 3 será utilizada para a análise deste e dos próximos resultados. Ao analisarmos



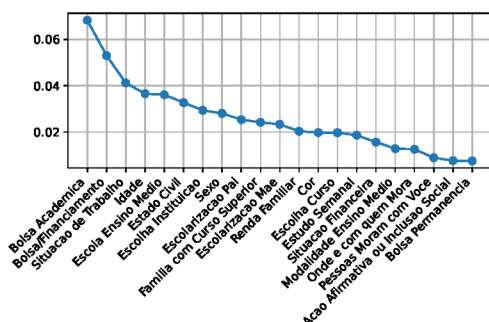
(a) ENADE 2005



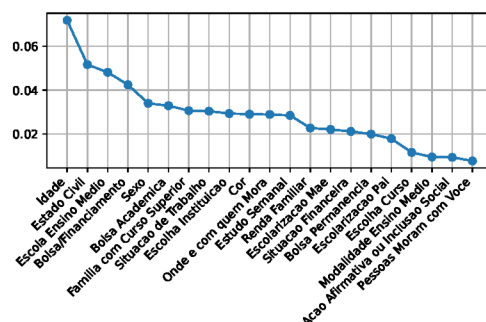
(b) ENADE 2008



(c) ENADE 2011



(d) ENADE 2014



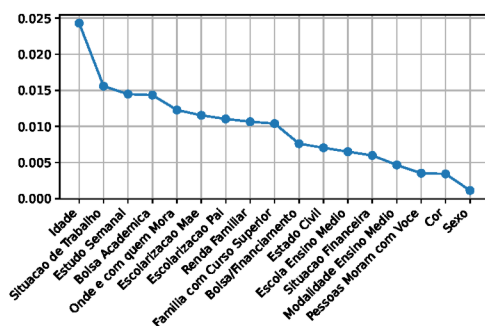
(e) ENADE 2017

Figura 3: Resultado da Correlação.

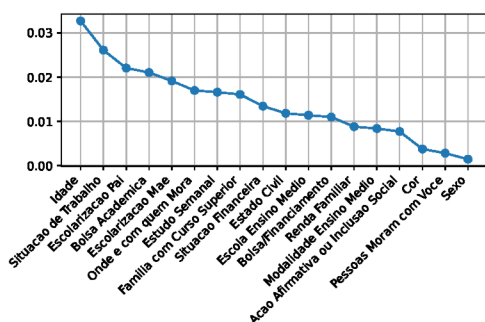
a Figura 4, percebemos que nos anos de 2005, 2011, 2014 e 2017, o primeiro atributo destoa muito acima dos outros. Já no ano de 2008, temos uma linearidade nos resultados.

O resultado do método de SA Taxa de Ganho pode ser visto na Figura 5. Ao analisarmos os anos de 2005 e 2008, percebemos um decaimento suave de todos os atributos. No ano de 2011, notamos que o atributo Bolsa/Financiamento possui uma maior taxa de ganho, comparada aos outros atributos. Nos anos de 2014 e 2017, os atributos com maiores valores são: Bolsa Acadêmica e Bolsa/Financiamento, nesta ordem nos dois anos.

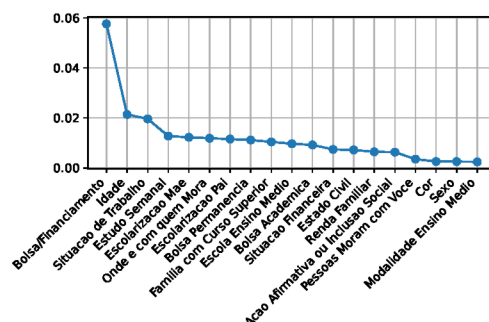
Na Figura 6, apresentamos o resultado do método de SA Incerteza Simétrica. No ano de 2008, há uma certa linearidade nos resultados. Já nos anos de 2005 e 2011, notamos o primeiro atributo destoando muito acima dos outros. E nos anos de 2014 e 2017, temos



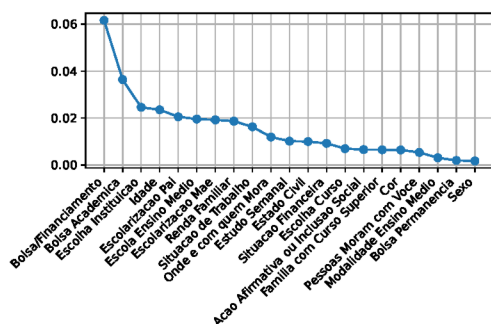
(a) ENADE 2005



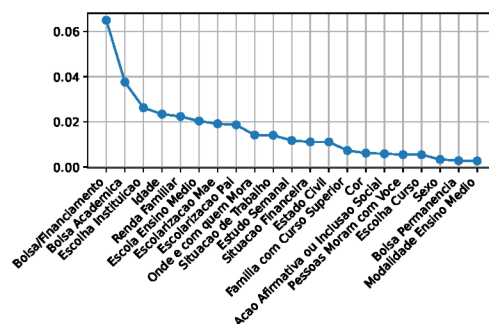
(b) ENADE 2008



(c) ENADE 2011



(d) ENADE 2014

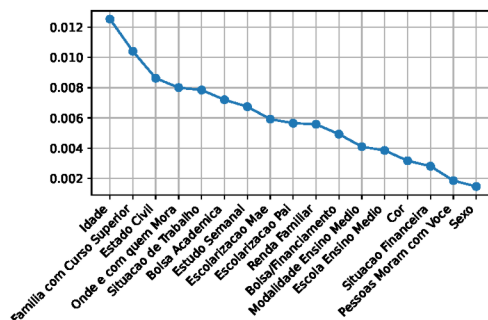


(e) ENADE 2017

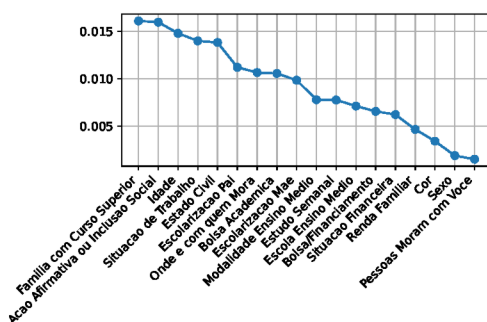
Figura 4: Resultado do Ganho de Informação.

dois atributos, com uma maior incerteza simétrica.

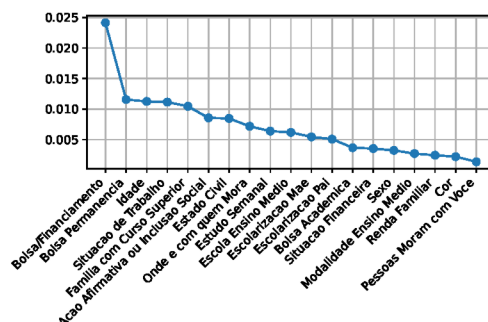
O resultado do último método de SA, o Alívio F, pode ser visto na Figura 7. Observando a Figura 7, percebemos o primeiro atributo (Estudo Semanal) um pouco destoado dos demais. Nos anos de 2008 e 2011, observamos um comportamento quase idêntico na curva dos gráficos, ambos apresentando uma linearidade. Por fim, nos anos de 2014 e 2017, dois atributos destacam-se com maior valor, sendo eles: Bolsa/Financiamento e Renda Familiar, nesta ordem nos dois anos.



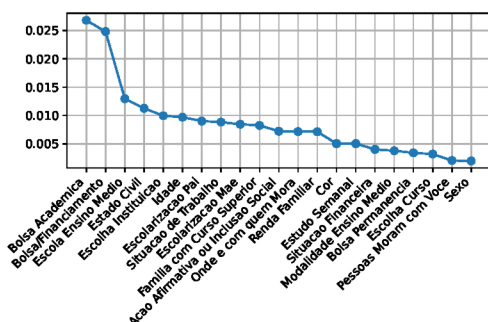
(a) ENADE 2005



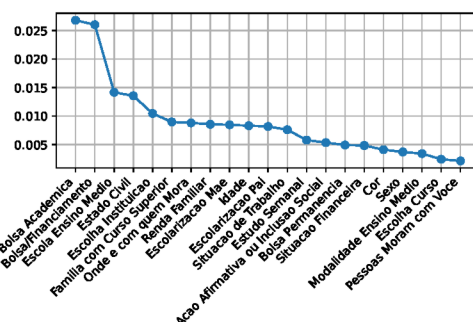
(b) ENADE 2008



(c) ENADE 2011



(d) ENADE 2014



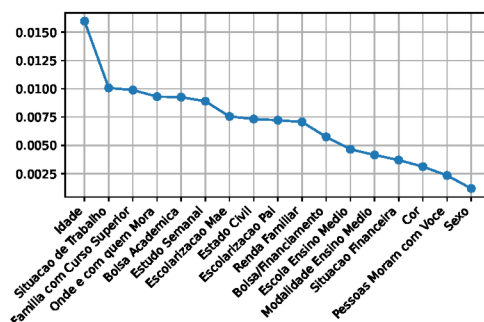
(e) ENADE 2017

Figura 5: Resultado da Taxa de Ganho.

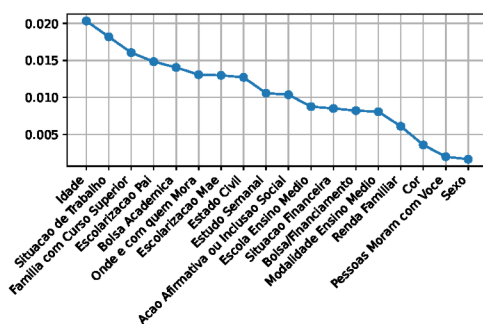
5.2 Resultados obtidos em cada ano do ENADE

Essa Seção apresenta os resultados obtidos quando as técnicas de SA são aplicadas nos diferentes anos do ENADE.

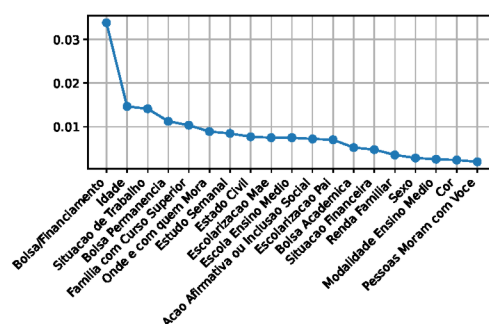
Os resultados dos cinco algoritmos de SA aplicados no ENADE 2005 estão disponíveis na Tabela 7, onde as linhas são os 17 atributos considerados (neste ano) e nas colunas, os diferentes métodos de SA. Os valores de cada célula referem-se à posição no *ranking* e ao resultado da operação obtidos pelo atributo. Os dez melhores valores, para cada método, são sublinhados. Além disso, estão destacados em **negrito** os atributos selecionados nas dez primeiras posições em pelo menos quatro dos algoritmos utilizados. Ao final da tabela, a coluna *Rank* contém a soma das quatro melhores posições obtidas pelos diferentes métodos de SA.



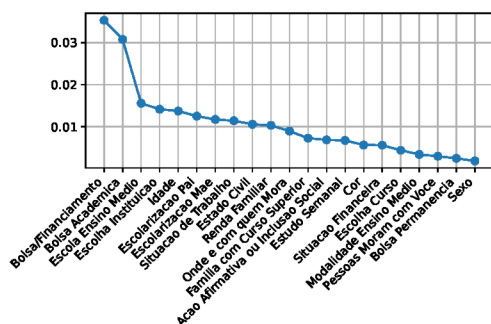
(a) ENADE 2005



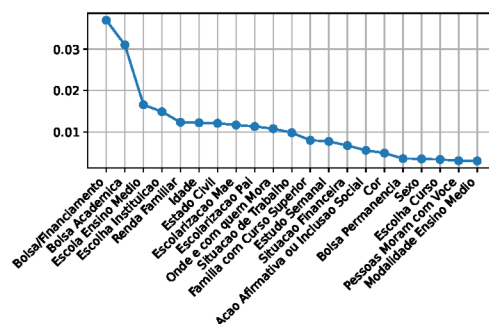
(b) ENADE 2008



(c) ENADE 2011



(d) ENADE 2014

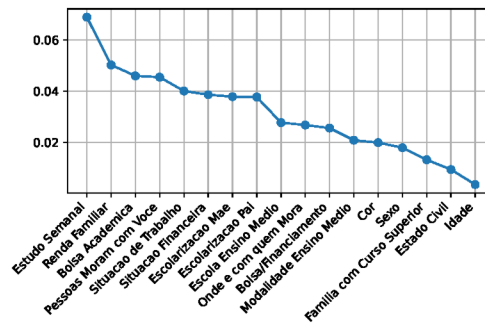


(e) ENADE 2017

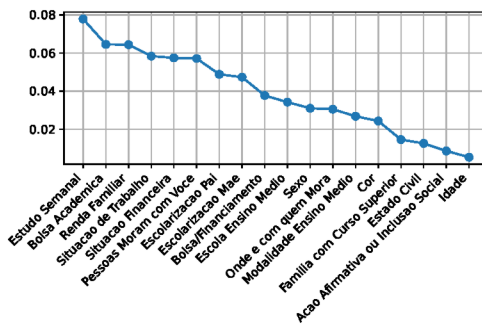
Figura 6: Resultado da Incerteza Simétrica.

Analisando a Tabela 7, podemos observar que os atributos que mais aparecem entre as duas primeiras posições do ranking são: Idade e Situação de Trabalho. Ainda podemos observar os atributos Estudo Semanal, Família com Curso superior e Bolsa Acadêmica se mantém nas primeiras colocações também. Por último, destacamos os atributos Onde e com quem mora, Escolarização da Mãe, Renda Familiar e Escolarização do Pai, fechando a seleção dos *top-10* considerados os mais importantes. É importante perceber que Estado Civil não foi considerado porque, apesar de bem classificado por três algoritmos, ficou em 11º lugar quando considerado o Ganho de Informação (GI) e 16º para o Alívio F (AL).

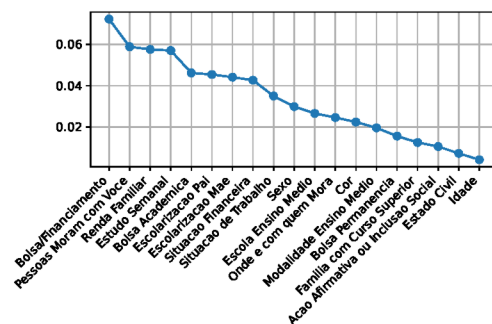
De forma análoga, os resultados dos cinco algoritmos de SA aplicados no ENADE 2008 estão disponíveis na Tabela 8. Observando os resultados, percebe-se que os atributos Idade, Situação de Trabalho, Bolsa Acadêmica, Família com Curso Superior e



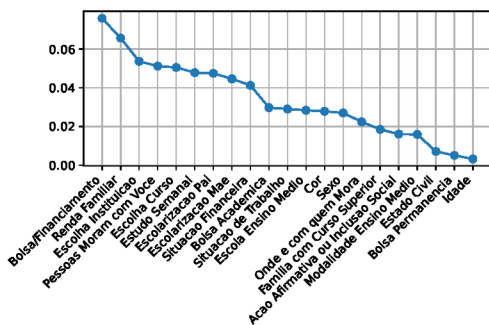
(a) ENADE 2005



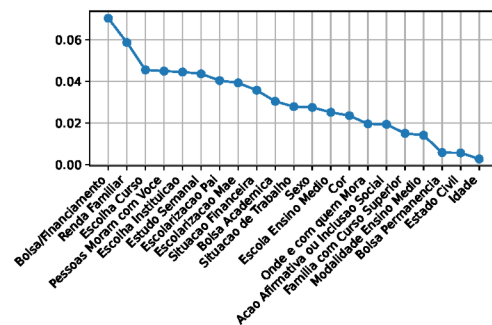
(b) ENADE 2008



(c) ENADE 2011



(d) ENADE 2014



(e) ENADE 2017

Figura 7: Resultado do Alívio F.

Escolarização Pai são os que mais se destacam. Por último, também aparecem como *top-10* em quatro dos algoritmos de SA, Escolarização Mãe e Estado Civil.

Na Tabela 9, são apresentados os resultados do ENADE 2011. Destaca-se, neste resultado, que o atributo Bolsa/Financiamento ocupa a primeira colocação em todos os cinco algoritmos, mostrando a sua importância neste ano de exame. Idade, Situação de Trabalho, Bolsa Permanência, Estudo Semanal, Família com Curso Superior e Escola Ensino Médio são os atributos que se mantêm no *top-10* em quatro das técnicas de seleção de atributos. Podemos notar outros atributos, antes considerados importantes, como Escolarização Mãe, Onde e com quem mora, mas estes não são considerados por não estarem no *top-10* em pelo menos quatro algoritmos de SA, estes atributos foram penalizados por Correlação de Pearson, Taxa de Ganho e pelo Alívio F.

Tabela 7: Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2005 para concluintes de Engenharia.

#	Atributos	Corr	GI	TG	IS	AL	Rank
1	Idade	4 0,02862	1 0,02431	1 0,01252	1 0,01597	17 0,00363	7
2	Situação de Trabalho	2 0,03612	2 0,01558	5 0,00784	2 0,01008	5 0,04007	11
3	Estudo Semanal	3 0,02992	3 0,01448	7 0,00674	6 0,00890	1 0,06886	13
4	Família com Curso Superior	1 0,03821	9 0,01040	2 0,01040	3 0,00989	15 0,01329	15
5	Bolsa Acadêmica	8 0,02307	4 0,01434	6 0,00720	5 0,00926	3 0,04598	18
6	Onde e com quem Mora	15 0,01169	5 0,01228	4 0,00800	4 0,00931	10 0,02682	23
7	Escolarização Mãe	11 0,01597	6 0,01154	8 0,00592	7 0,00756	7 0,03787	28
8	Estado Civil	7 0,02356	11 0,00704	3 0,00862	8 0,00733	16 0,00947	29
9	Renda Familiar	10 0,01744	8 0,01065	10 0,00558	10 0,00707	2 0,05023	30
10	Escolarização Pai	9 0,02201	7 0,01104	9 0,00565	9 0,00723	8 0,03773	33
11	Bolsa/Financiamento	6 0,02476	10 0,00761	11 0,00493	11 0,00575	11 0,02564	38
12	Escola Ensino Médio	5 0,02669	12 0,00650	13 0,00385	12 0,00466	9 0,02776	38
13	Situação Financeira	13 0,01345	13 0,00598	15 0,00281	14 0,00371	6 0,03864	46
14	Modalidade Ensino Médio	12 0,01516	14 0,00466	12 0,00410	13 0,00416	12 0,02089	49
15	Pessoas Moram com Você	16 0,01014	15 0,00352	16 0,00186	16 0,00235	4 0,04547	51
16	Cor	14 0,01305	16 0,00342	14 0,00317	15 0,00313	13 0,01996	56
17	Sexo	17 0,00945	17 0,00112	17 0,00147	17 0,00120	14 0,01800	65

Tabela 8: Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2008 para concluintes de Engenharia.

#	Atributos	Corr	GI	TG	IS	AL	Rank
1	Idade	3 0,03840	1 0,03271	3 0,01479	1 0,02033	18 0,00533	8
2	Situação de Trabalho	2 0,04780	2 0,02610	4 0,01399	2 0,01818	4 0,05835	10
3	Bolsa Acadêmica	1 0,05400	4 0,02106	8 0,01056	5 0,01404	2 0,06452	12
4	Família com Curso Superior	5 0,03780	8 0,01609	1 0,01609	3 0,01605	15 0,01459	17
5	Escolarização Pai	10 0,02470	3 0,02206	6 0,01121	4 0,01483	7 0,04880	20
6	Estudo Semanal	12 0,02230	7 0,01660	11 0,00777	9 0,01057	1 0,07785	28
7	Escolarização Mãe	8 0,02830	5 0,01914	9 0,00985	7 0,01298	8 0,04733	28
8	Onde e com quem Mora	11 0,02250	6 0,01699	7 0,01066	6 0,01307	12 0,03051	30
9	Ação Afirmativa ou Inclusão Social	4 0,03810	15 0,00771	2 0,01596	10 0,01036	17 0,00862	31
10	Estado Civil	9 0,02600	10 0,01183	5 0,01382	8 0,01271	16 0,01261	32
11	Situação Financeira	17 0,01750	9 0,01345	14 0,00623	12 0,00850	5 0,05744	40
12	Modalidade Ensino Médio	6 0,03160	14 0,00838	10 0,00779	14 0,00806	13 0,02681	43
13	Escola Ensino Médio	14 0,02110	11 0,01139	12 0,00713	11 0,00875	10 0,03422	44
14	Renda Familiar	15 0,02100	13 0,00880	15 0,00467	15 0,00609	3 0,06437	46
15	Bolsa/Financiamento	16 0,01850	12 0,01099	13 0,00657	13 0,00821	9 0,03770	47
16	Sexo	7 0,02830	18 0,00144	17 0,00191	18 0,00164	11 0,03101	53
17	Pessoas Moram com Você	18 0,01640	17 0,00281	18 0,00153	17 0,00198	6 0,05711	58
18	Cor	13 0,02130	16 0,00375	16 0,00343	16 0,00357	14 0,02434	59

A Tabela 10 apresenta os resultados dos cinco algoritmos de SA aplicados no ENADE 2014. Analisando a Tabela 10, podemos observar que os atributos Bolsa/Financiamento e Bolsa Acadêmica destacam-se por estarem em primeiro ou segundo no no *ranking* da maioria dos algoritmos. Escolha Instituição, Escola Ensino Médio, Idade, Escolarização Pai, Situação de Trabalho e Escolarização Mãe são os atributos que se mantêm no *top-10* em sua maioria, com exceção dos resultados do algoritmo Alivio F, que apresenta alguns resultados bem diferentes dos demais.

Por fim, os resultados dos cinco algoritmos de SA aplicados no ENADE 2017 estão disponíveis na Tabela 11. Nota-se que Bolsa/financiamento é o atributo que apresenta melhor aproveitamento entre todos os algoritmos. Ainda podemos destacar Bolsa Acadêmica, Escola Ensino Médio, Escolha Instituição, Renda Familiar e Idade como os atributos da seleção que são considerados os mais importantes. Para finalizar, o atributo

Tabela 9: Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2011 para concluintes de Engenharia.

#	Atributos	Corr	GI	TG	IS	AL	Rank
1	Bolsa/Financiamento	<u>1</u> 0,04934	<u>1</u> 0,05766	<u>1</u> 0,02415	<u>1</u> 0,03381	<u>1</u> 0,07220	4
2	Idade	<u>5</u> 0,03063	<u>2</u> 0,02141	<u>3</u> 0,01125	<u>2</u> 0,01463	19 0,00414	12
3	Situação de Trabalho	<u>4</u> 0,03107	<u>3</u> 0,01965	<u>4</u> 0,01114	<u>3</u> 0,01410	9 0,03495	14
4	Bolsa Permanência	<u>2</u> 0,04846	<u>8</u> 0,01117	<u>2</u> 0,01156	<u>4</u> 0,01122	15 0,01557	16
5	Estudo Semanal	<u>9</u> 0,02060	<u>4</u> 0,01276	<u>9</u> 0,00638	<u>7</u> 0,00844	<u>4</u> 0,05699	24
6	Família com Curso Superior	<u>7</u> 0,02955	<u>9</u> 0,01044	<u>5</u> 0,01045	<u>5</u> 0,01032	16 0,01249	26
7	Escolarização Mãe	14 0,01396	<u>5</u> 0,01229	11 0,00543	<u>9</u> 0,00747	<u>7</u> 0,04409	32
8	Onde e com quem Mora	18 0,00849	<u>6</u> 0,01190	<u>8</u> 0,00717	<u>6</u> 0,00887	12 0,02457	32
9	Bolsa Acadêmica	<u>6</u> 0,02963	11 0,00923	13 0,00366	13 0,00520	<u>5</u> 0,04614	35
10	Estado Civil	<u>8</u> 0,02694	13 0,00721	<u>7</u> 0,00846	<u>8</u> 0,00768	18 0,00722	36
11	Escolarização Pai	12 0,01468	<u>7</u> 0,01156	12 0,00507	12 0,00700	<u>6</u> 0,04538	37
12	Escola Ensino Médio	<u>10</u> 0,01744	<u>10</u> 0,00970	<u>10</u> 0,00617	<u>10</u> 0,00747	11 0,02653	40
13	Sexo	<u>3</u> 0,04429	18 0,00257	15 0,00322	16 0,00282	<u>10</u> 0,02985	44
14	Situação Financeira	13 0,01431	12 0,00742	14 0,00352	14 0,00474	<u>8</u> 0,04260	47
15	Ação Afirmativa ou Inclusão Social	16 0,01193	15 0,00631	<u>6</u> 0,00855	11 0,00717	17 0,01051	48
16	Renda Familiar	17 0,01102	14 0,00646	17 0,00243	15 0,00351	<u>3</u> 0,05750	49
17	Pessoas Moram com Você	19 0,00788	16 0,00353	19 0,00135	19 0,00194	<u>2</u> 0,05881	56
18	Cor	11 0,01599	17 0,00259	18 0,00219	18 0,00235	13 0,02238	59
19	Modalidade Ensino Médio	15 0,01291	19 0,00245	16 0,00269	17 0,00253	14 0,01956	62

Escolarização da Mãe mantém-se no *top-10* em quatro dos algoritmos, excetuando no algoritmo de Correlação.

5.3 Resultados obtidos de cada Atributo ao Longo dos Anos

Essa Seção, apresenta os resultados obtidos de cada atributo ao longo dos anos. De forma a analisar a variação do perfil dos estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos considerados neste estudo, foi calculada a média dos rankings para cada atributo e ano de aplicação do exame. Como visualizar todos os atributos de uma única vez dificultaria muito o entendimento, optou-se por dividi-los em três subconjuntos, apresentando os comportamentos de sete atributos em cada gráfico, agrupados de acordo com a ordem geral de classificação relativa a coluna *Rank*, conforme ilustrado nas Figuras 8, 9 e 10, com as partes 1, 2 e 3, respectivamente.

A Figura 8 mostra os sete primeiros atributos. É possível perceber que Idade e Situação de Trabalho vêm perdendo importância ao longo dos anos, assim como o atributo Família com Curso Superior, embora esse acene uma ligeira recuperação em 2017. Escolha Instituição é um atributo que surgiu somente em 2014, com um bom resultado (em terceiro), caindo para a quarta colocação em 2017. O atributo Bolsa/Financiamento inicia em 2005 com modesta importância mas a partir de 2011 torna-se o mais relevante. Bolsa Acadêmica ficou em quinto em 2005, caindo para nono em 2011, mas se torna o segundo atributo mais importante nos dois últimos anos considerados neste estudo. O último atributo analisado neste gráfico é Escolarização Mãe, que se mostra estável ao longo dos anos, ficando sempre entre o sexto e oitavo lugar.

Os atributos classificados da 8^a a 14^a posição do *ranking* geral são apresentados na Figura 9. Estudo Semanal e Onde e com quem Mora são atributos que perdem a im-

Tabela 10: Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2014 para concluintes de Engenharia.

#	Atributos	Corr	GI	TG	IS	AL	Rank					
1	Bolsa/Financiamento	<u>2</u>	0,05304	<u>1</u>	0,06162	<u>2</u>	0,02480	<u>1</u>	0,03530	<u>1</u>	0,07589	5
2	Bolsa Acadêmica	<u>1</u>	0,06830	<u>2</u>	0,03637	<u>1</u>	0,02679	<u>2</u>	0,03076	<u>10</u>	0,02973	6
3	Escolha Instituição	<u>7</u>	0,02936	<u>3</u>	0,02458	<u>5</u>	0,00996	<u>4</u>	0,01415	<u>3</u>	0,05364	15
4	Escola Ensino Médio	<u>5</u>	0,03613	<u>6</u>	0,01958	<u>3</u>	0,01297	<u>3</u>	0,01556	12	0,02837	17
5	Idade	<u>4</u>	0,03655	<u>4</u>	0,02353	<u>6</u>	0,00971	<u>5</u>	0,01372	21	0,00322	19
6	Escolarização Pai	<u>9</u>	0,02534	<u>5</u>	0,02054	<u>7</u>	0,00902	<u>6</u>	0,01251	<u>7</u>	0,04747	25
7	Situação de Trabalho	<u>3</u>	0,04121	<u>9</u>	0,01625	<u>8</u>	0,00884	<u>8</u>	0,01142	11	0,02904	28
8	Escolarização Mãe	11	0,02332	<u>7</u>	0,01923	<u>9</u>	0,00846	<u>7</u>	0,01172	<u>8</u>	0,04462	31
9	Estado Civil	<u>6</u>	0,03270	12	0,00999	<u>4</u>	0,01128	<u>9</u>	0,01056	19	0,00712	31
10	Renda Familiar	12	0,02032	<u>8</u>	0,01868	13	0,00715	<u>10</u>	0,01032	<u>2</u>	0,06569	32
11	Estudo Semanal	15	0,01861	11	0,01018	15	0,00505	14	0,00673	<u>6</u>	0,04782	46
12	Família com Curso Superior	<u>10</u>	0,02418	16	0,00651	<u>10</u>	0,00826	12	0,00726	16	0,01850	48
13	Onde e com quem Mora	18	0,01253	<u>10</u>	0,01192	12	0,00718	11	0,00894	15	0,02240	48
14	Escolha Curso	14	0,01970	14	0,00698	19	0,00320	17	0,00438	<u>5</u>	0,05051	50
15	Situação Financeira	16	0,01559	13	0,00923	16	0,00403	16	0,00560	<u>9</u>	0,04127	54
16	Cor	13	0,01973	17	0,00644	14	0,00505	15	0,00564	13	0,02783	55
17	Ação Afirmativa ou Inclusão Social	20	0,00759	15	0,00657	11	0,00723	13	0,00686	17	0,01604	56
18	Pessoas Moram com Você	19	0,00886	18	0,00537	20	0,00207	19	0,00298	<u>4</u>	0,05119	60
19	Sexo	<u>8</u>	0,02805	21	0,00174	21	0,00199	21	0,00185	14	0,02701	64
20	Modalidade Ensino Médio	17	0,01280	19	0,00311	17	0,00379	18	0,00340	18	0,01590	70
21	Bolsa Permanência	21	0,00747	20	0,00195	18	0,00342	20	0,00247	20	0,00505	78

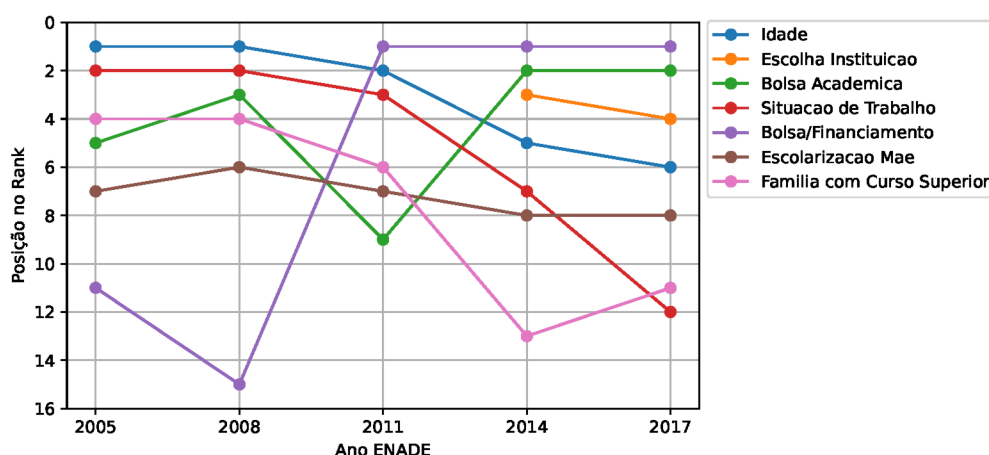


Figura 8: Evolução do perfil de estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos, parte 1/3.

portância ao longo dos anos. Escola Ensino Médio mantém-se estável na 12ª posição de 2005 a 2011. Depois, sobe no ranking até se tornar o terceiro atributo mais importante. Renda Familiar começa em sétimo, cai até a 16ª posição em 2011 e termina em quinto em 2017, mostrando uma forte recuperação. Três atributos apresentam certa variação, dependendo do ano, sendo eles: Escolarização Pai (variando de décimo a quinto dependendo do ano), Situação Financeira (varia entre as 11ª e 15ª posições) e Estado Civil (começa em oitavo, cai para décimo e volta a subir para o sétimo lugar).

Finalmente, a Figura 10 mostra os últimos atributos do *ranking* geral. Nota-se que Bolsa Permanência sai de quarto lugar em 2011 e despenca para a 20ª posição em 2017. Outros quatro atributos apresentam um comportamento decrescente ao longo dos anos,

Tabela 11: Resultados obtidos com diferentes algoritmos de seleção de atributos, considerando o ENADE 2017 para concluintes de Engenharia.

#	Atributos	Corr	GI	TG	IS	AL	Rank
1	Bolsa/Financiamento	4 0,04244	1 0,06501	2 0,02604	1 0,03696	1 0,07026	5
2	Bolsa Acadêmica	6 0,03290	2 0,03763	1 0,02681	2 0,03103	10 0,03045	11
3	Escola Ensino Médio	3 0,04810	6 0,02039	3 0,01418	3 0,01658	13 0,02518	15
4	Escolha Instituição	9 0,02935	3 0,02627	5 0,01047	4 0,01489	5 0,04450	17
5	Renda Familiar	13 0,02276	5 0,02242	8 0,00857	5 0,01233	2 0,05864	20
6	Idade	1 0,07175	4 0,02347	10 0,00833	6 0,01223	21 0,00280	21
7	Estado Civil	2 0,05161	13 0,01113	4 0,01357	7 0,01209	20 0,00572	26
8	Escolarização Mãe	14 0,02210	7 0,01918	9 0,00848	8 0,01168	8 0,03929	32
9	Escolarização Pai	17 0,01792	8 0,01870	11 0,00819	9 0,01132	7 0,04041	35
10	Onde e com quem Mora	11 0,02892	9 0,01419	7 0,00882	10 0,01079	15 0,01960	37
11	Família com Curso Superior	7 0,03066	14 0,00735	6 0,00898	12 0,00799	17 0,01507	39
12	Situação de Trabalho	8 0,03041	10 0,01409	12 0,00762	11 0,00982	11 0,02789	40
13	Estudo Semanal	12 0,02849	11 0,01178	13 0,00580	13 0,00772	6 0,04365	42
14	Situação Financeira	15 0,02121	12 0,01114	16 0,00485	14 0,00672	9 0,03576	50
15	Sexo	5 0,03398	19 0,00340	18 0,00372	18 0,00351	12 0,02752	53
16	Cor	10 0,02905	15 0,00620	17 0,00412	16 0,00491	14 0,02358	55
17	Escolha Curso	18 0,01166	18 0,00547	20 0,00244	19 0,00335	3 0,04551	58
18	Ação Afirmativa ou Inclusão Social	20 0,00944	16 0,00591	14 0,00536	15 0,00556	16 0,01940	61
19	Pessoas Moram com Você	21 0,00778	17 0,00552	21 0,00214	20 0,00306	4 0,04494	62
20	Bolsa Permanência	16 0,01997	20 0,00285	15 0,00494	17 0,00357	19 0,00585	67
21	Modalidade Ensino Médio	19 0,00958	21 0,00272	19 0,00343	21 0,00300	18 0,01415	77

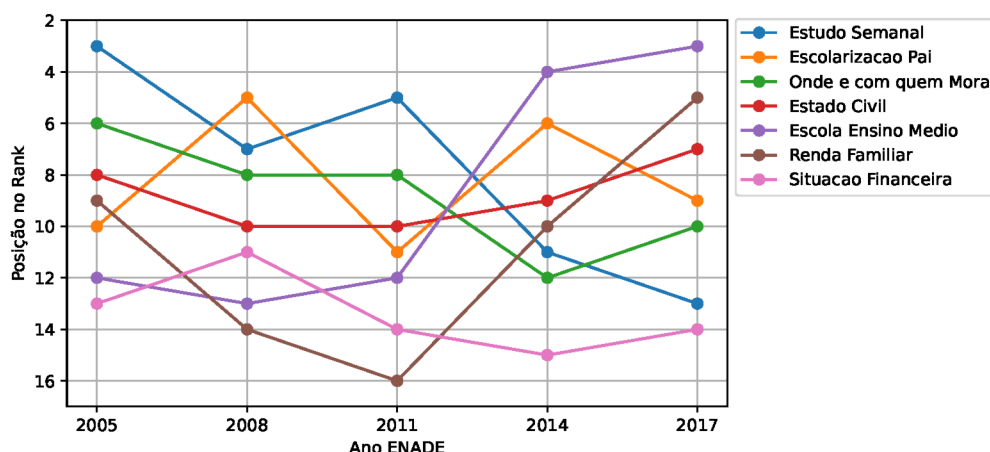


Figura 9: Evolução do perfil de estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos, parte 2/3.

sendo eles: Ação Afirmativa ou inclusão Social, Escolha Curso, Modalidade Ensino Médio, e Pessoas Moram com Você. Sexo apresenta bastante variação ao longo dos anos, oscilando da 13ª até a 19ª posição. Por fim, o atributo Cor é o mais estável, iniciando em 2005 e finalizando em 2017 na 16ª posição, tendo ficado em décimo oitavo em 2008 e 2011.

5.4 Resultados obtidos da Correlação de Classificação

Essa Seção, apresenta os resultados obtidos de cada algoritmo de correlação de classificação. Para aplicar os algoritmos de correlação de classificação, foi necessário retirar os atributos (bolsa permanência, escolha curso e escolha instituição) que não es-

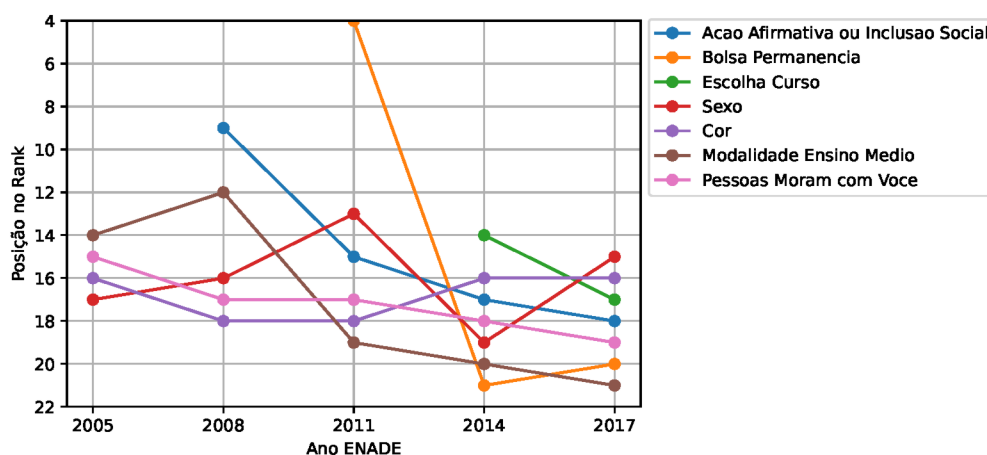


Figura 10: Evolução do perfil de estudantes de Engenharia com bom desempenho no ENADE ao longo dos anos, parte 3/3.

tavam presentes em todos os anos, pois os algoritmos usados só podem ser aplicados se todos os anos tiverem o mesmo número de atributos.

Nas Tabelas 12 e 13, podemos ver os resultados obtidos com a correlação de Spearman e a correlação de Kendall, respectivamente. Os resultados da correlação de classificação de Goodman e Kruskal e os D de Somers não são mostrados separados, pois os mesmos apresentam resultados idênticos aos de Kendall. O porquê destes resultados serem iguais pode ser visto na Subseção 2.4.3 e na Subseção 2.4.4. Estes resultados estão no formato de uma matriz de correlação, na qual exibe a semelhança entre os diferentes anos do exame. Quanto maior for o valor absoluto do coeficiente, mais forte é a relação entre as variáveis. Na diagonal principal, sempre vemos o valor 1 por ser a relação entre o próprio ano com ele mesmo. Na primeira linha e coluna, tem-se os anos e o resultado geral (classificação com a média das posições).

Observando os resultados da Tabela 12, percebemos que tanto no ano de 2005, quanto no ano de 2008, com o passar dos anos, o valor de semelhança vai decaindo. Enquanto que nos anos de 2014 e 2017, tem-se o inverso, com anos mais antigos apresentando os menores valores de semelhança. Nota-se 2011 como sendo o ano médio dessa mudança, pois nos anos 2008 e 2014 tem-se valores de correlação praticamente iguais, e o ano de 2005 apresenta-se o de maior correlação. Na coluna geral, praticamente todos os anos apresentam valores altos de correlação, exceto o ano de 2017, que é um resultado esperado por ser um Rank da média dos anos.

Na Tabela 13, temos os resultados da Correlação de Kendall, no qual nota-se resultados semelhantes aos da Correlação de Spearman, com a diferença nos valores de correlação que são menores. Esses valores não significam menor correlação, pois não são comparáveis numericamente, conforme explicado na Subseção 2.4.2. Nos anos de 2005 e 2008, com o passar dos anos, o valor de semelhança vai decaindo. Entretanto, nos anos de 2014 e 2017, tem-se o inverso, com anos mais antigos apresentando os menores valores

Tabela 12: Resultados obtidos com a Correlação de Spearman.

Ano	2005	2008	2011	2014	2017	Geral
2005	1	0.887	0.784	0.525	0.429	0.824
2008	0.887	1	0.650	0.473	0.301	0.787
2011	0.784	0.650	1	0.654	0.527	0.887
2014	0.525	0.473	0.654	1	0.897	0.824
2017	0.429	0.301	0.527	0.897	1	0.679
Geral	0.824	0.787	0.887	0.824	0.679	1

de correlação. Em 2011, nota-se novamente como o ano médio dessa mudança, pois os anos 2008 e 2014 tem valores de correlação praticamente iguais, e o ano de 2017 sendo o ano de menor correlação. Já a coluna geral apresenta valores próximos de correlação em praticamente todos os anos, exceto em 2017, que apresenta um valor um pouco mais baixo. Mas, este resultado é esperado por ser um Rank da média dos anos.

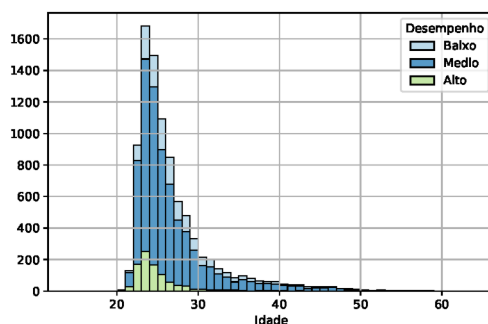
Tabela 13: Resultados obtidos com a Correlação de Kendall.

Ano	2005	2008	2011	2014	2017	Geral
2005	1	0.721	0.662	0.368	0.206	0.632
2008	0.721	1	0.559	0.353	0.191	0.647
2011	0.662	0.559	1	0.500	0.368	0.706
2014	0.368	0.353	0.500	1	0.750	0.647
2017	0.206	0.191	0.368	0.750	1	0.515
Geral	0.632	0.647	0.706	0.647	0.515	1

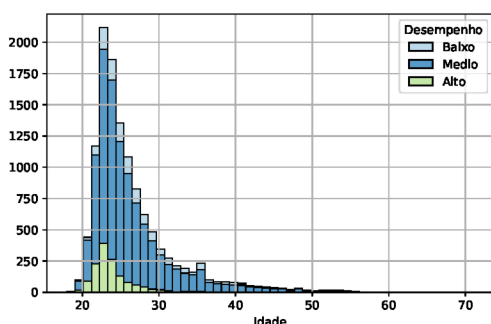
5.5 Resultados da Análise das Características mais Importantes

Essa Seção apresenta os resultados da análise das características mais importantes apontadas pelos algoritmos de Seleção de Atributos. Para a definição das características mais relevantes para explicar o desempenho no ENADE ao longo dos anos, consideramos os *top-10* atributos mais frequentes em pelo menos quatro dos algoritmos de seleção de características, conforme apresentado na Seção 5.2. As Figuras 11-24 apresentam a distribuição dos estudantes em relação a cada uma destas características considerando o desempenho baixo, médio ou alto, discretizados conforme descrito na Seção 4.3. Em cada figura, somente o ano que este atributo foi considerado importante é apresentado, salienta-se que as escalas das Figuras aumentam com o passar dos anos porque o número de participantes também aumenta.

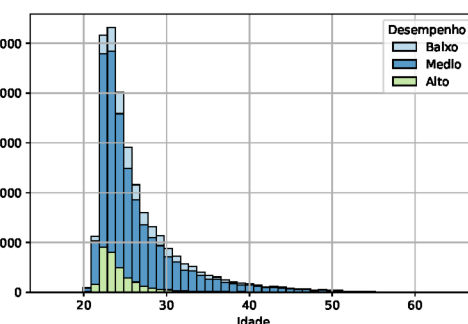
Analisando a Figura 11, observamos o atributo idade, estando a maioria dos concluintes de engenharia, independente do ano, na faixa etária de 20 a 30 anos. Nota-se que com o passar dos anos, o número de concluintes vai diminuindo. Aqueles que apresentam o desempenho alto, notamos que se concentram na faixa etária de 22 a 25 anos. Ainda podemos observar nos anos de 2011 e 2014, que o desempenho alto vai diminuindo de



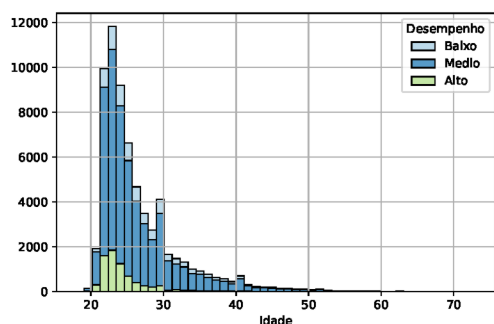
(a) ENADE 2005



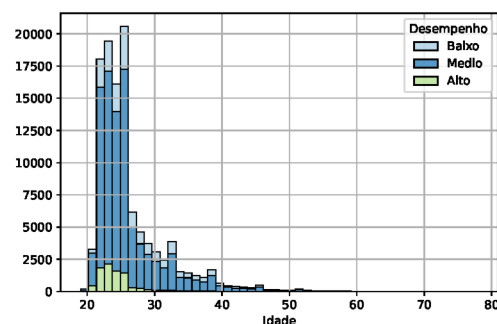
(b) ENADE 2008



(c) ENADE 2011



(d) ENADE 2014



(e) ENADE 2017

Figura 11: Relação entre Idade e Desempenho.

forma linear com o passar dos anos.

Considerando a Situação de Trabalho, na Figura 12, aqueles que trabalham 40 horas ou mais configuram a maioria em todos os anos analisados. Embora essa maioria tenha decaído nos anos de 2014 e 2017. Concluintes sem trabalho são o segundo em proporção e também são aqueles que apresentam os melhores resultados, referente ao desempenho alto. Destacamos que menos horas de trabalho, implica num melhor desempenho alto. O inverso acontece com o desempenho baixo, pois quanto mais horas trabalhadas, maior a proporção de desempenho baixo.

Observando a Figura 13, temos os concluintes que não receberam nenhum tipo de bolsa acadêmica, como grande maioria (47%, 47%, 73% e 72%, respectivamente para cada ano da figura). Eles destacam-se negativamente como sendo aqueles que nos quatro

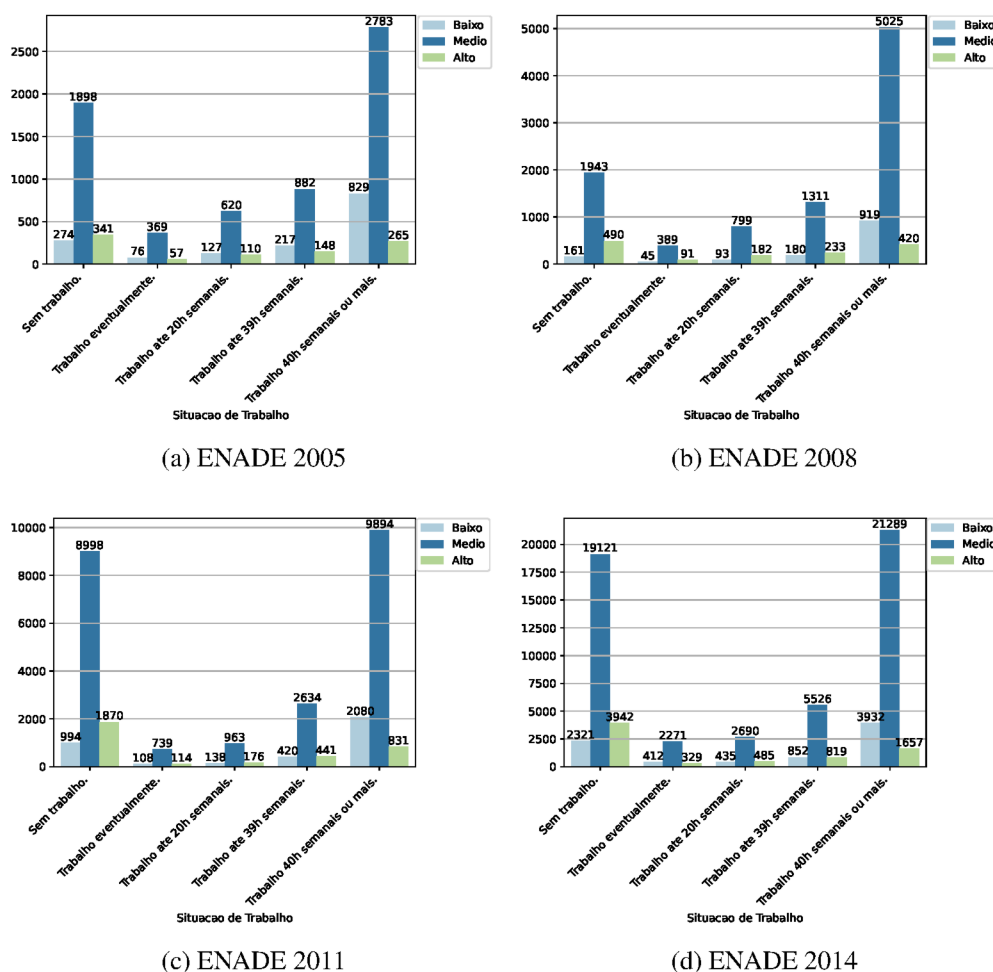


Figura 12: Relação da Situação de Trabalho e Desempenho.

anos apresentaram a maior taxa de desempenho baixo e a menor taxa de desempenho alto. Alunos com bolsa de Iniciação Científica ou Tecnológica apresentaram os melhores resultados nos anos de 2005 e 2011 (16% e 21% com desempenho alto), enquanto nos anos de 2014 e 2017 (24% e 18% com desempenho alto). Os alunos do PET (Programa de Educação Tutorial) são os alunos com os melhores aproveitamentos (29% e 20% de desempenho alto), mesmo sendo a minoria nos anos de 2014 e 2017.

Considerando a Escolarização da Mãe, na Figura 14, observamos uma minoria, em todos os anos, de alunos que a mãe não tem qualquer escolarização. Estes mesmos alunos são os que apresentam os piores desempenhos. Notamos que quanto menor a escolarização da mãe, maior a proporção de desempenho baixo. E quanto maior a escolarização da mãe, melhor o desempenho desse aluno, sendo menor a taxa de desempenho baixo e maior a taxa de desempenho alto.

Na Figura 15, temos uma distribuição quase igual, sendo 50,1%, 50,1%, e 51,2%, respectivamente, para cada ano, para concluintes que não tem na família alguém com curso superior. Quanto ao desempenho, percebemos que aqueles com família com curso

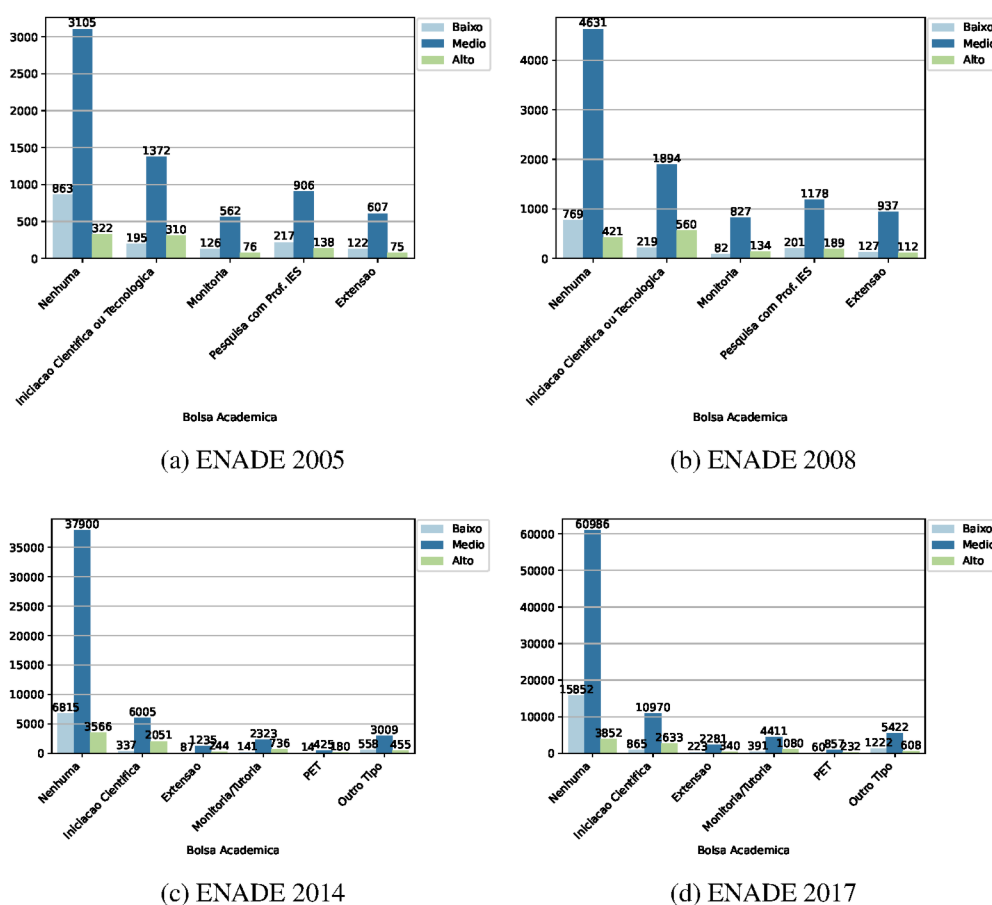


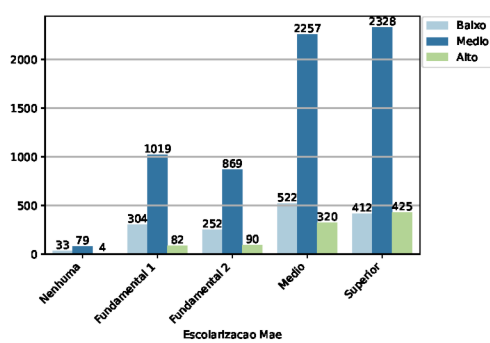
Figura 13: Relação entre Bolsa Acadêmica e Desempenho.

superior apresentam, nos três anos, o menor desempenho baixo e o maior desempenho alto.

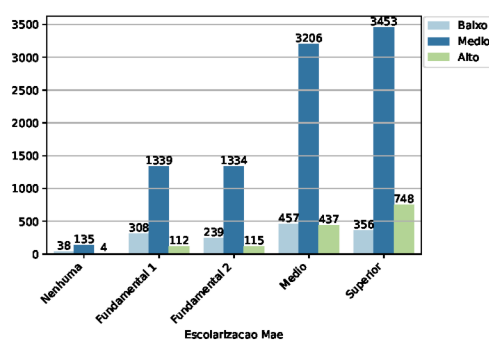
Analisando a Figura 16, observamos, em todos os anos, uma minoria de alunos que o pai não tem nenhuma escolarização, e estes são os que apresentam os piores desempenhos. Notamos que quanto menor a escolarização do pai, maior a proporção de desempenho baixo. Por outro lado, quanto maior a escolarização do pai, melhor o desempenho desse aluno, sendo menor a taxa de desempenho baixo e maior a taxa de desempenho alto. Estes mesmos resultados foram observados no nível de escolarização da mãe.

A maioria dos participantes não recebeu nenhum tipo de bolsa ou financiamento para cursar a graduação. A Figura 17 mostra os estudantes que realizaram o curso de forma gratuita, apresentando o melhor índice de desempenho alto (21%, 22%, e 16%, respectivamente para cada ano). Além de ficar entre os menores índices de desempenho baixo (5%, 5%, e 7%, respectivamente para cada ano). Ainda podemos destacar o ProUni Integral, que apresentou os melhores resultados dentre os tipos de bolsa.

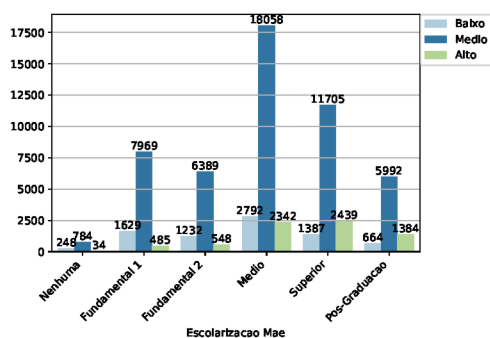
Na Figura 18, temos o tipo de escola que cursou o ensino médio. Em 2011, a maioria era proveniente de escola privada, enquanto que nos anos de 2014 e 2017, eram de escola pública. Destacamos negativamente os alunos que cursaram o ensino médio em escola



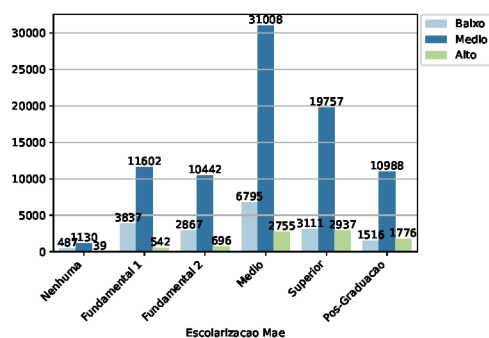
(a) ENADE 2005



(b) ENADE 2008

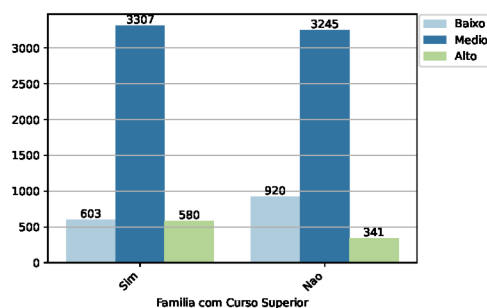


(c) ENADE 2014

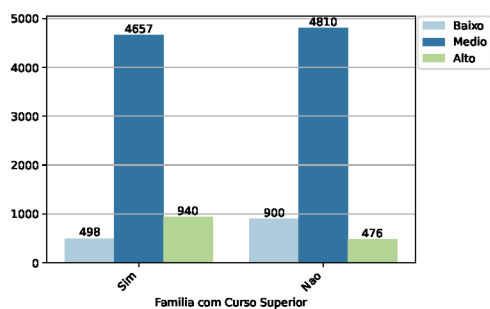


(d) ENADE 2017

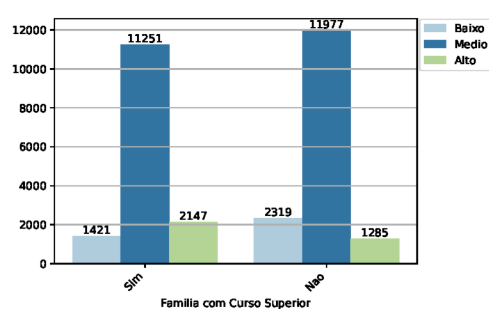
Figura 14: Relação entre Escolarização da Mãe e Desempenho.



(a) ENADE 2005

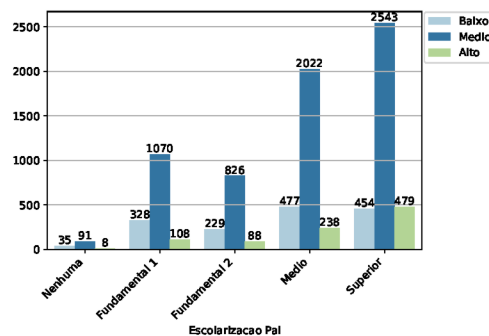


(b) ENADE 2008

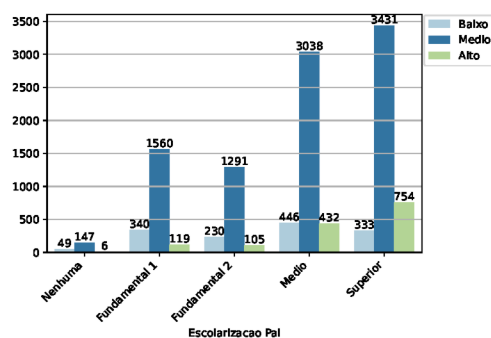


(c) ENADE 2011

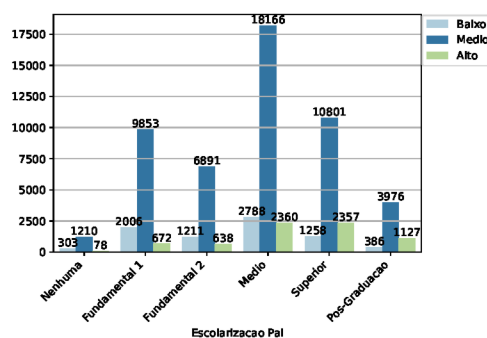
Figura 15: Relação entre Família com Curso Superior e Desempenho.



(a) ENADE 2005

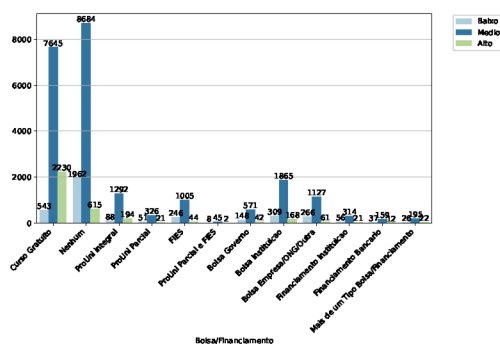


(b) ENADE 2008

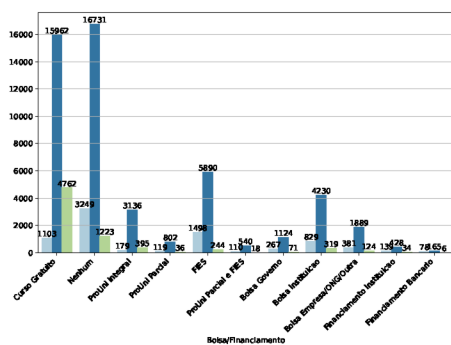


(c) ENADE 2014

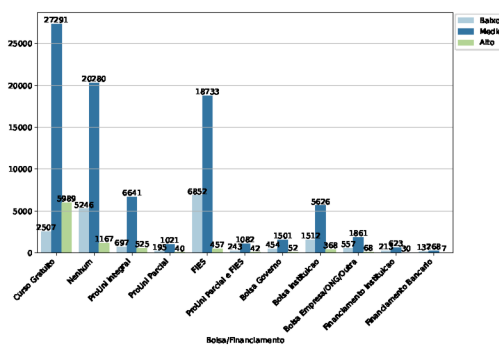
Figura 16: Relação entre Escolarização do Pai e Desempenho.



(a) ENADE 2011

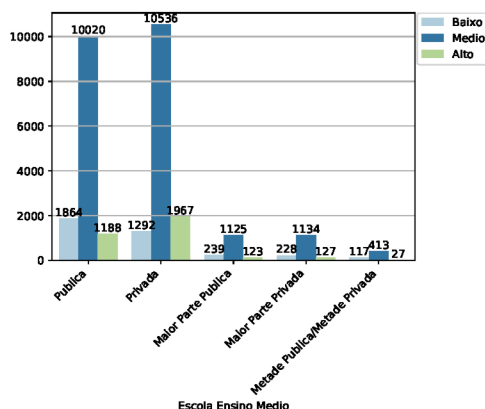


(b) ENADE 2014

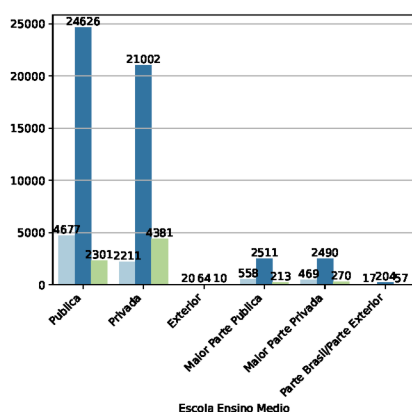


(c) ENADE 2017

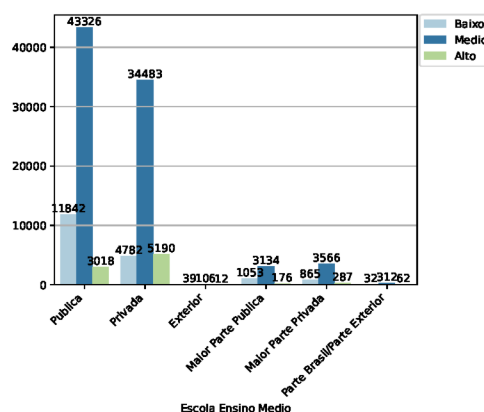
Figura 17: Relação entre Bolsa/Financiamento e Desempenho.



(a) ENADE 2011



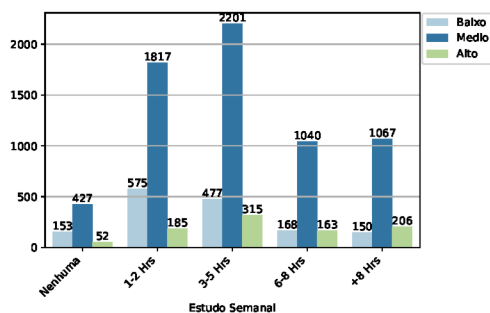
(b) ENADE 2014



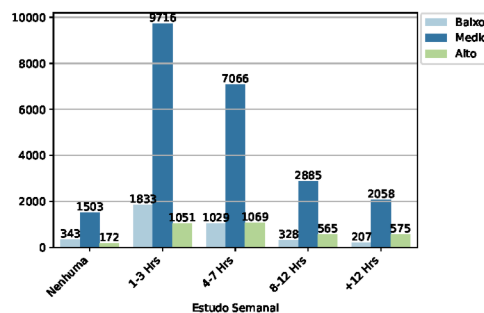
(c) ENADE 2017

Figura 18: Relação entre Escola Ensino Médio e Desempenho.

pública, mesmo que em parte, pois apresentaram os piores resultados. Aqueles que cursaram o ensino médio em escola privada, no geral, são alunos com melhor aproveitamento.



(a) ENADE 2005



(b) ENADE 2011

Figura 19: Relação entre Estudo Semanal e Desempenho.

Considerando o quantitativo de horas extra classe de estudo semanal, visto na Figura 19, percebemos que em 2005 a maioria estudou de 3 a 5 horas. Já em 2011, a maioria dos concluintes estudou entre 1 e 3 horas. Ainda observamos um aumento no número de alunos que estudaram 8 horas ou mais. Notamos que, em ambos os anos e conforme es-

perado, quanto maior o número de horas adicionais estudadas, maior o desempenho alto e menor o desempenho baixo.

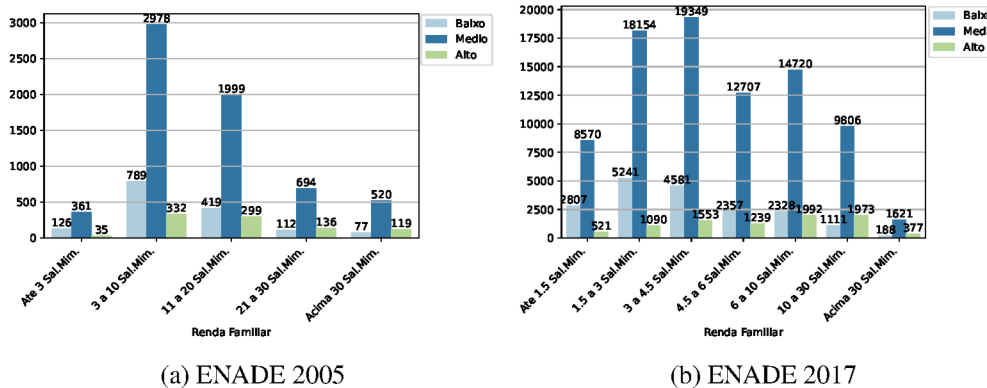


Figura 20: Relação da Renda Familiar e Desempenho.

Em relação a renda familiar, visualizada na Figura 20, observamos que a grande maioria de estudantes no ano de 2005 tiveram uma renda de 3 a 10 salários mínimos, representando 45% dos participantes. Em 2017, a maioria está entre 3 a 4,5 salários mínimos (22% dos participantes). Essa diferença entre os anos deu-se mais pela forma de perguntar, do que pelos dados em si. Se coletássemos os dados de 2011 e juntássemos aos de 2005, teríamos valores semelhantes. Outro fator que pode explicar isso seria o valor do salário mínimo que era muito baixo em 2005. Em 2017, verificou-se a diminuição em proporção de participantes com mais de 10 salários mínimos. Quanto ao desempenho, observamos um comportamento idêntico, pois em ambos os anos, quanto maior a renda, melhor foi o desempenho alto, e por outro lado, quanto menor a renda, maior o desempenho baixo.

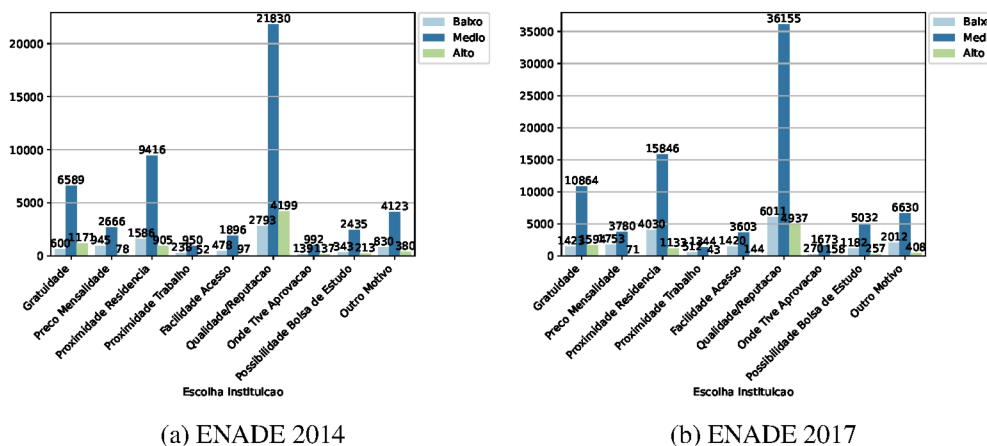


Figura 21: Relação entre Escolha da Instituição e Desempenho.

A maioria dos concluintes de engenharia escolhe o curso pela qualidade ou reputação, conforme a Figura 21. Depois, escolhe-se por proximidade da residência e em terceiro

lugar, por gratuidade. Em 2014 e 2017, podemos destacar em primeiro aqueles que escolheram pela gratuidade e em segundo pela qualidade ou reputação do curso, pois esses são os que apresentaram os melhores resultados. Negativamente, temos aqueles que escolheram pelo preço da mensalidade, apresentando o pior desempenho.

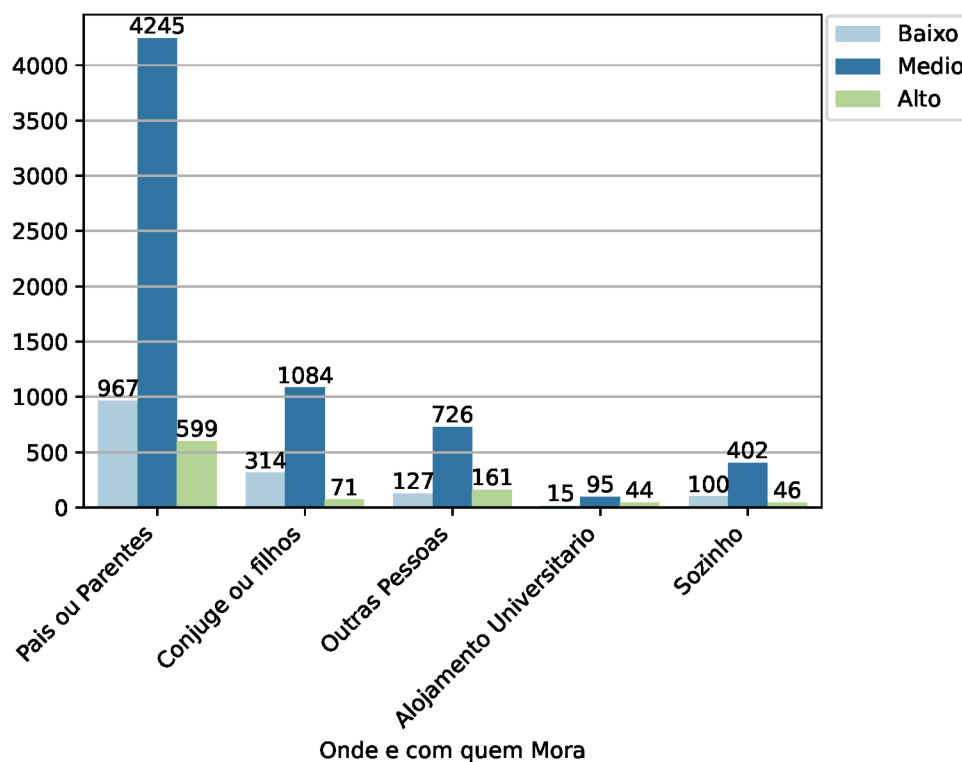


Figura 22: Relação entre Onde E Com Quem Mora e Desempenho. (ENADE 2005)

Na Figura 22, observa-se a grande maioria de participantes que moram com os pais ou parentes, tendo um bom desempenho alto. Os que moram com cônjuge ou filhos estão em segundo majoritariamente, mas se destacam negativamente por serem os que apresentam o maior desempenho baixo (21%) e o menor desempenho alto (5%). Aqueles que moram em alojamento universitário são minoria, mas apresentam os melhores resultados, pois apenas 10% apresentam desempenho baixo, e 28% possuem desempenho alto.

Considerando o estado civil, na Figura 23, observamos que a maioria dos concluintes de engenharia são solteiros, apresentando os melhores resultados (com desempenho alto de 13% e desempenho baixo de 10%). Os casados estão em segundo, em proporção de participantes, e destacam-se negativamente com o pior desempenho (desempenho alto de 5% e desempenho baixo de 17%), desconsiderando os viúvos por serem somente oito indivíduos.

Finalmente, analisando a Figura 24, verificamos que aqueles sem nenhuma bolsa permanência são os que apresentam o maior desempenho baixo (13%) e o segundo menor desempenho alto (9%), sendo eles a grande maioria de alunos. Em geral, os alunos com algum tipo de bolsa tendem a um desempenho melhor que aqueles sem bolsa. Como

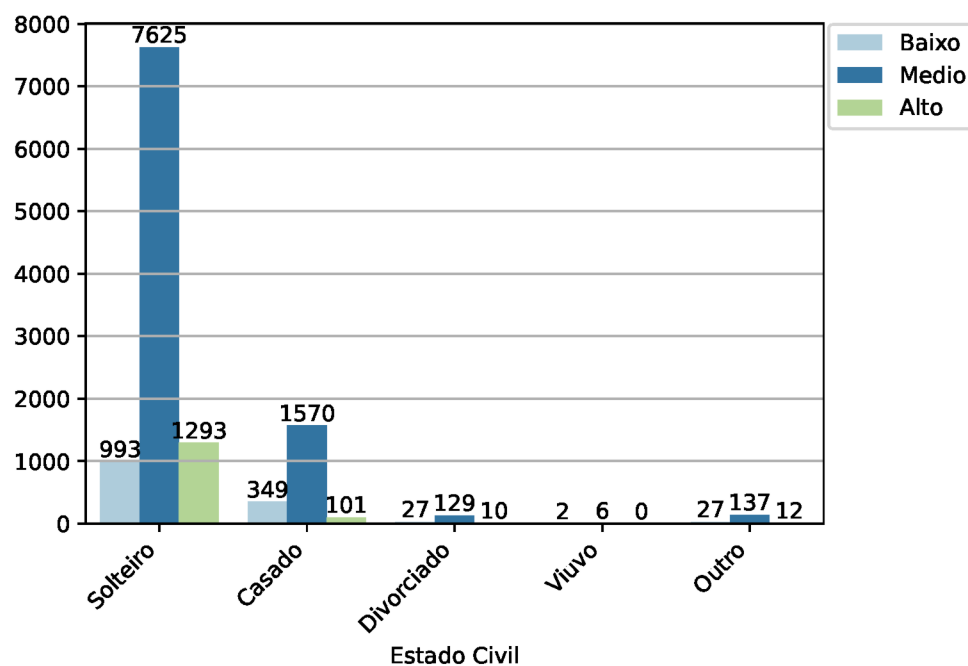


Figura 23: Relação entre Estado Civil e Desempenho. (ENADE 2008)

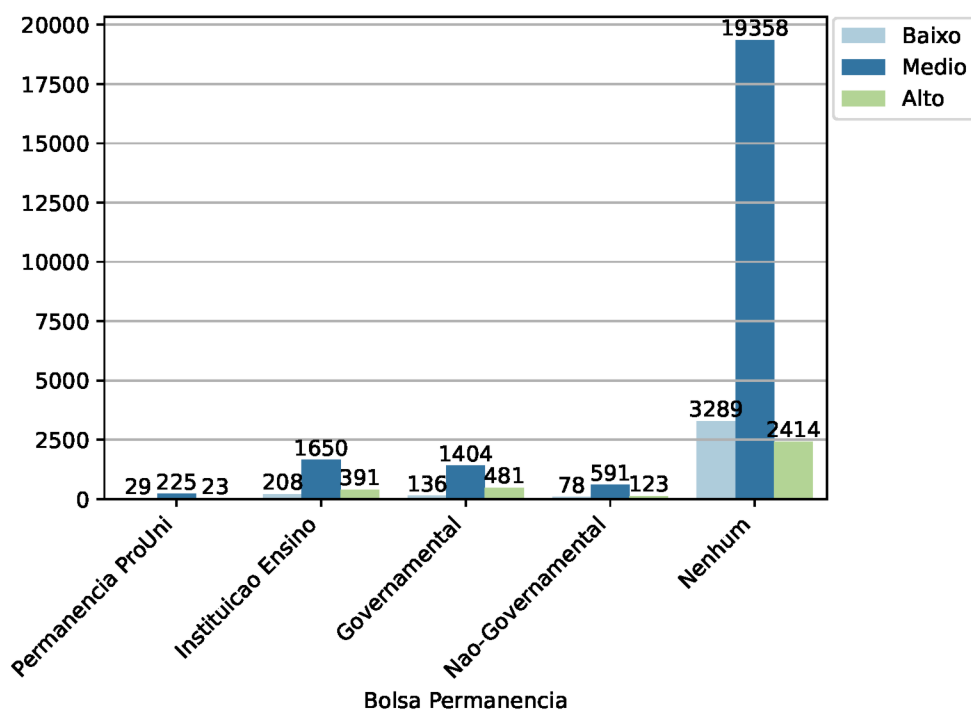


Figura 24: Relação entre Bolsa Permanência e Desempenho. (ENADE 2011)

destaque, podemos citar a bolsa governamental, que apresentou ótimo desempenho (com desempenho alto de 24% e desempenho baixo de 7%).

6 CONCLUSÃO

O objetivo deste trabalho foi explorar os dados do ENADE analisando como os aspectos pessoais e socioeconômicos são variáveis significativas na explicação da performance dos concluintes de graduação de engenharias, ao longo dos anos do ciclo avaliativo: 2005, 2008, 2011, 2014 e 2017. Cinco diferentes algoritmos para seleção de atributos foram aplicados, incluindo (dependendo do ano do exame) até 22 características pessoais e socioeconômicas pré-selecionadas. Ademais, aplicou-se quatro algoritmos de correlação de classificação, nas classificações dos resultados da seleção de atributos.

Primeiramente, ao longo dos anos de análise, notamos nos resultados de cada algoritmos de SA que alguns atributos destoam muito acima dos outros, mostrando sua maior importância. Outros atributos apresentaram uma linearidade nos resultados. Dependendo do ano ou algoritmo, também observamos um comportamento quase idêntico nas curvas dos gráficos, mostrando similaridade entre os anos ou resultados.

Analisando cada ano do ENADE, um total de 14 atributos diferentes foram considerados importantes, variando em cada ano do exame. Os atributos considerados relevantes em todos os anos (destacando o quantitativo de anos entre parênteses) foram: **Idade (5), Situação de Trabalho (4), Bolsa Acadêmica (4), Escolarização da Mãe (4), Família com Curso Superior (3), Escolarização do Pai (3), Bolsa/Financiamento (3), Escola Ensino Médio (3), Estudo Semanal (2), Renda Familiar (2), Escolha Instituição (2), Onde e Com Quem Mora (1), Estado Civil (1) e Bolsa Permanência (1).**

Dentre os resultados obtidos de cada atributo ao longo dos anos, observamos que os atributos Idade, Situação de Trabalho, Família com Curso Superior, Estudo Semanal, Onde e com quem Mora, Bolsa Permanência, Ação Afirmativa ou Inclusão Social e Modalidade Ensino Médio perderam importância ao longo dos anos do exame na avaliação das engenharias. Já os atributos Bolsa/Financiamento, Escola Ensino Médio, e Renda Familiar ganharam importância conforme o passar dos anos. Alguns atributos apresentaram bastante variação no decorrer dos anos analisados, enquanto outros mantiveram-se estáveis, com pouca variação no ranking de posições.

Analisando as características mais importantes apontadas pelos algoritmos de Seleção de Atributos, podemos concluir que o perfil do estudante de bom desempenho é jo-

vem, solteiro, não trabalha, mora com os pais, estuda mais horas extra classe, escolheu o curso por ser gratuito ou pela qualidade/reputação e recebeu algum tipo de bolsa (bolsa acadêmica ou bolsa permanência). A família ter curso superior, assim como a escolarização dos pais (pai e mãe) influenciou diretamente os resultados, pois quanto maior a escolarização, melhor o desempenho do concluinte. Renda familiar mostrou que tem relação direta e proporcional com o desempenho dos estudantes. Maior a renda implica melhora no índice de alto desempenho e menor é o índice de baixo desempenho. Alunos que cursaram o ensino médio em escola privada levam vantagem em relação a aqueles provenientes de escola pública. Estudantes que realizaram um curso gratuito tiveram melhor desempenho do que aqueles egressos de instituições privadas, evidenciando a qualidade das universidades públicas.

Por fim, os resultados apresentados nesse estudo, combinados com dados coletados de outras fontes, podem ser utilizados para melhorar a tomada de decisão de gestores da Educação em nível nacional. Por exemplo, garantir o financiamento para a atual demanda de projetos de ensino, pesquisa, extensão, desenvolvimento e inovação submetidos para as agências de fomento aumentará de forma significativa a oferta de quotas de bolsas acadêmicas nas instituições, as quais se mostraram essenciais no perfil dos estudantes com bom desempenho no ENADE. Outras ações importantes incluem direcionar a elaboração de políticas públicas que promovam a elevação da qualidade do ensino médio.

Para trabalhos futuros, a ideia é expandir a base de dados para mais anos do ENADE, analisar como os aspectos pessoais e socioeconômicos são variáveis significativas na explicação do desempenho do estudante ao longo do tempo, para outras áreas do conhecimento e cursos. Além de incluir comparações entre diferentes instituições de ensino superior, buscando entender como variáveis institucionais podem influenciar o desempenho dos estudantes. Pode-se ainda incluir métodos qualitativos mais aprofundados, como entrevistas com estudantes e professores.

REFERÊNCIAS

- [1] Araújo, R. A. et al. (2019). Análise dos microdados do Enade: Proposta de uma ferramenta de exploração utilizando mineração de dados. Dissertação de mestrado, Ciência da Computação, Universidade Federal de Goiás, Goiânia, Goiás.
- [2] Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.
- [3] Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining*, 1(1):3–17.
- [4] Barbosa, E., Cavalcanti, R., Mateus, J., and Ferreira, L. (2019). Identificação dos perfis de estudantes com nota elevada no ENADE. *RGCTI*, 3:17–29.
- [5] Bertolin, J. C. G., Fioreze, C., and Barão, F. R. (2022). Higher education and educational inequality in Brazil: Elitist heritage in a context of expanding access. *SciELO Preprints*.
- [6] Bramer, M. (2016). *Principles of data mining*, volume 180. Springer.
- [7] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall (Wadsworth and Inc.).
- [8] Capp, E. and Nienov, O. H. (2020). Bioestatística quantitativa aplicada. *lume*.
- [9] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.
- [10] Corder, G. W. and Foreman, D. I. (2011). *Nonparametric statistics for non-statisticians*. John Wiley & Sons, Inc.
- [11] Corso, G. and Resende, M. (2018). Mineração de Dados Aplicada ao Enade. Trabalho de conclusão de curso, Licenciatura em Ciência da Computação, Universidade de Brasília.

- [12] Costa, E., Baker, R. S., Amorim, L., Magalhães, J., and Marinho, T. (2013). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.
- [13] de Almeida, D. A., Almeida, S. P. N. d. C. e., and Amorim, M. M. T. (2021). Profile of graduates in mathematics degree: An analysis from Enade’s data (2005-2017). *SciELO Preprints*.
- [14] Downie, N. M. and Heath, R. W. (1974). Basic statistical methods. *Harper International Edition*, N.Y.
- [15] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience.
- [16] Eibe, F., Hall, M. A., and Witten, I. H. (2016). *The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques*. Morgan Kaufmann.
- [17] Faria, E. d. O. (2017). Uma análise dos fatores determinantes do desempenho dos alunos dos cursos superiores em Administração do Distrito Federal. Dissertação de mestrado, Administração, Universidade de Brasília, Brasília, DF.
- [18] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.
- [19] Fernandes, J. C. (2013). O mapa do apagão de engenheiros no Brasil. In *XLI Congresso Brasileiro de Educação em Engenharia*, volume 41, pages 1–9.
- [20] Fernando Choji, M., Diego N. Damasceno, C., Ibert Bittencourt, I., and Isotani, S. (2021). Mineração de dados do Enade de 2016 a 2018: uma análise sobre o município de Araçatuba/SP. *RENOTE*, 19(2):183–192.
- [21] Forthofer, R. N., Lehnen, R. G., Forthofer, R. N., and Lehnen, R. G. (1981). Rank correlation methods. *Public program analysis: a new categorical data approach*, pages 146–163.
- [22] Gondran, E., Lucca, G., Berri, R., Santos, H., and Borges, E. (2022). Analyzing the Determinant Characteristics for a Good Performance at ENADE Brazilian Exam Stratified by Teaching Modality: Face-to-face versus Online. In *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 234–242. INSTICC, SciTePress.
- [23] Gondran, E. L. (2021). Uma análise das características determinantes para o bom desempenho no ENADE estratificada por modalidade de ensino: EaD vs. Presencial. Trabalho de conclusão de curso, Universidade Federal do Rio Grande.

- [24] Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.
- [25] Goodman, L. A. and Kruskal, W. H. (1959). Measures of association for cross classifications. ii: Further discussion and references. *Journal of the American Statistical Association*, 54(285):123–163.
- [26] Goodman, L. A. and Kruskal, W. H. (1963). Measures of association for cross classifications iii: Approximate sampling theory. *Journal of the American Statistical Association*, 58(302):310–364.
- [27] Goodman, L. A. and Kruskal, W. H. (1972). Measures of association for cross classifications, iv: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67(338):415–421.
- [28] Gottardo, E., Kaestner, C., and Noronha, R. V. (2012). Avaliação de desempenho de estudantes em cursos de educação a distância utilizando mineração de dados. In *Anais do Workshop de Desafios da Computação Aplicada à Educação*, pages 30–39.
- [29] Gracioso, A. (2006). *Análise da eficácia escolar e do efeito-escola nos cursos de administração de empresas no Brasil*. PhD thesis, Fundação Getúlio Vargas.
- [30] Guimarães, P. R. B. (2017). Análise de correlação e medidas de associação. *Curitiba: Universidade Federal do Paraná*, pages 1–26.
- [31] Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.
- [32] Hall, M. A. and Smith, L. A. (1998). Practical feature subset selection for machine learning. In *Australasian Computer Science Conference ACSC'98*, volume 20, pages 181–191, Berlin. Springer.
- [33] Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12.
- [34] Inep (2020). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Acesso: 14 dezembro. 2022.
- [35] Jesus, J. K. L. d. (2018). Abordagens baseadas em teoria da informação para seleção automatizada de atributos. B.S. thesis, Universidade Federal do Rio Grande do Norte.
- [36] JÚNIOR, N. R. (2016). Método para otimização de janelas de tempo e discretização para classificação de movimentos futuros de ações da bolsa de valores. B.S. thesis, Pontifícia Universidade Católica do Paraná.

- [37] Karegowda, A. G., Manjunath, A., and Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277.
- [38] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
- [39] Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.
- [40] Kendall, M. G. (1948). Rank correlation methods. *Harvard Book*.
- [41] Kendall, M. G. (1955). Rank correlation methods: 1–18.
- [42] Kiefer, J. (1964). Maurice G. Kendall, Alan Stuart, The Advanced Theory of Statistics, Volume 2, “Inference and Relationship.”. *The Annals of Mathematical Statistics*, 35(3):1371–1380.
- [43] Kira, K. and Rendell, L. A. (1992). A Practical Approach to Feature Selection. In Sleeman, D. H. and Edwards, P., editors, *Ninth International Workshop on Machine Learning*, pages 249–256. Morgan Kaufmann.
- [44] Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, pages 171–182, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [45] Kretz, T., Bönisch, C., and Vortisch, P. (2010). Comparison of various methods for the calculation of the distance potential field. In *Pedestrian and evacuation dynamics 2008*, pages 335–346. Springer.
- [46] Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.
- [47] Laboratório de Gestão da Inovação - USP (2019). Em três anos, número de engenheiros formados já supera em 2,5 vezes demanda prevista para 2020. Acesso em: 14 de dezembro de 2023.
- [48] Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. PhD thesis, Universidade de São Paulo.
- [49] Lima, P. d. S. N., Ambrósio, A. P. L., Ferreira, D. J., and Brancher, J. D. (2019). Análise de dados do enade e enem: uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 24:89–107.

- [50] Lira, S. A. (2004). *Análise de correlação: abordagem teórica e de construção dos coeficientes com aplicações. 2004. 196f.* PhD thesis, Tese (Mestre em Ciências)-Setores de Ciências Exatas, Universidade Federal
- [51] Litchfield Jr, J. T. and Wilcoxon, F. (1955). Rank correlation method. *Analytical Chemistry*, 27(2):299–300.
- [52] Machado, L. d. S. and Francelino, W. L. (2020). Mineração de dados nos microdados Enade computação. *Ciência da Computação-Tubarão*.
- [53] Miao, J. and Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, 91:919–926.
- [54] Mitchell, T. (1997). *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill.
- [55] Morettin, P. A. and Bussab, W. O. (2017). *Estatística básica*. Saraiva Educação SA.
- [56] Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences. *Stata Journal*, 2(1):45–64(20).
- [57] Parmezan, A. R. S., Lee, H. D., Spolaôr, N., and Chung, W. F. (2012). Avaliação de métodos para seleção de atributos importantes para aprendizado de máquina supervisionado no processo de mineração de dados. Technical report, Universidade Estadual do Oeste do Paraná.
- [58] Paula, S. d. (2016). Um estudo exploratório sobre a criação de modelos de predição cruzada de defeitos apoiada por uma medida de correlação. B.S. thesis, Universidade Tecnológica Federal do Paraná.
- [59] Pereira, D. B., Plastino, A., Pereira, R. B., Zadrozny, B., Merschmann, L. H. d. C., and Freitas, A. A. (2011). Novas Medidas de Relevância para Seleção Lazy de Atributos. In *Congresso da Sociedade Brasileira de Computação*, volume 31.
- [60] Pornpanomchai, C., Supapattranon, C. K., and Siriwisesokul, N. (2011). Leaf and flower recognition system (e-botanist). *International Journal of Engineering and Technology*, 3(4):347.
- [61] Queiroga, R. M. (2005). *Uso de técnicas de data mining para detecção de fraudes em energia elétrica*. B.S. thesis, Universidade Federal do Espírito Santo.
- [62] REZENDE, V. O. and SANTOS, T. S. d. (2010). Geometria do táxi: A menor distância ente dois pontos nem sempre é como pensamos. In *Encontro de Produção Científica e Tecnológica*, volume 5.

- [63] Romero, C., Ventura, S., and García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384.
- [64] Rosa, E. R., Ferreira, D. J., Silva, N. F. F. d., and Assis, A. (2021). Estudo Exploratório através de Análises Longitudinais aplicado à Ciência da Computação a partir da Base de Dados do ENADE. *Revista Brasileira de Informática na Educação*, 29:1463–1486.
- [65] Saikumar, T., Yugander, P., Murthy, P. S., and Smitha, B. (2011). Colour based image segmentation using fuzzy c-means clustering. In *Proceedings of International Conference on Computer and Software Modeling, Singapore*, pages 180–5.
- [66] Salerno, M. S., Lins, L. M., Araújo, B. C., Gomes, L. A. V., Toledo, D., and Nascimento, P. A. M. M. (2014). Uma proposta de sistematização do debate sobre falta de engenheiros no Brasil. Technical report, Texto para Discussão.
- [67] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- [68] Siegel, S. and Castellan Jr, N. J. (2006). *Estatística não-paramétrica para ciências do comportamento*. Artmed Editora.
- [69] Smolentseva, A. (2018). *High Participation Systems of Higher Education*, chapter High Participation Society, pages 184–200. Oxford University Press.
- [70] Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American sociological review*, 27(6):799–811.
- [71] Souza, J. T. (2017). Métodos de seleção de atributos e análise de componentes principais: um estudo comparativo. Master’s thesis, Universidade Tecnológica Federal do Paraná.
- [72] Tan, P.-N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.
- [73] Tasca, M. (2015). *Contribuições ao Problema de Seleção de Atributos*. Tese de doutorado, Computação, Universidade Federal Fluminense, Niterói, Rio de Janeiro.
- [74] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203.
- [75] Vieira, A., Bertolini, D., and Schwerz, A. (2022). Análise do Desempenho no Enade dos Concluintes de Computação usando Técnica de Agrupamento. In *Anais do XXXIII*

Simpósio Brasileiro de Informática na Educação, pages 834–845, Porto Alegre, RS, Brasil. SBC.

- [76] Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 856–863. AAAI Press.

A ANO E ÁREA DE APLICAÇÃO DO ENADE

Ano I

- Cursos de bacharelado nas áreas de conhecimento de Ciências Agrárias, Ciências da Saúde e áreas afins;
- Cursos de bacharelado nas áreas de conhecimento de Engenharias e Arquitetura e Urbanismo.
- Cursos Superiores de Tecnologia nas áreas de Ambiente e Saúde, Produção Alimentícia, Recursos Naturais, Militar e Segurança.

Ano II

- Cursos de bacharelado nas áreas de conhecimento de Ciências Biológicas; Ciências Exatas e da Terra; Linguística, Letras e Artes e áreas afins.
- Cursos de licenciatura nas áreas de conhecimento de Ciências da Saúde; Ciências Humanas; Ciências Biológicas; Ciências Exatas e da Terra; Linguística, Letras e Artes;
- Cursos de bacharelado nas áreas de conhecimento de Ciências Humanas e Ciências da Saúde, com cursos avaliados no âmbito das licenciaturas.
- Cursos Superiores de Tecnologia nas áreas de Controle e Processos Industriais, Informação e Comunicação, Infraestrutura e Produção Industrial.

Ano III

- Cursos de bacharelado nas Áreas de Conhecimento Ciências Sociais Aplicadas e áreas afins;
- Cursos de bacharelado nas Áreas de Conhecimento Ciências Humanas e áreas afins que não tenham cursos também avaliados no âmbito das licenciaturas.
- Cursos Superiores de Tecnologia nas áreas de Gestão e Negócios, Apoio Escolar, Hospitalidade e Lazer, Produção Cultural e Design.