Universidade Estadual de Londrina

RONAN ANACLETO LOPES

ESTUDO DA MODELAGEM DO DESEMPENHO DISCENTE BASEADO NA AVALIAÇÃO CURRICULAR

Londrina
2017

RONAN ANACLETO LOPES

ESTUDO DA MODELAGEM DO DESEMPENHO DISCENTE BASEADO NA AVALIAÇÃO CURRICULAR

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Jacques Duílio Brancher.

Londrina
2017

Ronan Anacleto Lopes
Estudo da modelagem do desempenho discente baseado na avaliação curricular/ Ronan Anacleto Lopes. – Londrina–PR, 2017-
69 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Jacques Duílio Brancher

– Universidade Estadual de Londrina, 2017.

1. Curriculum. 2. Produção Científica. 3. Lattes. 4. Mestrado. 5. Mineração de Dados. I. Prof. Dr. Jacques Duílio Brancher. II. Universidade Estadual de Londrina. III. Mestrado em Ciências da Computação. IV. Estudo dos parâmetros de análise curricular baseado em Mineração de Dados com ênfase no desempenho discente

CDU 02:141:005.7

RONAN ANACLETO LOPES

ESTUDO DA MODELAGEM DO DESEMPENHO DISCENTE BASEADO NA AVALIAÇÃO CURRICULAR

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

Orientador: Prof. Dr. Jacques Duílio Brancher
Universidade Estadual de Londrina - UEL

Prof. Dr. Rodolfo Miranda de Barros
Universidade Estadual de Londrina - UEL

Prof. Dr. Pedro Paulo da Silva Ayrosa
Universidade Estadual de Londrina - UEL

Prof. Dr. Benjamin Luiz Franklin
Universidade Estadual de Londrina - UEL

Prof. Dr. Osvaldo Luiz de Oliveira
Faculdade Campo Limpo Paulista - FACCAMP

Londrina, 28 de julho de 2017.

LOPES, R. A.. **Estudo da modelagem do desempenho discente baseado na avaliação curricular**. 69 p. Dissertação de Mestrado (Mestrado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina–PR, 2017.

RESUMO

Em uma sociedade cada vez mais tecnológica, o papel dos pesquisadores tem ganho destaque. Da mesma forma, os cursos de Mestrado, responsáveis pela formação destes pesquisadores, almejam a melhoria e qualidade deste ensino. A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior realiza avaliações periódicas dos programas de pós-graduação, visando mensurar a qualidade desses cursos. Nesta avaliação, um dos elementos que se leva em conta é a produção científica dos discentes e docentes, portanto a produtibilidade destes reflete na nota atribuída ao curso. Nesta situação, as Universidades tendem a buscar discentes com alta produtibilidade científica, almejando melhores notas na avaliação. As instituições realizam processos de seleção almejando identificar discentes com o perfil, considerado pela instituição, de bom desempenho. Um dos meios avaliativos aplicados aos candidatos é a análise curricular, porém esta avaliação é realizada com base em pressupostos e experiências passadas dos avaliadores, o que pode tornar o processo tendencioso. Para solucionar esta problemática, é proposto neste trabalho um estudo das características curriculares e sua relação com a produtibilidade dos candidatos. Este estudo utilizou técnicas KDD para desenvolver um modelo classificatório capaz de predizer a produtibilidade dos candidatos. O desenvolvimento do KDD utilizou-se das técnicas de seleção de atributos *Qui-quadrado* e *Coeficiente de Gini* e, para a elaboração do modelo de regras, utilizou-se o algoritmo *Random Forest*. Este modelo e os processo que levaram a sua construção forneceram informações úteis ao processo de avaliação curricular. Por meio deste estudo, concluiu-se que é possível aplicar um modelo classificador da produtibilidade, todavia os dados utilizados não foram suficientes para gerar um modelo estatisticamente relevante, uma vez que o problema apresentou-se de alta complexidade, sendo necessário novos estudos.

Palavras-chave: Curriculum. Produção Científica. Lattes. Mestrado. Mineração de Dados.

LOPES, R. A.. **Study of modeling of student performance based on curriculum evaluation**. 69 p. Master's Thesis (Master in Science in Computer Science) – State University of Londrina, Londrina–PR, 2017.

ABSTRACT

In an increasingly technological society, the role of researchers has gained prominence. In the same way, the Master's courses, responsible for the training of these researchers, aim improving the quality of this teaching. The Coordination for the Improvement of Higher Education Personnel conducts periodic evaluations of the graduate programs, aiming to measure the quality of these courses. In this evaluation, one of the elements that is taken into account is the scientific production of the students and teachers, therefore the produtibility of these reflects in the note attributed to the course. In this situation, universities tend to seek students with high scientific productivity, aiming for better grades in the assessment. The Institutions carry out selection processes to identify students with the profile, considered by this institution, of good performance. One of the evaluation measures applied to candidates is curricular analysis, but this evaluation is carried out based on the assumptions and past experiences of the evaluators, which can make the process tendentious. To solve this problem, a study of the curricular characteristics and their relation with the produtibility of the candidates is proposed in this work. This study used KDD techniques to develop a classificatory model capable of predicting the producibility of the candidates. The development of KDD was based on the Chi-square and Gini Coefficient features selection techniques, and for the elaboration of the rules model it was used the algorithm Random Forest. This model and the processes that led to its construction provided useful information to the curriculum evaluation process. It was concluded that it is possible to apply a classifier model of the producibility, however the data used were not enough to generate a statistically relevant model, since the problem was highly complex and new studies were necessary.

Keywords: Curriculum. Scientific Production. Lattes. Master degree. Data Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e problemática

A sociedade contemporânea apoia sua evolução na ciência e tecnologia, sendo os pesquisadores, figuras de suma importância para o desenvolvimento da humanidade. Cabe aos cursos de pós-graduação *stricto-sensu* a responsabilidade de formar pesquisadores capazes de contribuir com a ciência.

No Brasil, existem atualmente 109 cursos de pós-graduação *stricto-sensu* na área da Computação, registrados junto a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[5]. Esta desempenha um papel fundamental na avaliação, promoção e fomento de cursos de pós-graduação, bem como na divulgação de produções científicas, por meio do Portal de Periódicos.

Para reconhecer e regular a qualidade dos cursos de Mestrado e Doutorado no Brasil, a CAPES executa periodicamente, uma avaliação para atribuir notas às instituições de ensino [6, 5]. De acordo com o Diretor Geral de Avaliação da CAPES, Renato Janine Ribeiro [6], o resultado da avaliação é usado tanto como base para a concessão de fomento, quanto como métrica para a continuidade ou fechamento dos cursos.

Um importante parâmetro avaliado pela CAPES é a contribuição científica do programa, a qual se baseia na produção e publicação de artigos em congressos e periódicos [5]. De acordo com as especificações da avaliação quadrienal executada pela CAPES, espera-se o destaque de $4 * N$ produções mais importantes, em que N é o número de professores ativos do programa. Estas produções são utilizadas como parâmetro da qualidade e distribuição das publicações do programa.

De acordo com o Documento de Área de Ciências da Computação [5], a produção intelectual avaliada possui peso de 40% da nota na avaliação de programas de Mestrado Acadêmico e Doutorado, enquanto os outros 60% são distribuídos entre: Corpo docente, com 20%; Corpo discente, teses e dissertações, representando 30% da nota; e Inserção social, com 10% de representatividade.

Dada essas proporções, percebe-se que a produção científica, inclusive as dos discentes tem impacto na avaliação do programa de Mestrado pela CAPES e consequentemente, nos benefícios financeiros que a instituição de ensino poderá receber [6].

Como resultado, as Universidades são estimuladas a compor seu corpo discente baseado na possibilidade de produção científica de seus candidatos. Portanto, o presente trabalho utilizará como métrica de *desempenho acadêmico*, a produção científica dos discentes, visando uma maior especificação do escopo estudado.

Nestes termos, compreende-se que o processo de seleção discente para programas de Mestrado e Doutorado tem impacto direto sob avaliação realizada pela CAPES, uma vez que este procedimento deve garantir o acesso ao programa somente aos candidatos com maior possibilidade de desempenho. Comumente a seleção é realizada de forma manual, através de análise curricular, entrevistas e avaliações teóricas, variando em cada instituição.

Este processo é realizado por um ou mais avaliadores humanos, os quais são influenciados por experiências passadas e conhecimentos intrínsecos. Dentre os meios de avaliação, a análise curricular fornece informações acadêmicas e profissionais nas quais os avaliadores buscam o perfil almejado pela universidade.

Ao avaliar os currículos, buscando o perfil dito apropriado para os objetivos da Universidade, busca-se selecionar os candidatos com maior probabilidade de sucesso no curso. Todavia, devido a influência que os critérios curriculares avaliados recebem dos preceitos do avaliador, questiona-se: Esta avaliação é realmente capaz de refletir a capacidade do candidato quanto a seu desempenho?

Para responder a este questionamento, o presente trabalho visa aplicar técnicas estatísticas de análise de dados conhecida na área da Ciência de Dados como Descoberta de Conhecimento em Base de dados ou KDD (*Knowledge Descovery in Database*).

1.2 Proposta do trabalho

O foco deste trabalho consiste na análise das características curriculares e sua relação com o desempenho acadêmico dos discentes. Através do desenvolvimento de um modelo classificatório da produção discente, busca-se informações que possam auxiliar na tomada de decisão quanto a avaliação curricular realizada nos processos de seleção em programas de Mestrado. Para tanto, busca-se os seguintes objetivos:

- **Objetivo Geral:** Realizar um estudo da modelagem da produtibilidade discente com base em dados curriculares, produzindo informações úteis à tomada de decisão do processo de avaliação curricular.
- **Objetivo Específico:**
  - Extrair dados curriculares de arquivos XML;
  - Aplicar técnicas de KDD com base nos atributos curriculares de mestres na área da Computação;
  - Obter resultados de fácil interpretação humana, através de tabelas, gráficos e estatísticas.

Por meio da conclusão destes objetivos, espera-se obter resultados capazes de contribuir para o meio científico. Destacam-se as principais contribuições como sendo:

- Demonstrar a relevância dos atributos curriculares no processo de previsão da produção intelectual discente.
- Apresentar uma metodologia de modelagem do desempenho discente;
- Destacar novos conhecimentos acerca do processo de avaliação curricular;

Para a obtenção destas contribuições, o presente trabalho está organizado como segue: Os capítulos 2 e 3 apresentam o referencial teórico do trabalho, quanto aos estudos de desempenho acadêmico e análises KDD; A Metodologia é apresentada no Capítulo 4; Em seguida, 5 apresenta os resultados obtidos e discute a relevância destes; Por fim são apresentadas as Conclusões do trabalho no capítulo 6.

2 PREVISÃO DE DESEMPENHO ACADÊMICO

O presente capítulo apresenta uma revisão literária acerca da previsão do desempenho acadêmico. Por meio desta revisão, são demonstrados as características do desempenho acadêmico e os reflexos de seu estudo às instituições de ensino. Também são apresentados artigos que utilizam técnicas KDD na modelagem deste desempenho.

2.1 Características do desempenho acadêmico

Para compreender o que é o desempenho acadêmico bem como suas características, é importante compreender o conceito de *desempenho* e o que ele significa na sociedade moderna. De acordo com Barbosa [7], o desempenho era uma métrica associada unicamente à produtibilidade, considerando bom desempenho, aqueles que produziam mais.

Esta visão era estritamente empresarial e industrial, onde o lucro proveniente da alta produção era tudo o que importava, fazendo com que a avaliação do desempenho fosse uma medida punitiva e controladora [7].

Todavia a partir da década de vinte, a forma como a avaliação do desempenho era visto e utilizada passou a ser alterada. Já por volta dos anos 80 e 90, a avaliação de desempenho perde o caráter punitivo e passa a ser um termômetro das necessidades para as organizações e seus recursos humanos [7].

Sob esta perspectiva, o desempenho passa a ser associado a uma série de fatores, como nível de aptidão e métodos de seleção utilizados. Sua avaliação passa a ser utilizada não só para identificar recursos humanos pouco produtivos, mas também para avaliar os próprios critérios de seleção, ensino ou treinamento [7]. O desempenho tornou-se uma importante métrica na melhoria da qualidade e identificação de falhas.

Magalhães e Andrade [8] descrevem que o desempenho em âmbito acadêmico, parte de um conjunto de critérios estabelecidos pela instituição com base no pressuposto perfil que esta almejou formar. Para estes autores, o desempenho é reflete o nível de habilidades alcançadas nas atividades avaliadas.

O método de avaliação investigado por estes autores, o vestibular, não apresenta relação com o desempenho dos alunos e recomendam que estudos sejam realizados, incluindo novas variáveis como características socioeconômicas [8].

Baseado no desempenho de um discente, também é possível realizar avaliação das instituições de ensino, em todos os níveis educacionais, quanto a qualidade do ensino empregado, os métodos de seleção discentes e outros fatores. A análise do desempenho pode considerar, além do desempenho acadêmico, características como barreiras sociológicas,

psicológicas e econômicas do sujeito avaliado [9].

Dutka [9] realiza um estudo visando os fatores socioeconômicos e outras variáveis externas que possam afetar o desempenho acadêmico. A avaliação do desempenho pode então ser utilizado como critério para bolsas, levando em consideração uma gama maior de informações.

Dutka ainda aponta a importância do estudo do desempenho para a seleção de candidatos, uma vez que o número de concorrentes capazes é maior do que o número de vagas das universidades, sendo necessário aplicar métodos apropriados para alocar os recursos educacionais de forma apropriada [9].

O desempenho acadêmico é portanto uma variável de grande relevância na compreensão e melhoria do processo educacional, todavia é também uma métrica complexa, composta por atributos sociais, culturais, e até mesmo emocionais. No trabalho de Pekrun [10] é abordado a influência do humor no desempenho acadêmico de adolescentes.

Constatou-se que existe uma influência real do humor na absorção de conhecimento e o reflexo deste no desempenho discente. Ao final concluíram que conservar boas emoções no ambiente de aprendizado melhora o rendimento estudantil [10].

Além de fatores psicológicos, o desempenho acadêmico é uma variável que pode ser influenciada por fatores biológicos. Heissel e Norris [11] pesquisaram a influência da incidência de sol, bem como os fatores hormonais, como a puberdade, nos adolescentes. Essa investigação verificou o desempenho dos alunos relacionando-o com o horário de início das aulas. Constatou-se que a alteração no início das aulas, atrasando-a em uma hora, melhorou o desempenho dos alunos em Matemática e Leitura.

Com base nos artigos analisados, percebeu-se que o desempenho acadêmico é composto por uma série de características, tais como: uma métrica de avaliação tanto do discente como de todo o ambiente escolar; elemento influenciado por fatores intrínsecos e extrínsecos do estudante, influenciado por elementos sociais, psicológicos e biológicos; auxilia na tomada de decisão para melhoria dos processos de ensino.

Tais fatores demonstram a relevância que os estudos do desempenho acadêmico tem. Porém a complexidade desta característica é seu maior desafio pois pode sofrer influência de todos os tipos de fatores e situações. Neste trabalho busca-se analisar o desempenho acadêmico sob o escopo específico da produtibilidade acadêmica.

Para realizar esta análise, optou-se por restringir o escopo de atributos avaliados. O presente trabalho busca estudar os dados utilizados no processo de avaliação curricular. Espera-se que os resultados obtidos com esta análise tragam reflexos positivos ao processo de seleção discente para programas de Mestrado.

2.2 Reflexos do estudo do desempenho acadêmico

A análise do desempenho acadêmico, suas características e reflexos na vida acadêmica, é essencial para que a instituição, os docentes e discentes possam tomar decisões acertadas quanto a melhoria do processo de ensino-aprendizagem. A aplicação de técnicas estatísticas, modelos preditivos e regressivos e outros meios de análise de dados, tem papel fundamental na melhoria dos processos da área acadêmica.

Chamillard [12], realizou o estudo do desempenho acadêmico de alunos de Ciências da Computação da Academia da Força Aérea Americana. Neste trabalho, foram elaborados modelos estatísticos preditivos de desempenho que possibilitam aos docentes identificar as matérias de maior relevância para o desempenho dos discentes, bem como as dificuldades desses, fornecendo base para uma melhor avaliação e implementação da grade curricular do curso.

A análise pessoal de cada aluno também pode fornecer informações relevantes ao desempenho acadêmico deste. Paireekreng e Prexawanprasut [13] estudaram os diferentes estilos de estudo dos discentes da Universidade de Bangkok. Através da classificação dos estilos dos alunos, elaborou-se um modelo de previsão do estilo de estudo de novos alunos. Com base neste modelo, foi possível a adaptação e recomendação de programas de estudos aos alunos, melhorando assim seu desempenho acadêmico.

Ainda que existam ferramentas que permitam identificar o provável desempenho de um aluno, é importante manter sua qualidade durante a sequência do curso. Lopez Guarin *et al* [14] analisam os dados acadêmicos dos 4 primeiros semestres de um curso de graduação. Constatou-se ao final que o método tradicional de aquisição de informação dos discentes não é suficiente para avaliar o desempenho acadêmico no decorrer do curso, sendo necessárias novas formas de manter este controle.

Observada a grande quantidade de informações necessárias na elaboração dos modelos preditivos, bem como a falta de padronização nelas, Kurniawan e Halim [15], elaboraram um estudo aplicando técnicas de mineração de dados e *Data Warehousing* nas informações obtidas. Com esta informação, a instituição toma decisões que auxiliem na prevenção da reprova dos discentes, melhorando a qualidade do ensino e conservando os recursos institucionais.

A previsão do desempenho acadêmico é uma ferramenta útil inclusive para a seleção dos discentes. Os trabalhos de Oliveira e Garcia [16] e Magalhães e Andrade [8], almejaram a identificação de características que permita a previsão do desempenho acadêmico com base nas informações colhidas ainda nos vestibulares.

Outras esferas de ensino também se valem das análises de desempenho. O processo de seleção de candidatos para Mestrado visa o ingresso dos discentes com o perfil que, segundo a instituição, obterá o melhor desempenho. Lamadrid-Figueroa *et al* [17] abordam

a análise do processo de seleção do programa de Mestrado, no qual obteve como resultado um modelo preditivo capaz de identificar possíveis atrasos na conclusão do curso.

Este atraso reflete as dificuldades do discente quanto a sua área de pesquisa, seja no desenvolvimento ou estudo desta. Uma forma de melhorar a qualidade das produções do estudante e desta forma, seu desempenho acadêmico, é garantir que sua linha de pesquisa esteja alinhada com suas afinidades de conhecimento.

Nesta situação, Ktona *et al* [18] apresentam um estudo do desempenho acadêmico da graduação dos alunos e seu reflexo nas diferentes áreas de Mestrado e linhas de pesquisa. Utilizando técnicas de classificação e agrupamento de dados, obteve-se cinco grupos de características e suas respectivas áreas de afinidade. Estes dados permitem orientar os graduandos aos cursos com maior probabilidade de sucesso, o que beneficia tanto o discente quanto a instituição e o programa de Mestrado como um todo.

2.3 Aplicação do KDD na previsão do desempenho acadêmico

A análise do desempenho acadêmico, as características que a influenciam e os benefícios que trazem à instituição, docente e discente tornou atrativa esta área de estudo. As literaturas relacionadas ao tema utilizam-se de diferentes metodologias aplicadas ao processo de predição do desempenho acadêmico. Um dos métodos em ascensão é a aplicação de técnicas de KDD.

El-Halees [19] em seu trabalho, aplica técnicas de KDD em uma base de dados de um curso EAD (Ensino à Distância). Com base nestas regras, o autor destaca a importância do KDD para a melhoria da tomada de decisão da instituição, que pode ser pro-ativa às dificuldades dos discentes, reduzindo taxas de reprova e influenciando positivamente na aprendizagem.

Dando continuidade ao trabalho de El-Halees [19], os autores Tair e El-Halees [20] aplicaram as as técnicas investigadas no trabalho anterior em uma base de dados histórica de alunos de graduação. Neste trabalho, os autores visaram corroborar o trabalho anterior, demonstrando que em diferentes bases e ambientes educacionais, presencial ou EAD, é possível aplicar técnicas de KDD para previsão e melhoria de desempenho acadêmico.

Devasia *et al* [21] propuseram o desenvolvimento de um sistema baseado na mineração de dados de uma base educacional, contendo o registro de 700 estudantes, 19 diferentes atributos na análise. Com este sistema, os autores foram capazes de prever o desempenho dos discentes e desta forma, permitindo que a instituição tomasse ações que melhoraram o desempenho geral de seus alunos.

Em seu trabalho, Ogor [22] também desenvolve um sistema que utiliza os conceitos de KDD voltado ao monitoramento do desempenho dos estudantes de graduação. Este

sistema foi capaz de predizer o desempenho acadêmico com aproximadamente 94% de sucesso, gerando relatórios que auxiliaram as instituições nas tomadas de decisão.

Com base nestes trabalhos, observa-se uma tendência quanto a sua aplicação no acompanhamento do desempenho acadêmico. Todavia, este desempenho pode ser utilizado também como métrica para outros fatores. A literatura apresenta trabalhos que utilizam o desempenho acadêmico como um meio, para justificar ações que melhorem o ensino, como redução na evasão ou reprova escolar.

No trabalho desenvolvido por Baradwaj e Pal [23], por exemplo, foram utilizados procedimentos de KDD para a detecção regras de desempenho de alunos de graduação. Por meio destas novas informações, foi possível auxiliar a identificação de alunos que, devido a chance de baixo desempenho, necessitam de atenção especial por parte dos docentes e instituição, favorecendo ações que reduzem a taxa de reprova.

O estudo do desempenho acadêmico pode ocorrer inclusive antes de seu ingresso, através de técnicas preditivas aplicadas no processo de seleção. Em programas de Mestrado, são utilizados como formas de seleção de candidatos: avaliação curricular, entrevistas, avaliações práticas e outros métodos.

As instituições de ensino são soberanas para escolher seus métodos de avaliação, porém é interessante que estes métodos favoreçam os candidatos com maior chance de sucesso no programa, ou seja, que tenham um bom desempenho de acordo com o esperado pela Universidade.

Um exemplo de estudo que aborda esta temática é o trabalho de Lamadrid-Figueroa *et al* [17]. Por meio de técnicas de regressão linear, analisaram 10 critérios utilizados no processo de admissão em um programa de Mestrado da Universidade do México. Nesta análise obteve-se correlações entre os critérios admissionais e o desempenho acadêmico, quanto ao sucesso nas disciplinas e conclusão do curso.

Nesta análise, os autores almejam melhorar o processo de seleção dos candidatos, determinando quais características avaliativas detém a maior correlação com a possibilidade de sucesso acadêmico dos avaliados. Seus resultados demonstram que a metodologia de seleção embasada em evidências científicas contribui para a seleção de melhores candidatos, de acordo com o esperado pela Universidade [17].

Outro exemplo é o trabalho de Bush [24], o qual utilizou-se de dados relacionados ao processo de seleção e de informações socioeconômicas e acadêmicas anteriores ao ingresso dos estudantes ao programa de Mestrado. O autor estudou os relacionamentos entre a idade, os critérios de admissão no programa de Mestrado, *status* socioeconômico e informações do ensino médico/técnico do candidato, comparando estes elementos com o desempenho acadêmico no Mestrado.

Os resultados deste trabalho apresentam evidências de que as conquistas acadê-

micas, ou seja, seu desempenho em atividades acadêmicas passadas, é um fator de real influência no futuro desempenho acadêmico durante o Mestrado. Estas novas informações permitem uma interpretação cientificamente fundamentada do processo de seleção, baseando seus os resultados na relevância de cada atributo para o possível desempenho do candidato [24].

Ao fim, observa-se que a aplicação da técnica de KDD voltado a análise e previsão do desempenho acadêmico gera bons resultados, permitindo que os cursos e instituições se beneficiem destes novos conhecimentos, melhorando seu processo de ensino. Este fator torna o processo de KDD atraente para ser utilizado no presente trabalho.

3 FUNDAMENTOS TEÓRICOS DO KDD

A aplicação dos processos de KDD dependem da utilização de algoritmos e formulas estatísticas, principalmente nas etapas de seleção de atributos e mineração de dados. As seções seguintes contém o embasamento teórico capaz de auxiliar na compreensão futura dos métodos utilizados. Nelas são apresentados os fundamentos do KDD, bem como funções e algoritmos aplicados nos processos de Pré-processamento e Mineração de Dados.

3.1 KDD

Fayyad [1] apresenta o KDD como um processo, ou seja, um conjunto de etapas que envolve a preparação dos dados, a busca por padrões de informação, avaliação do conhecimento obtido e por fim, os refinamentos necessários. Os procedimentos de forma geral são apresentados na Figura 1.

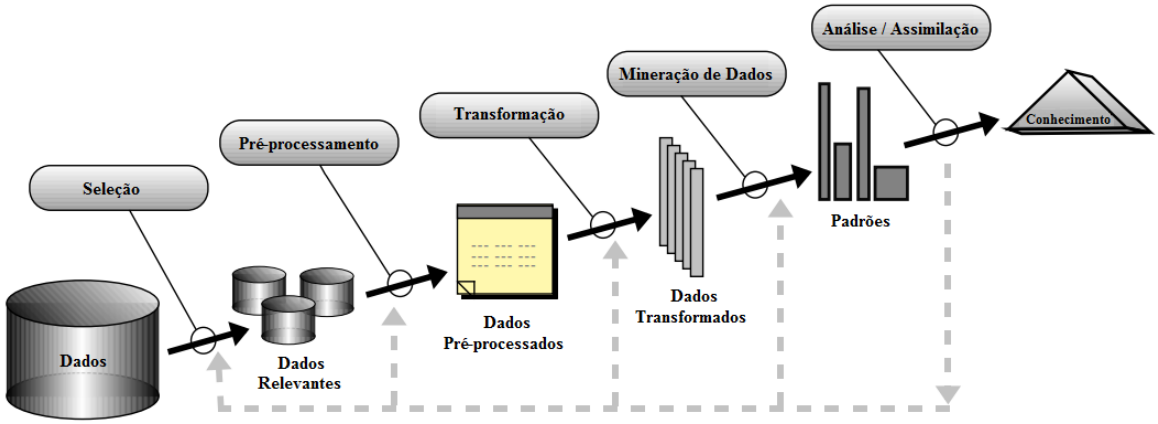

Figura 1 – Etapas do processo KDD [1].

As etapas apresentadas na Figura 1 são essenciais ao KDD. A *Seleção* é responsável pela obtenção dos atributos relevantes à análise. Por meio dos processos de *Pré-processamento* é realizado a limpeza dos dados, bem como sua organização em um conjunto de dados único. As *Transformações* nos dados os preparam, ajustando seus formatos e padrões de dados, para a aplicação dos algoritmos.

Na etapa de *Mineração de dados*, é realizada a análise com base nos algoritmos de extração de conhecimento. Seus resultados são apresentados na etapa de *Análise* ou *Pós-processamento*, gerando por fim uma nova gama de informações que, ao ser aplicado no processo de tomada de decisão, torna-se conhecimento.

Estas etapas podem ser compreendidas em 3 macro-etapas apresentadas por Garcia *et al* [25]: Pre-processamento; Mineração de dados; e Pós-processamento . Este trabalho

compreende que as definições de Garcia e Fayyad são complementares e se dispõe da seguinte forma:

- **Pré-processamento:** Compreende a seleção, limpeza, transformação e integração dos dados, portanto, todas as atividades que precedem o processamento dos dados pelos algoritmos de Mineração.
- **Mineração de dados:** Aplicação das técnicas de mineração de dados a base pré-processada.
- **Pós-processamento:** Etapa de interpretação dos modelos de regras detectados, bem como sua aplicação no processo de tomada de decisão.

Neste sentido, o KDD é um conjunto de técnicas e procedimentos que permitem ao analista de dados descobrir novos conhecimentos, dando sentido aos dados e auxiliando o processo de tomada de decisão. Estas características garantem ao KDD a presença nas mais variadas áreas de atuação, desde *marketing* à detecção de fraude [1].

Devido ás características do KDD, observa-se que sua aplicação se alinha com os objetivos deste trabalho. Através da etapa de Mineração de Dados é possível fornecer modelos que, ao serem analisados por um ser humano, podem resultar em novos conhecimentos relevantes [26, 27]. A escolha da aplicação do KDD também foi embasada na estrutura dos dados curriculares. Devido a sua amplitude e ruídos, é necessário aplicar etapas como a limpeza e seleção dos dados, etapas que fazem parte do Pré-processamento do KDD [2].

3.2 Diagrama de Caixa

O Diagrama de Caixa (*Box Plot*) é uma das técnicas utilizadas para analisar a distribuição dos dados, frequentemente utilizada no processo de detecção dos valores que excedem a variação padrão de um atributo, chamados de *outliers* [28].

Os *outliers* podem ser causados por problemas na aquisição dos dados, tais como defeitos em sensores, erro humano e dados inconsistentes/inválidos, porém também permite indicar elementos que se destacam entre os registros, como uma tentativa de invasão a uma rede ou um caso de inadimplência [26].

Este diagrama fornece uma perspectiva visual da distribuição dos dados em torno dos possíveis valores que um atributo pode assumir, portanto é um diagrama voltado a valores numéricos contínuos. Para que seja possível desenhar este diagrama, são necessários cálculos estatísticos que permitam determinar seus limiares [28].

Inicialmente deve-se calcular os pontos em que se divide a base de dados em proporções iguais, esses pontos são chamados de *Quantis*. Ao dividir a base em quatro *quantis*, a cada parte dá-se o nome de *Quartil*, e ao dividir a base em cem partes, chama-se *Percentil* [2]. Um exemplo da divisão em Quartis pode ser visto na Figura 2.

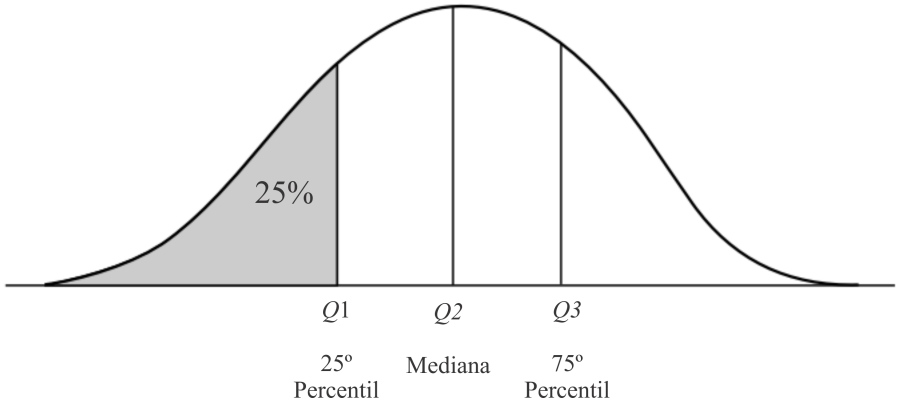

Figura 2 – Exemplo de divisão dos dados em *Quartis*. Adaptado de Han *et al* [2].

Como ilustra a Figura 2, o ponto de divisão primeiro quartil *Q1* é o equivalente ao 25º percentil, ou seja, 25% dos dados da base. O valor de *Q2* é o equivalente a mediana, portanto é o valor do atributo que dividi igualmente a base em duas partes, ou 50%. Por fim o valor de *Q3* separa a base em 75% dos registros menores que ele e 25% de registros maiores ou iguais [2].

Após obtido estes valores, é possível calcular o **Intervalo Interquartil** (IIQ), que representa uma simples medida de propagação dos dados entre os dados medianos. O calculo do IIQ é dado por: $IIQ = Q_3 - Q_1$

De acordo com esta equação do IIQ é a diferença entre o valor do terceiro quartil e o primeiro quartil. Todavia, uma medida numérica única de propagação é valida na aplicação em dados cuja a distribuição não é simétrica [2]. Portanto utiliza-se os valores encontrados nos *Quartis* e do *IIQ* para determinar a distribuição dos dados em um Diagrama de Caixa [3].

Para desenhar o diagrama, toma-se o ponto *Q2*, como ponto médio do gráfico, desenhando um símbolo para demarca-lo. Em seguida, realiza o mesmo processo com os pontos *Q1* e *Q3*, formando um retângulo em torno destes pontos. Dentro deste retângulo estão 50% dos dados da análise. Esse desenho de **caixa** dá nome ao diagrama [3]. Um exemplo pode ser visto na Figura 3.

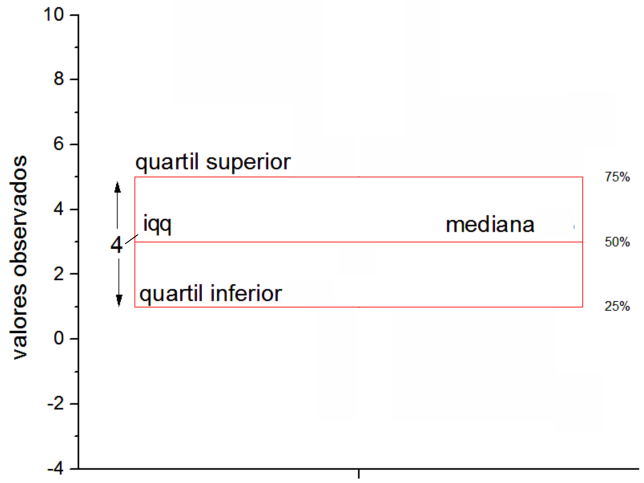

Figura 3 – Exemplo de Diagrama de Caixa. Adaptado de *e-Handbook of Statistical Methods* [3].

A Figura 3 apresenta um exemplo de um Diagrama de Caixa destacando os pontos do quartil superior $Q3$ e inferior $Q1$. Para detectar os *outliers*, utiliza-se um ponto de corte chamado Estimativa de Whisker, que determina um limiar onde os valores que extrapolem $c * IIQ$ (onde c é uma constante) além dos quartis externos, considera-se estes valores como *outliers* [2].

Para calcular o Limite Inferior (LI) e o Limite Superior (LS), utiliza-se as fórmulas: $LI = Q1 - c * IIQ$ e $LS = Q3 + c * IIQ$, sendo $c = 1,5$. Um exemplo mais completo do diagrama, contendo os pontos de Whisker é visto na Figura 4.

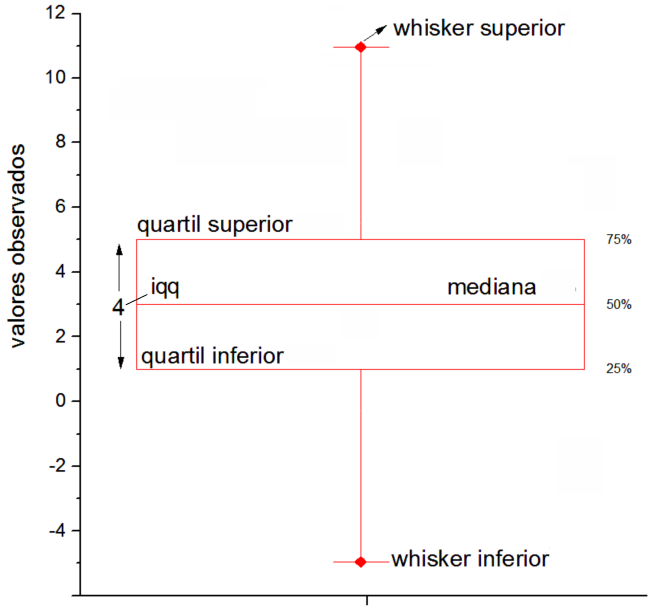

Figura 4 – Exemplo de Diagrama de Caixa com limiar da Estimativa de Whisker. Adaptado de *e-Handbook of Statistical Methods* [3].

Dado o diagrama da Figura 4, nota-se que os pontos anotados como *whisker superior* e *inferior* são mais próximos dos quartis externos do que a linha traçada no diagrama da Figura 3. Neste trabalho, os registros cujo valores estejam além deste limiar serão considerados inconsistentes e portanto, serão eliminados da análise.

3.3 Qui-quadrado

O Qui-quadrado (χ^2) é uma função estatística de teste de hipóteses, investigada pela primeira vez por Pearson em 1900, portanto também chamada de Qui-quadrado de Pearson. Esta função estatística permite realizar testes de independência entre variáveis, verificando a existência de dependência entre dois atributos, x e y, sendo y o atributo principal da análise [29].

De acordo com Vireira [30], a *hipótese* em que se considera as variáveis independentes é determinada por H_0, também chamada de *hipótese da nulidade*. Já a *hipótese* contrária, que se refere a dependência de uma variável com outra é indicada por H_1, chamada de *hipótese alternativa*. Para identificar a validade de uma destas hipóteses, calcula-se portanto o valor χ^2, dado pela seguinte fórmula:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \tag{3.1}$$

Onde r é a soma dos rótulos possíveis dos atributos x e y analisados, O_i são os valores de frequência observadas e E_i é frequência esperada de cada rótulo [30, 31]. Para obter os valores de O_i utiliza-se os dados da amostra pesquisada. Para a análise de dois atributos, cada um com dois possíveis rótulos, obtêm-se uma tabela 2x2 como pode ser vista na Tabela 1.

Tabela 1 – Tabela de Contingência.

Rótulos de x	Rótulos de y		Total
	y_1	y_2	
x_1	A_{11}	A_{12}	$A_{11} + A_{12}$
x_2	A_{21}	A_{22}	$A_{21} + A_{22}$
Total	$A_{11} + A_{21}$	$A_{12} + A_{22}$	$\sum A_{mn}$

Conforme a Tabela 1 apresenta, A_{mn} são as quantidades de registros que contém os rótulos correspondentes a linha m e coluna n da Tabela, também observa-se que tem-se $r = 4$ e $i = 1, ..., r$. Portanto a frequência observada $O_i = A_{mn}$ e utilizando esta tabela,

têm-se as frequências esperadas E_i calculadas com base na seguinte fórmula:

$$E_i = \frac{Total_{x_n} * Total_{y_m}}{Total_{xy_{mn}}} \tag{3.2}$$

De acordo com a fórmula, E_i conterá a distribuição que se espera de cada combinação de rótulos de x e y considerando-os independentes e aleatórios. Com estes dados é possível calcular o valor de χ^2. Para rejeitar ou não a *Hipótese de nulidade*, é necessário primeiro definir dois elementos da análise, o nível de significância α e o grau de liberdade g.

Os valores comumente atribuídos para α são 1%, 5% e 10%, para esta análise será utilizado o valor de 5% por ser o valor recomendado em [30]. Utilizando em conjunto com os graus de liberdade g, dado pelo seguinte cálculo $g = (m-1) * (n-1)$, tem-se um valor estimado de χ^2, chamado de Qui-quadrado de referência (chi^2_R). Com base nestes resultados, é possível interpretar os resultados com base em regras apresentadas na Tabela 2:

Tabela 2 – Condições de interpretação do Qui-quadrado.

Condição	**Resultado**
$\chi^2 < \chi^2_\alpha$	H_0
$\chi^2 \geq \chi^2_\alpha$	$NegaH_0$

De acordo com a Tabela 2, valores calculados de χ^2 menores que o valor de referência χ^2_α mantém-se a *Hipótese nula* (H_0) como possível, enquanto valores maiores ou iguais ao valor de referência, permite negar a existência desta hipótese, permitindo considerar que os dados obedecem a *Hipótese alternativa* (H_1) [30]. Para facilitar a análise, os principais valores de χ^2_α são apresentados na seguinte Tabela:

Tabela 3 – Valores de referência para χ^2_α.

g	**10%**	**5%**	**1%**
1	2,71	3,84	6,64
2	4,60	5,99	9,21
3	6,25	7,82	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09

A Tabela 3 demonstra os valores de referência χ^2_α para os níveis de significância α de 10%, 5% e 1% correlacionado com os graus de liberdade g de cada resultado [30]. Com base na comparação dos resultados calculados com a tabela de referência, é possível

determinar se existe dependência entre os atributos e H_0 é rejeitada, ou se a relação entre os dois atributos é independente e aleatória, corroborando com H_0.

Esta função é utilizada principalmente em análises que lidem com um grande número de atributos. Um exemplo de sua utilização ocorre no trabalho de Jin *et al* [31], em que foi comparado o desempenho de 5 classificadores quanto a sua capacidade de identificação de câncer cerebral e de mama, utilizando uma base com dados de genes das células.

As bases genéticas contém um número gigantesco de atributos, sendo necessário aplicar a função *Qui-quadrado* para selecionar aqueles atributos que apresentavam maior dependência com o problema abordado. Ao final este trabalho destaca que com o Qui-quadrado foi possível reduzir a base para classificação de câncer e diferencia-los de amostras saudáveis [31].

Por fim é preciso ter em mente, contudo, que: (1) na rejeição da hipótese nula implica que existe dependência entre duas variáveis; (2) na aceitação da hipótese nula, nada se pode afirmar sobre a dependência ou não entre as variáveis.

3.4 Árvore de Decisão

Os algoritmos de Mineração de Dados buscam adquirir conhecimento com base nos dados que lhe são apresentados. Dentre os diferentes tipos de algoritmos, existem aqueles baseados em Aprendizado de Máquina.

Baseado na experiência acumulada de soluções bem sucedidas, algoritmos de Aprendizagem de Máquina são capazes de tomar decisões. A capacidade de tomar decisões com base em conhecimentos adquiridos de suas experiências e análise de dados é chamada de aprendizado. Dentre os diferentes algoritmos desta categoria estão as Árvores de Decisão [32].

De acordo com Safavian e Landgrebe [33], Árvores de Decisão (AD) possuem a capacidade de transformar um complexo processo de tomada de decisão em um conjunto de regras. Todavia, uma estrutura em árvore possui nomenclaturas próprias de seus elementos, sendo elas[33]:

1. Um grafo $G = (V, E)$ é um conjunto finito de *vértices* ou *nós* (V) e *arestas* ou *galhos* (E). Grafos cujas arestas são apresentadas em pares ordenados (v, w) são chamados de Grafos diretos;

2. Um *caminho* é uma sequencia de arestas que indica a direção de um vértice v_1 para v_n.

3. Uma *árvore enraizada orientada* é definida como um grafo acíclico direto que obedece a certas regras:

    - Possui um vértice que é o ponto de partida de arestas mas não recebe nenhuma, chamado *raiz*.
    - Todo vértice, exceto o *raiz*, tem exatamente uma aresta de entrada.
    - Existe um *caminho* único entre a raiz e cada um dos vértices.

4. Se (v, w) é uma aresta em uma árvore, é dito que v é o vértice *pai* e w o filho.

5. Vértices que não possuem *filhos* são chamados de vértices *terminais* ou *folhas*. Todos os outros vértices (com exceção da *raiz*) são chamados de vértices *internos*.

Observada a constituição de uma árvore, Han, Kamber e Pei [2] explicam que os vértices (ou nós) denotam um teste condicional em um dos atributos da base, deste nó, cada aresta ou galho que dele deriva apresenta uma das possibilidades de resposta ao teste.

Vértices folhas contém uma possível resposta ao problema analisado. Esta resposta compreende em um possível rótulo da classe principal analisada, permitindo assim que seja realizado a classificação de novos registros.

O processo de criação de uma AD pode ser supervisionado ou não-supervisionado. Este trabalho foca na utilização de AD supervisionada, o que significa que dado um algoritmo, ou indutor, recebe um conjunto de dados de treino onde os rótulos da classe principal são conhecidos. Com base nos exemplos e contra-exemplos, é criado um modelo classificador, que busca determinar qual rótulo se encaixa em um novo registros, baseado em seu conjunto de características [32].

Além do conjunto de exemplos para *treinamento*, o algoritmo é submetido a um conjunto de exemplos de *teste*. Este segundo conjunto é disjunto quanto aos dados de treinamento e é utilizado para medir o grau de efetividade do conhecimento aprendido. O conjunto de *teste* é aplicado ao modelo gerado e compara-se os resultados do classificador com os rótulos reais de cada exemplo, formando assim dados estatísticos que possam validar o algoritmo [32].

Após a aplicação da base de *teste* os resultados de um classificador são entendidos como: Positivos Verdadeiros (PV); Negativos Verdadeiros (NV); Positivos Falsos(PF); Negativos Falsos (NF). Cada um destes representa um tipo de informação classificada [2].

- **PV**: Número de registros *positivos* cuja a classe principal foi corretamente classificado;
- **NV**: Número de registros *negativos* corretamente classificados;

- **PF**: Número de registros *negativos* incorretamente classificados como *positivos*;
- **NF**: Número de registros *positivos* incorretamente classificados como *negativos*;

Neste contexto, observa-se que registros *positivos* são aqueles em que o valor classificado é o mesmo do valor observado na base de teste. Valores *negativos* são os valores diferentes do valor observado [2]. A Tabela 4 demonstra a organização destes valores no que se chama Matriz de Confusão:

Tabela 4 – Matriz de confusão.

	Classificado		
Observado	*Positivo*	*Negativo*	**Total**
Positivo	PV	NF	P
Negativo	PF	NV	N
Total	P'	N'	P+N

Observa-se na Tabela 4 que o elemento P indica o Total de *Positivos* observado na tabela de teste, enquanto o N represente os valores *Negativos* desta. Os elementos P' e N' simbolizam respectivamente os *Positivos* e *Negativos* classificados.

A Tabela 4 demonstra a forma como os resultados de um classificador são compreendidos após a fase de teste. Com base nesta matriz, é possível aplicar funções de avaliação, permitindo aferir itens como a acurácia, sensibilidade e especificidade do classificador. A Tabela 5 apresenta as fórmulas utilizadas nestas métricas.

Tabela 5 – Fórmulas das métricas avaliativas de um classificador [2].

Métrica	**Fórmula**
Acurácia	$\frac{PV+NV}{P+N}$
Sensibilidade	$\frac{PV}{P}$
Especificidade	$\frac{NV}{N}$

A **Acurácia** determina a taxa de classificações corretas. A **Sensibilidade** apresenta a taxa de classificação correta dos *Positivos* (P') dentre todas os registros *Positivos* (P) da base. Por fim a **Especificidade** representa a taxa de classificação dos *Negativos* (N') quanto aos registros *Negativos* (N). Ainda que existam outras métricas, este trabalho se apoiará somente nestas.

3.4.1 Random Forest

Em uma perspectiva geral, o *Random Forest* é um classificador que utiliza o método de *Bootstrapping Aggregation* ou simplesmente *bagging*. Este método consiste em gerar

múltiplas versões de um classificador [34]. Por definição, "é um classificador que consiste em um conjunto de classificadores estruturados em árvore"[35].

Uma *Floresta* é o conjunto das AD geradas por este algoritmo [2]. Ao testar este algoritmo, cada árvore gera um voto no rótulo da classe principal que classificou, e aquele com maior número de votos é escolhido como rótulo classificado [36, 4, 34]. Este algoritmo deve garantir que as árvores sejam independentes e igualmente distribuídas.

O processo de *bagging* é fundamental para este algoritmo, pois uma vez que se fornece uma base de treino S para gerar o modelo de classificação, este método é responsável por criar subamostras de treino s com base em S, sendo $s < S$ [36]. Essas subamostras são geradas por um sorteio simples com reposição, gerando assim amostras aleatórias e independentes entre si [36, 34]. Para cada s é gerado uma nova AD A [34, 4].

Outra importante característica deste algoritmo é a aleatoriedade na seleção dos atributos para a construção das árvores. Dado um conjunto de atributos E, o algoritmo seleciona $e < E$ atributos aleatoriamente para cada subamostra s [36, 34]. Portanto cada AD da *Floresta* é gerada com base em seu conjunto de atributos e e base de treino s próprios, tornando-as independentes entre si [36, 35].

Por fim, durante o processo de teste das árvores construídas ao receber um novo registro, aplica-o a toda a floresta e cada árvore vota no rótulo da classe objetivo que foi capaz de classificar, ao final seleciona-se o rótulo com maior número de votos [35]. O algoritmo *Random Forest* pode ser compreendido em três etapas principais [36]:

- **Extração de amostra:** Ao receber a base de treino S, o algoritmo executa o processo de *bagging* gerando s_n subamostras, onde n é o número de árvores que serão geradas;

- **Geração da AD:** A cada árvore, seleciona-se um vetor e_n de atributos do conjunto total de atributos E, de forma aleatória. Aplica-se técnicas de seleção de atributos e constrói-se a árvore com base na respectiva subamostra s_n;

- **Predição:** No processo de classificação/predição de novos registros, não considera-se o tamanho da árvore, mas sim o voto que cada uma realiza, ou seja, a classe que cada árvore foi capaz de apontar como rótulo correto. Aquele com maior número de previsões será apresentado como classificação final.

A ordem de execução destes processos pode ser visto mais claramente através da Figura 5.

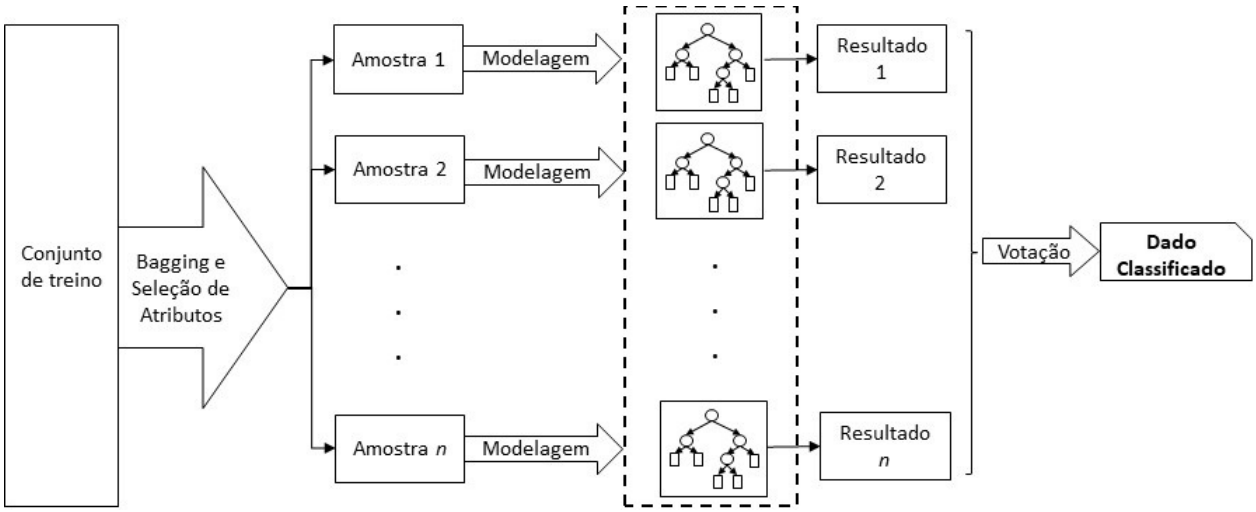

Figura 5 – Sequência de processos do *Random Forest*. Adaptado de [4].

A Figura 5 demonstra a sequência dos processos que devem ser realizados na execução de um algoritmo *Random Forest*. Observa-se que a modelagem das árvores pode variar para cada implementação, porém sempre são utilizados classificadores em estrutura de árvore.

Estes procedimentos tornam o *Random Forest* um algoritmo poderoso, com boa acurácia, relativamente robusto a ruídos e *outliers*, simples e fácil de ser executado de forma paralela, além de fornecer outras métricas implícitas como a importância dos atributos, correlação e erros estimados [35].

Como citado anteriormente, uma função interessante deste algoritmo é a medida de *importância da variável*. Esta métrica permite que o analista de dados tenha uma medida da relevância de cada atributo para a construção das árvores [37]. Esta métrica pode ser utilizada de forma recursiva para realizar uma nova seleção de atributos e melhorar a acurácia do algoritmo.

A literatura apresenta diversos trabalhos que demonstram as vantagens deste algoritmo e suas diferentes utilizações. Um exemplo prático da eficácia deste algoritmo é encontrada em [2], que comparou o desempenho da RF com o algoritmo AdaBoost, demonstrando que o primeiro apresenta acurácia comparável ou superior ao segundo, porém o RF é mais robusto quanto a erros e *outliers*.

Fernandez-Delgado *et al* [38] também realizaram um estudo comparativo entre classificadores, abordando 179 algoritmos de 17 famílias. Aplicaram a estes, 121 diferentes conjuntos de dados. A classificação dos algoritmos utilizou principalmente os critérios de acurácia média e máxima. Neste trabalho foram empregados diferentes variações dos algoritmos clássicos, sendo o *Parallel Random Forest* aquele que apresentou melhores resultados.

Utilizando a capacidade deste algoritmo de determinar a importância das variáveis, Han *et al* [39] realizaram um estudo sobre a aplicação das métricas disponíveis neste algoritmo para detecção da importância. Estas métricas, além de se mostrarem eficientes na seleção de atributos, demonstraram uma boa performance quando comparados a outras funções de seleção de atributos.

Comprovada sua eficiência, este algoritmo também apresenta casos de sucesso em aplicações no mundo real. Um exemplo de sua utilização ocorre no trabalho de Jin *et al* [4], em que é aplicado o *Random Forest* para garantir a segurança de uma rede em tempo real. Os resultados deste algoritmo foram comparados aos de outros classificadores e constatado que seu desempenho foi superior.

Devido a suas características e relevância na literatura, o *Random Forest* demonstra ser um algoritmo eficiente e poderoso, sendo selecionado como classificador deste trabalho, visando obter uma boa acurácia na classificação, bem como dados referente a importância das variáveis utilizadas.

3.4.2 Coeficiente de Gini

O Coeficiente de Gini é uma medida estatística de desigualdade, também chamado de medida de impureza. Esta distribuição é frequentemente vista em análises de renda, porém é compatível com qualquer tipo de distribuição [40]. Devido a sua capacidade de mensurar as impureza dos atributos, esta função pode ser utilizada como método para seleção de atributos, inclusive nos algoritmos de árvores de decisão como o CART (*Classification and Regression Trees*) [41, 40, 34] e o próprio *Random Forest* [34, 42, 43].

Algoritmos de Árvore de Decisão utilizam este coeficiente para identificar os atributos que permitem uma melhor separabilidade, ou seja, que sejam capazes de gerar nós mais "puros"o maior número de um único rótulo, sendo a situação ideal um nó que apresente somente um dos possíveis rótulos da classe principal [34]. O Coeficiente de Gini tem seu valor máximo 1 (pior) quando os registros são igualmente distribuídos nos rótulos possíveis da classe principal e tem seu valor mínimo 0 (melhor) quando todos os registros pertençam a somente um rótulo da classe principal [40].

O processo de seleção de atributos utilizando esta função se baseia na variação que ocorre no valor de Gini entre o nó pai e seus filhos, buscando a combinação que permita a maior redução média do valor desta função, chamado de Média do Decremento de Gini (MDG) [42]. Assim, o atributo que apresenta a maior MDG é aquele que melhor aproxima os dados de uma classificação correta, ou seja, que aproxima mais o Coeficiente Gini do valor 0 [43, 34].

Para calcular o nível de impuridade de um nó dado pelo Coeficiente de Gini é necessário observar algumas informações sobre a base avaliada. Sendo n_w o número de

registros de um nó w em uma árvore e n_w^l o número de registros do nó w que possua o rótulo r, sendo $l...L$ onde L seja o número de possíveis rótulos da classe principal. Ainda observa-se que p_w^l é a proporção de registros com rótulo l, dado pela fórmula $p_w^l = \frac{n_w^l}{n_w}$ [42, 34, 43]. Conhecendo estes parâmetros é possível calcular o Coeficiente de Gini através da seguinte fórmula:

$$G(w) = \sum_{l=1}^{L} p_w^l \left(1 - p_w^l\right) \tag{3.3}$$

A Equação 3.3 permite avaliar o valor de Gini de um nó específico, determinando o nível de desigualdade deste nó. Obtido o valor do nó, é necessário avaliar o decremento obtido pelos filhos deste nó, considerando uma árvore binária onde $w1$ é o filho à direita, que obedece a regra estabelecida e $w2$ o filho da esquerda, que contraria a regra [34]. Assim, calcula-se a variação do Coeficiente de Gini ΔG com a seguinte fórmula:

$$\Delta G(w) = G(w) - \frac{n_{w1}}{n_w} G(w1) - \frac{n_{w2}}{n_w} G(w2) \tag{3.4}$$

A Equação 3.4 deve ser utilizada para identificar o melhor limiar e atributo que resulte na maior variação do Coeficiente Gini ΔG [42, 44]. Durante o treinamento de uma árvore, busca-se o melhor atributo $v*$ e o melhor limiar de separabilidade dos registros $\eta*$, cujo o par de atributo e limiar $v * \eta*$ resulte no melhor valor de $\Delta G(w)$ [43]. Ao aplicar ao algoritmo *Random Forest*, este cálculo deve ser cumulativo em todas as árvores da floresta T [39, 43, 34]. Portanto, a importância da variável calculada pelo coeficiente de Gini $I_G(v)$ é dada pela seguinte equação:

$$I_G(v) = \sum_{T} \sum_{w} \Delta G(w)_v(w, T) \tag{3.5}$$

De acordo com a Equação 3.5 a importância de um atributo ($I_G(v)$) quanto a sua aplicação no algoritmo *Random Forest* é realizada pela soma da ótima variação do Gini do atributo v por todos os nós de todas as árvores da floresta em que foi utilizado ($\Delta G(w)_v(w, T)$). Ao final, obtém-se a MDG deste atributo, a qual quanto maior o valor, maior é sua importância para a classificação dos registros [43].

Por ser um método implícito no algoritmo do *Random Forest*, a métrica de importância de variável através do coeficiente de Gini é uma forma computacionalmente barata de se executar, uma vez que já será feito a sua execução no processo de crescimento das árvores, tornado-o uma opção viável de complemento ao processo de seleção de atributos [43].

4 METODOLOGIA

O presente capítulo tem por finalidade orientar quanto a ordem e quais procedimentos foram realizados neste trabalho. As seções seguintes descrevem a organização das ações de acordo com as divisões dos processos existentes em uma análise KDD. As seções se dividem em *Pré-processamento*, *Mineração de Dados* e *Pós-processamento*. O desenvolvimento da análise KDD foi dividido de acordo com as macro-etapas de García [25] e subdivididas segundo os processos descritos por Fayyad [1] e Han *et al* [2].

No presente trabalho, adotou-se como padrão a linguagem PHP para extração dos dados e o banco de dados MySQL. O processamento foi feito através da linguagem R [1], utilizando o programa RStudio. Detalhes sobre os programas e linguagens utilizadas constam no Apendice A. A Figura 6 descreve a organização dos procedimentos acerca da análise KDD realizada.

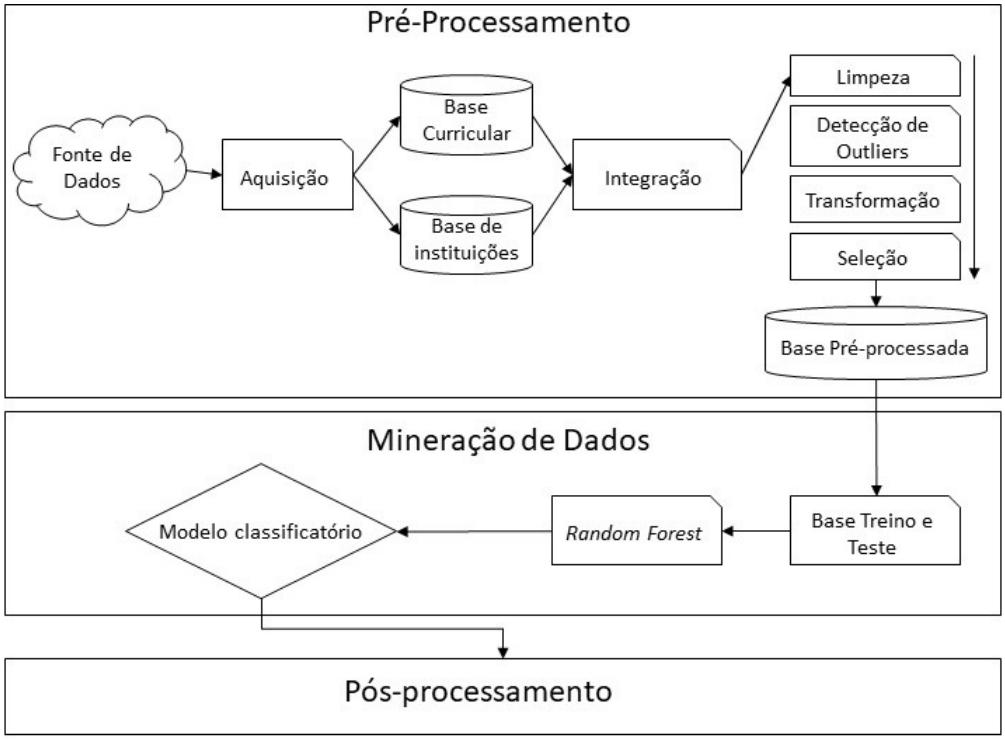

Figura 6 – Representação dos processos metodológicos utilizados.

Como observado na Figura 6, os processos do KDD podem ser agrupados em 3 grandes etapas. A primeira etapa é responsável pela aquisição e tratamento inicial dos dados, preparando-os para a etapa de Mineração. A segunda etapa se refere a aplicação do algoritmo selecionado para a Mineração de Dados (*Random Forest*), nesta etapa são realizados os procedimentos necessários a execução do algoritmo de forma apropriada.

[1] www.r-project.org

Por fim, gera-se o Modelo classificatório, o qual é testado e apresenta estatísticas utilizadas para a avaliação do modelo. Estas métricas estatísticas bem como as informações geradas pelo modelo são tratas e apresentadas de forma simplificada ao usuário na etapa de Pós-processamento.

4.1 Pré-processamento

O Pré-processamento compreende as etapas de Aquisição, Integração, Limpeza, Detecção de *Outliers*, Transformação e Seleção de atributos. Estes procedimentos visam garantir a constituição de uma base concisa e aplicável a todas as técnicas de análise de dados apresentadas. Futuros ajustes podem ser necessários para melhorar os resultados de algoritmos específicos, visto que cada um possui suas próprias particularidades.

4.1.1 Aquisição de dados

Para o bom desenvolvimento de um processo KDD é necessário que se tenha uma base considerável de dados. para o desenvolvimento deste trabalho, utilizou se a base curricular da Plataforma Lattes. A Plataforma Lattes é mantida pelo Conselho Nacional de Pesquisa (CNPq). Esta plataforma permite a integração de currículos acadêmicos de instituições públicas e privadas, fornecendo informações públicas individuais de acadêmicos [45].

No intuito de complementar alguns dados curriculares, obteve-se dados das características das Instituições de Ensino Superior onde foram realizados os cursos de graduação dos sujeitos observados na análise. Estes dados foram obtidos através do sitio eletrônico do Ministério da Educação (MEC). Estes dados contém informações como categoria da instituição, *status* de seu funcionamento e localização.

As informações obtidas destes repositórios foram organizadas em arquivos de formato XML (*Extensible Markup Language*). O XML por sua vez é considerado uma linguagem de marcação, que permite a criação de documentos estruturados que são facilmente interpretado por uma grande variedade de aplicações, bem como ser de fácil construção e interpretação [46]. Estes e outros fatores tornaram os arquivos XML eficientes para distribuição de informação.

Para realizar a interpretação e extração das informações de forma automatizada, optou-se pela utilização de um *script* desenvolvido em PHP, uma linguagem *open source* frequentemente utilizada em ambientes web devido a capacidade de se embutir ao HTML (*HyperText Markup Language*) [47, 48]. Os dados foram armazenados no banco de dados MySQL, um dos bancos de dados com maior comunidade de usuário, código aberto e de utilização gratuita [49]. A Tabela 6 apresenta os atributos adquiridos para base curricular.

Tabela 6 – Atributos selecionados da base de dados Curricular.

Atributo	**Tipo do Dado**
Instituição de Graduação	Texto
Curso de Graduação	Texto
Ingresso no Mestrado	Inteiro
Conclusão do Mestrado	Inteiro
Trabalho	Booleano
Bolsista	Booleano
Capítulos de Livro	Inteiro
Livros Publicados	Inteiro
Orientação de trabalhos	Inteiro
Pré-produção	Inteiro
Produção	Inteiro

A Tabela 6 descreve os atributos obtidos da base curricular Lattes. *Instituição de Graduação* é o nome da Universidade onde foi realizado o Curso de Graduação. Os atributos *Ingresso* e *Conclusão do mestrado* armazenam o ano desses eventos. *Trabalho* indica a existência de experiência profissional, e *Bolsista* identifica se houve participação do candidato em programa de bolsa, independente do tipo.

Os atributos *Capítulos de livros*, *Livros publicados* e *Orientação de trabalhos* contém a quantidade de cada um destes elementos registrados no currículo. O campo *Pré-produção* contém a quantidade de artigos publicados em congressos e periódicos antes do candidato ingressar no mestrado. O campo *Produção* apresenta as quantidades de publicações durante o mestrado.

Após a obtenção dos dados acadêmicos das amostras, observou-se a necessidade de informações das instituições propriamente ditas. Estas características visam criar um parâmetro de avaliação da qualidade de ensino de diferentes tipos e categorias de instituições e se existe influência no desenvolvimento acadêmico futuro do aluno e assim, o seu reflexo na produção. O resultado pode ser visto na Tabela 7.

Tabela 7 – Atributos da base de Instituições de Ensino Superior.

Atributo	**Tipo do Dado**
Nome da Instituição	Texto
Categoria da Instituição	Publica / Privada
Tipo da Instituição	Municipal Estadual Federal Especial Privada sem fins lucrativos Privada com fins lucrativos

Na Tabela 7, o atributo *Nome da Instituição* contém os nomes de cada Universidades Brasileiras. A *Categoria* classifica-as como mantida por iniciativa privada ou pelo Estado. O atributo *Tipo* descreve o tipo de seu mantenedor. Estes dados fornecem a base para aplicar as etapas seguintes da análise.

4.1.2 Integração dos Dados

A integração consiste em uma união de, neste caso, duas bases distintas porém com um atributo em comum. Para processar esta integração, utilizou-se a *Instituição de Graduação* constante no currículo *Lattes* que equivale-se ao *Nome da Instituição* da base de Instituições. Neste processo foram removidos os registros que não foram integrados por inconsistência entre os campos. A base obtida foi composta por 15 atributos, demonstrados na Tabela 8.

Tabela 8 – Base de dados Integrada.

Atributo	**Tipo do dado**
Instituição de Graduação	Texto
Curso de Graduação	Texto
Ingresso no Mestrado	Inteiro
Conclusão do Mestrado	Inteiro
Trabalho	Booleano
Bolsista	Booleano
Capítulos de Livro	Inteiro
Livros Publicados	Inteiro
Orientação de trabalhos	Inteiro
Pré-produção	Inteiro
Produção	Inteiro
Nome da Instituição	Texto
Categoria da Instituição	Publica / Privada
Tipo da Instituição	Municipal Estadual Federal Especial Privada sem fins lucrativos Privada com fins lucrativos

A Tabela 8 apresenta os atributos e seus respectivos tipos da composição da base integrada. Observa-se que esta base possui valores contínuos e discretos, sendo esta heterogeneidade de dados um possível problema às análises. Outra inconsistência que pode ser observada é a redundância dos atributos *Instituição de Graduação* e *Nome da instituição*. Os procedimentos de Limpeza, Detecção de *Outliers*, Transformação e Seleção de atributos são responsáveis por sanar essas inconsistências, melhorando a qualidade da base para a análise.

4.1.3 Limpeza dos dados

O processo de Limpeza consiste na remoção de registros e atributos que não favorecem o desenvolvimento da análise. Nesta etapa será garantido que os registros da base sejam relacionados com o estudo, portanto deve-se garantir que todos os elementos da base tenham concluído o mestrado.

É importante que seja realizado a redução da dimensão da base através da remoção de atributos desnecessários e duplicados. A redução do espaço amostral e da dimensão da base melhora o desempenho do processo de mineração bem como pode aumentar as métricas do processo de classificação.

4.1.4 Detecção de *Outliers*

Uma etapa crucial para análise de dados é a Detecção de *Outliers*, que consiste em identificar valores discrepantes dos atributos, que pode indicar um problema na etapa de aquisição dos dados. Estes problemas são causados por defeitos em sensores, informações inválidas, falha humana e diversas situações que podem comprometer a análise [26] .

Para realizar este tratamento, inicialmente é necessário identificar quais registros não obedecem a regra geral dos dados. A detecção destes registros será feita através da distribuição deles em um Diagrama de Caixa. Os elementos que forem detectados como *outliers* terão todo o registro eliminado da análise.

4.1.5 Transformação dos dados

O principal objetivo desda etapa é preparar a base, tornando-a compatível com as funções e algoritmos utilizados. É comum que as técnicas de análise de dados requisite um formato específico, ou que apresente limitações quanto a tratar diferentes tipos de dados. Esta etapa do Pré-processamento executa processos de conversão de atributos inteiros contínuos para rótulos discretos.

Quanto aos atributos que já possuem rótulos discretos, são verificadas as quantidade de registros em cada classe deste atributo, estabelecendo um limiar de 10% de representatividade no total da base, excluindo da análise os registros que estiverem abaixo deste limiar.

Quanto ao atributo Produção, este recebeu 4 rótulos discretos baseado nos valores de cada um dos *Quartis* deste atributo. Uma vez que o objetivo deste trabalho é modelar o desempenho dos discentes, são utilizados somente os registros dos quartis externos, ou seja, aqueles menores que o valor do primeiro quartil e maiores que o terceiro quartil. Os registros dos quartis internos, foram desconsiderados na análise.

4.1.6 Seleção de atributos

A Seleção de dados visa identificar os atributos que tem maior impacto no processo de modelagem do desempenho acadêmico, para tal, o atributo principal desta análise é a quantidade de artigos publicados durante o período do Mestrado. O Processo de KDD visa modelar regras e características que sejam capazes de auxiliar na previsão e classificação de novas amostras, utilizando os atributos da base como referência.

Devido a grande quantidade de atributos disponíveis para a análise, a modelagem das regras pode ser prejudicada por elementos que não estejam diretamente relacionadas com a produtibilidade científica dos discentes. Neste trabalho, optou-se por utilizar as funções Qui-quadrado e Coeficiente de Gini na seleção de atributos.

Esta etapa é responsável por desenvolver três versões da base de dados. Inicialmente é criada a base A, composta por todos os atributos disponíveis para a análise após as limpezas e transformações necessárioas. A segunda base é composta por elementos selecionados com base nos resultados da função Qui-quadrado, chamada base B. Por último será realizado uma nova seleção de atributos com base na importância das variáveis para o algoritmo *Random Forest*, detectado através do Coeficiente de Gini, gerando a base C.

4.2 Mineração de Dados

A etapa de Mineração de Dados é o procedimento do KDD em que, dado uma base de dados pré-processada, é aplicado um algoritmo de análise de dados que permite a extração de novas informações. Esta seção apresentará os procedimentos utilizados na aplicação do algoritmo de mineração de dados *Random Forest*, um algoritmo baseado na estrutura de árvore de decisão.

4.2.1 Treinamento e Teste

Algoritmos de AD supervisionados comumente exigem que sejam fornecidos dois conjuntos disjuntos de exemplos, a base de *treino* e a base de *teste*. Estas bases são subconjuntos da base de dados original e devem conter exemplos de todos os possíveis registros. A etapa de treinamento aplica a referida base ao algoritmo e obtém um modelo inicial. Em seguida este modelo é aplicado a base de teste, onde são obtidas as métricas de avaliação do algoritmo.

Para este trabalho, as bases de *treino* e *teste* utilizaram o método de amostragem *holdout* estratificado, em que, dado a porcentagem fixa de 60/40% da base original para os respectivos subconjuntos, são selecionados registros aleatórios sem reposição para a composição das bases, buscando manter a mesma proporção das classes [50, 32].

É importante observar que em procedimentos onde o algoritmo é influenciado

por alguma característica aleatória, este algoritmo apresenta uma característica não-determinística, ou seja, uma mesma entrada neste algoritmo pode produzir resultados diferentes.

Sendo o algoritmo *Random Forest* um algoritmo não-determinístico, para garantir que o resultado obtido não foi influenciado pela característica aleatória deste, é necessário realizar novas repetições, calculando a média das métricas de avaliação obtidas a cada repetição. Este trabalho validará os resultados com base em 3 sequências de repetições, 100, 250 e 500 repetições.

4.2.2 Aplicação do *Random Forest*

Após o pré-processamento dos dados e sua subdivisão em conjuntos de *treino* e *teste*, é executado o algoritmo responsável pela extração dos dados e modelagem das regras de classificação, neste trabalho é utilizado o algoritmo *Random Forest*, do pacote 'randomForest' [37].

O algoritmo foi executado com os parâmetros iniciais do próprio pacote, garantindo a criação de uma floresta com o valor padrão de 500 árvores. Neste contexto, o presente trabalho avaliará as diferentes formas de aplicação do algoritmo, visando extrair o melhor resultado. Baseado nas diferentes bases de dados a serem utilizadas, bem como a quantidade de repetição, será realizada as seguintes avaliações:

1. **Avaliação das Repetições:** Com os registros ajustados, será analisado qual a quantidade de repetições retorna o melhor resultado para o algoritmo *Random Forest*, avaliando a média das métricas de cada processamento.

2. **Avaliação dos Atributos:** Comparar as métricas do modelo gerado pelas bases de dados, qual delas apresenta o conjunto de atributos mais apropriado para gerar uma boa classificação.

3. **Avaliação Final:** Utilizando o melhor número de repetições e a base mais apropriada atributo, executa-se uma ultima vez o algoritmo, obtendo os melhores resultados e preparando a análise para o pós-processamento.

As avaliações 1 e 2 são realizadas em conjunto, executando as diferentes bases com os diferentes números de repetições e avaliando todos os resultados, detectando aquele que apresentar melhor resultado com base em sua Acurácia e Sensibilidade e Especificidade.Dado tais resultados, é realizado a análise dos dados com a configuração ótima obtida das execuções anteriores.

Os resultados, livres de influência humana na classificação dos candidatos, fornecem novos conhecimentos sobre o processo de avaliação curricular, relacionando-o com o

desempenho acadêmico. A melhoria da apresentação dos resultados é realizada na etapa de Pós-processamento.

4.3 Pós-processamento

Recebida as informações processadas pelo algoritmo de Mineração, a etapa de Pós-processamento é responsável por organizar estas informações de forma que pessoas interessadas nos resultados, mas sem conhecimento específico da área de análise de dados, possa compreender. Portanto nesta etapa os resultados são organizados em tabelas e gráficos, com dados padronizados e de forma clara para que os diferentes dados possam ser comparados.

Dado o conjunto de resultados, o Pós-processamento apresentará as matrizes de confusão do melhor e pior modelo, permitindo que seja possível visualizar as estatísticas geradas de forma clara. Com base nos resultados obtidos pelo Coeficiente de Gini, será apresentado a proporção entre as classes do atributo de maior importância e do atributo principal *Produção*. Estes dados fornecem a base para justificar as contribuições deste trabalho.

5 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados da metodologia abordada. Baseado nestes, são apresentadas as considerações acerca das contribuições almejadas: a relevância dos atributos curriculares para a previsão de desempenho; o modelo de previsão de produtibilidade; descoberta de conhecimento aplicável à avaliação curricular.

5.1 Pré-processamento

Esta seção visa apresentar os resultados do pré-processamento, descrevendo a organização dos registros utilizados nas análises de mineração. Por meio destas informações espera-se identificar as características curriculares que apresentam maior impacto na análise do desempenho discente.

5.1.1 Aquisição e Integração dos dados

O processo de aquisição obteve 5312 currículos da área de Computação e informações de 2640 Instituições de Ensino Superior. A integração das duas bases foi realizada com base no nome da Instituição de Graduação, excluindo os registros cadastrados incorretamente ou com instituições de outro país. A base integrada contava com 2090 registros válidos compostos por 14 atributos.

5.1.2 Limpeza de Dados

Foram removidos da base 525 registros que, mesmo iniciando o Mestrado, não o concluíram. Nesta etapa foi realizado a redução da dimensão da base, eliminando atributos desnecessários. Ao executar esta etapa, obteve-se a estrutura de dados apresentada na Tabela 9.

Tabela 9 – Base de dados após o processo de Limpeza.

Atributo	**Tipo dos dados**
Trabalho Bolsista	Binário
Capítulos de Livros Livros Publicados Orientação de Trabalhos Pré-produção	Inteiro
Tipo da Instituição	Municipal; Estadual; Estadual; Federal; Especial; Privada com fins lucrativos; Privada sem fins lucrativos
Produção	Inteiro

De acordo com a Tabela 9, os atributos *Instituição de Graduação, Nome da Instituição, Ingresso no Mestrado* e *Conclusão do Mestrado* foram removidos, bem como o atributo *Categoria da Instituição* que tem sua função substituída pelo atributo *Tipo*. A Figura 7 apresenta a distribuição dos registros quanto ao atributo *Curso de Graduação*.

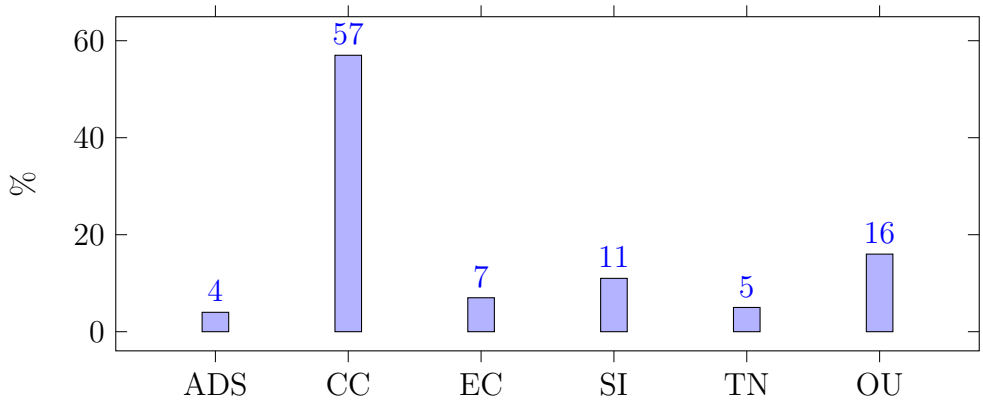

Figura 7 – Porcentagem de registros com cada rótudo de *Curso de Graduação*.

De acordo com a Figura 7, os elementos *ADS, CC, EC, SI, TN* e *OU* representam respectivamente os rótulos: *Análise de Sistemas, Ciências da Computação, Engenharia da Computação, Sistemas de Informação, Tecnologias* e *Outros*. Observa-se que o atributo está desbalanceado e, uma vez que não faz sentido agrupar as classes com menor representação, optou-se por remover o atributo.

O atributo *Tipo da Instituição* também é composto por rótulos discretos, portanto foi necessário verificar a representatividade de seus rótulos na base de dados como visto na Figura 8.

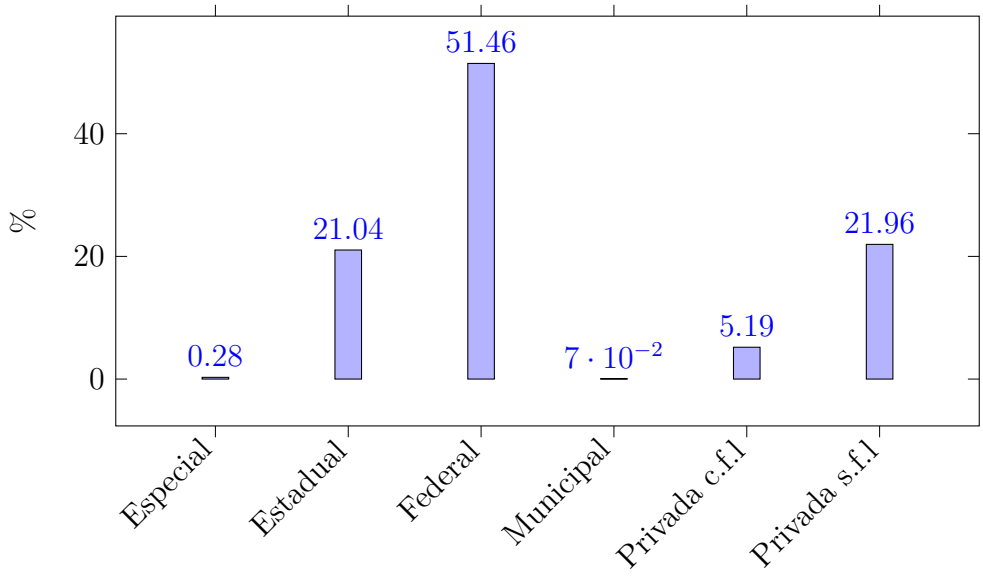

Figura 8 – Porcentagem de registros com cada rótudo de *Tipo da Instituição*.

A Figura 8 é composta pelas 6 classes do atributo *Tipo*, sendo *Privada c.f.l* o atributo *Privada com fins Lucrativos* e *Privada s.f.l* o atributo *Privada sem fins Lucrativos*. Com base nesta tabela, foram removidos os registros das classes com menos de 10% de representatividade. O atributo *Privada s.f.l* foi renomeado para *Privada*. Ao final obteve-se uma base com 1560 registros e 8 atributos.

5.1.3 Detecção de *Outliers*

O processo de detecção de *outliers* foi realizado com auxílio do Diagrama de Caixa (*Boxplot*) com base na divisão dos *Quartis* do atributo *Produção*, ao identificar valores fora do padrão, removeu-se os registros. O resultado pode ser visto na Figura 9.

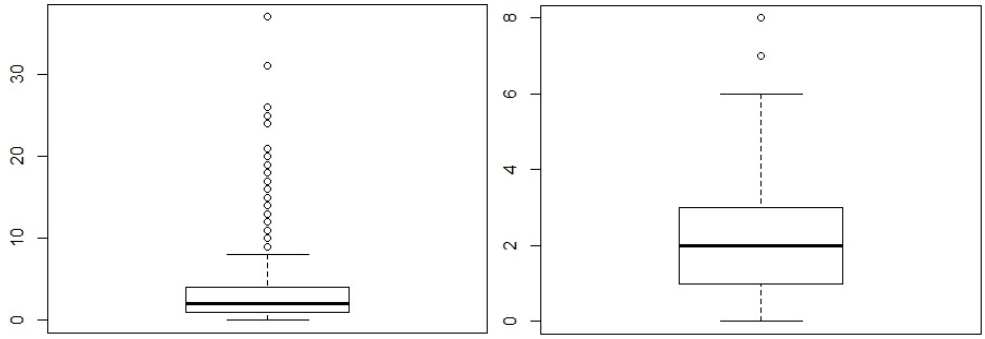

Figura 9 – Diagrama de Caixa do atributo *Produção.*. **Esquerda**: Primeira execução. **Direita**: Segunda execução.

Observando a Figura 9 à esquerda nota-se que a maioria dos registros possuem menos de 10 publicações durante o Mestrado. Os valores que excedem o limiar de Whisker, representados por círculos, são considerados *outliers*, resultando na exclusão de 97 registros.

Após a remoção é executado uma segunda vez, resultando na figura à direita, que ainda apresenta *outliers*. Foram detectados e removidos 61 *outliers*. Por fim, um novo diagrama de caixa é gerado, como demonstrado na Figura 10.

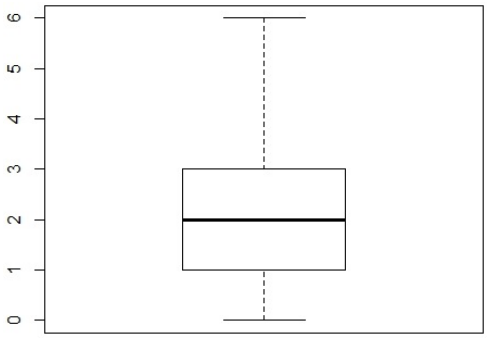

Figura 10 – Diagrama de Caixa do atributo *Produção* - iteração final.

A Figura 10 não apresentou nenhum registos fora dos limites da Estimativa de Whisker, portanto considera-se todos os registros com valores de produção dentro do padrão esperado pela base, resultando numa base com o total de 1402 registros.

5.1.4 Transformação dos dados

Os atributos *Trabalho* e *Bolsista* tiveram seus rótulos padronizados para os rótulos **SIM** e **NÃO**. Os atributos *Capítulos de Livro*, *Livros Publicados*, *Orientação de trabalhos* e *Pré-produção* cujo valores eram compostos por números inteiros contínuos, foram convertidos para **SIM** onde o valor fosse maior que 0 e **NÃO** para registros com valor igual a 0.

O atributo *Produção*, baseado na distribuição de seus *Quartis*, recebeu a alteração de seus rótulos, transformando os valores inteiros em classes que identifiquem a qual *quartil* o registro pertence, este resultado pode ser observado na Tabela 10.

Tabela 10 – Limiares dos quartis do atributo Produção.

Quartil	**Q1**	**Q2**	**Q3**
Número de Publicações	1	2	3

Dado os limiares de cada quartil demonstrados na Tabela 10, observa-se que 25% da base não possui produção científica ($Produção < Q1$). Entre uma e três produções, estão 50% dos registros ($Q1 \leq Produção < Q3$) e com três ou mais produções ($Q3 \leq Produção$), constam os demais 25%.

Considerando tais limiares, o presente trabalho visa classificar as amostras que apresentam a maior separabilidade, ou seja, os valores extremos da base. Os registros que apresentem valores de *Produção* medianos foram desconsiderados da análise, portanto os elementos entre $Q1$ e $Q3$ ($Q1 \leq Produção \leq Q3$) foram desconsiderados.

Os registros com valores externos mínimos, ou seja $Produção < Q1$, recebem o rótulo *Sem Produção* enquanto os registros com $Produção > Q3$ recebem o rótulo *Alta Produção*. Os resultados dos procedimentos de transformação de forma resumida podem ser vistos na Tabela 11.

Tabela 11 – Tabela das transformações de dados realizadas.

Atributo	**Tipo dos dados**	**Transformação dos Dados**
Trabalho Bolsista	Binário	SIM NÃO
Capítulos de Livro Livros Publicados Orientação de Trabalhos Pré-produção	Inteiro	SIM NÃO
Tipo da Instituição	Municipal Estadual Federal Especial Privada com fins lucrativos Privada sem fins lucrativos	Estadual Federal Privada
Produção	Inteiro	Sem Produção Alta Produção

Como é possível observar na Tabela 11, todos os atributos inteiros foram transformados, os discretos, baseado na representatividade de suas classes receberam transformações ou remoções. Ao final a base de dados possui 827 registros e 8 atributos, neste trabalho, esta configuração será chamada de Base A. A base *original* juntamente com a base A encontram-se disponíveis no Apêndice B.

5.1.5 Seleção de atributos

Inicialmente aplicou-se a fórmula Qui-quadrado à base A para identificar a independência ou dependência dos atributos quanto ao atributo principal *Produção*. O resultado da aplicação desta função pode ser observado na Tabela 12.

Tabela 12 – Tabela de resultados do Qui-quadrado.

Atributo	χ^2	**Graus de Liberdade**	χ^2_α	H_0
Trabalho	0,65	1	3,84	Não Rejeita
Bolsista	1,26	1	3,84	Não Rejeita
Capítulos de Livros	2,70	1	3,84	Não Rejeita
Livros Publicados	2,10	1	3,84	Não Rejeita
Orientação de Trabalho	1,12	1	3,84	Não Rejeita
Pré-produção	36,43	1	3,84	Rejeita
Tipo da Instituição	2,23	2	5,99	Não Rejeita

Os resultados da Tabela 12 utilizam o nível de significância α de 5% e e toma como base os valores de referência apresentados na Tabela 3. De acordo com esta tabela,

somente o atributo *Pré-produção* rejeita a Hipótese Nula H_0, fazendo-o o único atributo que apresenta dependência com o objetivo da pesquisa, *Produção*.

A base de dados B é formada pelos atributos que rejeitam a H_0 com base no resultado da função Qui-quadrado (Tabela 12), porém nota-se que somente um atributo (*Pré-produção*) rejeita a H_0. Modelar uma base de dados composta por somente um atributo é inviável, uma vez que isto impossibilita a etapa de análises de dados devido a falta de informações. Neste cenário observa-se que a base B é inutilizável no contexto de mineração de dados deste trabalho.

Sendo a função Qui-quadrado incapaz de realizar uma seleção de atributos apropriada para o contexto do trabalho, utilizou-se um segundo seletor de atributos, o Coeficiente de Gini. Através da aplicação da base A ao algoritmo *Random Forest*, realizando 500 repetições e criando uma floresta de 500 árvores, obteve-se os resultados observados na Figura 11.

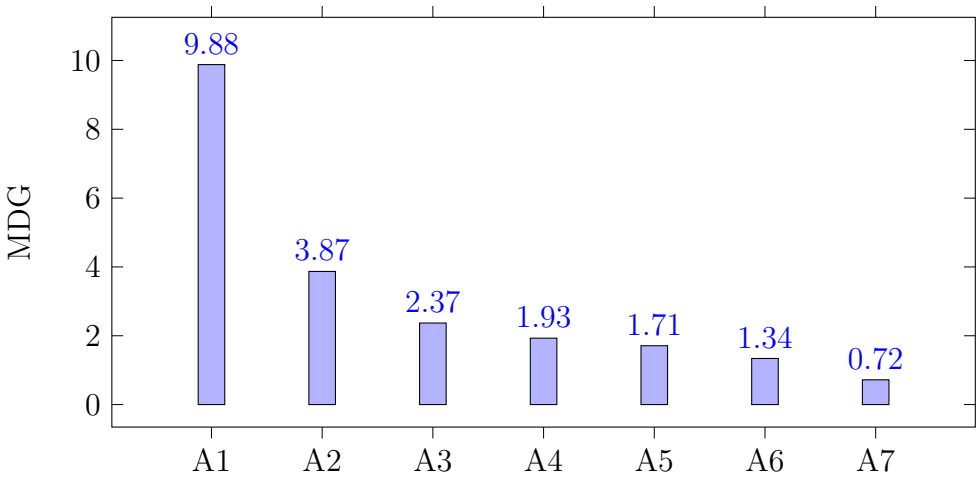

Figura 11 – Resultado do *Coeficiente de Gini*.

A Figura 11 apresenta os resultados do cálculo médio do *Coeficiente de Gini* dos atributos. De acordo com a Figura, os rótulos *A1, A2, A3, A4, A5, A6* e *A7* representam respectivamente os atributos *Pré-produção, Tipo da Instituição, Orientação de Trabalho, Trabalho, Capítulos de Livros, Bolsista* e *Livros Publicados*.

De acordo com a Figura 11 a *Pré-produção* ($A1$) foi o atributo que apresentou a maior importância para a classificação dos dados. Para realizar a seleção dos atributos, optou-se por selecionar aqueles que apresentem MDG maior que a média, que neste caso é de 3,12. Assim os atributos que compõe a base C, ou seja, que foram selecionados pelo Coeficiente de Gini, foram *Tipo da Instituição* e *Pré-produção* ($A2$ e $A1$ respectivamente).

Ao final tem-se as bases A,B e C, onde a base A e C serão aplicadas ao algoritmo de mineração de dados *Random Forest*. A base B não será aplicada pois a seleção de

atributos com Qui-quadrado não resultou em uma quantidade viável de atributos para o algoritmo.

5.1.6 Considerações sobre a relevância dos atributos

Após a conclusão do pré-processamento, ao observar as características das bases e os resultados dos algoritmos de seleção, pode-se realizar algumas considerações sobre os atributos que compõe estas bases. Analisando as bases geradas neste capítulo, busca-se identificar correlações entre os atributos e a *Produção* científica dos discentes.

Considerando as informações da Tabela 12 sobre os resultados do *Qui-quadrado*, o atributo Pré-produção foi o único que rejeitou a *Hipótese Nula*, portanto foi o único em que se considera a dependência com a *Produção*. De acordo com esta função, os demais atributos não rejeitam a *Hipótese Nula*, e portanto não de pode descartar sua independência da *Produção*.

Quanto aos resultados do *Coeficiente de Gini* da Figura 11, observa-se que o atributo *Pré-produção* possui o maior nível de importância. Este resultado corrobora com os dados obtidos através do *Qui-quadrado*, portanto considera-se que, nas condições apresentdas neste trabalho, a *Pré-produção* é o único atributo que apresentou relevância e dependência com a *Produção* discente.

5.2 Mineração de Dados

A mineração de dados consiste na aplicação de algoritmos capazes de extrair conhecimento relevante de bases com uma grande quantidade de informação. Neste trabalho serão utilizadas as duas bases organizadas anteriormente, a base *A* e a base *C*. Por meio da aplicação do algoritmo *Random Forest*, serão obtidos novas informações que poderão auxiliar na tomada de decisão.

O processamento das duas bases de dados pelo algoritmo *Random Forest* é realizado em diferentes conjuntos de repetições. Ao executar a análise com 100, 250 e 500 repetições e calculando a média das métricas de avaliação de cada resultado, permite que seja observado a influência das repetições no desempenho do algoritmo, bem como garante a aleatoriedade dos resultados. Os resultados da aplicação do algoritmo é visto na Tabela 13.

Tabela 13 – Média das métricas da aplicação do *Random Forest*.

	Base A			***Base C***		
Métricas	**Repetições**					
	100	*250*	*500*	*100*	*250*	*500*
Acurácia	58,09%	58,35%	58,51%	58,56%	**58,68%**	58,60%
Sensibilidade	61,93%	61,87%	61,81%	**62,16%**	**62,16%**	62,06%
Especificidade	50,83%	51,39%	**52,00%**	51,44%	51,48%	51,11%

A Tabela 13 apresenta os resultados obtidos com as execuções do algoritmo *Random Forest*. Os valores em destaque representam os melhores resultados daquela métrica avaliativa e portanto, indica qual base e repetição apresentou o melhor resultado. De acordo com a tabela, através da análise da base C com 250 repetições obteve-se a maior parte dos melhores resultados. Portanto será utilizado neste trabalho para a análise final, a base C com 250 repetições. A Tabela 14 apresenta os resultados desta análise.

Tabela 14 – Resultados do *Random Forest* com a configuração ótima.

Métricas	**Media dos Modelos**	**Melhor Modelo**	**Pior Modelo**
Acurácia	58,87%	62,42%	52,42%
Sensibilidade	62,25%	62,27%	57,14%
Especificidade	52,04%	63,16%	40,22%

A Tabela 14 apresenta as métricas de avaliação do algoritmo, contendo os resultados médios das 250 repetições, as métricas do melhor modelo encontrado e da pior modelagem. Estes resultados possibilitam averiguar a variação entre os modelos gerados, bem como o desempenho do melhor modelo detectado. Estes dados devem ser acessíveis e compreensíveis a todos os públicos interessados, portanto as informações obtidas do modelo serão organizadas e reapresentadas de forma simplificada na etapa de Pós-processamento.

5.2.1 Considerações sobre a modelagem de desempenho

Um dos objetivos deste trabalho foi modelar as características curriculares, aplicando regras que permitissem prever o desempenho acadêmico de novas amostras. Para realizar o treinamento do modelo utilizou-se as bases A e C e calculou-se a **Acurácia, Sensibilidade** e **Especificidade** do algoritmo para cada base e conjunto de repetição.

Com base no melhor modelo gerado pela configuração ótima, observou-se a **Acurácia** de 62,42%, portanto é possível a elaboração de um modelo capaz de auxiliar o processos de previsão do desempenho discentes. Todavia deve-se considerar que o problema em questão apresenta características complexas por sofrer influência de diferentes fatores sociais, ambientais e pessoais, o que fica visível nos resultados das métricas

Portanto ainda que este trabalho demonstre a possibilidade do desenvolvimento de um modelo, a baixa acurácia indica que os modelos gerados devem ser utilizados como um apoio a decisão, não deve-se acatar seus resultados de forma cartesiana, sendo necessário a intervenção humana. Esse fator ocorre devido a amplitude e complexidade do problema, que sofre influências de características além das curriculares.

5.3 Pós-processamento

Ao avaliar um classificador, as métricas de sua execução como **Acurácia, Sensibilidade** e **Especificidade** podem ser confusas para os interessados nos resultados que não tenham conhecimento específico, portanto é interessante demonstrar as informações obtidas com informações claras e de simples compreensão. A distribuição dos registros no melhor e pior modelo de classificação pode ser vista nas Tabelas 15 e 16.

Tabela 15 – Matriz de confusão do Melhor modelo.

	Classificado		
Observado	*Alta Produção*	*Sem Produção*	**Total**
Alta Produção	163	28	191
Sem Produção	110	29	139
Total	273	57	330

Tabela 16 – Matriz de confusão do Pior modelo.

	Classificado		
Observado	*Alta Produção*	*Sem Produção*	**Total**
Alta Produção	137	54	191
Sem Produção	101	38	139
Total	238	92	330

De acordo com as Tabelas 15 e 16 a classe positiva utilizada foi a **Alta Produção**, esta classe também recebeu o maior número de registros classificados, tanto na modelagem de maior acurácia quanto na de pior. Considerando o atributo *Pré-produção* como o maior responsável por esta classificação, é interessante observar a proporção de suas classes quando comparadas ao atributo *Produção*, esta comparação é vista na Figura 12.

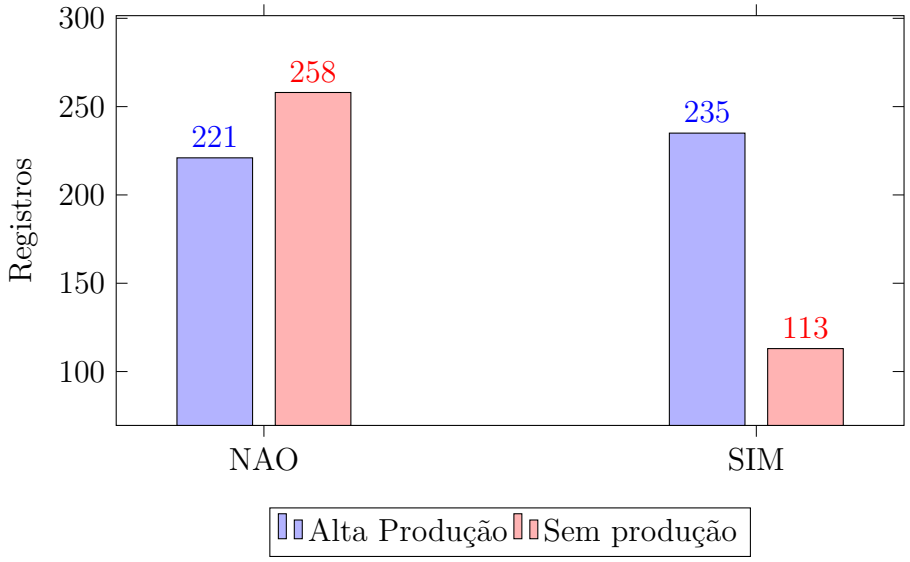

Figura 12 – Proporção entre as classes dos atributos Pré-produção e Produção.

Como observado na Figura 12 as classes do atributo *Pré-produção* foram distribuídas em duas colunas, uma indicando registros com **Alta produção** e outra, **Sem produção**. Registros com rótulo **Não** tem uma distribuição semelhante quanto ao seu desempenho produtivo após o ingresso, enquanto registros com *Sim* tem maior representação em **Alta produção**.

5.3.1 Considerações sobre o conhecimento gerado

Munido dos dados das análises, é possível gerar novos conhecimentos acerca da avaliação curricular, de forma que este procedimento seja melhorado e como reflexo, a qualidade dos discentes ingressantes. Para que seja facilitada interpretação dos resultados, foram formuladas regras que representam as informações geradas pelo KDD:

- Candidatos que apresentem experiência com escrita científica, tende a apresentar um melhor desempenho acadêmico.
- Com exceção da Pré-produção, os demais atributos curriculares não apresentam qualquer impacto na previsão de desempenho acadêmico.
- Dentre as amostras, 25% concluíram o Mestrado sem registrar nenhuma produção científica em seu currículo neste período.

6 CONSIDERAÇÕES FINAIS

Diante da sistemática avaliativa atual imposta pela CAPES, os programas tendem a valorizar a produção científica de seus recursos humanos, docentes e discentes. Dada esta problemática, a presente pesquisa realizou a modelagem do desempenho discente de programas de Mestrados em Computação, baseado em sua produção científica. Esta análise utilizou dados curriculares de Mestres e através da aplicação de técnicas de KDD, obteve-se novas informações que podem beneficiar o processo de avaliação curricular.

A execução do processo de KDD foi dividido em três etapas. Durante o Pré-processamento, analisou-se os relacionamentos dos atributos curriculares estudados com a produção científica das amostras. Com base nas funções de seleção de atributos, destacou-se a relevância da *Pré-produção* no processo de classificação, onde a experiência com escrita científica refletiu de forma positiva no desempenho discente.

Na etapa de Mineração de dados foi desenvolvimento de um modelo classificador através do algoritmo *Random Forest*, capaz de determinar a faixa de produção de novos registros. Os resultados estatísticos obtidos com o modelo gerado indica uma grande complexidade do problema da produtibilidade discente, uma vez que este sofre influência de um grande espectro de variáveis, além das curriculares.

É importante observar que um modelo que busque classificar elementos com esta complexidade não deve ser considerado como elemento decisivo no processo seletivo, mas sim como ferramenta que auxilie a análise e fundamentação das decisões dos avaliadores. Este modelo apresenta resultados pioneiros quanto a previsão de produção científica, fornecendo a base para novos estudos e melhorias no processo de seleção discente.

Na etapa de Pós-processamento são apresentadas as considerações sobre os conhecimentos gerados através do KDD dos atributos curriculares. Dada as informações deste trabalho considera-se que o problema é mais amplo do que o esperado, sendo as características curriculares, insuficientes para gerar um modelo confiável do desempenho discente.

Por fim, os estudos demonstra a possibilidade de modelagem da produtibilidade, de forma que, com a ampliação deste estudo, elementos como qualidade da produção, impacto na sociedade e registro de patentes sejam avaliados e modelados, auxiliando na detecção de talentos e ampliando a aplicabilidade do modelo.

6.1 Trabalhos futuros

O estudo realizado traz resultados pioneiros na análise de produtibilidade acadêmica, todavia existe uma grande gama de estudos a se realizar nesta área para tornar este modelo viável, exigindo uma série de novos estudos. Ainda que o trabalho tenha atingido seus objetivos, é importante apontar algumas dificuldades enfrentadas, as quais podem ser sanadas nestes estudos vindouros.

Inicialmente é importante que seja revisado o problema em si. Analisar quais as características envolvem a produção científica e os elementos que tem influência nesta, auxiliaria a fundamentar futuras pesquisa sobre a modelagem da produção. Também é importante que outras produções sejam investigadas, como os registros de patentes e produções técnicas.

Outro problema a ser sanado é a população avaliada. Neste critério duas propostas se veem válias: ampliar o espaço amostral, com dados consistentes e equilibrados; ou reduzir o escopo avaliado, analisando dados locais, em uma mesma universidade ou de um grupo fechado de pesquisadores, facilitando o controle das variáveis estudadas.

É recomendado que novos trabalhos, abordem diferentes metodologias para validar e melhorar este estudo. Ampliar as características avaliadas como histórico escolar, dados socioeconômicos, proficiência em línguas estrangeiras, qualidade das publicações e seu impacto destas sociedade. Por influenciarem a problemática, a análise de novos elementos pode resultar em melhorias da modelagem e do estudo como um todo.

Uma vez que este trabalho focou seu estudo na avaliação curricular para melhorar o processo seletivo de programas de Mestrado, sugere-se que seja estudada a influência de outros métodos de seleção, investigando quais são utilizados pelas universidades brasileiras. Avaliar a relevância destes métodos e seu impacto na modelagem do desempenho discente pode refletir de forma positiva inclusive em outros níveis de formação.

Espera-se que este trabalho possa fomentar o nicho de pesquisa voltado a detecção de novos talentos e quais características influenciam na sua formação. Estes estudos beneficiariam não somente programas de mestrado mas a sociedade como um todo.

REFERÊNCIAS

[1] FAYYAD, U.; UTHURUSAMY, R. Data mining and knowledge discovery in databases. *Communication of ACM*, v. 39, n. 11, p. 24–26, 1996. ISSN 0001-0782. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230>.

[2] HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. 3. ed. [S.l.]: Morgan Kaufmann, 2011. ISSN 09953914. ISBN 9788578110796.

[3] NIST/SEMATECH. *e-Handbook of Statistical Methods*. 2012. Disponível em: <http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm>.

[4] JIN, Y. et al. The Model of Network Security Situation Assessment Based on Random Forest. *7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, p. 977–980, 2016.

[5] NAVAUX, P. O. A.; CÁCERES, E. N.; ZORZO, A. F. *Documento de Área Ciência da Computação*. 2016. 31 p.

[6] RIBEIRO, R. J. *Para que serve a avaliação da Capes*. [S.l.], 2007. 2 p.

[7] BARBOSA, L. Meritocracia à brasileira: o que é desempenho no Brasil ? *Revista do Serviço Público*, v. 120, n. 3, p. 58–102, 1996.

[8] MAGALHÃES, F. A. C.; ANDRADE, J. X. Exame vesttibular, características demográficas e desempenho na Universidade: Em busca de fatores preditivos. In: . [S.l.: s.n.], 2002.

[9] DUTKA, C. F. The Prediction of Academic Performance. p. 184, 1975. Disponível em: <http://digitalcommons.mcmaster.ca/cgi/viewcontent.cgi?article=6529{&}context=opendissertati>.

[10] PEKRUN, R. et al. Achievement Emotions and Academic Performance: Longitudinal Models of Reciprocal Effects. *Child Development*, v. 00, n. 0, p. 1–18, 2017. ISSN 14678624.

[11] HEISSEL, J.; NORRIS, S. Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood Through Puberty. n. August, 2015. ISSN 0022-166X. Disponível em: <http://papers.ssrn.com/abstract=2674256>.

[12] CHAMILLARD, a. T. Using student performance predictions in a computer science curriculum. *Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education*, p. 260–264, 2006. ISSN 00978418. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.6585&rep=rep1&type=pdf>.

[13] PAIREEKRENG, W.; PREXAWANPRASUT, T. An integrated model for learning style classification in university students using data mining techniques. *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, p. 1–5, 2015.

Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7206951>.

[14] Lopez Guarin, C. E.; GUZMAN, E. L.; GONZALEZ, F. A. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Revista Iberoamericana de Tecnologias del Aprendizaje*, v. 10, n. 3, p. 119–125, 2015. ISSN 19328540.

[15] KURNIAWAN, Y.; HALIM, E. Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application and benefits). *Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2013*, n. August, p. 98–103, 2013.

[16] OLIVEIRA, A. G. de; GARCIA, D. F. Mineração da Base de Dados de um Processo Seletivo Universitário. *INFOCOMP Journal of Computer Science*, v. 3, n. 2, p. 38–43, 2004.

[17] LAMADRID-FIGUEROA, H. et al. Admissions Criteria as Predictors of Students' Academic Success in Master's Degree Programs at the National Institute of Public Health of Mexico. *Public health reports*, v. 127, n. 6, p. 605–611, 2012. ISSN 00333549. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true{&}db=a9h{&}AN=83362840{&}site=e>.

[18] KTONA, A.; XHAJA, D.; NINKA, I. Extracting Relationships between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques. *2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks*, p. 6–11, 2014. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7059136>.

[19] EL-HALEES, A. Mining Students Data To Analyze Learning Behavior : a Case Study Educational Systems. *Work*, n. February, 2008.

[20] TAIR, M. M. A.; EL-HALEES, A. M. Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information and Communication Technology Research*, v. 2, n. 2, p. 140–146, 2012.

[21] DEVASIA, T.; VINUSHREE, T.; HEGDE, V. Prediction of students performance using educational data mining. In: IEEE. *Data Mining and Advanced Computing (SAPIENCE), International Conference on*. [S.l.], 2016. p. 91–95.

[22] OGOR, E. N. Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques Department of Natural Sciences Turks & Caicos Islands Community College Visualization and Articulation. p. 0–5, 2007.

[23] BARADWAJ, B.; PAL, S. Mining educational data to analyze student's performance. *Internation Journal od Advamced Computer Science and Applications*, v. 2, n. 6, p. 63–69, 2012. ISSN 2156-5570. Disponível em: <http://arxiv.org/abs/1201.3417>.

[24] BUSH, J. Entry characteristics and academic performance of students in a master of pharmacy degree program in the United Kingdom. *American journal of pharmaceutical education*, v. 76, n. 7, p. 1–10, 2012. ISSN 15536467.

[25] GARCÍA, E. et al. A collaborative educational association rule mining tool. *Internet and Higher Education*, v. 14, n. 2, p. 77–88, 2011. ISSN 10967516.

[26] AMO, S. D. Técnicas de mineração de dados. *Jornada de Atualizaçao em Informatica*, 2004.

[27] CAMILO, C. O.; SILVA, J. C. da. Mineração de Dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, p. 29, 2009. ISSN 16113349.

[28] SILVA, M. R.; MOURA, F. P. de; JARDIM, C. H. O diagrama de caixa (Box Plot) aplicado à análise da distribuição temporal das chuvas em Januária, Belo Horizonte e Sete Lagoas, Minas Gerais-Brasil. *Revista Brasileira de Geografia Física*, v. 10, p. 023–040, 2017. ISSN 1984-2295.

[29] LI, T.; ZHANG, C.; OGIHARA, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, v. 20, n. 15, p. 2429–2437, 2004. ISSN 1367-4803. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9956{&}rep=rep1{&}type=pdf{\%}5Cnpapers2://publication/uuid/23DB36B5-2348-44C4-B831-DBDD6EC7702D{\%}5Cnhttp://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinform>.

[30] VIEIRA, S. *Introdução à Bioestatística*. [S.l.]: Elsevier Brasil, 2015.

[31] JIN, X. et al. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. *International Workshop on Data Mining for Biomedical Applications*, p. 106–115, 2006. ISSN 03029743.

[32] MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. *Sistemas inteligentes: fundamentos e aplicações*, p. 89–114, 2003.

[33] SAFAVIAN, S. R.; LANDGREBE, D. A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man and Cybernetics*, v. 21, n. 3, p. 660–674, 1991. ISSN 21682909.

[34] BASTOS, D. G. D.; NASCIMENTO, P. S.; LAURETTO, M. S. Proposta e Análise de Desempenho de Dois Métodos de Seleção de Características para Random Forests. p. 49–60.

[35] BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125.

[36] SONG, W.-j.; LI, B. A Method to Detect Machine Generated Domain Names Based on Random Forest Algorithm. p. 6–10, 2016.

[37] LIAW, a.; WIENER, M. Classification and Regression by randomForest. *R news*, v. 2, n. December, p. 18–22, 2002. ISSN 16093631.

[38] FERNÁNDEZ-DELGADO, M. et al. A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis*, v. 53, n. 4, p. 1208–1217, 2009. ISSN 01679473. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167947308004970>.

[39] HAN, H.; GUO, X.; YU, H. Variable Selection Using Mean Decrease Accuracy And Mean Decrease Gini Based on Random Forest. *7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, p. 219–224, 2016.

[40] PEREIRA, R. B. *Seleção lazy de atributos para a tarefa de classificação*. Tese (Doutorado) — Master's thesis, UFF-Universidade Federal Fluminense, Brazil, 2009.

[41] SATHYADEVI, G. Application of CART algorithm in hepatitis disease diagnosis. In: *International Conference on Recent Trends in Information Technology, ICRTIT 2011*. [S.l.: s.n.], 2011. p. 1283–1287. ISBN 9781457705885.

[42] LANGS, G. et al. Detecting Stable Distributed Patterns of Brain Activation Using Gini Contrast. *Motor Control*, v. 27, n. 4, p. 590–609, 2009. ISSN 09652140.

[43] MENZE, B. H. et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, v. 10, n. 1, p. 213, 2009. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/10/213>.

[44] GARCIA, R.; NIEVOLA, J. C.; PARAISO, E. C. Estudo Comparativo de Métodos de Seleção de Atributos na Predição de Matrizes de Conectividades em TDAH Obtidas Pela Técnica de Resting-State fMRI. p. 637–647, 2016.

[45] MENA-CHALCO, J.; DIGIAMPIETRI, L.; OLIVEIRA, L. Perfil de produção acadêmica dos programas brasileiros de pós-graduação em ciência da computação nos triênios 2004-2006 e 2007-2009. *Em Questão*, Universidade Federal do Rio Grande do Sul, v. 18, n. 3, 2012.

[46] BRAY, T. et al. Extensible markup language (XML). *World Wide Web Journal*, v. 2, n. 4, p. 27–66, 1997. Disponível em: <http://www.w3pdf.com/W3cSpec/XML/2/REC-xml11-20060816.pdf>.

[47] NIEDERAUER, J. *PHP com XML*. 3ª. ed. São Paulo: NOVATEC EDITORA LTDA., 2002. 15 p. ISBN 9788575221198.

[48] ACHOUR, M. et al. *Manual do PHP*. 2017.

[49] PIRES, C. E. S.; NASCIMENTO, R. O.; SALGADO, A. C. Comparativo de desempenho entre bancos de dados de codigo aberto. *Centro de Informatica-Universidade Federal de Pernambuco. Recife*, 2008.

[50] TANTITHAMTHAVORN, C. et al. An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. *IEEE Transactions on Software Engineering*, v. 5589, n. c, p. 1–16, 2016. ISSN 0098-5589.

Apêndices

APÊNDICE A – FERRAMENTAS UTILIZADAS

Neste capítulo são descritas as versões e características das ferramentas utilizadas no trabalho. Estas informações permitem que as experiências realizadas sejam replicadas em ambiente semelhante, permitindo verificar a validade dos resultados.

A.1 Configurações do Computador

- **Sistema Operacional:** Windows 10 Pro 64-bit
- **Idioma:** Português
- **Processador:** Intel(R) Core(TM) i5-2430M CPU @ 2.40GHz (4 CPUs), 2.4GHz
- **Memoria:** 4096MB RAM
- **Versão do DirectX:** DirectX 12
- **Placa de Vídeo:** NVIDIA GeForce GT 555M
- **HD:** Kingston SSD 120GB

A.2 Servidor Web

- **Servidor:** Apache/2.4.17 (Win32) OpenSSL/1.0.2d
- **Versão do PHP:** 5.6.21
- **Versão do Banco de Dados:** MySQL 5.5

A.3 Linguagem R

- **IDE:** RStudio - Versão: 1.0.136
- **Versão do R:** 3.3.1 para Windows

APÊNDICE B – MATERIAL DE REFERÊNCIA

Pasta contendo as bases de dados *Original* e *A*.

Bases de dados: https://goo.gl/lfzNhb

TRABALHOS PUBLICADOS PELO AUTOR

Trabalhos publicados pelo autor durante o programa (obrigatório somente para teses de doutorado e dissertações de mestrado no template DC/UEL).

1. Ronan Anacleto Lopes, Luiz Antônio Lima Rodrigues, Jacques Duílio Brancher, **Predicting Master's Applicants Performance Using KDD Techniques**, The 12º Iberian Conference on Information Systems and Technologies - CISTI, 2017