# Uso da Mineração de dados para predição de reprovação em seleção para Residência Médica.

Taciana Barbosa Duque[1], Seiji Isotani[2], Bruno Elias Penteado [3]

***Resumo***

*A residência médica é uma pós-graduação de excelência na formação profissional. Este estudo teve como objetivo identificar através da Mineração de dados se os resultados das avaliações dos estudantes durante a graduação podem predizer precocemente um desfecho desfavorável na prova de residência médica. O estudo foi realizado em uma faculdade de medicina na região nordeste do Brasil, utilizando-se o banco de dados de registro acadêmico da instituição. Foi utilizado o algoritmo de árvore de decisão Weka-classifiers-trees.J48.O modelo que apresentou melhor acurácia foi o correspondente às avaliações realizadas no quarto ano do curso (67.8947 %), sendo possível uma intervenção. Novos modelos devem ser testados buscando aprimorar o acompanhamento e gestão da aprendizagem .*

Palavras-chaves: Mineração de dados; WEKA; Medicina; Residência Médica

1 Pós-Graduanda em Computação Aplicada à Educação, USP, tacianaduque@usp.br.
2 Orientador, Prof. Titular Computação e Tecnologias Educacionais ICMC- USP sisotani@icmc.usp.br
3 Orientador, Doutor em Ciência da Computação, USP, brunopenteado@usp.br



Cite as: Duque, T. & Isotani, S. & Penteado, B. (2020). Uso da Mineração de dados para predição de reprovação em seleção para Residência Médica.. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

## 1. Introdução

A área da educação vem passando por grandes transformações.Novas metodologias e a inserção cada vez maior da tecnologia, traz diferentes demandas aos docentes, instituições de ensino e estudantes, seja na educação básica ou ensino superior. A crescente utilização de recursos computacionais na educação exige renovação e mudança de atitudes e do olhar sobre novos e antigos problemas [Ventura, 2010; Sesat, 2017].

Em ambientes educacionais, compreender o percurso de aprendizagem dos estudantes é de grande importância e é nesse aspecto que a Mineração de dados (MD) assume valioso papel. A MD é um conjunto de técnicas e procedimentos que podem ser realizados a partir de várias fontes de dados, analisando as informações existentes e tendo como saída um novo conhecimento sobre aqueles dados. MD é considerada, portanto, uma área interdisciplinar agregando conhecimento de diversas áreas como a estatística e aprendizagem de máquina **[**Sánchez-Guzmán A, & García, 2013; Villanueva, 2018].

Nas últimas décadas, com o avanço do uso de recursos computacionais, observa-se um aumento substancial de dados armazenados nas mais diversas áreas, não sendo diferente na educação, onde muitas das decisões que são tomadas, sejam na gestão, sejam no âmbito dos aspectos pedagógicos, não levam em consideração a análise das informações e dados coletados e armazenados [Carvalho, 2014; Ventura, 2010].

A mineração de dados educacionais se concentra em dados provenientes do contexto educacional; busca entender como se dá a aprendizagem e que aspectos podem contribuir. Previsões baseadas em desempenho de grupos específicos compõem o grande *pool* de pesquisas e aplicações da mineração de dados na educação [Villanueva, 2018].

A formação médica no Brasil segue as orientações das Diretrizes Curriculares Nacionais que orientam a formação de egresso com perfil generalista, crítico e reflexivo e reforçam a utilização de metodologias ativas de aprendizagem no projeto pedagógico dos cursos. Entre as metodologias ativas de aprendizagem, a aprendizagem baseada em problemas (ABP), iniciada na Universidade de McMaster na década de 70, vem sendo largamente utilizada em cursos médicos de diversos países, incluindo o Brasil [MEC, 2014; Dijkstra, 2009].

A construção do currículo dentro da metodologia ABP, se dá de maneira modular, flexível, integrado, com inserção precoce do estudante em atividades práticas e os conteúdos são apresentados não em aulas expositivas, mas, através de discussão de problemas, casos reais, contextualizados, e em pequenos grupos. Diversas estratégias de avaliações de caráter formativo e somativo são planejadas para promover a aprendizagem e a progressão do estudante ao longo do curso [Dijkstra, 2009; Troncon, 2014].

Também no ensino superior, existe um crescente interesse em identificar os principais fatores influenciadores da performance do estudante utilizando técnicas de mineração de dados. Antecipar a ocorrência de algum evento adverso através dessa estratégia, permite realizar intervenções educacionais que possam auxiliar a superar a dificuldade [Au Saa, 2019].

A predição é uma tentativa de antecipar o conhecimento sobre o que pode acontecer no futuro. Em mineração de dados, as informações contidas na base de dados servem de referência para a construção de um modelo que permita essa previsão [Han, 2006]. Existem três tipos de predição: a de classificação, regressão e estimação de densidade. As tarefas de predição têm o objetivo de antecipar o valor de determinado atributo (variável) baseado nos valores de outros atributos. A abordagem de predição tem sido a mais utilizada em Mineração de dados Educacionais (MDE), correspondendo a cerca de 60% das pesquisas publicadas [Au Saa, 2019].

O desafio encontra-se em explorar de maneira adequada os dados para auxiliar na tomada de decisão. A busca por fatores que interfiram na aprendizagem é o desafio diário de quem lida com educação, e as técnicas de mineração de dados devem se tornar cada vez mais acessíveis para todos os envolvidos [Ventura, 2010; Sesat, 2017] .

A residência médica é uma pós-graduação de excelência onde o médico pode aperfeiçoar suas competências na especialidade que deseja seguir [Ministério da Educação, 2006]. Estima-se que cerca de 70% dos concluintes do curso de medicina optem por realizar o processo seletivo de residência médica ao término do curso. O curso médico associa no seu projeto pedagógico variadas estratégias de ensino e aprendizagem, focadas no desenvolvimento das competências essenciais para a formação do médico que atenda aos anseios da população [Girardi, 2017].

A aprovação em processo seletivo não constitui o objetivo final direto do projeto pedagógico do curso. Entretanto, a residência médica é uma etapa importante no processo de formação continuada e da especialização do médico. A sua importância leva os estudantes a associarem às atividades curriculares nos dois últimos anos do curso, a realização de cursos preparatórios para esse processo seletivo da forma como fazem no ensino médio para acesso à graduação [Aragão, 2018; Epstein, 2007].

Pensar em estratégias de predição mais precoce de insucesso através de indicadores de avaliação de desempenho dos estudantes pode trazer benefícios por mudança no planejamento de ações educacionais específicas visando reverter esse resultado.

## 2. Objetivo do estudo

Este estudo tem como principal objetivo identificar se os resultados da performance dos estudantes de medicina obtidos do sistema de avaliação durante os primeiros quatro anos de curso são capazes de predizer o desfecho no processo de seleção para a residência médica.

## 3.Método

O estudo foi realizado com dados dos concluintes do curso de Medicina de uma faculdade localizada na região nordeste do Brasil. Trata-se de uma Faculdade privada sem fins lucrativos, especializada em cursos na área da saúde, possuindo ainda pós-graduação em nível de especialização e Mestrado em Educação para o ensino na área da saúde.

O software utilizado para a realização das tarefas de mineração de dados deste trabalho foi o WEKA 3.7 (Waikato Environment for Knowledge Analysis), que foi desenvolvido na Nova Zelândia, na Universidade de Waikato [Hall et al, 2009]. Anterior ao processo de mineração foi realizado o tratamento da base de dados. Esse tratamento foi baseado no processo KDD ( Knowledge Discovery in Databases). Etapas do processo KDD [Costa,Baker, Marinho,2012]:

### 3.1 Seleção

Os dados analisados foram obtidos do sistema de registro acadêmico da faculdade hospedados na base do Sistema Educacional Lyceum, que é utilizado pela instituição. Os dados foram selecionados tendo como critério de inclusão ter concluído o curso de medicina nos anos de 2018 e 2019, tendo como critérios de exclusão: tempo de integralização do curso superior a seis anos.

### 3.2 Pré-processamento

Nesta etapa, já com os dados da amostra a ser estudada, os atributos foram selecionados considerando a sua relevância para o objetivo do estudo. Foram selecionados atributos relacionados a características sociodemográficas dos estudantes e as médias obtidas nos oito primeiros períodos do curso tanto em atividades teóricas identificadas como módulos e atividades práticas. Foi selecionado também o atributo Teste de progresso, que trata-se de uma avaliação formativa, longitudinal que acontece a cada semestre durante os seis anos do curso. A codificação dos atributos foi mantida e criada uma legenda para auxiliar no processo de análise e interpretação dos dados. Foram incluídos mais dois atributos obtidos do banco de dados, das instituição responsável pelo processo seletivo de residência médica, que é de domínio público, onde ficam registradas as médias e a condição - aprovado ou reprovado - dos concluintes que realizaram o processo seletivo de Residência Médica.

### 3.3 Transformação

A fase de transformação, é anterior à mineração sendo a fase em que os dados devem ser formatados, agrupados, ou decompostos de forma que ofereça condições de um melhor resultado. Nessa etapa pode-se categorizar os valores, revisar se possuem códigos que comprometam a leitura e construção do modelo.

Na etapa de transformação, após a revisão dos dados, foi criado mais um atributo baseado nos atributos originais, considerado como importante para a etapa de mineração. Este novo atributo consistiu na média final de cada semestre, com base nas médias obtidas em cada módulo e atividade prática referente ao semestre, criando-se o atributo: média por período.

O banco de dados, ao final, ficou com 190 instâncias (número de concluintes) e 62 atributos. A relação dos atributos e respectivas legendas estão presentes no Quadro 3.1. De maneira progressiva foram identificados atributos mais relacionados ao desfecho mantendo-se uma proporção adequada em relação ao número de instâncias para a solicitação dos modelos.

## 3.4 Mineração de dados

As técnicas de mineração de dados podem ser descritivas e preditivas. Na sua categoria descritiva busca analisar os dados, descrever suas características, destacando aspectos de interesse. Na categoria preditiva a análise dos dados tem como finalidade a construção de um ou vários modelos que permitam predizer comportamentos futuros. Tomando como exemplo a classificação, busca-se encontrar um modelo que descreve e distingue classe de dados para predizer a qual pertence, um novo elemento. Existem diversos algoritmos classificadores e diversas formas de apresentar o conhecimento resultante da mineração dos dados [Cretton, 2018; Costa, 2012; Baker, 2011].

A qualidade e quantidade dos dados, têm importância no modelo resultante. Se os dados traduzem o que se deseja obter e/ou estão em quantidade suficiente para o aprendizado, são informações que devem ser consideradas não só na escolha do algoritmo, mas também na interpretação de seus resultados. Através do algoritmo de classificação, por exemplo, busca-se um modelo baseado nas características que mais se aproximam de uma dada classe [Cretton, 2018].

Algumas métricas são utilizadas para avaliar a eficácia do modelo criado. Uma delas é a acurácia, que traduz a precisão do modelo. Mostra a percentagem de acertos do modelo proposto. A taxa de erro e acerto é outra métrica utilizada, assim como o índice Kappa que traduz o grau de respostas concordantes, entre o previsto e o observado [Cretton, 2018].

No presente estudo, na etapa de mineração de dados, foi utilizada a técnica de classificação e árvore de decisões. A construção da árvore de decisão acontece a partir de um conjunto de dados de amostras já classificadas – dados de treinamento. Os modelos estatísticos do tipo Árvore de decisão realizam a classificação e predição de dados utilizando treinamento supervisionado.

A técnica de árvore de decisão apresenta o modelo em formato de árvore o que facilita a compreensão da relação entre os atributos. Utilizou-se o algoritmo de árvore de decisão Weka-classifiers-trees J48. O algoritmo J48 é considerado uma implementação em JAVA do algoritmo C4.5 no aplicativo de mineração de dados Weka. Em cada nó da árvore, o algoritmo identifica o atributo que mais efetivamente particiona o seu conjunto de dados em subconjuntos pertencentes a uma ou outra

categoria ou categorias. Nesse algoritmo, o atributo de maior significância aparece como raiz da árvore e o algoritmo gera árvore de decisão no sentido do topo para a base. As árvores de decisão constituem uma representação gráfica de um conjunto de regras. O algoritmo J48 procura identificar, portanto, o quanto informativo é um atributo para selecionar a separação ótima [WEKA].

No presente estudo os atributos foram compostos pelo conjunto de notas obtidas por estudantes de medicina que concluíram o curso nos anos de 2018 e 2019, como apresentados no Quadro 3.1. A reprovação na Residência Médica foi a variável preditiva. Inicialmente foi testado o algoritmo J48 para todos os atributos globalmente e posteriormente foram testados os atributos para as avaliações a cada ano do curso. Foram excluídas as avaliações do período do internato, que corresponde aos 2 últimos anos do curso. Foi considerado como precoce um modelo que possa fazer predição nos primeiros quatro anos de curso.

Não existiam dados faltantes e todas as variáveis numéricas (notas) estavam na mesma escala, de 0 a 10.Foram utilizados os parâmetros padrão do Weka versão 3.7: valor do *confidence factor* foi 0,25 , número mínimo de instâncias por nó (*minNumObj*) igual a 2, divisão entre treinamento e teste foi por cross-validation utilizando-se o método *K-fold* com 10 subconjuntos.

Na fase de pós-processamento foram analisados os padrões encontrados, buscando informações relevantes a partir das análises das informações obtidas. Os dados obtidos foram analisados inicialmente pela acurácia do modelo. Após a avaliação da acurácia foi analisada a coerência e embasamento teórico do modelo, para apresentação dos resultados.

**Quadro 3.1 Atributos e respectivas descrições e período de curso que estão relacionados**

| Atributo | Legenda | Período do curso |
|---|---|---|
| Residência | Variável binária: aprovado/reprovado | Não se aplica |
| Sexo | Masculino/feminino | Não se aplica |
| Tipo_ingresso | Vestibular/Transferência/Prouni | Não se aplica |
| Idade | Idade conclusão do curso em anos | Não se aplica |
| Formação_pai | código com grau de instrução | Não se aplica |
| Formação_mãe | código com grau de instrução | Não se aplica |
| Estado_civil | casado /solteiro | Não se aplica |
| Tempo_ingress_medicina | Ingresso no curso após conclusão ensino médio em anos | Não se aplica |
| MAB0100 | Nota das avaliações atividades práticas | 1º |
| MAB0101 | Nota módulo Epidemiologia | 1º |
| MAB0102 | Nota módulo Ética e Bioética | 1º |
| MAB0103 | Nota módulo Fisiologia e semiologia médica aplicada 1 | 1º |
| MAB0104 | Nota módulo Fisiologia e semiologia médica aplicada 2 | 1º |
| MP1 | Média do período | 1º |
| MAB0200 | Nota das avaliações atividades práticas | 2º |
| MAB0212 | Nota módulo Estudo dos fármacos | 2º |
| MAB0213 | Nota módulo O Sistema Brasileiro de Saúde | 2º |

| | | |
|---|---|---|
| MAB0214 | Nota módulo Concepção e nascimento | 2º |
| MAB0215 | Nota módulo Saúde da Criança Cresc. e desenvolvimento | 2º |
| MP2 | Média do período | 2º |
| MAB0300 | Nota das avaliações atividades práticas | 3º |
| MAB0323 | Nota módulo Saúde do Adolescente | 3º |
| MAB0324 | Nota módulo O método científico | 3º |
| MAB0325 | Nota módulo Doenças na infância 1 | 3º |
| MAB0326 | Nota módulo Doenças na infância 2 | 3º |
| MP3 | Média do período | 3º |
| MAB0400 | Nota das avaliações atividades práticas | 4º |
| MAB0434 | Nota módulo Saúde do adulto | 4º |
| MAB0435 | Nota módulo Deontologia médica | 4º |
| MAB0436 | Nota módulo que estuda Doenças na gestação | 4º |
| MAB0437 | Nota módulo que estuda Doenças infecciosas | 4º |
| MP4 | Média do período | 4º |
| MAB0500 | Nota avaliação prática cenário real | 5º |
| MAB0545 | Nota módulo Estudo desenvolvimento psíquico e das relações humanas | 5º |
| MAB0546 | Nota módulo Ética nas práticas de Saúde | 5º |
| MAB0547 | Nota módulo ques estuda doenças com edema, perda ou ganho de peso | 5º |
| MAB0548 | Nota módulo ques estuda Trauma, urgência e emergência | 5º |
| MP5 | Média do período | 5º |
| MAB0600 | Nota avaliação prática cenário real | 6º |
| MAB0652 | Nota módulo Dist. Psiquiátricos | 6º |
| MAB0653 | Nota módulo Diversidade e interculturalidade | 6º |
| MAB0654 | Nota módulo que estuda doenças que apresentam desconforto respiratório | 6º |
| MAB0655 | Nota módulo que estuda doenças que apresentam nódulos e tumores | 6º |
| MP6 | Média do período | 6º |
| MAB0700 | Nota avaliação prática cenário real | 7º |
| MAB0759 | Nota módulo Teoria e técnicas psicoterápicas | 7º |
| MAB0760 | Nota módulo Processos de aprendizagem | 7º |
| MAB0761 | Nota módulo que estuda alteração trânsito intestinal, vômito e icterícia | 7º |
| MAB0762 | Nota módulo estuda doenças q a principal manifestação clínica é Dor | 7º |
| MP7 | Média do período | 7º |
| MAB0800 | Nota avaliação prática cenário real | 8º |
| MAB868 | Nota módulo Saúde do idoso envelhecimento e terminalidade da vida | 8º |
| MAB0869 | Nota módulo Ética planetária e Saúde Global | 8º |
| MAB0870 | Nota módulo estuda doenças principal manifestação é anemia e sangramento | 8º |
| MAB0871 | Nota módulo estuda doenças com alterações dos sentidos e do Sistema nervoso | 8º |
| MAB0872 | Nota módulo estuda doenças que acometem os idosos | 8º |
| MP8 | Média do período | 8º |
| MAB0978 | Nota módulo Doenças prevalentes na assistência à criança | 9º |
| MAB1082 | Nota Módulo Doenças prevalentes na assistência à mulher | 10º |
| MAB1186 | Nota Módulo Temas prevalentes na Clínica médica | 11º |
| MAB1290 | Nota Módulo Temas prevalentes na Clínica cirúrgica | 12º |
| Teste de Progresso | Média Teste de progresso realizado | 1º ao12º |

## 4. Resultados

Os resultados são derivados da base de dados de registros acadêmicos da Faculdade, referente às notas e características sociodemográficas de concluintes do curso de medicina nos anos de 2018 e 2019, que realizaram o processo seletivo para a residência médica. Entre os 190 concluintes, 97 (51,05%) foram reprovados, sendo a classe majoritária para a qual buscou-se estudar um modelo preditivo.

A aplicação do algoritmo J48 quando foram incluídos todos os atributos não mostrou boa acurácia (57,7368%). Foram testados a partir daí atributos relacionados às notas do primeiro ao quarto ano do curso e na sequência, por ano do curso. Quando o algoritmo foi aplicado no conjunto total dos atributos, observou-se uma acurácia de 60%; quando analisamos cada ano do curso, observamos melhor acurácia (67.8947%) para os atributos notas obtidas do quarto ano. Os resultados desses testes com respectivas acurácias, sensibilidades e especificidades são observados no quadro 4.1. O comportamento da acurácia do modelo de acordo com o ano do curso pode ser melhor observado no gráfico 4.1, quando a inclusão das notas do quarto ano, já amplia a acurácia e observa-se o melhor desempenho com as notas exclusivamente deste ano.

A partir desse dado, analisamos a árvore de decisão relacionada ao modelo que incluíam as notas do quarto ano, podendo ser observado nas figuras 4.2. Observou-se uma inconsistência relacionada ao Módulo MAB0761 do sétimo período, mostrando aprovação com menor média neste módulo; na sequência analisamos a árvore do oitavo período em separado, figura 4.3, uma vez que são períodos independentes com conteúdos, avaliações e competências exigidas, também específicas. As acurácias do algoritmo J48 em cada um desses períodos foram semelhantes, sendo iguais a 65,26%.

**Quadro 4.1 Acurácia, sensibilidade e especificidade do algoritmo J48 testado em base de dados de registro acadêmico de estudantes de medicina para predição de reprovação na Residência Médica, por ano do curso.**

| Variáveis | J48 |
|---|---|
| Todas as variáveis (190 observações/62 atributos) | Acurácia: 54.7368 % |
| | Sensibilidade: 53,6% |
| | Especificidade: 55,9% |
| Todas as notas (190 observações/55 atributos) | Acurácia: 60,0 % |
| | Sensibilidade: 61,5%% |
| | Especificidade: 58,6% |
| Notas obtidas no 1º ano (190 observações/11 atributos) | Acurácia: 51,0526% |
| | Sensibilidade: 51,1% |
| | Especificidade: 26,0% |
| Notas obtidas no 2º ano (190 observações/11 atributos) | Acurácia:50,0% |
| | Sensibilidade:50,5% |
| | Especificidade:25,0% |
| Notas obtidas no 3º ano (190 observações/11 atributos) | Acurácia:47,3684 % |
| | Sensibilidade:49,1% |
| | Especificidade:37,9% |
| Notas obtidas no 4º ano (190 observações/12 atributos) | Acurácia: 67.8947 % |
| | Sensibilidade: 63,8% |
| | Especificidade:76,7% |
| Notas obtidas no 7º período (190 observações/6 atributos) | Acurácia: 65,26% |
| | Sensibilidade:61,0% |
| | Especificidade:77,6% |
| Notas obtidas no 8º período (190 obserações/7 atributos) | Acurácia:65,26 % |
| | Sensibilidade:61,0% |
| | Especificidade:77,6% |

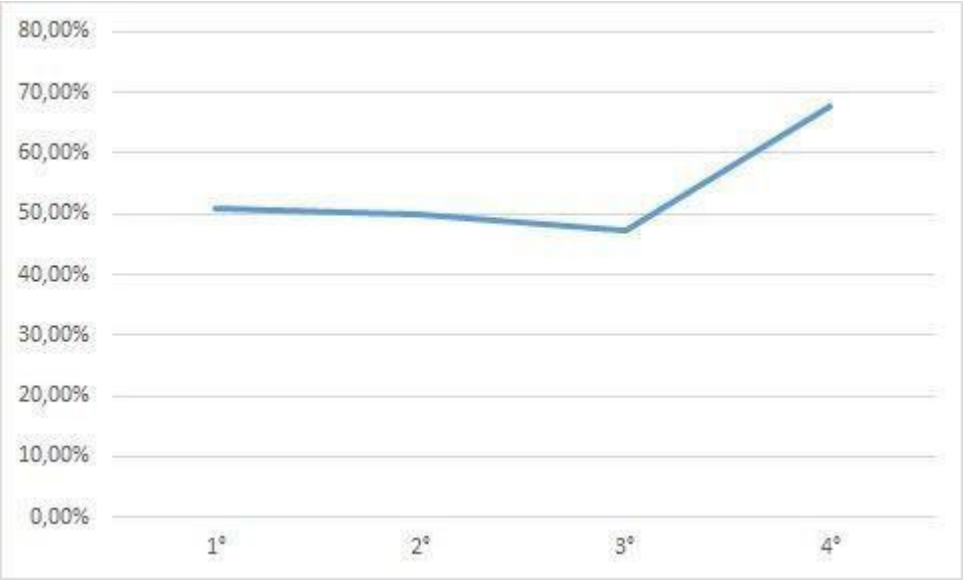


**Figura 4.1. Acurácia do algoritmo J48 testado em base de dados das médias por ano de curso de estudantes de medicina no período pré-internato para predição do desfecho reprovação no processo seletivo para Residência Médica.**

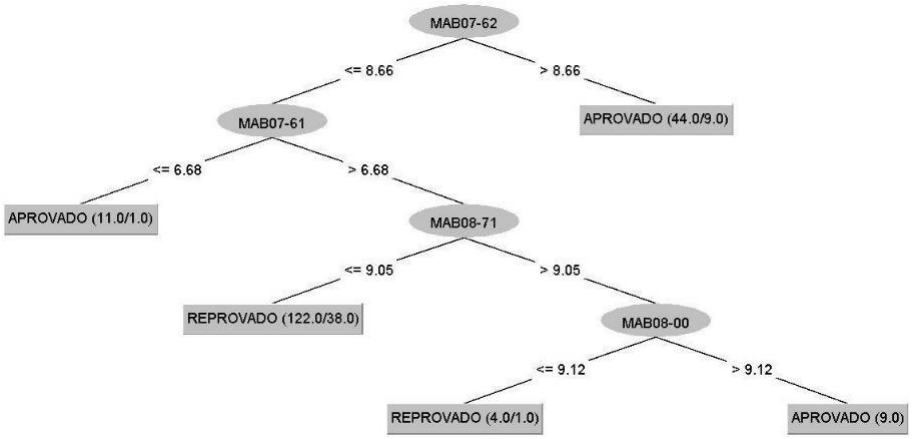


**Figura 4.2 Árvore de decisão referente às notas do 4º ano -7º e 8º períodos.**

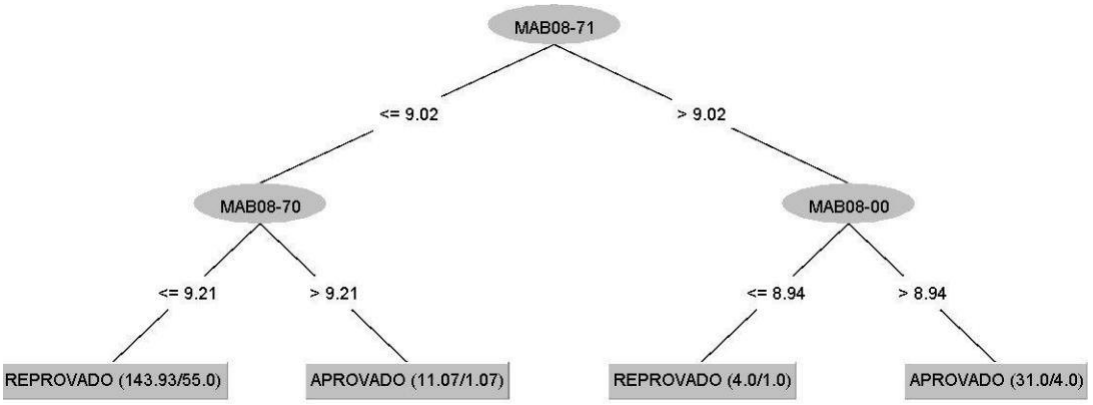


**Figura 4.3 Árvore de decisão final referente às notas do 4º ano - 8º período.**

## 5. Discussão

Estudos sobre MDE vêm aumentando consideravelmente nas últimas décadas. O crescente aumento de dados em ambientes virtuais de aprendizagem e em registros para gestão acadêmica de estudantes, sejam em espaços presenciais ou de forma remota, encontram nos modelos de mineração um caminho eficiente para auxílio na análise e tomada de decisão. O número de pesquisas publicadas sobre MDE teve seu aumento mais importante a partir de 2008, coincidindo com o ano de início da conferência internacional de MDE promovida pela *International Educational Data Mining Society*, estando os estudos relacionados a desempenho de estudantes e predição entre os mais frequentes [Villanueva et al , 2018].

A velocidade em que a MDE é incorporada nas práticas dos educadores e gestores na educação é variável, a depender da área. O interesse em buscar modelos que possam predizer precocemente fatores que contribuam para o baixo desempenho dos estudantes - reprovação, abandono - em diferentes etapas da formação, é evidente nas diversas publicações. Entretanto, em relação a estudantes de medicina, estudos brasileiros são mais frequentemente relacionados ao desempenho no ENADE, e muitos estudos nacionais ou internacionais sobre desempenho de estudantes de medicina são realizados com base em testes estatísticos, mas, não utilizando mineração de dados [ Pugh, 2016, Sesato, 2017;Yassin, 2018].

Alguns estudos relacionados à predição de reprovação, seja com estudantes de medicina ou outras áreas merecem ser destacados e contribuem com a análise dos resultados do presente estudo.

Costa et al (2017) realizaram estudo comparativo sobre técnicas de mineração de dados em curso introdutório de programação, disponível em uma Universidade pública brasileira, na modalidade presencial e à distância. Os autores estudaram 4 técnicas diferentes de MDE em predizer precocemente reprovação de estudantes tanto em curso presencial, quanto curso à distância. Observaram que após a primeira semana de curso, as técnicas de MDE são capazes de identificar com pelo menos 50% de eficácia, os estudantes com probabilidade de reprovação e observaram que as técnicas de árvore de decisão apresentaram as maiores eficácias, sejam nas fontes de dados de curso à distância sejam nos cursos presenciais.

O que chama a atenção no estudo de Costa é a tentativa de encontrar um modelo para identificar precocemente a reprovação, embora tenha obtido uma eficácia de 50% utilizando árvore de decisão. No presente estudo desenvolvido com estudantes de medicina, buscou-se identificar fatores que pudessem predizer a condição de reprovação no processo seletivo de residência médica. Focamos nas variáveis relacionadas a módulos (disciplinas) nos primeiros quatro anos. Os dois últimos anos, denominado de período de internato, são compostos por uma carga horária predominantemente prática. A identificação de um modelo preditivo até o quarto ano, permite uma intervenção nos dois últimos anos do curso, que poderá favorecer o resultado dos estudantes no processo seletivo de residência médica. Outro aspecto é que não podemos comparar a eficácia encontrada pelo estudo de Costa et al uma vez que diferentemente do presente estudo, não utilizaram a acurácia.

Amjed Abu Saa et al (2019) realizaram elegante estudo de revisão sistemática buscando identificar os principais fatores associados a estudos de performance em estudantes de nível superior utilizando mineração de dados e identificaram que a utilização de resultados de avaliação curricular foi o critérios mais frequentemente estudado, utilizando a técnica de árvore de decisão.

Creton e Gomes (2018) aplicaram a técnica de mineração de dados na base do ENADE relacionada aos resultados obtidos por estudantes de medicina com ênfase no nível de dificuldade apontado pelos estudantes e dados sociodemográficos e da Instituição de ensino superior (IES). Os autores separaram a característica da IES se pública, privada com ou sem fins lucrativos. Apresentaram um resultado global do que chamam de índice de confiança de 84%. O estudo de Creton e Gomes não permite uma comparação clara com o presente estudo, uma vez que não define se o índice de confiança utilizado foi a acurácia e por outro lado não apresenta os resultados obtidos em cada árvore de decisão. O estudo traz contribuições sobre o potencial existente na base de dados do ENADE para análises preditivas em cursos de graduação.

O presente estudo analisou avaliações curriculares de estudantes de medicina em uma Faculdade no nordeste do Brasil, que utiliza como metodologia de ensino a aprendizagem baseada em problemas, com o objetivo de predizer da maneira mais precoce possível a reprovação no processo seletivo de residência médica. A composição das notas utilizadas de cada módulo não se dá apenas por resultados em testes de conhecimento, mas, inclui avaliações de competências relacionais, estudo cooperativo e competências atitudinais.

A residência médica (RM) é uma pós graduação de excelência, ainda não obrigatória no Brasil, mas,é almejada pela maioria dos concluintes de medicina, para especialização na área de atuação escolhida. No presente estudo, utilizando árvore de decisão J48, observou-se que o modelo composto pelas notas obtidas no quarto ano do curso, forneceu a melhor acurácia (67,8947%). Os módulos componentes deste modelo referem-se ao estudo da clínica do adulto, relacionados a doenças que provocam dor incluindo doenças coronarianas, doenças abdominais aguda; módulos que estudam as doenças relacionadas ao aparelho digestório; módulos relacionados a doenças do sistema nervoso. Outro componente importante do modelo proposto é a presença do módulo MAB0800 que trata-se de um módulo de atividade prática, em cenário real, onde o estudante na fase pré-internato, realiza atendimento a pacientes no hospital de ensino, sob supervisão, e as avaliações se dão pela observação direta do desempenho do estudante nessa atividade. Pelas suas características, o desempenho de excelência nesta atividade prática pode traduzir um *proxi* do estudante bem preparado para as etapas seguintes, uma vez que essa avaliação acontece nos diferentes ambientes de aprendizagem, incluindo o atendimento à criança, à mulher, ao adulto e também em cenário de urgência e emergência.

O modelo do quarto ano, entretanto, apresentou uma inconsistência em relação ao módulo MAB0761. Trata-se de um módulo com médias baixas e especulamos se alguns estudantes de habitual bom desempenho possam ter tido desempenho insatisfatório neste módulo por problemas relacionados à elaboração do próprio testes. Diante dessa inconsistência, apresentamos também a árvore de decisão para os módulos do oitavo período, sendo o período mais próximo do internato, agregando na árvore de decisão, módulo MAB0870 que estuda as doenças hematológicas.

A prova de residência médica é composta de questões relacionadas às áreas da criança (pediatria), da mulher (ginecologia e obstetrícia), da saúde pública e do adulto que incluem a área de cirurgia e de clínica, sendo nesta última que encontram-se os conhecimentos presentes no modelo proposto. A prova é organizada de maneira igualitária em relação ao número de questões por área, entretanto estudo sobre dificuldade de questões em provas de RM, destacam a importância da avaliação da taxonomia das questões, que pode variar entre as áreas [Aragão, 2018].

O modelo proposto apresenta como melhor momento de predição o quarto ano do curso que corresponde ao período imediatamente anterior ao internato. Algumas escolas médicas nos EUA, planejam de forma curricular uma etapa de preparação dos seus concluintes para a prova de residência médica ou de licenciamento para o exercício da profissão. O planejamento do curso médico visa a atuação do profissional no

atendimento à população,mas, considerando a importância da etapa de RM nessa formação, pensar estratégias de acompanhamento para favorecer o êxito dos estudantes também em processos seletivos pode ser benéfico.

No Brasil, são os próprios estudantes que buscam cursos preparatórios, geralmente nos dois últimos anos do curso, quando deveriam estar mais focados no desenvolvimento de competências para a sua prática profissional, dividem o seu tempo de estudo e dedicação, algumas vezes até priorizando a realização de cursos preparatórios, em horários noturnos, comprometendo por vezes a saúde física e emocional.

O número de dados gerados em um período de doze semestres de um curso de medicina, pode responder a muitas indagações com a utilização da mineração de dados. Esse estudo trouxe algumas avaliações agrupadas em médias de módulos. Estudos posteriores podem ser realizados com as avaliações separadas por competências. Uma vez que outras provas de RM trazem no processo seletivo avaliação prática e não apenas de conhecimento por testes escritos, acreditamos que seja necessário outro olhar sobre as avaliações e desempenhos dos estudantes.

Esse estudo possui algumas limitações. Dentre elas destacamos tratar-se de dados de estudantes de uma única faculdade, o que não permite a generalização para outra população. Um outro aspecto limitante é sobre a composição das notas das atividades teóricas do banco de dados que agregam diferentes domínios; acreditamos que o desmembramento dessas notas, separando o componente relacionado aos aspectos cognitivos, talvez possa contribuir melhor com a definição de um modelo, diminuindo inconsistências. A análise da taxonomia das questões do processo seletivo de residência médica também podem auxiliar na melhor compreensão do modelo.

Como conclusão do estudo, o modelo proposto utilizou árvore de decisão para predizer a reprovação no processo seletivo de residência médica em concluintes de uma faculdade de medicina na região nordeste do Brasil. Uma melhor acurácia foi obtida utilizando os módulos do ano/período imediatamente anterior ao período do internato do curso médico. A busca de predição de baixa performance na educação é desejável por todos, e esse estudo espera contribuir também reforçando sobre a importância da utilização da mineração de dados com esse objetivo na gestão e acompanhamento de estudantes em cursos de medicina.

## 6.Referências


Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. *Tech Know Learn, 24*, 567–598 (2019). doi: 10.1007/s10758-019-09408-7

Aragão, J.C.S., Casiraghi, B., Coelho, O.C., Sarzedas, A. R. M., Peloggia, S.M.M., & Huguenin, T.F. (2018). Avaliação de Questões de Prova de Concursos de Residência Médica. *Revista Brasileira de Educação Médica*, *42*(2), 26-33. doi: 10.1590/1981-52712015v421n2rb20170016

Baker, R.S., Isotani, S., & Carvalho, A.M.J.B. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, *19*(2), 3-13. doi: 10.5753/rbie.2011.19.02.03

Carvalho, H.M. (2014). *Aprendizagem de máquina voltado para mineração de dados: árvore de decisão.* (Dissertação). Universidade de Brasília, Faculdade UnB Gama, Engenharia de Software, Brasília. Recuperado de: http://fga.unb.br/tcc/software/tcc-2014.1-engenharia-de-software/hialo-muniz-carvalho/ v1-tcc-hialo-muniz.pdf

Costa, E. B., Fonseca, B., Santana, M.A., Arajo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247-256. doi: 10.1016/j.chb.2017.01.047

Costa, E.B., Baker, R.S., Amorim, L.,Marinho, T.(2012).Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em Informática da Educação.

Cretton, N.N., & Gomes, G.R.R. Aplicação de técnicas de mineração de dados na base de dados do ENADE com enfoque nos cursos de medicina. (2016). *Acta Biomedica Brasiliensia, 7*(1),74-89. Doi: 10.18571/acbm.100

Dijkstra, J., Van der Vleuten, C. P., & Schuwirth, L. W. (2010). A new framework for designing programmes of assessment. *Advances in health sciences education: Theory and practice*,*15*(3), 379–393. doi: 10.1007/s10459-009-9205-z

Franco, R. S., Franco, C.A.G.S., Portilho, E.M.L., & Cubas, M.R. (2014). O conceito de competência: uma análise do discurso docente. *Revista brasileira de educação médica*, *38*(2), 173-181. doi: 10.1590/S0100-55022014000200003

Girardi, S.N., Carvalho, C.L., Maas, L. W., Araujo, J.F., Massote, A.W., Van Stralen, A.C. S., & Souza, O.A. (2017) Preferências para o trabalho na atenção primária por estudantes de medicina em Minas Gerais, Brasil: Evidências de um experimento de preferência declarada. *Cadernos de Saúde Pública*, *33*(8), e00075316. doi: doi.org/10.1590/0102-311x00075316

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. (2009). The Weka data mining sogtware: an update. SIGKDD Explor.News., 11(1): 10-18.

Han, J., Kamber, M., & Pei, J. (2006) *Data mining: Concepts and techniques*. Simon Fraser University.

Pugh, D., Bhanji, F., Cole, G., Dupre, J., Hatala, R., Humphrey-Murto, S., Touchie, C., ...Wood, T. J. (2016). Do OSCE progress test scores predict performance in a national high-stakes examination?. *Medical education*, *50*(3), 351–358. doi: 10.1111/medu.12942

*Resolução CNRM nº 02, de 17 de maio de 2006*. Dispõe sobre requisitos mínimos dos Programas de Residência Médica e dá outras providências. Recupera de: http://portal.mec.gov.br/docman/documentos-pdf/512-resolucao-cnrm-02-17052006

*Resolução nº 3, de 20 de junho de 2014.* Institui Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina e dá outras providências. Recuperado de: https://abmes.org.br/legislacoes/detalhe/1609

Roman, A.B., Sánchez-Guzmán, A., & García, R. (2013). Minería de datos educativa: Una herramienta para la investigación de patroness de aprendizaje sobre un contexto educativo. *Latin-American Journal of Physics Education*, *7*(4), 662-668.

Romero, C., & Ventura, S. (2010) Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618. doi: 10.1109/TSMCC.2010.2053532

Manjarres, A.V., Sandoval, L.G.M., & Suárez, M.S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review*, *33*,235-266. doi:10.1344/der.2018.33.235-266

Sesat, D.B., Milem, J.F., McIntosh, K.L., Bryan, W.P. (2017). Coupling Admissions and Curricular Data to Predict Medical Student Outcomes. *Researchin High Education 58*,295–312. doi: 10.1007/s11162-016-9426-y

Troncon, L.E.A., & Pinto, M.P.P. (2014). Avaliação do estudante: Aspectos gerais. Revista da Faculdade de Medicina de *Ribeirão Preto* *47*(3),314-23.doi:10.11606/issn.2176-7262.v47i3p314-323

Van der Vleuten, C. P.M., Schuwirth, L. W.T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best practice & research. Clinical obstetrics & gynaecology*, *24*(6), 703–719. doi:10.1016/j.bpobgyn.2010.04.001

Yassin, K., & Stefan, K. (2018). A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. *Medical Teacher*, *40*(11):1123-1129. doi: 10.1080/0142159X.2018.1472370

Weka.University of Wako Datamining Software in JAVA. Disponível em <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em 7 de dezembro de 2020.