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

Utilizando a Mineração de Dados como suporte à predição da reprovação em cursos técnicos integrados do CEFET

Gualberto Rabay Filho[1], Seiji Isotani[2], Carlos Diego Nascimento Damasceno[3]

Resumo

O baixo desempenho escolar está associado à reprovação e à evasão que representam um grave problema na educação. Quanto mais precoce se detectar os riscos de fracasso escolar, mais efetivamente se poderá tomar medidas para sua mitigação. Neste trabalho, o uso de técnicas de Mineração de Dados Educacionais estabelece um modelo de predição de reprovação para alunos do ensino técnico integrado de um campus do CEFET MG. Foram analisados dados referentes aos anos 2018/2019, utilizando-se o algoritmo de classificação J48 da ferramenta Weka. O primeiro ano do ensino médio é o que apresenta maior índice de reprovação e onde os resultados da predição mostraram uma acurácia de 88,6%.

Abstract

Poor performance in school is associated with failure and dropping out, which represents a serious problem in education. The earlier the risks of school failure are detected the more effectively measures can be taken to mitigate them. In this work the use of Educational Data Mining techniques establishes a model for predicting failure for students in the integrated technical education of a CEFET MG campus. It analyzed data referring to the years 2018/2019 using the classification algorithm J48 of the Weka tool. The first year of high school is the one that presents the highest rate of failure and in which the results of the prediction showed an accuracy of 88.6%.

[1] Pós-Graduando em Computação Aplicada à Educação, USP, rabay@cefetmg.br.

[2] Orientador, Universidade de São Paulo, sisotani@icmc.usp.br.

[3] Orientador, Universidade de São Paulo, damascenodiego@alumni.usp.br.

Cite as: Filho, G. & Isotani, S. & Damasceno, C. (2020). Utilizando a Mineração de Dados como suporte à predição da reprovação em cursos técnicos integrados do CEFET. 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

Uma educação pública e de qualidade é uma demanda básica da nossa sociedade e cabe às instituições de ensino envidar esforços para atender de forma eficaz esta demanda. No ensino médio o fracasso escolar é o maior desafio que gestores e professores enfrentam no seu dia a dia. Dois indicadores de fracasso escolar que são facilmente quantificados e tem uma correlação forte são a reprovação e a evasão escolar.

O fracasso escolar no ensino médio é um problema mundial como atesta documento elaborado pelo Departamento de Educação dos Estados Unidos [U.S, 2017]. Nesta cartilha são detalhadas quatro recomendações para se reduzir a evasão: monitoramento e intervenção proativa ao se verificar problemas acadêmicos; prover atendimento individualizado e intensivo aos estudantes em risco; engajar os estudantes em currículos e programas significativos e criar comunidades personalizadas para apoio no monitoramento e suporte dos alunos. Na Coréia do Sul, onde apenas 1,4% do total dos estudantes matriculados no ensino médio em 2016 evadiram, o baixo desempenho escolar com um percentual de 21,6% foi uma das principais causas desta taxa de evasão [Lee e Chung 2019].

De acordo com Desjardins et al. (1999), um bom desempenho escolar melhora a retenção e, portanto, o êxito escolar seria o melhor preditor de permanência dos alunos. Desta forma, segundo os autores, estima-se que menores índices de reprovação levarão também a uma redução da evasão. Veloso (2015) analisa os diversos fatores que causam a evasão escolar no ensino médio como: fatores sociais, fatores educacionais, fatores de localização e fatores econômicos.

A questão da evasão se torna premente considerando que a Lei de Diretrizes e Bases - LDB 9.394/96 (Brasil, 1996) tornou obrigatória a oferta e gratuidade deste nível de educação, colocando grande responsabilidade na escola sobre o acompanhamento e notificação aos órgãos responsáveis quando ocorra ausência mais prolongada do aluno sem uma devida justificativa. Do ponto de vista da gestão escolar quanto mais rapidamente forem apontados os indícios de uma futura reprovação e a consequente evasão mais rapidamente as ações preventivas e corretivas podem ser realizadas no sentido de mitigar o fracasso escolar.

A intervenção precoce junto ao aluno e docentes pode reduzir bastante a reprovação e consequentemente melhorar os índices de retenção. Um sistema desenvolvido na Universidade de Purdue, baseado em ferramentas analíticas da aprendizagem, realiza em tempo real estimativas de risco acadêmico e a partir da configuração pode disparar mensagens de texto, e-mails, alertas nas plataformas virtuais, entre outros [Arnold e Pistilli 2012]. Uma boa interação com uma comunicação fluida e rápida entre professores, gestores e alunos é o primeiro item dos 7 princípios das boas práticas de educação preconizados por Chickering e Ehrmann (1996).

Para se obter informações que agilizem o processo decisório, Manhães et al. (2011) apontam que os usos de recursos computacionais permitem identificar e classificar do ponto de vista de importância os principais fatores que levam à evasão. Somente o uso de

métodos computacionais pode permitir identificar padrões em grandes coleções de dados educacionais, o que de outra forma seria impossível de se obter [Romero e Ventura 2013].

O Centro Federal de Educação Tecnológica de Minas Gerais (CEFET MG), uma instituição pública multicampi, com oferta de cursos em diversos níveis e modalidades, sofre dos mesmos problemas de fracasso escolar já citados e uma ação efetiva que auxilie a gestão a melhorar seus indicadores será de grande utilidade. Uma boa política de inclusão social através de mecanismos de ofertas de bolsas é um elemento importante na melhoria dos índices de retenção dos alunos principalmente em municípios com baixa renda per capita como Nepomuceno [Ramalho 2013], onde se localiza a unidade que será o objeto de estudo deste trabalho. Com relação às taxas de aprovação, a autora mostra que não há grande diferença entre os alunos bolsistas e os não bolsistas [Ramalho 2013]. Mesmo com o apoio das bolsas e todos os projetos assistenciais, os índices de reprovação e evasão seguem elevados e, portanto, outras ações se fazem necessárias.

No biênio 2018-2019 apenas 36,98% dos alunos do primeiro ano do Ensino Médio do campus foram aprovados, o que representa um número preocupante. Uma das razões que provoca este baixíssimo índice de aprovação é a mudança que sofre o aluno da série inicial. Oriundo de um ensino público fundamental deficiente, que é ofertado em um turno ele abruptamente passa a estudar em tempo integral, tendo um número elevado de disciplinas e um nível de cobrança bem maior ao que estava acostumado. Outro dado observado é que os alunos repetentes do primeiro ano apresentam o mesmo índice de aprovação daqueles que estão fazendo o ano pela primeira vez. Isso indica que a recuperação de alunos com baixo desempenho não está ocorrendo de forma adequada e cabe mais atenção da escola a estes casos de repetição.

O que este trabalho propõe é definir um modelo de predição de reprovação para Ensino Médio Integrado do CEFET MG de Nepomuceno com o objetivo de nortear e agilizar as ações dos gestores com o fim de mitigar o fracasso escolar. Para isso utilizaram-se os conceitos de Mineração de Dados Educacionais com algoritmos de Aprendizagem de Máquina para analisar dados dos anos 2018 e 2019.

O restante deste trabalho está organizado da seguinte forma: na Seção 2 é apresentada a fundamentação teórica onde são abordados de forma resumida os assuntos que norteiam o projeto; para a Seção 3 foram selecionados trabalhos relacionados ao nível e à modalidade de ensino abordada; na Seção 4 se discute a metodologia utilizada; na Seção 5 são mostrados os resultados obtidos e a discussão; finalmente na Seção 6 são apresentadas as conclusões e propostas de trabalhos futuros.

2. Fundamentação Teórica

A Descoberta de Conhecimento em Bases de Dados (em inglês, *Knowledge Discovery in Database* - KDD), é definida como todo o processo envolvido na descoberta e construção do conhecimento a partir de dados obtidos de diversas fontes [Fayyad, Piatetsky-Shapiro, Smyth, 1996]. O autor separa este conceito do termo Mineração de Dados que representa uma das etapas do processo de aquisição do conhecimento. Outros autores como Baker et al. (2011), Baker e Yacef (2009) tratam os termos KDD e mineração de dados com sinônimos. A KDD possui aplicações em diversas áreas do conhecimento e ganha cada

vez mais relevância no auxílio à tomada de decisão considerando que atualmente se produz um enorme volume de dados através das redes sociais, dos sites de buscas, dos sensores, etc.

A aplicação na área da educação, conhecida como Mineração de Dados Educacionais (MDE), é relativamente recente e foi impulsionada pelo maior uso de recursos computacionais na educação e pela educação à distância. A MDE tem como objetivo principal desenvolver métodos que possam analisar um grande volume de dados gerados no ambiente educacional [Baker et al. 2011].

Uma revisão sistemática na área de MDE aplicada ao ensino superior, cobrindo o período de 2000 a 2017, buscou determinar como esta técnica juntamente com a Analítica da Aprendizagem (em inglês, *Learning Analytics*) poderiam auxiliar na resolução dos principais problemas da educação. Verificou-se, nesta revisão, quais métodos da MDE seriam os mais apropriados para cada tipo de problema e o método de classificação foi o mais aplicado em questões como predição de desempenho, evasão, comportamento em cursos on-line e aprendizagem à distância entre outros [Aldowah et al. 2019].

Um método de classificação tem como objetivo separar uma coleção de dados, denominada como conjunto de treinamento, em classes que constituem um modelo. A partir deste modelo é possível então se fazer predições de classificação ao se utilizar itens de outra coleção de dados, denominada conjunto de teste [Costa et al. 2012]. Diferentes métodos de classificação, como classificação estatística, árvore de decisão, redes neurais entre outros são citados por Romero et al. (2008).

O algoritmo de classificação J48 cria uma árvore de decisão cujo nó raiz representa o atributo mais significativo usando a estratégia de dividir um problema complexo em problemas menores. Este algoritmo é uma implementação em Java na ferramenta Weka do algoritmo C4.5 proposto por Quinlan (1993 apud Costa et al. 2012). Weka é uma ferramenta de mineração de dados *open source* que implementa diversos algoritmos e apresenta uma interface capaz de apresentar gráficos e tabelas além de ter uma API que possibilita incorporá-la como uma biblioteca a outros aplicativos.

Uma técnica muito usada com a classificação é a seleção de atributos, que traz vantagens como uma melhoria na visualização e interpretação dos dados, uma redução no tempo de processamento e no espaço de armazenamento [Guyon e Elisseeff 2006]. Esta técnica permite associar a cada atributo da coleção um valor, estabelecendo assim o seu grau de importância para o resultado da predição. A técnica *Information-Gain Attribute Ranking (InfoGain)* calcula o ganho de informação baseado na entropia para determinar a relevância individual do atributo [Parmezan et al. 2012]. *Infogain* é uma das técnicas disponíveis na ferramenta Weka e se caracteriza por sua simplicidade e por apresentar bons resultados [Paes et al. 2013].

Para se fazer a avaliação de um determinado método de classificação é necessário medir a qualidade e a precisão dos algoritmos usados ao se predizer os itens de dados de uma classe. Neste sentido, a matriz de confusão é uma ferramenta que permite a visualização do desempenho de um algoritmo. A matriz de confusão exibe para cada classe a quantidade de classificações corretas com relação ao total classificado de acordo com o modelo [Han et al. 2011]. Quando o conjunto de dados tem apenas duas classes, uma

delas se chama positivo e a outra negativa e a matriz de confusão tem as entradas mostradas na Tabela 2.1.

Tabela 2.1 Matriz de Confusão

Classificação Correta	**Classificado como**		**Total de Instâncias**
	Positivo	Negativo	
Positivo	Positivo Verdadeiro (TP)	Falso Negativo (FN)	Positivo (P)
Negativo	Falso Positivo (FP)	Verdadeiro Negativo (TN)	Negativo (N)

As entradas da matriz são descritas como:

- True Positive (TP) - é o número de instâncias positivas classificadas como positivas.
- True Negative (TN) - é o número de instâncias negativas classificadas como negativas.
- False Negative (FN) - é o número de instâncias positivas classificadas como negativas.
- False Positive (FP) - é o número de instâncias negativas classificadas como positivas.

A partir da matriz de confusão são determinadas as métricas mais usadas para medir o desempenho do classificador, que na área de recuperação da informação são [Bramer, 2016]:

- Acurácia - representa a porcentagem das instâncias corretamente classificadas (positivas ou negativas) sobre o total de instâncias,

á =(+)/(+)

- Precisão - ou confiabilidade positiva, representa a percentagem de instâncias classificadas como positivas que são realmente positivas,

ã = /(+)

- Sensibilidade (*recall*) - representa a percentagem de instâncias positivas que foram corretamente classificadas como positivas,

= /(+)

- *F measure* ou *F1 score* - é calculada como a média harmônica entre a Precisão e a Sensibilidade.

1 = 2. ã . /(ã +)

Para um modelo com duas classes, por exemplo, aprovado e reprovado, a acurácia apresenta bons resultados quando há um equilíbrio na distribuição das instâncias pelas duas classes. Quando há um desbalanceamento na quantidade de instâncias entre as classes, as medidas de precisão, sensibilidade e F1 são mais adequadas [Han et al. 2011].

As curvas ROC (do inglês, *Receiver Operating Characteristic*) permitem comparar de forma visual dois classificadores. Para as classes binárias elas mostram a razão entre o quanto um modelo pode reconhecer com acurácia os casos positivos e os casos identificados erroneamente como negativos [Han et al. 2011]. A área sob a curva ROC,

ou AUC (do inglês, *Area Under the Curve*), mostra-se como uma interessante métrica de desempenho atribuindo valores mais baixos para classificadores mais desbalanceados além de ser um indicativo da separação das classes positivas e negativas na classificação [Bradley 1997].

3. Trabalhos Relacionados

As técnicas de MDE têm sido utilizadas para analisar a reprovação ou evasão escolar em diversos níveis de ensino e nas modalidades presenciais ou à distância usando diferentes abordagens. Nesta seção foram revisados alguns trabalhos que trataram do ensino fundamental, do ensino médio e do ensino médio integrado com o técnico.

Com o objetivo de auxiliar a Prefeitura de Juiz de Fora na tomada de decisão sobre o problema da evasão escolar no ensino fundamental foram analisados dados de 43.672 alunos referentes aos anos 2017 e 2018 [Sales et al. 2019]. Os autores utilizaram o método de classificação Floresta Randômica Ponderada (WRF) obtendo uma média de 70% de precisão e uma cobertura de 97%.

Saraiva et al. (2019) apresenta uma proposta de predição de evasão considerando o desempenho acadêmico e a situação sócio econômica de alunos de um Curso Técnico em Informática do Instituto Federal do Ceará (IFCE). Foram analisados dados de período de 2009 a 2019, obtidos de 13 campi sob a ótica de diferentes algoritmos de classificação. O curso apresentou no mesmo período uma taxa de evasão de 49,5%.

Silva e Nunes (2015) analisam uma base de dados de alunos do ensino médio de uma escola particular, no período de 2011 a 2014 utilizando as técnicas de classificação com o algoritmo J48. A avaliação foi feita por cada série e o objetivo foi predizer a situação final de aprovado ou reprovado a partir dos atributos de notas, ser bolsista, cidade de origem etc. Os resultados indicaram que alunos bolsistas tendem a evadir menos e que normalmente os índices de reprovação são menores para o terceiro ano.

Neves Junior et al. (2019) apresentam um estudo usando técnicas de regressão quantílica com o objetivo de predizer a aprovação e reprovação de alunos do ensino médio no estado de Pernambuco. Foram utilizados dados referentes a vários indicadores educacionais do INEP do ano de 2016, entre eles a Taxa de Eficiência Escolar que é a razão entre as reprovações e as aprovações.

O algoritmo J48 foi aplicado numa base de mais de 300 mil alunos dos cursos Técnicos de Nível Médio do SENAI, no período de 2012 a 2014. Foram estabelecidos vários fatores relacionados com a evasão como: carga horária do curso, idade, situação ocupacional etc [Veloso 2015].

Baseado no censo escolar dos anos 2014, 2015 e 2016 dos estados do Ceará e de Alagoas, utilizaram-se as técnicas de Regressão Logística e Indução de Regras e a ferramenta SPSS da IBM para determinar as causas principais da evasão [Calixto et al. 2017]. Entre os fatores considerados se encontram: idade, etapa e modalidade de ensino, ausência de laboratórios, quantidade de alunos por turma, falta de atendimento adequado a alunos especiais entre outras.

Colpani (2018) utilizou os indicadores do Censo Escolar de 2017, referentes ao estado do Pará e aplicou as técnicas de regressão linear e correlação para analisar o problema da

evasão no ensino médio. Foram estabelecidas correlações entre diversos indicadores que apontou a Taxa de Distorção Idade-Série como sendo a variável mais associada à evasão.

Neste trabalho o objetivo é realizar predição de reprovação para alunos do nível médio integrado em uma unidade do CEFET MG e se diferencia dos anteriores por realizar esta predição de forma mais precoce a partir do desempenho do aluno por bimestre.

4. Metodologia

No Campus de Nepomuceno existem três cursos na modalidade Ensino Médio Técnico Integrado: Eletrotécnica, Mecatrônica e Redes de Computadores. O ingresso nos cursos é feito através de vestibular e cada curso oferece 35 vagas anuais. Um aluno pode ser reprovado no máximo uma vez em cada série, perdendo o direito à vaga na segunda reprovação. Para realizar o trabalho foram seguidos os passos sugeridos em Fayyad (1996) mostrados na Figura 4.1 na obtenção do conhecimento.

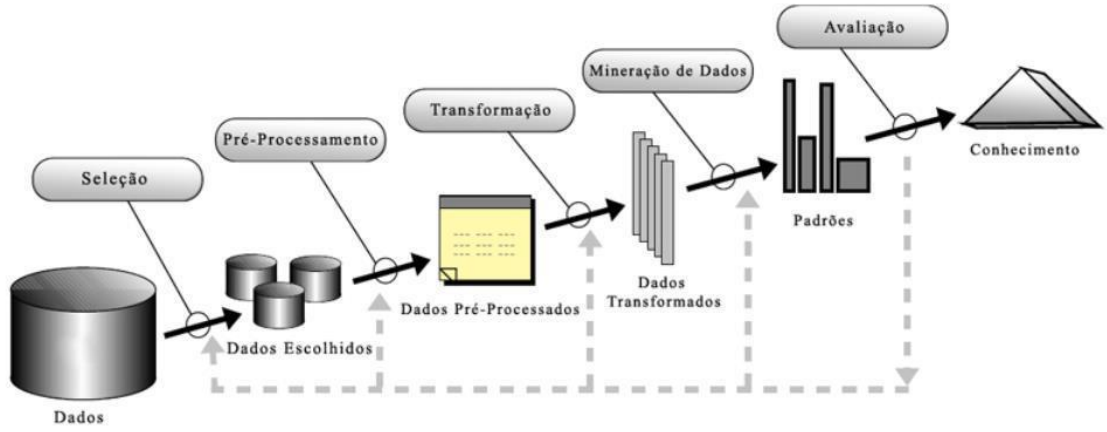

Figura 4.1. Passos que compõe a KDD [Camilo e Silva 2019 apud Fayyad et al, 1996]

4.1. Seleção de Dados

Para se cumprir os objetivos do trabalho foram selecionados dados referentes aos alunos matriculados nos três cursos técnicos integrados nos anos de 2018 e 2019, que são os anos a partir dos quais começou-se a realizar um registro mais consistente do desempenho dos alunos a cada bimestre letivo. Na Tabela 4.1 mostra-se o desempenho dos alunos separados por anos.

Entre os estudantes matriculados 52,81% eram do sexo feminino e 49,09% do sexo masculino. A distribuição de bolsas de permanência favoreceu 45,65% do total de estudantes. Observa-se um número elevado de alunos repetentes, que representa aqueles alunos que ao menos uma vez foram reprovados em algum dos anos escolares, com 35,32% do total dos alunos matriculados.

Tabela 4.1. Situação dos alunos

Série	Situação	Quant.	%

1a	Aprovado	115	36,98
	Reprovado	114	36,67
	Evadido	82	26,37
2a	Aprovado	104	80,0
	Reprovado	12	9,2
	Evadido	14	10,8
3a	Aprovado	105	93,75
	Reprovado	3	2,67
	Evadido	4	3,57

4.2. Pré-processamento

O fato de não existir um sistema integrado de informações no CEFET MG que tivesse todos os dados potencialmente relacionados à reprovação exigiu-se trabalhar com diversas fontes e formatos diferentes. Os dados referentes à frequência do aluno em 2018 foram obtidos de forma manual a partir de diários impressos de cada disciplina e totalizados em planilhas eletrônicas. Os dados sobre a informação socioeconômica (aluno com bolsa de permanência) foram obtidos de outra fonte em arquivos de texto. Da mesma forma, os dados sobre trancamento, cancelamento ou abandono que caracterizam a evasão foram fornecidos através de um relatório de situação escolar realizado ao final de cada ano. Os atributos de quantidade de notas abaixo da média em cada bimestre escolar são contabilizados a partir de um relatório anual em formato eletrônico realizado a partir de 2018. A partir desta etapa escolheram-se os seguintes atributos que potencialmente poderiam auxiliar na predição da reprovação:

1. Sexo
2. Série (1a, 2a e 3a)
3. Curso (Redes, Mecatrônica, Eletrotécnica)
4. Repetência (se o aluno possui reprovação escolar anterior)
5. Bolsa permanência (indica a fragilidade socioeconômica)
6. Faltas (quantidade de faltas anual)
7. Quantidade de notas abaixo da média em cada bimestre (4 atributos - do 1o ao 4o bimestre)
8. Situação final do aluno (aprovado, reprovado, trancado, cancelado, abandono)

Todos os dados coletados das diversas fontes foram agregados em planilhas eletrônicas e em seguida foram gerados arquivos CSV por cada série.

4.3. Transformação

A quantidade de alunos analisados representa aqueles que se mantiveram matriculados até o encerramento dos anos 2018 e 2019 e, portanto, tem situação final definida como aprovado ou reprovado. Na Tabela 4.2 é mostrada a quantidade de alunos no início do ano.

Tabela 4.2. Quantidade de Alunos (Início do Ano)

Série	Alunos
1a	311

2a	130
3a	112

Os casos registrados ao longo do ano de evasão (cancelamento, trancamento ou abandono) foram retirados do conjunto por não serem objeto da análise e serão discutidos em outra seção. Foram corrigidas algumas inconsistências nos dados e desconsiderado por exemplo o atributo ano que era irrelevante neste contexto já que as análises seriam feitas em separado por série. Nesta etapa também foram gerados os arquivos para o formato adequado ao processamento pela Weka.

4.4. Mineração de Dados

Para esta fase utilizou-se a ferramenta Weka aplicada a cada série em separado considerando que as taxas de evasão e reprovação na primeira série sempre foram bem maiores que as das séries seguintes. Foi feita uma primeira análise considerando todos atributos e então a partir de uma análise das árvores de decisão obtidas verificou-se que alguns destes atributos não tinham relevância para os objetivos de predição de reprovação e, portanto, poderiam ser removidos. Fez-se o uso da técnica de seleção *Infogain* para validar a exclusão dos atributos irrelevantes. Finalmente foi aplicado o J48 para cada série do curso em separado.

5. Resultados e discussões

Após as etapas iniciais do processo de descoberta do conhecimento para este problema são avaliados nesta seção os resultados da aplicação na etapa de mineração de dados.

Na Tabela 5.1 são mostrados os atributos e seus respectivos pesos de acordo com o algoritmo de seleção *Infogain* para a primeira série. Foram excluídos os atributos curso, repetente, bolsa e o sexo que obtiveram pontuação muito baixa no *rank*. A efetividade da seleção pode ser comprovada ao executar-se o J48 considerando ou não os atributos com baixa valoração e verificando-se que não há alterações na configuração da árvore de decisão. Este mesmo procedimento foi aplicado aos outros anos.

Tabela 5.1. Rank dos atributos *Infogain* do 1o. Ano

Atributos	Rank
notabaixa2	0,61275
notabaixa3	0,59426
notabaixa1	0,56617
notabaixa4	0,51815
faltas	0,34045
curso	0,01266
repetente	0,00634
bolsa	0,00496
sexo	0,00236

Considerando que a primeira série é a que apresenta a maior taxa de reprovação e consequentemente a maior taxa de evasão dedicaremos maior importância na análise da mesma. Conforme explicado na seção anterior foram analisados no primeiro ano 229 alunos, dos quais 115 foram aprovados e 114 foram reprovados. O algoritmo J48

apresentou 88,64% das instâncias corretamente classificadas. Observando a matriz de confusão na Tabela 5.2, identificaram-se 10 ocorrências erroneamente classificadas como reprovadas (falso negativo) e 16 ocorrências erroneamente classificadas como aprovadas (falso positivo). O valor obtido para a precisão é 0,868 e de 0,913 para a sensibilidade e um F1 de 0,89. Para a solução do problema de se fazer uma detecção precoce da reprovação é desejável um valor mais baixo de falso positivo já que ele representa a quantidade alunos que erroneamente foram identificados como aprovados mas que reprovaram. A taxa de falso positivo obtida foi de 14,0% que é um bom valor para o propósito deste trabalho.

Tabela 5.2. Matriz de Confusão do 1o. Ano

Situação	Classe=A	Classe=R
Aprovados	105	10
Reprovados	16	98

A árvore de decisão obtida da análise dos dados é mostrada na Figura 5.1.

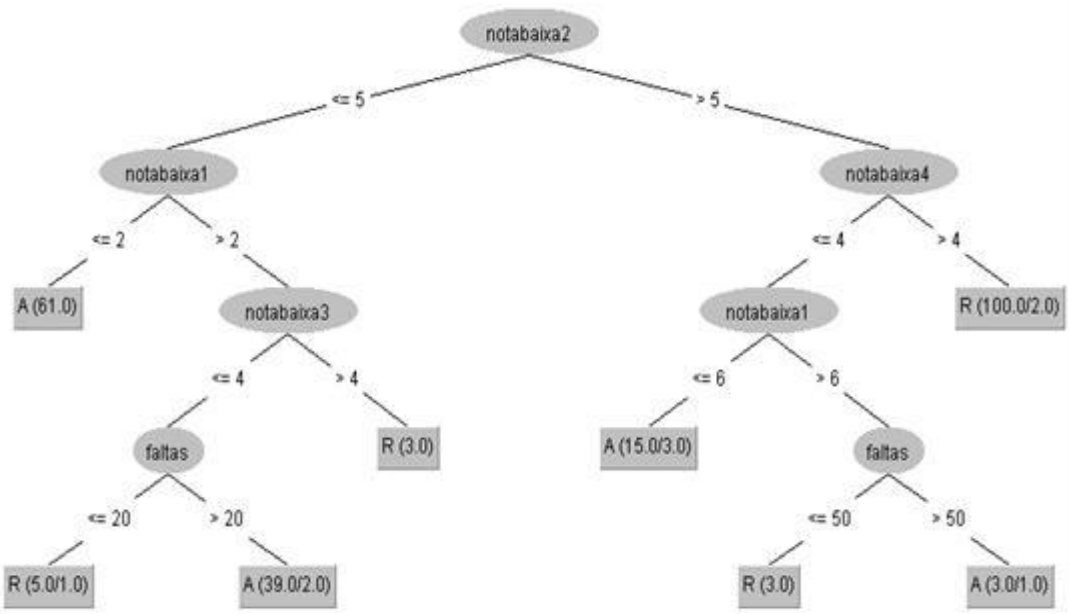

Figura 5.1. Árvore de Decisão 1o. Ano

A quantidade de notas abaixo da média em um bimestre representa um fator chave para se determinar uma situação final de reprovação. Na raiz da árvore aparece o atributo que representa a quantidade de notas baixas do 2º bimestre. A maior quantidade das reprovações (87,7%) ocorre para os alunos com mais de 5 notas abaixo da média no segundo bimestre e mais de 4 no quarto bimestre. A quantidade de faltas é um atributo de pouco peso e um fato interessante é que nestes nós da árvore observa-se que alunos aprovados são aqueles com mais faltas. Embora não tenha sido avaliada a distribuição das

faltas ao longo do ano letivo o que se observa na prática é que alunos que têm sua aprovação garantida no 3o bimestre tendem a faltar mais no último bimestre o que explica este resultado que parece ser uma contradição entre o bom desempenho e o número maior de faltas.

No segundo ano já se observa um índice de aprovação de 80,0% contra apenas 36,98% do primeiro ano. Os atributos com maior valor segundo a classificação do *Infogain* e que foram usados para a predição pelo J48 são mostrados na Tabela 5.3.

Tabela 5.3. Rank dos atributos *Infogain* do 2o Ano

Atributos	Rank
notabaixa1	0,2356
notabaixa4	0,19023
notabaixa2	0,16173
faltas	0,15498
notabaixa3	0,13569

O algoritmo classificou corretamente 93,1% das instâncias com uma precisão de 0,944. O valor obtido para a sensibilidade é de 0,981 e de 0,962 para o *F1*.

Tabela 5.4. Matriz de Confusão do 2o. Ano

Situação	Classe=A	Classe=R
Aprovados	102	2
Reprovados	6	6

A árvore de decisão gerada aponta a quantidade notas baixas do quarto bimestre seguida das notas baixas do primeiro bimestre como os elementos decisores da reprovação. Valem para o segundo ano as mesmas observações relativas às faltas feitas para o primeiro ano.

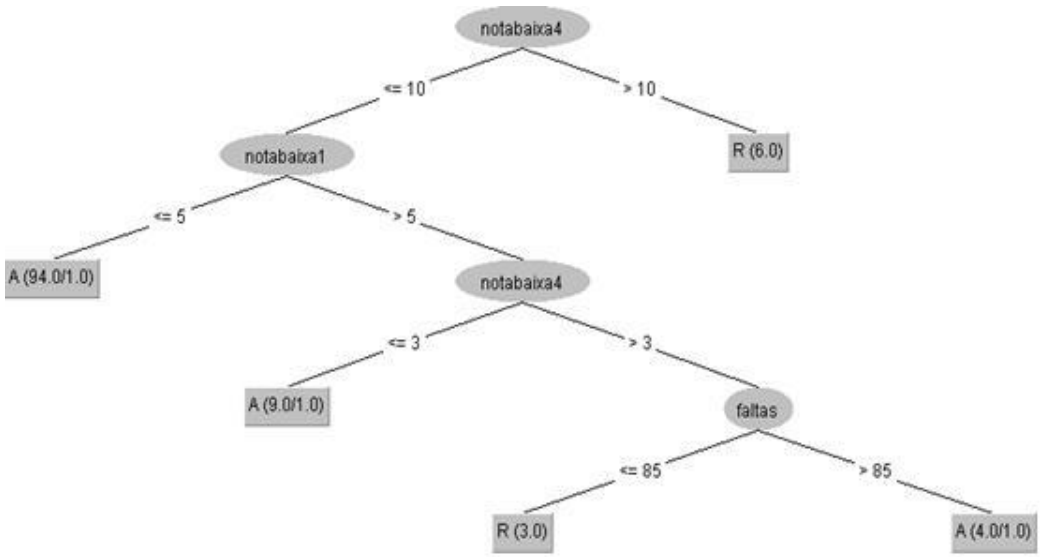

Figura 5.2. Árvore de Decisão 2o. Ano

Finalmente o terceiro ano tem o melhor índice de aprovação (93,75%) e também os mais baixos índices de evasão (3,57%) o que de certa forma corrobora a relação entre o bom desempenho e a permanência do aluno. Neste ano os atributos de notas abaixo da média foram considerados irrelevantes pelo *Infogain* e obtiveram pontuação nula. Apenas as faltas dos alunos contribuíram para a decisão da classificação sobre a aprovação, ainda que com um valor de 0,112423 que é bem abaixo dos valores atribuídos aos atributos selecionados do primeiro e do segundo ano. Na Tabela 5.5 são apresentados os ranks dos atributos *Infogain* do 3o ano.

Tabela 5.5. Rank dos atributos *Infogain* do 3o. Ano

Atributos	Rank
faltas	0,112423
curso	0,027619
bolsa	0,004123
sexo	0,003467
repetente	0,000273

Com 97,22% das instâncias classificadas corretamente, as 3 reprovações foram classificadas incorretamente como aprovadas (vide Tabela 5.6).

Tabela 5.6. Matriz de Confusão do 3o. Ano

Situação	Classe=A	Classe=R
Aprovados	105	0
Reprovados	3	0

Com estes números a precisão obtida foi de 0,972 com uma sensibilidade de 1,0 (0 falsos negativo) e um *F1* de 0,986 (Tabela 5.7). Estes resultados indicam que o uso da predição para o terceiro ano não agrega muito valor. As faltas são acumuladas ao longo do ano letivo e, portanto, não servem com uma boa variável de predição da reprovação como pode ser visto na árvore de decisão gerada (Figura 5.3).

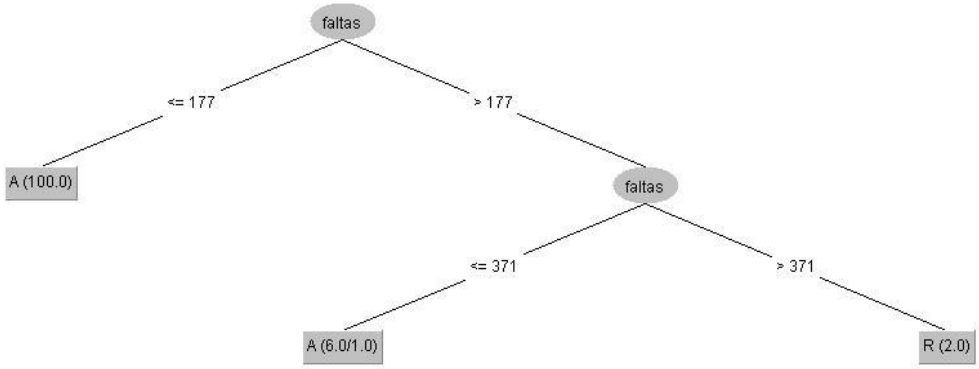

Figura 5.3. Árvore de Decisão 3o Ano

Com poucos itens pertencentes à classe de reprovados nos segundos e terceiros anos, os valores dos parâmetros avaliados são melhores que os do primeiro ano, mas há uma menor confiabilidade nos resultados. O valor do Área ROC (AUC) para os 3 anos mostrados na Tabela 5.7 indicam essa limitação provocada pelo desbalanceamento das classes.

Tabela 5.7. Parâmetros por ano

Série	Acurácia	Precisão	Sensibilidade	F1	Área ROC
1a	0,8864	0,868	0,918	0,890	0,941
2a	0,931	0,944	0,981	0,962	0,718
3a	0,9722	0,972	1,00	0,986	0,714

6. Conclusão

Foram analisados os dados de alunos do Ensino Médio Integrado no Campus de Nepomuceno do CEFET MG nos anos de 2018 e 2019 com o objetivo de criar um modelo de predição da reprovação. A maior dificuldade encontrada foi a diversidade de fontes de dados que tiveram seus dados extraídos de forma manual na maior parte dos casos. Usou-se o algoritmo J48 na fase mineração de dados após a seleção de atributos feita pelo algoritmo *Infogain* na suíte *Weka*. Os resultados obtidos foram satisfatórios e atenderam ao objetivo da criação de um modelo de predição da reprovação, ainda que este possa ser aperfeiçoado à medida que tenhamos um histórico de dados mais longo. No primeiro ano, no qual historicamente existem os piores indicadores de desempenho, o algoritmo classificou corretamente 88,64% das instâncias. Embora não fosse o objetivo deste trabalho estabelecer correlações entre as taxas de reprovação e evasão, diversos autores já sinalizaram este fato e, portanto, produzir ações para reduzir a reprovação poderá melhorar os indicadores de evasão. O modelo obtido servirá de apoio à tomada de decisões e permitirá que os gestores e professores possam atuar de forma mais objetiva e rápida junto aos alunos que apresentam maiores possibilidades de risco escolar.

Este trabalho pode ser ampliado para atender a outras modalidades e níveis de ensino como os cursos técnicos subsequentes noturnos, os cursos superiores e o ensino técnico à distância. Ainda poderá ser ampliado com dados dos outros oito campi do CEFET MG o que possibilitaria entender de forma mais abrangente como as diferentes unidades estão tratando o tema. Outros atributos podem ser considerados como por exemplo, a distribuição de faltas nos bimestres, o que certamente pode ser um fator importante na predição mais prematura. Uma maior integração dos sistemas acadêmicos com os cadastros da área de assistência estudantil e do registro escolar irá facilitar a elaboração de modelos que sejam mais eficazes na predição. Como consequência haverá agilidade nos alertas para que docentes e gestores atuem de forma mais rápida e personalizada no apoio aos alunos que apresentem risco escolar. Por último, ainda poderia ser sugerido uma ampliação do estudo ao se aplicar outras técnicas de classificação ao problema para definir a mais adequada em termos de desempenho.

Referências

Aldowah, H., Al-Samarraie, H., Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, v. 37, n. April 2018, p. 13–49, 2019.

Arnold, K. E. e Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. ACM International Conference Proceeding Series, n. April 2012, p. 267–270.

Baker, R., Isotani, S., Carvalho, A. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, v. 19, n. 02, p. 3–13, 2011.

Baker, R., Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. JEDM - Journal of Educational Data Mining (ISSN 2157-2100), Volume 1, Issue 1, October 2009.

Bramer, M. (2016). Principles of Data Mining. Third Edition. Springer-Verlag, 2016. DOI 10.1007/978-1-4471-7307-6.

Bradley, A. E. (1997). The Use of the Area under the Roc Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition, Vol. 30, No. 7, pp. 1145-1159, 1997.

Brasil (1996). Senado Federal. Lei de Diretrizes e Bases da Educação Nacional: nº 9394/96. Brasília: 1996.

Calixto, K., Segundo, C., and de Gusmão, R. P. (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. VI Congresso Brasileiro de Informática na Educação (CBIE 2017), Anais do SBIE 2017, volume 28, page 1447.

Camilo, C., Silva, J. Mineração de Dados: Conceitos, tarefas, métodos e ferramentas (2019). RT-INF_001-09 - Relatório Técnico, Instituto de Informática Universidade Federal de Goiás. Disponível em: http://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acessado em setembro de 2020.

Chickering, A. W. e Ehrmann, A. F. (1996). Implementing the seven principles for good practice in undergraduate education. American Association of Higher Education Bulletin, v. 49, n. 2, p. 2–4.

Colpani, R. (2018). Mineração de Dados Educacionais: um Estudo da Evasão no Ensino Médio com Base nos Indicadores do Censo Escolar. Informática na Educação: teoria & prática, Porto Alegre, v. 21, n. 3, p. 109-123, set./dez. 2018. Disponível em: https://seer.ufrgs.br/InfEducTeoriaPratica/article/view/87880/52130. Acessado em setembro de 2020.

Costa, E., Baker, R. S. J., Amorim, L. e Magalhães, J. (2012). Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. JAIE 2012, Congresso Brasileiro de Informática na Educação 2012, p. 1–29.

DesJardins, S.L., Ahlburg, D.A., McCall, B.P. (1999). An event history model of student departure. Economics of Education Review, vol. 18, issue 3, p. 375-390.

Fayyad, U; Piatetsky-Shapiro, G; Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230.

Guyon, I. and Elisseeff, A. (2006). An introduction to feature extraction. In Feature Extraction, Foundations and Applications. Springer, p. 124, 2006.

Han, J., Pei, J., Kamber, M. (2011). Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann, 2011 ISBN: 9780123814807.

Lee, S., Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. July 2019. Applied Sciences 9(15):3093. DOI: 10.3390/app9153093.

Manhães, L. M. B. et al. (2011). Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. Anais do XXII SBIE, XVII WIE, p. 150–159.

Neves Junior, R., Nascimento, R. L. S., Fagundes, R. A. de A. e Mattos Neto, P. S. G. (2019). Estimação de Índices de Aprovação e Reprovação Escolar do Ensino Médio. Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019), p. 339.

Paes, B. C., Plastino, A. e Freitas, A. A. (2013). Selection of attributes applied to hierarchical classification. 1st Symposium on Knowledge Discovery, Mining and Learning.

Parmezan, R., Lee, H., Spolaôr, N., Chung, W. F. , A. (2012). Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados. RT 002, Univ Estadual Oeste do Paraná. December, p. 58. Disponível em http://sites.labic.icmc.usp.br/aparmezan/publications/pdf/BIBLIOTECA_000_RT_002.pdf. Acessado em setembro de 2020.

Ramalho, L. G. (2013). Abordagem Avaliativa da Política de Assistência Estudantil em uma Instituição de Ensino Profissional. Dissertação do Mestrado Profissional CAEd/ FACED/ UFJF. Juiz de Fora, 2013.

Romero, C., Ventura, S., Espejo, G. P. e Hervás, C. (2008). Data mining Algorithms to Classify Students". In Proceedings of the 1st International Conference on Educational Data Mining, p.8-17.

Romero, C.; Ventura S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 3, n. 1, p. 12–27.

Sales, F., Mendes, Y., Dembogurski, B., et al. (2019). Evasão no Ensino Básico da Rede Pública Municipal de Juiz de Fora: uma Análise com Mineração de Dados. Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019), p. 1371.

Saraiva, D., Pereira, S., Gallindo, E., Braga, R. e Oliveira, C. (2019). Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática. Anais do XXVII WEI, p. 319–333.

Silva, J. L. D. e Nunes, I. D. (2015). Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. Anais do XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015), v. 1, n. Sbie, p. 1112.

U.S. Department of Education, Institute of Education Sciences and National Center for Education Evaluation and Regional Assistance (2017). Preventing Dropout in Secondary Schools. Technical Report NCEE 2017-4028. Disponível em https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/wwc_dropout_092617.pdf. Acessado em setembro de 2020.

Veloso, L. A. (2015). A Predição da Evasão Escolar dos Cursos Técnicos de Nível Médio: Um Estudo de Caso no Senai. Dissertação (Mestrado), Universidade Católica de Brasília, 2015.