6

Modelagem descritiva e preditiva da mudança de processo seletivo para ingresso no ensino superior

Joel Felipe de Oliveira Gaya [1], Sidnei da Fonseca Pereira Júnior [2], Paula Fernanda Schiavo [3], Eduardo Borges [4]. Silvia Silva da Costa Botelho [5].

Resumo

Um dos desafios para as instituições de ensino superior é obter métricas que auxiliem o acompanhamento do desempenho do discente tendo em vista as constantes transformações do ensino no país. A mineração de dados vem sendo uma das principais ferramentas utilizadas por conseguir extrair informação implícita, previamente desconhecida e potencialmente útil para tomada de decisão. Neste trabalho foi analisada uma base de dados de estudantes do curso de Engenharia de Computação da Universidade Federal do Rio Grande (FURG), onde houve uma mudança na forma de avaliação para o ingresso dos alunos no ensino superior. Foram gerados modelos para avaliar o impacto desta alteração os quais mostram que o desempenho dos alunos diminuiu. Utilizando modelos de dados preditivos, percebeu-se que a idade, as notas de ingresso e o número de repetições nas disciplinas são fatores preponderantes para que aluno conclua o curso.

Palavras-chave: Mineração de dados, Classificação, Regressão.

Abstract

One of the challenges for higher education institutions is to obtain metrics that help monitor students' performance in view of the constant transformations of teaching in the country. Data mining has been one of the main tools used to extract implicit, previously unknown and potentially useful information for decision making. In this work, a database of students of the Computer Engineering course of the Universidade Federal do Rio Grande (FURG) was analyzed, where there was a change in the form of evaluation for students' admission to higher education. Models have been generated to assess the impact of this change which show that student performance has declined. Using predictive data models, it was noticed that the age, the grades of entry and the number of repetitions in the disciplines are preponderant factors for the student to complete the undergraduate course.

Keywords: Data Mining, Classification, Regression.

[1]Universidade Federal do Rio Grande, E-mail: joelfelipe94@gmail.com
[2]Universidade Federal do Rio Grande, E-mail: sidnei.pereira@furg.br
[3]Universidade Federal do Rio Grande, E-mail: pfschiavo@furg.br
[4]Universidade Federal do Rio Grande E-mail: eduardoborges@furg.br
[5]Universidade Federal do Rio Grande, E-mail: silviacb.botelho@gmail.com

1 Introdução

Por muitos anos, cada uma das universidades públicas brasileiras possuía seu próprio método de processo seletivo. Esta realidade mudou recentemente com a adesão das universidades ao Exame Nacional do Ensino Médio (ENEM). O ENEM é uma avaliação realizada pelo Ministério da Educação (MEC) desde o ano de 1998. E tem como objetivo avaliar o conhecimento dos estudantes do ensino médio.

A partir de 2009, o ENEM passou a ser utilizado nos processos seletivos da grande maioria das instituições públicas do país. Na primeira edição desse modelo de seleção, o ENEM registrou cerca de 4,5 milhões de inscritos. Já em 2013, foram registradas mais de 7 milhões de inscrições. Assim como o número de inscrições, o número de vagas ofertadas em instituições públicas de ensino superior também cresceu de maneira significativa no país durante esse período.

Na Universidade Federal do Rio Grande (FURG), onde o estudo relatado neste artigo foi conduzido, a adesão ao exame foi feita no ano de 2011. Porém, como mencionado anteriormente esta não foi a única mudança que ocorreu na época, ela veio associada a uma série de mudanças como o aumento da oferta de vagas disponíveis na região, assim como no resto do país. Permitindo que estudantes de qualquer parte do país concorressem a uma vaga nas universidades, desde que ela aderisse ao SISU, sem a necessidade do estudante se deslocar até a universidade para prestar vestibular.

Apesar desta mudança ter afetado diversas universidades em todo o país, ainda não há muito estudo sobre o impacto no desempenho dos acadêmicos. Portanto, o presente trabalho busca verificar o impacto desta mudança em alguns aspectos do desempenho acadêmico.

Este artigo estende o trabalho prévio publicado na Conferência Sul em Modelagem Computacional (GAYA et al., 2016), o qual tem como principais objetivos avaliar o impacto da mudança do método de ingresso na aprovação e nas notas finais dos alunos e verificar se os auxílios solicitados pelos estudantes e atendidos pela Universidade contribuem de fato no desempenho acadêmico. Além disso, foram construídos modelos de aprendizado para predizer se o aluno conclui ou não o curso. Este artigo estende o trabalho anterior incluindo um referencial teórico sobre mineração de dados necessário para compreender os algoritmos utilizados nos experimentos, detalhando a estrutura do conjunto de dados utilizado, e incluindo um novo estudo de caso que apresenta o impacto do número de repetências em disciplinas introdutórias na conclusão do curso.

O restante deste artigo é organizado da seguinte forma. A Seção 2 apresenta um referencial teórico sobre as principais tarefas de mineração adotadas: associação e classificação. Além disso, introduz conceitos sobre alguns dos algoritmos utilizados nos experimentos. A Seção 3 mostra um conjunto de trabalhos relacionados, relatando as técnicas usadas e as conclusões dos autores. Na Seção 4 é feita uma descrição breve da base de dados e de como os dados foram pré-processados. A Seção 5 descreve a metodologia adotada neste trabalho. A Seção 6 apresenta os resultados obtidos com cada algoritmo e parametrização empregados, para cada tarefa de mineração. São apresentados os modelos de dados e as medidas diversas medidas de avaliação. Por fim, a Seção 7 resume as contribuições deste artigo e aponta a direção que será tomada nos trabalhos futuros.

2 Referencial Teórico

Descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD) é um processo não trivial de identificar padrões potencialmente úteis e compreensíveis em meio às observações presentes em uma base de dados (TAN; STEINBACH; KUMAR, 2005). Geralmente, esses padrões são extraídos de relacionamentos implícitos entre os dados analisados. Como resultado final, os padrões encontrados devem gerar conhecimento inteligível e imediatamente utilizável para o apoio às decisões.

A descoberta de conhecimento é dividia nas seguintes fases (HAN, 2005; PRASS, 2004):

- seleção de dados - escolha do conjunto de dados contendo todas as possíveis variáveis (atributos) e observações (registros) que farão parte da análise. Esta fase pode ser bem complexa, uma vez que os dados podem ser extraídos de fontes distintas e heterogêneas (bancos de dados, *data warehouses*, planilhas, textos, páginas *Web*, etc.) e ainda podem possuir os mais diversos formatos.

- pré-processamento - limpeza e normalização dos dados. Inclui outras tarefas como remoção de ruído (*outliers*), tratamento de registros incompletos e remoção de redundância.

- transformação dos dados - adequação dos dados em relação à técnica e algoritmo de mineração a serem utilizados. Esta fase inclui a escolha da representação dos dados e a redução de dimensionalidade (número de atributos).

- mineração - análise automática dos dados em busca de padrões utilizando algoritmos de mineração de dados.

- interpretação de resultados - fase final responsável pela geração de conhecimento baseada nos padrões encontrados.

Caso não sejam encontrados resultados relevantes em quaisquer fases, o processo deve retornar a uma das fases anteriores. Os experimentos realizados neste trabalho contemplam todas as fases do processo de descoberta de de conhecimento em bases de dados e são apresentados na Seção 6.

A mineração é a principal fase do processo de descoberta de conhecimento em bases de dados. Pode ser definida como a análise automática dos dados em busca de padrões (WITTEN; FRANK, 2011). É apoiada por algoritmos de aprendizado de máquina, estatísticas e técnicas de visualização. Dependendo do objetivo da mineração, diversas técnicas podem ser utilizadas (TAN; STEINBACH; KUMAR, 2005; HAN, 2005):

- associação - definição de regras que capturam a coocorrência de elementos em diferentes observações da bases de dados. Uma das aplicações clássicas da associação consiste na descoberta de produtos que são comprados juntos, largamente utilizada para recomendação em comércio eletrônico. A análise dos resultados pode determinar que ações devem ser tomadas para incrementar a venda de um produto ou que produtos são afetados por outros.

- classificação - método supervisionado que determina um modelo para um determinado atributo que é função dos valores dos outros atributos. Pode ser utilizado para predizer se uma nova instância fará parte de uma determinada classe.

- regressão - predição do valor de uma variável contínua baseado no valor de outras variáveis, considerando um modelo de dependência linear ou não linear.

Uma regra de associação tem o formato $A \rightarrow B$, onde A e B são elementos que co-ocorrem em diferentes observações da bases de dados. A é chamado de antecedente e B de consequente. Eles são conjuntos de itens de transações e a regra pode ser lida como: os atributos A frequentemente implicam nos atributos B.

O principal algoritmo da literatura para geração de regras de associação é o Apriori (AGRAWAL; SRIKANT et al., 1994), que foi projetado para trabalhar com grandes conjuntos de dados de transações armazenadas em bancos de dados e com um custo menor de processamento. O algoritmo é baseado na ideia que, se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes. Portanto, caso um item não seja frequente, não é necessário combiná-lo com outros itens para geração de regras.

Já a tarefa de classificação consiste no processo de encontrar, através de aprendizado de máquina, um modelo ou função que descreva diferentes classes de dados. Ou seja, o objetivo da classificação é rotular, automaticamente, novas instâncias da base de dados com uma determinada classe aplicando o modelo ou função aprendidos. Após a classificação, os dados originais estarão categorizados em classes.

Diversos classificadores foram propostos nos últimos anos (MITCHELL, 1997). Alguns utilizam árvores de decisão para rotular registros em determinada classe. São exemplos CART (BREIMAN et al., 1984), ID3 (QUINLAN, 1979) e C4.5 (QUINLAN, 1993). Outros algoritmos se baseiam em florestas (BREIMAN, 2001), redeis neurais artificiais (HAYKIN, 2007; BOSER; GUYON; VAPNIK, 1992; PLATT, 1999), modelos probabilísticos (JOHN; LANGLEY, 1995) ou regras (COHEN, 1995).

A regressão difere da classificação porque a função alvo f a ser aprendida mapeia os atributos de entrada em uma saída de valores contínuos. Geralmente é treinada usando o métodos dos mínimos quadrados. A regressão é dita linear quando a relação entre os atributos preditores e a resposta segue um comportamento linear. Modelos de regressão não-linear relacionam as variáveis usando outros tipos de funções, tais como polinômios ou modelos

baseados em árvores (QUINLAN, 1992).

3 Trabalhos Relacionados

Diversos estudos em aplicações e métodos de mineração de dados foram realizados em diversas áreas, tendo como objetivo extrair conhecimento de bases de dados. No entando, mesmo após anos de pesquisa na área, ainda não há um sistema de mineração genérico capaz de extrair conhecimento de uma base qualquer. Então, é esperado que a tarefa de projetar um sistema de mineração que realize tal tarefa usando métodos existententes de mineração, para uma base ainda não explorada seja um desafio ciêntifico (DESHPANDE; THAKARE, 2010).

Anteriormente, foram propostas diversas abordagens para predição de desempenho acadêmico como em (KHAN; KHIYAL; KHATTAK, 2015), e (PANDEY; PAL, 2011), que trazem ideias bastante úteis para o presente trabalho.

(KABAKCHIEVA, 2013) realizou um estudo em bases de dados de estudantes aplicando classificação Bayesiana para prever o desempenho de um indivíduo, baseado nos dados do ano anterior. Eles evidenciam, neste estudo, que não apenas o esforço do estudante, mas também influência familiar, hábitos do estudante e escolaridade dos pais influenciam dramaticamente na *performance* do mesmo. Vale ressaltar que este trabalho relata a importância dos hábitos do estudante em seu desempenho, mas esta informação ainda não está presente na base da universidade em questão. Porém, se de fato essa informação é relevante para o desempenho do acadêmico seria interessante que a instituição passasse então à armazená-la.

(YADAV; PAL, 2012) realizou um estudo com o intuito de prever a *performance* de estudantes de engenharia de universidades indianas, a partir de suas notas do ano anterior ou exame do primeiro ano. Eles testaram três algoritmos de classificação diferentes ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984), dentre os quais selecionaram o C4.5 como o melhor preditor neste caso. Os pesquisadores, implementaram o resultado do estudo de previsão alertando os alunos que tinham chances de serem reprovados e obtiveram um melhora no desempenho dos mesmos.

(PRASAD; BABU, 2013) usou mineração de dados em uma base de estudantes indianos para predizer a *performance* dos estudantes, com base em suas notas anteriores. Nele foi usado o algoritmo de classificação C4.5. Os autores concluíram que estudantes e professores podem melhorar seu desempenho com base em sua predição.

Vale ressaltar, entretanto, que a pesquisa apresentada não modela o desempenho de um acadêmico, mas sim procura entender quais são as variáveis que levam um aluno a obter um determinado desempenho. Mais especificamente, procuramos saber se a mudança na forma de ingresso tem afetado significativamente o desempenho dos acadêmicos deste curso.

4 Conjunto de Dados Analisado

O conjunto de dados fornecido pela FURG compreende alunos apenas do curso de Engenharia de Computação, de 1994, ano em que o curso teve inicio, até o final do ano de 2015.

A Figura 1 ilustra as tabelas do banco de dados relacional que foram utilizadas neste trabalho.

- *aluno* - armazena o ano de ingresso e a situação acadêmica de cada estudante. Esta tabela se relaciona com todas as demais.
- *nota* - contém, para cada disciplina e período letivo, o desempenho acadêmico dos alunos discriminando por uma série de notas bimestrais, frequência, notas do exame e finais.
- *ingresso_origem* - apresenta o modo de ingresso (vestibular, SISU, transferência, etc.), local e ano de nascimento dos estudantes.
- *apoio_social* - guarda as solicitações e os aceites dos auxílios alimentação, moradia, permanência e pré-escola.
- *vestibular* - armazena as notas de cada prova do vestibular e a classificação do estudante em relação aos demais.
- *enem* - contém as notas de cada prova do ENEM e a classificação na chamada do SISU.
- *socio_economico* - apresenta as respostas do questionário sócio-econômico respondido durante a prova do processo seletivo correspondente.

As informações interessantes para mineração se encontram dispersas em diversas tabelas, e além disso existe uma série de dados faltantes e mal cadastrados. Foram então realizadas operações para

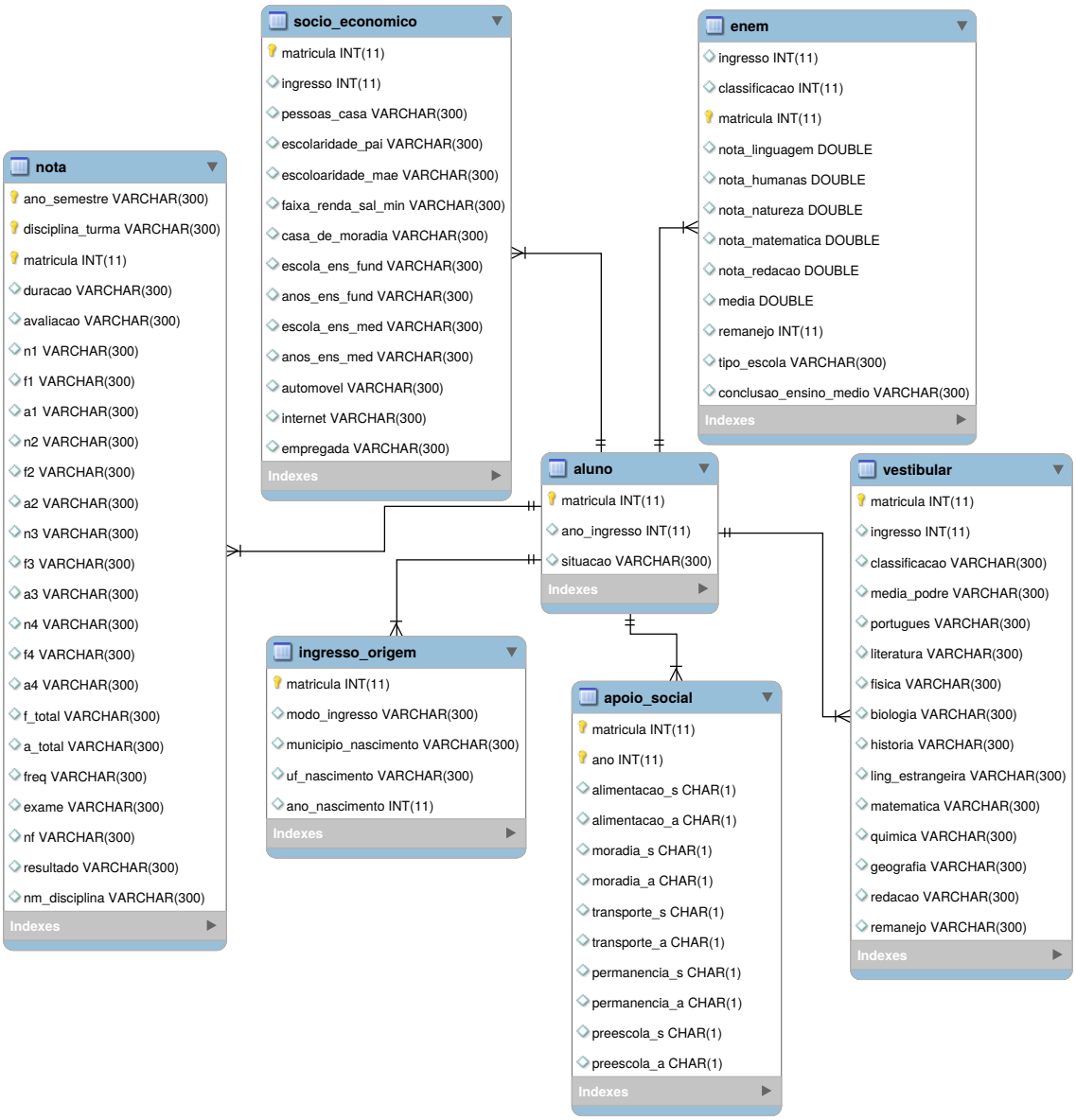

Figura 1: Esquema lógico das tabelas usadas no presente trabalho.

agrupar os dados, eliminar registros ruidosos e transformar a informação. A partir da base fornecida, foram realizadas consultas SQL (Structured Query Language) (MYSQL, 1995), que geraram as relações passíveis de serem mineradas. Ao fim desta transformação, foram identificados três estudos de caso.

4.1 Caso 1 - Impacto dos atributos no desempenho dos alunos

Os dados necessários para modelar o impacto dos atributos no desempenho dos estudantes é obtido a partir de duas tabelas. A primeira foi utilizada para as técnicas de classificação e regressão, a fim de verificar o impacto de cada atributo no desempenho dos alunos. A outra tabela foi utilizada para a associação, onde procura-se alguma inconsistência que possa vir a prejudicar os alunos de baixa renda e ou oriundos de outras regiões.

A tabela usada na classificação e na regressão, após as etapas de transformação e preprocessamento, contém os atributos idade de ingresso, nota obtida no exame admissional, nome da disciplina cursada, modo de ingresso, resultado e nota final. O modo de ingresso pode assumir s valores ENEM, vestibular ou outros. Resultado é o atributo que contém a informação se o aluno aprovou ou reprovou na disciplina. Cada instância representa o desempenho de um determinado aluno em uma disciplina.

Para a tarefa de regressão não é interessante que a tabela contenha o atributo resultado, já que ele se dá em função da nota final. Então este atributo foi removido para a execução desta tarefa. Da mesma forma para o atributo nota final na tarefa de classificação.

A medida de desempenho dos alunos foi feita por disciplina, devido ao fato da base em questão não conter turmas formadas que ingressaram através do novo método de ingresso (SISU) e de mudanças no currículo que mudaram o ano do curso no qual as disciplinas eram ministradas. E portanto, essa base não permite uma análise que envolva o número de formados ou desistências.

A tabela usada para associação, reúne informações provenientes de tabelas que continham dados dos benefícios concedidos aos alunos, e com questionários socioeconômicos respondidos no momento do ingresso. O preprocessamento dos atributos dos benefícios para associação foi feito da seguinte forma: se o benefício foi pedido e concedido o atributo recebe valor *sim*, se pedido e não concedido, o atributo recebe o valor *não* e caso ele não tenha sido pedido, o atributo não possui valor. Esta abordagem foi adotada para que benefícios que não fossem solicitados não formassem regras.

4.2 Caso 2 - Modelo para predizer impacto na conclusão do curso

No preprocessamento deste caso foram filtradas as tabelas de modo a se obter os atributos: número de matrícula, a idade de ingresso, as notas do ingresso (valores das notas das provas do vestibular correlacionados com as disciplinas utilizadas no ENEM conforme Tabela 1), a quantidade de vezes que o aluno cursou as disciplinas da primeira série (Cálculo 1, Física 1, Química e Computação), da segunda série, (Cálculo 2 e Física 2), e da terceira série (Métodos Numéricos e Teoria da Computação).

Ainda foram adicionadas as informações se o aluno estava *formado*, *cursando* ou *não formado*, o qual inclui qualquer motivo de evasão.

Após a consulta finalizada, foram identificados duas situações que precisaram de ajustes manuais. A primeira ocorreu quando percebeu-se que alguns alunos formados não possuíam a informação de quantidade de vezes que cursaram algumas matérias. Foi considerado que se a informação de Formado estiver correta, os discentes poderiam ter feito equivalência das disciplinas em outro curso ou instituição, por exemplo. Dessa forma, considera-se que todos os alunos que concluíram o curso fizeram pelo menos uma vez a cada disciplina consultada. A segunda situação foi que alguns alunos não possuíam a nota de ingresso. Esses cadastros foram excluídos da seleção, por se tratarem de poucos registros.

Por fim, restaram 820 instâncias com o perfil homogêneo, onde 305 registros estavam associados como formados, 322 como não formados e 193 como cursando.

4.3 Caso 3- Impacto do número de repetências em disciplinas iniciais na conclusão do curso

Para este caso de estudo foram consideradas o número de repetições em disciplinas cursadas nos dois primeiros anos do curso. Por se tratar de um curso seriado, esta análise foi feita apenas nas séries iniciais pois o aluno não pode adiantar disciplinas dos últimos anos.

Para este caso as etapas de transformação e preprocessamento resultam em uma tabela contendo os campos idade de ingresso, número de vezes que foram cursadas as disciplinas de Cálculo I, Introdução à Ciência da Computação, Fundamentos de Química, Física I, Álgebra Linear e Geometria Analítica (ALGA), Física II, Cálculo II, Mecânica, Sistemas de Computação, Estruturas de Dados e, por fim, a situação em que o aluno se encontra (formado ou não formado). Neste caso estudantes que ainda estão matriculados não foram considerados.

Ao término da etapa de preprocessamento a tabela contém 666 instâncias, onde 311 são formados e 355 são não formados.

5 Metodologia

A fim de realizar os experimentos, foi usada a ferramenta de mineração de dados *Waikato Environment for Knowledge Analysis* (WEKA) (HALL et al., 2009). Foram utilizadas técnica de classificação, regressão e associação, implementadas por vários algoritmos. Para cada uma destas técnicas é apresentado um algoritmo com duas parametrizações distintas. Os algoritmos usados foram escolhidos com base nos trabalhos relacionados e também com base em testes preliminares.

Vale ressaltar que o objetivo deste trabalho não é prever de fato os resultados de um aluno, mas sim analisar quais são os fatores que o levam à obter um

Tabela 1: Representação das notas de ingresso.

Disciplinas	Enem	Vestibular
Linguagem	Nota Linguagem	(Português + Literatura + Língua Estrangeira) / 3
Humanas	Nota Humanas	(História + Geografia) / 2
Natureza	Nota Natureza	(Física + Biologia + Química) / 3
Matemática	Nota Matemática	Matemática

Tabela 2: Medidas de avaliação obtidas do resultado do classificador J48, através da validação cruzada, para divisões não necessariamente binárias.

Instâncias corretamente classificadas	80.18 %
Instâncias incorretamente classificadas	19.82 %
Kappa	0.244
Erro absoluto médio	0.265
Raiz do erro quadrático médio	0.372
Erro relativo absoluto	80.98 %
Raiz do erro quadrático relativo	92.08 %

Tabela 3: Medidas de avaliação obtidas do resultado do classificador J48, através da validação cruzada, para divisões binárias.

Instâncias corretamente classificadas	82.08 %
Instâncias incorretamente classificadas	17.92 %
Kappa	0.349
Erro absoluto médio	0.252
Raiz do erro quadrático médio	0.363
Erro relativo absoluto	77.08 %
Raiz do erro quadrático relativo	89.92 %

determinado resultado. Porém, é de fundamental importância obter modelos dos dados que tenham boa acurácia, pois isso indica que o modelo representa de forma fiel os dados que compõem o seu treinamento. Por essas razões, este trabalho dá preferência à modelos de fácil interpretação.

6 Resultados Obtidos e Técnicas Empregadas

Esta seção apresenta em detalhes os resultados obtidos em cada estudo de caso. Também são detalhadas a s técnicas utilizadas e a parametrização dos algoritmos utilizados.

6.1 Caso 1 - Impacto dos atributos no desempenho dos alunos

No primeiro estudo de caso pretende-se analisar se houve impacto na aprovação dos alunos após a alteração do método utilizado no processo seletivo. Essa análise consiste na tarefa de classificação para a predição do atributo resultado, que indica aprovação ou reprovação. Foi utilizado o classificador J48 (BHARGAVA et al., 2013).

O experimento foi conduzido usando validação cruzada com 10 partições. Os resultados obtidos através desta análise podem ser observados na Tabela 2 e na Tabela 3. Ambos os resultados parecem representar bem o conjunto de dados usado no treinamento.

A Figura 2 mostra o modelo para a disciplina Fundamentos de Química, porém o resto do modelo se comporta de forma análoga. Levando em consideração o modelo gerado, pode-se observar de maneira clara três fenômenos. Primeiro, o impacto negativo da idade de ingresso nas notas, o que pode ser explicada pelo fato de alunos mais novos terem mais disponibilidade de tempo para as tarefas do

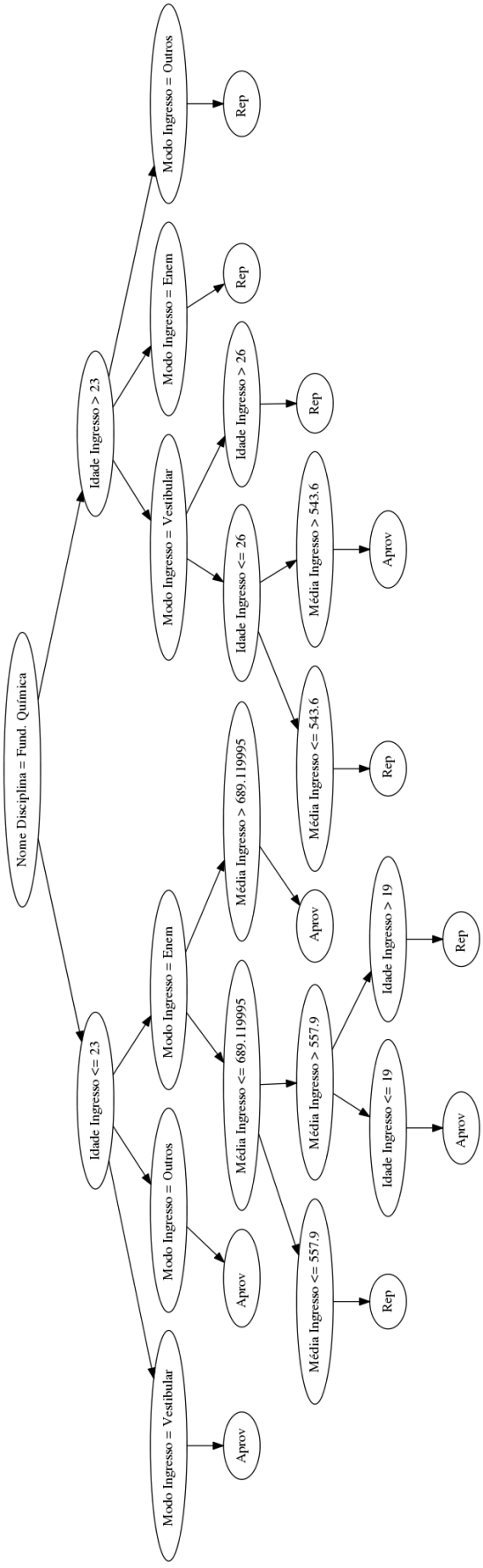

Figura 2: Árvore de decisão para a disciplina Fundamentos de Química.

Tabela 4: Medidas de avaliação obtidas do resultado da regressão linear, através da validação cruzada, usando o método M5 de seleção de atributos.

Coeficiente de correlação	0.512
Erro absoluto médio	2.019
Raiz do erro quadrático médio	2.670
Erro relativo absoluto	81.53 %
Raiz do erro quadrático relativo	85.90 %

Tabela 5: Medidas de avaliação obtidas do resultado da regressão linear, através da validação cruzada, sem o uso de métodos de seleção de atributos.

Coeficiente de correlação	0.512
Erro absoluto médio	2.018
Raiz do erro quadrático médio	2.670
Erro relativo absoluto	81.50 %
Raiz do erro quadrático relativo	85.91 %

$$Nota\ final = 0.0009 * Media\ de\ ingresso - 0.196 * Idade\ de\ ingresso + 0.5511 * [Modo\ ingresso = Vestibular] + 4.0567 + Ganho \quad (1)$$

curso. Em segundo lugar, o impacto positivo da nota obtida no ingresso, seja ele por vestibular ou SISU, fenômeno esse não estranho visto que alunos com bom desempenho em avaliações anteriores mantenham o bom desempenho. E finalmente, o impacto negativo da mudança de avaliação na aprovação dos alunos.

Para a análise de impacto nas notas finais obtidas pelos alunos em cada disciplina, foi utilizado o método de regressão linear para os mesmos atributos utilizados na análise anterior, conforme descrito na Seção 4.

De forma semelhante à tarefa anterior, os resultados obtidos foram avaliados usando validação cruzada com 10 partições. A análise dos resultados pode ser observada nas Tabelas 4 e 5. Neste caso, o uso de um método de seleção de atributos representa um ganho marginal.

Como esperado, a Equação 1 gerada por esse método, onde $Ganho$ é um fator aditivo definido de acordo com a disciplina, confirma os três fenômenos mencionados anteriormente. Entretanto, ela não trás nenhuma nova contribuição a não ser confirmar conclusões anteriormente obtidas.

O impacto negativo nas notas e no número de aprovações, faz pensar sobre o que leva estes alunos a terem um desempenho inferior ao dos alunos que ingressaram pelo antigo vestibular. Uma das possibilidades é que com o aumento dos alunos oriundos de outras regiões existe uma demanda maior por auxílio que pode não estar sendo atendida.

Para verificar a hipótese levantada, foi usada uma tabela que possui informações sócio-econômicas do aluno e dos pedidos de auxílio. Através dessas informações, procura-se identificar padrões que levem a universidade a acolher de forma mais eficaz os alunos. Neste trabalho optou-se por usar associação para encontrar tal padrão.

O algoritmo escolhido para a tarefa de associação foi o Apriori (AGRAWAL; SRIKANT et al., 1994). As regras obtidas e sua confiança são mostradas nas Tabelas 6 e 7. Nelas não observamos a existência de nenhuma anomalia aparente nos dados o que não é um resultado conclusivo. Porém, apesar de não conclusivo o resultado é uma indicativo de que a causa da queda de *performance* dos alunos que ingressaram pelo vestibular em relação aos que entraram pelo ENEM não foi causada por falta de auxílio da universidade.

Tabela 6: Regras obtidas pelo algoritmo Apriori, para suporte mínimo 0.05 e confiança 0.8.

Regra	**Confiança**
Alimentação=Sim, Moradia=Sim, Esc. Ens. Med.=Pública → resultado=Aprov	0.83
Alimentação=Sim, Moradia=Sim, Esc. Ens. Med.=Pública Internet=Sim → Resultado=Aprov,	0.83
Alimentação=Sim, Moradia=Sim, Esc. Ens. Fund.=Pública Esc. Ens. Med.=Pública → Resultado=Aprov	0.82
Alimentação=Sim, Moradia=Sim, → Resultado=Aprov	0.81
Alimentação=Sim, Moradia=Sim, Internet=Sim → Resultado=Aprov	0.80

Tabela 7: Regras obtidas pelo algoritmo Apriori, para suporte mínimo 0.013 e confiança 0.9.

Regra	**Confiança**
Transporte=Sim, Casa de Moradia=Própria, Esc. Ens. Fund.=Pública, Internet=Não → Resultado=Rep	1.00
Permanência=Não, Casa de Moradia=Própria, Esc. Ens. Fund.=Pública, Internet=Não → Resultado=Rep	1.00
Transporte=Sim, Permanência=Não, Casa de Moradia=Própria, Esc. Ens. Fund.=Pública, Internet=Não → Resultado=Rep	1.00
alimentacao=Sim, Transporte=Sim, Casa de Moradia=Própria, Esc. Ens. Fund.=Pública, Internet=Não → Resultado=Rep	1.00
Transporte=Sim, Casa de Moradia=Alugada, Esc. Ens. Fund.=Pública, Internet=Sim → Resultado=Rep	0.96
Moradia=Sim, Casa de Moradia=Alugada, Internet=Sim → Resultado=Rep	0.93
Casa de Moradia=Própria, Esc. Ens. Fund.=Pública, Internet=Não → Resultado=Rep	0.91

6.2 Caso 2 - Modelo para predizer impacto na conclusão do curso

Para encontrar um modelo que auxilie a predizer se o aluno conclui ou não o curso, retorna-se à tarefa de classificação, utilizando o algoritmo J48, com a validação cruzada com 10 partições.

Os resultados obtidos através desta análise podem ser observados nas Tabelas 8 e 9.

Em ambos os resultados, os dados usados no treinamento atingiram alto coeficiente de acertos, mostrando que idade de ingresso, médias das provas de ingresso e número de repetições nas disciplinas possuem grande relevância na predição sobre a conclusão do curso. Pode-se considerar o modelo muito especializado, mas a proposta é encontrar quais os atributos podem revelar a predição.

Analisando o modelo percebe-se também que o número de repetições de um aluno na disciplina Cálculo I é semelhante a disciplina de Física I. Ambas as disciplinas são ministradas no primeiro ano do curso. Alguns alunos que obtiveram notas muito altas no ingresso, independente se foi pelo Vestibular ou pelo Enem, não concluíram o graduação de Engenharia de Computação na FURG. Acreditamos que estas pessoas tenham se deslocado para outras universidades ou cursos. Como já havíamos abordado a situação de origem do aluno é um fator importante para a continuidade e desempenho no curso. Seria necessário uma avaliação específica destes casos para termos certeza.

Em um dos ensaios, onde os dados somente para a disciplina Calculo I foram usados (vide Figura 3), o modelo confirma a referência anterior que a idade de ingresso é um fator de grande influência. Mesmo errando bastante consideramos o modelo significativo (vide Tabela 10).

Tabela 8: Medidas de avaliação obtidas do resultado do classificador J48, através da validação cruzada, para divisões não necessariamente binárias.

Instancias corretamente classificadas	95.85 %
Instancias incorretamente classificadas	04.15 %
Kappa	0.917
Erro absoluto médio	0.071
Raiz do erro quadrático médio	0.198
Erro relativo absoluto	14.13 %
Raiz do erro quadrático relativo	39.67 %

Tabela 9: Medidas de avaliação obtidas do resultado do classificador J48, através da validação cruzada, para divisões binárias.

Instancias corretamente classificadas	96.33 %
Instancias incorretamente classificadas	03.67 %
Kappa	0.930
Erro absoluto médio	0.066
Raiz do erro quadrático médio	0.185
Erro relativo absoluto	13.22 %
Raiz do erro quadrático relativo	37.10 %

Tabela 10: Medidas de avaliação obtidas do resultado para ser aprovado em Cálculo I, através da validação cruzada.

Instancias corretamente classificadas	69.22 %
Instancias incorretamente classificadas	30.78 %
Kappa	0.267
Erro absoluto médio	0.409
Raiz do erro quadrático médio	0.463
Erro relativo absoluto	87.45 %
Raiz do erro quadrático relativo	95.63 %

6.3 Caso 3- Impacto do número de repetências em disciplinas iniciais na conclusão do curso

Neste caso de estudo a importância do método de avaliação é desconsiderada, e apenas será avaliada a quantidade de vezes que cada disciplina foi cursada em relação à conclusão do curso. É evidente que existe uma relação direta entre concluir as disciplinas das séries finais e portanto o curso. No entanto, para a presente análise apenas as disciplinas iniciais são consideradas. Com isso, pretende-se avaliar o impacto do número de vezes que um aluno tem de cursar uma disciplina em relação à evasão.

Esta análise assim como feito anteriormente na Seção 6.1, consiste em uma tarefa de classificação, tendo como atributo alvo a situação. Para isto foi novamente utilizado o classificador J48 (BHARGAVA et al., 2013).

Neste experimento foi usada a validação cruzada com 10 partições. Os resultados obtidos podem ser observados na Tabela 11. Visto que esta tabela mostra que o modelo representa de forma satisfatória os dados para os quais foi treinado, pode-se então passar à analise do modelo e com isso procurar compreender o volume de dados.

O modelo gerado é mostrado na Figura 4. De modo geral um número elevado de vezes que o aluno cursa essas disciplinas costuma implicar na não conclusão do curso. Porém, considerando a disciplina de Estruturas de Dados, o fato do aluno não cursa-la pode indicar que o mesmo desistiu na

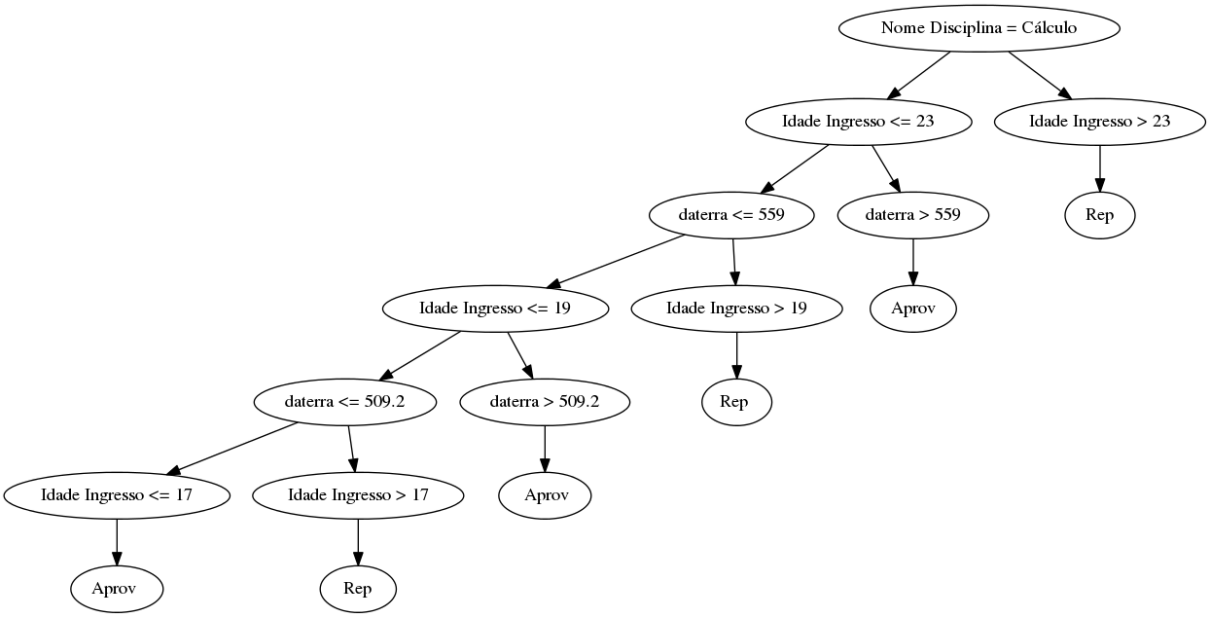

Figura 3: Árvore de decisão para a disciplina Cálculo I

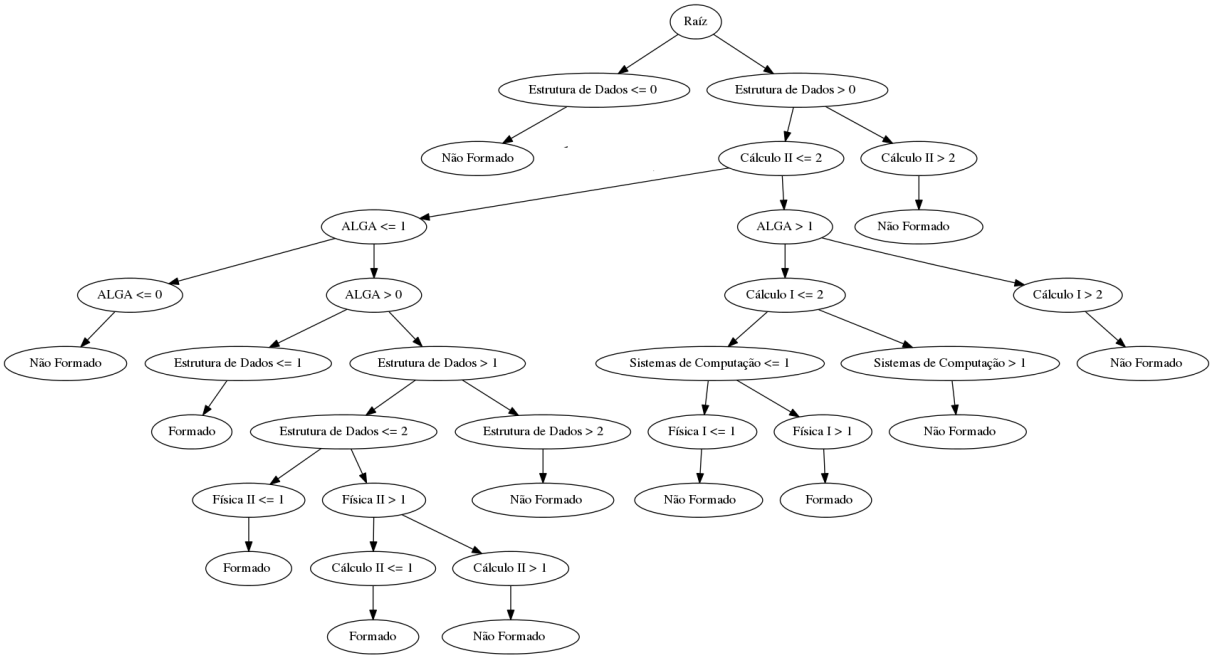

Figura 4: Árvore de decisão que modela a conclusão do curso em relação ao número de repetições em disciplinas introdutórias.

primeira série e portanto não conclui o curso. Já na disciplina de ALGA (obrigatória e ofertada na primeira série), não cursá-la indica que o estudante já foi aprovado em disciplina equivalente e foi dispensado. Observando o modelo, esta situação tem uma influência negativa ao longo do curso.

7 Conclusão e Trabalhos Futuros

Este trabalho mostrou que a mudança no método de ingresso afetou de forma significativa o desempenho dos alunos selecionados para o estudo. Além disso, o impacto no desempenho dos alunos foi negativo, ou seja, segundo os modelos obtidos, alunos que ingressaram pelo ENEM tem desempenho inferior quando comparados aos do vestibular.

Tabela 11: Medidas de avaliação obtidas do resultado do classificador J48, através da validação cruzada, para a análise do número do número de vezes que as disciplinas introdutórias foram cursadas em relação à evasão.

Instancias corretamente classificadas	82.13 %
Instancias incorretamente classificadas	17.87 %
Kappa	0.644
Erro absoluto médio	0.228
Raiz do erro quadrático médio	0.357
Erro relativo absoluto	45.79 %
Raiz do erro quadrático relativo	71.61 %

Esse é um dado bastante preocupante, visto que essa foi uma mudança ocorreu em escala nacional. Os resultados aqui obtidos não são conclusivos para todo país, mas indicam a necessidade de um estudo em maior escala. Um estudo dessa magnitude se justifica pelo investimento que foi feito por parte do governo federal tanto no próprio exame quanto no número de vagas.

O impacto negativo da mudança de avaliação para SISU/ENEM no desempenho dos acadêmicos pode ter diversas causas. Uma delas é que muitos estudantes chegam a universidade vindos de outras regiões, o que leva a pensar se existe algum aspecto socioeconômico que justifique essa queda. Uma das abordagens para lidar com isto são auxílios fornecidos pela universidade como, por exemplo, moradia para alunos com necessidades como essa. Outra questão importante é que metade das vagas da FURG pelo SISU são reservadas para estuantes provenientes de escola pública, muitos dos quais possuem médias de ingresso bastante inferiores aos da ampla concorrência. No antigo Vestibular não havia reserva de vaga deste tipo, apenas para necessidades específicas causadas por deficiências.

Além disso, identificamos que o número de repetições em disciplinas chaves tem relevância na formação do aluno no curso. Usando os dados de desempenho no ingresso e o número de repetições dos alunos nas disciplinas anteriores, os professores podem ajudar a concentrar esforço do aluno em áreas com problemas potenciais no referido curso. Os educadores também podem usar estas informação para orientar a sua implementação e avaliação das mudanças curriculares.

Neste trabalho, foi relatado o uso de técnicas de análise para construir modelos preditivos. Embora muitos dos modelos gerados não possuírem poder preditivo suficiente para ser útil, os modelos mais fortes e outras observações a partir da análise fornecem indicações úteis sobre as relações entre as disciplinas e o histórico de ingresso do aluno.

Para avaliar o impacto da situação socioeconômica dos alunos e os auxílios recebidos, foi realizada uma tarefa de associação em dados socioeconômicos, dados de auxílios fornecidos pela universidade e os resultados. O resultado desta tarefa, no entanto, não apontou nenhuma regra que indicasse uma falta de suporte, por parte da universidade, ao aluno impactando em seu resultado. E portanto, apesar de não ser conclusivo, esse resultado leva a pensar que de fato pode existir outra causa para essa queda de desempenho.

Portanto, considerando os resultados obtidos nesse trabalho, não foi possível inferir com confiança suficiente a causa exata para o impacto negativo da mudança de método de ingresso, visto que essa mudança não se deu isoladamente. No mesmo período ouve um aumento significativo de vagas em toda rede de ensino superior pública federal, e qualquer conclusão que tente apontar causas para este fenômeno deve levar isso em consideração. Além disso, a avaliação do processo ensino-aprendizagem deve considerar todos os atores que participam direta indiretamente, em todos os níveis da estrutura educacional, isto é, o Estado, a Instituição de Ensino, professores, técnicos e estudantes.

Contudo, o resultado obtido com este trabalho é um indício de que o mesmo fenômeno pode estar acontecendo em muitos cursos da FURG ou em muitas universidades em todo o país. E caso seja verificado que este não é um fenômeno isolado, devem ser tomadas as providencias cabíveis.

Referências

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499.

BHARGAVA, N. et al. Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, v. 3, n. 6, 2013.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. *Proceedings of the Annual Workshop on Computational Learning Theory*. [S.l.], 1992. p. 144–152.

BREIMAN, L. Random forests. *Machine learning*, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

COHEN, W. W. Fast effective rule induction. In: *Proceedings of the International Conference on Machine Learning*. [S.l.: s.n.], 1995. p. 115–123.

DESHPANDE, S.; THAKARE, V. Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS)*, v. 1, n. 1, p. 32–44, 2010.

GAYA, J. O. et al. Vision-based obstacle avoidance using deep learning. *Latim America Robotics Symposium*, 2016.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HAN, J. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 1558609016.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, USA: Prentice-Hall, Inc., 2007. ISBN 0131471392.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. [S.l.], 1995. p. 338–345.

KABAKCHIEVA, D. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, v. 13, n. 1, p. 61–72, 2013.

KHAN, B.; KHIYAL, M. S. H.; KHATTAK, M. D. Final grade prediction of secondary school student using decision tree. *International Journal of Computer Applications*, Foundation of Computer Science, v. 115, n. 21, 2015.

MITCHELL, T. *Machine Learning*. [S.l.]: McGraw Hill, 1997.

MYSQL, A. *MySQL: the world's most popular open source database*. [S.l.]: MySQL AB, 1995.

PANDEY, U. K.; PAL, S. Data mining: A prediction of performer or underperformer using classification. *arXiv preprint arXiv:1104.4163*, 2011.

PLATT, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Proceedings of the Advances in Large Margin Classifiers*. [S.l.]: MIT Press, 1999.

PRASAD, G. N. R.; BABU, A. V. Mining previous marks data to predict students performance in their final year examinations. In: ESRSA PUBLICATIONS. *International Journal of Engineering Research and Technology*. [S.l.], 2013. v. 2, n. (February-2013).

PRASS, F. S. *KKD: Processo de descoberta de conhecimento em bancos de dados*. 2004. 10–14 p. Grupo de Interesse Em Engenharia de Software, Florianópolis, v. 1.

QUINLAN, J. R. Discovering rules by induction from large collections of examples. In: MICHIE, D. (Ed.). *Expert Systems in the Micro-electronic Age*. Edinburgh: Edinburgh University Press, 1979.

QUINLAN, J. R. Induction of decision trees. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 1, n. 1, p. 81–106, 1986. ISSN 0885-6125.

QUINLAN, J. R. Learning with continuous classes. In: *Proceedings of the Australian Joint Conference on Artificial Intelligence*. Singapore: World Scientific, 1992. v. 92, p. 343–348.

QUINLAN, J. R. *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.]: Addison-Wesley, 2005.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2011.

YADAV, S. K.; PAL, S. Data mining: A prediction for performance improvement of engineering students using classification. *CoRR*, abs/1203.3832, 2012. Disponível em: <http://arxiv.org/abs/1203.3832>.