Predição de desempenho em um ambiente virtual de aprendizagem utilizando algoritmos baseados em árvores de decisão e em regras.

Igor Nóbrega dos Santos

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB Novembro - 2017

Igor Nóbrega dos Santos

Predição de desempenho em um ambiente virtual de aprendizagem utilizando algoritmos baseados em árvores de decisão e em regras.

Monografia apresentada ao curso Ciência da Computação
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em Predição de desempenho em um ambiente virtual de aprendizagem utilizando algoritmos baseados em árvores de decisão e em regras.

Orientadora: Thaís Gaudencio do Rêgo

João Pessoa, PB Novembro - 2017

Ficha Catalográfica elaborada por
Rogério Ferreira Marques CRB15/690

S237p Santos, Igor Nóbrega dos.
Predição de desempenho em um ambiente virtual de aprendizagem utilizando algoritmos baseados em árvores de decisão e em regras / Igor Nóbrega dos Santos. – João Pessoa, 2017.
41p. : il.

Monografia (Bacharelado em Ciência da Computação) – Universidade Federal da Paraíba - UFPB.
Orientadora: Profª. Dra. Thais Gaudencio do Rêgo.

1. Ciência da computação. 2. Ambiente virtual. 3. Educação à distância. 4. Mineração de dados. I. Título.

UFPB/BSCI CDU: 004(043.2)

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência da Computação intitulado ***Predição de desempenho em um ambiente virtual de aprendizagem utilizando algoritmos baseados em árvores de decisão e em regras.*** de autoria de Igor Nóbrega dos Santos, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Clauirton de Albuquerque Siebra
Centro de informática - UFPB

Profa. Dra. Thaís Gaudencio do Rêgo
Centro de informática - UFPB

Prof. Me. Daniel Miranda de Brito
Centro de informática - UFPB

João Pessoa, 30 de Novembro de 2017

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

"Seja humilde, pois até o sol com toda sua grandeza se põe e deixa a lua brilhar"

Bob Marley

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus.

A minha família, pois deram o suporte necessário para que eu pudesse entrar na universidade e escrever esse trabalho.

A muitos amigos da universidade, que sempre estavam comigo estudando e ajudando em diversos projetos e provas.

A minha orientadora Thaís por sempre estar disponível para ajudar, e pelos seus ensinamentos.

A todos que colaboraram de forma direta ou indireta para que isso fosse possível.

RESUMO

Mineração de dados é o processo que auxilia na tomada de decisão, realizando a exploração de grande quantidade de dados buscando encontrar anomalias, padrões e correlações. A utilização de técnicas de mineração de dados permite a construção de modelos preditivos para identificar precocemente um aluno com baixo desempenho acadêmico. Foi utilizado o Ambiente Virtual de Aprendizagem Moodle da UFPB Virtual, uma vez que possui dados reais de estudantes. Para prever o desempenho acadêmico, foram utilizados algoritmos de classificação baseados em regras, e em árvore de decisão. A decisão de tal escolha, é devido a relativa facilidade de entendimento e interpretação das regras de decisões e das árvores produzidas por esses algoritmos. Entre os algoritmos utilizados para a predição, o JRip obteve o melhor desempenho preditivo, com acurácia de 88,9%, a menor acurácia foi obtida pelo REPTree com valor de 80%. Foram também obtidas regras possivelmente úteis na predição de desempenho. A vantagem dessa abordagem é informar aos professores e tutores educacionais quando um aluno tem maior chance de apresentar um desempenho ruim, dessa forma medidas preventivas podem ser adotadas

Palavras-chave: Educação à Distância, AVA, Predição de Desempenho, Mineração de Dados.

ABSTRACT

Data mining is the process that assists in decision making, performing the exploration of large amounts of data seeking to find anomalies, patterns and correlations. The use of data mining techniques allows the construction of predictive models to identify early a student with low academic performance. We used the Virtual Moodle Learning Environment of the Virtual UFPB, since it has real data of students. In order to predict academic performance, classification algorithms based on rules and decision trees were used. The decision of such choice is due to the relative ease of understanding and interpretation of decision rules and trees produced by these algorithms. Among the algorithms used for prediction, JRip obtained the best predictive performance, with accuracy of 88.9 %, the lowest accuracy was obtained by REPTree with a value of 80 %. Possibly useful rules were also obtained in predicting performance. The advantage of this approach is to inform teachers and educational tutors when a student is more likely to perform poorly, so preventive measures can be adopted

Key-words: Distance Education, AVA, Performance Prediction, Data Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AVA – Ambiente Virtual de Aprendizagem

CRISP-DM – *Cross-Industry Process for Data Mining* (Processo entre indústrias para mineração de dados)

DM – *Data Mining* (Mineração de Dados)

EAD – Educação a Distância

EDM – *Educational Data Mining* (Mineração de Dados Educacionais)

KDD – *Knowledge Discovery in Databases* (Descoberta de Conhecimentos em Banco de Dados)

UFPB – Universidade Federal da Paraíba

UFPB Virtual - Unidade de Educação a Distância da Universidade Federal da Paraíba

TN – Especificidade

TP – Sensibilidade

Sumário

1 INTRODUÇÃO

A modalidade de ensino à distância vem crescendo não só no Brasil, mas no mundo inteiro. Especialistas consideram-na como o futuro da educação,tornando-se universal pelo fato de trazer em seu bojo maior comodidade e flexibilidade. Possibilitando ao estudante uma melhor integração entre estudo, trabalho, convívio familiar e demais atividades (Macedo, 2016).

Entretanto, esse crescimento traz consigo um problema bastante preocupante, o alto índice de evasão de alunos. De acordo com o Censo de Educação a Distância (EAD) Brasil 2015, as taxas de evasão nos cursos a distância são maiores que nos cursos presenciais, tais taxas registraram no período de 2015-2016 um percentual variando entre 26%-50% (Censo EAD, 2015). Na Figura 1 podemos ver uma relação entre a evasão nos cursos presenciais e a distância das universidades públicas e privadas no ano de 2014 (Capelato, 2016).

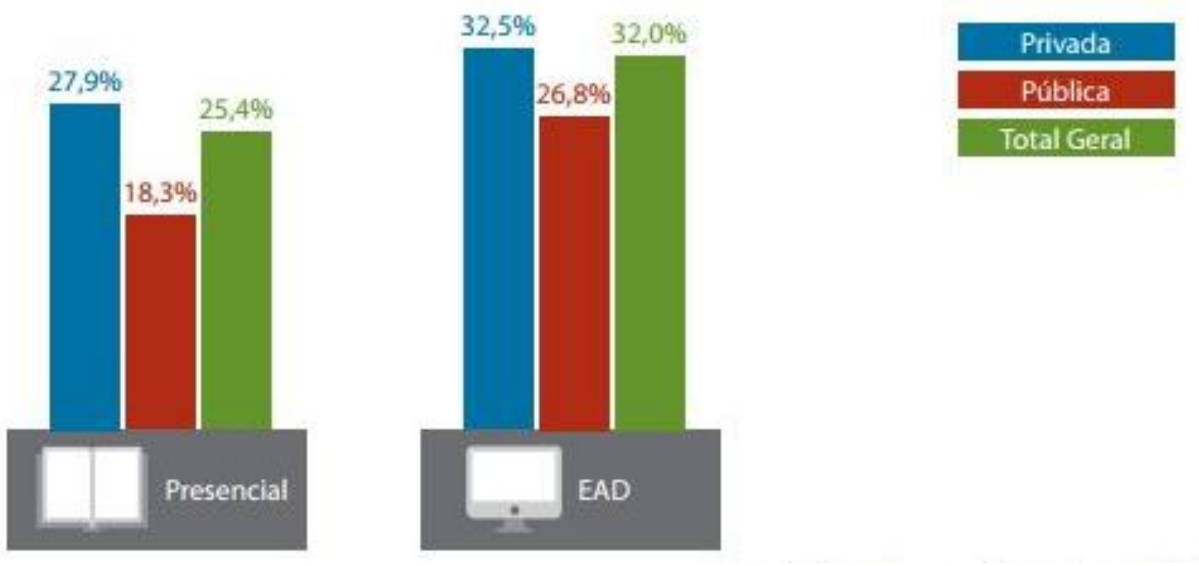


Figura 1: Relação entre a evasão nos cursos presenciais e a distância das universidades públicas e privadas no ano de 2014 Fonte: Sindata/Semesp

A busca por causas para este problema tem sido objeto de estudo de diversas pesquisas e trabalhos. Por exemplo, Bittencourt (2014) discute os fatores que causam evasão nos cursos de modalidade a distância da universidade federal de Alagoas. Entre esses fatores estão: falta de tempo, motivos pessoais, pouca motivação ou incentivo por parte do tutor, falta de didática do professor, insatisfação com o tutor, ausência de tutores nos polos, má infraestrutura, falta de contato com os colegas do curso, atividades complexas e com curto prazo de entrega, reprovação em disciplinas.

De acordo com Shahiri (2015) o Coeficiente de Rendimento Escolar (CRE) é o atributo mais importante na previsão do desempenho de um estudante. De cada 30 artigos

analisados, em 10, o CRE, foi o atributo mais significativo na previsão do desempenho do estudante. Além do CRE existem diversos atributos que podem ser importantes na predição de desempenho de estudantes de um Ambiente Virtual de Aprendizagem (AVA).

Umas das características mas importantes de um AVA é sua capacidade de coletar e armazenar uma grande quantidade de dados sobre estudantes. Nos últimos anos a comunidade científica tem se dedicado em pesquisas buscando criar ferramentas para acompanhar estudantes no AVA (Gottardo, 2013).

Uma solução importante para auxiliar na criação dessas ferramentas é a Mineração de Dados Educacionais (EDM), que se tornou uma área de pesquisa independente a partir de 2008, focada na criação de ferramentas e métodos que devem ser aplicados em dados educacionais, com o objetivo de descobrir informações importantes de como as pessoas aprendem, como esse aprendizado está relacionado com os atributos estudados nos banco de dados educacionais (Merceron, 2005, Qasem, 2017)

A utilização de técnicas de mineração de dados no contexto educacional para predição de desempenho foi feita com sucesso nos seguintes trabalhos Gottardo (2012a), Gottardo (2012b), Minaei-Bidgoli (2003), Oyerinde (2017), Kotsiantis (2012). Tais trabalhos utilizaram técnicas de mineração de dados para a classificação de estudantes, por exemplo, em aprovados ou reprovados de acordo com diversos critérios.

Neste trabalho busca-se realizar a predição de desempenho de estudantes através de um conjunto de atributos que representam um estudante em um AVA, de acordo com a Teoria de Interação em Educação a Distância (Gottardo, 2012b). Esta teoria destaca três tipos de interações relevantes. Essas interações podem ser vistas na Figura 2.

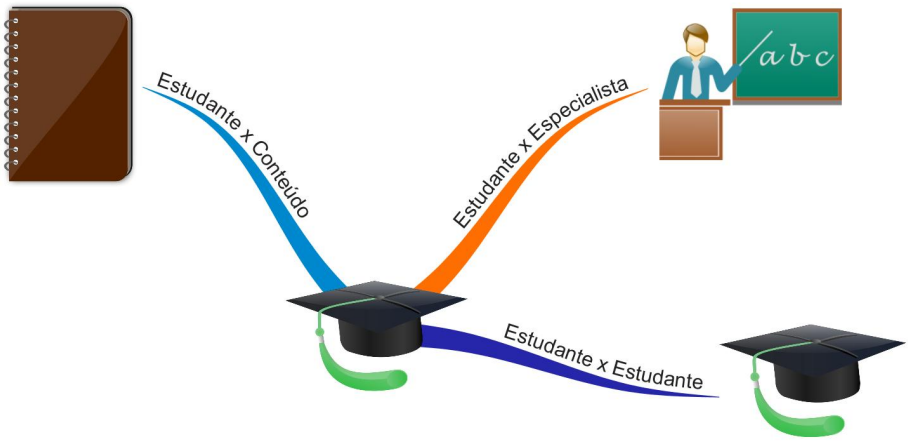


Figura 2: Interações importantes de acordo com a Teoria de Educação a Distância Fonte: Elaborada pelo autor

1- Interação entre o estudante e o conteúdo ou objeto de estudo.

2- Entre o estudante e o especialista que elaborou o material

3- Entre o estudante e outros estudantes, sozinho ou em grupo, mesmo que sem a presença em tempo real de um instrutor.

Com base nessas três interações da teoria de interação em educação a distância acima, definiram-se três dimensões para representar um estudante em um AVA: perfil de uso do AVA, interação estudante estudante, e interação estudante professor nas duas direções. Neste trabalho foram selecionados atributos que estão relacionados com as três dimensões que representam um estudante em um AVA.

1.1 Objetivo geral

O objetivo deste trabalho consiste na aplicação de técnicas de mineração de dados para predição de desempenho no Moodle.

1.2 Objetivos específicos

- Pré processamento e seleção de atributos utilizando a base de dados do Moodle da UFPB Virtual.
- Realizar um comparativo entre a influência de cada atributo sobre o desempenho final dos alunos, baseado no ganho de informação.
- Prever o desempenho de um aluno a partir da geração de regras de decisão

1.3 Estrutura da monografia

O restante deste trabalho está organizado da seguinte forma:

- Capítulo 2: Apresenta-se a fundamentação teórica necessária para o desenvolvimento do trabalho, conceitos de Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases, KDD), aprendizagem de máquina e métodos de classificação;
- Capítulo 3: Apresenta os trabalhos relacionados;
- Capítulo 4: Mostra-se a proposta da abordagem para a previsão de desempenho de estudantes em um AVA;
- Capítulo 5: Os resultados obtidos são apresentados e discutidos;
- Capítulo 6: Exibem-se as considerações finais e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta fundamentos básicos para facilitar o entendimento da abordagem utilizada. São abordados o processo de KDD, suas fases: seleção de dados, pré processamento, transformação dos dados, mineração dos dados, análise dos resultados e apresentação. Também são apresentados os conceitos básicos sobre classificação, indução por Árvores de Decisão que podem ser transformadas em regras. Regras baseadas em classificação. Por fim, são mostradas as métricas utilizadas para medir os desempenhos dos modelos preditivos utilizados como também a ferramenta de mineração de dados que foi utilizada para a construção dos modelos preditivos apresentados neste trabalho: Weka.

2.1 Descoberta de Conhecimento em Banco de Dados e Mineração de Dados

O termo KDD é muito utilizado para indicar o conjunto de passos realizados para a descoberta de conhecimento em banco de dados. Muitas pessoas tratam KDD como sinônimo de mineração de dados, ao mesmo tempo que outros enxergam mineração de dados como um passo essencial no KDD (Han, 2012). O KDD é o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis em dados (Fayyad, 1996).

Dados podem ser entendidos como um fragmento bruto da realidade que possui a capacidade de se transformar em informação. Por exemplo, podemos considerar instâncias de um banco de dados como sendo dados. Padrão por sua vez, pode ser um característica que todos os elementos de um determinado subconjunto do banco de dados possui. De tal forma que, encontrar um padrão nos dados significa achar um modelo que os descrevam. O termo "processo" indica que existe uma sequência de passos no KDD: seleção de dados, pré processamento, transformação de dados, mineração de dados e avaliação do conhecimento.

Por "não trivial" entende-se que não é um processo simples, e sim um processo que muitas vezes inclui o uso de diversos algoritmos presentes no processo de aprendizado de máquina. A extração de conhecimento em bases de dados é um processo dinâmico e que envolve diversas áreas como por exemplo: estatística, inteligência artificial, banco de dados.

2.2 Etapas do KDD

A seguir podemos observar na figura 3 as etapas do processo de KDD.

A primeira etapa do processo é a seleção, depois é realizado o pré processamento onde é feita a limpeza dos dados, ruídos e dados inconsistentes são removidos da base de

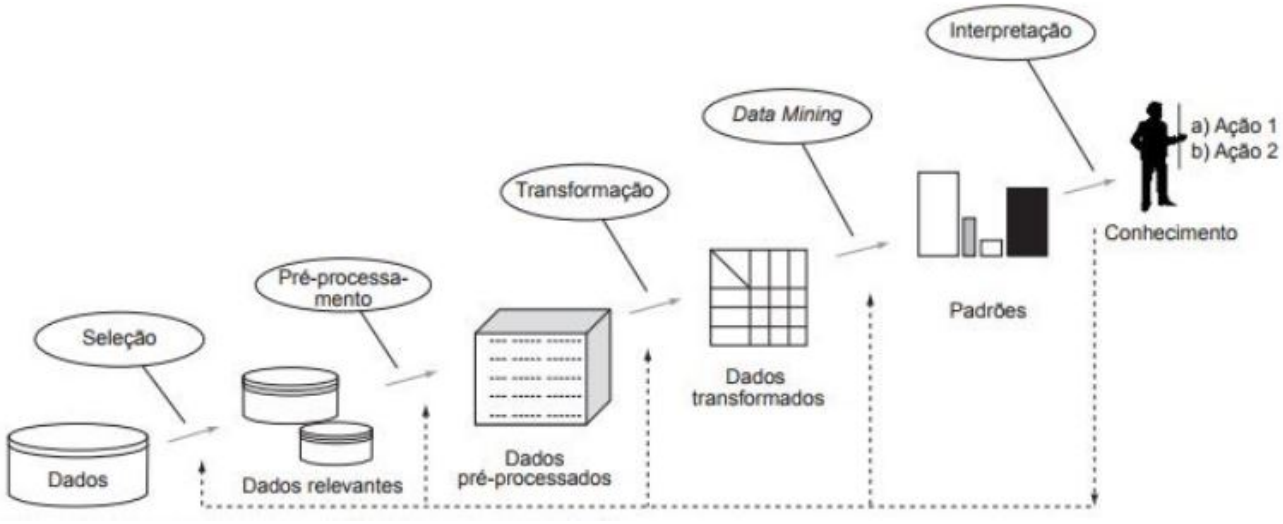


Figura 3: Etapas do processo de KDD Fonte: Fayyad (1996)

dados escolhida. Em seguida, acontece a integração dos dados, nessa etapa muitas bases de dados podem ser combinadas.

A seleção e transformação dos dados é onde acontece a seleção dos dados mais relevantes para a análise, como também a transformação desses dados, de forma que que eles fiquem apropriados para a aplicação dos algoritmos de mineração de dados. Depois de realizada estas etapas, podemos então realizar o processo de mineração de dados, ou seja, métodos inteligentes são aplicados para tentar encontrar padrões nos dados.

Por último, acontece a avaliação dos dados, é nessa etapa que acontece a interpretação dos resultados a fim de descobrir se os resultados realmente geraram "conhecimento". Essa avaliação é feita baseada em medidas de desempenho (Han, 2012).

2.3 Algoritmos de Classificação e Modelos Preditivos

A classificação, também conhecida como reconhecimento de padrão, discriminação, aprendizagem supervisionada ou previsão, é uma tarefa que envolve a construção de um procedimento que mapeia dado em uma das várias classes predefinidas (Chandra, 2012).

Uma importante parte no processo de construção de um classificador é a divisão dos dados em dados para treinamento e dados para teste. O classificador ou modelo de classificação é criado a partir dos dados de treinamento, depois de criado, o classificador deve ser utilizado para classificar os dados de teste, dessa forma é possível ter uma noção de o quão bom é o classificador construído.

No primeiro passo, um classificador é construído descrevendo um conjunto predeterminado de classes de dados ou conceitos. Isto é o passo de aprendizagem (ou fase de treinamento), onde um algoritmo de classificação constrói o classificador analisando ou

"aprendendo de" um conjunto de treinamento composto por tuplas de banco de dados e suas etiquetas de classe associadas. Uma instância X é representada por um vetor de atributos de n dimensões, X= (x1, x2,..., xn) que descreve as medições feitas nas instâncias de uma base de dados de n atributos, respectivamente, $a1$, $a2$,..., an. Cada instância X pertence a uma classe predefinida chamada rótulo de classe.

O atributo classe é discreto e não ordenado. Ele é categórico (ou nominal) em que cada valor serve como uma categoria ou classe. As instâncias individuais que compõem o conjunto de treinamento são conhecidas como tuplas de treinamento e são aleatoriamente selecionados a partir do banco de dados (Han, 2012).

É possível perceber na figura 4, que um conjunto de dados foi utilizado para treinamento e construção do modelo. Em seguida, o modelo construído é aplicado em um novo conjunto de dados, de teste. O modelo será capaz de classificar instâncias dos dados de teste, nas classes predefinidas, baseado no que foi aprendido.

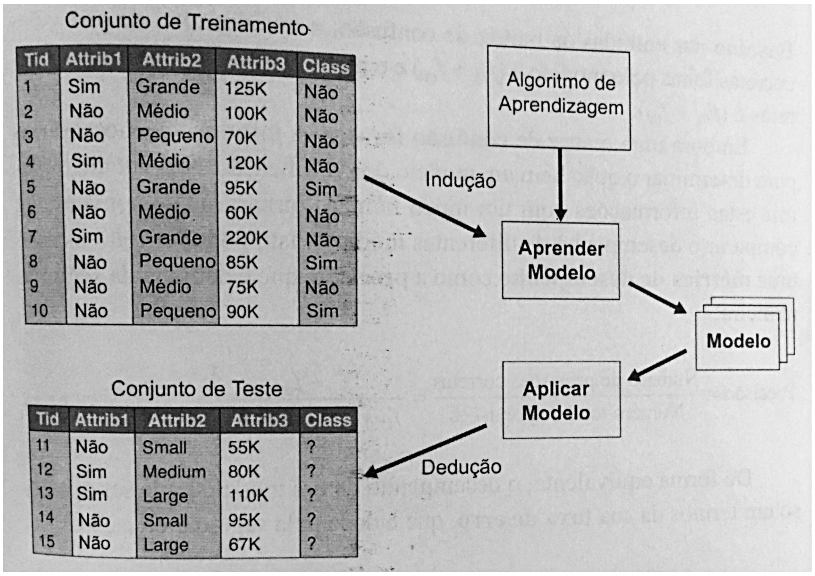


Figura 4: Exemplo de classificação genérico Fonte: Tan (2009)

A seguir, é apresentado o funcionamento de uma árvore de decisão, que está diretamente relacionado com a construção de regras de classificação,utilizadas nesse trabalho.

2.4 Indução por Árvores de Decisão

Indução por Árvores de Decisão é o aprendizado a partir do treinamento de tuplas com rótulos de classes. Uma árvore de decisão é um fluxograma em uma estrutura de árvore, onde cada nó interno (nó não filho) denota um conjunto de atributos, cada ramo representa a saída de teste e cada nó folha (nó terminal) possui um rótulo de classe. Nós internos são denotados por formas ovais e nós folhas são denotados por retângulos (Han, 2012)

O nó mais alto de uma árvore é chamado de nó raiz. Na Figura 5 é mostrada uma árvore de decisão, ela classifica o desempenho de um aluno baseado em alguns atributos, isto é, é predito se o aluno vai ter um desempenho bom ou ruim através do número de *quizz*(número de questionários respondidos por aluno no período analisado) e do número de *postagens*(número de postagens realizadas no fórum durante o período analisado). Podemos perceber que um estudante que respondeu um número <= a 108 questionários foi classificado como possuindo um desempenho ruim, foi possível classificar 26 estudantes corretamente a partir dessa regra. Caso o estudante tenha realizado mais de 108 questionários, a árvore segue para a regra relacionada com o número de postagens no fórum, se for superior a 83, o desempenho do aluno é ruim, se for <= 83 o número de questionários será avaliado novamente, caso seja <= 153, o estudante terá desempenho bom, caso contrário, desempenho ruim.

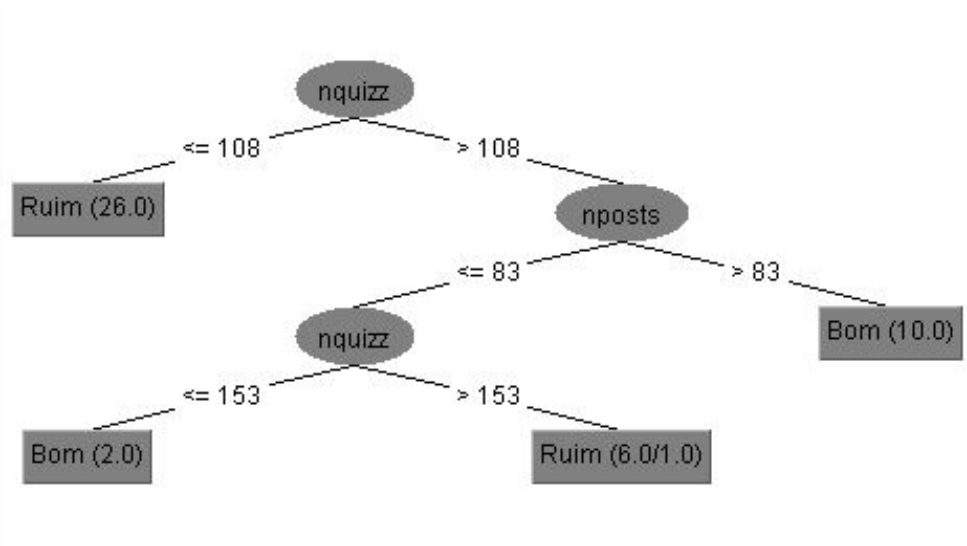


Figura 5: Exemplo de árvore de decisão Fonte: Elaborada Pelo Autor

As árvores de decisão podem realizar uma classificação da seguinte forma: dada uma tupla X, para qual é associada um rótulo de classe desconhecido, os valores dos atributos da tupla são testados contra uma árvore de decisão. Um caminho é traçado da raiz até um nó terminal, que detém a predição de classe para uma tupla. Árvores de decisão podem ser convertidas para regras de classificação (Han, 2012).

As árvores de decisão são populares, pois a construção de um classificador por árvore de decisão não exige nenhum conhecimento do domínio ou configuração de parâmetros e, portanto, é apropriado para descoberta de conhecimentos exploratórios. As árvores de decisão podem lidar com dados multidimensionais. Elas representam o conhecimento adquirido na forma de árvore, que geralmente tem fácil assimilação e uma representação intuitiva (Han, 2012).

Em geral são utilizadas medidas de seleção de atributos para selecionar o atributo

que melhor particiona as tuplas em classes distintas. Uma das medidas mais utilizadas para seleção de atributos é o ganho de informação. Quando as árvores de decisão são criadas, muitos dos ramos podem refletir ruídos nos dados de treinamento. Geralmente, a poda da árvore tenta identificar e remover tais ramos, com o objetivo de aumentar o desempenho do modelo preditivo (Santos, 2015).

Em seguida serão definidos os conceitos de cobertura e precisão de uma regra. Neste trabalho, optou-se por realizar a representação do conhecimento também como regras de classificação, pois dependendo do tamanho das árvores de decisão, podem possuir difícil interpretação.

2.5 Regras Baseadas em Classificação

Nesta seção são analisados os classificadores baseados em regras, onde um modelo de aprendizagem é representado na forma de regras SE ENTÃO. Na Seção 2.5.1 é visto como essas regras são utilizadas para realizar a classificação. Na Seção 2.5.2 é mostrada uma das formas que elas podem ser geradas, através de uma árvore de decisão.

2.5.1 Regras SE ENTÃO para a Classificação

As regras são uma boa forma de representar conhecimento. Um classificador baseado em regras usa um conjunto de regras SE ENTÃO para representar a classificação. Uma regra SE $->$ ENTÃO é expressa da seguinte forma:

SE condição ENTÃO conclusão. Um exemplo de regra: R1.

R1: SE número de tarefas realizadas $>= 80$ E número de postagens criadas no fórum < 50 ENTÃO desempenho = bom.

A parte "SE" (lado esquerdo) da regra é conhecida como antecedente da regra ou precondição. A parte "ENTÃO" (ou lado direito) é a parte consequente da regra. Na regra antecedente, a precondição consiste em um ou mais atributos de teste (ex: número de tarefas realizadas e número de postagens criadas). A regra consequente contém a predição da classe (neste caso, é predito se um aluno vai ter um desempenho bom ou ruim no curso de graduação). R1 pode ser escrita como:

R1: ($ntarefas >= 80$) E ($npostagens < 50$) ($desempenho$ = bom)

Se a condição de uma regra antecedente for verdadeira para uma dada tupla, diz-se que o antecedente da regra é satisfeito ou que a regra é satisfeita ou ainda que a regra cobre a tupla.

A cobertura e a precisão podem avaliar uma dada regra R. Dada uma tupla X, a partir de um conjunto de dados rotulados D, considere $ncobertos$ o número de tuplas

cobertas por R, *ncorretas* o número de tuplas corretamente classificadas por R; e $|D|$ o número de tuplas de D. São definidas a cobertura e a precisão de R como:

$$cobertura(R) = ncobertas/|D| Precisao(R) = ncorretas/ncobertos \quad (1)$$

Em outras palavras, a cobertura de uma regra é a porcentagem de tuplas que foram cobertas pela regra (isto é, todos os seus valores de atributos foram verdadeiros para o antecedente da regra). Para a precisão, é analisada primeiramente a cobertura das tuplas e vista a porcentagem delas que a regra pode classificar corretamente (Han, 2012).

2.5.2 Extração de Regras de uma Árvore de Decisão

Classificadores de Árvores de Decisão são um método popular de classificação,uma vez que é fácil o entendimento de como as árvores trabalham. Entretanto, Árvores de Decisão podem se tornar extensas e difíceis de interpretar. Em comparação com as árvores de decisão, as regras SE ENTÃO são mais fáceis de entender, principalmente se as árvores de decisão forem muito grandes.

Para extrair regras de uma árvore de decisão, uma regra é criada a partir de um caminho percorrido da raiz até um nó folha. Cada critério de divisão ao longo de um determinado caminho é construído a partir da regra antecedente (parte "SE"). O nó da folha contém a predição da classe, formando a regra consequente (parte "ENTÃO") (Han, 2012).

2.6 Métricas para Avaliar o Desempenho de um Classificador

Esta seção apresenta medidas que atestam o quão bom é um classificador para predizer instâncias em rótulos de classes predefinidas. Na maioria das vezes apenas a acurácia não é suficiente para atestar a qualidade de um classificador. Na Seção 2.6.1 é apresentada a Matriz de Confusão, nas Seções 2.6.2, 2.6.3 e 2.6.4 são mostradas, respectivamente, as medidas de acurácia, sensibilidade e especificidade.

2.6.1 Matriz de Confusão

Uma Matriz de Confusão é uma matriz que contém elementos relacionados ao desempenho de um classificador. A escolha do referencial a ser adotado para o modelo preditivo na presente abordagem é baseada na instância que se deseja identificar (classe de interesse) que é um aluno pertencente à classe dos alunos que possuem desempenho

ruim, ou seja, a classe dos alunos com desempenho ruim é considerada a classe positiva. A tabela da Matriz de Confusão é de grande utilidade, pois apresenta a distribuição dos dados entre as classes e possibilita o cálculo de diferentes métricas que atestam o quão bom ou não, é o modelo.

Como o objetivo principal do trabalho é identificar alunos com baixo desempenho, quando ocorre essa identificação, esse comportamento tem um caráter positivo. Dessa forma, se o aluno é classificado como pertencente à classe positiva significa que ele foi classificado com baixo desempenho. Já se um aluno é classificado como pertencente à classe negativa significa que ele foi classificado com um bom desempenho. Na matriz de confusão, quatro situações podem ocorrer. Essas situações são descritas e em seguida mostradas na Tabela 1.

***TP*:** número de instâncias positivas corretamente classificados como positivas.

***FP*:** número de instâncias negativas classificadas incorretamente como positivas.

***FN*:** número de instâncias positivas classificadas incorretamente como negativas.

***TN*:** número de instâncias negativas classificados corretamente como negativas.

***P*:** número de instâncias positivas.

***N*:** número de instâncias negativas.

***n*:** o número total de instâncias que é obtido pela pelo somatório das variáveis: $TP + FP + FN + TN$.

Tabela 1: Matriz de confusão

Classe	Classificado como A+	Classificado como A-
A+	Verdadeiro Positivo (TP)	Falsos Negativos (FN)
A-	Falsos Positivos (FP)	Verdadeiros Negativos (TN)

1. O exemplo pertence à A+ (desempenho ruim) e é predito pelo classificador como pertencente à classe A+ (desempenho ruim). Neste caso, a instância é um verdadeiro positivo (True Positive - TP). Neste trabalho, uma instância pertencente à TP é um aluno que possui um desempenho ruim e foi classificado corretamente como tal. Considere TP o número de instâncias true positives.

2. O exemplo pertence à classe A- (desempenho bom) e é predito pelo classificador como pertencente à classe A- (desempenho bom). Neste caso, a instância é um verdadeiro negativo (True Negative - TN). Neste trabalho, uma instância pertencente

a TN é um aluno que possui desempenho bom e foi classificado corretamente como tal. Considere TN o número de instâncias true negatives.

3. O exemplo pertence à classe A+ (desempenho ruim) e é predito pelo classificador como pertencente à classe A- (desempenho bom). Neste caso, a instância é um falso negativo (False Negative - FN). Neste trabalho, uma instância pertencente a FN é um aluno que possui um desempenho ruim, mas foi classificado incorretamente como possuindo desempenho bom. Considere FN o número de instâncias false negatives.

4. O exemplo pertence à classe A- (desempenho bom) e é predito pelo classificador como pertencente à classe A+ (desempenho ruim). Neste caso, a instância é um falso positivo (False Positive -FP). No contexto do trabalho, uma instância pertencente a FP é um aluno que que possui bom desempenho, mas foi classificado incorretamente como possuindo desempenho ruim. Considere FP o número de instâncias false positives.

2.6.2 Acurácia

A acurácia da classificação é a métrica que informa o desempenho geral de um modelo preditivo. Ela pode ser entendida como o número de predições corretas dividido pelo total de predições feitas, expresso em porcentagem. A acurácia é o número de instâncias corretamente classificadas ($TP + TN$) dividido pelo total de instâncias n ou ($TP+TN+FP+FN$). Foi mostrado na Tabela 1 a relação entre a acurácia e os elementos de uma Matriz de Confusão. Além da acurácia, outra análise deve ser realizada na comparação dos desempenhos preditivos dos algoritmos.

Um classificador pode se diferenciar de outro pelas taxas de acerto e erro na classificação dos exemplos positivos e negativos. Um classificador que possui uma elevada taxa de erro para o falso positivo não é adequado para a solução do problema. Neste caso, considera-se um erro grave do algoritmo classificar um aluno com desempenho ruim como possuindo um bom desempenho. O erro do algoritmo de classificar um aluno com desempenho bom como tendo um desempenho ruim, falso negativo, é considerado um erro menos grave. Assim, foram também utilizadas na análise as medidas: sensibilidade, especificidade.

2.6.3 Sensibilidade

Sensibilidade ou taxa de verdadeiros positivos representado neste trabalho pela a taxa de acerto dos alunos que possuem um desempenho ruim, aqui denominado em TP, é a probabilidade de um teste dar positivo dado que o indivíduo tenha a característica.

No contexto deste trabalho, a sensibilidade é a probabilidade de um aluno ser classificado como possuindo desempenho ruim, dado que o aluno realmente possui desempenho ruim. A sensibilidade é definida como o número de instâncias TP dividido pelo total de instâncias positivas P. Como o P é igual a $TP + FN$, a especificidade pode ser calculada como: $TP/(TP+FN)$. Também é chamado de Valor de Predição Positivo.

2.6.4 Especificidade

Especificidade ou taxa de verdadeiros negativos ou taxa de acerto dos alunos com desempenho bom, aqui denominado TN é a probabilidade de um teste dar negativo dado que o indivíduo não tem a característica. No contexto deste trabalho, a especificidade é a probabilidade de um aluno ser classificado como possuindo um bom desempenho, dado que o aluno possui um bom desempenho. A especificidade é definida como o número de instâncias TN dividido pelo total de instâncias negativas N. Como o N é igual a $TN + FP$, a especificidade pode ser calculada como:$TN/(\ TN+FP)$. Também é chamado de Valor de Predição Negativo.

2.7 Ferramenta WEKA

A ferramenta Weka (*Waikato Environment for Knowledge Analysis*) é uma coleção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados.

Weka foi escolhido para utilização neste trabalho devido a ser uma ferramenta gratuita que possui diversos algoritmos de mineração de dados e por já ter sido utilizado com sucesso em diversas áreas, incluindo Mineração de Dados Educacionais (Education Data Mining, EDM) (Garner,1995).

3 TRABALHOS RELACIONADOS

No Brasil, as oportunidades para a prática da EDM vêm crescendo muito. A criação dos cursos de EAD criou várias oportunidades para as pesquisas na área (Brito, 2014). Vários autores realizam análises e propõem ferramentas no contexto de EDM, embora ela tenha se tornado área de pesquisa oficial a partir de 2008.

Romero (2008) compara técnicas de EDM na classificação de estudantes com base em alguns atributos como: número de tarefas realizadas, número de *quizzes*(questionários respondidos pelo aluno), número de mensagens lidas. As notas dos alunos são utilizadas para definir a classe final. Além disso, uma ferramenta de suporte a tomada de decisão para uso dos instrutores foi construída. Os dados utilizados pela ferramenta são provenientes do Moodle.

Osmanbegović (2015) faz uma análise sobre quais são os fatores determinantes para definir o desempenho de um estudante. São apresentados diversos fatores, em seguida técnicas de filtragem de atributos baseadas no ganho de informação são aplicadas para definir os atributos mais relevantes. Entre eles estão tempo de estudado durante a semana, anos de estudo, tipo de escola que o aluno estudou.

Shahiri (2015) faz uma análise de quais são os algoritmos e também os atributos mais utilizados nas pesquisas que visam prever o desempenho de estudantes. Shahiria conclui ainda que o CRE é o atributo que possui maior relação com a predição de desempenho dos estudantes.

Santana (2014) realiza uma avaliação do perfil de uso do ambiente Moodle utilizando técnicas de mineração de dados. Os atributos analisados estão relacionados com a teoria de educação à distância, que essencialmente define três tipos de interações importantes em um Ambiente Virtual de Aprendizagem(AVA).

Gottardo (2013) analisa a influência de classes desbalanceadas para estimativa de desempenho acadêmico. Os atributos analisados no trabalho também estão relacionados com a teoria da educação à distância.

Baradwaj (2012) propuseram um modelo de Mineração de Dados(DM) baseado em árvore de decisão para analisar a performance de alunos do ensino superior, utilizando os dados de desempenho do fim do semestre.

Santos (2016) realiza uma abordagem para predição de desempenho de estudantes em um AVA utilizando séries temporais e seleção de atributos com o método cápsula, de acordo com Han (2012) o método cápsula já alcançou bons resultados em experimentos com outras bases de dados.

Neste trabalho, além da aplicação dos algoritmos de árvores de decisão e regras,

foram apresentadas regras possivelmente úteis na predição de desempenho baseadas nos atributos analisados. Ademais, os atributos capital(se o estudante vive ou não na capital), descriçãoPerfil(se o estudante possui ou não descrição no perfil) não havia sido citados em nenhum dos trabalhos relacionados.

4 ABORDAGEM PROPOSTA PARA PREDIÇÃO DE DESEMPENHO EM UM AVA

Essa pesquisa é baseada no modelo de Processo entre indústrias para mineração de dados (Cross-Industry Process for Data Mining, CRISP-DM) que é uma abordagem comum na resolução de problemas, que envolvem o descobrimento de conhecimento em banco de dados (Chapman, 2000). Essa metodologia é representada por um processo cíclico, que consiste de 6 fases básicas: entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento. A Figura 6 mostra uma representação do CRISP-DM.

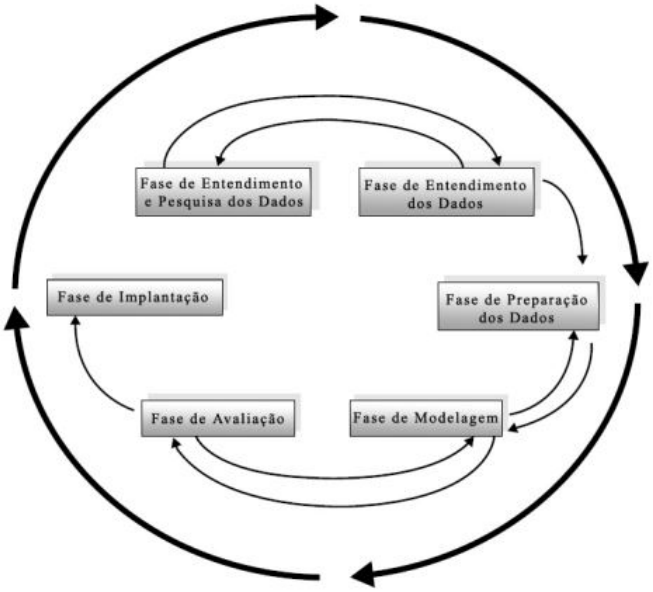


Figura 6: Fases do modelo CRISP-DM Fonte: Larose (2014) adaptado pelo autor

1. **Entendimento do Negócio** - Inicialmente foi realizada uma revisão na literatura disponível para estudar quais problemas existentes no ensino à distância que são resolvidos ou parcialmente resolvidos aplicando algoritmos de mineração de dados. Além disso, a investigação tentou encontrar problemas que ainda não foram resolvidos, e que são considerados muito importantes para o aprimoramento dos sistemas de educação à distância. O problema inicial é transformado em tarefa KDD, ou seja, foi feita uma análise de todas as tabelas do Moodle em busca de quais as informações ou atributos que mais poderiam influenciar no desempenho dos alunos. Em seguida foram usados algoritmos de seleção de atributos para selecionar os atributos mais relevantes para a predição de desempenho dos alunos. Os alunos são classificados em duas classes, que são criadas a partir da informação do CRE de

cada estudante. Seus rótulos podem ser bom ou ruim, referindo-se respectivamente a um bom desempenho (CRE>=5) ou a um desempenho ruim (CRE<5).

2. **Compreensão de Dados** - esta fase na realização da pesquisa começa com a coleta dos dados e continua com atividades de seleção e pré-processamento. É uma atividade chave em projetos de mineração de dados, e afeta, significativamente, a qualidade dos resultados finais. Depois de inspecionar, organizar e descrever todos os dados, é definido um procedimento para coletar e armazenar estes dados, que são principalmente organizados em um sistema de informação de banco de dados, Sistema de Informação de Gestão de Educação (Education Management Information System, EMIS), que no caso deste estudo é o Moodle.

3. **Preparação dos Dados** - Este é um processo muito importante que abrange todas as atividades para construção da base de dados final. Os dados extraídos da base de dados da UFPB Virtual foram inseridos em um banco de dados, que foi transformado em um formato adequado para a análise no software WEKA. Foram analisados os estudantes que ingressaram no curso de Computação da Universidade Federal da Paraíba na modalidade de ensino à distância. Optou-se por considerar apenas os estudantes que ingressaram no primeiro período de 2013, pois acredita-se que a abordagem de classificação apresentada tem melhores resultados para um grupo de alunos do mesmo curso e do mesmo período, porque de acordo com os atributos escolhidos pode haver uma variação considerável deles, de um curso para o outro. No total, após a eliminação de dados incompletos a amostra foi composta por 45 alunos(13 possuindo o desempenho bom e 32 possuindo desempenho ruim). A amostra foi descrita por 11 atributos que podem ser visualizados na Tabela 2.

   O desafio na presente pesquisa de KDD é construir um classificador que consiga prever qual a classe ou categoria de um estudante, uma nova instância, através de atributos que foram extraídos do Moodle. A escolha dos atributos foi baseado na representação de um estudante em um AVA de acordo com a Teoria de Interação em educação a distância, embora o trabalho não tenha se limitado a análise desses atributos, ou seja, mais atributos foram analisados. A classe alvo nesse estudo é o desempenho dos estudantes que pode ser classificado de acordo com a Tabela 3.

4. **Modelagem** - O conjunto de algoritmos escolhidos na etapa de modelagem irá classificar os alunos em duas classes (categorias), dependendo dos seus desempenhos e com base nos dados coletados e utilizados neste trabalho. Métodos de classificação de dados representam o processo de aprendizagem para criação da função que mapeia o dado para uma das classes predefinidas. Todos os algoritmos utilizados nesse modelo para classificação são baseados em aprendizagem indutiva. O conjunto de dados de entrada foi dado e consiste nos atributos selecionados e na classe corres-

Tabela 2: Descrição dos atributos

Atributo	**Descrição**	**Tipo**
id	Número de identificação do estudante	númerico
ntarefasrealizadas	Número de tarefas realizadas pelo aluno durante o período analisado	numérico
nposts	Número de posts realizados no fórum pelo o aluno durante o período analisado	numérico
nlogins	Número de logins feitos pelo estudante durante o período analisado	numérico
nquestionários	Número de questionários respondidos pelo estudante durante o período analisado	numérico
ndiscussões	Número de discussões criadas pelo estudante durante o período analisado	numérico
mtrocadas	Número de mensagens trocadas pelo estudante durante o período analisado (inclui mensagens trocadas com os tutores)	numérico
ndias	Dias decorridos entre o AVA está disponível para acesso e o primeiro acesso do estudante	numérico
idade	Idade do estudante ao final do período analisado	numérico
capital	Se o estudante reside na capital ou não	binário
descriçãoPerfil	Se o estudante possui uma descrição no perfil	binário

Tabela 3: Classificação do desempenho

	CRE>=5	**CRE<5**
Classe	Desempenho bom	Desempenho ruim

pondente. Vários algoritmos diferentes de classificação foram aplicados durante esta pesquisa. Eles foram escolhidos porque têm o potencial de dar bons resultados e por possuírem uma visualização de resultados intuitiva e com relativa facilidade de interpretação (Han, 2012). Classificadores WEKA populares (com suas configurações padrão) foram usados, a menos que indicado de outra forma. Foram utilizados os seguintes algoritmos observados na Tabela 4.

5. **Avaliação** - A avaliação dos resultados obtidos pelos algoritmos utilizados baseia-se principalmente na avaliação dos resultados experimentais. Para a avaliação da acurácia na classificação, foi utilizado o processo de validação cruzada. O processo de estudo e avaliação é repetido N vezes, a cada vez usa-se um conjunto como instância de teste e os demais como treinamento. Como em geral a acurácia não é suficiente para atestar um bom resultado, foram utilizadas também outras métricas

Tabela 4: Classificação dos algoritmos utilizados

Classe	**Algoritmo**
Árvore de decisão	J48
	RandomForest
	RepTree
	DecisionStamp
Regra	JRIP
	OneR
	PART
	DecisionTable

como: sensibilidade e especificidade.

6. **Desenvolvimento** - Esta etapa é mais importante, caso o modelo criado tenha uso diário no AVA. Os resultados adquiridos podem ser utilizados com o propósito de monitoramento de estudantes, apoio na tomada de decisão por parte dos tutores, entre outros. Em todos os casos, é importante que o contratante tenha plena consciência dos limites dos modelos e de todas as ações que são pré-requisito para sua implementação bem sucedida.

5 AVALIAÇÃO EXPERIMENTAL E DISCUSSÕES

Para os fins desta pesquisa, foi utilizado o software WEKA e o conjunto de dados descritos anteriormente. Para testar a precisão dos modelos de classificação adquiridos, foi utilizado o método de validação cruzada com 10 grupos. Os seguintes experimentos foram feitos:

- Seleção de Atributos
- Classificação e Avaliação

Depois da coleta e pré processamento dos dados, uma seleção mais rigorosa de atributos foi realizada. Essa seleção de atributos é necessária para se obter quais atributos tem mais relevância para o classificador, como também quais podem ser descartados. O objetivo do processo de seleção e avaliação dos atributos é mostrar quais os atributos são redundantes e sem importância para o classificador. Dessa forma, é possível eliminá-los e ainda trazer benefícios para a tarefa de classificação. Um aspecto importante para o bom desempenho das técnicas de classificação é a qualidade dos dados da base de treinamento. Atributos redundantes e/ou irrelevantes nas bases de dados de treinamento podem prejudicar a qualidade do classificador e, além disso, tornar o processo de construção do classificador mais demorado (Yang, 1997).

Existem três abordagens padrão para a seleção de atributos: internas, de filtro e envoltório. Neste trabalho escolheu-se a abordagem de filtro, onde os melhores atributos são selecionados antes que o algoritmo de mineração de dados seja executado, usando alguma técnica que seja independente da tarefa de mineração de dados (Tan, 2009). Por exemplo, decidiu-se selecionar os atributos com a menor correlação possível e com o maior ganho de informação.

Para essa tarefa, os seguintes métodos de filtragem foram utilizados *InfoGain* e *GainRatio* com o método de pesquisa Ranke. O *InfoGain* representa uma estimativa do valor do atributo, medindo seu ganho de informação em relação com a classe. *GainRatio* representa uma avaliação do valor do atributo, medindo sua relativa informatividade em relação à classe. Atributos com uma avaliação menor que 0,01 de acordo com os algoritmos utilizados foram descartados, pois eles não têm influência no resultado do classificador. Os resultados da avaliação e a ordem de classificação dos atributos baseados em seus valores individuais é mostrada na Tabela 5.

5.1 Classificação e avaliação

Na segunda fase, foi realizado alguns experimentos para avaliar o desempenho e a utilidade de diferentes algoritmos de classificação, esses algoritmos foram executados

Tabela 5: Ordem do ganho de informação dos atributos

Atributo	**GainRatio**	**InfoGain**
tarefasrealizadas	0,66	0,6197
nquizz	0,542	0,5263
nlogins	0,542	0,5263
nposts	0,542	0,4345
ndiscussões	0,363	0,3625
descriçãoperfil	0,126	0,0714
primeiroacesso	0	0
mtrocadas	0	0
capital	0	0
idade	0	0
sexo	0	0
id	0	0

sobre os dados.

Na Figura 7 a arquitetura geral do modelo é descrita, desde a extração dos dados até a construção do modelo de classificação. A base de dados utilizada é de um Ambiente Virtual de Aprendizagem. Na fase de pré processamento são realizadas consultas ao banco de dados para encontrar quais os atributos importantes para a tarefa de predição de desempenho. Todas as tabelas do moodle utilizadas para a coleta dos dados podem ser visualizadas na figura, são elas: *mdl_log, mdl_user, mdl_assign_submission, mdl_chat_message, mdl_quizz_attemps*. Os atributos mais relevantes podem ser encontrados na Tabela 5. Na próxima etapa, os dados são manipulados para serem usados na ferramenta Weka, e atributos com ganho de informação igual a 0 são desconsiderados. Em seguida, a etapa de mineração de dados é executada, onde são aplicados os algoritmos de árvore de decisão (J48, RandomForest, REPTree, DecisionStump) e regras (JRip, OneR, PART e DecisionTable). Ao final, tem-se a geração de regras e árvores de decisão para auxiliar no suporte educacional.

A Tabela 6 mostra a acurácia, a sensibilidade e a especificidade obtidas pelos algoritmos de classificação usando os atributos finais escolhidos, ou seja, tarefas realizadas, número de *quizzes*, número de *logins*, número de *pots*, número de discussões e descrição no perfil. Os atributos excluídos devido a falta de relevância na predição do desempenho foram: dias decorridos para acontecer o primeiro acesso ao AVA, número de mensagens trocadas, reside na capital, sexo e idade do aluno.

Analisando os resultados, podemos perceber que acurácia para todos os classificadores ficou entre 80%-88,8889%, com índices de sensibilidades variando entre 0,844-0,906 e de especificidade entre 0,692-0,923.

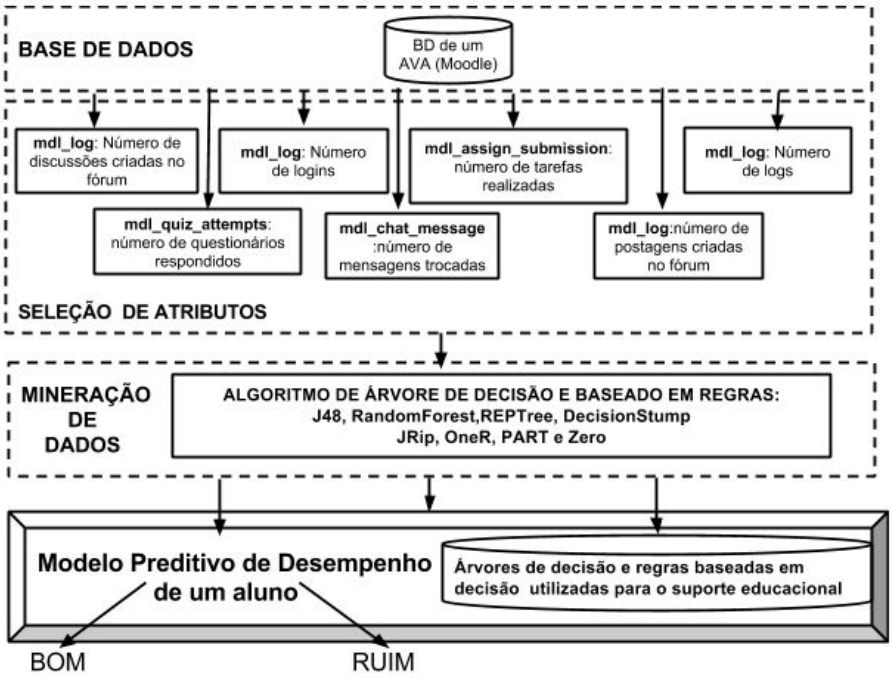


Figura 7: Arquitetura do modelo de predição de desempenho proposto Fonte: Elaborada pelo autor

Tabela 6: Resultados

Tipo do Algoritmo	Algoritmo	Acurácia	Sensibilidade	Especificidade
Árvore de Decisão	J48	88,8889%	0,875	0,923
	RandomForest	86,6667%	0,906	0,769
	REPTree	80,000%	0,844	0,692
	DecisionStump	88,8889%	0,875	0,923
Regras	JRIP	88,8889%	0,906	0,923
	OneR	88,8889%	0,875	0,923
	PART	88,8889%	0,875	0,923
	DecisionTable	86,6667%	0,875	0,846

5.2 Análise Gráfica dos Resultados

A Figura 8 mostra a acurácia dos classificadores em porcentagem. Como pode ser observado todos os métodos obtiveram um percentual acima dos 80% ou seja, bons níveis de acurácia foram registrados por todos os classificadores.

A Figura 9 mostra a relação entre *Sensibilidade* (TP) e *Especificidade* (TN) para todos os classificados. O RandomForest, o REPTree e o Decision Table obtiveram melhores resultados para o TP, que foi bem superior em relação ao TN nos casos do RandomForest e o REPTree, já o restante dos algoritmos tiveram melhores resultados para o TN, porém com o TP não muito distante. Como a predição de um aluno com baixo desempenho é mais importante nesse trabalho, o JRIP obteve o melhor resultado, considerando que teve uma acurácia superior a 80% e sensibilidade e especificidade respectivamente

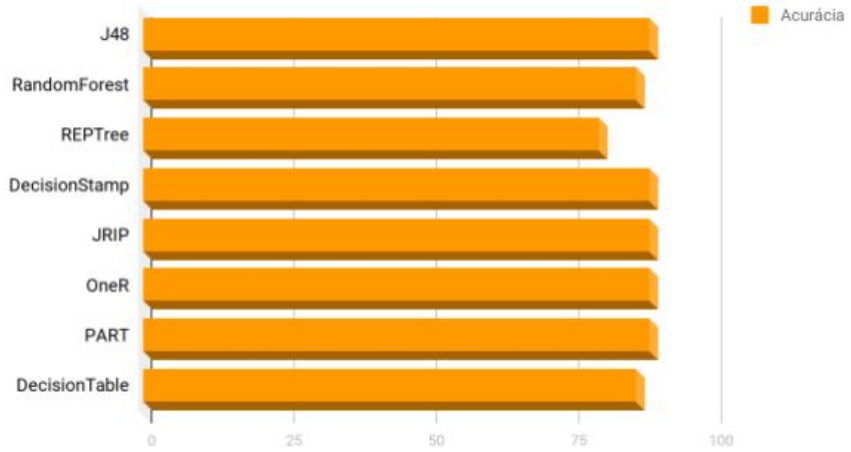


Figura 8: Acurácia dos classificadores em porcentagem Fonte: Elaborada pelo autor

iguais a 0,906 e 0,923. É importante destacar que esses resultados mostram uma boa predição tanto de estudantes com bom desempenho como também de estudantes com um desempenho ruim.

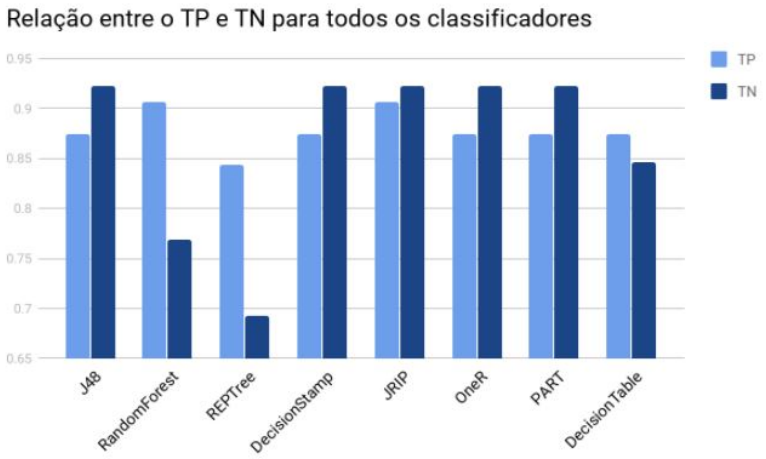


Figura 9: Relação entre TP e TN para todos os classificadores Fonte: Elaborada pelo autor

5.3 Regras de Decisão Geradas

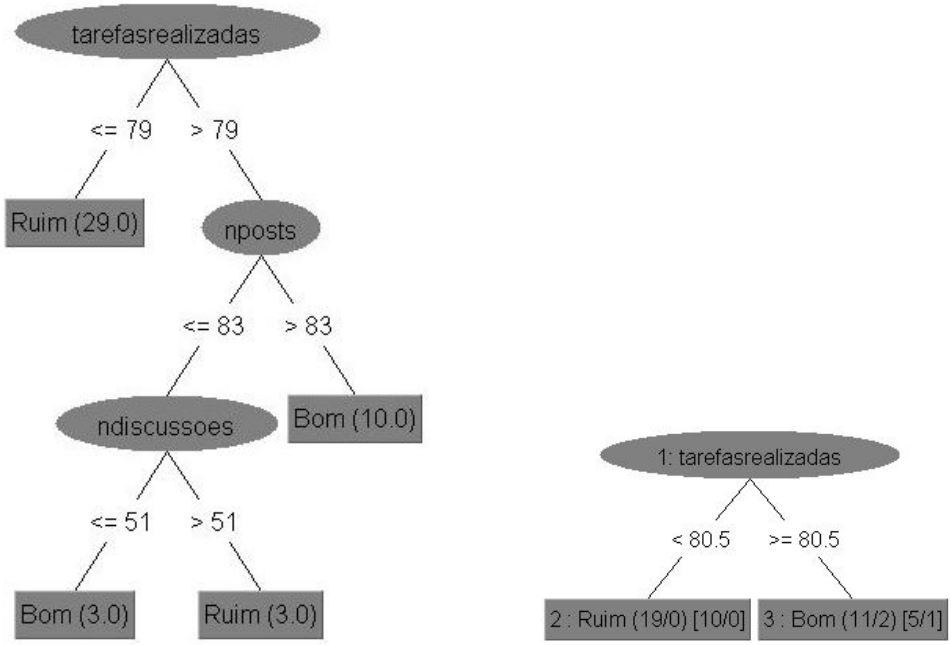
Dos quatro algoritmos de regras utilizados (JRIP, OneR, PART, DecisionTable), três geraram regras que podem ser vistas na Tabela 7. O atributo número de tarefas esteve presente em todas as regras, e o PART apresentou o maior número de regras, associadas com o número de *posts* nos fóruns, número de tarefas realizadas e número de discussões criadas. Além disso, obteve bons resultados de sensibilidade, especificidade, acurácia e cobertura.

Tabela 7: Regras geradas pelos classificadores

JRIP	(tarefasrealizadas >= 82) =>desempenho=Bom (16,0/3,0) =>desempenho=Ruim (29,0/0,0)
OneR	tarefasrealizadas: <80,5 ->Ruim >= 80,5 ->Bom
PART	tarefasrealizadas <= 79: Ruim (29,0) nposts >83: Bom (10,0) ndiscussoes <= 51: Bom (3,0) : Ruim (3,0)

5.4 Árvores Geradas pelos classificadores de Árvore de Decisão

A Figura 10 mostra as árvores de decisão geradas pelos classificadores. Dos quatro classificadores de árvore de decisão, dois geraram árvores que podem ser vistas nas Figuras 10a (J48) e 10b (REPTree). Podemos interpretar a árvore da Figura 10a da seguinte forma: 29 alunos que realizaram um número de tarefas <=79 foram classificados corretamente como possuindo um desempenho ruim. Outros 10 alunos que realizaram mais que 79 tarefas e fizeram mais de 83 postagens foram classificados corretamente como tendo um desempenho bom. Por fim, 6 alunos que realizaram um número de tarefas >=79 e número de postagens <= 83 foram classificados utilizando - se o número de discussões, 3 possuindo desempenho bom(realizaram um número de discussões <= 51) e 3 possuindo desempenho ruim(realizaram um número de discussões >51). Já a Fig 10b, utilizou apenas o atributo número de tarefas para realizar a classificação, os alunos que realizaram um número de tarefas <= 80,5 foram classificados como possuindo um desempenho ruim, com cobertura de 42 %, ou seja, 19 alunos foram classificados corretamente e nenhum de forma incorreta, no conjunto de 45 alunos. Já os alunos que fizeram um número de tarefas >= 80.5 foram classificados como possuindo um desempenho bom, nesse caso a cobertura foi de 24 % dado que dois alunos foram classificados incorretamente, ou seja, possuindo um desempenho ruim e 11 corretamente(possuindo um desempenho bom), no total de 45 alunos. Os valores [10/0] e [5/1] da Figura 10b não foram analisados, pois não são objeto de estudo do trabalho, uma vez que as métricas utilizadas foram, sensibilidade, especificidade, acurácia e cobertura.



(a) Árvore gerada pelo J48 **(b) Árvore gerada pelo REPTree**

Figura 10: Árvores de decisão geradas pelos classificadores Fonte: Elaborado pelo autor

6 CONCLUSÕES

Pode-se afirmar que a qualidade de um curso está relacionada com a qualidade no processo de acompanhamento de estudantes. Entretanto, geralmente, os AVAs disponíveis atualmente não dispõem de ferramentas adequadas para suportar um processo eficiente de acompanhamento de estudantes. Os impactos negativos dessa limitação podem ser potencializados em cursos EAD, que continuamente agregam uma grande quantidade de estudantes. Diante disso, esforços têm sido feitos pelos pesquisadores no desenvolvimento de soluções tecnológicas que forneçam informações relevantes para auxiliar a gestão do processo de ensino desses cursos. Diversos trabalhos têm investigado a aplicação de técnicas de mineração de dados para geração de informações para apoiar o processo de gestão dos cursos. Neste trabalho, essas informações são deduções relativas ao desempenho de estudantes. Essas deduções poderiam ser úteis para professores:

- Acompanhar de maneira individual os estudantes que utilizam um AVA;
- Para definir estratégias pedagógicas específicas que busquem maximizar o desempenho e minimizar reprovações.

Nesta pesquisa, investigou-se quais os atributos de um AVA, especificamente do Moodle, possuem maior relevância na tarefa de prever o desempenho dos estudantes,

como também realizou-se a aplicação de algoritmos de classificação baseados em regras e em árvore de decisão na tarefa de predição de desempenho acadêmico. Esses algoritmos obtiveram uma taxa média de acurácia entre 80%-90%, sensibilidade entre 0,875-0,906 e especificidade 0,692-0,923.

Dos algoritmos analisados, o JRip obteve os melhores resultados na classificação. Além disso, dos atributos investigados, idade, sexo, residir ou não na capital, número de dias decorridos para o primeiro acesso ao AVA, número de mensagens trocadas obtiveram nenhuma relação com a predição de desempenho, portanto, foram desconsiderados pelos classificadores.

Este trabalho busca contribuir com a EDM, onde espera-se que novos esforços sejam feitos em futuras pesquisas na tentativa de trazer para os AVA ferramentas que permitam um acompanhamento efetivo dos estudantes nesses ambientes. Esses recursos poderiam ser de monitoramento e adaptação de conteúdo.

6.1 Contribuições

Análise de classificadores baseados em regra e árvore de decisão para predição de desempenho de estudantes.

Análise da influência de um conjunto de atributos na predição de desempenho dos estudantes.

Fornecer informações importantes para auxiliar trabalhos futuros na área de EDM, como por exemplo, quais tabelas do Moodle foram utilizadas no processo de KDD.

6.2 Trabalhos Futuros

A seguir são apresentadas algumas melhorias que podem ser realizadas em trabalhos futuros:

- Analisar uma quantidade maior de classificadores.
- Desenvolver um sistema de monitoramento para o Moodle, que seja capaz de armazenar alguns atributos utilizados nesse de forma semanal ou em menores períodos.
- Realizar a mesma análise feita por esse trabalho para uma quantidade maior de alunos e em diferentes cursos.

REFERÊNCIAS

Brijesh Kumar Baradwaj and Saurabh Pal. Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*, 2012.

Rodrigo Capelato and et. al. Mapa do ensino superior no brasil 2016. *SEMESP, São Paulo*, 2016.

EAD Censo. Br: Relatório analítico da aprendizagem a distância no brasil 2015= censo ead. *BR: Analytic Report of Distance Learning in Brazil*, 2015.

N. Subhash Chandra, B Uppalaiah, G Charles Babu, K Naresh Kumar, and P Raja Shekar. General approach to classification: Various methods can be used to classify x-ray images. *IJCSET*, 2(3):933–937, 2012.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

Daniel Miranda de Brito, Iron Araújo de Almeida Júnior, Eduardo Vieira Queiroga, and Thaís Gaudencio do Rêgo. Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 25, page 882, 2014.

Ramon Nóbrega dos Santos, Clauirton de Alburqueque Siebra, and EstêvãoDomingos Soares Oliveira. Uma abordagem temporal para identificação precoce de estudantes de graduação a distância com risco de evasão em um ava utilizando árvores de decisão. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3, page 262, 2014.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

Stephen R Garner et al. Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand computer science research students conference*, pages 57–64, 1995.

Ernani Gottardo, Celso Kaestner, and Robinson Vida Noronha. Previsao de desempenho de estudantes em cursos ead utilizando mineraçao de dados: uma estratégia baseada em séries temporais. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 23, 2012a.

Ernani Gottardo, Celso Kaestner, and Robinson Vida Noronha. Avaliação de desempenho de estudantes em cursos de educação a distância utilizando mineração de dados. In

Anais do Workshop de Desafios da Computação Aplicada à Educação, pages 30–39, 2012b.

Ernani Gottardo, Celso Kaestner, and Robinson Vida Noronha. Aplicação de técnicas de mineração de dados para estimativa de desempenho acadêmico de estudantes em um ava utilizando dados com classes desbalanceadas. In *ICBL2013–International Conference on Interactive Computer aided Blended Learning*, 2013.

Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2012.

Sotiris B. Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.

Daniel T Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

Rosely Pereira Costa Macedo and Tarcísio Mendel Almeida. O processo de educação à distancia no ensino profissionalizante. um estudo de caso no noroeste fluminense. *SIED: EnPED-Simpósio Internacional de Educação a Distância e Encontro de Pesquisadores em Educação a Distância*, 2016.

Ibsen Mateus Bittencourt and Luis Paulo Leopoldo Mercado. Evasão nos cursos na modalidade de educação a distância: estudo de caso do curso piloto de administração da ufal/uab. *Ensaio: Avaliação e Políticas Públicas em Educação*, 22(83), 2014.

Agathe Merceron and Kalina Yacef. Educational data mining: a case study. In *AIED*, pages 467–474, 2005.

Behrouz Minaei-Bidgoli, Deborah A Kashy, Gerd Kortemeyer, and William F Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in education, 2003. FIE 2003 33rd annual*, volume 1, pages T2A–13. IEEE, 2003.

Edin Osmanbegović, Mirza Suljić, and Hariz Agić. Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija*, 16(34):147–158, 2015.

O.D. Oyerinde and P.A. Chia. Predicting students' academic performances–a learning analytics approach using multiple linear regression. *perception*, 157(4), 2017.

Mais Haj Qasem, Raneem Qaddoura, and Bassam Hammo. Educational data mining (edm): A review. *New Trends in Information Technology*, page 149, 2017.

Cristóbal Romero, Sebastián Ventura, Pedro G Espejo, and César Hervás. Data mining algorithms to classify students. In *Educational Data Mining 2008*, 2008.

Leandro C Santana, Alexandre MA Maciel, and Rodrigo L Rodrigues. Avaliação do perfil de uso no ambiente moodle utilizando técnicas de mineração de dados. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 25, page 269, 2014.

Rodrigo Santos, Cristiano Pitangui, Luciana Assis, and Alessandro Vivas. Uso de séries temporais e seleção de atributos em mineração de dados educacionais para previsão de desempenho acadêmico. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 27, page 1146, 2016.

Amirah Mohamed Shahiri, Wahidah Husain, et al. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422, 2015.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introdução ao datamining: mineração de dados*. Ciência Moderna, 2009.

Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, pages 412–420, 1997.