Aplicação de Árvore de Decisão para a Determinação do Perfil de Reprovação de Participantes de um Curso a Distância

Thiago de Oliveira Madeira

JUIZ DE FORA

JULHO, 2018

Aplicação de Árvore de Decisão para a Determinação do Perfil de Reprovação de Participantes de um Curso a Distância

Thiago de Oliveira Madeira

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Heder Soares Bernardino

JUIZ DE FORA

JULHO, 2018

Aplicação de Árvore de Decisão para a Determinação do Perfil de Reprovação de Participantes de um Curso a Distância

Thiago de Oliveira Madeira

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Heder Soares Bernardino
D.Sc. em Modelagem Computacional

Luciana Brugiolo Gonçalves
D.Sc. em Ciência da Computação

Victor Stroële de Andrade Menezes
D.Sc. em Engenharia de Sistemas e Computação

JUIZ DE FORA
06 DE JULHO, 2018

Resumo

Com o advento da informática e da *internet*, o ensino a distância se popularizou e instituições de ensino estão cada vez mais adotando plataformas *web* capazes de auxiliar os alunos em locais fora da sala de aula. Com a utilização destas plataformas, diferentes tipos de dados podem ser coletados. Esses dados são provenientes de cadastros, formulários e até mesmo coletados pela plataforma. Neste último caso, os dados podem indicar a maneira como os usuários utilizam estes ambientes virtuais de aprendizagem e interagem com seus colegas ou instrutores. Contudo, este modelo de ensino apresenta um problema referente à retenção de estudantes, sendo comum observar altas taxas de reprovação em cursos aplicados a distância.

Neste contexto, é possível extrair informações a partir de dados para classificar estudantes participantes de cursos a distância. Modelos de classificação podem revelar informações não triviais a respeito dos motivos que levam um estudante a ser reprovado. Um exemplo de modelo de classificação é o de Árvore de Decisão. Este modelo é capaz de classificar um estudante, através de seus dados, e revelar quais são os principais fatores que afetaram o seu desempenho ao final de um curso. Por esta razão, e por apresentar fácil interpretabilidade e alta acurácia, o modelo de classificação de Árvore de Decisão foi escolhido para investigar uma base de dados referente ao "Curso de Prevenção do Uso de Drogas para Educadores de Escolas Públicas" e identificar quais foram os fatores que mais influenciaram no resultado final do desempenho dos participantes deste curso.

Palavras-chave: Mineração de Dados, Árvores de Decisão, Descoberta de Conhecimento, Ensino a Distância

Abstract

With the advent of computing and internet, distance learning has become popular and educational institutions are increasingly adopting web-based platforms that can assist students in places outside the classroom. With the use of these platforms, different types of data can be collected. This data comes from records, forms and even collected by the platform. In the latter case, the data may indicate how users use these virtual learning environments and interact with their colleagues or instructors. However, this teaching model presents a problem regarding student retention, and it is common to observe high failure rates in distance courses.

In this context, it is possible to extract information from data to classify students participating in distance learning courses. Classification models can reveal non-trivial information about the reasons that lead a student to fail. An example of a classification model is the Decision Tree. This model is able to classify a student using their data and reveal which are the main factors that affected their performance at the end of a course. For this reason, and due to its easy interpretability and high accuracy, the Decision Tree classification model was chosen to investigate a database referring to the "Drug Use Prevention Course for Public School Educators" and to identify which were the factors that most affected the performance of the participants of this course.

Keywords: Data Mining, Decision Trees, Knowledge Discovery, Distance Learning

Agradecimentos

A toda a minha família e amigos, pelo encorajamento e apoio.

Ao professor Heder pela orientação, amizade e principalmente, oportunidade de ter aprendido tanto sobre uma área extremamente interessante.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

"The Universe is under no obligation to make sense to you."

Neil deGrasse Tyson

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computução
UFJF	Universidade Federal de Juiz de Fora
KDD	*Knowledge Discovery in Databases*
EAD	Educação a Distância
AVA	Ambiente Virtual de Aprendizagem
kNN	*k-Nearest Neighbors*

1 Introdução

1.1 Apresentação do Tema

Através da utilização dos primeiros meios de comunicação em massa, como correios, rádio, televisão, audiocassete e videocassete, o ensino passou a ser difundido em locais diferentes da sala de aula (VIDAL, 2002). Atualmente, estas tecnologias já se tornaram obsoletas e instituições de ensino estão cada vez mais adotando Ambientes Virtuais de Aprendizagem (AVA) para auxiliar os seus cursos aplicados à distância (MINAEI-BIDGOLI et al., 2003).

Com a propagação da utilização de AVA por parte das instituições de ensino, grandes quantidades de dados referentes ao padrão de utilização destas plataformas passaram a ser geradas e coletadas (MINAEI-BIDGOLI et al., 2003). Em geral, dados podem ser amplamente explorados a fim da descoberta de conhecimentos que sejam úteis e possam ser aplicados para a melhoria e desenvolvimento de áreas objetos de estudo (HAN; PEI; KAMBER, 2011).

Segundo Norton (1999), análises e classificações utilizando tais dados podem ser feitas manualmente; contudo, este método é trabalhoso e dificulta a descoberta de informações não triviais, o que é suficiente para torná-lo inviável. Para auxiliar neste processo, é possível utilizar-se de algoritmos de aprendizado de máquinas, em especial, aqueles ditos supervisionados. Um algoritmo de aprendizado de máquina supervisionado tem como principal característica a capacidade de encontrar regras e padrões que classificam um conjunto de dados apresentado como entrada e, através destas regras, gerar um modelo de classificação de elementos, com base em suas características e percepção de padrões (RUSSELL; NORVIG, 1995).

Um exemplo de algoritmo de aprendizado de máquina supervisionado é a Árvore de Decisão, cujas características estão descritas no capítulo 3 deste trabalho. Ao aplicar um modelo de classificação baseado em árvores de decisão em uma base de dados de um AVA, é possível classificar os alunos que participam de cursos à distância suportados por tais plataformas *web* e identificar quais são os atributos que mais tiveram impacto nesta

classificação (MINAEI-BIDGOLI et al., 2003).

1.2 Problema

Netto, Guidotti e Santos (2017) diz em seu trabalho que a Educação a Distância (EAD) tem como maior virtude o livre arbítrio, por parte do educando, na escolha do local e horário para estudar. Contudo, esta possibilidade pode transformar-se em um problema se o aluno não apresentar disciplina para cumprir as rotinas de estudo. Tais facilitadores trazem consigo um importante efeito colateral, as altas taxas de reprovação, muitas vezes associadas à evasão (NETTO; GUIDOTTI; SANTOS, 2017). Faz-se necessário entender e minimizar quais são os principais indicadores de que um estudante de EAD é propenso a reprovar, de forma a melhorar a eficácia da aplicação de cursos a distância por parte de instituições de ensino.

Metodologias de ensino à distância estão sendo cada vez mais utilizadas, portanto, avaliar se elas estão sendo seguidas por bons resultados faz-se necessário para a otimização de investimentos e identificação de pontos onde os métodos utilizados possam ser melhorados (MONTEIRO et al., 2016).

1.3 Justificativa

Segundo Pereira e Zambrano (2017) e Monteiro et al. (2016), os conhecimentos obtidos através da aplicação de técnicas de descoberta de informações são capazes de auxiliar na economia de recursos, como tempo e dinheiro, tanto por parte da instituição quanto por parte do aluno. Uma vez feita a classificação dos perfis de estudantes de um curso à distância, é possível viabilizar a adoção de medidas pedagógicas interventivas, com o intuito de reduzir as taxas de reprovação e, até mesmo, taxas de evasão.

Tais conhecimentos podem ser obtidos através da escolha de Árvore de Decisão como modelo de classificação. Este tipo de modelo oferece fácil interpretabilidade e visualização de resultados, alta acurácia e é capaz de operar com dados multidimensionais (HAN; PEI; KAMBER, 2011).

1.4 Objetivos

O objetivo geral deste trabalho é identificar o perfil de um participante do "Curso de Prevenção do Uso de Drogas para Educadores de Escolas Públicas" com potecial de reprovação. Dentre os objetivos secundários, estão:

1. Realizar o tratamento dos dados coletados da plataforma *web* de apoio ao curso;
2. Gerar diferentes modelos de classificação baseados em Árvore de Decisão, utilizando diferentes grupos de atributos como entrada;
3. Avaliar a acurácia dos modelos de classificação gerados;
4. Identificar a ordem de relevância dos atributos para o desempenho final de um participante do curso;
5. Gerar um modelo de classificação baseado em Árvore de Classificação, capaz de predizer o desempenho final de participantes de futuras edições do "Curso de Prevenção do Uso de Drogas para Educadores de Escolas Públicas"
6. Identificar estudantes com potencial de reprovação em futuras edições do curso, possibilitando assim a tomada de medidas interventivas em tempo hábil.

1.5 Organização do Texto

O primeiro capítulo introduz o tema do trabalho, nele são apresentados o problema a ser estudado e os objetivos pretendidos. O capítulo seguinte contém um levantamento da literatura sobre trabalhos relacionados, onde é descrito como outros autores conduziram seus projetos e obtiveram os seus resultados. O terceiro a estrutura e as etapas de construção da árvore de decisão utilizada neste trabalho. No quarto capítulo, são detalhadas as etapas de pré-processamento da base de dados utilizada e, no próximo capítulo, os experimentos realizados e resultados são obtidos e comparados com o da literatura levantada. Por fim, o sexto capítulo traz as considerações finais.

O apêndice A apresenta os atributos mantidos da base de dados, os seus tipos, os seus potenciais valores e a sua descrição. O apêndice B apresenta os atributos excluídos

da base e a razão da exclusão. O apêndice C apresenta uma descrição mais detalhada de um subgrupo de atributos do apêndice A, referentes a um questionário social.

2 Revisão Bibliográfica

A EAD, através de plataformas *web*, é uma ferramenta de aprendizagem que é cada vez mais adotada por instituições de ensino. A evolução e disseminação desta tecnologia possibilita que alunos possam ter acesso aos conteúdos que estão estudando, e que também possam ser avaliados, em locais diferentes das salas de aulas. Porém, esta facilidade traz consigo um efeito colateral: os cursos de EAD apresentam taxas de reprovação maiores que os modelos de ensino tradicionais (MINAEI-BIDGOLI et al., 2003).

Para identificar potenciais fatores que levam à reprovação de um estudante, técnicas de Mineração de Dados podem ser aplicadas em dados provenientes de plataformas *web* de EAD. A fim de compreender melhor este fenômeno, e identificar quais características classificam um estudante como propenso, ou não, a reprovar, as seções seguintes descrevem trabalhos já publicados na área.

2.1 Qualidade dos Dados

Objetivando classificações com alta acurácia, é necessário utilizar dados refinados por técnicas de pré-processamento (HAN; PEI; KAMBER, 2011), como foi feito no trabalho de Pereira e Zambrano (2017). Foram removidos da base de dados registros caracterizados como *outliers* e atributos que possuíam alta porcentagem de valores nulos. Igualmente, também foram removidos atributos que não apresentassem representatividade sobre os resultados de classificação desejados. Infelizmente, o autor não cita quais foram os atributos removidos da base de dados, a estratégia utilizada para a substituição de valores ausentes e o limiar utilizado para excluir atributos com muitos valores nulos.

2.2 Modelos de Classificação

Para o desenvolvimento deste referencial bibliográfico, foram levantados trabalhos que utilizam, ao menos, Árvores de Decisão como modelos de classificação para prever o de-

sempenho de estudantes ao final de um curso EAD. Alguns trabalhos optaram por mais de um modelo de classificação, como por exemplo, o de Dias et al. (2008), que além de árvores de decisão, utiliza-se também de Redes Bayesianas para realizar as suas classificações. Já o modelo de classificação proposto por Minaei-Bidgoli et al. (2003), destaca-se pela utilização da Combinação de Múltiplos Classificadores (CMC). O resultado desta abordagem é obtido através da combinação dos resultados de diferentes classificadores. Em particular, o trabalho deste autor utiliza seis modelos de classificação, sendo eles: o Classificador Bayesiano, o 1 vizinho mais próximo (1-NN), o k vizinhos mais próximos (k-NN), Janelas de Parzen, Perceptron Multicamadas e Árvore de Decisão. Este mesmo trabalho também aplicou algoritmos genéticos para encontrar os melhores parâmetros a serem utilizados pelos algoritmos de classificação. O uso desta abordagem garantiu melhorias de aproximadamente 10% na acurácia dos resultados obtidos através da CMC. Além de utilizar os modelos para classificar o desempenho de um estudante ao final de um curso, utilizou também estas abordagens para classificar problemas que eram submetidos aos estudantes durante a aplicação do EAD. Desta maneira, seria capaz de identificar quais são as atividades que mais influenciam no desempenho de um estudante e também seriam capazes de auxiliar os produtores de conteúdo do curso.

O trabalho de Pereira e Zambrano (2017) utiliza apenas um modelo de árvore de decisão para realizar as classificações e, além de conseguir predizer com boa acurácia o desempenho final de estudantes ao final de um curso, também é capaz de identificar quais foram os fatores que mais influenciaram na obtenção destes desempenhos.

Durante a criação dos modelos de classificação, todos os trabalhos utilizaram como ferramenta de análise de acurácia a Validação Cruzada com 10 Dobras. Esta abordagem de avaliação tende a particionar a base em tamanhos iguais e realiza uma sequência de classificações intercalando as partições criadas ora como conjunto de treinamento, ora como conjunto de testes. O número de dobras (partições) igual a 10 é utilizado devido à sua tendência de gerar resultados com viés e variâncias relativamente baixos (HAN; PEI; KAMBER, 2011).

A fim de gerar modelos capazes de classificar o desempenho de estudantes ao final de um curso de EAD, os algoritmos de classificação dos trabalhos levantados foram

alimentados com dados pertencentes a dois grandes grupos: dados sociodemográficos e dados de utilização de plataformas *web*.

2.2.1 Dados sociodemográficos

São os dados coletados no momento inicial dos cursos, comumente obtidos através de cadastros ou formulários; devido à natureza desta coleta, é comum apresentarem grandes quantidades de valores ausentes e inconsistentes (PEREIRA; ZAMBRANO, 2017). Exemplos deste tipo de dado são sexo, data de nascimento, se mora ou não com os pais, local de residência, faixa de renda e se possui ou não familiares cursando o mesmo curso.

2.2.2 Dados de utilização de plataformas *web*

Estes dados são coletados durante e ao final da aplicação dos cursos; devido à natureza desta coleta, é comum estarem presentes em menor quantidade que os dados sociodemográficos, refletindo assim a tendência de evasão, e consequentemente de reprovação, do EAD (NETTO; GUIDOTTI; SANTOS, 2017). Exemplos deste tipo de dado são número de acesso aos materiais de estudo, tempo de permanência na plataforma, número de interações com outros estudantes e instrutores, tempo gasto para realizar atividades, desempenho e quantidade de atividades realizadas.

2.3 Padrões Encontrados

Os resultados de Pereira e Zambrano (2017), obtidos através de árvores de classificação, indicam que é possível gerar modelos de classificação capazes de refletir a realidade. Através dos dados coletados de plataformas *web* de apoio ao EAD, chegaram à conclusão de que os atributos sociodemográficos são os principais fatores que levam um estudante a ser reprovado. Alguns dos padrões encontrados são altos índices de evasão entre estudantes que pagaram taxas mais altas de matrícula na universidade, estudantes provenientes do sul da Colômbia (país onde se localiza a universidade que aplicou o EAD e também realizou este trabalho), estudantes que vivem com a mãe e estudantes solteiros.

Por outro lado, Minaei-Bidgoli et al. (2003) e Dias et al. (2008) apresentam dados

de utilização da plataforma *web* como sendo os mais relevantes. O primeiro, identifica os seguintes atributos, listados por ordem de relevância, como principais indicadores de propensão à reprovação: quantidade de atividades corretas, quantidade de tentativas até obter um acerto, se acertou ou não uma atividade na primeira tentativa, tempo gasto para resolver as atividades, tempo de permanência na plataforma e comunicação com outros alunos e instrutores. Já Dias et al. (2008), identificou que os estudantes mais motivados a permanecer no curso de EAD que forneceu a base de dados estudada, são aqueles que completam atividades com um grau de dificuldade acima da média e aqueles que mais cedo começam a utilizar a plataforma, em relação ao dia de matrícula. E, como principal fator que leva à reprovação, estudantes que completam poucas atividades e acessam menos vezes a plataforma de apoio *web*.

Todos os trabalhos levantados apresentam modelos de classificação com acurácias maiores do que 80%.

3 Árvore de Decisão

Um modelo de classificação é uma tentativa de extrair algum tipo de conhecimento a partir de uma base de dados (FUCHS, 2017) e são capazes de auxiliar em problemas de classificação (ALPAYDIN, 2010). Um problema de classificação consiste em identificar em qual conjunto de categorias um novo elemento observado pertence, baseado em dados de um conjunto de treinamento cujas categorias são conhecidas (ALPAYDIN, 2010). Exemplos de modelos de classificação são o de Regressão Logística, o Classificador Bayesiano, o k vizinhos mais próximos (k-NN), Janelas de Parzen, Perceptron Multicamadas e Árvore de Decisão (MINAEI-BIDGOLI et al., 2003). Cada um destes modelos possui suas próprias características e problemas para os quais são mais recomendados serem utilizados (HAN; PEI; KAMBER, 2011).

O modelo de classificação utilizado neste trabalho é o de Árvore de Decisão. Esta escolha foi feita pois modelos de classificação baseados em árvores de decisão são capazes de gerar uma representação visual altamente intuitiva para o ser humano, geralmente apresentam boa acurácia e são a base de vários sistemas de indução de regras comerciais (HAN; PEI; KAMBER, 2011).

3.1 Estrutura de uma Árvore de Decisão

Uma árvore de decisão é um fluxograma em forma de árvore. Cada nó folha desta árvore representa alguma das classes alvo do modelo e cada nó interno representa um teste em um atributo. As possíveis respostas para estes testes determinam o caminho a ser percorrido na árvore. É possível utilizar-se de uma árvore de decisão para classificar uma fruta, através das questões “o peso é menor ou igual que 500 gramas?” ou “a cor é verde?” Após cada teste, este registro irá percorrer a árvore e terminar em um nó folha, que indicará a sua classificação. Em uma árvore de decisão, quanto mais próximo da raiz da árvore o nó correspondente ao teste de um atributo estiver, maior será o peso deste atributo para o resultado de classificação deste modelo (HAN; PEI; KAMBER, 2011). Caso este teste

estivesse relacionado a um valor inteiro, a importância em relação à aproximação à raiz é do atributo e do valor, não apenas do atributo (HAN; PEI; KAMBER, 2011). Através do caminho que for percorrido a partir da raiz até um nó folha, regras de classificação podem ser facilmente extraídas, como por exemplo, caso a respostas para a primeira pergunta citada acima seja "sim" e para a segunda seja "não", pode-se inferir que a fruta cujo peso for menor ou igual que 500 gramas e não for verde, será uma maçã.

Para gerar um modelo de classificação, é necessária a definição de dois conjuntos de dados (HAN; PEI; KAMBER, 2011): um contendo os atributos que serão utilizados para realizar a classificação (como por exemplo, "quantidade de sementes", "cor" e "peso") e outro contendo as possíveis classes de saída do modelo (como por exemplo, "maçã" e "melância"). A Figura 3.1 apresenta um exemplo de árvore de decisão. Neste exemplo, deseja-se identificar se uma fruta desconhecida é uma maçã ou uma melancia e considera-se que o atributo mais importante para a classificação da fruta é o seu peso, logo, o teste para o peso está na raiz da árvore. Em seguida, verifica-se a cor, a quantidade de sementes e obtém-se a classificação da fruta entre maçã ou melancia.

3.2 Construção de uma Árvore de Decisão

A árvore de decisão utilizada neste trabalho particiona a base de dados através de testes binários, que apresentam "sim" ou "não" como resposta. Os nós desta árvore são construídos através de uma heurística gulosa, ou seja, o nó raiz irá possuir como teste aquele atributo que dividirá os dados em partições o mais pura possíveis. Uma partição é considerada pura se todos os seus registros pertencem a uma única classe (HAN; PEI; KAMBER, 2011). Para medir a impuridade de uma partição, pode-se utilizar o índice Gini (HAN; PEI; KAMBER, 2011). Quanto menor for o índice Gini de uma partição, mais pura ela será (HAN; PEI; KAMBER, 2011). O índice Gini de uma partição D é calculado por

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2, \tag{3.1}$$

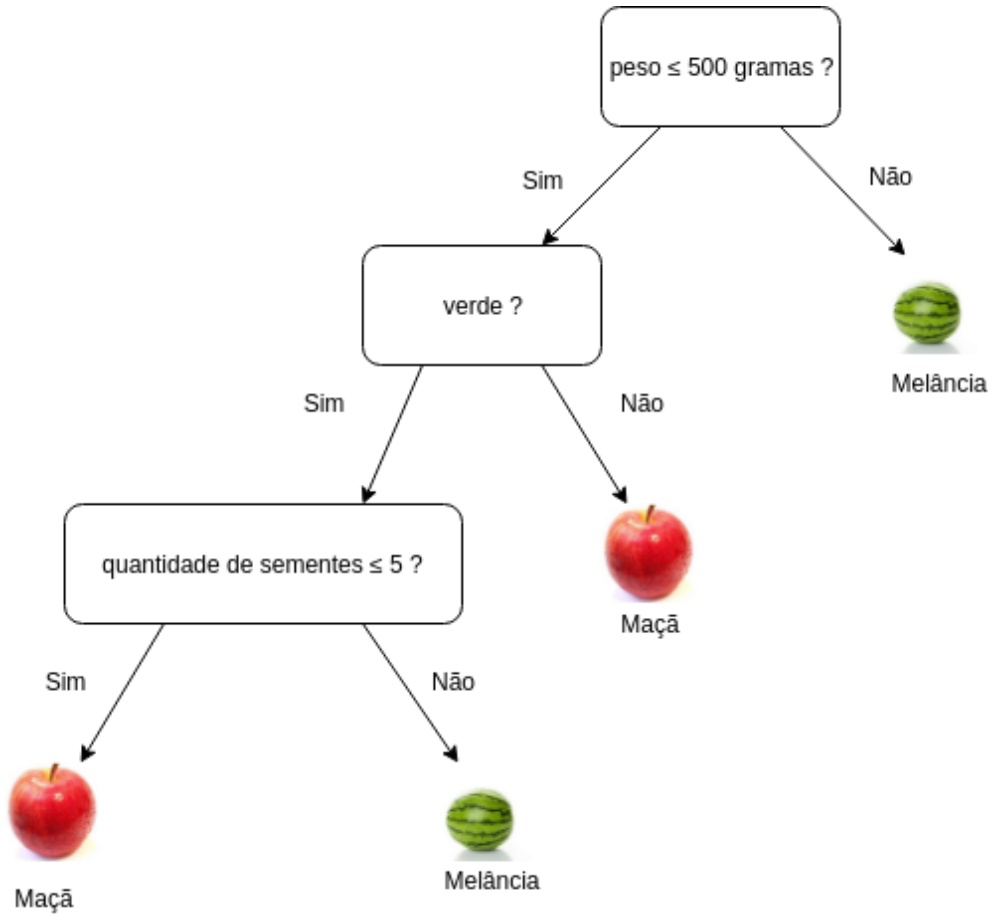

Figura 3.1: Exemplo de uma Árvore de Decisão capaz de classificar uma fruta entre maçã ou melancia.

onde p_i é a probabilidade de um registro da partição D pertencer a uma classe C_i e m é a quantidade de classes utilizadas na classificação (HAN; PEI; KAMBER, 2011).

Após a seleção do nó raiz, os dados estarão subdivididos em dois grupos: um grupo que responde "sim" para o teste da raiz e um grupo que responde "não". Para cada um destes grupos, seleciona-se novamente um atributo que dividirá os dados em partições o mais pura possíveis. Este processo é repetido até que obtenha-se nós com índices Gini igual a 0, ou seja, partições completamente puras.

4 Pré-processamento dos dados

Para obter modelos com alta acurácia, é desejável que os dados utilizados na classificação sejam tratados por técnicas de pré-processamento. Bases de dados podem apresentar diversos elementos prejudiciais para os modelos de classificação, como por exemplo, valores inconsistentes, valores ausentes, existência de *outliers* e registros duplicados (HAN; PEI; KAMBER, 2011). As seções seguintes descrevem os passos seguidos para a preparação dos dados e a descrição detalhada dos atributos utilizados e removidos pode ser encontrada nos apêndices A e B, respectivamente. Existe um subgrupo de atributos no apêndice A, referentes a um questionário social, que está melhor descrito no apêndice C.

Os dados utilizados neste trabalho foram coletados da plataforma *web* que serviu de apoio para o "Curso de Prevenção do Uso de Drogas para Educadores de Escolas Públicas", que foi aplicado em 2012 para cerca de 10 mil educadores da rede pública dos estados de Minas Gerais, Rio de Janeiro e Paraná. Segundo Monteiro et al. (2016), a maioria dos participantes deste curso eram mulheres com ensino superior completo; e esta informação verificou-se verdadeira através da investigação da base de dados. Ao final do processamento, a base de dados que continha inicialmente 10974 registros e 202 atributos, passou a ter 1516 registros e 118 atributos. Destes 1516 registros, 868 são participantes reprovados e 648 participantes aprovados. A Figura 4.1 apresenta um fluxograma representando as etapas de pré-processamento dos dados.

4.1 *Outliers*

Segundo John (1995), registros classificados como *outliers* representam um problema para a base de dados e os mesmos devem ser tratados através de técnicas de mineração de dados. Um registro é dito *outlier* caso apresente valores discrepantes em relação a maioria dos registros da base de dados. Porém, Dimensionless (2016) afirma que modelos de classificação baseados em árvores são insensíveis a *outliers*. Logo, os *outliers* encontrados na base de dados deste trabalho foram apenas ignorados.

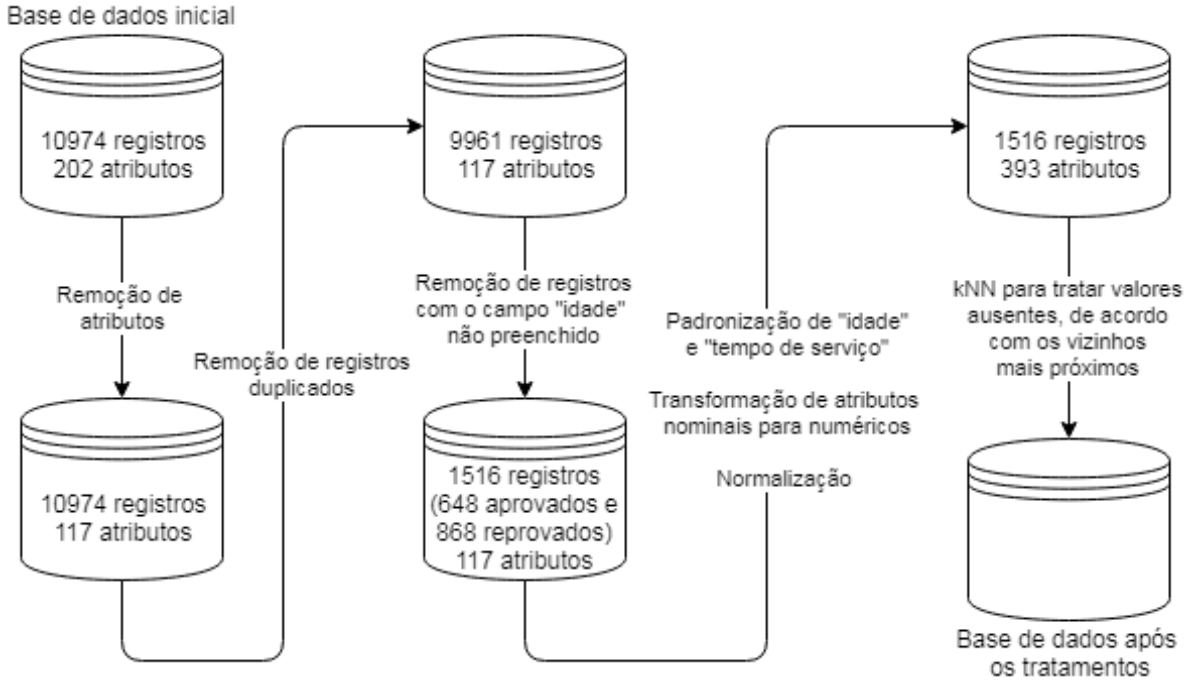

Figura 4.1: Fluxograma representando as etapas do pré-processamento dos dados

4.2 Eliminação de linhas e colunas da base de dados

Os atributos descritos no apêndice B foram eliminados da base de dados, pois não apresentariam nenhuma melhoria na classificação caso fossem utilizados. Alguns dos atributos excluídos apresentavam baixa relevância para o modelo de classificação, como por exemplo, atributos contendo CPF e RG. Também existiam atributos redundantes, como por exemplo, múltiplas colunas com as informações de nome, sobrenome e *e-mail*. Outro tipo de atributo removido da base de dados são os ilegíveis. Este trabalho classifica um atributo como ilegível se, através da investigação do seu nome ou potenciais valores, não for possível identificar o que ele representa. Também é classificado como ilegível aquele atributo que foi preenchido pelos próprios participantes, em campos de texto livre, e não se mostrou possível extrair algum conhecimento para ser aplicado no modelo de classificação. Além destes, removeu-se também aqueles atributos que não apresentavam uma quantidade suficiente de valores presentes, definida arbitrariamente como 300 (aproximadamente 20% da quantidade de registros utilizados na classificação). Por último, também foram excluídos atributos que foram obtidos de forma equivocada, como por exemplo, `sumTime`, que deveria exibir o tempo que um participante permaneceu conectado no AVA utilizado pelo curso. Contudo, ele apenas exibe o tempo de permanência de uma seção, cuja iden-

tificação não se mostrou possível. Esta informação foi obtida através da investigação do cruzamento de atributos correlacionados. O atributo `sumTime` é obtido através da diferença entre os atributos `iniciado.y` e `completo.y`, que apresentavam a data e hora de início e término de uma atividade, que também não se mostrou possível ser identificada. Ao final deste processo, foram removidos 84 atributos da base de dados.

Os dados também apresentavam registros duplicados. Para realizar a identificação destes registros, `cod` foi definido como atributo primário. A regra definida para escolher qual registro seria mantido, dentre aqueles que apresentavam um mesmo `cod`, é permanecer com aquele que apresenta a maior quantidade de atributos preenchidos. Outro filtro que foi aplicado na base de dados é sobre o atributo `idade`. Durante a investigação dos dados, foi verificado que apenas se este atributo estivesse preenchido, os atributos seguintes também estariam preenchidos (exceto quando omitidos explicitamente pelo participante do curso): `sexo`, `escolaridade`, `estadocivil`, `ocupacao`, `tempodeservico`, `religiao`, `contatoanterior`, `lidadiretamente`, `lida.onde`, `materialdidatico`, `prazoatividades`, `interacaopares`, `organizacaocurso`, `import.ajud.tutor`, `autoavaliacao.x`, `part.outrocurso`, `pp001`, `pp002`, `pp003`, `pp004`, `pp005`, `pp006`, `pp007`, `pp008`, `pp009`, `pp010`, `pp011`, `pp012`, `pp013`, `pp014`, `pp015`, `pp016`, `pp017`, `pp018`, `pp019`, `pp020`, `pp021`, `pp022`, `pp023`, `pp024`, `pp025`, `pp026`, `pp027`, `pp028`, `pp029`, `pp030`, `pp031`, `pp032`, `pp033`, `pp034`, `pp035`, `pp036`, `pp037`, `motivopart`, `barreiras` e `facilitadores`. Também foi verificado que aproximadamente 85% dos registros apresentavam o atributo `idade` ausente; e consequentemente os demais atributos citados acima. Logo, devido a esta quantidade de valores ausentes, todos os registros com o campo `idade` ausente (e, consequentemente os campos dos demais atributos do grupo acima citado também ausentes) foram excluídos da base de dados.

4.3 Padronização dos dados

Um problema encontrado na base de dados diz respeito ao padrão de entrada dos campos `idade` e `tempodeservico`, por parte dos participantes do curso. Nestes campos, os participantes tinham a liberdade de preenchê-lo com um texto. Esta característica inviabiliza a extração destes valores para a aplicação no modelo de classificação. Por exemplo,

o atributo `idade` apresenta para alguns registros os seguintes valores: “28”, “43 anos”, “DEZENOVE”, “trinta dois anos”, “23/09/1969”, “1977”. E o atributo `tempodeservico`: “2”, “1 ano e 6 meses”, “4 meses”, “2011”, “42 anos / 6 anos”, “15 escola e 3 meses Ong”, “mais de 31 anos”. Estes são apenas alguns casos e a base de dados está repleta de exemplos que dificultam a automatização da padronização destes dois atributos. Devido a esta natureza, a padronização dos dados foi feita manualmente. Foi também imposto um limite inferior, para tempo de serviço, equivalente a 1 ano. Logo, participantes que afirmaram ter experiência de trabalho inferior a 1 ano, tiveram estes valores ajustados para cima. Uma série de outras pequenas decisões também foram tomadas sobre estes dois atributos, para auxiliar o processo de padronização.

4.4 Transformação de atributos

Outra transformação aplicada na base de dados do curso, foi a conversão de atributos nominais em atributos numéricos. Desta maneira, foi possível adequar os dados ao padrão exigido pela biblioteca utilizada neste trabalho. Para realizar a transformação, colunas binárias foram adicionadas na base de dados, representando os potenciais valores dos atributos `sexo, escolaridade, estadocivil, ocupacao, religiao, contatoanterior, lidadiretamente, lida.onde, materialdidatico, prazoatividades, interacaopares, organizacaocurso, import.ajud.tutor, autoavaliacao.x, part.outrocurso, pp001, pp002, pp003, pp004, pp005, pp006, pp007, pp008, pp009, pp010, pp011, pp012, pp013, pp014, pp015, pp016, pp017, pp018, pp019, pp020, pp021, pp022, pp023, pp024, pp025, pp026, pp027, pp028, pp029, pp030, pp031, pp032, pp033, pp034, pp035, pp036, pp037, motivopart, barreiras` e `facilitadores`. Por exemplo, para transformar o atributo `sexo`, cujos potenciais valores são “Masculino” ou “Feminino”, duas colunas colunas com estes nomes foram adicionadas na tabela. Registros com o atributo `sexo` preenchido com o valor “Masculino” tiveram os seus novos atributos `Masculino` preenchido com “1” e `Feminino` com “0”.

4.5 Tratamento de valores ausentes

Para definir os valores ausentes do atributo `tempodeservico`, calculou-se a média deste atributo entre aqueles participantes que apresentavam a mesma `escolaridade` e uma diferença inferior a 5 (para mais ou para menos) para o atributo `idade`. Este cálculo é uma boa maneira de predizer o tempo de serviço de um participante, pois pessoas que exercem funções similares, possuem a mesma escolaridade e a mesma idade, tendem a possuir o mesmo tempo de serviço.

O atributo `sexo` apresentava apenas um valor ausente. Para preenchê-lo, foi consultado o nome do participante e definido manualmente. Uma vez que foram mantidos na base de dados apenas registros que não apresentavam valores ausentes para o atributo `idade`, estes dois atributos foram utilizados em conjunto com o atributo `tempodeservico` para descobrir quais eram os vizinhos mais próximos de cada registro (em relação a estes três atributos).

Para substituir os demais valores ausentes, foi-se utilizado o *k-Nearest Neighbors* (kNN). Ao encontrar um valor ausente, descobre-se quais são os valores deste atributo para os vizinhos do registro em questão. Então, para substituir o valor ausente, calcula-se a média aritmética dos vizinhos (caso se trate de um atributo numérico) ou calcula-se a moda (valor mais frequente) dos vizinhos (caso se trate de um atributo nominal). Foram testados valores entre 3 e 20 para a quantidade de vizinhos levados em consideração para a substituição, mas este número não apresentou diferença maior do que 1% na acurácia do modelo de classificação. Ao final dos testes, foi definido o valor 5 para a quantidade de vizinhos. Ao utilizar-se de algoritmos de mineração baseados em distância, como por exemplo o *kNN*, melhores resultados serão encontrados caso os dados fornecidos como entrada estejam normalizados. O processo de normalização consiste em transformar os dados para que possam ser representados em intervalos menores ou iguais, como por exemplo, [-1, 1] ou [0, 1]. (HAN; PEI; KAMBER, 2011). Logo, os atributos `idade` e `tempodeservico` foram normalizados entre o intervalo [0, 1].

Os atributos `assignment.view, course.view, feedback.view, folder.view, forum.add.post, forum.update.post, forum.user.report, forum.view.discussion, forum.view.forum, quiz.attempt, quiz.continue.attempt, quiz.view,`

`quiz.view.summary`, `resource.view`, `url.view`, `user.view`, `user.view.all`, `blog.view`, `forum.unsubscribe`, `user.update`, `discussion.mark.read`, `forum.add.discussion`, `forum.mark.read`, `forum.delete.post`, `forum.view.forums`, `quiz.review`, `forum.subscribe`, `forum.search`, `quiz.view.all` e `user.change.password` nunca apresentam o valor zero para seus registros e frequentemente apresentam valores ausentes. Devido a esta característica, e através da investigação sobre a base de dados, foi-se imputado que estes valores ausentes deveriam conter o valor zero. Por exemplo, a quantidade de vezes que uma pessoa alterou a sua senha. É esperado que ao menos um participante não tenha alterado a sua senha de `login` nenhuma vez, e esta característica não é refletida na base de dados. Portanto, conclui-se que os valores dos atributos acima citados, que deveriam ser originalmente preenchidos com zero, estão definidos na base de dados como apenas ausentes. Uma hipótese que justifica a ocorrência deste fenômeno, é a existência de algum equívoco durante a coleta dos dados a partir do AVA utilizado pelo curso.

5 Experimentos e resultados

Este capítulo contém o detalhamento dos experimentos realizados, a descrição dos resultados obtidos e a comparação com os resultados da literatura. Após realizado o pré-processamento dos dados, como descrito anteriormente, a base foi utilizada juntamente com os algoritmos de classificação da biblioteca *scikit-learn*.

Para a criação dos modelos, os atributos da base de dados foram classificados em três grupos:

Grupo 01 Todos os atributos da base de dados.

Grupo 02 Atributos relacionados à utilização da plataforma *web* de apoio ao curso, sendo eles: `assignment.view`, `course.view`, `feedback.view`, `folder.view`, `forum.add.post`, `forum.update.post`, `forum.user.report`, `forum.view.discussion`, `forum.view.forum`, `quiz.attempt`, `quiz.continue.attempt`, `quiz.view`, `quiz.view.summary`, `resource.view`, `url.view`, `user.view`, `user.view.all`, `blog.view`, `forum.unsubscribe`, `user.update`, `discussion.mark.read`, `forum.add.discussion`, `forum.mark.read`, `forum.delete.post`, `forum.view.forums`, `quiz.review`, `forum.subscribe`, `forum.search`, `quiz.view.all` e `user.change.password`.

Grupo 03 Atributos sociodemográficos coletados ao início do curso, sendo eles: `idade`, `sexo`, `escolaridade`, `estadocivil`, `ocupacao`, `tempodeservico`, `religiao`, `contatoanterior`, `lidadiretamente`, `lida.onde`, `part.outrocurso`, `pp001`, `pp002`, `pp003`, `pp004`, `pp005`, `pp006`, `pp007`, `pp008`, `pp009`, `pp010`, `pp011`, `pp012`, `pp013`, `pp014`, `pp015`, `pp016`, `pp017`, `pp018`, `pp019`, `pp020`, `pp021`, `pp022`, `pp023`, `pp024`, `pp025`, `pp026`, `pp027`, `pp028`, `pp029`, `pp030`, `pp031`, `pp032`, `pp033`, `pp034`, `pp035`, `pp036`, `pp037`, `motivopart`, `barreiras`, `facilitadores` e `aprovado`.

Para cada um destes três grupos, foram geradas árvores de decisão com profundidades máxima variando entre 1 a 20. Segundo Han, Pei e Kamber (2011), árvores com

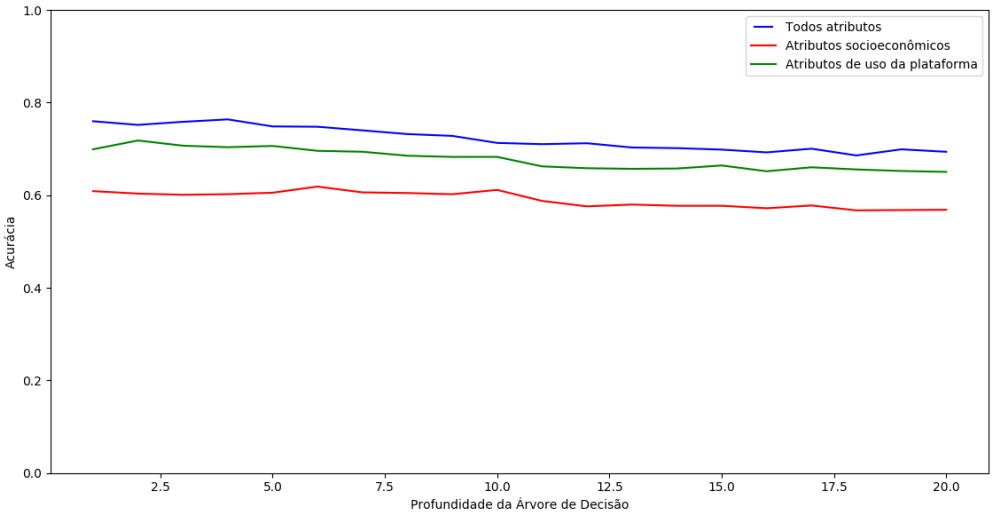

Figura 5.1: Profundidade x Acurácia

alturas menores tendem a classificar dados mais rapidamente e com maior acurácia. O gráfico da Figura 5.1 relaciona a acurácia encontrada pelas árvores de decisão, utilizando como dados de entrada cada um dos grupos acima citados, com a variação da profundidade máxima. É possível observar um decaimento da acurácia a medida em que se aumenta a profundidade da árvore, indo de acordo com a afirmação de Han, Pei e Kamber (2011). Para verificar a precisão dos modelos gerados, foi utilizada como ferramenta de análise a Validação Cruzada com 10 Dobras. Os modelos serão apresentados e discutidos nas seções seguintes.

Os resultados obtidos através dos dois primeiros grupos podem ser considerados triviais. Em suma, as árvores de decisão criadas a partir destes dois conjuntos de dados indicam que o participante que não fizer as atividades ou não acessar os conteúdos disponibilizados pela plataforma, tenderão a ser reprovados. Porém, o terceiro grupo de atributos apresenta informações não triviais. A árvore gerada por este conjunto de atributos revela correspondências menos previsíveis entre o perfil dos participantes e o seu desempenho ao final do curso.

5.1 Resultados obtidos

5.1.1 Todos atributos

Ao utilizar todos os atributos da base de dados para gerar a Árvore de Decisão, é possível observar que os atributos `quesm3, forum15` e `ativcolm3` foram os que mais influenciaram na classificação dos alunos. `quesm3` é um atributo que representa se um participante do curso realizou ou não uma das atividades propostas com este mesmo nome. Investigando a base de dados, supõe-se que esta seja a terceira atividade enviada para os alunos, de um total de vinte e oito. Também é possível perceber que a idade dos participantes influenciou na classificação. Aqueles com idade inferior, ou igual, a 31 anos tendem a ser aprovados e aqueles com idade superior a 31 anos tendem a ser reprovados.

A Figura 5.2 exibe a representação gráfica da Árvore de Decisão gerada, aplicando uma poda de profundidade máxima igual a quatro. A intensidade das cores dá uma ideia da probabilidade de aprovação ou reprovação de um aluno e a proximidade de um atributo com a raiz é proporcional à sua capacidade de separar os dados em conjuntos puros, segundo o índice Gini.

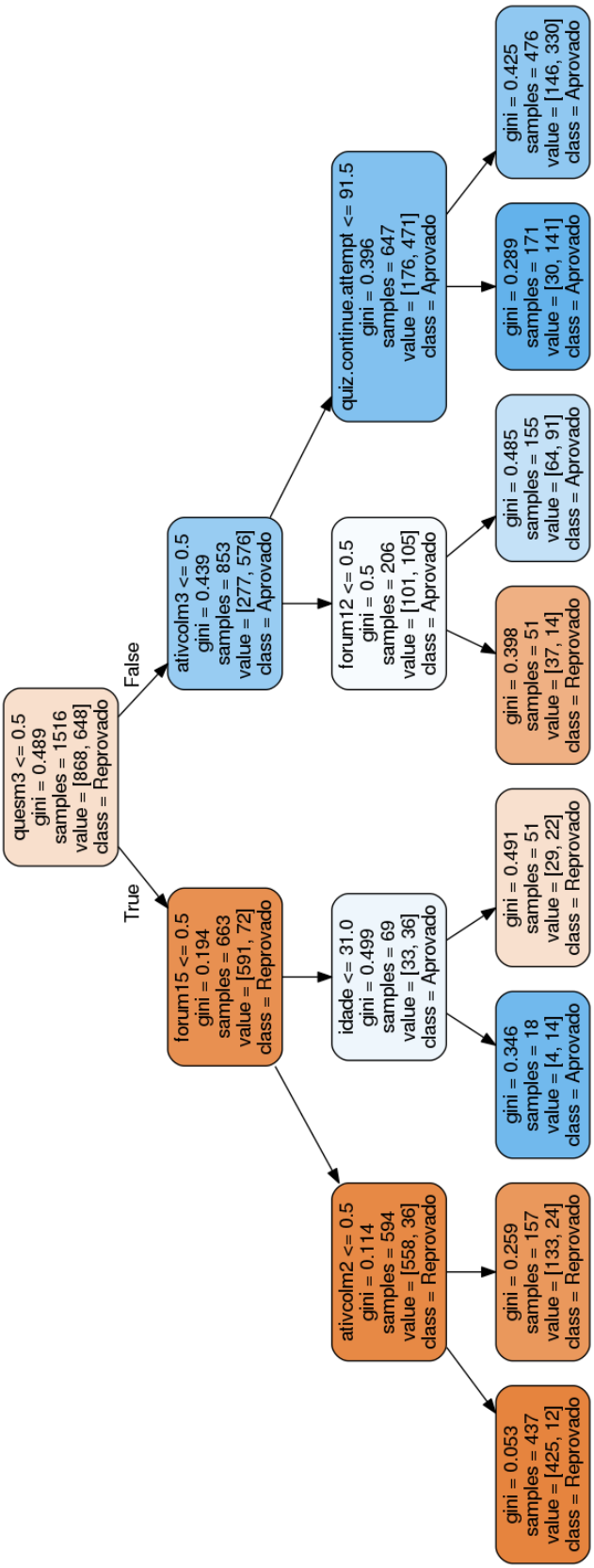

Figura 5.2: Árvore de Decisão utilizando todos atributos da base de dados

A Figura 5.3 apresenta uma matriz de confusão normalizada para esta árvore de decisão. Através desta matriz, é possível avaliar quão bem o modelo foi capaz de classificar os registros da base de dados, através da árvore de decisão com profundidade máxima igual a quatro. Segundo Han, Pei e Kamber (2011), a acurácia deste modelo é dada pela soma dos casos em que o modelo acertou dividido pelos demais casos, como mostra a equação 5.1.

$$\frac{0.78 \times 868 + 0.84 \times 648}{1516} = 0.81 \tag{5.1}$$

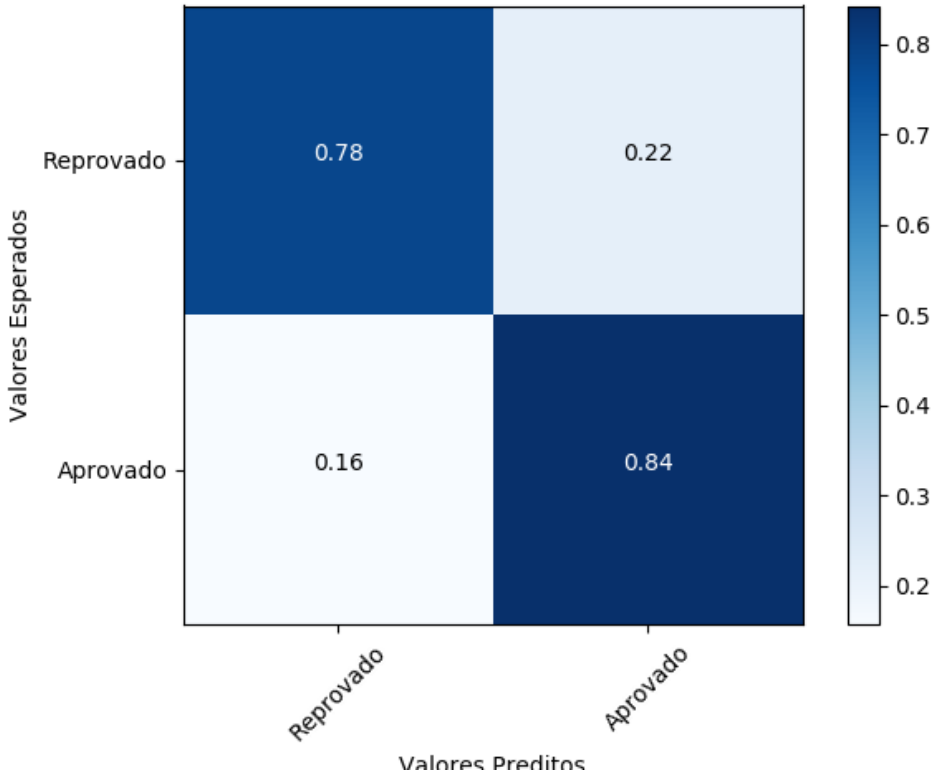

Figura 5.3: Matriz de confusão para a árvore de altura quatro gerada com todos os atributos

5.1.2 Atributos de utilização do AVA

A fim de gerar uma nova Árvore de Decisão, foram selecionados apenas os atributos referentes aos padrões de utilização da plataforma *web*. Neste cenário, os atributos mais relevantes para o desempenho final dos alunos foram `quiz.attempt`, `course.view` e `forum.add.post`. Participantes que apresentaram poucas tentativas de responder os questionários e que visualizaram poucas vezes a página do curso, tendem a ser reprovados.

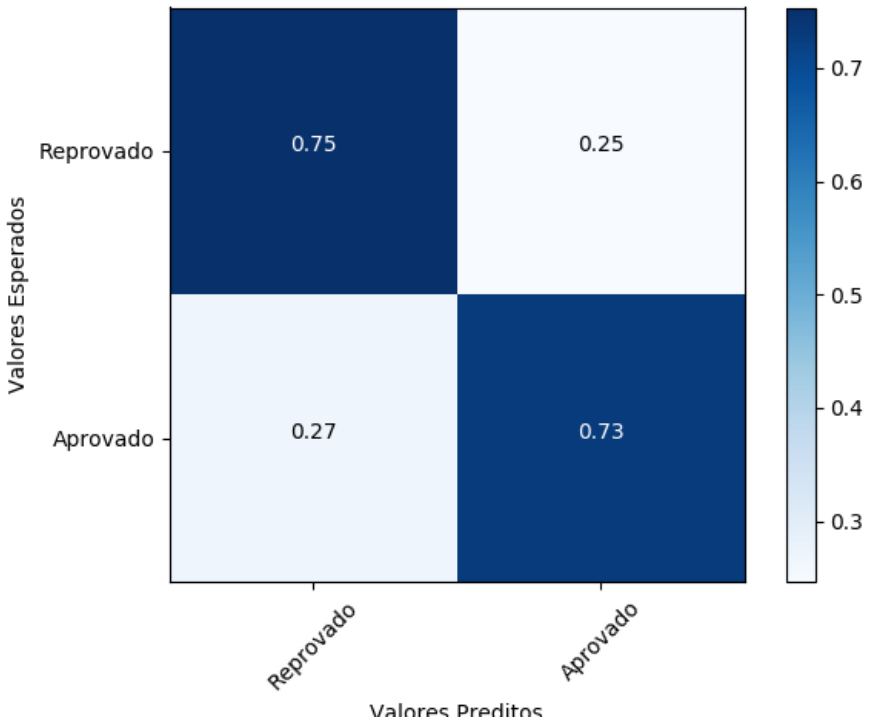

Figura 5.4: Matriz de confusão para a árvore de altura quatro gerada com os atributos do AVA

Outro fator associado às reprovações é a visualização de tarefas, pois participantes que não visualizam as tarefas também não as cumprem. Por outro lado, participantes que criaram tópicos no fórum tendem a ser aprovados. Pode-se supor que estes participantes criaram tópicos para tirarem dúvidas com instrutores ou interagir com colegas de curso.

A Figura 5.5 exibe a representação gráfica da Árvore de Decisão gerada e a Figura 5.4 exibe a sua matriz de confusão, cuja acurácia é dada pela equação 5.2:

$$\frac{0.75 \times 868 + 0.73 \times 648}{1516} = 0.74 \tag{5.2}$$

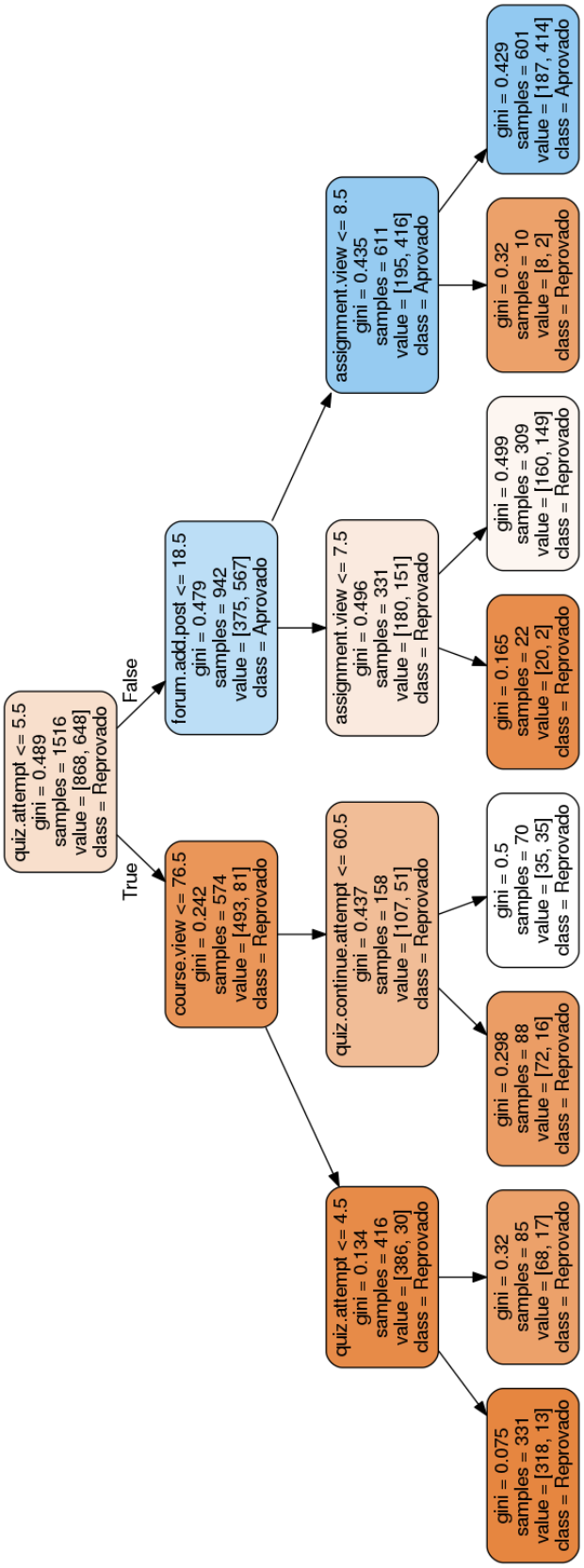

Figura 5.5: Árvore de Decisão gerada apenas com atributos de utilização do AVA

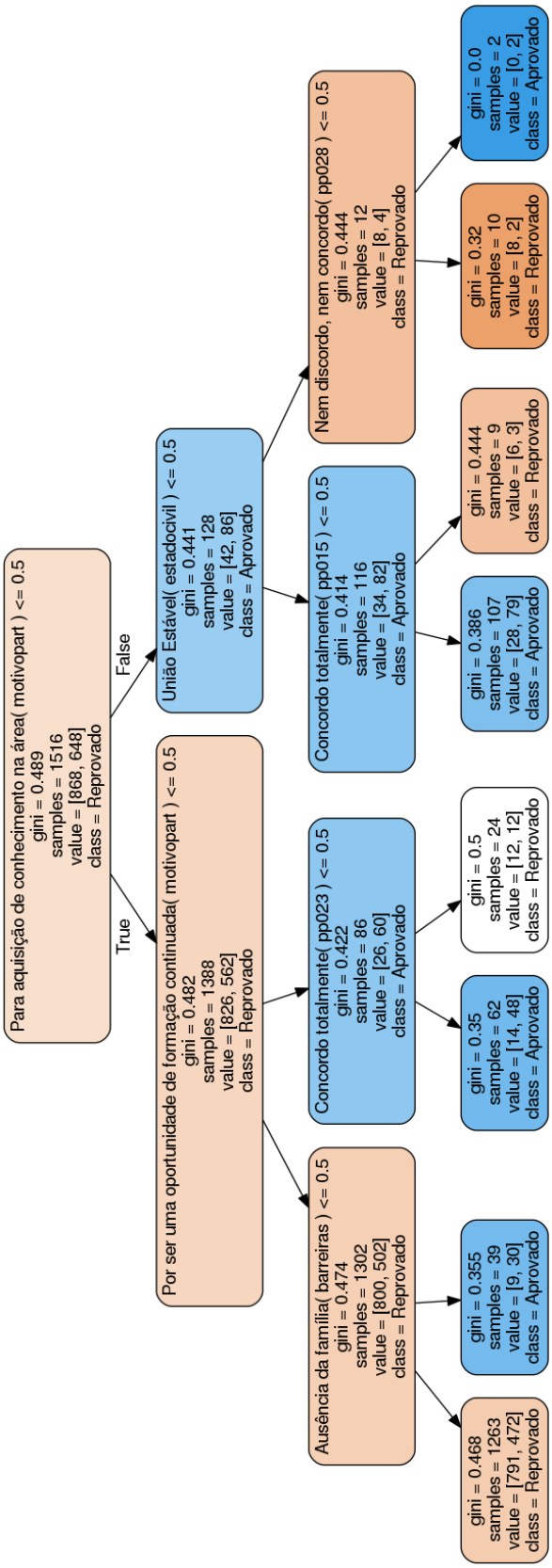

Figura 5.6: Árvore de Decisão gerada apenas com os atributos sociodemográficos

5.1.3 Atributos sociodemográficos

A Figura 5.6 exibe a representação gráfica da Árvore de Decisão gerada para o terceiro grupo de atributos, os atributos sociodemográficos. Este conjunto de atributos revela informações que possibilitam a aplicação de medidas interventivas logo ao início do curso, porém, infelizmente, é o conjunto de atributos que apresentou a acurácia mais baixa dentre os três grupos sugeridos.

Dentre os resultados obtidos, é possível observar que participantes que tiveram como motivação de se inscrever no curso "a aquisição de conhecimento na área", não estão em uma "união estável" e não concordam totalmente com o atributo `pp015` (As principais causas do uso de drogas são a falta de disciplina e autocontrole) tendem a ser aprovados no curso. Em contrapartida, aqueles participantes que não participaram do curso "para adquirir conhecimento na área" ou "por ser uma oportunidade de formação continuada" e também não acreditam que a "ausência da família" dificulta o tratamento de usuários de álcool ou drogas, apresentam as maiores taxas de reprovação no curso.

A Figura 5.7 exibe a matriz de confusão desta árvore, cuja acurácia é dada por:

$$\frac{0.93 \times 868 + 0.31 \times 648}{1516} = 0.66 \tag{5.3}$$

5.2 Análise dos resultados

Os resultados obtidos neste trabalho apresentam acurácia inferior às obtidas nos trabalhos levantados na revisão bibliográfica. Conforme relata Han, Pei e Kamber (2011), a qualidade dos resultados obtidos pelos modelos de classificação está diretamente relacionado à qualidades dos dados utilizados e ao tratamento aplicado a eles. É seguro dizer que este trabalho se beneficiaria com um estudo mais aprofundado para o pré-processamento dos dados. Ainda assim, os atributos que mais influenciaram no desempenho final dos alunos do curso puderam ser identificados, viabilizando assim a tomada de medidas interventivas em suas edições futuras.

Dentre os atributos levantados como mais influenciadores no desempenho final,

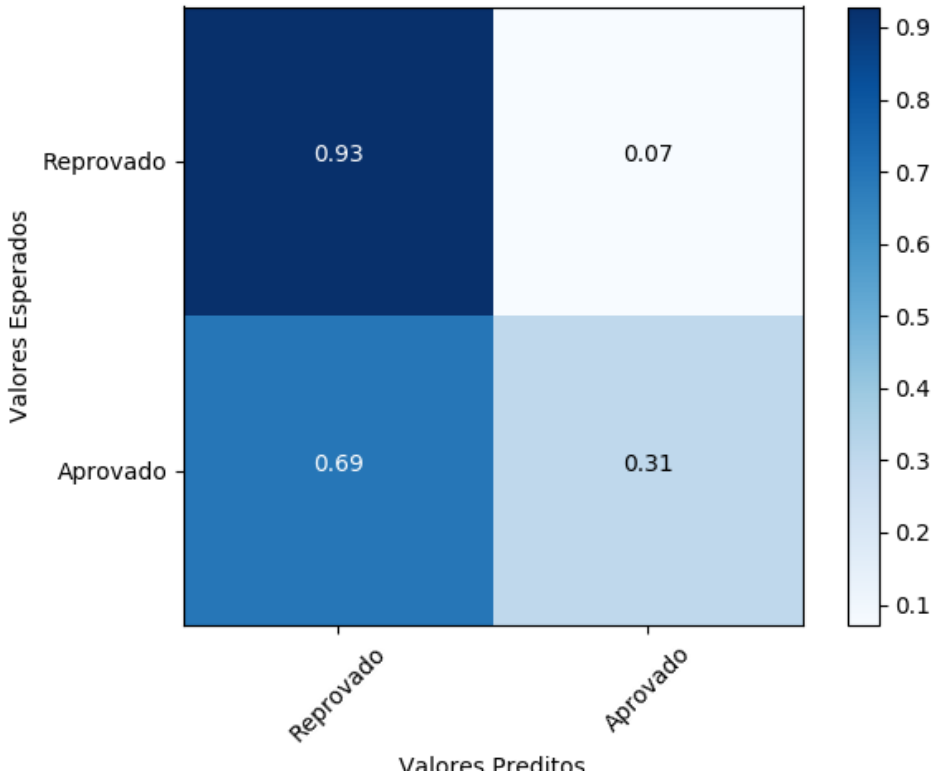

Figura 5.7: Matriz de confusão para a árvore de altura quatro gerada com os atributos sociodemográficos

estão a quantidade de tentativas de realização das atividades e a frequência de acesso das páginas da plataforma. Minaei-Bidgoli et al. (2003) encontrou em seu trabalho que a quantidade de atividades corretas e a quantidade de tentativas até obter um acerto são os principais indicadores de aprovação e Dias et al. (2008) encontrou que os fatores mais relevantes para a reprovação são poucos acessos na plataforma e pouca quantidade de atividades realizadas. Os resultados destes dois autores vão de encontro com os resultados obtidos neste trabalho. Por sua vez, Pereira e Zambrano (2017) encontrou atributos sociodemográficos como principais influenciadores, porém os atributos que ele utilizou não existem para a base de dados do curso analisado neste trabalho.

6 Considerações Finais

Neste trabalho foram propostos conjuntos de atributos da base de dados coletada do "Curso de Prevenção do Uso de Drogas para Educadores de Escolas Públicas" para a criação de modelos de classificação baseados em árvores de decisão. Estes modelos tem como objetivo revelar o perfil de estudantes com potencial de reprovação neste curso e identificar os principais fatores que influenciam em seu desempenho ao final do curso.

6.1 Conclusões

Apesar do modelo de classificação obtido utilizando todos os atributos da base de dados possuir a maior acurácia dentre os três grupos de atributos propostos, ele não apresenta aplicabilidade para identificar o perfil de um participante propenso a reprovar. Isto ocorre pois, dizer que um estudante reprovou no curso por não realizar atividades é um conhecimento trivial. Contudo, este modelo é capaz de auxiliar na identificação de atividades que impactaram diretamente no desempenho de um participante. Por exemplo, é possível que estudantes que não realizaram uma atividade em específico, apresentem alto índice de aprovação, enquanto estudantes que a realizaram apresentem alto índice de reprovação. Esta hipótese, embora não tenho sido observada neste trabalho, é um conhecimento não trivial capaz de ser obtido através das árvores de decisões geradas e poderia auxiliar os produtores de conteúdo do curso.

Por outro lado, os resultados obtidos utilizando o terceiro grupo de atributos (os sociodemográficos) são aplicáveis logo ao início do curso. Identificar em um estágio inicial estudantes propensos a reprovar pode aumentar as taxas de aprovação do curso, pois haverá tempo hábil para a aplicação de medidas pedagógicas interventivas sobre tais estudantes com potencial de reprovação.

O segundo grupo de atributos também revela conhecimento aplicável, como por exemplo, alunos que apresentam baixas visualizações de atividades tendem a ser reprovados. Este grupo apresenta aplicabilidade quanto a identificação de estudantes propensos

a reprovar. O modelo utilizando este grupo de atributos pode ser gerado em um momento anterior ao final do curso, ainda em tempo hábil para a aplicação de medidas interventivas.

6.2 Trabalhos futuros

Para trabalhos futuros, é sugerido realizar outros tratamentos na base de dados, aplicar outros algoritmos de classificação e utilizar novos subconjuntos de dados para gerar os modelos, de forma a melhorar a descoberta de informações não triviais.

Os resultados obtidos neste trabalho podem auxiliar os produtores de conteúdo do curso estudado a reformular algumas atividades enviadas para os participantes e também identificar aqueles alunos que tem potencial de reprovação, possibilitando a adoção de medidas interventivas.

A Atributos Mantidos da Base de Dados

Tabela A.1: Atributos mantidos

Atributo	Tipo	Potenciais Valores	Descricao
cod	Inteiro	Código aleatório com 10 ou 11 digitos	Identificador primario
quesm1 – quesm3	Inteiro	{0, 1}	Atividades
quesm1r – quesm2r	Inteiro	{0, 1}	Atividades
forum1 – forum15	Inteiro	{0, 1}	Atividades
ativcolm1 – ativcolm4	Inteiro	{0, 1}	Atividades
ativcolm1r – ativcolm2r	Inteiro	{0, 1}	Atividades
forum1r – forum2r	Inteiro	{0, 1}	Atividades
idade	Inteiro	[19, 70]	
sexo	Nominal	Masculino, Feminino	
escolaridade	Nominal	Pós-graduação Ensino Superior Completo Ensino Médio Completo Ensino Superior Incompleto Ensino Fundamental Incompleto Ensino Fundamental Completo	
estadocivil	Nominal	Casado (a) União Estável Divorciado (a) Solteiro (a) Viúvo (a) Outros	
ocupacao	Nominal	Professor (a) Diretor (a) Supervisor (a) Outros Orientador (a) Coordenador (a) Pedagógico Estudante	
tempodeservico	Inteiro	[1, 48]	

Tabela A.2: Atributos mantidos

Atributo	Tipo	Potenciais Valores	Descrição
religiao	Nominal	Evangélica Católica Espírita Sem religião Outras Budismo Candomblé Umbanda	
contatoanterior	Nominal	Não Sim	Com a temática álcool e drogas, antes do curso
lidadiretamente	Nominal	Sim Não	Conhece usuários de álcool ou drogas
lida.onde	Nominal	Outros Família Escola Comunidade Amigos Serviços de saúde Serviços de atuação Não Lida	
materialdidatico	Nominal	Adequado Muito adequado Pouquíssimo adequado Pouco adequado	
prazoatividades	Nominal	Flexível Muito flexível Pouco flexível Pouquíssimo flexível	
interacaopares	Nominal	Muito importante Importante Pouco importante Pouquíssimo importante	Interações com colegas ou instrutores
organizacaocurso	Nominal	Muito organizado Organizado Desorganizado Muito desorganizado	
import.ajud.tutor	Nominal	Sempre Às vezes Raramente Nunca	
autoavaliacao.x	Nominal	Sim, considero Sim, considero, porém, poderia estar me esforçando mais Não, não considero	

Tabela A.3: Atributos mantidos

Atributos Mantidos			
Atributo	Tipo	Potenciais Valores	Descrição
part.outrocurso	Nominal	Não Sim	Participa ou participou de outros cursos com a mesma temática
assignment.view	Inteiro	[0.0, 212.0]	Visualização de tarefas
course.view	Inteiro	[0.0, 1061.0]	Visualização da página inicial do curso
feedback.view	Inteiro	[0.0, 39.0]	Visualização de feedback das atividades feitas
folder.view	Inteiro	[0.0, 59.0]	Visualização de materiais do curso
forum.add.post	Inteiro	[0.0, 85.0]	Tópicos adicionados ao fórum.
forum.update.post	Inteiro	[0.0, 86.0]	Atualização de tópicos adicionados.
forum.user.report	Inteiro	[0.0, 77.0]	Denúncia de outros usuários
forum.view.discussion	Inteiro	[0.0, 999.0]	Visualização de discussões nos fórums
forum.view.forum	Inteiro	[0.0, 599.0]	Visualização da página inicial do fórum
quiz.attempt	Inteiro	[0.0, 28.0]	Tentativas de responder um questionário
quiz.continue.attempt	Inteiro	[0.0, 348.0]	Tentativas até obter uma respota certa
quiz.view	Inteiro	[0.0, 160.0]	Visualização dos questionários
quiz.view.summary	Inteiro	[0.0, 41.0]	Visualização dos resumos dos questionários
resource.view	Inteiro	[0.0, 304.0]	Visualização dos recursos que o AVA oferecia
url.view	Inteiro	[0.0, 324.0]	Visualização de links postados nos fórums
user.view	Inteiro	[0.0, 238.0]	Visualização de perfil de usuários
user.view.all	Inteiro	[0.0, 296.0]	Visualização da lista de perifl de todos usuários
blog.view	Inteiro	[0.0, 10.0]	Visualização do blog do curso
forum.unsubscribe	Inteiro	[0.0, 17.0]	Parou de seguir um fórum
user.update	Inteiro	[0.0, 80.0]	Atualizou informações de usuário

Tabela A.4: Atributos mantidos

Atributos Mantidos			
Atributo	Tipo	Potenciais Valores	Descrição
discussion.mark.read	Inteiro	[0.0, 52.0]	Discussões marcadas como lidas
forum.add.discussion	Inteiro	[0.0, 2.0]	Discussões adicionadas ao fórum
forum.mark.read	Inteiro	[0.0, 5.0]	Fórums marcados como lidos
forum.delete.post	Inteiro	[0.0, 17.0]	Tópicos deletados em fórums
forum.view.forums	Inteiro	[0.0, 48.0]	Visualização de uma página com todos fórums
quiz.review	Inteiro	[0.0, 105.0]	Revisões de um questionário
forum.subscribe	Inteiro	[0.0, 19.0]	Inscrições em um fórum
forum.search	Inteiro	[0.0, 35.0]	Pesquisas no fórum
quiz.view.all	Inteiro	[0.0, 6.0]	Visualização de uma página com todos questionários
user.change.password	Inteiro	[0.0, 2.0]	Quantidade de trocas de senha
motivopart	Nominal	“Identificação pessoal com o tema” “Identificação profissional com o tema” “Para aquisição de conhecimento na área” “Pelo fato de o curso ser gratuito” “Pelo fato de o curso estar vinculado à Universidade” “Por ser um curso à distância” “Por ser uma oportunidade de formação continuada”	Motivo que os participantes afirmaram para participar do curso

Tabela A.5: Atributos mantidos

Atributos Mantidos			
Atributo	Tipo	Potenciais Valores	Descrição
barreiras	Nominal	"Ausência da familia" "Pouca comunicação com os pais" "Uso de substâncias por familiares" "Presença de drogas ilícitas no ambiente escolar" "Proximidade da rede de distribuição de drogas" "Ausência de limites dos alunos" "Ausência de colaboração da equipe escolar" "Ausência de regras"	Barreiras que os participantes afirmaram existir que dificultam a parada do consumo de álcool e drogas
facilitadores	Nominal	"Possuir alunos interessados na temática" "Presença de uma equipe para trabalhar a temática" "Estímulo aos alunos" "Desenvolvimento de projetos na escola" "Apoio aos projetos em desenvolvimento" "Presença de regras no ambiente escolar Promoção de compromisso e confiança" "Valorização do ambiente escolar" "Participação da comunidade e dos pais no trabalho de prevenção"	Facilitadores que os participantes afirmaram existir para ajudar na conclusão e aplicação do conhecimento obtido no curso

B Atributos Excluídos da Base de Dados

Tabela B.1: Atributos excluídos

Atributos Excluídos	
Atributo	Motivo
nomecompleto	Relevância
email	Relevância
sobrenome.x	Relevância
nome.x	Relevância
instituicao	Quantidade
depto	Quantidade
e.mail.x	Relevância
estado.x	Erro
iniciado.x	Ilegível
completo.x	Erro
tempo.x	Erro
nome.completo	Relevância
rg	Relevância
e.mail.1	Relevância
termo.x	Ilegível
estadocivil.outros	Ilegível
formacao	Ilegível
ocupacao.outra	Ilegível
nomeservico	Ilegível
religiao.outra	Ilegível
lida.outros	Ilegível
motivo.outros	Ilegível
X	Ilegível
assignment.upload	Erro
assignment.view.all	Redundância
course.recent	Ilegível
feedback.startcomplete	Ilegível
feedback.submit	Ilegível
quiz.close.attempt	Redundância
assignment.view.submission	Redundância
notes.view	Quantidade
quiz.report	Relevância
assignment.update.grades	Quantidade
quiz.preview	Quantidade

Tabela B.2: Atributos excluídos

Atributos Excluídos	
Atributo	Motivo
course.report.participation	Quantidade
feedback.view.all	Quantidade
course.report.outline	Quantidade
forum.start.tracking	Quantidade
forum.stop.tracking	Quantidade
grade.update	Quantidade
assignment.update	Quantidade
calendar.add	Quantidade
calendar.edit	Quantidade
course.add.mod	Quantidade
course.update.mod	Quantidade
forum.delete.discussion	Quantidade
quiz.add	Quantidade
quiz.editquestions	Quantidade
course.update	Quantidade
forum.subscribeall	Quantidade
course.report.log	Quantidade
notes.add	Quantidade
forum.view.subscribers	Quantidade
forum.update	Quantidade
forum.move.discussion	Quantidade
resource.update	Quantidade
quiz.update	Quantidade
course.delete.mod	Quantidade
label.update	Quantidade
quiz.manualgrade	Quantidade
sumTime	Erro
nsessions	Quantidade
turma	Quantidade
sobrenome.y	Relevância
nome.y	Relevância
id	Relevância
insti	Relevância
dpto	Relevância
e.mail.y	Relevância
estado.y	Relevância
iniciado.y	Erro

Tabela B.3: Atributos excluídos

Atributos Excluídos	
Atributo	Motivo
completo.y	Erro
tempo.y	Erro
nota	Quantidade
termo.y	Relevância
motivopartOutros	Ilegível
material	Redundante
flexprazo	Redundante
interacaocol	Redundante
organizado	Redundante
freqauxitutor	Redundante
autoavaliacao.y	Redundante
partoutrocurso	Redundante

C Questionário Social

Todas as perguntas desta tabela foram respondidas com uma destas opções:

- Discordo totalmente
- Discordo
- Nem discordo, nem concordo
- Concordo
- Concordo totalmente

Tabela C.1: Questionário social

Questionário social	
Atributo	Motivo
pp001	Usuários de drogas não tem forca de vontade.
pp002	Usuários de drogas tem menor destaque na sociedade.
pp003	Usuários de drogas não podem ocupar cargos que exigem maior responsabilidade.
pp004	O uso de drogas representa uma fraqueza de carater.
pp005	Usuários de drogas não se preocupam com si mesmos.
pp006	Usuários de drogas são pessoas moralmente fracas.
pp007	Usuários de drogas são pessoas sem determinação.
pp008	Usuários de drogas não querem parar de usa-las.
pp009	Usuários de drogas raramente prejudicam alguém a não ser a si próprios.
pp010	A maioria dos usuários de drogas estão desempregada.
pp011	O tratamento raramente ajuda o usuário de drogas.
pp012	Nao se deve ter grandes expectativas na relação com os usuários de drogas
pp013	Quem abusa de drogas pode aprender a diminuir o uso, tendo-o sob controle novamente.
pp014	Usuários de drogas podem ser ajudados antes de chegarem ao fundo do poco.
pp015	As principais causas do uso de drogas são a falta de disciplina e autocontrole.
pp016	A melhor forma de controlar os usuários de drogas é mantê-los isolados.

Tabela C.2: Questionário social

Questionário social	
Atributo	Descrição
pp017	Existem caracteristicas que diferenciam os usuários de drogas das pessoas normais.
pp018	Uma pessoa deve ser hospitalizada assim que apresentar sinais de uso de drogas.
pp019	A dependência de drogas é uma doença.
pp020	Os usuários de drogas são pessoas indesejáveis na sociedade.
pp021	A sociedade não deveria se preocupar em proteger-se dos usuários de drogas.
pp022	Os usuários de drogas são responsáveis pelos problemas associados ao uso de drogas.
pp023	Os usuários de drogas devem ser isolados da sociedade.
pp024	Uma pessoa seria ingênua em se casar com alguém que tenha sido usuário de drogas, mesmo que estivesse recuperado.
pp025	As pessoas não gostariam de morar próximo a alguém que tenha sido usuário de drogas.
pp026	Alguem que tenha um histórico de uso de drogas deve ser impedido de assumir qualquer cargo público.
pp027	Os usuários de drogas devem ser privados de seus direitos individuais.
pp028	Usuários de drogas devem ser encorajados a assumir sua responsabilidade por suas atividades diárias.
pp029	Ninguém tem o direito de excluir os usuários de drogas de sua vizinhanca.
pp030	Os usuários de drogas oferecem mais perigo do que as pessoas imaginam.
pp031	Os usuários de drogas ainda são ridicularizados.
pp032	Devem ser gastos mais recursos públicos financeiros no tratamento dos usuários de drogas.
pp033	A sociedade precisa ser mais tolerante com os usuários de drogas.
pp034	A sociedade tem a responsabilidade de fornecer o melhor tratamento possível aos usuários de drogas.
pp035	Os usuários de drogas merecem nossa simpatia.
pp036	Os usuários de drogas são um peso para a sociedade.
pp037	Aumentar o investimento nas politicas de drogas é um desperdício de dinheiro publico.
pp038	O número de serviços de tratamento é suficiente para o número de usuários de drogas.
pp039	É melhor evitar alguém que tenha problemas com drogas.

Bibliografia

ALPAYDIN, E. *Introduction to Machine Learning. [Sl]*. [S.l.]: The MIT Press, 2010.

DIAS, M. M. et al. Aplicação de técnicas de mineração de dados no processo de aprendizagem na educação a distância. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2008. v. 1, n. 1, p. 105–114.

DIMENSIONLESS. Why are tree-based models robust to outliers? 2016. (Acessado em: 28-Junho-2018). Disponível em: ⟨https://www.quora.com/Why-are-tree-based-models-robust-to-outliers⟩.

FUCHS, K. Machine learning: Classification models. 2017. (Acessado em: 30-Junho-2018). Disponível em: ⟨https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529⟩.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011.

JOHN, G. H. Robust decision trees: Removing outliers from databases. In: *KDD*. [S.l.: s.n.], 1995. p. 174–179.

MINAEI-BIDGOLI, B. et al. Predicting student performance: an application of data mining methods with an educational web-based system. In: IEEE. *Frontiers in education, 2003. FIE 2003 33rd annual*. [S.l.], 2003. v. 1, p. T2A–13.

MONTEIRO, É. P. et al. Curso de prevenção ao uso de drogas: Descrição e avaliação de satisfação. *Estudos de Psicologia (Natal)*, SciELO Brasil, v. 21, n. 3, p. 328–336, 2016.

NETTO, C.; GUIDOTTI, V.; SANTOS, P. K. D. A evasão na ead: investigando causas, propondo estratégias. In: *Congresos CLABES*. [S.l.: s.n.], 2017.

NORTON, M. J. Knowledge discovery in databases. *Library Trends*, v. 48, n. 1, p. 9–21, 1999.

PEREIRA, R. T.; ZAMBRANO, J. C. Application of decision trees for detection of student dropout profiles. In: IEEE. *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. [S.l.], 2017. p. 528–531.

RUSSELL, S.; NORVIG, P. Intelligent agents. *Artificial intelligence: A modern approach*, v. 74, p. 46–47, 1995.

VIDAL, E. Ensino a distancia vs ensino tradicional. *Universidade Fernando Pessoa, Porto*, 2002.