UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Jhillian Bianchi

PREVISÃO DE EVASÃO EM CURSOS DE ENSINO SUPERIOR ATRAVÉS DE APRENDIZADO DE MÁQUINA ASSOCIADO À ANÁLISE DE DISCIPLINAS APROVADAS

Santa Maria, RS
2017

Jhillian Bianchi

PREVISÃO DE EVASÃO EM CURSOS DE ENSINO SUPERIOR ATRAVÉS DE APRENDIZADO DE MÁQUINA ASSOCIADO À ANÁLISE DE DISCIPLINAS APROVADAS

Trabalho de Conclusão de Curso apresentado ao Bacharelado em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do grau de **Bacharel em Ciência da Computação**

Orientador: Prof. Dr. Sergio Luis Sardi Mergen (UFSM)

442
Santa Maria, RS
2017

Jhillian Bianchi

PREVISÃO DE EVASÃO EM CURSOS DE ENSINO SUPERIOR ATRAVÉS DE APRENDIZADO DE MÁQUINA ASSOCIADO À ANÁLISE DE DISCIPLINAS APROVADAS

Trabalho de Conclusão de Curso apresentado ao Bacharelado em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do grau de **Bacharel em Ciência da Computação**

Aprovado em 14 de dezembro de 2017:

Sergio Luis Sardi Mergen, Dr. (UFSM)
(Presidente/Orientador)

Márcia Pasin, (Dr. UFSM)

Ana Trindade Winck, (Dr. UFCSPA)

Santa Maria, RS
2017

DEDICATÓRIA

Dedico este trabalho aos meus pais, minha irmã e toda a minha família que sempre me apoiaram, me socorreram nas horas de angústia, não permitiram que eu desistisse e nunca mediram esforços para que eu chegasse no lugar que estou hoje.

AGRADECIMENTOS

Agradeço ao professor Dr. Sergio Luis Sardi Mergen, por toda a orientação e tempo despendido em meu auxílio, principalmente com relação a este trabalho de graduação.

À professora Dr. Márcia Pasin e à professora Dr. Ana Trindade Winck por aceitarem meu convite e constituírem parte da banca avaliadora, prestigiando este trabalho.

Aos meus amigos, que demonstraram muita paciência e me auxiliaram muito, principalmente nos momentos de ansiedade, que sempre foram compreensivos e sempre encontravam uma forma de me fazer sorrir, me dando forças para continuar, meu sincero muito obrigado.

A minha prima Vanessa pelo grande apoio prestado desde o dia em que entrei nesta universidade, seja no âmbito emocional, seja na própria revisão deste trabalho.

Agradeço à equipe do Centro de Processamento de Dados da UFSM (CPD) por ter disponibilizado os dados da instituição, o que viabilizou o desenvolvimento deste trabalho.

Aos meus pais e irmã, por sempre me apoiarem em todos os momentos, inclusive com os puxões de orelha, que são imprescindíveis.

Por fim, ao Grêmio de FootBall Porto Alegrense, o imortal tricolor, rei de copas e dono da América, por uma das minhas maiores alegrias como torcedor, o tri-campeonato da Copa Libertadores da América, que me motivou ainda mais nessa reta final de graduação.

"Não vale a pena mergulhar nos nossos sonhos e esquecer de viver."

(ALVO DUMBLEDORE - HARRY POTTER E A PEDRA FILOSOFAL - J.K.ROWLING)

RESUMO

PREVISÃO DE EVASÃO EM CURSOS DE ENSINO SUPERIOR ATRAVÉS DE APRENDIZADO DE MÁQUINA ASSOCIADO À ANÁLISE DE DISCIPLINAS APROVADAS

AUTOR: JHILLIAN BIANCHI
ORIENTADOR: SERGIO LUIS SARDI MERGEN (UFSM)

O problema da evasão escolar no âmbito do ensino superior atingiu valores alarmantes. Segundo estudos de DAVOK; BERNARD (2016), cerca de 50% dos alunos da área de ciências exatas e da terra não concluem seus respectivos cursos. Para buscar uma solução viável para tal problema, o uso de algoritmos de aprendizado de máquina têm sido cada vez mais frequente, a fim de buscar padrões e identificar alunos com potencial de evasão. Porém, encontrar uma combinação de atributos adequada para treinamento não é tarefa tão simples. A grande expansão do poder computacional possibilitou o desenvolvimento de técnicas de mineração de dados mais eficientes, além do aumento de informações processadas e armazenadas atualmente. Assim, a finalidade principal deste trabalho é a partir de dados referentes a disciplinas aprovadas do aluno, que são dados de cunho não sensível, identificar um conjunto de atributos que, por meio de um classificador, alcance taxas expressivas de acerto acerca de indivíduos com potencial de evasão. Uma alta taxa de acerto na identificação de alunos com esse potencial é primordial para que se tomem medidas eficazes no que diz respeito ao controle da evasão, podendo ser uma destas medidas o acompanhamento pedagógico do aluno em questão.

Palavras-chave: Evasão. Classificação. Aprendizado de máquina. Mineração de dados.

ABSTRACT

ABSTRACT TITLE

AUTHOR: JHILLIAN BIANCHI
ADVISOR: SERGIO LUIS SARDI MERGEN (UFSM)

The scholar dropout on higher education context reaches dangerous levels. As studies DAVOK; BERNARD (2016), about 50% of exact science students don't finish their respective courses. To help finding a useful solution, machine learning algorithms has been used more frequently in order to search for education patterns and to identify students with dropout potential. However, finding a attribute combination which is efficient for the training step is not a simple assignment. Moreover, the big rise of system computational powerful made possible the development of better data mining techniques, and from the increasingly amount of generated and processed data. Thereby, this work's main purpose is to make an attribute combination based on information about approved disciplines from a student, with a classifier technique, to achieve expressive accuracy levels on students dropout potential. A good dropout student hit level is primordial to improve the searches for effective measures on dropout control, so it can be used for the student pedagogical monitoring.

Keywords: Dropout. Classification. Machine learning. Data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CRISP-DM	*CRoss Industry Standard Process for Data Mining*
WEKA	*Waikato Environment for Knowledge Analysis*
SIE	Sistema de Informações para o Ensino
UFSM	Universidade Federal de Santa Maria
KDD	*Knowledge Discovery in Databases*
CPD	Centro de Processamento de Dados
IES	Instituição de Ensino Superior
LMT	*Logistic Model Tree*

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O volume elevado de abandono no ensino superior tem sido um dos principais motivos de preocupação no âmbito da educação brasileira. As taxas de evasão atualmente contam com números muito expressivos, e, de forma genérica, independe de área do conhecimento ou localização geográfica. Segundo DAVOK; BERNARD (2016), entre 2008 e 2010, estudos concluíram que 40% dos alunos da área de Engenharia e mais de 50% dos estudantes da área de Ciências Exatas e da Terra não conseguem concluir seus respectivos cursos.

A evasão em diversas áreas é imensamente prejudicial para discentes. Muitos alunos encontram dificuldades de adaptação, seja ao ambiente ou às disciplinas (MORAIS et al., 2017). Um combinado destes fatores faz com que as experiências acadêmicas do aluno tornem-se mais complexas, dificultando dessa forma a formação considerada ideal e inspirando à desistência. A consequência de desistências prejudica, além de docentes, as próprias instituições, que têm uma queda na regularidade e qualidade de ensino.

Sendo assim, o desenvolvimento de técnicas que possibilitem a otimização das experiências dos alunos enquanto da sua permanência no ensino superior é de extrema importância, para que o discente sinta-se mais confortável e incluso, de forma natural. Partindo destas características, os níveis de evasão podem ser mitigados, resultando em uma melhora na qualidade do ensino e do profissional formado, buscando sempre níveis de excelência.

1.2 PROBLEMA

Os primeiros estudos mais elaborados sobre a evasão, realizados no período de 2000 a 2005, já retratavam um quadro preocupante. Segundo os estudos, além dos níveis de evasão alterados que já chegavam a uma média de 22% nas Instituições de Ensino Superior (IES), o combate à evasão era falho ou inexistente. Nas palavras do autor,

> [...] são raríssimas as IES brasileiras que possuem um programa institucional profissionalizado de combate à evasão, com planejamento de ações, acompanhamento de resultados e coleta de experiências bem-sucedidas. (SILVA FILHO et al., 2007)

Atualmente, a grande maioria dos cursos de graduação tem como meta o combate ao abandono, mesmo que não tenham um plano de acompanhamento bem desenvolvido. Com o objetivo de formar um profissional cada vez mais capacitado e apto a desempenhar suas funções

e assim assegurar aos alunos uma maior inserção no mercado de trabalho evitando que vagas financiadas pelo dinheiro público permaneçam inocupadas, instituições financeiras têm apoiado o combate à evasão.

A grande expansão do poder computacional, que possibilitou um desenvolvimento maior das técnicas de mineração de dados, além do aumento de informações processadas e armazenadas atualmente são dois fatores que auxiliam no desenvolvimento das ações de combate à evasão. Uma das ações é a identificação de alunos que tenham maior potencial à evasão, buscando padrões para a evasão. Desta forma, o uso de recursos computacionais vem auxiliando no processo de redução dos altos níveis de evasão universitária, e, a partir disso, ações em vários aspectos podem ser aplicadas, como acompanhamento pedagógico ao potencial evasor.

O combate ao abandono já é objetivo de desenvolvimento de trabalhos na Universidade Federal de Santa Maria (UFSM). KANTORSKI et al. (2016) realizaram um estudo que buscou identificar alunos com potencial de evasão na UFSM. Como forma de parceria com a universidade, utilizaram-se de dados sensíveis, como renda e recebimento de benefício socioeconômico, além de dados acadêmicos do alunos, disponibilizados pelo centro de dados da instituição (CPD-UFSM).

1.3 PROPOSTA

Haja vista o poder computacional, a variedade de técnicas de mineração de dados e a grande quantidade de informação registrada digitalmente que se tem à disposição na atualidade, é imprescindível que essas técnicas sejam utilizadas para auxiliar a conter a evasão universitária.

Tendo como objeto de estudo os alunos matriculados na IES, busca-se uma combinação de características provenientes de dados não sensíveis do discente, que seja capaz de representar sua vida acadêmica. Além disso, optou-se pelo uso de um classificador necessariamente binário, pois dividirá os alunos em dois grupos: com potencial para evasão e sem potencial para evasão.

A partir dessas decisões iniciais, investigou-se quais características dos alunos são relevantes para a classificação e quais as técnicas que podem ser úteis. O resultado final alcançado é o desenvolvimento de uma modelo que realiza a classificação de forma precisa e que auxilia na tomada de decisões voltadas à redução da evasão, possibilitando medidas como alteração curricular ou acompanhamento pedagógico do aluno em questão.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 DESCOBERTA DE CONHECIMENTO

A quantidade de dados criada pelo ser humano nesta era digital aumenta consideravelmente a cada dia. Acompanhando a evolução do volume de dados gerado, a capacidade de armazenamento também cresceu muito. Visto que o acesso à informação é facilitado, e que o poder computacional já possibilita análises mais complexas sobre essa gigantesca quantidade de informação, é imprescindível o uso de ferramentas computacionais no auxílio à interpretação, análise e relacionamento dos dados.

Com o intento de atender a essas necessidades, existe uma área denominada KDD (*Knowledge Discovery in Databases*), que tem foco voltado principalmente à extração de conhecimento, análise e transformação de dados. A KDD atua principalmente sobre os conceitos de dado, informação e conhecimento, por isso é importante que se saliente as diferenças e regras hierárquicas entre eles.

Os três conceitos da pirâmide hierárquica da Figura 2.1 foram mostrados em GOLDSCHMIDT; PASSOS; BEZERRA (2015) da seguinte forma: os **dados**, na base da pirâmide, podem ser interpretados como itens elementares captados por recursos de Tecnologia da Informação, cadeias de símbolos que não possuem semântica. As **informações** representam os dados processados, com contextos bem definidos.O **conhecimento** corresponde a um padrão ou conjunto de padrões que pode envolver e relacionar dados e informações.

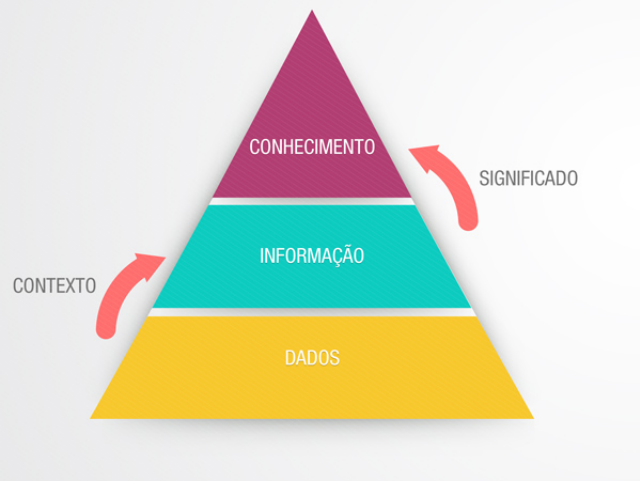

Figura 2.1: A hierarquia dado, informação, conhecimento (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Geralmente, o topo dessa pirâmide é o que tem mais relevância na análise e na busca de padrões. Sendo assim, informação e conhecimento são assumidos como bases para que se possa tomar decisões. Sistemas de apoio à decisão necessitam que o dado bruto disponível seja processado e modelado, para que dessa forma a maior quantidade de informação acerca do assunto seja obtida. Essa informação, associada à opinião de especialistas da área, forma o conjunto de fontes de conhecimento, base dos sistemas de apoio à decisão.

Assim, a sequência de procedimentos e técnicas empregadas para transformação de dados em conhecimento pode ser entendida como processo de KDD. Em uma definição formal, KDD é um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O processo de KDD é dividido em algumas fases, conforme a Figura 2.2, adaptada e traduzida de FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996). Embora os passos devam ser executados na dada ordem, o fato de o processo ser iterativo e interativo permite que o usuário intervenha para controlar o andamento das etapas, uma vez que cada etapa depende do resultado das anteriores.

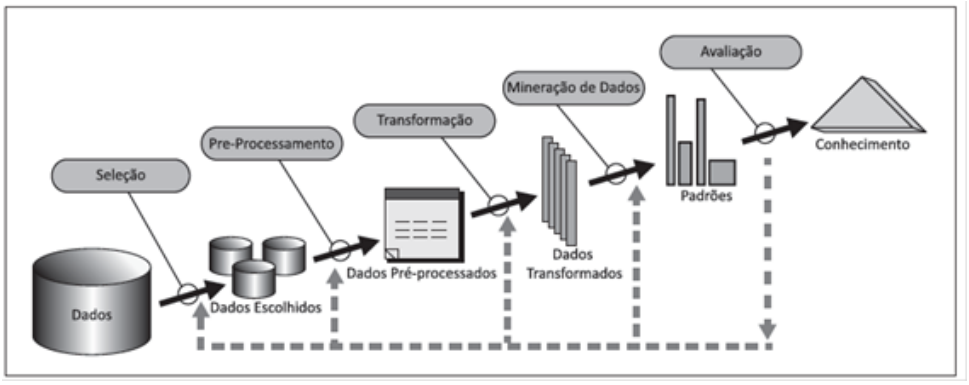

Figura 2.2: Fases do processo de KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A etapa inicial de seleção corresponde à fase em que são escolhidos os conjuntos de dados, as variáveis e os registros que serão utilizados durante o processo. Geralmente é a fase com maior complexidade e despende maior tempo de processamento, já que os dados muitas vezes provêm de fontes diferentes e possuem diferentes formatações e padrões.

A seguir, vem a fase de pré-processamento e limpeza. É nessa fase que os dados mais relevantes são processados, a fim de remover redundâncias, dados incompletos ou ainda dados fora da curva (*outliers*). É uma etapa de crucial importância pois pode determinar a eficiência

dos algoritmos nas etapas subsequentes.

A etapa de transformação de dados é responsável por fazer com que os dados estejam em um padrão, para que sejam aceitos pelo algoritmo de mineração. Geralmente é nesta fase que alguns atributos são combinados, a fim de gerar um novo atributo que possa representar melhor a informação desejada ou mesmo complementar dados faltantes. Um exemplo clássico de transformação de dados é a obtenção da idade de um indivíduo, a partir de sua data de nascimento. Esses dados gerados no processo são chamados de dados ou atributos derivados.

A mineração dos dados, etapa subsequente, é a mais abordada e abrangente da literatura. Essa fase concentra-se em técnicas e algoritmos que sejam capazes de fazer dos dados informação de alguma valia. Também é nessa fase que se costuma buscar padrões e regras associativas, para que se estabeleça uma correlação entre os dados, incluindo regras de classificação ou árvores de decisão, regressão ou agrupamento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Por fim, a etapa de análise e avaliação de resultados é aquela na qual os especialistas da área dão a devida interpretação à informação que acabou de ser gerada. É a etapa em que a informação se transforma em conhecimento e passa a auxiliar no decorrer das atividades, normalmente servindo como auxílio em uma tomada de decisão.

2.2 APRENDIZADO DE MÁQUINA

Programas de computador que aprendem ou melhoram seu desempenho a partir suas experiências sempre foram de grande interesse da área da Inteligência Artificial, desde seus primórdios (MITCHELL, 1997). Com o intuito de atender a esses requisitos, surgiu uma subarea da inteligência artificial chamada de Aprendizado de Máquina (ou *Machine Learning*, no inglês), que foca principalmente em conjuntos de técnicas e algoritmos com capacidade de aprendizado automático ou com poder de previsão de resultados.

MOHRI; ROSTAMIZADEH; TALWALKAR (2012) definiram aprendizado de máquina como um conjunto de métodos computacionais que usa de experiências para melhorar seu desempenho ou fazer previsões precisas. Para os autores, experiência se refere a toda informação passada que seja de alguma forma útil para tomadas de decisões futuras e que seja possível de ser obtida e analisada.

Programas da área de aprendizado de máquina podem utilizar-se dessas experiências coletadas na tomada de decisões. O programa embasa-se no fato de que uma decisão que foi

correta no passado continua sendo uma decisão correta no presente, para um caso similar. Por exemplo, em registros passados, vários pacientes que foram diagnosticados com a doença X apresentavam sintomas A, B, C e D. Hoje, um paciente que apresenta os sintomas A, B, C e D teria, para o sistema, grandes chances de possuir a doença X, pois os casos são similares.

Esse mesmo processo pode ser aplicado também a outras áreas totalmente distintas, o que caracteriza uma grande abrangência e inter-relação entre áreas do conhecimento. A área de educação, por exemplo, também se utiliza do aprendizado de máquina para tentar reduzir um dos seus principais problemas atualmente, que é a evasão escolar. Nesse caso, parte-se da ideia de que existe um padrão nas evasões dos discentes e que descobri-lo facilita muito a identificação de alunos que têm potencial de evasão, para os quais medidas anti desistência poderiam ser tomadas.

Ou seja, uma vez que se tenha um padrão a ser descoberto, regras de correlação entre os dados e decisões a serem tomadas, um algoritmo de aprendizado de máquina pode ser a solução para ambas as questões. A previsão que esses algoritmos possibilitam de um evento com alguma antecedência, como a evasão, no caso acima, pode ser importantíssima para que se encontre uma maneira mais eficiente de lidar com o problema.

De forma geral, há algumas maneiras distintas de um programa baseado em aprendizado de máquina de fato "aprender". A busca por padrões e tendências nos dados - de forma a possibilitar a separação dos registros em grupos - pode ser chamada de aprendizado indutivo e é feita através de duas estratégias principais. No âmbito de mineração de dados, as formas de aprendizado de um programa são conhecidas como aprendizagem supervisionada e aprendizagem não-supervisionada.

A aprendizagem não-supervisionada ocorre quando não se sabe de antemão quantos são os grupos diferenciáveis de dados que estão presentes. O resultado dessa abordagem é uma descrição das possíveis classes, que são os diferentes grupos que foram formados nos dados com base em uma função de similaridade. Geralmente, é uma abordagem que necessita de uma observação mais detalhada para que se consiga entender quais são as características que são semelhantes em cada um dos grupos gerados. Alguns exemplos de aprendizagem não-supervisionada são os métodos de agrupamento(*Clustering*) e as regras de associação.

A aprendizagem supervisionada ocorre quando se sabe previamente quais são os grupos em que os registros devem ser separados. O sistema então fica encarregado de gerar uma descrição para as classes e posteriormente formular a regra de classificação para cada uma delas.

Nessa abordagem, a aprendizagem é feita por meio de exemplos, como se fossem experiências passadas, que já devem conter o valor da classe predefinido. Muitas vezes, um especialista da área de interesse é necessário para que se consiga construir um sistema que proporcione uma melhor precisão de previsão. Classificação e regressão, vistos a seguir, são exemplos de métodos de aprendizagem supervisionada.

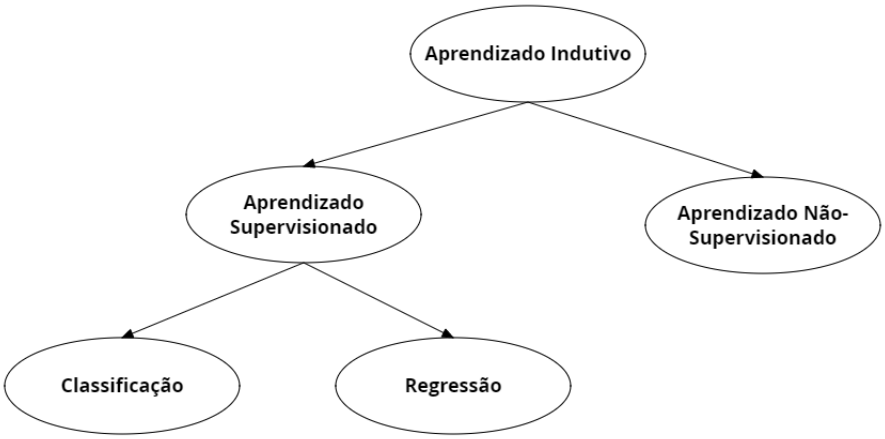

Figura 2.3: A separação das abordagens indutivas de aprendizado. Fonte: Autor

Considerando-se que este trabalho foca principalmente em tarefas de classificação, nas quais já se tem rótulos predefinidos, são demonstradas de forma mais aprofundada, a partir de agora, as técnicas de aprendizagem supervisionada.

2.2.1 Regressão

Regressão, também conhecida como aproximação de funções, é o processo de determinar um modelo que produza uma saída Y, dado um conjunto de entradas X, tal que $Y = f(X_n)$. O objetivo do processo de regressão é encontrar, nesse conjunto de saídas gerado também chamado espaço de hipóteses, a função que mais se assemelhe a função original f.

De forma geral, o modelo irá utilizar-se de n variáveis independentes $(X_1, X_2, ..., X_n)$ e uma variável dependente Y, a fim de estabelecer um valor para os coeficientes $a, \beta_1, \beta_2, ..., \beta_n$. Além disso, resulta também um valor de erro E, que consiste na variação que o modelo não consegue prever, de forma que o modelo resuma-se a uma função como:

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + E \tag{2.1}$$

Um modelo de regressão linear simples, por exemplo, concentra-se em encontrar uma

função que consiga aproximar da melhor maneira possível um conjunto de dados dispostos, por meio de uma equação linear (uma reta). A função linear encontrada assume um valor Y contínuo para cada uma das coordenadas do plano de projeção dos pontos referentes aos dados. A Figura 2.4 apresenta um exemplo de regressão linear genérico.

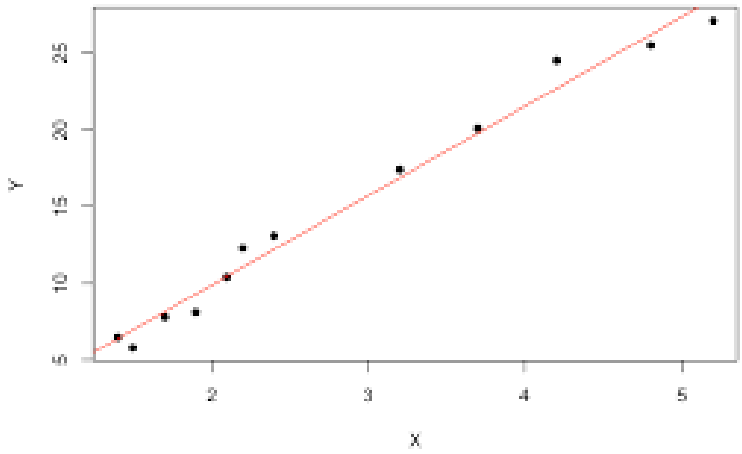

Figura 2.4: Um exemplo de regressão linear. Fonte: Autor.

Outra forma de regressão amplamente utilizada é a regressão logística. Muitos autores utilizaram-se da regressão logística em diversas áreas do conhecimento, tais como ciências sociais (PEARL; REED, 1920), (OLIVER, 1982), ciências biológicas e saúde pública (DYKE; PATTERSON, 1952), (GRIZZLE, 1961). É um tipo especial de regressão, que utiliza regressão linear generalizada com uma função de ligação, que no caso é a função *logit*. A grande diferença do modelo de regressão logística para os demais modelos de regressão é a capacidade de, por meio da função *logit*, prever o valor de uma variável discreta, geralmente binária.

Dessa forma, a regressão logística se compara às demais tarefas de classificação que ainda serão vistas nessa seção, principalmente pelo fato de ambas possuírem a variável resposta do tipo categórico. Isso permite que os analistas de dados possam escolher o modelo mais indicado para o problema.

2.2.2 Classificação

Classificação é a tarefa de mineração de dados que consiste em organizar os dados em classes predefinidas, por meio de um modelo de classificação. Esse modelo de classificação deve analisar o conjunto de atributos de entrada que corresponde ao registro e decidir qual(is)

da(s) classe(s) é mais semelhante a ele. A grande diferença entra um classificador e um modelo de regressão é o tipo da variável de saída. Enquanto no modelo de regressão a variável de saída é contínua, no classificador ela é uma variável discreta.

Geralmente, as bases de dados utilizadas para tarefas de classificação são divididas em duas partes. Uma primeira parte contém os dados de treinamento do classificador, os quais serão utilizados pelo algoritmo de aprendizagem para geração do modelo de classificação. A segunda parte corresponde aos dados de teste, que farão a validação e avaliação do modelo gerado pelo algoritmo.

Classificadores são as técnicas mais comuns e com maior leque de aplicabilidades dentre as abordagens de aprendizado de máquina. São variadas as opções de classificadores que estão à disposição, dentre os quais as árvores de decisão, os classificadores bayesianos, as redes neurais artificiais, os classificadores baseados em regras, as máquinas de vetores de suporte, os modelos logísticos e outros.

2.2.2.1 Árvore de decisão

Árvore de decisão é um dos métodos de inferência indutiva mais amplamente utilizado, além de ser um dos mais práticos e resistentes a dados ruidosos (MITCHELL, 1997). Basicamente, uma árvore de decisão é composta de dois tipos distintos de nós: os nós folha, e os nós de decisão. Nós folha correspondem a um dos rótulos predefinidos pelo usuário, enquanto os nós de decisão sempre contém um teste sobre um atributo relevante. O resultado é uma nova subárvore com a mesma estrutura da árvore principal para cada uma das possíveis soluções do teste.

Visualizando a Figura 2.5, percebe-se que há no problema duas classes possíveis, doente e saudável. No exemplo, aparecem nos nós folha, retratados de forma retangular. Os demais nós da árvore são os atributos que foram utilizados no modelo de classificação, tais como dor, temperatura do corpo e modo como se sente o paciente. É possível perceber ainda que quando há um atributo contínuo, a divisão das subárvores será sempre binária. No exemplo, a temperatura tem sua divisão entre acima de 37^{o}C e abaixo de 37^{o}C.

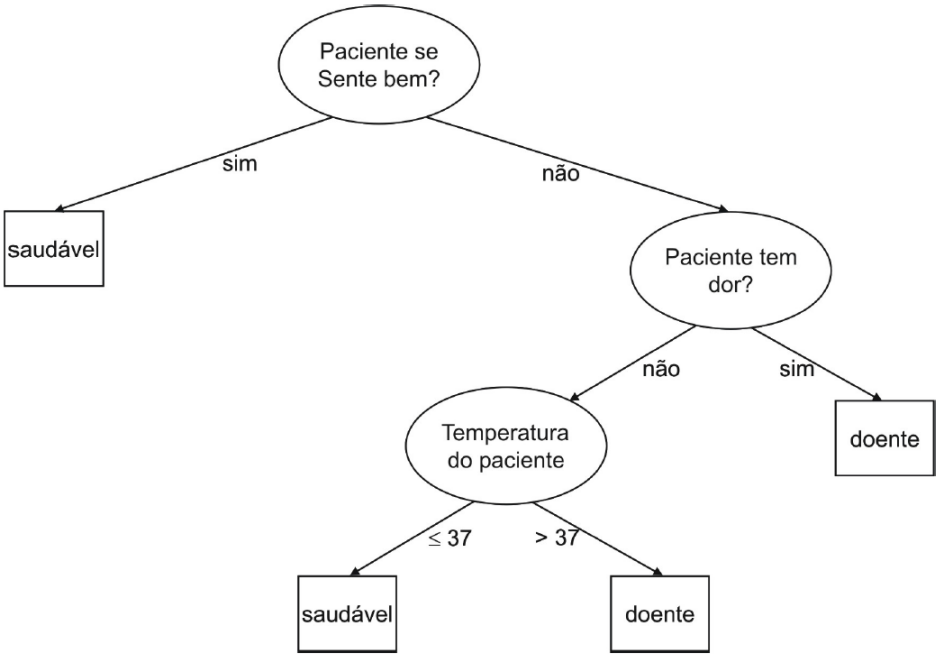

Figura 2.5: Um exemplo simples de árvore de decisão para diagnóstico de um paciente (MONARD; BARANAUSKAS, 2003).

Outra forma de interpretação de uma árvore de decisão é aquela feita por meio de um conjunto de regras. Para MITCHELL (1997), as árvores representam uma disjunção de conjunções de regras, porque todos os caminhos a partir da raiz até uma folha representam uma conjunção de atributos testados, e a própria árvore é uma disjunção dessas conjunções.

2.3 SELEÇÃO DE ATRIBUTOS

A seleção de atributos entra como uma técnica de auxílio às tarefas de criação de modelos preditivos. É uma de suas tarefas escolher quais dos atributos são mais ou menos relevantes dentre um leque de opções, mantendo ainda um nível aceitável de precisão. Os objetivos principais da técnica são três: melhorar o desempenho dos preditores, prover preditores mais rápidos e economicamente viáveis, e proporcionar um melhor entendimento das entrelinhas do processo de geração dos dados (GUYON; ELISSEEFF, 2003).

LIU; MOTODA (1998) definiram-na como o processo de escolha de um grupo de características ótimo, de acordo com um certo critério. É este critério que especifica todos os detalhes de moldagem do conjunto de dados e características, e, dessa forma, uma escolha não adequada pode impactar no resultado final. Por isso, a escolha de um subgrupo de características ótimo

nem sempre implica escolher o subgrupo mínimo, mas sim aquele que atende a uma melhor proporção entre dimensão e acurácia preditiva.

Pode ser utilizada também para identificar atributos que são irrelevantes para o problema disposto ou mesmo que interfiram de forma negativa na criação de um modelo de predição. Uma vez que existam dados redundantes ou ruidosos na hora da criação do modelo, pode ocorrer uma queda na precisão do mesmo.

Existem basicamente duas formas principais de selecionar quais são os atributos relevantes: a redução de dimensionalidade e a escolha de atributos. A diferença básica entre os dois métodos é que a redução de dimensionalidade geralmente combina alguns atributos para a formação de novos atributos, enquanto a outra trabalha sem alterar os dados, apenas realizando uma escolha sobre qual dos atributos manter ou remover (BROWNLEE, 2014).

2.3.1 Redução de dimensionalidade

Muitas vezes, trabalha-se com bases de dados que possuem um número bem considerável de atributos. Dessa forma, a dimensionalidade do problema, que é proporcional ao número destes atributos, também aumenta. Assim, a redução de dimensionalidade é uma técnica que visa minimizar a complexidade do problema, facilitando dessa forma a execução de algoritmos de mineração de dados.

A redução de dimensionalidade consiste principalmente em combinar um ou mais atributos em um novo atributo. Dentre as várias técnicas existentes, as principais são: remoção de atributos não valorados, filtros de baixa variância e alta correlação. Os estudos realizados sobre essas técnicas alcançam resultados de redução da dimensionalidade acima de 60% (SILIPO et al., 2015).

2.3.2 Escolha de atributos

A escolha de atributos consiste em desprezar atributos que são menos relevantes ou que são duplicados. Desenvolvidos para aumentar a eficácia e o desempenho dos classificadores, os métodos de escolha de atributos costumam ser diferenciados entre duas abordagens: *filters* e *wrappers*.

2.3.2.1 Wrappers

É a forma mais simples de escolha de atributos, e utiliza-se de um algoritmo de classificação simples como caixa-preta para atribuir um escore ao conjunto de características que está em treino no momento. Esse score é baseado no poder preditivo, ou seja, na acurácia do classificador (KOHAVI; JOHN, 1997).

Conforme a Figura 2.6, a abordagem *wrapper* possui duas fases distintas. A primeira fase consiste basicamente em execuções consecutivas dos algoritmos que geram os subgrupos de características e dos algoritmos de aprendizado. O processo se encerra assim que um subgrupo é considerado adequado, geralmente em termos de acurácia na classificação. A segunda fase corresponde ao treinamento do classificador, por meio do algoritmo de aprendizado, seguido da validação e da medição de acurácia.

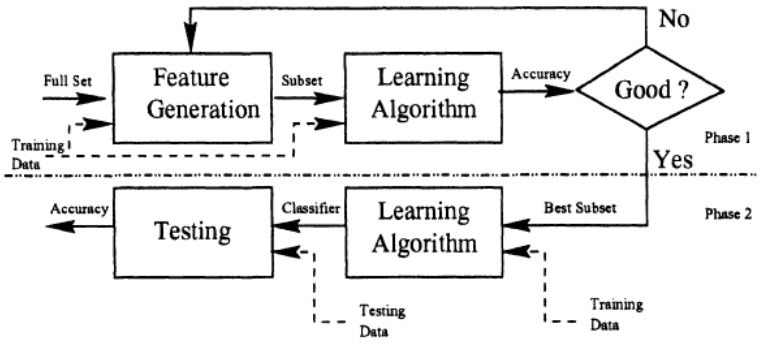

Figura 2.6: A abordagem *wrapper* (LIU; MOTODA, 1998).

Dessa forma, por utilizar um classificador na forma de caixa-preta, a abordagem *wrapper* torna-se universal. Com esse modelo é possível analisar uma grande variedade de dados, nas mais variadas formas, apenas escolhendo um classificador para verificação de acurácia e escolha de um subgrupo de características consideradas adequadas. Porém, vários estudos desenvolvidos por outros pesquisadores como DASH; LIU (1997), analisaram a performance desse modelo com relação a seleção de atributos e sugeriram outra forma de seleção. Essa abordagem ficou conhecida como *filters*, cuja descrição do funcionamento é objetivo da próxima subseção.

2.3.2.2 Filters

Muitas vezes, quando se possui um conjunto de dados de um tamanho excessivo, ou quando há muitos atributos a serem considerados, contando ainda com as limitações impostas

pelo classificador escolhido, não é viável aplicar um classificador diretamente sobre os dados, tornando-se necessário pré-processar os dados de outra maneira (LIU; MOTODA, 1998).

É assim que surge a abordagem *filters*, propondo uma mudança que ataca a primeira fase do modelo visto na Figura 2.6. Basicamente, o modelo sugere que sejam utilizadas medições como informação, as quais costumeiramente correspondem a coeficientes de correlação entre atributos ou medida de entropia, conforme a Figura 2.7. Dessa forma, não é necessário o uso de classificadores diretamente sobre os dados, um fato que atrapalhava a performance do modelo *wrapper* quando do tamanho excessivo da base dados utilizada.

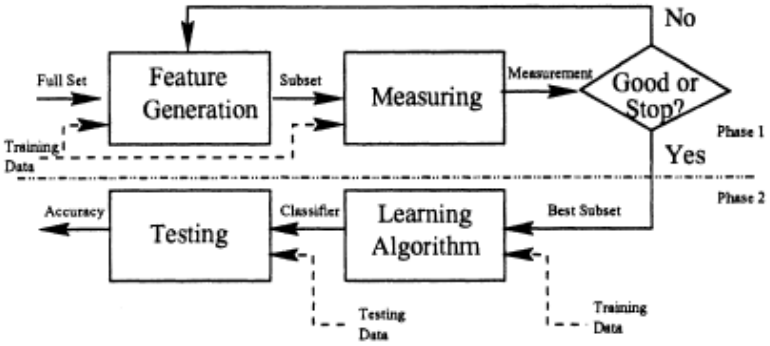

Figura 2.7: A abordagem *filter* (LIU; MOTODA, 1998).

Estabelecendo um breve comparativo, *wrappers* costumam trazer um resultado melhor, porém estão sempre suscetíveis e dependentes ao classificador e ao tamanho da base dados. O tempo também é um fator a ser considerado, sempre que se aumenta a base de dados. Já os *filters* nem sempre retornam um resultado excelente, mas são a solução quando se trata de grande porte de dados. Em termos de metodologia, a grande diferença entre as abordagens reside no processo de escolha do subgrupo de características. Enquanto nos *wrappers* o processo de escolha é embutido no processo de aprendizado, nos *filters* a escolha ocorre antes do aprendizado.

2.3.3 Atributos não valorados

Não raramente, quando se está a construir uma base de dados, nem todas as informações necessárias para a construção do modelo estão presentes. Por vezes pode não saber-se qual o valor de uma das características para alguns dos casos presentes na base de dados. Estes atributos que não estão presentes para alguns dos exemplos são comumente denominados na literatura como atributos não valorados ou valores desconhecidos.

Muitas vezes, a origem de um valor desconhecido na base de dados deve-se ao preenchi-

mento falho da base por parte de quem a construiu. Nesse caso, cabe ao analisador desses dados decidir quais as medidas a serem tomadas, ou seja, definir se este atributo que está faltando é relevante ou não, e excluí-lo caso não o seja.

O uso de atributos desconhecidos deve ser previsto e tratado pelo algoritmo de classificação. Em árvores de decisão, por exemplo, a maioria das técnicas simples de indução assume que todos os atributos estarão presentes em todos os casos da base (WU et al., 2008). Para contornar tal problema, algumas técnicas foram desenvolvidas e são utilizadas até então.

Para ilustrar estas técnicas, suponha um atributo A. Se o atributo A é um atributo de decisão da árvore e é não valorado para um caso X do conjunto de teste, as opções principais que são utilizadas pelo algoritmos hoje são: considerar apenas os casos de treino que contenham um valor para A (BREIMAN et al., 1984) ou preencher o valor faltante com um valor em termos dos demais valores do atributo A em outros casos por meio de árvores de decisão (QUINLAN, 1986).

2.4 MÉTRICAS DE AVALIAÇÃO

Na mesma proporção de importância que é dada à escolha de um modelo correto e adequado, por meio das técnicas já mencionadas, é imprescindível que se saiba as métricas corretas para avaliá-lo da melhor forma possível. Fatores como desproporcionalidade da distribuição dos registros entre as classes podem interferir na qualidade de avaliação que algumas das métricas trazem. Desse modo, é importante saber qual métrica usar e para qual finalidade.

A grande maioria das métricas de avaliação baseia-se em quatro métricas fundamentais: verdadeiros positivos(VP), verdadeiros negativos(VN), falsos positivos(FP) e falsos negativos(FN). As demais métricas, como precisão, acurácia, cobertura, entre outras, que buscam melhor representar uma característica do modelo que foi induzido e são métricas derivadas daquelas. As quatro métricas fundamentais costumam aparecer na forma de uma matriz de confusão, conforme a Figura 2.8.

Para facilitar a compreensão das métricas fundamentais, utiliza-se o exemplo da Figura 2.8, com apenas duas classes possíveis, POSITIVO e NEGATIVO. Além disso, suponha que existam 100 registros disponíveis para teste, e sabe-se que 40 deles são POSITIVO e os outros 60 são NEGATIVO. Ao passar por todo o processo de aprendizado do classificador e posterior teste, os resultados na matriz de confusão estão descritos na Figura 2.8.

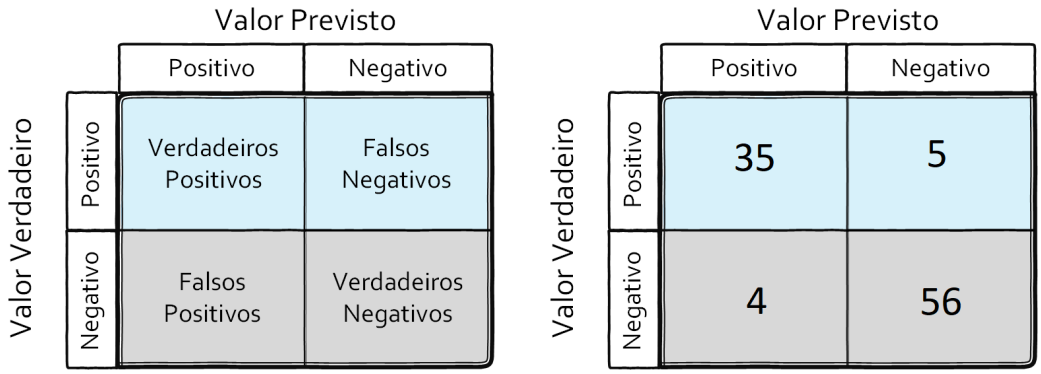

Figura 2.8: Um exemplo de matriz de confusão. Adaptado de ANDREONI (2014).

Verdadeiros positivos e verdadeiros negativos são as duas métricas que dizem respeito aos acertos do classificador. Os verdadeiros positivos ocorrem quando o valor verdadeiro da classe é positivo e o valor previsto também é positivo; no exemplo, 35. Os verdadeiros negativos ocorrem então quando o valor verdadeiro é negativo e o valor previsto também é negativo; no exemplo, 56.

Falsos positivos e falsos negativos são as duas métricas que representam os erros do classificador. Tratam-se como falsos positivos todos os registros que eram verdadeiramente negativos, mas foram classificados como positivos; 4, no exemplo. Já os falsos negativos são os registros que eram verdadeiramente positivos, mas erroneamente foram classificados negativos; 5, no exemplo.

Seguindo pelo mesmo exemplo, exemplificam-se as demais métricas. A acurácia, que é uma métrica generalista, é utilizada principalmente para obter-se uma visão geral de quantas instâncias estão sendo classificadas corretamente. Não representa uma medida precisa, uma vez que, se as classes estiverem desbalanceadas, a porcentagem de acertos geral acaba sendo mascarada pelos acertos de uma única classe. É calculada da seguinte forma:

$$A = \frac{VP + VN}{VP + VN + FP + FN}$$
$$A = \frac{35 + 56}{35 + 56 + 5 + 4}$$
$$A = \frac{81}{90} = 90\%$$

A precisão, outra medida de avaliação, já é mais objetiva. Mede os acertos sobre uma das classes apenas. O intuito principal da precisão é saber a pureza do resultado que foi obtido, ou seja, saber qual o percentual de positivos verdadeiros dentre os positivos classificados. A

fórmula de cálculo da precisão e a precisão sobre os positivos é demonstrada abaixo:

$$P = \frac{VP}{VP + FN}$$
$$P = \frac{35}{35 + 4}$$
$$P = \frac{35}{39} = 89.7\%$$

Outra importante métrica que será utilizada neste trabalho é a cobertura, conhecida mais amplamente na literatura como *recall*. A cobertura mede principalmente a capacidade de abrangência do modelo ao oferecer como resultado a porcentagem do total de registros daquela classe que foi classificada corretamente. O valor de cobertura para a classe de positivos é dado abaixo, seguindo a fórmula:

$$C = \frac{VP}{VP + FP}$$
$$C = \frac{35}{35 + 5}$$
$$C = \frac{35}{40} = 87.5\%$$

2.5 A METODOLOGIA CRISP-DM

Dividida em seis fases (conforme a Figura 2.9) a metodologia *CRoss Industry Standard Process for Data Mining* (CRISP-DM) (SHEARER, 2000) é amplamente utilizada quando da necessidade da criação de um modelo de aprendizado, por meio de problemas que envolvam mineração de dados e aprendizado de máquina. Pela grande semelhança, principalmente com o modelo de construção do conhecimento proposto por FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996), essa metodologia torna-se eficiente e representa de forma coerente todos os passos necessários para a solução do problema em questão.

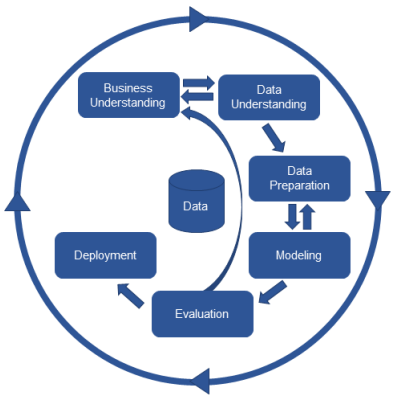

Figura 2.9: O modelo de processo CRISP-DM. (TAYLOR, 2017)

Business Understanding representa a primeira fase do modelo, que corresponde ao estudo do problema. Inevitavelmente, estudar o problema em que se está inserido é etapa essencial no processo de criação de um modelo indutivo. É estudando o problema que se propõem as possíveis soluções que irão nortear o desenvolvimento do modelo. Então, mesmo que FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996) omitam ou não deem tanto foco a essa fase no seu modelo de construção de conhecimento, um melhor entendimento do problema, por consequência, geralmente converge para soluções melhores, tornando dessa forma imprescindível a execução dessa fase.

Data Understanding é o passo seguinte ao entendimento do problema, que corresponde à escolha dos dados a serem utilizados e ao entendimento dos mesmos. Nessa fase, geralmente são definidas as fontes e métodos de obtenção dos dados, além da sua análise de relevância. Não raramente, retorna-se desta para a primeira fase por diversos motivos, seja por impossibilidade de obtenção dos dados, seja por inviabilidade de execução da solução proposta, ou outros ainda. Uma vez que os dados foram obtidos, realiza-se um estudo a fim de compreender a significância dos dados para a solução do problema.

Também é importante ressaltar que, devido ao uso da metodologia CRISP-DM, é possível retornar deste ponto ao ponto inicial de aquisição de dados e entendimento do negócio quantas vezes seja necessário. Ou seja, se os dados adquiridos não estiverem de acordo com a proposta inicial ou não se pode obtê-los, pode-se sempre retornar ao estágio anterior, selecionar novos dados ou mudar de proposta. Com os dados obtidos e detalhados, com a devida importância já identificada, o próximo passo a ser dado é a preparação do modelo de dados.

Data Preparation é a etapa responsável por todo o preprocessamento necessário a fim de padronizar a base de dados, optar por redução de dimensionalidade, criação de novos atributos, fusão de atributos em um único, entre outros. Esta etapa, resumidamente, define o modelo de dados que irá ser utilizado. Isto se relaciona também com a próxima fase da metodologia, que preza pela definição de qual estratégia de mineração de dados irá ser utilizada, definida de acordo com a proposta de solução do problema que se quer atingir. É uma etapa que geralmente se repete muito durante o processo, já que o modelo como um todo normalmente não alcança avaliações de excelência durante os primeiros testes. Logo, é importante que se retorne a esta fase e se proponha um novo modelo de dados para tentar incrementar o desempenho.

Modeling é a fase que contempla todas as técnicas de mineração de dados que podem ser colocadas em prática, dado o modelo planejado até agora. Porém, muitas vezes o planejamento

realizado na primeira fase é realizado pensando em antecipar fase de modelagem. Um problema de classificação, muito comumente, por exemplo, é planejado desde os estágios iniciais. A fase de modelagem é executada proporcionalmente à fase de *Data Preparation*, sendo necessária a calibragem dos parâmetros e refinamento de atributos, que ocorrem muitas vezes durante todo o processo.

Evaluation é a fase que determina se o modelo está no caminho certo ou não. Obtendo-se resultados que não são os desejados na fase de avaliação, é necessário reiniciar todo o processo. É necessário que seja modificado o modelo proposto, a fim de que os resultados passem a ser mais satisfatórios.

3 METODOLOGIA

Neste capítulo são abordadas questões relacionadas a levantamento de requisitos, ferramentas utilizadas, aquisição de dados, planejamento e criação do modelo indutivo proposto. Baseado na metodologia conhecida popularmente por CRISP-DM, são abordados nas seções subsequentes os tópicos principais acerca do modo como o modelo preditivo foi desenvolvido.

3.1 FERRAMENTAS UTILIZADAS

Antes de descrever os passos executados seguindo a metodologia CRISP-DM, esta seção descreve as ferramentas utilizadas no preprocessamento dos dados e na fase posterior de mineração dos dados. A linguagem Python foi a escolhida para as tarefas de preprocessamento, enquanto que a tarefa de mineração ficou a cargo da ferramenta WEKA.

A linguagem Python foi escolhida por proporcionar uma grande facilidade no que diz respeito à manipulação dos dados. Uma vez que os métodos automatizados de seleção de atributos vistos na seção 2.3.2 não foram utilizados, devido à grande demanda de preprocessamento e transformação dos dados, a seleção dos atributos foi realizada de forma manual, por meio de um *script* desenvolvido, que faz a limpeza e o processamento necessário, conforme é descrito na seção 3.4.

A ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) é amplamente utilizada no âmbito de mineração de dados. A ferramenta possui um vasto leque de algoritmos e funcionalidades, inclusos os algoritmos de mineração propriamente ditos, além de filtros, seletores de atributos, visualizadores de dispersão e de árvores de decisão, entre outras. O WEKA foi escolhido principalmente por ser de fácil compreensão e por possibilitar uma análise detalhada dos resultados.

Como a tarefa principal deste trabalho é a proposição de um conjunto de atributos que maximize o desempenho de classificação, o WEKA cumpre este papel principal de proporcionar a execução dos algoritmos de classificação sobre o modelo proposto e a posterior análise de desempenho por meio das métricas de avaliação.

3.2 ENTENDIMENTO DO PROBLEMA E LEVANTAMENTO DE REQUISITOS

O problema da evasão escolar de nível superior, como já dito, é de grande dimensionalidade e gera prejuízos, independentemente de área do conhecimento ou localização geográfica. Como já comentado na seção 1.1, os índices de não conclusão são alarmantes, por volta de 40% na área das engenharias e chegando a taxas de 50% em ciências exatas e da terra (DAVOK; BERNARD, 2016).

Porém, diversos são os motivos para que um aluno possa vir a evadir. Muitas vezes, por não se sentir à vontade, ele têm dificuldade na adaptação (MORAIS et al., 2017). Esse fato é imensamente prejudicial, tanto para a instituição de ensino, como para os docentes, ou ainda para o mercado de trabalho, que vê candidatos em potencial abandonarem suas chances, deixando cargos em aberto.

Para fins deste trabalho, estudaram-se os possíveis motivos de registro de uma desistência dentro da Universidade Federal de Santa Maria (UFSM). A UFSM cita diversos motivos em seus registros, dentre os quais os principais estão presentes na Tabela 3.1, representados por descrição e código.

Tabela 3.1: Motivos principais de evasão na UFSM.

Descrição	**Código**
Transferido	2
Falecimento	3
Formado	4
Transferência Interna	5, 6, 20
Jubilamento	7
Cancelamento/Trancamento	8, 12, 26
Abandono	9
Classificado e não matriculado	17
Transferência/Reativação Vínculo	18

Visto que o objetivo inicial deste trabalho é a correta identificação de alunos em potencial para evasão, ficou decidido previamente do uso de um classificador, de forma que consiga fazer a diferenciação entre alunos que evadiram e alunos que graduaram. Para tal, o modelo de registro de evasões facilita a obtenção deste tipo de dado. Conforme a Tabela 3.1, pode-se perceber os principais tipos de evasão encontrados, dando atenção especial ao "FORMADO"(cod. 4). Pode-se perceber que este é o único da natureza de conclusão, enquanto que os outros são de evasão.

3.3 ENTENDIMENTO DOS DADOS

No começo do processo de busca e construção do conhecimento, ocorre a decisão sobre o teor dos dados a serem adquiridos. Muitas vezes uma escolha ruim dos dados interfere, e muito, nos resultados finais. Foi decidido então que, por facilidade de obtenção e também pela não existência de invasão de informações privativas, seria feito o uso de dados não-sensíveis, como aprovações e reprovações em disciplinas da grade curricular.

Uma vez decidido que se utilizasse de dados não-sensíveis dos discentes, disponíveis via Sistema de Informações para o Ensino (SIE-UFSM), foi necessário que se investigasse quais os dados que estão disponíveis e que podiam ser utilizados no desenvolvimento do modelo. Os dados teriam que representar da melhor forma possível a vida acadêmica do aluno. A proposta então foi a de moldar o modelo sobre as características fundamentais de um discente regular: a análise das aprovações nas diferentes disciplinas.

Verificada a disponibilidade da proposta inicial, configurou-se o modelo de dados a ser adquirido: a base de dados deveria conter registros de aprovação dos alunos. Sendo assim, é importante destacar alguns atributos principais e outros de menor relevância. Existem alguns atributos que conseguem, mais que os outros, representar melhor a vida acadêmica do discente. Isso significa que eles têm um maior poder de representação e serão mais trabalhados nas fases seguintes.

Atributos como 'ano de conclusão' e 'período de conclusão' da disciplina, além de 'ano ideal' e 'período ideal' para conclusão, são as características que mais trazem consigo uma ideia de progresso, mesmo que indiretamente, de cada discente respectivamente. Além desses, o registro traz um dado fundamental, uma vez que a ideia é classificar alunos com potencial de evasão. O atributo 'forma de evasão' traz diferentes valores, que condizem com as possíveis razões para um discente evadir de determinado curso, incluindo colação de grau.

Dessa forma, os atributos chave utilizados são:

- anoConclusao, que representa o ano em que o discente concluiu determinada disciplina, sendo que para este trabalho varia entre 1992 e 2017;
- anoIngresso, que representa o ano em que o aluno ingressou na instituição de ensino, que para este trabalho varia entre 1992 e 2016;
- periodoConclusao, que remete ao semestre em que foi concluída a disciplina, com valor possível 1 ou 2, fazendo menção à primeiro ou segundo semestre do ano;

- periodoIdeal, que faz menção ao semestre considerado ideal para conclusao da disciplina e varia entre 1 e o número de semestres correspondentes ao respectivo curso.

Sendo assim, a proposta desenvolvida a partir dos dados disponíveis é a seguinte: utilizar-se dos dados de aprovação nas disciplinas por cada aluno para realizar um mapeamento de evolução no curso no que se refere ao adiantamento ou ao atraso do discente em relação às condições ideais. Mais especificamente, a proposta é guardar um registro para cada aluno, sendo que o número de atributos que este registro contém é exatamente o número de disciplinas que são possíveis de serem cursadas por esse discente.

No que diz respeito ao valor desse atributo para as respectivas disciplinas, ele é um coeficiente que representará, em número de semestres, qual o tempo de adianto ou atraso do aluno. Vale ressaltar que um valor positivo indicará que o aluno está atrasado, enquanto um valor negativo representa que o aluno está adiantado. Por consequência, quando o valor atribuído for zero, indica que o discente concluiu a disciplina no semestre ideal.

3.4 PREPARAÇÃO DOS DADOS

Como citado na seção anterior, algumas informações importantes para que seja possível representar a vida acadêmica do aluno estão apresentadas de forma indireta. Para alcançar a melhor forma do modelo proposto na seção anterior, um preprocessamento de grande porte é necessário. É preciso transformar todos os registros referentes a um aluno em um único registro, que contenha todas as disciplinas que ele cursou preenchidas com o coeficiente de atraso ou adiantamento correspondente.

A primeira tarefa é desenvolver um coeficiente que consiga suprir a necessidade de representar a linha do tempo que é o atraso ou adiantamento de um aluno em uma dada disciplina. Para tal, foi desenvolvido um cálculo que tem como resultado o número de semestres em que se está defasado ou à frente do ideal.

A fórmula desenvolvida, com base nos atributos chave descritos na seção 3.3, é a dada pela fórmula 3.1.

$$Coeficiente = 2 * (anoConclusao - anoIngresso) + (periodoConclusao - periodoIdeal) \tag{3.1}$$

De acordo com a equação 3.1, toma-se, por exemplo, um aluno que tenha ingressado na

instituição de ensino em 2010. Este concluiu uma disciplina A, cujo período ideal é o segundo, no ano de 2011, período 2. Pela fórmula, chega-se à conclusão de que este aluno concluiu a disciplina com 2 semestre de atraso. Da mesma forma ocorreria se este discente concluísse a disciplina no tempo correto, recebendo assim como coeficiente o valor zero, e ocorreria também se ele concluísse antes do recomendado, recebendo um coeficiente negativo proporcional ao número de semestres que está adiantado.

Sendo assim, aplica-se a fórmula 3.1 a todas as disciplinas que o discente obteve aprovação. Ao final desse processo, obtém-se um valor para cada uma das disciplinas cursadas. No que diz respeito às disciplinas não cursadas, o valor será preenchido com o caractere '?', simplesmente por exigência de sintaxe da ferramenta utilizada na realização dos testes. Para o WEKA, esse caractere representa um atributo não valorado.

Outro ponto importante dessa fase é a discretização das possíveis classes em apenas duas: evadidos ou graduados. Dessa forma, todos os alunos que não têm por natural a classificação de graduado, automaticamente estão agora inseridos na classe de evadidos. Essa técnica faz com que o modelo se torne mais genérico e menos suscetível ao embaralho entre classes.

O resultado final desta etapa é um único registro para cada aluno presente no relatório. Nele está contido um valor para cada uma das disciplinas em que este discente obteve aprovação, sendo que as disciplinas não aprovadas são representadas com o caractere identificador de atributo não valorado '?'.

Para tal tarefa de processamento e criação do novo modelo, foi desenvolvido um script automatizado, em linguagem Python, que recebe o banco de dados no formato '.csv', com ordem de atributos predefinidos e retorna dois arquivos em formato '.arff', próprios da ferramenta WEKA. Um deles contém os dados de treinamento, e o outro contém os dados de validação.

Um arquivo no formato '.arff' é um arquivo com padrão especial, desenvolvido para uso exclusivo da ferramenta WEKA. O cabeçalho desse arquivo sempre contém o nome dado ao modelo, além dos atributos de cada um dos registros, tratados respectivamente pelas tags '@RELATION' e '@ATTRIBUTE', sendo que esta última aparece tantas vezes quantos forem os atributos do problema. Para finalizar o cabeçalho, indica-se o início da seção de dados com a tag '@DATA'. Abaixo um exemplo de cabeçalho e dados em um simples arquivo '.arff'.

```
@RELATION evasao

@ATTRIBUTE disc_1 real
@ATTRIBUTE disc_2 real
@ATTRIBUTE disc_3 real
@ATTRIBUTE disc_4 real
@ATTRIBUTE disc_5 real

@DATA

0,1,2,4,?
0,-1,0,0,0
0,0,0,0,0
1,1,3,3,5
0,0,?,?,?
```

Figura 3.1: Um exemplo de arquivo '.arff'.

Como se pode observar na Figura 3.1, vários coeficientes são atribuídos para disciplinas, numeradas de 1 a 5. Cada registro inserido na seção de dados do arquivo '.arff' é representado em uma linha do arquivo, com obrigatoriamente um valor designado para cada um dos atributos descritos no cabeçalho. Pode-se perceber no exemplo acima que há neste arquivo 5 instâncias representando alunos em 5 disciplinas diferentes. Nota-se que, no caso de o discente não ter concluído uma das disciplinas, é inserido um '?' para representar o valor desconhecido.

Ainda é importante comentar que, da forma como foi proposto este modelo, a dimensionalidade do problema é diretamente proporcional ao número de disciplinas possíveis. Mesmo assim, técnicas de seleção de dados como *wrappers* e *filters* são dispensáveis neste caso em específico, neste momento, já que o objetivo do modelo em questão é detalhar ao máximo a vida acadêmica do aluno a fim da obtenção de uma precisão máxima.

3.5 MODELAGEM

Objetos de estudo da seção 2.3.3, atributos não valorados terão parte importante neste trabalho pela necessidade de representar disciplinas que não foram cursadas pelo aluno. Assim,

o algoritmo de classificação a ser escolhido precisa necessariamente ser compatível com esse tipo de representação.

No caso deste trabalho, a fase de modelagem contempla a escolha do classificador, uma vez que o fato de a tarefa ser de classificação já fora decidida ainda nas fases iniciais. Após uma análise mais delicada dos dados e do problema, optou-se por realizar a classificação utilizando um modelo de classificação híbrido, que representa um misto entre árvore de decisão e regressão logística, conhecido como *Logistic Model Tree (LMT)* (LANDWEHR; HALL; FRANK, 2005).

O LMT utiliza regressão logística nas folhas de uma árvore de decisão induzida pelo algoritmo C4.5 (QUINLAN, 2014) utilizando o algoritmo de boosting LogitBoost, que se trata do algoritmo generalizado AdaBoost (FRIEDMAN; HASTIE; TIBSHIRANI, 1998) aplicando a função de regressão logística como função de custo. Em outras palavras, os níveis superiores da árvore direcionam, via C4.5, os novos registros de classificação para nós folha, nos quais se utiliza regressão logística para determinar a classe do novo registro.

O modelo do LMT é de grande valia para esta tarefa de classificação por um motivo especial. O LMT foi desenvolvido de forma que aceita muito bem trabalhar com atributos faltantes e além disso, consegue reduzir muito a ocorrência de *overfitting* (LANDWEHR; HALL; FRANK, 2005). *Overfitting* é o termo que se usa, em estatística, para denominar o fato de o classificador se ajustar muito bem ao grupo de testes, mas ser ineficaz quando novos dados são inseridos. Pelo fato de o modelo proposto neste trabalho contar com dados diferentes para treino e teste, o LMT se mostra eficiente e preciso.

3.6 AVALIAÇÃO E DESENVOLVIMENTO

A seção de avaliação e desenvolvimento, como comentado na seção 3.2, cabe aos resultados e trabalhos futuros. A partir disso, a avaliação de desempenho do modelo proposto é mostrada no Capítulo 4, que explana todo o conhecimento gerado acerca do processo de identificação de alunos com potencial para evasão.

Durante os experimentos, foram utilizados dados de diversos cursos da UFSM, a fim de comprovar se os resultados se mantinham quando da troca das disciplinas utilizadas. No capítulo de experimentos, são vistos todos os resultados gerados a partir desses modelos, bem como a forma com que esses experimentos foram conduzidos.

4 EXPERIMENTOS E RESULTADOS

Este capítulo trata dos experimentos realizados sobre o modelo de classificação proposto no capítulo anterior, bem como dos resultados obtidos, medidos através das métricas de avaliação.

4.1 O CONCEITO DE MARCO

Quando se trabalha com dados que contêm apenas registros de alunos evadidos, como é o caso, é necessário que se criem alternativas a fim de adquirir dados de validação do modelo. Assim, foi definido o conceito de marco, um ano e período específicos, como se fossem uma barreira temporal, segregando os registros em 3 grupos principais: evadidos (incluem formandos), regulares e desconsiderados.

Os evadidos representam o grupo dos registros, neste caso alunos, que iniciaram e concluíram seus estudos antes do marco. Suponha-se um aluno que tenha ingressado em 2001 e concluído em 2006. Sabendo que o marco foi definido como 2010/1, ou seja, ano de 2010 no primeiro semestre, tem-se que esse aluno em específico faz parte do grupo de evadidos/concluintes. A Figura 4.1 representa onde estão alocados os registros que fazem parte desse grupo, utilizando o marco como 2010/1.

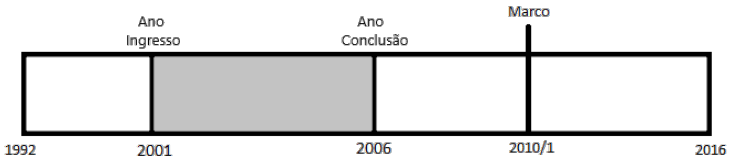

Figura 4.1: O grupo de alunos evadidos/concluintes.

Os registros rotulados como regulares são o principal objetivo desta abordagem utilizando marco, pois compõem o grupo de dados de validação do modelo. O grupo dos alunos regulares é composto por todos os estudantes que iniciaram seus estudos antes do marco, mas concluíram ou evadiram após o mesmo. Suponha-se que dado aluno ingressou em 2008 e evadiu em 2012. Sabendo que o marco estabelecido era 2010/1, esse aluno é designado ao grupo dos regulares, por ter iniciado antes de 2010 e concluído/evadido após 2010. A Figura 4.2 representa graficamente onde se alocam os registros componentes desse grupo.

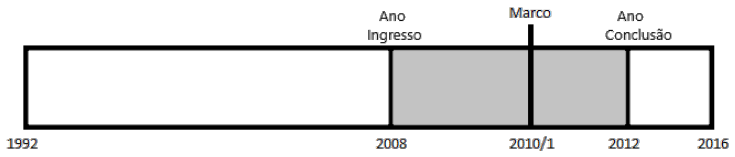

Figura 4.2: O grupo de alunos regulares.

Para estes alunos, a fórmula de cálculo do coeficiente (3.1) sofre algumas mudanças. A fim de considerar o conceito de barreira temporal, os períodos de conclusão das disciplinas que foram realizadas após o marco são substituídos pelo ano e período do marco. Sendo assim, a fórmula de cálculo para estes alunos é dada pela equação:

$$Coeficiente = 2 * (anoMarco - anoIngresso) + (periodoMarco - periodoIdeal) \quad (4.1)$$

Os demais alunos não são considerados. Sendo assim, o grupo de alunos desconsiderados contém todos os alunos que têm ingresso posterior ao marco. Suponha que um aluno ingressou em 2011, sabendo-se que o marco foi estabelecido em 2010/1. Esse aluno é desconsiderado pois iniciou seus estudos dois anos após o marco. A Figura 4.3 mostra os registros desse grupo.

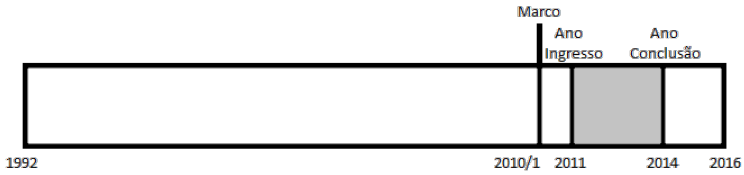

Figura 4.3: O grupo de alunos desconsiderados.

Concluindo, pode-se perceber que dois grupos são amplamente utilizados, evadidos e regulares. O grupo de evadidos constitui para o algoritmo de classificação o que é chamado de grupo de treinamento, enquanto o grupo de regulares é imprescindível, pois compõe o grupo de validação. O grupo de treinamento é aquele que é analisado pelo classificador a fim de construir o modelo de classificação, ao qual o grupo de validação é submetido.

4.2 EXPERIMENTOS

A condução dos experimentos se iniciou pela divisão dos registros entre o arquivo de treino e o arquivo de validação. Para tal, o desenvolvimento de um *script* que automatizasse o cálculo de coeficiente e a seguinte separação dos dados foi necessário. O seu funcionamento funcionamento é simples, e se inicia com a realização do cálculo de coeficiente para cada um dos registros. Posteriormente analisa-se a data de ingresso e de conclusão dos alunos para realizar a divisão baseando-se no conceito de marco explanado na seção anterior e realizando a operação de seleção e cálculo uma vez para cada ano que o marco varia.

Uma vez concluído esse processo, obtém-se como resultado um par de arquivos para cada ano do marco: um contendo dados de treino, o outro contendo dados de validação. Estes são carregados na ferramenta WEKA, de forma separada. Inicialmente carrega-se o arquivo referente ao treinamento, faz-se a escolha do algoritmo de classificação, bem como as suas configurações. Logo após, utiliza-se a opção *Supplied Test Set* da ferramenta para carregar o arquivo de validação.

O resultado deste processo é uma matriz de confusão, que apresenta os erros e acertos de classificação para o modelo em questão. Várias métricas de avaliação estão disponíveis também, tais como acurácia, precisão, cobertura, ROC, F-measure. Uma visão geral do processo de aquisição dos resultados é mostrada abaixo na Figura 4.4.

Figura 4.4: O processo de aquisição dos resultados. Fonte: Autor.

Outra informação que por vezes é útil e que está disponível ao fim do processo é a

visualização da árvore de decisão gerada. Para todos os algoritmos classificados como árvore de decisão - incluindo o LMT -, o WEKA gera a árvore de decisão correspondente ao modelo que foi induzido. Nesta árvore, costumam-se encontrar os padrões de decisão que guiam os registros até os nós folha, assim como a distribuição dos registros classificados.

4.2.1 Os *Datasets*

Após a conclusão da etapa de preprocessamento para cada ano do marco, dois *datasets* são gerados. Um deles corresponde aos dados de treino, sendo que o outro corresponde aos dados de teste. Existe, dessa forma, um par de *datasets* que representa o conjunto de dados utilizados para criação do modelo e validação dos resultados referente a cada ano do marco, variando este de 1992 a 2016.

Uma vez que a dimensionalidade é alta, devido à forma de preprocessamento, a quantidade de atributos é um fator relevante na modelagem. Modelos com mais atributos tendem a demorar mais tempo para serem construídos em comparação aos modelos menores. Outro fator importante é a quantidade de instâncias de treinamento, pois geralmente um modelo criado com mais instâncias de treino consegue se adequar melhor aos dados de validação. Além desses, o balanceamento de classes também costuma ter relevância. A Tabela 4.1 apresenta esses fatores para os cinco cursos que foram utilizados nos experimentos, com ano do marco 2016.

Tabela 4.1: Volume de dados para treinamento por curso para o ano de 2016.

Curso	**Nr. Atributos**	**Instâncias**	**Nr. Formados**	**Nr. Evadidos**
Zootecnia	315	845	586	259
Ciência da Computação	243	629	409	220
Administração Diurno	336	673	543	130
Administração Noturno	311	718	434	284
Pedagogia	225	789	522	267

A Tabela 4.2 apresenta os dados referentes às instâncias de teste, para os mesmos cinco cursos no ano de 2016. Pode-se perceber que, em comparação ao volume de dados da Tabela 4.1, este é bem menor.

Estes dados de validação trazem consigo uma informação importante, que comprova o alto índice de evasão. Pode-se perceber que no caso de Ciência da Computação, por exemplo, entre o início de 2016 e a metade de 2017, 36 alunos evadiram, sendo que por ano, iniciam os estudos 40 alunos.

Tabela 4.2: Volume de dados para teste por curso para o ano de 2016.

Curso	Instâncias	Nr. Formados	Nr. Evadidos
Zootecnia	43	28	13
Ciência da Computação	59	23	36
Administração Diurno	21	10	11
Administração Noturno	38	28	10
Pedagogia	25	3	22

4.3 RESULTADOS

Para este trabalho, são analisados dados de cinco cursos da UFSM: Administração Diurno, Administração Noturno, Ciência da Computação, Pedagogia e Zootecnia. Esses dados foram disponibilizados via SIE-UFSM, com apoio do CPD-UFSM. As métricas de avaliação utilizadas para medir o desempenho de classificação foram a cobertura e a precisão.

Percebe-se, na Figura 4.5, referente aos dados do curso de Administração Diurno, que a cobertura ficou em 100% para o curso de Administração Noturno. Isso quer dizer que quando se toma o marco no ano de 2016, por exemplo, dos 10 registros de alunos evadidos que existiam na base de validação, os 10 foram classificados corretamente. Ainda tomando o ano de 2016, podemos perceber a precisão em 90%, o que quer dizer nesse caso que dos 11 alunos que foram classificados como evadidos, 10 eram realmente evadidos e aquele que sobrou trata-se de um falso-positivo.

Sobre o gráfico da Figura 4.6, representando o curso de Administração Diurno, é possível visualizar que os índices altos de precisão e cobertura se mantêm. Pode-se perceber ainda que a precisão está sempre acima de 95%, o que reflete uma grande pureza nos dados, enquanto um nível acima de 87% na cobertura garante a identificação de quase todos os alunos com potencial para evasão.

É prudente ressaltar também que quanto mais antigo o ano do marco, maior a possibilidade de se estar tentando prever uma evasão que irá acontecer em 2 ou 3 anos. Por exemplo, ao tomar o marco como 2013, pode ser que se esteja tentando prever a evasão de um aluno que ingressou em 2012 e vai evadir apenas em 2015 ou 2016, ou seja, 2 a 3 anos depois. Este é um fator que aumenta o nível de dificuldade da classificação, de uma forma geral.

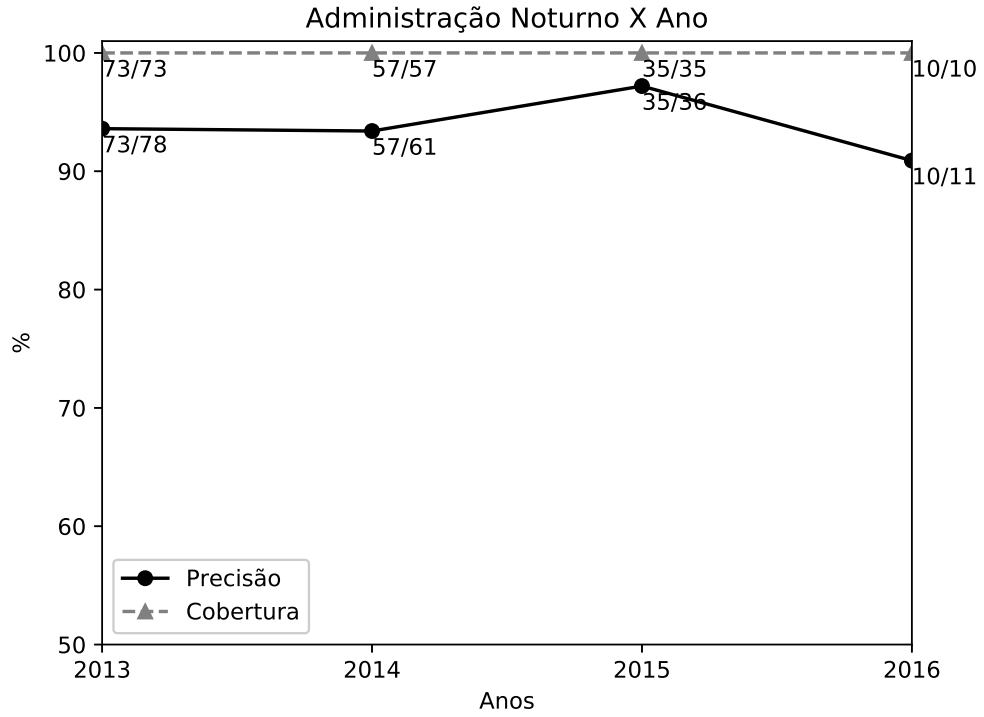

Figura 4.5: Precisão e cobertura sobre evadidos no curso de Administração Noturno.

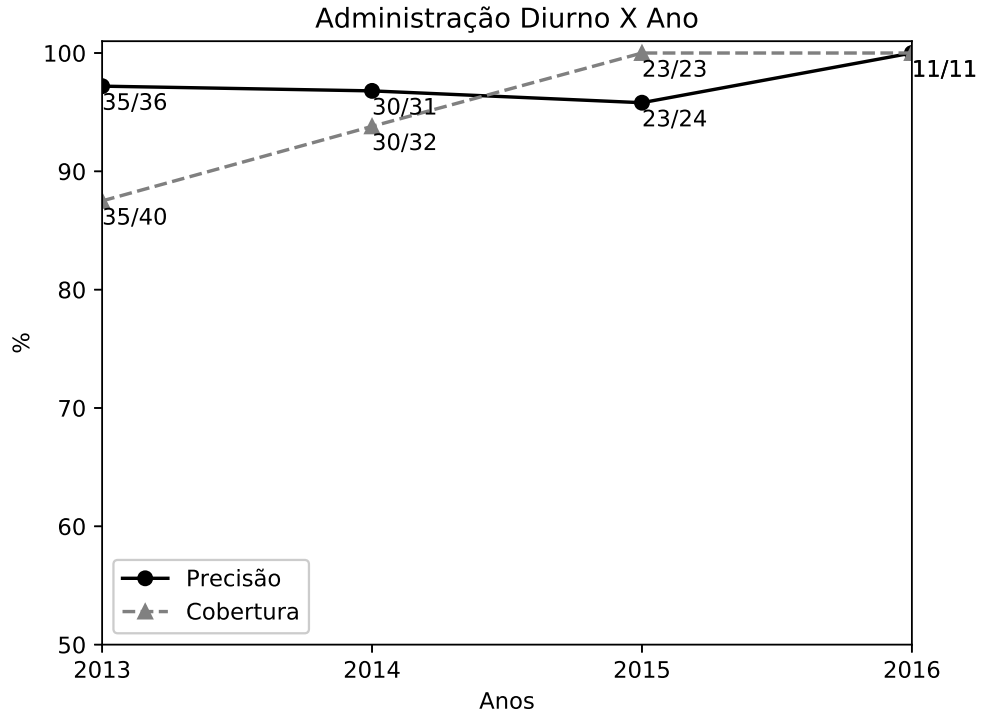

Figura 4.6: Precisão e cobertura sobre evadidos no curso de Administração Diurno.

A Figura 4.7 representa os resultados do modelo para o curso de Ciência da Computação. É interessante ressaltar aqui que a cobertura é muito alta, sempre acima dos 90%. Isso favorece, como supracitado, a identificação do maior número possível de alunos que realmente evadiram. Porém, o fato de a precisão ter níveis mais baixos em comparação aos gráficos da Figura 4.5 e 4.6 faz com que os dados fiquem um pouco menos puros.

O fato de os dados terem baixo nível de pureza, mas alta cobertura revela que os alunos de Ciência da Computação têm uma linha de diferenciação tênue entre evadidos e formados, ou seja, os alunos que se formam são muito parecidos, em relação à vida acadêmica, com aqueles que evadem.

Inclusive, o fato de a cobertura ser alta para evadidos nos revela, por meio da matriz de confusão, que a precisão para formados é alta. Observando o gráfico da Figura 4.7 e tomando o ano de 2013, por exemplo, vemos que 68 dos 98 alunos classificados estavam corretos. Isso quer dizer que 30 deles são na verdade alunos que graduaram, mas foram tomados pelo classificador com evadidos, muito provavelmente pelo fato de ser comum a realização de disciplinas com muitos semestres de atraso.

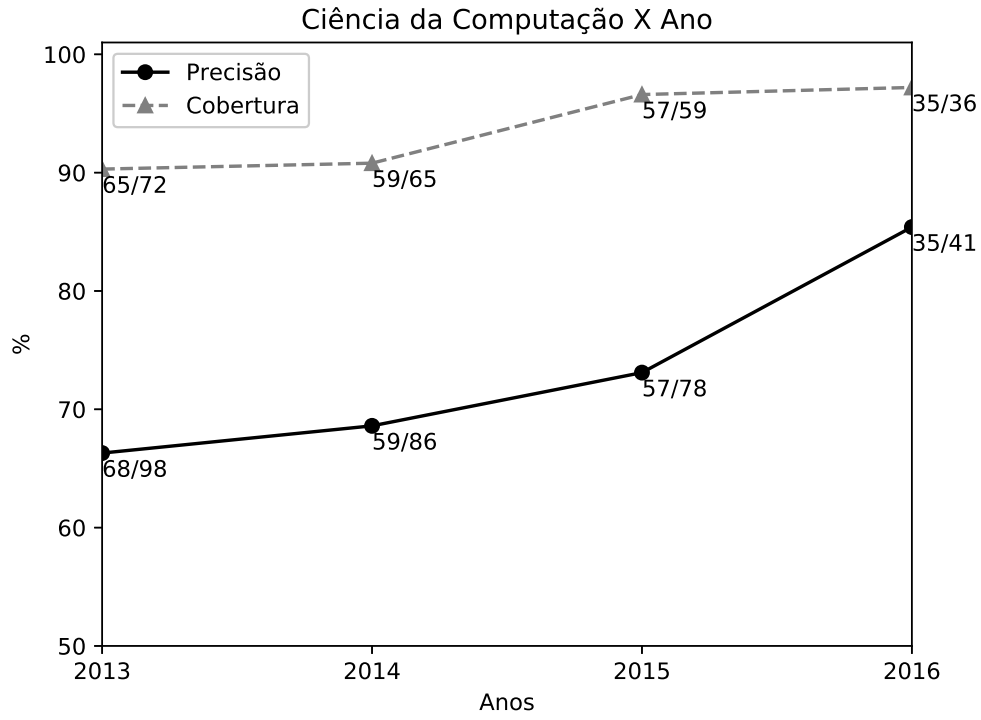

Figura 4.7: Precisão e cobertura sobre evadidos no curso de Ciência da Computação.

Para os cursos de Zootecnia e de Pedagogia, conforme as Figuras 4.8 e 4.9 respectiva-

mente, pode-se perceber uma diminuição leve nas taxas de desempenho do classificador. Muito dessa queda de desempenho se deve ao fato acima citado de que identificar possíveis evasões com 2 ou 3 anos de antecedência é uma tarefa um tanto complexa.

Além disso, Zootecnia e Pedagogia são cursos bem generalistas, com um vasto leque de opções para escolha de disciplinas. É curioso analisar que muitas das evasões desses dois cursos ocorrem logo no início da graduação, o que deixa os registros dos alunos com muito poucas conclusões de disciplinas, colaborando para a difícil identificação de evasões, principalmente quando não se tem muitos dados sobre o aluno em questão.

O fato de alunos desistirem em semestres iniciais e assim ficarem com poucas disciplinas concluídas também induz o modelo criado a classificar todos os alunos com poucos dados como evadidos, o que nem sempre é verdade, justificando dessa forma a grande quantidade de erros nos anos de 2013 e 2014.

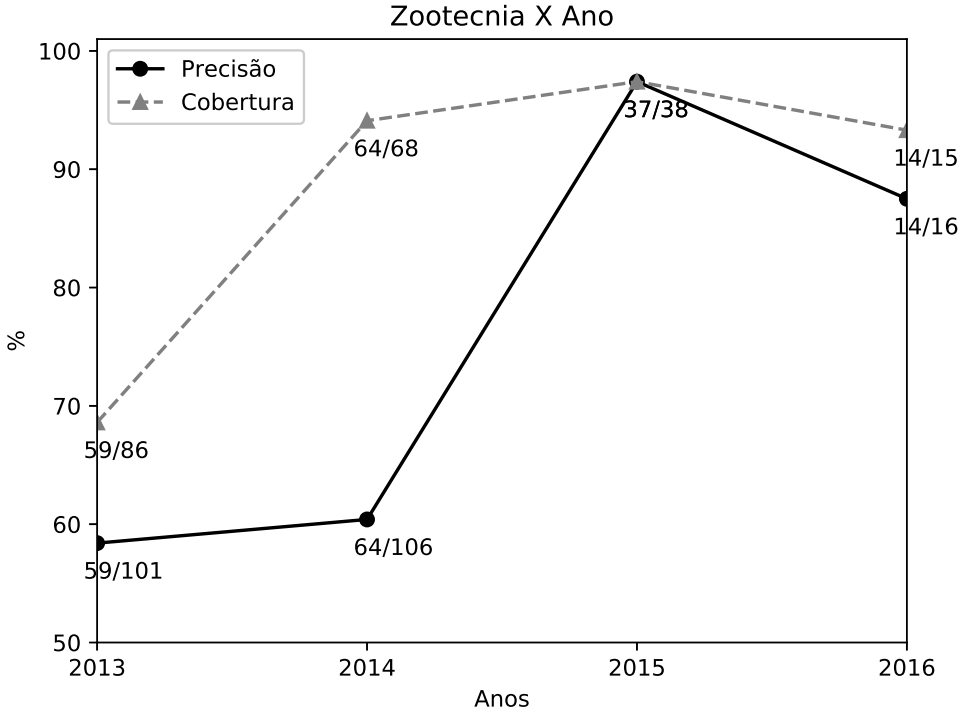

Figura 4.8: Precisão e cobertura sobre evadidos no curso de Zootecnia.

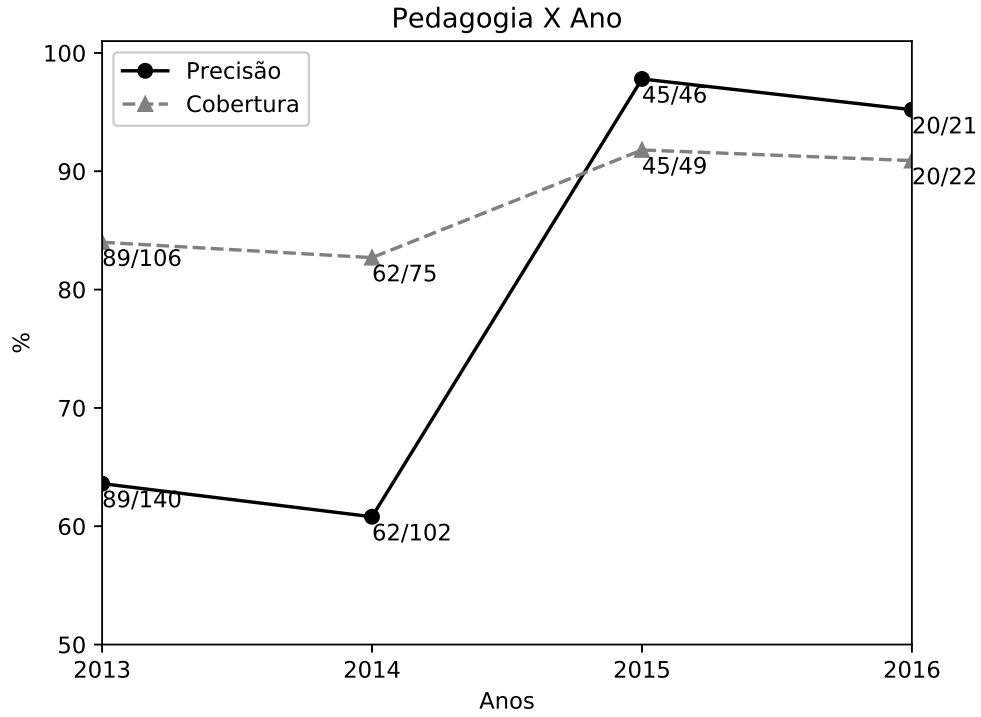

Figura 4.9: Precisão e cobertura sobre evadidos no curso de Pedagogia.

Como comparação de desempenho do modelo gerado, pode-se tomar o trabalho de KANTORSKI et al. (2016), realizado com dados dos cursos de Administração e de Zootecnia da UFSM, em que os classificadores testados chegaram a uma taxa de acerto na previsão de evadidos em torno de 70%. No presente trabalho, as taxas de precisão e cobertura encontram-se na grande maioria das vezes acima dos 70%, o que representa uma evolução na eficácia do modelo quando se faz uso do modelo de atributos e classificação proposto neste trabalho.

5 CONCLUSÃO

O modelo de dados e características proposto para este trabalho possuía a finalidade de identificar alunos com potencial para evasão. Para tal objetivo, foram utilizados dados referentes à aprovação dos alunos, transformados por meio de um *script* para que resultassem em apenas um registro por alunos, contendo um atributo por disciplina que o aluno poderia cursar. O modelo de classificação utilizado neste trabalho foi o LMT, que comprovou sua eficácia no que diz respeito a atributos faltantes e à redução do *overfitting*, conforme foi relatado LANDWEHR; HALL; FRANK (2005).

Os resultados obtidos são animadores, pois se percebe um grande número de acertos ao identificar evadidos. Na grande maioria dos casos, precisão e cobertura estão em níveis acima de 90% para os anos de 2015 e 2016, o que indica uma eficiência altíssima na previsão de evasão escolar em uma janela de tempo futuro que chega a 2 anos.

As taxas sempre muito altas de cobertura são entusiasmantes, independente do período de tempo analisado. Para o caso do problema de evasão escolar, especificamente, a métrica de cobertura é mais importante que a própria métrica de precisão. Isso porque para o caso da evasão importa principalmente a capacidade de detecção dos alunos com potencial. Se junto a estes aparecer algum aluno que não é considerado em risco, não há muita influência. Assim, essa necessidade é atendida em sua totalidade pelas boas taxas de cobertura, aliadas às razoáveis taxas de precisão.

Além do mais, a proposta de desenvolver um modelo de desempenho máximo, desconsiderando a alta dimensionalidade do problema, conseguiu realmente trazer os bons resultados esperados. Dada a grande dimensionalidade de alguns dos modelos, que chegam a quase 400 atributos por registro, uma pesquisa futura poderia concentrar-se em utilizar algum método de remoção de atributos menos impactantes, como forma de reduzir o tempo de processamento para criação do modelo.

Outra possibilidade de trabalho futuro é a análise mais aprofundada e criteriosa sobre os motivos da redução de desempenho em alguns casos específicos. Sabe-se que a identificação de alunos com potencial de evasão usando uma janela de futuro acima de 3 anos é de difícil solução, uma vez que alunos com este perfil tendem a apresentar dados escassos para uma avaliação mais concreta. Logo, uma forma de contornar esse problema é necessária para se alcançar um desempenho ainda melhor.

Uma outra proposta de sequência deste trabalho é a utilização de atributos com valores fixos a fim de melhorar o desempenho. Entende-se por atributo com valor fixo um atributo cujo valor não mude a cada mudança de marco, por exemplo. Os atributos que fazem parte do modelo criado neste projeto são totalmente mutáveis e voláteis; seu valor varia constantemente conforme se altera ano e período do marco. A utilização de valores fixos pode auxiliar na identificação de padrões pelo algoritmo de classificação.

REFERÊNCIAS

ANDREONI, M. **AN INTRUSION DETECTION AND PREVENTION ARCHITECTURE FOR SOFTWARE DEFINED NETWORKING**. 2014. 11p. — Mestrado em Engenharia Elétrica - Universidade Federal do Rio de Janeiro.

BREIMAN, L. et al. **Classification and regression trees**. [S.l.]: CRC press, 1984.

BROWNLEE, J. **An Introduction to Feature Selection**. https://machinelearningmastery.com/an-introduction-to-feature-selection/: [s.n.], 2014.

DASH, M.; LIU, H. Feature selection for classification. **Intelligent data analysis**, [S.l.], v.1, n.1-4, p.131–156, 1997.

DAVOK, D. F.; BERNARD, R. P. Avaliação dos índices de evasão nos cursos de graduação da Universidade do Estado de Santa Catarina–UDESC. **Avaliação: Revista da Avaliação da Educação Superior**, [S.l.], v.21, n.2, p.503–521, 2016.

DYKE, G.; PATTERSON, H. Analysis of factorial arrangements when the data are proportions. **Biometrics**, [S.l.], v.8, n.1, p.1–12, 1952.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, [S.l.], v.17, n.3, p.37, 1996.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive Logistic Regression: a statistical view of boosting. **Annals of Statistics**, [S.l.], v.28, p.2000, 1998.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining**: conceitos, técnicas, algoritmos, orientações e aplicações. 2.ed. [S.l.]: Elsevier, 2015.

GRIZZLE, J. E. A new method of testing hypotheses and estimating parameters for the logistic model. **Biometrics**, [S.l.], v.17, n.3, p.372–385, 1961.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, [S.l.], v.3, p.1157–1182, 2003.

KANTORSKI, G. et al. Predição da Evasão em Cursos de Graduação em Instituições Públicas. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE). **Anais...** [S.l.: s.n.], 2016. v.27, n.1, p.906.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial intelligence**, [S.l.], v.97, n.1-2, p.273–324, 1997.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. **Machine learning**, [S.l.], v.59, n.1-2, p.161–205, 2005.

LIU, H.; MOTODA, H. **Feature Selection for Knowledge Discovery and Data Mining**. 1.ed. [S.l.]: Springer US, 1998. (The Springer International Series in Engineering and Computer Science 454).

MITCHELL, T. M. **Machine Learning**. 1.ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: MIT press, 2012.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. **Sistemas Inteligentes. Rezende, SO Editora Manole Ltda**, [S.l.], p.115–140, 2003.

MORAIS, A. de et al. Assunção de responsabilidade e reflexão dirigida no curso de pedagogia: implicações para a adaptação e formação no ensino superior. **Educação Temática Digital**, [S.l.], v.19, n.2, p.482, 2017.

OLIVER, F. Notes on the logistic curve for human populations. **Journal of the Royal Statistical Society. Series A (General)**, [S.l.], p.359–363, 1982.

PEARL, R.; REED, L. J. On the rate of growth of the population of the United States since 1790 and its mathematical representation. **Proceedings of the national academy of sciences**, [S.l.], v.6, n.6, p.275–288, 1920.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, [S.l.], v.1, n.1, p.81–106, 1986.

QUINLAN, J. R. **C4. 5**: programs for machine learning. [S.l.]: Elsevier, 2014.

SHEARER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing**, [S.l.], v.5, n.4, p.13–22, 2000.

SILIPO, R. et al. **Seven Techniques for Data Dimensionality Reduction**. https://mineracaodedados.files.wordpress.com/2015/06/knime_seventechniquesdatadimreduction.pdf: [s.n.], 2015.

SILVA FILHO, R. L. L. et al. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, [S.l.], v.37, n.132, p.641–659, 2007.

TAYLOR, J. **Four Problems in Using CRISP-DM and How to Fix Them**. https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html: [s.n.], 2017.

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and information systems**, [S.l.], v.14, n.1, p.1–37, 2008.