



Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação
Programa de Pós-Graduação em Ciência da Computação

Abelardo Vieira Mota

Aplicação de aprendizado de máquina ao problema de evasão de discentes da UFC

Fortaleza
2015

Abelardo Vieira Mota

Aplicação de aprendizado de máquina ao problema de evasão de discentes da UFC

João Paulo Pordeus Gomes

Fortaleza

2015

Dedico este trabalho.

Essentially, all models are wrong, but some are useful.
(George E. P. Box)

Resumo

A evasão de discentes de cursos de ensino superior é um problema que tem recebido atenção do governo federal, dadas as altas taxas observadas nos últimos anos e por implicar em desperdício de recursos e de tempo. Esse problema tem sido objeto de estudos que buscam, a partir de dados, construir modelos de predição que possam ser utilizados para identificar discentes com grande probabilidade de abandonarem seus cursos, de forma que ações possam ser executadas a fim de diminuir essa probabilidade. Uma das ferramentas utilizadas para construção desses modelos é Aprendizado de Máquina, subárea de Inteligência Artificial composta por algoritmos e técnicas que permitem que um programa melhore sua performance a partir de dados. Este trabalho objetiva avaliar a aplicabilidade de Aprendizado de Máquina ao problema de evasão de discentes da UFC.

Palavras-chaves: Aprendizado de máquina. Evasão. Mineração de dados.

Abstract

The ideal abstract will be brief, limited to one paragraph and no more than six or seven sentences, to let readers scan it quickly for an overview of the paper's content.

Key-words: Machine Learning. Drop-out. Data Mining.

Lista de ilustrações

Figura 1 – Taxa de Sucesso na Graduação - UFC	18
Figura 2 – Fases do CRISP-DM	32
Figura 3 – Detalhamento da fase Análise de negócio	34
Figura 4 – Detalhamento da fase Análise dos dados	37
Figura 5 – Detalhamento da fase Preparação dos dados	39
Figura 6 – Detalhamento da fase Modelagem	42

Lista de tabelas

Tabela 1 – Taxa de sucesso da graduação por curso na UFC em 2013 - 10 piores resultados, em ordem crescente	18
---	----

Sumário

1	INTRODUÇÃO	17
1.1	O que é o problema	17
1.2	Soluções	19
1.3	Objetivos	20
2	APRENDIZADO DE MÁQUINA	21
2.1	Definição	21
2.2	Modelagem	21
2.3	Modelos	25
2.4	Aplicações	25
3	EVASÃO DE DISCENTES	27
3.1	Definição	27
3.2	Dados de ocorrência	27
3.3	Causas	27
3.4	Soluções	27
3.5	Aplicação de aprendizado de máquina	27
4	PONTOS DE PARTIDA	29
4.1	Desempenho no primeiro semestre	29
5	METODOLOGIA	31
5.1	Processo CRISP-DM 1.0	31
5.2	Ferramentas utilizadas	45
6	RESULTADOS	47
7	CONCLUSÃO	49
	REFERÊNCIAS	51

Todo list

como citar?	17
quando sai o próximo?	17
referencia	17
manual de indicadores do TCU, p.5	17
qual a definição?	17
referencia	18
outros tipos de ações - artigos do instituto Lobo	19
expandir	21
traduzir	21
discutir	22
qual a escolhida no Mitchell? Lembrar de apresentar o exemplo	22
definir o que quer dizer por modelo	25
melhorar	25
citar a coefficient for agreement for nominal scale	28
Educational Data Mining with Focus on Dropout Rates	28
Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação	28
Analysis of Student Data for Retention Using Data Mining Techniques	28
O capítulo todo é um resumo do CRISP-DM 1.0... fico dando (??)m todo parágrafo?	31
Por que um processo?	31
Por que CRISP-DM?	31
História do CRISP-DM	31
E hoje em dia?	31
figura com diagrama	31
Overview	31
traduzir ou não?	31
tabela com exemplos de valores para as dimensões -> p.7	32
figura com representação do processo	32
qual a tradução para deployment?	33
report techniques...	37

1 Introdução

1.1 O que é o problema

O problema de evasão de discente consiste no abandono, pelo discente, de um processo de aprendizado antes de sua conclusão. De acordo com o escopo do processo de aprendizado, essa definição pode ser detalhada: a evasão pode ser de um curso, de uma instituição de ensino, de cursos de uma determinada área, do sistema de ensino etc. Independente do escopo da definição de evasão, ela representa desperdício de recursos de todos os envolvidos: discente, docente, instituição de ensino e sociedade.

No Brasil, a redução da ocorrência desse problema faz parte de uma das diretrizes do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais(REUNI), instituído pelo DECRETO Nº 6.096, DE 24 DE ABRIL DE 2007

como citar?

:

I - redução das taxas de evasão, ocupação de vagas ociosas e aumento de vagas de ingresso, especialmente no período noturno;

Na Universidade Federal do Ceará(UFC), de acordo com o Anuário estatístico

quando sai o próximo?

de 2014, ano base 2013

referencia

, o indicador "Taxa de sucesso na graduação", definido como a proporção entre número de discentes diplomados e número de discentes ingressantes da graduação

manual de indicadores do TCU, p.5

, esteve em 2013 com o menor valor desde 2008(Figura 1). Já o indicador "Taxa de sucesso da graduação por curso", em 2013

qual a definição?

, possuiu valor mínimo igual a 6.8%, referente ao curso Ciências Sociais, habilitação em licenciatura(Tabela 1).



Figura 1 – Taxa de Sucesso na Graduação - UFC

Tabela 1 Taxa de sucesso da graduação por curso na UFC em 2013 - 10 piores resultados, em ordem crescente

Curso	Período	Taxa de Sucesso
Ciências Sociais - Licenciatura	Noturno	6.8%
Redes de Computadores - Quixadá	Noturno	13.3%
Geografia - Bacharelado	Diurno	15.3%
Letras - Português-Alemão	Diurno	17.6%
Engenharia Metalúrgica	Diurno	18.3%
Ciências Econômicas - Sobral	Noturno	20.9%
Sistemas de Informação - Quixadá	Diurno	22.0%
Filosofia - Bacharelado	Noturno	24.3%
Matemática - Bacharelado	Diurno	24.4%
Engenharia Elétrica - Sobral	Diurno	25.0%

Em DIRETRIZES GERAIS DO PROGRAMA DE APOIO A PLANOS DE REESTRUTURAÇÃO E EXPANSÃO DAS UNIVERSIDADES FEDERAIS REUNI p.4

referencia

é ressaltado que o indicador "Taxa de conclusão dos cursos de graduação", cuja definição é igual à do "Taxa de Sucesso na graduação":

(...) não expressa diretamente as taxas de sucesso observadas nos cursos da universidade, ainda que haja uma relação estreita com fenômenos de retenção e evasão.

Esse indicador não é sensível, por exemplo, à ocorrência de uma greve, fenômeno que causa atraso na formação dos discentes, podendo diminuir relevantemente a quantidade

de diplomados em determinado ano. Funciona bem considerando o modelo em que, necessariamente, após a diplomação de uma turma de discentes, turma de igual tamanho irá ingressar. A realidade mostra que o processo de diplomação, iniciado com o ingresso do discente e finalizado com a emissão e recebimento de seu diploma, é mais complexo. Se adicionarmos o indivíduo que assume o papel de discente, a análise das causas da evasão torna-se mais complexa, trazendo à tona dimensões como a social e a econômica, além da vida pré universidade do indivíduo. Em (??), por exemplo, informações sobre o trabalho do discente e se ele é casado foram os fatores mais relevantes na classificação dos discentes analisados com relação à evasão do curso.

1.2 Soluções

outros tipos de ações - artigos do instituto Lobo

Para diminuir as taxas de evasão, uma das estratégias adotadas é a identificação precoce de discentes com grande tendência para abandonarem seus cursos e a execução de ações que minimizem tal tendência. A identificação pode ser conduzida por observação do comportamento e resultados dos discentes, de forma subjetiva, pelos docentes e coordenadores de cursos, por exemplo. Em estudo realizado no departamento de engenharia elétrica da Eindhoven University of Technology (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009), é relatado que em dezembro os discentes desse departamento recebem um aviso informando se são ou não aconselhados a continuarem no curso. Esse aviso é baseado na performance do discente no curso e em informações obtidas de professores do primeiro semestre e de discentes monitores. É relatado que o aviso parece ter bastante acurácia: geralmente discentes aconselhados a continuarem têm sucesso no próximo ano do curso, enquanto aqueles desaconselhados geralmente não continuam no curso. Dois problemas decorrem dessa forma de identificação: sendo conduzida por pessoas, essa forma de identificação é limitada pelo conjunto de observações as quais o observador tem acesso; sendo subjetiva, seus resultados podem sofrer resistência para serem aceitos. A utilização de técnicas de aprendizado de máquina como forma de identificação pode contornar esses problemas, por, primeiro, fazer uso de dados registrados por sistemas de informação, provavelmente contendo informações mais amplas que as que uma pessoa pode observar; segundo, por fazer maior uso de dados registrados, sendo aceita mais facilmente como identificação objetiva. Nesse estudo foram utilizados diversos algoritmos de aprendizado de máquina com o objetivo de tentar detectar que um estudante irá abandonar seu curso. Foram utilizadas informações de discente referentes tanto ao período anterior ao seu ingresso na universidade, quanto ao posterior.

Tendo a UFC como escopo o estudo (??) foi desenvolvido e objetivou entender o problema da evasão de discentes na UFC a partir da opinião de discentes que abandonaram seus cursos. A pesquisa foi realizada com 86 discentes que evadiram entre os anos 1999 e 2000. Seguem os motivos e as frequências com que foram apresentados:

- Incompatibilidade entre horários de trabalho e de estudo - 39,4%
- Aspectos familiares e desmotivação com os estudos - 20%
- Precariedade das condições físicas do curso ou inadequação curricular - 10%

1.3 Objetivos

O presente trabalho objetiva avaliar a aplicabilidade de técnicas de aprendizado de máquina ao problema de evasão de discentes na UFC, utilizando os dados que seus sistemas de informação, como o Sistema Integrado de Gestão de Atividades Acadêmicas(SIGAA), gerenciam.

2 Aprendizado de máquina

2.1 Definição

Aprendizado de máquina é uma subárea de Inteligência Artificial que agrupa conhecimentos sobre algoritmos e técnicas que permitam que um programa melhore sua performance a partir de dados. Mais formalmente:

Um programa aprende a partir de uma experiência E, com relação a uma classe de tarefas T e a uma medida de performance P, se sua performance em tarefas da classe T, medida por P, melhora com a experiência E. (MITCHELL, 1997, p.2, tradução nossa)

expandir

2.2 Modelagem

(MITCHELL, 1997), para exemplificar a modelagem de um programa com uma abordagem de aprendizado de máquina, apresenta uma sequência de passos para desenvolver um programa que aprenda a jogar xadrez, a ser utilizado para disputar um campeonato mundial de xadrez.

Escolha da experiência

O primeiro passo é escolher a experiência a partir da qual o programa irá aprender, denominada experiência de treinamento. (MITCHELL, 1997) classifica os tipos de experiência a partir de três atributos: feedback, se direto ou indireto com relação a como o programa será utilizado; nível de controle que há sobre a experiência; e representatividade. É ressaltado que o tipo de experiência utilizada pelo programa pode ter impacto significativo no sucesso ou falha em seu aprendizado.

O atributo feedback representa quão direta é a informação fornecida pela experiência para o problema em questão. Com relação a esse atributo, o elemento experiência pode ser classificado como experiência de feedback direto e experiência de feedback indireto. Por exemplo, a tupla estado do tabuleiro e melhor movimento possível a partir desse estado é classificada como experiência de feedback direto: o programa irá atuar realizando movimentos e esse tipo de experiência informa diretamente qual o melhor movimento. Já a tupla sequência de movimentos de uma partida e seu resultado final é classificada como experiência de feedback indireto: o resultado final da partida não fornece informação direta sobre a qualidade dos movimentos que nela foram executados. A atividade de determinar o grau de influência que elementos de uma experiência de feedback indireto têm sobre o resultado é denominada credit assignment.

traduzir

O atributo nível de controle representa quanto de controle é possível ter sobre a captura da experiência. Com relação a esse atributo, o elemento experiência pode

ser classificado como experiência selecionada por especialista, experiência sugerida pelo programa e analisada por um especialista e experiência selecionada e analisada pelo programa. Por exemplo, a experiência será do tipo selecionada por especialista se houve um jogador experiente de xadrez que selecionou estados de tabuleiro e indicou que melhores movimentos poderão ser feitos; será do tipo sugerida pelo programa e analisada por um especialista se o próprio programa selecionar estados de tabuleiro para serem analisadas por um jogador experiente de xadrez; será do tipo selecionada e analisada pelo programa se o programa utilizar o resultado de partidas que disputar consigo mesmo.

O atributo representatividade indica quão bem a experiência reflete a realidade. Com relação a esse atributo, o elemento experiência pode ser classificado em representativo, se sua distribuição representar a distribuição dos exemplos com os quais o programa efetivamente será utilizado, e não representativo, caso contrário. Por exemplo, a experiência não irá representar a realidade caso esteja limitada ao conjunto de partidas de apenas um jogador: considerando que o programa será utilizado em um campeonato mundial, do qual participam jogadores diversos, com estilos de jogo diversos, é capaz de o programa, treinado com essa experiência, depare com estados de tabuleiro que não encontrou no treinamento. (MITCHELL, 1997) ressalta que muito da teoria de aprendizado de máquina depende da assunção de que a experiência utilizada no treinamento reflete a realidade.

Escolha da função alvo

O próximo passo é a escolha do tipo de conhecimento que deverá ser aprendido, representado por uma função denominada função alvo, e como ele será utilizado pelo programa. Considerando que o programa irá atuar como um jogador de xadrez, uma possível função a ser considerada é uma cujo domínio seja o conjunto de estados de tabuleiro e que retorne o melhor movimento a partir do estado de tabuleiro informado. Esse tipo de conhecimento depende da assunção de que, dado um estado de tabuleiro, existe um melhor movimento a ser executado. O problema de aprendizado dessa função depende, portanto, do problema de determinar quão um movimento influencia no resultado final de uma partida.

discutir

qual a escolhida no Mitchell? Lembrar de apresentar o exemplo

Outro tipo de conhecimento é uma função que tenha como domínio o conjunto de estados de tabuleiro e retorne um número real, indicando quão bom o estado de tabuleiro informado é.

O programa irá jogar verificando qual estado de tabuleiro maximiza o valor da função, considerando o conjunto de estados de tabuleiro que podem ser alcançados a partir do estado atual do tabuleiro e de todas jogadas válidas.

Escolha de uma representação para a função alvo

Após a escolha do tipo de conhecimento que deverá ser aprendido, é necessário definir como esse conhecimento será representado. A função que associa um estado de tabuleiro a um número real pode assumir diversas formas: pode ser uma matriz contendo uma célula com um número real para cada estado de tabuleiro possível; pode ser um conjunto de regras que associe atributos do estado do tabuleiro a números reais; pode ser

uma função polinomial de atributos do estado do tabuleiro em números reais etc. Para darmos continuidade ao detalhamento dos passos, escolhemos aqui uma representação de função simples: denominaremos por V a função que associa um estado de tabuleiro a um número real, calculada como combinação linear dos seguintes atributos do estado do tabuleiro:

- x_1 : número de peças pretas no tabuleiro
- x_2 : número de peças brancas no tabuleiro
- x_3 : número de reis pretos no tabuleiro
- x_4 : número de reis brancos no tabuleiro
- x_5 : número de peças pretas ameaçadas por peças brancas no tabuleiro
- x_6 : número de peças brancas ameaçadas por peças pretas no tabuleiro

A função V pode ser representada então por:

$$V(t) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6$$

Os coeficientes w_0 a w_6 são parâmetros do programa, a serem ajustados no aprendizado. Resumindo os passos até aqui realizados, temos:

- Tarefa: jogar xadrez
- Medida de performance: proporção de jogos do campeonato mundial de xadrez ganhos
- Experiência de treinamento: partidas disputadas pelo programa contra ele mesmo
- Função alvo: $V : EstadosDeTabuleiro \rightarrow \mathbb{R}$
- Representação da função alvo: $V(t) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6$

Escolha de um algoritmo de aproximação

O próximo passo consiste na escolha de um algoritmo que, a partir de um conjunto de experiências, irá ajustar os parâmetros da representação da função a fim de aproximá-la da função alvo. Para tanto, é necessário um conjunto de dados de treinamento, composto por um par de estado de tabuleiro e o valor que a função deverá atribuir. Denotamos por $\langle b, V_{treino}(b) \rangle$ um dado para treinamento: b é uma tupla contendo os atributos de um estado de tabuleiro e $V_{treino}(b)$ o valor a ele atribuído. Por exemplo:

$$\langle \langle x_1 = 3, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 0 \rangle, +100 \rangle$$

Determinar os valores dos estados de tabuleiros utilizados no treinamento é tarefa fácil para os estados finais: pode-se definir dois valores, $max < min$, e atribuir aos estados finais de vitória o valor max e aos de derrota o valor min . Já a determinação dos valores dos tabuleiros intermediários não é tão simples. O fato de uma partida ter sido ganha não implica necessariamente que todos os estados de tabuleiro nela percorridos devem

receber um valor alto. Para definir tais valores pode-se utilizar uma regra de estimação. Por exemplo:

$$V_{treino}(b) \leftarrow V(sucessor(b)) \quad (2.1)$$

Essa regra atribui a um estado de tabuleiro de treinamento o valor que a função alvo estimada retorna para o estado de tabuleiro após a próxima jogada do oponente. Apesar de parecer estranho fazer uso de V , a função que se está estimando, para determinar os valores a serem utilizados para refiná-la, intuitivamente essa abordagem parece fazer sentido, por atribuir a um estado de tabuleiro um valor que é função de um estado que foi possível alcançar a partir dele.

Após a determinação dos valores dos estados de tabuleiros contidos na experiência de treinamento, o conjunto de dados de treinamento pode ser utilizado. Para estimar a função alvo a partir dos dados de treinamento, dada a representação da função escolhida, é necessário agora determinarmos seus pesos, w_0 a w_6 . Para tanto, primeiro definimos como mensurar quão bem a função estimada se adequa aos dados de treinamento. Uma medida comum é o erro quadrático, assim definido:

$$E = \sum_{(b, V_{treino}(b)) \in \text{dados de treinamento}} (V_{treino}(b) - V(b))^2$$

O problema de estimar a função alvo pode ser modelado então como o problema de encontrar os pesos w_0 a w_6 que minimizem o erro quadrático sobre os dados de treinamento. Um dos algoritmos que incrementalmente ajusta os pesos aos dados de treinamento, minimizando o erro quadrático é o Método dos Mínimos Quadrados. Esse algoritmo funciona ajustando, para cada dado de treinamento, os pesos da função na direção que minimiza o erro quadrático. Para tanto, atualiza iterativamente, para cada dado de treinamento, os pesos da função, incrementando com um valor proporcional a $(V_{treino}(b) - V(b))$: se o valor da função aplicado ao dado de treinamento, $V(b)$, é igual ao valor do dado de treinamento, $V_{treino}(b)$, o incremento será nulo; se o valor da função é maior que o do dado de treinamento, o incremento será negativo, fazendo com que, para o dado de treinamento em questão, o valor da função diminua; se o valor da função é menor que o do dado de treinamento, o incremento será positivo, fazendo com que, para o dado de treinamento em questão, o valor da função aumente.

Conclusão

Nessa sequência de passos podem ser identificados quatro elementos de um programa que aprende:

- Sistema de performance: elemento responsável pela utilização do conhecimento aprendido para resolver uma tarefa. No exemplo, será responsável por determinar qual a próxima jogada, dado um estado de tabuleiro, utilizando a função que foi aprendida.
- Crítico: elemento responsável por receber como entrada um dado de treinamento e informar a que valor deve ser associado. No exemplo, o crítico é representado pela equação 2.1.

- Generalizador: elemento responsável por receber como entrada um conjunto de dados de treinamento e retornar uma função estimativa de uma função alvo. No exemplo, o generalizador é o Método dos Mínimos Quadrados.
- Representação da função alvo: elemento que define a estrutura da função que será utilizada como estimativa da função alvo. No exemplo, foi utilizada como representação uma combinação linear.

Outras configurações para esses elementos foram desenvolvidas. Por exemplo, como representação da função alvo pode-se utilizar um grafo em estrutura de árvore, denominado árvore de decisão. Cada nó seu que não seja folha possui uma regra que associa um dado a um de seus nós filhos. Os nós folhas são associados a um valor. Seu funcionamento consiste em apresentar um dado à regra de um nó, inicialmente o nó raiz, e recursivamente aplicar esse procedimento ao nó filho ao qual a regra associa o dado, até que seja alcançado um nó folha, cujo resultado associado é então retornado como o valor da função.

2.3 Modelos

definir o que quer dizer por modelo

representation + evaluation + optimization ([DOMINGOS, 2012](#))

2.4 Aplicações

melhorar

De acordo com (??), o conhecimento sobre aprendizado de máquina pode ser aplicado a diversas áreas, como, por exemplo, a reconhecimento de voz; a visão computacional, sendo utilizado no desenvolvimento de sistemas de reconhecimento facial; a controle de robôs.

3 Evasão de discentes

3.1 Definição

3.2 Dados de ocorrência

3.3 Causas

3.4 Soluções

3.5 Aplicação de aprendizado de máquina

Trabalhos relacionados

Em (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009) são aplicados algoritmos de aprendizado de máquina a dados de discentes do Electrical Engineering department, Eindhoven University of Technology, considerando o período de 2000 a 2009, com o objetivo de identificar discentes em grupos de risco de evasão. É relatado que esse departamento já avaliava os discentes com relação ao risco de evasão, mas de forma subjetiva. Os dados dos discentes são particionados em pré universidade e pós universidade, gerando três bases de dados de treinamento, a primeira consistindo nos dados pré universidade, a segunda nos dados pós universidade e a terceira com todos os dados. São utilizados os algoritmos OneRule, CART, C4.5, BayesNet, SimpleLogistic, JRip e Random Forest. É apresentado um resultado de 68% de acurácia com o algoritmo OneRule aplicado à primeira base de dados, sem diferença significativa na performance dos demais algoritmos. O mesmo resultado repete-se com as demais bases, mudando apenas o valor da acurácia alcançada, sendo igual a 76% para a segunda base de dados e 75% para a terceira. O estudo ressalta o maior custo da ocorrência de falsos negativos que de falsos positivos na identificação de discentes com risco de evasão. Ocorre que, argumenta-se, há prejuízo maior em não oferecer apoio a um discente com risco de evasão do que oferecer, desnecessariamente, apoio a um discente sem tal risco. O estudo faz uso então de uma matriz de custo, com o algoritmo CostSensitiveClassifier, obtendo melhores diminuição na ocorrência de falsos negativos, mas com perdas de acurácia.

Em (MANHÃES; CRUZ; ZIMBRÃO,) são aplicados algoritmos de aprendizado de máquina a dados de discentes de seis cursos da Universidade Federal do Rio de Janeiro(UFRJ), com o objetivo de identificar discentes que não terão pelo menos uma aprovação no segundo semestre de seus cursos. Os cursos considerados foram: Direito, Farmácia, Física, Engenharia Civil, Engenharia Mecânica, Engenharia de Produção. É indicado que esses cursos foram escolhidos por pertencerem a departamentos distintos, com perfis de discentes distintos. É observado também que tais cursos diferem com relação à quantidade de discentes ingressantes, à taxa de evasão registrada e à efetividade de certas práticas de ensino. Para cada curso são desenvolvidas uma base de dados de treinamento e uma base de dados de teste, composta pelos dados de seus discentes de primeiro semestre. Para a base de dados de treinamento foram utilizados os dados

dos anos pares(de 1994.1 a 2008.1), já para a de teste foram utilizados os dados dos anos ímpares(de 1995.1 a 2009.1). O algoritmo de aprendizado utilizado foi o Naïve Bayes. São apresentados os resultados utilizando as medidas acurácia, taxa de verdadeiros positivos, taxa de verdadeiros negativos e Kappa

citar a coefficient for agreement for nominal scale

. A acurácia, por exemplo, varia de 70% a 100% entre as bases nas quais o modelo desenvolvido foi testado.

Os estudos analisados fizeram uso de apenas uma definição de evasão, a evasão do curso, utilizando experiência de feedback indireto. O uso de experiência de feedback direto, considerando a definição de evasão no curso, torna mais complexa a coleta de dados para treinamento: considerando que um curso possa ser concluído com uma duração máxima de 10 anos, por exemplo, apenas após 10 anos do ingresso de um discente é que seus dados poderão ser utilizados. Outros fatores podem afetar esse prazo, como o trancamento do curso, a ocorrência de greves etc.

Considero que os estudos analisados não realizaram um estudo mais criterioso sobre o problema em questão, focando os esforços mais na utilização de algoritmos de aprendizado de máquina que na análise do problema, de suas diversas definições, dos atributos utilizados, de como utilizar os resultados obtidos para diminuir o problema etc.

(VILLWOCK; APPIO; ANDRETA, 2015)

Educational Data Mining with Focus on Dropout Rates

(MANHÃES et al., 2012)

Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação

(SHERRILL; EBERLE; TALBERT, 2011)

Analysis of Student Data for Retention Using Data Mining Techniques

Taxonomia de features

Modelos de aprendizado de máquina

Avaliação dos resultados

Ferramentas

Resultados

Conclusões

4 Pontos de partida

Como pontos de partida, consideraremos:

- Analisar o problema de predição de discente com alta probabilidade de evasão do curso a partir de dados referentes a sua vida pré universidade e dados de seu desempenho no primeiro semestre do curso.
- Analisar as múltiplas definições de evasão de discentes, suas peculiaridades, as causas atribuídas pela literatura e a aplicabilidade de técnicas de aprendizado de máquina.
- Analisar o problema da predição do desempenho de um discente em uma disciplina a partir de dados sobre seu histórico como discente. Esse problema é relevante àqueles interessados na melhoria do desempenho de discentes em uma determinada disciplina a partir de atividades a serem iniciadas antes de o discente efetivamente cursá-la.
- Analisar o problema da predição da taxa de evasão de um curso a partir de seus dados, como, por exemplo, sua estrutura curricular. Esse problema é relevante ao processo de desenho ou redesenho de um curso, sua solução podendo ser utilizado como guia para decisões acerca do curso.

4.1 Desempenho no primeiro semestre

Especificação do problema

Benefícios

5 Metodologia

O capítulo todo é um resumo do CRISP-DM 1.0... fico dando (??)m todo parágrafo?

5.1 Processo CRISP-DM 1.0

Por que um processo?

(DOMINGOS, 2012) (DRUMMOND, 2008) (SCULLEY et al., 2014) (DRUMMOND, 2009)

Por que CRISP-DM?

(AZEVEDO, 2008)

História do CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) é um processo de aplicação de Mineração de Dados, desenvolvido pelo CRISP-DM Special Interest Group e publicado em 2000. Foi concebido em 1996 por três empresas que utilizavam Mineração de Dados: DaimlerChrysler (à época Daimler-Benz), SPSS (à época ISL) e NCR; motivadas pela incerteza com relação à qualidade de seus trabalhos, pelo questionamento de se toda nova empresa que quisesse aplicar Mineração de Dados terá que passar pelo aprendizado que passaram, baseado em tentativa e erro, e como garantir, para seus clientes, que Mineração de Dados era uma área suficientemente madura para ser incorporada a seus processos de negócio. Em 1999 foi publicado um draft do CRISP-DM versão 1.0, sendo aplicado pela DaimlerChrysler, SPSS e NCR a vários tipos de aplicações, indústrias e problemas de negócio, sendo considerado, então, validado suficientemente para ser publicado e distribuído (CHAPMAN et al., 2000).

E hoje em dia?

figura com diagrama

Overview

CRISP-DM segue uma estrutura hierárquica, composta de quatro níveis de abstração (do mais genérico ao mais específico): fase, tarefa genérica, tarefa especializada e instância de processo. Os dois primeiros níveis, fase e tarefa genérica, foram modelados a fim de serem genéricos o suficiente para atenderem às todas aplicações de Mineração de Dados; completos, abrangendo todo o processo de Mineração de Dados; e estáveis, sendo aplicáveis tanto para as técnicas de Mineração de Dados existentes, quanto às que venham a ser desenvolvidas. O terceiro nível, tarefa especializada, é composto pelas tarefas a serem executadas em situações específicas para alcançar os objetivos das tarefas genéricas. Exemplificando, seja a tarefa genérica Limpar dados: a ela relacionadas estão as tarefas especializadas Limpar dados numéricos e Limpar dados categóricos. O quarto nível, instância de processo, é composto pelos registros de ações, decisões e resultados de uma

traduzir
ou
não?

execução do processo.(CHAPMAN et al., 2000)

Apesar de a representação do processo sugerir que ele é composto por uma sequência fixa de fases, na prática as tarefas podem ser executadas seguindo outras ordens: é o caso de, por exemplo, na tarefa Avaliação do modelo ser verificado que são necessários mais dados, a serem adquiridos através de tarefas que, de acordo com o diagrama, já foram executadas.

Para o mapeamento do modelo em uma instância do processo, a especificação do CRISP-DM 1.0 identifica como relevantes quatro dimensões do contexto de Mineração de Dados: domínio de aplicação, tipo de problema de Mineração de Dados, aspectos técnicos e ferramentas e técnicas. Os valores dessas dimensões são utilizados nas decisões sobre que tarefas específicas podem ou devem ser executadas.

O processo de mapeamento do CRISP-DM a uma instância do processo é, de acordo com o CRISP-DM 1.0, composto pelas etapas:

1. Analisar o contexto, identificando os valores para as dimensões domínio de aplicação, tipo de problema de Mineração de Dados, aspectos técnicos e ferramentas e técnicas
2. Remover do modelo CRISP-DM os detalhes não aplicáveis ao contexto analisado
3. Adicionar detalhes específicos do contexto analisado
4. Especializar(ou instanciar) elementos genéricos do modelo de acordo com características concretas do contexto
5. Possivelmente renomear elementos genéricos do modelo a fim de tornar mais explícito seu significado, de acordo com o contexto

tabela com exemplos de valores para as dimensões -> p.7

figura com representação do processo



Figura 2 – Fases do CRISP-DM

A seguir segue uma descrição breve de cada uma das fases:

1. **Análise do negócio:** O objetivo desta fase é entender os requisitos e objetivos do projeto sob uma perspectiva de negócio, traduzí-los para requisitos e objetivos sob uma perspectiva de Mineração de Dados e então traçar um plano preliminar para alcançá-los.
2. **Análise dos dados:** Esta fase inicia com uma coleta inicial de dados e segue para o estudo dos dados a fim de identificar problemas de qualidade, obter insights e detectar possíveis subconjuntos de dados que permitam desenvolver hipóteses sobre informações que não estejam presentes.

3. **Preparação dos dados:** Esta fase é composta por atividades necessárias para gerar, a partir dos dados inicialmente coletados, um conjunto de dados a ser utilizado pelas ferramentas de modelagem. Inclui atividades como seleção de tabelas, registros, atributos, transformações e limpeza de dados.
4. **Modelagem:** Nesta fase são aplicadas técnicas de modelagem e os modelos desenvolvidos são otimizados. Normalmente aplicam-se ao problema mais de uma técnicas de modelagem. Como algumas técnicas de modelagem podem exigir que os dados estejam em dado formato, pode ser necessário voltar para a fase Preparação dos dados.
5. **Avaliação:** Esta fase é iniciada quando já foi desenvolvido um modelo com alta qualidade, do ponto de vista da Mineração de Dados. Nela são avaliados a adequação do modelo como ferramenta para alcançar o objetivo de negócio que motivou o projeto e a qualidade da instância do processo. Ela termina com a decisão pela utilização ou não dos resultados obtidos.
6. **Instalação:** Após o desenvolvimento de um modelo, faz-se necessário que ele seja disponibilizado para os usuários finais, seja na forma de relatórios, seja na forma de sistemas de apoio à tomada de decisão, para que seja efetivamente utilizado, auxiliando no alcance dos objetivos de negócio que motivaram a criação do projeto.

qual a
tradução
para
deployme

1 - Análise do negócio

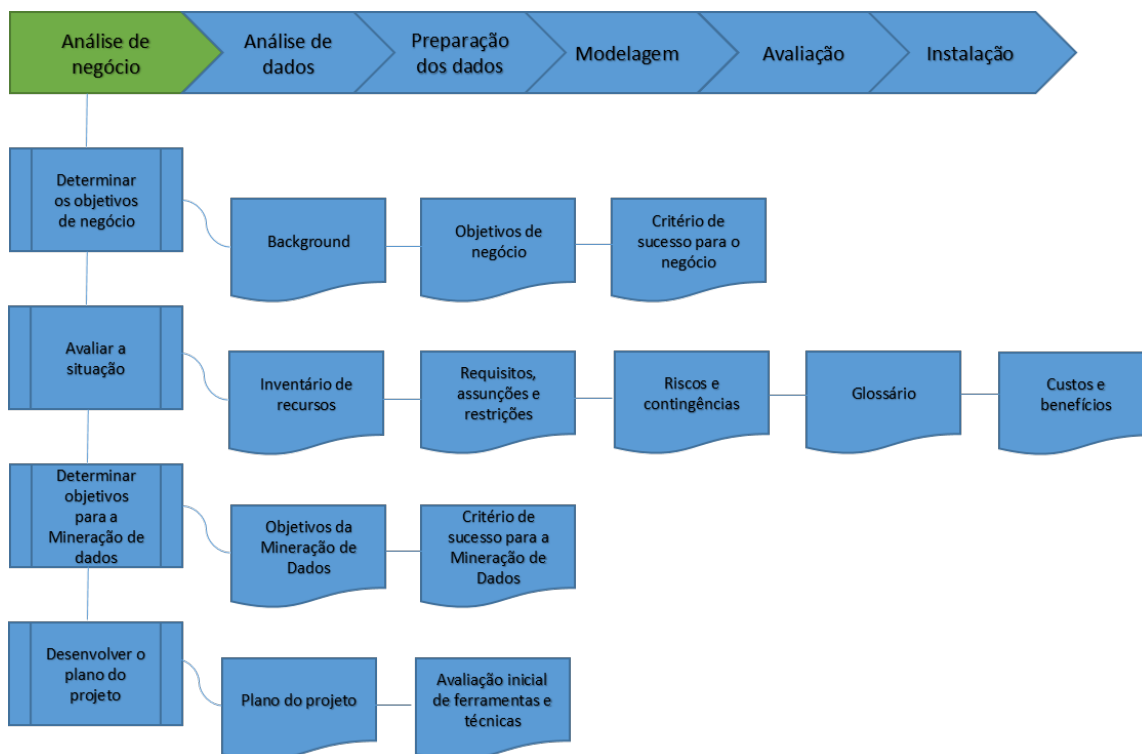


Figura 3 – Detalhamento da fase Análise de negócio

1.1 - Analisar os objetivos de negócio

O primeiro passo em um projeto de Mineração de Dados é analisar, sob uma perspectiva de negócio, o que realmente o cliente deseja alcançar. Normalmente o cliente possui vários objetivos concorrentes e restrições, que devem ser balanceados. O objetivo desta tarefa é descobrir fatores importantes do negócio que possam influenciar no resultado do projeto. Uma das consequências de se negligenciar esse passo é o projeto finalizar respondendo corretamente a perguntas erradas.

Background

Registro das informações sobre o estado do negócio no início do projeto.

Objetivos do negócio

Registra o objetivo primário do projeto, sob uma perspectiva de negócio, além de outras questões de negócio que o cliente objetiva esclarecer.

Critério de sucesso, sob uma perspectiva de negócio

Registra os critérios para que o projeto seja considerado um sucesso, sob uma perspectiva de negócio.

1.2 - Avaliar a situação

O objetivo desta tarefa é analisar mais detalhadamente de informações importantes para determinar os objetivos da Mineração de Dados e desenvolver um plano para o projeto. São analisadas informações como quais os recursos disponíveis, quais as restrições e quais as suposições.

Inventário de recursos

Registra os recursos disponíveis para o projeto, como recursos humanos(especialistas do negócio, analistas de dados, técnicos de suporte), dados(arquivos, bases de dados operacionais, data warehouses), hardwares e softwares.

Requisitos, suposições e restrições

Registra os requisitos do projeto, incluindo prazos, níveis de qualidade, segurança e aspectos legais. Registra as suposições do projeto, sejam suposições que poderão ser verificadas a partir dos dados utilizados pelo projeto, sejam suposições que não poderão ser verificadas. No caso de suposições que não poderão ser verificadas, é importante que sejam registradas, principalmente caso possam afetar a validade dos resultados do projeto. Registra as restrições do projeto, sejam restrições na disponibilidade de recursos, sejam restrições tecnológicas.

Riscos e planos de contingência

Registra os eventos que, caso ocorram, poderão afetar os prazos ou a qualidade do projeto, bem como os planos de contingência, detalhando que ações devem ser executadas caso esses eventos ocorram.

Glossário

Registra o conjunto de termos e seus significados que são relevantes para o projeto. Inclui tanto termos pertencentes à terminologia do negócio, quanto termos pertencentes à terminologia de Mineração de Dados.

Custos e benefícios

Registra uma análise dos custos do projeto comparados com os potenciais benefícios para o negócio, caso o projeto alcance sucesso. Essa comparação deve ser o mais específico possível. Por exemplo, pode-se utilizar valores monetários em contextos comerciais.

1.3 - Analisar os objetivos de Mineração de Dados

O objetivo desta tarefa é traduzir para objetivos, numa perspectiva de Mineração de Dados, os objetivos de negócio analisados na tarefa 1.1.

Objetivos da Mineração de Dados

Registra os resultados objetivados pelo projeto para auxiliar no alcance dos objetivos de negócio.

Critério de sucesso, sob uma perspectiva de Mineração de Dados

Registra, em termos técnicos, os critérios para determinar se o projeto alcançou sucesso.

1.4 - Desenvolver um plano do projeto

Registra um plano para alcançar os objetivos de Mineração de Dados e então os objetivos de negócio. Inclui tanto as atividades a serem desenvolvidos, quanto as ferramentas e técnicas a serem utilizadas.

Plano do projeto

Registra as atividades a serem desenvolvidas, incluindo duração, recursos necessários, entradas, saídas e dependências. É importante que sejam registradas as dependências e riscos das atividades e como podem impactar nos prazos. Dado o aspecto iterativo de um projeto de Mineração de Dados, o plano de projeto é um documento dinâmico, sendo recomendado que ao fim de cada fase ele seja revisado e atualizado.

Avaliação inicial de ferramentas e técnicas

Registra a avaliação de ferramentas e técnicas para a escolha de quais serão utilizadas no projeto. É importante que essa análise seja realizada no início do projeto, dado que a escolha das ferramentas e técnicas podem influenciar todo o resto do projeto.

2 - Análise dos dados

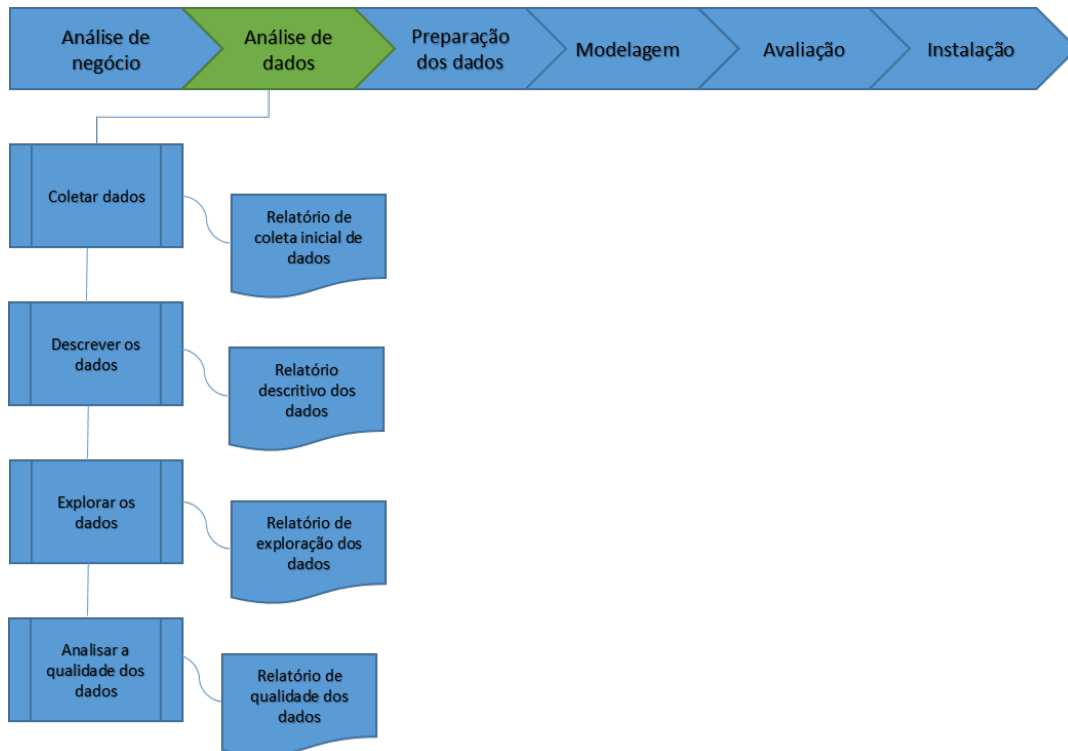


Figura 4 – Detalhamento da fase Análise dos dados

2.1 - Coletar dados

O objetivo desta tarefa é realizar a coleta dos dados indicados nos recursos do projeto. Nesta tarefa estão inclusos tanto o trabalho de extração quanto de integração dos dados, caso provenham de fontes de dados diferentes.

Relatório da coleta inicial de dados

Registra os conjuntos de dados coletados, suas localizações, os métodos utilizados na coleta e problemas, com respectivas soluções adotadas, que nela tenham ocorrido.

2.2 - Descrever os dados

O objetivo desta tarefa é realizar uma análise estrutural dos dados, avaliando se eles satisfazem os requisitos do projeto.

Relatório descritivo dos dados

Registra informações sobre os dados coletados, como formato, quantidade de registros e nomes de atributos.

2.3 - Explorar os dados

O objetivo desta tarefa é realizar uma análise da distribuição dos dados, através de consultas, visualizações e técnicas de report. Nela estão inclusas análise da distribuição

de atributos dos dados, análise do relacionamento entre pares de atributos, análise de subpopulações e análise estatística.

Relatório de exploração dos dados

Registra as descobertas dessa tarefa e impacto que causarão no projeto.

2.4 - Analisar a qualidade dos dados

O objetivo desta tarefa é analisar a qualidade dos dados, verificando, por exemplo, se são completos(há registros para todos os casos necessários), se são corretos(ausência de erros), se há dados ausentes; e analisar soluções para os problemas de qualidade encontrados.

Relatório de qualidade dos dados

Registra os resultados da análise de qualidade dos dados, indicando, os problemas de qualidade e as possíveis soluções.

3 - Preparação dos dados

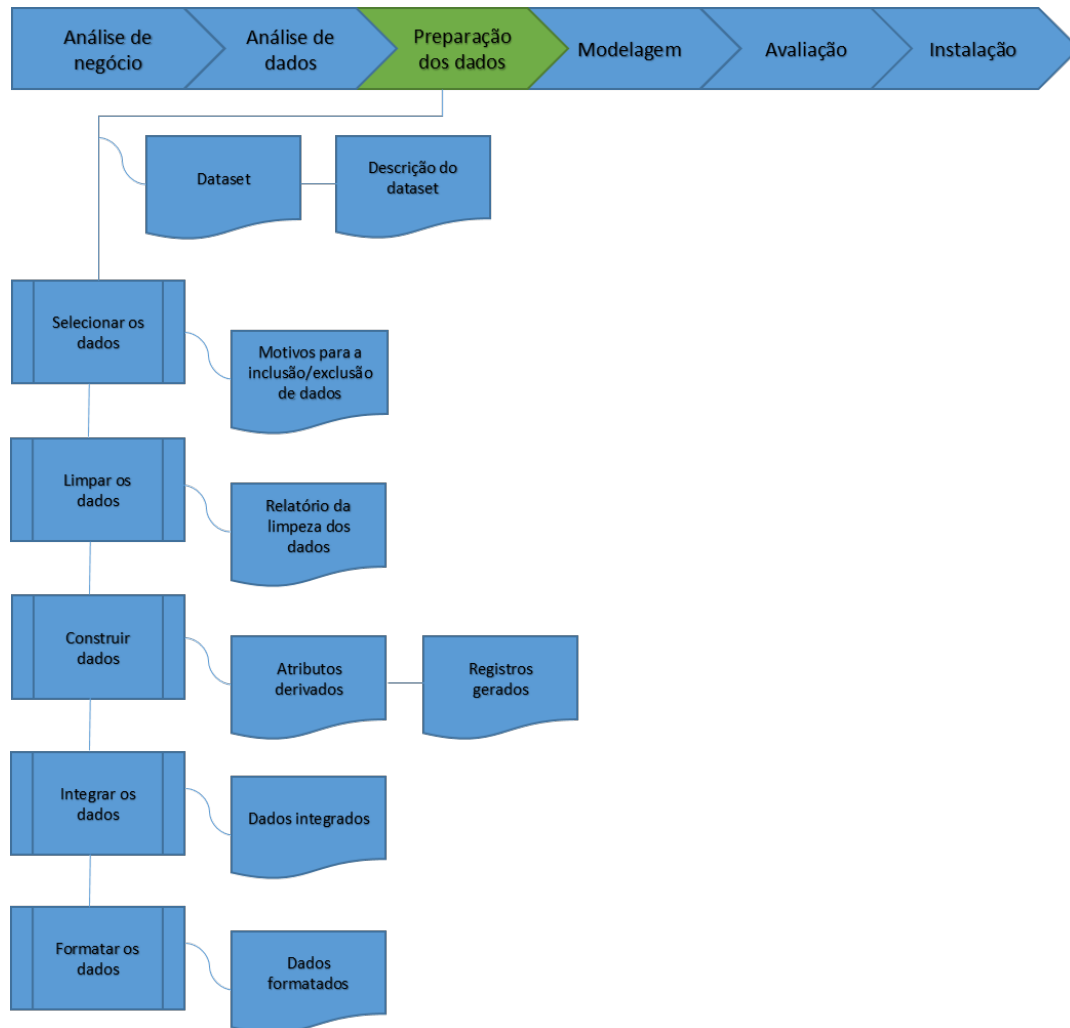


Figura 5 – Detalhamento da fase Preparação dos dados

Dataset

Datasets produzidos por esta fase, a serem utilizados no desenvolvimento de modelos ou em análises.

Descrição do dataset

Registra informações sobre o dataset produzido por esta fase.

3.1 - Selecionar dados

O objetivo desta tarefa é selecionar um subconjunto de dados a serem utilizados nas análises posteriores. Essa seleção envolve tanto a seleção de registros quanto a seleção de atributos. A lista de critérios para essa seleção inclui relevância dos dados para os objetivos de Mineração de Dados, qualidade e restrições técnicas, como limite no volume dos dados ou nos tipos dos dados.

Motivos para inclusão/exclusão de dados

Registra os motivos para inclusão e exclusão de dados.

3.2 - Limpar os dados

O objetivo desta tarefa é produzir um dataset com nível de qualidade adequado para a aplicação das técnicas e modelos selecionados pelo projeto, resolvendo os problemas de qualidade analisados na tarefa Analisar a qualidade dos dados. Para tanto, atividades como seleção de subconjunto dos dados, inserção de valores default e estimação de valores ausentes poderão ser necessárias.

Relatório da limpeza dos dados

Registra as alterações realizadas nos dados para resolver problemas de qualidade, indicando os motivos e possíveis consequências.

3.3 - Construir dados

O objetivo desta tarefa é a criação de novos dados, sejam derivados de dados já existentes, sejam novos registros, criados, por exemplo, através de interpolação de outros registros já existentes.

Atributos derivados

Registra os atributos que foram construídos a partir de outros já existentes. Por exemplo, $\text{área} = \text{altura} * \text{largura}$.

Registros gerados

Registra a geração de novos registros.

3.4 - Integrar os dados

O objetivo desta tarefa é criar novos dados através da integração de dados de fontes diversas.

Dados integrados

Esta saída é composta tanto pelos dados que foram gerados a partir da integração de dados de fontes diversas, quanto dados agregados.

3.5 - Formatar os dados

O objetivo desta tarefa é realizar transformações nos dados que não alterem seus significados, necessárias para que os dados possam ser utilizados pelas ferramentas. Exemplos de transformações são mudança do formato do arquivo onde estão os dados, alteração na ordem das colunas ou alteração na ordem dos registros.

Dados formatados

Registra as transformações realizadas nos dados, indicando motivos e possíveis consequências.

4 - Modelagem

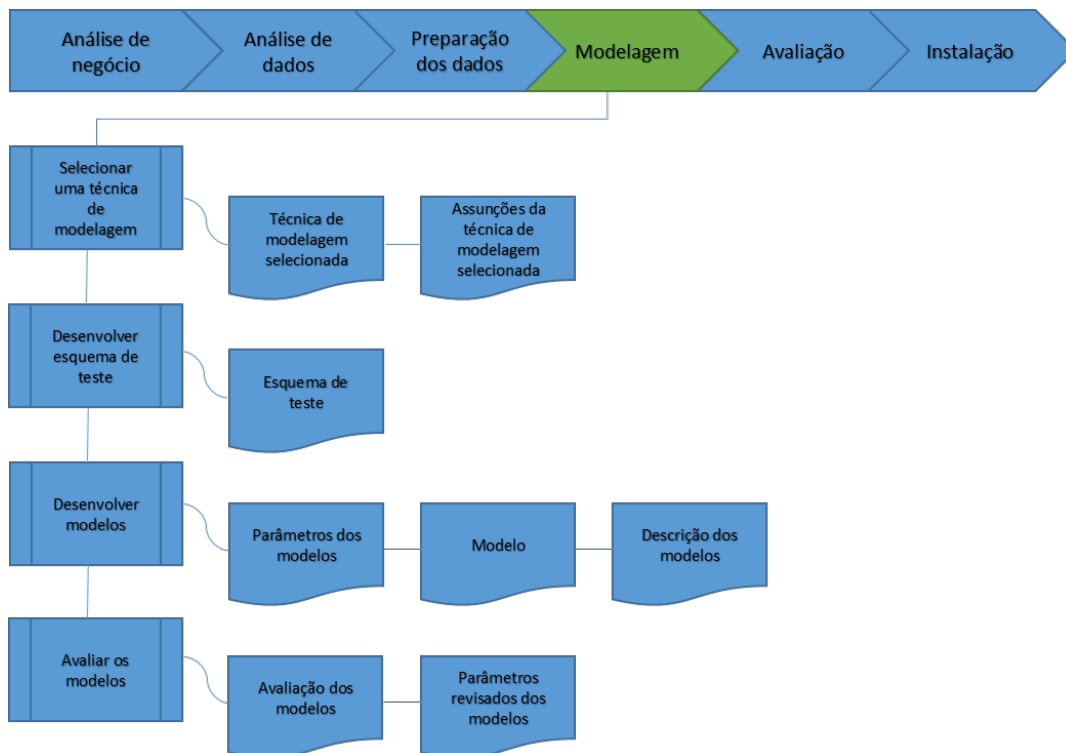


Figura 6 – Detalhamento da fase Modelagem

4.1 - Selecionar uma técnica de modelagem

O objetivo desta tarefa é selecionar uma técnica de modelagem a ser aplicada a um dataset gerado. No caso de várias técnicas de modelagem terem sido escolhidas para serem aplicadas, esta tarefa deve ser executada para cada uma delas.

Técnica de modelagem selecionada

Registra informações sobre a técnica de modelagem selecionada.

Assunções da técnica de modelagem selecionada

Registra as assunções feitas pela técnica de modelagem selecionada. Por exemplo, que todos os registros são independentes ou que todos os atributos possuem distribuição uniforme.

4.2 - Desenvolver esquema de teste

O objetivo desta tarefa é desenvolver um procedimento ou mecanismo para testar a qualidade e validade do modelo a ser desenvolvido. Para tanto, deve-se decidir, por exemplo, sobre como os dados serão particionados em subconjuntos de treinamento e de teste e quais métricas serão utilizadas para avaliar o desempenho.

Esquema de teste

Registra um plano para treinamento, teste e avaliação do modelo a ser desenvolvido.

4.3 - Desenvolver modelos

O objetivo desta tarefa é aplicar a técnica de modelagem escolhida ao dataset desenvolvido na fase anterior.

Parâmetros dos modelos

Registra os parâmetros utilizados pelos modelos desenvolvidos, bem como os motivos para suas escolhas.

Modelos

Modelos desenvolvidos.

Descrição dos modelos

Registra informações sobre os modelos desenvolvidos, como, por exemplo, como interpretá-los.

4.4 - Avaliar os modelos

O objetivo desta tarefa é avaliar os modelos desenvolvidos sob uma perspectiva de Mineração de Dados, verificando se os critérios de sucesso de Mineração de Dados foram satisfeitos, se os resultados do testes foram satisfatórios. Os modelos desenvolvidos devem ser então comparados e ordenados de acordo com critérios de avaliação.

Avaliação dos modelos

Registra os resultados da tarefa Avaliar o modelo, como a performance dos modelos desenvolvidos e uma ordem dos modelos de acordo com critérios de qualidade.

Parâmetros revisados dos modelos

Registra alterações propostas em parâmetros dos modelos desenvolvidos de acordo com a avaliação dos modelos. Os parâmetros revisados servem para serem utilizados no desenvolvimento, em uma nova iteração, de novos modelos.

Avaliação

Instalação

5.2 Ferramentas utilizadas

6 Resultados

7 Conclusão

Referências

- AZEVEDO, A. I. R. L. Kdd, semma and crisp-dm: a parallel overview. 2008.
- CHAPMAN, P. et al. Crisp-dm 1.0 step-by-step data mining guide. 2000.
- DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, ERIC, 2009.
- DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012.
- DRUMMOND, C. Finding a balance between anarchy and orthodoxy. In: *Proceedings of the Twenty-Fifth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning III (4 pages)*. [S.l.: s.n.], 2008.
- DRUMMOND, C. Replicability is not reproducibility: nor is it good science. 2009.
- MANHÃES, L. M. B. et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*, 2012.
- MANHÃES, L. M. B.; CRUZ, S. M. S. da; ZIMBRÃO, G. Evaluating performance and dropouts of undergraduates using educational data mining.
- MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.
- SCULLEY, D. et al. Machine learning: The high interest credit card of technical debt. In: *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. [S.l.: s.n.], 2014.
- SHERRILL, B.; EBERLE, W.; TALBERT, D. Analysis of student data for retention using data mining techniques. 2011.
- VILLWOCK, R.; APPIO, A.; ANDRETA, A. A. Educational data mining with focus on dropout rates. *International Journal of Computer Science and Network Security (IJCSNS)*, International Journal of Computer Science and Network Security, v. 15, n. 3, p. 17, 2015.