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.

Agradecimentos

Agradeço à Cris. Vlw!

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 ou 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

Lista de tabelas

Sumário

Todo list

1 Introdução

O fenômeno evasão de discente consiste na interrupção de um processo de aprendizado de um discente antes de sua conclusão. Por exemplo, um discente que abandonou o curso de Computação, na UFC, no qual estava matriculado havia dois anos, pois precisou trabalhar para sustentar sua família e os horários das disciplinas eram incompatíveis com os horários do trabalho. Deste exemplo pode-se observar alguns dos atributos do fenômeno: o agente que interrompeu o processo, o discente, o curso, a instituição de ensino superior(IES), o tempo cursado e o motivo.

Sob a perspectiva de que o processo de aprendizado é um investimento e que o resultado esperado é a sua conclusão, o fenômeno evasão de discente pode ser considerado um problema: para a sociedade, com a frustração da expectativa de formação de profissionais e pesquisadores qualificados; para a instituição de ensino, caso tenha realizado investimentos em infraestrutura e em recursos humanos para atender a uma quantidade esperada de discentes ativos maior que a real, ocorrendo desperdício de recursos; caso seu orçamento seja ou função da quantidade de discentes ativos, no caso das instituições de ensino particulares, ou função da quantidade de discentes diplomados, no caso das instituições de ensino superior públicas; para o indivíduo que investiu tempo, dinheiro e dedicação, mas não terá os benefícios da conclusão da graduação, crítica no caso das profissões que exigem, para serem exercidas, diploma de graduação.

validar o impacto da evasão no orçament das IFES

O estudo da evasão de discentes é motivado não apenas pelos problemas que dela podem decorrer mas também por diretrizes dos diversos níveis administrativos envolvidos com o processo.

No nível federal, redução da ocorrência desse fenômeno 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(REPúBLICA, 2007):

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

Na UFC, o instrumento de planejamento Plano de Desenvolvimento Institucional(UFC, 2013), para o período de 2013 a 2017, apresenta como um dos objetivos da política de assistência estudantil a redução da evasão; o programa de gestão da chapa eleita para reitoria no período de 2015 a 2019(CUSTóDIO, 2015) apresenta um conjunto de propostas que possuem como um dos objetivos a redução dos índices de evasão; o planejamento estratégico do Centro de Tecnologia da UFC para o período de 2015 a 2025 propõe a criação de uma equipe de apoio pedagógico para atuar no combate a problemas relacionados à evasão de discentes.

Em (LOBO, 2012) são apresentadas sete ações que ajudam a diminuir a ocorrência de evasão de discentes:

1. Estabelecer um grupo de trabalho encarregado de reduzir a evasão
2. Avaliar as estatísticas da evasão

3. Determinar as causas da evasão
4. Estimular a visão da IES centrada no aluno
5. Criar condições que atendam aos objetivos que atraíram os alunos
6. Tornar o ambiente e o trânsito na IES agradáveis aos alunos
7. Criar programa de aconselhamento e orientação dos aluno

Estas ações podem ser beneficiadas pela utilização de ferramentas de Aprendizado de Máquina, subárea de Inteligência Artificial, que estuda o desenvolvimento de programas cujas performances melhorem a partir de dados.

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 departamente 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.

O presente trabalho objetiva avaliar a aplicabilidade de técnicas de aprendizado de máquina ao problema de evasão de discentes na UFC.

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)

adicionar exemplos de aplicação

2.2 Modelagem

Para exemplificar a modelagem de um programa com uma abordagem de aprendizado de máquina, (MITCHELL, 1997) apresenta uma sequência de passos para desenvolver um programa que aprenda a jogar xadrez, a ser utilizado para disputar um campeonato mundial de xadrez, tendo sua performance medida pela frequência de partidas que vencer.

Escolha da experiência

O primeiro passo consiste na escolha da experiência a partir da qual o programa irá aprender, denominada experiência de treinamento. (MITCHELL, 1997) apresenta três atributos da experiência de treinamento que devem ser levados em consideração por impactarem no sucesso do aprendizado: feedback, nível de controle sobre os exemplos de treinamento e quão bem a a distribuição dos exemplos de treinamento representam a distribuição dos exemplos sobre os quais a performance final do programa será mensurada.

O atributo feedback representa quão direta é a informação fornecida pela experiência para o problema em questão, podendo assumir os valores feedback direto e feedback indireto. Por exemplo, a tupla estado do tabuleiro e o 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 a ser executado. 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 influência 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.

é necessári traduzir credit assignme

Com relação ao nível de controle sobre os exemplos de treinamento, a 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 foi

selecionada por um jogador experiente de xadrez, que selecionou estados de tabuleiro e indicou que melhores movimentos poderiam ser feitos, a partir desses estados; 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.

Com relação a quão bem a a distribuição dos exemplos de treinamento representam a distribuição dos exemplos sobre os quais a performance final do programa será mensurada, a experiência pode ser classificado como representativa, se a distribuição de seus exemplos representar a distribuição dos exemplos com os quais o programa efetivamente será utilizado, e não representativo, caso contrário. Por exemplo, a experiência é não representativa 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 esperado que 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 é representativa.

Para o problema em análise, (MITCHELL, 1997) escolhe como experiência de treinamento um conjunto de partidas jogadas pelo programa contra ele mesmo.

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 exemplo de função alvo, a ser utilizada no exemplo, é uma 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. Por exemplo, 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; um conjunto de regras que associe atributos do estado do tabuleiro a números reais;

uma função polinomial de atributos do estado do tabuleiro em números reais etc. Para dar continuidade ao detalhamento dos passos, (MITCHELL, 1997) escolhe a seguinte representação de função, denominada V:

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

Os coeficientes w_0 a w_6 são parâmetros do programa, a serem ajustados no aprendizado. As variáveis x_1 a x_6 possuem as seguintes definições:

- 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

Resumindo os passos até aqui realizados, temos:

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

Escolha de um algoritmo de aproximação

O próximo passo consiste na escolha de um algoritmo que irá realizar o treinamento da função, isto é, que, a partir da experiência de treinamento, irá ajustar seus coeficientes a fim de aproximá-la da função alvo. Para tanto, é necessário um conjunto de exemplos de treinamento composto por tuplas, $\langle b, V_{treino}(b)\rangle$, de estado de tabuleiro e o valor da função alvo para esse estado, respectivamente, b e $V_{treino}(b)$. Por exemplo, o seguinte exemplo representa um estado de tabuleiro onde o jogador com peças pretas ganhou:

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

Para o desenvolvimento dos exemplos de treinamento, é necessário estimar um valor para cada estado de tabuleiro alcançado nas partidas disputadas pelo programa contra ele mesmo(experiência de treinamento escolhida), indicando quão bom aquele estado de tabuleiro é. A abordagem adotada no exemplo consiste em atribuir a $V_{treino}(b)$, caso seja um estado final, um valor arbitrário, caso seja um estado intermediário, o valor da função estimada, $\bar{V}(t)$, aplicado ao próximo estado do tabuleiro para o jogador ativo em b, $Sucessor(b)$:

$$V_{treino}(b) \leftarrow \bar{V}(t)(Sucessor(b)) \tag{2.1}$$

(MITCHELL, 1997) indica que, apesar de parecer estranho atribuir aos exemplos de treinamento valores estimados pela função que está sendo treinada, sob certas condições, pode ser provado que essa abordagem converge para uma estimativa perfeita de V_{treino}.

ponto muito confuso do capítulo! qual o valor inicial dos coeficientes? random? qual o valor de V_treino para o último estado do tabuleiro? usa um valor arbitrário? qual? acho que +100 win -100 loss e intuitivamente parece mesmo que a tendência é convergir

Para estimar a função alvo a partir dos dados de treinamento é necessário estimar seus pesos, w_0 a w_6. Para tanto, primeiro é necessário definir uma medida de quão bem a função estimada se ajusta aos exemplos de treinamento. Uma medida comum é o erro quadrático, assim definido:

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

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 Least Mean Square. Esse algoritmo funciona atualizando iterativamente, para cada exemplo de treinamento, os pesos da função, ajustando-os com um valor proporcional a $(V_{treino}(b) - V(b))$: se o valor da função aplicado ao exemplo de treinamento, $V(b)$, for igual ao valor do exemplo de treinamento, $V_{treino}(b)$, o ajuste será nulo; caso seja maior, o ajuste será negativo, fazendo com que, para o dado de treinamento em questão, o valor da função diminua; caso seja menor, o incremento será positivo, fazendo com que, para o dado de treinamento em questão, o valor da função aumente.

continuar a discussão dessa parte

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

apresentar a classificação de modelos em supervised - classification, regression, unsupervised, reinforcement etc

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

O fenômeno evasão de discente consiste na interrupção de um processo de aprendizado de um discente antes de sua conclusão. Ele pode ser caracterizado por um conjunto de atributos: o agente que interrompeu o processo, o discente, o escopo, o tempo cursado e o motivo.

A iniciativa de interromper o processo de aprendizado pode ter sido tanto do discente, nos casos de desistência e transferência, ou da instituição de ensino, nos casos de jubilamento.

Com relação ao atributo escopo, a evasão pode ser classificada em:

- Evasão de curso: refere-se à evasão de um curso.
- Evasão de área de conhecimento: refere-se à evasão de um curso sem posterior ingresso em curso da mesma área de conhecimento.
- Evasão de IES: refere-se à evasão de um curso sem posterior ingresso em curso da mesma instituição de ensino.
- Evasão do ensino: refere-se à evasão de um curso sem posterior ingresso em outro curso.

Com relação ao tempo cursado, (FILHO et al., 2007) indica que, em todo o mundo, a taxa de evasão no primeiro ano de curso é duas a três vezes maior que a dos anos seguintes.

3.2 Dados de ocorrência

A fim de quantificar a ocorrência de evasão de discentes, é necessária a utilização de métricas. A métrica que quantifica com maior exatidão o fenômeno, de acordo com (LOBO, 2012) é a Acompanhamento de Coorte, consistindo na análise individualizada de cada discente. Para realizar tal análise são necessárias informações do histórico do discente durante o processo de aprendizado. Caso essas informações não estejam disponíveis, como quando a análise é feita sobre dados agregados, faz-se necessária a utilização de métricas de mais alto nível. Em (TODO, 2012a) são definidas as métricas Taxa de Titulação e Evasão Anual.

Definido com o nome Taxa de Sucesso em (TCU, 2009), a Taxa de Titulação é a razão entre a quantidade de discentes diplomados em um ano e a quantidade de diplomados esperados para aquele ano, considerando o prazo esperado de conclusão(por exemplo, caso tenham ingressados 60 discentes em Computação em 2008 e diplomados

Tabela 1 Taxa de Sucesso na UFC, ano 2013 - 5 menores e 5 maiores resultados

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%
Educação Física - Licenciatura	Diurno	124.0%
Odontologia - Sobral	Diurno	135.0%
Pedagogia	Diurno	143.6%
Geografia - Licenciatura	Diurno	219.0%
Filosofia - Licenciatura	Noturno	223.1%

30 em 2012, considerando 4 anos o tempo esperado de conclusão do curso, o valor dessa métrica para o ano 2012 será 0.5).

$$Taxa\ de\ Sucesso = \frac{N^o\ de\ diplomados}{N^o\ de\ diplomados\ esperados} \tag{3.1}$$

Este indicador pode gerar interpretações equivocadas por abstrair detalhes do fenômeno, como a ocorrência de ocupação, via transferência, de vagas ociosas, e por considerar apenas as conclusões no prazo esperado, podendo gerar dados inesperados, como percentuais acima de 100%, indicando haver se formado no ano em questão mais discentes que o esperado, consequência, por exemplo, de discentes atrasados ou de discentes adiantados.

Em 1 são apresentados os cinco maiores e cinco menores valores de Taxa de Sucesso para cursos da UFC no ano 2013, de acordo com (**??**).

Definido em (TODO, 2012b) e (TODO, 2012a), é a razão entre a quantidade de rematrículas realizadas e a quantidade de rematrículas esperadas.

$$1 - \frac{M(n) - Ig(n)}{M(n-1) - Eg(n-1)} \tag{3.2}$$

- $M(n)$: quantidade de matrículas na unidade de tempo n
- $Eg(n)$: quantidade de egressos na undiade de tempo n
- $Ig(n)$: quantidade de ingressantes na unidade de tempo n

Para derivar essa taxa, basta notarmos as seguintes equações e respectivos significados:

1. $$M(n) = E(n) + X(n) + Eg(n) \tag{3.3}$$

   significando que a quantidade de matrículas em determinada unidade de tempo n é composta pelas quantidades de matrículas de discentes que evadirão em n, de discentes que persistirão, isto é, se matricularão em $n + 1$, e de discentes que se diplomarão ao fim de n.

2.
$$M(n) = X(n-1) + Ig(n) \tag{3.4}$$
significando que a quantidade de matrículas em determinada unidade de tempo n é composta pelas quantidades de matrículas de discentes que se matricularam em n e em $n-1$, ou seja, que persistiram de $n-1$ a n, e de discentes que ingressaram em n.

Fazendo as devidas manipulações nas equações, chegamos na equação de quantidade de evasões ao fim de determinada unidade de tempo:

$$E(n) = [M(n) - Eg(n)] - [M(n+1) - Ig(n+1)] \tag{3.5}$$

Podendo ser interpretada como a diferença de quantidade máxima de discentes persistindo e a quantidade efetiva de discentes que persistiram.

A fórmula da evasão anual pode ser então definida como a proporção entre a quantidade de evasões em uma unidade de tempo e a quantidade máxima de discentes persistindo naquela unidade de tempo.

A definição de evasão anual até então analisada refere-se à evasão de cursos, de forma que um discente que tenha mudado de curso na mesma IES terá ao mesmo tempo contabilizados sua matrícula no período anterior($M(n-1)$) e como ingressante($Ig(n)$).

Para a evasão anual de IES, é proposta a seguinte definição:

$$1 - \frac{M(n) - [Ig(n) - ITC(n)]}{M(n-1) - Eg(n-1)} \tag{3.6}$$

- $ITC(n)$: quantidade de ingressantes no ano n com forma de ingresso transferência interna

analisar o ciclo de vida de discente e derivar os indicador

Já para a evasão anual do sistema, a ser aplicada a dados de várias, se possível todas, IES, a definição proposta é:

$$1 - \frac{M(n) - [Ig(n) - ITC(n) - ITIES(n)]}{M(n-1) - Eg(n-1)} \tag{3.7}$$

- $ITIES(n)$: quantidade de ingressantes no ano n com forma de ingresso transferência externa

apresentar dados do brasil

apresentar dados da UFC

3.3 Consequências

Sob a perspectiva de que o processo de aprendizado é um investimento e que o resultado esperado é a sua conclusão, o fenômeno evasão de discente pode ser considerado um problema: para a sociedade, com a frustração da expectativa de formação de profissionais e pesquisadores qualificados; para a instituição de ensino, caso tenha realizado investimentos

em infraestrutura e em recursos humanos para atender a uma quantidade esperada de discentes ativos maior que a real, ocorrendo desperdício de recursos; caso seu orçamento seja ou função da quantidade de discentes ativos, no caso das instituições de ensino particulares, ou função da quantidade de discentes diplomados, no caso das instituições de ensino superior públicas; para o indivíduo que investiu tempo, dinheiro e dedicação, mas não terá os benefícios da conclusão da graduação, crítica no caso das profissões que exigem, para serem exercidas, diploma de graduação.

validar o impacto da evasão no orçamento das IFES

Em (PRESTES; FIALHO,) foi estimado um prejuízo econômico, decorrente da evasão de discentes de graduação no período de 2007 a 2012, para a UFPB, de R$ 415.032.704,52. A estimativa considera perdidos os recursos financeiros investidos para manutenção do discente que não concluiu a graduação, utilizando a fórmula:

$$Perda\ Anual = n\_evadidos \times t\_permanencia \times v\_aluno$$

onde:

copiado da introdução - na introdução, resumir

- $n\_evadidos$ representa a quantidade, por ano, média de discentes que evadiram no período, considerada a média aritmética do total de discentes que evadiram no período pela quantidade de anos do período.
- $t\_permanencia$ representa o tempo de permanência esperado de um discente antes de evadir.
- v_aluno representa o custo corrente com hospital universitário por aluno corrente, indicador definido pelo TCU(TCU, 2009).

estimar o prejuízo econômico da evasão para a UFC

Aplicando a fórmula para os dados da UFC, obtemos o resultado .

3.4 Causas

(LOBO, 2012) indica um conjunto de causas mais comuns de evasão do sistema de ensino superior no Brasil, dentre elas:

1. Baixa qualidade da educação básica brasileira.
2. Limitação das políticas de financiamento ao estudante.
3. Escolha precoce da especialidade profissional.
4. Dificuldade de mobilidade estudantil, indicada como a dificuldade de transferência entre IES e de aproveitamento de créditos.
5. Falta de pressão para combater a evasão.
6. Enorme quantidade de docentes despreparados para o ensino e para lidar com o aluno real.

No trabalho (ANDRIOLA; ANDRIOLA; MOURA, 2006) são apresentados os resultados de pesquisa feita com uma amostra(tamanho 86 de um universo de tamanho 412) de discentes evadidos no período de 1999 e 2000 acerca dos motivos que os levaram a abandonar os cursos. Os seguintes motivos foram apresentados:

1. Incompatibilidade entre horários de trabalho e de estudo(destacado por 39.4% dos evadidos).

2. Aspectos familiares(por exemplo: necessidade de dedicar-se aos filhos menores) e desmotivação com os estudos(justificado por 20% dos evadidos).

3. Precariedade das condições físicas do curso ou inadequação curricular(mencionado por 10% dos evadidos).

3.5 Soluções

Em (LOBO, 2012) são apresentados "Sete pontos para baixar a evasão":

1. Estabelecer um grupo de trabalho encarregado de reduzir a evasão.

2. Avaliar as estatísticas da evasão.

3. Determinar as causas da evasão.

4. Estimular a visão da IES centrada no aluno.

5. Criar condições que atendam aos objetivos que atraíram os alunos.

6. Tornar o ambiente e o trânsito na IES agradáveis aos alunos.

7. Criar programa de aconselhamento e orientação dos alunos.

3.6 Aplicação de aprendizado de máquina

falar das oportunidades, diante de análises de causa e das soluções

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 significante 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 Método

O capítulo todo é um resumo do CRISP-DM 1.0... fico dando
cite em todo parágrafo? o mesmo para o capítulo sobre aprendizado de máquina

Por que um processo?

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

O processo de aplicação de técnicas de Mineração de Dados para atender a um problema real demanda não apenas conhecimentos e execução de atividades relacionadas, estritamente, a aprendizado de máquina, mas também conhecimentos e execução de atividades relacionadas ao entendimento do problema em questão. A fim de organizar as atividades envolvidas nesse processo, foram desenvolvidas especificações de processos que orientam quais os passos devem ser seguidos para partir do entendimento de um problema e chegar a uma solução baseada em Mineração de Dados, como o CRISP-DM, o KDD e o SEMMA(AZEVEDO, 2008).

(AZEVEDO, 2008)

por que crisp-dm

4.1 Processo 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 teria que passar pelo aprendizado pelo qual passaram, baseado em tentativa e erro, e como garantirem, 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).

falar sobre mineração de dados como super área no capítulo sobre aprendiza de máquina

CRISP-DM segue uma estrutura hierárquica, composta por 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 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 anteriores, de acordo com o diagrama do processo.

Para o mapeamento do modelo em uma instância do processo, a especificação do CRISP-DM 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, 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

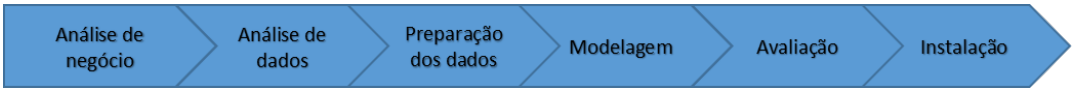

Figura 1 – Fases do CRISP-DM

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, de registros, de atributos e transformação de dados.

4. **Modelagem**: Nesta fase são desenvolvidos e otimizados modelos. Normalmente aplicam-se ao problema mais de uma técnicas de modelagem. Como algumas técnicas de modelagem podem possuir pré requisitos sobre os dados, 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.

A seguir segue o detalhamento de cada fase, especificando suas tarefas genéricas e os documentos que são gerados.

1 - Análise do negócio

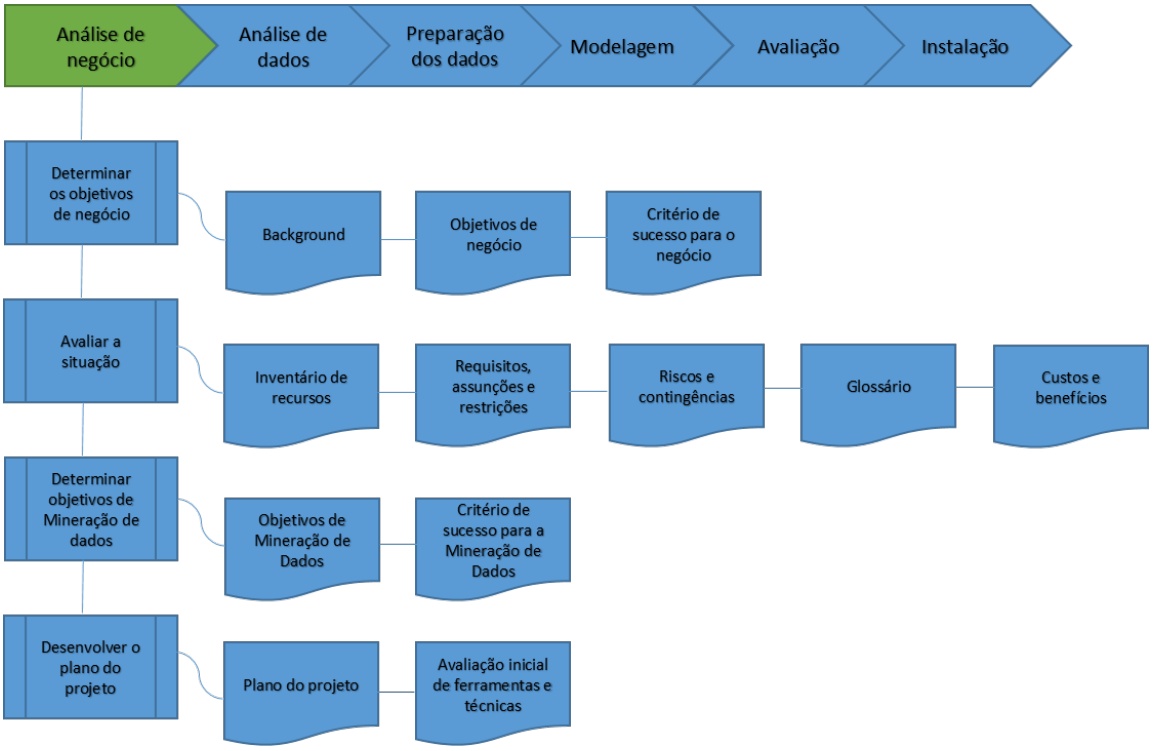

Figura 2 – Detalhamento da fase Análise de negócio

1.1 - Determinar os objetivos de negócio

O primeiro passo em um projeto de Mineração de Dados é analisar, sob uma perspectiva de negócio, o que o cliente deseja alcançar. Normalmente o cliente possui várias restrições e objetivos concorrentes, 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

esqueci de traduzir Background...

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

Objetivos do negócio

Registra os objetivos de negócio que motivaram a criação do projeto, além de questões a eles relacionadas que o cliente deseje responder.

Critério de sucesso para o 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 informações importantes para determinar os objetivos da Mineração de Dados e desenvolver um plano para o projeto. São analisadas informações como quais os recursos estão disponíveis, quais as restrições impostas, quais as assunções e outros fatores que forem relevantes para a especificação dos objetivos de Mineração de Dados e para o desenvolvimento do plano do projeto.

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, assunções e restrições

Registra os requisitos do projeto, incluindo prazos, níveis de qualidade, segurança e aspectos legais; as assunções do projeto, sejam assunções que poderão ser verificadas a partir dos dados utilizados pelo projeto, sejam assunções que não poderão ser verificadas, que devem ser registradas, visto que podem afetar a validade dos resultados do projeto; e as restrições do projeto, sejam restrições na disponibilidade de recursos, sejam restrições tecnológicas.

Riscos e contingências

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 o custo estimado do projeto e a economia esperada, em termos monetários.

1.3 - Determinar os objetivos de Mineração de Dados

O objetivo desta tarefa é traduzir para objetivos de Mineração de Dados os objetivos de negócio analisados na tarefa Determinar os objetivos de negócio.

Objetivos de Mineração de Dados

Registra os objetivos a serem alcançados pelo projeto para auxiliar no alcance dos objetivos de negócio.

Critério de sucesso para a Mineração de Dados

Registra, em termos técnicos, os critérios para determinar se o projeto alcançou sucesso, sob uma perspectiva de Mineração de Dados.

1.4 - Desenvolver o plano do projeto

O objetivo desta tarefa é desenvolver um plano para alcançar os objetivos de Mineração de Dados e então os objetivos de negócio, analisando que atividades serão executadas e que ferramentas e técnicas serão 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 seja revisado e atualizado.

Avaliação inicial de ferramentas e técnicas

Registra a avaliação de um conjunto de ferramentas e técnicas que poderão ser 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.

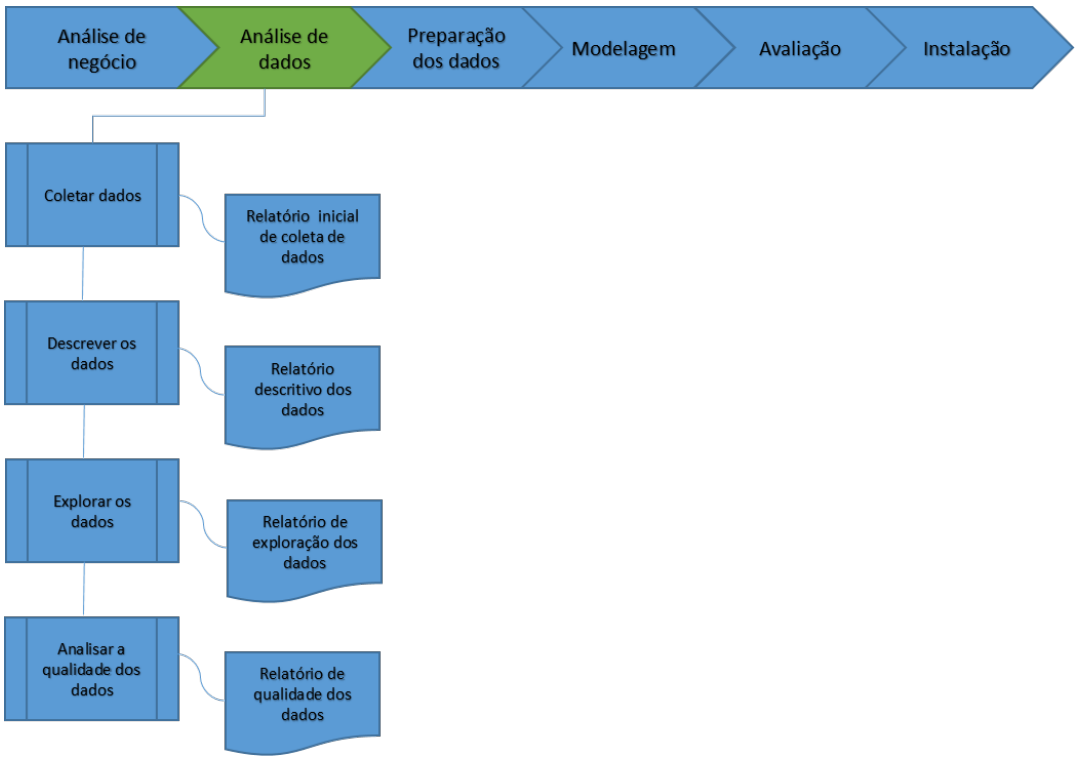

Figura 3 – Detalhamento da fase Análise de dados

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 diferentes, e carregamento dos dados em ferramenta específica, caso necessário.

Relatório inicial de coleta 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 estruturais 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

qual a tradução para report

de atributos dos dados, análise do relacionamento entre pares de atributos, análise de subpopulações e análise estatística. Essa análise serve tanto para suportar diretamente os objetivos de Mineração de Dados, quanto para refinar as informações sobre a estrutura e a qualidade dos dados.

Relatório de exploração dos dados

Registra as informações descobertas na tarefa Explorar os dados e o impacto que poderão causar 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(frequência 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

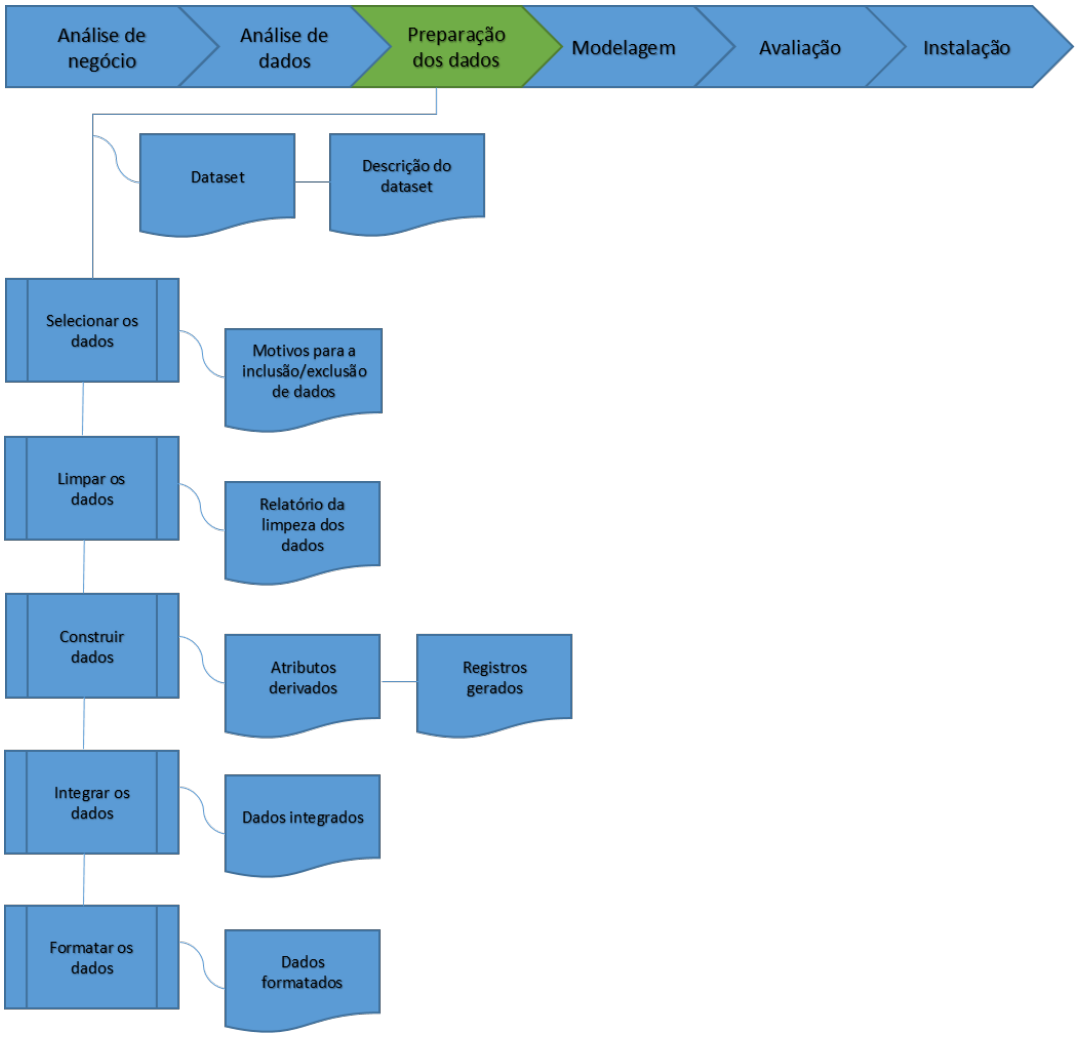

Figura 4 – Detalhamento da fase Preparação dos dados

Datasets

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

Descrição dos datasets

Registra informações sobre os datasets produzido nesta fase.

3.1 - Selecionar dados

O objetivo desta tarefa é selecionar datasets a serem utilizados em 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.

Motivos para inclusão/exclusão de dados

Registra os motivos para inclusão e exclusão de dados realizadas na tarefa Selecionar 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 padrão 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, através da derivação, a partir dos dados disponíveis, de novos registros ou atributos.

Atributos derivados

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

Registros gerados

pensar em exemplo de construção de registro

Registra os registros que foram construídos a partir de outros já existentes.

3.4 - Integrar os dados

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

Dados integrados

Dados resultantes da tarefa Integrar os dados, indicando quais fontes foram utilizadas e de que forma foi realizada a integração.

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

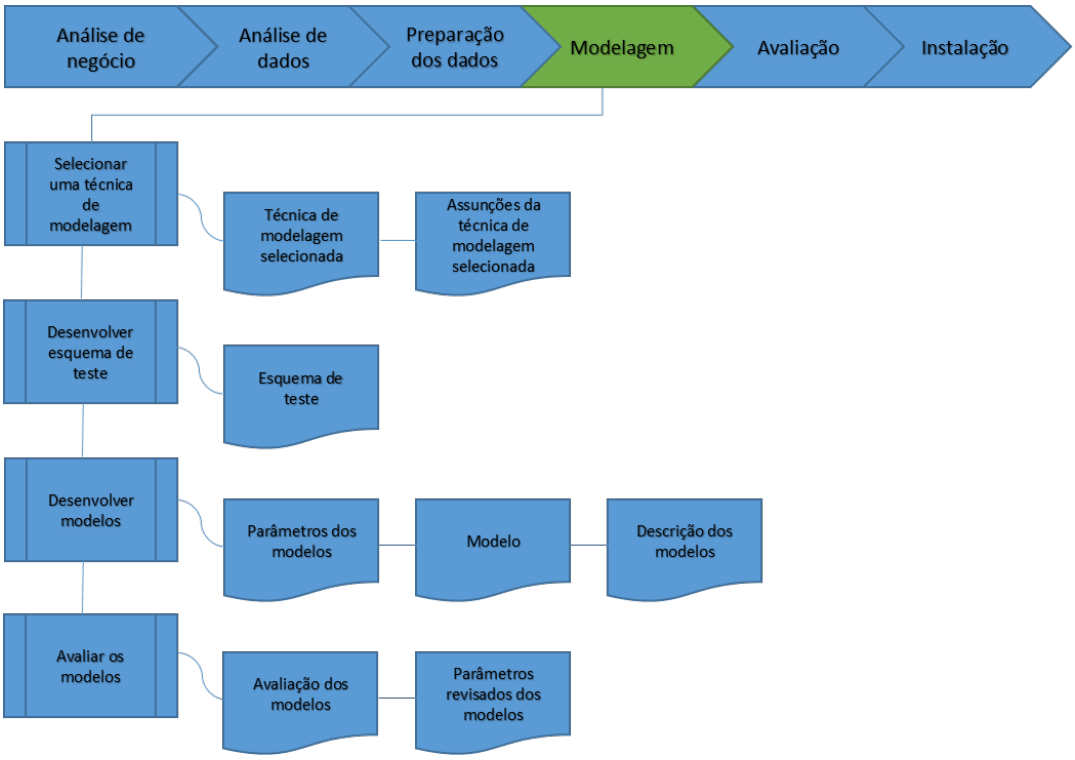

Figura 5 – Detalhamento da fase Modelagem

4.1 - Selecionar uma técnica de modelagem

O objetivo desta tarefa é selecionar uma técnica de modelagem a ser aplicada a um dos datasets gerados. No caso de várias técnicas de modelagem terem sido escolhidas para serem aplicadas, a fase Modelagem deve ser aplicada a cada uma delas.

Técnica de modelagem selecionada

Registra informações sobre a técnica de modelagem selecionada, como seu funcionamento e formato de dados de entrada.

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 a um dataset desenvolvido na fase anterior.

Parâmetros dos modelos

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

falar no capítulo aprendizado sobre parâmetros de modelos

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 e 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 os modelos, como a performance dos modelos desenvolvidos e uma ordenação 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.

5 - Avaliação

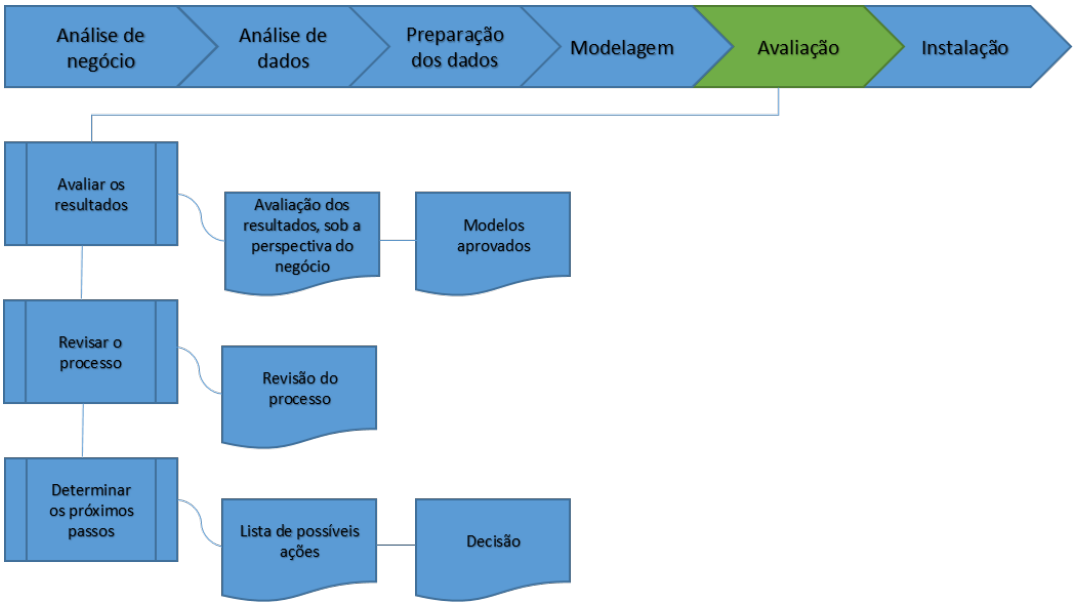

Figura 6 – Detalhamento da fase Avaliação

5.1 - Avaliar os resultados

O objetivo desta tarefa é avaliar em que medida os resultados alcançados com a Mineração de Dados, sejam modelos desenvolvidos, sejam informações extraídas, auxiliam no alcance dos objetivos de negócio e, se possível, testar os modelos desenvolvidos em aplicações reais.

Avaliação dos resultados, sob a perspectiva do negócio

Registra a avaliação dos resultados, sob a perspectiva do negócio, indicando se o projeto obteve sucesso em suportar os objetivos de negócio.

Modelos aprovados

Modelos que, na avaliação dos resultados, apresentaram resultados satisfatórios.

5.2 - Revisar o processo

A partir deste ponto do processo os modelos desenvolvidos já apresentam resultados satisfatórios e torna-se apropriada a realização de uma revisão do processo, a fim de verificar a qualidade das atividades até então desenvolvidas.

Revisão do processo

Registra os resultados da revisão do processo, indicando atividades que não foram desenvolvidas com a qualidade esperada e que deverão ser repetidas.

5.3 - Determinar os próximos passos

O objetivo desta tarefa é definir quais as próximas atividades a serem desenvolvidas, de acordo com os resultados da avaliação dos resultados, da revisão do processo e dos

recursos disponíveis para o projeto. Pode-se decidir pela implantação dos modelos desenvolvidos, pela realização de uma nova iteração ou a finalização do projeto.

Lista de possíveis ações

Registra as potenciais ações a serem executadas e os respectivos motivos para executá-las.

Decisão

Registra a decisão sobre quais os próximos passos a serem seguidos e a respectiva motivação.

6 - Instalação

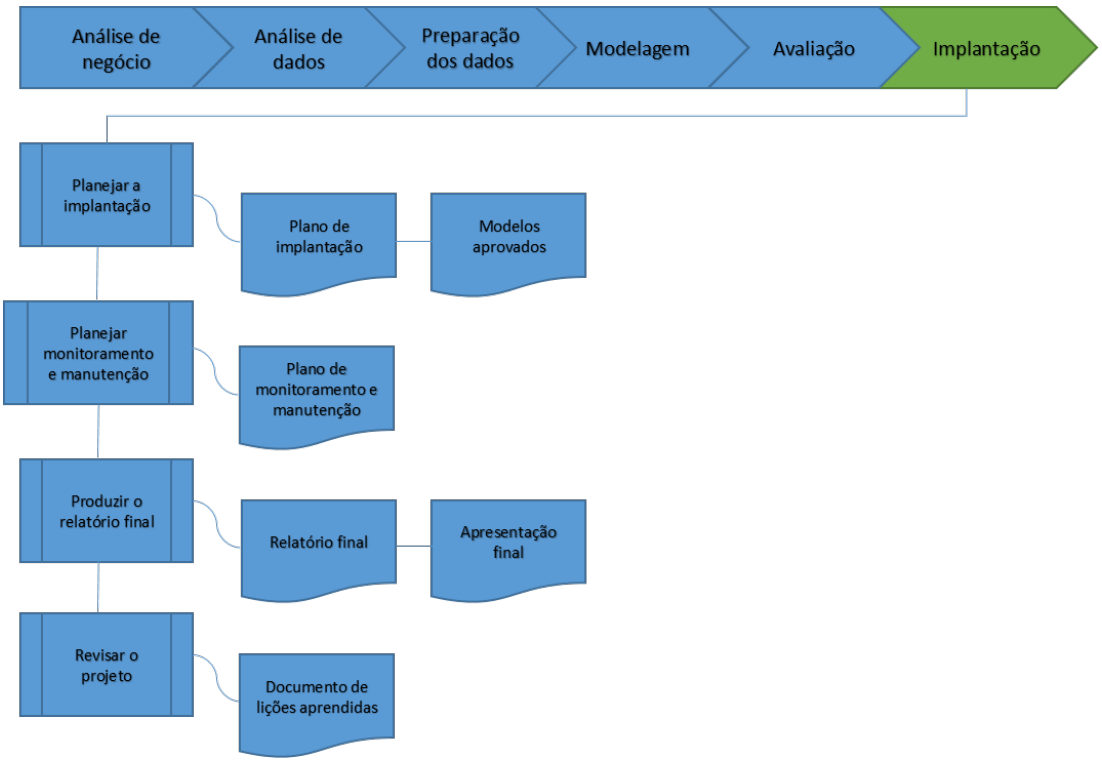

Figura 7 – Detalhamento da fase Implantação

6.1 - Planejar a implantação

O objetivo desta tarefa é produzir um plano de implantação dos modelos aprovados na fase anterior, tornando-os acessíveis aos usuários finais, seja como sistemas de apoio à decisão, seja via relatórios.

Plano de implantação

Registra as atividades necessárias para a implantação dos modelos aprovados.

6.2 - Planejar monitoramento e manutenção

O objetivo desta tarefa é desenvolver planos de monitoramento e manutenção, objetivando verificar os resultados dos modelos implantados e evitar, através de monitoramento e manutenção, que sejam utilizados indevidamente ou tornem-se obsoletos.

Plano de monitoramento e manutenção

Registra as estratégias de monitoramento e de manutenção.

6.3 - Produzir relatório final

O objetivo desta tarefa é produzir um relatório registrando o histórico do projeto, indicando as atividades que foram desenvolvidas, os atores responsáveis e os resultados alcançados.

Relatório final

Registra o histórico do projeto.

Apresentação final

Registra informações sobre os resultados alcançados pelo projeto, a ser apresentado aos clientes.

6.4 - Revisar o projeto

O objetivo desta tarefa é analisar o que foi feito correta e incorretamente no projeto.

Documento de lições aprendidas

Registra as lições aprendidas no projeto, indicando que ações foram executadas corretamente, para que sejam replicadas, e que ações foram executadas incorretamente, para que sejam corrigidas em futuros projetos.

4.2 Ferramentas utilizadas

(BUITINCK et al., 2013) (MCKINNEY, 2010)

5 Pontos de partida

Como pontos de partida, consideraremos a análise de dados para avaliação de um conjunto de assunções e o desenvolvimento de modelos de aprendizado de máquina úteis no entendimento do fenômeno e combate de seus problemas.

Assunções a serem avaliadas

As assunções a serem avaliadas são listadas indicando em que são baseadas e como serão avaliadas:

1.
   - As informações disponíveis ao fim do primeiro ano de um curso permitem desenvolver modelos de predição de evasão de discentes no primeiro ano.
   - **Baseado em**: a quantidade de discentes que evadem no primeiro ano ser maior que a quantidade de discentes que evadem nos anos subsequentes.
   - **Como será avaliada**: serão desenvolvidos e avaliados com dados de ingressantes de 2015 modelos de predição de evasão no primeiro ano.

2.
   - O desempenho de um discente no ENEM em uma área de conhecimento permite predizer seu desempenho em disciplinas dessa mesma área ou de área similar.
   - **Baseado em**: a ideia de que os conhecimentos testados no ENEM são pré requisitos para o aprendizado dos conhecimentos do curso.
   - **Como será avaliada**: serão desenvolvidos e avaliados com dados de ingressantes de 2015 modelos de predição, a partir do desempenho no ENEM, do desempenho em disciplinas do primeiro ano do curso.

3.
   - Conhecer as causas da evasão não implica em conhecer as causas da persistência.
   - **Baseado em**: (TINTO, 2012) afirma que, apesar de a evasão e a persistência serem fenômenos relacionados, o entendimento das causas da evasão não implica, necessariamente, no entendimento das causas da persistência.
   - **Como será avaliada**:

4.
   - Atributos do currículo de um curso, proporção de componentes curriculares obrigatórios, sinergia entre disciplinas obrigatórios de cada semestre, distribuição de disciplinas teóricas e práticas, distribuição de turmas nos turnos do dia estão correlacionados com a taxa de evasão anual do curso.
   - **Baseado em**: a intuição de que quanto maior a proporção de componentes curriculares obrigatórios, menor a liberdade de escolha, pelo discente, de sua formação; de que é mais difícil o aprendizado conteúdos relacionados que não relacionados; de que quão mais distantes forem cursadas disciplinas teóricas e práticas, de mesma conteúdo ou de conteúdos relacionados, mais difícil será o aprendizado; de que quanto mais distribuído nos turnas as turmas de disciplinas obrigatórias de um semestre de um currículo forem, menor a liberdade, do discente, de realizar outras atividades, como estágio.

formular melhor

analisar melhor. a ideia aqui é de dois fenômeno aparente um contrário ao outro, onde a remoção das

- **Como será avaliada**: serão desenvolvidos e avaliados modelos de predição da taxa de evasão anual de um curso a partir de dados de seus currículos.

5.
- Quão mais distante forem o histórico de um discente e o currículo ao qual ele está vinculado, maior a probabilidade de ele abandonar o curso.
- **Baseado em**: a intuição de que desvios do currículo são indícios de problemas para o discente, como é o caso de reprovação em uma disciplina obrigatória.
- **Como será avaliada**: serão desenvolvidas e avaliadas métricas para mensurar a distância entre o histórico de um discente e o currículo ao qual ele está vinculado.

6.
- A taxa de evasão anual no primeiro ano de curso é maior para discentes que ingressaram via 2º opção do Sisu que para aqueles que ingressaram via 1º opção do Sisu.
- **Baseado em**: o fato de que o discente que ocupa vaga de 2º opção continua no processo de seleção do Sisu, concorrendo por vagas com sua inscrição de 1ª opção.
- **Como será avaliada**: serão calculadas as taxas de evasão anual para discentes ingressantes via Sisu via 1ª opção e via 2ª opção.

Modelos de aprendizado de máquina a serem desenvolvidos

Os modelos de aprendizado de máquina a serem desenvolvidos são:

indicar como serão utilizados para alcançar os business goals

1. Modelo de predição de evasão por um discente no primeiro ano a partir de informações de seu desempenho em disciplinas obrigatórias do 1º semestre.
2. Modelo de predição de retençãode um discente a partir de informações de seu desempenho em disciplinas obrigatórias do 1º semestre.
3. Modelo de predição do desempenho de um discente em disciplinas obrigatórias do primeiro ano a partir do seu desempenho no ENEM.
4. Modelo de predição do desempenho de um discente em disciplinas a partir de seu desempenho em disciplinas pré requisito.

definir no capítulo sobre evasão

6 Resultados

7 Conclusão

Referências

ANDRIOLA, W. B.; ANDRIOLA, C. G.; MOURA, C. P. Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da universidade federal do ceará (ufc). *Ensaio: aval. pol. públ. Educ*, SciELO Brasil, 2006.

AZEVEDO, A. I. R. L. Kdd, semma and crisp-dm: a parallel overview. 2008.

BUITINCK, L. et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

CHAPMAN, P. et al. Crisp-dm 1.0 step-by-step data mining guide. 2000.

CUSTóDIO, H. e. *Programa de gestão 2015-2019*. 2015. <http://www.henry-custodio.com/programa.pdf>.

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.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

LOBO, M. B. d. C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, n. 25, 2012.

MANHÃES, L. M. B. et al. Identificaçao dos fatores que influenciam a evasao em cursos de graduaçao através de sistemas baseados em mineraçao 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.

MCKINNEY, W. Data structures for statistical computing in python. In: *Proceedings of the 9th*. [S.l.: s.n.], 2010. v. 445, p. 51–56.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

PRESTES, E. M. D. T.; FIALHO, M. G. D. A evasão no ensino superior globalizado e suas repercussões na gestão universitária.

REPúBLICA, P. da. *Decreto nº 6096*. 2007. <http://legislacao.planalto.gov.br/legisla/legislacao.nsf/Viw_Identificacao/DEC%206.096-2007?OpenDocument>.

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.

TCU. *Orientações para o cálculo dos indicadores de gestão - TCU*. 2009. <http://portal.mec.gov.br/sesu/arquivos/pdf/indicadores.pdf>.

TINTO, V. *Completing college: Rethinking institutional action*. [S.l.]: University of Chicago Press, 2012.

TODO. *COMO A MUDANÇA NA METODOLOGIA DO INEP ALTERA O CÁLCULO DA EVASÃO*. 2012. <http://www.institutolobo.org.br/imagens/pdf/artigos/art_079.pdf>.

TODO. *ESCLARECIMENTOS METODOLÓGICOS SOBRE OS CÁLCULOS DE EVASÃO*. 2012. <http://www.institutolobo.org.br/imagens/pdf/artigos/art_078.pdf>.

UFC. *Plano de Desenvolvimento Institucional*. 2013. <http://ufc.br/a-universidade/documentos-oficiais/313-plano-de-desenvolvimento-institucional-pdi>.

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.