



**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 . . . . .	15
---	----



# Lista de tabelas

Tabela 1 – Taxa de sucesso da graduação por curso na UFC em 2013 - 10 piores resultados, em ordem crescente . . . . .	16
---	----



# Sumário

1	INTRODUÇÃO	15
1.1	O que é o problema	15
1.2	Soluções	16
1.3	Objetivos	17
2	APRENDIZADO DE MÁQUINA	19
2.1	Definição	19
2.2	Modelagem	19
2.3	Modelos	23
2.4	Aplicações	23
3	EVASÃO DE DISCENTES	25
3.1	Definição	25
3.2	Dados de ocorrência	25
3.3	Causas	25
3.4	Soluções	25
3.5	Aplicação de aprendizado de máquina	25
4	PONTOS DE PARTIDA	27
4.1	Dados pré universidade e do primeiro semestre	27
5	METODOLOGIA	29
5.1	Visão geral do processo	29
5.2	Análise do problema	29
5.3	Busca dos dados	29
5.4	Análise dos dados	29
5.5	Pré processamento dos dados	29
5.6	Desenvolvimento, teste e avaliação de modelos	29
5.7	Finalização da solução	29
5.8	Ferramentas utilizadas	29
6	RESULTADOS	31
7	CONCLUSÃO	33
	REFERÊNCIAS	35





# 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 de 2014, ano base 2013, o indicador "Taxa de sucesso na graduação", definido como a proporção entre número de discentes diplomados e número de discentes ingressantes da graduação, esteve em 2013 com o menor valor desde 2008 (Figura 1). Já o indicador "Taxa de sucesso da graduação por curso", em 2013, possuiu valor mínimo igual a 6.8%, referente ao curso Ciências Sociais, habilitação em licenciatura (Tabela 1).

quando  
sai o  
próximo?

referência

manual  
de  
indicador  
do  
TCU,  
p.5

qual a  
definição

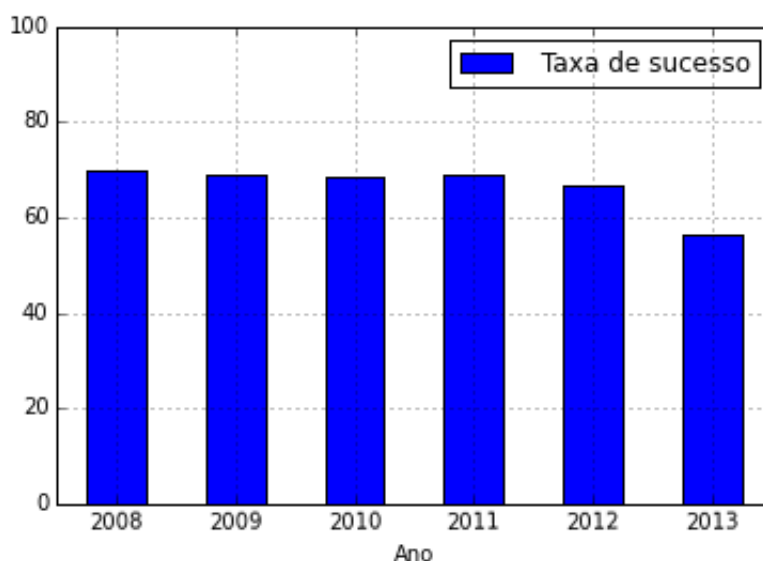


Figura 1 – Taxa de Sucesso na Graduação - UFC

**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%

referência

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

representation + evaluation + optimization (DOMINGOS, 2012)

## 2.4 Aplicações

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.

definir  
o que  
quer  
dizer  
por  
modelo

melhorar



## 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. 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.

citar a  
coeficiente  
for  
agreement  
for  
nominal  
scale

Educational  
Data  
Mining  
with  
Focus  
on  
Dropout  
Rates

Identificação  
dos  
Fatores  
que  
Influenciam  
a  
Evasão  
em  
Cursos  
de  
Graduação

Analysis  
of  
Student  
Data  
for  
Retention  
Using  
Data  
Mining  
Techniques

([VILLWOCK; APPIO; ANDRETA, 2015](#))

([MANHÃES et al., 2012](#))

([SHERRILL; EBERLE; TALBERT, 2011](#))

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 Dados pré universidade e do primeiro semestre

Especificação do problema

Benefícios



## 5 Metodologia

5.1 Visão geral do processo

5.2 Análise do problema

5.3 Busca dos dados

5.4 Análise dos dados

5.5 Pré processamento dos dados

5.6 Desenvolvimento, teste e avaliação de modelos

5.7 Finalização da solução

5.8 Ferramentas utilizadas





## 6 Resultados



## 7 Conclusão



# Referências

- 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.
- 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.
- SHERILL, 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.