Identificando Estudantes em Risco de Evasão Através de um Sistema Preditivo Baseado em Aprendizagem de Máquina

Gabriel Ferreira Silva[1], Marcelo Ladeira[1]
mladeira@unb.br

[1]Departamento de Ciência da Computação
Universidade de Brasília

2 a 4 de Outubro de 2017

KDMiLe2017

5th Symposium on Knowledge Discovery, Mining and Learning

October 2nd-4th, 2017

Sumário

Problemas Causados pela Evasão

Evasão em universidades brasileiras traz desperdícios acadêmicos, sociais e econômicos.

A Universidade de Brasília teve prejuízo estimado em 95,6 milhões em 2014 [1].

Abordagem da UnB - Alunos em Condição

Critérios para um aluno estar em condição:

- Duas reprovações na mesma disciplina obrigatória.
- Não ser aprovado em 4 disciplinas do curso em 2 períodos regulares consecutivos.
- Chegar ao último período do curso sem a possibilidade de concluí-lo.

Abordagem da UnB e Problemas

A UnB adota a seguinte abordagem:

- Separar alunos em condição dos demais.
- Ter alunos em condição supervisionados por orientador.

Problemas da abordagem da UnB:

- Alunos (amostra diversificada) separados em apenas dois grupos.
- UnB age apenas quando aluno já está em condição.

Proposta de Solução

Utilizar dados descaracterizados para criação de um sistema previsor capaz de identificar alunos em risco de serem desligados.

Entendendo a Saída

Sistema previsor fornece uma tripla (v_1, v_2, v_3) de valores entre 0 e 1 que somam 1.

v_1, v_2, v_3 indicam respectivamente a chance do aluno se graduar, ser desligado ou migrar de curso.

Vantagens

Sistema permitiria à UnB agir com antecedência e flexibilidade:

- Ações podem ser tomadas antes de um aluno entrar em condição.
- Ações podem ser tomadas de acordo com o risco apresentado por um aluno.

Cursos Estudados

Cursos considerados:

- Ciência da Computação
- Engenharia de Computação
- Engenharia de Controle e Automação (Engenharia Mecatrônica)
- Engenharia de Redes
- Engenharia de Software
- Licenciatura em Computação

Algoritmos de Aprendizagem de Máquina

Os seguintes algoritmos de aprendizagem de máquina (veja [2] ou [3]) foram utilizados:

- ANN
- *Naive Bayes*
- *Random Forest*
- Regressor Linear
- SVR

Levantamento do Estado da Arte e CRISP-DM

Levantamento do estado da arte, de modo a compreender:

- Quais fatores considerar [4].
- Como aplicar técnicas de aprendizagem de máquina [5].
- Peculiaridades da UnB [6].

Na pesquisa, utilizou-se o modelo de referência para mineração de dados CRISP-DM.

Obtenção e Utilização dos Dados

Informações descaracterizadas de alunos de graduação da UnB. Dados pessoais e de desempenho acadêmico.

Considerou-se apenas alunos que ingressaram a partir de 2000 e saíram até 2016.

Seleção Preliminar de Atributos

Atributos pessoais considerados em uma análise inicial:

- Cotista (ou não)
- Curso
- Forma de Ingresso
- Idade ao Ingressar na Universidade
- Raça
- Sexo
- Tipo da Escola

Seleção Preliminar de Atributos

Atributos relativos ao desempenho acadêmico:

- Coeficiente de Melhora Acadêmica
- Indicador de Aluno em Condição
- Média do Período
- Posição em relação ao semestre que ingressou
- Quantidade de créditos já integralizados
- Taxa de Aprovação, Reprovação e Trancamento
- Taxa de aprovação na disciplina mais difícil do semestre

Eliminação de Atributos Devido a Missing Values

Optou-se por eliminar atributos com percentagem de *missing values* maior que 40%.

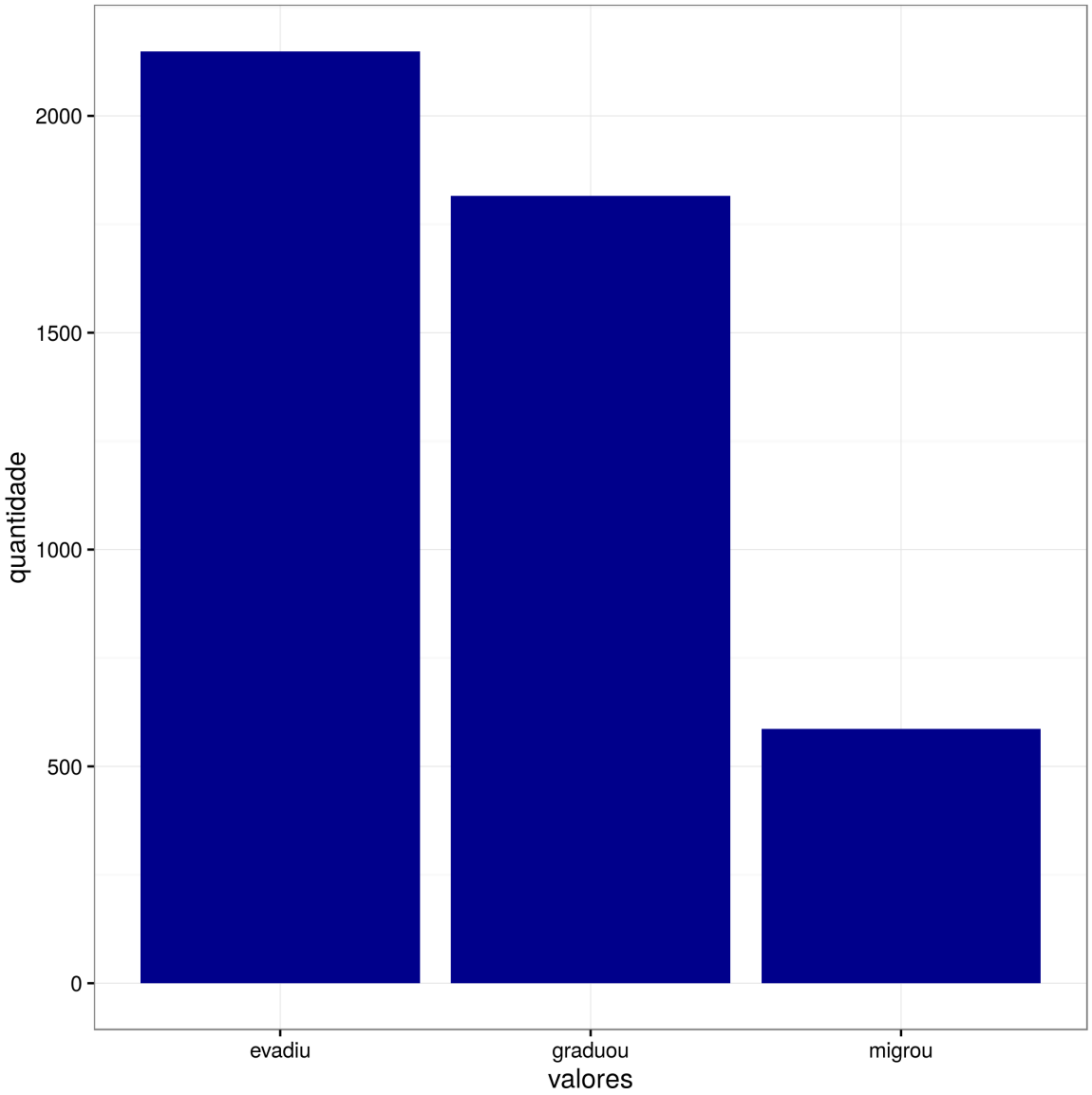
Assim, atributos raça e tipo da escola foram eliminados.

Mudança na Base de Dados

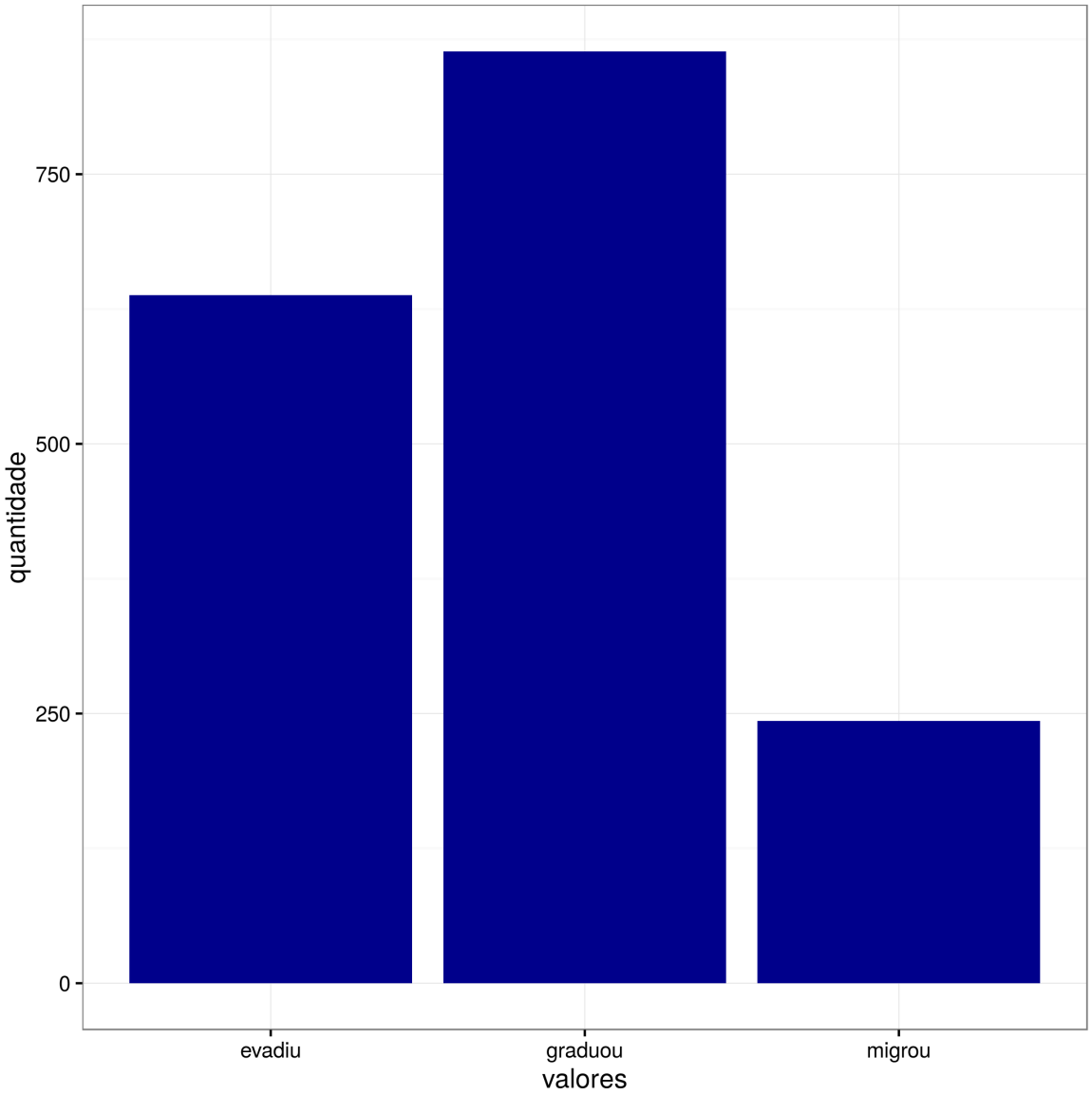
Através da análise de tabelas de contingência, decidiu-se particionar a base de dados original:

- Alunos Jovens da FT
  - Engenharia de Redes
  - Engenharia Mecatrônica
- Alunos Jovens de Licenciatura
  - Licenciatura em Computação
- Alunos Jovens de Computação
  - Ciência da Computação
  - Engenharia de Software
  - Engenharia de Computação
- Alunos Seniores (mais de 30 anos)
  - Todos os cursos

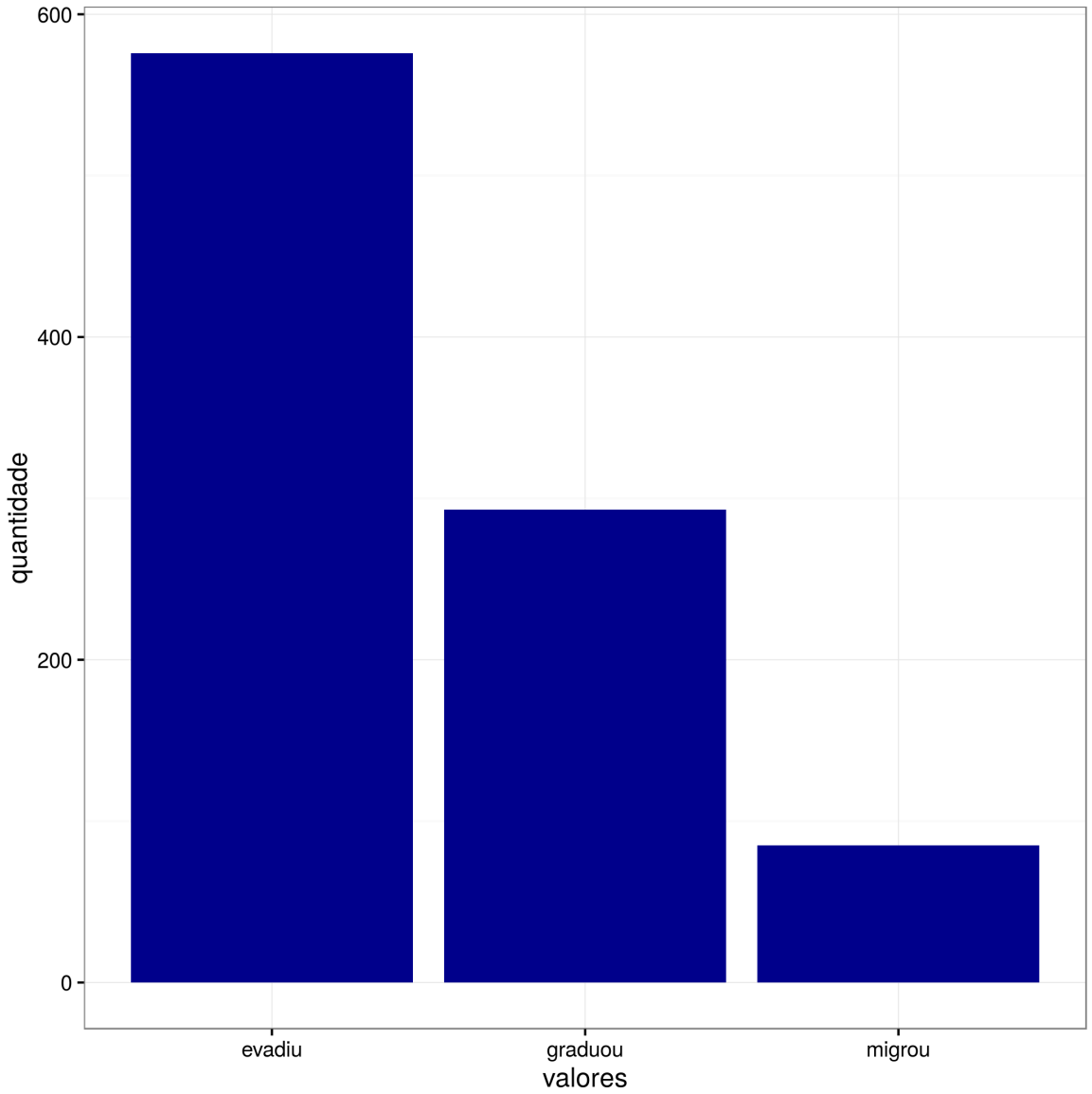
Alunos - Forma de Saída



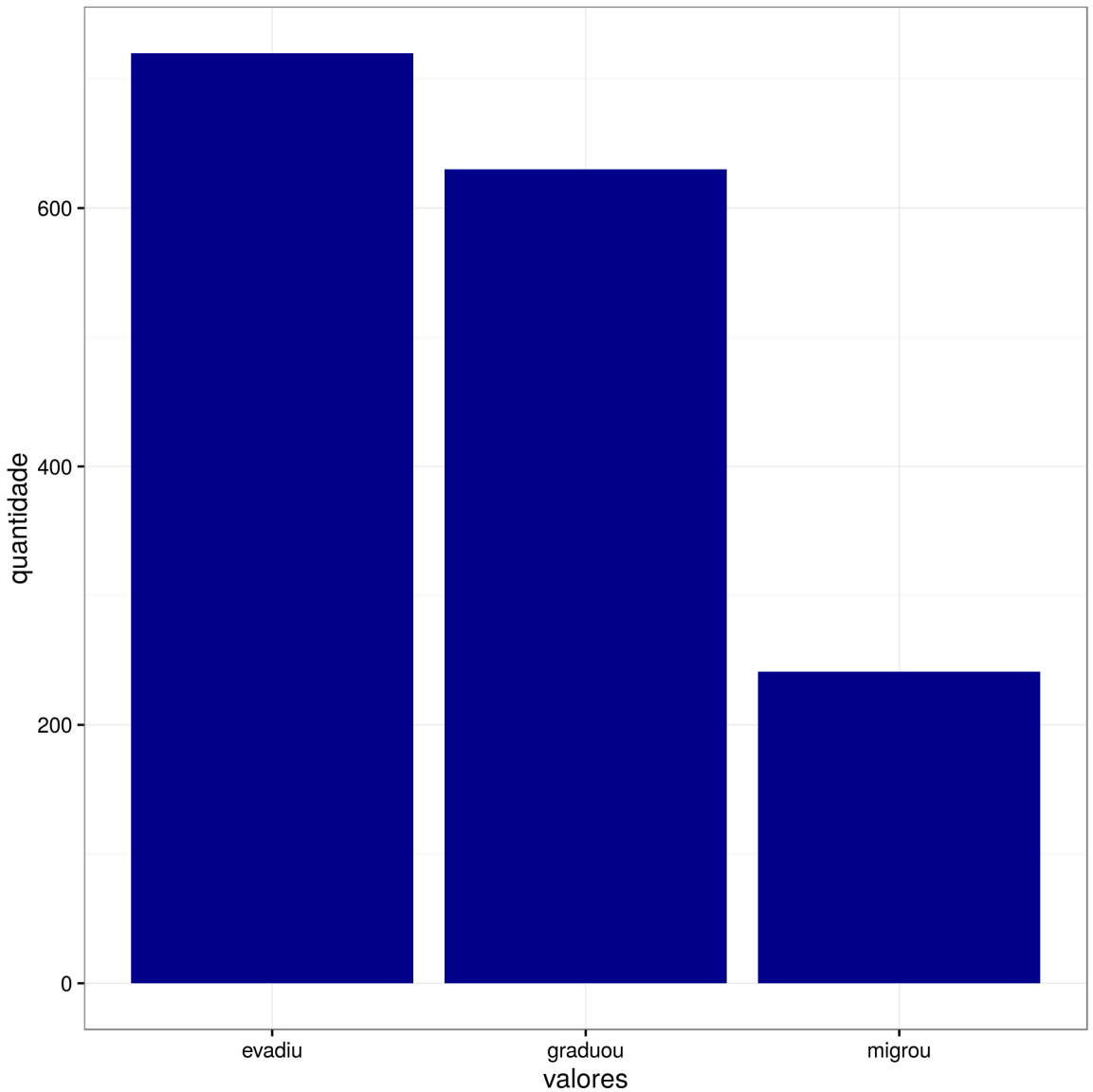
Alunos Jovens da FT - Forma de Saída



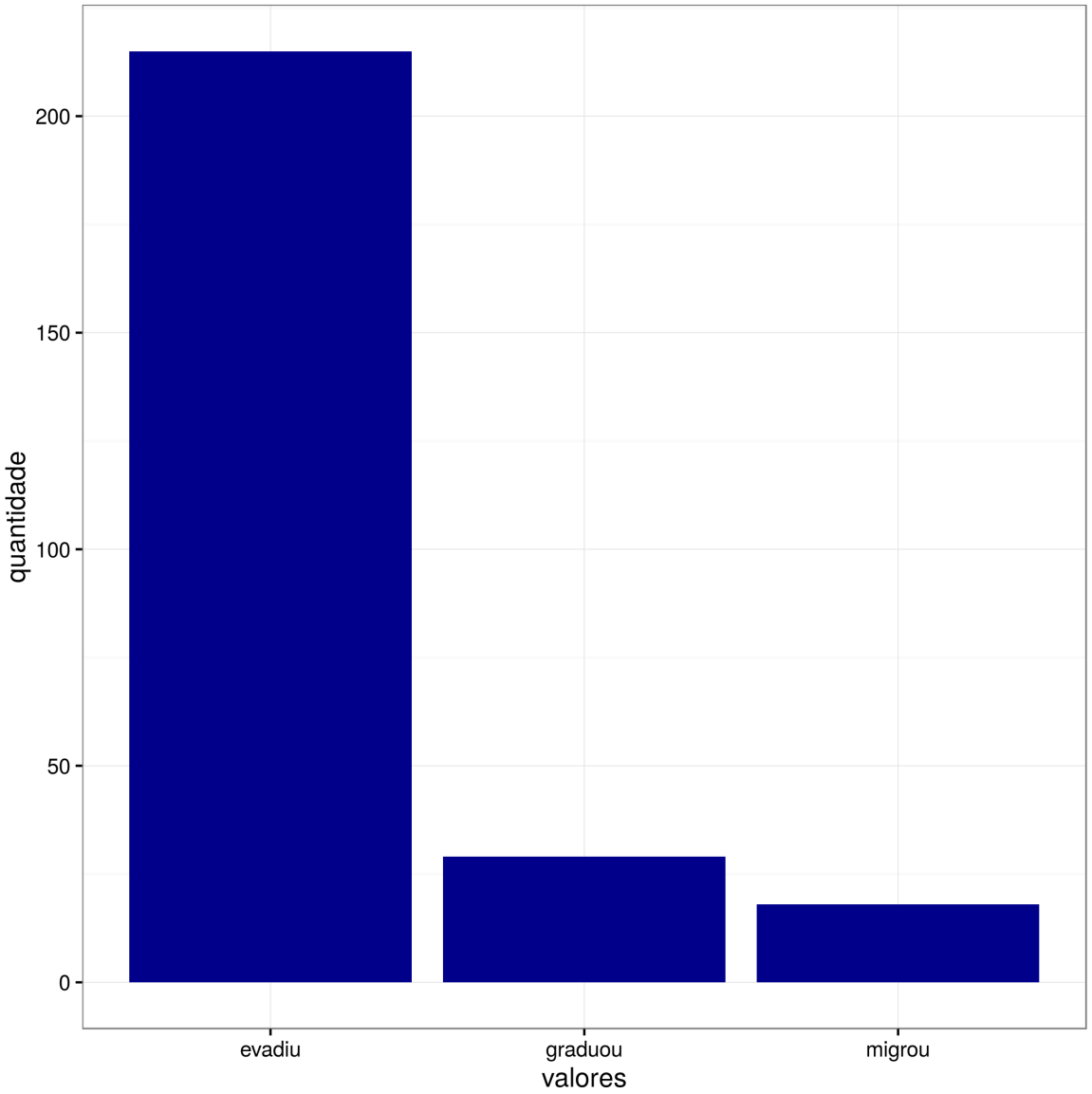
Alunos Jovens da Licenciatura - Forma de Saída



Alunos Jovens da Computação - Forma de Saída



Alunos Seniores - Forma de Saída



Eliminação de Atributos Relacionados

Utilizou-se o teste de Kendall para eliminar atributos que apresentassem mais de 80% de correlação.

Taxa de aprovação e taxa de reprovação estavam fortemente relacionados, de modo que eliminou-se o atributo taxa de reprovação.

Eliminação de Atributos Irrelevantes

Eliminação de atributos irrelevantes usando árvores de decisão:

- Para alunos jovens da licenciatura, atributo curso não era relevante.
- Para alunos seniores, os atributos curso, cota e taxa de trancamento não foram considerados relevantes.
- Para outras bases de dados, nenhum atributo foi classificado como irrelevante.

Ajuste de Parâmetros

Fez-se ajuste de parâmetros para:

- ANN
- *Naive Bayes*
- SVR

Para os demais algoritmos, utilizou-se o *default* da biblioteca `scikit-learn`.

Divisão em Treino e Teste

Dados de Treino: Alunos que ingressaram de 2000 até 2009.

Dados de Teste: Alunos que ingressaram de 2010 até 2016.

Necessidade da Divisão em Semestres

Para o problema de negócio considerado, sistema previsor deve ser capaz de calcular o risco de alunos evadirem tanto para alunos no início do curso quanto para estudantes mais adiantados.

Alguns atributos dos alunos, como a taxa de aprovação, mudam a cada semestre.

Funcionamento da Divisão em Semestres

Modelos são induzidos separadamente para cada semestre:

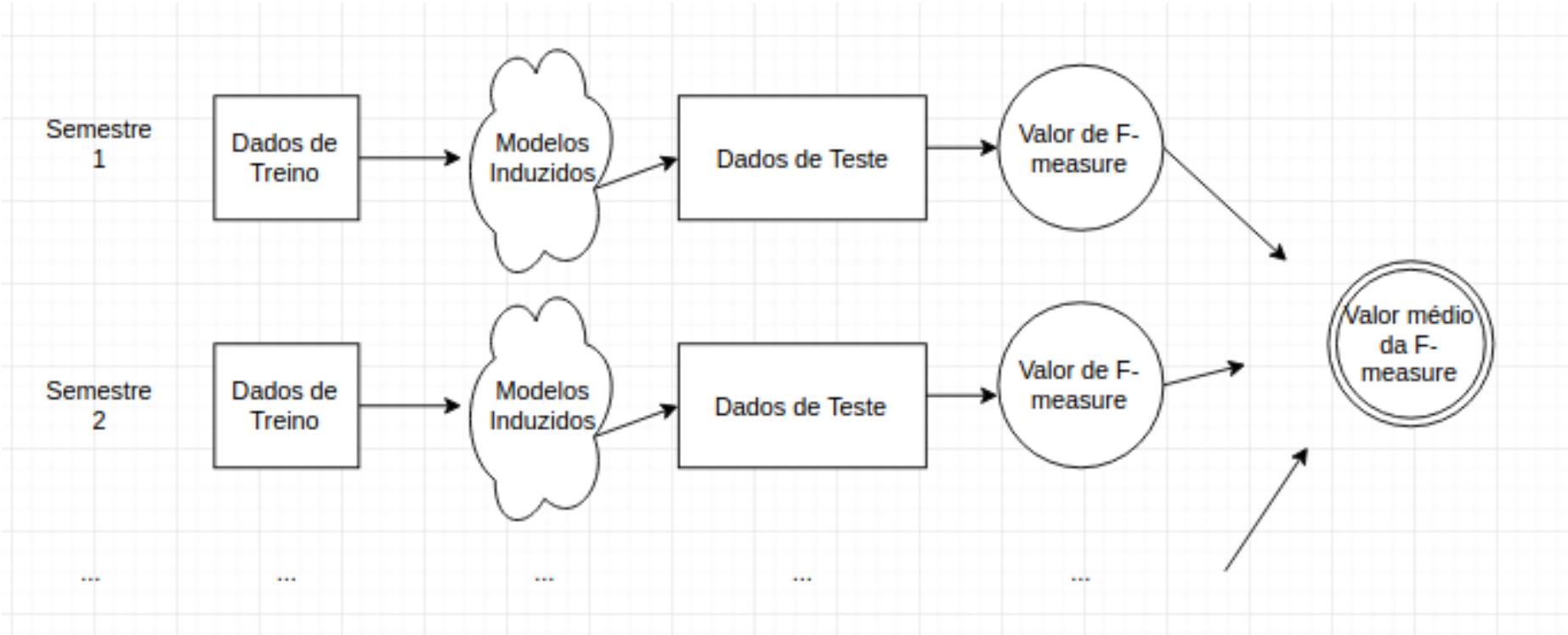
1. Inicialmente, modelos induzidos com dados do 1º semestre dos aluno do conjunto de treino e avaliados com dados do 1º semestre do conjunto de teste.
2. Repete-se o procedimento para os dados do 2º semestre dos alunos e assim por diante.

Processo de Avaliação de Desempenho

Processo:

1. Cada modelo induzido gera, para cada aluno em cada semestre ativo, uma tripla que indica a possibilidade do aluno concluir, evadir ou migrar.
2. Maior valor da tripla é usado como sendo a previsão do modelo.
3. Compara-se a previsão com o que realmente aconteceu com o aluno, de modo a verificar se o modelo acertou ou errou.

Diagrama para Avaliação de Desempenho



Avaliação de Desempenho

Para sumarizar o desempenho do algoritmo, calculou-se a média das *F-measures* de cada semestre:

$$\text{F-measure média} = \frac{\sum_{1 \leq i \leq n} \text{Fmeasure}_i}{n} \quad (1)$$

Avaliação de Desempenho

Modelos tem desempenho comparado entre si e com o ZeroR.

ZeroR é um classificador simples que sempre prevê a classe majoritária.

Resultados do Ajuste de Parâmetros para ANN

Melhores configurações para ANN de acordo com a base de dados:

Base de Dados	**Neurônios**	**Aprendizagem**
Jovens - FT	100	0.001
Jovens - Licenciatura	36	1.0
Jovens - Computação	36	0.001
Seniores	24	0.7

Resultados do Ajuste de Parâmetros para SVR

Melhores configurações para SVR de acordo com a base de dados:

Base de Dados	**Kernel**	**Penalização**
Jovens - FT	linear	1.0
Jovens - Licenciatura	linear	1.0
Jovens - Computação	rbf	1.0
Seniores	linear	1.0

Resultados do Ajuste de Parâmetros para Naive Bayes

Melhores configurações para Naive Bayes de acordo com a base de dados:

Base de Dados	**Distribuição dos Atributos**
Jovens - FT	Gaussiana
Jovens - Licenciatura	Bernoulli
Jovens - Computação	Multinomial
Seniores	Gaussiana

Resultados - Síntese

F-measure dos modelos de acordo com a base de dados:

Algoritmo	**Sen**	**J - FT**	**J - Lic**	**J - Comp**
ANN	0.62	0.76	0.85	0.74
Naive Bayes	0.28	0.56	0.76	0.65
Random Forest	0.70	0.73	0.85	0.76
Regressor Linear	0.75	**0.80**	**0.86**	**0.77**
SVR	**0.79**	0.76	0.82	0.70
ZeroR	0.61	0.64	0.70	0.60

Mau Desempenho do Naive Bayes

Mau desempenho do Naive Bayes justifica-se pelos atributos passados não poderem ser considerados condicionalmente independentes, dada a classe.

Bom Desempenho dos Modelos, Especialmente o Regressor Linear

De forma geral, algoritmos conseguiram ter melhor resultado que o ZeroR.

Bom desempenho obtido pelo regressor linear: *F-measure* em torno de 0.795.

Conclusão

Pesquisa aponta a viabilidade de usar aprendizagem de máquina para análise preditiva de alunos em risco de evasão na UnB.

Metodologia usada na pesquisa pode ser aplicada para outros cursos de graduação da UnB ou de outras universidades.

Trabalhos Futuros

Possíveis trabalhos futuros:

1. Implementação de ações na UnB com base no indicado pelo sistema previsor.
2. Testar desempenho do sistema em outras universidades.
3. Melhorar o sistema previsor, por exemplo incluindo novos atributos.

Agradecimentos e Resolução de Dúvidas

Obrigado por virem!

Alguma dúvida?

Referências I


http://www.correiobraziliense.com.br/app/noticia/cidades/2015/10/10/interna_cidadesdf,501999/evasoes-na-universidade-de-brasilia-causam-prejuizo-de-shtml.


John D Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press, 2015.


Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

Referências II


Päivi Kinnunen and Lauri Malmi.
Why students drop out cs1 course?
In *Proceedings of the second international workshop on Computing education research*, pages 97–108. ACM, 2006.


Hadautho Roberto Barros da Silva and Paulo Jorge Leitão Adeodato.
A data mining approach for preventing undergraduate students retention.
In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

Referências III


Cursos superiores que requerem maior capacidade de abstração algorítmica e conhecimento matemático apresentam maiores taxas de evasão? um estudo dos cursos de computação no brasil.
Technical report.