

Educação a Distância e Ciência de Dados: Desenvolvimento de Modelos Preditivos no Reconhecimento da Evasão Estudantil

Paulo R. V. do Carmo, Alan H. Costa, Sandro Rautenberg, Maria A. C. Knüppel,
Marta C.R. Anciutti

Universidade Estadual do Centro-Oeste
{pauloviviurka4, alanhenschel2, sandro.rautenberg, knuppelc,
martanciutti}@gmail.com

Agenda

Introdução

Objetivo

Materiais e Métodos

Resultado e Discussão

Obtenção e Ingestão de Dados

Exploração de Dados

Definição de Parâmetros

Implementação do Modelo

Tomada de Decisão

Considerações Finais

Trabalhos Futuros

Referências

Agradecimentos

Introdução

Com o aumento massivo de dados disponíveis, surgiram métodos computacionais para extrair informações úteis a respeito destes dados: **Ciência de Dados e Tomada de Decisão**.

Como as organizações voltadas ao Ensino a Distância (EaD) fazem uso de **Tecnologias da Informação e Comunicação** (TICs), estas podem se beneficiar de métodos da Ciência de Dados.

A **Evasão Estudantil** é um dos maiores problemas nas instituições EaD, chegando a 25% nacionalmente (TAMARIZ; SOUZA, 2015). E o Núcleo de Educação a Distância da Universidade Estadual do Centro-Oeste (NEaD/UNICENTRO) possui uma Evasão em torno de 54%.

Objetivo

Silva (2017) pontua a construção de **Modelos Preditivos** para minimizar a Evasão Estudantil.

No contexto da EaD estes modelos deveriam identificar previamente indícios de Evasão de seus alunos, suportando o processo de Tomada de Decisão.

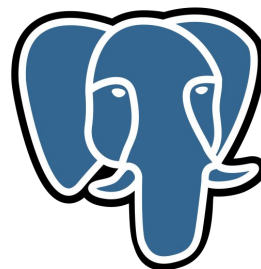
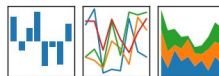
Neste trabalho investiga-se o desenvolvimento de Modelos Preditivos aplicados ao auxílio dos processos decisórios a Evasão Estudantil no NEaD/UNICENTRO.

Materiais e Métodos

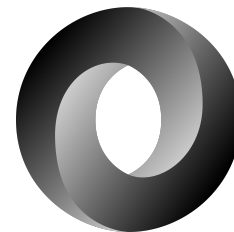


pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

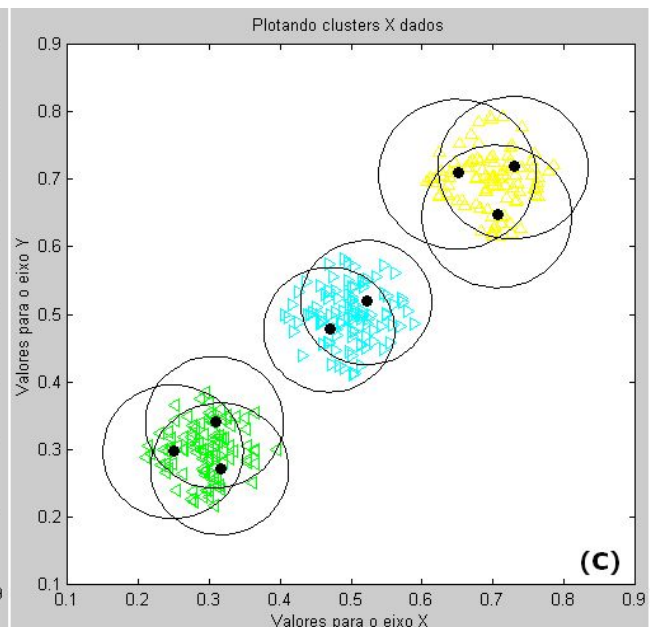
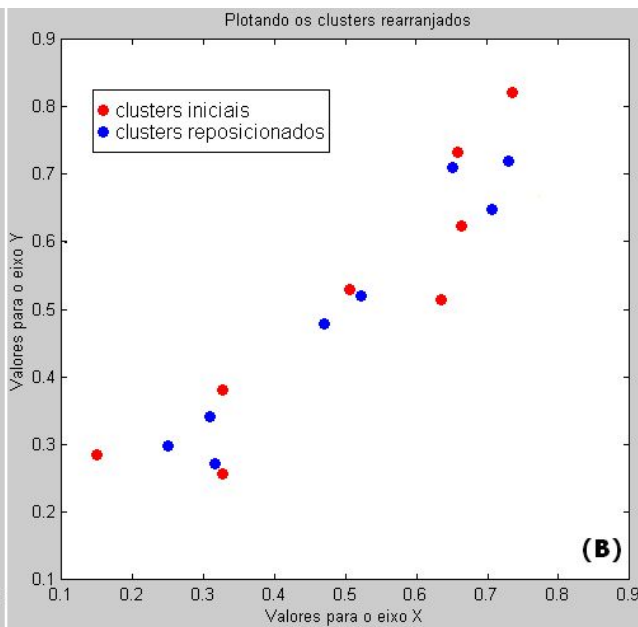
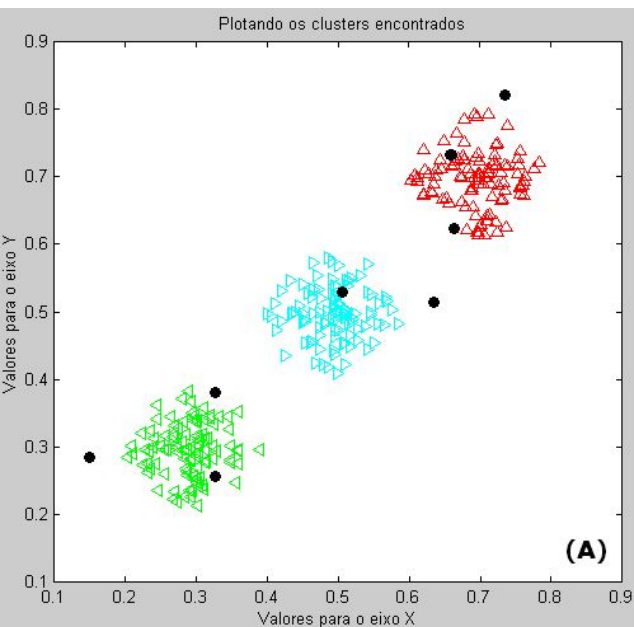


mongoDB



Materiais e Métodos

K-Means: Simples, eficiente e com bons resultados para agrupamentos de volumosas bases de dados (JAIN, 2010).



Resultado e Discussão

Obtenção e Ingestão de Dados

Obteve-se um arquivo de **backup dos registros** da base de dados do MOODLE junto à Coordenadoria de Tecnologia da Informação da UNICENTRO (COORTI/UNICENTRO), então estes dados foram importados em uma instância do SGBD **PostgreSQL**.

Exploração de Dados

Criação e análise de vários cenários. Estes resultaram na seleção da tabela de *log* do MOODLE, com as colunas: *actionid*, *targetid*, *userid* e *timecreated*.

Esta tabela de *log* personalizada foi então, importada para uma instância do MongoDB.

Resultado e Discussão

Definição de Parâmetros

Após a realização de duas *queries* no MongoDB, os dados foram reorganizados em 57 ações realizadas no MOODLE/NEaD/UNICENTRO filtradas por um dendograma para melhor definir os comportamentos de um aluno.

#	Ação	Descrição
1	<i>Created Discussion Subscription</i>	O total de respostas em postagens de fóruns
2	<i>Created Post</i>	O total de criações de postagens em um fórum
3	<i>Created Submission</i>	O total de submissões de tarefas
4	<i>Total Action</i>	O total de ações
5	<i>Total Created</i>	O total de ações de envio
6	<i>Total Viewed</i>	O total de ações de visualização
7	<i>Viewed Attempt</i>	O total de visualizações em tentativas de <i>quizz</i>
8	<i>Viewed Course</i>	O total de visualizações do curso
9	<i>Viewed Course Module</i>	O total de visualizações de tarefas
10	<i>Viewed Discussion</i>	O total de visualizações de postagens do fórum
11	<i>Viewed Grade Report</i>	O total de visualizações da grade de notas
12	<i>Viewed Message</i>	O total de visualizações de mensagens
13	<i>Viewed Submission Form</i>	O total de visualizações dos dados de envio de um <i>post</i> ou tarefa
14	<i>Viewed Submission Status</i>	O total de visualizações da situação de uma tarefa enviada

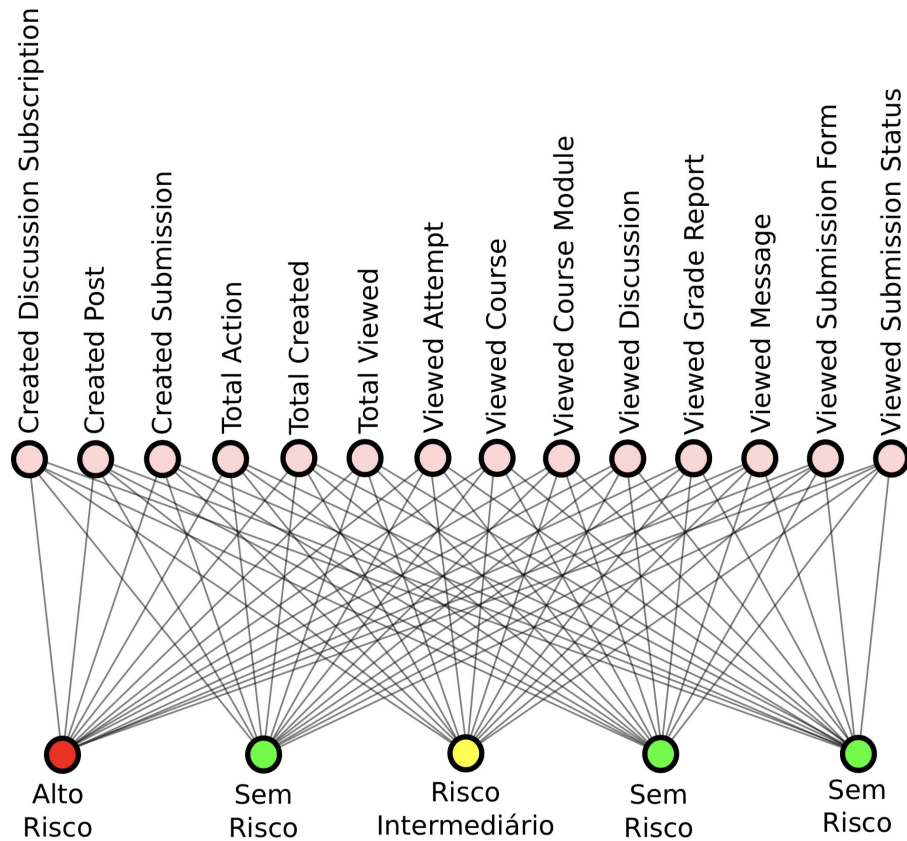
Resultado e Discussão

Implementação do Modelo

Foi desenvolvido um *script* em Python que utiliza o algoritmo *k-means* da biblioteca *Scikit-Learn*. Então ele foi executado **50** vezes para cada configuração testada com objetivo de obter a melhor e mais estável.

#	Taxa de Acerto (%)	Variância
02	62.483	5.0390e-08
03	78.626	4.6148e-08
04	89.382	1.0582e-06
05	95.822	3.8399e-07
06	89.067	7.4852e-06
07	84.964	8.6147e-05
08	84.650	5.6599e-05
09	85.076	1.6210e-04
10	87.509	2.7443e-04
11	89.490	1.1840e-03
12	95.687	5.3677e-05
13	93.973	2.2468e-05
14	93.918	8.9078e-06
15	91.284	2.3249e-04
16	91.643	2.2567e-04
17	94.153	3.8435e-04
18	94.933	3.1329e-04
19	95.829	1.0323e-04
20	95.783	3.3889e-05

Resultado e Discussão



Tomada de Decisão

Com a melhor configuração selecionada já é possível obter previsões a respeito dos discentes.

	Nome	Sobrenome	Evasão
1		CARPI	Alto Risco
2		RIBAS	Alto Risco
3		PEREIRA	Alto Risco
4		ZANDONA	Risco Intermediário
5		GIROTTI	Risco Intermediário
6		SOUZA	Risco Intermediário
7		QUADROS	Sem Risco
8		SILVA	Sem Risco
9		NASCIMENTO	Sem Risco

Considerações Finais

A partir do objetivo de criar modelos preditivos para auxiliar na Tomada de Decisão a respeito da Evasão Estudantil, do Ciclo de Vida da Ciência de Dados e das tecnologias utilizadas foi possível criar um modelo preditivo baseado no algoritmo *k-means*, que agrupa os alunos em três faixas: (i) alto risco; (ii) risco intermediário; e (iii) sem risco.

Trabalhos

Estudos e uso de outros algoritmos de agrupamento ou classificação para comparação;

Implementação de uma interface customizada para o *workflow* de gestão do NEaD/UNICENTRO.

Futuros

Referências

BOCK, T. What is a Dendogram? 2019. Disponível em: <<https://www.displayr.com/what-is-dendrogram>> . Acesso em: 11 jun 2019.

BUGNION, P.; MANIVANNAN, A.; NICOLAS, P. R. Scala: Guide for Data Science Professionals. Birmingham: Packt Publishing, 2017.

ECONOMIST. The world's most valuable resource is no longer oil, but data. Disponível em: <<https://goo.gl/AW4XsF>>. Acesso em: 09 jun 2019.

GARTNER. What is Big Data? – Gartner IT Glossary – Big Data. Disponível em: <<https://goo.gl/GwQWLA>>. Acesso em: 23 mai 2019.

JAIN, A. K. Data clustering: 50 years beyond k-means. Pattern Recognition Letters, v.31, n. 8, p. 651–666, 2010.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C. B.; HUNG, A. Big data: The next frontier for innovation, competition, and productivity. Disponível em: <<https://goo.gl/Vg2G2U>>. Acesso em: 23 mai 2019.

MOODLE. Features – MoodleDocs. Disponível em: <<https://docs.moodle.org/35/en/Features>>. Acesso em: 23 mai 2019.

Referências

PYTHON.ORG. What is Python? Executive Summary | Python.org. Disponível em: <<https://goo.gl/QYghos>>. Acesso em: 28 mai 2019.

RAUTENBERG, S. et al. Evasão Estudantil e Ciência de Dados: primeiros passos de uma pesquisa aplicada no contexto da Educação a Distância da Universidade Estadual do Centro-Oeste. In: DIAS, G. A.; DUTRA, M. L. (Ed.) Anais do WIDAT'2018 - II Workshop de Informação, Dados e Tecnologia. João Pessoa: PPGCI-UFPB, 2018. p. 368.

SILVA, F. C. da. Gestão da Evasão na EaD: Modelo Estatístico Preditivo para os Cursos de Graduação a Distância da Universidade Federal de Santa Catarina. Florianópolis, 2017. 137 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico. Programa de PósGraduação em Administração.

TAMARIZ, A. D. R.; SOUZA, M. de. Educação a distância no Brasil: perspectivas para redução na evasão de alunos matriculados. Revista Científica Linkania Master, v. 5, n.1, 2015.

TAURION, C. O que é um modelo preditivo? Disponível em: <https://cio.com.br/o-quee-um-modelo-preditivo>. Acesso em: 09 jun 2019.

van der AALST, W. Data Scientist: The Engineer of the Future. In: Interoperability of Enterprises Systems and Applications Conference (I-ESA'2014), 2014, Albi-France, Proceedings... Heidelberg: Springer, 2014.

Agradecimentos

À Secretaria da Ciência, Tecnologia e Ensino Superior (SETI/PR) pelo suporte financeiro (Projeto - Implementação da Universidade Virtual do Paraná – UVPR/SETI, Termo de Cooperação nº 145/2017, vinculado a unidade gestora do Fundo Paraná).

E a Universidade Aberta do Brasil (UAB), por ter financiado a viagem a Brasília para o evento.



pauloviviurka4@gmail.com