Data Warehouse Educacional: Uma visão sobre a Evasão no Ensino Superior

G.A.S.Santos[1], A.L.Bordignon[2], D.B.Haddad[1], D.N.Brandão[1], L.Tarrataca[1], K.T.Belloze[1]

[1] PPCIC - Programa de Pós-graduação em Ciências da Computação (CEFET-RJ)
Caixa Postal 20271-204 – 455, Maracanã, Rio de Janeiro - RJ

[2]Instituto de Matemática, Universidade Federal Fluminense, Niterói - RJ

gustavo.santos@eic.cefet-rj.br, {diego.brandao, kele.belloze}@cefet-rj.br

Abstract. *Dropout is one of the main challenges of educational institutions. In this sense, this paper presents the implementation of a Data Warehouse for data analysis and decision making in a higher education institution in Brazil. The presented Data Warehouse allows integrated views that assist in analysis such as: 1) distribution of students' performance coefficient; 2) identification of student profiles and 3) insight into student achievement by locality. These analyzes are intended to assist academic management in identifying patterns that lead to dropout and thus to promote directions for preventive actions and mainly to expand the use of this analytical database developing new solutions, such as predictive models.*

Resumo. *A evasão mostra-se como um dos principais desafios das instituições de ensino. Nesse sentido, este trabalho apresenta a implementação de um Data Warehouse para análise de dados e auxílio à tomada de decisão em uma instituição de ensino superior do Brasil. O Data Warehouse apresentado permite visões integradas que auxiliam em análises de: 1) distribuição do coeficiente de desempenho do alunos; 2) identificação dos perfis dos estudantes e 3) um dashboard sobre o rendimento dos alunos por localidade. Essas análises têm como propósito auxiliar a gestão acadêmica na identificação de padrões que acarretam na evasão e, desta forma, promover direcionamentos para medidas preventivas e, principalmente, expandir o uso deste banco de dados analítico para desenvolver novas soluções, como por exemplo, modelos preditivos.*

1. Introdução

A gestão acadêmica das instituições de ensino superior compreende diversas atividades. No caso das Instituições Federais de Ensino Superior (IFES) no Brasil, essas atividades consistem em ensino, pesquisa e extensão. Além disso, as IFES, em sua maioria, possuem políticas e programas socioeconômicos com o objetivo de auxiliar e dar suporte a essas atividades. É importante notar que há vários desafios para a gestão acadêmica, muitos dos quais requerem cuidadosa atenção como evasão, retenção e vagas ociosas [Speller et al. 2012].

Especificamente, a evasão é uma situação que ocorre quando os estudantes ocupam vagas e se desassociam das universidades sem concluir o curso em que se matricularam. O custo decorrente devido a evasão é muito maior do que o desejado pelo

governo [dos Santos Baggi and Lopes 2011]. Dados do Censo de Educação Superior no Brasil, em 2016, mostraram que dentre as 10,6 milhões de vagas oferecidas nos cursos de graduação, 26,0% dessas foram oriundas de evasão [INEP 2016]. Por outro lado, o relatório da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) [OECD 2016] revela que o custo médio anual de um estudante do ensino público superior no Brasil, em 2013, foi de US$13.539,90.

Na busca de respostas para o problema de evasão, algumas IFES adotam soluções tecnológicas, baseadas em sistemas de apoio a tomada de decisão. Tais sistemas, dentre os quais o *Data Warehouse* assume um papel de destaque [Olszak and Ziemba 2007], contribuem com a análise de dados históricos e modelos de predição. Esses podem apoiar significativamente na compreensão dos problemas institucionais e, em particular, suas causas e consequências [Shim et al. 2002].

Este trabalho descreve a implementação de um *Data Warehouse* Educacional (EDW - *Educational Data Warehouse*), de modo a fornecer um banco de dados analítico capaz de apresentar análises integradas em uma IFES. O objetivo deste EDW é que as análises geradas apoiem a gestão acadêmica na identificação de padrões que acarretam evasão e desta forma sejam desenvolvidas medidas preventivas em relação ao tema.

Além desta introdução, este artigo está estruturado da seguinte forma: a seção 2 apresenta a implementação do EDW. A seção 3 apresenta o estudo de caso sobre a evasão juntamente com uma apresentação de gráficos gerados pelo sistema. Finalmente, a seção 4 apresenta as considerações finais e trabalhos futuros.

2. O *Data Warehousing* Educacional

Na maioria dos projetos de DW, o desafio consiste em planejar de maneira eficiente o desenvolvimento geral do sistema. A Figura 1 apresenta o fluxo de dados para o processo que incorpora o contexto informacional necessário para a construção do EDW.

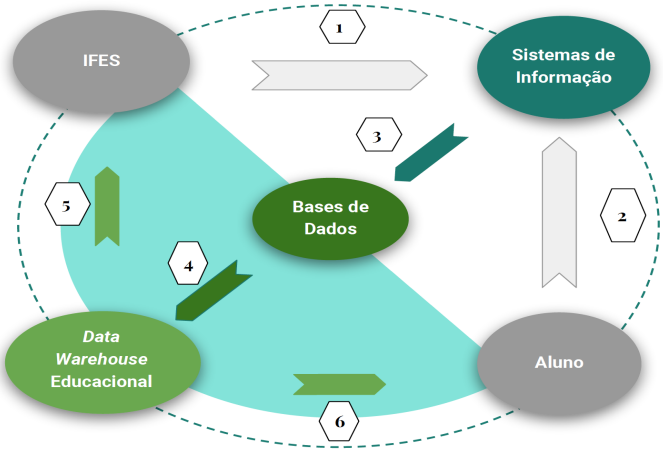

Figura 1. Fluxo de Dados para o *Data Warehouse* Educacional.

Neste processo, estão representadas as seguintes entidades: IFES, Aluno, Sistemas de Informação, Bases de Dados e *Data Warehouse* Educacional. As etapas (1) e (2) lidam, respectivamente, com as interações entre estudantes e funcionários da universidade. Isso é feito no sistema de informação (SI) da universidade por meio de atividades

com foco no gerenciamento acadêmico dos cursos e, consequentemente, nos alunos. Na etapa (3), as informações inseridas no SI são armazenadas em bases de dados. A etapa (4) representa o objetivo principal deste trabalho que será descrito nas subseções a seguir. Finalmente, as etapas (5) e (6) se relacionam com a aquisição de conhecimento que é obtida por meio do banco de dados analítico implementado pelo EDW. A etapa (5) fornece as informações analíticas que apoiam o processo de tomada de decisão. A etapa (6) fornece informações que podem ser consultadas, bem como recomendações baseadas no perfil do aluno, no curso, e nas interações que são armazenadas no EDW. O desenvolvimento do EDW, focado no tema evasão, é baseado nos requisitos funcionais apresentados na Tabela 1.

Tabela 1. Requisitos funcionais para o EDW

Item	Descrição
01	Analisar a evasão com relação ao desempenho acadêmico
02	Analisar a evasão com base nos perfis dos estudantes
03	Analisar a evasão considerando a localidade de curso

2.1. Fontes de dados: descrição dos Sistemas de Informação

O conjunto de SI escolhido para o EDW é apresentado a seguir. O sistema de identificação única é uma aplicação da IFES que visa centralizar os dados das pessoas que têm ou tiveram algum vínculo com a universidade e também todas as informações acadêmicas referentes aos cursos de graduação. O sistema de iniciação científica foi desenvolvido para administrar: (i) o processo de submissão de projetos de pesquisa; (ii) seleção de projetos e candidatos; (iii) concessão de bolsas de iniciação científica e (iv) avaliação de projetos. O sistema de bolsas foi desenvolvido com o objetivo de administrar bolsas de assistência estudantil. O sistema de monitoria foi desenvolvido com o objetivo de facilitar o processo de submissão das candidaturas para monitoria de disciplinas. Todos esses sistemas foram integrados em um banco de dados analítico por meio dos procedimentos de modelagem multidimensional: Extração, Transformação, Carregamento (*Extract, Transform, Load* - ETL) e visualização de dados.

2.2. Construção do EDW

O EDW é baseado em um modelo multidimensional, o qual permite que os dados sejam integrados e visualizados sob várias dimensões [Inmon and Linstedt 2014]. A fim de permitir que várias questões de negócio possam ser respondidas por meio da evidência dos fatos, é necessário consolidar outras perspectivas de informação, as quais são caracterizadas como dimensões. Para atender aos requisitos de negócios descritos na Tabela 1, foram criadas oito tabelas de dimensões, representando as seguintes entidades: 'Aluno', 'Histórico do Aluno', 'Bolsa', 'Bolsista', 'Curso', 'Acompanhamento do Aluno', '*Status* do Aluno' e 'Tempo'.Além disso, foram criadas duas tabelas de fatos representando as entidades 'Evasão' e 'Histórico Acadêmico'.

Para popular as dimensões e fatos, aplicou-se o processo ETL, o qual ocorreu em várias etapas. Neste trabalho, a primeira etapa é referente à extração dos dados dos sistemas acadêmicos, a qual gera arquivos no formato `.csv` (*comma-separated values*. Em seguida, é executado um procedimento que carrega esses arquivos em um banco de

dados relacional. Feito isso, os dados são ajustados e corrigidos com base nos requisitos e transformados em um banco de dados multidimensional.

3. Estudo de Caso com Foco na Evasão

Nesta seção, alguns resultados obtidos por meio da análise de dados são apresentados. O conjunto de dados produzido pelo EDW contém informações sobre os estudantes como as notas do Exame Nacional do Ensino Médio (ENEM), histórico acadêmico, registro de ação afirmativa (políticas sociais), cor da pele e dados sociodemográficos. O EDW contêm informações sobre aproximadamente 80.000 alunos dos 106 cursos de graduação que são oferecidos pela IFES. Tais dados são fornecidos pela tabela "FATO_EVASAO" do modelo EDW. Este conjunto de dados compreende alunos desde os anos de 2005 até 2018.

Na Figura 2, é possível identificar a distribuição dos alunos por localidade (diferentes *campus* da instituição), considerando o percentual referente à faixa de coeficiente de rendimento (CR) do estudante e localidade. É apresentado também o percentual de CR total da Instituição, o qual mostra que um pouco mais de 37% de alunos apresenta CR abaixo de 4,0.

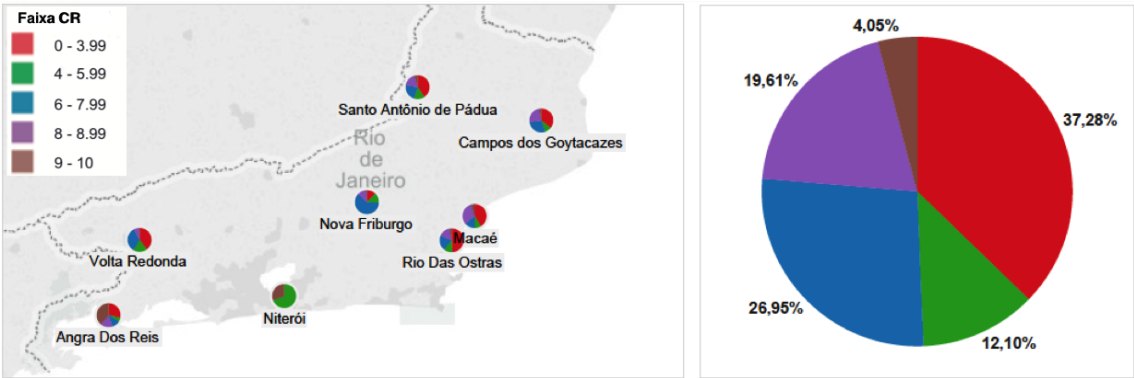

Figura 2. Distribuição do rendimento dos alunos por localidade e total.

Durante a análise exploratória dos dados, algumas perspectivas foram observadas. A primeira perspectiva foi de que o conjunto de dados continha uma representação de classes desbalanceadas, com 76% de estudantes evadidos e 24% graduados. Dessa forma, a Tabela 2 apresenta detalhes quantificando a amostra ("Qtd.") e exibindo para cada classe a mediana dos valores dos atributos: "Idade", "SemestreFinal" (último semestre do ano cursado), "CR (Coeficiente de Rendimento)", "CargaHorariaCursada"(CargaHor) e "TempoPermanencia". Com base nos resultados, é possível observar que o padrão de evasão destaca-se no primeiro semestre de cada ano letivo por alunos com idade próxima a 25 anos, CR mediano de 3,4 e com um total de carga horária cursada próximo 240 horas de currículo (3 anos).

A segunda perspectiva pode ser identificada por meio do histograma apresentado na Figura 3. Nele é possível perceber que há uma frequência maior da amostra de alunos com baixo desempenho nas regiões de CR variando de '0' até '3', sendo boa parte desta frequência de 34% constituída por alunos evadidos, em geral, ocorrendo esta evasão no primeiro ano de ingresso. Essa observação foi possível a partir das análises realizadas no EDW.

Tabela 2. Mediana dos atributos baseada nos perfis “Evadido”e “Graduado”.

Classe	Qtd.	Idade	SemestreFinal	CR	CargaHor	TempoPermanencia
Evadido	9852	25	1º	3.4	240	3 anos
Graduado	3117	24	2º	8.3	3199	5 anos

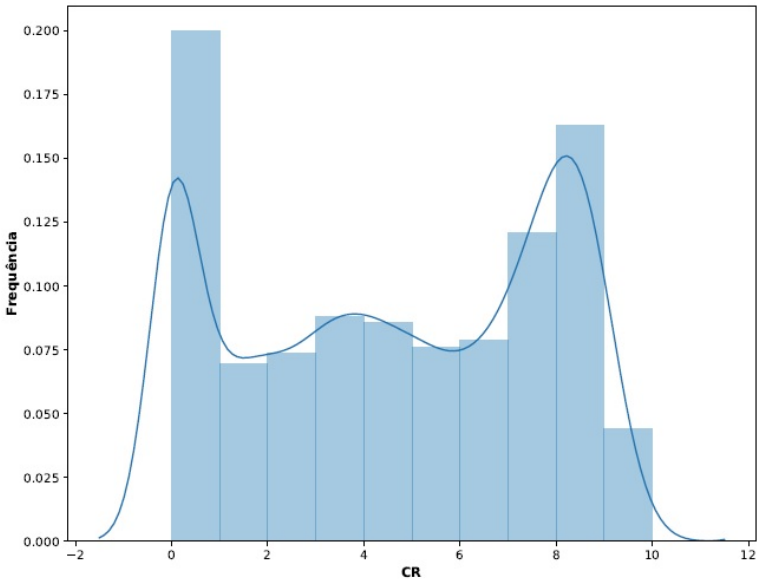

Figura 3. Histograma do CR de toda amostra, incluindo graduados e evadidos.

A partir das perspectivas anteriores, avaliou-se também a necessidade de medir a correlação entre os atributos gerados pelo conjunto de dados observado. Nesse caso, foram priorizados os atributos de ‘CR’ e ‘StatusFormacao’, a fim de avaliar a relação de influência dos demais atributos no aspecto de desempenho acadêmico e no *status* de formação do aluno (‘evadido’ ou ‘graduado’). A Tabela 3 apresenta as correlações, observa-se que as notas do ENEM tem uma correlação baixa com relação ao CR do aluno bem como se ele vai evadir ou não (‘StatusFormacao’).

4. Considerações Finais

Neste trabalho foi apresentado um *Data Warehouse* Educacional para atender a demanda de um sistema de apoio à tomada de decisão sobre uma visão integrada referente ao tema evasão. A proposta mostrou-se como solução capaz de conceder subsídios que podem auxiliar a gestão acadêmica na identificação de padrões que impactam na evasão.

Por meio das análises realizadas foi percebido que as notas do ENEM têm uma influência significativa no CR e *status* de formação do aluno, assim como a carga horária cursada e o tempo de permanência do aluno. Esses são indicadores altamente relevantes na análise de evasão. Além disso, percebeu-se a necessidade de ajustes quanto à entrada de novas informações consideradas importantes, mas que ainda não existem no sistema de gestão acadêmica. Essas informações são referentes ao histórico do ensino fundamental e médio dos alunos de graduação. Foi evidenciado que este tipo de informação implicará em uma melhor percepção quanto ao perfil do aluno de graduação, pois representa boa parte de sua formação escolar e com isso, pode ser possível efetuar uma análise preditiva no primeiro período do aluno.

Tabela 3. Correlação dos atributos “CR” e “StatusFormacao” entre os demais atributos.

Atributo	CR	StatusFormacao
EnemLinguagem	0.141993	0.099262
EnemHumanas	0.094311	0.018340
EnemCiencias	0.103915	0.040897
EnemMatematica	0.067701	0.036235
EnemRedacao	0.139197	0.095648
IdTurno	0.076495	−0.014528
IdTurnoAtual	0.008767	−0.057659
CR	1.000000	0.633797
AnoIngresso	−0.093145	−0.208442
SemestreIngresso	−0.108587	−0.101707
Idade	−0.145779	−0.067144
CargaHorCursada	0.702870	0.901757
Trancamento	0.037985	−0.008887
TempoPermanencia	0.539021	0.480889
StatusFormacao	0.633797	1.000000

Como propostas de trabalhos futuros, é factível utilizar técnicas de aprendizado de máquina para identificar o conjunto de atributos mais relevante para um aluno, e também classificar os alunos baseando-se numa estratégia de análise de risco de evasão, conforme o perfil dos estudantes. Essas propostas de trabalhos podem possibilitar uma maior efetividade na redução dos impactos da evasão e, consequentemente, uma melhor gestão de recursos.

Referências

dos Santos Baggi, C. A. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16(2).

INEP (2016). Censo da educação superior - notas estatísticas 2016. In *Diretoria de Estatísticas Educacionais (DEED)- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)*. Acessado em 13 de Outubro de 2018.

Inmon, W. H. and Linstedt, D. (2014). *Data architecture: a primer for the data scientist: big data, data warehouse and data vault*. Morgan Kaufmann.

OECD (2016). *Education at a Glance 2016*.

Olszak, C. M. and Ziemba, E. (2007). Approach to building and implementing business intelligence systems. *Interdisciplinary Journal of Information, Knowledge, and Management*, 2(1):135–148.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., and Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision support systems*, 33(2):111–126.

Speller, P., Robl, F., and Meneghel, S. M. (2012). Desafios e perspectivas da educação superior brasileira para próxima década. *Oficina de Trabalho. p. 164, 2012. ISBN: 978-85-7652-171-6*.