Evasão Estudantil e Ciência de Dados: primeiros passos de uma pesquisa aplicada no contexto da Educação a Distância da Universidade Estadual do Centro-Oeste

Student Evasion and Data Science: first steps of an applied research to the Distance Learning context of the Universidade Estadual do Centro-Oeste

Sandro Rautenberg[1], Alan H. Costa[1], Paulo R. V. do Carmo[1], Renan A. M. Nutse[1], Maria A. C. Knüppel[1], Marta C. R. Anciutti[1]

(1) Universidade Estadual do Centro-Oeste, Alameda Élio Antonio Dalla Vecchia, 838,
Vila Carli, Guarapuava-PR, CEP 85040-167,
{sandro.rautenberg, alanhenschel2, pauloviviurka4, renanmnutse, knuppelc, martanciutti}@gmail.com

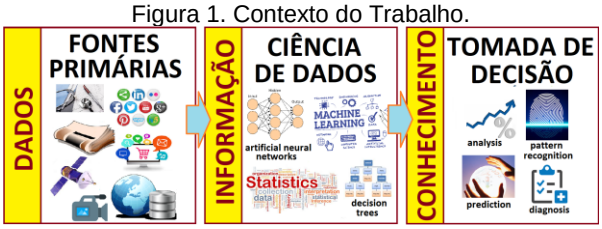
Resumo: A Educação a Distância é uma alternativa de ensino para indivíduos que necessitam horários e locais de estudo flexibilizados. Dentre as instituições que oferecem essa modalidade de ensino, a Universidade Estadual do Centro-Oeste (UNICENTRO) ministra cursos de graduação a distância por intermédio do Núcleo de Educação a Distância (NEaD). Em uma análise realizada, o NEaD destaca um alto índice de evasão de alunos. Para melhor entender essa realidade, institucionalmente, formalizou-se um projeto de extensão empregando tecnologias da Ciência de Dados. Iniciado em Junho/2018, o projeto visa inferir alguns elementos de análise e apoiar decisões pedagógicas a partir dos microdados do Ambiente Virtual de Aprendizagem MOODLE. Neste sentido, este artigo tem como objetivo relatar os primeiros passos do referido empreendimento. Decorridos seis meses de execução, dentre os resultados preliminares alcançados, pontua-se: **i)** o estabelecimento de um *workflow* para desenvolver os cenários de exploração e visualização de dados; **ii)** a concepção de uma camada para ingestão de dados, mitigando os efeitos indesejáveis de versionamento de modelos de dados do MOODLE; e **iii)** a formalização de um meio de comunicação para com os gestores do NEaD/UNICENTRO, quando da prospecção de cenários de exploração de dados.
Palavras-chave: Ciência de Dados; Educação à Distância; Evasão Estudantil; MOODLE; Ambientes Virtuais de Aprendizagem.

Abstract: Distance Learning is a teaching alternative for individuals who need flexible schedules and places for studying. Among the institutions that offer this kind of education, the Universidade Estadual do Centro-Oeste (UNICENTRO) teaches undergraduate distance courses through the Núcleo de Educação a Distância (NEaD). In an initial analysis, the NEaD highlights a high evasion rate among students. To better understand this reality, institutionally, a project was formalized using Data Science technologies. Initiated in June/2018, the project aims to infer some elements from the microdata of the MOODLE Learning Management System for supporting pedagogical decisions. In this sense, this article aims to report the first steps of this effort. After six months of execution, the preliminary results achieved include: **i)** the establishment of a workflow for developing data exploitation scenarios; **ii)** the design of a data layer, mitigating the undesirable effects of versioning of the MOODLE's data model; and **iii)** the formalization of a communication model with the NEaD/UNICENTRO managers, when prospecting data exploitation scenarios.
Keywords: Data Science; Distance Learning; Student Evasion; Moodle; Learning Management System.

1 Introdução

Com o avanço do uso da Internet, produz-se cada vez mais dados em várias plataformas digitais. Diversos aplicativos e dispositivos interconectados (computadores, *smartphones*, etc.) relacionam uma série de eventos (van der AALST, 2014), armazenando enormes quantidades de registros, sinais, imagens, vídeos e *posts*. Por conseguinte, os dados são abundante e velozmente produzidos, podendo servir como matéria-prima à Tomada de Decisão (ECONOMIST, 2018). Por isso, o desenvolvimento de soluções computacionais que obtém insumos de conhecimento de volumes de dados torna-se foco de investimento das organizações.

Figura 1. Contexto do Trabalho.



Fonte: Dados da Pesquisa.

Atualmente, esse contexto relaciona os conceitos de Ciência de Dados e Tomada de

Decisão como segue. Conforme ilustrado na Figura 1, tem-se as fontes de dados cujas características dificultam a captura, o armazenamento, o gerenciamento, a análise e a exploração de dados por parte de ferramentas computacionais tradicionais (GARTNER, 2018; MANYIKA *et al*., 2011). Para auxiliar a análise e a exploração de dados, pode-se recorrer à Ciência de Dados (do inglês, *Data Science*). A Ciência de Dados é caracterizada como uma camada de métodos devotados à extração de informação útil a partir de complexas e dinâmicas bases de dados (BUGNION; MANIVANNAN; NICOLAS, 2017). E, por conseguinte, ao recuperar informação útil, pode-se auxiliar os gestores no desempenho de suas atividades decisórias.

Diante dessa visão, este trabalho parte do pressuposto que as organizações voltadas ao Ensino a Distância (EaD)[1] podem se beneficiar dos métodos e tecnologias da Ciência de Dados para melhorar seus processos. Principalmente, em face destas organizações adotarem Tecnologias da Informação e Comunicação (TICs), produzindo conjuntos de dados a serem explorados, a partir dos: **i)** gerenciamento da vida acadêmica de seus discentes; e **ii)** uso de meios digitais para o compartilhamento de material educacional. No tocante às referidas organizações, um dos problemas principais enfrentados é a Evasão Estudantil[2], fenômeno cuja taxa média anual alcança 25% dos alunos matriculados (ABED, 2014).

Pontualmente, esse problema também é enfrentado pelo Núcleo de Educação à Distância da Universidade Estadual do Centro-Oeste (NEaD/UNICENTRO). O NEaD/UNICENTRO tem como incumbência a difusão do conhecimento tecnológico e científico para os diferentes segmentos sociais, por intermédio de projetos, atividades e programas de EaD. Mediante seus polos avançados, essa organização é capilarizada em 51 municípios no Estado do Paraná (NEAD, 2018), circunvizinhando quase todos os 399 municípios da referida unidade federativa. Institucionalizado no ano de 2005, atualmente, o NEaD/UNICENTRO enfrenta o desafio de minimizar seus índices de Evasão Estudantil, que está em torno de 54%.

Tecnologicamente, para minimizar os índices de Evasão Estudantil, Silva (2017) pontua a construção de modelos preditivos. Tais modelos deveriam identificar preventivamente a probabilidade de um aluno evadir. Segundo a autora, com a utilização de modelos preditivos, pode-se abstrair informações confiáveis para suporte ao processo de Tomada de Decisão dos gestores, aumentando os índices de retenção discente. Portanto, admite-se que o emprego de métodos e tecnologias da Ciência de Dados é pertinente ao estabelecimento de modelos preditivos da Evasão Estudantil.

Diante disso, para melhor entender e tratar a realidade da Evasão Estudantil inerente ao NEaD/UNICENTRO, um Projeto de Extensão[3] foi formalizado em Junho de 2018. Em sua essência, tal empreendimento visa criar modelos de análise e visualização de dados e informações da evasão, baseando-se em metodologias e tecnologias da Ciência de Dados. Cabe ressaltar que objetivos do referido projeto permeiam, principalmente: **i)** a abstração dos motivos da desistência dos alunos matriculados em cursos de graduação na modalidade EaD; e **ii)** o amparo em ações afirmativas para aumentar os índices de retenção de estudantes.

2 Objetivo

Diante o exposto, o objetivo deste capítulo de livro é relatar os resultados parciais alcançados no tocante aos primeiros meses

[1] "[...] a EaD pode ser entendida como agente de inovação dos processos de ensino e aprendizagem, incentivando a incorporação de novas Tecnologias da Informação e Comunicação (TICs) aos métodos didático-pedagógicos, e possibilitando ao cidadão o acesso à educação superior pública e de qualidade, a partir da democratização do acesso à educação [...]" (MORÉ *et al.*, 2010).

[2] A Evasão Estudantil é compreendida como "a saída definitiva do aluno de seu curso de origem, sem concluí-lo" (ANDIFES; ABRUEM; SESu/MEC, 1996).

[3] RESOLUÇÃO Nº 022 - PROEC/UNICENTRO, DE 16 DE JULHO DE 2018 - Aprova o Projeto de Extensão Data Science e evasão estudantil: um estudo do caso no contexto do Núcleo de Educação a Distância da Universidade Estadual do Centro-Oeste, na modalidade de Ação de Extensão, na categoria de Projeto de Extensão (UNIVERSIDADE ESTADUAL DO CENTRO-OESTE - UNICENTRO, 2018).

de atividades realizadas do projeto de extensão institucionalizado.

3 Materiais e Métodos

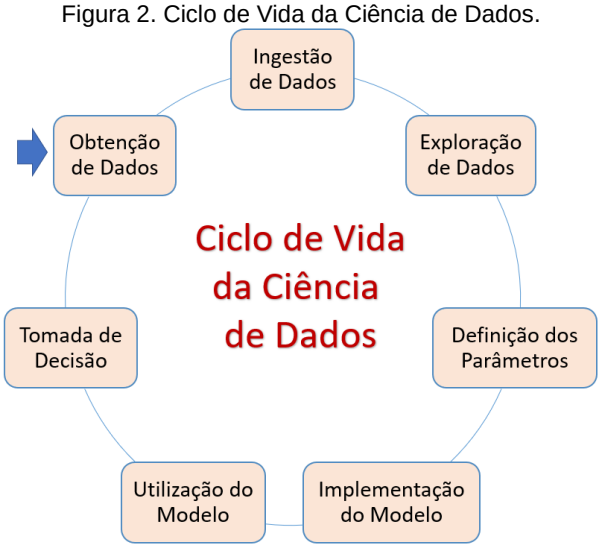
Para subsidiar as atividades de EaD, o NEaD/UNICENTRO utiliza o Ambiente Virtual de Aprendizagem MOODLE (*Modular Object Oriented Distance LEarning*) como plataforma de suporte (versão 3.x). Este é um sistema gratuito que permite a gestão de cursos *online*, mediante o gerenciamento da comunicação entre atores (docentes, discentes, tutores, etc.) e dos objetos de aprendizagem (*ebooks*, vídeo aulas, atividades, fóruns, mensagens, etc.). Assim, o MOODLE oferece um ambiente dinâmico que oportuniza o aprendizado a qualquer momento e em qualquer lugar, ao utilizar a Internet como plataforma de comunicação (MOODLE, 2018).

Atualmente, a base de dados do MOODLE do NEaD/UNICENTRO comporta cerca de 16 *gigabytes* de dados, contendo os registros de 5.785 alunos de EaD (formados, desistentes ou em formação). De acordo com seu modelo relacional, na referida base de dados, os registros da interação dos usuários para com os objetos disponibilizados são armazenados minuciosamente em suas 300 tabelas de dados. Diante dessa riqueza de detalhes em seus registros, a base de dados do MOODLE é um insumo pertinente ao discernimento da Evasão Estudantil inerente ao NEaD/UNICENTRO.

Neste sentido, para promover a análise e a exploração de dados/informação a partir da base de dados do MOODLE, adota-se a proposta de Bugnion; Manivannan e Nicolas (2017) como procedimento metodológico da Ciência de Dados. Os referidos autores sugerem sete passos (Figura 2), conforme segue:

- **Obtenção de Dados**. Preconiza as tarefas de avaliação e seleção de dados primários e seus metadados, por exemplo, a partir: do processamento de arquivos de texto; do monitoramento de uma rede de sensores; de consultas a bases de dados de sistemas legados; ou de API (*Application Programming Interface*) da web.
- **Ingestão de Dados**. Trata das transformação e carga dos dados primários advindos de fontes diferentes e formatos diversificados em uma base de dados centralizada. Isso implica em organizar, representar e inserir dados pré-processados em um repositório de dados principal, mitigando os esforços futuros na geração de informação relevante.

Figura 2. Ciclo de Vida da Ciência de Dados.



Fonte: baseado em (BUGNION; MANIVANNAN; NICOLAS, 2017) [tradução dos autores]

- **Exploração de Dados**. Privilegia a execução de estudos preliminares de modo a estabelecer conjecturas iniciais e entendimentos superficiais dos dados disponibilizados em relação à informação requisitada. Assim, estabelece-se um fluxo de trabalho (*workflow*) para relacionar os dados em busca da informação relevante.
- **Definição dos Parâmetros**. Está intimamente ligado às escolhas necessárias para o emprego do(s) algoritmo(s) de Aprendizado de Máquina. Nesta atividade, por exemplo: converte-se os dados de entrada conforme os requisitos de manipulação do algoritmo de aprendizado; transforma-se os dados de saída de modo a refletir uma saída legível aos seres humanos; estabelece-se os intervalos dos parâmetros de entrada a serem considerados; define-se os critérios de parada do algoritmo de aprendizado ou o nível de confiabilidade exigido da resposta gerada.
- **Implementação do Modelo**. Prima pela utilização dos algoritmos de Aprendizado de Máquina para estabelecer modelos a partir dos dados de entrada e saída. Iterativamente, envolve o emprego de estratégias de treinamento e testes dos algoritmos para a definição dos parâmetros mais

adequados. Como resultado, um modelo é estabelecido que melhor represente as características dos dados utilizados.

- **Utilização do Modelo**. Uma vez estabelecido um modelo, pode-se utilizá-lo para inferir informações sobre dados não utilizados na etapa anterior. Isso confirmará o poder de generalização do modelo perante situações do mundo real.
- **Tomada de Decisão**. Em situações reais, combinando o resultado gerado pelo modelo com o conhecimento particular do domínio, os gestores tomam suas decisões. Uma parte fundamental nesta etapa envolve a customização da apresentação/visualização dos dados e informações através de relatórios e gráficos. Desta forma, é possível gerar *insights* mais claros e convincentes em um processo de Tomada de Decisão.

Como suporte tecnológico, salienta-se que no Projeto de Extensão é previsto o uso da Linguagem de Programação Python. No contexto do desenvolvimento de softwares, Python destaca-se por ser uma linguagem de programação de alto nível, interpretada, baseada no paradigma de orientação a objetos e com dinamicidade semântica (PYTHON.ORG, 2018). Por dinamicidade semântica entende-se o poder de extensão da linguagem com a adição de novas funcionalidades mediante a incorporação de bibliotecas. Neste sentido, a este projeto poderão ser incorporadas as bibliotecas:

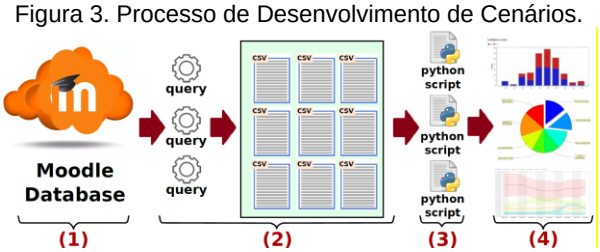
- **Pandas[4].** É uma biblioteca *open source* de alta performance que facilita o uso de estruturas e ferramentas de análise de dados em soluções computacionais desenvolvidas em Python.
- **Psycopg2[5].** Em aplicações desenvolvidas com Python, permite o acesso rápido e seguro a bases de dados mantidas no Sistema Gerenciador de Banco de Dados PostgreSQL.
- **Matplotlib[6].** É uma biblioteca para desenvolvimento de gráficos estatísticos, podendo ser utilizada em soluções computacionais Python.
- **Seaborn[7].** Assim como a Matplotlib, a Seaborn também permite o desenvolvimento de visualizações estatísticas de dados em aplicações Python.
- **NLTK (Natural Language ToolKit)[8].** É uma biblioteca feita para processamento de texto em linguagem natural. Possui ferramentas estatísticas e gráficas e *corpus* textuais para testes em vários idiomas.
- **VADER (Valence Aware Dictionary and sEntiment Reasoner)[9].** Juntamente com a NLTK, provê uma ferramenta léxica de suporte à análise de sentimentos. É baseada em regras sintáticas e semânticas, podendo ser utilizada na detecção de tendências positivas, neutras ou negativas em *post* de mídias sociais ou mensagens entre usuários, por exemplo.
- **Scikit-learn[10].** É uma biblioteca *open source* que implementa vários algoritmos de Aprendizado de Máquina em Python. Inclui algoritmos de classificação, regressão e agrupamento, por exemplo.

4 Discussão e Resultados Parciais

Considerando o ciclo de vida da Ciência de Dados, o estágio atual do projeto institucionalizado envolve a Ingestão de Dados e a Exploração de Dados. Ressalta-se que estes passos são relacionados aos dados e metadados da base de dados do MOODLE, dispendendo esforço significativo por parte dos cientistas de dados quanto ao: **i)** entendimento da complexidade do modelo de dados; e **ii)** pré-processamento e limpeza dos dados.

A Figura 3 representa as atividades desempenhadas junto aos gestores do NEaD/UNICENTRO. As atividades envolvem:

Figura 3. Processo de Desenvolvimento de Cenários.



Fonte: Dados da Pesquisa.

[4] Acesso: https://pandas.pydata.org/.
[5] Acesso: http://initd.org/psycopg/docs/.
[6] Acesso: https://matplotlib.org/.
[7] Acesso: https://seaborn.pydata.org/.
[8] Acesso: https://www.nltk.org/.
[9] Acesso: https://github.com/cjhutto/vaderSentiment.
[10] Acesso: http://scikit-learn.org.

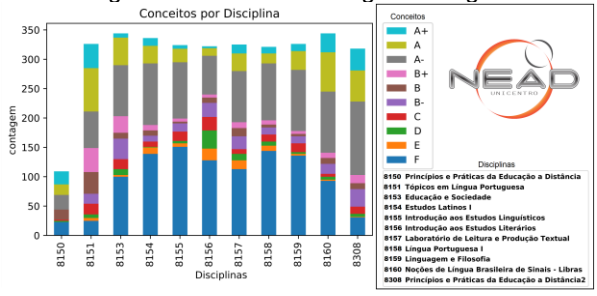
(1) A Obtenção de Dados é realizada junto à Coordenadoria de Tecnologia e Informação da UNICENTRO. A base de dados disponibilizada advém de um arquivo de *backup* do Ambiente Virtual de Aprendizagem MOODLE. Os dados obtidos são estagiados em uma instância local do Sistema Gerenciador de Banco de Dados PostGreSQL.

(2) A Ingestão de Dados é realizada de forma incremental, à medida que os cenários de exploração ou visualização de dados/informação e os respectivos *workflows* são definidos em conjunto aos gestores do NEaD/UNICENTRO. A cada cenário definido, uma *query* é desenvolvida para organizar, coletar e armazenar um subconjunto de dados relevante no repositório centralizado.

(3) Para cada subconjunto e cenário estabelecido, um *script* Python é desenvolvido para: **i)** gerar uma apresentação de dados ou visualização informações através de relatórios e gráficos, respectivamente (fase de Exploração de Dados); ou **ii)** generalizar um modelo abstrato a partir dos dados, mediante o emprego de algoritmos de Aprendizagem de Máquina (fases de Definição de Parâmetros e Implementação de Modelo).

(4) Como meio de comunicação para com os gestores do NEaD/UNICENTRO, gráficos e relatórios são apresentados, fomentando a geração de *insights* em um processo decisório (fases de Utilização do Modelo e Tomada de Decisão).

Figura 4. Gráfico para comunicação com os gestores. Cenário proposto para auxiliar a análise da distribuição dos conceitos alcançados, disciplina a disciplina, por alunos em formação - curso de graduação EaD “Letras Português e Literatura de Língua Portuguesa”.



Fonte: Dados da Pesquisa.

Como prova de conceito das atividades estabelecidas, na forma de um ensaio elementar, um cenário de visualização de dados (Figura 4) foi proposto pelos gestores do NEaD/UNICENTRO. Para com os objetivos deste trabalho, o propósito deste ensaio de visualização foi: **i)** ratificar a forma de trabalho entre os cientistas de dados; e **ii)** sensibilizar os gestores quanto à forma de comunicação.

5 Considerações Parciais

Este capítulo de livro está circunscrito a um projeto institucional, o qual objetiva a produção de elementos de análise e visualização voltados à compreensão do fenômeno da Evasão Estudantil nos cursos de graduação EaD do NEaD/UNICENTRO. Considerando a evolução dados → informação → conhecimento, o projeto institucionalizado utiliza os microdados do Ambiente Virtual de Aprendizagem MOODLE para abstrair informação útil sobre os padrões de desistência dos alunos matriculados. Consequentemente, com as informações abstraídas submetidas ao conhecimento de domínio dos gestores, pode-se amparar as ações que contribuem ao aumento dos índices de retenção dos estudantes. Em face disso, baseando-se em tecnologias da Ciência de Dados, pretende-se subsidiar um processo profícuo de Tomada de Decisão guiada por Dados no ambiente organizacional do NEaD/UNICENTO.

Neste contexto, considerando o ciclo de vida da Ciência de Dados (Bugnion; Manivannan; Nicolas, 2017) e o período de execução, o projeto encontra-se nas fases de ingestão e exploração de dados. Cronologicamente, decorreram seis meses de atividade. Apesar da infância do projeto, alguns resultados preliminares são de grande relevância ao NEaD/UNICENTRO, sendo:

i) O estabelecimento do *workflow* padrão de desenvolvimento de cenários (Figura 3).

ii) A concepção de uma camada independente para obtenção e ingestão de dados (as *queries* do passo 2 – Figura 3). Como uma camada de tradução de dados, sua utilização isola a ingestão de dados com *queries*, mitigando os efeitos indesejáveis de versionamento do modelo de dados do MOODLE. Tal feita padroniza a utilização dos dados na visualização da informação ou na generalização de modelos.

iii) A formalização de um meio de comunicação para com os gestores do NEaD/UNICENTRO. Mediante isso, incrementalmente, a prospecção de novos cenários e a compreensão do fenômeno da Evasão Estudantil podem ser facilitadas.

Ademais, como trabalhos futuros para com o projeto institucionalizado, pretende-se avançar à atividade de Utilização de Modelos de Aprendizagem de Máquina. Neste sentido, é prevista a implementação de soluções baseadas em Agrupamento, Análise de Sentimentos e Modelos Preditivos. Consequentemente, almeja-se aprimorar a Tomada de Decisão nas ações pedagógicas dos gestores do NEaD/UNICENTRO.

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).

Referências

ABED – Associação Brasileira de Educação a Distância. **Censo EAD Brasil 2014 - Relatório Analítico da Aprendizagem a Distância no Brasil**. Curitiba, 2015.

ANDIFES, A.; ABRUEM, A.; SESu/MEC, S. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. **Avaliação - Revista Da Avaliação Da Educação Superior**, v. 1, n. 2, p. 55-65, 1996.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. Sebastopol: O'Reilly Media, Inc., 2009.

BÜCHNER, A. **Moodle 3 Administration**. 3ª ed. Birmingham: Packt Publishing, 2016.

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

Disponível em: <https://goo.gl/kbFxEJ> . Acesso em: nov. 2015.

ECONOMIST. **The world's most valuable resource is no longer oil, but data**. Disponível em: <https://goo.gl/AW4XsF>. Acesso em: 28 jul. 2018.

GARTNER. **What is Big Data? – Gartner IT Glossary – Big Data**. Disponível em: <https://goo.gl/GwQWLA>. Acesso em: 28 jul. 2018.

JOHN, T.; MISRA, P. **Data Lake for Enterprises: Leveraging Lambda Architecture for building Enterprise Data Lake**. Birmingham: Packt Publishing, 2017.

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: 28 jul. 2018.

MOODLE. Features – MoodleDocs. Disponível em: <https://docs.moodle.org/35/en/Features>. Acesso em: 06 set 2018.

MORÉ, R. P. O.; MORITZ, G. de O.; PEREIRA, M. F.; MELO, P. A. de. Modelo de Gestão para Educação a Distância: o Sistema de Acompanhamento ao Estudante – SAE. **RAI - Revista de Administração e Inovação**, v. 7, n. 2, p. 104-125, 2010.

NEaD. **Núcleo de Ensino à Distância da Unicentro - PR » Polos**. Disponível em: <https://goo.gl/wPYJqe>. Acesso em: 05 set. 2018.

PYTHON.ORG. **What is Python? Executive Summary | Python.org**. Disponível em: <https://goo.gl/QYghos>. Acesso em: 06 set. 2018.

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 Pós-Graduação em Administração.

UNIVERSIDADE ESTADUAL DO CENTRO-OESTE (UNICENTRO). **RESOLUÇÃO Nº 022 - PROEC/UNICENTRO, DE 16 DE JULHO DE 2018**. Disponível em: <https://goo.gl/7VcZX4>. Acesso em: 06 set. 2018.

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.