PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Pós-graduação em Ciência de Dados e Big Data

PROJETOS DE DATA SCIENCE:
a construção de portfólio de projetos a partir das disciplinas do curso de Ciência de Dados

Aluna: Rejane Corrêa de Oliveira
Orientador: Bruno Laporais Pereira

Belo Horizonte
2019

RESUMO

Este trabalho trata de aspectos de projetos de Ciência de Dados (*Data Science*) e a construção de um portfólio profissional. A metodologia utilizada foi uma revisão sobre a bibliografia disponível sobre o tema e a organização dos trabalhos desenvolvidos durante as disciplinas do curso de pós-graduação em Ciência de Dados e Big Data da PUC Minas, com o objetivo de construção de um portfólio de projetos para o desenvolvimento de carreira e apresentação profissional da aluna. Adicionalmente, este trabalho propõe um modelo para a avaliação dos projetos sobre uma representação visual, onde a motivação é possibilitar a caracterização dos projetos constituintes de um portfólio em uma única imagem, e, por conseguinte, facilitar sua compreensão.

SUMÁRIO

1. Introdução

Um projeto de Ciência de Dados (*Data Science*) tem como objetivo extrair conhecimento e *insights* de dados coletados. Para o profissional que desenvolve projetos desta área – o cientista de dados, o portfólio de projetos se apresenta como um recurso apropriado para a busca de novas ou melhores posições no mercado de trabalho.

Um portfólio profissional reflete o desenvolvimento de conhecimentos e habilidades ao longo do tempo, apresentando evidências das competências do profissional. O portfólio contém materiais que documentam as experiências e habilidades do profissional, e ilustra a sua trajetória de carreira. Ele provê amostras de produtos de trabalho que possuem significado tanto para o próprio profissional quanto para o meio em que atua. Desta forma, o portfólio de projetos deve retratar a *expertise* do profissional.

Os projetos de Data Science podem ser caracterizados por aspectos técnicos do problema que abordam (SALTZ, SHAMSHURIN e CONNORS, 2017) ou pela forma com que foram estruturados, baseados em metodologias e boas práticas para seu desenvolvimento e comunicação posterior (RYBICKI, 2019). A primeira abordagem possibilita uma avaliação da complexidade de tais projetos. No entanto, não há um modelo que avalie conjuntamente a complexidade intrínseca ao desenvolvimento dos projetos de Data Science e a sua estruturação com o objetivo de usabilidade e comunicação. Este trabalho propõe um modelo de avaliação de projetos de Data Science que permita uma mensuração de sua complexidade e legibilidade/usabilidade posterior. Além disso, este trabalho contribui com uma proposta de visualização para tal modelo.

Os objetivos deste trabalho são, portanto:

• uma revisão de boas práticas para desenvolvimento de Projetos de *Data Science* e portfólios profissionais;

• a proposição de um modelo (métrica) de avaliação de projetos de *Data Science* e sua representação visual;

• a construção de um portfólio pessoal de projetos de *Data Science* seguindo o modelo de avaliação proposto, utilizando os trabalhos desenvolvidos nas disciplinas do curso Ciência de Dados e Big Data do programa de Pós-graduação da PUC Minas.

As perguntas que se deseja responder com este estudo são:

Q1: É possível desenvolver uma métrica capaz de combinar os conceitos, existentes e/ou ainda não explorados na literatura, para caracterização de projetos de *Data Science*?

Q2: É viável representar as dimensões que caracterizam um projeto de *Data Science* em uma única visualização?

Q3: Os trabalhos desenvolvidos nas disciplinas do curso Ciência de Dados e Big Data da PUC Minas, quando reunidos, são apropriados para a apresentação em um portfólio profissional?

Com uma revisão da literatura existente sobre projetos de *Data Science*, um levantamento de suas características principais e de seus desafios para implantação, foi possível construir um modelo para representar as dimensões que caracterizam projetos desta natureza e responder às duas primeiras perguntas de pesquisa. A resposta à terceira pergunta de pesquisa se dá com a aplicação do modelo sobre os trabalhos apresentados pela autora durante o referido curso.

2. Referencial Teórico

2.1. Aspectos de Projetos de *Data Science*

O sucesso no ambiente de negócios orientado a dados (*data-driven*) de hoje exige ser capaz de pensar sobre como alguns conceitos fundamentais do pensamento analítico aplicam-se aos problemas de negócios em questão. Há uma extensa coleção de técnicas de mineração de dados, algoritmos e ferramentas que vem sendo largamente utilizados, juntamente com princípios básicos e conceitos que fundamentam essas técnicas e também o pensamento sistemático que promove o sucesso na tomada de decisão orientada a dados. Esta área de conhecimento é nomeada “Ciência de Dados” (PROVOST e FAWCETT, 2013).

Segundo Demchenko *et al*. (2016), Ciência de Dados (*Data Science*) é uma área emergente da ciência, que possui uma abordagem multidisciplinar e uma estreita ligação com as tecnologias de Big Data, assumindo um papel transformador nos campos da pesquisa e da indústria.

De maneira simplificada e intuitiva, pode-se dizer que o objetivo da Ciência de Dados é extrair conhecimento de dados coletados. Para isto, são aplicados métodos oriundos de diversos campos do conhecimento, como ciência computacional, estatística, e gestão de dados, entre outros (RYBICKI, 2019). Além da produção acadêmica sobre o assunto, muito do conhecimento e das práticas do setor são divulgados em blogs especializados, devido à natureza de mercado da área.

Tipicamente, um projeto de Ciência de Dados (*Data Science*) é descrito como um projeto que utiliza técnicas de estatística e aprendizado de máquina (*Machine Learning*) em

grandes volumes de dados estruturados e/ou não-estruturados originados por diversas fontes como sistemas, pessoas, sensores (SALTZ, SHAMSHURIN e CONNORS, 2017).

Paruchuri (2016a) descreve dois tipos principais projetos de Data Science: *data storytelling* e operacional. *Data storytelling* (isto é, "contar uma história com os dados") define projetos que envolvem profundidade nas análises exploratórias, a fim de justificar e induzir alguém a perceber de forma simplificada o *insight*[1] observado e, principalmente, produzir conhecimento e novos *insights*. Os componentes principais deste tipo de projeto são: compreender e definir o contexto, explorar múltiplos ângulos, utilizar visualizações convincentes, enriquecimento de dados, e apresentar uma narrativa consistente.

O segundo tipo é o que impacta as operações cotidianas de uma empresa e poderá ser utilizado mais de uma vez, e frequentemente por muitas pessoas (PARUCHURI, 2016b). Para este tipo, *storytelling* é menos importante que a competência técnica. Este tipo de projeto é construído de ponta a ponta (do início ao fim ou *end-to-end*), sendo necessário buscar um conjunto de dados, entendê-lo, e então criar uma solução que processe tais dados, transformando-os e possibilitando responder a perguntas por vezes dinâmicas. É importante que esta solução seja executada tão rápido quanto necessário e possível para um bom resultado, e utilize um mínimo de recursos de sistema, como memória. É comum que esta solução seja executada muitas vezes, então o produto entregável (*deliverable*) é toda a solução computacional, e não apenas uma apresentação com visualizações das descobertas da investigação, sendo esta última apenas uma parte do processo. Os componentes principais deste tipo de projeto são: compreender o contexto, explorar os dados e descobrir as nuances, criar um projeto bem estruturado, de modo que seja fácil sua integração em fluxos operacionais, escrever um código de alta performance, e documentar bem a implementação e uso de seu código, para sua reusabilidade (PARUCHURI, 2016b).

O trabalho de Saltz, Shamsrurim e Connors (2017) busca identificar quais são as características-chave que melhor descrevem projetos de *Data Science*, e como tais características podem ser integradas em um modelo coerente de projeto, de modo a construir um modelo geral que possibilite escolher a melhor maneira de gerenciar os projetos desta natureza. Segundo estes autores, não é suficiente a tradicional descrição dos projetos por meio dos "4 V's", isto é, com base nas características intrínsecas aos dados: Volume (tamanho dos dados/ necessidade de se utilizar técnicas de Big Data), Variedade (número de fontes e tipo dos dados – estruturado/não-estruturado), Velocidade (velocidade da geração/coleta dos dados

[1] Do inglês, significa uma compreensão ou solução de um problema pela súbita captação mental dos elementos e relações adequados.

que necessitam ser analisados) e Veracidade (confiabilidade dos dados). Os autores argumentam que há ainda outros aspectos que caracterizam os projetos de *Data Science*, como a incerteza quanto à entrada/saída da solução (quais dados são relevantes, o que se poderá descobrir com tais dados), bem como os desafios durante seu desenvolvimento. Isso difere os projetos de *Data Science* de projetos de outra natureza, como desenvolvimento de software, por exemplo. A proposta dos autores é a inclusão de aspectos como tamanho da organização, tamanho das equipes e sua virtualização, cultura da organização (se é orientada a dados, se tem foco em retorno do investimento ou P&D, etc). Assim, para avaliar a complexidade dos projetos, os autores consideram:

- o contexto dos dados (4V's),

- o contexto analítico (se é um projeto para geração de hipóteses – escopo aberto, ou para teste de hipóteses – finalidade definida, ou ambos),

- o contexto da equipe (tamanho e virtualização das equipes),

- o contexto organizacional (tamanho e cultura da organização).

Uma abordagem complementar é a descrita no trabalho de Rybicki (2019), que investiga as melhores práticas para estruturação de projetos de *Data Science*, utilizadas no meio acadêmico e/ou na indústria. O autor argumenta que a estrutura do projeto de *Data Science* deve ser vista também como mais um dos recursos de comunicação das descobertas decorrentes do projeto, permitindo sua reprodutibilidade e uso posterior. Em seu trabalho, o autor resume as estruturas de práticas oriundas de:

- metodologias de *Data Science*, como KDD (*Knowledge Discovery in Databases*), CRISP-DM (*Cross-industry Standard Process for Data Mining*), TDSP (*Team Data Science Process*), *Cookiecutter Data Science*;

- estabelecidas pela comunidade, como hierarquia de diretórios para dados e códigos, documentação e controle de versão,

- plataformas de compartilhamento de dados e projetos, como a plataforma Kaggle (KAGGLE, 2018).

Paruchuri (2016b) enfatiza a necessidade de uma boa estruturação do projeto, no caso de projetos do tipo operacional, pois nesse caso é necessário trabalhar com múltiplos arquivos, de diferentes tipos (códigos Python, markdown, arquivos csv, json, etc.) e alternar entre eles. Trabalhar de uma forma estruturada permite realizar mudanças no fluxo de processamento de dados sem que seja necessário recalcular tudo, e isto é importante no caso de projetos muito grandes. Um projeto bem estruturado, segundo o autor, segue alguns princípios, considerados boas práticas:

- separa os arquivos de dados e arquivos de código;

- separa dados brutos dos dados processados;

- tem um arquivo de orientação para *deploy* e utilização da solução;

- mantem arquivos de requisitos técnicos (como bibliotecas) para execução da solução;

- tem um arquivo que centraliza as configurações que serão usadas em outros arquivos;

- possui controle de acesso e segurança dos arquivos, evitando acesso por outra pessoa não autorizada (ex.: .gitignore);

- modulariza as tarefas principais do código como leitura, geração de atributos, predições, etc.;

- versiona a saída de cada etapa, a fim de permitir reutilizar parte de um fluxo já executado.

A estruturação do projeto envolve delimitar as etapas de seu desenvolvimento. Os fluxos para ingestão e preparação de dados são etapas importantes do projeto de Ciência de Dados, e consideradas etapas críticas para análises mais complexas. Yu, Wang e Lai (2006) propõem um modelo para a preparação de dados, constituído de três fases:

i) pré-análise, no qual os dados de interesse são identificados e coletados: inclui análise de requisitos de dados, coleta de dados, seleção de dados e integração de dados;

ii) pré-processamento de dados, no qual os dados são examinados e analisados e onde alguns dados podem ser reestruturados ou transformados para torná-los mais úteis;

iii) pós-análise de dados, em que alguns dados são validados e reajustados, para retroalimentar o processo de modelagem.

Assim como para construir outros tipos de software, as equipes que constroem os fluxos de dados (*pipelines*) precisam de testes automatizados para gerenciar a complexidade envolvida na tarefa. Mas em vez de apenas testar o código, os testes devem ser feitos sobre os dados, pois neles é que reside grande parte da complexidade dos projetos de Ciência de Dados e Big Data (Great Expectations, 2018). Sculley *et al*. (2014) enumeram vários exemplos em que a dependência de dados tem grande impacto sobre um sistema em produção que utiliza modelos de *Machine Learning*.

O sucesso de um projeto de *Data Science* envolve atenção para todos estes aspectos durante seu desenvolvimento.

2.2. Aspectos do Portfólio Profissional

Em muitas profissões, as competências desenvolvidas como resultados da experiência de um profissional são apresentadas sob a forma de um portfólio, que contém materiais capazes de documentar as experiências e competências do profissional, e ilustram a sua trajetória na carreira. É um documento particularmente útil quando o profissional se apresenta ao mercado de trabalho pela busca de novas ou melhores posições, uma vez que reflete o desenvolvimento de conhecimento e habilidades ao longo do tempo, apresentando evidências dos trabalhos anteriormente desenvolvidos e das competências adquiridas (OERMANN, 2002).

Oermann (2002) cita dois tipos de portfólio: '*best-work*' (melhores trabalhos) e '*growth and development*' (crescimento e desenvolvimento), assim descritos:

- Portfólios do tipo '*best-work*' são utilizados como documentação que ampara a candidatura a uma vaga de emprego ou a busca de uma promoção na carreira. Contém materiais cuidadosamente selecionados, rotulados, organizados e preparados para uma avaliação externa. Este tipo de portfólio demonstra ao potencial empregador o conhecimento, habilidades e experiências relevantes do profissional.

- Portfólios do tipo '*growth and development*' é um documento de trabalho do próprio profissional que auxilia o monitoramento de seu progresso em seus objetivos de aprendizagem e desenvolvimento de carreira. Pode ser usado como um plano de desenvolvimento profissional, onde são listadas as atividades de aprendizagem concluídas, materiais que a comprovem, sua eficácia em promover a aquisição de competências profissionais de modo crescente e continuado, e especificar objetivos a serem alcançados. O profissional pode selecionar materiais deste portfólio para incluir no portfólio do tipo '*best-work*', que é apresentado a outros.

Sobre portfólios profissionais especificamente elaborados para o meio acadêmico, Froh, Gray e Lambert (1993) sugerem em primeiro lugar uma reflexão sobre a sua construção, buscando respostas a perguntas como:

> "Qual é a importância relativa das diversas atividades e, portanto, seu peso ou influência no reconhecimento do trabalho (recompensas)? Quais são os critérios para julgar os vários níveis de qualidade e quantidade destas atividades? Qual é a evidência aceitável de qualidade e quantidade em relação às diversas atividades?" (FROH, GRAY e LAMBERT, 1993)

Estas mesmas perguntas podem ser usadas como ponto de partida para a construção de portfólios em qualquer área profissional. Os autores também argumentam que um foco mais

amplo no desenvolvimento dos portfólios profissionais pode ajudar o profissional em todas as fases de sua carreira.

Segundo Froh, Gray e Lambert (1993), o desenvolvimento de portfólio de carreira não é uma atividade feita em uma única etapa, mas sim um processo cumulativo e refletido que se estende ao longo da carreira profissional. O portfólio profissional pode ser visto como a base para o avanço para a próxima fase da carreira. Em cada fase, o desenvolvimento do portfólio pode ajudar os profissionais a refletir sobre realizações e atividades passadas, traçar objetivos profissionais futuros e fornecer documentação selecionada aos recrutadores/avaliadores quando em busca de novas ou melhores posições. Documentação selecionada é um aspecto que os autores destacam: um grande desafio na construção da carteira profissional é decidir "o quanto é o suficiente" quando o portfólio é usado para tomada de decisão, de modo a não torná-lo sobrecarregado, ou, ao contrário, muito esparso.

Quanto à avaliação dos portfólios por outros, considerando o contexto acadêmico, Froh, Gray e Lambert (1993) sugerem que os portfólios profissionais podem ser avaliados seguindo algumas diretrizes: (1) o trabalho é fundamentado e refletido; a apresentação demonstra experiência em fazer escolhas em um determinado contexto, capacidade de resposta a desenvolvimentos não previstos e criatividade no desenvolvimento de uma abordagem acadêmica; (2) o trabalho resulta em novo conhecimento de uma situação específica e demonstra a sua validade e importância; e (3) o trabalho envolve a comunicação de novos conhecimentos para os outros.

2.3. Portfólio de Projetos de *Data Science*

Assim como em outros campos profissionais, as empresas olham para portfólios quando tomam decisões de contratação de cientistas de dados, pois é uma das possibilidades de informação sobre o profissional, o que torna o portfólio uma boa maneira de avaliar alguém por suas habilidades reais demonstradas nos projetos já desenvolvidos (PARUCHURI, 2016a). Galarnyk (2018) define um portfólio como “uma evidência pública de suas habilidades em ciência de dados”, e acrescenta que projetos são talvez os melhores substitutos para experiência profissional, e por este motivo podem ser particularmente úteis para iniciantes na carreira.

De forma similar, o blog Analytics Vidhya (2018) recomenda, em tom enfático:

> “Projetos de ciência de dados oferecem-lhe um caminho promissor para dar o pontapé inicial sua carreira neste campo. Não só você vai aprender a ciência de dados aplicando-a, você também terá projetos para mostrar no seu currículo! Hoje em dia, os recrutadores avaliam o

> potencial de um candidato pelo seu trabalho e não colocam muita ênfase em certificações. Não importa se você diga quanto você sabe se você não tem nada para mostrar-lhes!".

Segundo Vasconcellos (2017), "Para um *Data Scientist*, um portfólio é tão importante quanto as experiências e habilidades que este adquiriu nos últimos anos."

De acordo com Paruchuri (2016a), as principais habilidades que as empresas procuram em cientistas de dados, e, portanto, devem ser as habilidades demonstradas em um portfólio, são: habilidade de comunicação, habilidade de colaboração com outros profissionais, competência técnica, habilidade de raciocinar sobre os dados, motivação e ter iniciativa. Assim, um bom portfólio deve ser composto de múltiplos projetos, cada um destes destacando uma ou duas destas habilidades.

Aconselhamentos sobre a elaboração de projetos de *Data Science* podem ser encontrados com facilidade nos blogs especializados na área, como *Towards Data Science* (GOODMAN, 2016), *KDNuggets* (GALARNYK, 2018), *DataQuest* (PARUCHURI, 2016), *Analytics Vidhya* (2018), *Data Science Central* (HIGDON, 2014), entre outros. De maneira geral, a recomendação é que exista uma escolha sobre um conjunto de dados interessantes e que seja possível analisá-los por múltiplas perspectivas, com especial atenção à forma de comunicação dos resultados. Há também quem apresente conselhos sobre o que não incluir no portfólio de projetos de *Data Science*: segundo Harris (2018), é recomendado não incluir projetos muito comuns, uma vez que isso não só não destacaria o profissional, mas o desmereceria frente a um recrutador. Exemplos de projetos muito comuns são os que incluem a classificação de sobreviventes com dados do Titanic, a identificação de dígitos manuais da base MNIST, e a classificação de espécies de flores utilizando o *dataset* "iris". Galarnyk (2018) também enfatiza a importância de não citar projetos muito comuns. O Quadro 1 apresenta uma síntese destas recomendações.

Quanto à exposição do trabalho, Galarnyk (2018) argumenta que muito do trabalho de *Data Science* reside na comunicação e apresentação de dados, sendo por este motivo recomendado que o profissional mantenha perfis *online* onde possa mostrar o seu trabalho. Galarnyk (2018), Shinde (2018) e Vasconcellos (2017) ressaltam a importância de divulgar e compartilhar os projetos, e recomendam as plataformas Github e Kaggle, assim como blogs (ex: Medium) e mídias sociais (Twitter, LinkedIn), além de outros espaços como Stack Overflow, Tableau Public, Quora, Youtube, etc. Segundo Shinde (2018), construir o perfil na plataforma Kaggle (KAGGLE, 2018) é algo a colocar no currículo, pois os empregadores podem verificar as habilidades de um candidato apenas pesquisando seu nome. Eles podem

ver de quantas competições o profissional participou, os tipos de modelos que ele construiu, ou análise que realizou com os conjuntos de dados disponíveis.

No entanto, entre as várias opções de repositório para o portfólio de projetos de *Data Science*, os autores são unânimes quanto a uma delas: o Github (GITHUB, 2018). O Github é um repositório de hospedagem de projetos que utilizam a ferramenta de versionamento de arquivos *git* (RATAMERO, 2016). Os vários autores (ver Quadro 1) enfatizam que os especialistas ou recrutadores de cientistas de dados muitas vezes verificam o Github para ver quem está postando o quê, e o quão precisos são seus códigos e modelos:

> "um perfil no Github é um sinal poderoso que você é um cientista de dados competente. Na seção de projetos de um currículo, as pessoas muitas vezes deixam *links* para o GitHub, onde o código é armazenado para seus projetos. (...). GitHub permite que as pessoas possam ver o que você construiu e como você construiu. (...) Se você tomar o tempo para desenvolver seu perfil no GitHub, você pode ser melhor avaliado do que outros candidatos. Vale ressaltar que você precisa ter algum tipo de arquivo (...) com uma descrição do seu projeto, uma vez que muito ciência de dados é sobre a comunicação de resultados. Verifique se o arquivo de README.md descreve claramente qual é o seu projeto, o que faz e como executar seu código." (GALARNYK, 2018).

Quadro 1 – Síntese das recomendações para projetos de *Data Science* divulgadas em blogs especializados.

Blog	Autor	Datasets	Visualização	Texto	Repositório/ outras recomendações	Data da publicação
Towards Data Science (HARRIS, 2018)	Jeremie Harris	- Não utilizar datasets muito conhecidos, como Titanic, Iris, MNIST - não utilizar projetos de MOOCs			- apresentar habilidades de versionamento (Git), Devops e bancos de dados (SQL, NoSQL)	Jun/2018
Medium (GOODMAN, 2016)	Jason Goodman	-Utilizar dados reais e não dados já limpos de bases como Kaggle -buscar seus próprios dados em API's de acesso livre, web scraping - usar dados interessantes é mais importante do que uma técnica sofisticada de modelagem - a análise deve ser interessante seja qual for o resultado, confirmando ou não a hipótese	- investir tempo criando visualizações interessantes e interativas de preferência - gráficos simples de linhas e barras também são bons pois são de fácil compreensão	- manter o texto curto (explicações adicionais devem ser colocadas em um apêndice)	- Github - comentar e organizar bem o código - fazer com que todas as etapas sejam reprodutíveis (desde extração de dados até visualização) - são ferramentas úteis R Markdown e IPython Notebooks	Dez/2016
Dataquest, Projeto tipo Storytelling (PARUCHURI, 2016a)	Vik Paruchuri	- usar dados em que esteja verdadeiramente interessado, e não apenas para compor o projeto - compreender e definir o contexto, - explorar múltiplos ângulos, enfatizando um tópico escolhido - complementar com dados de outras fontes relacionadas	- utilizar visualizações convincentes	- apresentar uma narrativa consistente	- apresentar no Github - notebooks são uma ferramenta interessante de apresentação pois permitem seguir o raciocínio da análise	Jun/2016

Blog	Autor	Datasets	Visualização	Texto	Repositório/ outras recomendações	Data da publicação
Dataquest, projeto tipo operacional (PARUCHURI, 2016b)	Vik Paruchuri	- usar um conjunto de dados grande que necessite transformação e para o qual possa responder a perguntas dinâmicas (para predição de algum fenômeno)		- Explicar de maneira sucinta do que se trata o projeto - Explicar como reproduzir o projeto	- Github - organizar bem a estrutura do projeto, de modo que todas as etapas sejam reprodutíveis	Jul/2016
KDNuggets (GALARNYK, 2018)	Michael Galarnyk	- usar dados interessantes	- incluir algumas visualizações	- texto bem escrito, mostrando resultados novos e interesssantes	- divulgar e compartilhar o trabalho Github, Kaggle, blogs, mídias sociais	Jul/2018
Data Science Central (HIGDON, 2014)	Peter Higdon	- dados que mostre habilidade de transformar (curadoria de dados)	- visualização de gráficos (“*Visualization zoo*”)		- não há necessidade de mostrar grande conhecimento matemático (sic), pois o mais importante é o código que funcione e a interseção com o negócio	Ago/2014
Analytics Vidhya (2018)	vários	- 24 conjuntos de dados sugeridos, agrupados nos níveis iniciante, intermediário e avançado -usar pelo menos um caso de conjunto de dados grande			- usar problemas de diferentes áreas com técnicas variadas - compartilhar no Github assim que tiver 2 ou 3 projetos para apresentar	Mai/2018

Quadro 1, continuação da página anterior.

2.4. O Curso de Formação em Ciência de Dados e Big Data

O curso de pós-graduação *latu sensu* de CIÊNCIA DE DADOS E BIG DATA oferecido pela PUC Minas foi criado para atender a uma demanda efetiva do profissional denominado Cientista de Dados “com competências e habilidades para entender bem as estratégias e necessidades do negócio e gerenciar, projetar e desenvolver soluções de análise em grandes volumes de dados” (PUC MINAS, 2018).

O curso tem como público-alvo os profissionais com formação superior em Ciência da Computação, Sistemas de Informação e cursos correlatos, com atuação em Inteligência de Negócios e TI, gestão de projetos, análise de redes sociais e mídias, entre outros, e que possuam interesse em tecnologias para análise de bases de dados para apoio à tomada de decisão (PUC MINAS, 2018).

O curso de CIÊNCIA DE DADOS E BIG DATA da PUC Minas é vinculado ao ICEI – Instituto de Ciências Exatas e Informática / Departamento de Engenharia de Software e Sistemas de Informação, insere-se na Área do conhecimento: 1.03.00.00-7 (Ciência da Computação) e Sub-área: 1.03.03.04-9 (Sistemas de Informação) (PUC MINAS, 2018). A oferta do segundo semestre de 2017 possui projeto pedagógico com as ementas das disciplinas descritas no Quadro 2.

Entre os objetivos do curso, estão: 1) “Formar profissionais capazes de analisar o estado da arte de *Big Data* e *Business Analytics* com forte embasamento conceitual e prático; 2) Capacitar os participantes na análise dos problemas empresariais e a projetar, desenvolver e gerenciar projetos que demandam técnicas atuais para análise de grandes volumes de dados, de maneira a apoiar a empresa para que ela alavanque sua competitividade; 3) Mostrar a importância dos dados no âmbito da organização, bem como elaborar e executar o processo de garantia de qualidade dos mesmos e desenvolver os conceitos relacionados à sua governança. (PUC MINAS, 2018).

Como parte do processo avaliativo, são solicitados aos alunos trabalhos/projetos que demonstrem a aquisição dos conhecimentos apresentados em cada disciplina.

Segundo Saltz e Heckman (2015), um curso de formação em *Big Data* baseado em projetos é apropriado para estudantes com origens acadêmicas diversas, de modo que conquistem ao longo do curso um conjunto de habilidades para coletar, analisar e comunicar os resultados de um projeto de *Data Science*.

A formação acadêmica e treinamento necessário para o estabelecimento da profissão de Cientista de Dados é objeto do projeto EDISON, desenvolvido no âmbito da Comunidade Europeia (DEMCHENKO *et al*., 2016). Este projeto estabeleceu um *framework* com componentes que definem as competências necessárias para o exercício bem sucedido da profissão, as áreas de conhecimento para construir grades curriculares que atendam ao desenvolvimento destas competências, um corpo de conhecimento que incorpora as melhores práticas do setor (*DS-BoK – Data Science Body of Knowledge*), entre outros.

Segundo a definição estabelecida no âmbito do projeto EDISON (DEMCHENKO *et al*., 2016), as áreas de conhecimento no domínio da Ciência de Dados são:

- *Data Analytics* (inclui *Machine Learning*, métodos estatísticos e *Business Analytics*);
- *Data Science Engineering* (inclui Engenharia de Software e Infraestrutura);
- *Data Management* (inclui *data curation*, *preservation* e *data infrastructure*);
- *Scientific or Research Methods;*
- *Business process management;*
- *Data Science Domain Knowledge* (inclui conhecimento específico sobre *Data Science*).

Quadro 2 – Ementário do curso de CIÊNCIA DE DADOS E BIG DATA, PUC Minas (oferta 2º semestre/2017).

Disciplina	Ementa
AM- Machine Learning	Metodologia para descoberta de conhecimento em banco de dados. Exploração do espaço problema e espaço solução. Técnicas de aprendizado supervisionado e não-supervisionado. Regras de associação, agrupamento (clustering) e classificação. Rede neural, Agrupamento com K Means. Classificador Naïve Bayesian. Árvore de decisão. Outros algoritmos
GQD- Arquitetura e Qualidade de Dados	Arquitetura de dados. Arquitetura de Dados no DMBOK e Open Group Togaf. A relação do papeis: AD, DBA, Arquiteto Corporativo e Arquiteto de Dados. Conceitos e motivações para governança de dados. Maturidade em governança de dados. Conceitos de qualidade de dados. Atividades e técnicas para qualidade de dados. Avaliação da qualidade de dados. Metadados. Master Data Management
IBD- Ciência de Dados e Big Data em Negócios	Importância da informação no negócio. Necessidades em decisões de negócio. Conceitos de Big Data. Big Data em relação a outras disciplinas. Ciência dos dados. Ciclo de vida do processo de ciência de dados. Papeis dos envolvidos em projetos de Ciência de dados e Big Data. Computação em nuvens. Arquitetura de Big Data. Modelos de entrega e distribuição de serviços de Big Data. Principais plataformas de Cloud Computing para Big Data.
NSQ- Banco de dados Não Relacionais	Bancos de Dados NoSQL: definição; motivação; modelo de Transações. Modelos Nosql. Propriedades Modelo Relacional x Propriedades Modelos Nosql. Principais SGBD's. Soluções para Big Data
HD- Soluções para Processamento Paralelo e Distribuído de Dados	Princípios de processamento e de volumes de dados massivos. Conceitos básicos de sistemas distribuídos. Modelo de Computação MapReduce: definição e motivação. Hadoop. Spark. Outros ambientes de processamento. Aplicações
HIV- Tecnologias para o Ecossistema de Big Data	Frameworks sobre Hadoop: PIG, Hive, Impala e outras soluções. Introdução linguagem SQL. Programação de aplicações. Conexão de clientes
ETL - Integração e Fluxos de Dados	Conceitos. Identificação de requisitos. ETL, ELT e ELTL. Estrutura de dados ETL. Projeto e desenvolvimento de aplicação ETL. Plano de Teste. Operação. Ferramentas de ETL. Parametrização e configuração. Cloud Dataflow
ILE- Introdução às Linguagens Estatísticas	Aspectos básicos da programação em linguagem Python e R. Programação de aplicações
AED - Estatística Geral – Teoria e Aplicações	Estatística descritiva. Probabilidade e distribuições de probabilidade. Inferência: estimação pontual e intervalar e testes de hipóteses. Utilização de software para análises estatísticas e análise de casos aplicados à gestão
AP - Técnicas Estatísticas de Predição: Teoria e Aplicações	Modelos Preditivos e tipos de análise. Abordagens para análise preditiva. Séries temporais. Regressão Linear simples e múltipla. Regressão logística. Modelos preditivos na plataforma hadoop
RI- Recuperação da Informação na Web e em Redes Sociais	Conceito de inteligência. Conceito de inteligência coletiva. Conceito de crowdsourcing. Ferramentas de análise, monitorização e benchmark. Web mining. Algoritmos e soluções para problemas de busca e extração de informação da Web. Algoritmos e soluções para a análise de redes sociais online e em sites de conteúdo. Web crawling. Text Mining.
PAF- Processamento e Análise de Fluxos Contínuos de Dados	Data Streaming e dados em tempo real. Conceitos de eventos e sua topologia. Identificação e processamento de eventos complexos. Sistemas de gestão de fluxo de dados. Principais ferramentas e tecnologias. Projeto de solução para stream analytics
VIS - Data Discovery, OLAP e Visualização de Dados	Fundamentos e requisitos de aplicações de suporte a decisão. Projeto, construção e tecnologias de aplicações OLAP. Dashboards. Data Storytelling. Fundamentos da descoberta de dados. Projetos em design de informação. Métodos e técnicas de visualização de dados. Self Service BI. Geoanálises
GPT - Gerência De Projetos	Introdução a Gerenciamento de Projetos. Ciclo de vida de projetos. Áreas de conhecimento do PMI. Boas práticas. Definição do problema e dos requisitos do projeto integrado. Planejamento do projeto integrado. Ferramentas de gerenciamento de projetos
APL- Projeto Integrado – Construção Aplicação Big Data e Analytics	Revisão de padrões. Aspectos arquiteturais de uma solução de big data e ciência de dados. Problemas comuns de big data e ciência de dados. Visão geral de Produtos em Analytics (Data discovery, Data Services,). Utilização de serviços de dados em nuvem. Projetos aplicativos utilizando big data e ciência de dados. Criação de Data Products

3. Metodologia

Este trabalho propõe a criação de um modelo de avaliação de projetos de *Data Science*, uma representação visual para esse modelo, e sua aplicação para avaliação dos projetos desenvolvidos durante as disciplinas do curso de Ciência de Dados e Big Data. Adicionalmente, o portfólio de projetos de *Data Science* da aluna é criado no repositório Github.

3.1. Modelo de avaliação de projetos de *Data Science*

Avaliar um portfólio de projetos é uma tarefa custosa e muitas vezes subjetiva. A proposta deste trabalho é construir um modelo que permita suportar essa avaliação, com foco em projetos de *Data Science*. Obter um "retrato" do projeto e possibilitar a sua comparação com outros projetos de maneira direta é o objetivo deste modelo.

Baseado nos estudos de Saltz *et al*. (2017) e Rybicki (2019), é possível retratar projetos de *Data Science* em duas dimensões: (i) a ***complexidade***, como características inerentes ao projeto; (ii) a ***estruturação***, como características relacionadas à abordagem usada para seu desenvolvimento. O modelo proposto busca caracterizar o projeto de maneira a contemplar aspectos tanto da complexidade intrínseca a este, como também da abordagem realizada para seu desenvolvimento e implementação (estruturação). Nossa proposta é desenvolver uma visualização que combine as duas dimensões supracitadas, proporcionando uma percepção visual mais rápida e direta sobre quais tipos de projeto constituem aquele portfólio.

Sobre o eixo da ***complexidade***, podem existir projetos de *Data Science* que variam de baixa complexidade a alta complexidade. Uma adaptação do trabalho de Saltz *et al*. (2017) é aqui utilizada para descrever a complexidade dos projetos com base em características de volume, variedade, velocidade, e veracidade dos dados, assim como a intensidade computacional para pré-processamento e/ou modelagem. Ainda como adaptação do estudo destes autores, o tipo de análise requerida também dá indicações sobre a complexidade do projeto, e é incluído na métrica de avaliação deste atributo: geração de hipóteses ("escopo aberto"), ou teste de hipóteses ("escopo fechado"), ou ambas as situações ("escopo aberto com teste de hipóteses").

O projeto também pode ser avaliado quanto à sua ***estruturação***, e esta dimensão em particular seria um indicativo do uso de boas práticas de determinada técnica no projeto. Rybicki (2019) descreve como boas práticas em projetos de *Data Science* a existência dos

seguintes aspectos: o uso de alguma metodologia para seu desenvolvimento (CRISP-DM, TDSP, KDD2, etc.), controle de versão, um *script* para o *pipeline* ou *workflow description* - isto é, uma forma de fazê-lo funcionar de ponta-a-ponta, e a própria hierarquia de diretórios e subdiretórios de arquivos (*readme, data - metadata, raw, processed -, results, scripts, documentation*, etc.). Segundo este autor, um projeto com boa estruturação possibilita o reuso e legibilidade, sendo indicativo de sua boa qualidade.

Ainda quanto à dimensão ***estruturação***, um bom projeto de *Data Science* deve contar com o controle de versão dos dados, de desenvolvimento de testes sobre os dados e preparação dos dados, *pipeline* para as fases de preparação dos dados, bem como garantir a reprodutibilidade dos experimentos (BAKER, 2016, e HINSEN, 2015). A atenção especial sobre os dados e seus *pipelines* é sugerida por trabalhos como Sculley *et al*. (2014), Garga et. al (2016), Yu, Wang e Lai (2006) e Eichelberger *et al*. (2017). Sculley *et al*.(2014) destacam que a dependência de dados custa mais do que a dependência de código em sistemas de Machine Learning, e é fonte importante de débito técnico ("*technical debt* ") que não deve ser ignorada. Os *pipelines* desenvolvidos para a etapa de preparação de dados também são difíceis de gerenciar, assim como detecção de erros, recuperação de falhas, e testes (SCULLEY, 2014). A relevância destes aspectos inspirou a criação de bibliotecas com a finalidade de testar *pipelines* de dados ("Great Expectations", 2018) e de código ("mltest", para *machine learning*, citado por Roberts, 2018), o que em parte atende à questão levantada por Hinsen (2015) sobre a dificuldade da reprodutibilidade em experimentos em ciência computacional. Por esse motivo, o modelo proposto para este trabalho inclui uma avaliação de atributos relacionados aos dados e preparação dos dados. A qualidade do código também é aspecto a ser considerado, sendo possível o uso de ferramentas como Better Code Hub (BETTER CODE HUB, 2018) ou Codacy (CODACY, 2018).

Por fim, o projeto também pode ser classificado quanto à área do conhecimento a que pertence principalmente. De acordo com a nomenclatura do *Data Science Framework* proposto pelo projeto EDISON (DEMCHENKO *et al*., 2016), as áreas do conhecimento podem ser separadas em: *Data Analytics*; *Data Science Engineering*; *Data Management*; *Scientific or Research Methods*; *Business process management*; e *Data Science Domain Knowledge*. A área de conhecimento à qual pode ser atribuída o projeto é a terceira dimensão que é empregada no modelo proposto de avaliação de projetos desenvolvido neste trabalho.

O Quadro 3 resume os critérios de avaliação e propõe uma escala numérica para os níveis de cada critério das duas dimensões principais. Para os itens mais críticos para o sucesso do projeto, sugere-se um peso maior.

Quadro 3 – Critérios para modelo de avaliação de projetos de *Data Science*.

COMPLEXIDADE (mínimo = 6; máximo =18)			
Atributo / Nota	**1**	**2**	**3**
Dados - Volume	GB	TB	
Dados - Variedade	Estruturados	Não estruturados	Estruturados + Não Estruturados
Dados - Velocidade	Mini *Batch* (processamento de poucos registros)	*Batch* (processamento de muitos registros)	*Near Real-time*
Dados - Veracidade	Organizados *(Tidy)*	Organizados *(Tidy*(-)	Desorganizados *(Messy)*
Intensidade Computacional	Irrelevante	Relevante na fase de pré-processamento OU de modelagem	Relevante em ambas as fases pré-processamento E modelagem
Tipo de análise	Escopo fechado	Escopo aberto	Escopo aberto com teste de hipóteses
ESTRUTURA (mínimo = 0; máximo =10)			
Atributo/ Nota	**0**	**1**	**2**
Metodologia para o desenvolvimento (ex: CRISP-DM)	não	sim	
Controle de versão de dados	não		sim
Pipeline para construção do datalake	não		sim
Desenvolvimento de testes (sobre dados)	não	sim	
Métrica de qualidade do código	não	sim	
Hierarquia de pastas e organização dos dados (e scripts, resultados, documentação)	não	sim	
Reprodutibilidade de experimentos	não		sim
ÁREA DO CONHECIMENTO			
Data Analytics			
Data Science Engineering			
Data Management			
Scientific or Research Methods			
Business process management			
Data Science Domain Knowledge			

3.2. Representação Gráfica para Avaliação de Projetos de *Data Science*

Como recurso para uma visualização direta e adequada dos projetos caracterizados segundo os critérios do Quadro 3, uma representação gráfica é proposta e descrita a seguir. Segundo Cleveland e McGill (1985), quando um gráfico é construído, informação quantitativa e categórica é codificada por meio de posição, forma, tamanho, símbolos e cor. Ao ver um gráfico, a informação é decodificada pelo sistema visual humano. A tarefa de extrair informação de um gráfico por meio da decodificação visual é chamada de percepção gráfica. Segundo os autores, muito do poder dos gráficos vem da chamada capacidade pré-atentiva do sistema visual, que permite a percepção instantânea da informação por meio da detecção de padrões e magnitude, como tendências e valores atípicos, sem aparente esforço mental. O objetivo da visualização é auxiliar a compreensão dos dados, já representações visuais bem projetadas podem substituir cálculos cognitivos com simples inferências perceptuais e melhorar a compreensão, a memória e tomada de decisão (HEER, BOSTOCK e OGIEVETSKY, 2010).

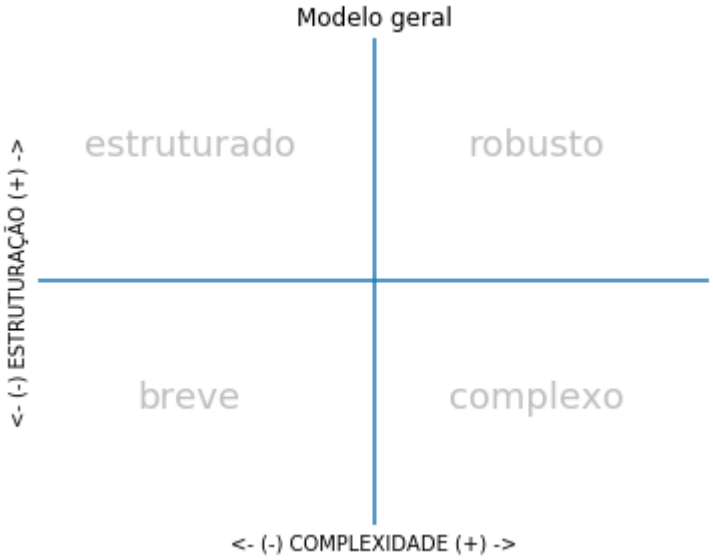
Os pontos são os objetos de codificação de dados com a forma mais simples possível. No contexto de um gráfico de eixo 2-D (eixos X e Y), pontos codificam valores por sua posição em associação com a escala ao longo de cada eixo. São chamados gráficos de dispersão, e suas duas escalas quantitativas permitem exibir a correlação (ou a falta desta) entre dois conjuntos de medidas. Cada ponto em um gráfico de dispersão codifica dois valores quantitativos: um ao longo do eixo X e outro ao longo do eixo Y (FEW, 2006). Experimentos de percepção gráfica demonstram que a posição espacial (como em um gráfico de dispersão ou gráfico de barras) conduz à decodificação mais exata dos dados numéricos e é geralmente preferível a outras opções visuais como ângulo, comprimento unidimensional, área (bidimensional), volume (tridimensional), ou saturação de cor (HEER, BOSTOCK e OGIEVETSKY, 2010).

Assim, pela simplicidade e rapidez na comunicação visual, a representação gráfica escolhida para o modelo proposto neste trabalho é a posicional, com pontos indicando a correlação entre as duas dimensões que se quer representar. A Figura 1 ilustra a representação gráfica para avaliação de projetos de *Data Science* segundo os critérios do Quadro 3 sobre as duas dimensões ***complexidade*** (eixo horizontal) e ***estruturação*** (eixo vertical), dividida em quatro quadrantes, possibilitando a leitura das seguintes subdivisões:

- BREVE, denominação para o quadrante inferior esquerdo: projetos pouco complexos e pouco estruturados;

- ESTRUTURADO, denominação para o quadrante superior esquerdo: projetos pouco complexos e mais estruturados;
- COMPLEXO, denominação para o quadrante inferior direito: projetos complexos e pouco estruturados;
- ROBUSTO, denominação para o quadrante superior direito: projetos complexos e bem estruturados.

Figura 1 – Modelo de Representação Gráfica para Avaliação de Projetos de *Data Science*.



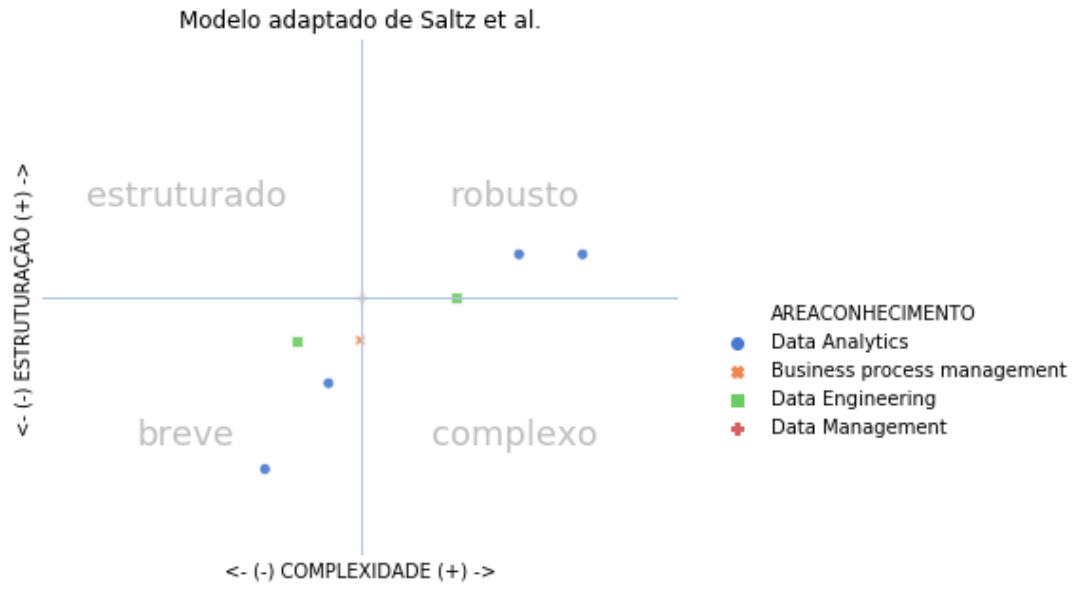
4. Resultados

4.1. Modelo de avaliação de projetos de *Data Science* aplicado a um caso geral

O modelo de avaliação de projetos de *Data Science* proposto em 3.1 é aqui aplicado a um caso geral. A referência é o trabalho de Saltz *et al*.(2017), onde são detalhadas características de oito projetos. As métricas de avaliação segundo o Quadro 3 foram aplicadas ou inferidas para os projetos estudados por aqueles autores, e estão detalhadas no Anexo A. Nos casos de informação não existente para alguns dos requisitos, os valores inferidos estão com a formatação em itálico; os demais são oriundos de informação tal como mencionada na fonte referida (Anexo A). As áreas de conhecimento segundo a nomenclatura do Projeto EDISON (DEMCHENKO *et al.*, 2016) foram acrescentadas conforme as características descritas.

A representação gráfica resultante da aplicação do modelo de avaliação de projetos adaptado ao estudo de Saltz *et al*. (2017) é apresentada na Figura 2. Os oito projetos representados se posicionam ao longo do espaço que vai do quadrante “BREVE” ao quadrante “ROBUSTO”, pelo incremento tanto de complexidade quanto de estruturação. As áreas de conhecimento foram representadas com diferentes cores e símbolos, de modo a facilitar a compreensão. Há um distinto contraste entre o projeto representado no extremo canto inferior esquerdo e o projeto mais distante do canto superior direito: enquanto o primeiro representa um projeto com baixa velocidade de geração de dados estruturados e limpos da ordem de Gigabytes, escopo fechado de análise (teste de hipóteses definido) e intensidade computacional pouco relevante nas várias etapas (pré-processamento e modelagem), o último projeto representa um caso de dados estruturados e não-estruturados, gerados em alta velocidade, de maneira desorganizada (*messy*), em volume de Terabytes, com análise de escopo aberto (geração e teste de hipóteses), e intensidade computacional importante na fase de pré-processamento. Certamente este último projeto necessita de uma solução mais “robusta” para a confiabilidade dos resultados, e isso mostra que o modelo de avaliação aqui proposto é capaz de representar distintamente projetos com características tão diferentes.

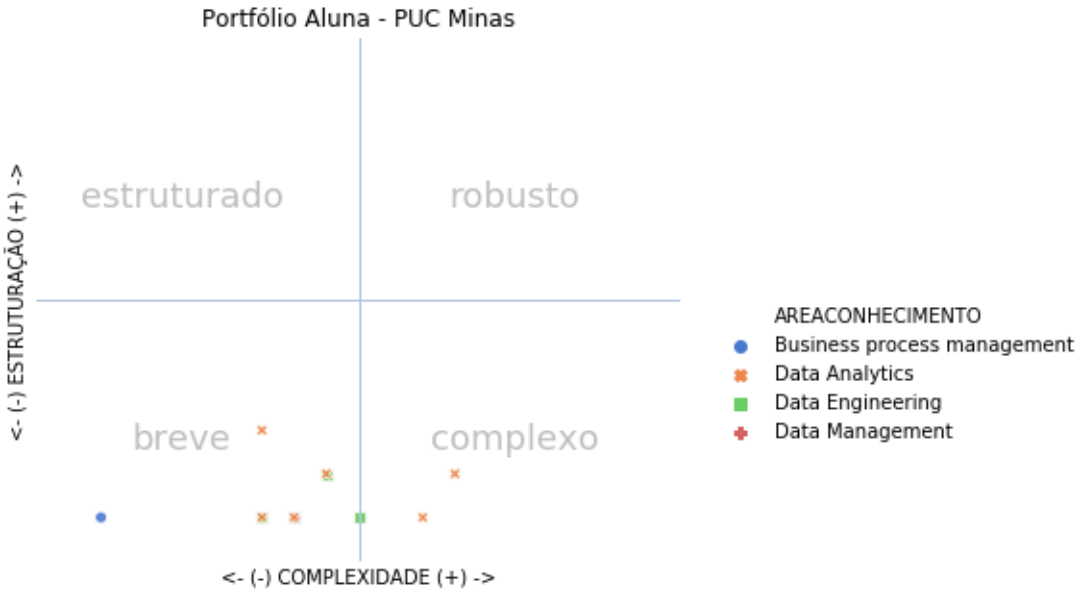
Figura 2 – Representação gráfica do Modelo de avaliação de projetos de *Data Science* (adaptado ao caso de Saltz *et al*., 2017).



4.2. Modelo de avaliação de projetos de *Data Science* aplicado aos trabalhos do curso de Ciência de Dados e Big Data

O Quadro 4 apresenta os trabalhos desenvolvidos durante as disciplinas do curso de Ciência de Dados e Big Data da PUC Minas, com as técnicas utilizadas. Os trabalhos foram associados às áreas de competência referidas no âmbito do projeto EDISON (DEMCHENKO *et al*., 2016). Em algumas disciplinas, o processo avaliativo foi no formato de prova de conhecimentos e não na forma de projetos; tais disciplinas não foram incluídas no Quadro 4, portanto. Os trabalhos apresentados no Quadro 4 foram avaliados conforme o modelo proposto no tópico anterior. As métricas de avaliação utilizadas estão detalhadas no Anexo B. A visualização resultante é apresentada na Figura 3.

Figura 3 – Modelo de avaliação de projetos de *Data Science* aplicados ao portfólio da autora.



Observa-se na Figura 3 que a maior parte dos projetos recai sobre o quadrante esquerdo inferior, indicando baixa a média complexidade e quase nenhuma prática de estruturação – ou seja, projetos da categoria BREVE. Esse resultado é compatível com trabalhos desenvolvidos em um curso de formação, onde a abrangência dos trabalhos não ultrapassa muito o cronograma e o conteúdo programático das disciplinas. Isso também reflete o início de uma jornada de aprendizado, que poderá evoluir no decorrer da vida profissional da aluna. Estima-se que a trajetória natural de desenvolvimento de um profissional de Ciência

de Dados seja migrar de projetos “breves” para “complexos”, uma vez que o desafio adicional a ser enfrentado é mais claro neste último caso do que nos projetos do quadrante “estruturado”. Por fim, deseja-se que o profissional em desenvolvimento, ao somar experiências ao longo de sua carreira, seja capaz de atuar em projetos do quadrante “robusto”.

4.3. Construção do portfólio profissional

A plataforma Github foi escolhida como repositório para receber os projetos da aluna, uma vez que é a mais recomendada entre os diversos autores (Quadro 1). O *link* de referência para o repositório no Github é https://github.com/rejaneol.

Os projetos descritos no Quadro 4 e representados na Figura 3 são elegíveis para incorporar o portfólio no segmento ‘*growth and development*’ (“crescimento e desenvolvimento”), segundo a sugestão de Oermann (2002) como um documento de trabalho para o desenvolvimento de carreira da presente autora.

Os projetos que não possuíam código ou que foram desenvolvidos em plataforma específica (ex: Databricks) não foram adicionadas ao repositório do Github. Entre os demais, os que forem escolhidos para compor o portfólio do tipo “*best-work*” serão marcados com uma estrela (“*star*”, funcionalidade disponível no Github).

5. Conclusões

A partir de uma ampla revisão da literatura existente sobre projetos de *Data Science*, um levantamento de suas características principais e desafios e boas práticas para sua implantação, foi proposto um modelo para representar as dimensões que caracterizam projetos desta natureza. O modelo proposto introduz uma métrica, combinando conceitos existentes e outros ainda não explorados na literatura, para caracterizar os projetos de *Data Science*, o que atende à primeira pergunta de pesquisa. Em seguida, foi proposta uma representação gráfica para as dimensões que definem os projetos de *Data Science*, permitindo uma única e imediata visualização do conjunto de projetos e suas características, o que atende à segunda pergunta de pesquisa. A utilização da métrica de avaliação e sua representação visual conduzem a uma melhor avaliação de projetos.

A aplicação do modelo proposto, e sua representação visual, aos trabalhos desenvolvidos nas disciplinas do curso Ciência de Dados e Big Data da PUC Minas, possibilita a construção de um portfólio pessoal de projetos de *Data Science* e demonstra que, quando reunidos, estes trabalhos do curso de formação são apropriados para apresentação das habilidades e conhecimentos adquiridos enquanto ingressante à profissão de cientista de

dados, respondendo à terceira pergunta de pesquisa. Neste primeiro momento, como profissional em formação, o conjunto de projetos recai principalmente sobre o quadrante denominado “Breve”, pois foram trabalhos desenvolvidos de modo a atender ao conteúdo e ao prazo de cada disciplina. É o início de uma trajetória profissional que poderá ser documentada com a aplicação continuada do modelo proposto neste trabalho.

Quadro 4 – Trabalhos desenvolvidos nas disciplinas do curso de CIÊNCIA DE DADOS E BIG DATA, PUC Minas (oferta 2º semestre/2017).

Disciplina	Título do projeto	Técnicas utilizadas/ ferramentas	Área(s) de competência
AM- Machine Learning	Modelo de Previsão de Eficiência Energética em Edificações	(Python, Scikit-Learn)	*Data Analytics*
IBD- Ciência de Dados e Big Data em Negócios	Proposta de solução de Big Data Analytics para um caso de uso.	Modelo de negócio	*Business process management*
NSQ- Banco de dados Não Relacionais	E-commerce	Bancos de dados SQL (SQLite) e NoSql (Mongo DB. Redis e Neo4j), Python	*Data Management, Data Science Engineering*
HD- Soluções para Processamento Paralelo e Distribuído de Dados	1) Informação bibliográfica: termos, autores 2) Promoções e vendas realizadas em uma rede de lojas	1) MapReduce 2) Spark Dataframe (Databricks)	*Data Science Engineering, Data Analytics*
HIV- Tecnologias para o Ecossistema de Big Data	Predição de venda de produtos	Análise exploratória de dados (SQL, R, Zeppelin)	*Data Analytics, Data Science Engineering*
ETL - Integração e Fluxos de Dados	Copas do Mundo FIFA: análise de dados históricos	Extração de dados da web, transformação e carga, análise exploratória de dados (Power BI)	*Data Management*
ILE- Introdução às Linguagens Estatísticas	Diversidade profissional em uma empresa. Análise exploratória de dados de perfis profissionais	Base de dados json, Análise exploratória de dados: Estatística descritiva, gráficos (Python)	*Data Analytics*
AP - Técnicas Estatísticas de Predição: Teoria e Aplicações	Análise Estatística do acidente do Titanic.	Regressão logística (linguagem R)	*Data Analytics*
RI- Recuperação da Informação na Web e em Redes Sociais	Interação entre os atores do ecossistema brasileiro de inovação por meio da análise de mídias sociais.	Twitter, Processamento de Linguagem Natural (Knime)	*Data Analytics, Data Science Domain Knowledge*
VIS - Data Discovery, OLAP e Visualização de Dados	Educação no Brasil. Visualização dos Resultados do Saeb 2015 e 2017	Power BI	*Data Analytics*
APL- Projeto Integrado – Construção Aplicação Big Data e Analytics	SISTEMAS DE RECOMENDAÇÃO: um estudo de caso	Algoritmos KNN e filtragem colaborativa ALS	*Data Analytics, Data Science Engineering*

REFERÊNCIAS

Analytics Vidhya. 24 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely). [S. l.], Jul. 2018. Disponível em https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/. Acesso em: dezembro/2018.

BAKER, M. 1,500 scientists lift the lid on reproducibility. Nature News. Springer Nature, Maio, 2016.

Better Code Hub. [S. l.]. Disponível em: https://github.com/marketplace/better-code-hub. Acesso em: dezembro/2018.

CLEVELAND, W.S., McGILL, R. Graphical Perception and Graphical Methods for Analyzing Scientific Data. Science, New Series, v. 229, n. 7416, p. 828-833, Ago. 1985.

Codacy. [S. l.]. Disponível em: https://github.com/marketplace/codacy. Acesso em: dezembro/2018.

DEMCHENKO, Y., BELLOUM, A.S.Z., LOS, W. *et al*. EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry. 2016 IEEE International Conference on Cloud Computing Technology and Science, Luxemburgo, Dez. 2016. Disponível em: https://www.researchgate.net/publication/312963890. Acesso em: dezembro/2018.

EICHELBERGER, H., QIN, C., SCHMID, K. Experiences with the Model-based Generation of Big Data Pipelines. Lecture Notes in Informatics (LNI), Bonn, 2017.

FEW, Stephen. Data Visualization: Rules for Encoding Values in Graph. The Perceptual Edge. [S. l.], Jan. 2006.

FROH, Robert C., GRAY, PeterJ., LAMBERT, Leo M. Representing Faculty Work: The Professional Portfolio. New Directions for Higher Education, n. 81, Spring, 1993.

GARGA, N., SINGLAB, S., JANGRAC, S. Challenges and Techniques for Testing of Big Data. Procedia Computer Science, v. 85, p. 940 – 948, 2016.

GALARNYK, Michael. How to Build a Data Science Portfolio. [S. l.], Jul. 2018. Disponível em: https://www.kdnuggets.com/2018/07/build-data-science-portfolio.html. Acesso em: dezembro/2018.

Github: the world's leading software development plattform. [S. l.]. Disponível em: https://github.com/. Acesso em: dezembro/2018.

GOODMAN, Jason . Advice on Building Data Portfolio Projects. [S. l.], dez. 2016. Disponível em: https://medium.com/@jasonkgoodman/advice-on-building-data-portfolio-projects-c5f96d8a0627. Acesso em: dezembro/2018.

Great Expectations. Down with Pipeline debt / Introducing Great Expectations. Disponível em https://medium.com/@expectgreatdata/down-with-pipeline-debt-introducing-great-expectations-862ddc46782a. Acesso em: dezembro/2018.

HARRIS, Jeremie. The 4 fastest ways not to get hired as a data scientist. [S. l.], jun. 2018. Disponível em: https://towardsdatascience.com/the-4-fastest-ways-not-to-get-hired-as-a-data-scientist-565b42bd011e. Acesso em: dezembro/2018.

HEER, J., BOSTOCK, M., OGIEVETSKY, V. A Tour through the Visualization Zoo: A survey of powerful visualization techniques, from the obvious to the obscure. ACMQueue, Graphics, v. 8, n.5, Mai. 2010.

HIGDON, Peter. Your Data Science Portfolio: Math Skills Don't Matter. Data Science Central. [S. l.], Ag. 2014. Disponível em: https://www.datasciencecentral.com/profiles/blogs/your-data-science-portfolio-math-skills-don-t-matter. Acesso em: dezembro/2018.

HINSEN, K. Technical Debt in Computational Science. Computing in Science and Engineering, v., n.17, p.103-107, nov. 2015.

Kaggle: Your Home for Data Science. [S. l.]. Disponível em: https://www.kaggle.com/, Acesso em: dezembro/2018.

OERMANN, Marilyn H. Developing a Professional Portfolio in Nursing. Orthopaedic Nursing, v 21, n. 2, Março/Abril 2002.

PROVOST, F., FAWCETT, T. Data Science And Its Relationship To Big Data And Data-Driven Decision Making. Big Data, v.1, n.1, Mar. 2013.

PARUCHURI, Vik. Building a data science portfolio: Storytelling with data. [S. l.], Jun. 2016a. Disponível em: https://www.dataquest.io/blog/data-science-portfolio-project . Acesso em: dezembro/2018.

PARUCHURI, Vik. Building a data science portfolio: Machine learning project. [S. l.], Jul. 2016b. Disponível em https://www.dataquest.io/blog/data-science-portfolio-machine-learning. Acesso em: dezembro/2018.

PUC MINAS. Curso de Ciência de Dados e Big Data. [S. l.]. Disponível em: https://www.pucminas.br/Pos-Graduacao/IEC/Cursos/Paginas/Ciencia-de-Dados-e-Big-Data-Pra%C3%A7a%20da%20Liberdade_5.aspx?moda=5&polo=7&area=79&curso=138&situ=1 Acesso em: Dezembro/2018

RATAMERO, Luciano. git e github parte 1: o que são e como usar? [S. l.], fev. 2016. Disponível em: https://www.ratamero.com/blog/git-e-github-parte-1-o-que-sao-e-como-usar/. Acesso em: dezembro/2018.

ROBERTS, C. mltest: Automatically test neural network models in one function call. [S. l.], fev. 2018. Disponível em: https://medium.com/@keeper6928/mltest-automatically-test-neural-network-models-in-one-function-call-eb6f1fa5019d. Acesso em: dezembro/2018.

RYBICKI, Jedrzej. Best Practices in Structuring Data Science Projects. Structuring Data Science Projects. Z. Wilimowska *et al*. (Eds.): Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018, Parte 3, AISC 854, p. 348–357, 2019.

SALTZ, J., HECKMAN R. Big Data science education: A case study of a project-focused introductory course. Themes in Science & Technology Education, v.2, n.8, p. 85-94, 2015.

SALTZ, J., SHAMSHURIN, I., CONNORS, C.. Predicting Data Science Sociotechnical Execution Challenges by Categorizing Data Science Projects. Journal Of The Association For Information Science And Technology, v.12, n.68, p.2720–2728, 2017.

SCULLEY, D., HOLT, G., GOLOVIN, D., DAVYDOV, E., PHILLIPS,T., EBNER, D., CHAUDHARY, V., YOUNG, M. Machine Learning: The High-Interest Credit Card of Technical Debt. Proceedings SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop), 2014.

SHINDE, Manali. How to Construct a Data Science Portfolio from Scratch. [S. l.], Abr. 2018. Disponível em https://medium.com/one-datum-at-a-time/how-to-construct-a-data-science-portfolio-from-scratch-de0b70e58bc1. Acesso em: dezembro/2018.

VASCONCELLOS, Paulo. Como criar seu portfólio de Data Scientist e divulgar seus trabalhos. Medium, [S. l.], Jan. 2017. Disponível em: https://paulovasconcellos.com.br/como-criar-seu-portfolio-de-data-scientist-cc7e6b23b996. Acesso em: dezembro/2018.

YU, L., WANG S., LAI K.K. An integrated data preparation scheme for neural network data analysis. IEEE Transactions On Knowledge And Data Engineering, v. 18, n. 2, fev. 2006.

ANEXO A – Aplicação do Modelo de Avaliação de Projetos de *Data Science*.
Avaliação dos projetos citados por Saltz *et al* (2017) - adaptado.

Titulo do projeto	dadosVolume	dadosVariedade	dadosVelocidade	dadosVeracidade	intensidadeComputacional	tipoAnalise	**COMPLEXIDADE**	metodologia	controleVersaoDados	dataLakePipeline	dataTesting	codeQuality	hierDirSubdir	teprodutibilidade	**ESTRUTURA**	**AREACONHECIMENTO**
1	2	3	1	3	2	3	14	*1*	*0*	*2*	*0*	*0*	*1*	*2*	6	Data Analytics
2	1	1	1	1	3	2	9	*1*	*0*	*0*	*0*	*0*	*1*	*2*	4	Business process management
3	2	1	2	1	3	3	12	*1*	*0*	*2*	*1*	*0*	*1*	*0*	5	Data Engineering
4	1	1	1	1	2	1	7	*0*	*2*	*0*	*1*	*0*	*1*	*0*	4	Data Engineering
5	1	1	1	1	1	1	6	*0*	*0*	*0*	*0*	*0*	*1*	*0*	1	Data Analytics
6	2	3	3	3	2	3	16	*0*	*2*	*2*	*1*	*0*	*1*	*0*	6	Data Analytics
7	1	2	1	2	1	2	9	*0*	*2*	*0*	*1*	*1*	*1*	*0*	5	Data Management
8	1	1	2	1	1	2	8	*0*	*0*	*2*	*0*	*0*	*1*	*0*	3	Data Analytics

Obs: Valores inferidos em itálico.

ANEXO B - Avaliação dos trabalhos das disciplinas do curso de Ciência de Dados e Big Data segundo o modelo proposto de avaliação de projetos de *Data Science*.
(Referência Quadro 4).

Titulo do projeto	dadosVolume	dadosVariedade	dadosVelocidade	dadosVeracidade	intensidadeComputacional	tipoAnalise	**COMPLEXIDADE**	metodologia	controleVersaoDados	dataLakePipeline	dataTesting	codeQuality	hierDirSubdir	teprodutibilidade	**ESTRUTURA**	**AREACONHECIMENTO**	**Disciplina**
Modelo de Previsão de Eficiência Energética em Edificações	1	1	1	1	1	1	6	1	0	0	0	0	1	0	2	Data Analytics	AM- Machine Learning
Proposta de solução de Big Data Analytics para um caso de uso.						1	1	0	0	0	0	0	0	0	0	Business process management	IBD- Ciência de Dados e Big Data em Negócios
E-commerce	1	3	1	1	1	1	8	0	0	0	0	0	1	0	1	Data Engineering	NSQ- Banco de dados Não Relacionais
1) Informação bibliográfica: termos, autores	1	1	1	1	1	1	6	0	0	0	0	0	0	0	0	Data Engineering	HD- Soluções para Processamento Paralelo e Distribuído de Dados
2) Promoções e vendas realizadas em uma rede de lojas	1	1	1	1	1	2	7	0	0	0	0	0	0	0	0	Data Engineering	
Predição de venda de produtos	1	2	1	2	1	2	9	0	0	0	0	0	0	0	0	Data Engineering	HIV- Tecnologias para o Ecossistema de Big Data
Copas do Mundo FIFA: análise de dados históricos	1	1	1	1	1	2	7	0	0	0	0	0	0	0	0	Data Management	ETL - Integração e Fluxos de Dados
Diversidade profissional em uma empresa. Análise exploratória de dados de perfis profissionais	1	2	1	3	1	3	11	0	0	0	0	0	0	0	0	Data Analytics	ILE- Introdução às Linguagens Estatísticas
Análise Estatística do acidente do Titanic.	1	1	1	2	1	1	7	0	0	0	0	0	0	0	0	Data Analytics	AP - Técnicas Estatísticas de Predição: Teoria e Aplicações
Interação entre os atores do ecossistema brasileiro de inovação por meio da análise de mídias sociais.	1	2	1	3	2	3	12	0	0	0	0	0	1	0	1	Data Analytics	RI- Recuperação da Informação na Web e em Redes Sociais
Educação no Brasil. Visualização dos Resultados do Saeb 2015 e 2017	1	1	1	2	1	2	8	0	0	1	0	0	0	0	1	Data Analytics	VIS - Data Discovery, OLAP e Visualização de Dados
SISTEMAS DE RECOMENDAÇÃO: um estudo de caso	1	1	1	1	1	1	6	0	0	0	0	0	0	0	0	Data Analytics	APL- Projeto Integrado – Construção Aplicação Big Data e Analytics

ANEXO C – Código Python para visualização do modelo de avaliação de projetos de *Data Science.*

```
# importando as bibliotecas
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

# carregando os dados
with pd.ExcelFile('dsprojectsevaluation.xlsx') as xls:
    df_rejane = pd.read_excel(xls, 'portfolio Rejane')
    df_geral = pd.read_excel(xls, 'modelo geral')

# visualização com seaborn
fig, ax = plt.subplots(figsize = (6,5))

# marca d'água
fig.text(0.30, 0.3, 'breve',
         fontsize=18, color='gray',
         ha='center', va='center', alpha=0.5)

fig.text(0.30, 0.65, 'estruturado',
         fontsize=18, color='gray',
         ha='center', va='center', alpha=0.5)

fig.text(0.7, 0.3, 'complexo',
         fontsize=18, color='gray',
         ha='center', va='center', alpha=0.5)

fig.text(0.7, 0.65, 'robusto',
         fontsize=18, color='gray',
         ha='center', va='center', alpha=0.5)

# grafico da Figura 2
sns.scatterplot(x="COMPLEXIDADE", y="ESTRUTURA", hue="AREACONHECIMENTO",
style="AREACONHECIMENTO", sizes=(150, 50), alpha=.9, palette="muted",
data=df_geral)

plt.axis([-1, 19, -1, 11])
plt.axhline(5, linewidth=1, color='lightsteelblue')
plt.axvline(9, linewidth=1, color='lightsteelblue')

plt.title("Modelo adaptado de Saltz et al.")
plt.legend(bbox_to_anchor=(1.05, 0.5), loc=2, borderaxespad=0.,
frameon=False)

plt.xlabel('<- (-) COMPLEXIDADE (+) ->')
plt.ylabel('<- (-) ESTRUTURAÇÃO (+) ->')

plt.xticks([])
plt.yticks([])

ax.spines["right"].set_visible(False)
ax.spines["top"].set_visible(False)
ax.spines["left"].set_visible(False)
ax.spines["bottom"].set_visible(False)

plt.show()
```