Avaliação de Projetos de Data Science

Rejane Oliveira

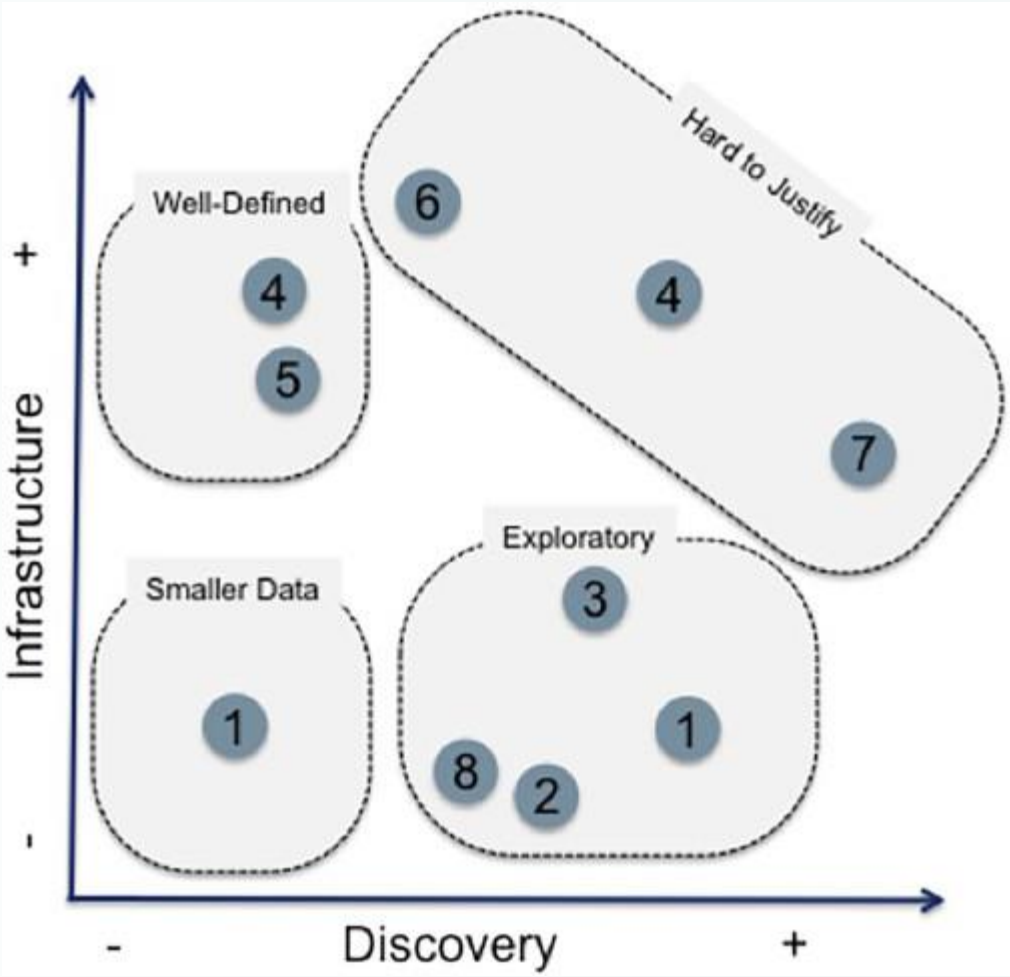
A3Tech #27

Avaliação de Projetos de Data Science

- motivação
- referências
- o modelo 3DSE:
  - visual
  - critérios para avaliação
- aplicação

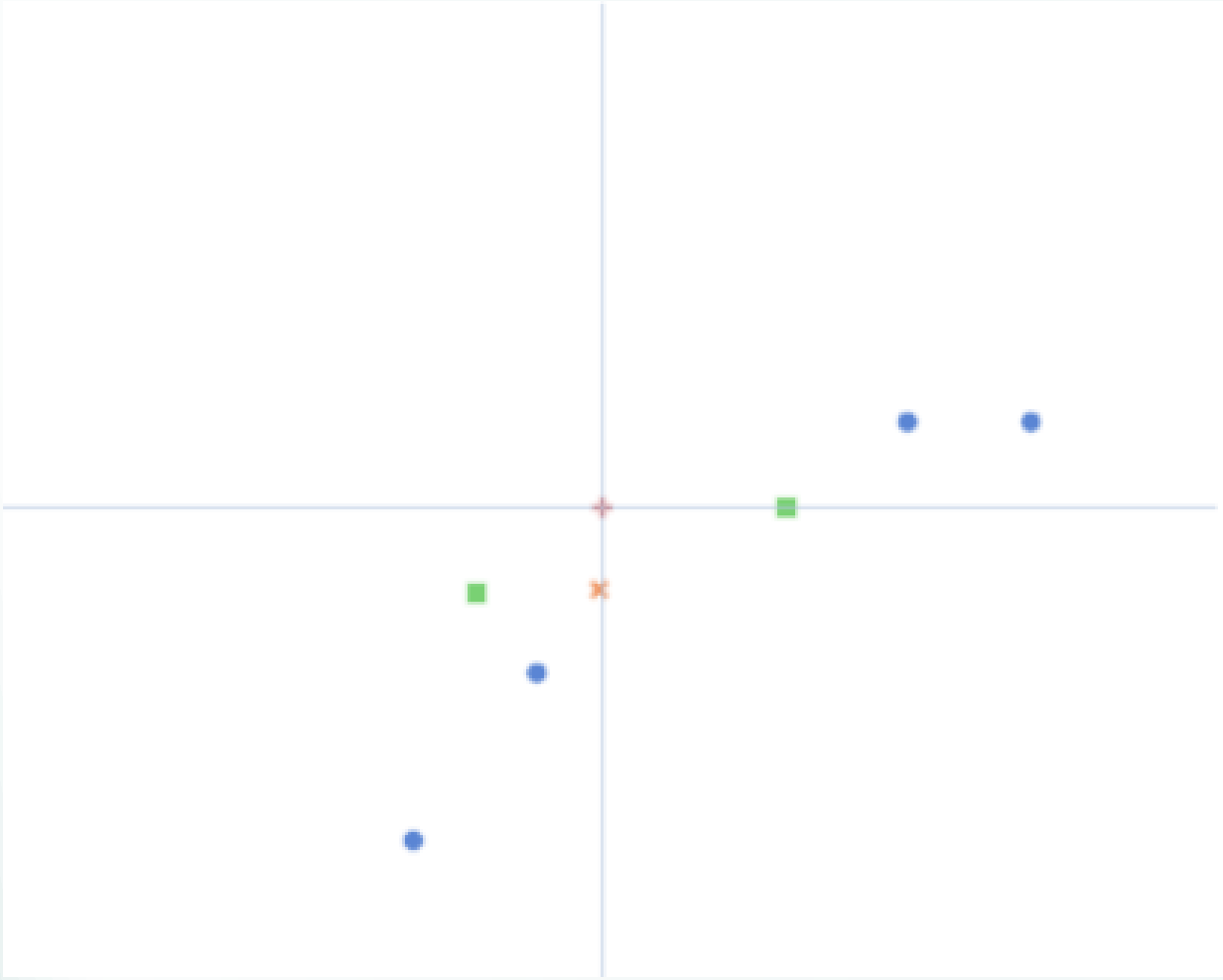
- motivação

- referências



Referência: 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.

- o modelo 3DSE:
visual



- o modelo 3DSE:

critérios para avaliação

COMPLEXIDADE

(inerente ao projeto)

- 4 V's: volume, variedade, velocidade, e veracidade dos dados
- sensibilidade dos dados
- intensidade computacional para pré-processamento e/ou modelagem
- tipo de análise requerida:
  - geração de hipóteses ("escopo aberto"),
  - ou teste de hipóteses ("escopo fechado"),
  - ou ambas as situações ("escopo aberto com teste de hipóteses").

COMPLEXIDADE

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
Dados – sensibilidade			Presente

(pontos: min – 6; max - 21)

- o modelo 3DSE:
critérios para avaliação

ESTRUTURAÇÃO

(boas práticas)

- uso de alguma metodologia para seu desenvolvimento (CRISP-DM, por exemplo)
- controle de versão
- qualidade do código
- hierarquia de diretórios e subdiretórios de arquivos
- controle de versão dos dados
- desenvolvimento de testes sobre os dados e preparação dos dados
- *pipeline* para as fases de preparação dos dados
- reprodutibilidade dos experimentos

ESTRUTURAÇÃO

Atributo/ Nota	0	1	2
Metodologia para o desenvolvimento (ex: CRISP-DM)	não	sim	
Controle de versão de dados	não		sim
Pipeline para 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

(pontos: min – 0; max - 10)

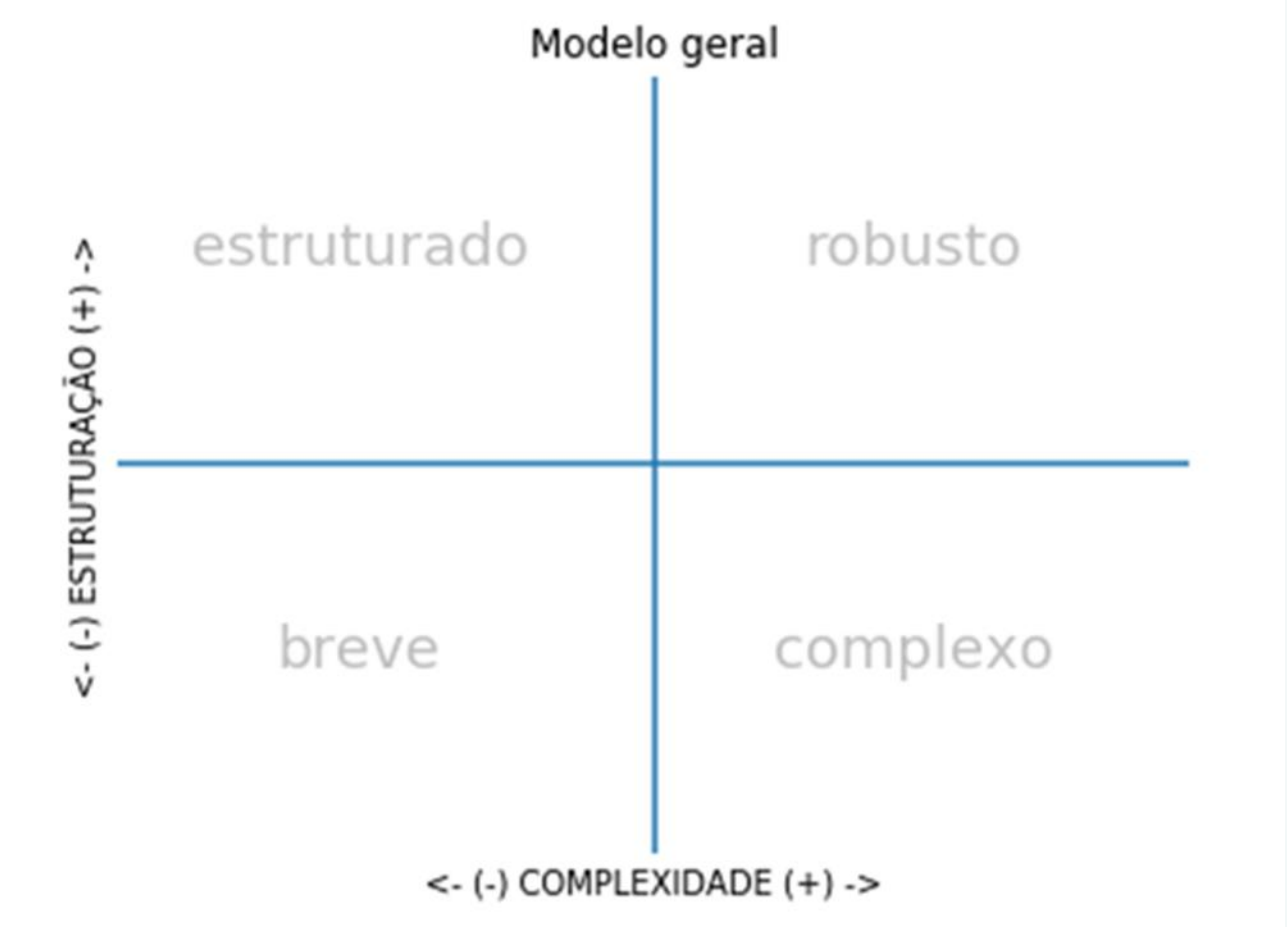
- o modelo 3DSE:
critérios para avaliação

ÁREAS DO CONHECIMENTO

- Data Analytics
- Data Engineering
- Data Management
- Scientific or ResearchMethods
- Business Process Management
- Data Science Domain Knowledge

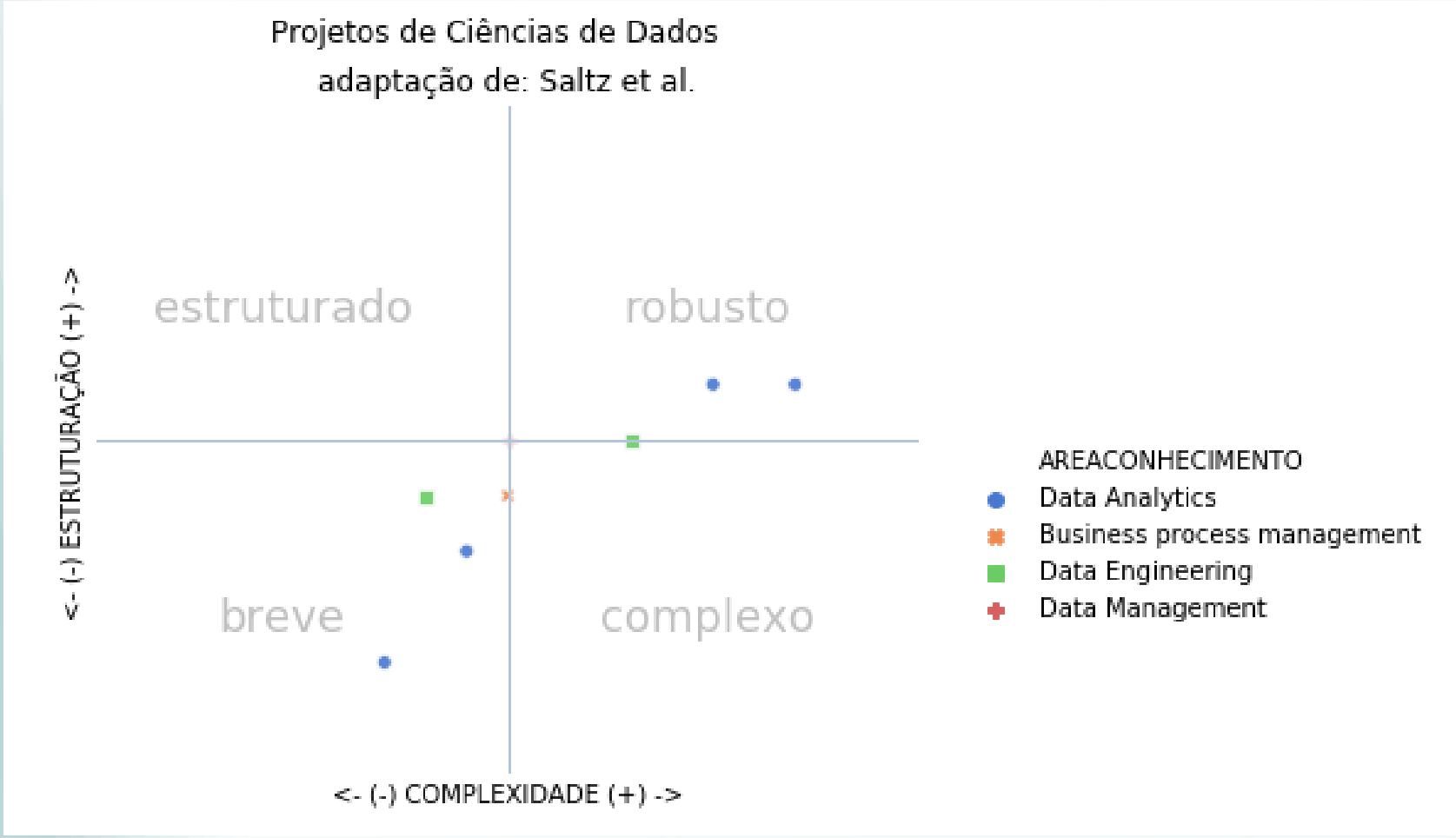
Referência: Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., and Brewer, S. (2016). Edison data science framework: a foundation for building data science profession for research and industry. In 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), pages 620–626. IEEE, 2016.

- o modelo 3DSE:
visual



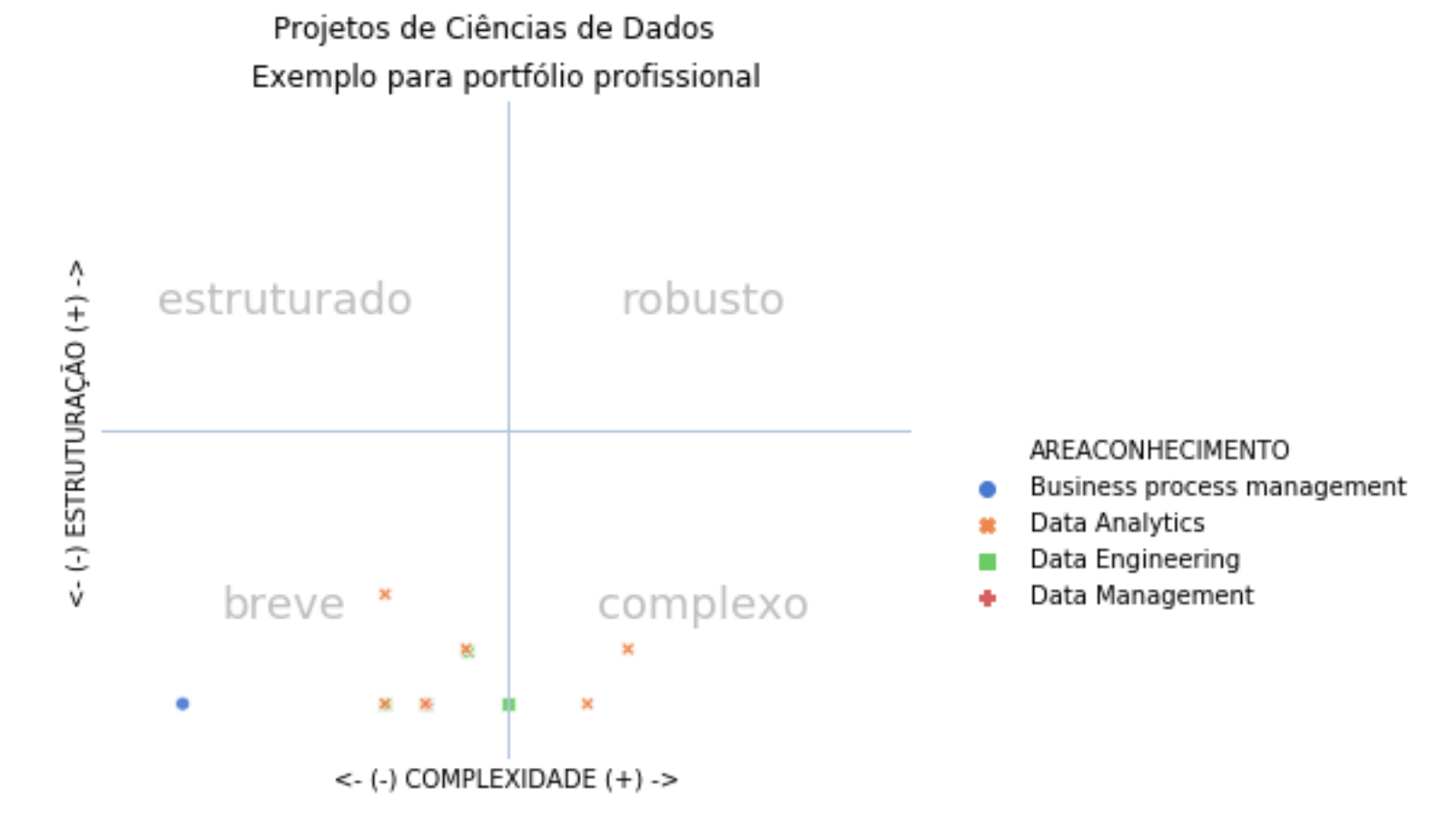
- aplicação

portfólios de projetos de um time ou empresa



- aplicação

portfólio profissional



Modelo 3DSE: Referências principais

Complexidade

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.

Estruturação

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.

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.

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.