A Review of Processes and Platforms for Developing AI-ML-Based Systems

Bachelor Thesis

# Organization

|  |  |
| --- | --- |
| Examiner: | Prof. Dr. Stefan Wagner |
| Supervisor: | Dr. Justus Bogner |
| Student: |  |
| Timeframe: | 2020-05-18 – 2020-11-18 |

# Context & Motivation

More and more software systems rely on techniques from machine learning or artificial intelligence to support decision making or to automate complex information processing. Such AI-ML-based systems [1] are still primarily software applications with a lot of traditional components [2]. Nonetheless, these systems require specialized software development methods, as traditional practices may not be feasible in several areas [3][4]. Over the years, several high-level processes have been proposed to guide the development of AI-ML-based systems, e.g. KDD [5], CRISP-DM [6], CommonKADS [7], ASUM-DM [8], SEMMA [9], TDSP [10], or CRISP-ML(Q) [11]. Likewise, a number of platforms was created to ease the development of applications and data pipelines, such as Model Governance [12], Amazon SageMaker [13], TFX [14], or the Microsoft AI Platform [15]. This plethora of options makes it difficult to compare available approaches and their different scopes, strengths, and weaknesses. Even though a small number of reviews exist [16][17][18], they are far from comprehensive. Additionally, no review so far covers both processes and platforms. Combining these concepts could provide insights into which platforms best support which processes.

# Objectives

The goal of this study is therefore to systematically collect and analyze existing processes and platforms for the development of AI-ML-based systems. The results should further the understanding of the general development process and available platform support for such systems and allow a comparison of existing options. More fine-grained research questions should be defined by the student.

# Methods

Develop a suitable methodology to answer the chosen research questions. Methods to collect and analyze existing literature such as systematic literature reviews or grey literature reviews may be a fitting choice. Conceptualize a theoretical framework for the comparison of identified processes and platforms. For the platforms, it could also be interesting to compare them with a small case study example project.

# References

[1] SE4ML - Software Engineering for AI-ML-based Systems, 2020, <https://www.dagstuhl.de/20091>

[2] T. Menzies, “The Five Laws of SE for AI,” *IEEE Softw.*, vol. 37, no. 1, pp. 81–85, Jan. 2020.

[3] S. Amershi *et al.*, “Software Engineering for Machine Learning: A Case Study,” in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 2019, pp. 291–300.

[4] Z. Wan, X. Xia, D. Lo, and G. C. Murphy, “How does Machine Learning Change Software Development Practices?,” *IEEE Trans. Softw. Eng.*, pp. 1–1, 2019.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “The KDD process for extracting useful knowledge from volumes of data,” *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[6] R. Wirth and J. Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000.

[7] G. Schreiber *et al.*, *Knowledge Engineering and Management: The CommonKADS Methodology*. Cambridge, MA, USA: MIT Press, 2000.

[8] ASUM-DM, 2018, <https://developer.ibm.com/technologies/artificial-intelligence/articles/architectural-thinking-in-the-wild-west-of-data-science/>

[9] Semma, 2017, <https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en>

[10] TDSP, 2020, <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/>

[11] S. Studer *et al.*, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology,” *Arxiv Prepr.*, Mar. 2020.

[12] V. Sridhar, S. Subramanian, D. Arteaga, S. Sundararaman, D. Roselli, and N. Talagala, “Model Governance: Reducing the anarchy of production ML,” in *Proceedings of the 2018 USENIX Annual Technical Conference, USENIX ATC 2018*, 2018, pp. 351–358.

[13] Amazon SageMaker, <https://aws.amazon.com/sagemaker/>

[14] D. Baylor *et al.*, “TFX: A TensorFlow-Based Production-Scale Machine Learning Platform,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17*, 2017, pp. 1387–1395.

[15] M. Salvaris, D. Dean, and W. H. Tok, “Microsoft AI Platform,” in *Deep Learning with Azure*, Berkeley, CA: Apress, 2018, pp. 79–98.

[16] R. Isdahl and O. E. Gundersen, “Out-of-the-Box Reproducibility: A Survey of Machine Learning Platforms,” in *2019 15th International Conference on eScience (eScience)*, 2019, pp. 86–95.

[17] L. A. Kurgan and P. Musilek, “A survey of Knowledge Discovery and Data Mining process models,” *Knowl. Eng. Rev.*, vol. 21, no. 1, pp. 1–24, Mar. 2006.

[18] A. Azevedo and M. F. Santos, “KDD, SEMMA and CRISP-DM: a parallel overview,” in *Proceedings of the IADIS European Conference on Data Mining*, 2008, pp. 182–185.