Exploratory study of Machine Learning Tools on Big Data platforms

2nd SYMPOSIUM OF APPLIED RESEARCH

Master in Computer Engineering

Filipe Ferreira da Ascenção, Prof. Nuno Lopes, Prof. Joaquim Silva

Instituto Politécnico do Cávado e do Ave

BACKGROUND

90% of all the data that has ever been created in the history, have been created in the last two years (SINTEF, 2013):

- These “Big data” comes from the public, private, personal, collective and business sectors
- Today everything can create data (e.g. smartphone, sensor, Web browsing, etc.)

BACKGROUND

Big Data processing platforms aim to aid on Data analysis and can provide to people and organizations

- a better way to understand their universe…
- a wider perspective to take better decisions

Machines learn by analyzing the processed data and allows people and organizations to take better decisions

- By Identifying easily new opportunities
- Avoiding possible risks

OBJECTIVES & METHODOLOGY

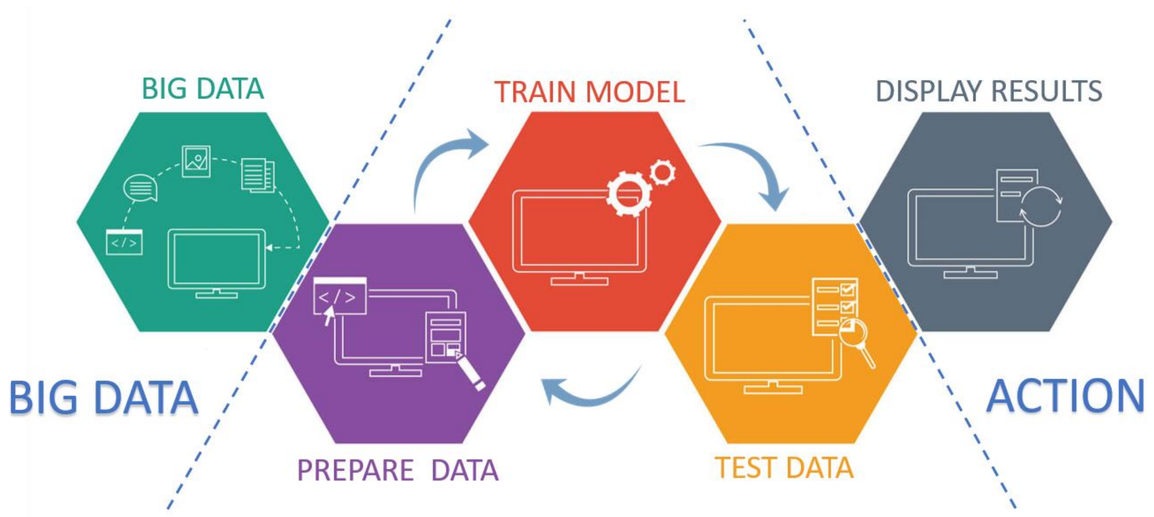
The personal motivation is to understand ML and Big Data concepts, and their importance in supporting decision making.

There are two main goals for this study:

- Identify available ML tools for data processing and analysis in big data platforms
- Assess a selected group of ML tools based on explicit criteria

The methodology adopted will consist of:

- literature review
- Identification and selection of existing ML tools,
- the evaluation of ML tools
- the writing of essays and conclusions.



RESULTS AND CONCLUSIONS

The expected results for this study, that will be documented in the dissertation, are two:

- List of existing ML tools for data processing and analysis in big data platforms
- Assessment of a selected group of ML tools based on explicit criteria

These results can help to do an introduction to ML tools for big data platforms and to select the most suited ML tool for a specific job.

BIBLIOGRAPHY

Aldhous, P. (2011). Peter Norvig: Google’s data junkie. New Scientist. https://doi.org/10.1016/S0262-4079(11)60987-1

Dean, B. Y. J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. Communications of the ACM, 53(1), 72–77. https://doi.org/10.1145/1629175.1629198

Nilsson, N. J. (2005). Introduction to Machine Learning. Machine Learning, 56(2), 387–99. https://doi.org/10.1016/j.neuroimage.2010.11.004

Provost, F., & Fawcett, T. (2013). Data Science for Business. Book. https://doi.org/10.1007/s13398-014-0173-7.2

SINTEF. (2013). Big Data, for better or worse: 90% of world’s data generated over last two years. Retrieved from http://www.sciencedaily.com/releases/2013/05/130522085217.htm

Chart adapted from https://www.class-central.com/report/best-machine-learning-courses/

IPCA INSTITUTO POLITÉCNICO DO CÁVADO E DO AVE

ESCOLA SUPERIOR DE TECNOLOGIA IPCA

digarc Digital Games Research Center

CONTACTS

A12749@alunos.ipca.pt

nlopes@ipca.pt

jpsilva@ipca.pt