Contribution ID: **35** Type: **Oral Presentation**

A Data Science framework in the INCD

Tuesday, September 24, 2019 3:00 PM (15 minutes)

INCD - National Distributed Computing Infrastructure is a Portuguese digital infrastructure designed to support the national scientific community, providing computing and storage services to the national scientific and academic community in all areas of knowledge. LNEC – National Laboratory for Civil Engineering is one of the partners that collaborate in this initiative, developing use cases that take advantage from the available infrastructure. This work reports a Data Science framework based on Conda that was developed as part of this collaboration. The use of this framework allows researchers to benefit from the INCD infrastructure, running their research scripts, using the several Conda packages available, including Jupyter Notebook. To showcase the framework, two case studies were implemented, demonstrating the use of Machine Learning algorithms applied to data generate from the dam safety monitoring systems.

The first case study presents a prediction setting, implemented in Python, in which Multiple Linear Regression (MLR) and Neural Networks (NN) are trained and used to predict dam behavior in manually collected data. Environmental variables are used as predictors for both the MLR and the NN. Both predictions are evaluated and compared, also using the develop framework. Note that such predictions heavily depend on the specific properties of each data set. Thus, the capabilities of this environment on top of INCD infrastructure enable a flexible adaptation of each prediction that can be easily tuned to each specific case.

A classification task is proposed in the second case study, implemented in Python and R, using the DBSCAN (Density-based spatial clustering of applications with noise) clustering algorithm to identify outliers in automatically collected data from sensors installed on Portuguese dams. Together with the dam response variable, environmental variables are used to obtain the clusters and detect outliers. Afterwards, PCA (Principal Component Analysis) is used to obtain a 2D plot to visualize outliers identified by the DBSCAN.

Other pieces of research were also developed using this framework, including the use of Deep Learning (more specifically, Recurrent NN) to improve prediction of dam behavior, using Keras and TensorFlow, which benefited from the INCD infrastructure for improved computation times.

Finally, it is important to remark that framework was recently presented to several LNEC researchers and was received with a large interest, with most of the researchers already starting to use INCD to create and run their scripts in a cloud environment.

Primary authors: ANTUNES, António (Laboratório Nacional de Engenharia Civil); MARTINS, Tiago (Laboratório Nacional de Engenharia Civil); Dr BARATEIRO, José (Laboratório Nacional de Engenharia Civil); Dr OLIVEIRA, Anabela (Laboratório Nacional de Engenharia Civil); Dr AZEVEDO, Alberto (Laboratório Nacional de Engenharia Civil)

Presenter: ANTUNES, António (Laboratório Nacional de Engenharia Civil)

Session Classification: IBERGRID Contributions

Track Classification: Enabling Research Applications in advanced Digital Infrastructures