
Contribution ID: **258** Type: **Orale**

# Software Defect Prediction on Unlabelled Dataset using Machine Learning

*Wednesday, 5 June 2019 17:30 (30 minutes)*

Machine Learning (ML) has proven to be of great value in a variety of Software Engineering tasks, such as software defects prediction and estimation and test code generation. To accomplish these tasks, datasets (e.g. features represented by software metrics) have to be collected for the various modules (such as files, classes and functions) and properly preprocessed before the application of machine learning techniques. These activities are essential to manage missing values and/or removal inconsistencies amongst data.

Typically, new projects or projects with partial historical data may lack some features' data, e.g. defect data are not included. Their datasets are called unlabelled datasets and are the vast majority of software datasets. The extraction of the complete set of features (defectiveness included) and the labelling of the various instances imply effort and time. In literature there exist various approaches to build a prediction model on unlabelled datasets that entail a high number of permutations that is extremely time consuming. Cloud computing infrastructure, GPU-equipped resources and adequate ML framework can give the chance to overcome this problem.

In this study, we are going to present the analysis of existing software unlabelled datasets by implementing models in different available frameworks, such as TensorFlow and Keras, and running in Python and R. Recently, as a work in progress, we started to explore the application onto a large code base like the full software stack of the CMS experiment at the LHC collider at CERN. We have evaluated these frameworks by considering three aspects: extensibility, hardware utilization and speed. We intend to reduce the distance between theory and practice by providing strengths and limitations of the considered frameworks to enable users to assess suitability according to their requirements.

**Primary authors:** RONCHIERI, Elisabetta (CNAF); CANAPARO, Marco (CNAF); SALOMONI, Davide (CNAF)

**Presenter:** RONCHIERI, Elisabetta (CNAF)

**Session Classification:** Tecnologie Software e ML

**Track Classification:** Machine Learning