Within-project Defect Prediction of Infrastructure-as-Code Using Product and Process Metrics

Stefano Dalla Palma
Jheronimus Academy of Data Science
Tilburg University
Tilburg, The Netherlands
s.dallapalma@uvt.nl

Fabio Palomba
Software Engineering (SeSa) Lab
University of Salerno
Salerno, Italy
fpalomba@unisa.it

***Abstract*—Infrastructure-as-code (IaC) is the DevOps practice enabling management and provisioning of infrastructure through the definition of machine-readable files, hereinafter referred to as *IaC scripts*. Similarly to other source code artefacts, these files may contain defects that can preclude their correct functioning. In this paper, we aim at assessing the role of *product* and *process* metrics when predicting defective IaC scripts. We propose a fully integrated machine-learning framework for IaC Defect Prediction, that allows for repository crawling, metrics collection, model building, and evaluation. To evaluate it, we analyzed 104 projects and employed five machine-learning classifiers to compare their performance in flagging suspicious defective IaC scripts. The key results of the study report RANDOM FOREST as the best-performing model, with a median AUC-PR of 0.93 and MCC of 0.80. Furthermore, at least for the collected projects, product metrics identify defective IaC scripts more accurately than process metrics. Our findings put a baseline for investigating IaC Defect Prediction and the relationship between the product and process metrics, and IaC scripts' quality.**[1]

***Index Terms*—Infrastructure-as-code; Defect Prediction; Empirical Software Engineering.**

I. OBJECTIVE

This work aims *to help software practitioners prioritize their inspection efforts for IaC scripts by proposing prediction models of failure-prone IaC scripts and investigating the role of product and process metrics for their prediction*. To this end, we propose the RADON FRAMEWORK FOR IAC DEFECT PREDICTION, a fully integrated Machine-Learning-based framework that allows for repository crawling, metrics collection, model building, and evaluation. The framework assessment led to the definition of the following research questions:

- **RQ$_1$** *To what extent does the classifier selection impact the performance of Machine-Learning models to **predict the failure-proneness of IaC scripts**?*
- **RQ$_2$** *How is the prediction performance affected by the choice of the **metric sets**?*
- **RQ$_3$** *Which metrics are **good defect predictors**? That is, what are the most selected predictors and their combinations?*

[1]Full paper available at: https://ieeexplore.ieee.org/document/9321740

II. RQ1

RQ1 aims at identifying the effect that the choice of classifiers (e.g., Naive Bayes and Random Forest) has on the prediction performance. We gathered a comprehensive and meaningful set of failure-prone IaC scripts and metrics to implement and assess different classifiers for predicting the failure-proneness of an IaC script. Afterward, we compared their performance and focused on RANDOM FOREST as the best performing model. The contribution is a *set of classifiers suitable for the detection of suspicious failure-prone IaC scripts*.

RQ$_1$ summary: *The models trained using* RANDOM FOREST *perform statistically better than those relying on the remaining classifiers. The difference is statistically different with large effect size.*

III. RQ2

RQ2 aims at identifying the effect that the choice of metric sets (i.e., code and process metrics, and groups thereof) has on the prediction performance.

RQ$_2$ summary: *The models which feature IaC-Oriented metrics perform statistically better than those relying on the remaining metric sets. The difference is statistically significant with large effect size.*

IV. RQ3

Finally, **RQ3** aims to identify and rank the measures that highly affect the prediction performance. A recursive feature selection method is performed to find the optimal number of features and to rank them according to their importance for the prediction. The contribution is *a set of metrics for the detection of suspicious failure-prone IaC scripts* that DevOps engineers and researchers can use to further understand and assess the quality of IaC scripts.

RQ$_3$ summary: *IaC-oriented metrics tend to maximize the prediction performance. In particular, the number of tokens, text entropy and number of code lines are the most occurring predictors.*