# ZADÁNÍ BAKALÁŘSKÉ PRÁCE

## I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Sahula** Jméno: **Lukáš** Osobní číslo: **435008**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Softwarové inženýrství a technologie**

## II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

**Neúplná Data a Rozhodovací Lesy v Úloze Klasifikace Šifrovaného síťového provozu**

Název bakalářské práce anglicky:

**Handling Missing Values in Decision Forests in the Encrypted Network Traffic**

Pokyny pro vypracování:

The thesis addresses the problem of training a classifier from data with missing values. More specifically, a Random Forest is to be trained to classify malware from proxy log data. The main focus is on implementing missing-valuehandling method(s) that would lead to improved classification performance, measured by precision and recall.
The concrete goals are:
1) Review prior art in the area of handling missing values in sparse datasets that are applicable to decision forest based classifiers and choose methods suitable for implementation in the thesis´ context.
2) Consider if a modification of existing tools or proposal of entirely new method would be beneficial for the specific case of proxy log classification.
(optional)
3) Implement the chosen methods in Python and connect them to a framework for model training and evaluation (developed prior as part of the software engineering project).
4. Evaluate and compare the classification performance (precision, recall) of implemented methods on a provided test set to the baseline method based on replacing missing values by a constant.

Seznam doporučené literatury:

[1] Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114, 5(6), 12.
[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning 2nd edition.
[3] Stekhoven, D. J., & Bühlmann, P. (2011). MissForest?non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
[4] Feelders, A. (1999). Handling missing data in trees: surrogate splits or statistical imputation?. Principles of Data Mining and Knowledge Discovery, 329-334.
[5] Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). When Should We Ignore Examples with Missing Values?. International Journal of Data Warehousing and Mining (IJDWM), 13(4), 53-63.
[6] Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. Journal of machine learning research, 8(Jul), 1623-1657.

 © ČVUT v Praze, Design: ČVUT v Praze, VIC

Jméno a pracoviště vedoucí(ho) bakalářské práce:

**Ing. Jan Brabec, katedra počítačů FEL**

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **09.02.2018** Termín odevzdání bakalářské práce: **25.05.2018**

Platnost zadání bakalářské práce: **30.09.2019**

| Ing. Jan Brabec<br>podpis vedoucí(ho) práce | podpis vedoucí(ho) ústavu/katedry | prof. Ing. Pavel Ripka, CSc.<br>podpis děkana(ky) |
|---|---|---|

## III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. Datum převzetí zadání

Podpis studenta

 © ČVUT v Praze, Design: ČVUT v Praze, VIC