POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Operating System Detection from Network Traffic**
Jméno autora:	**Bc. Anastasiia Kuznetsova**
Typ práce:	diplomová práce
Fakulta:	Fakulta jaderná a fyzikálně inženýrská (FJFI)
Katedra:	Katedra softwarového inženýrství
Oponent práce:	Ing. Jiří Franc Ph.D.
Pracoviště oponenta práce:	Katedra matematiky FJFI ČVUT

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

The thesis explores the passive identification of device operating systems (OS) in a network by analyzing visited hostnames and proposes data transformations and machine learning models for OS detection, which are then tested and compared for effectiveness.
The assignment appears to be a complex task as it requires a understanding of network traffic data, data transformations, and machine learning models. The author needs to not only propose and implement these models but also conduct rigorous testing and comparison to determine the most effective approach.
However, since the author did not have to collect the data herself and could use implemented methods in available libraries for classification, I rate the assignment as moderately difficult.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

All criteria from the assignment were met, though it's worth noting that these requirements were somewhat broad. Consequently, certain methodologies were employed that may not have been entirely suitable or convincing, but they were nevertheless utilized.

Zvolený postup řešení **vhodný s výhradami**

Posuďte, zda student zvolil správný postup nebo metody řešení.

The solution procedure is correct and common, necessitating data preprocessing, appropriate feature selection, and the application of a classification method. Regrettably, the selected methodologies and their implementation, as depicted in the work, exhibit traces of imprecision.

Odborná úroveň **průměrná**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

The primary contribution of the work lies in detailing the process of preparing and processing the data sample for analysis. While I find the exploration of basic methods grounded in decision trees to be adequate, their application is presented inadequately, leaving several fundamental questions unanswered.

Formální a jazyková úroveň **výborná**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

The thesis is written in English, demonstrating a good proficiency in language. The majority of the mathematical content is confined to the research section on classification models, where it should be noted that formulas are integral parts of the sentences. Unfortunately, the diagrams representing the decision trees aren't directly related to the researched problem; they seem to pertain more to bananas and blueberries.

However, plotting graphs for the dataset under investigation is straightforward with the assistance of the Scikit-learn library.

Výběr zdrojů, korektnost citací **průměrné**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

The number of cited sources used in the work is sufficient. Due to the uniqueness of the dataset, we cannot expect many articles on similar topics. Thus, most of the citations relate to the general task of machine learning classification.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

My primary concerns with the work arise from the omission of certain critical details that are essential both for understanding the work and potentially replicating the results.

In this work, I find the absence of raw data samples, as they were initially received by the student, and the corresponding pipeline detailing how the data was progressively transformed into its final form for the classifier. The provided statistics are appreciable, yet they do not permit readers to form their own judgment regarding the procedure's correctness.

Another significant deficiency I perceive is in the explanation of how the classification was executed. Within this work, I am unable to determine how the data was divided into testing and training subsets. It is unclear whether GridSearchCV was solely employed for hyperparameter tuning or if the results subsequently presented are already the output from the cross-validation. The majority of my defense questions will pertain to this particular aspect.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

During the defense, I require answers to the following questions:

- Present part of the raw data as you received it, how it was divided into training and testing parts, and how gradually the transformations and on the basis of which data were performed.
- Present the tabulated data that entered the classifier and confirm that the response distribution and some features are the same for both data samples.
- How did you check if the model was not overtrained? If you have no tree depth set, min_samples_split = 2 and min_samples_leaf =1 what will each resulting tree look like?
- Have you tried different objective function or combining the results of multiple models together?
- Where can I find the code used in the thesis and is it possible to replicate it?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **C - dobře.**

Datum: 25.5.2023 Podpis: