POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Neúplná Data a Rozhodovací Lesy v Úloze Klasifikace Šifrovaného síťového provozu**
Jméno autora:	**Lukáš Sahula**
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Ing. Martin Svatoš
Pracoviště oponenta práce:	Katedra počítačů

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Student se musel seznámit s pojmy z oblasti strojového učení, rozhodovacích stromů, dat s chybějícími hodnotami a nevyváženými datasety. Naimplementování a porovnání stávajících metod na datasetu hodnotím jako průměrné.

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.

Všechny body zadání byly diskutovány.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Student zvolil správný přístup pro řešení daného problému. Velmi kladně hodnotím analýzu datasetu v sekci 3.2, která předchází výběru metod (tedy stěžejní části práce).

Odborná úroveň **C - dobře**

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

Výběr metod, návrh experimentů i metodologie působí rozumně, avšak má k nim následující výhrady. V práci není uvedeno nastavení experimentů (počet stromů, hloubky stromů, nastavení generátoru pro metodu OTFI,…); to vše kvůli reprodukovatelnosti experimentů.
Dále bych očekával při vyhodnocení precision (sekce 5.2.1) v obrázku č. 9 alespoň směrodatné odchylky. Vzhledem k porovnání nedeterministických algoritmů bych však daleko více uvítal odhad testovací chyby z několika běhů s různou konfigurací (například jiný seed generátoru náhodných čísel).

Formální a jazyková úroveň, rozsah práce **B - velmi dobře**

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

Práce je psaná angličtinou s pár překlepy (*email*, *corellation*). Autor používá pro odkazování citované literatury i algoritmů, obrázků a tabulek čísla v hranatých závorkách, i když se tento způsob zpravidla používá pouze odkazy na literatury. Číslování v práci začíná už od první strany, nikoliv až od první kapitoly, jak bývá zvykem**.**
V druhém odstavci sekce 2.1.4 má být patrně průnik dat levého a pravého potomka roven prázdné množině, nikoliv θ (která v následujícím algoritmu značí nejlepší rozdělení dat ve vnitřním uzlu).
V sekci 2.1.7 rovnice definice entropie (2) obsahuje zbytečně navíc horní mez sumy *n*.
První odstavec poslední kapitoly: pojem *škálovat* se používá ve smyslu asymptotické složitosti. Pro vyjádření toho, že přesnost (precision, recall,…) je dobrá i na datech s velkým množstvím chybějících dat, by bylo lepší použít jiný termín, například robustnost.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

Výběr zdrojů, korektnost citací	**B - velmi dobře**
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.	
Práce cituje relevantní zdroje v oblastí náhodných a rozhodovacích stromů, avšak bylo by hezčí uvést více metod v alespoň základních rysech, popřípadě rozdíly, namísto prostého odkazu na článek, který je popisuje ([3]). Sekce 2.1.4 popisuje obecný algoritmus TDIDT a bylo by proto hezčí citovat některou z prvních prací popisující tento algoritmus (například publikace od J. R. Quinlan z 80. let) namísto pár let staré diplomové práce.	

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.
Implementace je psána čistým kódem s dostatečným množstvím komentářů. Zdařile vypadá i využití Pythoních knihoven (např. scipy, numpy,…). Velice kladně také hodnotím instalační skript, který se postará o instalaci potřebných knihoven. Jedinou poznámku mám ke třídě *DecisionTree* a vnitřní implementace metody *fit*, protože jednotlivé přístupy (OTFI, MIA,…) se liší hledáním rozdělení (*splitu*, tvorbou pravého a levého potomka), zatímco zbylá část metody (koncová podmínka apod.) je obecnou charakteristikou TDIDT.

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.

Předložená práce porovnává různé techniky nahrazování náhodných hodnot pro úlohu *multi-class* klasifikaci v nevyváženém datasetu, což zahrnuje především nastudování příslušné odborné literatury.

První kapitoly práce jsou zdařilé, popisují základní kameny přístupu, motivaci a velmi správná je analýza datasetu před dalším postupem. Z následujících kapitol bych uvítal rozsáhlejší popis metod a článků zabývajících se problémem doplňování chybějících dat na jiných typech datasetů. V experimentech samotných bych pak uvítal jednak alespoň nástin parametrů, se kterými byly experimenty puštěny. Hodnoty precision se občas liší o několik málo procent, proto by bylo lepší prezentované výsledky nedeterministických algoritmů byly například průměrem vícero běhů s různými konfiguracemi.

Některé diskutované metody nebyly otestovány, protože, jak je psáno v textu, je potřeba aby výsledný model škálovala. Jsou nějaké výrazné rozdíly v rychlostech klasifikace mezi testovanými metodami?

I přes zmíněné nedostatky hodnotím práci jako zdařilou a to především přihlédnutí k tématu, které je poměrně specifické, analýze datasetu a přiložené implementaci.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **B - velmi dobře.**

Datum: 6.6.2018 Podpis: