FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Michal Štepanovský, Ph.D.
Student:	Vojtěch Chvojka
Název práce:	Detekce těžby kryptoměn na základě periodického chování síťové komunikace
Obor / specializace:	Bezpečnost a informační technologie
Vytvořeno dne:	13. června 2023

Hodnotící kritéria

1. Splnění zadání

- ▸ [1] **zadání splněno**
  [2] zadání splněno s menšími výhradami
  [3] zadání splněno s většími výhradami
  [4] zadání nesplněno

2. Písemná část práce

50 / 100 (E)

Písemné části práce mohla být věnována větší pozornost. Oceňuji, že student odkazuje na použité zdroje. Přestože, celkový počet citovaných zdrojů by pro bakalářskou práci mohl být dostatečný, není zcela obvyklé čerpat z jednoho zdroje tak rozsáhlé pasáže textu, jak to lze dohledat v této práci (viz například odstavec "2.1.1 Co to jsou kryptoměny", který téměř celý vychází z jediného zdroje). Vhodnější by bylo čerpat a ověřovat informace z několika zdrojů a ty společně citovat. Jako do očí bijící příklad nevhodnosti zvoleného přístupu (tedy použití jediného zdroje a bezhlavé převzetí tam uvedených informací) mohu zmínit například odstavec "2.4 Strojové učení", ve kterém student dle literatury [35] člení rozhodovací stromy na 4 kategorie: "1. Klasifikační strom", "2. Regresní strom", "3. Stromové lesy" a "4. Klasifikační a regresní strom". Překvapuje mě, že si student nevšiml nelogičnosti tohoto členění, ve kterém se míchají stromy a lesy. Dalším indikátorem nevhodnosti literatury [35] mohlo být opakované použití termínu "klasifikační strom" pro první a čtvrtou kategorii. Jako poslední indikátor nevhodnosti literatury [35], kterého si student měl všimnout, mohu zmínit převzaté popisy těchto stromů/lesů. Například, co znamená tvrzení: "Aby byl výsledek co nejlogičtější"? V této souvislosti mě rovněž překvapuje, že student naštěstí nepřevzal kategorii "5. K Means Clustering" z výše uvedené literatury, která k-means clustering mylně ztotožňuje s rozhodovacím stromem.

Dále, pokud uvážím algoritmy, které student analyzoval v praktické části (viz Tabulka 4.1), očekával bych, že právě tyto algoritmy budou předmětem odstavce "2.4 Strojové učení". Čtenář se tedy v teoretické části o těchto algoritmech nic nedozví.

Student rovněž v některých případech odkazuje velmi nepřehledně na obrázky, tabulky či rovnice. Jako příklad uvádím první větu odstavce 3.3.2: "Jak je vidět na obrázcích 3.4 3.5 3.6"... (chybí čárka a spojka a). Přestože lze pozorovat snahu, někdy vůbec není zřejmé, na co se vlastně odkazuje. Jako příklad uvádím větu z odstavce 4.3: "Když zobrazíme významnost vlastností podle skupiny vlastností 4.2 vidíme, že...". Odkazuje se na Tabulku 4.2, Obrázek 4.2, Odstavec 4.2? Dále mi v práci překáží nedůslednost dodržování názvosloví. Například student používá tři termíny "miner", "težař" a "horník". Jako další příklad mohu uvést Tabulku 4.1 a text pod ní. V tabulce se uvádí HitGradientBoostingClasifier, v textu pod ní pouze HitGradientBoosting. V textu se uvádí, že si nejlépe vedl XGBoost. V tabulce 4.1 jej však nenacházím. Má se jednat o XGBClasifier? Podobně, student cituje dvojím způsobem - pomocí číselného odkazu a nedůsledně harvardským stylem (viz například popisek Obrázku 3.9).

3. Nepísemná část, přílohy 70 /100 (C)

Nepísemná část práce představuje několik souborů v jazyce Python (celkem přibližně tisíc řádků kódu - předpokládám, že student je autorem všech souborů, ve kterých v hlavičce není uveden autor). Student rovněž využívá programy dodané vedoucím práce. Ocenil bych podrobnější dokumentaci vytvořených funkcí. V některých případech dokumentace chybí zcela. Studentem odvedená práce je dokumentována/podložena v písemné části v Kapitole 4 "Detekce těžby kryptoměn". Písemná a nepísemná část práce mohly být lépe provázány.

4. Hodnocení výsledků, jejich využitelnost 100 /100 (A)

Výsledky práce jsou v praxi využitelné. Deklarovaná úspěšnost klasifikace je povzbuzující. Nicméně, z principu tvorby datové sady je možné, že některé těžby kryptoměn nebudou vůbec detekovány. V odstavci 2.3.1 je uvedeno: "Každý účastník provozu (IP adresa + port), který byl podezříván z provozování poolu byl požádán, aby poskytl práci pro Plného těžaře, čímž mu potvrdil, že je to pool a jeho komunikace byla označena jako miner [22]." Z toho plyne, že pouze již podezřelá komunikace bude po potvrzení označena jako miner. Tento nedostatek však nepřikládám studentovi.

Celkové hodnocení 55 /100 (E)

Mé celkové hodnocení je velmi ovlivněno nekvalitně zpracovanou písemnou částí práce. Vzhledem k omezenému prostoru jsem nevyjmenovával všechny nalezené nedostatky. Na druhou stranu, student splnil zadání práce. Proto předloženou práci doporučuji k obhajobě a hodnotím stupněm E.

Otázky k obhajobě

Použitý dataset je tvořen 2 soubory: decrypto_dataset_design.csv a decrypto_dataset_evaluation.csv. Z práce není zcela zřejmé, kdy byl který soubor použit. Mám dojem, že při výběru vhodného klasifikátoru (viz Tabulka 4.1) student použil oba soubory. Rovněž si nejsem jist, jak student postupoval při hledání nejlepších parametrů

pro klasifikátor (odstavec 4.2). Testovací data totiž nesmí být použita při hledání vhodného klasifikátoru a jeho parametrů. Testovací data se mohou použít pouze pro finální odhad úspěšnosti již vyladěného klasifikátoru. Může student při obhajobě vysvětlit, jak pracoval s trénovacími a testovacími daty?

Instrukce

Splnění zadání

Posuďte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti.

Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Nepísemná část, přílohy

Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce – opakovatelnost experimentů.

Hodnocení výsledků, jejich využitelnost

Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Celkové hodnocení

Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.