FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Magda Friedjungová, Ph.D.
Student:	Jan Čáp
Název práce:	Automatická explorační analýza dat pro binární klasifikaci pomocí knihovny pandas profiling
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	10. 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

Zadání bylo splněno.

2. Písemná část práce

88 / 100 (B)

Práce je psána v češtině, je velmi dobře čitelná a má rozumný rozsah. Jednotlivé kapitoly jsou informačně vyvážené, jejich členění je logické. Práce obsahuje několik překlepů a pár gramatických chyb. Student místy používá výroky, ke kterým chybí reference, např. "je nejrozšířenější knihovnou" apod. Jinak je práce s referencemi dobrá.

Teoretická část práce začíná rešerší, která obsahuje ukázky výstupů jednotlivých knihoven na zadaném datasetu. V sekci 1.1 výrok "Metoda pracuje pouze s číselnými veličinami, textové a kategoriální veličiny do tabulky nezahrnuje." pro metodu describe() není úplně přesný, metoda poskytuje popisné statistiky i pro příznaky typu object. Trochu rozkol vnímám v podkapitole 1.6, kdy student vybírá vhodnou knihovnu pro svá rozšíření, nicméně v zadání práce je již určeno, že se má jednat o knihovnu Pandas Profiling (kterou student v dané podkapitole také zvolí). Nicméně je zde provedena pěkná diskuze. Dále následuje teoretická část, ve které se student věnuje statistickým textům a vizualizacím. Při uvádění matematických vzorců by bylo vhodné zadefinovat všechny použité proměnné. Vzhled vizualizací by bylo vhodné sjednotit - některé grafy mají nadpisy, některé mají osy popsané v češtině, některé v angličtině. Student dále navrhuje konkrétní úpravy, které do knihovny následně implementuje. 1. Před popisem úprav je v kapitole 3 také popsáno vyšetření jednotlivých proměnných a typy upozornění v případě "zajímavostí" v dané proměnné. Studentovo rozšíření implementované v kapitole 4

umožňuje mj. zadat cílovou proměnnou. Z textu však není jasné, zda je i cílová proměnná stejně vyšetřena jako ostatní proměnné v reportu. Tzn. také u ní probíhá vyšetření chybějících hodnot, konstant apod.? Z odevzdaného kódu a ukázek je pak patrné, že vyšetřena je, do textu by to však bylo vhodné doplnit. Dále mi chybí zmínka při implementaci, jakým způsobem student do knihovny přispívá. Zda např. vytvořil pull requesty do knihovny na GitHubu. Jinak je praktická část pěkně popsaná a nemám k ní další připomínky.

3. Nepísemná část, přílohy 89 /100 (B)

Student odevzdal jak vstupní a výstupní data (reporty), tak knihovnu s implementovanými úpravami, jejíž spuštění přehledně popsal v readme. Doplnila bych pouze informaci, jak probíhalo reálné zadávání úprav do knihovny, abychom byli schopni ověřit reálný přínos práce, kterým je právě ono poskytnutí data science komunitě. Jinak student zvolil vhodné technologie a popsané experimenty jsou dostatečné.

4. Hodnocení výsledků, jejich využitelnost 92 /100 (A)

V rámci práce student popsal a realizoval rozšíření knihovny Pandas Profiling o nový datový typ, vizualizace, statistické testy a zohlednění binární klasifikace včetně evaluace pomocí modelu. Zmíněné úpravy zřejmě ještě čekají na zapracování do ostré verze knihovny, nicméně práce je přínosná a užitečná pro data science komunitu.

Celkové hodnocení 90 /100 (A)

Práci navrhuji hodnotit klasifikačním stupněm A a to z výše zmíněných důvodů.

Otázky k obhajobě

1. Mohl byste prosím stručně popsat, jakým způsobem probíhal návrh vašich úprav autorům knihovny? A v jakém stavu je integrace vašich úprav nyní?
2. Zkoušel jste ze zvědavosti i jiný referenční model než rozhodovací strom s gradientním boostingem? Přeci jen při robustnosti rozhodovacího stromu je otázkou, zda budou změny v klasifikační přesnosti po provedení transformací dostatečně znatelné.

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.