FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Daniel Vašata, Ph.D.
Student:	Tomáš Jungman
Název práce:	Čištění dat pomocí pravděpodobnostního programování
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	6. června 2022

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 bez výhrad.

2. Písemná část práce 75 /100 (C)

Práce je logicky strukturovaná. Po jazykové stránce je práce na průměrné úrovni a obsahuje relativně malé množství nedokonalostí a poněkud nestandardních obratů (např. část 3.2 - především poslední odstavec; odkazování na "kapitoly" i u podsekcí - např. poslední věta v části 4.2.5). Po typografické stránce je práce až na nepatrné množství nedostatků v pořádku. Zdroje jsou relevantní a správně citované.
Kromě výše uvedeného mám ještě několik dalších výtek. Z pohledu informační bohatosti by první kapitolou měla být až kapitola 3 a kapitoly 1 a 2 by měly být spíše nečíslované sekce. Velmi zvláštní jsou obsahy kapitol psané pod názvem kapitoly kurzívou. V tabulce 5.6 není uvedeno, jak se liší ty tři skupiny výsledků.

3. Nepísemná část, přílohy 80 /100 (B)

Nepísemnou částí práce bylo použití knihovny PClean pro sérii experimentů na oficiálních příkladech použití PClean a pak na reálné datové sadě pro predikci cen automobilů. V příloze práce jsou uvedené poměrně jednoduché zdrojové kódy, které těmto experimentům odpovídají. Na první pohled tyto kódy vypadají značně triviálně, ale věřím, že vzhledem k nedozrálosti knihovny PClean dalo jejich zprovoznění poměrně značnou práci.

4. Hodnocení výsledků, jejich využitelnost 80 /100 (B)

Výsledkem je jednak úvod do problematiky doplňování chybějících hodnot s přihlédnutím k využití pravděpodobnostního programování a pak také samotné experimenty, které reálně ukazují výhody a nedostatky prezentované knihovny PClean. Z pohledu prezentovaných výsledků se mi zdá poměrně zvláštní, že počet tříd v tabulce 5.1 nebo typ apriorního rozdělení v tabulce 5.3 mají vliv na úplnost doplnění. To v práci není dobře popsáno. Také bych očekával při evaluaci experimentů porovnání s nějakým dalším méně triviálním přístupem k doplňování - tj. např. s doplňováním pomocí metody nejbližších sousedů, která je v scikit knihovně implementována ve třídě KNNImputer a umožňuje velmi snadné použití. Na druhou stranu ale vnímám netriviálnost automatického doplňování a kladně hodnotím schopnost autora získat rozumné výsledky na reálné datové sadě.

Celkové hodnocení 78 /100 (C)

Celkově je práce na průměrné úrovni a vzhledem k výše uvedeným nedostatkům navrhuji její hodnocení stupněm C.

Otázky k obhajobě

Proč v tabulkách 5.1 a 5.3 nejsou stejné hodnoty úplnosti doplňování?
Zkoušel jste měřit vliv doplňování chybějících hodnot při procesu učení (s tím, že by evaluace proběhla na nepoškozených datech)?

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.