# Posudek oponenta diplomové práce

Autor/autorka práce: **Marek Lovčí**

Název práce: **Návrh metodiky pro vyhodnocení kvality datových sad**

**Obsah práce**

Práce je členěna do pěti hlavních kapitol. Autor nejprve popisuje existující metodologie ke stanovení úrovně kvality dat (2.1–2.5), náznakem zmiňuje problém se zabezpečením dat (2.6) a uvádí běžné chyby v datech (2.7). V následující kapitole je navržen metodologický rámec, jehož obecnost je demonstrována na třech hypotetických případech. V kapitole 4 autor definuje šest dimenzí datové kvality a náznakem popisuje způsob stanovení celkového hodnocení. V této kapitole je dále uveden dotazník sloužící k ohodnocení kvality dat v nemocničním systému, jehož modifikaci autor použil ve své případové studii, o které pojednává kapitola 5. V kapitole 6 je stručné zhodnocení.

Text práce je navenek strukturován do vhodných logických celků, ale při čtení působí nesourodě. Autor se zřejmě snažil zeširoka zachytit několik souvisejících problémů a je těžké vypozorovat jednu hlavní myšlenku, která by vše propojovala. Jedna z vedlejších myšlenek je například bezpečnost dat, kterou autor rozvíjí zmínkou o blockchain technologiích a v jednom modelovém scénáři. Nicméně do praktické části toto nijak nezasahuje. Podobně je tomu s výčtem typů datových chyb.

**Kvalita řešení a dosažených výsledků**

V práci identifikuji tři výsledky: 1) teoretický návrh obecné metodologie; 2) teoretický návrh objektivního způsobu určení celkové kvality pomocí šesti kvantitativních dimenzí; 3) praktické vyhodnocení kvality otevřené datové sady Covid-19 poskytované Ministerstvem zdravotnictví ČR.

Teoretické návrhy (1, 2) dávají smysl a dokáži si představit jejich praktickou aplikaci.

Zvolený způsob zkoumání kvality vybrané datové sady (3) považuji za **nevhodný**. Autor používá dotazník publikovaný v 2004, který má za cíl vytipovat problémové oblasti datové kvality z pohledu „sběrače dat“, což by v kontextu zvolené sady byly nemocnice nebo správce datových registrů (ÚZIS). Autor práce, který je v pozici pouhého uživatele dat, není z principu schopný zodpovědět na desítky otázek.

Dotazník v původní verzi pracuje se čtyřmi odpověďmi (neznámé, nesplněné, splněné) a otázky jsou seskupené do několika kategorií. Pokud je na jednu z otázek v sadě odpovězeno „neznámo“, celá kategorie je také hodnocena jako „neznámo“. Autor ale vytvořil modifikaci tohoto systému, kdy odpovědím přiřazuje hodnoty 0, 2, 3 a celou kategorii pak hodnotí jako aritmetický průměr hodnot položek. Neznámé položky (0) tak silně ovlivňují výsledek směrem k horší kvalitě a nemají charakter neutrálního prvku. Myslím, že použití funkce modus nebo zobrazení histogramu hodnot, by poskytlo lepší souhrnnou informaci o stavu datasetu.

Několik vyhodnocení otázek je dle mě diskutabilních. Například uživatelskou přívětivost rozhraní pro zadávání dat a její dokumentaci autor hodnotí jako nedostatečné, i když to je pro něj neznámá informace. Naopak u otázky ohledně kontroly duplicit a chyb uvádí hodnotu neznámo, i když je na datovém portálu odkazováno na výsledky datového auditu, tudíž tato činnost je nějakým způsobem vykonávána.

Aby autorem zvolený postup měl smysl, bylo by nutné obrátit se na zdravotnický personál, který data do systému zadává a na základě strukturovaného hovoru nebo dotazníkového šetření chybějící hodnoty vyplnit.

**Formální úroveň**

V práci jsem narazil na několik překlepů, což je pochopitelné vhledem k rozsahu textu. Překlep v referenčním datu v kapitole 5 (pravděpodobně mělo být 1. 4. 2021 nikoliv 1. 4. 2020) je nešťastný, protože zásadně mění faktickou informaci.

Spíše estetický problém spatřuji ve zpracování tabulek kritérií (tab. 5.1, 5.3–5.6). Autor v jejich hlavičkách zbytečně uvádí kritéria v plném rozsahu a–j, i když zachycují menší počet sledovaných kritérií.

Čitelnosti práce by prospělo, kdyby autor uvedl přehled používaných zkratek. V textu se vyskytuje relativně velké množství zkratek a některé z nich nejsou ani řádně uvedené. Zkratky jako DQ (data quality, první výskyt v kapitole 1) je možné s trochou bystrosti či znalosti problému odvodit, ale například AIMQ či TQDM (oboje kapitola 2.1, první odstavec) jsou pro čtenáře nic neříkající zkratky názvů technik. První z technik je popsána až v následující kapitole 2.2, druhou jsem musel dohledat v původním zdroji, z kterého autor čerpal.

**Práce s literaturou**

Zmiňovaný odstavec v kapitole 2.1 také dobře demonstruje nestandardní postupy citování a nejednotnost citačních značek v textu. Autor se v první větě odkazuje celým názvem díla bez použití příslušné citační značky (v tomto případě [40]). Ve výčtu technik používá referenční značky v harvardském stylu. V textu se majoritně používají číselné značky, které ale neodpovídají pořadí prvního výskytu.

Použitá literatura je k řešenému tématu relevantní.

**Splnění zadání**

Splněno s výhradou.

V rámci bodu *„(3) Ověřte možnost automatické klasifikace zvolených datových sad z hlediska kvality“* bych očekával hlubší diskuzi překážek a problémů, na které autor narazil. V kapitole 6.1 jen stroze konstatuje, že při snaze použít kvantitativní metody se dostal do slepé uličky.

**Dotazy k práci**

Na jaké problémy jste narazil při snaze použít metriky definované v kapitole 4.1?

Proč jste u modifikované verze dotazníku dle kapitoly 4.3 použil numerické hodnoty 0, 2, 3 pro reprezentaci odpovědí a hodnoty 0, 1, 2, 3 pro popis souhrnných stavů?

Navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni 6. 6. 2021 Ing. Martin Kryl