

## I. IDENTIFIKAČNÍ ÚDAJE

<b>Název práce:</b>	<b>Dolování znalostí z dat v oblasti silniční nehodovosti</b>
<b>Jméno autora:</b>	<b>Bc. lic. Krzysztof Paweł Urbaniec</b>
<b>Typ práce:</b>	diplomová
<b>Fakulta/ústav:</b>	Fakulta dopravní (FD)
<b>Katedra/ústav:</b>	Ústav dopravní telematiky
<b>Oponent práce:</b>	doc. Ing. Milan Šimůnek, Ph.D.
<b>Pracoviště oponenta práce:</b>	Katedra informačního a znalostního inženýrství, FIS VŠE Praha

## II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

<b>Zadání</b>	<b>náročnější</b>
<i>Hodnocení náročnosti zadání závěrečné práce.</i>	
Zadání práce je mezioborové – vyžaduje kombinaci znalostí z domény dopravy (konkrétně silniční nehodovosti), práce z databázemi a geografickými informačními systémy a aplikaci metod dobývání znalostí z databází (DZD).	

<b>Splnění zadání</b>	<b>splněno</b>
<i>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.</i>	
Autor splnil zadání i tři v něm specifikované zásady pro vypracování DP.	

<b>Zvolený postup řešení</b>	<b>vynikající</b>
<i>Posuďte, zda student zvolil správný postup nebo metody řešení.</i>	
Postup i zvolené metody jsou správné. Autor postupoval podle metodiky CRISP-DM a provedl všechny doporučené kroky. Zároveň věnoval dostatečný čas klíčovému kroku <i>Předzpracování dat</i> a zároveň správně aplikoval nutné návraty z pozdějších kroků do předchozích za účelem dosažení ještě lepších výsledků.	

<b>Odborná úroveň</b>	<b>A – výborně</b>
<i>Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.</i>	
Autor prokázal, že je schopen využít znalosti získané studiem a samostatně nastudovat literaturu z jiného oboru (konkrétně DZD). Vše prakticky aplikoval na reálný problém a reálná data, včetně nutnosti zvládnutí dalších dovedností při práci s několika SW nástroji (databázový systém PostgreSQL, systém QGIS a systém LISp-Miner) a jejich vzájemného provázání. Zde se zákonitě vyskytlo několik technických problémů, které autor úspěšně vyřešil.	

<b>Formální a jazyková úroveň, rozsah práce</b>	<b>A – výborně</b>
<i>Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.</i>	
Práce je napsána srozumitelně a čtivě, s minimem překlepů nebo typografických chyb. V celé práci jsem našel pouze jeden překlep („Ttentokrát“ na str. 50), jedno chybějící slovo „jsou“ (ve větě „..., kterými 4ft-Miner, CF-Miner...“ na str. 19), chybějící pravou závorku (str. 17, 2x) a pár opomenutých neslabičných předložek „v“ a „k“ na koncích řádků (např. str. 38). V daném rozsahu textu je to zanedbatelné a s přihlédnutím k tomu, že čeština není patrně autorovým mateřským jazykem, by takto kvalitně napsaný text mohl závidět i leckterý student – rodilý mluvčí. Trochu nestandardní je používání termínu „badatel“, místo obvyklejšího „analytik“. Na str. 18 použité spojení „GUHA-procedury <b>aplikované</b> v systému LM“ by bylo přesnější jako „GUHA-procedury <b>implementované</b> v systému LM“. U SD-procedur je místo „ <b>dvě</b> hypotézy stejného tvaru“ (str. 25) asi srozumitelnější hovořit o „ <b>jedné</b> hypotéze ověřované na <b>dvou</b> podmnožinách dat“. Při výkladu frekvenční analýzy atributu (str. 36) by pomohlo vložení obrázku s histogramem kategorií zmiňovaného atributu.	

**Výběr zdrojů, korektnost citací**

**A – výborně**

*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.*

Výběr zdrojů i jejich citování je v pořádku. Pouze občas došlo k prohození citací na zdroje [7] a [8] (např. popiska obrázku na str. 16), které je však pochopitelné vzhledem k dostupnosti nové publikace [8] až krátce před termínem odevzdání DP.

**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.*

Kromě vlastní analýzy dat o nehodovosti ve Středočeském kraji (první cíl) prokázal autor i vhodnost využití geografických dat v DZD (druhý cíl). Za splněný se dá považovat i třetí cíl – vytvoření návodu pro začínající analytiku, i když je přece jenom asi příliš ambiciózní hovořit o „kompletní příručce pro DZD“.

Závažné věcné připomínky k textu nemám. Pouze upřesňuji, že na str. 42 ve větě „Například z **nižšího** relativního počtu daných nehod...“ má být spíš „**vyššího**“. Metabáze systému LISp-Mineru nemusí být pouze lokálně na disku v MDB souboru, ale může být na jakémkoliv síťovém databázovém serveru (podporujícím ODBC), tedy přístupná i více uživatelům zároveň. Je však pravda, že primárně je předpokládáno využití LISp-Mineru jako jedinouživatelské aplikace. K drobné nepřesnosti došlo na str. 79 v cestě k nástroji pro definici ODBC datových zdrojů – v případě Windows 7 x64 je správná 32-bitová aplikace C:\Windows\SysWOW64\odbcad32.exe, nikoliv C:\Windows\System32\odbcad32.exe.

Při případném pokračování v analýzách se mohou hodit i další techniky předzpracování dat (např. úplné vyřazení příliš četné kategorie; slučování kategorií; vytváření odvozených hodnot přímo v LISp-Mineru; a v neposlední řadě definice doménových znalostí a jejich následné automatické odfiltrování z výsledků). Určitě také doporučuji i autorem zmiňované přidání dalších externích dat – jako jsou data o počasí a velmi pěkný nápad má autor s údaji o společenských událostech.

**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.*

Na práci se mi líbí kombinace metod dobývání znalostí z databází s GISem, který je využit pro získání dalších odvozených údajů. Za velmi důležitou považuji potenciální možnost následného využití alespoň některých objevených závislostí v praxi (viz též doprovodná otázka).

V neposlední řadě bych chtěl zdůraznit i přínos, který mají takovéto studentské práce na rozvoj systému LISp-Miner. Zvláště v případě, že jde o práce mezioborové, přinášející odlišný a „nezávislý“ pohled na problematiku DZD. Již bakalářská práce tohoto autora přinesla inspiraci pro rozšíření systému LISp-Miner o některé funkce na podporu analýzy geodat. Stejně tak je inspirující i diplomová práce (např. zmíněná logaritmická diskretizace).

Doplňující otázky:

- Uvažoval jste o způsobu využití alespoň některých z výsledků nalezených v datech? Je reálná možnost konzultace těchto výsledků s majitelem dat (resp. správci silnic, investory atp.) a dosažení zlepšení?
- V práci zmiňujete možné rozšíření analýz o data z dalších krajů. S tím souvisí i průběžná aktualizace dat o další roky. Zkuste se zamyslet nad možností automatické (resp. polo-automatické) opakované analýzy. Bylo by možné využít dosavadní předzpracování dat a zadání úloh na nová/aktualizovaná data? Nebo by bylo nutné vždy provést nějaké úpravy? Dal by se při prvotní analýze volit způsob předzpracování a zadání úloh tak, aby se co nejvíce usnadnila následná opakovaná analýza?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně**.

Datum: 13. 6. 2015

Podpis: