

Využitie systému WEKA v CAQ

Ing. Renata Janošcová, PhD.
Vysoká škola manažmentu v Trenčíne
rjanoscova@vsm.sk

Ing. Marián Hricko, PhD.
Hella Slovakia Front-Lighting, s.r.o.
Marian.Hricko@hella.com

Seminar on Data mining tools and CAQ (S1)

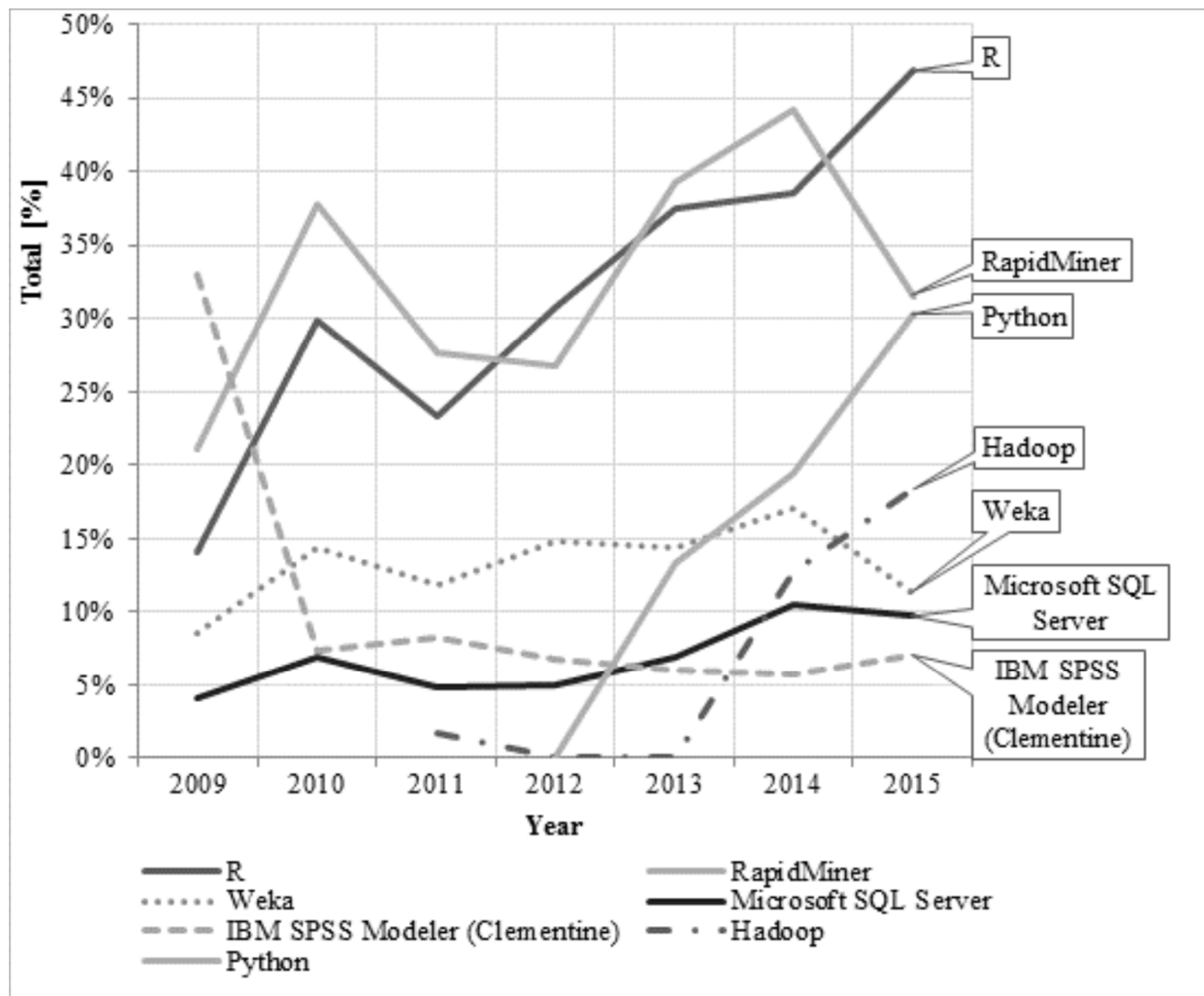
Vysoká škola manažmentu v Trenčíne
International Workshop on Knowledge Management
IWKM'2017

Trenčín 12 - 13. 10. 2017

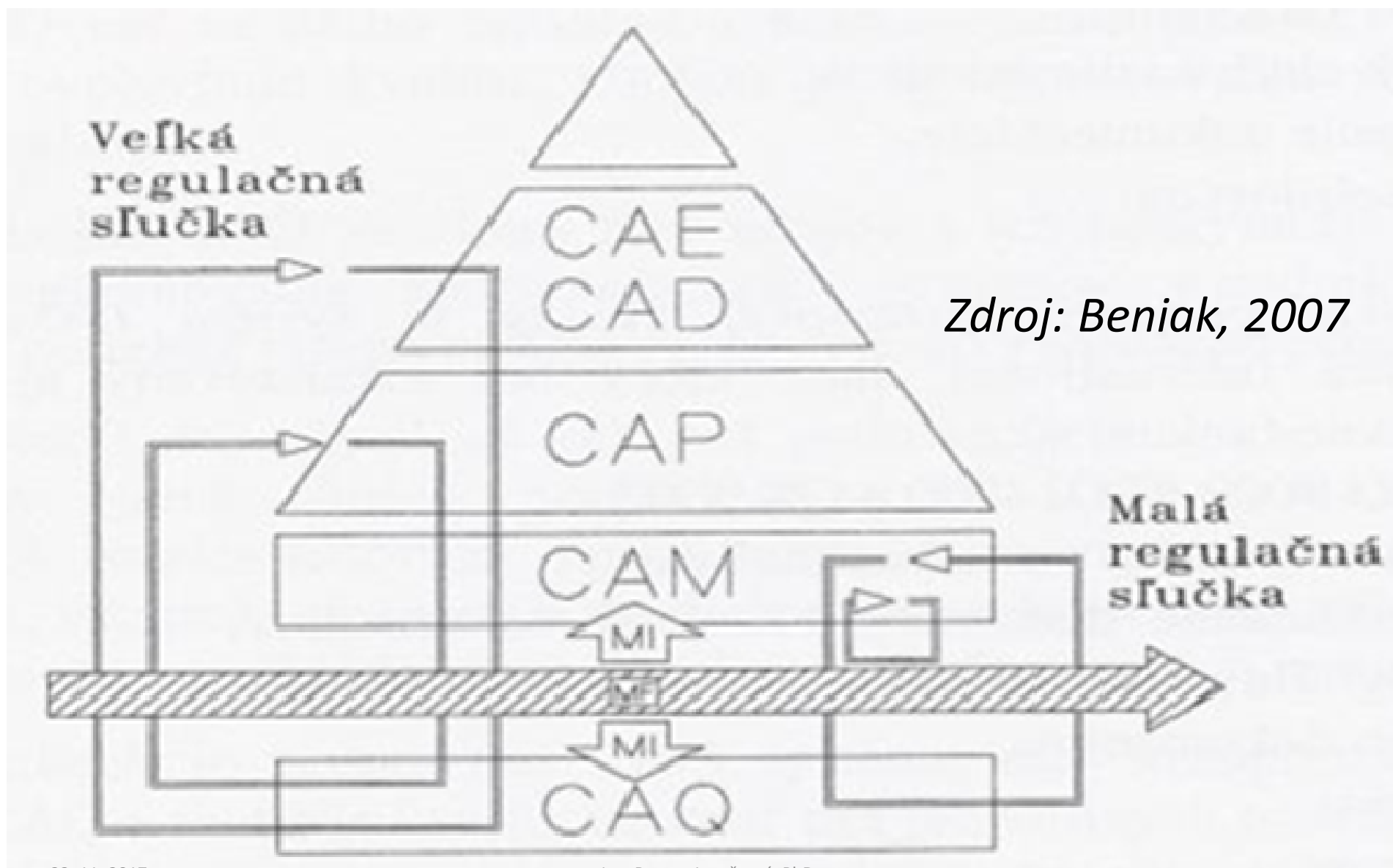
Data mining

- **Stretávame sa s rôznymi výrazmi pre data mining:**
- “dobývání znalostí z databází” (Berka, 2003), “hlbková analýza údajov” (Terek, 2009), “objavovanie znalostí v databázach” (Paralič, 2003), “získavanie poznatkov z databáz” (Janošcová a Jurišová, 2012), “data mining” (Turčínek et al., 2013)
- Zaujímavou definíciou data miningu, s ktorou sa stotožňuje prof. M. Terek (EU Bratislava) je:
- "Data mining je proces objavovania rôznych modelov, zhrnutie a odvodzovanie hodnôt z danej kolekcie dát" (Kantardzic, 2011)
- **Dnešný trh ponúka mnoho nástrojov (softvérových systémov) pre analytiku, Big data a data mining**

Trendy využívania vybraných softvérových nástrojov pre DM (2009 – 2015)



CAQ ako súčasť CAx systémov



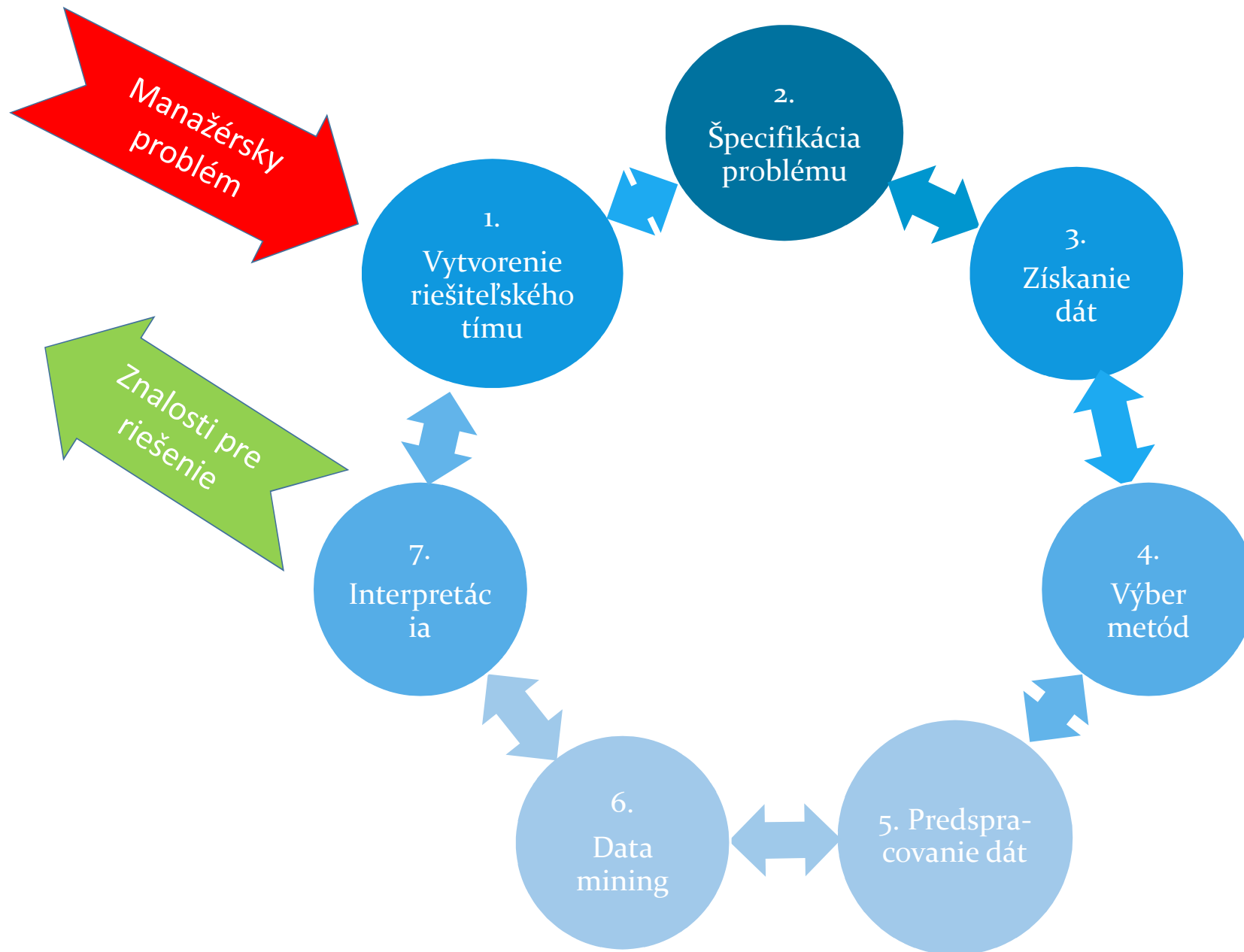
CAQ ako nástroj kvalitného riadenia

- J. Palán (2003) predpovedal, že „**řízení kvality je vlastně kvalitní řízení...**“
 - Môžeme pritom pozorovať, že CAQ sa tu postupne **mení na kvalitné manažérstvo**, a to s podporou IS/IKT
- **CAQ v širšom ponímaní** chápeme ako **IS pre podporu SMK vo všetkých etapách životného cyklu a hodnotového reťazca produktu**
 - Takto chápaný systém **CAQ prechádza naprieč CIM** ako aj **naprieč integrovanými podnikovými informačnými systémami**
- Takéto systémy CAQ riadia procesy tak, aby sa výskyt nekvalitných produktov minimalizoval - **nereagujú až po vzniku nekvalitnej produkcie, ale jej predchádzajú, a dokonca ju predikujú**
 - (ISO 9000)

Data mining a QM

- Harding, J. A. a kol. (2006, s. 970-976) poukazuje na relevantnosť dolovania z dát **v oblasti výroby a kvality produkcie**:
 - Vo výrobných procesoch, operáciách
 - **Pri detekcii porúch, údržbe**
 - **Pri podpore rozhodovania**
 - **Pri zlepšovaní kvality**
 - Pri riadení vzťahov so zákazníkmi
 - Pri plánovaní budúcej výroby, a td'.
- „Techniky dolovania z dát majú tendenciu byť oveľa **pokročilejšie** než jednoduché **SPC** nástroje“ - uvádza SPC expert (Pyzdek, 2013)
- **KDD – Knowledge Discovery in Databases**

KDD – manažérsky pohľad (Anand a kol., 1996)



Špecifikácia problému

- V procese výroby predmetného výrobku ($n \cdot 1000$ ks/zmenu) **vznikajú rôzne nezhody**, odhalené až na výstupnej 100 % vizuálnej kontrole
- Prebieha testovanie rizikového dielca výrobku (v prístrojovom teste **sa skúma podrobne 13 charakteristík** rizikového dielca , testovanie trvá cca týždeň, testuje sa 1 model za mesiac)
- **V prípade nezhody na testovanom dielci/dielcoch** výrobca pristupuje k **nápravným opatreniam** (100 % kontrola a roztriedenie celej mesačnej produkcie predmetného výrobku)
- V rizikovej dávke býva **30 – 50 % nezhodných kusov** výrobkov
- **Zdroj:** HRICKO, M. Využívanie hĺbkovej analýzy údajov pri zlepšovaní kvality produkcie. Dizertačná práca. EU v Bratislave. FHI. Školiteľ: prof. Ing. Milan Terek, PhD. Bratislava: FHI EU, 2014.

Cieľ - predikovania nekvalitnej výroby

- **Analýza historických dát z reálneho výrobného podniku**, pôsobiaceho v oblasti automobilovej výroby s cieľom predikovať nekvalitnú výrobu
- **Úloha:** Pomocou DM v procese KDD zistiť, či výrobca môže na výrobku, označenom ako zhodný, odhaliť použitie rizikovej súčiastky z potencionálne rizikovej dávky
- **Výstup:** Vytvoriť mechanizmus upozornenia, ktorý by v predstihu informoval o tom, že **riziková súčiastka výrobku môže byť z rizikovej výrobnnej dávky** a včasné informovanie materskej firmy o **možnej nezhodnej dávke výrobku**

Získanie dát

- **Bolo potrebné zistiť hraničné hodnoty**, ktoré budú slúžiť na identifikáciu toho, či je **výrobná dávka** použitej rizikovej súčiastky **riziková**, alebo **zhodná**
- Každý parameter má **cieľovú hodnotu, MIN a MAX**
- **Výsledok testovania** výrobkov je binárny: **0 (obsahuje zhodnú súčiastku)**, **1 (obsahuje súčiastku z rizikovej dávky)**
- **Historické dáta** – vzorka zložená len zo zhodných produktov
- **Rozdelenie do dvoch množín** (riziková, vyhovujúca)
- Pomer trénovacej a testovacej množiny: **50 % : 50 %**

Data-WEKA.xlsx - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

AU380

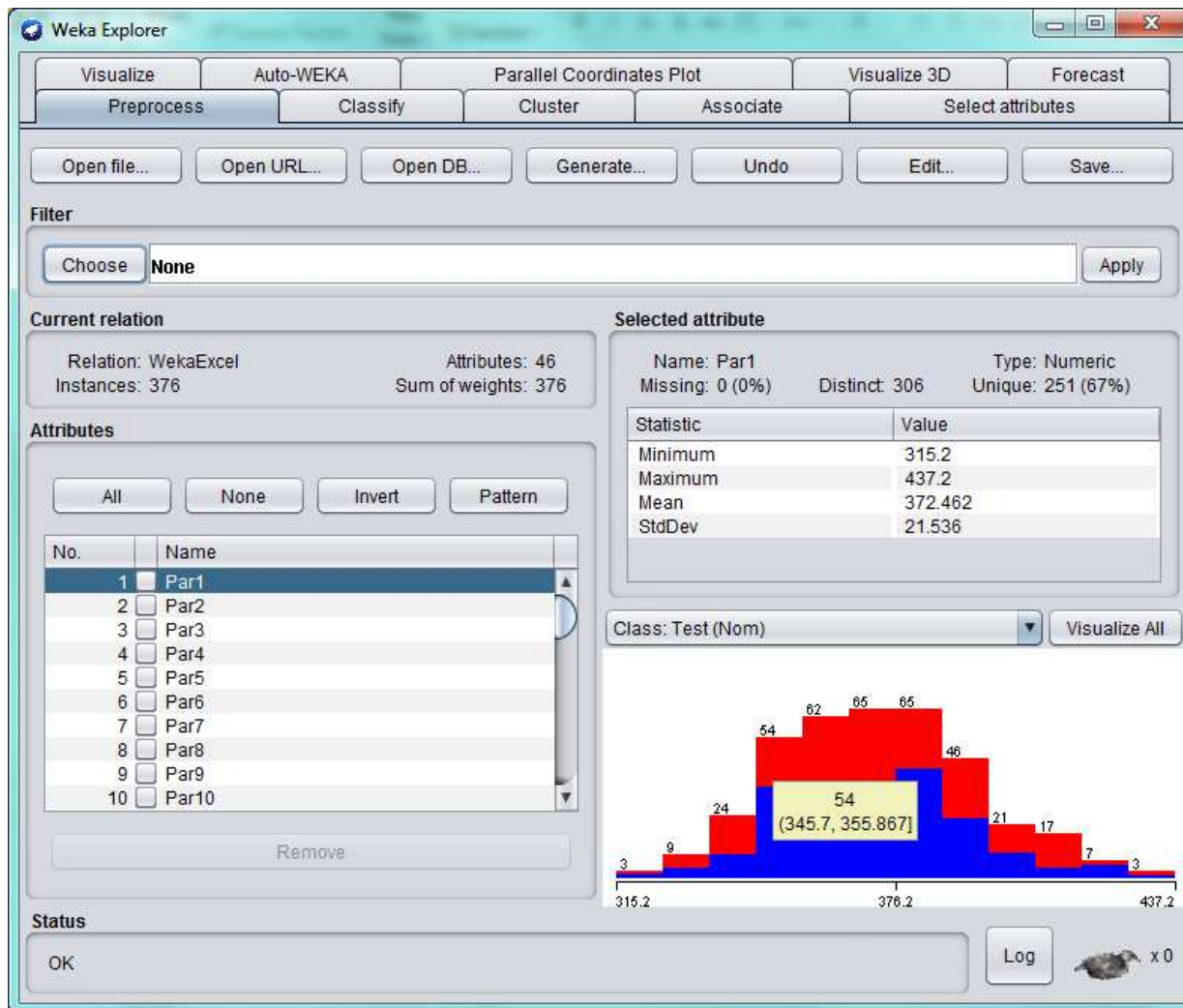
	A	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU
1	Par1	Par26	Par27	Par28	Par29	Par30	Par31	Par32	Par33	Par34	Par35	Par36	Par37	Par38	Par39	Par40	Par41	Par42	Par43	Par44	Par45	Test	
2	437,2	75,78	83,54	0,1	0,8	4,9	11,3	22,2	25,1	46,5	76,9	81,3	113,2	151,3	159,4	218,8	241,5	273,7	342,4	398,7	2,129	OK	
3	351,1	70,11	81,96	0,1	0,9	2,4	12,2	18,9	28,5	38,7	57,1	90,5	113,6	150,0	175,9	204,2	263,5	262,1	372,9	384,5	2,163	OK	
4	382,4	72,27	85,85	0,1	1,0	3,9	8,5	18,0	32,9	43,3	62,5	82,3	116,0	133,6	149,0	193,8	213,6	275,5	314,3	364,7	2,161	OK	
5	396,4	71,21	95,55	0,1	1,2	4,1	9,8	20,7	29,1	45,1	57,0	81,0	115,9	148,2	150,6	220,9	237,9	299,5	340,6	341,1	2,265	RISK	
371	347,0	69,06	86,80	0,1	1,9	4,5	10,1	17,6	33,0	43,1	66,2	88,2	111,8	146,4	140,2	194,1	259,7	287,8	390,6	395,9	2,417	OK	
372	396,9	70,85	78,82	0,1	1,0	2,0	10,4	17,1	29,8	50,7	71,9	75,2	104,0	129,2	159,0	188,2	239,7	278,4	344,2	338,8	2,054	RISK	
373	338,3	76,50	74,51	0,1	1,4	4,9	11,6	15,7	37,2	44,6	65,2	92,6	105,9	135,9	167,6	216,8	242,3	283,0	344,6	342,3	2,219	RISK	
374	371,1	70,71	88,61	0,1	0,8	2,9	12,7	20,0	30,5	41,0	68,6	74,3	104,7	131,5	159,0	210,2	240,8	288,0	355,2	383,0	2,232	OK	
375	402,7	73,22	83,74	0,1	1,6	3,1	9,0	16,6	29,3	46,4	62,3	79,1	107,9	125,4	171,3	210,3	236,3	249,3	288,8	354,2	2,239	RISK	
376	409,8	72,86	76,63	0,1	0,8	3,9	9,0	21,4	33,1	41,9	65,2	94,0	114,3	124,5	181,1	222,8	270,4	284,7	351,5	368,8	2,078	RISK	
377	397,0	76,78	80,31	0,1	0,7	4,1	8,4	15,7	29,8	47,9	65,8	72,4	113,8	132,5	167,6	199,3	268,7	277,1	344,9	368,8	2,042	RISK	
378																							
379																							
380																							

Data meranie (TWKM)

READY 100 %

- Skúmaná vzorka má **45 atribútov** (parametrov)
- Počet prípadov pre DM – **376**
- Výsledok testovania – v závislom atribúte **Test**

Popisné štatistiky (preskúmanie dát)



- Minimum
- Maximum,
- Priemer,
- Smerodajná odchýlka
- Medián
- Počet
- Počet chýbajúcich hodnôt
- Typ
- Počet unikátnych hodnôt
- ...

Výber metód - WEKA

- 1999, New Zealand, „free software“
- Základné funkcionality
 - **Príprava dát** (podpora rôznych formátov, filtre, ...)
 - **Klasifikácia** (rozhodovacie/klasifikačné stromy)
- **Klastering**
- **Asociácie**
- **Výber atribútov**
- **Vizualizácia dát** (závislosti, NS, BS, 2D, 3D,...)
- **Forecast**

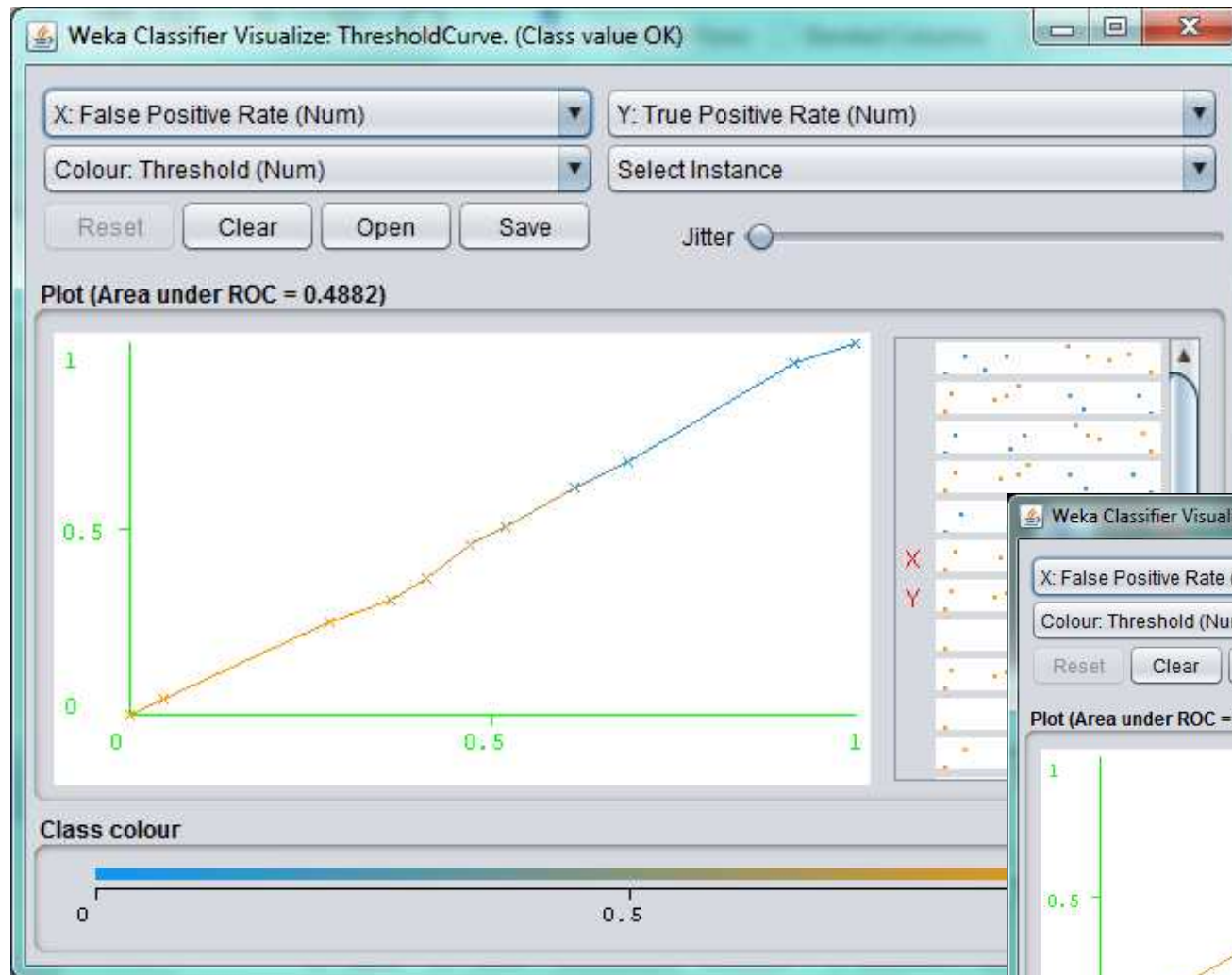


Data mining

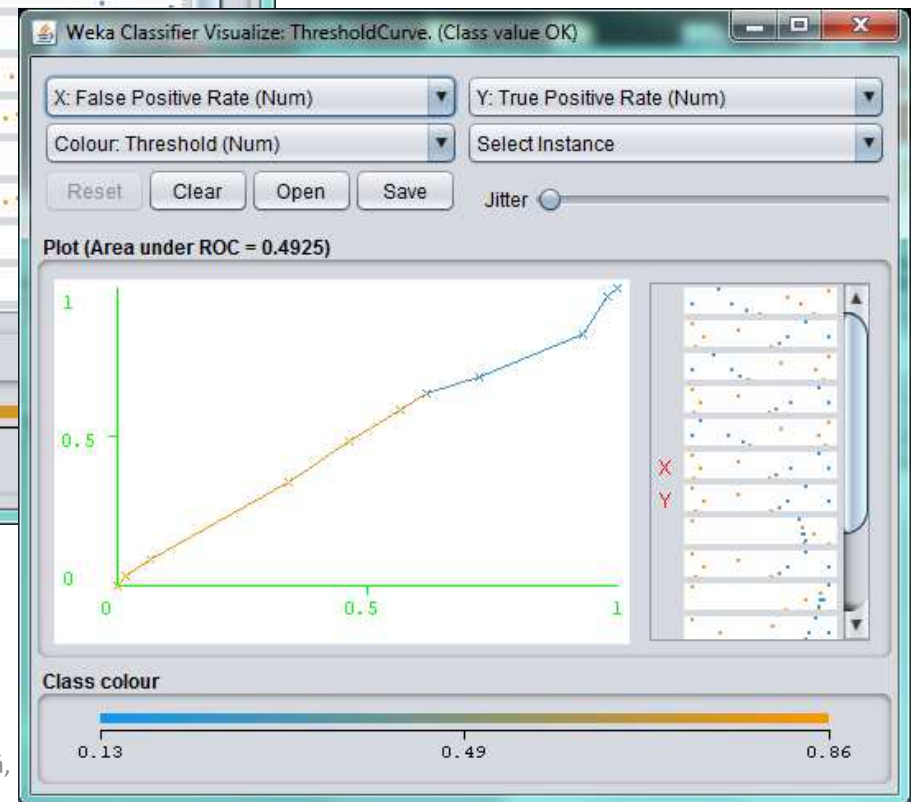
- Vytvoríme 3 modely rozhodovacích stromov (RS):
- **CART** (*classification and regression tree*) – pri výbere atribútu vetvenia RS využíva hodnotu **Gini indexu** (min) (**SimpleCART**)
- **C4.5** – na vetvenie využíva hodnotu **informačného zisku** (*information gain*) – (**J48**)
- **User classifier**

- **Kritériá:**
- premenné pri vetvení nebudú viac krát použité
- maximálny počet vetvení na jednej úrovni - 2
- maximálna hĺbka stromu (teda počet úrovní mimo koreňového uzla - 10)
- vytvoriť najmenší rozhodovací strom s najlepším hodnotením

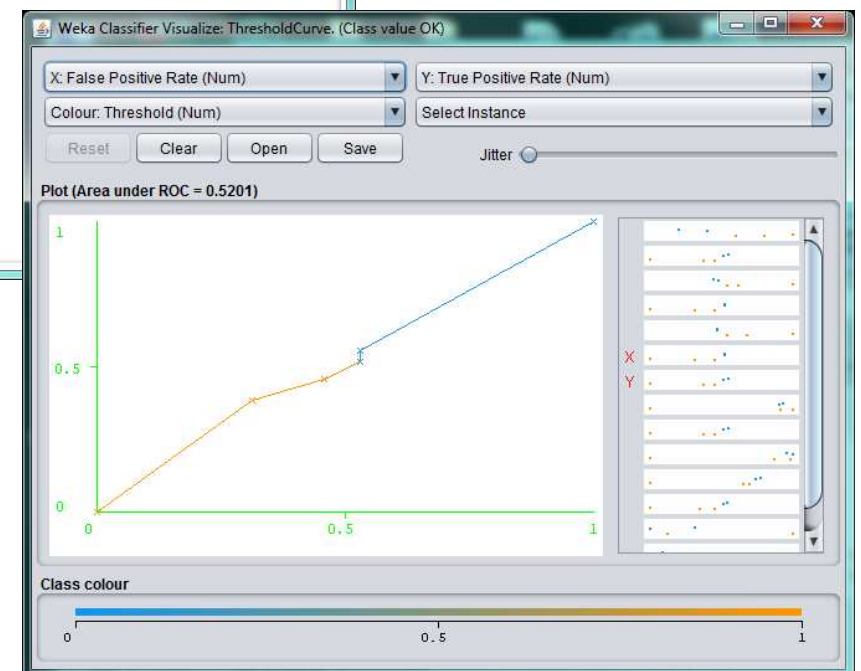
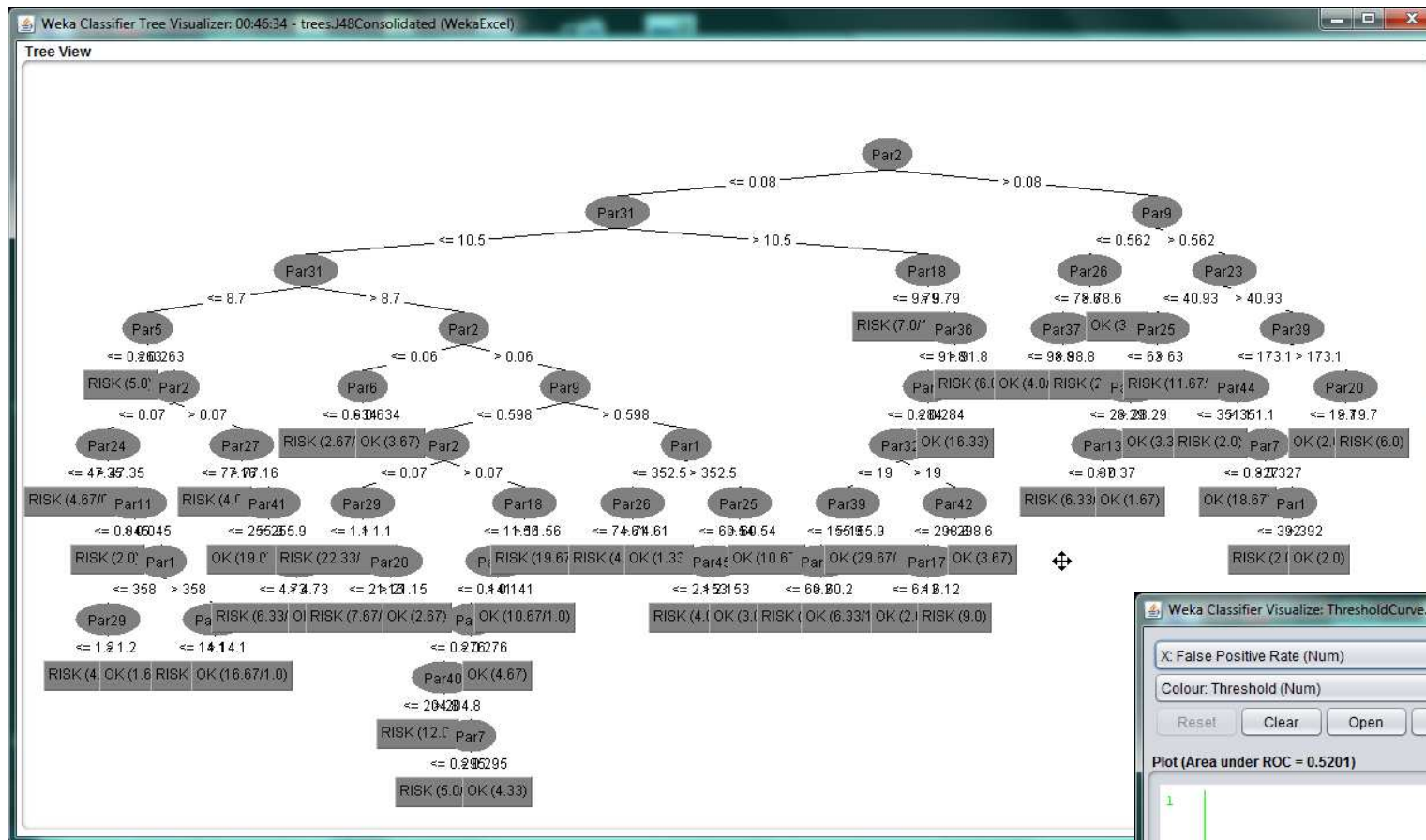
CART Decision Tree



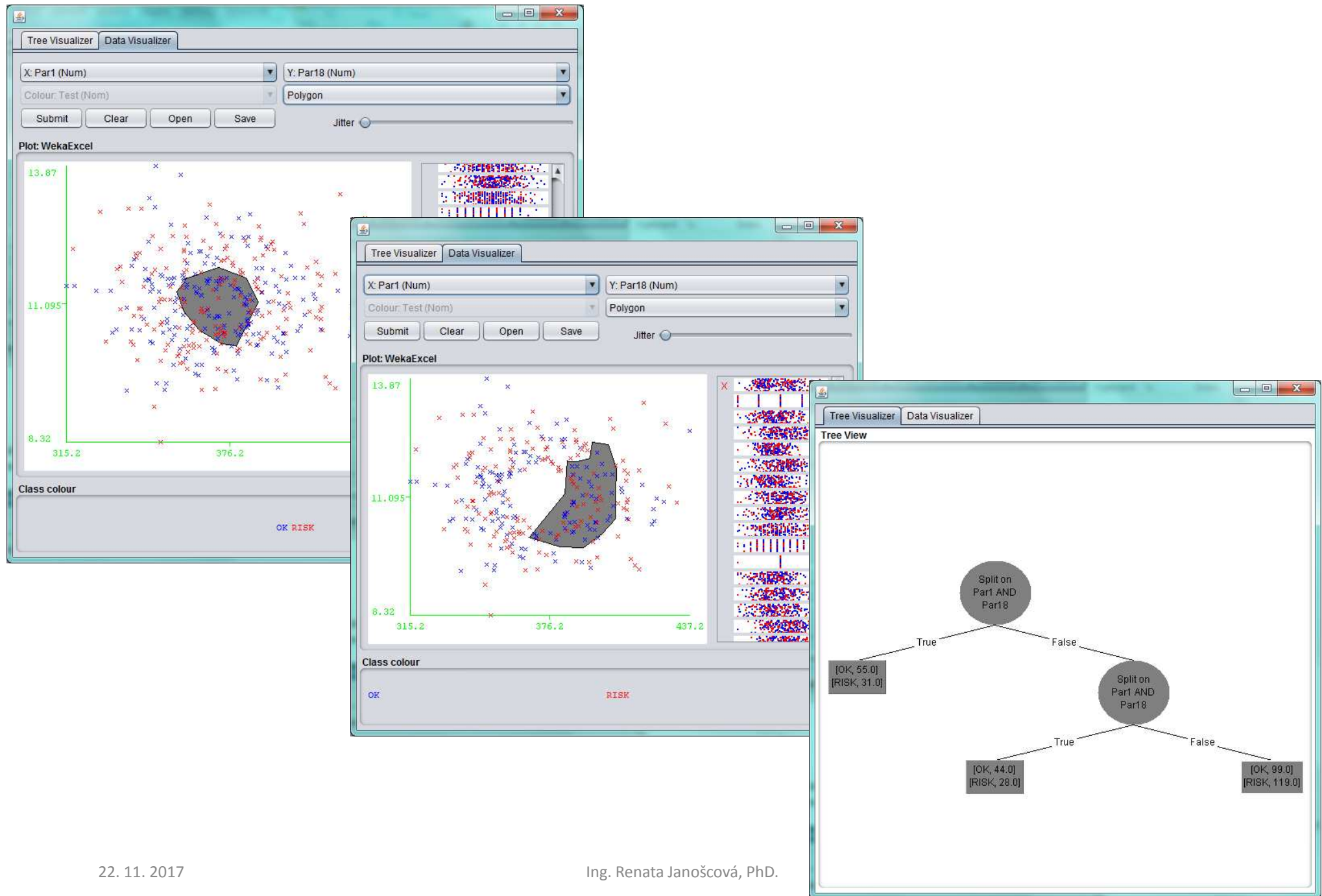
66 % : 44 %



J48 Consolidated pruned tree



User Classifier



Porovnanie modelov

Algoritmus (trénovanie : validácia)	Number of Leaf Nodes:	Size of the Tree:	Correctly Classified Instances	Root mean squared error
CART Decision Tree (50 % : 50 %)	23	45	49.4681 %	0.6453
J48 Consolidated pruned tree (50 % : 50 %)	49	97	49.4681 %	0.6949
User Classifier (50 % : 50 %)	2	2	44.1489 %	0.5007
CART Decision Tree (66 % : 34 %)	23	45	51.5625 %	0.5687

Interpretácia a implementácia

- Prevod „najlepšieho“ rozhodovacieho stromu na pravidlá

Súbor podmienok č. 1		
Par4	$\geq$	0,265
Par21	$<$	29,69
Par1	$<$	398,5
Par10	$\geq$	0,147

Počet súborov podmienok = počtu listových uzlov

- Model bol následne zapracovaný do šablóny meracieho protokolu pomocou logických funkcií (**IF - THEN**)

Zdroje

- BERKA, P. 2003. *Dobývání znalostí z databází*. Praha : Academia
- HRICKO, M. 2014. *Využívanie hĺbkovej analýzy údajov pri zlepšovaní kvality produkcie*. Dizertačná práca. Školiteľ: prof. Ing. Milan Terek, PhD. Bratislava: FHI EU, 101s.
- KELEMEN, J. a kol. 2007. Pozvanie do znalostnej spoločnosti.
- JANOŠCOVÁ, R. & JURIŠOVÁ, E. 2012. *Databázy - zdroj pre znalostný manažment*. Trenčín : Vysoká škola manažmentu.
- KANTARDZIC, M. 2011. *Data Mining. Concepts, Models, Methods, and Algorithms*. s. 1. J. Wiley and Sons.
- PARALIČ, J. 2003. *Objavovanie znalostí v databázach*. Košice : Elfa.
- TEREK, M. 2009. Možnosti využívania hĺbkovej analýzy údajov v ekonomike. *Ekonomické rozhľady: vedecký časopis Ekonomickej univerzity v Bratislave*, 38(1), pp. 18-27.
- TURČÍNEK, P., MOTYČKA, A. & KOSOVÁ, A. 2013. *Economic Data Classification by Means of Machine Learning Methods*. Paris, Recent Advances in Information Science, Proceedings of the 4th European Conference of Computer Scienc, pp. 1-6.

Ďakujem za pozornosť