Analýza dat z PISA 2009

Dobývání znalostí

Tomáš Hřebejk
Otakar Trunda

PISA

- Testování studentů základních a středních škol
  - Reading, Math, Science + vyplnění dotazníku
  - Obrovské množství záznamů
- Stovky atributů, často silně korelované
  - Kvalitativní i kvantitativní
- Problémy reálných dat
  - Chybějící hodnoty, nevhodný formát, …
- Váhy u jednotlivých záznamů
- Použili jsme pouze data týkající se ČR

Hledané znalosti a použité metody

- Hledání podobností v datech
  - Klastrování pomocí k-means
- Predikce výsledku test Reading
  - Pomocí lineární regrese
  - Pomocí rozhodovacího stromu
- Další pokusy
  - Asociační pravidla
  - SVM, Logistická regrese
- Použitý nástroj: Rapidminer

Získané výsledky

- Shlukování
- Lineární regrese
- Rozhodovací strom

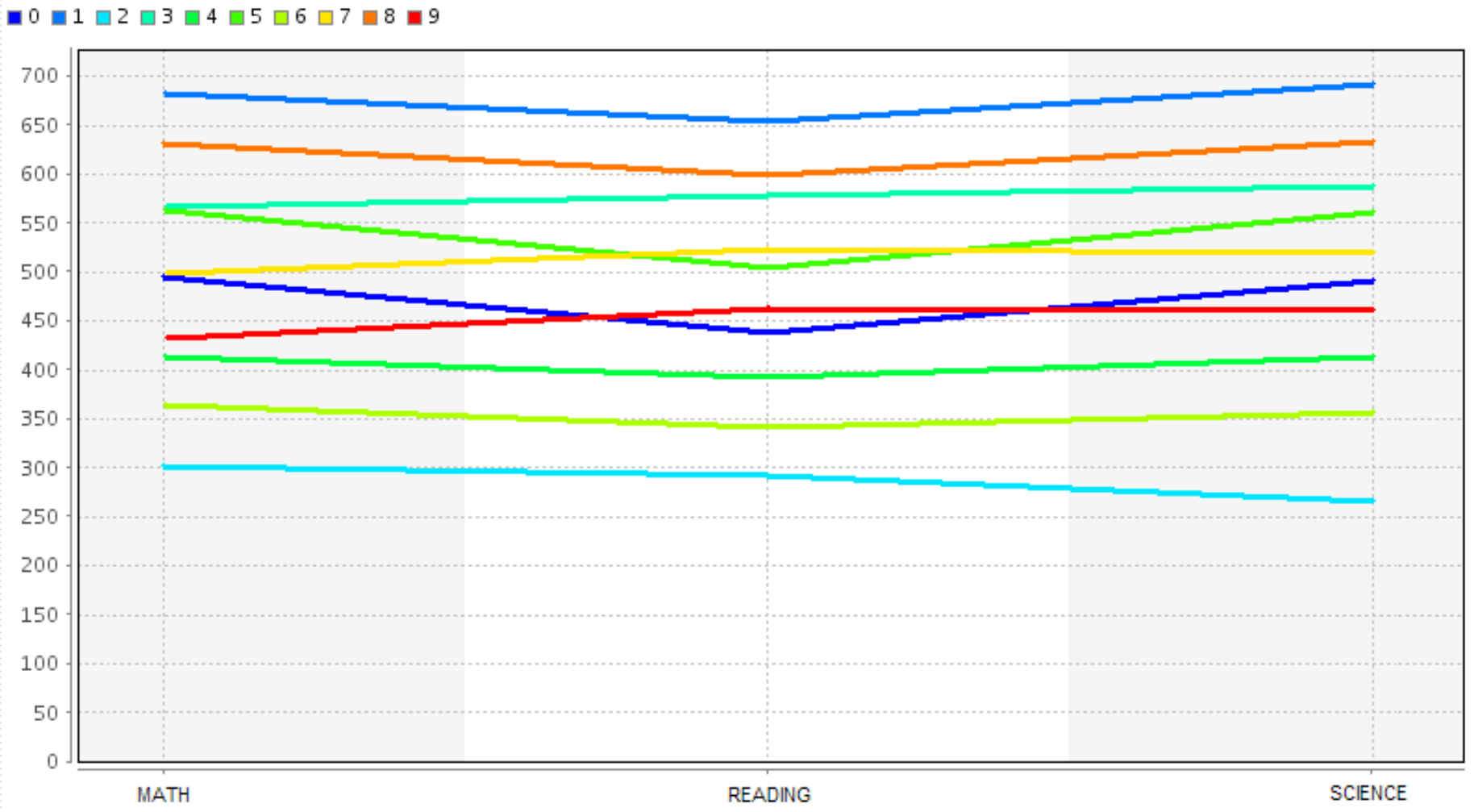
Shluková analýza

- Cíl:

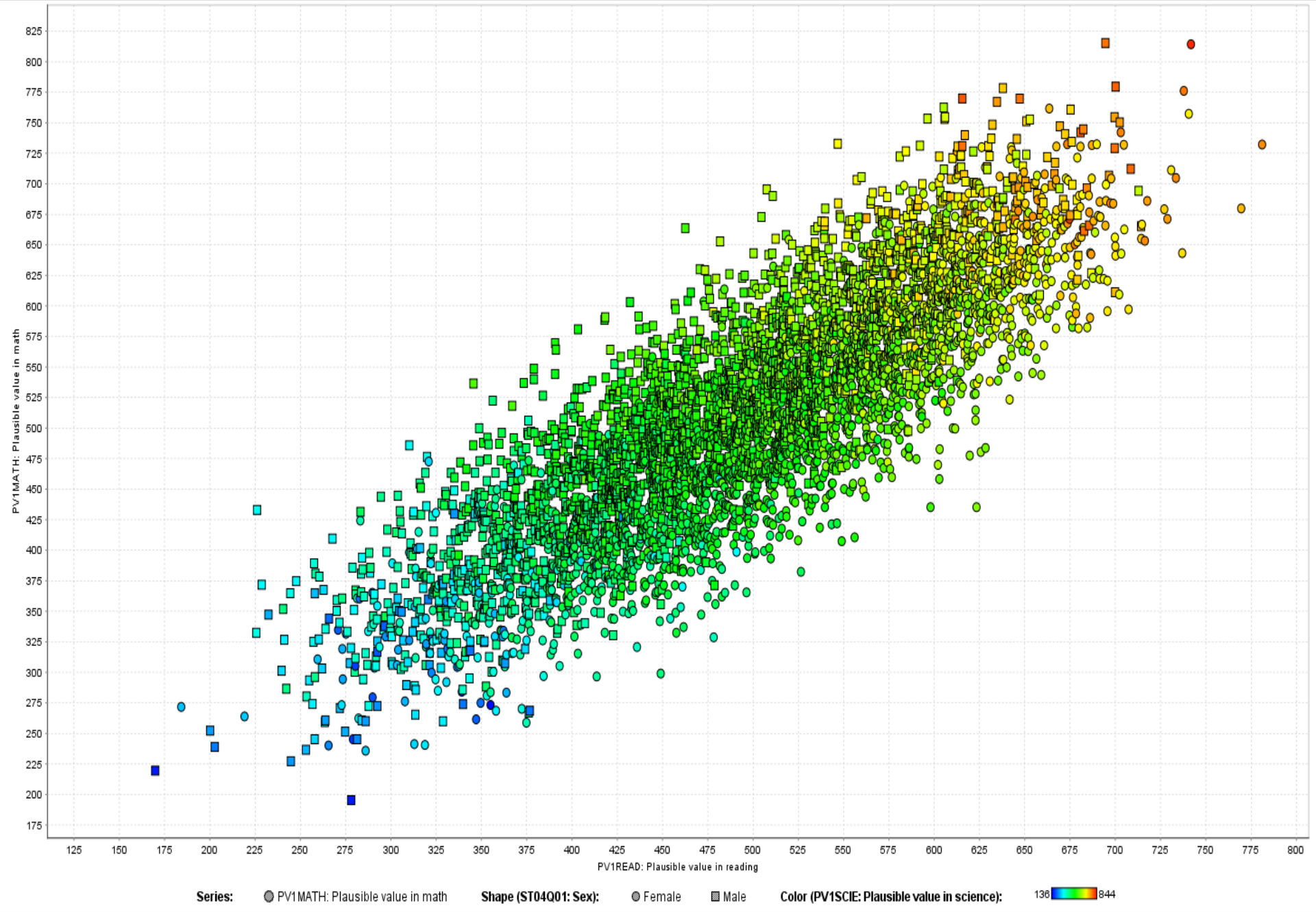
Najít skupiny podobných záznamů v datech

- Pouze na třech atributech: výsledky testů Reading, Math, Science
- Předem daný počet shluků - 10
- Algoritmus K-means, doba běhu: minuta
- Hypotéza: Budou existovat skupiny studentů vynikajících v jednom oboru ale horších v ostatních oborech

Shluková analýza - výsledky



Shluková analýza - pohled na data


PV1MATH: Plausible value in math
825
800
775
750
725
700
675
650
625
600
575
550
525
500
475
450
425
400
375
350
325
300
275
250
225
200
175
125
150
175
200
225
250
275
300
325
350
375
400
425
450
475
500
525
550
575
600
625
650
675
700
725
750
775
800
PV1READ: Plausible value in reading
Series:
PV1MATH: Plausible value in math
Shape (ST04Q01: Sex):
Female
Male
Color (PV1SCIE: Plausible value in science):
136
844

Získané výsledky

- Shlukování
- Lineární regrese
- Rozhodovací strom

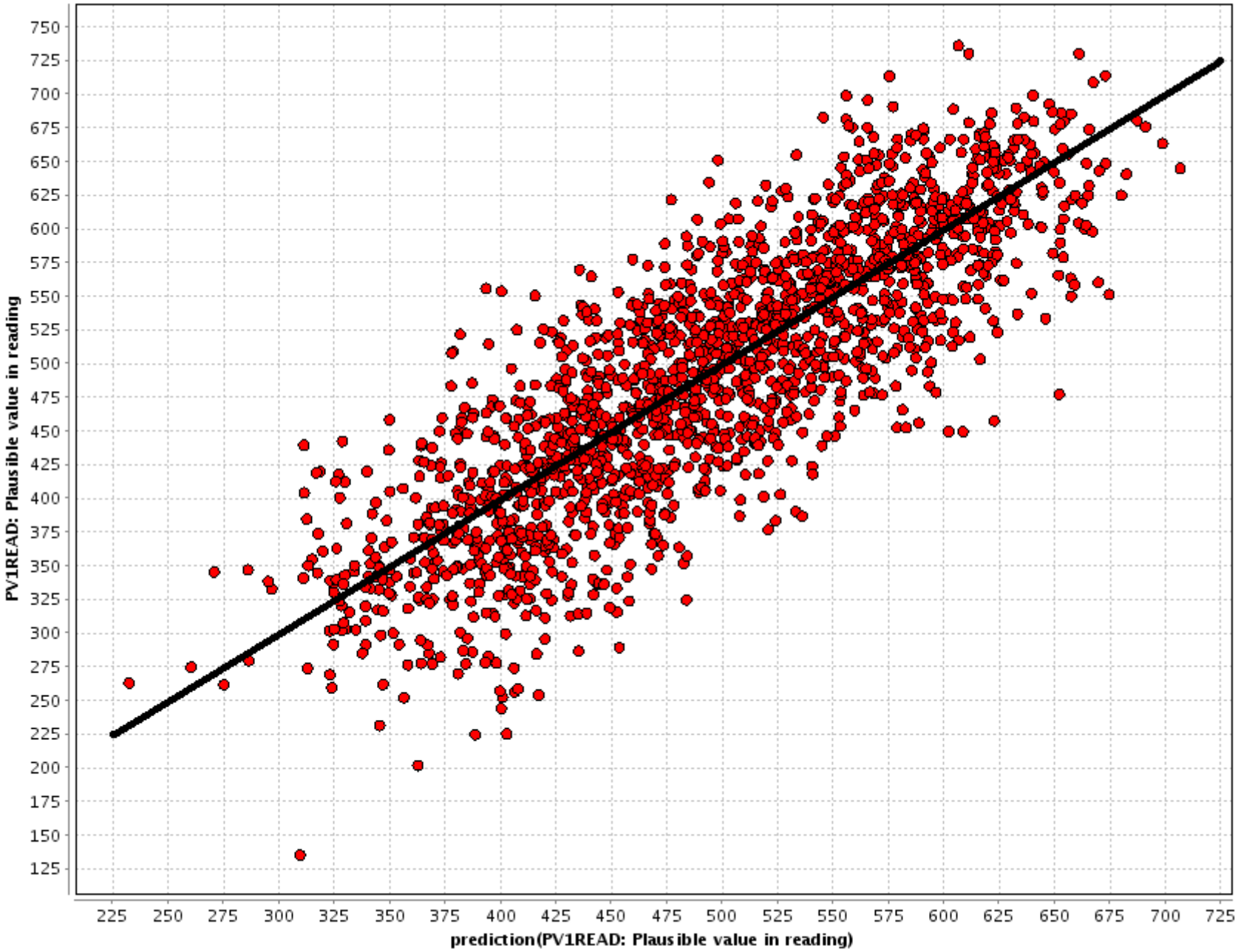
Lineární regrese

- Cíl:

Predikce úspěšnosti v testu Reading na základě odpovědí v dotazníku

- Vícehodnotové atributy nahrazeny jednohodnotovými - 0,1
- Normalizace číselných atributů na <0,1>
- Hledání význačných atributů pomocí míry korelace
- „Prořezání“ koeficientů s malými hodnotami pomocí t-testu
- Čas běhu: několik hodin

Úspěšnost predikce:


PV1READ: Plausible value in reading
prediction(PV1READ: Plausible value in reading)
750
725
700
675
650
625
600
575
550
525
500
475
450
425
400
375
350
325
300
275
250
225
200
175
150
125
225
250
275
300
325
350
375
400
425
450
475
500
525
550
575
600
625
650
675
700
725

Hodnoty některých koeficientů

How many books at home = 0-10	-15.80730308937991
How many books at home = 11-25	-7.639428695993201
How many books at home = 201-500	6.738374526960074
How many books at home = More than 500	0.0
Possessions poetry = No	-13.208824859237467
Possessions <technical reference books> = Yes	21.820230887022284
Read Attitude - Favourite hobbies = Strongly agree	19.013428102778434
Read Attitude - Favourite hobbies = Agree	10.640325880641866
Sex = Female	5.131042497544162

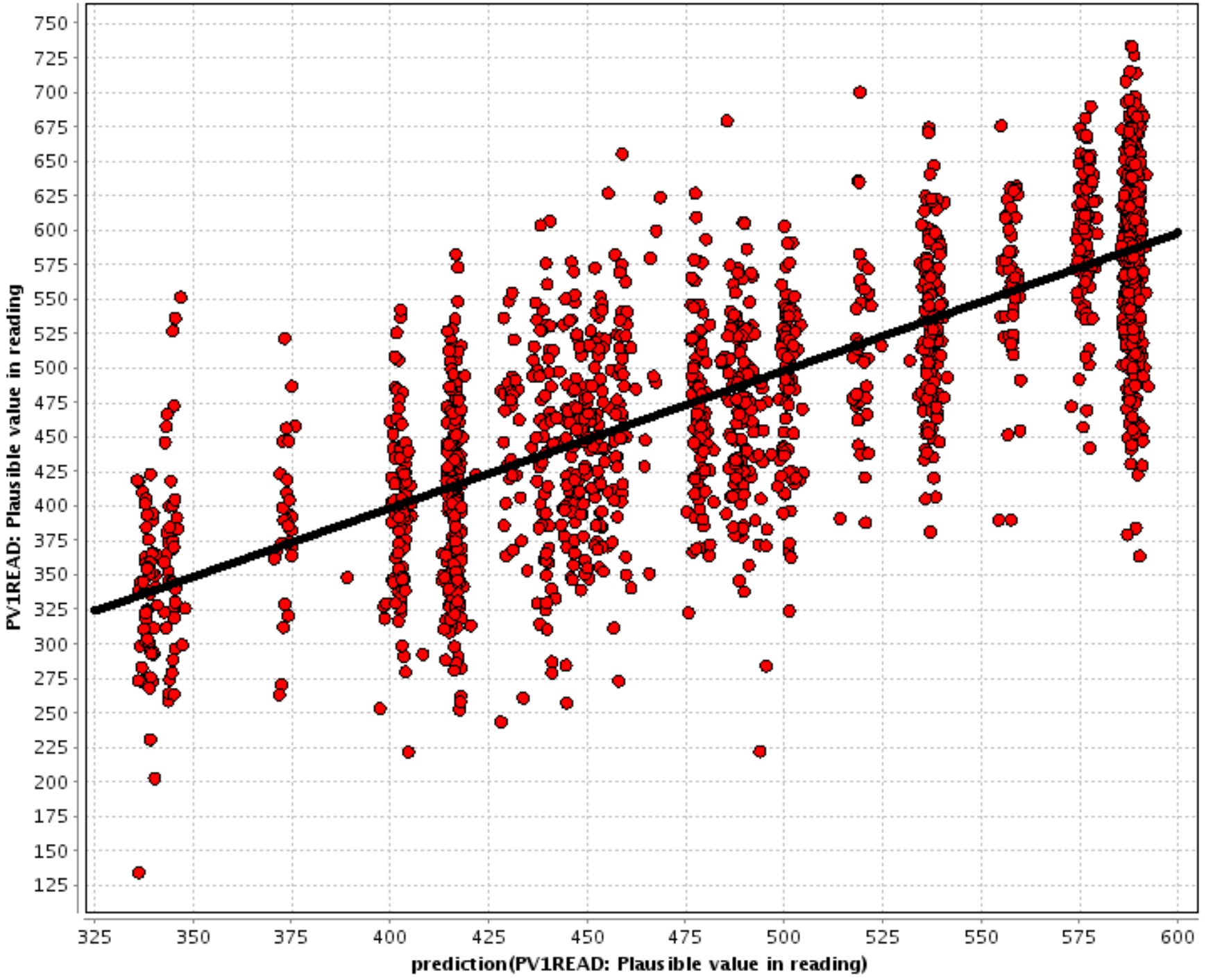
Získané výsledky

- Shlukování
- Lineární regrese
- Rozhodovací strom

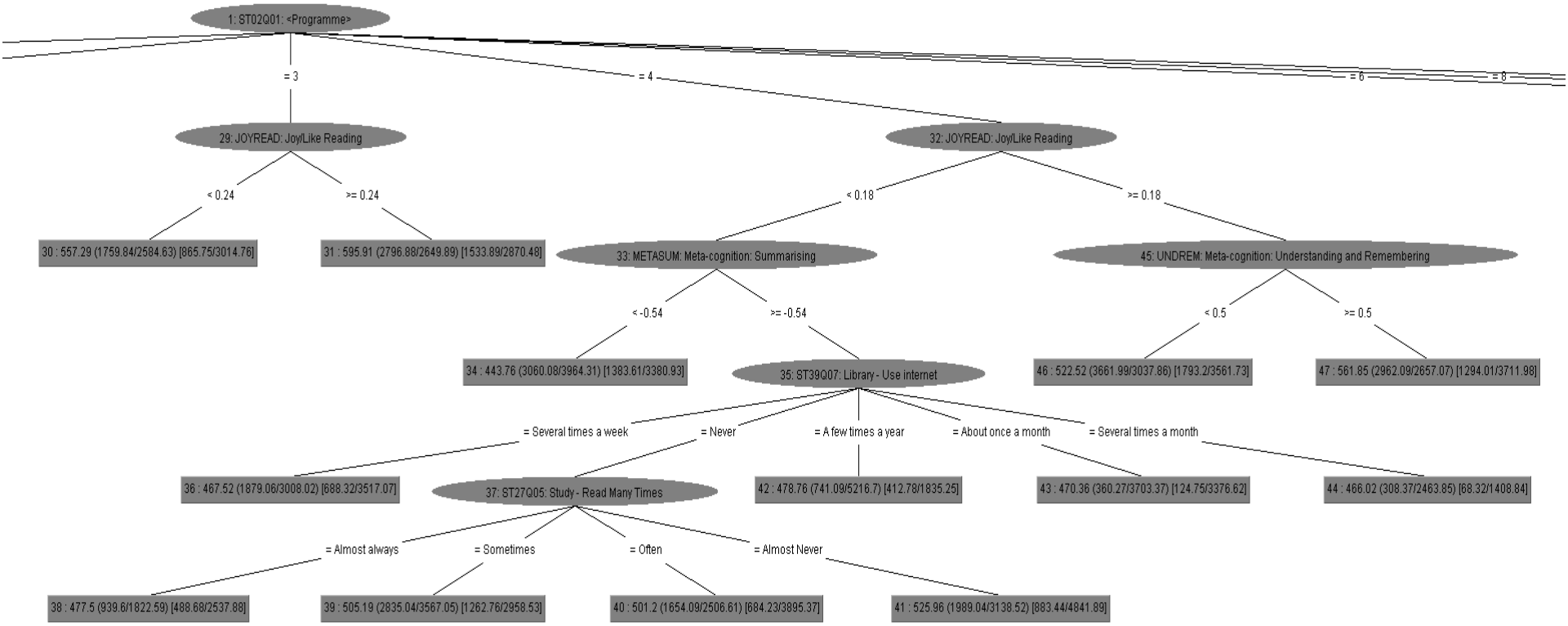
Rozhodovací strom

- Predikce výsledku testu Reading
- Algoritmus typu CART
  - Predikce hodnoty číselné veličiny
  - Zdrojové atributy kvalitativní i kvantitativní
- Data rozdělena na trénovací a testovací
  - 70% - 30%
  - Měření chyby: součet čtverců odchylek
- V listu predikuji jedinou hodnotu
- Doba běhu: sekundy

Úspěšnost predikce:


PV1READ: Plausible value in reading
750
725
700
675
650
625
600
575
550
525
500
475
450
425
400
375
350
325
300
275
250
225
200
175
150
125
325
350
375
400
425
450
475
500
525
550
575
600
prediction(PV1READ: Plausible value in reading)

Rozhodovací strom - výsledek



Srovnání výsledků

Clasifier	root mean squared error	absolute error	relative error	correlation
Zero clasifier	94.042	77.054	17.89%	0
W-REPTree	63.177	49.798	11.50%	0.742
Linear regression	55.568	43.82	10.00%	0.807

- Regrese dosahuje vyšší úspěšnosti než rozhodovací strom
- Ale také mnohonásobně vyšší čas pro vytvoření modelu

Další pokusy

- Logistická regrese pro predikci pohlaví
  - Úspěšnost přes 80%
- SVM pro predikci pohlaví a výsledků testu
  - Horší výsledky než lineární regrese
- Hledání asociačních pravidel
  - Úspěšná pravidla většinou nezajímavá (očekávaná)
  - *(At_School_Use_Email = Almost every day)* ➔ *(At_Home_Use_Email = Almost every day)*
  - *(Read_Attitude_Favourite_hobbies = Strongly agree)* ➔ *(Read_Attitude_Waste_of_time = Strongly disagree)*
  - Až na výjimky: *(At_Home_One_Player_Games = Almost every day)* ➔ *(Sex = Male)*

Shrnutí výsledků

- Zajímavá data
  - Mnoho skrytých „znalostí“
  - Velký potenciál k dalšímu zkoumání
- Zkušenosti s programem Rapidminer:
  - Jednoduché intuitivní použití
  - Mnoho implementovaných metod
  - Problémy s větším množstvím dat