Educational Data Mining: Preliminary results at University of Porto

Pedro Strecht

João Mendes Moreira

Carlos Soares

EUNIS Business Intelligence Conference

March 6th-7th 2014

Paris

Summary

- Data Analysis in Education
- … at the University of Porto
- An illustrative example of an EDM task
- Conclusions and Future work

Data Analysis in Education

- For a few decades higher education institutions manage their data using University Information Systems (UIS)

- The growing adoption of UIS allowed research to move towards automatic knowledge discovery from academic databases

- Over the past 10 years there has been an increase on research using data mining techniques to discover phenomena in the data

- An example of application of data mining is:
    - Predicting the success or failure of student enrolled in a course
    - Learning the reasons behind it

University of Porto

U.PORTO

- Founded in 1911
- 14 faculties, 1 business school
- ~700 study programs
- ~32 000 students, ~2 000 teachers and researchers, ~1 800 administrative staff
- University Information Systems began being developed in-house and explored since 1992
- The SIGARRA system had a major improvement in 2012 which prompts the University to improve their processes using BI and DM

Data Analysis in Education

Educational big data

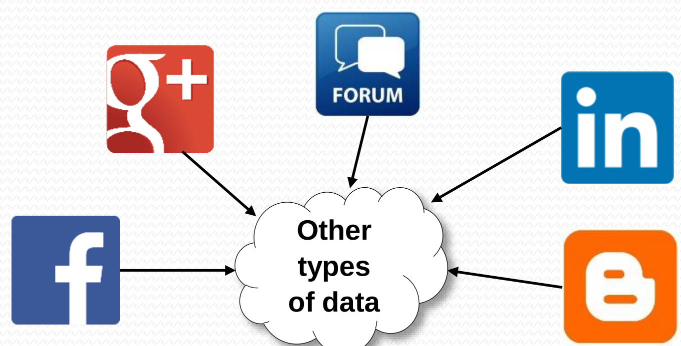
- Academic information

Administrative data

- Teaching and learning environments

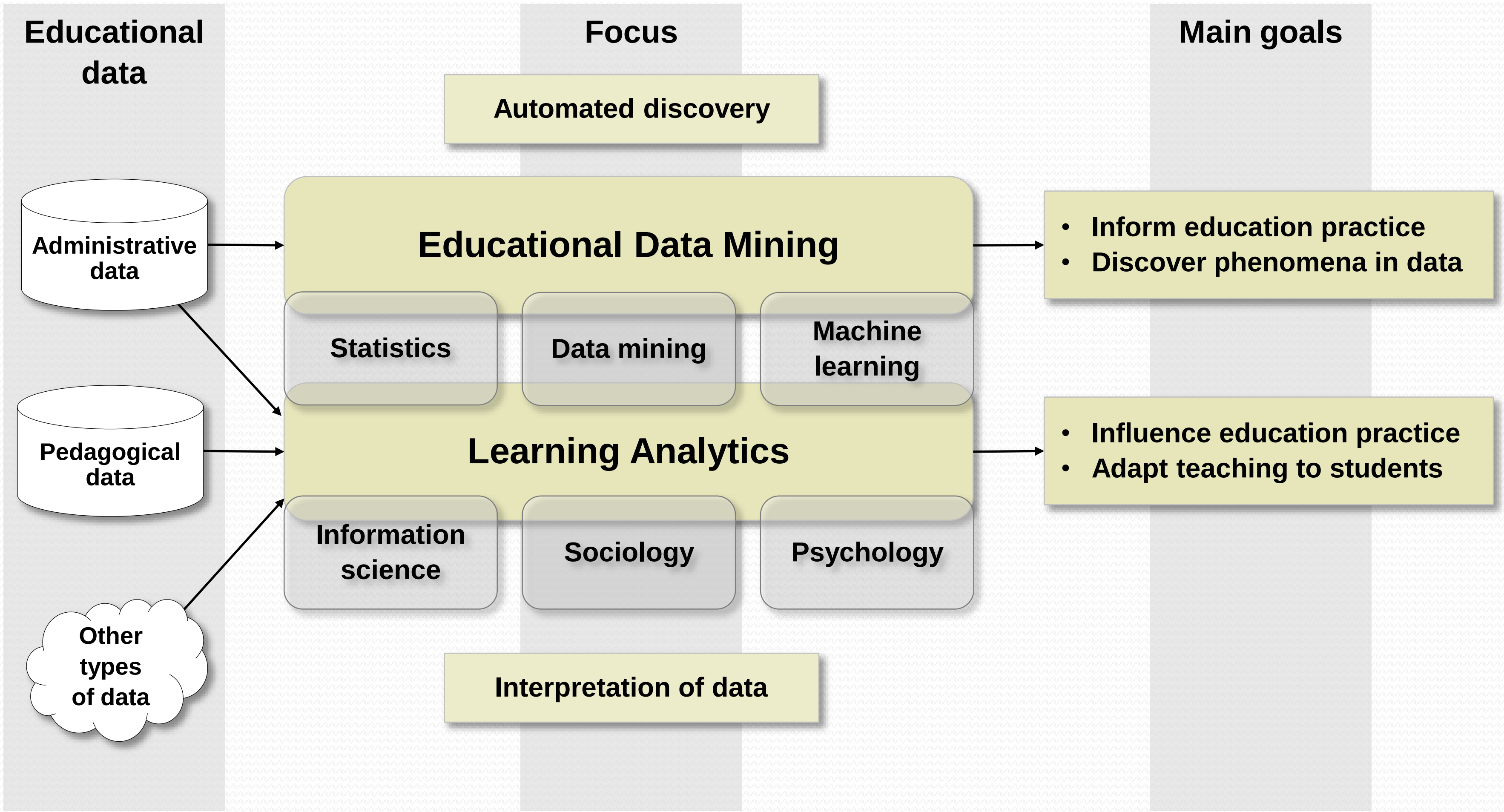
Pedagogical data

- Others sources

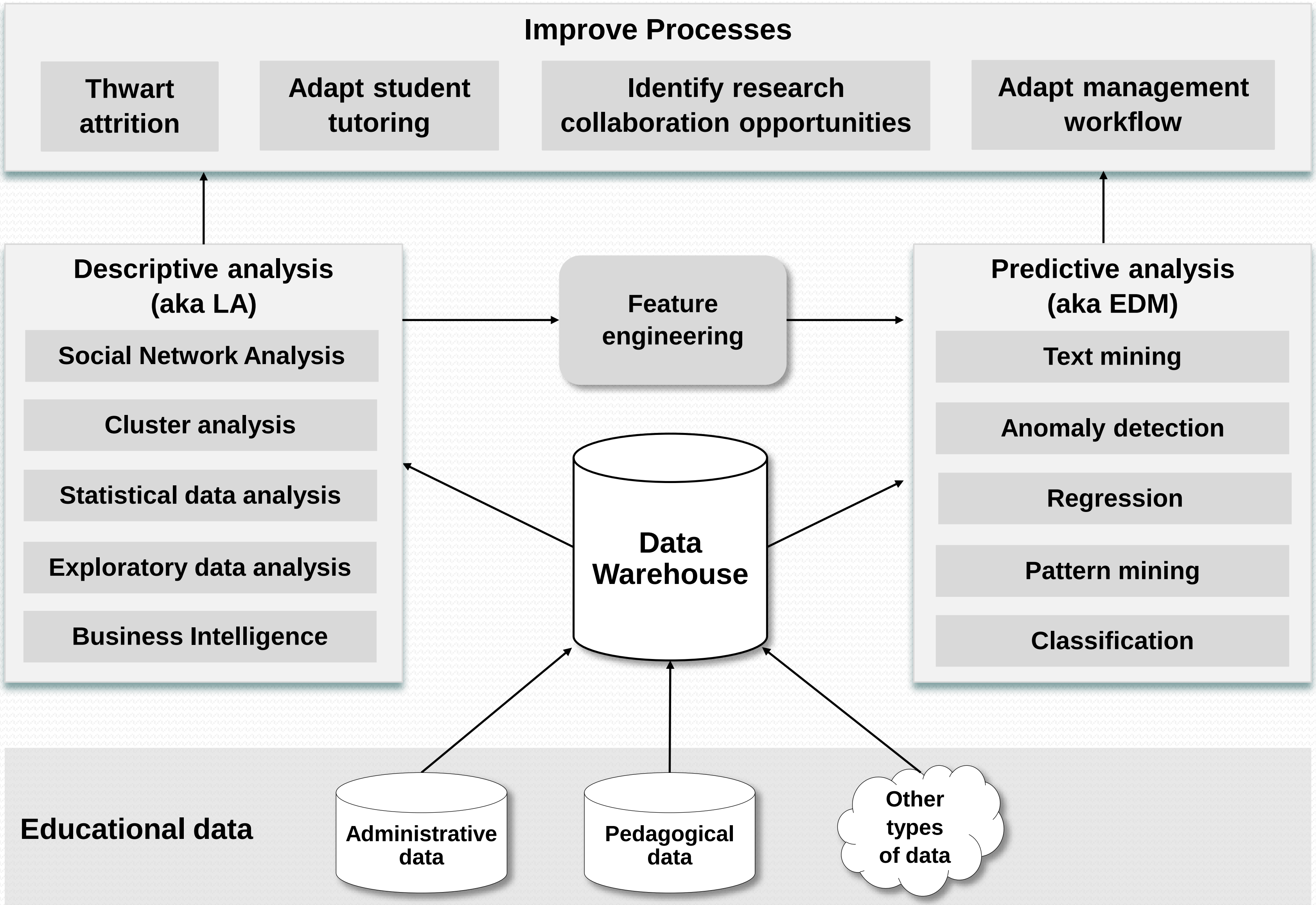


Data Analysis in Education

Overview of research

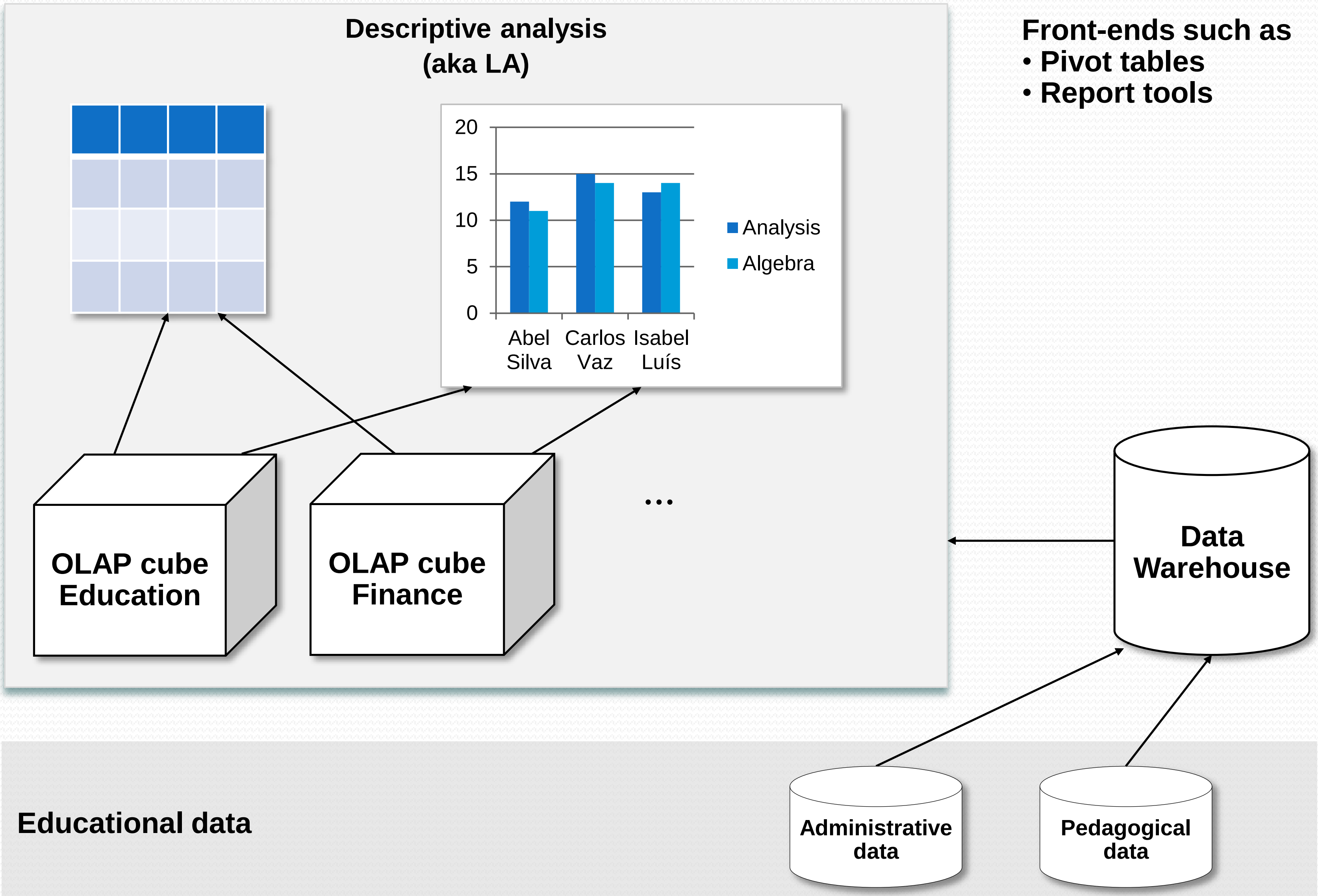


Educational DM & Learning Analytics at U.Porto: general perspective

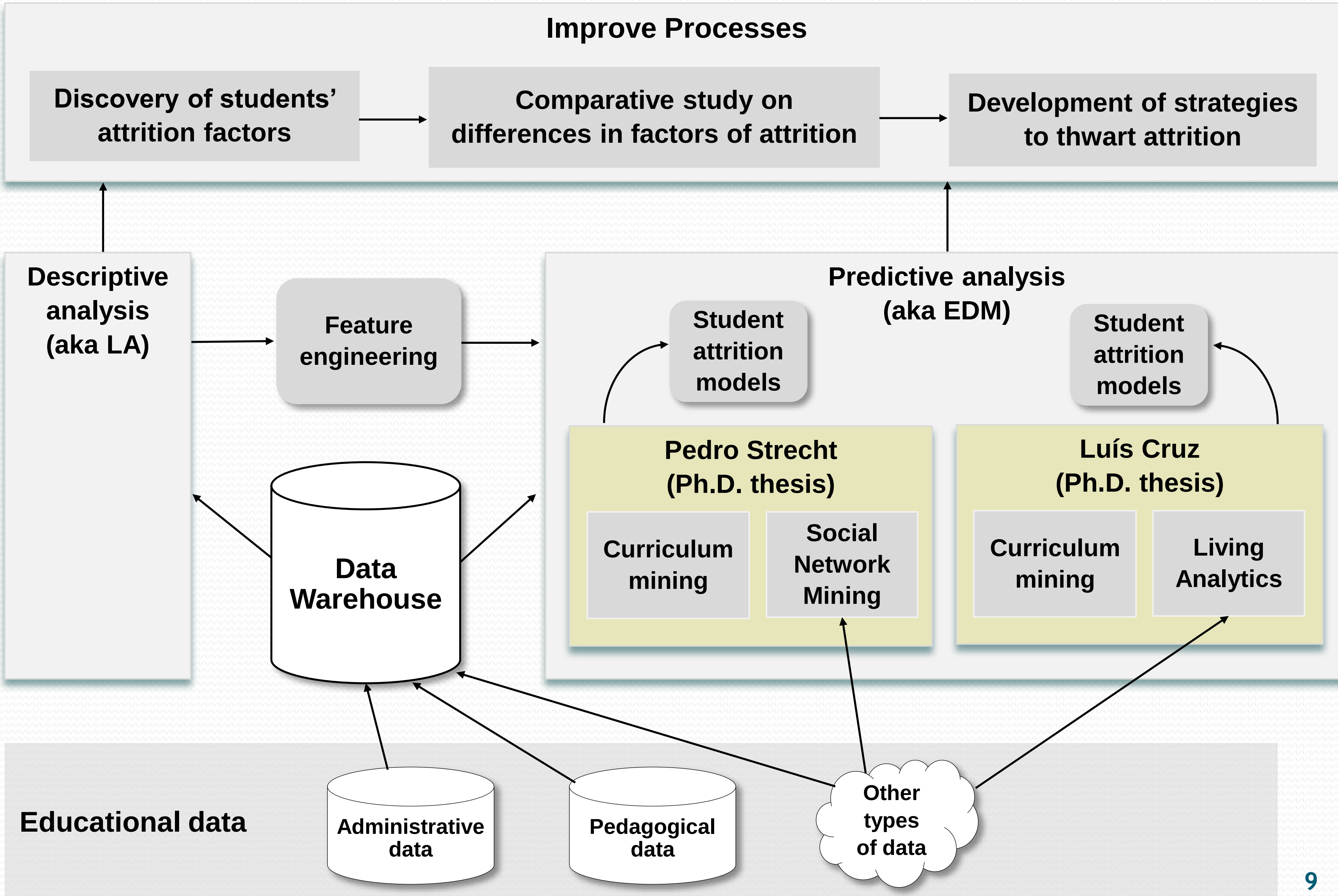


Learning Analytics at U.Porto: current work

Front-ends such as
- Pivot tables
- Report tools



Educational DM at U.Porto: current work

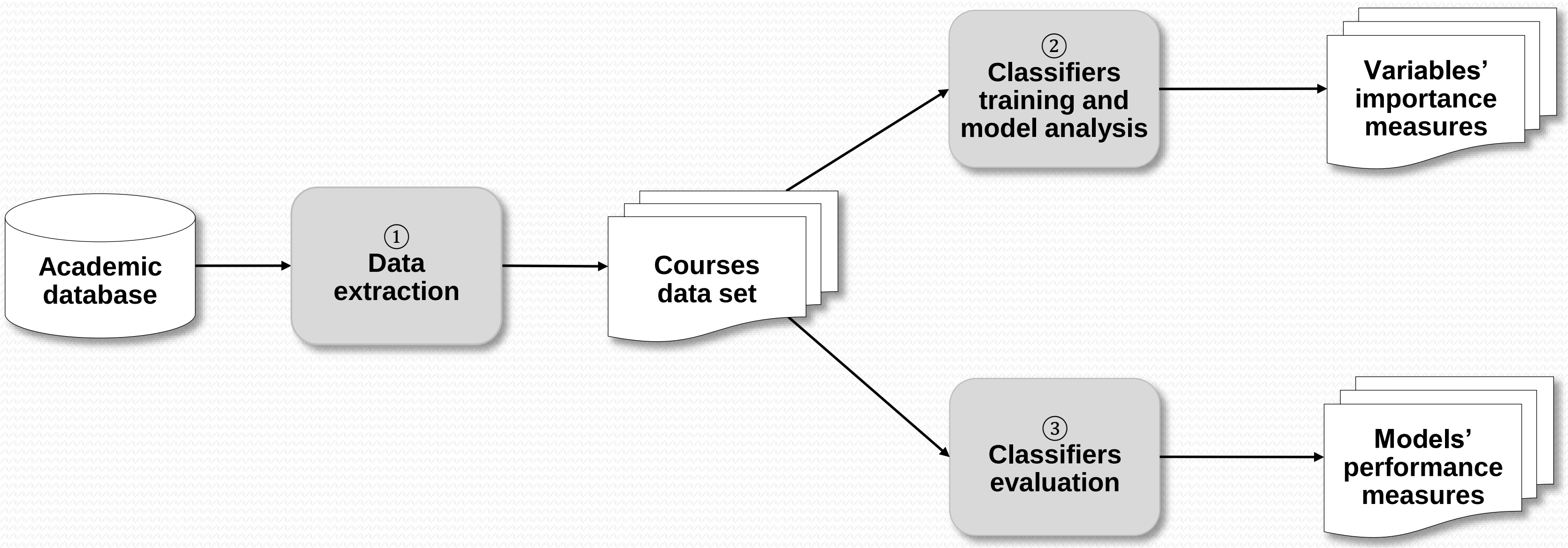


Summary

- Data Analysis in Education
- … at the University of Porto
- An illustrative example of an EDM task
- Conclusions and Future work

An illustrative example of an EDM task

- System to predict if a student will pass or fail a course
- Using administrative data from UIS
- Three different processes



Data extraction

- 14 variables extracted relating to each student

Group	Variable
Socio-demographic information	Age
	Sex
	Marital status
	Nationality
	Displaced
	Scholarship
	Special needs
Admission information	Type of admission
Enrollment information	Type of student
	Status of student
	Years of enrollment
	Delayed courses
	Type of dedication
Financial information	Debt situation

- 8 courses were selected

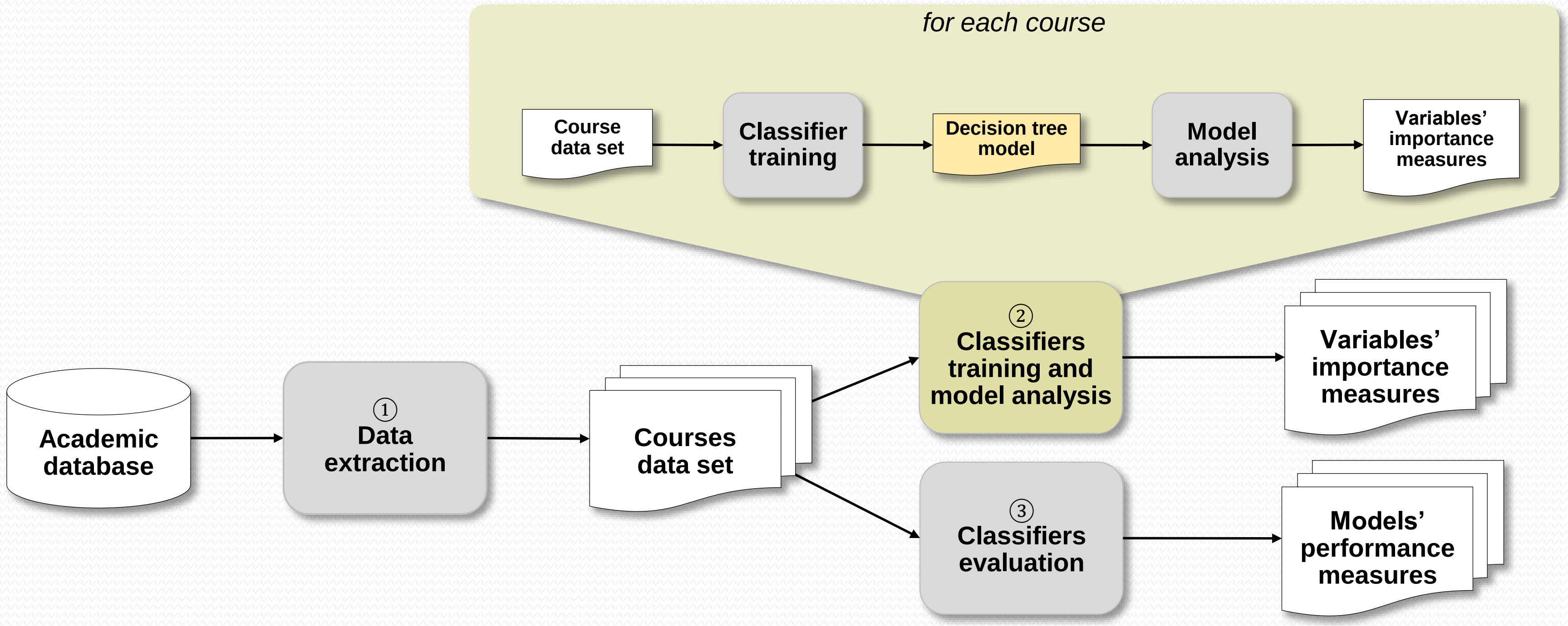
Data extraction

- Data set sample for course Mathematics II

Age	Sex	Marital status	Nationality	Displaced	Scholarship	Special needs	Type of admission	Type of student	Status of student	Years of enrollment	Delayed courses	Type of dedication	Debt situation	Approval
18	m	s	pt	y	n	n	r	r	o	0	0	f	n	n
32	m	m	pt	n	n	n	tcs	r	o	8	12	p	n	n
18	f	s	pt	y	n	n	r	r	o	0	0	f	n	y
18	m	s	pt	n	n	n	r	r	o	0	0	f	n	y
22	m	s	br	n	n	n	to	r	o	1	0	f	n	y

Classifiers training and model analysis

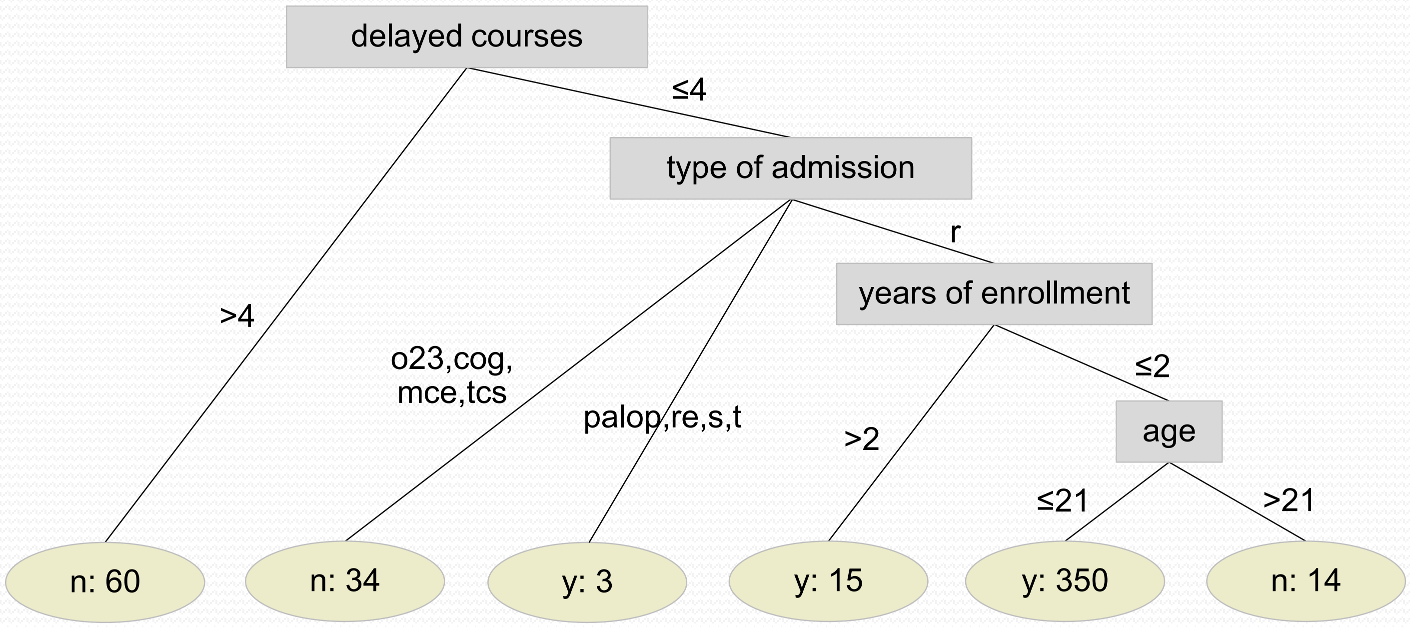
Experimental setup



Classifiers training and model analysis

Classifier training

- Classifiers predict categorical class labels
- Students are classified as either having as either having passed or failed
- Example of decision tree for course Mathematics II:

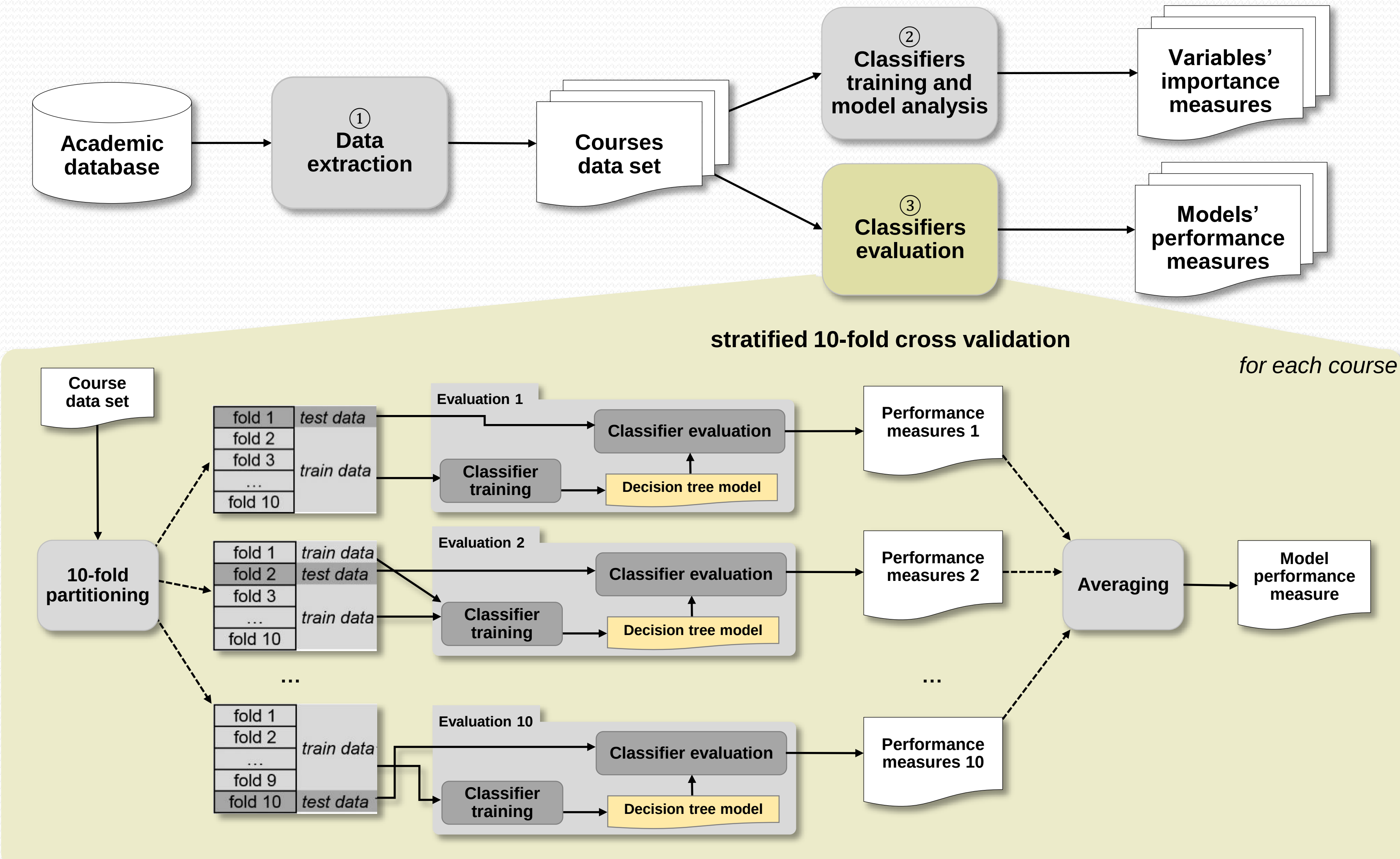


Classifiers training and model analysis
Preliminary results

- Variables’ importance measure for each course

Course	#P	I_p (%)													
		Age	Sex	Marital status	Nationality	Displaced	Scholarship	Special needs	Type of admission	Type of student	Status of student	Years of enrollment	Delayed courses	Type of dedication	Debt situation
Economic History	1								100.0						
Organic Chemistry II	3								11.5			72.1	100.0		
Neuroanatomy	2											11.8	100.0		
Marketing	1											100.0			
Anatomy I	1												100.0		
Anatomy II	4								36.7		20.9	18.4	100.0		
Mathematics II	4	76.4							87.4			79.6	100.0		
Introduction to Linear Signals and Systems	3	100.0							93.1				83.9		

Classifiers evaluation
Experimental setup



Classifiers evaluation

Performance results

- Model performance for each course (10 experiments)

Course	Number of examples	Category distribution (%)		F1 (avg ± std.dev)
		y	n	
Economic History	656	72	28	0.83 ± 0.003
Organic Chemistry II	562	21	79	0.10 ± 0.030
Neuroanatomy	542	94	6	0.96 ± 0.001
Marketing	519	90	10	0.95 ± 0.002
Anatomy I	518	73	27	0.85 ± 0.003
Anatomy II	477	73	27	0.84 ± 0.004
Mathematics II	476	61	39	0.78 ± 0.005
Introduction to Linear Signals and Systems	475	55	45	0.71 ± 0.099

Conclusions

- There is a global effort of University of Porto to improve their processes using BI and DM
- This work presents the preliminary experiments on Educational Data Mining
  - Using administrative data
  - Collecting 14 variables from students enrolled in 8 courses
  - Interpreting results from decision tree models
- Results indicate that
  - Decision trees are quite different from one another
  - Delayed courses is the most important variable
    - Will this pattern hold if more courses are used?
  - Model performance is quite acceptable overall

Future work

- Study the reasons for the variability of variables in each course
- Alternatives to combine decision trees into
  - a single consensual tree
  - small set of trees

  that represent the general knowledge about the success/failure behavior across all the University
- Although the focus is on EDM, such an approach will be interesting for other areas of application

Questions

?