

EVALUATION OF STUDENT PERFORMANCE WITH DATA MINING: AN APPLICATION OF ID3 AND CART ALGORITHMS

Manawin Songkroh (Ph.D)

***College of Arts, Media and Technology, Chiang Mai
University, Chiang Mai, Thailand***

Andrea Kő (Ph.D)

Corvinus University of Budapest, Hungary



Outline:

- Purpose
- Business Objective
- Data Mining Objective
- Data Description
- Data Quality Assessment
- Model Selection
- Model Evaluation
- Conclusions & Implications

Purposes:

- General Purpose: to **classify students** into successful and marginal groups, in order to **find better ways to advise** them and
- To assist university admission officials in **identifying students** that are likely to be successful in a graduate program
- Data Mining Purpose: To **create classification model** : CART & ID3

Business Objective

- **Retain & ensure** the graduation in appropriate time frame

Data Mining Objectives

- Data Acquisition
- *Data Preparation*
- Build Classification Models
- *Model Evaluation*

Data Preparation

A	B	C	D	E	F	G	H	I	J	K	L
ID	TRM	E_CREDIT	C_CREDIT	A_CREDIT	GPA	POINT	OCE	OCC	OCA	OPN	OGPA
492112079	491	15	15	15	2.3	34.5	15	15	15	34.5	2.3
492112079	492	30	30	30	2.15	64.5	15	15	15	30	2.15
492112079	501	45	45	45	2.33	105	15	15	18	40.5	2.7
492112079	502	60	60	60	2.3	138	15	15	15	33	2.2
492112079	511	78	78	78	2.37	184.5	18	18	18	46.5	2.58
492112079	512	93	93	93	2.45	228	15	15	24	43.5	2.9
492112079	521	120	111	120	2.51	279	27	18	27	51	2.83
492112080	491	15	15	15	2.6	39	15	15	15	39	2.6
492112080	492	30	30	30	2.5	75	15	15	15	36	2.4
492112080	501	45	45	45	2.7	121.5	15	15	18	46.5	3.1
492112080	502	60	60	60	2.7	162	15	15	15	40.5	2.7
492112080	511	78	78	78	2.6	202.5	18	18	18	40.5	2.25
492112080	512	93	93	93	2.61	243	15	15	21	40.5	2.7
492112080	521	120	111	120	2.7	300	27	18	27	57	3.17

example of data

A	B	C	D	E	F
492112079	511	702100	INTRO TO ENTRE FINA	3	C
492112079	511	703202	PRINCIPLES OF MANAGEMENT	3	D+
492112079	511	703204	BUSINESS ETHICS	3	C+
492112079	512	255323	INTRO TO QA	3	A
492112079	512	705211	MARKETING PRINCIPLES	3	C
492112079	512	954242	WEB PROGRAMMING	3	B
492112079	512	954260	KM SYSTEM	3	I
492112079	512	954342	DATA MANAGEMENT	3	B
492112079	512	954380	CO-OP I	3	P
492112079	512	954443	MULTIMEDIA PROGRAMMING	3	W
492112079	512	954460	INTRO TO MANU INFO SYS	3	C+
492112079	521	006206	HOME MANAGEMENT	3	B
492112079	521	253300	INTRO TO ENVI MANGT	3	B+
492112079	521	703231	HUMAN RESOURCE MANAGEMENT	3	C
492112079	521	954441	E-COMMERCE	3	C+
492112079	521	954481	CO-OP II	6	S
492112079	521	954491	SELECTED TOPIC IN MMIT	3	B
492112079	521	954380	CO-OP I	3	S
492112079	521	954260	KM SYSTEM	3	B
492112080	491	001103	FOUNDATION ENGLISH I	3	C
492112080	491	050100	USAGE OF THAI LANGUAGE	3	B
492112080	491	159101	FUN IN SOCIO AND ANTHRO	3	C+
492112080	491	208100	INTRO TO STAT CONCEPT	3	B
492112080	491	954140	IT LITERACY	3	C+
492112080	492	001104	FOUNDATION ENGLISH II	3	D+
492112080	492	011157	SOCIAL PHILOSOPHY	3	A

Data Recipe

A	B	C	D	E	F		DD	DE
Prefix	Program	Province	C001	C002	C003	C008	S2008_3	S2009_1
r.	Special	Chiang0Mai	0	0	D	F	2.29	2.41
r.	Special	Lamphun	0	0	C+	C+	2.56	2.61
r.	Special	Chiang0Mai	0	0	D	F	2.09	2.18
iss	Special	Chiang0Mai	0	0	B+	C+	2.75	2.71
iss	Special	Payao	0	0	D	F	2.05	1.99
r.	Special	Chiang0Mai	0	0	C+	D+	2.15	2.14
r.	Special	Chiang0Mai	0	0	D+	D	2.59	2.76
iss	Special	Lampang	0	0	D	F	2.55	2.69
iss	Special	Chiang0Mai	0	0	B	B	3	3.09
iss	Special	Other	0	0	D	D	2.39	2.5
iss	Special	Chiang0Mai	0	0	D	F	2.04	2.15
r.	Special	Lampang	0	0	C	D+	2.41	2.55
r.	Special	Chiang0Mai	0	0	B	C+	3.06	3.05

Model Selection

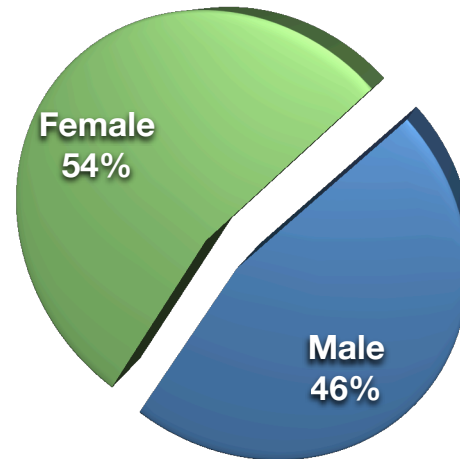
- CART & ID3 assumes ***nonparametric***, algorithms selected ***automatically, categorical/continuous variables.***

Model Evaluation

- Cross Validation
- 80:20 (Training and Evaluation Test Set)

Data Description

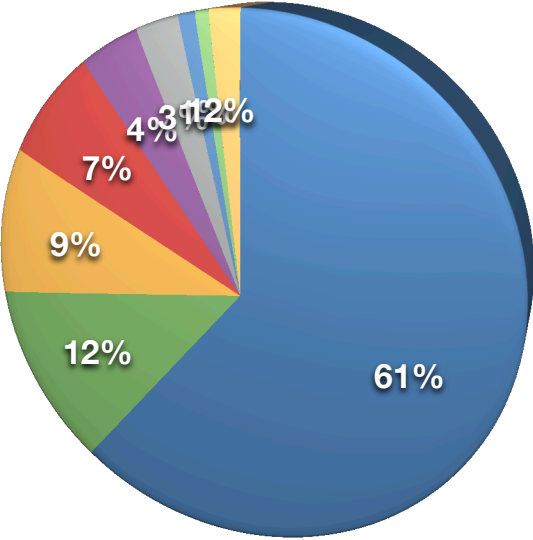
Number of Students by Gender



● Male ● Female

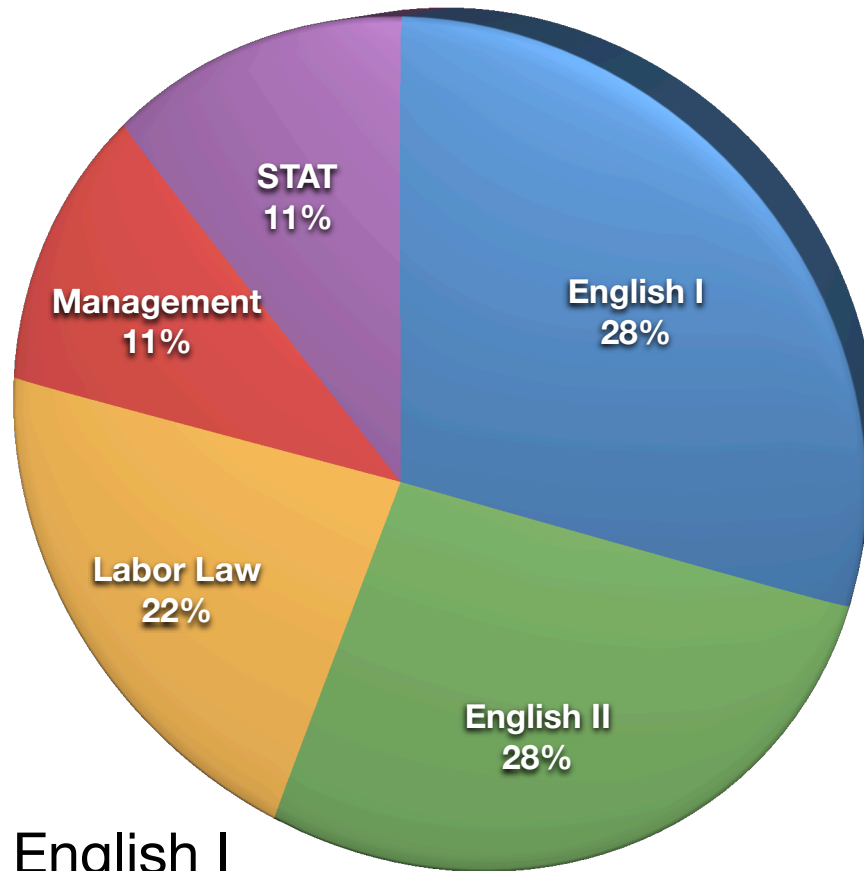
M= 235 F= 272

By Province



- Chiang Mai
- Chiang Rai
- Payao
- Nan
- Other
- Lamphun
- Lampang
- Prae
- Bangkok

Grade F Frequency



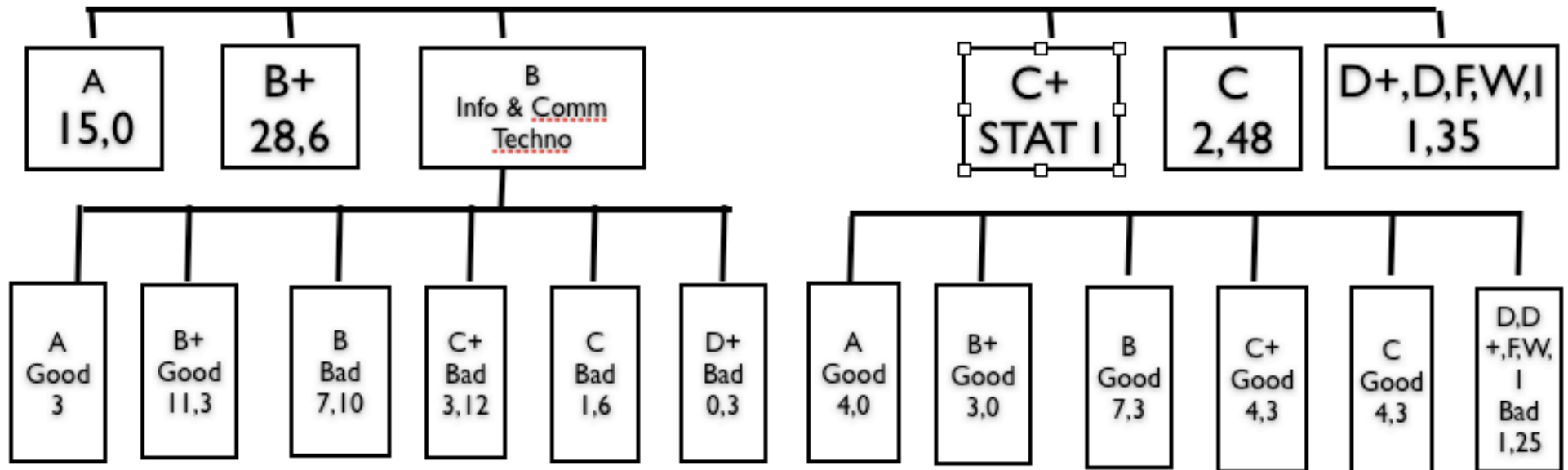
- English I
- English II
- Labor Law
- Management
- STAT

Data Quality Assessment

- No outliers
- 10 missing data
- $\text{GPA} > 2.5 = \text{Good}$,
- $\text{GPA} < 2.5 \rightarrow \text{Bad}$

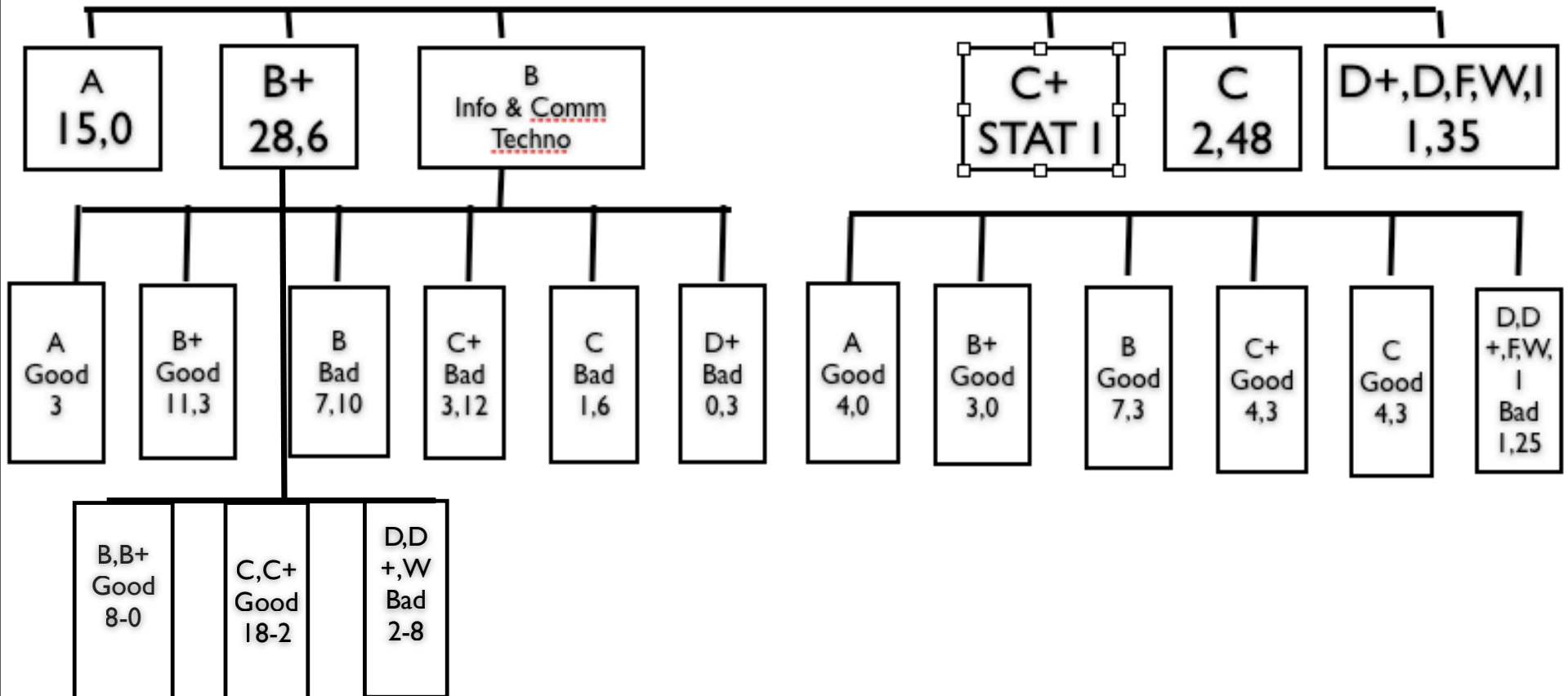
ID3 Model

English 3



CART Model

English 3

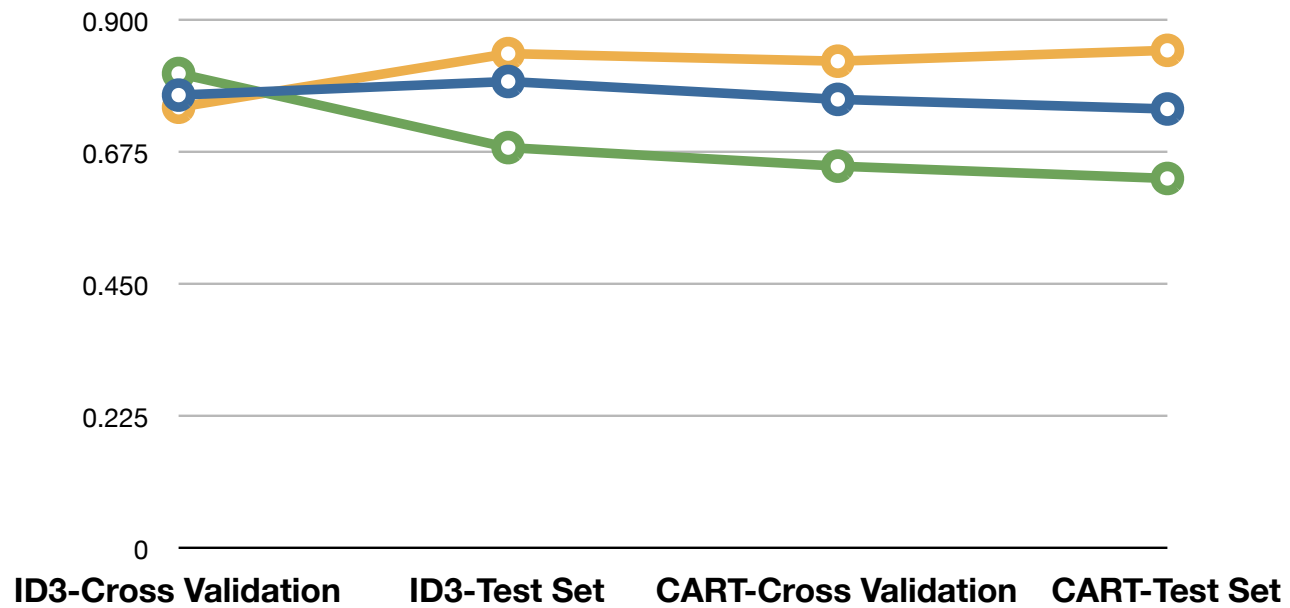


Model Evaluation

Evaluator	Overall Accuracy	Accuracy for Good	Accuracy for Bad
ID3-Cross Validation	77.37%	81.05%	75.30%
ID3-Test Set	79.69%	68.42%	84.44%
CART-Cross Validation	76.64%	65.26%	83.15%
CART-Test Set	75%	63.16%	85%

Comparison of Model Evaluation

Comparison of Model Evaluation



Implications:

- **English, Statistics, and Information and Communication Technology** are the key determinant subjects.
- The results of Classification are **congruent with** the frequency data as many students receive F in these classes.
- **English and statistics** should be the subject used to screen students during **admission.**

Implications (2)

- **Info & Comm Technology** is the Major mandatory subject that is required special attention as it will determine the academic performance of the other related subjects.

Conclusions

- This presentation **outlined the features of a classification technique** to evaluate student performance in their undergraduate programs
- **Classification technique** holds the promise as an evaluation tool to classify students into successful and marginal categories and supports to identify students that are likely to be successful in a graduate program

Conclusion (2)

- The use of a classification model can **support and potentially improve decision making** by program directors and dean.

Q & A