



Institutional Research

*From a retrospective view using business intelligence to
a future view using predictive analytics*

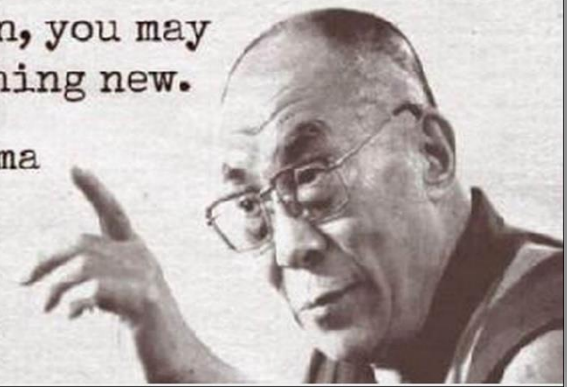
Amanda and Thys Lourens
28 September 2015

Outline

- Background
- The key issue – student dropout
- Evolvment of IR - Insight layers of PowerHEDA BI
- Retrospective views of data
- Future views – predictive analytics case study
- Summary and future developments

When you talk, you are
only repeating what
you already know. But
if you listen, you may
learn something new.

- Dalai Lama



Background

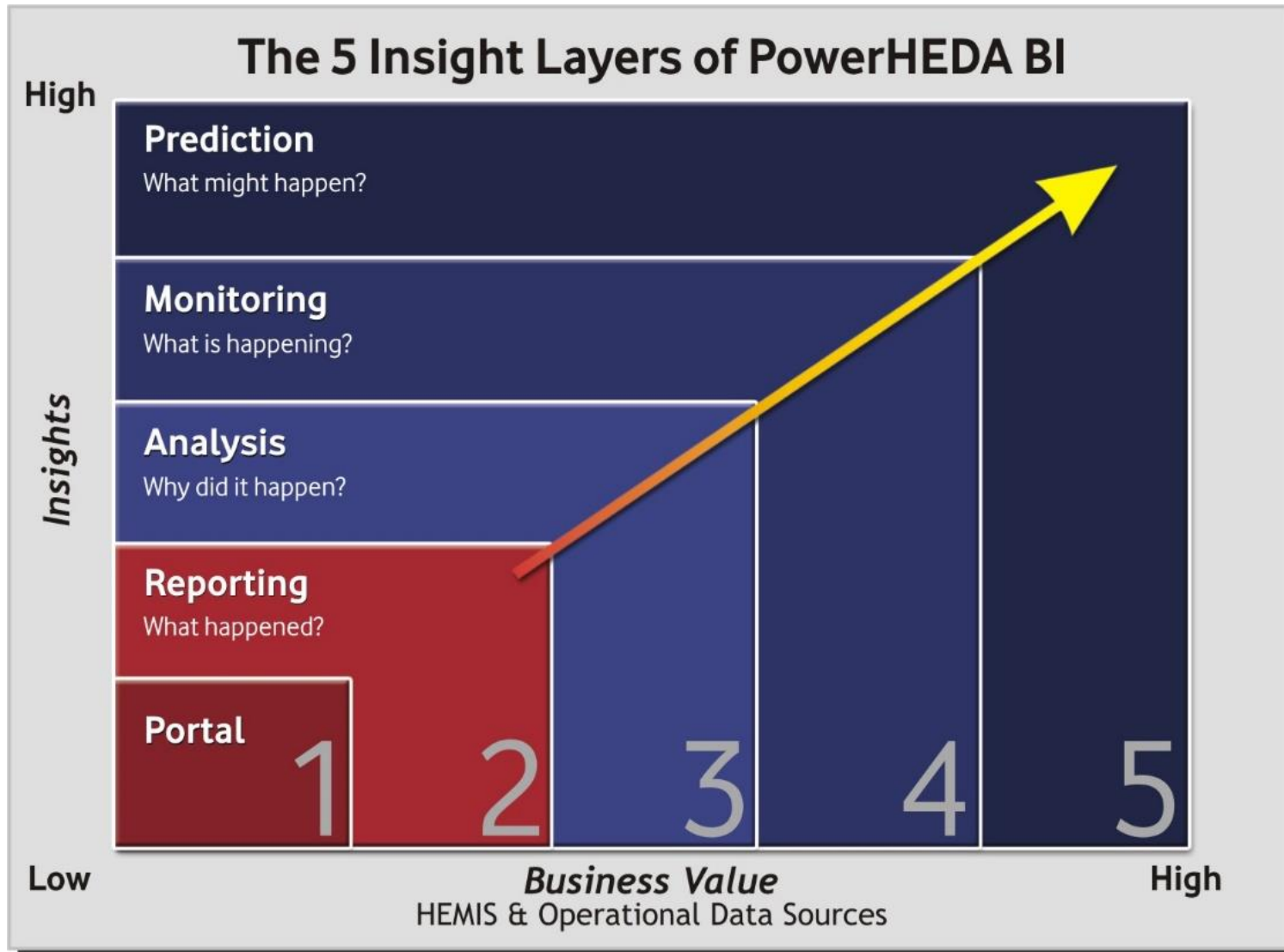
- Evolvement of IR - from '**looking backward**' to identifying '**what is next**' or moving beyond trends
- Retrospective and future views of information
- **IR practitioners** - from an administrative role to academic and scholarly roles
- Illustrating the changing focus with one single **key** issue
- Identifying solutions to assist **students** and the **institution**

Business understanding:

Key issue

- High drop out by **second year** of study (Scott et.al. 2007, CHE 2010 & 2014, NDP 2011)
- Higher education literature (internationally and locally) provides a wide range of theory about **reasons** for students leaving (Pike et al. 2014, Biswas 2007, Woodhead 2002, Herzog 2005, Hess 2008, Liu 2000, Dekker et al. 2009, Tinto 1975, Pascarella & Terenzini 1983, Letsaka & Maile 2008, van Zyl et al. 2012, Lourens & Smit 2003, Murray 2014)
- The need exists for a **practical** contribution to student retention to enable an institution to implement student-specific intervention strategies
- Gaining more insight in relation to the issue of **second year** student **dropout**

Evolverment of IR: Gaining more insight



Top 10 Qualification's with highest second year drop-out rate

Year: 2014
Faculty: INFORMATICA
Qualification Type: NATIONAL DIPLOMA
Extended Flag: Normal
Formal: F

<u>APPROVED QUALIFICATION NAME</u>	<u>2014 COHORT SIZE</u>	<u>2YR DROP OUT %</u>
ND:DIPLOMA B	16	50.00%
ND:DIPLOMA A	187	46.52%
ND: BUSINESS ANALYTICS	181	37.02%
ND:DIPLOMA C	36	33.33%
ND:DIPLOMA D	16	31.25%
ND:DIPLOMA E	30	30.00%
ND:DIPLOMA F	28	28.57%
ND:DIPLOMA G	42	23.81%
ND:DIPLOMA H	34	23.53%
ND:DIPLOMA I	34	23.53%

Data Definitions

Data	ITS Operational M01V
Second Year Drop-out	Student that enrolled in first year but did not returned for the same Qualification in the following year
FTEN Status	First-time Entering
BLOCK_ONLY_FOR_EXAMS	N
SUBSIDY_TYPE	<> C

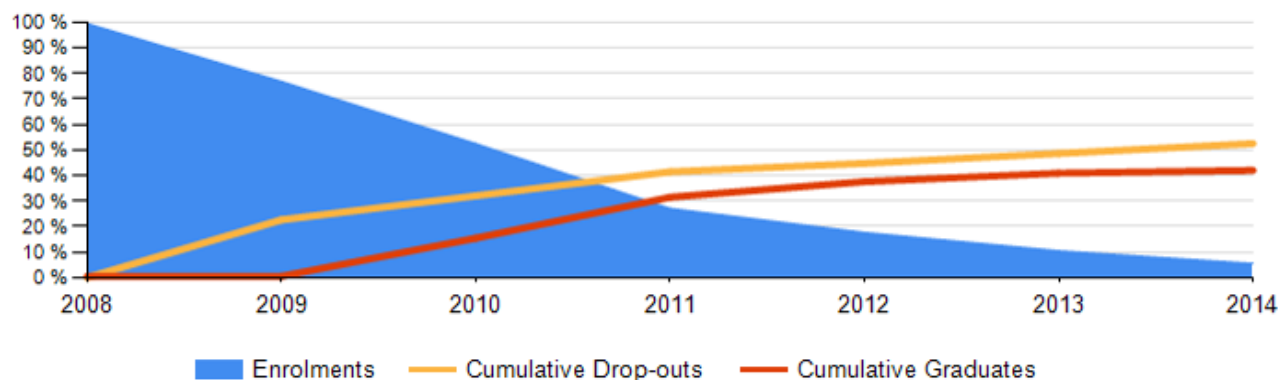
Retrospective views

What
happened?

Report Parameters

Cohort Year	2008						
Cohort Definition	First-time Cluster Enrolment						
Approved Qualification Description	ND: BUSINESS ANALYTICS						
Entrance category	F						
Include Related Clusters	Yes						
Entering Term	1st Year 2008	2nd Year 2009	3rd Year 2010	4th Year 2011	5th Year 2012	6th Year 2013	7th Year 2014
2008 Baseline Enrolment	181	181	181	181	181	181	181
Enrolments (Retained)	180	113	72	38	22	13	10
% Enrolments	99%	62%	40%	21%	12%	7%	6%
Stop-outs (included in Enrolments)	0	11	3	7	8	2	0
Drop-outs	0	67	14	12	5	3	1
% Drop-outs	0%	37%	8%	7%	3%	2%	1%
Cumulative Drop-outs	0	67	81	93	98	101	102
% Cumulative Drop-outs	0%	37%	45%	51%	54%	56%	56%
Graduates	1	0	27	22	11	6	2
% Graduates	1%	0%	15%	12%	6%	3%	1%
Cumulative Graduates	1	1	28	50	61	67	69
% Cumulative Graduates	1%	1%	15%	28%	34%	37%	38%

2008 Baseline Enrolment



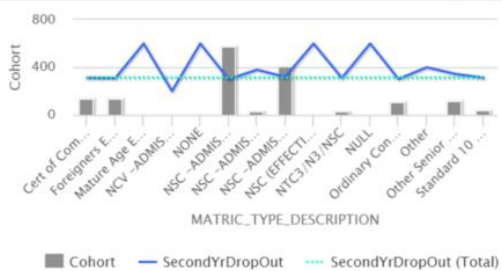
What happened?

Retrospective views: Why did it happen?

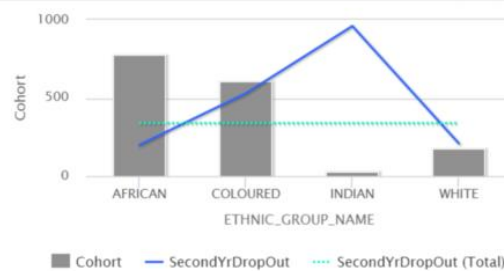
Related Dashboards

- Student Cohort Analyses
- Current FTEN Enrolments
- Second Year Drop Out Statistics
- Second Year Drop-out Statistics over Time

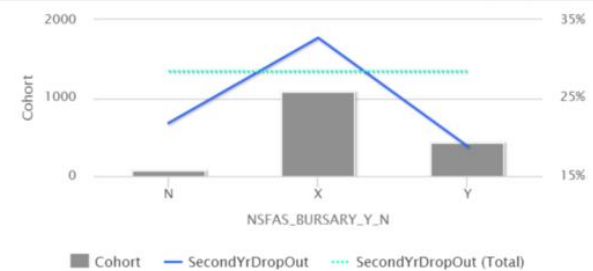
by Matric Type



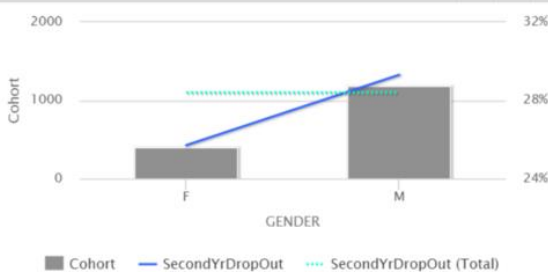
by Ethnic Group



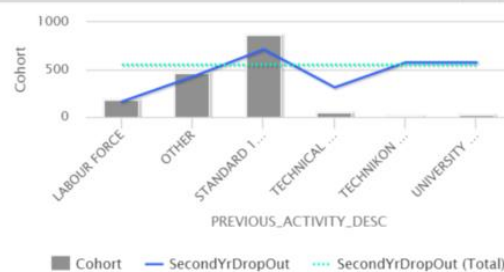
by NSFAS Status



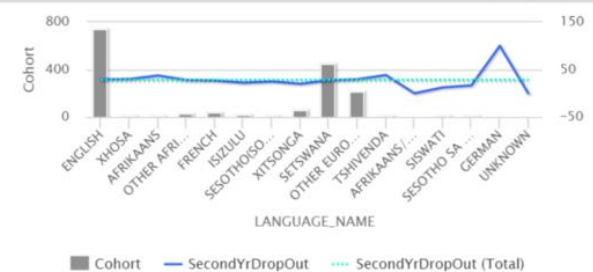
by Gender



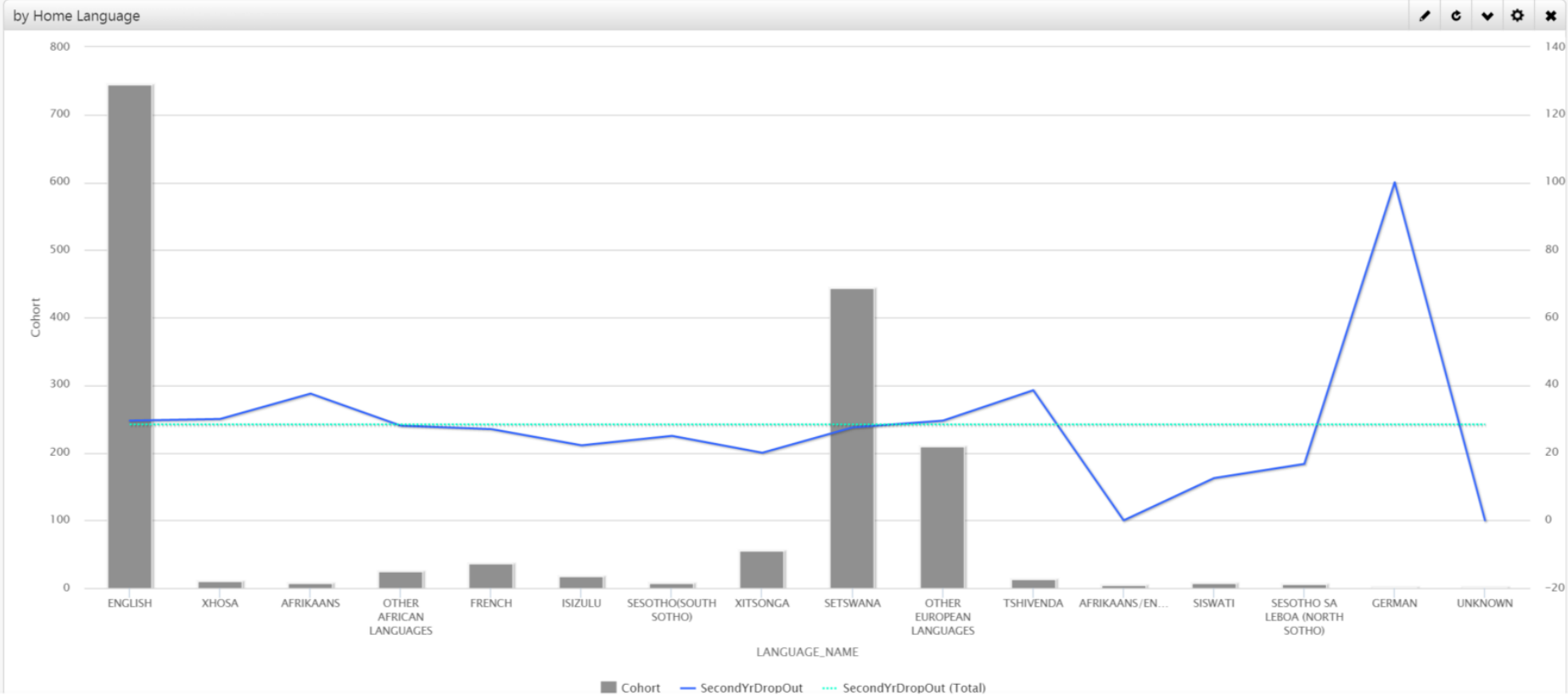
by Previous Year Activity



by Home Language



Retrospective views: Why did it happen?

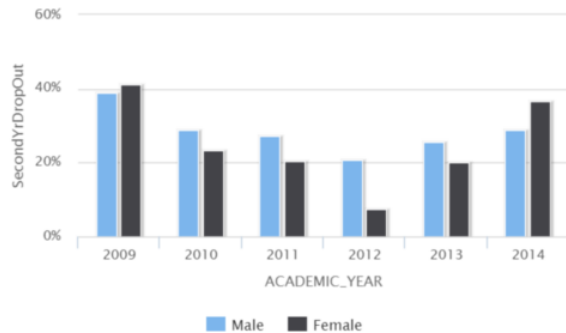


Retrospective views: Why did it happen?

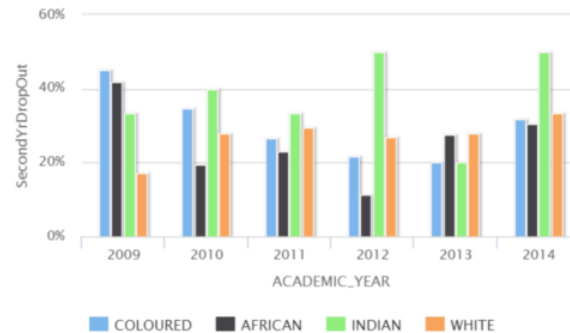
Related Dashboards

- Student Cohort Analyses
- Current FTEN Enrolments
- Second Year Drop Out Statistics
- Second Year Drop-out Statistics over Time

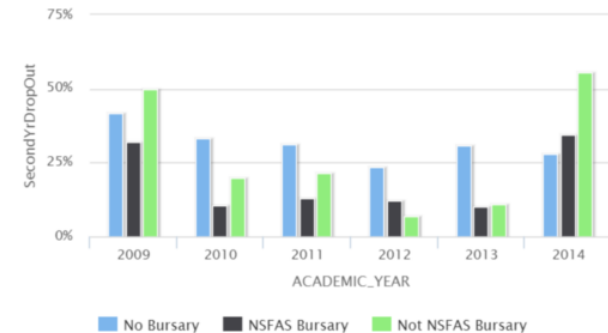
by Gender



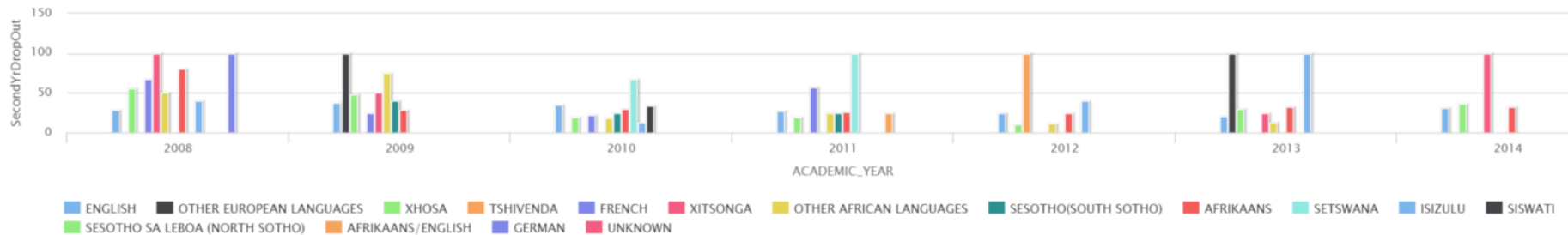
by Ethnic Group



by NSFAS Status



by Home Language



What is happening?

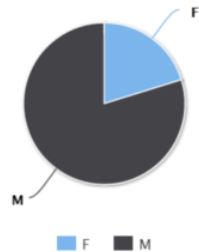
National Diploma: Business Analytics

2015 Overview

by Full/Part Time

	2013	2014	2015	
	Cohort	Cohort	Cohort	Cohort
FT	200	158	148	
PT	17	23	30	
TOTAL	217	181	178	

2015 by Gender



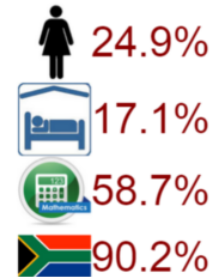
Related Dashboards

- Student Cohort Analyses
- Current FTEN Enrolments
- Second Year Drop Out Statistics
- Second Year Drop-out Statistics over Time

by Matric Type

	2013	2014	2015	
	Cohort	Cohort	Cohort	Cohort
Ordinary Conditional Exemption	9	4	9	
Other Senior Certificates	10	18	21	
Cert of Complete Exemption	4	2	2	
NSC - ADMISSION TO BACHELOR	113	91	98	
Foreigners Exemption	16	11	12	
NSC - ADMISSION TO DIPLOMA	53	43	25	
NTC3/N3/NSC	5	6	3	
NSC - ADMISSION TO CERTIFICATE	5	5	2	
NCV - ADMISSION TO CERTIFICATE	2	0	0	
Standard 10 Practical	0	1	1	
NONE	0	0	1	
NSC (EFFECTIVE FROM	0	0	3	

Key Stats in Images



Key Stats

100.0%	Enrolled for Development Software
50.0%	Less than 30% in Dev Software
5 030	Avg Bursary Amount
17.4%	% Afrikaans
85.4%	% Black
3.4%	% Matric Math ABC
52.8%	% Direct from matric

What is happening?

HEDA 2011 Student Tracker - First test mark at risk (Mr B George) - Records (PREVIEW)

Reset filter

Preview email (group)

Run & Email Notifications

View Status

Export Excel \ SMS

CALENDAR YEAR	ACADEMIC BLOCK CODE	STUDENT NUMBER	SUBJECT CODE	QUALIFICATION CODE	MARK TYPE CODE	MARK 1	MARK 2	DEPARTMENT CODE
2013	2	21201955	ADPH11	ADPH1A	TM	45		6500
2013	2	21208449	ABASC11	ABASC1A	TM	0	29	6500
2013	2	21200574	ADPH11	ADPH1A	TM	40		6500
2013	2	21206462	ABASC11	ABASC1A	TM	47	35	6500
2013	2	21206462	ADPH11	ADPH1A	TM	26		6500
2013	2	21216736	ADPH11	ADPH1A	TM	38		6500
2013	2	21206411	ADPH11	ADPH1A	TM	34		6500
2013	2	21201817	ADPH11	ADPH1A	TM	47		6500
2013	2	21207454	ADPH11	ADPH1A	TM	45		6500
2013	2	21206687	ADPH11	ADPH1A	TM	37		6500
2013	2	21208449	ADPH11	ADPH1A	TM	13		6500
2013	2	21210277	ADPH11	ADPH1A	TM	16		6500
2013	2	21206461	ADPH11	ADPH1A	TM	37		6500
2013	2	21207080	ADPH11	ADPH1A	TM	47		6500
2013	2	21207128	ADPH11	ADPH1A	TM	46		6500
2013	2	21218150	ADPH11	ADPH1A	TM	21		6500
2013	2	21201913	BLPH421	BLPH42D	TM	48		6100
2013	2	21201254	BLPH421	BLPH42D	TM	28		6100
2013	2	21208449	COB231	COB23D	TM	20		6100
2013	2	21200715	BLPH421	BLPH42D	TM	28		6100

But, we need to look forward...

- Being able to ***predict*** more ***accurately*** which students might potentially drop out would enable institutions to focus on intervention strategies and will improve enrolment planning.
- Aim of case study is to provide a ***list of student names*** with high probability of dropping out by the second year of study

Predictive analytics: What might happen?

- Predictive analytics is the process of discovering *interesting* and *meaningful patterns* in data. It draws from related disciplines including statistics, machine learning and data mining (Abbott, 2014).
- **CRISP-DM** (Cross-Industry Standard Process Model for Data Mining)
 - Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment
- Konstanz Information Miner (**KNIME**) – free open source platform for data analysis
- Supervised learning methods

Statistics vs. Predictive Analytics (Abbott, 2014)

	Statistics	Predictive Analytics
View of “other” field	“data dredging”	“we can do that ... and more!”
Emphasis	Theory, optimum solutions	“Good” Heuristics
Approach	Parametric/non-parametric	Non-parametric
Key metrics of performance	P-values, R^2 , SE ...	Lift, ROC
What is King?	Model	Data

Also see David J. Hand, *“Statistics and Data Mining: Intersecting Disciplines”*

Predictive modelling: Case study

- Institutional data *for first-time entering* (FE) *undergraduate contact* students in a particular programme
- Want to predict student dropouts *in or before 2nd year of study*
 - 2nd year dropout = FE students who did not register for the 2nd year of study (dependent variable)
- Variety of background information (pre-university and performance-linked data) as *independent* variables
- Three *algorithms* in KNIME used in predictive modelling (most commonly used)
 - Logistic regression
 - Decision trees
 - Naïve Bayes

Logistic Regression

- Linear classification technique for binary dependent variable and categorical/continuous independent variables.
- Test for collinearity important

Classification/Decision Trees

- Very popular – (Rexer Analytics Data Miner Surveys)
- Easy to build and understand (typical “if-then-else” rules)
- Can handle nominal and continuous inputs.
- Build-in variable selection and non-parametric (i.e NO assumptions about distributions for inputs or the target variable)
- Handle missing data automatically

Naïve Bayes

- Based on Bayes’ theorem with independence assumptions between predictors
- Can handle an arbitrary number of independent variables whether continuous or categorical (Ng & Jordan, 2002)

Predictive analytics:

Data preparation

- Data records for FE contact students from **2008 -2014**
- **1593** records used in dataset, data automatically imported from PowerHEDA to KNIME
- **2nd year dropout** (Yes=1/No=0) as dependent/target variable
- **27** variables in dataset – test for collinearity
- **21** independent variables used in Naïve Bayes and Decision Tree models
- **8** independent variables used in Logistic Regression model after backward feature elimination method used
- First-year module marks **clustered** and **binned** in categories

Predictive analytics:

Data understanding

First-time entering students (2008 – 2014)	Frequency	%
2 nd year dropout students	452	28
Students in residence	280	18
Students with MATH = 1 (Math taken in Gr 12)	916	58
Students with DS = bin 1 (mark < 30%)	382	24
Students with ITS = bin 1 (mark < 30%)	143	9
Students with SS = bin 1 (mark < 30%)	135	9
Students with NSFAS = X (no bursary)	1 081	68
Full-time students (Offering type = FT)	1 395	88
Male students	1 188	75
Home Language = English	742	47
Number of subjects taken = 6	1059	67
Students with Matric type = B (NSC - Bachelor)	573	36

Predictive analytics: Training, Testing and Validation

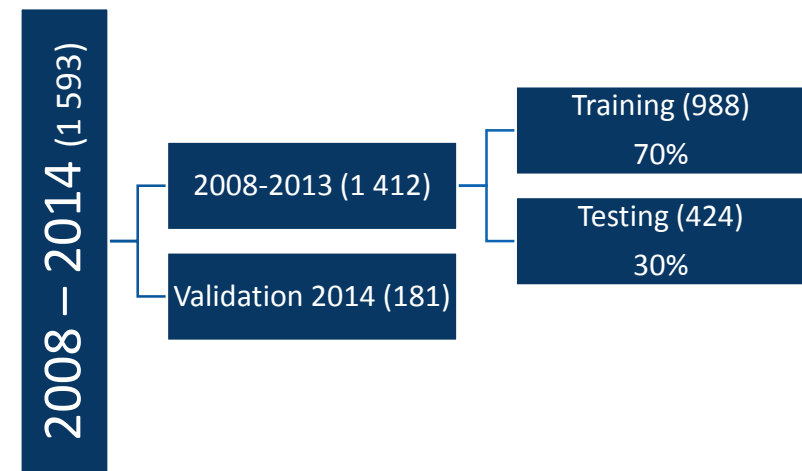
- **2008 to 2013** dataset randomly subdivided into **70%** training and **30%** testing datasets

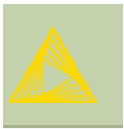
- **2014** dataset kept aside to use later as validation dataset

- **988** records used to build the three models

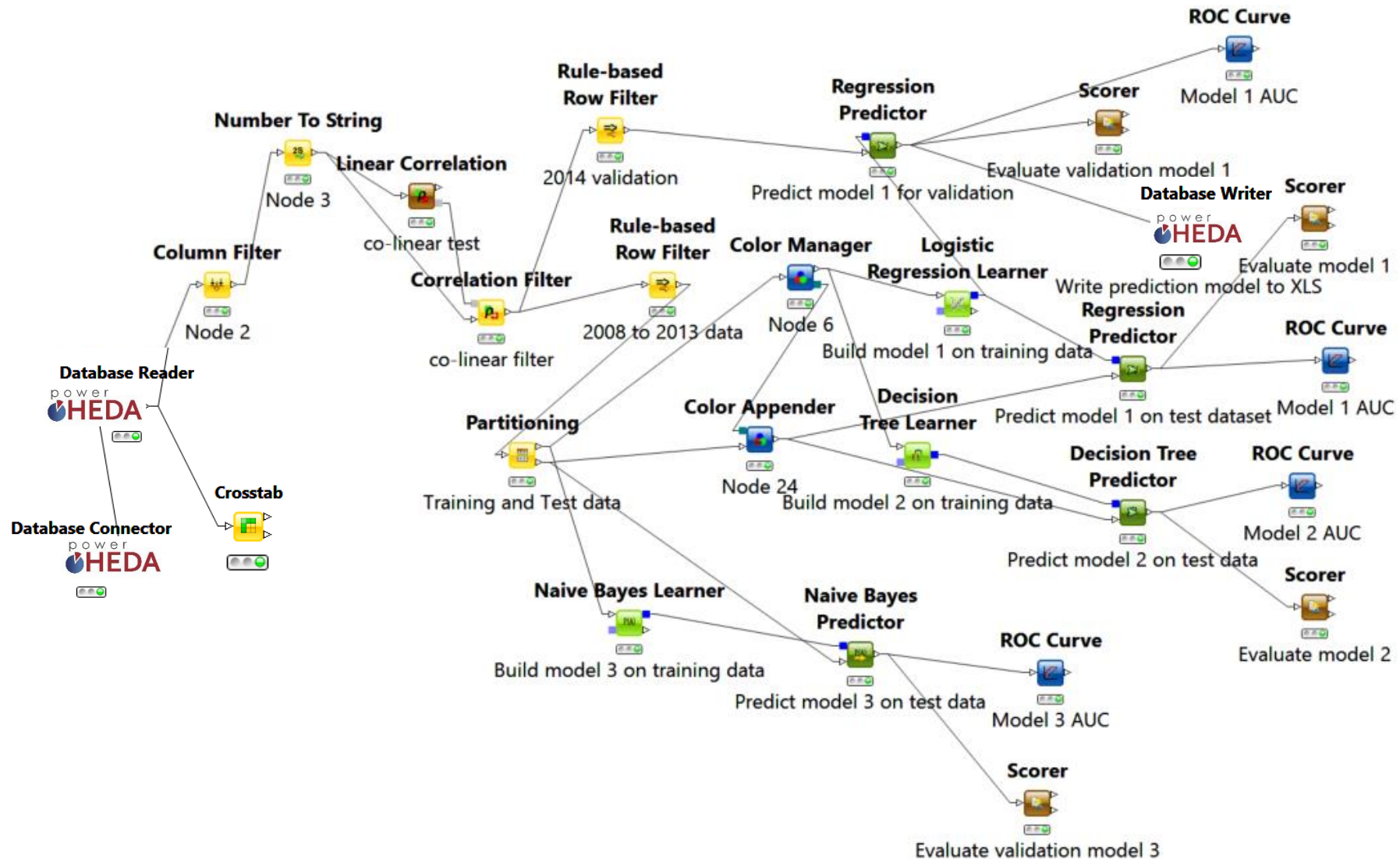
- **424** records used as testing dataset to assess accuracy of the models

- **181** records from the **2014** data used to predict the outcome based on the selected model



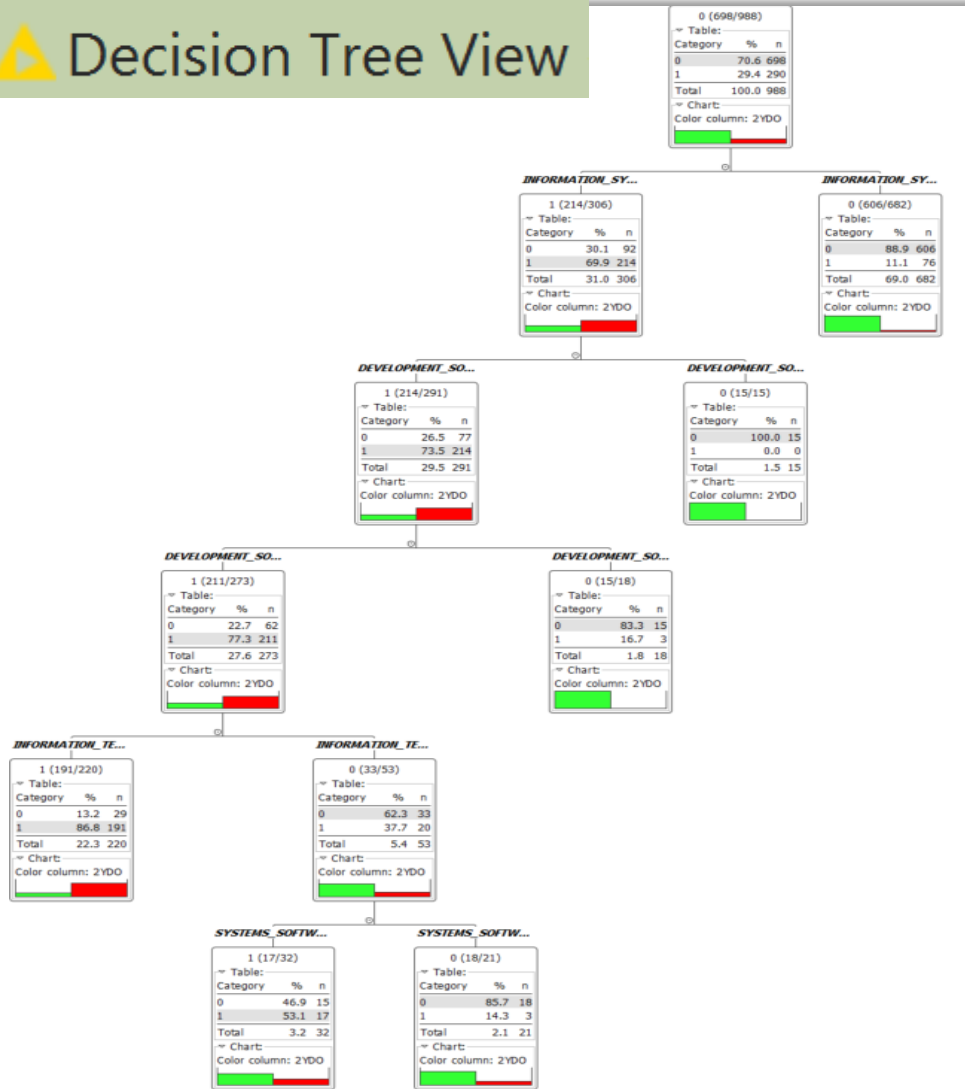


KNIME workflow example



Decision tree and Naïve Bayes: Output examples

Decision Tree View



Naive Bayes Learner View

Class counts for 2YDO

Class:	0	1
Count:	698	290

Total count: 988

P(GENDER | class=?)

Class/GENDER	F	M
0	180	518
1	59	231
Rate:	24%	76%

P(HOME_LANGUAGE | class=?)

Class/HOME_LANGUAGE	A	AE	E	FR	G	N	NS	OA	OE	SS	SW	TS	TW	V	X	Z
0	97	3	331	16	0	1	2	18	3	9	3	6	4	3	189	13
1	35	0	148	8	1	0	0	9	3	4	1	3	1	0	72	5
Rate:	13%	0%	48%	2%	0%	0%	0%	3%	1%	1%	0%	1%	1%	0%	26%	2%

Gaussian distribution for INFORMATION_SYSTEMS_Mark_Bin per class value

	0	1
Count:	698	290
Mean:	4.29226	2.23103
Std. Deviation:	1.15286	1.64024
Rate:	71%	29%

Gaussian distribution for INFORMATION_TECHNOLOGY_SKILLS_Mark_Bin per class value

	0	1
Count:	698	290
Mean:	3.82378	1.65517
Std. Deviation:	1.56326	1.57787
Rate:	71%	29%

P(NSFAS_BURSARY_Y_N | class=?)

Class/NSFAS_BURSARY_Y_N	N	X	Y
0	42	448	208
1	6	241	43
Rate:	5%	70%	25%

Logistic regression: Output example

Statistics on Logistic Regression

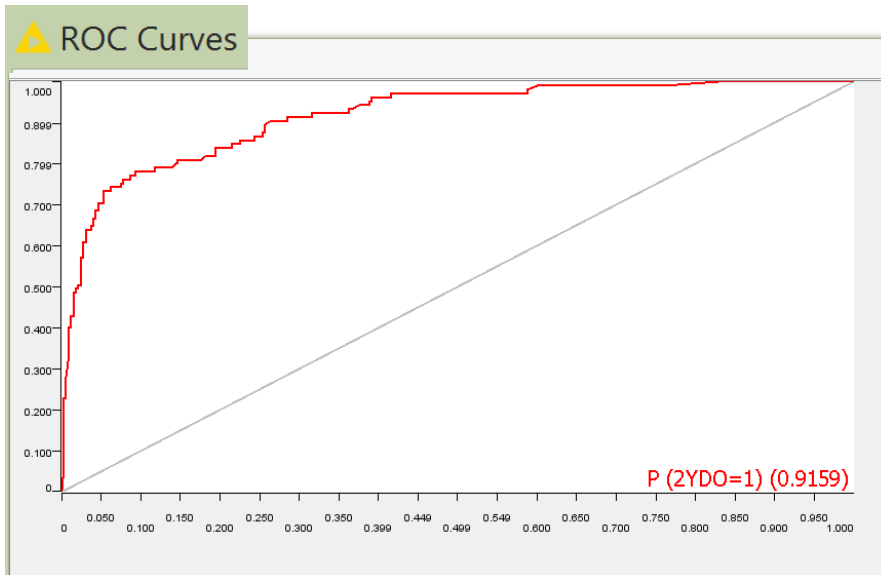
▲ Logistic Regression Result View

Logit	Variable	Coeff.	Std. Err.	z-score	P> z
0	NSFAS_BURSARY_Y_N=X	-1.4202	0.5309	-2.6749	0.0075
	NSFAS_BURSARY_Y_N=Y	-0.8435	0.5537	-1.5233	0.1277
	Res=Y	0.3783	0.2869	1.3186	0.1873
	DEVELOPMENT_SOFTWARE_Mark_Bin	0.6076	0.0788	7.7129	1.05E-14
	INFORMATION_SYSTEMS_Mark_Bin	0.3928	0.0773	5.0829	3.72E-7
	INFORMATION_TECHNOLOGY_SKILLS_Mark_Bin	0.3313	0.0715	4.6312	3.63E-6
	SYSTEMS_SOFTWARE_Mark_Bin	0.0243	0.0709	0.3429	0.7317
	TECHNICAL_PROGRAMMING_Mark_Bin	0.0802	0.0766	1.047	0.2951
	MATH	0.0742	0.1039	0.7138	0.4753
	Constant	-1.849	0.5617	-3.2919	0.001

Evaluation methods

Assess model accuracy using

- Confusion matrix (breakdown of classification errors – actual vs predicted)
- Receiver Operating Characteristic (ROC) curves with Area under Curve(AUC)
- Percentage correctly classified (PCC) and Error rates



Statistic	Logistic Regression	Decision Tree	Naïve Bayes
AUC	0.9159	0.8457	0.9194
Accuracy (PCC %)	88.6	87.5	87.7
Error %	11.3	12.5	12.3

Confusion matrix:

Logistic regression model

Training Records (n = 424)	True Positive	False Positive	True Negative	False Negative	Sensitivity	Specificity	F-measure	Accuracy	Cohens Kappa
Dropout = 0	299	28	77	20	0.937	0.733	0.926		
Dropout = 1	77	20	299	28	0.733	0.937	0.762		
Overall								0.887	0.688

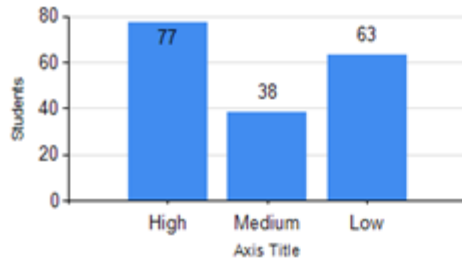
- **True positive**
 - Actual and predicted value = 1 (77 correctly classified as drop-out)
- **False positive**
 - Actual value = 0, predicted value = 1 (20 incorrectly classified as drop-out)
- **True negative**
 - Actual and predicted value = 0 (299 correctly classified as returning)
- **False negative**
 - Actual value = 1, predicted value = 0 (28 incorrectly classified as returning)
- **Sensitivity** = Actual drop-outs classified correctly (73.3%)
- **Specificity** = Actual returning students classified correctly (93.7%)
- **Accuracy** = Overall model accuracy (88.7%)

Predictive analytics: Deployment of model

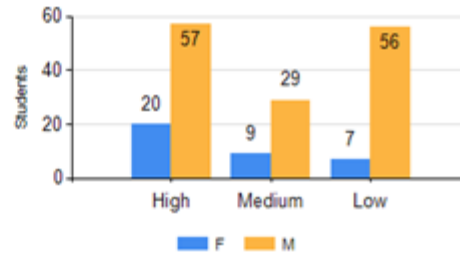
- Logistic Regression model ***deployed*** to score the data
- Probabilities automatically exported from ***KNIME*** to ***PowerHEDA***
- PowerHEDA ***integrated*** the KNIME output with institutional data
- PowerHEDA ***report*** sent to programme owner with details of students with ***high probability of not returning in 2nd year of study***

This report list First-time Entering students who are enrolled for **APP 101 IN APPLICATIONS DEVELOPMENT** in 2015 with an indication of their probability to drop-out in their second year.

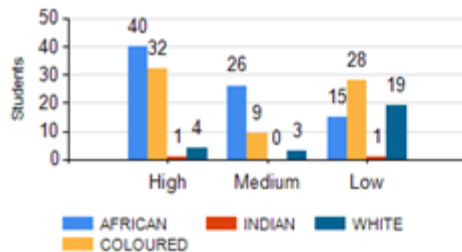
By Probability



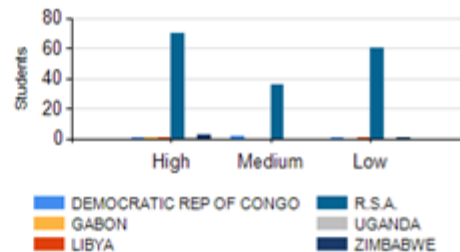
By Gender



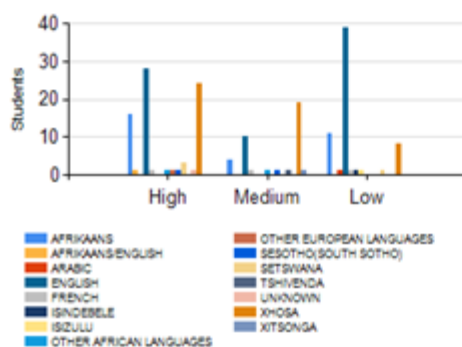
By Ethnic Group



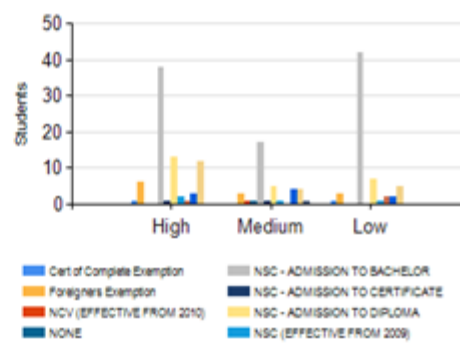
By Country



By Home Language



By Matric Type



PREDICTIVE LEARNING ANALYTICS

N Dip: Business Analytics

Future views

List of 2015 students and their second-year dropout probability

<u>Student Name</u>	<u>Student Number</u>	<u>Cell Phone</u>	<u>Email</u>	<u>%</u>	
MR A. ABRAHAM	21001076	077776667	21001076@power.ac.za	21.08%	●
MR B. ABRAHAM	21001077	077776667	21001077@power.ac.za	3.48%	●
MR BOYF. ABRAHAM	21001712	077776667	21001712@power.ac.za	8.93%	●
MR BO. ABRAHAM	21001713	077776667	21001713@power.ac.za	21.08%	●
MR DE. ALBERTIN	21001687	077776667	21001687@power.ac.za	21.08%	●
MR DU. ALLEN	21001206	077776667	21001206@power.ac.za	42.11%	▲
MR E. APPELS	21001717	077776667	21001717@power.ac.za	63.08%	◆
MR E. APPEL	21001718	077776667	21001718@power.ac.za	42.11%	▲
MR F. BARR	21001719	077776667	21001719@power.ac.za	66.45%	◆
MR G. BARR	21001720	077776667	21001720@power.ac.za	63.08%	◆
MR H. BARR	21001721	077776667	21001721@power.ac.za	3.48%	●
MR I. BARR	21001722	077776667	21001722@power.ac.za	42.11%	▲
MR J. BARR	21001723	077776667	21001723@power.ac.za	66.45%	◆
MR K. BARR	21001724	077776667	21001724@power.ac.za	42.11%	▲
MR L. BARR	21001725	077776667	21001725@power.ac.za	66.45%	◆
MR M. BARR	21001726	077776667	21001726@power.ac.za	3.48%	●
MR N. BARR	21001727	077776667	21001727@power.ac.za	42.11%	▲
MR O. BARR	21001728	077776667	21001728@power.ac.za	66.45%	◆
MR P. BARR	21001729	077776667	21001729@power.ac.za	3.48%	●
MR Q. BARR	21001730	077776667	21001730@power.ac.za	42.11%	▲
MR R. BARR	21001731	077776667	21001731@power.ac.za	66.45%	◆
MR S. BARR	21001732	077776667	21001732@power.ac.za	42.11%	▲
MR T. BARR	21001733	077776667	21001733@power.ac.za	66.45%	◆
MR U. BARR	21001734	077776667	21001734@power.ac.za	3.48%	●
MR V. BARR	21001735	077776667	21001735@power.ac.za	42.11%	▲
MR W. BARR	21001736	077776667	21001736@power.ac.za	66.45%	◆
MR X. BARR	21001737	077776667	21001737@power.ac.za	42.11%	▲
MR Y. BARR	21001738	077776667	21001738@power.ac.za	66.45%	◆
MR Z. BARR	21001739	077776667	21001739@power.ac.za	3.48%	●

What might happen?

Summary and future developments

- **Scoring** needed for early identification of students – statistical results must be practical
- **Integration** of BI tool (PowerHEDA) and statistical software package (KNIME) very important
- Predictive models should be **modified** periodically
- No magic “**one generic**” answer!
- Future developments:
 - ‘First-year experience’ data should be included in future studies – need large samples
 - Student portal and mobile application for students

THANK YOU

Selected References...

- Abbot 2014, *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. John Wiley & Sons Inc. Indiana.
- Biswas 2007, *Accelerating Remedial Math Education: How Institutional Innovation and State Policy Interact*. Boston, MA: Jobs for the Future.
- Council on Higher Education (CHE). 2014. *VitalStats: Public Higher Education 2012*. Pretoria: CHE
- Council on Higher Education (CHE). 2010. *Access and throughput in South African Higher Education: Three case studies*. Pretoria: CHE.
- Dekker et al. 2009, *Predicting Students Drop Out: A Case Study*. Educational Data Mining.
- Herzog 2005, *Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen*. Research in Higher Education 46(8): 883-928.
- Hess 2008, *Still At Risk: What Students Don't Know, Even Now*. Washington, DC: Common Core.
- Letsaka & Maile 2008, *High university dropout rates: A threat to South Africa's future*. Human Science Research Council, 2008, P1-7.
- Liu 2000, *Institutional Integration: An Analysis of Tinto's Theory*. Paper presented at the 40th Annual Forum of the Association for Institutional Research Cincinnati, Ohio, May 21 - 24, 2000.

Selected References

- Lourens & Smit 2003, *Retention: Predicting first-year success*. South African Journal for Higher Education Vol.17 (2) p169 - p176.
- Murray, M. 2014. *Factors affecting graduation and student dropout rates at the University of KwaZulu-Natal*. South African Journal of Science. 2014; 110(11/12), Art.
- Ng & Jordan, 2002, *On discriminative vs generative classifiers: A comparison of logistic regression and naïve Bayes*. University of California, Berkeley.
- Pascarella & Terenzini 1983, *Predicting voluntary freshman year persistence/withdrawal behaviour in a residential university: a path analytic validation of Tinto's model*. Journal of Educational Psychology. 75(2), pp 215-226.
- Pike et al. 2014, *NSSE benchmarks and institutional outcomes: A note on the importance of considering the intended uses of a measure in validity studies*. Research in Higher Education, 54, 149-170.
- Scott et al. 2007, *A case for improving teaching and learning in South African Higher Education*. Higher Education Monitor No 6. Pretoria: CHE 2007.
- Tinto 1975, *Dropout from Higher Education: A theoretical synthesis of recent research*. Review of Educational Research 45 pp 89-125.
- van Zyl et al. 2012, *To what extent do pre-entry attributes predict first year student academic performance in the South African context?* South African Journal of Higher Education. Vol 18 (1), pp. 1095-1111.
- Woodhead 2002, *The Standards of Today and How to Raise Them to the Standards of Tomorrow*. London, UK: Adam Smith Institute.