ADSA H6011: Data Pre-Processing and Exploration

Short Title:	Data Pre-Processing and Exploration APPROVED
Full Title:	Data Pre-Processing and Exploration

Module Code:	ADSA H6011

ECTS credits:	10

NFQ Level:	9

Module Delivered in	1 programme(s)

Module Contributor:	Geraldine Gray

Module Description:	This module aims to investigate the properties of data, how to visualise data, and how pre-processing can enhance the information content of data.

Learning Outcomes:

On successful completion of this module the learner will be able to

1. Discuss in depth a variety of data preparation techniques, and their applicability to various problem domains
2. Research current trends in data visualisation, and select the appropriate graphical representation for data and results
3. Understand the links between data and necessary pre-processing algorithms to improve as well as prepare the data for modelling purposes
4. Evaluate appropriate techniques to improve data quality, and be aware of the limitations of such techniques
5. Analyse a data set to assess what data preparation is required to both clean the data set and expose its information content
6. Highlight information content using data visualisation techniques
7. Independently research current trends and developments in data preparation techniques

Module Content & Assessment

Indicative Content
Data Exploration Summary statistics, probability and random variables, data distributions, bias, correlation
Data Pre-processing Data selection: data characteristics and quality, sampling variability, assessing if sample is representative. Data cleaning: filling missing values, errors and outliers, data inconsistencies. Data transformations: aggregation, sampling, attribute construction, discretisation, scaling. Dimensionality reduction: attribute weighting, principal component analysis. Outlier detection: density based outlier detection, distance based outlier detection. Working with time series data
Data Visualisation Explore visual design techniques in terms of graphical representation types (e.g. perceptual grouping, columns, grids, volumes, relations, n-dimensional data), application of colour in visualisation, transparency, etc. Appropriate visualisation techniques for various data types from simple elements to more complex data relations Graph interpretation and manipulation

Indicative Assessment Breakdown	%
Course Work Assessment %	100.00%

Course Work Assessment %

Assessment Type	*Assessment Description*	*Outcome addressed*	*% of total*	*Assessment Date*
Reflective Journal	Students must prepare a portfolio of literary reviews and analysis covering a range of topics across all areas of the syllabus. Topics to include current trends in data exploration, data cleaning, data pre-processing and data visualisation.	1,2,3,4,7	50.00	n/a
Practical/Skills Evaluation	Work on a project to apply appropriate data exploration and pre-processing techniques on a dataset selected by the student. A project report is required outlining all aspects of the project and the steps undertaken to improve the dataset for modelling purposes.	3,4,5,6	50.00	n/a

No Final Exam Assessment %

Indicative Reassessment Requirement
Coursework Only *This module is reassessed solely on the basis of re-submitted coursework. There is no repeat written examination.*

ITB reserves the right to alter the nature and timings of assessment

Indicative Module Workload & Resources

Indicative Workload: Full Time	
Frequency	**Indicative Average Weekly Learner Workload**
Every Week	2.00
Every Week	2.00
Every Week	6.00

Resources

Recommended Book Resources

Salvador García, Julián Luengo, Francisco Herre 2014, *Data Preprocessing in Data Mining*, Springer [ISBN: 978331910246]

Markus Hofmann, Ralf Klinkenberg 2013, *Rapidminer: Data Mining Use Cases and Business Analytics Applications*, Chapman and Hall/CRC

Supplementary Book Resources

Jiawei Han, Micheline Kamber, Jian Pei 2011, *Data Mining: Concepts and Techniques, Third Edition*, Morgan Kaufmann [ISBN: 0123814790]

Myatt, G. J., Johnson, W. P. 2014, *Making Sense of Data*, 2nd Ed., Wiley [ISBN: 1118407415]

Alan Hashimoto, Mike Clayton,, *Visual Design Fundamentals* [ISBN: 1584505818]

Roxy Peck 2011, *Statistics*, Cengage Learning [ISBN: 0495552992]

This module does not have any article/paper resources

Other Resources

Journal: Science Direct*Computational Statistics and Data Analysis*

Journal: IEEE*Transactions on Knowledge and Data Engineering*

Module Delivered in

Programme Code	Programme	Semester	Delivery
BN_KADSA_R	Master of Science in Computing in Applied Data Science & Analytics	2	Mandatory