SCHOOL OF COMPUTING AND ENGINEERING SCIENCES

DATA MINING, STORAGE AND RETRIEVAL

COURSE OUTLINE

CODE: DSA 8102

Lecturer :	Dickson Owuor	**Phone :**	0728 558 811
Email:	dowuor@strathmore.edu		
Consultation:	Thu-Fri @5:30PM – 8:30PM	**Module Leader:**	

Purpose:

To develop an understanding of data mining principles, methods, implementation techniques applications. The course will also explore efficient storage and retrieval of data for analysis and data discovery.

Intended Learning Outcomes:

At the end of this course, the learner should be able to:

(1) Apply data mining techniques on data using R or Python.

(2) Use engaging, real-world examples to build a theoretical and practical understanding of key data mining methods.

(3) Develop predictive models for classification and prediction.

(4) Discuss data mining concepts over structured and unstructured data with special emphasis on practical applications of this important research area.

Contact Hours: 45

Prerequisite: Basic knowledge of R/Python language

Lecture time: 5:30PM – 8:30PM

Content, Outcomes and Activities

Data mining concepts and algorithms; mining data streams; data extraction and discovery from raw data; data Feature selection and data cleaning; exploring concepts of data discovery and efficient scalable pattern discovery methods, pattern-based classification and its applications; data exploration to enable forecasting and prediction; databases and data warehouses; data discretization methods and variable association rules; rule algorithms and their hybrids; indexing and query languages, data compression multimedia storage and retrieval, high dimensional data modelling and query processing; retrieval models; dictionaries, Term vocabulary and postings lists; vector space models and classifications.

Week	**Topic**	**Intended Learning Outcomes**	**Learning Activities**	**Assessment**
Week 1	Welcome	At the end of this topic, students will know: - Mode of course delivery, - topics, exercises and tutorials to be covered, - expectations.	- Welcome exchanges - Expectations	
Week 2	Data mining concepts and algorithms	At the end of this topic, students will be able to: - determine why data mining is in high demand, - describe the data mining algorithms and areas of applications, - explore various data mining technologies, and discuss issues in data mining.	- Introduction to R	- Practice questions
Week 3	Mining data streams	At the end of this topic, students will be able to: - describe the data stream concept and data stream sources and application areas, - describe the streaming model, sliding window and lossy count algorithm.	- Data streams with R	- R tutorial
Week 4	Data extraction	At the end of this topic, students will be able to: - describe data extraction processes and specifications, - describe the importance of data extraction within the ETL process, and - compare the staging and checkpoint restart logic.	- Web scrapping with R	- R tutorial

Week 5	Data cleansing	At the end of this topic, students will be able to: - discuss data quality concept, - describe data preprocessing, - discuss data cleansing and explain causes of errors, - describe various data cleansing techniques	- Introduction to Python - Introduction to Jupyter app - Data cleansing with Python	- Jupyter tutorial - Assignment submission
Week 6	**Review and CAT 1**			
Week 7	Feature selection	At the end of this topic, students will be able to: - describe feature selection concept, - discuss various techniques for feature selection.	- Feature selection with Python	- Jupyter tutorial
Week 8	Feature transformation	At the end of this topic, students will be able to: - describe feature transformation concept, - discuss the feature engineering concept, and - describe common methods of variable transformation.	- Feature transformation with Python	- Jupyter tutorial
Week 9	Data mining	At the end of this topic, students will be able to: - describe key pattern mining concepts, - examine sequential pattern mining, constraint-based mining, and - describe key classification and clustering algorithms.	- Data mining with R	- R tutorial - Assignment submission
Week 10	Databases and query languages	At the end of this topic, students will be able to: - discuss key database concepts, - explain the entity-relational models, - describe the SQL commands concepts, and - demonstrate how to write correct SQL queries.	- E-R modelling with R	- R tutorial

Week 11	Data warehouses and multidimensional data modelling	At the end of this topic, students will be able to: - explain the data warehouse concept, - identify the differences between operational database systems and data warehouses, - explain the data warehouse models, - determine the models for high-dimensional query processing, - describe the OLAP data modelling, - describe various indexing methods for data warehouses.	- Multi-dimensional data modelling with Python - Binary tree indexing exercise	- Jupyter tutorial
Week 12	**Review and Assignment/Exercise**			
Week 13	Dictionaries and postings	At the end of this topic, students will be able to: - describe the basic indexing concept, - demonstrate how to preprocess to form the term vocabulary in documents, tokenize and show what terms to put in an index, - describe phrase queries and positional postings, and	- Information retrieval with Python	- Jupyter tutorial
Week 14	Information retrieval models	At the end of this topic, students will be able to: - describe the different retrieval models, - distinguish the probabilistic approaches, and - describe the language models.	- Vector space modelling	- Computing vector space models
Week 15	**Revision & Final Exam**			

Course Delivery Methodology

1. Lectures will be used to introduce material on the formal aspects of the unit. Course materials will be made available digitally.
2. Class discussions, group work, video, online class group discussions, lab sessions

Academic Assessment

Type	**Weighting (%)**
Examination	**60**
Coursework	
CAT 1	10
Assignments	10
Group work	10
Exercises	10
Total	**100 %**

Core Reading Materials

1) Gazi, B. (2010). Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski and Lukasz A. Kurgan. *Data Mining: A Knowledge Discovery Approach*. Springer ((2007)). ISBN: 978-0387333335. £ 55.99. 606 pp. Hardcover. The Computer Journal, 53(4), 489-490.
2) Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
3) Manning, C. D., Raghavan, P., & Schütze, H. (2008). Boolean retrieval. *Introduction to information retrieval*, 1-18.

Further Reading

1) Pang-Ning Tan, Michael Steinbach and Vipin Kumar, (2014). *Introduction to Data Mining*. Pearson Education Limited, Essex.
2) Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.
3) Piegorsch, W. W. (2015). *Statistical data analytics: foundations for data mining, informatics, and knowledge discovery*. John Wiley & Sons.
4) Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
5) G Casella and RL Berger, *Statistical Inference; 2nd Edition*; Cengage Learning (2001); ISBN: 0534243126, 978-0534243128.
6) Sinai, Y. G. (2013). *Probability theory: an introductory course*. Springer Science & Business Media.

Video(s)

1) Simple Linear Regression in SAS. [ONLINE] Available at: https://www.youtube.com/watch?v=GKJREoop13s&t=459s. [Accessed 08 May 2021].
2) A Feature Selection Approach with Ensembles in SAS Enterprise Miner. [ONLINE] Available at: https://www.youtube.com/watch?v=jeZX31XsmFs&list=PLVV6eZFA22Qzg_1teSQ77qRRhprXg9EiU. [Accessed 08 May 2021].
3) Data Mining with R and SAS enterprise miner. [ONLINE] Available at: https://www.youtube.com/watch?v=fx8q2HJDPSk. [Accessed 08 May 2021].

Assignments/Homework

Students will be given assignments during the semester most of which will be part of the semester project which counts as part of coursework evaluation.

Location and Time

Please refer to the official timetable. Lecture rooms may change from time to time and it is important to keep track of communication from the module leader.

Policies

Punctuality is fundamental. Active **participation** in class discussions is required. It is prohibited to sign the attendance register on behalf of colleagues who are not present. **Plagiarism** is a serious offence. If detected in any form in course work and assignments, the following will apply:

a. In partial or non-serious cases (such as not citing whole word-for-word quotes), half the total possible marks of the assignment are duly struck off.
b. In serious cases (such as whole duplication of a paper), a zero policy will apply i.e., all offending assignments will be awarded a mark of zero.

Note: The level of seriousness referred to above is at the discretion of the lecturer.

Appeals are certainly possible through the relevant channels

Communication Channel

E-mail, phone, Module Leader