The University of Texas at El Paso
Industrial Data Analytics
Spring 2019 – Course Syllabus

Professor : Dr. Jose F. Espiritu
e-mail : jfespiritu@utep.edu

Class meets (Education Building 112): Mondays 6:00 pm – 8:50 pm

Office hours (A240): Tuesdays and Thursdays 10:00 am – 11:30 am and by appointment.

Course web page:
https://blackboard.utep.edu/

Course Motivation:
Data mining (Data analytics), also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. Data mining has been applied in a great number of fields, including retail sales, bioinformatics, and counter-terrorism. Data analytics is an important, fast-growing field that has quickly become a key basis of productivity growth, innovation, and consumer surplus. There is an increasing need for analytics-savvy employees who can think uniquely *across* disciplines to transform data into relevant insights for making better business decisions. Briefly, data analytics is computationally intelligent extraction of interesting, useful and previously unknown knowledge from large databases. It is a highly *inter-disciplinary area* representing the confluence of machine learning, statistics, operations research, database systems and high-performance computing, among other fields.

Course Description:
This course is an introductory course open to both graduate and senior undergraduate students. As an introductory course on data mining, this course introduces the concepts, algorithms, techniques, and systems of data mining as well as an introduction to the concept of Internet of Things, including (1) an introduction to the Internet of Things and data analytics, (2) data preprocessing, (3) mining frequent patterns and association, (4) classification, (5) cluster analysis, and (6) learning about software used in data mining and (7) demonstration of how to apply data analytics techniques using R. The course will provide students with basic understanding of common data mining and analytics techniques

Prerequisites:
Basic engineering statistics and basic calculus and linear algebra. However, this is not a course of mathematical statistics or advanced calculus.

Course objectives:
- Be capable of confidently applying common data analytic algorithms in practice;
- Be capable of performing data mining experiments using R-software;
- Be capable of performing experiments in using real-world data.
- Understand the importance of Data Analytics
- Students will learn about fundamental data analytic techniques and tools used analyze data

Learning Outcomes:

At the end of this course, students should be able to

- Understand modern views in data analysis and the Internet of Things concept;
- Explain the data mining methodology.
- Use visual techniques to describe data.
- Explain the assumptions of K-Means Clustering.
- Segment data using Cluster Analysis, and interpret the output
- Explain the assumptions of various techniques such as Cluster Analysis, Multiple Regression, Discriminant Analysis, Logistic Regression, Decision trees

Useful references:

- Data Mining: Concepts and Techniques, 3rd edition, written by Jiawei Han, Micheline Kamber, Jian Pei.
- An Introduction to Statistical Learning, written by Daniela Witten, Trevor Hastie, Robert Tibshirani
- An Introduction to R by W. N. Venables, D. M. Smith and the R Core Team
- Charu C. Aggarwal, Data Mining: The Textbook, Springer, 2015
- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005 (2nd ed. 2016)
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005
- Mohammed J. Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms 2014.

Grading

ITEM	**Percentage**
Case study/Journal Presentation	10%
Class attendance and participation	5%
Lab participation	10%
Homework assignments	10%
Project update 1, presentation	5%
Final project presentation	10%
Final Project Report	10%
Exam 1	20%
Exam 2	20%

1, Journal critique and presentation:

Each team of 3 students (***individual assignment for graduate students***) is required to present a journal article or case study, and prepare a 12-minute presentation. The team must understand and explain the model being presented to solve the problem addressed. Sample application areas are

- Agriculture
- Manufacturing
- Homeland Security
- Healthcare
- Finance
- Transportation
- Energy
- Environment, etc

2. Class attendance and Participation:

Attendance to all sessions is strongly recommended. Students are responsible for all of the material covered in the class. Participation in class discussions and activates are part of your grade in this course. You are expected to be on time and to participate you must attend class having prepared the assigned materials for the day.

3, Homework policy:

Homeworks will be announced in class. Homework will be due at the beginning of the class. Late homework or reports will not be accepted, unless certified medical proof is given. If you are unable to attend the class at which the homework is due, it is your responsibility to submit it earlier.

4. Project Description

The purpose of the term project is to provide an opportunity to apply and/or further explore a topic area related to the course. Some examples of types of projects that are acceptable are:

- Literature review of data mining methods used in a problem area that demonstrates breath and organization of the literature
- Research proposal incorporating engineering analysis that demonstrates creativity and originality
- Apply different models to a specific problem and compare results (interest is in models not covered in class, maybe use some from the class and present a new one)
- Further exploration and case study presentation of the Internet of Things as applied to different fields

Each team is required to identify a topic of interest and submit a one page abstract by Feb/18/19. It is expected that the group will review recently published literature on the topic of interest. The project will be evaluated based on the technical content (30%), quality of documented references (40%) and quality of technical writing (40%). It is recommended that the students review the following journals for selecting a topic:

- ACM Transactions on Knowledge Discovery in Data (TKDD).

- SIGKDD Explorations, a magazine of the SIGKDD, the data miners professional group.
- Data Mining and Knowledge Discovery journal (now published by Springer).
- Analytics magazine from INFORMS.
- Big Data, open access peer-reviewed journal, provides a forum for world-class research exploring the challenges and opportunities in collecting, analyzing, and disseminating vast amounts of data. Liebert Publishers.
- Case Studies In Business, Industry And Government Statistics, electronic journal, Bentley University.
- Chance, a quarterly magazine for people interested in analysis of data, from American Statistical Association and Springer.
- Data Science Journal, published by the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU).
- EPJ Data Science Journal, SpringerOpen.
- IEEE Transactions on Knowledge and Data Engineering
- Intelligent Data Analysis journal (IOS Press).
- International Journal of Data Mining and Bioinformatics (IJDMB), ISSN (Online): 1748-5681 - ISSN (Print): 1748-5673
- Journal Of Big Data, a SpringerOpen Journal.
- Journal of Data Mining and Knowledge Discovery, tri-monthly, ISSN: 2229–6662 , 2229–6670, Bioinfo publications, India.
- Journal of Data Science, an international journal devoted to applications of statistical methods at large.
- Journal of Intelligent Information Systems.
- Journal of Machine Learning Research
- KAIS: Knowledge and Information Systems: An International Journal (Springer-Verlag)
- Machine Learning and Machine Learning Online
- Michael Ley comprehensive list of Computer Science and Database Journals
- Predictive Modeling News, the montly newsletter for healthcare professionals involved with predictive modeling.
- Statistical Analysis and Data Mining, Wiley journal, editors: Arnold Goodman, Chandrika Kamath, Vipin Kumar
- Transactions on Machine Learning and Data Mining, IBAI journal, Editor: Petra Perner, ISSN: 1865-6781.

Final project report format: (written in word, 1.5 spacing, Times New Roman 11, 20-25pages) due (May/13/19)

Possible outline for project reports

Prepare your final report for an interdisciplinary group of decision makers. Your written reports must be formatted into different section headings (shown below)

1. **Introductory Pages**
   - Cover Page: title, authors
   - Table of Contents, List of Tables, and List of Figures
2. **Executive Summary** (brief summary of the overall project)
3. **Introduction**

- Mention the importance of the problem/area being studied
- Ideally provide specific information from reports related to the problem and its significance (Cite at least 15 references)
- Explain benefits of solving the problem, including motivation economics
- Present a clear and concise definition of the problem

4. **Literature Review**
   - Present an up-to date literature review from Journal publications and conference proceedings
   - Write a small paragraph from each article indicating what the authors did to solve the problem presented in the publication
   - End the literature review section mentioning a specific idea about how you may expand the work presented by them
5. **Model Development**
   - Mention the data mining model/algorithm which will be for the analysis of the problem indicated in the previous section
6. **Case study**
   - Clearly state the problem where the algorithm will be applied
   - Apply the data mining algorithm(s)
7. **Conclusions and Future Work**
   - Write a paragraph mentioning the final conclusions for the proposed work
   - Mention final recommendations and Future work
8. **Bibliography**

Students with disabilities:
If you have or suspect a disability and need accommodations you should contact Disabled Student Services Office (DSSO) at 747-5148 or at dss@utep.edu or come by Room 106 Union East Building.

Cell phone policy- In the past I have left it up to students to be courteous to the class and use their own discretion with cell phones and pagers. However, there are simply too many disturbances and distractions caused by a few cell phone abusers. Cell phones and pagers are to be off during class. You are not to take calls during class. I find that leaving and returning during class is very disturbing and rude. If you are in a group discussion it is especially rude. Exceptions would be emergency situations which you can notify me about before class.

Software

A) R

Over the recent years, R has become the leading software tool for statistical computing and graphics. The software is greatly enhanced by numerous contributed packages submitted by users. The majority of computing in the leading applied statistical journals is done in R, and it is used almost exclusively in some of the leading-edge applications, such as in genetics and data mining. The purpose of this course is to set a foundation for using of the statistical language for computing and graphics R.

The course will introduce students to the syntax and inner workings of R, to become proficient in everyday computational tasks with datasets, skilled in applications of elementary data minig methods, with an emphasis on (initial) data exploration and simple graphics R.

B) Matlab

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment. A proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python.

Academic Honesty

It is expected that the students will conduct with integrity in all course areas. Do not attempt to engage in a dishonest activity such as copying, plagiarism, falsifying information, etc. The professor will take measures to prevent such instances and will bring a case to the university authorities. Information about University wide policies could be found in the Dean of Students Web page at http://studentaffairs.utep.edu/Default.aspx?alias=studentaffairs.utep.edu/dos

Tentative Course Schedule (it may change, based on feedback or progress):

Lecture #	Date	Topics
	01/21	Dr. Martin Luther King, Jr. Holiday, No Class
1	01/28	Class syllabus and Introduction to data analytics, Internet of Things
2	02/04	Internet of Things – Lab practice and assignment
3	02/11	Getting to know your data, Data preprocessing and Journal paper selection
4	02/18	Introduction to R and Lab examples - Linear regression Final project proposal selection due
5	02/25	Linear regression
6	03/04	Regression additional topics
7	03/11	Exam 1
	03/18	Spring Break, No Class
8	03/25	Clustering
9	04/01	Project update 1- Presentation
10	04/08	Clustering
11	04/15	Classification
12	04/22	Classification
13	04/29	Neural Networks
14	05/06	Exam 2 - Additional models
15	05/13	Final Project presentations and report due