PROBABILITY & STATISTICS FOR DATA MANAGEMENT AND ANALYSIS

Bachelor in Data and Business Analytics BDBA SEP-2023
PSDMA-DBA.2.M.A

Area Operations and Business Analytics
Number of sessions: 35
Academic year: 23-24
Degree course: SECOND
Number of credits: 6.0
Semester: 1º
Category: BASIC
Language: English

Professor: **RAFIF SROUR DAHER**

E-mail: rsrour@faculty.ie.edu

Experience in teaching, analytical and empirical research, and data analysis. Demonstrated ability to work in international/multicultural environments (Lebanon, USA, Spain). Life-long learner; both academically and personally. Advocate of women in STEM empowerment - Breaking stereotypes one at time. Multiple times winner of Best Professor award and strong advocate of using coaching and mentoring to help students improve their academic performance and overall college experience. Lately, nominated among the top 183 Leading Data Academics of 2021 by CDO magazine. Also nominated among 55 leading women in the #technology sector in Spain, under the category of "Yo, Jefa", 2021. Driven by a passion to use technology as a disrupter in higher education, constantly working on innovating curricula and teaching methodologies.

Experience

- Executive Vice Dean, IE School of Science and Technology, May 23 - Present.
- Vice Dean of Undergraduate Programs, IE School of Science and Technology, Jun 2022 - Present.
- Acting Dean, IE School of Science and Technology.

  Academic Director of BSc in Data and Business Analytics, School of Science and Technology, Dec 2017 - May 2022.
- Adjunct Faculty, IE university, Sep 2014 - Present.

  rsrour@faculty.ie.edu; before or after class or by appointment.

  Office: Tower, 4.19

SUBJECT DESCRIPTION

Nowadays, all companies are striving to become more profitable, to reach customers quicker, and to offer higher-quality products and services. In addition, businesses want to reach these objectives with less fewer and at lower costs. It seems like an impossible task, but it is not. An essential requirement in this process is effective knowledge creation and management. There is no lack of information, but there is a dearth of knowledge. As Tom Peters said in his book Thriving on Chaos, "we are drowning in information and starved for knowledge".

In recent year, there has been an exponential increase in the amount of information available for decision making. Big data, as it is commonly known is currently being collected and stored in data warehouses, ready to be mined for actionable insights. Some of that data can be analyzed and understood with simple statistics, but much of it requires more complex, multivariate statistical techniques to convert into knowledge. Even more, some data requires to use new approaches such as the Bayesian statistics to transform it into usable data.

This course is divided into two main sections: Exploratory data analysis and multivariate analysis, both dependent and independent techniques.

LEARNING OBJECTIVES

The main objective of this course is to explore, in depth, both from a practical and theoretical standpoints, a number of advanced statistical techniques needed for data analysis and management. Each dataset is unique and necessitates a special treatment. Understanding the nature of the dataset, its underlying structure and distribution, the objectives for which it has been collected are essential in determining the appropriate multivariate analysis tool.

At the end of the course; students should be able to:

- Explain what multivariate analysis is and when its applications are appropriate;
- Understand the six-step approach to multivariate model building;
- Determine which multivariate technique is suitable and for what purpose;
- Study and understand the theoretical as well as the practical aspects behind some common multivariable analysis techniques including; principle component analysis, multivariable regression, logistics regression, cluster analysis, decision trees, among others.
- Apply learned concepts to specific case study, using Python as programming tool;
- Analyze new multivariate statistical techniques through hands-on project;
- Share their knowledge with their classmates as well as with the IE community (professors, master students, students from other degrees) by presenting their group projects as part of an established initiative at the Bachelor of Data and Business Analytics: The Speaker Series.

TEACHING METHODOLOGY

IE University teaching method is defined by its collaborative, active, and applied nature. Students actively participate in the whole process to build their knowledge and sharpen their skills. Professor's main role is to lead and guide students to achieve the learning objectives of the course. In this course, we will use both machine learning and AI technology to improve the learning process and the overall student experience. This is done by engaging in a diverse range of teaching techniques and different types of learning activities such as the following:

Learning Activity	**Weighting**	**Estimated time a student should dedicate to prepare for and participate in**
Lectures	33.33 %	50.0 hours
Discussions	10.0 %	15.0 hours

Exercises in class, Asynchronous sessions, Field Work	13.33 %	20.0 hours
Group work	16.67 %	25.0 hours
Individual studying	26.67 %	40.0 hours
TOTAL	100.0 %	150.0 hours

PROGRAM

The theoretical content of this course consists of two main parts and each part is divided into several modules. The first part covers data assessment and evaluation with special emphasis on multivariate profiling and data quality issues. The second part dive into multivariable analysis techniques such as principle component analysis, multiple regression, LASSO and RIDGE regression, multiple discriminant analysis, cluster analysis and decision trees.

In this context, the course is divided into 7 modules, each module comprises 5 sessions:

- Module 1: INTRODUCTION TO MULTIVARIABLE STATISTICAL ANALYSIS
- Module 2: DATA ASSESSMENT AND EVALUATION
- Module 3: PRINCIPLE COMPONENT ANALYSIS AND CLUSTER ANALYSIS
- Module 4: MULTIPLE REGRESSION ANALYSIS
- Module 5: DECISION TREES
- Module 6: MULTIPLE DISCRIMINANT ANALYSIS
- Module 7: ASSESSMENT, CONSOLIDATION, AND BEYOND!

All the required readings are from the compulsory textbook "*Multivariate data Analysis", Hair, Black, Babin & Anderson, Pearson New International Edition*. Extra handouts will be provided to students to cover specific topics.

Disclaimer: The following description of the material covered is tentative. An attempt will be made to cover all listed topics. However; the pace in the class depends on the group performance.

Important: In the next section, "*Multivariate data Analysis*" textbook is referred to by the ACRONYM MDA.

MODULE 1/7: INTRODUCTION TO MULTIVARIABLE STATISTICAL ANALYSIS

Sessions 1 – 5

SESSION 1 (LIVE IN-PERSON)

Topics: Introduction and presentation of the course syllabus, objectives and pedagogy. Introduction of module 1 topics and deliverables. Multivariate analyses techniques: an overview. Statistical significance versus statistical power.

Book Chapters: Multivariate Data Analysis; Chapter 1; pp 1 – 11 (See Bibliography)

SESSION 2 (LIVE IN-PERSON)

Topics: Guidelines for selection and interpretation of multivariate techniques. Six-step approach to multivariate model building. Hands-on activity: the HBAT dataset.

INTRODUCING THE GROUP PROJECT.

Book Chapters: Multivariate Data Analysis; Chapter 1; pp 11 – 24 (See Bibliography)

SESSION 3 (LIVE IN-PERSON)

Discussion plus activity.

Activity 1. Model-building approach to multivariate analysis; Approximate time of completion: < 50 min; type of work: individual.

Learning objective: This activity aims at testing your capacity to define each step within a multivariate model and to identify the major issues within each step.

Description: You will be provided with a model schema for multivariable analysis. Define each stage and discuss one or two of its major limitations. Details regarding the platform that will be used to submit your work will be provided in due time.

Please limit your words to the space provided.

Evaluation: Continuous evaluation. Total points: 10 points.

SESSION 4 (LIVE IN-PERSON)

Topics: Understanding data: Graphical presentation. Missing data: Types and impact.

Book Chapters: Multivariate Data Analysis; Chapter 2, pp 34 – 42 (See Bibliography)

SESSION 5 (LIVE IN-PERSON)

Topics: Identifying Missing Data.

Quiz 1. Total points: 10 points.

MODULE 2/7: DATA ASSESSMENT AND EVALUATION

Sessions 6 – 10

SESSION 6 (LIVE IN-PERSON)

Topics: Introducing module 2, topics, and deliverables. Outliers.

Graded Assignment 1. Introducing the assignment and setting deadlines: SESSION 16.

Book Chapters: Multivariate Data Analysis; Chapter 2, pp 62 – 68 (See Bibliography)

SESSION 7 (LIVE IN-PERSON)

Topics: Identify the most common types of assumptions involved in major multivariable techniques. Learn both graphical and statistical methods of testing assumptions.

Book Chapters: Multivariate Data Analysis; Chapter 2, pp 68 – 75 (See Bibliography)

SESSION 8 (ASYNCHRONOUS)

Practical: Data Assessment and Evaluation: Part 1.

SESSION 9 (LIVE IN-PERSON)

Topics: Learn how to transform data to meet build-in model assumptions.

Book Chapters: Multivariate Data Analysis; Chapter 2, pp 75 – 84 (See Bibliography)

SESSION 10 (ASYNCHRONOUS)

Practical: ***Data Assessment and Evaluation: Part 2.***

MODULE 3/7: PRINCIPLE COMPONENT ANALYSIS AND CLUSTER ANALYSIS

Sessions 11 – 15

SESSION 11 (LIVE IN-PERSON)

Topics: Introducing module 3, topics, objectives and deliverables. Exploratory factor analysis.

Book Chapters: Multivariate Data Analysis; Chapter 3, pp 89 – 104 (See Bibliography)

SESSION 12 (LIVE IN-PERSON)

Topics: Principle Component Analysis (Handout).

Introducing asynchronous activities for session 13.

SESSION 13 (ASYNCHRONOUS)

Practice: *Principle Component Analysis.*

Dataset: Crime.

SESSION 14 (LIVE IN-PERSON)

Topics: Cluster analysis: Hierarchical methods versus *K* means. Introducing asynchronous activities for session 15.

SESSION 15 (ASYNCHRONOUS)

Practice: *Cluster Analysis.*

Dataset: Children.

MODULE 4/7: MULTIPLE REGRESSION ANALYSIS

Sessions 16 – 21

SESSION 16 (LIVE IN-PERSON)

Topics: Introducing module 4, content and objectives. Multivariable linear regression: Objectives, Design and Assumptions.

Activity: Guided discussion on the assumptions involved in the multivariable linear regression: Normality, Linearity, Homoscedasticity and independence of the error terms.

Deadline: submit assignment 1.

Book Chapters: Multivariate Data Analysis; Chapter 4, pp 163 – 172; 177 – 182 (See Bibliography)

SESSION 17 (LIVE IN-PERSON)

Topics: Multivariable linear regression: Model selection, interpretation and validation. Introducing the case study.

Book Chapters: Multivariate Data Analysis; Chapter 4, pp 182 – 190 and pp 202 – 227 (See Bibliography)

SESSION 18 (LIVE IN-PERSON)

Topics: The multivariable linear regression – The model; Creating Additional Variables: Transformations to meet assumptions; Influential Observations: Assessments and Remedies.

Learning Objectives:

- Learn the various techniques for modeling multivariable linear regression.
- Stress the importance of assumptions in model building; introduce the concept of dummy variables for use of nonmetric variables, teach students what kind of transformation is needed for curvilinear relationships, etc.
- Re-inforce the concepts developed in class; identify various types of influential observations and choose the best remedy.

SESSION 19 (LIVE IN-PERSON)

Q & A session. Meet Thiago - The Robot.

Analyzing the case study, comparison with lasso and ridge.

SESSION 20 (LIVE IN-PERSON)

Prior to lecture: Read case study notes. Watch the screencast and practice.

In class Practice: Generative models and coding.

Topic: Multiple Regression Analysis

SESSION 21 (LIVE IN-PERSON)

Review of Assignment 1.

MODULE 5/7: DECISION TREES

Sessions 22 – 25

SESSION 22 (LIVE IN-PERSON)

Topics: Introducing module, objectives and deliverables. Introducing NeoGrocer Casebook: Retaining valuable customers.

SESSION 23 (LIVE IN-PERSON)

Topics: NeoGrocer casebook. Linear regression: Lasso and Ridge.

SESSION 24 (LIVE IN-PERSON)

Topics: Non-linear regression: feature space, decision trees. training a decision tree.

SESSION 25 (ASYNCHRONOUS)

Topics: Training of decision trees. Practice.

MODULE 6/7: DISCRIMINANT ANALYSIS

Sessions 26 – 30

SESSION 26 (LIVE IN-PERSON)

Topics: Introducing module 6, topics and deliverables. Logistic regression: When to use? Likelihood and probability measures. Logistic model: Objectives, Design and assumptions.

ASSIGNMENT 2.

Book Chapters: Multivariate Data Analysis; Chapter 6, pp 322 – 338 (See Bibliography)

SESSION 27 (LIVE IN-PERSON)

Topics: Logistic regression: model estimation, interpretation and validation.

Book Chapters: Multivariate Data Analysis; Chapter 6, pp 322 – 338 (See Bibliography)

SESSION 28 (LIVE IN-PERSON)

Practice: *logistic Analysis.*

SESSION 29 (LIVE IN-PERSON)

Topics: Inquiry-based learning: Discriminant Analysis. group work

Objectives:

1. Define discriminant analysis and to compare it to Linear and Logistic regression as well as to MANOVA.
2. Understand the concepts behind discriminant analysis especially centroids, discriminant Z scores and the discriminant function.
3. Be able to differentiate between various types of discriminant analysis and understand the applications of each.

SESSION 30 (LIVE IN-PERSON)

Prior to class - Watch the following videos on Discriminant analysis: in class discussion of the content covered in video.

Other / Complementary Documentation: Discriminant Analysis (Youtube)

SESSION 31 (LIVE IN-PERSON)

MODULE 7: ASSESSMENT, CONSOLIDATION, AND BEYOND!

This module is different from all the rest, both in content, teaching approach and learning outcomes. Students will be provided with discussion topics, hands on data sets, short exercises and puzzles to test their capacity to synthesize all what they have learned in the previous modules.

Learning objectives:

- Consolidate conceptual content;
- Discuss the importance of data assessment and evaluation;
- Summarize and evaluate multivariate techniques discussed in previous modules;
- Hands-on real-life problems;
- Evaluate and guide student project work.

SESSION 32 (LIVE IN-PERSON)

Speaker Series

SESSION 33 (LIVE IN-PERSON)

Speaker Series

SESSIONS 34 - 35 (LIVE IN-PERSON)

Final exam

EVALUATION CRITERIA

Your final grade in the course will be based on both individual and group work of different characteristics that will be weighted in the following way:

criteria	percentage	Learning Objectives	Comments
Final Exam	25 %		
Quizzes	25 %		
Individual Work	15 %		
Workgroups	25 %		
Class Participation	10 %		

RE-SIT / RE-TAKE POLICY

A. Class participation

Class participation will be evaluated based on the following criteria:

- *Quality* (not quantity) of your participation in class discussion and forums: The most important dimension of participation concerns what it is that you are saying. A high quality comment reveals depth of insight, rigorous use of case evidence, consistency of argument, and realism. Frequency refers to the attainment of a threshold quantity of contributions that is sufficient for making a reliable assessment of comment quality. The logic is simple: if contributions are too few, one cannot reliably assess the quality of your remarks. However, once threshold quantity has been achieved, simply increasing the number of times you talk does not automatically improve your evaluation. Beyond the threshold, it is the quality of your comments that must improve. In particular, one must be especially careful that in claiming more than a fair share of “airtime”, quality is not sacrificed for quantity. Finally, your attempts at participation should not be such that the instructor has to “go looking for you”. You should be attempting to get into the debate on a regular basis.

You might want to avoid being classified as one of the following types of students:

- Repeaters, i.e., students that, consciously or unconsciously, make comments that are really just repeats/rephrasing of what has already been said (by other students, or you). This wastes time and adds nothing to learning.
- Ramblers, i.e., students that take a lot of time to say simple things or they may tell long personal/professional stories, or they roam into topics that are not relevant, or simply make low-quality comments just to participate. They waste valuable time and prevent other students from being able to participate.
- Students that have been distracted (by Facebook, etc.) or who have stopped paying attention and then, later on, when they realized they have missed a term or concept, they ask you about

it.

B. Group project_Speaker Series

The group project is an integral part of this course. Each group (randomly composed of 2 – 5 students) will be asked to choose an advanced topic in Statistics, to write a paper on that topic and to prepare a presentation. These presentations will be delivered as part of an initiative entitled: Speaker Series. This initiative is used to promote students communication and presentations skills in conveying complex concepts. Specifics of the group project including learning objectives, deliverables, evaluation and rubrics will be posted on a separate page for the group project on Blackboard ultra.

C. Individual Assignments

A total of 1 individual assignment will be given to the students throughout the semester. in this assignment, students will be given real datasets (Github database) and asked a series of questions in which they will need to use Python to solve. Detailed rubrics will be provided in due time.

D. Continuous Evaluation

In the new liquid learning environment, special emphasis is given to student overall performance and evolution throughout the semester. Several deliverables will be taken into account in grading the students (refer to the detailed session content above), but also the overall student quality of work and efforts will be considered.

E. Quizzes

Throughout the semester, students are given online-quizzes based on materials covered in class. These quizzes will help the students assess their overall understanding of the subject being studied, identify any caveat in their learning and address it with the professor in due time.

F. Final Exam

There will be one final exam. For these exams, you must bring your own simple calculator (phones, tablets, laptops and other electronic devices are not allowed). You are also allowed to bring up 3 one-sided A4 sheets to the final exam. The sheets can only contain formulae that you think could be helpful. NO QUESTIONS ARE ALLOWED DURING THE EXAMS. THE CHEAT-SHEET ALONG WITH ANY SCRAP PAPER WILL BE COLLECTED AND STAPLED TO YOUR EXAMS.

In order to pass the course, you need **a minimum grade of 3.5 in the final exam**. If your grade in the final exam does not reach the threshold value of 3.5, you will fail the course, even in the case in which your weighted average (computed using the table above) exceeds 5.0.

BIBLIOGRAPHY

Compulsory

- Joseph, F. Hair Jr, William C. Black, Barry J. Babin & Ralph E. Anderson. (2014). *Multivariate Data Analysis.* 7th edition. Pearson International Edition. Pearson Education Limited. ISBN 9781292021 (Printed)

BEHAVIOR RULES

Please, check the University's Code of Conduct here. The Program Director may provide further indications.

ATTENDANCE POLICY

Please, check the University's Attendance Policy here. The Program Director may

provide further indications.

ETHICAL POLICY

Please, check the University's Ethics Code here. The Program Director may provide further indications.