ISM UNIVERSITY OF MANAGEMENT AND ECONOMICS

Data Mining for Business Intelligence

Course code	*IT101*
Compulsory in the programmes	*Economics and Data Analytics*
Level of studies	*Undergraduate*
Number of credits and	*6 ECTS (48 contact hours + 6 consultation hours, 106 individual work hours)*
Course coordinator (title and name)	*Paulius Rauba*
Prerequisites	*Statistical Data Analysis, Mathematical Analysis, Computer Programming*
Language of instruction	*English*

THE AIM OF THE COURSE:

The rapidly increasing amount of data generated each year makes extracting useful information from that data ever more important. Companies are making use of various data techniques to answer business questions and power their decision-making. This data is often stored in data warehouses and databases which has to be extracted, pre-processed, and analyzed before it can be modeled using statistical techniques.

The goal of this course is to provide the necessary technical expertise for extracting and exploring data stored in databases as well as building and evaluating common statistical models. Specifically, the course covers four broad topics: (1) Extracting data from databases using SQL and performing exploratory data analysis; (2) Building data-based supervised statistical models to find a predictive function; (3) Evaluating the results of the models, and (4) Working with non-tabular data and unsupervised models. The main statistical methods covered include shrinkage methods (L1 and L2 regularization), maximum margin classifiers, tree-based bagging and boosting algorithms, and clustering methods. Students are also equipped with the tools to perform model evaluation using cross-validation approaches, bootstrapping for estimating uncertainty, utilizing common classification evaluation metrics (e.g. F1-Score, ROC-AUC), and extracting feature importance, among others.

The course heavily focuses on predictive modeling using Python. Foundational knowledge in statistics, mathematical analysis, and Python programming is assumed.

MAPPING OF COURSE-LEVEL LEARNING OUTCOMES (OBJECTIVES) WITH DEGREE LEVEL LEARNING OBJECTIVES (See Annex), ASSESSMENT AND TEACHING METHODS

Course level learning outcomes (objectives)	**Degree level learning objectives (Number of LO)**	**Assessment methods**	**Teaching methods**
CLO1. Understand the key elements of relational databases and data storage. Extract data from databases by writing SQL queries.	ELO1.1 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work
CLO2. Perform exploratory data analysis on tabular data using relevant Python packages. Visualize the relationship between variables.	ELO1.1 ELO3.1 ELO3.2 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work
CLO3. Understand the differences between supervised and	ELO1.1	Final exam, Mid-	Lectures,

unsupervised models, the bias-variance trade-off, and train-test splits. Critically evaluate which models are best suited for specific tasks.	ELO4.3	term exam, group project	seminars, independent work
CLO4. Understand and apply linear regression-based shrinkage methods for variable selection and addressing high-dimensional data problems.	ELO1.1 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work
CLO5. Understand and apply multiple classification models for modeling binary and non-binary response variables. Understand the concepts of decision boundaries and maximum margin classifiers.	ELO1.1 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work
CLO6. Understand and apply tree-based methods for both classification and regression problems. Understand the differences between bagging and boosting algorithms.	ELO1.1 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work
CLO7. Evaluate models for regression and classification problems. Understand the different choices of metrics available for imbalanced data. Extract key features from statistical models.	ELO1.1 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work
CLO8. Write clear, reproducible, and well-documented code in Python using the Jupyter Notebook environment. Be able to use the most relevant Python packages for data wrangling.	ELO1.1 ELO3.1 ELO3.2 ELO4.3	Final exam, Mid-term exam, group project	Lectures, seminars, independent work

ACADEMIC HONESTY AND INTEGRITY

The ISM University of Management and Economics Code of Ethics, including cheating and plagiarism, is fully applicable and will be strictly enforced in the course. Academic dishonesty and cheating can and will lead to a report to the ISM Committee of Ethics. With regard to remote learning, ISM reminds students that they are expected to adhere to and maintain the same academic honesty and integrity that they would in a classroom setting.

COURSE OUTLINE

Week	Topic	In-class hours	Readings
I. Extracting and Exploring Data			
1.	1. Introduction to Data Mining and SQL.	4	Malik, Goldwasser, & Johnston (2019), Ch2
2.	2. Exploratory data analysis.	4	Will be provided during the lectures
3.	3. Statistical learning: An introduction.	4	James et al. (2013), Ch2
II. Building Statistical Supervised Models			
4.	4. Linear regression models: Selection and regularization methods.	4	James et al. (2013), Ch3, Ch6
5.	5. Classification models: Linear Discriminant Analysis and Support Vector Machines.	4	James et al. (2013), Ch4, Ch9
6.	6. Tree-based methods and ensemble learning.	4	James et al. (2013), Ch8

7.	*Midterm exam*	2	
8.	7. Moving beyond linearity.	4	James et al. (2013), Ch7
III. Evaluating Statistical Supervised Models			
9.	8. Model evaluation and resampling methods.	4	James et al. (2013), Ch5
10.	9. Feature importance and A/B testing.	4	Will be provided during the lectures
IV. Moving Beyond Supervised Models and Tabular Data			
11.	10. Unsupervised learning.	4	James et al. (2013), Ch12
12.	11. Heterogeneous data.	4	James et al. (2013), Ch10
13.	Course review	2	
		Total: 48 hours	
	CONSULTATIONS	6	
	FINAL EXAM	2	

FINAL GRADE COMPOSITION

Type of assignment	**%**
Group Components 30%	
Group project	30%
Individual Components, 70%	
Homework	20%
Mid-term exam	25%
Final exam	25%
Total:	**100**

DESCRIPTION AND GRADING CRITERIA OF EACH ASSIGNMENT

Group project. In the group project, students will have to prepare a report detailing the analysis they have performed. The analysis should include (i) data extraction from a database, (ii) predictive classification of regression model(s), (iii) evaluation of the model(s), and (iv) presentation of the results through insights-driven, clear, and clean data visualization. The framework and instructions for the task will be provided by the lecturer during class. The group sizes are expected to be between 3-4 people. There will be one such project worth 30% of the final grade.

Homework. Students will be assigned homework tasks to be completed with Python. The homework tasks will include applying the theoretical knowledge gained during the class in the Jupyter Notebook environment. This will include completing tasks such

as performing exploratory data analysis, fitting supervised and unsupervised statistical models, fine-tuning the models and evaluating results. It counts towards 20% of the final grade.

Mid-term exam. The mid-term exam will be held during the midterm exam session. It counts towards 25% of the final grade. The midterm will be based on topics 1-6. The midterm will consist of theoretical questions, practical and coding problems.

Final exam. The final exam counts towards 25% of the final grade. The final exam includes multiple-choice questions and open questions. It tests conceptual, analytical, and numerical skills. The exam will be primarily based on topics 7-11; it might include questions from the previous chapters as well. The final examination will take place during the final examination session.

Retake exam. Students who receive a failing final grade shall have the right to the retake exam, which will comprise 50% of the final grade and cover all topics of the course. Midterm exam and final exam results will be annulled.

REQUIRED READINGS

James, Gareth, et al. (2013). An Introduction to Statistical Learning: with Applications in R. 2nd ed., Springer.

ADDITIONAL READINGS

Malik, U., Goldwasser, M., & Johnston, B. (2019). SQL for Data Analytics : Perform Fast and Efficient Data Analysis with the Power of SQL. Packt Publishing.

Jake VanderPlas (2016). Python Data Science Handbook: Essential Tools for Working with Data (1st. ed.). O'Reilly Media, Inc.

ANNEX

DEGREE LEVEL LEARNING OBJECTIVES

Learning objectives for the Bachelor of Business Management
Programmes:
International Business and Communication,
Business Management and Marketing,
Finance,
Industrial Technology Management,
Entrepreneurship and Innovation

Learning Goals	Learning Objectives
Students will be critical thinkers	BLO1.1. Students will be able to understand core concepts and methods in the business disciplines
	BLO1.2. Students will be able to conduct a contextual analysis to identify a problem associated with their discipline, to generate managerial options and propose viable solutions
Students will be socially responsible in their related discipline	BLO2.1. Students will be knowledgeable about ethics and social responsibility
Students will be technology agile	BLO3.1. Students will demonstrate proficiency in common business software packages
	BLO3.2. Students will be able to make decisions using appropriate IT tools
Students will be effective communicators	BLO4.1. Students will be able to communicate reasonably in different settings according to target audience tasks and situations
	BLO4.2. Students will be able to convey their ideas effectively through an oral presentation
	BLO4.3. Students will be able to convey their ideas effectively in a written paper

Learning objectives for the Bachelor of Social Science
Programmes:
Economics and Data Analytics,
Economics and Politics

Learning Goals	Learning Objectives
Students will be critical thinkers	ELO1.1. Students will be able to understand core concepts and methods in the key economics disciplines
	ELO1.2. Students will be able to identify underlying assumptions and logical consistency of causal statements
Students will have skills to employ economic thought for the common good	ELO2.1.Students will have a keen sense of ethical criteria for practical problem-solving
Students will be technology agile	ELO3.1. Students will demonstrate proficiency in common business software packages
	ELO3.2. Students will be able to make decisions using appropriate IT tools
Students will be effective communicators	ELO4.1.Students will be able to communicate reasonably in different settings according to target audience tasks and situations
	ELO4.2.Students will be able to convey their ideas effectively through an oral presentation
	ELO4.3. Students will be able to convey their ideas effectively in a written paper