DT332 APPLIED DATA SCIENCE
IES Abroad Madrid

DESCRIPTION:
Data science is the discipline that data scientists use to answer meaningful business questions using available data. The mastery of this subject is essential for managers who need to understand the insights that can be obtained from data, and for individuals who want to apply the scientific method to data.

In this course, students will start by exploring how to connect business and data. A particular focus of the course will be on the basics of applied data science, such as domain knowledge, visualization, data engineering, database management, and statistical modeling. The course's most salient goals will be to provide students with an in-depth understanding of the most effective ways to communicate data insights as they are created.

In particular, the course content covered during the semester will include Introduction to Python, Knowledge Discovery in Databases, ETL (Extract, Transform, Load), Data Exploration, Feature Engineering, delivering actionable insights, Visualization, and Data Storytelling.

CREDITS: 3 credits

CONTACT HOURS: 39 hours

LANGUAGE OF INSTRUCTION: English

PREREQUISITES: Recommended Linear Algebra, Basic Programming / Instructor's consent

ADDITIONAL COST: None

METHOD OF PRESENTATION:
This course will consist of one session per week, with three blocks each lasting 50 minutes. During these sessions, the instructor will present theoretical concepts, provide examples, and lead practical exercises.

REQUIRED WORK AND FORM OF ASSESSMENT:
- Weekly Assignments - 10%
- Class preparation and participation - 10%
- Midterm exam - 40%
- Final Exam - 40%
- Bonus questions on Final Exam - 10%

Course Element

Weekly assignments
The assignments will be related to the materials delivered during the course and the development of a data science project, to be selected by the student in the 8th Week. No late assignments will be accepted. Assignments are due at the beginning of the lecture and, in addition to uploading it to the course website. The lowest homework score will be deleted for grading purposes.

Class Preparation and Participation
The in-class participation, including answering questions, contribution to discussions and participation in group activities will be evaluated.

Midterm Exam and Final Exam
The exam format can vary, it could be multiple-choice, fill-in-the-blank, short answer or it could require the students to write some code to solve a problem.

The exams would be designed to evaluate the student's understanding of the concepts and their ability to apply them to solve data science problems. It can include a combination of theoretical and practical questions, such as a question that describes a scenario and asks the student to select a specific technique or method to analyze or visualize the data, or questions that ask the student to write code to perform a specific data science task.

Bonos questions on Final Exam
The final exam is comprehensive and will include a bonus question worth 10/100 extra points that will allow the students to make up a low midterm score.

LEARNING OUTCOMES:
By the end of the course, students will be able to:

- To understand the data science process by defining and exploring popular terms such as data mining, machine learning or big data to provide an overall framework to applied data science.
- To identify business key facts by deploying data paradigms, such as Kimball's data warehouse modelling, to structure the data in a meaningful way.
- To extract, transform and load raw data following the appropriate data paradigms by using state-of-the-art frameworks to populate the databases.
- To conduct initial exploratory analysis by applying a variety of statistical techniques to identify the appropriate features for more advanced modelling.
- To formulate adequate hypothesis by analyzing and transforming real business questions to understand what is needed from the data.
- To apply supervised and unsupervised modelling by applying advance machine learning algorithms to obtain knowledge from data.
- To apply different visualization techniques by using state-of-the-art visual suites (such as Tableau) to appropriately report data insights to the business managers.

ATTENDANCE POLICY:
As a member of our class community, you are expected to be present and on time every day. Attending class has an impact on your learning and academic success. For this reason, attendance is **mandatory for all IES Madrid classes**, including course-related excursions. Absences will only be excused in cases of documented medical or family emergencies. Any assessed work, including exams, tests, and presentations, will be rescheduled if the absence is justified; otherwise, it will be graded as 0. You are allowed a maximum of 1 unexcused absence for once-a-week courses and 2 unexcused absences for twice-a-week courses. Each additional absence will deduct 5 points from your final grade. **Accumulating 7 or more absences in twice-a-week courses, or 4 or more in once-a-week courses, may lead to a failing grade.** Students arriving more than 11 minutes after the scheduled start time will be marked as 'late.' Three 'late' arrivals will be counted as 1 absence. Any student arriving 30 minutes or more after the scheduled start time will be marked as 'absent.' For more information, please refer to the full Attendance and Punctuality Policy document.

CONTENT:

Choose an item	Content	Assignments
Week 1	**Data Science: An introduction** • What is Data Science. Difference between Data Mining, Big Data, Machine Learning and Artificial Intelligence. • The Knowledge Discovery Process in Databases • Type of Data Professions. The Data Scientist.	• Reading: Chapter 1 of **Data science and its relationship to big data and data-driven decision making**, Foster Provost and Tom Fawcett, 2013, DOI: 10.1089/big.2013.1508
Week 2	**Introduction to the Data Science Framework** • Introduction to Python and key libraries (Pandas, Seaborn, Numpy, Sklearn, etc.) • Introduction to Talend • Introduction to SQL Server • Introduction to Tableau	• Preparation of the class: setup of the different tools to be used during the course.

Choose an item	Content	Assignments
Week 3	**Data Science Projects** • Project Definition • Structuring your Data • Data Exploration. Basic Descriptive Analysis	• Assignment 1: Conduct a basic Descriptive Analysis of an open dataset (such as Boston Housing or Titanic Dataset) and answer a set of questions about the dataset.
Week 4	**Supervised Machine Learning** • Regression (Simple Linear Regression, Multiple Linear Regression, Polynomial Regression, Logistic Regression) • Metrics, Gradient Descent, Bias-Variance Trade-off and Regularization	• Assignment 2: • Using a given dataset, build regression models to predict the target variable based on the predictor variables, using a Machine Learning library. • Make predictions and conduct a critical comparison between the predicted values and actual values, using appropriate evaluation metric(s) explained in the lecture. • The student should experiment with different subsets of the predictor variables and subsets of the data and analyze how it impacts the model performance.

Week 5	**Supervised Machine Learning (Classification I)** • Logistic Regression • Metrics (binary classification & multiclassification) • Tree-Based Classifiers	• Assignment 3: • Using a given dataset, build a basic classification model to predict the class label based on the features, using a Machine Learning library. • Make predictions and evaluate the model performance using appropriate evaluation metric(s) such as Accuracy, Precision, Recall, F1-score or AUC-ROC Curve. • The student should experiment with different subsets of the features and different feature selection techniques, and analyze how it impacts the model performance.
Week 6	**Supervised Machine Learning & Feature Engineering** • K-Nearest Neighbors • Feature Transformation (scale, standardize, normalize)	
Week 7	**Midterm Exam**	None

Choose an item	Content	Assignments
Week 8	**Understanding your Business** • Understanding the Business from a Data Perspective • Data Warehousing & analytical processing • Example of schemas (Star, Snowflake) • Case of Study	• Reading: Annex A "Proposal Review Guide" of **Data science and its relationship to big data and data-driven decision making**, Foster Provost and Tom Fawcett, 2013, DOI: 10.1089/big.2013.1508 • Assignment 5: • Elaborate a basic Project Plan, including the specific problem or field of application, project objectives and goals and milestones. • Provide a summary of the scope of the project including specific objectives and what the project will and will not cover. • Define the data model to be used, including the type of data, data sources, and how the data will be obtained and structured. • Outline the methodology to be used, including the techniques and tools that will be employed, and how the project is expected to be executed.

Week 9	**Extract, Transform and Load Processes** • Principles of Extract, Transform and Load tools. • Description of the Use Case that will be used during the second half of the course. • Practical Exercises.	• Assignment 6: • Using a ETL tool, construct a basic data model related to the previous case of study. The tasks will include: • Connect to the specified data sources and extract the necessary data. • Transform the data to match the data model structure. • Load the transformed data into the data model.
Week 10	**Unsupervised Learning (I)** • Basic of Clustering, Evaluation metrics and Challenges • Partitioning approach (k-means. K-medoids) • Hierarchical Approach (Agglomerative clustering, divisive clustering)	• Assignment 7: • Using a given dataset, build a basic classification model to predict the class label based on the features, using a Machine Learning library. • Make predictions and evaluate the model performance using appropriate evaluation metric(s) such as Accuracy, Precision, Recall, F1-score or AUC-ROC Curve. • The student should experiment with different subsets of the features and different feature selection techniques, and analyze how it impacts the model performance.

Choose an item	Content	Assignments
Week 11	**Unsupervised Learning (II) & Visualization** • Density-Based Clustering (DBSCAN and introduction to OPTICS, HDBSCAN*) • Data Storytelling. Types of visualization	• Assignment 8: • Using a given dataset, apply different unsupervised learning techniques, such as k-means, hierarchical clustering and density-based clustering to find groups of similar observations. • Evaluate the performance of the clustering methods using appropriate evaluation metric(s) such as the Silhouette coefficient. • The student should experiment with different clustering methods and number of clusters, and compare the results. • Interpret the results of the clustering by examining the characteristics of the clusters and identifying potential insights

Week 12	**Visualization (II) & Reporting of Final Results** • Visualization in Python • Visualization in Tableau • Review of contents for Final Exam	• No assignment (preparation of the Final Exam)
Week 13	**Final Exam**	

COURSE-RELATED TRIPS:

- There are no course-related trips associated to this course.

REQUIRED READINGS:

- Textbook: Data Mining: Concepts and Techniques. Jiawei Han, Micheline Kamber, Jian Pei. – 3rd Edition. ISBN 978-0-12-381479-1

RECOMMENDED READINGS:

- All of Statistics: A Concise Course. Larry Wasserman. ISBN 978-1-4419-2322-6. DOI 10.1007/ 978-0-387-21736-9
- Data science and its relationship to big data and data-driven decision making, Foster Provost and Tom Fawcett, 2013, DOI: 10.1089/big.2013.1508
- Introduction to Linear Algebra (5th Edition). Gilbert Strang, Wellesley-Cambridge Press, 2016.