MASTER IN DATA SCIENCE

Main Language of Instruction:
French ○ English ☑ Arabic ○

Campus Where the Program Is Offered: CST

OBJECTIVES

This Master aims to train:

- Specialists at a high level who can design and use new tools for collecting massive data and process them using appropriate algorithms.
- Researchers with expertise in computer science, applied mathematics, and statistics.
- Designers of database management systems that can ensure the quality, security, and accessibility of information.
- Multidisciplinary consultants capable of transforming information into decision support tools within a company.

PROGRAM LEARNING OUTCOMES (COMPETENCIES)

1. Acquire and apply advanced knowledge appropriate to the discipline
1.1. Acquire theoretical and practical concepts appropriate to the discipline
1.2. Demonstrate proficiency in applying theoretical concepts to practical problems within the discipline
2. Solve critical issues and demonstrate expertise in key areas in the field of study
2.1. Identify and evaluate key challenges in the field
2.2. Solve critical issues by using advanced mathematics and sciences
2.3. Exhibit depth of knowledge in specialized areas
3. Apply new and diversified theoretical and experimental methods as appropriate to the discipline
3.1. Demonstrate the ability to learn and apply new methods and technologies
3.2. Utilize advanced analytical tools and techniques to solve complex issues in the field
3.3. Integrate new technologies into existing systems to improve performance
4. Communicate, at an advanced level, in oral and written form
4.1. Prepare clear, concise, and well-organized written reports on complex topics
4.2. Deliver effective oral communications, demonstrating mastery of the subject matter

ADMISSION REQUIREMENTS

Admission of students is based on their file and an interview might be required.

- For applications to the first semester of the Master's program, students must hold a Bachelor in Computer and Communications Engineering, Computer Science, or Mathematics, or hold an equivalent degree recognized by USJ.
- For applications to the third semester of the Master's program, students must hold a Bachelor of Engineering in Computer and Communications Engineering, be a CCE student at ESIB and have earned at least 120 credits in the Engineering Cycle, or hold a Master in Computer Science, or Computer and Communications, Mathematics or an equivalent degree recognized by USJ.

COURSES/CREDITS GRANTED BY EQUIVALENCE

ESIB CCE graduates, holders of a Bachelor's or professional Master's degree in Computer Science or Mathematics, fifth-year ESIB CCE students, and holders of a recognized equivalent diploma, can validate up to 60 credits of the program by equivalence, depending on the courses they already passed in their previous or current curriculum, and depending on the decision of the admission jury.

PROGRAM REQUIREMENTS

120 credits: Required courses (116 credits), Institution's elective courses (4 credits).

Required courses (116 Cr.):
Cloud and Digital Transformation (4 Cr.), Graph Theory and Operational Research (4 Cr.), Inferential Statistics (6 Cr.), Natural Language Processing (4 Cr.), Optimization for AI (4 Cr.), Programming for Data Science and Artificial Intelligence (4 Cr.), Data Visualization and Communication (4 Cr.), Enterprise Data Management (5 Cr.), Machine Learning and Deep Learning (4 Cr.), Marketing Data Science (5 Cr.), Mining Massive Data Set (4 Cr.), Regression Models (4 Cr.), Social Big Data (4 Cr.), AI for Business and Marketing (6 Cr.), Applied Regression and Time Series Analysis (4 Cr.), Big Data Frameworks (4 Cr.), Generative AI (4 Cr.), Legal, Political and Ethical Considerations for Data Scientists and AI (4 Cr.), Software Engineering for AI (4 Cr.), Theoretical Guidelines for High-Dimensional Data Analysis (4 Cr.), Master Dissertation (30 Cr.).

Institution's elective courses (4 Cr.):
Students will choose one of the following two courses: Mathematics for Data Science and AI (4 Cr.) or Relational Database (4 Cr.).

SUGGESTED STUDY PLAN

Semester 1

Code	Course Name	Credits
020CTDIM1	Cloud and Digital Transformation	4
048DSTGM1	Graph Theory and Operational Research	4
048DSSIM1	Inferential Statistics	6
048DSPMM1	Programming for Data Science and Artificial Intelligence	4
020OPAIM1	Optimization for AI	4
020IANLM3	Natural Language Processing	4
020IAMAM1 020BDREM1	Mathematics for Data Science and AI (Elective) Or Relational Database (Elective)	4
	Total	**30**

Semester 2

Code	Course Name	Credits
020INTDM2	Enterprise Data Management	5
020DVCOM3	Data Visualization and Communication	4
020MLDLM3	Machine Learning and Deep Learning	4
020MADSM2	Marketing Data Science	5
020FOBDM2	Mining Massive Data Sets	4
048DSSBM1	Social Big Data	4
048MBCMM2	Regression Models	4
	Total	**30**

Semester 3

Code	Course Name	Credits
048DSARM3	Applied Regression and Time Series Analysis	4
020IABMM3	AI for Business and Marketing	6
020BDFRM3	Big Data Frameworks	4
020GAIES5	Generative AI	4
020IALPM3	Legal, Political and Ethical Considerations for Data Scientists and AI	4
020IAIDM2	Software Engineering for AI	4
048DSTGM3	Theoretical Guidelines for High-Dimensional Data Analysis	4
	Total	**30**

Semester 4

Code	Course Name	Credits
020STGEM4	Master Dissertation	30
	Total	**30**

COURSE DESCRIPTION

020CTDIM1 Cloud and Digital Transformation 4 Cr.

Cloud computing and big data are currently the two main technological developments driving companies' growth in the digital sector. Big data, achieved through the collection and analysis of large amounts of data, represents the potential for new activities in many sectors. Cloud computing allows anywhere and on-demand access to digital services, resulting in significant cost reductions. These two subjects are closely linked: cloud computing is the only technology capable of supporting the computation of problems defined by big data.
This course introduces cloud-based big data solutions, such as AWS's big data platform. Students will learn how to utilize existing cloud services to process data using the vast ecosystem of tools, how to create big data environments and apply the best practices to secure those environments in an economical approach.

048DSTGM1 Graph Theory and Operational Research 4 Cr.

This course introduces students to graph theory and operational research as modeling and decision-making tools for the data scientist. By the end of the course, students will be able to make mathematical and computer representations of graphs, apply graph traversal algorithms, calculate the shortest path, maximize flow problems, analyze complex networks, use the NetworkX Python library, use Markov chains to solve real-world problems, understand the Simplex algorithm and linear programming, and use numerical tools for solving optimization problems.

048DSSIM1 Inferential Statistics 6 Cr.

Statistical inference consists of predicting the unknown characteristics of a population based on a sample drawn from this population. Thus, the objective of statistics is symmetrical to that of probability. By the end of this course, students will be able to conduct a complete statistical study: spanning from selecting appropriate statistical models, to estimating unknown quantities and making informed decisions. The applications attributed during this course are led using the R language software mainly for data manipulation, implementing statistical procedures, plotting graphics and functions and presenting results in a comprehensible way.

048DSPOM1 Programming Languages for Data Science and Artificial Intelligence 4 Cr.

This course equips students with the necessary tools for developing advanced-level programs understanding the Object-Oriented Programming (OOP) approach. The first part of the course focuses on the C++ language while the second part delves into Python and its functionalities relevant to data science. In the final part, students are introduced to machine learning examples using Python, allowing exploration of the power of the libraries provided by the Python community.

020IAMAM1 Mathematics for AI and Machine Learning 4 Cr.

Artificial Intelligence has gained importance in the last decade with a lot depending on the development and integration of AI in our daily lives. The progress that AI has already made is astounding with the self-driving cars, medical diagnosis and even beating humans at strategy games like Go and Chess.
The future for AI holds tremendous promise, potentially leading to the creation of robotic companions. Consequently, many developers are now diving into AI and ML programing. However, mastering AI and ML algorithms demands a strong understanding of mathematics.
Mathematics plays an important role as it builds the foundation for programming for these two streams. This course will help students master the mathematical foundation required for writing programs and algorithms for AI and ML.
The course covers three main mathematical theories: Linear Algebra, Multivariate Calculus and Probability Theory.

020IANLM3 Natural Language Processing 4 Cr.

This course goes beyond the phase of gathering large amounts of data by focusing on how machine learning algorithms can be rewritten and scaled to work on petabytes of both structured and unstructured data simultaneously, to generate sophisticated models used for making predictions. Conceptually, the course is divided into two parts.
The first part covers deep learning and key network architectures, including: convolutional neural networks, autoencoders, recurrent neural networks, and long short-term memory (LSTM) networks. This part also addresses stochastic networks, conditional random fields, Boltzmann machines, stochastic and mixed deterministic models, as well as deep reinforcement learning.
The second part focuses on natural language processing (NLP): Research in automatic natural language processing is a subfield of artificial intelligence aimed at developing automated techniques for manipulating linguistic data. Immediate applications of these techniques include the development of more natural textual interfaces, automatic document translation, spam detection, information retrieval from document collections based on queries, question/answer systems, and more. This part introduces students to the following topics: Introduction to the problem of automatic natural language processing and its applications. The relationship between natural language and formal languages: the problem of ambiguity. An overview of current linguistic theories. Speech analysis and synthesis. Morphological analysis: structure of the dictionary and suffix analysis. Syntactic analysis: ATN parsers, unification grammars, and the representation of the semantics of natural languages: formal logic and frameworks. Semantic interpretation. World knowledge and speech context. Applications.

020OPAIM1 Optimization for AI 4 Cr.

This course delves into the mathematical optimization techniques essential for developing and refining machine learning algorithms and AI applications. Focusing on theoretical foundations, this course explores deep neural network initialization, gradient descent techniques, automatic differentiation and backpropagation, and adaptive learning rate algorithms such as Adam and RMSProp. Additionally, it covers principal component analysis (PCA), density estimation algorithms, and support vector machines (SVM). Students will learn to solve unconstrained and constrained optimization problems, apply these methods to neural networks, and enhance model performance. The course provides a comprehensive understanding of optimization's role in AI, equipping students with the theoretical knowledge to tackle complex challenges in various AI domains.

020BDREM1 Relational Database 4 Cr.

This course introduces the design, creation and management of databases. It allows students to master the concept of "Database," designing a database for a given Information System (IS), understanding the Relational Model, acquiring skills in creating and managing a database using SQL language, and understanding the techniques of database management systems.

020INTDM2 Enterprise Data Management 5 Cr.

"Enterprise Data Management (EDM) is the ability of an organization to precisely define, easily integrate and effectively retrieve data for both internal applications and external communication. EDM focuses on the creation of accurate, consistent, and transparent content." (Wikipedia)

This course addresses the challenges of enterprise data management at scale, primarily focusing on data architecture, data modeling and data integration, both on-premise and on the cloud. It covers different enterprise data architectures such as Data Warehouses, and Data Lakes. Additionally, it details various data models (including structured, semi-structured (XML), unstructured, and semantic data with RDF/OWL/SPARQL. The course also describes various NoSQL databases (key-value, column, document or graph-oriented databases), as well as various big data formats (Avro, ORC and Parquet). The course explains different data integration approaches: integration according to a materialized view (Data Warehouses/OLAP) and integration according to a virtual view (Mediators/GAV-LAV).
This course also covers Stream, and Batch processing using big data architectures such as Lambda architecture as well as integration and processing pipelines, using appropriate tools such as Talend Big Data Integration Studio, and Azure Data Factory.

020DVCOM3 Data Visualization and Communication 4 Cr.

Access to data is exponentially increasing while human capacity to manage and understand it remains constant. Clearly and effectively communicating about the models found in the data is a key skill for a successful data scientist. This course introduces basic concepts of visualization, analysis, and visual representation of data, necessary for the creation of suitable applications and tools that allow students to manage and analyze big data flows. It involves designing and implementing complementary visual and verbal representations of patterns and analyses to convey results, answer questions, drive decisions, and provide convincing evidence supported by data.

020MADSM2 Marketing Data Science 5 Cr.

This course provides students with foundations in various modeling methodologies commonly used in marketing, including Attribution Modeling, Marketing Mix Modeling (MMM), and other advanced techniques such as Bayesian methods. Through a combination of theoretical lectures, hands-on practical exercises, and real-world case studies, students will develop the skills necessary to analyze marketing data, derive actionable insights, and make data-driven decisions to optimize marketing strategies.

020MLDLM3 Machine Learning and Deep Learning 4 Cr.

This course goes beyond the phase of collecting large volumes of data by focusing on how machine learning algorithms can be rewritten and extended to scale for petabytes of structured and unstructured data. Also, sophisticated models for predictions are included. The course is divided into three main parts.
The first part deals with the design and development of algorithms allowing the behavior of computers to evolve based on empirical data, such as databases or sensory data. We also define supervised, unsupervised and reinforcement learning.
The second part introduces deep learning as well as key network architectures including: convolutional neural networks, autoencoders, recurrent neural networks, long-term short-term networks "LSTM". This part also covers deep reinforcement learning.
The third part deals with the processing of natural languages. Indeed, research in the automatic processing of natural languages is a field of artificial intelligence aiming at the development of automated techniques for the manipulation of language data, in textual or sound forms. Immediate applications include developing more natural textual interfaces, automatic document translation, spam detection, information retrieval, question-answering systems, among others. This part introduces students to the following subjects: Introduction to the problem of automatic processing of natural language and its applications.

020FOBDM2 Mining Massive Data Sets 4 Cr.

This course covers the fundamentals of designing dedicated software systems for big data analytics.
The course begins with principles of designing relational database systems for analyzing business data, including declarative queries, query optimization and transaction management. It also covers the evolution of basic data systems to support complex analytical problems and scientific data management.
The course then explores fundamental architectural changes necessary for processing data beyond the limits of a single computer. This includes parallel databases, "MapReduce", column storage, distributed key value, and enabling low-latency analytical results from real-time data streams. Finally, this course examines advanced data management systems supporting various data models including tree structure (XML and JSON), structured data

graphics (RDF), new workloads such as machine learning tasks (Spark), and mixed workloads (such as Google Cloud Dataflow).

048MBCMM2 Regression Models 4 Cr.

This course covers the fundamentals of regression, including linear regression, its approach, and its applications in practical studies. It also includes ANOVA techniques and logistic regression. The course alternates between theoretical presentations and computer exercises, utilizing the R language.

048DSSBM1 Social Big Data 4 Cr.

This course introduces the structures and data types found on social networks (such as Facebook, Twitter, Instagram, etc.). It covers various methods of data collection and analysis based on application areas under the R language. Students gain proficiency in utilizing different application programming interface services (API) to collect data, analyze and explore social media data for research and development purposes. Ultimately, students will use the data drawn and analyzed to improve their presence and strategy on social networks.

020IABMM3 AI for Business and Marketing 6 Cr.

This course explores the integration of artificial intelligence tools and techniques in business and modern marketing practices. Students will delve into the utilization of AI algorithms, machine learning models, and data analytics to optimize marketing strategies across various digital channels and business decision-making. Through real-world applications and hands-on experience, students will learn to personalize content, enhance customer engagement, and drive ROI through targeted advertising and dynamic pricing. The course emphasizes ethical considerations and responsible AI usage, empowering businesses to effectively leverage technology while maintaining integrity and trust.

048DSARM3 Applied Regression and Time Series Analysis 4 Cr.

This course introduces the following subjects: Visualization techniques for time series data, key concepts in probability and mathematical statistics, classical linear regression models, variable transformation, model specification, causal inference, variable estimation, autoregressive (AR) instrumental models, moving average, autoregressive moving average (ARMA), integrated average autoregressive (ARIMA), GARCH models, vector autoregression (VAR), statistical forecast, and regression with time series data.

020BDFRM3 Big Data Frameworks 4 Cr.

This course is conceptually divided into two parts.
The first part covers the fundamental concepts of MapReduce parallel computing, focusing on Hadoop, MrJob and Spark. It delves deeply into Spark, data frames, Spark Shell, Spark Streaming, Spark SQL, MLlib. Students use MapReduce for industrial applications and deployments across various fields, including advertising, finance, health, and search engines.
The second part focuses on algorithmic design and development in parallel computing environments (Spark). It covers algorithmic development (learning decision tree), graphics processing algorithms (such as PageRank and short path), Newton algorithms, and support vector machines.

020GAIES5 Generative AI 4 Cr.

Generative AI is a course designed to offer a comprehensive understanding of generative models in AI, like ChatGPT and diffusion models, and the practical application of these technologies. The course emphasizes open-source models, training techniques, and fine-tuning practices.
Specific objectives of the course:

- Understanding the architecture and principles of generative models.
- Learning the process of training and fine-tuning AI models.
- Exploring open-source AI models and their applications.
- Developing hands-on experience in building AI-driven solutions.

020IALPM3 Legal, Political and Ethical Considerations for Data Scientists and AI 4 Cr.

This course introduces ethics, politics, and ethical implications of data, including personal data. It examines the legal, political, and ethical issues that arise throughout the entire lifecycle of the science of data collection, storage, processing, analysis and use, including privacy, surveillance, security, classification and discrimination. Additionally, a brief introduction will be provided about law and Labor law in general. Case studies will be used to explore these issues in various areas such as criminal justice, national security, health, marketing, politics, education, automotive, employment, athletics, and development. Particular attention will be paid to legal and political constraints and considerations specific to each area.

020IAIDM2 Software Engineering for AI 4 Cr.

By the end of this course, students will have acquired a comprehensive understanding of industrial AI, enabling them to effectively apply AI techniques in real-world industrial settings. They will be equipped with practical skills in MLOps, AI deployment, and XAI, making them valuable contributors to the rapidly evolving field of industrial artificial intelligence. In the final part of the course, students will explore the emerging field of Explainable AI (XAI). They will learn techniques to interpret and explain the decisions made by AI models, with an emphasis on their application in industrial scenarios.

048DSTGM3 Theoretical Guidelines for High-Dimensional Data Analysis 4 Cr.

This course introduces the different types of quantitative research methods and statistical techniques for analyzing data. It starts with an emphasis on measurement, statistical inference and causal inference. Next, it explores a range of statistical techniques and methods using the language of open-source statistics (using R or Python). Different techniques for data analysis and visualization are introduced, with a focus on applying this knowledge to real-world data problems. The techniques included are descriptive and deductive statistics, sampling, experimental design, parametric and non-parametric difference tests, least squares regression, and logistic regression.

020STGEM4 Master Dissertation 30 Cr.

During the last semester, students must complete a professional internship in a company or research work in a laboratory for 4 months.
The internship can take place in Lebanon or abroad. The scientific responsibility for the internship is provided jointly by the company and an instructor from USJ or a partner university. This internship, of a minimum of one semester, aims to develop all the skills necessary in the data science field.
Students can also choose to contribute to an academic research project. It takes place in a laboratory either in Lebanon or in an external institution.
The internship or research work is the subject of a dissertation and a public defense in the presence of USJ professors.
Only students who have validated all the courses of the first year and the first semester of the second year of the Master's program are authorized to submit their internship report or present their research dissertation.
The dissertation or the report includes a bibliographic part and a technical part.
The evaluation of the internship work considers three elements:

- Evaluation of the trainee's scientific initiative.
- Evaluation of the report.
- Evaluation of the oral defense.