# Big Data and predictive analytics

## Lecturer: Maria Chiara Debernardi

### Language

English

### Course description and objectives

In the contemporary digital landscape, Big Data represents a transformative approach to information processing and strategic analysis. The exponential proliferation of digital technologies and advanced storage media has generated unprecedented volumes of complex, multidimensional data across diverse organizational domains. While fundamental statistical methodologies remain consistent, Big Data demands sophisticated technological infrastructures to effectively capture, process, and derive meaningful insights.
This course explores the Big Data phenomenon through a practical lens, emphasizing value extraction using advanced predictive analytics techniques. Participants will leverage the KNIME Analytics Platform – a robust, open-source graphical interface (no coding required) – to transform raw data into knowledge.

Upon successful completion of this course, students will:

- critically differentiate between traditional and Big Data paradigms
- appreciate the potential use of data in a corporate environment
- effectively leverage KNIME Analytics Platform for data analysis and Machine Learning
- develop predictive modeling capabilities

### Audience

The course is open to all Bocconi's Master of Science students. In particular it is targeted at:

- those who want to understand what Big Data really is, and how to exploit it
- those who want to gain some practical, analytical skills and confidence with the Data Scientist Toolkit

The course is part of the Enhancing Experience - Curricular Integrative Activities. Upon successful completion of the course (attendance of at least 75% of the scheduled lessons and passing the final exam), students will get **2 credits** and an **Open Badge**, sharable across the web (LinkedIn) or personal CV.

## Prerequisites

No prior programming experience or familiarity with KNIME Analytics Platform is required.
Participants should have a thorough understanding of:

- descriptive and inferential statistical methodologies, equivalent to foundational university-level Statistics exam
- advanced mathematical concepts, with specific emphasis on function optimization techniques (e.g., finding the minimum of a loss function)

## Duration

16 hours

## Teaching mode

**Distance learning**. Lessons will take place in **synchronous remote mode**.

The **final test** on the last day of class, however, can **only** be taken **in physical presence**. Online mode will not be provided.

## Calendar

| Lecture | Date | Time | Room |
|---|---|---|---|
| 1 | Mon 24/03/2025 | 18.15 - 19.45 | Virtual room |
| 2 | Thu 27/03/2025 | 18.15 - 19.45 | Virtual room |
| 3 | Mon 31/03/2025 | 18.15 - 19.45 | Virtual room |
| 4 | Thu 03/04/2025 | 18.15 - 19.45 | Virtual room |
| 5 | Mon 07/04/2025 | 18.15 - 19.45 | Virtual room |
| 6 | Thu 10/04/2025 | 18.15 - 19.45 | Virtual room |
| 7 | Mon 14/04/2025 | 18.15 - 19.45 | Virtual room |
| 8 | Thu 17/04/2025 | 18.15 - 19.45 | InfoAS04/05 |

## Syllabus of the course

| Lesson | Topics | Book reference |
|---|---|---|
| 1 | **Introduction**<br>- Big Data: definition(s) and taxonomy<br>- Predictive analytics<br>- The KNIME environment<br>- Building a KNIME workflow<br>***Exercises*** | **Parr. 1.1, 1.3.3, 12.1 + slides** |
| 2 | **Business understanding and Data preparation**<br>- CRISP-DM: how to efficiently create a predictive analytics model<br>- Data preparation: the ETL step<br>- Exploring the dataset<br>***Exercises*** | **Parr. 3.3, 4.2, 4.3, 2.2, 2.3, 6.1, 9.2 + slides** |
| 3 | **Predictive analytics techniques**<br>- Predictive analytics algorithms: characteristics and taxonomy<br>- When to use which model<br>- Sampling: train, test, and cross-validation<br>- Quantitative prediction: regression<br>***Exercises*** | **Parr. 11.5, 1.3, 1.4, 7.1, 7.2, 7.3, 9.1 + slides** |
| 4 | **The classification problem**<br>- Data preparation for classification tasks<br>- Setting up a classification model<br>- Model performance and evaluation<br>- Confusion matrix<br>***Exercises*** | **Parr. 9.6, 9.7, 7.4, 7.5 + slides** |
| 5 | **Classification algorithms**<br>- Naïve Bayes<br>- k-NN<br>- Decision tree<br>- Support vector machine (SVM)<br>- Comparing different models: ROC curve<br>- Hyperparameter tuning | **Parr. 11.1, 10.2,11.2, 11.4, 7.4 + slides** |
| 6 | **Model ensembles**<br>- Bagging<br>- Boosting<br>- Random forest<br>- Stacking<br>***Exercises*** | **Parr. 11.3, 11.2.3 + slides** |

| Lesson | Topics | Book reference |
|---|---|---|
| 7 | **Neural Networks**<br>- From Linear regression to Neural Network models<br>- From Machine learning to Deep learning (*hints only*)<br>***Exercises*** | **Parr. 11.6, 9.4 + slides** |
| 8 | **Q&A and final test – *in presence only***<br>- Guided recap exercise<br>- Last doubts and clarifications<br>- **Exam** | |

## Software used

KNIME Analytics Platform (knime.com): latest version available (5.4.2 or higher)
Download it from: download-knime
Please note that no registration is required, but to download your specific OS KNIME version you must accept the *terms and conditions* of the open-source license.

## Suggested bibliography

The reference bibliography for the final exam is only based on slides and commented exercises provided by the Lecturer.
Additional bibliography:

- Skiena S. S., *The Data Science Design Manual*, Springer, 2017

## Available seats

This activity is limited to **110** participants and reserved for **students of the Master of Science Programs**.
Registration cannot be carried out once this number has been reached or after the registration period closes.