Big Data and predictive analytics

Lecturer: Maria Chiara Debernardi

Language

English

Course description and objectives

Big Data is the hot new buzzword in ICT circles. The proliferation of digital technologies, along with digital storage and recording media, has created massive amounts of diverse data, which can be strategically exploited by companies in all sectors. Big Data, which can take up terabytes and petabytes of storage space in diverse formats including text, video, sound, images, and more, can be analyzed with the same methods as more traditional data are, with the main difference being the tools required to support the specific characteristics defining Big Data.
The course gives an overview of the Big Data phenomenon, focusing on how to extract value from data using predictive analytics techniques inside the KNIME Analytics Platform (a free and open-source graphical user interface tool).

Upon successful completion of this course, students should be able to:

- understand the Big Data phenomenon, and the differences between Big Data and Traditional Data
- appreciate the potential use of data in a corporate environment
- use one of the easiest Predictive Analytics tools (no coding required) to extract valuable information from raw data

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 coding experience is needed, nor knowledge of KNIME Analytics Platform.
It is essential to have a good knowledge of the fundamentals of descriptive and inferential statistics, corresponding to the first Statistics exam of one's study plan, along with the mathematical concepts of function optimization (e.g., finding the minimum of a loss function).

Duration

16 hours

Teaching mode

Distance learning. Lessons will take place **exclusively** 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	Tue 04/06/2024	14.45 - 16.15	Virtual room
2	Tue 04/06/2024	16.30 - 18.00	Virtual room
3	Thu 13/06/2024	16.30 - 18.00	Virtual room
4	Tue 18/06/2024	16.30 - 18.00	Virtual room
5	Fri 21/06/2024	16.30 - 18.00	Virtual room
6	Tue 25/06/2024	16.30 - 18.00	Virtual room
7	Thu 27/06/2024	16.30 - 18.00	Virtual room
8	Tue 02/07/2024	16.30 - 18.00	InfoAS04/05

Syllabus of the course

Lesson	Topics	Book reference
1	**Introduction** - Big Data: definition(s) and taxonomy - Predictive analytics - The KNIME environment - Building a KNIME flow ***Exercises***	**Parr. 1.1, 12.1 + slides**
2	**Business understanding and Data preparation** - CRISP-DM: how to efficiently create a predictive analytics model - Data preparation: the ETL step - Exploring the dataset ***Exercises***	**Parr. 3.3, 4.2, 4.3, 4.4, 2.2, 2.3, 6.1 + slides**
3	**Predictive analytics techniques** - Predictive analytics algorithms: characteristics and taxonomy - When to use which model - Sampling: train, test, and cross-validation - Quantitative prediction: regression ***Exercises***	**Parr. 11.5, 1.3, 1.4, 7.1, 7.2, 7.3 + slides**
4	**Classification algorithms** - Data preparation for classification tasks - Setting up a classification model - Model performance and evaluation - Comparing different models ***Exercises***	**Parr. 9.6, 9.7, 10.2, 11.1, 7.4, 7.5 + slides**
5	**Model ensembles** - Bagging - Boosting - Random forest - Stacking - Hyperparameter tuning ***Exercises***	**Parr. 11.3, 11.2.3 + slides**
6	**From rule-based to black box and back** - From Linear regression to Neural Network models - From Machine learning to Deep learning (*hints only*) - XAI: how to extract rules from results ***Exercises***	**Parr. 9.1, 9.2, 9.4, 11.6 + slides**

Lesson	Topics	Book reference
7	**Unsupervised models** - Clustering - Recommendation systems (*hints only*) ***Exercises***	**Parr. 10.1, 10.5 + slides**
8	**Q&A and final test – *in presence only*** - Guided recap exercise - Last doubts and clarifications - **Exam**	

Software used

KNIME Analytics Platform (knime.com): latest version available (5.2.3 or higher)
Download it from: download-knime
Please note that no registration is required, but you must accept the terms and conditions of the open-source license to download your specific OS KNIME version.

Suggested bibliography

The reference bibliography for the final exam is 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 to **students of the Master of Science Programs**.
Registrations cannot be carried out once this number has been reached or after closing of the registration period.