Guia docent
270964 - DAKD - Anàlisi de Dades i Descobriment del Coneixement

Última modificació: 23/11/2023

Unitat responsable: Facultat d'Informàtica de Barcelona
Unitat que imparteix: 723 - CS - Departament de Ciències de la Computació.

Titulació: MÀSTER UNIVERSITARI EN CIÈNCIA DE DADES (Pla 2021). (Assignatura optativa).

Curs: 2023 **Crèdits ECTS:** 6.0 **Idiomes:** Anglès

PROFESSORAT

Professorat responsable: ALFREDO VELLIDO ALCACENA

Altres:

CAPACITATS PRÈVIES

Students are expected to have at least some basic background in the area of artificial intelligence and, more specifically, with the areas of Machine Leaning and Computational Intelligence.
Some basic knowledge of probability theory and statistics would be beneficial.
Other than this, the course is open to students and researchers of all types of background.

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE12. Aplicar la ciència de dades en projectes multidisciplinaris per resoldre problemes en dominis nous o poc coneguts per la ciència de dades i que siguin econòmicament viables, socialment acceptables, i d'acord amb la legalitat vigent
CE13. Identificar les principals amenaces en l'àmbit de l'ètica i la privacitat de dades en un projecte de ciència de dades (tant en l'aspecte de gestió com d'anàlisi de dades) i desenvolupar i implantar mesures adequades per esmorteïr aquestes amenaces.
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.

Genèriques:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents

Transversals:
CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

Bàsiques:
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

METODOLOGIES DOCENTS

This course will build on different teaching methodology (TM) aspects, including:
TM1: Expositive seminars
TM2: Expositive-participative seminars
TM3: Orientation for individual assignments (essays)
TM4: Individual tutorization

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

1.Presenting DM as a process that should involve a methodology id applied at its best.
2.Introducing the students to the new concept of DM for processes, called Process Mining.
3.Delving into some detail in one of the stages of DM: data exploration.
4.Dealing in detail with the problem of data visualization for exploration as a key issue in DM.
5.Introducing the students to the basics of probability theory as applied in Data Analysis and Knowledge Discovery (DAKD)
6.Introducing the students to the probabilistic variant of DAKD in the form of Statistical Machine Learning, both for supervised and unsupervised learning models.
7.Dealing in detail with different unsupervised models for data visualization, including case studies.
8.Approaching the multi-faceted concept of data mining (DM) from different perspectives.

HORES TOTALS DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup gran	45,0	30.00
Hores activitats dirigides	9,0	6.00
Hores aprenentatge autònom	96,0	64.00

Dedicació total: 150 h

CONTINGUTS

Introduction to the concept of data mining (DM).

Descripció:
DM is a multi-faceted concept that requires discussion and clarification. We will do this at the beginning of the course.

DM as a methodology.

Descripció:
We argue that DM should not be focused on the concept of data analysis/modeling, but, instead, should be treated as a methodology with diverse inter-related stages.

DM for processes: Process Mining.

Descripció:
A new development in DM methodologies is that which deals with one specifically suited for processes. It is called Process Mining and will be described and discussed in this course.

Data exploration in DM.

Descripció:
One of the main stages of well-structures DM methodologies is Data exploration. It will be discussed as a preamble to data visualization.

Data visualization for exploration.

Descripció:
One of the aspects of the problem of data exploration is data visualization. It has a research 'life' of its own as it involves not only computer-based mathematical models, but also natural perception and processing.

Basics of probability theory in Data Analysis and Knowledge Discovery (DAKD)

Descripció:
For a long time in the last half-century, multivariate statistics and artificial intelligence (mostly in the field of machine learning) have developed in parallel without fully meeting. Statistical machine learning has bridged that field over the last two decades. We introduce it by first providing some basic principles of probability theory (Bayesian inference).

Statistical Machine Learning for DAKD: supervised models.

Descripció:
Once the basics of Bayesian inference are set, we will delve into the field of Statistical Machine Learning for IDA, starting with supervised learning models, with an emphasis on feed-forward artificial neural networks.

Statistical Machine Learning for DAKD: unsupervised models.

Descripció:
Once the basics of Bayesian inference and of Statistical Machine Learning for IDA in supervised models are set, we will continue with unsupervised models, focusing on self-organizing maps and related models.

Unsupervised models for data visualization, with case studies.

Descripció:
In the final item of the contents of the course, we will bring statistical machine learning and data visualization together by discussing some probabilistic unsupervised learning models for data visualization, including some case studies as an example.

ACTIVITATS

Essay on DAKD for DM

Descripció:
Students will have to write a research essay on the topic of DAKD for DM, with different options:
1. State of the art on an specific DAKD-DM topic
2. Evaluation of an DAKD-DM software tool with original experiments
3. Pure research essay, with original experimental content

Objectius específics:
1, 2, 3, 4, 5, 6, 7, 8

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE13. Identificar les principals amenaces en l'àmbit de l'ètica i la privacitat de dades en un projecte de ciència de dades (tant en l'aspecte de gestió com d'anàlisi de dades) i desenvolupar i implantar mesures adequades per esmorteïr aquestes amenaces.
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE12. Aplicar la ciència de dades en projectes multidisciplinaris per resoldre problemes en dominis nous o poc coneguts per la ciència de dades i que siguin econòmicament viables, socialment acceptables, i d'acord amb la legalitat vigent
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 3h
Activitats dirigides: 3h

Introduction to Data Mining and its Methodologies

Descripció:
Introduction to Data Mining as a general concept and to its methodologies for practical implementation

Objectius específics:
1

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 23h
Grup gran/Teoria: 9h
Activitats dirigides: 1h
Aprenentatge autònom: 13h

Process Mining

Descripció:
Introduction to the novel concept of Process Mining and its application within the DM framework.

Objectius específics:
2

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE12. Aplicar la ciència de dades en projectes multidisciplinaris per resoldre problemes en dominis nous o poc coneguts per la ciència de dades i que siguin econòmicament viables, socialment acceptables, i d'acord amb la legalitat vigent
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.

Dedicació: 9h
Grup gran/Teoria: 3h
Activitats dirigides: 1h
Aprenentatge autònom: 5h

Data Visualization

Descripció:
As part of the DM stage of Data Exploration, we focus in the problem of Data Visualization.

Objectius específics:
3, 4

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 16h
Grup gran/Teoria: 6h
Activitats dirigides: 1h
Aprenentatge autònom: 9h

Basics of probability theory for intelligent data analysis

Descripció:
Introduction to probability theory for intelligent data analysis, with a focus on Bayesian statistics

Objectius específics:
5

Competències relacionades:
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 16h
Grup gran/Teoria: 6h
Activitats dirigides: 1h
Aprenentatge autònom: 9h

Statistical Machine Learning methods

Descripció:
The meeting of statistics and machine learning: Statistical Machine Learning methods, from the point of view of both supervised and supervised learning

Objectius específics:
5, 6

Competències relacionades:
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 31h
Grup gran/Teoria: 12h
Activitats dirigides: 1h
Aprenentatge autònom: 18h

SML in data visualization, with case studies

Descripció:
We merge the topics of SML and data visualization, illustrating its use with some real case studies

Objectius específics:
4, 7, 8

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE13. Identificar les principals amenaces en l'àmbit de l'ètica i la privacitat de dades en un projecte de ciència de dades (tant en l'aspecte de gestió com d'anàlisi de dades) i desenvolupar i implantar mesures adequades per esmorteïr aquestes amenaces.
CE10. Identificar els mètodes d'aprenentatge automàtic i modelització estadística a utilitzar per resoldre un problema específic de ciència de dades, i aplicar-los de forma rigorosa
CE12. Aplicar la ciència de dades en projectes multidisciplinaris per resoldre problemes en dominis nous o poc coneguts per la ciència de dades i que siguin econòmicament viables, socialment acceptables, i d'acord amb la legalitat vigent
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE8. Extreure informació de dades estructurades i no estructurades, tenint en compte la naturalesa multivariant de les mateixes.
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 25h
Grup gran/Teoria: 9h
Activitats dirigides: 1h
Aprenentatge autònom: 15h

SISTEMA DE QUALIFICACIÓ

The course will include two evaluation tasks:
The first one will be a data science purely analytical task performed according to data mining principles.
The second one will involve writing an essay according to one of these three modalities:
1. State of the art on an specific IDA-DM topic
2. Evaluation of an IDA-DM software tool with original experiments
3. Pure research essay, with original experimental content

BIBLIOGRAFIA

Bàsica:
- MacKay, D.J.C. Information theory, inference, and learning algorithms. Cambridge UK: Cambridge University Press, 2003. ISBN 0521642981.
- Hand, D.; Mannila, H.; Smyth, P. Principles of data mining. Cambridge: MIT Press, 2001. ISBN 026208290X.
- Bishop, C.M. Pattern recognition and machine learning. New York: Springer, 2006. ISBN 0387310738.

Complementària:
- Hand, D.J. Statistics: a very short introduction. New York: Oxfrod University Press, 2008. ISBN 9780199233564.

- Spence, R. Information visualization: design for interaction. 2nd ed. Harlow [etc.]: Pearson/Prentice Hall, 2007. ISBN 9780132065504.
- Yau, N. Visualize this: the flowing data guide to design, visualization, and statistics. Indianapolis, NY: Wiley, 2011. ISBN 9780470944882.