Guía docente
270964 - DAKD - Análisis de Datos y Descubrimiento del Conocimiento

Última modificación: 23/11/2023

Unidad responsable: Facultad de Informática de Barcelona
Unidad que imparte: 723 - CS - Departamento de Ciencias de la Computación.

Titulación: MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (Plan 2021). (Asignatura optativa).

Curso: 2023 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: ALFREDO VELLIDO ALCACENA

Otros:

CAPACIDADES PREVIAS

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.

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.

Genéricas:
CG2. Identificar y aplicar métodos de análisis, extracción de conocimiento y visualización de datos recogidos en formatos muy diversos.

Transversales:
CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

Básicas:
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios

METODOLOGÍAS DOCENTES

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

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

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.

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas grupo grande	45,0	30.00
Horas aprendizaje autónomo	96,0	64.00
Horas actividades dirigidas	9,0	6.00

Dedicación total: 150 h

CONTENIDOS

Introduction to the concept of data mining (DM).

Descripción:
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.

Descripción:
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.

Descripción:
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.

Descripción:
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.

Descripción:
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)

Descripción:
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.

Descripción:
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.

Descripción:
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.

Descripción:
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.

ACTIVIDADES

Essay on DAKD for DM

Descripción:
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

Objetivos específicos:
1, 2, 3, 4, 5, 6, 7, 8

Competencias relacionadas:
CG2. Identificar y aplicar métodos de análisis, extracción de conocimiento y visualización de datos recogidos en formatos muy diversos.
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 3h
Actividades dirigidas: 3h

Introduction to Data Mining and its Methodologies

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

Objetivos específicos:
1

Competencias relacionadas:
CG2. Identificar y aplicar métodos de análisis, extracción de conocimiento y visualización de datos recogidos en formatos muy diversos.
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 23h
Grupo grande/Teoría: 9h
Actividades dirigidas: 1h
Aprendizaje autónomo: 13h

Process Mining

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

Objetivos específicos:
2

Competencias relacionadas:
CG2. Identificar y aplicar métodos de análisis, extracción de conocimiento y visualización de datos recogidos en formatos muy diversos.
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.

Dedicación: 9h
Grupo grande/Teoría: 3h
Actividades dirigidas: 1h
Aprendizaje autónomo: 5h

Data Visualization

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

Objetivos específicos:
3, 4

Competencias relacionadas:
CG2. Identificar y aplicar métodos de análisis, extracción de conocimiento y visualización de datos recogidos en formatos muy diversos.
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 16h
Grupo grande/Teoría: 6h
Actividades dirigidas: 1h
Aprendizaje autónomo: 9h

Basics of probability theory for intelligent data analysis

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

Objetivos específicos:
5

Competencias relacionadas:
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 16h
Grupo grande/Teoría: 6h
Actividades dirigidas: 1h
Aprendizaje autónomo: 9h

Statistical Machine Learning methods

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

Objetivos específicos:
5, 6

Competencias relacionadas:
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 31h
Grupo grande/Teoría: 12h
Actividades dirigidas: 1h
Aprendizaje autónomo: 18h

SML in data visualization, with case studies

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

Objetivos específicos:
4, 7, 8

Competencias relacionadas:
CG2. Identificar y aplicar métodos de análisis, extracción de conocimiento y visualización de datos recogidos en formatos muy diversos.
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE10. Identificar los métodos de aprendizaje automático y modelización estadística a utilizar para resolver un problema específico de ciencia de datos y aplicarlos de forma rigurosa
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE8. Extraer información de datos estructurados y no estructurados, teniendo en cuenta la naturaleza multivariante de los mismos.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 25h
Grupo grande/Teoría: 9h
Actividades dirigidas: 1h
Aprendizaje autónomo: 15h

SISTEMA DE CALIFICACIÓN

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

BIBLIOGRAFÍA

Básica:
- 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.

Complementaria:
- 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.