UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMATICA

Ficha del curso: 2022-2023

Grado: GRADO EN INGENIERÍA DE COMPUTADORES	**Curso**: Optativas 3º y 4º **(** 2C **)**	Idioma: **Español**
Asignatura: 803347 - Minería de datos y el paradigma Big Data **Asignatura en Inglés**: Data mining and the Big Data paradigm	**Abrev:** MIN **Carácter:** Optativa	6 **ECTS**
Materia: Complementos de gestión y procesamiento de la información		24 **ECTS**
Otras asignaturas en la misma materia: Análisis de redes sociales Bases de Datos noSQL Gestión de la información en la web		 6 ECTS 6 ECTS 6 ECTS
Módulo: Optativo		
Departamento: Sistemas Informáticos y Computación	**Coordinador:** Requeno Jarabo, José Ignacio	

Descripción de contenidos mínimos:

-Conceptos básicos y/o claves de la inteligencia del negocio o inteligencia empresarial (business intelligence)

-Paradigmas complementarios al modelo relacional para la representación de datos, información y conocimiento: modelo multidimensional y almacenes de datos (data warehouse); datos abiertos y/o enlazados; datos textuales, semiestructurados y/o no estructurados;grandes volúmenes de datos ("big data").

-Principales técnicas de obtención, representación, procesamiento y explotación de los paradigmas mencionados anteriormente.

-Minería de datos: Metodologías, procesos, técnicas y modelos principales; campos de aplicación.

Programa detallado:

La expansión de la WWW y el crecimiento exponencial en la capacidad de almacenamiento y procesamiento de los sistemas de información actuales han abierto nuevas vías para la representación y explotación de datos impensables hasta no hace muchos años.

De esta forma, han surgido recientemente paradigmas y conceptos como "data warehouse", datos abiertos, datos enlazados, datos no estructurados o textuales, o "big data", que se han unido a los tradicionales datos del paradigma relacional en las necesidades de almacenamiento y procesamiento de los datos y la información, sobre todo del mundo empresarial y de los negocios. Estas necesidades han venido impulsadas en gran medida, por ejemplo, por los requisitos que plantea una gestión de la Inteligencia del Negocio más moderna y automatizada.

Acompañando a esta evolución en los paradigmas de representación y/o almacenamiento de los datos y la información han surgido también nuevas formas de procesamiento y/o explotación de los mismos. Entre ellas figura, de manera privilegiada, la minería de datos, con sus propias metodologías (como KDD, SEMMA o CRISP-DM) y técnicas (como clustering o asociación) relativamente consolidadas, para la deducción y/o extracción automática de información y conocimiento de estos.

La asignatura contextualiza y presenta todos estos nuevos paradigmas de representación de datos y de la información, junto con los problemas que plantean y las soluciones provistas hasta la fecha para su solución, en forma de metodologías, técnicas y buenas prácticas para el desarrollo y explotación de sistemas de información (es decir, de minería de datos y de textos) apropiados.

OBJETIVOS:

El objetivo de esta asignatura es familiarizar al alumnado con los distintos paradigmas actuales de representación de datos, prestando especial atención a aquellos que posibilitan o implican el manejo de grandes volúmenes de datos (big data), así como con las principales metodologías, técnicas y buenas prácticas existentes para su manejo, procesamiento y/o explotación dentro del campo de la minería de datos.

TEMARIO:

1. Introducción y conceptos básicos: inteligencia del negocio; paradigmas de representación de datos, información y conocimiento (evolución de las bases de datos); campos de aplicación de la minería de datos; técnicas de minería de datos.

2.- Almacenamiento de datos: paradigmas complementarios al modelo relacional.
2.1.- El paradigma de los "data warehouse".
2.1.1.- Introducción a los almacenes de datos.
2.1.2.- El modelo multidimensional: esquemas multidimensionales, cuboides y operaciones.
2.1.3.- Arquitecturas multidimensionales: OLAP versus OLTP; ROLAP y otras formas de gestión.
2.2.- El paradigma de los datos abiertos y/o enlazados: archivos CSV, XML, RDF y derivados.
2.4.- El paradigma de los datos textuales y/o no estructurados: problemas y procesos clave en el procesamiento del lenguaje natural (PLN).
2.5.- El paradigma de los grandes volúmenes de datos ("big data"). Bases de datos distribuidas y/o NoSQL: introducción a MapReduce, Hadoop, MongoDB y/o BigTable.

3.- Minería de datos.
3.1.- Introducción y conceptos básicos.

Fecha: ____ de _____ de _____

Firma del Director del Departamento:

3.2.- Metodologías de desarrollo de sistemas de minería de datos: KDD, SEMMA y/o CRISP-DM.
3.3.- Procesos y técnicas claves en el desarrollo de sistemas de minería de datos.
3.3.1.- Importación, preprocesamiento y exportación de datos; herramientas ETL (extract, transfer and load).
3.3.2.- Exploración y visualización básica de datos.
3.3.3.- Modelos de transformación y/o procesamiento de datos.
3.3.3.1.- Árboles de decisión y bosques aleatorios.
3.3.3.2.- Regresión.
3.3.3.3.- Agrupamiento.
3.3.3.4.- Reglas de asociación.
3.3.4.- Minería de textos.
3.3.4.1.- Adquisición de datos textuales: web crawling y minería de páginas web y redes sociales.
3.3.4.2.- Procesamiento automático de textos: procesos de lematización, radicalización, etc.
3.3.5.- Minería de grafos: explotación y aplicaciones de los repositorios de datos enlazados. Introducción a SPARQL.
3.3.6.- Evaluación.

Programa detallado en inglés:

The expansion of the WWW and the exponential growth in storage capacity and processing capabilities of current information systems has given rise to the creation of new data representation formats and data exploitation means that were inconceivable some years ago.

In this way, some new data-related concepts and their related paradigms have recently emerged, such as "data warehouse", open data, linked data, unstructured or textual data, or "big data". All of them, together with the traditional relational data paradigm, are currently involved in the data and information storage and processing needs of business and enterprises. These needs have been driven, to a large extent, for example, by the requirements posed by a more modern and automated Business Intelligence management.

In order to make up for this evolution in the paradigms and formats of representation and/or storage of data and information, some new means for their processing and/or exploitation have also emerged. The main one is, most probably, data mining, which has already developed its own methodologies (such as KDD, SEMMA or CRISP-DM) and techniques (such as clustering or association rules) for the deduction and/or automatic extraction of information and knowledge from these data formats.

This course seeks to contextualize and present all these new data and information representation formats and paradigms, as well as the problems they pose and the solutions found to date to solve them, by means of suitable methodologies, techniques and good practices for the development and exploitation of information systems (i.e., data and text mining).

GOALS:

The main goal of this course is to make students aware of
- the current data representation paradigms, paying special attention to those that involve big data management and/or processing, as well as
- the main methodologies, techniques and good practices identified so far for their management, processing and/or exploitation within the data mining field.

CONTENTS:

1. Introduction and basic concepts: business intelligence; data, information and knowledge representation paradigms (evolution of databases); areas of data mining applications; data mining techniques.

2.- Data storage: relational model supplementary paradigms.
2.1.- The data warehouse paradigm.
2.1.1.- Introduction to data warehouses.
2.1.2.- The multidimensional model: multidimensional schemes, cuboids and operations.
2.1.3.- Multidimensional architectures: OLAP vs. OLTP; ROLAP and other management techniques.
2.2.- The open and/or linked data paradigm: CSV, XML and RDF files (and derivatives).
2.4.- The textual and/or unstructured data paradigm: problems and key processes in natural language processing (NLP).
2.5.- The big data paradigm. Distributed and/or NoSQL databases: introduction to MapReduce, Hadoop, MongoDB and/or BigTable.

3.- Data mining.
3.1.- Introduction and basic concepts.
3.2.- Methodologies for data mining system development: KDD, SEMMA and/or CRISP-DM.
3.3.- Key processes and techniques in data mining system development.
3.3.1.- Data import, preprocessing and export; ETL tools (extract, transfer and load).
3.3.2.- Data exploration and basic visualization.
3.3.3.- Data transformation models and/or processing.
3.3.3.1.- Decision trees and random forests.

Fecha: ____ de _____ de _____

Firma del Director del Departamento:

3.3.3.2.- Regression.
3.3.3.3.- Clustering.
3.3.3.4.- Association rules.
3.3.4.- Text mining.
3.3.4.1.- Text acquisition: web crawling, web page and social network mining.
3.3.4.2.- Automatic word and text processing: lemmatization, stemming, etc.
3.3.5.- Graph mining: exploitation and applications of linked data repositories. Introduction to SPARQL.
3.3.6.- Assessment.

Competencias de la asignatura:

Generales:

CG6-Conocimiento adecuado del concepto de empresa, marco institucional y jurídico de la empresa. Organización y gestión de empresas.

CG17-Conocimiento y aplicación de las características, funcionalidades y estructura de las bases de datos, que permitan su adecuado uso, y el diseño y el análisis e implementación de aplicaciones basadas en ellos.

CG18-Conocimiento y aplicación de las herramientas necesarias para el almacenamiento, procesamiento y acceso a los Sistemas de información, incluidos los basados en web.

CG23-Conocimiento y aplicación de los principios fundamentales y técnicas básicas de los sistemas inteligentes y su aplicación práctica.

Específicas:

CE_C1-Capacidad para tener un conocimiento profundo de los principios fundamentales y modelos de la computación y saberlos aplicar para interpretar, seleccionar, valorar, modelar, y crear nuevos conceptos, teorías, usos y desarrollos tecnológicos relacionados con la informática.

CE_C4-Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas en cualquier ámbito de aplicación.

CE_C5-Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.

CE_C7-Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.

CE_TI1-Capacidad para comprender el entorno de una organización y sus necesidades en el ámbito de las tecnologías de la información y las comunicaciones.

Básicas y Transversales:

CT1-Capacidad de comunicación oral y escrita, en inglés y español utilizando los medios audiovisuales habituales, y para trabajar en equipos multidisciplinares y en contextos internacionales.

CT2-Capacidad de análisis y síntesis en la resolución de problemas.

Resultados de aprendizaje:

Comprensión de los aspectos básicos empresariales y del negocio que afectan al diseño y desarrollo de sistemas de información y/o de gestión de (bases de) datos, información y conocimientos. (CG6)

Capacidad para identificar y analizar los elementos de la lógica del negocio que afectan al proceso de diseño y desarrollo de un sistema de inteligencia empresarial (business intelligence), minería de datos y/o de explotación de big data e integrarlos, con una metodología adecuada, en dicho proceso. (CG6, CT2, CE_C5, CE_TI1)

Representar y procesar datos, información y conocimientos de forma integrada y conveniente usando metodologías, procesos, modelos propios de las áreas de la minería de datos (incluyendo la minería de textos) y del campo de los big data. (CG17, CG18, CG23, CT2, CE_C1, CE_C4, CE_C5, CE_C7)

Conocer, aplicar y evaluar distintas técnicas y modelos de minería de datos para resolver un problema concreto. (CG23, CT2, CE_C1, CE_C7)

Entender las componentes básicas de un sistema de inteligencia empresarial en general y, más concretamente, de un sistema de minería de datos, para su aplicación en el desarrollo de sistemas y aplicaciones en este ámbito. (CG17, CG18, CE_C5, CE_TI1)

Elegir la(s) representación(es) más adecuada(s) del problema para aplicar las técnicas y modelos de minería de datos que lo resuelvan. (CG17, CG18, CT2)

Fecha: ____ de _____ de _____

Firma del Director del Departamento:

Diseñar e implementar (preferentemente, mediante grupos de trabajo) un sistema de minería de datos y/o de procesamiento de grandes volúmenes de datos (big data) utilizando metodologías, procesos, técnicas y modelos adecuados de estas áreas. (CG17, CG18, CG23, CT1, CT2, CE_C1, CE_C7)

Aumento y/o mejora de la capacidad de trabajo en equipo y de realización de presentaciones orales. (CT1, CT2)

Evaluación detallada:

En ambas convocatorias (ordinaria y extraordinaria) la realización de las prácticas y/o proyectos es obligatoria. La(s) práctica(s) se entregarán individualmente o en grupo. La realización y entrega de la(s) práctica(s) puede conllevar su defensa (parcial y/o total) pública o en presencia del profesor.

La nota final se calculará de acuerdo a los siguientes porcentajes:

* 20% participación en clase: proactividad y colaboración en el desarrollo de las sesiones presenciales y de las tareas colectivas, resolución de ejercicios y cuestionarios, etc.
* 80% práctica(s) obligatoria(s).

- No entregar las prácticas en el plazo establecido supondrá el suspenso en la asignatura (no se calculará la media con el resto de elementos de la evaluación) en la convocatoria ordinaria.
- Existe la posibilidad de entregar las prácticas en la convocatoria extraordinaria, manteniéndose para la convocatoria extraordinaria la calificación correspondiente a la participación en clase y el resto de tareas aprobadas.

Actividades docentes:

Reparto de créditos:
Teoría: 3,00
Problemas: 0,00
Laboratorios: 3,00

Otras actividades:
No tiene

Bibliografía:

- Russell, Matthew A. (2014) Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More (2nd Edition). Sebastopol (CA, EE.UU.): O'Reilly.

- Witten, Ian H., Frank, Eibe (2005) Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition). San Francisco: Morgan Kaufmann Publishers (Elsevier).

- Zhao, Yanchang (2013) R and Data Mining: Examples and Case Studies. San Diego, Waltham, Londres, Amsterdam: Academic Press (Elsevier).

- Dietrich, D. (2015). Data science and big data analytics: discovering, analyzing, visualizing and presenting data. John Wiley & Sons.

Ficha docente guardada por última vez el **11/06/2021 16:50:00** por el departamento: **Sistemas Informáticos y Computación**

Fecha: ____ de _____ de _____

Firma del Director del Departamento: