PLAN DOCENTE DE LA ASIGNATURA

Curso académico: 2024/2025

Identificación y características de la asignatura				
Código	502302		Créditos ECTS	6
Denominación (español)	Minería de Datos y Almacenes de Datos			
Denominación (inglés)	*Data Mining and Data Warehouses*			
Titulaciones	Grado en Ingeniería Informática en Ingeniería del Software			
Centro	Escuela Politécnica			
Semestre	8	Carácter	Optativa	
Módulo	Optatividad en Ingeniería del Software			
Materia	Ingeniería Multimedia			
Profesor/es				
Nombre	Despacho		Correo-e	Página web
Félix Rodríguez Rodríguez	23 (Edif.Telecos)		felixr@unex.es	Web EPCC (felixr) y madiba.unex.es
Área de conocimiento	Lenguajes y Sistemas Informáticos			
Departamento	Ingeniería de Sistemas Informáticos y Telemáticos			
Profesor coordinador (si hay más de uno)				

Competencias

Competencias básicas:

- **CB1**: Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.
- **CB2**: Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio.
- **CB3**: Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.
- **CB4**: Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.
- **CB5**: Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

Competencias generales:

- **CG01**: Capacidad para concebir, redactar, organizar, planificar, desarrollar y firmar proyectos en el ámbito de la ingeniería en informática que tengan por objeto, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería

del Software, la concepción, el desarrollo o la explotación de sistemas, servicios y aplicaciones informáticas.

- **CG02**: Capacidad para dirigir las actividades objeto de los proyectos del ámbito de la Informática.
- **CG03**: Capacidad para diseñar, desarrollar, evaluar y asegurar la accesibilidad, ergonomía, usabilidad y seguridad de los sistemas, servicios y aplicaciones informáticas, así como de la información que gestionan.
- **CG04**: Capacidad para definir, evaluar y seleccionar plataformas hardware y software para el desarrollo y la ejecución de sistemas, servicios y aplicaciones informáticas, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería del Software.
- **CG05**: Capacidad para concebir, desarrollar y mantener sistemas, servicios y aplicaciones informáticas empleando los métodos de la ingeniería del software como instrumento para el aseguramiento de su calidad, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería del Software.
- **CG06**: Capacidad para concebir y desarrollar sistemas o arquitecturas informáticas centralizadas o distribuidas integrando hardware, software y redes.
- **CG07**: Capacidad para conocer, comprender y aplicar la legislación necesaria durante el desarrollo de la profesión de Ingeniero Técnico en Informática y manejar especificaciones, reglamentos y normas de obligado cumplimiento.
- **CG08**: Conocimiento de las materias básicas y tecnologías, que capaciten para el aprendizaje y desarrollo de nuevos métodos y tecnologías, así como las que les doten de una gran versatilidad para adaptarse a nuevas situaciones.
- **CG09**: Capacidad para resolver problemas con iniciativa, toma de decisiones, autonomía y creatividad. Capacidad para saber comunicar y transmitir los conocimientos, habilidades y destrezas de la profesión de Ingeniero Técnico en Informática.
- **CG10**: Conocimientos para la realización de mediciones, cálculos, valoraciones, tasaciones, peritaciones, estudios, informes, planificación de tareas y otros trabajos análogos de informática, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería del Software.
- **CG11**: Capacidad para analizar y valorar el impacto social y medioambiental de las soluciones técnicas, comprendiendo la responsabilidad ética y profesional de la actividad del Ingeniero Técnico en Informática.
- **CG12**: Conocimiento y aplicación de elementos básicos de economía y de gestión de recursos humanos, organización y planificación de proyectos, así como la legislación, regulación y normalización en el ámbito de los proyectos informáticos.

Competencias Ingeniería del Software:

- **CIS01**: Capacidad para desarrollar, mantener y evaluar servicios y sistemas software que satisfagan todos los requisitos del usuario y se comporten de forma fiable y eficiente, sean asequibles de desarrollar y mantener y cumplan normas de calidad, aplicando las teorías, principios, métodos y prácticas de la Ingeniería del Software.
- **CIS02**: Capacidad para valorar las necesidades del cliente y especificar los requisitos software para satisfacer estas necesidades, reconciliando objetivos en conflicto mediante la búsqueda de compromisos aceptables dentro de las limitaciones derivadas del coste, del tiempo, de la existencia de sistemas ya desarrollados y de las propias organizaciones.
- **CIS03**: Capacidad de dar solución a problemas de integración en función de las estrategias, estándares y tecnologías disponibles.

- **CIS04**: Capacidad de identificar y analizar problemas y diseñar, desarrollar, implementar, verificar y documentar soluciones software sobre la base de un conocimiento adecuado de las teorías, modelos y técnicas actuales.
- **CIS05**: Capacidad de identificar, evaluar y gestionar los riesgos potenciales asociados que pudieran presentarse.
- **CIS06**: Capacidad para diseñar soluciones apropiadas en uno o más dominios de aplicación utilizando métodos de la ingeniería del software que integren aspectos éticos, sociales, legales y económicos.

Contenidos

Breve descripción del contenido

Fundamentos de la extracción automática de conocimiento. Visualización de la información. Tratamiento masivo de datos. Preparación de datos. Técnicas de extracción y minado de datos. Minado de datos complejos. Almacenes de datos.

Temario de la asignatura

1. Introducción al Descubrimiento de Conocimiento, KDD.

1.1. Motivación. Modelo de Descubrimiento de Conocimiento en Almacenes y Bases de Datos (KDD). Evolución en el tiempo y áreas de aplicación.
1.2. Almacenes de Datos, *Data Warehouses* (DW): objetivos, definiciones, arquitecturas y visiones. Sistemas Gestores de Bases de Datos (SGBD) *vs* DW. *Data Marts, Data Lakes,* Almacenes de datos especializados. Roles, estructuras, integración de datos, coste.
1.3. Minería de Datos, *Data Mining* (DM): orígenes, motivación, objetivos y tareas. Áreas de aplicación. Disciplinas involucradas.
1.4. Fases KDD: (i) Técnicas ETL: Selección, Limpieza, Transformación (integración, reducción, enriquecimiento, refinamiento) y Carga de datos; (ii) Minado de datos; (iii) Evaluación, Interpretación y Visualización de resultados.
1.5. Taxonomía y descripción general de las técnicas de *Data Mining*. Técnicas descriptivas *vs* predictivas. Aprendizaje Supervisado, No Supervisado y por Refuerzo (*Supervised*, *Unsupervised*, *& Reinforcement Learning*).
1.6. *Data Mining* desde el punto de vista de la Recuperación Basada en el Contenido.
1.7. Retos de la Minería de datos.

2. *Big Data*. Bases de Datos NoSQL.

2.1. *Big Data*. Definición, extensión, arquitecturas y uso. Esquema *MapReduce* para aplicaciones distribuidas y *Big Data*. *Apache Hadoop* y *Spark*.
2.1. Almacenamiento de datos *NoSQL*. Definición, uso y esquemas. *ACID* vs *BASE*. Teorema *CAP* de Sistemas Distribuidos. Persistencia políglota en aplicaciones.
2.2. NoSQL *Data Warehousing*. Arquitecturas. Modelos de datos *NoSQL*, Agregados. Familias *NoSQL*: clave-valor, de columnas, documental, grafo, multimodelo y otras.
2.3. Desarrollo de una aplicación específica bajo un modelo de almacenamiento *NoSQL* (*Aerospike, ArangoDB, ArcadeDB, Cassandra, CouchBase y CouchDB, DynamoDB y AWS, Elastic, Firestore, HBase, MongoDB, Neo4J, Redis, RethinkDB, Riak* y otras).

Laboratorio: Estudio completo de dos BD *NoSQL* por grupos de dos estudiantes y desarrollo de una aplicación dedicada. Pipeline distribuido con *Hadoop* (o *Spark*).

3. Visualización de Datos: *Data Visualization* & *Visual Data Mining*.

3.1. Visualización y Análisis Exploratorio de Datos. Percepción visual humana y Visualización de Datos.
3.2. Excelencia Gráfica y Factor Mentira de Tufte.

3.3. Representación de datos 1D, 2D y 3D. Histogramas, suavización, diagramas de cajas (*boxplots*), gráficas de dispersión (*scatterplots*) y dispersión homocedástica, mezclado, trazado trasparente, trepidación (*jittering*), funciones, contornos, y proyecciones 2D y 3D, entre otras.
3.4. Representación de datos temporales, espaciales y espacio-temporales. Sistemas de Información Geográfica (SIG).
3.5. Representación de datos de alta dimensionalidad. Matrices de gráficas de dispersión, coordenadas paralelas, gráficas en estrella, icónicas, de caras, mosaicos, redes, grafos, entre otras.
3.6. Implementaciones.
3.7. Visualización de Datos con *WEKA*, *Python* y otros entornos.

Laboratorio: Implementación de al menos dos métodos de visualización por grupos de dos estudiantes.

4. Aprendizaje Supervisado: Clasificación (*Classification*).

4.1. Aprendizaje supervisado y el problema de la clasificación. Conceptos básicos de la clasificación. Clasificación lineal simple.
4.2. Modelos de evaluación y validación. Comparación de métodos: Precisión predictiva; velocidad y escalabilidad; robustez; interpretabilidad.
4.3. Clasificadores de vecindad próxima. *K-NN*. Medidas de distancia. Distancia editada.
4.4. Clasificación mediante Árboles de decisión. Criterios de partición: Ganancia de Información. Algoritmos *ID3*, *C4.5* y *J48*. Eliminación de sobrecargas. Ventajas e inconvenientes.
4.5. Métodos de Clasificación Bayesianas. Clasificadores *Naïve-Bayes*. Clasificación bayesiana multiclase.
4.6. Clasificación basada en reglas. Inducción de reglas de árboles de decisión y por métodos de cobertura secuencial.
4.7. Comparación de algoritmos de Clasificación basadas en precisión predictiva, velocidad y escalabilidad, robustez e interpretabilidad.
4.8. Perceptrones. Caso del *spam*. Algoritmo de *Winnow*. Perceptrón multiclase. Limitaciones.
4.9. Máquinas de Vectores de Soporte (*Support Vector Machines*, SVM). *SVM* multiclase: métodos *OvO* y *OvA*.
4.10. Redes Neuronales (*Neural Networks*, NN) y Aprendizaje Profundo (*Deep Learning*, DL). Aprendizaje y retropropagación (*backpropagation*). Esquemas *FFNN*, *CNN*, *RNN*, *LSTM* y otros. Subajuste (*undeerfitting*) y sobreajuste (*overfitting*). Ventajas e inconvenientes. Implementaciones *WEKA*, *Scikit*, *TensorFlow* y *Keras*.
4.11. Modelos de Lenguaje de Gran Tamaño (*Large Language Models*, *LLM*) y Redes Generativas Adversas (*Generative Adversarial Networks*, *GAN*).
4.12. Métodos ensamblados: *Bagging, Boosting, Random Forest*. Problemas de muestreo : sobremuestreo (*oversampling*) y submuestreo (*undersamplig*),
4.13. Métodos de clasificación con *WEKA*, *Python* y otros entornos.

Laboratorio: Implementación de al menos un método de clasificación por grupos de dos estudiantes orientados a una aplicación específica propuesta.

5. Aprendizaje no Supervisado: Agrupación o *Clustering*.

5.1. Aprendizaje no supervisado y el problema del *Clustering*. Conceptos básicos del *Clustering*. Similitud, espacios y medidas de distancia. Medidas de calidad y evaluación.
5.2. *Clustering* Jerárquico. Dendogramas. Métodos divisivos y aglomerativos. Algoritmos *Diana* y *Agnes*.

5.3. *Clustering* mediante Particionamiento. Algoritmos *K-Means*, *K-Medoids* (*PAM*), *Clarans*, *EM*, *CURE*. Partición mediante *R-trees*: algoritmo de *Birch*.
5.4. Métodos basados en la densidad. Algoritmos *DBScan*, *Optics*.
5.5. *Clustering NN*, agrupaciones mediante el vecino más próximo.
5.6. Métodos de *Clustering* con *WEKA*, *Python* y otros entornos.
Laboratorio: Implementación de al menos un método de clustering por grupos de dos estudiantes orientados a una aplicación específica propuesta.

6. Minado de Patrones Frecuentes (*Frequent Patterns Mining*). Asociaciones.
6.1. Patrones Frecuentes (*Frequent Patterns*, FP). Definición. Motivación. Aplicaciones.
6.2. Análisis de *FP*. Conceptos básicos: Asociaciones y Reglas de asociación, patrones cerrados y max-patrones. Complejidad.
6.3. Algoritmo *A-priori*. Mejoras: *A-priori* por partición, por *DHP*, muestreo, por *DIC* .
6.4. Algoritmo de crecimiento de patrones frecuentes, *FP-Growth*.
6.5. *FP-Growth* vs *A-priori*.
6.6. Métodos de asociación con *WEKA*, *Python* y otros entornos.
Laboratorio: Implementación de al menos un método de FPM por grupos de dos estudiantes orientados a una aplicación específica propuesta.

7. *Data Warehousing* con Bases de Datos Relacionales: *OLAP*.
7.1. Sistemas de ayuda a la toma de decisiones y Almacenes de datos *DW*.
7.2. Procesamiento Analítico en Línea y Análisis Multidimensional de los Datos. *OLTP vs OLAP*. Arquitectura multicapa. Modelos. *ROLAP, MOLAP* y *HOLAP*.
7.3. ETL: Extracción, Transformación y Carga de datos. Metadatos y Repositorios.
7.4. Modelado de *DW OLAP*: Cubos de datos (*data cubes*). Tablas de dimensiones y de hechos. Esquemas de modelización. Atributos, medidas y jerarquías. Modelos de construcción de cubos de datos. Operaciones típicas *OLAP*. Generalización de datos mediante inducción orientada a atributos. Implementación y Administración. *Data Mining* vs *OLAP*.
7.5. Alta Dimensionalidad. Reducción de la dimensionalidad. Peligros.
7.6. *DW* con *MS SQL Server Data Tools* (*SSDT*).
Laboratorio: Desarrollo opcional de un cubo de datos.

Actividades formativas

Horas de trabajo del alumno por tema		Horas teóricas	Actividades prácticas				Actividad de seguimiento	No presencial
Tema	**Total**	**GG**	**PCH**	**LAB**	**ORD**	**SEM**	**TP**	**EP**
1	4	2		0			0	2
2	46	2		8			1	35
3	6	2		2			0	2
4	54	12		12			0	30
5	21	4		4			1	12
6	6	2		2			0	2
7	9	2		2			1	4
Evaluación **	4	4						
TOTAL	150	30		30			3	87

GG: Grupo Grande (85 estudiantes).
PCH: prácticas clínicas hospitalarias (7 estudiantes)
LAB: prácticas laboratorio o campo (15 estudiantes)
ORD: prácticas sala ordenador o laboratorio de idiomas (20 estudiantes)

** Número total de horas de evaluación de esta asignatura.

SEM: clases problemas o seminarios o casos prácticos (40 estudiantes).
TP: Tutorías Programadas (seguimiento docente, tipo tutorías ECTS).
EP: Estudio personal, trabajos individuales o en grupo, y lectura de bibliografía.

Metodologías docentes

Clases teórico-prácticas en el aula:
Clases expositivas para el desarrollo de los contenidos fundamentales de las materias. Se incluyen actividades breves, individuales o en grupo que permitan aplicar los conceptos expuestos y resolver problemas, facilitando la participación activa de los estudiantes.

Sesiones de laboratorio:
Actividades prácticas, sesiones de laboratorio guiadas, seminarios de resolución de problemas, etc., en grupos bajo la dirección del profesor. Se podrán incluir actividades previas y posteriores a las sesiones de laboratorio y seminario que ayuden a conseguir los objetivos propuestos.

Realización de actividades, trabajos y estudio por parte del estudiante, de manera autónoma, individualmente o en grupo:
Las actividades que el estudiante desarrollará de manera no presencial estarán orientadas principalmente a la adquisición de conocimientos básicos en el ámbito de la Informática y al desarrollo de los proyectos y trabajos solicitados, bien individualmente o en grupo.

De manera más específica, la metodología docente se centra en:

- Clases teóricas con apoyo de material audiovisual, presentándose los contenidos básicos de cada tema. Tras una breve introducción de los contenidos que se pretenden desarrollar en la clase, y de los comentarios oportunos de los conceptos asumidos en clases anteriores, el desarrollo docente se realiza apoyado con transparencias, textos bibliográficos y artículos de interés y videos disponibles de expertos en el campo a tratar. Se motiva a los estudiantes a intervenir en cualquier momento para hacer las clases más dinámicas y facilitar el aprendizaje. Los temas terminan con la exposición de las conclusiones más relevantes de los conceptos tratados.
- Trabajos de investigación realizados por los estudiantes sobre conceptos de *Data Mining*, *Data Warehousing*, *Big Data* y Bases de datos NoSQL. Los trabajos se desarrollarán en grupos de dos estudiantes generalmente.
- Presentación oral (máximo 15 minutos por grupo) y defensa de algunos de los trabajos propuestos por el profesor por parte de todos los componentes de los grupos de trabajo.
- Tutorías (con opción a ser virtuales) para consultar cualquier duda relacionada con los contenidos, organización y planificación de la materia.
- Campus Virtual y Blogs personales de los estudiantes y de la asignatura. En el Campus Virtual se dejará todo el material utilizado y necesario para el desarrollo del curso, así como las indicaciones de los trabajos a realizar, sus cometidos y plazos de entrega en el Blog de la asignatura. Asimismo, cada estudiante irá plasmando la evolución de sus trabajos y conocimientos en un blog individual dependiente del blog principal de la asignatura. Estos blogs de los estudiantes servirán para ser evaluados y comentados por el resto de los estudiantes como un pilar fundamental de la metodología docente de esta asignatura.

Resultados de aprendizaje

Al finalizar el curso satisfactoriamente, el alumno podrá:

- Reconocer el potencial en el análisis de los sistemas de información para la ayuda a la toma de decisiones.
- Conocer los fundamentos del almacenamiento de datos, sus diferentes arquitecturas, modelos e implementaciones, diferenciándose entre el almacenamiento relacional tradicional, la utilización tabular y cubos de datos y el nuevo paradigma NoSQL para datos a gran escala.
- Conocer y diferenciar distintas técnicas de aprendizaje supervisado y no supervisado, centradas primordialmente en las técnicas más importantes de *Clasificación* y en las de *Clustering o agrupamiento*. También conocerá técnicas de asociación y de aprendizaje por refuerzo.
- Reconocer la importancia de la visualización de datos para la interpretación de los resultados, la excelencia gráfica y el *factor mentira* así como las propuestas fundamentales de visualización de datos 1D, 2D y 3D, espaciales y espacio-temporales, así como los de alta dimensionalidad.
- Distinguir entre el procesamiento analítico online y el análisis multidimensional de los datos.

Sistemas de evaluación

Según la normativa de evaluación vigente en la Universidad de Extremadura, los estudiantes podrán superar la asignatura mediante la modalidad de evaluación continua o mediante la modalidad de evaluación global.
Cada estudiante deberá decidir a qué tipo de evaluación se acoge en cada una de las convocatorias del curso y comunicarlo al profesorado durante el primer cuarto del periodo de impartición de la asignatura. Dicha comunicación se realizará mediante una consulta que estará disponible en el aula virtual. Si no se realiza esta comunicación en el plazo establecido se entenderá que el estudiante opta por la evaluación continua.

Evaluación continua:
La opción deseable para llevar a cabo plenamente la metodología docente utilizada. Se valora la asistencia presencial a las clases teórico-prácticas en el aula. En las mismas y en las sesiones de laboratorio se propondrán la realización de trabajos prácticos (individuales y en grupo) y la presentación y exposición de ciertos trabajos específicos planteados.

El trabajo de cada estudiante quedará reflejado en un *blog* docente individual ligado a la asignatura, que se irá construyendo a lo largo del curso. Paralelamente, los blogs de los estudiantes servirán para que los estudiantes coevalúen el trabajo realizado entre ellos. La evaluación final se decidirá por parte del profesor basándose en la ponderación de todos los trabajos realizados, su presentación, el blog construido y en el contenido justificado de las coevaluaciones de los estudiantes. Cada uno de estos trabajos será evaluado individualmente y la calificación será la proporcionada por la media de todas las calificaciones obtenidas.

Evaluación global:
Tanto si el profesor estima muy baja la participación por parte del estudiante a lo largo del curso, como si el estudiante decide ser evaluado sin utilizar evaluación continua durante el curso ordinario, así como para cualquier convocatoria extraordinaria, cabe la evaluación basada en el resultado de calificar el 50% de su nota mediante un examen final. El otro 50% de la nota vendrá dada por la realización y entrega de todos los trabajos obligatorios propuestos durante el curso. El estudiante deberá realizar y

entregar todos estos trabajos antes de la fecha de examen indicada en el calendario de exámenes fijado por la dirección del centro.

Sistema de revisión de exámenes y pruebas de evaluación continua:
La revisión de exámenes y pruebas se realizará de acuerdo a la normativa de evaluación vigente.
Para el examen realizado en cada convocatoria oficial, el alumno podrá comentar y revisar los resultados del mismo en las fechas y horarios que se indiquen en la publicación de las calificaciones provisionales.
Para el resto de pruebas de evaluación continua que se realicen durante el semestre, la revisión se realizará en el plazo máximo de 10 días hábiles después de la publicación de la calificación. Dependiendo del tipo de prueba, esta revisión podrá realizarse en el horario de tutorías de libre acceso del profesor o en un horario especificado de manera particular para la revisión de dicha prueba.

Recomendación importante:
Se recomienda la asistencia a todas las clases presenciales y a las tutorías, dado el alto contenido práctico de la asignatura. Sobretodo porque comprender bien todos los conceptos impartidos van a permitir realizar los ejercicios prácticos de manera más aplicada. La participación continuada en el estudio y el desarrollo de los contenidos, sin dejar para el final la materia, se considera la vía idónea para lograr los objetivos de aprendizaje.

Bibliografía (básica y complementaria)

Básica: (por orden de recomendación)

[1] Witten, I., Frank, E., Hall, M. A., & Pal, C. J. (2017). "*Data Mining. Practical Machine Learning Tools and Techniques*", 4th ed., Morgan Kaufmann.
[2] Leskovec, J., Rajaraman, A. & Ullman, J. D. (2020) "*Mining of Massive Datasets*", 3th ed., Cambridge.

Complementaria:

[3] Chollet, F. (2022). "*Deep Learning with Python*", 2nd ed., Manning.
[4] Géron, A. (2022). "*Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*", 3rd ed. O'Reilly.
[5] Torres, J. (2020). " *Python Deep Learning with Python. Introducción práctica con Keras y TensorFlow*", 2º ed., Marcombo.
[6] White, T. (2015). "*Hadoop: The Definitive Guide*", 4th ed., O'Reilly.

Otros recursos y materiales docentes complementarios

Recursos de laboratorio y trabajo no presencial dejados o enlazados en el Campus Virtual y en el Blog principal de la asignatura.