

Grado en Ingeniería Informática

DATOS DE LA ASIGNATURA

| | | | | |
|--------------------------------|-------------------------|---|---------------------------|----------------------------|
| Nombre: | | | | |
| Minería de Datos | | | | |
| Denominación en inglés: | | | | |
| Data Mining | | | | |
| Código: | | Carácter: | | |
| 606010307 | | Optativo | | |
| Horas: | | | | |
| | Totales | Presenciales | No presenciales | |
| Trabajo estimado: | 150 | 60 | 90 | |
| Créditos: | | | | |
| | Grupos reducidos | | | |
| Grupos grandes | Aula estándar | Laboratorio | Prácticas de campo | Aula de informática |
| 3 | 0 | 0 | 0 | 3 |
| Departamentos: | | Áreas de Conocimiento: | | |
| Tecnologías de la Información | | Ciencia de la Computación e Inteligencia Artificial | | |
| Curso: | | Cuatrimestre: | | |
| 4º - Cuarto | | Primer cuatrimestre | | |

DATOS DE LOS PROFESORES

| | | | |
|----------------------------|----------------|------------------|-----------------------------|
| Nombre: | E-Mail: | Teléfono: | Despacho: |
| *López Gómez, Luis Ignacio | nacho@uhu.es | 87683 | Edif. Torreumbria, Dcho. 57 |

*Profesor coordinador de la asignatura

1. Descripción de contenidos

1.1. Breve descripción (en castellano):

Estudio de las diferentes métodos y técnicas para para la construcción de modelos predictivos y descriptivos a partir del análisis de bases de datos de alta cardinalidad y alta dimensionalidad, como ayuda a la toma de decisiones.

1.2. Breve descripción (en inglés):

Study of different methods and techniques for building predictive and descriptive models based on analysis of high cardinality and high dimensionality databases, as an aid to decision making.

2. Situación de la asignatura

2.1. Contexto dentro de la titulación:

Existen distintas asignaturas en el grado relacionadas con Minería de Datos (MD). Especialmente complementaria con ella, es la de Aprendizaje Automático que tiene la orientación a la parte algorítmica y de implementación (programación) de cada técnica que se utiliza en minería de datos. En MD, el enfoque está más orientado al uso de éstas técnicas, independientemente de la implementación particular. Es decir, consideraremos el conjunto de técnicas como una biblioteca de componentes reutilizables, cada uno de los cuales será seleccionada de acuerdo a las características de la tarea que se requiere resolver. En otros casos, esta asignatura puede servir de introducción a otras asignaturas del programa de posgrado. Asignatura de carácter práctico y de importancia emergente dentro del campo denominado Soft Computing. Hoy se ha de contemplar a la MD cómo una parte del análisis de datos dentro del contexto del denominado BIG DATA.

2.2. Recomendaciones:

- Se deben tener los necesarios conceptos de aprendizaje automático, tanto a nivel simbólico cómo subsimbólico
- Tener conocimientos de Bases de Datos, fundamentalmente las basadas en el modelo relacional
- El alumno debe tener conocimientos medios-avanzados de estadística.

3. Objetivos (Expresados como resultados del aprendizaje):

- Un acercamiento al campo para el estudio de tareas y técnicas, necesarias para la toma de decisiones de problemas complejos: problemas sin solución inicial, ni algorítmica ni analítica, con datos estructurados, semiestructurados y no estructurados.
- Descripción del estado del arte de la Minería de Datos en el contexto del Big Data.
- Nuevos enfoques y técnicas de Minería de Datos para el tratamiento masivo de datos estructurados y no estructurados.
- Estudio introductorio de las principales arquitecturas para el tratamiento de datos masivos : HADOOP y SPARK.

4. Competencias a adquirir por los estudiantes

4.1. Competencias específicas:

4.2. Competencias básicas, generales o transversales:

- **CG0:** Capacidad de análisis y síntesis: Encontrar, analizar, criticar (razonamiento crítico), relacionar, estructurar y sintetizar información proveniente de diversas fuentes, así como integrar ideas y conocimientos.
- **G02:** Capacidad de comunicación oral y escrita en el ámbito académico y profesional con especial énfasis, en la redacción de documentación técnica
- **G03:** Capacidad para la resolución de problemas
- **G04:** Capacidad para tomar decisiones basadas en criterios objetivos (datos experimentales, científicos o de simulación disponibles) así como capacidad de argumentar y justificar lógicamente dichas decisiones, sabiendo aceptar otros puntos de vista
- **G05:** Capacidad de trabajo en equipo.
- **G06:** Capacidad para el aprendizaje autónomo así como iniciativa y espíritu emprendedor
- **G07:** Motivación por la calidad y la mejora continua, actuando con rigor, responsabilidad y ética profesional.
- **G08:** Capacidad para adaptarse a las tecnologías y a los futuros entornos actualizando las competencias profesionales.
- **G09:** Capacidad para innovar y generar nuevas ideas.
- **G11:** Respeto a los derechos fundamentales y de igualdad entre hombres y mujeres
- **T02:** Conocimiento y perfeccionamiento en el ámbito de las TIC's

5. Actividades Formativas y Metodologías Docentes

5.1. Actividades formativas:

- Sesiones de Teoría sobre los contenidos del Programa.
- Sesiones Prácticas en Laboratorios Especializados o en Aulas de Informática.
- Actividades Académicamente Dirigidas por el Profesorado: seminarios, conferencias, desarrollo de trabajos, debates, tutorías colectivas, actividades de evaluación y autoevaluación.

5.2. Metodologías docentes:

- Clase Magistral Participativa.
- Desarrollo de Prácticas en Laboratorios Especializados o Aulas de Informática en grupos reducidos.
- Planteamiento, Realización, Tutorización y Presentación de Trabajos.
- Evaluaciones y Exámenes.

5.3. Desarrollo y justificación:

El desarrollo de la asignatura se basará en clases teóricas y prácticas al uso y actividades académicamente dirigidas. Se planterá el ciclo de vida de un proyecto, para su análisis, implementación y discusión durante el desarrollo de la asignatura.

En las clases prácticas se utilizará el entorno Weka para el desarrollo de los distintos trabajos. Así mismo, se presentarán entornos para la introducción de las arquitecturas y software de Minería de Datos en el contexto de Big Data: HADOOP y SPARK.

6. Temario desarrollado:

1. INTRODUCCIÓN

- 1.1. El concepto de Minería de Datos
- 1.2. La minería de datos y el proceso de descubrimiento de conocimiento a partir de datos
- 1.3. Relación con otras disciplinas
- 1.4. Aplicaciones
- 1.5. Fases del proceso de extracción de conocimiento a partir de datos

2. PREPARACIÓN DE DATOS

- 2.1. Consideraciones previas generales. Los almacenes de datos.
- 2.2. Técnicas sencillas de preprocesado
 - 2.2.1. Compleción (datos faltantes)
 - 2.2.2. Limpieza de errores
 - 2.2.3. Transformación de atributos
 - 2.2.4. Escalado
 - 2.2.5. Discretización
 - 2.2.6. Numerización
- 2.3. Técnicas de reducción de la dimensionalidad I: Análisis de Componentes Principales.
- 2.4. Técnicas de reducción de la dimensionalidad II: Métodos de Filtrado y Envoltura

3. TAREAS Y TÉCNICAS DE MINERÍA DE DATOS

- 3.1. Tareas en minería de datos.
- 3.2. Correspondencia entre métodos y tareas.
- 3.3. Caracterización de las técnicas de minería de datos.
- 3.4. Técnicas de Minería de Datos
 - 3.4.1. Métodos estadísticos.
 - 3.4.2. Reglas de asociación y dependencia.
 - 3.4.3. Métodos Bayesianos.
 - 3.4.4. Árboles de Decisión y sistemas de reglas.
 - 3.4.5. Redes Neuronales Artificiales.
 - 3.4.6. Máquinas de vectores soporte.
 - 3.4.7. Extracción de conocimiento con algoritmos evolutivos y reglas difusas.
 - 3.4.8. Métodos basados en casos y vecindad.

4. EVALUACIÓN

- 4.1. Consideraciones generales.
- 4.2. Técnicas básicas de evaluación de clasificadores
 - 4.2.1. Medidas de la calidad de un clasificador: la tasa de errores
 - 4.2.2. La descomposición del error en sesgo y varianza: el concepto de generalización
 - 4.2.3. El sobreentrenamiento
 - 4.2.4. Repetibilidad estadística: la validación cruzada.
- 4.3. Aspectos específicos de la evaluación de los diferentes clasificadores estudiados
- 4.4. Técnicas estadísticas de comparación de clasificadores
- 4.5. Medidas de calidad de agrupamiento
- 4.6. Interpretación, difusión y uso de modelos

5. IMPLANTACIÓN E IMPACTO DE LA MINERÍA DE DATOS

- 5.1. Implantación de un Programa de Minería de Datos (PMD) en una organización
 - 5.1.1. Cuándo implantar un PMD: Necesidades y objetivos
 - 5.1.2. Fases de un PMD: Estándar CRISP-DM
 - 5.1.3. Integración de un PMD dentro de una organización
 - 5.1.4. Recursos necesarios
 - 5.2. Repercusiones y retos de la minería de datos
 - 5.2.1. Impacto social
 - 5.2.2. Cuestiones éticas y legales
 - 5.2.3. Problemas y soluciones: Tendencias futuras
- 6.0 Nuevos enfoques. Big Data

7. Bibliografía

7.1. Bibliografía básica:

Manual básico: Introducción a la Minería de Datos de Hernández Orallo et al. Ed. Pearson, 2004.
Data Mining, Concepts and techniques. Jiawei Han et al. Morgan Kaufmann, 2012.

7.2. Bibliografía complementaria:

Data Mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems). Ian H. Witten y Eibe Frank, 2011.
Data Mining: A Heuristic Approach. Hussein A. Abbass, Ruhul A. Sarker and Charles S. Newton. University of New South Wales, Australia (2002).

8. Sistemas y criterios de evaluación.

8.1. Sistemas de evaluación:

- Examen de teoría/problemas
- Defensa de Prácticas
- Defensa de Trabajos e Informes Escritos

8.2. Criterios de evaluación y calificación:

Examen teórico-práctico (50%), prácticas de laboratorio(40%), defensa de trabajos (10%). Es necesario aprobar con un cinco cada apartado evaluable, para el cálculo de la calificación final. La nota de prácticas de guardará hasta Setiembre, si bien habrá en su caso, un examen de la parte práctica.

9. Organización docente semanal orientativa:

| | Semanas | Grupos Grandes | Grupos Reducidos Aula Estándar | Grupos Reducidos Aula de Informática | Grupos Reducidos Laboratorio | Grupos Reducidos prácticas de campo | Pruebas y/o actividades evaluables | Contenido desarrollado |
|-----|---------|----------------|-----------------------------------|---|---------------------------------|--|---------------------------------------|------------------------|
| #1 | 2 | 0 | 2 | 0 | 0 | | Practica 1 | Tema1 |
| #2 | 2 | 0 | 2 | 0 | 0 | | | Tema1 |
| #3 | 2 | 0 | 2 | 0 | 0 | | | Tema2 |
| #4 | 2 | 0 | 2 | 0 | 0 | | Practica 2 | Tema2 |
| #5 | 2 | 0 | 2 | 0 | 0 | | Practica2 | Tema2 |
| #6 | 2 | 0 | 2 | 0 | 0 | | Practica 2 | Tema3 |
| #7 | 2 | 0 | 2 | 0 | 0 | | Test semanal | Tema3 |
| #8 | 2 | 0 | 2 | 0 | 0 | | Practica 3 | Tema4 |
| #9 | 2 | 0 | 2 | 0 | 0 | | Practica 3 | Tema4 |
| #10 | 2 | 0 | 2 | 0 | 0 | | Practica 3 | Tema4 |
| #11 | 2 | 0 | 2 | 0 | 0 | | Test semanal | Tema5 |
| #12 | 2 | 0 | 2 | 0 | 0 | | Practica 4 | Tema5 |
| #13 | 2 | 0 | 2 | 0 | 0 | | Practica 4 | Tema5 |
| #14 | 2 | 0 | 2 | 0 | 0 | | Practica 4 | Tema6 |
| #15 | 2 | 0 | 2 | 0 | 0 | | Test semanal | Tema6 |
| | 30 | 0 | 30 | 0 | 0 | | | |