Guía Docente

DATOS DE IDENTIFICACIÓN

Titulación:	Grado en Ingeniería Matemática
Rama de Conocimiento:	Ingeniería y Arquitectura
Facultad/Escuela:	Escuela Politécnica Superior
Asignatura:	Minería de Datos
Tipo:	Obligatoria
Créditos ECTS:	6
Curso:	3
Código:	4968
Periodo docente:	Sexto semestre
Materia:	Ciencia de Datos
Módulo:	Matemáticas Avanzadas y Computación
Tipo de enseñanza:	Presencial
Idioma:	Castellano
Total de horas de dedicación del alumno:	150

Equipo Docente	Correo Electrónico
María Fernanda Acosta García	mf.acosta@ufv.es

DESCRIPCIÓN DE LA ASIGNATURA

La asignatura de Minería de Datos o Data Mining proporciona al alumno las bases teóricas y prácticas para afrontar con éxito el análisis de información diversa y la extracción de información fiable a través de la exploración del contenido. Sus bases teóricas provienen de la inteligencia artificial y el aprendizaje estadístico. Además de introducir a los alumnos a la calidad del dato, se estudian las técnicas de aprendizaje estadístico y minería de datos más habituales.

OBJETIVO

Conocer las técnicas y principios metodológicos para la extracción de información relevante a partir de distintos tipos de datos con el fin de diseñar soluciones que apoyen el proceso de toma de decisiones y permitan la realización de tareas descriptivas y predictivas.

Los fines específicos de la asignatura son:

Saber explorar un conjunto de datos para extraer la información más relevante mediante técnicas estadísticas y de visualización de la información.

Saber usar técnicas predictivas de clasificación y de regresión aplicadas a un conjunto de datos.

Saber usar técnicas descriptivas de análisis de asociación y clustering aplicadas a un conjunto de datos.

Aplicar una metodología orientada al Data Mining, como es CRISP-DM, a un problema de extracción del conocimiento.

Conocer, saber usar y saber cuándo aplicar diferentes técnicas de aprendizaje estadístico y Data Mining.

CONOCIMIENTOS PREVIOS

Se recomienda haber cursado la asignatura de "Estadística" de segundo curso.

Es deseable que el alumno tenga cierta competencia lectora en inglés, ya que uno de libros recomendados de la bibliografía básica (Tan & Steinbach) está escrito en dicho idioma. Se utilizarán las transparencias en inglés del autor de dicho libro. No se requerirán competencias de expresión oral o escrita en inglés.

CONTENIDOS

INTRODUCCIÓN A LA MINERÍA DE DATOS O DATA MINING:

Tema 1. Introducción a la asignatura.
Definición de Minería de Datos. Motivación y orígenes. Tareas del Data Mining. Proceso y fases de extracción del conocimiento. Metodología CRISP-DM.

Tema 2. Preparación de datos.
Tipos de datos. Calidad de los datos. Técnicas de recopilación, almacén, limpieza y transformación de datos. Reducción de la dimensión. Técnicas de exploración y selección de datos. Estadística Descriptiva: frecuencias, percentiles, medidas de centralidad, dispersión y asimetría. Principales distribuciones de probabilidad. Selección y muestreo de datos.

Tema 3. Visualización de datos.
Histograma, diagrama de tallo y hojas, diagrama de barras, diagrama de caja, diagrama de dispersión, pie chart,

diagrama de burbujas. Otras formas de visulalización: mapa coroplético (diagrama geográfico), diagrama de contornos, coordenadas paralelas, diagrama radial, diagrama treemap, diagrama de Sankey, diagrama de velocímetro (gauge), etc.

Tema 4. Análisis de asociación.
Definición del problema. Algoritmo Apriori. Extracción de reglas. Evaluación de los patrones de asociación. Manejo de atributos categóricos y continuos. Patrones secuenciales.

Tema 5. Clasificación.
Definición del problema. Clasificación basada en reglas. Árboles de decisión y algoritmos. Clasificación basada en k-Nearest-Neighbors. Probabilidad condicional, Teorema de Bayes y clasificadores. Suppor Vector Machines. Comparación y evaluación de clasificadores.

Tema 6. Predicción.
Definición del problema. Regresión lineal: coeficientes, bondad del ajuste, análisis de residuos, contraste de hipótesis. Regresión lineal multivariante. Regresión logística.

Tema 7. Clustering.
Definición del problema. Medidas de distancia. Algoritmo k-Means. Evaluación de clusters.

Tema 8. Herramientas y casos de uso.
Data Mining con Weka. Social Network Analysis con Pajek. Casos de uso.

ACTIVIDADES FORMATIVAS

La enseñanza-aprendizaje del módulo Tecnologías Específicas atiende a una combinación armónica entre el trabajo presencial y autónomo del alumno.

La asignatura de Aprendizaje Estadístico y Data Mining es eminentemente práctica. La toma de contacto con cada tema se realizará a través de metodologías expositivas por parte del profesor, que esencialmente están dirigidas a presentar brevemente los fundamentos teóricos de la asignatura.

Asimismo se plantearán prácticas de aplicación de los contenidos teóricos implementadas en Python y mediante el uso de diversas herramientas (Pajek y Weka). Dichas prácticas se comenzarán a realizar en clases de laboratorio de informática con el profesor delante, pero deberán completarse de manera autónoma por parte de los alumnos fuera del aula. Para la resolución de dudas y dificultades compartidas por diversos alumnos, el profesor podrá organizar tutorías individuales o grupales.

Para favorecer la adquisición de habilidades de comunicación oral y escrita, así como del vocabulario específico de la asignatura, los alumnos se organizarán para realizar un trabajo teórico/práctico de la asignatura, del que entregarán una memoria y que tendrán que defender.

Finalmente, con el fin de facilitar al alumno el acceso a los materiales y la planificación de su trabajo, así como la comunicación con el profesor y el resto de alumnos, se empleará el Aula Virtual, que es una plataforma de aprendizaje online que ofrece diferentes recursos electrónicos para complementar, de forma muy positiva, el aprendizaje del alumno. Allí se colgarán los materiales de la asignatura, se plantearán las tareas que deben entregar los alumnos, se añadirán enlaces de interés sobre la asignatura y se habilitará un foro de comunicación entre el profesor y los alumnos.

Las actividades formativas, así como la distribución de los tiempos de trabajo, pueden verse modificadas y adaptadas en función de los distintos escenarios establecidos siguiendo las indicaciones de las autoridades sanitarias.

DISTRIBUCIÓN DE LOS TIEMPOS DE TRABAJO

ACTIVIDAD PRESENCIAL	TRABAJO AUTÓNOMO/ACTIVIDAD NO PRESENCIAL
60 horas	90 horas
Lección expositiva 18h Clase práctica 11h Laboratorio 20h Tutorías 5h Evaluación 6h	Estudio teórico y práctico 40h Trabajos individuales o en grupo 42h Trabajo virtual en red 8h

COMPETENCIAS

Competencias básicas

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

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

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

Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado

Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía

Competencias generales

Capacidad de resolver problemas con iniciativa, toma de decisiones, creatividad, razonamiento crítico y rigor de pensamiento, y de comunicar y transmitir conocimientos, habilidades y destrezas en el campo de la Ingeniería Matemática.

Capacidad para aplicar técnicas, modelos y herramientas matemáticas y computacionales, así como las metodologías de gestión y planificación, a la resolución de proyectos en entornos reales, en diferentes ámbitos de aplicación.

Competencias específicas

Capacidad para utilizar las técnicas matemáticas y algorítmicas necesarias para el tratamiento de datos masivos con el fin de generar conocimiento a partir de la información que ayude en la toma de decisiones.

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.

RESULTADOS DE APRENDIZAJE

Conocerá los fundamentos teóricos y metodológicos de las técnicas de aprendizaje y saber aplicarlos a la resolución de problemas.

Conocerá los principales métodos y algoritmos de la minería de datos.

Conocerá los conceptos subyacentes a los algoritmos probabilistas analizando en qué situaciones de la vida real es preciso acudir a este tipo de procedimientos.

SISTEMA DE EVALUACIÓN DEL APRENDIZAJE

ELEMENTOS DE EVALUACION:

[1] Examen parcial teórico-práctico escrito:
- Se realizarán a mitad de cuatrimestre y al final del curso.
- Evaluarán la adecuación de los conocimientos adquiridos por el alumno respecto de los objetivos de aprendizaje.
- Cada examen se puntuará de 0 a 10, siendo necesario obtener una nota superior o igual a 5,0 para superarlo.
- Se evaluarán el planteamiento completo, la resolución correcta, la interpretación adecuada, la justificación del resultado, la calidad de la presentación (vocabulario específico, ortografía, gramática) de cada pregunta.

[2] Prácticas no guiadas en parejas (excepcionalmente individuales):
- Están repartidas a lo largo del cuatrimestre.
- Evaluarán•la aplicación de los distintos algoritmos estudiados en clase a diferentes conjuntos de datos.
- Cada práctica puntuará de 0 a 10, siendo necesario obtener una nota superior o igual a 5,0 en cada práctica.
- Se evaluarán de la memoria explicativa el planteamiento completo, la resolución correcta, la interpretación adecuada, la justificación del resultado, las fuentes de información consultadas, la calidad de la presentación (vocabulario específico, ortografía, gramática). Se valorará la calidad del código y su comentario. El profesor podrá pedir que la práctica sea presentada por la pareja o individualmente, en cuyo caso indicará la ponderación de la presentación y de la memoria previamente y valorará el grado de conocimiento de cada alumno por la soltura en la exposición del contenido, por la adecuación de las respuestas a las preguntas, por la calidad del soporte empleado (transparencias, gráficos, etc.)

[3] Proyecto por parejas de un tema relacionado con la asignatura:
- Se entregará una memoria escrita con antelación a la defensa ante la clase de ambos miembros, y la defensa se realizará antes de la finalización de las clases.
- El proyecto puntuará de 0 a 10, siendo la memoria el 50 % y la defensa el 50 % que se aplicará de forma individual a cada miembro según realice cada uno la presentación. Para que el proyecto contabilice en laasignatura es necesario tener una nota igual o superior a 5,0 en la memoria y 5,0 en la defensa.
- Se evaluarán la dificultad del trabajo seleccionado, la claridad del planteamiento, el desarrollo, la corrección demlos resultados, la capacidad crítica de las conclusiones, las fuentes consultadas y la calidad formal (expresiones oral y escrita, material empleado para la presentación, gráficos, etc.). En la defensa, además, se valorarán
individualmente las respuestas a las preguntas que haga el profesor y los compañeros).

[4] Participación activa en la asignatura:
- Participación y calidad de las intervenciones
- Presentación temprana de los trabajos
- Ampliación de los contenidos de clase
- La puntuación será entre 0 y 10.
- Este elemento no es recuperable.
Para puntuar en el apartado de participación en clase, es necesario asistir al menos a un 80% de las clases.

Evaluación en caso de no poder ser presencial:
- El examen teórico-práctico se realizará a través del Campus Virtual. Se dividirá en dos partes: la primera serán preguntas relacionadas con la teoría que se responderán razonando (en total 5 preguntas) y la segunda parte serán dos ejercicios (uno fácil y otro mas difícil) que se resolverán en papel subiendo una foto a una entrega. Por último habrá que defender los ejercicios mediante una grabación en video donde se explicará brevemente el proceso.
- La defensa de los trabajos no guiados se realizará a través del Campus Virtual mediante un test que se resolverá individualmente.
- La defensa del proyecto se realizará online.

NOTA FINAL DE LA ASIGNATURA:
El baremo de cada elemento de evaluación en la calificación final de la asignatura es el siguiente:
- Exámenes parciales: 30%
- Prácticas no guiadas: 40%
- Proyecto final: 20%
- Participación activa: 10%
Condición necesaria para aprobar:
[1] >= 5,0
[2] >= 5,0
[3] >= 5,0
Cálculo:
0,3 * [1] + 0,4 * [2] + 0,2 * [3] + 0,1 * [4]
La nota numérica de los exámenes, prácticas, trabajos y ejercicios se redondeará a una cifra decimal

RECUPERACIÓN EN CONVOCATORIA ORDINARIA:
Los alumnos que no hayan alcanzado la nota mínima en alguno de los apartados anteriores, podrán optar a su recuperación al final del cuatrimestre:
- Las notas de las partes aprobadas a lo largo del curso se guardan.
- [1] Realizarán un examen global que incluirá los dos exámenes parciales el mismo día del segundo examen parcial. Será necesario obtener una calificación superior o igual a 5,0. El primer parcial no superará la calificación tope de 6,0.
- [2] Entrega de todas la prácticas (suspensas y aprobadas) antes del examen. Es necesario que cada una de las prácticas sea superior o igual a 4,0 y que la media de todas sea superior o igual a 5. La media de las practicas no

superará la calificación tope de 6,0

RECUPERACIÓN EN CONVOCATORIA EXTRAORDINARIA:
Los alumnos que no hayan alcanzado la nota mínima en alguno de los apartados anteriores tras la convocatoriamordinaria y su recuperación o deseen mejorar nota, podrán optar a su realización al final del segundo cuatrimestre:
- Las notas de las partes aprobadas a lo largo del curso se guardan.
- [1] Realizarán un examen global que incluirá los dos exámenes parciales el día señalado. Será necesario obtener una calificación superior o igual a 5,0.
- [2] Entrega de todas la prácticas (suspensas y aprobadas) antes del día señalado para el examen. Es necesario que cada una de las prácticas sea superior o igual a 4,0 y que la media de todas sea superior o igual a 5.
En caso de suspender definitivamente la asignatura, no se conservan las notas para el año siguiente.

CONSUMO DE CONVOCATORIAS:
A efecto de cómputo de convocatorias en una asignatura, solamente se contabilizarán como consumidas aquellas en las que el alumno se haya presentado a todas las pruebas de evaluación, o a una parte de las mismas, siempre que su peso en la nota final supere el 50%, aunque no se presente al examen final. Se entenderá que un alumno se ha presentado a una prueba aunque la abandone una vez comenzada la misma. La condición de No Presentado en la convocatoria extraordinaria estará ligada a la no asistencia o entrega de ninguna prueba, práctica o trabajo que esté pendiente

DISPENSA ACADÉMICA:
- Aquellos alumnos que estén exentos de la obligación de asistir a clase, bien por segunda matrícula en la asignatura o sucesivas, bien por contar con autorización expresa de la Dirección del Grado, serán evaluados por el mismo tipo de pruebas.
- Estos alumnos tienen la obligación de realizar los exámenes, las prácticas y el trabajo en los mismos plazos que el resto de sus compañeros.
- Además realizarán la defensa de forma presencial el día estipulado
- La Participación [4] se medirá con su iniciativa, participación en el foro y tener tutorías.

NORMATIVA ANTIPLAGIOS Y COPIA:
Se considerará "plagio" cualquier tipo de copia de ejercicios en un examen, de memorias de prácticas, de código fuente de prácticas, de memorias de trabajos (incluida la presentación oral), de ejercicios para casa, etc., ya sea de manera total o parcial, con el engaño de hacer creer al profesor que son propios del alumno. Cualquier tipo de fraude o plagio por parte del alumno en una elemento evaluable será sancionado e implicará un 0 en la calificación de esa parte de la asignatura, anulando la convocatoria en curso. La situación, además, será comunicada a la Dirección, que, a su vez comunicará a Secretaría General, siguiendo el protocolo establecido en la universidad.

Los exámenes se realizarán de manera presencial.

En caso de que las autoridades sanitarias obliguen a volver a un escenario donde la docencia haya que impartirla exclusivamente en remoto, se respetarán los porcentajes de evaluación anteriormente detallados.

BIBLIOGRAFÍA Y OTROS RECURSOS

Básica

* P.-N. Tan y M. Steinbach, Introduction to Data Mining, 1.ª edición. Pearson Education, 2013. ISBN: 978-1292026152. Las transparencias oficiales del libro pueden encontrarse en: http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Witten, I., Frank, E., Hall, M. Data mining: practical machine learning tools and techniques. 3 ed. 2011. ISBN 978-0-12-374856-0

* J. Hernández Orallo, M.ª J. Ramírez Quintana y C. Ferri Ramírez, Introducción a la minería de datos, 1.ª edición. Pearson Educación, 2007. ISBN: 978-84-205-4091-7.

* Material docente del profesor disponible en el Aula Virtual.

Complementaria

* A. de Vries y J. Meys, R for dummies, 2.ª edición. Hungry Minds, 2015. ISBN: 978-1119055808.

* Z. Chen, Data Mining and Uncertaining Reasoning. Wiley, 2001. ISBN: 978-0-471-38878-4.

* C. Westphal y T. Blaxton, Data mining solutions: Methods and tools for solving real-world problems. Wiley, 1998. ISBN: 978-0471253846.

* W. N. Venables, D. M. Smith and the R Core Team, An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics, Version 3.3.0, 3 de mayo de 2016. Disponible en la web: http://cran.r-project.org/doc/manuals/R-intro.pdf

* M. J. A. Berry y G. S. Linoff, Mastering Data Mining: The Art and Science of Customer Relationship Management, 1.ª edición. Wiley, 1999. ISBN: 978-0471331230.

* X.-S. Yang, Nature-inspired Metaheuristic Algorithms, 2.ª edición. Luniver Press, 2010. ISBN: 978-1905986286.

* Y. Zhao, R Reference Card for Data Mining. 2013. http://www.rdatamining.com/docs/r-reference-card-for-data-mining

* D. Hiebeler, MATLAB / R Reference. 2010. Disponible en la web: http://cran.r-project.org/doc/contrib/Hiebeler-matlabR.pdf

* R. R. Bouckaert, E. Frank, M. Hall, et al., Weka Manual for Version 3-9-0. University of Waikato, 2016. http://www.cs.waikato.ac.nz/ml/weka/documentation.html

* RStudio (entorno de desarrollo para R): http://www.rstudio.com/products/RStudio/

* R Project for Statistical Computing: http://www.r-project.org/

* Weka, University of Waikato (New Zealand), http://www.cs.waikato.ac.nz/ml/weka/