ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

GUIA DOCENTE

CURSO 2020/2021

Máster en Ingeniería Informática (Plan 2018)

DATOS DE LA ASIGNATURA			
Nombre:			
Técnicas Escalables de Análisis de Datos			
Denominación en inglés:			
Scalable Data Analysis Techniques			
Código:		**Carácter:**	
1180418		Optativo	
Horas:			
	Totales	**Presenciales**	**No presenciales**
Trabajo estimado:	75	30	45
Créditos:			
	Grupos reducidos		
Grupos grandes	**Aula estándar**	**Laboratorio**	**Prácticas de campo**	**Aula de informática**
1.5	0	0	0	1.5
Departamentos:		**Áreas de Conocimiento:**	
Tecnologías de la Información		Ciencias de la Computación e Inteligencia Artificial	
Curso:		**Cuatrimestre:**	
1º - Primero		Segundo cuatrimestre	

DATOS DE LOS PROFESORES			
Nombre:	**E-Mail:**	**Teléfono:**	**Despacho:**
Peregrín Rubio, Antonio	peregrin@uhu.es	959217653	ETSI 156
*Rodriguez Roman, Miguel Angel	miguel.rodriguez@dti.uhu.es	959217372	134 / Escuela Técnica Superior de Ingeniería / El Carmen

*Profesor coordinador de la asignatura

Consultar los horarios de la asignatura

DATOS ESPECÍFICOS DE LA ASIGNATURA

1. Descripción de contenidos

1.1. Breve descripción (en castellano):

Consisten en un estudio de los modelos más relevantes de *Knowledge Discovery in DataBases* disponibles, especialmente en el ámbito de la Computación Inteligente, a disposición para su uso tanto en bibliotecas de código reutilizable como en publicaciones recientes que los describan en detalle, para los entornos de Big Data, con objeto de que el alumno amplíe sus conocimientos y capacidad para resolver problemas específicos sobre tratamiento de datos en situaciones determinadas del mundo real, basado en el estudio de ejemplos de modelos ya creados y las posibilidades de adaptación y/o ampliación.
Sus contenidos incluyen, técnicas escalables para:

- Preprocesamiento: transformaciones, filtrado, reducción y aumento de datos, representación y visualización, etc.
- Análisis descriptivo y predictivo: clasificación, regresión, asociación, etc.
- Tratamiento en tiempo real y *data streams*.
- Aplicaciones específicas en seguridad de transacciones bancarias, sanidad, etc.

1.2. Breve descripción (en inglés):

It deals with the study of the most relevant models of Knowledge Discovery in DataBases available, especially in the field of Computational Intelligence, to be used both in libraries of reusable code and also in recent publications that describe them in detail, for Big Data environments, in order to provide the student the chance to expand his knowledge and ability to solve specific problems on data processing in certain situations of the real world, based on the study of some examples of models already created and the options to be adapted and expanded.
Its contents include, scalable techniques for:

- Preprocessing: transformations, filtering, data reduction/increase, representation and visualization, etc.
- Desc.riptive and predictive analysis: classification, regression, association, etc.
- Real-time processing and data streams.
- Specific applications in security of banking transactions, health, etc

2. Situación de la asignatura

2.1. Contexto dentro de la titulación:

Segundo cuatrimestre, segunda mitad del mismo. Materia recomendada para un perfil de formación completo para un Analista de Datos.

2.2. Recomendaciones:

Haber cursado la asignatura del primer cuatrimestre Big Data, y también recomendable haber cursado Infraestructuras para Big Data en la primera parte del segundo cuatrimestre.

3. Objetivos (Expresados como resultados del aprendizaje):

El alumno debe aprender a emplear técnicas avanzadas para el tratamiento de datos en entornos de Big Data, así como a adaptarlas para resolver nuevos problemas en entornos aplicados. Específicamente, los objetivos que se alcanzan en esta asignatura son: conocer y comprender modelos para procesamiento previo de datos (reducción, limpieza, discretización, etc), búsqueda de patrones, relaciones entre datos, extracción de conocimiento, etc., métodos en general orientados a la extracción de conocimiento, es decir, a obtener valor y proporcionar comprensión de los mismos para empresas y organizaciones.

4. Competencias a adquirir por los estudiantes

4.1. Competencias específicas:

4.2. Competencias básicas, generales o transversales:

- **CB7:** 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
- **CB9:** Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades
- **CB10:** Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo
- **CG1:** Capacidad para proyectar, calcular y diseñar productos, procesos e instalaciones en todos los ámbitos de la ingeniería informática
- **CG8:** Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y mulitidisciplinares, siendo capaces de integrar estos conocimientos
- **CT1:** Gestionar adecuadamente la información adquirida expresando conocimientos avanzados y demostrando, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en el campo de estudio.
- **CT2:** Dominar el proyecto académico y profesional, habiendo desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro su ámbito temático, en contextos interdisciplinares y, en su caso, con un alto componente de transferencia del conocimiento.
- **CT4:** Comprometerse con la ética y la responsabilidad social como ciudadano y como profesional, con objeto de saber actuar conforme a los principios de respeto a los derechos fundamentales y de igualdad entre hombres y mujeres y respeto y promoción de los Derechos Humanos, así como los de accesibilidad universal de las personas discapacitadas, de acuerdo con los principios de una cultura de paz, valores democráticos y sensibilización medioambiental.
- **CT5:** Utilizar de manera avanzada las tecnologías de la información y la comunicación, desarrollando, al nivel requerido, las Competencias Informáticas e Informacionales ('CI2).

5. Actividades Formativas y Metodologías Docentes

5.1. Actividades formativas:

- Sesiones de Teoría sobre los contenidos del Programa.
- Sesiones de Resolución de Problemas.
- 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. Metologías docentes:

- Clase Magistral Participativa.
- Desarrollo de Prácticas en Laboratorios Especializados o Aulas de Informática en grupos reducidos.
- Resolución de Problemas y Ejercicios Prácticos.
- Tutorías Individuales o Colectivas. Interacción directa profesorado-estudiantes.
- Planteamiento, Realización, Tutorización y Presentación de Trabajos.
- Conferencias y Seminarios.
- Evaluaciones y Exámenes.

5.3. Desarrollo y justificación:

- En las Sesiones presenciales de Teoría se empleará en algunos casos la metodología "Clase magistral", cuando se trate de orientar y situar la asignatura en su contexto, y encuadrar las distintas líneas de la misma.
- En las Sesiones presenciales de Resolución de Problemas se propondrán, bien ejercicios, o bien completar casos de uso con ampliaciones propuestas por el profesor.
- Sesiones presenciales de Practicas en Laboratorio de Informatica, en las que el alumno adquirirá experiencia en el manejo de herramientas y en la programación de algoritmos que debe presentar y defender para su evaluación por el profesor.
- Las Actividades Académicamente Dirigidas consistirán en trabajos propuestos por el profesor para ser desarrollados por los alumnos de forma autónoma, pero con un control periódico del profesor, y la presentación final de una memoria y/o una exposición en clase por parte del alumno.
- Se programarán seminarios, cuando sea posible, y se promocionará la exposición de trabajos por parte de los alumnos para que adquieran destrezas en la presentacion de los materiales que elabore.
- Las actividades no presenciales de Lectura de Contenidos consisten en la lectura propiamente de material facilitado por el profesor para este fin (a través de la plataforma Moodle) de recursos que que permiten al alumno profundizar y extender su conocimiento en diferentes áreas de la materia.
- Las Tutorías Colectivas serán tutorías online a través de la plataforma Moodle en las que intervendrán tanto los alumnos, colaborativamente, como el profesor para aclarar y conducir el debate cuando sea necesario. La metodología no presencial "Tutoría en Línea" complementará, mediante sesiones de chat interactivo, las posibles necesidades de tutoría de los alumnos sin necesidad de desplazamiento al centro.
- A través de la plataforma Moodle, se pondrá a disposiclón de los alumnos también, cuando sea posible, de entrevistas a expertos y vídeos de sesiones magistrales de especialistas en la materia. Estas actividades no sólo son no presenciales, sino que puden también ser comentadas en las Sesiones Teóricas presenciales en algunos casos, cuando el profesor estime que el diálogo y debate sobre las mismas pueda ser relevante.
- El Trabajo Individual Autónomo del Estudiante incluye aquellas actividades no recogidas específicamente en otras actividades, y que forman parte de las actividades que lleva a cabo no presencialmente para completar su formación, a instancias de las líneas marcadas en la Sesiones Magistrales como en las indicadas en la plataforma Moodle. Ejemplo: búsqueda documental, elaboración de esquemas, etc.
- Se podrán proponer actividades que conlleven el Trabajo Colaborativo de los estudiantes, es decir, la organización, distribución de tareas, combinación del trabajo individual o en subgrupos, todo ello orientado a la conformación final de un trabajo que precise dicha distribución entre distintos estudiantes, formentando así el trabajo en equipo para un objetivo común. En este tipo de ejercicios, cabe también la aplicación de metodologías basadas en acción, es decir, la actualización de los objetivos por los propios alumnos con el visto bueno del profesor, en función de la evolución del ejercicio llevado a cabo por el grupo de estudiantes.
- La evaluación de la asignatura se realizará, como se cita en su apartado correspondiente de esta guia, atendiendo a distintas partes de la misma y con diferente nivel de influencia en la nota final. Esto incluye, en su correspondiente porcentaje, a las sesiones presenciales de Actividades de Evaluación (Ej: exámenes) y las Actividades de Autoevaluación no presenciales (Ej: ejercicios recogidos por la plataforma Moodle puntuables).

6. Temario desarrollado:

Tema 1: Preparación de los Datos en Big Data.
- Modelos para la selección de instancias y características. *Rankings* y pesos de características. Sobremuestreo y submuestreo. Discretización. Generación de prototipos.

Tema 2: Impacto de la División de Datos y Modelos de Fusión de Información.
- Datos no balanceados y *small disjunts*. Modelos para particionar datos.

Tema 3: Modelos Escalables para Reglas de Clasificación.
- Métodos para aprender modelos de clasificación. Aprendizaje con datos no balaceados. EDGAR-MR.

Tema 4: Modelos Escalables para Sistemas basados en Reglas Difusas.
- Modelos para clasificación, regresión y agrupamiento.

Tema 5: Modelos Escalables para Reglas de Asociación.
- Tipos de modelos para aprender reglas de asociación. Áboles de patrones frecuentes, PARMA, BigFIM, Dist-Eclat, FiDoop, etc.

Tema 6: Modelos Escalables para Descubrimiento de Subgrupos.
- Estrategias, modelos implementados, Apriori-K, FP-Tree, etc.

7. Bibliografía

7.1. Bibliografía básica:

Libros:
- Data Preprocessing in Data Mining,Vol. 72 of Intelligent Systems Reference Library. Springer International Publishing AG, 2015,ISBN 978-3-319-10246-7
- Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. *Integr. Ser. Inf. Syst*, *36*, 1-12.
- Berson, A., & Smith, S. J. (1997). Data warehousing, data mining, and OLAP. McGraw-Hill, Inc..

Artículos:
- S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-based Systems 98 (2016) 1-29, doi:10.1016/j.knosys.2015.12.006.
- Fernández, A., del Río, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced Big Data classification: outcomes and challenges. Complex & Intelligent Systems, 3(2), 105-120.
- García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 9.
- García-Pedrajas, N., & de Haro-García, A. (2012). Scaling up data mining algorithms: review and taxonomy. Progress in Artificial Intelligence, 1(1), 71-87.
- Cieslak, D. A., & Chawla, N. V. (2007, October). Detecting fractures in classifier performance. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on (pp. 123-132). IEEE.
- Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
- Río S., López V., Benítez J.M., Herrera F. (2015). A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules. International Journal of Computational Intelligence Systems, 8 (3), 422-437.
- Peralta D., Río S., Ramírez-Gallego S., Triguero I., Benítez J.M., Herrera F. (2015). Evolutionary Feature Selection for Big Data Classification: A MapReduce Aproach. Mathematical Problems in Engineering, doi: 10.1155/2015/246139.
- Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

7.2. Bibliografía complementaria:

- Big data : a revolution that will transform how we live, work, and think, Viktor Mayer-Scho?nberger, Kenneth Cukier, HoughtonMifflin Harcourt, 2013 (Versión Castellano: Big Data, La revolución de los Datos Masivos)
- Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. IBM. Paul Zikopoulos, Chris Eaton. McGraw Hill Professional, 29/05/2015. http://www-01.ibm.com/software/data/bigdata/
- Analítica predictiva: Predecir el futuro utilizando Big Data / Eric Siegel. Madrid : Anaya Multimedia, 2013
- Del Cloud Computing al Big Data: Visión introductoria para jovenes emprendedores. Jordi Torres i Vinnals. Editorial UOC - PID_00194204. Primera edición: septiembre 2012. Todos los derechos reservados de esta edición, FUOC, 2012. v.3.0 Espanña de Creative Commons.
- Hadoop, Soluciones Big Data,, Boris Lublinsky, Kevin Smith, Alex Yakubovich, Anaya Multimedia 2014
- Big Data, Técnicas, Herramientas y Aplicaciones, María Pérez Márquez, 2015

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:

Por defecto, todos los alumnos (salvo que soliciten la Evaluación Única) serán evaluados con el sistema de Evaluación Continua en las convocatorias ordinarias (I a la III, es decir, Junio, Septiembre y Diciembre respectivamente), el cual, es el recomendado especialmente para esta materia. No obstante, aquellos alumnos que no puedan (o no deseen) acogerse al sistema recomendado de Evaluación Continua, pueden solicitar acogerse al sistema de Evaluación Única mediante escrito firmado y entregado a través del Registro General de la Universidad (presencial o telemático) dirigido al Departamento de Tecnologías de la Información y a la atención del profesor coordinador de la materia, Miguel Ángel Rodríguez (recomendándose, enviar también copia de dicho escrito al profesor por correo electrónico miguel.rodriguez@dti.uhu.es) sólo como medida complementaria, y por agilidad administrativa), donde claramente consten los datos del alumno y de la asignatura para la que solicita la Evaluación Única. La convocatoria extraordinaria para la finalización del título (Noviembre) sólo se regirá mediante el sistema de Evaluación Única, y por tanto, no se necesita realizar solicitud previa para ello.
--
Evaluación Continua (recomendada para las Convocatorias I a la III):
- Los conocimientos teóricos (teoría y problemas) de la materia se evaluarán mediante dos tipos de pruebas:
- 1) **un examen escrito**presencial e individual, de preguntas largas, cortas y problemas combinados, según la convocatoria del Centro para esta materia. En dicho examen no se permitirá el uso de ningún dispositivo electrónico, se le proporcionará papel y el estudiante empleará bolígrafo azul o negro propio, y no se empleará ningún elemento documental externo. El peso de este examen será de un 25%, y el alumno debe obtener al menos un 4 sobre 10 puntos para el cómputo de la nota final considerando esta parte superada.
- 2) **pruebas de evaluación mediante la plataforma de enseñanza virtual**Moodle, que consistirán en ejercicios planteados por el profesor durante el curso, generalmente de resolución breve o media, para incentivar la autonomía del estudiante en la resolución de cuestiones y problemas de la materia, para resolver de forma individual y no presencial, con plazos de entrega relativamente breves dictados en el momento de cada propuesta particular), con un peso en la evaluación final del estudiante del 20% en la nota final, requiriendo que el alumno alcance al menos una puntuación de 4 sobre 10 para que compute en la nota final considerando esta parte superada.
- Los conocimientos prácticos se evaluarán mediante **prácticas de laboratorio**, las cuales consisten en una serie de enunciados en los que se indican los objetivos a alcanzar, los medios y describen con detalle los entregables que el alumno debe enviar a la plataforma Moodle en el formato y plazo estipulados en el propio enunciado, y defendidas presencialmente en clase de prácticas de laboratorio (el alumno debe mostrar su trabajo, y responder a las cuestiones sobre el mismo que el profesor le plantee). Todas las prácticas obligatorias propuestas deben ser entregadas y defendidas presencialmente. En cada práctica, el profesor indicará si su realización es individual o por grupos (en tal caso, el tamaño y componentes del grupo debe ser supervisado y aceptado por el profesor al inicio de la práctica). No es necesario asistir presencialmente a las sesiones de prácticas que no sean las de entrega y defensa de prácticas, pero es muy recomendable hacerlo regularmente, pues se trata de una actividad diseñada para ser presencial. El peso en la evaluación de esta parte será de un 40%, y debe obtener una calificación mínima de 4 puntos sobre 10, para computar en la nota final considerando esta parte superada.
- La participación activa del estudiante en las actividades propuestas (seguimiento del estudiante), es decir, su contribución a los debates en la plataforma, foros de ideas, respuesta a ejercicios cortos, actividades académicas, etc., será también un elemento que se valorará con un peso de un 15%, a criterio de los profesores de teoría y prácticas de laboratorio. Esta parte no tiene un mínimo de puntuación establecido para considerarse superada y computar en la media ponderada de la nota final.

--
Los alumnos que no aprueben en la convocatoria ordinaria I (Junio), se pueden presentar a la convocatoria ordinaria II (Septiembre) o a la convocatoria ordinaria III (Diciembre) de un mismo curso académico, y realizar la parte correspondiente a la no superada (examen escrito, pruebas de evaluación mediante la plataforma moodle, y las prácticas). Este criterio no se mantiene entre distintos cursos académicos, es decir, las partes aprobadas, no se mantienen de un curso a otro. Asimismo, la participación activa del estudiante en las actividades propuestas se mantendrá durante las tres convocatorias ordinarias del curso académico, pero no entre distintos cursos académicos.
--
Evaluación Única (para la Convocatoria extradordinaria para la finalización del título, y disponible para las Convocatorias I a la III):
La evaluación única consistirá en un examen presencial individual, con acreditación previa del alumno mediante DNI que deberá mostrar tantas veces se le solicite, que se celebrará en la fecha y lugar que la convocatoria del Centro fije para la asignatura. Este examen estará compuesto por una serie de partes diferentes, que se describen a continuación:
- Para valorar los conocimientos teóricos, se plantearán una serie de cuestiones (largas y cortas combinadas) escritas sobre la materia, a las que el alumno debe responder por escrito, en un tiempo máximo de 1:30 minutos. El peso de esta parte en la nota final será de un 45% (equivalente a la suma del 25% de examen escrito más un 20% de las pruebas de evaluación mediante Moodle de la modalidad de Evaluación Continua).
- Para valorar los conocimientos prácticos, se planteará la implementación de algunos algoritmos expuestos en la materia (el alumno debe hacer estas implementaciones por escrito. La duración para esta parte del examen será de 2h como máximo. El peso de esta parte en la nota final será de un 40%.
- La capacidad de tener criterio y aportar del alumno en la materia, será evaluada por el profesor oralmente, es decir, el profesor propondrá a cada alumno de forma individual (haciendo pasar a cada alumno individualmente a la sala de examen) un tema de discusión, que el alumno mantendrá con el profesor durante un tiempo máximo de 10 minutos, en los que el profesor evaluará la destreza del alumno argumentando sobre la materia. El peso en la nota final que tendrá esta parte es de un 15%.

Las partes escritas anteriormente aludidas, es decir, las cuestiones para valorar los conocimientos teóricos y prácticos, se llevarán a cabo facilitando al alumno el papel que debe rellenar empleando para ello un bolígrafo azul o negro (de su propiedad), en las condiciones ambientales más favorables que permita el aula en la que se fije la convocatoria, no permitiéndose el uso de dispositivos electrónicos (teléfonos móviles, tabletas, auriculares, ordenadores, relojes inteligentes,

wearables, etc.). Como excepción, estarán permitidos los dispositivos electrónicos fijos tales como implantes auditivos internos, marcapasos, etc., y ante la duda, siempre se recomienda preguntar antes de llevarlos al examen al profesor. Los apuntes, libros, hojas de distinto tamaño y textos camuflados en general, no estarán permitidos puesto que no se consideran material didáctico ni documentación a utilizar admisible. La duración total del examen completo, con todas sus partes, no excederá nunca las 4 horas.
--
En general, tanto para la Evaluación Continua como para la Evaluación Única, se garantiza la adquisición de las competencias de la siguiente forma: mediante la evaluación (con defensa) de las prácticas de laboratorio de la Evaluación Continua, y de las preguntas del examen sobre conocimientos prácticos en el caso de la Evaluación Única, las competencias CB9, CG8, CT1, CT2, CT5; mediante el examen escrito y las pruebas de evaluación mediante la plataforma de enseñanza virtual en la Evaluación Continua, y en las preguntas sobre conocimientos teóricos del examen en la Evaluación Única, las competencias: CG1, CB7, CB10, CT1, CT2; y mediante el seguimiento individual del estudiante en la Evaluación Continua, y en la parte oral del examen de la Evaluación Única, las competencias: CB7, CT2, CT4.
--
Nota sobre la calificación “Matrícula de Honor”: si existieran más alumnos con una calificación que les permita aspirar a la matrícula de honor (10 sobre 10 puntos en la media ponderada), es decir, en caso de equidad, la calificación de Matrícula de Honor se asignará basándose en la mayor participación en clase e implicación del alumno en la asignatura (en todo tipo de sesiones), a juicio de los profesores de la asignatura.

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	0	0	0	0	0		
#2	0	0	0	0	0		
#3	0	0	0	0	0		
#4	0	0	0	0	0		
#5	0	0	0	0	0		
#6	0	0	0	0	0		
#7	0	0	0	0	0		
#8	1	0	1	0	0		Tema 1
#9	2	0	2	0	0		Tema1
#10	2	0	2	0	0		Tema 2
#11	2	0	2	0	0		Tema 3
#12	2	0	2	0	0	Entrega Práctica 1	Tema 3
#13	2	0	2	0	0	Actividad de clase	Tema 4
#14	2	0	2	0	0	Actividad de clase	Tema 5
#15	2	0	2	0	0	Entrega Práctica 2 y Examen	Tema 6
	15	0	15	0	0		