Guía docente de la asignatura

Asignatura	TÉCNICAS ESCALABLES DE ANÁLISIS DE DATOS		
Materia	SISTEMAS INTELIGENTES Y BASADOS EN CONOCIMIENTO		
Módulo			
Titulación	MÁSTER EN INGENIERÍA INFORMÁTICA		
Plan	510	**Código**	53198
Periodo de impartición	S2	**Tipo/Carácter**	OB
Nivel/Ciclo	MÁSTER	**Curso**	1
Créditos ECTS	6		
Lengua en que se imparte	ESPAÑOL		
Profesor/es responsable/s	CARLOS J. ALONSO GONZÁLEZ, BELARMINO PULIDO JUNQUERA, PEDRO C. ÁLVAREZ ESTEBAN		
Datos de contacto (E-mail, teléfono…)	calonso@infor.uva.es 983 185602; belar@infor.uva.es 983 185606; pedroc@eio.uva.es 983 423930		
Horario de tutorías	Véase la información actualizada en la web: www.uva.es		
Departamento	INFORMATICA (ATC, CCIA, LSI), ESTADÍSTICA E INVESTIGACIÓN OPERATIVA		

1. Situación / Sentido de la Asignatura

1.1 Contextualización

La asignatura “Técnicas Escalables de Análisis de Datos” introduce los elementos necesarios para aplicar técnicas de Aprendizaje Automático a grandes volúmenes de datos como lo son los procedentes de aplicaciones web o móviles, la Internet de las Cosas y las redes de sensores, así como procedentes de servicios financieros, sanidad u otros campos científicos.

El conjunto de datos que se puede usar en estos campos es enorme y el conjunto de técnicas de aprendizaje a aplicar muy variado. Estos datos puede ser propiedad de una organización o pueden proceder de múltiples fuentes, pero en todos los casos su volumen puede ser tan grande que no se puedan procesar en un único ordenador, por lo cual será necesario recurrir posiblemente a un almacenamiento distribuido, a un procesamiento distribuido o a ambos.

Además, la gran cantidad de datos a procesar hará necesario analizar con cuidado el tipo de técnicas o algoritmos aplicables, ya que los requisitos de memoria pueden hacer inviables la utilización de técnicas o aplicaciones más convencionales.

1.2 Relación con otras materias

Existe relación con las asignaturas de la especialidad de Big Data como son la “Tecnología para el Big Data”, “Almacenamiento Escalable” y “Big Data: Inteligencia de Negocios”, donde se tratarán problemas asociados a la gestión y almacenamiento de grandes cantidades de datos en entornos distribuidos y su posterior uso en los procesos de negocio para la extracción de información con técnicas similares a la Minería de Datos o el Descubrimiento de Conocimiento en Bases de Datos.

El aprendizaje de modelos a partir de grandes cantidades de datos también está relacionado con el aprendizaje de modelos gráficos probabilísticos como son las Rede Bayesianas o las Redes de Markov que se estudian en la asignatura obligatoria del máster “Métodos avanzados de razonamiento y representación del conocimiento”.

1.3 Prerrequisitos

Se recomienda que el alumno haya cursado estudios de grado con un contenido medio de competencias en Inteligencia Artificial y en Matemática Discreta. En relación con los Grados de Informática hasta ahora vigentes en los planes de estudio de la UVa, se recomienda que el alumno haya cursado la asignatura de “Técnicas de Aprendizaje Automático”.

2. Competencias

2.1 Generales

Código	Descripción
CG1	Capacidad para proyectar, calcular y diseñar productos, procesos e instalaciones en todos los ámbitos de la ingeniería informática.
CG3	Capacidad para dirigir, planificar y supervisar equipos multidisciplinares.
CG4	Capacidad para el modelado matemático, cálculo y simulación en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la Ingeniería en Informática.
CG7	Capacidad para la puesta en marcha, dirección y gestión de procesos de fabricación de equipos informáticos, con garantía de la seguridad para las personas y bienes, la calidad final de los productos y su homologación.
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 multidisciplinares, siendo capaces de integrar estos conocimientos.
CG9	Capacidad para comprender y aplicar la responsabilidad ética, la legislación y la deontología profesional de la actividad de la profesión de Ingeniero en Informática.

2.2 Específicas

Código	COMPETENCIAS ESPECÍFICAS Descripción
CET5	Capacidad para analizar las necesidades de información que se plantean en un entorno y llevar a cabo en todas sus etapas el proceso de construcción de un sistema de información.
CET9	Capacidad para aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados en el conocimiento.

3. Objetivos

Código	Descripción
SI-BC 5	Conocer los métodos básicos de aprendizaje automático y Minería de Datos y entender los problemas de escalabilidad en entornos de grandes almacenes de datos.
SI-BC 6	Conocer cómo se pueden implementar estas técnicas en frameworks específicos, así como sus limitaciones.
SI-BC 7	Conocer y ser capaz de aplicar técnicas de análisis de datos mediante clasificación (tanto supervisada como no supervisada), agrupamiento (clustering) o asociación.
SI-BC 8	Ser capaz de aplicar técnicas avanzadas de análisis de datos recomendadores, clustering o clasificación.
SI-BC 9	Ser capaz de implementar estas técnicas en distintos ámbitos de aplicación, utilizando las tecnologías adecuadas

4. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
Clases teórico-prácticas (T/M)	**15**	Estudio y trabajo autónomo individual	30
Clases prácticas de aula (A)			
Laboratorios (L)	**30**	Estudio y trabajo autónomo individual	24
Prácticas externas, clínicas o de campo			
Seminarios (S)	**15**	Estudio y trabajo grupal dirigido	28
Tutorías grupales (TG)			
Evaluación*			8
Total presencial	**60**	**Total no presencial**	**90**

* Evaluación: Se incluyen en las actividades de Laboratorio y Seminarios.

5. Bloques temáticos

Bloque 1: Técnicas Escalables de Análisis de Datos.

Carga de trabajo en créditos ECTS: 6

a. Contextualización y justificación

Véase apartado 1.1.

b. Objetivos de aprendizaje

Véase apartado 3.

c. Contenidos

1. Introducción a Apache Spark.
    a. Spark y sus componentes.
    b. RDD y Scala.
2. Conceptos generales sobre Aprendizaje Automático y Grandes Volúmenes de Datos.
    a. Arquitectura de un sistema de aprendizaje automático.
    b. Acceso, procesamiento y filtrado de datos.
3. Métodos de aprendizaje sobre MLLib / Spark.
    a. Recomendadores con MLLib.
    b. Aprendizaje supervisado: Clasificadores y Modelos Predictivos con MLLib.
    c. Aprendizaje no supervisado: Técnicas de agrupamiento (*clustering*) con MLLib.
    d. Procesado de texto avanzado con MLLib.

d. Métodos docentes

Clase magistral participativa para discutir los contenidos básicos de la asignatura.
Laboratorios para la experimentación con las ideas básicas del bloque temático.
Realización de proyectos.

e. Plan de trabajo

Se proporcionará al comienzo de la asignatura.

f. Evaluación

Véase apartado 7.

g. Bibliografía básica

- Nick Pentreath. Machine Learning with Spark. Packt Publishing. 2015. ISBN: 9781783288519. http://www.packtpub.com/

- Petar Zečević y Marko Bonaći. Spark in Action. Manning Publications. 2016. ISBN: 9781617292606. https://www.manning.com/books/
- Ian H. Witten, Eibe Frank y Mark A. Hall. Data Mining: practical machine learning tools and techniques (third Edition). Morgan Kaufmann, 2011.

h. Bibliografía complementaria

- Apache Organization. Apache Spark. http://spark.apache.org/
- Apache Organization. Apache MLLib. http://spark.apache.org/mllib/
- Kaggle. Kaggle in class. https://inclass.kaggle.com/
- Rishi Yadav. Spark Cookbook. Packt Publishing 2015.
- Jure Leskovek, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. Second edition. Cambridge University Press, 2014.

i. Recursos necesarios

Notas de la asignatura.

Guiones de cuestiones y problemas.

Curso Moodle de soporte a la asignatura.

Software gratuito para el desarrollo de análisis de datos escalable como Cloudera y Spark.

6. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO*
Técnicas Escalables de Análisis de Datos	6	Semanas 1-15

7. Tabla resumen de los instrumentos, procedimientos y sistemas de evaluación/calificación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Proyectos	90%	Se realizarán varios mini-proyectos para evaluar cada una de las partes del Bloque 1.
Participación en clases, cuestionarios, seminarios prácticas y tutorías.	10%	

La participación en clases, seminarios, prácticas y tutorías se evalúa a partir de las entregas opcionales y la participación en clase.
Recuérdese que aunque en ningún caso la asistencia a clase es evaluable, los profesores responsables pueden excluir de alguna actividad formativa evaluable a aquellos alumnos que no participen en las actividades presenciales, que incluyen las tutorías activas, los seminarios y las prácticas de laboratorio, especialmente, aunque no limitado a, en aquellas actividades de carácter grupal.

En la convocatoria extraordinaria la prueba consistirá en la realización de un proyecto que permitiría obtener el 100% de la calificación en esta convocatoria. No obstante, aquellos estudiantes que quieran conservar las calificaciones obtenidas en las partes de alguno de los MiniProyectos podrán solicitarlo con antelación y en ese caso sólo tendrían que realizar la parte proporcional del examen teórico.

8. Consideraciones finales

9. Cronograma de actividades:

Se proporcionará un cronograma detallado de las actividades antes del inicio del curso.