Proyecto/Guía docente de la asignatura

Asignatura	TÉCNICAS ESCALABLES DE ANÁLISIS DE DATOS EN ENTORNOS BIG DATA: REGRESIÓN Y DESCUBRIMIENTO DE CONOCIMIENTO		
Materia	APRENDIZAJE AUTOMÁTICO DE ALTAS PRESTACIONES		
Módulo	MÓDULO 2: TECNOLOGÍAS INFORMÁTICAS		
Titulación	MÁSTER EN INGENIERÍA INFORMÁTICA		
Plan	639	**Código**	54926
Periodo de impartición	S2	**Tipo/Carácter**	OB
Nivel/Ciclo	MÁSTER	**Curso**	1
Créditos ECTS	3		
Lengua en que se imparte	ESPAÑOL		
Profesor/es responsable/s	**J. BELARMINO PULIDO JUNQUERA Y PEDRO C. ÁLVAREZ ESTEBAN,** ALEJANDRO RODRÍGUEZ COLLADO Y ADRIÁN ARROYO CALLE		
Departamento	INFORMATICA (ATC, CCIA, LSI), ESTADÍSTICA E INVESTIGACIÓN OPERATIVA		
Datos de contacto (E-mail, teléfono…)	b.pulido@uva.es, (983185606), pedrocesar.alvarez@uva.es (983423930), alejandro.rodriguez.collado@uva.es, adrian.arroyo@uva.es		

1. Situación / Sentido de la Asignatura

1.1 Contextualización

La asignatura “Técnicas Escalables de análisis de datos en entornos big data: regresión y descubrimiento de conocimiento” es la continuación de la asignatura “Técnicas escalables de análisis de datos en entornos Big Data: Clasificadores” que 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, el 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.

En esta asignatura se analizarán grandes cantidades de datos mediante técnicas como la regresión, el clustering y los recomendadores.

1.2 Relación con otras materias

Es una continuación de “Técnicas escalables de análisis de datos en entornos Big Data: Clasificadores” y junto con la asignatura “Deep learning y sus aplicaciones”, constituyen la materia “Aprendizaje automático de altas prestaciones”.

Existe también relación con las asignaturas de la Materia “Tecnologías de Gestión de Información” como son “Arquitecturas Big Data” y “Tecnologías distribuidas y BlockChain”, 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.

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. Se recomienda vivamente también que el alumno haya cursado la asignatura “Técnicas escalables de análisis de datos en entornos Big Data: Clasificadores”, en tanto que la introducción a Spark/Scala se trata en esa asignatura. 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.
CG2	Capacidad para la dirección de obras e instalaciones de sistemas informáticos, cumpliendo la normativa vigente y asegurándola calidad del servicio.
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.
CG10	Capacidad para aplicar los principios de la economía y de la gestión de recursos humanos y proyectos, así como la legislación, regulación y normalización de la informática.

2.2 Específicas

Código	Descripción
CET1	Capacidad para modelar, diseñar, definir la arquitectura, implantar, gestionar, operar, administrar y mantener aplicaciones, redes, sistemas, servicios y contenidos informáticos.
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.
CET7	Capacidad para comprender y poder aplicar conocimientos avanzados de computación de altas prestaciones y métodos numéricos o computacionales a problemas de ingeniería.
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
1	Conocer los métodos de inducción de regresores en el ámbito del Aprendizaje Automático y Minería de Datos y entender los problemas de escalabilidad en entornos de grandes almacenes de datos.
2	Conocer los métodos principales de descubrimiento de conocimiento en el ámbito Big Data: Clustering y Recomendadores, atendiendo también a los problemas de escalabilidad.
3	Conocer cómo se pueden implementar estas técnicas en frameworks como Hadoop, Spark, etc., así como sus limitaciones.
4	Conocer y ser capaz de aplicar técnicas no supervisadas de análisis de datos en el contexto de Big Data.
5	Ser capaz de desarrollar aplicaciones Big Data utilizando técnicas no supervisadas de análisis de datos, en distintas áreas de aplicación utilizándolas tecnologías adecuadas.

4. Contenidos y/o bloques temáticos

Bloque 1: Técnicas Escalables de análisis de datos en entornos big data: regresión y descubrimiento de conocimiento.

Carga de trabajo en créditos ECTS: 3

a. Contextualización y justificación

Véase apartado 1.1

b. Objetivos de aprendizaje

Véase apartado 3.

c. Contenidos

1. Métodos de regresión
2. Métricas y selección de modelos
3. Métodos de clustering
4. Sistemas recomendadores

d. Métodos docentes

Ver Apartado 5: Métodos docentes y principios metodológicos.

e. Plan de trabajo

Se les comunicará a los alumnos al inicio del período docente mediante la plataforma Moodle que se utilice en la asignatura.

f. Evaluación

Ver apartado 7: Sistema y características de la evaluación.

g. Material docente

- Se utilizará el Moodle de la Escuela de Ingeniería Informática: aulas.inf.uva.es, donde se colocará el material de la asignatura, tanto resúmenes de los contenidos en formato PDF (transparencias) como programas en Scala.
- Las herramientas de comunicación serán:
    - Asíncronos: foros, emails y mensajes directos dentro del Moodle.
    - Síncronos: Salas de videoconferencia Webex, proporcionadas por la UVa. Las direcciones de las salas concretas se proporcionarán en la página del curso en Moodle.

g.1 Bibliografía básica

- Rajdeep Dua, Manpreet Sing Ghotra, Nick Pentreath. Machine Learning with Spark. Second Edition. Packt Publishing Ltd. 2017.
- Petar Zečević y Marko Bonaći. Spark in Action. Manning Publications. 2016. ISBN: 9781617292606. https://www.manning.com/books/
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An introduction to statistical learning: with applications in R. Springer, 2013.
- Ian H. Witten, Eibe Frank y Mark A. Hall. Data Mining: practical machine learning tools and techniques (third Edition). Morgan Kaufmann, 2011.

g.2. Bibliografía complementaria

- Nick Pentreath. Machine Learning with Spark. Packt Publishing. 2015. ISBN: 9781783288519. http://www.packtpub.com/
- 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/
- Lindsey, James K. Applying Generalized Linear Models. Springer, 1997.
- Jure Leskovek, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. Second edition. Cambridge University Press, 2014.

g.3 Otros recursos telemáticos (píldoras de conocimiento, blogs, videos, revistas digitales, cursos masivos (MOOC), …)

Vídeos con explicación sobre la materia de la asignatura.

Se proporcionarán enlaces a otros contenidos abiertos en internet como blogs o vídeos relacionados con Spark y Aprendizaje Automático.

h. Recursos necesarios

- Notas de la asignatura.
- Guiones de ejercicios y proyectos.
- Curso Moodle de soporte a la asignatura.
- Software gratuito para el desarrollo de análisis de datos escalable como Spark.

i. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
3	Semana 1 – 8

5. Métodos docentes y principios metodológicos

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.

6. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES o PRESENCIALES A DISTANCIA(1)	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
Clases teórico-prácticas (T/M)	8	Estudio y trabajo autónomo individual	15
Clases prácticas de aula (A)		Estudio y trabajo grupal dirigido	30
Laboratorios (L)	15		
Prácticas externas, clínicas o de campo			
Seminarios (S)	7		
Tutorías grupales (TG)			
Evaluación*			
Total presencial	**30**	Total no presencial	**45**
		TOTAL presencial + no presencial	**75**

(1) Por "actividad presencial a distancia" nos referimos a que un grupo siga una videoconferencia de forma síncrona a la clase impartida por el profesor para otro grupo presente en el aula.

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

7. Sistema y características de la evaluación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Proyectos y Mini-proyectos	90%	Se realizarán tres mini-proyectos relacionados con la materia de la asignatura. Cada mini-proyecto tendrá un peso del 30 % de la nota de la evaluación continua.
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.

CRITERIOS DE CALIFICACIÓN
• **Convocatoria ordinaria:** o Se realizará evaluación continua con las actividades y pesos indicados en la tabla anterior. Se exigirá obtener un mínimo de tres puntos en las distintas pruebas para poder hacer la nota media en la evaluación continua. • **Convocatoria extraordinaria:** o La calificación del 100% de la nota de la convocatoria extraordinaria se obtendrá mediante la realización de un único proyecto relacionado con alguna de las técnicas vistas.

Recuerde 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.

8. Consideraciones finales