Proyecto/Guía docente de la asignatura

Asignatura	TÉCNICAS ESCALABLES DE ANÁLISIS DE DATOS EN ENTORNOS BIG DATA: CLASIFICADORES		
Materia	APRENDIZAJE AUTOMÁTICO DE ALTAS PRESTACIONES		
Módulo			
Titulación	MÁSTER EN INGENIERÍA INFORMÁTICA (No presencial)		
Plan	693	**Código**	55115
Periodo de impartición	1er Cuatrimestre	**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 Carlos J. Alonso González		
Datos de contacto (E-mail, teléfono…)	belar@infor.uva.es, ext. 5606 y calonso@infor.uva.es, ext. 5602		
Departamento	INFORMÁTICA (ATC, CCIA y LSI)		

1. Situación / Sentido de la Asignatura

1.1 Contextualización

La asignatura “Técnicas Escalables de Análisis de Datos en entornos Big Data: Clasificadores” introduce los elementos necesarios para aplicar técnicas de Aprendizaje Automático supervisado, concretamente técnicas de clasificación básicas y avanzadas, 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 la asignatura “Arquitecturas Big Data” de la materia “Tecnologías de Gestión de Información”, donde se tratarán problemas asociados a la gestión y almacenamiento de grandes cantidades de datos en entornos distribuidos.

También existe relación con la asignatura “Técnicas escalables de análisis de datos en entornos Big Data: regresión y descubrimiento de conocimiento” de la materia “Aprendizaje Automático de Altas Prestaciones” con la que compartirá la plataforma tecnológica Spark, así como sus bibliotecas ML y MLLIB, además del lenguaje de programación Scala.

1.3 Prerrequisitos

Aunque la asignatura será auto-contenida, 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 en Ingeniería 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”.

También es conveniente que el estudiante sea capaz de leer en inglés técnico y programar en un lenguaje de programación orientado al objeto o funcional (que se estudian en la asignatura “Paradigmas de Programación” del actual grado en Ingeniería Informática de la UVa).

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

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

Comunes a la materia de "Aprendizaje Automático de Altas Prestaciones":

- Comprender los métodos matemáticos, estadísticos y de inteligencia artificial que permiten abordar un problema a partir de datos para obtener conocimiento que permita a un agente la toma de decisiones y/o la interacción con su entorno.
- Ser capaz de aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar y diseñar sistemas inteligentes y sistemas basados en el conocimiento capaces de abordar los aspectos mencionados en el párrafo anterior.
- Conocer, comprender y utilizar plataformas computacionales que den soporte a las técnicas y métodos necesarios para el desarrollo de sistemas inteligentes y sistemas basados en el conocimiento capaces de abordar los aspectos mencionados en los párrafos anteriores.
    - Conocer, comprender y aplicar metodologías para el desarrollo de aplicaciones basadas en datos conducentes al desarrollo de sistemas inteligentes y sistemas basados en el conocimiento capaces de abordar los aspectos mencionados en los párrafos anteriores.

Específicos de esta asignatura:

- Conocer los métodos de inducción de clasificadores en el ámbito del Aprendizaje Automático y Minería de Datos y entender los problemas de escalabilidad en entornos de grandes almacenes de datos.
- Conocer cómo se pueden implementar estas técnicas en frameworks como Apache Hadoop, Apache Spark, etc., así como sus limitaciones.
- Conocer las metodologías para el desarrollo de proyectos de Minería de Datos y Big Data.
- Conocer y ser capaz de aplicar técnicas supervisadas de análisis de datos en el contexto de Big Data.
  Ser capaz de desarrollar aplicaciones Big Data utilizando técnicas supervisadas de análisis de datos, en distintas áreas de aplicación utilizando las tecnologías adecuadas

4. Contenidos y/o bloques temáticos

Bloque 1: "Técnicas Escalables de Análisis de Datos en entornos Big Data: clasificadores"

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. Introducción a Apache Spark.
    a. Spark y sus componentes.
    b. Introducción a Scala y los Dataframes.

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.
    c. Metodología experimental de evaluación y Selección de modelos.

3. Métodos de aprendizaje sobre MLLib de Spark.
    a. Clasificadores básicos.
    b. Clasificadores avanzados.

d. Métodos docentes

Véase el apartado 5.

e. Plan de trabajo

El plan detallado de actividades se proporcionará a los estudiantes en el campus virtual durante la primera semana del curso, una vez se conozcan los detalles precisos de asignación de aulas y recursos.
En general el/la estudiante debe realizar la siguiente actividad para cada tema:
- estudiar el material (visualizar vídeo o leer resumen, estudiar la presentación de diapositivas (comentadas)) del tema,
- revisar el material complementario sobre el tema (si lo hubiese),
- realizar los ejercicios de auto-evaluación asociados a cada tema.
Durante el cuatrimestre se realizarán varias pruebas de evaluación de carácter sumativo. Al final del cuatrimestre se realizará otra.

f. Evaluación

Véase el apartado 7.

g Material docente

g.1 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/
- Mohamed Guller. Big Data Analytics with Spark. Apress. 2015.
- 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

- 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.
- C. Bishop. Pattern Recognition and Machine Learning. Springer, N.Y., 2005
- Jure Leskovek, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. Second edition. Cambridge University Press, 2014.
- L. Kuncheva, Combining pattern classifiers, Second edition. Wiley, 2014.

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

Se proporcionarán vídeos introductorios de cada tema, así como un resumen del mismo destacando los aspectos relevantes.

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 cuestiones y problemas.

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	Semanas 1 a 8 del cuatrimestre

5. Métodos docentes y principios metodológicos

Realización de proyectos.
Laboratorios para la experimentación con las ideas básicas del bloque temático: los estudiantes disponen de Máquinas Virtuales que pueden usar de forma remota y las dudas se resuelven mediante el uso de mensajes en Moodle o tutorías online.

Los contenidos necesarios para la realización de los proyectos se organizan en base a los temas que se han presentado en el apartado 4.c)
Estos contenidos serán proporcionados para la consulta y estudio de los estudiantes:

- Realización de vídeos introductorios a los puntos relevantes de cada tema.
- Resumen detallado de los objetivos de cada tema.
- Notas de la asignatura: presentaciones de diapositivas en formato PDF con comentarios, complementadas por material multimedia disponible online para cada tema.
- Guiones de cuestiones y problemas para autoevaluación.

Se realizarán semanalmente tutorías por video-conferencia. Se programa inicialmente una video-conferencia al terminar cada tema y tras la corrección de un entregable, para resolver dudas de los estudiantes.
Se usarán las herramientas disponibles en el campus virtual de videoconferencias para las tutorías individuales o grupales, asociadas a las consultas de los estudiantes.

Se dispondrá de dos foros en el campus virtual: uno de avisos, donde sólo pueden publicar los profesores y donde se va informando de las novedades de la asignatura (nuevo material, enunciados de entregables, rúbricas, resultados, etc.) y otro de dudas (donde los estudiantes interactúan con los profesores para resolver sus dudas).

Se usará software de libre disposición para el desarrollo de análisis de datos escalable como Spark.

6. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES o PRESENCIALES A DISTANCIA(1)	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
		Visualización y estudio del material depositado	26
		Estudio y trabajo autónomo individual	30
		Estudio y trabajo grupal dirigido	15
		Evaluación	4
Total presencial		Total no presencial	0
		TOTAL presencial + no presencial	**75**

(1) Se considera "Actividad presencial a distancia" si un grupo sigue una videoconferencia de forma síncrona a la clase impartida por el profesor para otro grupo presente en el aula.

7. Sistema y características de la evaluación

Se realizará evaluación continua.

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Cuestiones teórico/prácticas	5%	Realización de ejercicios de programación
Realización de un proyecto	90%	Se realizarán en varias fases, asociadas a evaluar los conocimientos y competencias adquiridos tras los temas 2.a), 2.b) y 2.c) enfocados como las etapas iniciales del proyecto. El proyecto final contendrá las anteriores etapas, añadiendo los conocimientos de los bloques 3.a) y 3.b)
Participación en clases, cuestionarios, seminarios prácticas y tutorías.	5%	La participación en clases, seminarios, prácticas y tutorías se evalúa a partir de las entregas opcionales y la participación en clase.

CRITERIOS DE CALIFICACIÓN

Se mantienen los criterios de evaluación dado que ya se usaba evaluación continua y todas las entregas se realizaban ya online.
 	
• **Convocatoria ordinaria:**
  ○ Se realizará evaluación continua con las actividades y pesos indicados en la tabla anterior. Se exigirá obtener un mínimo de cuatro en las distintas pruebas para poder hacer la nota media.
 	
• **Convocatoria extraordinaria:**
  ○ La calificación del 100% de la nota de la convocatoria extraordinaria se obtendrá mediante la realización de un proyecto. No obstante, los estudiantes podrán solicitar que se conserve en la evaluación extraordinaria aquellas pruebas superadas (nota >= 5.0) durante la convocatoria ordinaria. En ese caso el peso del proyecto en la calificación (y en consecuencia su dificultad) se verá reducida de forma proporcional.

8. Consideraciones finales