# Proyecto docente

| | | | |
|---|---|---|---|
| **Asignatura** | Técnicas de Aprendizaje Automático Escalables | | |
| **Materia** | Ciencia de Datos | | |
| **Titulación** | Máster Universitario en Inteligencia de Negocio y Big Data en Entornos Seguros | | |
| **Plan** | **621 (UVA)** | **Código** | **54549 (UVA)**<br>**8103#90º (UBU)** |
| **Periodo de impartición** | S2 | **Tipo/Carácter** | Obligatoria |
| **Nivel/Ciclo** | Máster | **Curso** | 1 |
| **Créditos ECTS** | 3 | | |
| **Lengua en que se imparte** | Castellano | | |
| **Profesor/es responsable/s** | Carlos J. Alonso González, J. Belarmino Pulido Junquera | | |
| **Datos de contacto (e-mail, teléfono…)** | calonso@infor.uva.es; (Tel. 983185602) belar@infor.uva.es (Tel. 983185606) | | |
| **Horario de tutorías** | **Consultar página web de www.uva.es** | | |
| **Coordinador** | Carlos J. Alonso González | | |
| **Departamento** | Informática (ATC, CCIA y LSI) | | |
| **Web** | | | |
| **Descripción General** | La asignatura cubre las etapas necesarias para la creación de un clasificador o de un recomendador en un entorno Big Data, aplicando las metodologías de Análisis (exploración, limpieza y transformación) de datos en entornos Big Data, y después siguiendo la metodología de selección y evaluación de modelos para elegir los mejores modelos. | | |

## 1. Situación / Sentido de la asignatura

### 1.1 Contextualización

La asignatura “Técnicas de Aprendizaje Automático Escalable” 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.

Esta asignatura se centra en los conceptos básicos del aprendizaje automático, que presenta en el contexto de Big Data, así como en dos tipos de técnicas de aprendizaje supervisado: los clasificadores y los recomendadores.

### 1.2 Relación con otras asignaturas

La asignatura se imparte en el segundo semestre del curso y forma parte de la Materia “Ciencia de Datos” que se centra en el procesamiento escalable de datos. Al igual que para el almacenamiento, la aplicación de técnicas de análisis de grandes volúmenes de datos, se tiene que apoyar en tecnologías adecuadas a la forma de almacenar, el tipo y el volumen de los datos con los que se está tratando.

La asignatura forma parte de esta Materia junto con las asignaturas “Aprendizaje sobre flujos de datos” y “Aprendizaje no supervisado”.

La asignatura proporciona conocimientos y habilidades necesarias para abordar los contenidos de las otras dos asignaturas de la Materia “Ciencia de Datos”, pero principalmente para las técnicas de “Aprendizaje sobre flujos de Datos”, aunque los conocimientos básicos de procesados y limpieza de datos, así como la validación de los modelos es general para la Materia.

Además, los conocimientos y habilidades adquiridas en esta asignatura pueden utilizarse dentro del contexto de la Materia “Inteligencia de Negocio” y las dos asignaturas que allí se imparten: “Inteligencia de negocio aplicada” I y II.

### 1.3 Prerrequisitos

Aunque la asignatura será autocontenida se recomienda haber cursado y superado las asignaturas de la Materia “Tecnologías Informáticas para el Big Data”, especialmente las de “Modelos de programación para el Big Data” y “Arquitecturas Big Data”. Sería recomendable también conocer los contenidos de la asignatura “Almacenamiento Escalable”.

Se supone que el estudiante tiene conocimientos básicos sobre aprendizaje, más concretamente sobre la metodología CRISP-DM para la creación de modelos de aprendizaje automático y que conoce clasificadores básicos como los árboles de decisión o el método Naïve Bayes.

## 2. Competencias

### 2.1 Generales del título

- CG1. Adquisición de competencias teóricas y prácticas para el análisis y diseño de soluciones empresariales en Big Data (almacenamiento y procesamiento de grandes volúmenes de información heterogénea).
- CG3. Capacidad de diseñar e implementar sistemas capaces de extraer conocimiento práctico de grandes volúmenes de datos aplicado al mundo de la empresa (Inteligencia de Negocio/Business Intelligence)

### 2.2 Específicas materia

- CDS1. Capacidad de aplicar, validar y evaluar métodos de Ciencia de Datos/Data Science e Inteligencia Artificial sobre conjuntos y flujos de datos masivos y complejos.
- CDS2. Capacidad de dirigir proyectos para la extracción de conocimiento basados en métodos eficientes de análisis de datos.
- CDS3. Capacidad para el análisis, exploración y síntesis de conjuntos complejos de datos no estructurados y de diseñar soluciones que permitan extraer de los mismos información relevante y valiosa para el soporte a la toma de decisiones.

## 3. Resultados de aprendizaje

Al finalizar la asignatura, el o la estudiante será capaz de:

- Comprender y aplicar métodos de inducción de clasificadores escalables a grandes conjuntos de datos.
- Comprender y aplicar métodos de generación de recomendaciones escalables a grandes conjuntos de datos.
- Comprender y ser capaz de aplicar las metodologías experimentales para la selección de modelos y evaluación de clasificadores.
- Comprender y ser capaz de aplicar la metodología para realizar un proyecto de aprendizaje automático en el contexto de Big Data.
- Conocer y utilizar algunas de las plataformas tecnológicas que permiten desarrollar proyectos de aprendizaje automático en Big Data

## 4. Contenido / Programa de la asignatura

### Bloque 1: Aprendizaje automático escalable en entornos Big Data

**Carga de trabajo en créditos ECTS:** 3

### a. Contextualización y justificación

Véanse los apartados 1.1 y 1.2.

### b. Objetivos de aprendizaje

- Comprender y ser capaz de aplicar métodos de inducción de clasificadores escalables a grandes conjuntos de datos.
- Comprender y ser capaz de aplicar métodos de generación de recomendaciones escalables a grandes conjuntos de datos.
- Comprender y ser capaz de aplicar las metodologías experimentales para la selección de modelos y evaluación de clasificadores.
- Comprender y ser capaz de aplicar la metodología para realizar un proyecto de aprendizaje automático en el contexto de Big Data.
- Conocer y ser capaz de utilizar algunas de las plataformas tecnológicas que permiten desarrollar proyectos de aprendizaje automático en Big Data

### c. Contenidos y materiales de aprendizaje

1. Introducción a las plataformas tecnológicas (Apache Spark, Scala u otras)
2. Metodología de análisis Big Data. Preparación de datos.
3. Clasificadores escalables básicos
4. Metodología experimental de evaluación y selección de modelos
5. Recomendadores

### d. Métodos docentes

Los especificados en el apartado 5.

### e. 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/
- 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.

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


## g. Recursos necesarios

- Se utilizará el Moodle del campus virtual: ubuvirtual.ubu.es, donde se colocará el material de la asignatura
- Las herramientas de comunicación serán:
  - Asíncronos: foros, emails y mensajes directos dentro del Moodle del campus virtual
  - Síncronos: Salas de videoconferencia LifeSize proporcionadas por la UVA o bien salas Webex. Las direcciones de las salas concretas se proporcionarán en la página del curso en Moodle.

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

- Se proporcionarán vídeos introductorios de cada tema, así como un resumen de cada uno, destacando los aspectos relevantes
- Se proporcionarán enlaces a otros contenidos abiertos en internet como blogs o vídeos relacionados con Spark y Aprendizaje Automático

## h. Temporalización

| BLOQUE 1 | CARGA ECTS | PERIODO PREVISTO DE DESARROLLO |
|---|---|---|
| Introducción a las plataformas tecnológicas (Apache Spark, Scala u otras) | 0,5 | Semana 1 |
| Metodología de análisis Big Data. Preparación de datos. | 0,8 | Semanas 1-2-3 |
| Clasificadores escalables básicos | 0,8 | Semanas 3-4 |
| Metodología experimental de evaluación y selección de modelos | 0,7 | Semanas 4-5 |
| Recomendadores | 0,2 | Semana 5 |

## 5. Metodología de enseñanza y dedicación del estudiante a la asignatura

| Actividad Formativa | Horas | Presencialidad (%) | Competencias Relacionadas |
|---|---|---|---|
| Clases, conferencias y técnicas expositivas | 12 | 0 | CG1, CG3, CDS1, CDS3 |
| Actividades autónomas y en grupo (trabajos y lecturas dirigidas) | 45 | 0 | CG1, CG3, CDS2, CDS3 |
| Pruebas de seguimiento y exposición de trabajos | 10 | 50 | CG1, CG3 |
| Tutoría individual, participación en foros y otros medios colaborativos | 8 | 0 | CG1, CG3, CDS1 |

## 6. Tabla de dedicación del estudiante a la asignatura

| ACTIVIDADES | HORAS |
|---|---|
| Horas de tutoría síncrona o asíncrona | 8 |
| Horas de lectura y reproducción materiales | 22 |
| Horas de trabajo autónomo individual | 30 |
| Horas de trabajo en grupo | 15 |
| Total | 75 |

## 7. Temporalización (por bloques temáticos)

Véase el apartado 4.f, ya que sólo hay un bloque temático de 3 ECTS que se desarrollará en 5 semanas.

## 8. Evaluación

| Instrumento / Procedimiento | Peso primera convocatoria | Peso segunda Convocatoria |
|---|---|---|
| Evaluación sumativa I | 20% | 20% |
| Realización de trabajos, proyectos, resolución de problemas y casos | 70% | 70% |
| Participación en foros y otros medios participativos | 10% | 10% |

**Criterios / Comentarios a la evaluación**

- **Primera convocatoria:** Será necesario realizar todos los trabajos asignados, obteniendo un mínimo de 4 puntos en todos ellos, para optar a la nota promediada.
- **Segunda convocatoria:** Será necesario realizar todos los trabajos asignados, que serán diferentes de los de la primera convocatoria, para optar a la nota promediada.
- La calificación de participación en foros y otros medios se obtendrá durante el período de la primera o de segunda convocatoria.

**Participación en los foros.**

Los foros de la asignatura serán moderados a posteriori por los profesores responsables de la misma, eliminando cualquier mensaje que no se considere adecuado.

Por supuesto, el lenguaje será correcto en todo momento, no admitiéndose ningún tipo de descalificaciones o comentarios que puedan resultar ofensivos para cualquier participante.

El foro tiene una naturaleza científico-técnica. No se admitirá ningún comentario que no se ciña a los aspectos científico-técnicos relacionados con la asignatura.

No se admitirá ningún tipo de opinión personal más allá de las opiniones que con base científica-técnica se pueden presentar como conclusiones de los trabajos realizados.

La moderación de los comentarios tendrá la siguiente penalización sobre los alumnos autores del comentario:

Tras el primer comentario moderado se enviará un aviso a los alumnos implicados.

Tras el segundo comentario moderado se restará un punto de la nota final de los alumnos implicados.

Tras el tercer comentario moderado se suspenderá el acceso al foro a los alumnos implicados y se les restarán 3 puntos adicionales de la nota final.

## 9. Consideraciones / Comentarios adicionales