Máster en Técnicas Estadísticas

DATOS IDENTIFICATIVOS DE LA MATERIA

Nombre de la materia: Tecnologías de Gestión de Datos

Número de créditos ECTS: 5

Curso académico: 2019/2020

Profesorado:

Guillermo López Taboada (coordinador, 3.5 créditos)

Rubén Fernández Casal (1.5 créditos)

OBJETIVOS DE LA MATERIA

En esta materia se pretende familiarizar al alumnado con las tecnologías de gestión de datos, tanto las tradicionales (datos estructurados) como las más modernas (bases de datos NoSQL y Big Data/Hadoop). Los objetivos a alcanzar como resultado del aprendizaje son:

• Manejar de forma autónoma y solvente el software necesario para acceder a conjuntos de datos en entornos profesionales y/o en la nube.

• Saber gestionar conjuntos de datos masivos en un entorno multidisciplinar que permita la participación en proyectos profesionales complejos que requieran el uso de técnicas estadísticas.

• Saber relacionar el software de diseño y gestión de bases de datos con el específicamente implementado para el análisis de datos.

CONTENIDOS DE LA MATERIA

Tema 1. **Introducción al lenguaje SQL.**
1.1 Bases de datos relacionales
1.2 Sintaxis SQL
1.3 Conexión con bases de datos desde R

Tema 2. **Introducción a tecnologías NoSQL**
2.1 Conceptos y tipos de bases de datos NoSQL (documental, columnar, clave/valor y de grafos)
2.2 Conexión de R a NoSQL

Tema 3. **Tecnologías para el tratamiento de datos masivos**
3.1 Tecnologías Big Data (Hadoop, Spark, Hive, Rspark, Sparklyr)
3.2 Visualización y generación de cuadros de mando
3.3 Introducción al análisis de datos masivos.

BIBLIOGRAFÍA BÁSICA Y COMPLEMENTARIA

Básica

Daroczi, G. (2015). Mastering Data Analysis with R (1st edition). Packt Publishing
Grolemund, G.; Wickham, H. (2016). R for Data Science (1st edition). O'Reilly
Bajar Silberschatz, A.; Korth, H.; Sudarshan, S. (2014) Fundamentos de Bases de Dato. Mc Graw Hill
Fernández Casal, R. (2019) Ayuda y Recursos para el Aprendizaje de R https://rubenfcasal.github.io/post/ayuda-y-recursos-para-el-aprendizaje-de-r/

Complementaria

McKinney, W. (2017) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2ª ed.) O'Reilly
White, T. (2015) Hadoop: The Definitive Guide (4ª ed.). O'Reilly
Holmes, A. (2014) Hadoop in practice (2ª ed.). Manning
Centro de Supercomputación de Galicia (2019) Servicio de Big Data del CESGA https://bigdata.cesga.es/

COMPETENCIAS

En esta materia se trabajarán las competencias básicas, generales y transversales recogidas en la memoria del título. Se indican a continuación cuáles son las competencias específicas, que se potenciarán en esta materia:

E1 - Conocer, identificar, modelar, estudiar y resolver problemas complejos de estadística e investigación operativa, en un contexto científico, tecnológico o profesional, surgidos en aplicaciones reales.
E2 - Desarrollar autonomía para la resolución práctica de problemas complejos surgidos en aplicaciones reales y para la interpretación de los resultados de cara a la ayuda en la toma de decisiones.
E3 - Adquirir conocimientos avanzados de los fundamentos teóricos subyacentes a las distintas metodologías de la estadística y la investigación operativa, que permitan su desarrollo profesional especializado.
E6 - Adquirir conocimientos teórico-prácticos avanzados de distintas técnicas matemáticas, orientadas específicamente a la ayuda en la toma de decisiones, y desarrollar capacidad de reflexión para evaluar y decidir entre distintas perspectivas en contextos complejos.

E8 - Adquirir conocimientos teórico-prácticos avanzados de las técnicas destinadas a la realización de inferencias y contrastes relativos a variables y parámetros de un modelo estadístico, y saber aplicarlos con autonomía suficiente un contexto científico, tecnológico o profesional.
E9 - Conocer y saber aplicar con autonomía en contextos científicos, tecnológicos o profesionales, técnicas de aprendizaje automático y técnicas de análisis de datos de alta dimensión (big data).
E10 - Adquirir conocimientos avanzados sobre metodologías para la obtención y el tratamiento de datos desde distintas fuentes, como encuestas, internet, o entornos “en la nube".

METODOLOGÍA DOCENTE

La enseñanza constará de clases expositivas e interactivas, así como de la tutorización del aprendizaje y de las tareas encomendadas al alumnado.

En las clases expositivas e interactivas se resolverán ejemplos mediante el software R, por lo que es necesario que el alumnado disponga en el aula de un ordenador.

Asimismo, se llevarán a cabo seminarios en cuestiones específicas de R así como de otras tecnologías como Python y el servicio de Big Data del Centro de Supercomputación de Galicia.

Se propondrán actividades para el alumnado, que consistirán en la realización tanto de ejercicios prácticos como trabajos tutelados de ingesta, manipulación, visualización y análisis exploratorio de datos para poder aplicar técnicas estadísticas sobre un rango cada vez mayor de fuentes de datos, tanto estructuradas/SQL como no estructuradas/NoSQL e incluso sobre lagos de datos (Hadoop).

Se proporcionarán los apuntes de la materia, así como otro material orientativo del aprendizaje del software. Los apuntes y otros instrumentos didácticos estarán disponibles a través de alguna herramienta de acceso por vía web.

CRITERIOS Y MÉTODOS DE EVALUACIÓN

Evaluación continua (40%): la evaluación continua se realizará en base a la realización de prácticas de laboratorio, tanto presenciales o no presenciales, usando principalmente R y documentando las conclusiones extraídas, por parte del alumnado (30%) así como la realización de trabajos tutelados (10%). La calificación obtenida se conservará entre las oportunidades (ordinaria y extraordinaria) dentro de la convocatoria de cada curso. Con las prácticas de laboratorio se valorará el nivel de adquisición de las competencias básicas CB7 y CB8, las específicas CE1, CE3, CE9 y CE10, y las transversales CT2 y CT3. Así mismo, con los trabajos tutelados se valorará el nivel de adquisición de las competencias básicas CB6-CB10, las generales CG1-CG5, las específicas CE1-CE2, CE6, CE8-CE10, y las transversales CT1-CT5.

Examen final (60%): el examen final constará de varias cuestiones teórico-prácticas sobre los contenidos de la materia, dentro de las que se podrá incluir la propuesta de técnicas de tratamiento de datos y la interpretación de resultados, tanto conceptualmente como

de forma aplicada utilizando el lenguaje R. En el examen, se evaluarán las competencias específicas: CE1, CE3, CE6, CE8-CE10, las competencias básicas CB6-CB9, las competencias generales CG1 y CG2, y las competencia transversal CT13.
Para poder aprobar la asignatura en la primera oportunidad será necesario obtener como mínimo el 30% de la nota máxima de la suma de las prácticas de laboratorio y trabajos tutelados e, igualmente, el 30% de la nota máxima final de la Prueba mixta (examen), y tener una nota total (prácticas más trabajos tutelados más prueba mixta) igual o superior al 50% de la nota máxima.

En la segunda oportunidad solamente se podrá recuperar la nota del examen. Las notas de prácticas y de trabajos tutelados serán las obtenidas durante el curso. Para los alumnos que utilicen la oportunidad adelantada de diciembre se utilizarán las notas de prácticas y trabajos tutelados que obtuvieran en su último curso. En esta oportunidad solo será necesario para aprobar obtener una nota total igual o superior al 50% de la nota máxima.

Una vez que un estudiante es evaluado en una práctica de laboratorio o en un trabajo tutelado implica que será calificado. Por tanto, la calificación "No Presentado" no es posible una vez que una práctica/trabajo ha sido evaluada.

TIEMPO DE ESTUDIO Y DE TRABAJO PERSONAL QUE DEBE DEDICAR UN ESTUDIANTE PARA SUPERAR LA MATERIA

Cada crédito ECTS se traduce en 7 horas de clase de tipo presencial. Se estima que el alumnado necesitará aproximadamente unas 2,5 horas de trabajo autónomo por cada hora presencial para la comprensión global de los contenidos, incluyendo las actividades asociadas a ejercicios y otras tareas. En total resultan 25 horas por crédito ECTS.

RECOMENDACIONES PARA EL ESTUDIO DE LA MATERIA

Debido al fuerte componente práctico es recomendable ir haciendo las actividades prácticas y trabajos académicamente dirigidos de forma regular a lo largo del cuatrimestre.

RECURSOS PARA EL APRENDIZAJE

Bibliografía y apuntes. Uso del campus virtual de la universidad y del sitio web del Máster en Técnicas Estadísticas como soporte para el material del programa.

Las herramientas software utilizadas en esta materia son generalmente open-source o tienen licencia gratuita para estudiantes.

OBSERVACIONES

El desarrollo de los contenidos de la materia se realizará teniendo en cuenta que las competencias a adquirir por el alumnado deben cumplir con el nivel MECES3. En esta asignatura se intentará que cualquier alumno, independientemente de su formación previa, adquiera un sólido conocimiento de las tecnologías de gestión de bases de datos, tanto relacionales como no relacionales. Asimismo, se buscará una familiarización con las

principales técnicas computacionales para la gestión práctica de datos masivos. Esto dotará al alumno de una gran autonomía a la hora de procesar y estudiar datos, independientemente de su formato y origen.