MINERÍA DE DATOS II

1- DESCRIPCIÓN

La asignatura de Minería de Datos II continúa el trabajo iniciado de Minería de Datos I al alumno en el conocimiento y la aplicación de técnicas avanzadas de predicción y de minería de textos para extraer información clave. El curso ofrece formación básica en la aplicación de algoritmos avanzados mediante software R para la toma de decisiones empresariales.

2- OBJETIVO

Formar profesionales capaces de comprender la información contenida en los datos almacenados por las organizaciones, o la información contenida en los ficheros de texto, Internet, etc. para encontrar patrones o reglas que una vez validados puedan ser incorporados en los sistemas de análisis de información de las empresas.

Los fines específicos de la asignatura son:

Conocer técnicas avanzadas de predicción que utiliza la minería de datos para obtener unos resultados.

Comprender la información contenida en una base de datos con el uso de técnicas de Machine Learning.

Explotación de técnicas de minería de textos, con el fin de obtener valor sobre una base de datos de texto.

Conocer y saber usar el programa R para el análisis de datos y su procesamiento mediante técnicas de minería de datos.

3- CONOCIMIENTOS PREVIOS

Se recomienda haber cursado la asignatura de "Minería de Datos I" en el primer primer cuatrimestre de segundo curso, un nivel de comprensión de lectura medio en el idioma inglés y un conocimientos básicos de programación en R.

4- CONTENIDOS

1. MODELOS AVANZADOS
    - 1.1 Introducción data mining. Análisis supervisado y no supervisado.
    - 1.2 Modelos ensamblados
    - 1.3 Bagging
    - 1.4 Random forest
    - 1.5 Boosting

2. TEXT MINING
    - 2.1 Introducción al text mining. Definiciones y aplicaciones reales.
    - 2.2 Extracción de datos de Twitter
    - 2.3 Limpieza de texto y tokenización
    - 2.4 Term Frequency e Inverse Document Frequency
    - 2.5 Wordcloud

5- ACTIVIDADES FORMATIVAS

La asignatura de Minería de Datos II tiene un enfoque eminentemente práctico.

La materia se desarrolla en dos tipos de sesiones:
Sesiones teórico-prácticas, en las que, por un lado, se llevará a cabo la exposición de los contenidos del programa mediante el uso de clases magistrales por parte del profesor, así como con el uso de metodologías ágiles que favorecen el proceso de enseñanza-aprendizaje; y por otro lado, se realizarán ejercicios, problemas y casos sencillos relacionados con los conocimientos impartidos en la misma sesión. En estas sesiones se fomentará el uso de la metodología flipped learning o aprendizaje inverso.
Sesiones prácticas que se desarrollarán en el aula, individualmente o en grupos de alumnos, con el programa R donde se analizarán y procesarán datos obtenidos de distintas fuentes aplicando los algoritmos desarrollados en clase.

Por último, se organizarán grupos de alumnos que deberán desarrollar un trabajo que recoja todo el proceso de minería de datos: selección del conjunto de datos, análisis de las propiedades de los datos recopilados, selección y aplicación de la técnica de minería de datos (construcción de un modelo de predicción y/o de clasificación), extracción del conocimiento y validación del modelo y de que las conclusiones son satisfactorias. Estos trabajos se elaborarán siguiendo la metodología del aprendizaje basado en problemas.

El alumno podrá hacer uso de tutorías tanto individuales como grupales para obtener unos mejores resultados del aprendizaje.
Los trabajos en equipo exigen una supervisión continuada por parte del profesor por lo que los alumnos deberán concertar tutorías a lo largo de todo el semestre para garantizar el cumplimiento de los tiempos de realización de cada parte del proceso, así como para orientar y resolver dudas sobre el trabajo a desarrollar.

Las actividades formativas, así como la distribución de los tiempos de trabajo, pueden verse modificadas y adaptadas en función de los distintos escenarios establecidos siguiendo las indicaciones de las autoridades sanitarias*.*

6- DISTRIBUCIÓN DE LOS TIEMPOS DE TRABAJO

ACTIVIDAD PRESENCIAL: 30 horas
TRABAJO AUTÓNOMO/ACTIVIDAD NO PRESENCIAL: 45 horas

7- COMPETENCIAS

COMPETENCIAS BÁSICAS
CB1 - Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.
CB2 - Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio

CB3 - Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética
CB4 - Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado
CB5 - Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía

COMPETENCIAS GENERALES

CG1 - Capacidad de organización, sistematización y planificación en la identificación de problemas, pautas y modelos en el contexto del big data
CG10 - Compromiso ético en la sociedad de la información
CG2 - Capacidad para el cumplimiento de objetivos, resolución de problemas y toma de decisiones en un entorno de datos masivos tanto cuantitativos como cualitativos
CG3 - Capacidad para analizar datos a gran escala procedentes de diferentes fuentes: audiovisuales, textos y numéricas
CG4 - Capacidad de diseñar e implementar proyectos e informes, utilizando con naturalidad los canales digitales
CG5 - Capacidad de liderazgo y de trabajar en equipo en la sociedad de la información
CG7 - Capacidad de pensamiento crítico, autocrítico, analítico y reflexivo
CG8 - Capacidad de aprendizaje autónomo en la sociedad de la información

COMPETENCIAS ESPECÍFICAS

CE10 - Saber identificar y resolver problemas reales de la empresa, a través del análisis avanzado de datos y de la elección de técnicas adecuadas para la toma de decisiones
CE17 - Saber manejar herramientas cuantitativas e informáticas para la toma de decisiones
CE24 - Conocer los fundamentos de la estadística multivariante (data mining) así como su aplicación en el mundo del big data
CE27 - Conocer y saber utilizar a nivel de analista herramientas de estadística avanzada, almacén de datos, bases de datos relacionales, bases de datos no relacionales y sistemas de gestión de big data.

8- RESULTADOS DEL APRENDIZAJE

Conoce y selecciona métodos y algoritmos de la minería de datos para identificar y resolver problemas reales de la empresa.
Utiliza con soltura el programa R en el análisis de datos para la toma de decisiones empresariales.
Conoce y aplica las técnicas de text mining aprendidas durante el curso.
Sabe extraer, comprender y analizar la información contenida en grandes volúmenes de datos.

9- SISTEMAS DE EVALUACIÓN DEL APRENDIZAJE

El sistema de evaluación del aprendizaje será mediante evaluación continua, examen teórico-práctico, realización de ejercicios, realización y ejecución de problemas y prácticas, valoración de trabajos y de la participación en clase. La asistencia a clases expositivas y prácticas es obligatoria para someterse a este sistema.

Todos los alumnos deberán realizar un examen final independientemente de la calificación obtenida en exámenes parciales, evaluación continua, pruebas, etc. Para hacer media, la calificación obtenida en dicha prueba final debe ser de al menos un 5.

En ningún caso, la superación de la asignatura podrá reducirse a la aprobación de un examen.

Todas las pruebas susceptibles de evaluación estarán supeditadas a lo establecido en la normativa de evaluación de la universidad francisco de vitoria.

Las conductas que defrauden el sistema de comprobación del rendimiento académico, tales como plagio de trabajos o copia en exámenes son consideradas faltas graves según el artículo 7 de la normativa de convivencia de la ufv y serán aplicadas las sanciones oportunas como recoge el artículo 9 del mismo documento.

Los exámenes se realizarán de manera presencial siempre y cuando la situación sanitaria lo permita, pudiendo ser modificados con el objetivo de cumplir las indicaciones dadas por las autoridades.

SISTEMA DE EVALUACIÓN PARA ALUMNOS DE PRIMERA MATRÍCULA

1. Bloque 1: modelos avanzados
    a. Actividades diarias, ejercicios y trabajos individuales: 12.5%
    b. Exámenes parciales: 12.5%
2. Bloque 2, minería de textos: Elaboración de trabajos grupales: 25%
3. Examen final: 50%

ALUMNOS CON DISPENSA ACADÉMICA

1. Trabajos presentados a requerimiento del profesor: 30%
2. Examen final (será el mismo que los alumnos de asistencia normal): 70%

ALUMNOS DE SEGUNDA Y SUCESIVAS MATRÍCULAS

En este caso los alumnos podrán optar entre cualquiera de los dos sistemas anteriores, previa comunicación al docente al inicio del semestre.

CONVOCATORIA EXTRAORDINARIA

Se aplicará el mismo criterio que en los casos anteriores.
El tiempo fijado para los trabajos prácticos está programado para ejecutarse en ese tiempo, por lo que las fechas de entrega se han de cumplir. El calendario de entrega de estos trabajos se proporcionará al comienzo del curso.

Se mantendrá el sistema de evaluación en caso de confinamiento.

NOTA SOBRE PLAGIO

Cualquier fraude o plagio (*) por parte del alumno en una actividad evaluable será sancionado e implicará un 0 en la calificación de esa parte de la asignatura, anulando la convocatoria en curso. Este comportamiento, además, será comunicado a la Dirección de la Carrera que a su vez comunicará a la Dirección General, siguiendo el Protocolo establecido por la Universidad Francisco de Vitoria.

(*) Se considera “plagio” cualquier tipo de copia de cuestiones o ejercicios de examen, memorias de trabajos, prácticas, etc., ya sea de manera total o parcial, de trabajos ajenos al alumno con el engaño de hacer creer al profesor que son propios.

10- BIBLIOGRAFÍA Y OTROS RECURSOS

- An Introduction to statistical learning, **James**, G., **Witten**, D., **Hastie**, T., **Tibshirani**, R.. Enlace: http://www-bcf.usc.edu/~gareth/ISL/
- Kuhn, Max and Johnson, Kjell. **Applied predictive modeling.**
- Julia Silge and David Robinson, **Text mining with R**