

**MINERIA DE DATOS USANDO SISTEMAS
INTELIGENTES****Año 2023****Carrera/Plan:**Licenciatura en Sistemas Planes 2021/2015/2012
Licenciatura en Informática Planes
2021/2015/2012**Área:** Algoritmos y Lenguajes**Año:** 4º o 5º año**Régimen de Cursada:** *Semestral***Carácter:** Optativa**Correlativas:** Algoritmos y Estructuras de Datos
– Matemática 3**Profesora:** Dra. *Laura Lanzarini***Hs semanales:** 2 hs de teoría – 4 hs de práctica**FUNDAMENTACIÓN**

La Minería de Datos reúne un conjunto de técnicas que, a partir de grandes volúmenes de datos, poseen la capacidad de adquirir conocimiento nuevo, novedoso y potencialmente útil. El resultado de la aplicación de estas técnicas es un modelo de la información disponible capaz de resumir las relaciones existentes entre los datos. Ejemplos de este tipo de modelos son: redes neuronales, árboles de decisión, reglas de clasificación, entre otros.

Habitualmente, ante la presencia de grandes volúmenes de información, lo que se hace es contrastar una hipótesis predeterminada, por ejemplo, a través de consultas SQL. En Minería de Datos, el proceso es totalmente inverso llegando a obtener relaciones entre los datos sin tener ninguna hipótesis preestablecida.

Este curso tiene como eje central la resolución de problemas concretos utilizando estas técnicas. Se espera que el alumno adquiera los conceptos necesarios para analizar un problema y presentar los resultados obtenidos de una manera útil para la toma de decisiones.

OBJETIVOS GENERALES

Introducir al alumno en las técnicas de Minería de Datos. Se analizarán modelos basados en regresión, árboles, reglas, redes neuronales y técnicas de agrupamiento. Se cubrirán las distintas etapas del proceso de Extracción de Conocimiento como herramienta de ayuda a la toma de decisiones. El énfasis está puesto en la resolución de problemas de clasificación y predicción.

COMPETENCIAS

- LI-CE4- Planificar, dirigir, realizar y/o evaluar proyectos de relevamiento de problemas del mundo real, especificación formal de los mismos, diseño, implementación, prueba, verificación, validación, mantenimiento y control de calidad de sistemas de software/sistemas de información que se ejecuten sobre equipos de procesamiento de datos, con capacidad de incorporación de tecnologías emergentes del cambio tecnológico. Capacidad de análisis, diseño y evaluación de interfaces humano computador y computador-computador.

- LS-CE1- Planificar, dirigir, realizar y/o evaluar proyectos de relevamiento de problemas del mundo real. Especificación formal, diseño, implementación, prueba, verificación, validación, mantenimiento y control de calidad de sistemas de software que se ejecuten sobre sistemas de procesamiento de datos, con capacidad de incorporación de tecnologías emergentes del cambio tecnológico. Capacidad de análisis, diseño y evaluación de interfaces humano computador y computador-computador.

CONTENIDOS MINIMOS (de acuerdo al Plan de Estudios)

Introducción a la Minería de Datos.

Técnicas

- Árboles de decisión.
- Reglas de clasificación y asociación
- Técnicas de Agrupamiento
- Redes Neuronales

Evaluación de Modelos

PROGRAMA ANALÍTICO

- Introducción. Obtención de conocimiento a partir de los datos. El concepto de patrón. El proceso KDD. Fases del proceso de extracción del conocimiento. La Minería de Datos como fase del proceso KDD. Relación con otras disciplinas.
- Recuperación de información vs recuperación de datos. Proceso de recuperación de información.
- Preparación de Datos. Metadatos. Análisis de la información de entrada. Medidas estadísticas básicas. Construcción y análisis de representaciones gráficas. Medidas de similitud entre atributos y entre ejemplos. Limpieza y transformación. Transformación y creación de atributos. Discretización y Numerización, Normalización de rango, escalado y centrado. Exploración mediante visualización y selección de datos.
- Técnicas de Minería de Datos. Extracción de Patrones. Introducción. Tareas y Métodos. Tareas predictivas y descriptivas. Aprendizaje supervisado y aprendizaje no supervisado. La Minería de Datos y el aprendizaje inductivo. Comparación de las técnicas de Minería de Datos.
- Técnicas de Agrupamiento. Métricas de calidad del agrupamiento. Tipos de agrupamiento: Jerárquico, particivo y probabilista. Medidas de distancia y de conectividad. Proceso de agrupamiento. Clustering particivo. Algoritmo k-medias. Algoritmos de clustering jerárquicos aglomerativos y divisivos. Dendrogramas. Algoritmo probabilista EM (Expectation - Maximization)
- Árboles de decisión. Métricas de selección de atributos. Entropía. Ganancia de Información. Tasa de Ganancia. Índice Gini. Poda y Sobreajuste. Algoritmos Id3, C4.5 y Random Forest. Algoritmos de Boosting. AdaBoost.
- Clasificadores bayesianos. Teorema de Bayes. Hipótesis máxima a posteriori. Clasificador Naïve Bayes. Ejemplos.
- Reglas de clasificación. Partición vs cobertura. Métodos ZeroR, OneR, PRISM, PART y CN2. Métricas de una regla: soporte, cobertura, confianza, interés y convicción.
- Reglas de asociación. Calidad de las reglas. Algoritmo A priori. Concepto de ítem frecuente. Mejoras del algoritmo a priori: FP-Growth y FP-Tree.
- Redes Neuronales Feedforward. Descripción de la arquitectura. Regla delta generalizada. Algoritmo de entrenamiento backpropagation. Incorporación del término de momento. Capacidad de generalización de la red. Resolución de problemas de clasificación y predicción.
- Análisis y difusión del modelo obtenido. Evaluación de modelos. Comparación de técnicas de aprendizaje. Evaluación y mejora del modelo obtenido. Performance del modelo. Matriz de confusión. Sensibilidad, especificidad, precisión y recall. F measure. Visualización utilizando curvas ROC.

BIBLIOGRAFIA BASICA

- Hernández Orallo, Ramírez Quintana, Ferri Ramírez. *Introducción a la Minería de Datos*. Prentice Hall. 2004. ISBN 84-205-4091-9.
- Ian H. Witten, Eibe Frank, Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*, (Fourth Edition). Morgan Kaufmann. 2017. ISBN 978-0-12-804291-5.

BIBLIOGRAFIA COMPLEMENTARIA

- Mehmed Kantardzic. *Data Mining. Concepts, Models, Methods, and Algorithms*. (Third Edition). Wiley. IEE Press. 2020. ISBN: 978-1-119-51604-0
- Mohammed Zaki, Wagner Meira, Jr. *Data Mining and Machine Learning. Fundamental Concepts and Algorithms* (Second Edition). Cambridge University Press. 2020. ISBN: 978-1-108-47398-9.
- Jiawei Han, Jian Pei, Hanghang Tong. *Data Mining: Concepts and Techniques*, (Fourth Edition). Morgan Kaufmann. 2022. ISBN- 978-0-128-11761-3.

METODOLOGÍA DE ENSEÑANZA

La asignatura tiene como objetivo central capacitar al alumno en el uso de técnicas de Minería de Datos para la modelización y análisis de la información disponible, por tal motivo, el dictado del curso presenta una fuerte coordinación entre la teoría y la práctica. Esta articulación se lleva a cabo mediante la ejemplificación y resolución de situaciones concretas similares a las planteadas en las actividades prácticas.

Al inicio del curso se publica el cronograma de actividades. En dicho cronograma se detallan los temas a desarrollar en cada clase tanto teórica como práctica.

ACTIVIDADES PRACTICAS

Durante el desarrollo del curso se publicarán autoevaluaciones, en forma periódica, con el objetivo de que los alumnos comprueben los conocimientos adquiridos en forma voluntaria.

También se definirán entregas de ejercicios prácticos y la elaboración de un trabajo práctico integrador que deberá ser entregado al finalizar el curso.

EVALUACIÓN

Cada alumno puede optar por una de las siguientes formas de aprobación:

a) Régimen de promoción

Los alumnos que elijan esta opción deberán:

- Entregar durante la cursada la resolución de algunos ejercicios seleccionados de cada enunciado de práctica.
- Entregar al final del curso la resolución de un trabajo práctico integrador y
- Aprobar con nota mayor o igual a 6 (seis) puntos el examen que se tomará al final del curso. Este examen cuenta con 2 (dos) recuperatorios.

b) Régimen convencional

Los alumnos que elijan esta opción deberán:

- Entregar durante la cursada la resolución de algunos ejercicios seleccionados de cada enunciado de práctica.
- Aprobar con nota mayor o igual a 4 (cuatro) puntos el examen que se tomará al final del curso. Este examen cuenta con 2 (dos) recuperatorios. Quienes aprueben con nota mayor o igual a 4 (cuatro) puntos obtendrán la cursada de la asignatura debiendo luego rendir examen final.

CRONOGRAMA DE CLASES Y EVALUACIONES

Semana		Contenidos/Actividades
1	06-mar	Introducción. Proceso KDD. Análisis y visualización de atributos..
2	13-mar	Preprocesamiento de atributos. Normalización.
3	20-mar	Agrupamiento Partitivo. Algoritmo k-means. Métricas de calidad.
4	27-mar	Agrupamiento Jerárquico. Algoritmo probabilista EM
5	03-abr	Clasificador Naive Bayes
6	10-abr	Análisis de modelos predictivos. Métricas para selección de atributos
7	17-abr	Arboles de Clasificación. Algoritmos ID3, C4.5
8	24-abr	Arboles de Clasificación (cont). Random Forest
9	01-may	Reglas de clasificación
10	08-may	Reglas de asociación.
11	15-may	Redes Neuronales. Perceptrón y multiperceptrón
12	22-may	Análisis de resultados y modelos obtenidos
13	29-may	Trabajo Integrador y consultas para el parcial
14	05-jun	1ra. Fecha de parcial
15	12-jun	Muestra de exámenes de la 1ra. Fecha. Consultas referidas al trabajo integrador y a la 2da. Fecha de parcial
16	19-jun	2da. Fecha de parcial
17	26-jun	Muestra de exámenes de la 2da. Fecha. Consultas referidas al trabajo integrador y a la 3era. Fecha de parcial
18	03-jul	3ra. Fecha de parcial
19	10-jul	Muestra de exámenes de la 3ra. Fecha

Evaluaciones previstas	Fecha (a confirmar)
1ra.Fecha de examen	07/06/2023
2da. Fecha de examen	21/06/2023
3ra. Fecha de examen	05/07/2023

Contacto de la cátedra

- e-mail: midusi.unlp@gmail.com
- WEB : weblidi.info.unlp.edu.ar/catedras/md_si
- Plataforma virtual de gestión de cursos: IDEAS.



Dra. Laura Lanzarini