MINERIA DE DATOS USANDO SISTEMAS INTELIGENTES	**Carrera/Plan:** Licenciatura en Sistemas Planes 2015/2012/2003-07 Licenciatura en Informática Planes 2015/2012/2003-07
Año 2021	**Área:** Algoritmos y Lenguajes **Año:** 4º o 5º año **Régimen de Cursada:** *Semestral* **Carácter:** Optativa **Correlativas**: Algoritmos y Estructuras de Datos – Matemática 3 **Profesor**: Dra. *Laura Lanzarini* **Hs semanales**: 6 hs

FUNDAMENTACIÓN

La Minería de Datos reúne un conjunto de técnicas que, a partir de grades 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 poder 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 interfases 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 interfases humano computador y computador-computador.

CONTENIDOS MINIMOS (de acuerdo al Plan de Estudios)

Introducción a la Minería de Datos.
Técnicas
- Arboles 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, partitivo y probabilista. Medidas de distancia y de conectividad. Proceso de agrupamiento. Clustering partitivo. Algoritmo k-medias. Algoritmos de clustering jerárquicos aglomerativos y divisivos. Dendrogramas. Algoritmo probabilista EM (Expectation - Maximization)
- Arboles de decisión. Métricas de selección de atributos. Entropía. Ganancia de Información. Tasa de Ganancia. Indice Gini. Poda y Sobreajuste. Algoritmos Id3, C4.5 y Random Forest. Construcción de árboles para grandes volúmenes de datos.
- 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. Aprendizaje profundo.
- 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

- Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques, (Third Edition). Morgan Kaufmann. 2013. ISBN-13: 978-0123814791.
- Nong Ye . *Data Mining: Theories, Algorithms, and Examples.* CRC Press. 2013. ISBN 9781439808382.
- Kohonen, T. *Self-Organizing Maps*. 2nd Edition. Springer. ISSN 0720-678X. 1997.
- Karray and De Silva. *Soft Computing and Intelligent Systems Design Theory, tools and Applications.* Peason Education. 2004. ISBN 0-321-11617-8

METODOLOGÍA DE ENSEÑANZA

La materia se dictará bajo la modalidad de taller mediante clases virtuales sincrónicas en las que, además de ver los temas teóricos, se resolverán y analizarán ejemplos sencillos y concretos. Dichas clases serán grabadas y puestas a disposición de los alumnos para posterior consulta y revisión de los temas.

MATERIAL DEL CURSO Y COMUNICACION

Todo el material del curso estará disponible a través de la plataforma de educación a distancia *Ideas*. Se utilizará únicamente la cartelera disponible en *Ideas* para dar difusión a las novedades del curso. Los alumnos podrán comunicarse con los docentes a través del servicio de mensajería provisto por la plataforma.

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 teórico-práctico que se tomará al final del curso. Este examen cuenta con 2 (dos) recuperatorios.
  Quienes obtengan una nota mayor o igual a 4 (cuatro) pero inferior a 6 (seis) sólo obtendrán la cursada.

b) Régimen convencional
Esta modalidad de aprobación no posee requisitos de entregas.
Al finalizar el curso se debe rendir un examen referido a los aspectos prácticos de la materia. Este examen cuenta con 2 (dos) recuperatorios. Quienes lo 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	Introducción. Proceso KDD. Minería de Datos. Aplicaciones relacionadas.
2	Metadatos. Representación gráfica de atributos.
3	Transformación de atributos
4	Análisis de atributos. Correlación.
5	Técnicas de Agrupamiento. Agrupamiento partitivo
6	Agrupamiento Jerárquico. Algoritmo probabilista EM
7	Arboles de Clasificación. Algoritmos ID3, C4.5
8	Arboles de Clasificación (cont). Random Forest
9	Análisis de modelos predictivos
10	Clasificador Naïve Bayes
11	Reglas de clasificación
12	Reglas de asociación.
13	Redes Neuronales. Perceptrón y multiperceptrón
14	Comparación de modelos predictivos y consultas para la 1ra. fecha de parcial
15	1ra. Fecha de parcial
16	Muestra de exámenes de la 1ra. Fecha. Consultas referidas al trabajo integrador y a la 2da. Fecha de parcial
17	2da. Fecha de parcial
18	Muestra de exámenes de la 2da. Fecha. Consultas referidas al trabajo integrador y a la 3era. Fecha de parcial
19	3ra. Fecha de parcial

Contacto de la cátedra (mail, sitio WEB, plataforma virtual de gestión de cursos):

La cátedra cuenta con una página web de acceso público en la siguiente dirección
http://weblidi.info.unlp.edu.ar/catedras/md_si/
Allí se indica la manera de contacto con la cátedra y la forma de acceder al material publicado en el entorno virtual de enseñanza y aprendizaje **IDEAS**

Dra. Laura Lanzarini