PROGRAMA DE CURSO
MINERÍA DE DATOS

A. Antecedentes generales del curso:

Departamento	Ciencias de la Computación					
Nombre del curso	Minería de Datos	Código	CC5205	Créditos	6	
Nombre del curso en inglés	*Data Mining*					
Horas semanales	Docencia	3	Auxiliares	--	Trabajo personal	7
Carácter del curso	Obligatorio	X		Electivo		
Requisitos	CC3001: Algoritmos y estructura de datos					

B. Propósito del curso:

Al finalizar este curso se espera que los/las estudiantes generen conocimiento a partir de diversos tipos de datos, aplicando el proceso de *"knowledge discovery in databases"*, con énfasis en las técnicas principales de Minería de Datos (limpieza de datos, clasificación, clustering, análisis de asociación, etc).

Los y las estudiantes tendrán oportunidad de identificar y seleccionar las técnicas básicas de análisis que, según los criterios presentados en el curso, mejor se apliquen al objetivo de generación de conocimiento, según el dominio del problema planteado. Para esto se analizarán casos que incorporen problemáticas y datos de áreas tan diversas como, por ejemplo, Astronomía, Economía, Medicina, Marketing, Redes Sociales, entre otras. Se espera, además, acercar a los y las estudiantes a la problemática del análisis de grandes volúmenes de datos y que puedan adquirir técnicas más avanzadas sobre estos temas.

Es un curso introductorio, donde los ejemplos son referenciales y los y las estudiantes para su trabajo semestral eligen un área de interés, al que se le aplica la metodología estándar de la minería de datos.

Las y los estudiantes identificarán el problema de sobreajuste en modelos de procesamiento de información que impiden un análisis objetivo de los datos. Finalmente, desarrollarán habilidades éticas de desempeño profesional en el ámbito de Ingeniería de datos.

El curso tributa a las siguientes competencias específicas (CE) y genéricas (CG):

CE4: Extraer información relevante, utilizando el proceso de descubrimiento de conocimiento de datos.

CG1: Comunicación académica y profesional

Comunicar en español de forma estratégica, clara y eficaz, tanto en modalidad oral como escrita, puntos de vista, propuestas de proyectos y resultados de investigación fundamentados, en situaciones de comunicación compleja, en ambientes sociales, académicos y profesionales.

CG2: Comunicación en inglés

Leer y escuchar de manera comprensiva en inglés variados tipos de textos e informaciones sobre temas concretos o abstractos, comunicando experiencias y opiniones, adecuándose a diferentes contextos de acuerdo a las características de la audiencia.

CG3: Compromiso ético

Actuar de manera responsable y honesta, dando cuenta en forma crítica de sus propias acciones y sus consecuencias, en el marco del respeto hacia la dignidad de las personas y el cuidado del medio social, cultural y natural.

CG4: Trabajo en equipo

Ejecutar con su equipo, de forma estratégica, diversas actividades formativas propuestas, considerando la autogestión de sí mismo y la relación con el otro, asumiendo diversos roles: de líder, colaborador u otros, según requerimientos y objetivos, sin discriminar por género u otra razón.

C. Resultados de aprendizaje:

Competencias específicas	Resultados de aprendizaje
CE4	RA1: Formula y valida hipótesis de generación de conocimiento de acuerdo al dominio o área de los problemas de procesamiento de información, aplicando un conjunto de técnicas básicas de análisis de minería de datos, así como modelos, algoritmos y librerías de programación.
	RA2: Interpreta los resultados obtenidos a partir de las hipótesis iniciales, distinguiendo patrones novedosos y significativos de información en el contexto de la minería de datos.
	RA3: Detecta cuando se genera un problema de sobreajuste en los modelos de procesamiento de información con el cual se puede generar un razonamiento erróneo sobre los datos, a fin de aplicar criterios para evitar dicho tipo de problema.

	RA4: Resuelve un problema de minería de datos sobre un área de interés, considerando el planteamiento de hipótesis iniciales con su reformulación, el uso de *knowledge discovery in databases* y metodologías de análisis estudiadas.
Competencias genéricas	Resultados de aprendizaje
CG1	RA5: Reporta, en forma oral y escrita, resultados del análisis de minería de datos, la ejecución de un proyecto y sus avances, considerando hipótesis iniciales con su reformulación si corresponde, el uso de *knowledge discovery in databases* y metodologías de análisis.
CG2	RA6: Lee en inglés, de manera analítica y comprensiva, textos y artículos sobre la minería de datos, su desarrollo, evolución y aplicaciones.
CG3, CG4	RA7: Ejecuta de forma colaborativa, responsable y organizada las actividades comprometidas con sus pares, propiciando un ambiente de respeto por el otro, a través de la búsqueda de acuerdos para cumplir con las tareas encomendadas.

D. Unidades temáticas:

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
1	RA1, RA2, RA3, RA4, RA5	Introducción a la minería de datos	3 semanas
Contenidos		Indicador de logro	
1.1. Introducción a la minería de datos. 1.2. El proceso KDD (*knowledge discovery in Databases).* 1.3. Fuentes y tipos de datos (categóricos, numéricos). 1.4. Preprocesamiento y limpieza de datos. 1.5. Introducción a las técnicas principales de MD y sus aplicaciones (a. supervisado, b. no-supervisado, Reglas de Asociación). 1.6. Probabilidades y estadística: variables aleatorias, teorema central del límite, estimación puntual, por intervalo y test de hipótesis.		La/el estudiante: 1. Describe la minería de datos, en cuanto a su función, pasos metodológicos y principales técnicas. 2. Identifica los pasos del proceso de extracción de conocimiento KDD desde diversas bases de datos. 3. Clasifica variables en categóricas, numéricas, entre otrass, según fuentes y tipos de dato a procesar. 4. Utiliza, a nivel básico y con ejemplos, técnicas con las que limpia y pre procesa datos. 5. Identifica y resuelve problemas simples de minería de datos, utilizando técnicas estadísticas básicas de análisis de datos. 6. Define de manera inicial el proyecto de minería de datos, considerando los datos elegidos, hipótesis inicial más pregunta de investigación e información preliminar.	

	7. Expone brevemente, un avance de proyecto, considerando los datos elegidos, hipótesis inicial más pregunta de investigación e información preliminar sobre el análisis a desarrollar. 8. Reporta, de manera concisa, avances del proyecto, informando sobre el repositorio, el código fuente y un resumen de las tareas y pasos desarrollados.
Bibliografía de la unidad	[1] Cap.1, Cap.2, Cap. 3. [3] Cap. 1, Cap. 2. [4] Cap. 1., Cap. 4.7.

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
2	RA1, RA2, RA3, RA4, RA5, RA6	Aprendizaje supervisado	4 semanas
Contenidos		Indicador de logro	
2.1. Metodologías clásicas de clasificación (árboles de decisión, SVM, Naive Bayes, KNN, otros). 2.2. Evaluación del desempeño de clasificadores (bootstrap, crossvalidation). 2.3. Métricas de evaluación (accuracy, precision, recall, F1, ROC). 2.4. Comparación de modelos y clasificadores.		La/el estudiante: 1. Define qué es el aprendizaje supervisado, de acuerdo a su función y propiedades. 2. Utiliza metodologías clásicas (árboles de decisión, SVM, *Naive Bayes,* KNN, otros) para la clasificación de un conjunto de datos. 3. Interpreta los resultados obtenidos de diferentes clasificadores, a partir de ejemplos. 4. Utiliza librerías de programación en el contexto del aprendizaje supervisado. 5. Trabaja de manera organizada en las actividades comprometidas por el equipo 6. Trabaja en su proyecto de minería de datos, definiendo resultados preliminares y ajustes considerados necesarios. 7. Reporta, de manera concisa, avances del proyecto, informando, en forma oral y escrita, los resultados del proceso de minería de datos inicial, un resumen de las tareas y pasos desarrollados. 8. Lee de manera analítica diversos textos y artículos en inglés sobre minería de datos.	
Bibliografía de la unidad		[1] Cap. 4, Cap. 5 [3] Cap. 3., Cap. 4, Cap. 5, Cap. 6.	

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
3	RA1, RA2, RA3, RA4, RA5, RA6	Aprendizaje no supervisado	4 semanas

Contenidos	Indicador de logro
3.1. Introducción y conceptos básicos de clustering. 3.2. Algoritmos clásicos de *clustering* (k-means, *clustering* aglomerativo jerárquico, DBSCAN, otros métodos). 3.3. Elección de métodos de clustering. 3.4. Evaluación de *clusters*.	La/el estudiante: 1. Aplica algoritmos de *clustering* sobre conjuntos de datos, interpretando los resultados obtenidos. 2. Resuelve problemas de aprendizaje no supervisado, utilizando librerías de programación. 3. Trabaja en una actividad de laboratorio sobre el aprendizaje no supervisado, considerando las técnicas de clustering usadas y los datos obtenidos. 4. Elabora, con su equipo, las tareas de forma colaborativa, responsable y organizada. 5. Reporta de manera concisa los resultados sobre el aprendizaje no supervisado, las técnicas de *clustering* usadas y los datos obtenidos. 6. Lee de manera comprensiva diversos textos y artículos en inglés sobre minería de datos, determinando sus ideas centrales.
Bibliografía de la unidad	[1] Cap. 8, Cap. 9 [3] Cap. 3.6, Cap. 4.8, Cap. 6.8

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
4	RA1, RA2, RA3, RA4, RA5, RA6, RA7, RA8	Análisis de asociación	2 semanas

Contenidos	Indicador de logro
4.1. Generación de conjuntos de elementos frecuentes. 4.2. Reglas de asociación. 4.3. Algoritmo de regla de asociación: Apriori. 4.4. Métricas en reglas de asociación (soporte, confianza, lift).	La/el estudiante: 1. Determina las propiedades del análisis de asociación, considerando sus ventajas. 2. Utiliza la minería de reglas de asociación sobre conjuntos de datos. 3. Interpreta los resultados obtenidos a partir de diferentes reglas de asociación. 4. Utiliza librerías de datos y librerías de programación en la resolución de problemas de análisis de asociación.

	5. Selecciona algoritmos, según el tipo de problema en el contexto de la minería de datos. 6. Trabaja, de manera colaborativa y organizada en las actividades del equipo. 7. Reporta en textos concisos los resultados de los diferentes análisis de asociación sobre un conjunto de datos 8. Lee de manera comprensiva y analítica textos y artículos en inglés sobre minería de datos.
Bibliografía de la unidad	[1] Cap. 6, Cap. 7 [3] Cap 3.4, Cap. 4.5

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
5	RA1, RA2, RA3, RA4, RA5	Tópicos adicionales de minería de datos	2 semanas
Contenidos		Indicador de logro	
5.1. Modelos lineales para regresión y clasificación. 5.2. Introducción a las redes neuronales. 5.3. Privacidad de datos. 5.3.1. Anonimización: k-anonymity. 5.3.2. Differential Privacy. 5.4. Selección de atributos. 5.5. Método de filtro. 5.6. Método wrapper. 5.7. Reducción de Atributos. 5.8. PCA. 5.9. Multidimensional Scaling.		La/el estudiante: 1. Utiliza modelos lineales en ejemplos de regresión y clasificación, considerando las limitaciones de estos modelos. 2. Describe el funcionamiento de redes neuronales para regresión y clasificación. 3. Analiza ejemplos sobre el tratamiento de la privacidad de la información, considerando tipos de problemas asociados en el contexto de la minería de datos. 4. Selecciona métodos y enfoques de minería de datos con los cuales preservar la privacidad de las personas. 5. Analiza el problema de atributos irrelevantes o redundantes en el aprendizaje supervisado. 6. Resuelve problemas en minería de datos, utilizando librerías de programación. 7. Aplica a nivel básico técnicas para la selección y reducción de atributos en ejemplos. 8. Reporta en forma oral y escrita los resultados finales del proyecto, informando, de manera concisa, sobre su ejecución, las metodologías de análisis usadas, entre otros.	
Bibliografía de la unidad		[1] Cap. 5, Apéndice… [3] Cap 3.2, 4.6, Cap. 7.1, Cap 7.3.	

E. Estrategias de enseñanza -aprendizaje:

El curso considera las siguientes estrategias de enseñanza:

- **Clases expositivas**, en donde el estudiante identifica los problemas fundamentales en minería de datos, así como modelos y técnicas para abordarlos.
- **Casos de estudio,** en donde se discuten aplicaciones reales (problemas bien definidos) de los métodos enseñados. Discusión en clases de ejemplos (1 problema para las unidades 2, 3, 4).
- **Laboratorios:** en las respectivas unidades, el estudiante es expuesto a librerías de programación que permiten implementar modelos de solución a problemas en minería de datos. Se trabaja sobre la base de aprendizaje activo donde se revisan los conceptos esenciales de la minería de datos.
- **Trabajo en base a un proyecto grupal a realizar en clases:** se trabajará en el desarrollo de un proyecto real de minería de datos durante el semestre. Se crearán instancias de trabajo en clases donde los estudiantes recibirán apoyo del equipo docente.

F. Estrategias de evaluación:

Al inicio del semestre el equipo docente informará el tipo de evaluación a realizar, la cantidad, así como las ponderaciones correspondientes.

Para esta propuesta, el curso considera las siguientes instancias de evaluación:

- **Laboratorios:**
  - »Tres (3) laboratorios prácticos acotados (divididos en dos clases), a resolverse en grupos de 2 personas y que requieren manejo teórico de la minería de datos. Se acompaña el trabajo del/la estudiante.
    Se eliminará el laboratorio de peor nota para los estudiantes que entreguen todos sus trabajos. Se evalúan los resultados de aprendizaje RA1, RA2, RA3 y RA5.
- **Proyecto grupal:**
  - »Se evalúa en 3 hitos (o etapas):
    - Hito 1: se presentan los datos elegidos, hipótesis iniciales y posibles técnicas (mide la construcción y validación de hipótesis). Entregables: presentación + informe con codigo utilizado y principales observaciones.
    - Hito 2, se presentan los resultados del proceso de minería de datos inicial (mide interpretación + reformulación). Entregables: presentación + codigo de avance con principales observaciones).
    - Hito 3, se presentan los resultados finales. Entregables: presentación + informe final.

- El informe es incremental y se construye a partir de reportes parciales. Cada entrega es retroalimentada.
- Exposiciones orales breves de 7 minutos en las unidades correspondientes y recibe la retroalimentación respectiva por parte de los académicos.

El proyecto evalúa los resultados de aprendizaje RA1, RA2, RA3, RA4, RA5, RA6, RA7, RA8.

G. Recursos bibliográficos:

Bibliografía obligatoria:

1. Tan, P., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining.* Addison-Wesley.
2. Samet, H. (2006). *Foundations of Multidimensional and Metric Data Structures.* Morgan Kaufmann.
3. *Data Mining: Practical Machine Learning Tools and Techniques*, Third Edition (Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten, Eibe Frank and Mark Hall.
4. Zezula, P., Amato, G., Dohnal, V., Batko, M. (2006). *Similarity Search: The Metric Approach. Springer.*

H. Datos generales sobre elaboración y vigencia del programa de curso:

Vigencia desde:	Otoño, 2021
Elaborado por:	Bárbara Poblete y Felipe Bravo
Validado por:	Enviado a revisión y validación académico par: Jorge Pérez. Validado por CTD Departamento de Ciencias de la Computación
Revisado por:	Área de Gestión Curricular