Análisis de grandes volúmenes de datos

1. Características generales

Nombre:	Análisis de grandes volúmenes de datos
Sigla:	CI-0163
Créditos:	4
Horas lectivas:	5 horas de teoría
Requisitos:	CI-0127 Bases de datos, CI-0128 Proyecto Integrador de Ingeniería y Bases de Datos
Correquisitos:	Ninguno
Clasificación:	Curso propio
Ciclo de carrera:	I o II ciclo, 4to año
Docente(s):	Allan Berrocal Rojas
Datos de contacto:	Correo: allan.berrocal@ucr.ac.cr
Grupo:	01
Semestre y año:	I ciclo 2022
Horario de clases:	K de 13:00 a 15:50 aula 205 V de 13:00 a 14:50 virtual en este Enlace de Zoom Aula en Mediación Virtual CI0163-22a Space en Matrix
Horario de consulta:	L de 11 - 13 y J 13 – 15 (previa cita) Enlace de Zoom para consulta personal
Asistente:	Por definir
Modalidad:	Bimodal (50/50), K presencial, V virtual Exámenes presenciales

2. Descripción

Las capacidades de generación y recopilación de los datos han aumentado rápidamente debido a varios factores: automatización de los negocios, diversificación de las transacciones en medios electrónicos, uso de dispositivos electrónicos especializados en recopilación de datos (sensores), crecimiento del uso de redes sociales y aumento de las conexiones entre dispositivos (internet de las cosas), entre otros. Este crecimiento de datos genera una necesidad de contar con las técnicas y herramientas automatizadas que permiten la transformación de grandes volúmenes de datos en la información o

conocimiento útil para la toma de decisiones con el objetivo de mejorar la situación de los negocios u organizaciones/instituciones. Este análisis permite descubrir patrones interesantes (no triviales, implícitos, previamente desconocidos y potencialmente útiles) que un ser humano no es capaz de encontrar. La aplicación de técnicas de análisis de grandes volúmenes de datos requiere un aprendizaje sobre la preparación de datos, la capacidad de poder seleccionar una técnica adecuada al problema a solucionar y a la interpretación de resultados.

3. Objetivos

Objetivo general

El objetivo *general* del curso es que cada estudiante desarrolle las habilidades necesarias para el pre-procesamiento de datos y el uso de técnicas adecuadas de análisis, con el fin de descubrir el conocimiento en grandes volúmenes de datos, mediante estrategias que integren lo teórico y lo práctico, incluyendo un fuerte componente de actividades en el laboratorio.

Objetivos específicos

Durante este curso, cada estudiante desarrollará habilidades para:

1. Identificar las necesidades de aplicar las técnicas automatizadas de descubrimiento de conocimientos en grandes volúmenes de datos, con el fin de encontrar los patrones que permiten la transformación de los datos en información o conocimiento útil para la toma de decisiones, a través de estrategias declarativas.
2. Pre-procesar los datos, incluyendo su limpieza, transformación, integración y reducción, para asegurar descubrimiento de conocimiento veraz, a través de estrategias declarativas y prácticas.
3. Utilizar y contrastar diferentes técnicas de análisis de grandes volúmenes de datos para seleccionar la técnica más apropiada al problema y tipo de datos en cuestión, mediante el uso práctico de estas técnicas.
4. Interpretar y evaluar los resultados obtenidos al aplicar las técnicas automatizadas para asegurar el descubrimiento de patrones no triviales, implícitos, previamente desconocidos y potencialmente útiles.
5. Ampliar los conocimientos a los métodos o las técnicas novedosas usadas para distintos conjuntos de datos, para enfrentar los cambios continuos en el manejo y análisis de datos.

4. Contenidos

Objetivo específico	Eje temático	Desglose
1	Conceptos introductorios de grandes volúmenes de datos y diferentes métodos de análisis	Grandes volúmenes de datos: sus características y diferentes formas de almacenamiento. Métodos tradicionales de análisis de datos, minería de datos, aprendizaje automático, métodos emergentes.
2	Pre-procesamiento de datos	Diferentes tipos de atributos y conjuntos de datos. Exploración de datos. Limpieza, transformación e integración de los datos. Reducción de la dimensionalidad de datos. Muestreo.
3, 4	Clasificación y predicción	Conceptos básicos, árboles de decisión, redes neuronales, algoritmos genéticos, máquinas de soporte vectorial. Medidas de exactitud y error. Criterios para selección del modelo.
3, 4	Asociación	Análisis de canasta básica, algoritmo de Apriori. Métricas para evaluar las reglas de asociación.
3, 4	Segmentación	Medición de distancia entre datos de diferentes tipos. Método k-means, métodos jerárquicos (de aglomeración), métodos basados en densidad, método SOM (*self-organizing maps*). Evaluación de segmentos.
5	Nuevas tendencias en análisis de grandes volúmenes de datos	Análisis de flujo de datos: El problema de muestreo, ventanas deslizantes, el filtro Bloom para seleccionar flujos de interés, algoritmos de clasificación y segmentación de flujo de datos. Minería de patrones frecuentes en flujo de datos
		Análisis de grafos: Segmentación de grafos, métodos de particionamiento. Minería de grafos aplicada a redes sociales.
		Otros: análisis de datos de serie de tiempo, secuenciales, espaciales, entre otros.

5. Metodología

Se propone aplicar una metodología híbrida que combina clases magistrales con lecturas extra-clase, discusiones y ejercicios prácticos. Las lecciones pueden intercalarse entre lecciones magistrales sincrónicas presenciales, a distancia y aula invertida.

En las lecciones magistrales, tanto presenciales como a través de la plataforma Zoom, se utilizan recursos audiovisuales para ilustrar los conceptos que se estudien.

También se realizarán trabajos grupales que incluyan participación y discusión activa de los estudiantes.

En el caso de aula invertida, los estudiantes realizan el estudio independiente de contenidos en materiales provistos en el aula virtual, y las lecciones se dedicarán al acompañamiento de estudiantes en el desarrollo de ejercicios y actividades relacionados a dichos contenidos. Durante las horas extra clase los estudiantes estudiarán el material del curso y resolverán los ejercicios planteados. Las soluciones a los ejercicios serán presentadas según se indique en los enunciados incluyendo la discusión en clase.

La consulta será mediante videoconferencia a través de Zoom. Se requiere cita previa.

6. Evaluación

%	Rubro	Descripción
20 %	Prácticas	Serán asignadas en Mediación Virtual y serán evaluadas mediante un quiz o reportes escritos según lo indique el profesor. Todas las prácticas tendrán el mismo peso para la nota, excepto aquellas para las cuales se especifique algo distinto en el enunciado.
15 %	Tareas cortas	Serán asignadas en Mediación Virtual y serán evaluadas según especifique el profesor en cada caso (por ejemplo, mediante una presentación corta, discusión en clase, reunión con el profesor o reunión con el asistente). Todas las prácticas tendrán el mismo peso para la nota, excepto aquellas para las cuales se especifique algo distinto en el enunciado.
10 %	Quices	Se podrán realizar en cualquiera de las clases y al momento de la clase que el profesor lo considere apropiado. Estos se realizarán en Mediación Virtual. Los quices serán de igual ponderación y la duración de cada quiz será especificada en su enunciado.
10 %	Reportes	Consiste en presentar reportes durante el semestre resumiendo y discutiendo sobre los temas estudiados en el curso. Los reportes serán evaluados por su calidad, completitud y profesionalismo.
15 %	Investigación de un tema	Consiste en investigar un tema relevante en el contexto del curso y compartir los resultados mediante una presentación en clase. La evaluación consiste en una presentación en clase y una discusión en un foro en MV sobre el tema presentado por cada equipo.
30 %	Proyecto de investigación aplicado	Consiste en proponer un problema en el contexto de análisis de datos, realizar investigación respecto del problema, proponer una solución e implementarla documentando los resultados para finalmente presentarlos en la clase. En la solución se deben utilizar los conceptos vistos en clase e incorporar nuevas herramientas que faciliten el

		desarrollo de la solución propuesta al problema. La evaluación consiste en una presentación en clase.

Lineamientos:

1. Toda asignación se considera tardía si se entrega después de las 6 a.m. posterior a la fecha de entrega. Si la entrega tiene 24 o menos horas de retraso se le aplicará una penalización de 50% del valor. Después de 24 horas de retraso, la asignación será calificada con cero.
2. Todas las evaluaciones deberán ser entregadas al profesor el día propuesto en el enunciado, por los medios indicados en su enunciado.
3. Los quices se realizan durante las lecciones y en cualquier momento durante el transcurso de la lección, y solo se repondrán en los casos que establece el Reglamento de Régimen Académico Estudiantil en su Articulo 24. Se recomienda a los estudiantes contar con medios adicionales que garanticen el acceso a Mediación Virtual en caso de anomalías como fallos eléctricos o de telefonía fija, como disponer de la aplicación Moodle o un navegador para dispositivos móviles pre-configurados.
4. Todas las evaluaciones son estrictamente individuales excepto aquellas en cuyo enunciado se especifique algo distinto.
5. Todo trabajo debe ser entregado de forma digital.
6. En todos los trabajos y las evaluaciones de los estudiantes, se calificará la redacción, ortografía, estructura y contenido.
7. En toda asignación se evaluarán las buenas prácticas de programación, como identificadores significativos, indentación, apego a una convención de estilo, documentación de interfaces e implementaciones de subrutinas, y reutilización de código. Serán castigadas malas prácticas de programación como redundancia de código, accesos inválidos de memoria, condiciones de carrera, espera activa, y otras prácticas estipuladas durante las lecciones del curso.
8. Es ilegal presentar como propio, material parcial o totalmente creado por otras personas u obtenido de fuentes de información, como por ejemplo de libros o de Internet, sin la debida referencia al autor de la propiedad intelectual. En cualquier asignación en que se sospeche de plagio se aplicará el debido proceso estipulado en el Reglamento de Orden y Disciplina de los Estudiantes de la Universidad de Costa Rica.

7. Cronograma

La tabla de abajo muestra las fechas tentativas de cobertura de los contenidos, que pueden ser reajustadas de acuerdo con el avance durante el ciclo lectivo. Las fechas

de entrega de los ejercicios y demás asignaciones se comunicarán oportunamente durante clases.

Semanas	Temas
1	Conceptos introductorios de grandes volúmenes de datos y diferentes métodos de análisis
2	Pre-procesamiento de datos
4	Clasificación y predicción
2	Segmentación
2	Asociación
4	Nuevas tendencias en análisis de grandes volúmenes de datos

8. Bibliografía

1. Berry M. y Linoff G. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management, 3a edición. Wiley Publishing, 2011.
2. Han J. y Kamber M. Data Mining: Concepts and Techniques, 3a edición. Morgan Kaufman Publishers, 2011.
3. Larose D. y Larose Ch. Data Mining and Predictive Analytics, 2a edición. Wiley Publishing, 2016.
4. Leskovec J., Rajaraman A. y Ullman J. Mining of Massive Datasets. Standford University, 2014.
5. Loshin D. Big Data Analytics: from Strategic Planning to Entreprise Integration with Tools, Techniques, NoSQL, and Graphs. Mogran Kaufmann, 2013.
6. Tan P.N., Steinbach M. y Karpatne A. Introduction to Data Mining, 2a edición. Pearson, 2018.
7. Vercellis C. Business Intelligence: Data Mining and Optimization for Decision Making. Wiley Publishing, 2009.
8. Witten I.H., Frank W., Hall M. y Pal Ch. Data Mining: Practical Machine Learning Tools and Techniques, 4a edición. Morgan Kaufmann Publishers, 2016.
9. Andreas C. Müller and Sarah Guido. An Introduction to Machine Learning with Python. O'Reilly, primera edición, 2016.

9. Recursos estudiantiles

Para información sobre recursos estudiantiles disponibles en la UCR, incluyendo el Sistema de bibliotecas y la normativa universitaria vigente, favor visitar la página: https://www.ecci.ucr.ac.cr/vida-estudiantil/servicios-institucionales-para-estudiantes/guia-de-recursos-estudiantiles-de-la-ucr.