PRONTUARIO

TÍTULO DEL CURSO	**:**	Introducción a la Ciencia de Datos
CODIFICACIÓN	**:**	CDAT 4001
CANTIDAD DE HORAS/CRÉDITO	**:**	45 horas por semestre Tres horas contacto semanal[1] / Tres créditos
PRERREQUISITOS	:	COMP 3081 (Introducción a la Programación y Ciencia de Cómputos I), COMP 3083 (Laboratorio de Introducción a la Programación y Ciencia de Cómputos I) o autorización de Directora Departamental
CORREQUISITOS	:	Se sugiere tener familiaridad con SQL y con álgebra lineal es conveniente, pero no imprescindible.

DESCRIPCIÓN DEL CURSO:

La ciencia de datos es una disciplina emergente de carácter multidisciplinario que deriva sus herramientas principales de la intersección de la matemática, la estadística y la ciencia de cómputos. Su objeto de estudio es la extracción sistemática de conocimiento partiendo de datos. En este curso el o la estudiante conocerá los fundamentos de la disciplina y ensayará los distintos pasos del proceso de la ciencia de datos: colección, limpieza y procesamiento de datos a gran escala, análisis exploratorio, modelado, visualización y la generación de un producto de datos. El énfasis es brindar una visión panorámica de la disciplina y cómo se integran sus partes en la práctica. La profundización en algunos aspectos, en particular, en la colección y la preparación de datos y el aprendizaje automático, se aborda en otros cursos.

Este curso se ofrecerá bajo la modalidad presencial.

[1] Una hora contacto equivale a cincuenta (50) minutos.

OBJETIVOS DE APRENDIZAJE:

Al terminar la discusión de cada sección del bosquejo de contenido del curso el o la estudiante podrá:

1) Introducción

   a) Discutir qué es Ciencia de Datos como campo emergente dentro de la informática.
   b) Mencionar áreas de aplicación de la Ciencia de Datos en la informática, las ciencias (naturales, comerciales y sociales), así como en las redes sociales.
   c) Identificar fuentes de datos para análisis.
   d) Describir el proceso de ciencia de datos identificando sus componentes y cómo interaccionan.

2) Datos a gran escala (*Big Data*)

   a) Identificar retos que presenta el manejo de datos a gran escala en cuanto a su volumen, velocidad y veracidad.
   b) Comparar diversas plataformas de las cuales se pueden obtener datos a gran escala.
   c) Utilizar al menos un lenguaje de programación dotado de bibliotecas numéricas, estadísticas y redes neuronales como *Python* o *R*.
   d) Utilizar *Hadoop* para sistemas de archivos distribuidos y cómputos con el paradigma *Map- Reduce* y *Spark*.
   e) Utilizar sistemas de almacenamiento de datos distribuidos (como *Cassandra*) no- estructurados y no-*SQL* para almacenamiento y análisis de datos.
   f) Utilizar *API*s de herramientas para extraer datos de la web (*Data Wrangling*).

3) Análisis exploratorio de datos (AED)

   a) Enumerar los objetivos del análisis exploratorio de datos.
   b) Generar e interpretar gráficos asociados al AED como diagrama de caja, histograma, gráfica dependiente de tiempo, diagrama de Pareto, diagrama de dispersión, diagrama de tallos y hojas, coordenadas paralelas, gráficas de estrella, entre otros.
   c) Utilizar estadística descriptiva (media, mediana, varianza) y relacionarlas con las técnicas gráficas.

4) Extracción de información de datos (*data mining*)

   a) Definir características (variables) partiendo de datos crudos sin estructurar
   b) Enumerar cómo se pueden utilizar filtros, árboles de decisión, bosques aleatorios en la extracción de información a partir de datos.

5) Visualización

a) Utilizar técnicas para visualizar patrones espaciales en un conjunto de datos.
b) Utilizar técnicas para visualizar patrones en un conjunto de datos que fluctúan en el tiempo.
c) Reconocer los principios que determinan la efectividad de una herramienta de visualización versus otra.
d) Utilizar técnicas para realizar comparaciones e identificar diferencias entre conjuntos de datos.
e) Utilizar técnicas para visualizar relaciones entre muchas variables.
f) Utilizar técnicas para reducir la cantidad de variables en un conjunto de datos.
g) Detectar los valores atípicos en un conjunto de datos.

6) Estudios de casos

a) Definir sistema de recomendación y proveer ejemplos notables.
b) Distinguir entre adquisición implícita o explícita de datos.
c) Aplicar estrategias de filtrado de datos para generar recomendaciones.
d) Representar con grafos las relaciones entre individuos.
e) Aplicar técnicas de análisis de grafos para determinar conjuntos de individuos de interés común.

7) Gobernanza

a) Citar debates actuales sobre asuntos sociales, éticos, legales y políticos relacionados con el uso de analítica de datos a gran escala.
b) Definir ética de datos.
c) Enunciar ejemplos de leyes, reglamentos y estándares aplicables al manejo de datos.

LIBRO DE TEXTO PRINCIPAL: NO TIENE

BOSQUEJO DE CONTENIDO Y DISTRIBUCIÓN DEL TIEMPO:

Tema	Distribución del tiempo		
	Presencial	**Híbrida**	**En línea**
Tema 1: Introducción 1. La ciencia de datos como disciplina emergente 2. Fuentes de datos 3. El proceso de ciencia de datos 4. Repaso: herramientas de programación básicas, matrices, tablas y limpieza de datos.	3 horas	No aplica	No aplica

Tema 2: Datos a gran escala (Big Data) 1. Definiciones y conceptos 2. Plataformas para el almacenamiento de datos y cómputos a gran escala 3. Manejo de datos por *SQL* y *no SQL* 4. *API*s de herramientas para extraer datos de la web	6 horas	No aplica	No aplica
Tema 3: Análisis exploratorio de datos 1. Definición y propósitos 2. Métodos visuales 3. Medidas estadísticas básicas	6 horas	No aplica	No aplica
Tema 4: Extracción de información de datos (data mining) 1. Definición de características (variables) partiendo de datos crudos sin estructurar 2. Selección de características: proceso y algoritmos (filtros, árboles de decisión, bosques aleatorios)	6 horas	No aplica	No aplica
Tema 5: Visualización 1. De patrones espaciales a) Sistemas de coordenadas b) Visualización de datos en mapas geográficos 2. De patrones en el tiempo a) Selección de escala b) Normalización c) Identificación de tendencias d) Identificación de estacionalidad e) Identificación de periodicidad f) Remoción de ruido 3. De relaciones entre dos variables a) Concepto de correlación b) La correlación como herramienta predictiva 4. De muchas variables a) Técnicas de visualización de datos de muchas variables b) Utilización de técnicas de reducción de dimensiones c) Análisis de componentes principales d) Identificación de grupos y valores atípicos	12 horas	No aplica	No aplica

Tema 6: Estudio de caso: Sistemas de Recomendación básicos 1. Objetivos 2. Colección de datos implícita y explícita 3. Filtrado colaborativo 4. Filtrado basado en contenido 5. Filtrado basado en conocimiento 6. Recomendaciones basadas en demografía	3 horas	No aplica	No aplica
Tema 7: Estudio de caso: Redes sociales 1. Representación con grafos 2. Agrupamiento 3. Descubrimiento de comunidades 4. Particionamiento 5. Vecindades en grafos	3 horas	No aplica	No aplica
Tema 8: Gobernanza 1. Aspectos sociales, éticos, legales y políticos 2. Ética de datos 3. Algunas leyes, reglamentos y estándares aplicables al manejo de datos	3 horas	No aplica	No aplica
La suma de las horas sugeridas es de 42. Las tres horas restantes serán utilizadas para la ejecución de las **estrategias instruccionales** afines con el curso y **evaluaciones** (ver secciones más adelante). Los tópicos en este bosquejo de contenido son aptos para ser ordenados de otras maneras a juicio del docente que imparta el curso en el ejercicio de su libertad de cátedra.			
Total de horas contacto	**45 horas**	No aplica	No aplica

ESTRATEGIAS INSTRUCCIONALES: Con miras a lograr los objetivos del curso el profesor o la profesora seleccionará entre las siguientes técnicas instruccionales:

Presencial	Híbrido	En línea
• introducción breve de los temas • demostraciones • trabajo de programación en el laboratorio (individual o en grupo)	No aplica	No aplica

RECURSOS MÍNIMOS DISPONIBLES O REQUERIDOS:

Recurso	Presencial	Híbrido	En línea
La Universidad debe proveer un laboratorio para trabajo independiente de los estudiantes, el equipo electrónico que necesita el profesor o la profesora para impartir la clase, el programado apropiado para el curso (intérpretes y compiladores de todos los lenguajes que se utilizarán, herramientas para el desarrollo de aplicaciones) y acceso a la Internet. Se sugiere al estudiante poseer una computadora portátil con capacidad de ejecutar los compiladores, intérpretes u otros programados utilizados en este curso.	Institución	No aplica	No aplica

TÉCNICAS DE EVALUACIÓN:

Presencial	Híbrida	En línea
Las evaluaciones consisten de al menos tres exámenes, trabajos de programación, presentaciones por estudiantes, y un proyecto final. Los pesos relativos de estas actividades en la nota final deberán ser discutidos y acordados el primer día de clases. **Total 100%**	No aplica	No aplica

ACOMODO RAZONABLE:

MODIFICACIÓN RAZONABLE (Acomodo razonable)

La Universidad de Puerto Rico (UPR) reconoce el derecho que tienen los estudiantes con impedimentos a una educación post secundaria inclusiva, equitativa y comparable. Conforme a su política hacia los estudiantes con impedimentos, fundamentada en la legislación federal y estatal, todo estudiante cualificado con impedimentos, tiene derecho a la igual participación de aquellos servicios, programas y actividades que están disponibles de naturaleza física, mental o sensorial y que por ello se ha afectado, sustancialmente, una o más actividades principales de la vida como lo es su área de estudios post secundarios, tiene derecho a recibir acomodos o modificaciones razonables. De usted requerir acomodo o modificación razonable en este curso, debe notificarlo al profesor sobre el mismo, sin necesidad de divulgar su condición o diagnóstico. De manera simultánea, debe solicitar a la Oficina de Servicios a Estudiantes con Impedimentos (OSEI) de la unidad o Recinto, en forma expedita, su necesidad de modificación o acomodo razonable.

INTEGRIDAD ACADÉMICA:

La Universidad de Puerto Rico promueve los más altos estándares de integridad académica y científica. El Artículo 6.2 del Reglamento General de Estudiantes de la UPR (Certificación Núm. 13, 2009-2010, de la Junta de Síndicos) establece que "la deshonestidad académica incluye, pero no se limita a: acciones fraudulentas, la obtención de notas o grados académicos valiéndose de falsas o fraudulentas simulaciones, copiar total o parcialmente la labor académica de otra persona, plagiar total o parcialmente el trabajo de otra persona, copiar total o parcialmente las respuestas de otra persona a las preguntas de un examen, haciendo o consiguiendo que otro tome en su nombre cualquier prueba o examen oral o escrito, así como la ayuda o facilitación para que otra persona incurra en la referida conducta". Cualquiera de estas acciones estará sujeta a sanciones disciplinarias en conformidad con el procedimiento disciplinario establecido en el Reglamento General de Estudiantes de la UPR vigente.

NORMATIVA SOBRE HOSTIGAMIENTO SEXUAL

«La Universidad de Puerto Rico prohíbe el discrimen por razón de sexo y género en todas sus modalidades, incluyendo el hostigamiento sexual. Según la Política Institucional contra Hostigamiento Sexual, Certificación 130 (2014-15) de la Junta de Gobierno, si un estudiante es o está siendo afectado por conductas relacionadas a hostigamiento sexual, puede acudir a la Oficina de la Procuraduría Estudiantil, el Decanato de Estudiantes o la Coordinadora de Cumplimiento con Titulo IX para orientación o para presentar una queja».

«The University of Puerto Rico prohibits discrimination based on sex, sexual orientation, and gender identity in any of its forms, including that of sexual harassment. According to the Institutional Policy Against Sexual Harassment at the University of Puerto Rico, Certification 130 (2014-2015) from the Board of Governors, any student subjected to acts constituting sexual harassment, may turn to the Office of the Student Ombudsperson, the Office of the Dean of Students, or the Coordinator of the Office of Compliance with Title IX for an orientation or formal complaint».

SISTEMA DE CALIFICACIÓN:

Se adjudicará la calificación A, B, C, D, o F, según el nivel de competencia demostrado en las evaluaciones. El profesor o profesora podrá usar la escala 100-85 A, 84-75 B, 74-60 C, 59-50 D, 49-0 F u otra que resulte más apropiada para asignar las calificaciones del curso y la informará, las primeras semanas de clases, y en la Guía del Estudiante.

BIBLIOGRAFÍA:

Referencias electrónicas:

ACM/IEEE-CS Task Group on Information Technology Curricula. (2013). *Information Technology Curricula 2017: Curriculum Guidelines for Baccalaureate Degree Programs in Information Technology*. Final report. ACM Press and IEEE Computer Society Press. ISBN 978-1-4503-6416-4, DOI: 10.1145/3173161

Blum, A., Hopcroft, J., and Kannan, R. (2016). *Foundations of Data Science*. https://www.cs.cornell.edu/jeh/book2016June9.pdf

Chatfield, C. (2013). *The Analysis of Time Series*. Chapman & Hall.

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing data science: Big data, machine learning, and more, using Python tools*. Manning Publications Co. ISBN-13: 978-1633430037

Dinu, J. (2019). *Foundations of Data Science: A Practical Introduction to Data Science with Python*. Addison-Wesley Data & Analytics Series. ISBN-13: 978-0134398808

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and Tensor Flow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc. ISBN-13:978-1491962299

Godsey, B. (2017). *Think Like a Data Scientist: Tackle the data science process step-by-step*. Manning Publications, Co. ISBN 9781633430273

Jannert, P. K. (2011) *Data Analysis with Open Source Tools*. O'Reilly Media, Inc.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press. ISBN 978-1-107-07723-2

Loukides, M. and Mason, H. (2018). *Ethics and Data Science*. O'Reilly Media, Inc. ISBN 978- 1-492-04388-1

Marz, N. & Warren, J. (2015). *Big Data: Principles and best practices of scalable real-time data systems*. New York; Manning Publications Co. ISBN-13: 978-1617290343

O'Neil, C. & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc. ISBN: 978-1-449-35865-5

Stanton, J. (2012) *Introduction to Data Science*. Syracuse University https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

Tufte, E. P. (2001). *The visual display of quantitative information*. Graphic Press LLC.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*.

O'Reilly Media, Inc. ISBN-13: 978-149191

Viktor, M. S. & Kenneth, C. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Yau, N. (2011). *Visualize this: The Flowing Data Guide to Design, Visualization, and Statistics*. John Wiley & Sons.

Yau, N. (2013). *Data points: visualization that means something*. John Wiley & Sons.

Zaki, M. J., Meira Jr. W. & Meira, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge University Press. ISBN-13: 978-0521766333

PLAN DE CONTINGENCIA EN CASO DE SURGIR UNA EMERGENCIA O INTERRUPCIÓN DE CLASES:

En caso de surgir una emergencia y haya alguna interrupción en las clases, el profesor o la profesora de este curso se comunicará con sus estudiantes para informar los acuerdos departamentales que se seguirán al respecto y poder dar continuidad a la sesión académica.

CRÉDITOS:

Redacción inicial por Dr. Elio Ramos Colón y Prof. José O. Sotero Esteva, 2 de enero de 2019. Revisado por Dr. Ollantay Medina Huamán, Dr. Elio Ramos Colón, Profa. Idalyn Ríos Díaz y Prof. José O. Sotero Esteva, 15 de enero de 2019.

Revisión por Comité de Currículo, abril 2019. Endosado por el Departamento de Matemáticas en Reunión Ordinaria, 12 abril de 2019. Revisión general por Profa. Bárbara L. Santiago-Figueroa, 31 de julio de 2019.

Algunos de los objetivos de aprendizaje han sido adaptados de las recomendaciones contenidas en el *Information Technology Currícula 2017* de las organizaciones profesionales *Association for Computing Machinery* y el *Institute of Electrical and Electronic Engineers – Computer Society* (Ver Bibliografía). Se reconoce la influencia del prontuario *Introduction to Data Science: CptS 483-06 – Syllabus, First Offering: Fall 2015* por Assefaw Gebremedhin de Washington State University.

Este bosquejo y los objetivos que le preceden cubren las siguientes unidades de conocimiento de las recomendaciones curriculares *Information Technology Curricula 2017* de las organizaciones profesionales *ACM* e *IEEE-Computing Society* según la siguiente tabla:

Dominio de conocimiento	**Horas cubiertas**
	Suplementario
ITS- DSA-(01, 02, 05) Data Scalability and Analytics	15

Rev. Marzo 2022/ Certificación 2021-2022-10/DM

Rev. Abril 2022 (Certificación Núm. 33, 2020-2021 JG-anejo 9)