

Unidad Académica Responsable: Departamento de Informática y Ciencias de la Computación

Programa: Magíster en Ciencias de la Computación

I.- IDENTIFICACIÓN

Nombre: Data Science 1: Introducción a la Ciencia de Datos			
Código:	Créditos: 3	Créditos SCT: 6	
Modalidad: presencial	Calidad: especialidad	Duración: semestral	
Trabajo Académico: 160			
Horas Teóricas: 32		Horas Prácticas: 10	Horas Laboratorio: 22
Horas de otras actividades: 96			
Horas presenciales: 64		Horas no presenciales: 96	

II.- DESCRIPCIÓN

En esta asignatura se asume que el estudiante tienen conocimientos fundamentales en probabilidades y estadística. Esta asignatura proporciona conocimientos introductorios a la ciencia de datos. Aporta conocimientos básicos en el manejo de datos masivos, aprendizaje computacional, visualización y computación de alto rendimiento.

Esta asignatura aporta a la siguiente competencia del perfil de egreso:

- Mostrar un manejo profundo y actualizado en Ciencias de la Computación, centrándose en conocimientos fundamentales en teoría de computación.

III.- RESULTADOS DE APRENDIZAJE ESPERADOS

Al finalizar el curso los alumnos deben ser capaces de:

1. Adquirir, explorar, administrar y muestrear datos.
2. Identificar y comparar distintos modelos de aprendizaje supervisado y no supervisado entrenados sobre grandes volúmenes de datos.
3. Visualizar grandes volúmenes de datos interpretar los conocimientos extraídos de ellos y comunicarlos.
4. Implementar técnicas de manejo de datos en arquitecturas de computación de alto rendimiento.
5. Desarrollar un proyecto básico de ciencia de datos en colaboración utilizando herramientas computacionales.

IV.- CONTENIDOS

1. Repaso de probabilidades y estadística
 - a. Probabilidades y variables aleatorias
 - b. Estadística descriptiva, teorema del límite central, correlaciones
 - c. Inferencia
2. Data wrangling
 - a. Tipos de datos
 - b. Exploración de datos
 - c. Limpieza de datos
3. Análisis de datos y aprendizaje computacional

- a. Dimensionalidad y su reducción
- b. Aprendizaje supervisado:
 - i. Regresión paramétrica: regresión lineal, Ridge, LASSO, etc.
 - ii. Regresión no paramétrica: Kernel regression, Gaussian process regression, etc.
 - iii. Clasificación: support vector machines, árboles de decisión, random forest, etc.
- c. Aprendizaje no supervisado:
 - i. Clustering: K-means, hierarchical clustering, DBSCAN, etc.
 - ii. Estimación de densidad: histogramas, kernel density estimation, Gaussian mixture models, etc.
- 4. Interpretación de resultados
 - a. Técnicas básicas de visualización
 - b. Presentación de conocimiento extraído de datos
- 5. Introducción a la computación de alto rendimiento
 - a. Computación distribuida
 - b. MapReduce

V.- METODOLOGÍA

El curso contará con clases teóricas, prácticas y laboratorios. Se requerirá la participación activa de los alumnos mediante la realización de tareas orientadas a la implementación de distintos algoritmos, el desarrollo de proyectos, la discusión de materiales y la presentación de temas afines.

VI. EVALUACIÓN

La evaluación de la asignatura se calculará como el promedio ponderado de tareas, certámenes, cuestionarios periódicos, presentación de artículos científicos y un proyecto grupal que se llevará a cabo durante el transcurso del curso.

VII.- BIBLIOGRAFÍA Y MATERIAL DE APOYO

Básica

Christopher M. Bishop: Pattern Recognition and Machine Learning. Springer, 2007. ISBN-10: 0387310738, ISBN-13: 978-0387310732.

Tan P., Steinbach M., Kumar V.: Introduction to Data Mining. Addison-Wesley, 2006. ISBN-10: 0321321367, ISBN-13: 978-0321321367.

Complementaria

Trevor Hastie, Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer; 2nd edition 2016. ISBN-10: 0387848576, ISBN-13: 978-0387848570.