
Syllabus - Introduction to Data Science

Curso Gratuito

Meses: Mayo/Junio 2020 – Lugar: Internet

El material pertinente del curso, incluido este syllabus, y los recursos relacionados con el curso estarán disponibles en el repositorio del curso. Además, el sitio (https://github.com/patriciomalleag/DS_1) se utilizará para publicar material de sesiones, anuncios, etc.

Curso	Introduction to Data Science.
Carga Horaria	3 sesiones semanales de 80 minutos.
Descripción	Data Science es el estudio de la extracción generalizable de conocimiento a partir de datos, que en su gran parte resultan ser masivos (Big Data). Se presentará un conjunto de habilidades integradas que abarcan matemáticas, estadísticas, aprendizaje automático, bases de datos y otras ramas de la informática, junto con una buena comprensión del arte de la formulación de problemas para diseñar soluciones efectivas. Este curso introducirá a los participantes a este campo de rápido crecimiento y los equipará con algunos de sus principios y herramientas básicos. Además se espera que se aprenderán los conceptos, técnicas y herramientas que necesitan para lidiar con diversas facetas de la práctica de la ciencia de datos, incluyendo recopilación e integración de datos, análisis exploratorio de datos, modelos predictivos, modelos descriptivos, creación de productos de datos, evaluación y comunicación efectiva. El enfoque en el tratamiento de estos temas estará en la amplitud, en lugar de la profundidad, y se hará hincapié en la integración y síntesis de conceptos, además en su aplicación para resolver problemas.
Objetivos	<p>Se espera que al final del curso se pueda:</p> <ul style="list-style-type: none">■ Describir conceptos como Data Science, Big Data y los conjuntos de habilidades necesarios para ser un científico de datos.■ Explicar la importancia de organizar bien el proceso previo a un proyecto de análisis de datos, e identificar los pasos a seguir.■ Describir el proceso de Data Science y cómo interactúan sus componentes.■ Utilizar SQL para realizar consultas en bases de datos relacionales.■ Introducirse en los modelos de bases de datos no relacionales (NoSQL).■ Utilizar Python para llevar a cabo análisis y modelos estadísticos básicos.■ Aplicar algoritmos básicos de Machine Learning para el modelamiento predictivo.■ Identificar algoritmos básicos de clasificación.■ Identificar y explicar algoritmos fundamentales que constituyen un motor de recomendación.
Audiencia:	El curso es adecuado para participantes de pregrado (o posgrado) con interés en el área de análisis de datos. No hay restricciones con el área en el que se desempeñe o estudie el participante, puesto que el campo de aplicación es amplio; además tampoco es requisito que se encuentre estudiando mientras tenga conocimientos en los prerrequisitos que se describen mas adelante en este syllabus.

¹patricio.mallea@ug.uchile.cl

Prerequisitos	Se espera que los participantes tengan conocimientos básicos de algoritmos, una experiencia de programación razonable, y cierta familiaridad con el álgebra lineal básica (p. Ej., Solución de sistemas lineales y cálculo de valores y vectores propios) y una buena base en probabilidad y estadística. Si está interesado en tomar el curso, pero no está seguro de tener los conocimientos necesarios, es posible que aún pueda sobrellevarlo si está dispuesto a hacer un esfuerzo adicional para llenar los vacíos.
Modalidad	El curso consta de sesiones tres veces a la semana, 80 minutos cada una, de las cuales 2 serán teóricas y otra práctica, e involucra un conjunto de laboratorios externos (herramientas online) y un proyecto opcional para los participantes. El proyecto podría adoptar varias formas: analizar un conjunto de datos interesante utilizando métodos y herramientas de software existentes; construyendo su propio producto de datos; o crear una visualización de un conjunto de datos complejo. Se alienta a los participantes a trabajar en equipos de dos o tres para el proyecto.

Tópicos del curso

1. Data Science y Big Data: Conceptos básicos. *(1 semana)*
 - a) ¿Por qué utilizar Data Science y Big Data?
 - b) Formatos de datos existentes.
 - c) Acercamiento al proceso de Data Science.
 - d) Frameworks y Herramientas disponibles.
2. Procesos de Data Science. *(1 semana)*
 - a) Definiendo metas y creando el proyecto.
 - b) Obteniendo los datos.
 - c) Limpiando los datos.
 - d) Análisis de datos exploratorio.
 - e) Construyendo modelos.
 - f) Presentando descubrimientos y construyendo aplicaciones sobre eso.
3. Structured Query Language. [SQL] *(2 semanas)*
 - a) ¿Qué es SQL?
 - b) Tipos de datos y estructuras.
 - c) Consultas más comunes.
 - d) Funciones para el análisis de datos.
 - e) Introducción a NoSQL.
4. Python para Data Science. *(2 semanas)*
 - a) Repaso de programación.
 - b) Librerías importantes.
 - c) Clouds y sus herramientas.
 - d) Visualización de datos.
5. Machine Learning. *(2 semanas)*
 - a) Conceptos básicos.
 - b) El proceso de modelamiento.
 - c) Modelos de regresión.
 - d) Modelos de clasificación.
 - e) Modelos de clusterización.
 - f) Modelos de recomendación.
6. Big Data y como manejarlo. *(1 semana)*
 - a) Problemas con datos masivos.
 - b) Técnicas y tips para lidiar con datos masivos.
 - c) Frameworks: Hadoop y Spark.
7. Text Mining: Analizando textos. *(1 semana)*
 - a) Aplicaciones de la minería de datos.
 - b) Técnicas para la minería de datos.

Bibliografía

- Ali, M., Cielen, D. & Meysman, A. (2016). *Introducing Data Science*. [Manning]
- Emberak, O. (2018). *Data Analysis and Visualization Using Python*. [Apress]
- Goldwasser, M., Johnston, B. & Malik, U. (2019). *SQL for Data Analytics*. [Packt]
- Dubovikob, K. (2019). *Managing Data Science*. [Packt]
- Irizarry, R. (2019). *Introduction to Data Science*.
- Kishore, V. (2018). *Pro Machine Learning Algorithms*. [Apress]