Programa del curso:

“Introducción a la ciencia de datos en R”

-Año 2020-

Docente a cargo: Lic. Andrés Rabosto

Contacto: andres.rabosto87@gmail.com

Cargar horaria: 64 hs / dos semestres independientes de 32 hs

Fecha de inicio: 27/03/2020

Objetivo:

El curso tiene como objetivo introducir a los participantes en diversas áreas de la llamada “ciencia de datos”, utilizando el software R para importar, limpiar, analizar, modelar y visualizar diferentes tipos de datos. Se hará énfasis en las tareas básicas de lectura, escritura y manipulación de datos, análisis exploratorio y estadístico.

El curso brinda una introducción a diversos temas y herramientas para el análisis sistemático de datos y la utilización de modelos explicativos y predictivos. Se busca que los participantes adquieran conocimientos y herramientas que les permitan luego continuar avanzando por su cuenta.

Fundamentos:

R comenzó como un software y un lenguaje de programación para análisis gráfico y estadístico de datos. Con los años su uso y aplicaciones fueron extendiéndose a prácticamente a todas las disciplinas, y hoy es ampliamente utilizado no solo por estadísticos sino también por sociólogos, economistas, politólogos, consultores, analistas financieros, comunicadores, geógrafos, biólogos, etc. Junto con Pyton, constituye el entorno más utilizado en la naciente de disciplina de la ciencia de datos.

Esta popularidad se debe en buena medida a que se trata de software libre: una plataforma gratuita de código abierto que permite que numerosos grupos de usuarios aporten sus desarrollos (librerías o paquetes) para realizar diferentes tipos de tareas, como por ejemplo, el análisis de textos, sonidos, imágenes, etc. Esta gran versatilidad requiere un uso intensivo de la línea de comandos en la interfaz de usuario, por la cual muchos potenciales usuarios evitan aprender R, ya que en una primera aproximación lo encuentran difícil de usar.

Dirigido a: Todo aquel interesado en el análisis de datos.

Curso dinámico, con fundamentos teóricos y aplicaciones prácticas en computadora, evitando tecnicismos. Aunque es recomendable tener alguna noción de estadística y

análisis de datos, no se necesitan conocimientos previos. Cada punto se irá desarrollando teóricamente junto con la implementación práctica.

Si bien está pensado integralmente, el curso se divide en dos bloques independientes. No es necesario cursar ambos.

Bloque 1: introducción a R y al análisis de datos.

Entorno de R y RStudio, programación básica en R, análisis exploratorio y descriptivo, visualización de datos, correlación y covarianza, modelado clásico, informes y comunicación de resultados.

Bloque II: Introducción a modelos de Aprendizaje Automático (Machine Learning) y datos no estructurados

Redes neuronales, árboles de decisión, k-means, Support Vector Machines. Análisis de opiniones, procesamiento de imágenes.

Bloque I:

1. Introducción a R y RStudio

    1. Entorno R y RStudio
    2. Programación básica
    3. El paquete tydiverse

2. Limpieza de datos, análisis exploratorio y descriptivo

    1. Importación de datos. Tipos de datos y variables.
    2. Data frames. Tratamiento de datos faltantes.
    3. Distribuciones, frecuencias, tendencia central y dispersión
    4. Covariancia y correlación

3. Visualización de datos

    1. Gráficos en Rbase
    2. Gráficos en capas: ggplot2
    3. Gráficos animados: gganimate

4. Introducción al modelado clásico

    1. Regresión lineal
    2. Regresión logística

5. Documentación en R y comunicación de resultados

    1. RMarkDown
    2. Shinyapps

Bloque II

6. Introducción a métodos de clasificación de Aprendizaje automático

    1. árboles de decisión
    2. k-means.

7. Clasificación con redes neuronales

    1. Modelado con redes neuronales
    2. Matriz de confusión y performance del modelo
    3. Introducción a Support Vector Machine

8. Regresión con redes neuronales

    1. Machine learning
    2. Testeo: Train/test
    3. Cross validation
    4. Medidas de ajuste

9. Análisis computacional de textos

    1. Obtener datos de la API de twitter
    2. Bolsa de palabras y corpus de texto
    3. Medidas de resumen y nubes de palabras

10. Procesamiento de imágenes