ASOCIACIÓN ARGENTINA
DE BIOINFORMÁTICA
Y BIOLOGÍA COMPUTACIONAL

Introducción al análisis de datos con R

Del 29 de noviembre al 10 de diciembre de 2021, de 9 am a 12 pm.

Calendario:
-Lunes 29 de noviembre.
-Miércoles 1 de diciembre.
-Viernes 3 de diciembre.
-Lunes 6 de diciembre.
-Martes 7 de diciembre.
-Viernes 10 de diciembre.

Objetivos:
El curso propone una introducción a la programación con el lenguaje R orientada al análisis de datos en general, con énfasis en el análisis de datos biológicos a partir de ejemplos.

Audiencia:
El curso está pensado para investigadores y estudiantes de grado y posgrado que tengan interés en hacer sus primeros pasos en la programación y en el análisis de datos. No se requieren conocimientos previos de programación aunque es aconsejable tener conocimientos mínimos de computación, de estadística y de biología para comprender mejor los ejemplos (sin embargo, esto último no es excluyente).
Duración y modalidad:
18 horas, con 9 horas de teoría y 9 horas de práctica en total. Las clases se dictarán de forma virtual.

Certificados:

Se tomará un examen no obligatorio al finalizar el curso y se emitirá certificado de aprobación con nota.
Se emitirán certificados de participación para aquellas personas que no rindan el examen.

Docentes:
Lic. Andres Rabinovich
Tomas Vega Waichman
Lic. Fernando Orti
Lic. Maximiliamo S. Beckel

Programa:
Clase 0) Preparación del entorno de trabajo antes de asistir al curso.
Instalación de R y RStudio.

29/11 Clase 1) Introducción a R y RStudio.
Qué es R. La consola y el entorno Rstudio.
Variables y tipos de datos.
Introducción a estructuras de control (IF y FOR) y funciones.
Cómo utilizar las ayudas de R (y google).
Repositorios de librerías (CRAN y Bioconductor).
Cómo instalar y acceder a librerías.

01/12 Clase 2) Introducción a R, continuación.
Importación, exportación, visualización y limpieza de datos.
Subsetting de vectores, listas, data.frames, matrices.
Vectorización.

03/12 Clase 3) Estadística descriptiva.
Elementos básicos de estadística descriptiva (media, mediana, moda, desvío estándar, varianza, IQR).
Distribuciones.
Visualización de datos: scatterplot, histograma, boxplot, gráfico de coordenadas paralelas.
PCA para poder visualizar espacios de alta dimensionalidad.

06/12 Clase 4) Breve introducción a las pruebas de hipótesis (con R).
Qué es y para qué sirve una prueba de hipótesis.
Tipos de errores.
P-valores e Intervalos de confianza.
Cómo realizar algunas pruebas de hipótesis útiles en R (t.test, ANOVA, chi cuadrado)

07/12 Clase 5) Introducción a Machine learning: aprendizaje supervisado.
Formas de medir distancias o similitudes en los datos.
KNN para clasificar datos discretos.
Partición del conjunto de datos en entrenamiento y testeo para mejorar la validación.
Cómo realizar una regresión lineal en R. Analizar la salida y utilizar el modelo para predecir nuevos datos.

10/12 Clase 6) Introducción a Machine learning: aprendizaje no supervisado.
Partición de los datos en grupos (clustering).
Clustering jerárquico, dendrogramas y alturas de corte.
Clustering por k-means. Métodos para elegir K.
Silhouette para validar los clusters obtenidos.
Problemas con la forma de los clusters.

Bibliografía recomendada (manuales, tutoriales y libros):
An introduction to R. [online]
McDonald J. (2014). HANDBOOK OF BIOLOGICAL STATISTICS, SPARKY HOUSE PUBLISHING. [pdf]
Mangiafico S. (2015), AN R COMPANION FOR THE HANDBOOK OF BIOLOGICAL STATISTICS, New Brunswick, Rutgers University [pdf]
James, Witten, Hastie & Tibshirani, "An Introduction to Statistical Learning with Applications in R", 2nd ed, Springer, 2017 [pdf]