FORMATO PARA PROPONER CURSOS

SEMESTRE 2020-1

EL CURSO YA HA SIDO EVALUADO E IMPARTIDO SI _____ NO _X_

1. Indicar la modalidad: Tópico Selecto	
2. Título: *Ciencia de Datos*	
3. Tutor responsable	
Nombre completo	Luis Antonio Mendoza Sierra
Entidad Académica	Instituto de Investigaciones Biomédicas
Teléfono	5556229210
RFC	MESL7003113N5
Correo electrónico	lmendoza@biomedicas.unam.mx
4. Profesores invitados	
Nombre completo	Carlos Ramírez Álvarez
Entidad/Adscripción	Hospital Infantil de México Federico Gómez
Teléfono	55 1678 1174
RFC	RAAC871020AA4
Correo electrónico	cramireza@ciencias.unam.mx
5. Características para la impartición del Curso	
Indique:	
El lugar en donde se realizará	Sala de cómputo del Instituto de Investigaciones Biomédicas, nueva sede.
Los días de la semana en los que se impartirá	martes y viernes
El horario propuesto	17-19 horas
El número de sesiones	18
La duración en horas por sesión (mínimo 36 horas)	2 horas

El número total de alumnos que puede aceptar	15 alumnos (Importante: enviar correo previamente a cramireza@ciencias.unam.mx para verificar disponibilidad)
El número de alumnos del PDCB que puede aceptar	15 alumnos (Importante: enviar correo previamente a cramireza@ciencias.unam.mx para verificar disponibilidad)
La disponibilidad de impartirlo por videoconferencia	SI

6. Método de evaluación

Por favor incluya en este apartado él % de la contribución relativa de:

Exámenes (número)	30%
Participación en clase	10%
Presentación de un proyecto	30%
Trabajos	30%
Otros	0%

7. Introducción/justificación del Curso

Los resultados de los experimentos biológicos requieren ser ordenados, clasificados y visualizados en forma de gráficas y tablas, además de someterse a pruebas estadísticas y de ser posible ajustarse a modelos matemáticos predictivos, todo lo anterior con el objetivo de interpretar, dar sentido biológico y generar nuevas hipótesis que den pie a nuevos descubrimientos. La presentación de los datos también permite una mejor comunicación de los resultados experimentales y a resaltar la importancia del trabajo experimental.

En este curso se expondrán una serie de herramientas para el tratamiento y análisis de datos. En la actualidad a este conjunto de habilidades se le ha denominado Ciencia de Datos y se encuentra en tendencia en muchos campos desde el ámbito científico, en las ciencias sociales y en el sector financiero. Se hará especial énfasis en el análisis de datos resultados de experimentos biológicos que incluyen desde ensayos de casos *versus* controles, dosis-respuesta, búsqueda de predictores en datos clínicos, hasta el estudio de datos transcriptómicos en experimentos de secuenciación de nueva generación. El curso también pretende ser un taller para asesorar al estudiante en el análisis de sus propios datos experimentales.

8. Temario

1. Panorama general de la Ciencia de Datos [2 horas, Luis Mendoza].
2. Lenguaje bash (Linux) [2 horas, Carlos Ramírez].
3. Introducción al lenguaje de programación R (Rstudio) [2 horas, Carlos Ramírez].
4. Importe y manejo de tablas (txt, csv, tsv, vcf) [2 horas, Carlos Ramírez].
4.1. Manejo de data frames y matrices (dplyr) [2 horas, Carlos Ramírez].
4.2. Manipulación de cadenas de caracteres [2 horas, Carlos Ramírez].
5. Gráficas y análisis exploratorio de datos (R base) [2 horas, Carlos Ramírez].

6. Creación de gráficas utilizando (ggplot) [2 horas, Carlos Ramírez].
7. Pruebas estadísticas y simulaciones de eventos aleatorios (R base) [2 horas, Carlos Ramírez].
8. Uso de herramientas para el manejo de versiones de código (git, github, gitlab)
9. Creación de reportes desde R (markdown) [2 horas, Carlos Ramírez].
10. Caso de estudio: patrones de expresión en datos transcriptómicos de células sanguíneas [2 horas, Carlos Ramírez].
11. Principios de minería de datos (conexión via API) [2 horas, Carlos Ramírez],
12. Interfaces interactivas de R y python (Jupyter) [2 horas, Carlos Ramírez].
13. Introducción al lenguaje python [2 horas, Carlos Ramírez].
14. Caso de estudio: Data mining de la literatura de Pubmed (E-utilities) [2 horas, Carlos Ramírez].
15. Introducción a lenguajes de bases de datos (sql) [2 horas, Carlos Ramírez].
16. Caso de estudio: manejo de archivos fasta, sam, bam, vcf [2 horas, Carlos Ramírez].
17. Introducción al aprendizaje de máquina (clustering, modelos de regresión, ridge regression, LASSO, random forest, svm) [2 horas, Carlos Ramírez].
18. Caso de estudio: desarrollo de un clasificador de datos transcriptómicos [2 horas, Carlos Ramírez].

9. Bibliografía

Roger D. Peng. *R programming for Data Science* (2015). Leanpub.
Gareth James. *An introduction to Statistical Learning: with Applications in R* (2017). Springer Texts in Statistics.