OPR319: Introducción a la Ciencia de Datos Aplicada con RStudio

Roberto Cantillan Pablo Jiménez

2024-01-01

E-mail: TBD
Office Hours: [Horario de oficina]
Office: [Número de oficina]

Web: Repositorio link
Class Hours: [Horario de clases]
Class Room: [Sala de clases]

Identificación de la Actividad Curricular

- **Nombre**: Ciencia de Datos Aplicada con RStudio
- **Código**: OPR 319 (Equivalente a SLG-522)
- **Semestre lectivo**: X
- **Horas**: Presencial: 54 | Autónomas: 96 | TOTAL: 150
- **Créditos SCT**: 5
- **Duración**: Semestral
- **Modalidad**: Presencial
- **Área de Formación**: Profesional
- **Requisito**: Todas las actividades curriculares aprobadas hasta el VIII semestre

Descripción y Caracterización de la Actividad Curricular

La actividad curricular de Electivo III (Ciencia de datos aplicada) se ubica en el X semestre de la carrera de Sociología y pertenece al área de formación profesional.

En la era actual de big data, la capacidad de analizar datos y extraer información valiosa se ha convertido en una habilidad esencial para los científicos sociales y profesionales en diversos campos. En este contexto, el propósito fundamental de este curso es introducir a los estudiantes en las herramientas básicas para extraer información valiosa a partir de datos sin procesar, a menudo generados fuera de los tradicionales diseños de investigación científica. Esta habilidad es crucial no solo para la investigación académica, sino también para la toma de decisiones informadas en el sector público y privado.

El curso busca establecer una base sólida para el dominio de las herramientas necesarias en Ciencia de Datos, utilizando R y Tidyverse, un conjunto de software estadístico de código abierto ampliamente reconocido en la investigación tanto académica como aplicada. Se abordarán cada etapa del proceso de manipulación y análisis de datos, incluyendo importación, limpieza, transformación, validación, visualización, modelado y creación de informes automáticos y reproducibles.

La programación en R, utilizando la interfaz de RStudio, es un componente central de este curso. Aprender a programar no solo permite a los estudiantes realizar análisis de datos más eficientes y reproducibles, sino que también desarrolla habilidades de pensamiento lógico y resolución de

problemas que son valiosas en cualquier campo profesional. Además, el dominio de R y RStudio proporciona a los estudiantes herramientas poderosas y flexibles que son ampliamente utilizadas tanto en la academia como en la industria.

Al finalizar el curso, los estudiantes serán capaces de:

1. Analizar bases de datos de complejidad intermedia a avanzada.
2. Ejecutar tareas de programación para procesamiento de datos, relacionando niveles de estratificación con fenómenos sociales sustantivos.
3. Modelar con bases de datos simples y complejas, relacionando resultados con fenómenos sociales sustantivos.
4. Comunicar resultados de análisis mediante la creación de informes automatizados y reproducibles.

Estas habilidades no solo mejorarán la capacidad de los estudiantes para realizar investigaciones rigurosas, sino que también aumentarán significativamente su competitividad en el mercado laboral, donde la demanda de profesionales con habilidades en análisis de datos está en constante crecimiento.

Competencias del Perfil de Egreso Asociadas a la Actividad Curricular

Competencias Profesionales

1. Gestionar organizaciones, proyectos e intervenciones orientándose al trabajo colaborativo y con apertura a la diversidad social.
    - 1.3. Gestionar proyectos e intervenciones sociales, promoviendo el trabajo en equipo interdisciplinar y colaborativo.
2. Proponer iniciativas pertinentes a las demandas y necesidades de entidades diversas, en base a una comprensión integral de los fenómenos y los contextos sociales.
    - 2.3. Proponer iniciativas pertinentes, relacionando críticamente conceptos y teorías contemporáneas de la sociología y las ciencias sociales.

Competencias Genéricas

5. Realizar investigaciones que contribuyan al desarrollo del conocimiento científico y aplicado en contextos propios de su proceso formativo.
    - 5.3. Desarrollar investigación aplicada, implementando los pasos del método científico y articulando conclusiones adecuadas y coherentes al proceso investigativo.

Resultados de Aprendizaje - Aprendizajes Esperados

1. Ejecutar tareas de programación para procesamiento de datos relacionando niveles de estratificación con fenómenos sociales sustantivos.
2. Modelar con bases de datos simples y complejas, relacionando resultados con fenómenos sociales sustantivos.
3. Comunicar resultados de análisis mediante la creación de informes automatizados y reproducibles.

Unidades de Aprendizaje y Ejes Temáticos

Semana 01, 08/08 - 12/08: Introducción a R y RStudio

- Instalación de R y RStudio

- Interfaz de RStudio
- Conceptos básicos de R: objetos, funciones, paquetes
- Tipos de datos en R: numéricos, caracteres, lógicos, factores
- Operaciones aritméticas y lógicas
- Vectores y matrices

Semana 02, 15/08 - 19/08: Jueves 15 feriado

Semana 03, 22/08 - 26/08: Operaciones Básicas y Estructuras de Datos

- Operaciones aritméticas y lógicas
- Vectores y matrices
- Listas y data frames
- Indexación y subconjuntos

Semana 04, 29/08 - 02/09: Importación y Exportación de Datos

- Lectura de archivos CSV, Excel, y otros formatos
- Escritura de datos en diferentes formatos
- Conexión con bases de datos

Semana 05, 05/09 - 09/09: Introducción a `Tidyverse`

- Filosofía de `Tidyverse`
- Pipes (%>%) y su uso
- Introducción a `dplyr`: `select()`, `filter()`, `mutate()`

Semana 06, 12/09 - 16/09: Manipulación de Datos con `dplyr` I

- Agrupación y resumen de datos: `group_by()` y `summarize()`
- Ordenamiento de datos: `arrange()`
- Creación de nuevas variables (avanzado): `mutate()`, `case_when()`

Semana 07, 19/09 - 23/09: Feriado Fiestas Patrias

Semana 08, 26/09 - 30/09: Manipulación de Datos con `dplyr` II

- Joins: `inner_join()`, `left_join()`, etc.
- Operaciones de conjunto: `union()`, `intersect()`, `setdiff()`
- Manejo de datos faltantes

Semana 09, 03/10 - 07/10: Transformación de Datos con `tidyr`

- Datos tidy y su importancia
- Funciones `pivot_longer()` y `pivot_wider()`
- Separación y unión de columnas: `separate()` y `unite()`

Semana 10, 10/10 - 14/10: Iteración y Automatización con `purrr`

- Conceptos de programación funcional
- Familia de funciones `map()`
- Uso de `purrr` con `dplyr`

Semana 11, 17/10 - 21/10: Visualización de Datos I: Fundamentos de `ggplot2`

- Gramática de gráficos
- Capas en `ggplot2`
- Tipos básicos de gráficos: dispersión, líneas, barras

Semana 12, 24/10 - 28/10: Visualización de Datos II: `ggplot2` Avanzado

- Faceting
- Temas y personalización de gráficos
- Combinación de múltiples gráficos

Semana 13, 31/10 - 04/11: Feriado

Semana 14, 07/11 - 11/11: Introducción al Modelado Estadístico

- Regresión lineal con `lm()`
- Interpretación de resultados
- Diagnósticos básicos de modelos

Semana 15, 14/11 - 18/11: Modelos Avanzados con `tidymodels`

- Introducción a `tidymodels`
- Preprocesamiento de datos
- Entrenamiento y evaluación de modelos

Semana 16, 21/11 - 25/11: Reportes Automatizados I: Introducción a `RMarkdown`

- Sintaxis básica de `Markdown`
- Chunks de `R` en `RMarkdown`
- Generación de reportes en diferentes formatos (PDF, HTML)

Semana 17, 28/11 - 02/12: Reportes Automatizados II: `Quarto`

- Parámetros en `Quarto`
- Creación de presentaciones con `Quarto`
- Introducción a `Quarto`

Semana 18, 05/12 - 09/12: Proyecto Final - Trabajo en Clase

- Sesión de trabajo guiado para el proyecto final
- Consultas y asesorías individuales

Semana 19, 12/12 - 16/12: Proyecto Final - Presentaciones

- Presentación de proyectos finales por parte de los estudiantes
- Retroalimentación y discusión

Estrategias de Enseñanza y Aprendizaje

- Clases expositivas dialogadas
- Aprendizaje Basado en Problemas
- Aprendizaje Colaborativo o Cooperativo
- Talleres
- Trabajo en laboratorio de computación

Procedimientos de Evaluación de Aprendizajes

1. Tarea de programación I (35%)
   - Instrumento: Rúbrica e Informe (Escala de apreciación)
   - Contenidos: Importación y exportación, agrupación y resumen de datos y manipulación de datos de datos con `dplyr`, manipulación de datos II (unión), transformación de datos (`tidyr`), manejo de datos faltantes.
2. Tarea de programación II (35%)
   - Instrumento: Rúbrica y Diagrama de flujo (Pauta de cotejo)
   - Contenidos: Iteración y automatización (`purrr`), creación de gráficas (`ggplot`).
3. Tarea de programación III (30%)
   - Instrumento: Rúbrica y Autoevaluación de proceso (Pauta de cotejo)
   - Contenidos: Modelado estadístico con `tidymodels` (procesamiento y uso de resultados), reportes con `Quarto`.

Recursos de Infraestructura

- Sala de clases con proyector y audio
- Sala de clases con computadores con R y RStudio instalado
- Acceso a UCM Virtual (Plataforma Web LMS)

Recursos Bibliográficos

Básica Obligatoria

- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for data science. O'Reilly Media, Inc. Versión en español: R para ciencia de datos: https://es.r4ds.hadley.nz/
- Ismay, C., & Kim, A. Y.-S. (2020). Statistical inference via data science: A ModernDive into R and the Tidyverse. CRC Press / Taylor & Francis Group. (versión gitbook de libre acceso: https://moderndive.com/)

Complementaria

- Imai, K., & Williams, N. W. (2022). Quantitative Social Science: An Introduction in Tidyverse. Princeton University Press.
- Wickham, H. (2019). Advanced r. CRC press. (link a versión on line de libre acceso https://adv-r.hadley.nz/)

Otros Recursos

- UCM Virtual (Plataforma Web LMS)
- Artículos científicos (Digital)
- Repositorio SciELO (Página Web)
- SIBIB (Página Web)
- Repositorio GitHub (Página Web)
- Base de datos de acceso público (Página Web de distintas organizaciones)