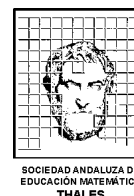


19 CONGRESO DE ENSEÑANZA Y APRENDIZAJE DE LAS MATEMÁTICAS.

Matemáticas y Multiculturalidad
Aproximación a través de una matemática dinámica



TÍTULO: CIENCIA DE DATOS. MODELOS GENERATIVOS Y PREDICTIVOS CON CODAP

AUTOR/ES:

María Lucia Vargas, *Colegio La Salle-Buen Consejo, Puerto Real (Cádiz)*

Ana Serradó Bayés, *Colegio La Salle-Buen Consejo, Puerto Real (Cádiz)*

e-mail: marialuciavargas@lasallebuenconsejo.es

RESUMEN

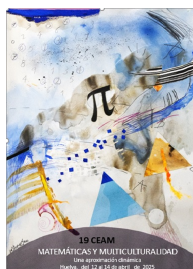
CODAP es un software educativo gratuito para el análisis de datos. Esta herramienta web para el estudio de la ciencia de datos está diseñada como plataforma para desarrolladores y como aplicación para alumnado de Primaria, Secundaria, Bachillerato y Universidad. Para ello, el taller se iniciará con una introducción teórica sobre la necesidad de herramientas tecnológicas que permitan desarrollar las prácticas en ciencias de datos. Para posteriormente, poner en práctica la ciencia de datos en situaciones de aprendizaje concretas que permitan: (a) conocer las formas básicas de interactuar con los datos (visualización, exploración y comparación), (b) desarrollar los movimientos en ciencias de datos (filtrado, agrupación, resumen, cálculo, fusión y jerarquización), (c) analizar modelos generativos de datos masivos (parámetros de posición, centralización y dispersión), (d) establecer asociaciones de variables; (e) aplicar del algoritmo Arbor y los principios del “aprendizaje automático” en el desarrollo de modelos predictivos.

Nivel educativo: Primaria, Secundaria Obligatoria, Bachillerato y Universidad

INTRODUCCIÓN

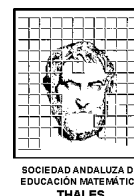
La comprensión y el manejo de datos se ha convertido en un aprendizaje imprescindible hoy en día porque en muchos contextos de toma de decisiones en el mundo, incluidos los negocios, la política y la sociedad, se apoyan en pruebas basadas en datos. Para apoyar al alumnado desde los primeros niveles de Educación Primaria en su desarrollo como ciudadanos, es crucial que el razonamiento sobre los datos se estimule a nivel escolar (Frischemeier y otros, 2021). Una forma de animar al alumnado a razonar sobre los datos es ofrecerles la oportunidad de participar en el análisis exploratorio de datos (AED). El análisis exploratorio de datos fue introducido por primera vez por Tukey en el año 1962 en el ensayo “*El futuro del análisis de datos*”, donde argumentaba la importancia del análisis de datos por encima de otros aspectos del estudio estadístico (Tukey, 1962). Este documento se considera el precursor de la denominada Ciencia de datos, definida por Yan y Davis (2019) como “*la ciencia del aprendizaje de los datos*”. La ciencia de datos en educación se concibe como un campo fuertemente interdisciplinar, y se define como “*un conjunto de habilidades y técnicas que incluyen la estadística, la informática (codificación, visualización, computación de datos, aprendizaje automático, etc.), las matemáticas y la experiencia en el área temática de la que proceden los datos*” (Engel, 2017, pág. 1).

Por sus características, su aprendizaje implica la realización de conexiones entre las diferentes disciplinas que la integran para dar respuesta a problemas sociales a partir del estudio de los datos. Al intentar dar respuesta a estas preguntas nos enfrentamos a la concepción de las matemáticas como



19 CONGRESO DE ENSEÑANZA Y APRENDIZAJE DE LAS MATEMÁTICAS.

Matemáticas y Multiculturalidad Aproximación a través de una matemática dinámica



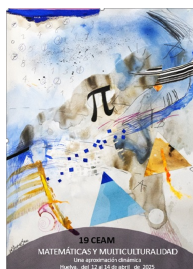
un lenguaje que nos permita explicar hechos ya acaecidos, realizar inferencias y predicciones para tomar decisiones sobre hechos reales. En este sentido el taller favorecerá que los participantes reflexionen sobre las siguientes cuestiones incluidas en documento de presentación del XIX CEAM: qué matemática es necesaria en una sociedad matematizada, cómo otras ciencias recurren constantemente al lenguaje matemático para expresar sus informaciones (como la geografía o la geología) o, si sirven los procesos de matematización de las prácticas con datos para entender el quehacer matemático en el aula que tiene origen en las necesidades sociales y tecnológicas del siglo XXI.

Se presentará el papel de la estadística en la ciencia de datos como una disciplina matemática necesaria para una sociedad cada vez más matematizada. La complejidad del papel que adquieren las matemáticas en esta sociedad se entiende por las interacciones que realiza constantemente con otras áreas STEM/STEAM para expresar las conclusiones que se obtienen de los ciclos de investigación estadística. Dichos ciclos se presentarán en este taller: (a) como los organizadores de las acciones a realizar con CODAP, los movimientos necesarios para enfrentarse a grandes tablas de datos (“big data”); (b) como procesos de matematización de las prácticas con datos que permiten entender una nueva forma de quehacer matemático en el aula de primaria, secundaria y universidad que tiene origen en las necesidades sociales actuales y se fundamenta en los avances tecnológicos del siglo XXI. En relación con esta última pregunta, la sociedad actual necesita del uso de la tecnología, en muchos casos de software estadístico y dinámico. En este taller presentamos la plataforma *Common Online Data Analysis Platform* (CODAP, <https://codap.concord.org/>) para evidenciar el potencial que tiene como entorno de aprendizaje que favorece el desarrollo del razonamiento estadístico (Mojica, Barker y Azmy, 2019).

1. INVESTIGANDO BAJO LOS PRINCIPIOS DE LA CIENCIA DE DATOS

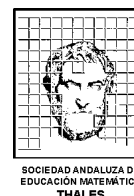
La literatura en el campo de la ciencia de datos desde una perspectiva STEM/STEAM indica que existen, entre otras, dos metodologías básicas: (a) *un aprendizaje basado en problemas donde el alumnado se enfrenta a uno o varios ciclos de investigación estadística en datos, como los propuestos por Silva-Hormazábal, Rodríguez-Silva y Alsina (2022)*, y (b) *un aprendizaje basado en ingeniería donde el alumnado debe diseñar, implementar, evaluar y mejorar artefactos que le permitan recoger datos estadísticos para evaluar su adecuación al objetivo del mismo, y modelizar, si procede su uso* (English, 2017).

Consideramos los diferentes ciclos de investigación estadística como los organizados cognitivos del estudio de datos tanto en el campo de la estadística como en el trabajo de la misma en la educación en la ciencia de datos. Entre las diferentes propuestas de ciclos de investigación estadística, en este taller usaremos el ciclo de seis fases desarrollado por Lee y otros (2022). Este ciclo de seis fases se describe en la figura 1. A diferencia de otros ciclos de tres o cuatro fases, la propuesta de Lee y otros (2022) introduce la modelización y la defiende como una destreza básica a la hora de enfrentarse al estudio de los datos. Breiman (2001) establece que hay dos culturas en el uso de la modelización estadística que permiten obtener conclusiones a partir de los datos: *la cultura de la modelización de los datos y la cultura de la modelización algorítmica*. Donoho (2015) en el campo de la ciencia de datos, interpreta estas dos culturas estableciendo dos objetivos al análisis de datos:



19 CONGRESO DE ENSEÑANZA Y APRENDIZAJE DE LAS MATEMÁTICAS.

Matemáticas y Multiculturalidad Aproximación a través de una matemática dinámica



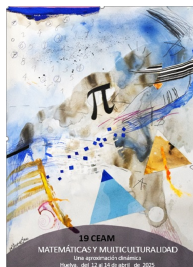
- ❑ La generativa: inferir cómo la naturaleza asocia las variables de respuesta a las variables de entrada.
- ❑ La predicción: poder predecir cuáles van a ser las respuestas a futuras variables de entrada.

En los últimos años, esta cultura de modelización generativa se ha popularizado a nivel escolar al introducir el concepto de inferencia informal, que en nuestro currículum se relaciona con tres saberes: la formulación de preguntas adecuadas que permitan conocer las características de interés de una población, datos relevantes para dar respuesta a cuestiones planteadas en investigaciones estadísticas: selección y presentación de la información procedente de una muestra mediante herramientas digitales, y estrategias de deducción de conclusiones a partir de una muestra con el fin de emitir juicios y tomar decisiones adecuadas. Mientras que la cultura de modelización predictiva se ha asociado a la introducción de principios del *machine learning* o *aprendizaje automático* -entendido como la deducción automática de patrones- que permiten la construcción de un primer modelo entrenándolo, ajustándolo a los datos en términos estadísticos para, a continuación, comprobar con datos de prueba el ajuste y mejorar la generalización del modelo con el fin de predecir datos futuros desde el pasado.



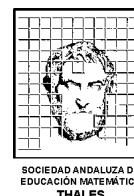
Figura 1. Ciclo de investigación en ciencia de datos. Lee et al. (2022)

Estas prácticas en la ciencia de datos se simplifican al usar CODAP para explorar los datos con seis movimientos básicos, que pasamos a describir en la siguiente sección.



19 CONGRESO DE ENSEÑANZA Y APRENDIZAJE DE LAS MATEMÁTICAS.

Matemáticas y Multiculturalidad
Aproximación a través de una matemática dinámica



2. LOS MOVIMIENTOS DE DATOS COMO MARCO PARA EXPLORAR DATOS CON CODAP

Erickson y otros (2019) definen los movimientos de datos como *“acciones que alteran los contenidos de un conjunto de datos, su estructura, o valores”* (pág. 4). Es decir, un movimiento de datos es una acción que el alumnado realiza a los datos. Los conjuntos de datos más simples, seguramente no necesitan de ningún movimiento. Pero cuando hablamos de datos reales, de “big data” debido a la complejidad de dichos conjuntos de datos, se necesita el uso de estrategias para hacer frente a los mismos. En la ciencia de datos que podemos realizar con CODAP hay seis movimientos de datos a llevar a cabo:

- ❑ *Filtrado*: crear un subconjunto de los datos y centrarse en este subconjunto de los datos.
- ❑ *Agrupación*: división de los datos en subconjuntos (relacionado con el filtrado)
- ❑ *Resumen*: agregar sus datos, por ejemplo, calcular los parámetros de posición, centralización y dispersión como paso previo a la comparación de grupos.
- ❑ *Cálculo*: crear nuevos atributos en la base de datos ya existentes. Por ejemplo, agregar nuevas columnas a una tabla.
- ❑ *Fusión/Unión*: conectar dos bases de datos de forma conjunta.
- ❑ *Reorganizar* el conjunto de datos mediante la creación de una estructura de datos jerárquica.

Todos estos movimientos se realizan de forma simultánea a las diferentes fases de los ciclos de investigación estadística y se refieren a las siguientes acciones que se realizan en CODAP:

- ❑ importar datos y crear un gráfico sencillo,
- ❑ filtrar, reescalar y añadir más atributos a los gráficos (frecuencia, porcentajes, parámetros de centralización, posición y dispersión),
- ❑ explorar mapas,
- ❑ conocer los ejes de los gráficos y la herramienta de dibujo (puntos, histogramas, cajas y bigotes),
- ❑ mostrar y ocultar valores en un gráfico,
- ❑ jerarquizar datos y crear modelos predictivos,

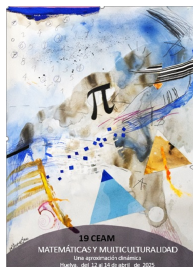
3. ACTIVIDAD DEL TALLER

El objetivo de la actividad del taller es dar respuesta a la pregunta: *¿Qué probabilidad hay de que acontezca un terremoto destructivo en la ciudad de Huelva durante el CEAM-19?*

Se iniciará el taller a partir de establecer aquellas preguntas de investigación estadística que nos permitan responder a la pregunta anterior.

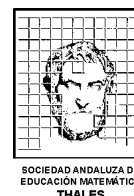
Como fuente de datos usaremos el catálogo de terremotos del Instituto Geográfico Nacional de España con el fin de descargarnos los datos sobre terremotos en la península ibérica desde XI CEAM que se celebró en Huelva en 2004.

Para dar respuesta a la pregunta y sub-preguntas de investigación usaremos la plataforma online CODAP y realizaremos las siguientes acciones:



19 CONGRESO DE ENSEÑANZA Y APRENDIZAJE DE LAS MATEMÁTICAS.

Matemáticas y Multiculturalidad Aproximación a través de una matemática dinámica



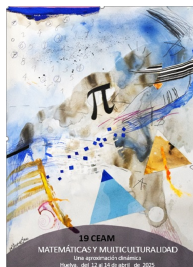
- ❑ importaremos la tabla de datos obtenida a CODAP,
- ❑ seleccionaremos una muestra aleatoria de datos y valoraremos la fiabilidad de la misma para sacar conclusiones sobre la pregunta principal que guiará nuestra investigación, y las preguntas secundarias que hayamos establecido.
- ❑ con dicha muestra, analizaremos las principales características de la tabla (variables, casos...) y fijaremos los atributos correspondientes a cada una de las variables, filtraremos los resultados de la provincia de Huelva,
- ❑ construiremos un mapa que refleje dónde se han producido terremotos en estos 10 años,
- ❑ modelizaremos los datos a partir de la generación de los principales parámetros de centralización, dispersión y posición,
- ❑ compararemos los datos según la intensidad máxima del terremoto
- ❑ valoraremos las posibles asociaciones entre las variables profundidad y magnitud,
- ❑ y realizaremos predicciones sobre las variables presentadas y valoraremos la probabilidad de que haya un terremoto destructivo en función de la profundidad y la intensidad del mismo.

Todo ello con el fin de inferir sobre la probabilidad que hay de que acontezca un terremoto destructivo en la ciudad de Huelva durante el CEAM-19.

Reconocimientos: Trabajo realizado en el marco del Proyecto “Data Science Education in STEAM for Civic Engagement and Social Justice from Early Years” [DataScEd4CiEn]–2023-1-CY01-KA220-SCH-000164724, diciembre 2023-junio 2026.

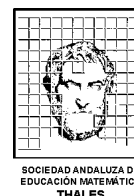
4. REFERENCIAS

- Breiman, L. (2001). Statistical modeling: The Two Cultures. *Statistical Science*, 16(3), 199-231.
- Donoho, D. (2017). 50 years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4). <https://doi.org/10.1080/10618600.2017.1384734>
- Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, 16(1), 44–49. <https://doi.org/10.52041/serj.v16i1.213>
- English, L. (. (2017). Manufacturing licorice: Modeling data in third grade. En J. Newton, y E. Galindo (Eds.), *Proceedings of the 39th Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA)*. Hossier Association of Mathematics Teacher Educators. <https://eprints.qut.edu.au/113850/>
- Erickson, T., Wilkerson, M., Finzer, W., y Reichsman, F. (2019). Data Moves. *Technology Innovations in Statistics Education*, 12(1). <https://doi.org/10.5070/T5121038001>
- Frischemeier, D., Biehler, R., Podworny, S., y Budde, L. (2021). A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data. *Teaching Statistics*, 43, 182-189. <https://doi.org/10.1111/test.12283>
- Lee, H. S., Mojica, G. M., Thrasher, E. P., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, 21(2), <https://doi.org/10.52041/serj.v21i2.41>
- Mojica, G. F., Barker, H., y Azmy, C. N. (2019). Instrumented learning in a CODAP-enabled learning



19 CONGRESO DE ENSEÑANZA Y APRENDIZAJE DE LAS MATEMÁTICAS.

Matemáticas y Multiculturalidad Aproximación a través de una matemática dinámica



environment. In J. M. Contreras, M. M. Gea, M. M. López-Martín, & E. Molina-Portillo (Eds.), *Proceedings of the III International Virtual Congress on Statistical Education*. <https://www.ugr.es/~fqm126/civeest/mojica.pdf>

Silva-Hormazábal, M., Rodríguez-Silva, J., y Alsina, A. (2022). Conectando matemáticas e ingeniería a través de la estadística: una actividad STEAM en educación primaria. *Revista Electrónica de Conocimientos, Saberes y Prácticas*. <https://doi.org/0.5377/recsp.v5i1.15118>

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, pp. 1-67.

Yan, D., y Davis, G. E. (2019). A First Course in Data Science. *Journal of Statistics Education*, 27(2), 99-109. <https://doi.org/10.1080/10691898.2019.1623136>