

Correlatividades: Introducción a la Estadística y Ciencia de Datos

Carácter: optativa

Carga horaria: 6 hs semanales - **48 hs totales**

Carreras: Licenciatura en Ciencia de Datos

Programa analítico del curso

Objetivos : El principal objetivo de este curso es que los alumnos y las alumnas se familiaricen con diferentes técnicas de modelado estadístico (no abarcadas por otras materias del área) y comprendan que la elección de la misma dependerá de la naturaleza del fenómeno bajo análisis, de los objetivos del estudio y de las particularidades del set de datos con el que se cuenta. Cada técnica será presentada mediante el planteo de un problema biológico, simulando una situación real en cuanto a preguntas de investigación y obtención de los datos. De esta manera, los conceptos teóricos involucrados detrás de cada análisis se entrelazan con la aplicación e interpretación de los resultados en contexto.

1- Conceptos básicos del diseño experimental: aleatorización, replicación, control del error. Distintos diseños experimentales. Bases de datos: adecuación, formato, limpieza y curado. Exploración de datos. Datos faltantes. Tratamientos posibles.

2- Métodos de validación cruzada. Aplicaciones.

3- Concepto de Robustez. Regresión robusta. Aplicaciones.

4- Métodos de regularización. Regresión Lasso. Aplicaciones.

5- Modelos aditivos generales y generalizados. Aplicaciones.

6- Métodos basados en árboles. Aplicaciones.

Bibliografía:

Alberti, J., & Massone, O. (2022). Tired of losing valuable data? Build your lab ecological database as a cornerstone for long-term approaches. *Ecología Austral*, 32(1), 151–157.

Donders, A. R. T., Van Der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10), 1087–1091.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *An Introduction to Statistical Learning*. Springer.

Holmes, S. H., & Huber, W. (2018). *Modern statistics for modern biology*. Cambridge University Press. (<http://web.stanford.edu/class/bios221/book/Chap-Supervised.html>)

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, 5(9), 961–970.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14.

Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, 7(6), 636–645.