Línea de Trabajo fin de Máster

(Fecha última actualización: 01/02/2021)

Máster en Estadística. CURSO ACADÉMICO 2020-21	
Título	Técnicas de minería de datos para mitigar el sesgo de selección en encuestas no probabilísticas.
Profesor(es)	María del Mar Rueda García Ramón Ferri
Descripción	Esta línea de trabajo fin de Máster se presenta como profundización del contenido del programa de Máster Oficial en Estadística Aplicada. Se trata en particular de profundizar en los métodos existentes para reducir el sesgo de selección en encuestas no probabilísticas y en el uso de técnicas de minería de datos para aumentar su eficiencia.
Objetivos particulares	- Introducción a los métodos para el tratamiento de sesgo de selección (calibración, Propensity Score Adjustment, Matching, estimadores basados en modelos). - Estudio de técnicas de minería de datos enfocadas en la selección de variables (Extracción Recursiva de Características, filtros de información mutua o basados en los resultados predictivos) y en los algoritmos de regresión y clasificación para predecir valores objetivo y probabilidades (k vecinos más cercanos, XGBoost, LASSO, redes neuronales). - Propuesta de una estrategia de estimación de parámetros poblacionales y aplicación a conjuntos de datos reales.
Prerrequisitos y recomendaciones	Se recomienda poseer un nivel intermedio de conocimiento en muestreo estadístico y en modelos de regresión.
Plan de trabajo	Se realizará un estudio inicial de los métodos disponibles para el tratamiento del sesgo de selección de cara a que el alumno/a tenga una idea de los métodos existentes. Se realizará una búsqueda de bibliografía sobre los algoritmos más relevantes utilizados para la selección de variables y la regresión y clasificación. Se estudiará una estrategia para seleccionar el mejor subconjunto de variables para un problema de predicción de forma que se optimice la reducción del sesgo de selección. El alumno deberá buscar conjuntos de datos reales sobre los que probar los diferentes modelos para hacer un estudio comparativo.
Competencias generales y específicas	CG1, CG2, CG3, CG4, CG10, CE2, CE8, CE15, CE16, CE27,CE29
Bibliografía	- Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis, 143, 106839. - Breidt, F. J., &Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. Statistical Science, 32(2), 190-205. - Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532), 2011-2021. - Deville, J. C., &Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382. - Ferri-García, R., & Rueda, M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS one, 15(4). - Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods & Research, 37(3), 319-343.