UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Propuesta TFG. Curso 2025/2026

GRADO: Grado en Estadística

CÓDIGO DEL TFG: 223-034-2025/2026

1. DATOS BÁSICOS DEL TFG:

Título: Integración de grandes volúmenes de datos no probabilísticos

Descripción general (resumen y metodología):

Los avances en las ciencias de la computación han permitido automatizar la recogida y almacenamiento de datos en muchas aplicaciones prácticas, ayudando a crear bases de datos de gran volumen de las que extraer conocimiento mediante diversos tipos de técnicas de minería de datos. Sin embargo, en muchas ocasiones estos grandes volúmenes de datos no cubren a la totalidad de la población objetivo que se desea estudiar en un cierto problema, por lo que estas bases de datos, a pesar de sus dimensiones, se podrían considerar una muestra de una población obtenida mediante un muestreo no probabilístico, ya que en la mayoría de casos la probabilidad de participar de las unidades que han quedado fuera de la base de datos es nula o bien la extracción de datos se ha realizado por conveniencia (p. ej. por accesibilidad a distintas APIs, provedoores, “data lakes”, etc.). En este trabajo se estudiarán las técnicas disponibles para mitigar los sesgos de selección que pudieran surgir en este tipo de situaciones, utilizando información auxiliar de muestras probabilísticas obtenidas de las mismas poblaciones o incluso censos que existan para algunas variables auxiliares. Estas técnicas se pondrán en práctica en un caso de estudio empleando datos reales de aplicaciones donde se generan datos a gran velocidad y de gran variedad.

Tipología: Estudio de casos, teóricos o prácticos, relacionados con la temática del Grado.

Objetivos planteados:

El objetivo principal de este trabajo es el estudio y la aplicación de las técnicas de integración de datos probabilísticos y no probabilísticos. Los objetivos específicos son los siguientes:
OE1. Estudio de la literatura referente a la integración de datos: tipos de muestreo (probabilístico y no probabilístico), sesgos en el muestreo no probabilístico, técnicas de integración de datos según disponibilidad de la variable de interés.
OE2. Estudio de la literatura referente a grandes volúmenes de datos: definición, características y métodos de almacenamiento y tratamiento de datos.
OE3. Aplicación de las técnicas de integración de datos en un problema real que involucre bases de datos de gran volumen.

Bibliografía básica:

- Yang, S., & Kim, J. K. (2020). Statistical data integration in survey sampling: A review. Japanese Journal of Statistics and Data Science, 3(2), 625-650.
- Chlebicki, P., Chrostowski, Ł., & Beręsewicz, M. (2024). Data integration of non-probability and probability samples with predictive mean matching. arXiv preprint arXiv:2403.13750.
- Beaumont, J. F., & Rao, J. N. K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies. The Survey Statistician, 83, 11-22.
- Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wiśniowski, A., & Struminskaya, B. (2024). Bayesian integration of probability and nonprobability samples for logistic regression. Journal of Survey Statistics and Methodology, 12(2), 458-492.
- Kim, J. K., & Tam, S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review, 89(2), 382-401.
- Chen, S., Yang, S., & Kim, J. K. (2022). Nonparametric mass imputation for data integration. Journal of survey statistics and methodology, 10(1), 1-24.

- Rueda, M. D. M., Pasadas-del-Amo, S., Rodríguez, B. C., Castro-Martín, L., & Ferri-García, R. (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. Biometrical Journal, 65(2), 2200035.

Recomendaciones y orientaciones para el estudiante:

Se recomienda que el estudiante haya cursado las materias Muestreo Estadístico, Diseño de Encuestas, Modelos Lineales y Minería de Datos del Grado en Estadística.

Plazas: 1

2. DATOS DEL TUTOR/A:

Nombre y apellidos: RAMÓN FERRI GARCÍA

Ámbito de conocimiento/Departamento: ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Correo electrónico: rferri@ugr.es

3. COTUTOR/A DE LA UGR (en su caso):

Nombre y apellidos: JORGE LUIS RUEDA SÁNCHEZ

Ámbito de conocimiento/Departamento: ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Correo electrónico: jorgerueda@ugr.es

4. COTUTOR/A EXTERNO/A (en su caso):

Nombre y apellidos:

Correo electrónico:

Nombre de la empresa o institución:

Dirección postal:

Puesto del tutor en la empresa o institución:

Centro de convenio Externo:

5. DATOS DEL ESTUDIANTE:

Nombre y apellidos: JUAN JOSE CECILLA MORALES

Correo electrónico: juanjoceci@correo.ugr.es