



UNIVERSIDAD NACIONAL DE LUJÁN

# Clasificación de flujos de datos continuos y multi etiquetados

Tesina de grado presentada para optar al título de  
Licenciado en Sistemas de Información

Juan Cruz Cardona

Director: Santiago Banchero

2021



# CLASIFICACIÓN DE FLUJOS DE DATOS CONTINUOS Y MULTI ETIQUETADOS

La clasificación multi-etiquetas es un paradigma de aprendizaje supervisado que generaliza las técnicas clásicas de clasificación para abordar problemas en donde cada instancia de una colección se encuentra asociada a múltiples etiquetas. La mayor parte de los trabajos de investigación han sido realizados en contextos de aprendizaje por *batch*. Los ambientes de flujo continuo de datos (o *streaming*) presentan nuevos desafíos a esta área debido a las limitaciones de tiempo de respuesta y almacenamiento que acarrearán. A esto se agrega la naturaleza evolutiva de este tipo de escenarios, que obligan a los algoritmos a adaptarse a cambios de concepto. En la presente investigación se aplican algoritmos de clasificación multi-etiquetas a colecciones estructuradas y no estructuradas. Los experimentos se llevarán a cabo en ambientes simulados de *streaming* de datos para conocer el impacto que produce este contexto sobre los resultados de la clasificación y acoplar el modelo a escenarios del mundo real. A su vez, se partirá de estas colecciones de datos para generar instancias sintéticas y así producir flujos potencialmente infinitos. Por último, se abordarán estrategias de ensambles de algoritmos en búsqueda de una mejora en la calidad de la tarea de predicción de objetos no observados por el modelo. De esta manera, se proveerá a la comunidad de nuevos estudios experimentales sobre algoritmos y colecciones ya conocidos del área de clasificación multi-etiquetas, de manera tal de extender el conocimiento sobre su rendimiento bajo escenarios evolutivos y de naturaleza variable.

**Palabras claves:** clasificación, multi-etiquetas, *streaming*, algoritmos, flujos.



## AGRADECIMIENTOS

Acá, agradezco...



## Índice general

1.. Introducción . . . . .	1
----------------------------	---





## 1. INTRODUCCIÓN

En los últimos años ha habido un aumento considerable de datos de diversa índole y generados por fuentes heterogéneas. Según los autores Gantz y Reinsel, el volumen total de datos creados y replicados en el mundo durante el año 2011 supera los 1.8 ZB (zettabytes) y se ha estimado que duplica cada dos años [1]. Los avances en el área de tecnología de la información (IT) han contribuido a una continua producción de datos y expansión del campo digital, tal es el caso para la red social *Facebook*, la cual recibe cada hora un flujo de 10 millones de fotos que publican sus usuarios [2].



## BIBLIOGRAFÍA

- [1] J Gantz y D Reinsel. «Extracting value from chaos». En: *IDC IView* (2011), págs. 1-12.
- [2] V Mayer-Schonberger y K Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. An Eamon Dolan book. Houghton Mifflin Harcourt, 2013.