Propuesta de técnicas de validación para la calidad de datos abiertos e identificación de patrones para predicciones con Machine Learning

Roxana Martínez, Christian Parkinson, Martín Caruso, Diego López,
Rocío Vargas, Nayiby Rojas

CAETI - Centro de Altos Estudios en Tecnología Informática
Universidad Abierta Interamericana (UAI)
Montes de Oca 745, Ciudad Autónoma de Buenos Aires, Argentina

{roxana.martinez, christian.parkinson, rocio.vargas}@uai.edu.ar

{martin.caruso, diego.lopez, nayibi.rojas}@alumnos.uai.edu.ar

RESUMEN

La política de los datos abiertos busca promover la innovación y transformar la actividad gubernamental para brindar mejores servicios y generar mayores niveles de transparencia en la sociedad. Por lo que, mantener la calidad en las fuentes de datos disponibilizadas es fundamental para su tratamiento y obtener así, un conocimiento de éstas. Actualmente, son pocos los trabajos realizados en aspectos de validaciones, análisis de contenidos internos de estos datasets, herramientas de datos, identificación de patrones en su estructura y demás. En base a esto, esta línea de investigación se enfoca en el análisis, diseño y desarrollo de herramientas de software que utilicen técnicas y propuestas para la validación de la calidad de los datos públicos abiertos en el contexto de Gobierno Abierto. Además de detectar el "estado de salud" de estos datos (grado de integridad, redundancia y otros), se trabaja en el estudio de patrones con el fin de incorporar modelos de predicción para aportar un valor agregado a la información otorgada por los gobiernos desde aspectos de Machine Learning, lo que ofrecerá un mejor análisis para la toma de decisiones gubernamentales.

Palabras clave: Datos Abiertos, Gobierno Abierto, Métricas de Calidad de Datos, Machine Learning, Predicciones a partir de Patrones.

CONTEXTO

El presente trabajo es parte del proyecto denominado "Investigación y desarrollo de software para la validación de la calidad de datos abiertos e identificación de patrones para predicciones", que tiene inicio en el mes de marzo 2022. Este proyecto pertenece a la línea de investigación de Ingeniería de Software del Centro de Altos Estudios en Tecnología Informática (CAETI) de la Facultad de Tecnología Informática de la Universidad Abierta Interamericana (UAI), el cual contribuye al desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC's) en Argentina y en el mundo, llevando adelante la investigación básica y aplicada en diversas áreas. El proyecto es financiado y evaluado por la Secretaría de Investigación de la Universidad, tiene una duración de 2 años, y cuenta con la participación de docentes y estudiantes de grado y posgrado en diversas carreras de la Facultad de Tecnología Informática.

1. INTRODUCCIÓN

El Gobierno Abierto se enfoca en la "construcción de estados transparentes, participativos, que rindan cuentas de manera adecuada e innovadores, poniendo al ciudadano en el centro de la toma de las decisiones públicas, como una forma de fortalecer el Estado democrático" [1]. Este contexto implica la utilización de diversas tecnologías abiertas con el fin de fomentar la

Innovación Pública. Por otra parte, se sostiene que este término "es una doctrina política que surge a partir de la adopción de la filosofía del movimiento del software libre a los principios de la democracia. Este paradigma tiene como objetivo que la ciudadanía colabore en la creación y mejora de servicios públicos y en el robustecimiento de la transparencia y la rendición de cuentas" [2]. Por lo que el Estado Nacional debe incorporar las técnicas para gestionar dicha cantidad de datos con un diseño tecnológico y por sobre todo enfocado en el ciudadano promedio para fomentar la inclusión social. "Al ampliar el acceso a la información pública se fortalece la rendición de cuentas y se enriquece el debate público, a la vez que se crean nuevas oportunidades para generar valor agregado" [3]. El enfoque de formato abierto es el formato de archivo no propietario, cuya especificación debe estar documentada públicamente, es de libre conocimiento e implementación y libre de patentes o de cualquier otra restricción legal o económica para su uso. Para ello, es necesaria una licencia abierta, que es un acuerdo de provisión de datos para que cualquier persona los utilice, los reutilice y los distribuya, estando sujeto a las condiciones de dicha licencia.

"Los datos que se pueden reutilizar y redistribuir sin ninguna restricción se denominan datos abiertos" [4]. Estos datos deben estar en formatos digitales, con un modelo estándar abierto. Por otra parte, es importante destacar que, "no toda la información pública disponible o publicada en la web es información abierta válida para su reutilización. No sólo se trata de publicar los datos, sino que hay que garantizar el acceso a ellos, razón por la que debe recurrirse a formatos digitales, estandarizados y abiertos, siguiendo una estructura clara que permita su comprensión y reutilización" [5]. La gestión de la información en formatos abiertos, "datos abiertos", consiste en el acceso y uso de la información pública por parte de terceros para entregar nuevos servicios a los ciudadanos, esto permite acceder a una gran cantidad de datos procedentes de diferentes organizaciones del ámbito de la administración pública [6].

Por lo anteriormente explicado, es fundamental que personal dedicado trabaje en el tratamiento de las fuentes de datos abiertos, para que los organismos estatales y ciudadanos, tengan un mejor conocimiento sobre un determinado tema público, como ser: economía, transporte, etc.

Magallón afirma que "la cultura de datos abiertos trata de obtener un valor añadido de la información. A diferencia de lo ocurrido hasta ahora esta información no genera sólo su valor por estar reservada a unos pocos, sino que lo hace por su disponibilidad para ser interpretada y traducida por cualquier actor interesado en trabajar con ella" [7], es decir, un dato abierto es aquel que puede ser accedido, y conlleva un formato que permite la interoperabilidad con otros softwares. Diversos organismos estatales que ofrecen una gran cantidad de fuentes de datos de varios temas gubernamentales con criterios preestablecidos en sus portales, que brindan datasets que son utilizados como insumo fundamental de información y servicios.

Existen algunos trabajos enfocados a las mediciones en aspectos de calidad [8] [9] [10], por ejemplo, el Barómetro de Datos Abiertos, ODB [11] de la World Wide Web Foundation, es una medida global de cómo los gobiernos publican y utilizan los datos abiertos para la rendición de cuentas, la innovación y el impacto social. Otras mediciones fueron desarrolladas por The Global Open Data Index [12] que es el punto de referencia mundial anual para la publicación de datos gubernamentales abiertos, gestionado por la Open Knowledge Foundation [13], funciona como una encuesta de crowdsourcing, que mide la apertura de los datos gubernamentales a través de la metodología GODI [14]. Otro proyecto es el Open Government Data de la Organisation for Economic Co-operation and Development, OECD [15], su objetivo es avanzar en la evaluación de impacto de Open Government Data (OGD), para este caso, su índice evalúa los esfuerzos de los gobiernos para implementar datos abiertos en las tres áreas críticas: apertura, utilidad y reutilización de

los datos gubernamentales; otras medidas son las puntuaciones otorgadas a la calidad de los datos abiertos, este es el caso del esquema modelo de cinco estrellas de Berners-Lee [16], o informes sobre el estado de los datos elaborados por fundaciones comprometidas con el impacto de estos datos públicos [17].

Gracias al gran impulso de la tecnología en el área de datos, el concepto de Machine Learning cada vez se encuentra más en auge, "el aprendizaje automático se refiere al proceso por el cual los ordenadores desarrollan el reconocimiento de patrones o la capacidad de aprender continuamente y hacer predicciones basadas en datos tras lo cual realizan ajustes sin haber sido programados específicamente para ello. Como forma de inteligencia artificial, el aprendizaje automático automatiza el proceso de creación de modelos analíticos y permite que las máquinas se adapten a nuevas situaciones de manera independiente" [18]. Este paradigma, tiene implicaciones para el descubrimiento científico, ya que la complejidad de los patrones que las máquinas son capaces de identificar no es fácilmente lograda por los procesos cognitivos humanos [19].

Machine Learning apunta a la utilización de algoritmos que analizan datos y, a partir de éstos, logran determinar el comportamiento del software. Para lograr un correcto análisis, son necesarios algoritmos que sean alimentados por datos que sostengan estos sistemas automatizados, es decir, una mayor disponibilidad y calidad de los datos abiertos servirá para alimentar esos algoritmos y a la vez poder también mejorarlos y auditar su correcto funcionamiento [20]. "Uno de los puntos importantes para tener en cuenta en el proceso de Machine Learning es el preprocesado y preparación de las variables que componen los conjuntos de entrenamiento y test de algoritmos, ya que será una condicionante esencial" [21] para un buen análisis.

En lo relativo al dato, es necesario contar con técnicas de validación de calidad, debido a que hoy por hoy, son pocos los trabajos realizados en aspectos de enfoque [22], por lo que, contar con las diversas técnicas de Machine Learning y que éstas puedan ser utilizadas en datos públicos, podría generar un gran beneficio a la sociedad, y, por otro lado, mantener la calidad y la apertura de los datos públicos, ayudará a los gobiernos y a los diferentes actores de la sociedad civil a tomar mejores decisiones, ya que tienen una visión e información de la realidad más precisa.

Uno de los problemas actuales es que, en los portales de datos abiertos, la disponibilidad de los datos no necesariamente coincide con que tengan calidad, lamentablemente, hoy sigue siendo una dificultad y es un gran desafío para las políticas públicas. El análisis de muchos de los conjuntos de datos públicos representa un problema crucial, ya que está disperso, no estandarizado y en muchos casos desactualizado.

2. LÍNEAS DE INVESTIGACIÓN DESARROLLO

Este proyecto pertenece a la línea de investigación de Ingeniería de Software del Centro de Altos Estudios en Tecnología Informática (CAETI). Los ejes principales del tema que se está investigando son:

- Analizar las falencias actuales en cuestiones de calidad de datos abiertos públicos.
- Identificar las mejores técnicas de validación de calidad de datos abiertos disponibilizados.
- Diseñar y desarrollar algoritmos para identificar patrones en datos públicos.
- Analizar modelos predictivos orientados a los datos abiertos.
- Efectuar predicciones sobre nuevos datos encontrados con técnicas de Machine Learning en datos públicos.

3. RESULTADOS OBTENIDOS/ESPERADOS

Enfoque del proyecto:

Los datasets abiertos gubernamentales en contexto de gobierno abierto, tienen falencias en aspectos de calidad de datos, éstos pueden ser detectados y analizados a través de

herramientas de validación y medición para mejorar su interoperabilidad a nivel software, como así también, pueden utilizar algoritmos para identificar patrones y realizar predicciones para lograr beneficios orientados a la comunidad social.

Objetivos principales:
Realizar el análisis, diseño y desarrollo de herramientas de software para la gestión y validación de la calidad de los datos públicos en el contexto de Gobierno Abierto. Detectando el "estado de salud" de las diversas fuentes de datos provenientes de casos de aplicación gubernamentales con los prototipos desarrollados, incorporando algoritmos para identificar patrones que logren predicciones sobre nuevos datos.

Objetivos específicos esperados:
Estudiar el alcance de los conceptos implicados en este contexto de estudio; Analizar los trabajos relacionados en la temática de Gobierno Abierto relativos al tratamiento de datos abiertos y públicos; Analizar las técnicas de Machine Learning para datos abiertos; Relevar los últimos trabajos relacionados en cuanto a los aspectos de calidad de datos abiertos y públicos; Relevar trabajos enfocados a Técnicas de Machine Learning con datos abiertos; Analizar los criterios de los portales de datos abiertos para efectuar la publicación de éstos; Relevar los distintos tipos y formatos disponibles de datos abiertos que existen en los dataset actuales de los sitios más relevantes y gubernamentales de Argentina; Analizar las falencias de los datasets disponibles en los sitios de Gobierno; Analizar herramientas de manejo con Machine Learning que permitan análisis de patrones con enfoque predictivo; Establecer criterios estándar de calidad de datos; Desarrollar una herramienta de validación de datasets gubernamentales basada en métricas de calidad de datos; Generar una guía de buenas prácticas para las técnicas de Machine Learning en la utilización de datos abiertos; Definir y recolectar la muestra de datasets para ser testeados por la herramienta propuesta como validadora de calidad; Análisis de casos predictivos con casos reales; Análisis de casos de aplicación con los prototipos desarrollados.

Metodología y Técnicas:
El proceso metodológico de investigación empleado para el presente proyecto de investigación se define con un proceso sistemático cualitativo, que implementa una forma evolutiva incremental en cada una de las etapas involucradas, siendo estas: identificación del problema; revisión teórica; recolección de datos, clasificación y análisis de datos; estudio de escenarios de identificación de patrones sobre éstos con enfoques predictivos sobre nuevos datos; desarrollo de prototipos (análisis cuantitativo para cada una de las métricas de calidad de datos); visualización de la predicción de patrones; validación de la solución; identificación de las limitaciones del trabajo.

4. FORMACIÓN DE RECURSOS HUMANOS

Este proyecto se compone por 3 (tres) docentes con estudios de posgrado: uno de ellos Magíster en Tecnología Informática y a la espera de respuesta por parte del jurado que se encuentra en revisión de su tesis doctoral en Ciencias Informáticas en la Universidad Nacional de La Plata (UNLP), y dos docentes que se encuentran realizando su tesis de maestría en Tecnología Informática en la Universidad Abierta Interamericana (UAI). El equipo también cuenta con la participación de estudiantes de grado y de posgrado de la UAI.
En relación directa con la línea de I/D presentada para el proyecto, los miembros del equipo se encuentran en realización de: 1 tesis doctoral (docente), 2 tesis de maestría (un docente-estudiante y un estudiante) y 2 tesinas de grado (estudiantes) en la UAI.

5. BIBLIOGRAFÍA

[1] Arroyo Chacón, J. (2017). La Innovación Abierta Como Pilar Del Gobierno Abierto

(Open Innovation as a Pillar of Open Government). Revista Enfoques, 15(27), 13-41.

[2] Sosteniblepedia.org. "*Gobierno Abierto*". Disponible en:

https://www.sosteniblepedia.org/index.php?title=Gobierno_abierto

[3] Buenos Aires provincia (2017). "Kit de Apertura Municipal". Disponible en: http://escueladefiscales.com/Kit%20de%20Apertura%20Municipal%202017%20-%20provincia%20de%20buenos%20aires.pdf

[4] Oviedo, E., Mazón, J. N., & Zubcoff, J. J. (2013). Hacia un modelo de calidad de datos para portales de datos abiertos. In XXXIX Latin American Computing Conference (CLEI), Naiguata (pp. 1-8).

[5] Garriga-Portolà, M. (2011). ¿Datos abiertos? Sí, pero de forma sostenible. Profesional de la Información, 20(3), 298-303.

[6] Naser, A., & Ramírez Alujas, Á. (2017). Plan de gobierno abierto: una hoja de ruta para los gobiernos de la región.

[7] Magallón Rosa, R. (2017). Datos abiertos y acceso a la información pública en la reconstrucción de la historia digital.

[8] ISO 25012 (2008). "*Ingeniería de software - Requisitos de calidad y evaluación de productos de software (SQuaRE) - Modelo de calidad de datos*". Disponible en: https://www.iso.org/obp/ui/es/#iso:std:iso-iec:25012:ed-1:v1:en

[9] Martínez, R. et al. (2021). Metrics proposal to measure the quality of governmental datasets. IEEE Latin America Transactions, Vol. 100. ISSN 1548-0992.

[10] de España, G. (2017). Manual práctico para mejorar la calidad de los datos abiertos. Madrid. Disponible: en:

https://datos.gob.es/sites/default/files/doc/file/manual_practico_para_mejorar_la_calidad_de_los_datos_abiertos_1.pdf

[11] Open Data Barometer. "*The Open Data Barometer*". Disponible es:

https://opendatabarometer.org/?_year=2017&indicator=ODB

[12] Open Knowledge Foundation (2020). "*Global Open Data Index - Argentina*". Disponible en: https://index.okfn.org/place/ar/

[13] Open Knowledge Foundation (2020). "*A fair, free and open future*". Disponible en: https://okfn.org/

[14] Open Knowledge Foundation Network (2017). "*Methodology - Global Open Data Index*". Disponible en:

https://index.okfn.org/methodology/

[15] OECD Better policities for better lives. "*Open Government Data*". Disponible en: http://www.oecd.org/internet/digital-government/open-government-data.htm

[16] Open Data (2012). "*5 Open Data*". Disponible en: https://5stardata.info/en/

[17] ODI Open Data Institute. "*The 2019 Data Skills Framework*". Disponible en: https://theodi.org/article/open-data-skills-framework/

[18] Herrera Carrasco, J. (2020). Evaluación de Modelos de Transporte mediante datos abiertos y técnicas de Aprendizaje Automático.

[19] Boulton, G., Hodson, S., Babini, D., Li, J., Marwala, T., Musoke, M. G., … & Wyatt, S. (2017). Datos abiertos en un mundo de grandes datos: Un acuerdo internacional ICSU-IAP-ISSC-TWAS. Revista iberoamericana de ciencia tecnología y sociedad, 12(34), 267-272.

[20] Datos.gob.es (2017). "*El futuro de los datos abiertos y sus múltiples caras*". Disponible en: https://datos.gob.es/es/noticia/el-futuro-de-los-datos-abiertos-y-sus-multiples-caras

[21] Gómez, C. E. J., & Roma, J. C. (2018). Análisis predictivo de datos abiertos sobre el uso turístico del servicio de alquiler compartido de bicicletas de Nueva York. Universidad Oberta de Catalunya, Master Universitario en Ciencia de Datos.

[22] Kumar, V. D., & Alencar, P. (2016, December). Software engineering for big data projects: Domains, methodologies and gaps. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 2886-2895). IEEE.