Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Una Visión General de los Datos Abiertos y su Análisis con Aprendizaje Automático para Ciudades Inteligentes

An Overview of Open Data and its Analysis with Machine Learning for Smart Cities

Estrada-Guzmán, E.[1]*, Martínez-Vargas, M.P.[2], Rodríguez-Pimienta, C.G.[1], Peña-Pérez Negrón, A.[1]

[1] Centro Universitario de Ciencias Exactas e Ingenierías; 44430. Universidad de Guadalajara, Guadalajara, Jalisco.
[2] Centro Universitario de Ciencias Económico Administrativas; 45100. Universidad de Guadalajara, Guadalajara, Jalisco.
elsa.estrada@academicos.udg.mx*; martha.mvargas@academicos.udg.mx; cristina.rodriguez@academicos.udg.mx; adriana.pena@cucei.udg.mx

Innovación tecnológica: Se presenta un método de evaluación de la calidad de Open Data sets en relación con las dimensiones Smart City.

Área de aplicación industrial: Gobierno, Empresa Privada, Medioambiente, Educación, Salud, Económica y Movilidad.

Recibido: 16 junio 2023
Aceptado: 14 noviembre 2023

Abstract

Much has been said about the transparency of the information that public organizations must provide, but little about its possible use and quality. Open Data resources have been increasing and promoted by Smart Cities to improve the quality of life, due to its potential for decision-making on its various subdimensions. The objective of this work is to highlight the importance of quality of Open Data, and the applicability to data science for the evaluation of prediction models in Smart Cities. For this purpose, the methodology applied consists of four phases: “open data review on websites”, “selection, downloading and classification”, “preprocessing”, and “data science treatment”. The results showed that out of 100 cases after the cleaning process the volume of records was reduced by 40% on average. Also, the smart government open data sets require greater preprocessing effort, and only a few provide dictionaries to interpret their contents; therefore, low quality is evident. After applying the data science treatment to six representative cases, the results showed that only in one case of smart living, the sensitivity measurement was 93%, and in the rest

of the cases were less of 80%, which shows the absence of conclusive patterns in the remaining 5 cases that means that an applicability of some methods of data science is not guaranteed.

Key Words: Data Science, Machine Learning, Open Data, Smart Cities, Supervised Learning.

Resumen

Mucho se ha dicho sobre la transparencia de la información que deben proporcionar los organismos públicos, pero poco sobre su posible uso y calidad. Los recursos de Open Data han ido en aumento y promovidos por las Smart Cities para mejorar la calidad de vida, debido a su potencial para la toma de decisiones en sus diversos subdimensiones. El objetivo de este trabajo es resaltar la importancia de la calidad del Open Data, y la aplicabilidad de la ciencia de datos para la evaluación de modelos de predicción en las Smart Cities. Para ello, la metodología aplicada consta de cuatro fases: "revisión de open data en sitios web", "selección, descarga y clasificación", "preprocesamiento", y "tratamiento data science". Los resultados mostraron que de 100 casos después del proceso de limpieza, el volumen de registros se redujo en un 40% en promedio. Además, los conjuntos de datos abiertos de smart government requieren un mayor esfuerzo de preprocesamiento y solo unos pocos proporcionan diccionarios para interpretar su contenido; por lo tanto, se evidencia una baja calidad. Después de aplicar el tratamiento de data science para seis casos representativos, los resultados mostraron que solo en un caso de smart living, la medición de la sensibilidad fue de 93%, y en el resto de los casos fue menor de 80%; demostrando la ausencia de patrones concluyentes en los 5 casos restantes lo que significa que no se garantiza una aplicabilidad de ciertos métodos con la ciencia de datos.

Palabras Clave: Ciencia de Datos, Aprendizaje Máquina, Datos Abiertos, Ciudades Inteligentes, Aprendizaje Supervisado.

1. Introducción

El término Open Data se aplica a datos legibles por una computadora, libremente compartidos, usados y construidos sin restricciones para su utilización [1], pueden ser no estructurados, pero se recomienda que sean estructurados para facilitar el acceso universal por el manejo de los formatos. Estos datos son una herramienta útil para diversas áreas de la sociedad en los sectores: público, privado, o comercial, entre otros [2]. En el Open Data, se destaca la utilización del Machine Learning (ML) y otros algoritmos que potencializan los datos [3]. Es por esto que gobiernos de diversas partes del mundo están optando por una cultura abierta que promueva la transparencia, participación y colaboración de todos los ciudadanos [4]. Actualmente, existen múltiples aplicaciones relacionadas con la vida en la ciudad que involucran el Open Data para resolver problemas personales o de comunidades, planteando sugerencias o brindando información, por ejemplo, tráfico, estacionamiento, clima, dónde comer, etc. Algunas funcionan como aplicaciones de abastecimiento colectivo que recopilan datos de sus usuarios de forma voluntaria, por ejemplo, informar de un evento en una ubicación específica que luego reside en una base de datos, otras de modo más sofisticado proveen toda una infraestructura. En el 2015,

ya se reconocía una aplicación llamada "PulsePoint" [5] que alertaba a las personas de un paro cardiaco repentino en su vecindario, posteriormente en el año 2019 surgió el concepto "open data soft" [6], que comprende herramientas y una plataforma para las organizaciones que deseen crear portales de Open Data de forma rápida, y a la vez recopilar y procesar sus datos.

Requerimientos de calidad

Las expectativas de calidad para estos esquemas son muy altas, ya que se espera información verídica, completa y significativa, dada la relevancia e impacto que pueda tener en la toma de decisiones. En un mapeo sistemático de exploración del Open Data realizado a principios del año 2021 [7], en el que se revisaron 893 artículos publicados entre 2006 y 2019, se identificaron temas sobre limitantes de reúso de datos por problemas de calidad, su falta de acceso al público por necesidad de tecnologías para su publicación y con ello su poca explotación, así mismo se hizo latente la falta de legislación y estandarización para su entendimiento universal.

Requerimientos de análisis

Se ha demostrado también la necesidad de formación de los profesionales en la analítica para el caso de grandes volúmenes de datos, y su explotación, resaltando la importancia de incluir cuestiones legales para información clínica de pacientes [8]. Por otro lado, la necesidad de análisis es vital. En lo que se refiere al gobierno transparente el análisis predictivo es una herramienta de apoyo, en el trabajo [9] se usaron datos abiertos de impuestos para demostrar que con ML es posible predecir los impuestos y la categoría.

El propósito de este trabajo es ofrecer un panorama a los analistas de la industria e investigadores, sobre la calidad del Open Data, sus temas principales, desde la perspectiva de las dimensiones de las Smart Cities, y su viabilidad de extracción de patrones para su posible explotación. Para evaluar la calidad de los datos se aplicó un método que abarcó 4 fases, que fueron la visita a sitios web, descarga, selección y clasificación, preprocesamiento, y el tratamiento del data science. Se aplicó el aprendizaje supervisado a una muestra de 6 casos, y se utilizó como estimador de veracidad de la sensibilidad estadística.

Este documento está estructurado de la siguiente forma, la sección II presenta la relación entre Open Data y Smart Cities, la sección III Antecedentes, la sección IV el Método, la sección V muestra los Resultados, la sección VI Discusión de hallazgos, VII Conclusiones y trabajos futuros, finalmente la sección VIII los Apéndices.

2. Open Data y Smart Cities

El Open Data ha contribuido al desarrollo de las ciudades proporcionando los esquemas de servicios inteligentes, producto de la aplicación de las Smart Cities [10], y de las tecnologías de la información y la comunicación. A su vez, las Smart Cities han dado empuje al Open Data porque proporcionan el enfoque y objetivo, que es el garantizar el desarrollo sustentable, para una distribución equitativa de recursos con la participación ciudadana entre comunidad y el gobierno. Del impacto de este fenómeno aparecen las dimensiones en las que se buscan soluciones a temas de salud, educación, movilidad, medio ambiente, ciudadanía, gobernanza y economía.

En [11] se hace una revisión para verificar el uso de smart data y Open Data para soluciones Smart Cities. Se destaca que el conocimiento popular puede contribuir a la generación de datos, posibilitar el producir y usar datos públicos para propósitos privados,

y producir iniciativas para resolver problemas de diferentes dominios de las Smart Cities.

3. Antecedentes

Algunos trabajos previos han presentado propuestas de estructuras para la evaluación y monitoreo de las Smart Cities usando como referencia el Open Data y el ML, cuyo objetivo es medir el desarrollo de las Smart Cities, pero no al Open Data, [12],[13],[14],[15]. Otros trabajos relacionados son revisiones sistemáticas de la literatura para identificar técnicas ML para Smart Cities [16], [17], [18], pero en su mayoría el big data es el tema central. Referente a la calidad de Open Data existen varios trabajos, pero no tienen alcance al ámbito de las Smart Cities [19], [20], [21].

La obtención de patrones a través de la observación de fenómenos usando la ciencia de datos ha sido una de las estrategias clave para la toma de decisiones en las Smart Cities, el Open Data provee a la vez de una guía hacia ese objetivo. En [22] se describe el típico ciclo del Open Data que consiste en la *"Recolección de datos","publicación","análisis","interpretación" y "uso";* su finalidad es hacer inferencias acerca de la realidad y enfatizar la importancia de la validación del análisis para buscar fenómenos e interpretarlos. Este marco inicia por una situación existente, que debe ser analizada, cuyas conclusiones muestran acciones que deben encaminar a cambios masivos.

Desde la perspectiva de [23], el Open Data se refiere a la necesidad de añadir un valor más allá de publicar solo data sets, por lo que este recomienda un proceso de tres capas: *"infraestructura", "métodos" y "análisis".* La infraestructura se trata de software para que los datos estén disponibles, los métodos son paquetes o librerías como CRAN Python CONDA que contienen los modelos para implementar las técnicas de análisis, y en el análisis se lleva a cabo la transformación de los datos originales a datos con valor, para hacer toma de decisiones.

En ambos enfoques juegan un rol esencial el análisis y la interpretación para la toma de decisiones. Este trabajo, se centra en el análisis y su interpretación con datos que ya han sido recolectados y publicados en sitios abiertos. En cuanto a la infraestructura y la publicación se revisa la disponibilidad de los archivos, los formatos y contenidos, y se evalúa su completez para seleccionar solo aquellos que contienen por ejemplo 2 o más variables.

4. Método

Tomando como referencia el ciclo del Open Data, el proceso de tres capas descritos anteriormente y las dimensiones Smart Cities, el método empleado para esta investigación consta de las siguientes fases: *"exploración", "selección, descarga y clasificación", "preprocesamiento" y "tratamiento de las smart cities",* ver figura 1.

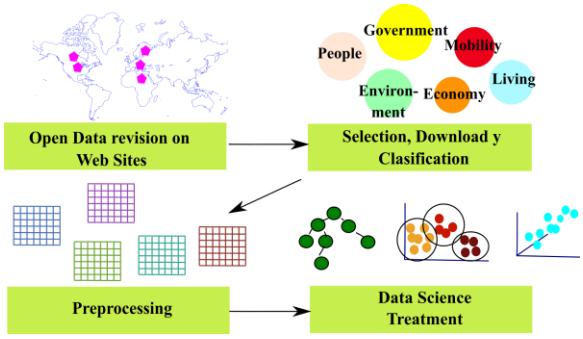


Figura 1. Método evaluación de calidad para el Open Data.

4.1. Fase revisión de open data en sitios web

En esta fase se llevó a cabo una exploración de open data sets (ODS) en diversos sitios web, y luego se revisó literatura sobre la usabilidad del Open Data en el contexto de Open government, para identificar los países con más interacción en plataformas de

participación ciudadana. Se destaca especialmente el estudio [24] que analiza la usabilidad centrada en el usuario en la exploración de 41 portales. Describe un desglose de la solicitud de conjunto de datos abiertos por país, mide la participación e involucramiento de los usuarios con los ODS, e identifica tres grupos. El primero, con países de más bajo promedio de calificación de solicitudes de acceso a la información, el segundo grupo con un nivel mediano y el grupo tres con más alto promedio en el proceso de involucramiento del usuario. De dichos grupos, se tomaron los países representativos (Dinamarca, Austria, Canadá, China, Francia, Reino Unido y USA) como referencia para este trabajo de exploración.

4.2. Fase selección, descarga y clasificación

En la fase *selección, descarga y clasificación* se seleccionaron diversos ODS de los portales de los países representativos, ver Apéndice A. Se generó un registro de cada uno de los ODS con la información según su contenido y se clasificaron de acuerdo a las dimensiones Smart City. El formato de registro se muestra en la tabla 1.

Tabla 1. Variables de control para ODS descargados.

País
Año
Continente
Total de registros
Número de registros vacíos, incompletos o inválidos
Dimensión Smart City

En esta fase, las tareas fueron iterativas, ya que no todos los archivos descargados cumplían con los criterios del Open Data. Los criterios propios de la evaluación a ODS según [25] son los siguientes:

- Usabilidad - Estructura Open Data para una fácil reutilización.
- Accesibilidad - Los datos pueden ser accesibles para cualquier persona y para cualquier propósito).
- Comprensibilidad - La descripción de los datos es fácil de entender para el público.
- Exactitud - Los datos son correctos.
- Perdurable - Actualización oportuna de datos.

Esta inspección sirvió también para identificar la completitud, se revisó la cantidad de datos ausentes ya que un gran número de valores faltantes afectaría al análisis o éste sería parcialmente erróneo. En esta actividad también se verificó que los datos fueran entendibles y correctos para aplicar el análisis fácilmente, y se tomó en cuenta la fecha a la que pertenecían para hacer pruebas con hechos relacionados en su tiempo de obtención.

4.3. Fase preprocesamiento

La fase de preprocesamiento hace referencia a la preparación de los datos ya que estos serán empleados por los métodos del data science en la siguiente fase. Si no se depuran los resultados no serán confiables o bien fallará la ejecución del algoritmo. Es decir, en este paso se asegura la calidad de los datos. Ya que las fuentes emisoras de los datos pueden fallar, es posible que la información generada sea parcial o totalmente inválida. Por lo tanto, con la finalidad de construir sistemas de alta exactitud en el área de la data science, se deben realizar tareas correspondientes a la detección de anomalías. En esta fase, se realizó la detección de anomalías e inconsistencias y se hizo una evaluación de calidad de los ODS, la detección se hizo según los lineamientos de las tareas de preprocesamiento de [26]. Las actividades fueron las siguientes:

a) Estandarización y normalización.
b) Removida de registros duplicados, tratamiento de valores faltantes,

tratamiento de valores fuera de rango usando valores normales en los diccionarios de los data sets.

c) Codificación de características categóricas (En su lugar se ubicaron las variables categóricas incluyendo las numéricas candidatas a ser variables dependientes).
d) Selección y reducción de variables, en la que se seleccionaron las variables independientes y su dependiente para cada inferencia construida.

4.4. Fase tratamiento del data science

Ya que el objetivo de este trabajo es la aplicación del data science para recuperar patrones de predicción y evaluar su ajuste a modelos ML, esta fase se dividió en dos grupos de tareas, las actividades para la construcción de las hipótesis y las actividades dedicadas a la investigación de los métodos ML para probar las hipótesis, y la aplicación del método.

Construcción de inferencias

Con la finalidad de elegir las variables que pueden probar la existencia de un patrón, se plantean las hipótesis, en la que al menos dos variables tienen relación de dependencia. Pero esta dependencia no siempre ocurre, por eso es solo un modelo de prueba. Representan también funciones, donde se mapea una entrada (una o varias variables independientes) a una salida (variable dependiente).

El aprendizaje supervisado es una tarea de ML que usa etiquetas (variable dependiente) para entrenar a los datos, y usa también una colección de ejemplos de entrenamiento para inferir una función. Por lo tanto, las hipótesis son necesarias para saber qué modelos deben ser entrenados.

En la tabla 2, se muestra el formato de la selección de variables y su inferencia. En este formato se especificó para cada ODS, su nombre, tal como se identificó en la fuente de origen, el nombre de las variables dependiente e independientes escritas conforme al estándar de nombramiento de variables, las inferencias o hipótesis definidas según la expresión genérica de una función de relación, esto es, a la izquierda la variable dependiente, a la derecha las independientes y estas separadas por una tilde; ya que es una convención usual en la llamada a las funciones de ML de aprendizaje supervisado, la descripción complementa la frase de la hipótesis para comprender mejor el patrón a demostrarse en la prueba del análisis.

Tabla 2. Esquema de selección de variables y construcción de inferencias en ODS.

Nombre del ODS
Nombre de variable dependiente
Nombre de variable(s) independiente(s)
Inferencia (Realidad a fenómeno a comprobar o hipótesis) f(x)~x
Descripción del supuesto o hipótesis ML

Búsqueda de métodos ML para dimensiones Smart Cities

Dado que la eficiencia de una solución de ML depende no solo de la naturaleza y características de los datos sino también del desempeño de los algoritmos de aprendizaje [26]. Se realizó una revisión de estudio de efectividad de las diferentes técnicas para seleccionar la técnica apropiada para cada caso. Esta actividad, consistió en identificar las técnicas ML para las seis distintas dimensiones smart city, que son medio ambiente (smart environment), formas de vida (smart living), economía (smart economy), movilidad (smart mobility), gobierno (smart government) y sociedad (smart people). Se construyeron cadenas de búsqueda formadas con la frase “data science” combinada a cada una de esas dimensiones, y en algunos casos ligando la cadena a un key performance indicator (kpi), que está comprendido en alguna dimensión

según la rueda de *Cohen* [27]. Los kpi's han sido utilizados para evaluar el desarrollo de las Smart Cities, estos son puntos de referencia para medir el bienestar y la calidad de vida de los ciudadanos, con estos se mide el desarrollo de una población. En la tabla 3, se resumen los métodos de predicción encontrados. Por cada dimensión Smart City, se muestran los kpi's, que son sujetos de análisis en las revisiones estudiadas (variables o métricas).

Tabla 3. Métodos utilizados en las dimensiones de las Smart Cities.

Smart City dimension	Variables o métricas	Métodos de predicción identificados	Referencia
Smart environment	Aire, agua, suelos. Toxicidad.	Regresión. Modelos de clasificación.	[28]
Smart living	Manejo de recursos en edificios y casas.	Máquina de vectores de soporte. Red neuronal artificial. Bosques aleatorios.	[29]
	COVID-19.	Bosques aleatorios. Polinomios fraccionarios multiples.	[30]
Smart economy	Identificación de comportamientos, gustos o hábitos que identifican a los consumidores. Productos con patrones inusuales o poco comunes que podrían ser fraudulentos, como en los sectores financiero o de seguros.	Modelos de agrupación. Modelos ensamblados, redes neuronales de aprendizaje profundo, vision artificial, procesamiento de lenguaje natural, comercio algorítmico.	[31]
Smart mobility	Tráfico. Gestión de fallas.	Aprendizaje supervisado. Aprendizaje no supervisado. Aprendizaje semisupervisado.	[32]
Smart government	Participación del ciudadano en la gobernación. Toma de decisiones políticas. Manejo de infraestructura nacional.	Aprendizaje supervisado y no supervisado.	[33]
Smart people	Rendimiento de los alumnos y detección de comportamientos de los alumnos, entre otros.	Modelos de clasificación. Estimación de densidad de regresión.	[34]

Para ***smart environment***, la cadena de búsqueda fue "*environmental data science*", de la cual se encontró un trabajo que explora el potencial de ML para revolucionar el data analysis en el campo del enviromental science and enginnering, este sumariza en cuatro los mejores tipos de aplicaciones de ML que es el hacer predicciones, identificar la importancia de las características con regression or classification modeling [28].

Para ***smart living*** se usó la cadena "*smart building data science*". El enfoque predominante es el uso de modelos clasicos

como máquina de vectores de soporte, redes neuronales artificiales o sus ensambles enfoques boosting o bagging, y bosques aleatorios. Pero se tiende a un aprendizaje profundo sugiriendo redes convolucionales de memoria recurrente a corto plazo, o incluso redes adversarias generativas para imitar el sistema real [29].

Otra búsqueda fue con la cadena *"health data science"*. Un trabajo importante encontrado es [30], que se enfoca a predecir el número acumulado de infecciones confirmadas por COVID-19. Utiliza técnicas de y polinomios fraccionarios múltiples.

La cadena de búsqueda para ***smart economy*** fue *"data science applied to economy and marketing"*, con esta búsqueda se encontró el trabajo [31], que hace una revisión sistemática de métodos de análisis utilizados en el marketing digital. Los métodos son agrupación, ensemble de modelos, redes neuronales de aprendizaje profundo, visión artificial, procesamiento de lenguaje natural y comercio algorítmico.

Para ***smart mobility*** se construyeron dos cadenas, "*data science applied to traffic congestion*" y "*data science applied to traffic monitoring*". De los resultados de la búsqueda, se destaca la investigación [32], enfocada en tecnologías que manejan el monitoreo y tráfico de grandes redes conocido como big data. Este involucra basicamente al análisis de grandes datos conocido como big data analytics; se mencionan las categorias de algoritmos implicados, como supervisados, no supervisados y semisupervisados y se alude a que con frecuencia se citan enfoques novedosos tales como redes neuronales profundas y el aprendizaje por refuerzo.

La cadena de búsqueda para ***smart government*** fue *"data science applied to data governance"*, se identifica un artículo que describe en modo general la Smart Cities que implementan los gobiernos de todo el mundo [33]. Señala a tecnologías de inteligencia artificial (IA) que se utilizan: asistentes, o advisors y se dividen en dos ramas knowledge-based systems y ML. De esta última se desglosa el aprendizaje supervisado y el aprendizaje no supervisado. Y otras tecnologías importantes para la interacción gobierno-ciudadano son el procesamiento de lenguaje natural y el análisis de sentimientos.

Finalmente, para ***smart people*** la cadena de búsqueda fue "data science applied to education", de lo cual se recuperó el trabajo [34]. Los métodos más populares fueron la minería de datos y el análisis de aprendizaje en ambientes de educación, resaltando la predicción, clasificación, regression y estimación de densidades [34].

En resumen, existe una variedad de métodos utilizadas en las Smart Cities, incluso en una sola dimensión suelen aplicarse diversos algoritmos. Los métodos más frecuentes son de las categorías aprendizaje supervisado y no supervisado, el aprendizaje profundo se menciona para casos de grandes volúmenes de datos, para todos ellos la motivación del análisis es la toma de decisiones.

Selección del método ML y método de evaluación

Para seleccionar el tipo de técnica en este estudio, se consideraron además otros criterios, tales como los contenidos de los ODS, la presencia o no de la variable target o dependiente, el tipo de dato y de las variables independientes involucradas en la hipótesis, que podrían ser numéricas (discretas o continuas) o texto. Ya que los tipos de técnicas asociadas al aprendizaje supervisado, según [35] son la técnica de clasificación cuando la variable dependiente es discreta y regresión cuando la variable dependiente es continúa. Aunque existen

algunos tipos de clasificadores que aceptan tanto valores discretos (numéricos) como variables categóricas con etiquetas (texto) como valores. Para aplicar los algoritmos se emplearon matlab y lenguaje R.

La evaluación de los modelos se llevó a cabo utilizando la matriz de confusión que se basa en la sensibilidad estadística, consiste en sumar los elementos clasificados correctamente y dividir dicha suma entre el total de elementos del ODS, luego la cifra se multiplica por 100 para obtener un porcentaje, es una prueba para calcular la probabilidad de que un modelo clasifique correctamente, se usó r^2 para los casos de regresión lineal.

5. Resultados

De acuerdo al objetivo de esta investigación que es resaltar la importancia de la calidad del Open Data y su aplicabilidad de la ciencia de datos, para la evaluación de modelos de predicción para las Smart Cities, en esta sección se presentan los resultados de acuerdo a la aplicación de las fases de la metodología. En el primer apartado se presenta la clasificación de una muestra de ODS en las Smart Cities para apreciar la cobertura de información en todas las dimensiones, así mismo se visualiza el rango de cantidad de datos de los ODS, básico para aplicar la ciencia de datos; con ello el profesional o autoridad de la toma de decisiones obtiene un panorama de factibilidad de análisis sobre el kpi de interés. En el segundo apartado se hace una evaluación de pérdida de datos durante la limpieza de la muestra de ODS, indicando su impacto directamente en la calidad. En el tercer apartado se especifican y aplican las técnicas ML y se evalúa el ajuste de los datos a dicha prueba por medio del cálculo de sensibilidad estadística, con esta se verifica si es factible la técnica utilizada, y confiable el modelo obtenido para explotar un posible patrón que encamine a la toma de decisiones.

5.1. Resultados de las fases 1 y 2: exploración del open data en sitios web, selección, descarga y clasificación

De la visita a los portales de Dinamarca, Austria, Canadá, China, Francia, Reino Unido y USA se seleccionaron y almacenaron 100 ODS, estos se encuentran en el repositorio del Apéndice A. Los formatos son CSV y XLSX, abarcan un período del 2015 al 2020.

De acuerdo, a las Smart City se obtuvo una clasificación de cada ODS en cada dimensión, véase la figura 2.

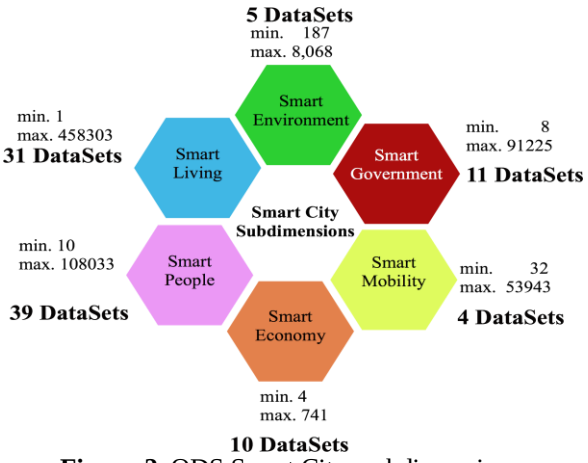


Figura 2. ODS Smart Cites subdimensions.

En la figura 2, aparecen las 6 dimensiones, que son smart environment, smart economy, smart living, smart people, smart mobility, y smart government. Donde se muestra el total de ODS clasificados en cada dimensión, y el mínimo (Min.) y máximo (Max.) de registros (filas) que son los tamaños de los ODS: más pequeño y mayor con respecto de la propia dimensión smart city. Se observa que, smart living y smart people contienen más ODS en comparación con el resto de las dimensiones. El tema sobresaliente de los ODS de smart

people es la cobertura de la educación, y de smart living es la seguridad ciudadana.

Del análisis de la clasificación se resalta que smart living posee el archivo de mayor tamaño, 458,303 registros. En contraste, un archivo de smart economy con cantidad máxima de registros de 741, es menor a todos los máximos del resto de las dimensiones, su contenido es referente a servicios de exportaciones.

5.2. Fase preprocesamiento: evaluación de calidad de los ODS

Se observa que el 80% de ODS requirieron limpieza a causa de contener registros o filas incompletas (conjuntos de celdas vacías), caracteres inválidos y sin estandarizar. Menos del 10% adjuntan diccionarios de datos, lo que afecta su interpretación. Además, algunos no mantienen la estructura de un ODS convencional (una variable por columna y un registro por fila).

En la tabla 4, se muestra el listado de 6 ODS, cada uno se encuentra clasificado en su dimensión smart environment (SE), smart economy (SC), smart living (SL), smart people (SP), smart mobility (SM), smart government (SG). Se observa también el país de procedencia, fecha de los datos, y total de celdas de calidad con respecto al total general. Los tipos de fuentes más comunes son encuestas, censos, reportes e informes. El contenido de los ODS abarcó una amplia diversidad de temas desde lo más esencial como impuestos, salud, o medioambiente, hasta diversión y entretenimiento, entre otros. En cuanto al formato, todos utilizan los universales como CSV y Excel. En el repositorio del Apéndice B se muestra el análisis de los 100 casos.

Tabla 4. Casos de estudio para las seis dimensiones smart city.

Orden	Clasificación	Año datos	Nombre del data set	País	Fecha de los datos	Total de variables (columnas)	Total celdas (filas* columnas)	Total celdas(filas*columnas) con calidad
1	SC	2016	Ministry of Trade and Industry - Department of statistics, *exports of services by major trading partner, annual*, Singapure: June 22, 2016. [1]	China	2016, jun. 22	3	2223	741
2	SG	2018	Canada revenue agency, *final table 01 – General statement by province and territory of taxation.*[2]	Canadá	2018, jan. 31	12	540	494
3	SL	2020	Allegheny County / City of Pittsburgh / Western PA Regional Data Center, *Allegheny County Depression Medication.*[3]	USA	2017, mar. 06	7	2737	2737
4	SM	2020	City of Bloomington, *Monroe County Crash Data_2003 to 2015.*[4]	USA	2017, sep. 06	12	647316	645801
5	SP	2014	Canada Border Services Agency, *volume of travelers and conveyances.*[5]	Canadá	2014, feb. 26	4	280	280
6	SE	2020	Saint-Denis town hall (Reunion Island)/Commune of Saint-Denis de la Réunion, Average monthly concentrations of pollutants in ambient air from the permanent network of automatic measurements at cinor.[6]	Francia	2020, oct. 16	22	4114	4056

[1] https://data.gov.sg/dataset/exports-of-services-by-major-trading-partner-annual
[2] https://open.canada.ca/data/en/dataset/db43ce04-e010-44f3-bb8b-11461dfc6996
[3] https://catalog.data.gov/dataset/allegheny-county-depression-medication
[4] https://catalog.data.gov/dataset/traffic-data
[5] https://open.canada.ca/data/en/dataset/fae0ec36-c1ab-4ee6-a65e-4d26cf3671bc
[6] https://www.data.gouv.fr/fr/datasets/concentrations-moyennes-mensuelles-de-polluants-dans-lair-ambiant-issues-du-reseau-permanent-de-mesures-automatiques-sur-la-cinor-5/

5.3. Fase tratamiento del data science: casos de estudio

En esta sección se muestran los resultados de la fase data science treatment. Aquí se aplicaron técnicas ML para evaluar 6 casos, estos corresponden a los que se muestran en la tabla 4. Se eligieron aleatoriamente 6 entre los 100, para probar un ODS de cada dimensión smart city. A continuación, se muestra el detalle del planteamiento de la hipótesis de cada caso.

Caso 1: Smart environment

- *Fuente.* The pest management centre's minor use pesticides program, iniciativa para mejorar la disponibilidad de productos de riesgo reducido para los productores agrícolas y mejorar el acceso a plaguicidas de uso menor.
- *Análisis.* Demostrar si el estatus del proyecto depende del cultivo y del ingrediente del pesticida. *Hipótesis:*ProjectStatusEn_EtatProjetAn ~ CropNameFr_NomCultureFr + ActiveIngrdntNameEn_NomIngrdntActifAn. *Método*: Clasificación.

Caso 2: Smart government

- *Fuente.* Declaraciones fiscales.
- *Análisis.* Se evalúa si existe una clasificación del estado del reembolso que dependa del número de declaraciones, del impuesto federal neto y del impuesto provincial neto.
- *Hipótesis:* ReturnType~Number_of_returns + Netfederaltax. *Método:* Clasificación.

Caso 3: Smart living

- *Fuente.* Departamento de salud acerca de la medicación para los casos de depresión y ansiedad detectados.
- *Análisis.* Se prueba si el total de mujeres enroladas (internadas) por al menos 90 días por desorden de ansiedad depende del número total de mujeres con al menos una solicitud de medicación para el desorden de ansiedad.
- *Hipótesis:* EWAD ~ EWAN. *Método:* Regresión.

Caso 4: Smart mobility

- *Fuente.* Datos originados de un sistema automatizado de reportes e intercambio de información de automóviles con choques provocados por colisiones, vehículos e individuos. País de origen. Estados Unidos.
- *Análisis.* Se prueba si el tipo de lesión en un accidente depende del día, fin de semana o no y el tipo de colisión (cuántos vehículos y de qué tipo o peatones involucrados). *Hipótesis:* InjuryType ~ Day + CollisionType. Método: Clasificación.

Caso 5: Smart people

- *Fuente.* Agencia de servicios fronterizos que proporciona el número anual de viajeros y medios de transporte que llegan a cada región de una nación. País de origen. Canadá.
- *Análisis.* Se prueba si la región está determinada por el volumen de viajeros que lleguen y el medio de transporte en que lo hacen. *Hipótesis:* Region ~ Traveller Volume + ConveyanceVolume. *Método*: Clasificación.

Caso 6: Smart environment

- *Fuente.* Muestras de concentraciones medias mensuales de contaminantes en el aire en una región. País de origen. Francia.
- *Análisis.* Se prueba si el tipo de contaminante se debe al tipo de zona (urbana, industrial, periurbano o de tráfico) y la estación de muestreo.

Hipótesis: nom_poll ~ typologie + code_station. *Método*: Clasificación.

En la tabla 5, se observa cada caso probado, aparace la dimensión Smart City y la sensibilidad estadística, el método de evaluación y el algoritmo de aprendizaje aplicado, se selecciona máquinas de vectores de soporte para clasificación porque todas las variables de la hipótesis son numéricas, y árboles de decisión debido a que al menos una de las variables es no numérica. Se elige regresión lineal para los casos de predicción numérica.

Tabla 5. Resultados de la aplicación del ML.

Dimension Smart City	Sensibilidad estadística	Método de evaluación	Algoritmo aplicado
1. Economy	8%	R^2	Regresión lineal simple
2. Government	54%	Matriz de confusión	Máquina de vectores de soporte (función kernel, multiclase)
3. Living	93%	R^2	Regresión lineal simple
4. Mobility	77%	Matriz de confusión	Árboles de decisión
5. People	22%	Matriz de confusión	Árboles de decisión
6. Environment	27%	Matriz de confusión	Árboles de decisión

En esta aplicación del ML, se resalta que solo el caso 3 rebasa el 80% de sensibilidad o confiabilidad del modelo. El caso 4, se acerca a este valor con 77%, y los cuatro restantes son menores al 80%. Se usa como valor de referencia el 80% porque es el porcentaje aceptable más común para reconocer un patrón en los modelos; significa que no se observan patrones en 5 de los 6 casos, debido a que el porcentaje es menor al aceptable.

De lo anterior se concluye que al ser probadas las inferencias para evaluar el kpi de interés con ML, la mayoría de los casos no muestran presencia de patrones de acuerdo al algoritmo utilizado, esto indica que no se asegura que se encuentre algún modelo, pero no se elimina la posibilidad de usar otras técnicas.

6. Hallazgos

En cuanto a la descarga y clasificación se descubre que el Open Data abarca diversidad de temas que engloban a las dimensiones de las Smart Cities, y los archivos extraidos fueron accesibles en formatos universales. Por otro lado, fue tardada y complicada la búsqueda de un dataset asociado a un tema y a una dimensión Smart City. Durante el preprocesamiento se evidencia un alto esfuerzo de limpieza y la búsqueda de diccionarios para su interpretación, lo que afecta al tamaño de la muestra de los datos y remite en la aparición de sesgos.

Se evidenció que no solo es complejo el encontrar un tema específico sino también el encontrar un patrón en los datos, ya que implica el construir hipótesis con variables que muchas veces se desconoce su naturaleza. Con respecto, a la aplicación de algoritmos se encontró solo un ODS con un patrón, cuya sensibilidad fue del 93%, esto indica que el resto de los ODS no obtuvieron un patrón, porque o bien, eran datos incompletos o las hipótesis fueron mal planteadas, o no existe patrón en esa prueba.

Se considera que, aunque los modelos obtenidos mostraron bajo porcentaje de sensibilidad, el ciudadano común puede explotar su contenido, mediante la ejecución de diversas consultas básicas, siempre y cuando tenga acceso a la información que le brinde dominio del contexto. Además, es muy frecuente que los administradores de los sitios eliminen los ODS y ya no se pueda tener

acceso a ellos, por eso se debe contar siempre con un repositorio seguro.

7. Conclusiones

En este estudio se propuso resaltar la importancia y calidad del Open Data y su aplicabilidad de la ciencia de datos para la evaluación de modelos de predicción para las Smart Cities, donde se identificó una gran variedad de ODS para todas sus dimensiones. Se extrajo una muestra de 100 ODS y se observó que en la mayoría de estos hubo alta reducción de datos después del preprocesamiento, lo que evidencia una baja calidad. Para 6 casos, se aplicó ML conforme a las inferencias previamente establecidas y se encontró solo un caso de existencia de patrón para predicción, demostrando que no se garantiza una aplicabilidad de ciertos métodos con la ciencia de datos.

Trabajos futuros

Un trabajo a futuro es investigar el cómo enlazar los ODS a las métricas que evalúan a una Smart City ya que actualmente se enfrenta la dificultad de relacionar a los kpi's con esta. Así mismo, se espera trabajar en el estudio del ciclo de vida del Open Data para que este sea centrado en la calidad de los datos y en el proceso de la visualización de herramientas alternativas para la toma de decisiones.

8. *Apéndices*

Apéndice A: *En la siguiente dirección de acceso a la nube se muestran los 100 open data sets descargados en formatos CSV y XLS:* *https://drive.google.com/drive/folders/16ij_6WgIsbnCffYV6iNFRb2Z1dIEQvf*

Apéndice B: *En la dirección se encuentra una tabla con los 100 ODS clasificados en las dimensiones Smart Cities en formato.* *https://docs.google.com/document/d/1h5-72YwTXVNwqd6oT9IPfRV5CPx0K6JT/edit?usp=sharing&ouid=106418425807417936772&rtpof=true&sd=true*

Apéndice C: *En la segunda parte del documento se encuentran las referencias bibliográficas de los 100* ODS. *https://docs.google.com/document/d/1h5-72YwTXVNwqd6oT9IPfRV5CPx0K6JT/edit#bookmark=id.q2enl1p50xno*

9. Referencias

[1] R. J. Miller, "Open data integration," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 2130–2139, Aug. 2018, doi: 10.14778/3229863.3240491.

[2] L. Akerreta Escribano and J. Moyano Collado, "Contar historias con los datos: Aragón Open Data Focus, una experiencia innovadora de reutilización de los datos del sector público," *Scire: representación y organización del conocimiento*, vol. 27, no. 1, pp. 31–43, Jun. 2021, doi: 10.54886/scire.v27i1.4777.

[3] K. R. Dalal, "Review on Application of Machine learning Algorithm for Data Science," in *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, IEEE, Nov. 2018, pp. 270–273. doi: 10.1109/ICICT43934.2018.9034256.

[4] T. Siebenlist and A. Mainka, "Culture of openness," in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, New York, NY, USA: ACM, May 2018, pp. 1–6. doi: 10.1145/3209281.3209365.

[5] E. M. Thompson, "What makes a city 'smart'?," *International Journal of Architectural Computing*, vol. 14, no. 4, pp. 358–371, Dec. 2016, doi: 10.1177/1478077116670744.

[6] R. Duan and T. Noda, "The classification and potential of business archetypes by using open data," in *Proceedings of the 15th International Symposium on Open Collaboration*, New York, NY, USA: ACM, Aug. 2019, pp. 1–9. doi: 10.1145/3306446.3340830.

[7] R. Enriquez-Reyes, S. Cadena-Vela, A. Fuster-Guillo, J.-N. Mazon, L. D. Ibanez, and E. Simperl, "Systematic Mapping of Open Data Studies: Classification and Trends From a Technological Perspective," *IEEE Access*, vol. 9, pp. 12968–12988, 2021, doi: 10.1109/ACCESS.2021.3052025.

[8] A. Núñez Reiz, M. A. Armengol de la Hoz, and M. Sánchez García, "Big Data Analysis and Machine Learning in Intensive Care Units," *Medicina Intensiva (English Edition)*, vol. 43, no. 7, pp. 416–426, Oct. 2019, doi: 10.1016/j.medine.2019.06.012.

[9] T. Cha, "Open Government Data for Machine Learning Tax Recommendation," in *The 21st Annual International Conference on Digital Government Research*, New York, NY, USA: ACM, Jun. 2020, pp. 331–333. doi: 10.1145/3396956.3397002.

[10] S. I. H. Shah, V. Peristeras, and I. Magnisalis, "Government (Big) Data Ecosystem: Definition, Classification of Actors, and Their Roles," *International Journal of Computer and Information Engineering*, vol. 14, no. 4, pp. 102–114, 2020.

[11] M. Guimarães Belizario and R. C. Galarraga Berardi, "Use of Smart and Open Data in Smart Cities," in *Twenty-fifth Americas Conference on Information Systems AMCIS 2019 Proceedings. 7*, Cancún, 2019, pp. 1–10.

[12] E. Estrada, R. Maciel, A. Ochoa, B. Bernabe-Loranca, D. Oliva, and V. Larios, "Smart City Visualization Tool for the Open Data Georeferenced Analysis Utilizing Machine Learning," *International Journal of Combinatorial Optimization Problems and Informatic*, vol. 9, no. 2, pp. 1–16, Jun. 2018.

[13] E. Estrada, A. P. Pérez Negrón, G. Lara López, and L. Casillas, "Algoritmos para el control de indicadores clave de desempeño para Smart Cities," *Research in Computing Science*, vol. 147, no. 8, pp. 121–133, Dec. 2018, doi: 10.13053/rcs-147-8-9.

[14] E. Estrada, R. Maciel, A. P. P. Negrón, G. L. López, V. Larios, and A. Ochoa, "Framework to support the Data Science of smart city models for decision-making oriented to the efficient dispatch of service petitions," *IET Software*, vol. 14, no. 2, pp. 159–164, Apr. 2020, doi: 10.1049/iet-sen.2019.0044.

[15] F. T. Neves, M. de Castro Neto, and M. Aparicio, "The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring," *Cities*, vol. 106, p. 102860, Nov. 2020, doi: 10.1016/j.cities.2020.102860.

[16] L. Hurbean, D. Danaiata, F. Militaru, A.-M. Dodea, and A.-M. Negovan, "Open Data Based Machine Learning Applications in Smart Cities: A Systematic Literature Review," *Electronics (Basel)*, vol. 10, no. 23, p. 2997, Dec. 2021, doi: 10.3390/electronics10232997.

[17] E. Badidi, Z. Mahrez, and E. Sabir, "Fog Computing for Smart Cities' Big Data Management and Analytics: A Review," *Future Internet*, vol. 12, no. 11, p. 190, Oct. 2020, doi: 10.3390/fi12110190.

[18] S. E. Bibri, "On the sustainability of smart and smarter cities in the era of big data: an interdisciplinary and transdisciplinary literature review," *J Big Data*, vol. 6, no. 1, p. 25, Dec. 2019, doi: 10.1186/s40537-019-0182-7.

[19] R. Zhang, M. Indulska, and S. Sadiq, "Discovering Data Quality Problems," *Business & Information Systems Engineering*, vol. 61, no. 5, pp. 575–593, Oct. 2019, doi: 10.1007/s12599-019-00608-0.

[20] T. Olsson and P. Runeson, "Open data collaborations," in *Proceedings of the 15th International Symposium on Open Collaboration*, New York, NY, USA: ACM, Aug. 2019, pp. 1–4. doi: 10.1145/3306446.3340832.

[21] A. Colborne and M. Smit, "Characterizing Disinformation Risk to Open Data in the Post-Truth Era," *Journal of Data and Information Quality*, vol. 12, no. 3, pp. 1–13, Sep. 2020, doi: 10.1145/3328747.

[22] M. Janssen and G. Kuk, "Big and Open Linked Data (BOLD) in research, policy, and practice," *Journal of Organizational Computing and Electronic Commerce*, vol. 26, no. 1–2, pp. 3–13, Apr. 2016, doi: 10.1080/10919392.2015.1124005.

[23] D. Arribas-Bel, M. Green, F. Rowe, and A. Singleton, "Open data products-A framework for creating valuable analysis ready data," *J Geogr Syst*, vol. 23, no. 4, pp. 497–514, Oct. 2021, doi: 10.1007/s10109-021-00363-5.

[24] A. Nikiforova and K. McBride, "Open government data portal usability: A user-centred usability analysis of 41 open government data portals," *Telematics and Informatics*, vol. 58, p. 101539, May 2021, doi: 10.1016/j.tele.2020.101539.

[25] Jiang, Shao, Liou, Shao, and Shi, "Improving the Sustainability of Open Government Data," *Sustainability*, vol. 11, no. 8, p. 2388, Apr. 2019, doi: 10.3390/su11082388.

[26] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Comput Sci*, vol. 2, no. 5, p. 377, Sep. 2021, doi: 10.1007/s42979-021-00765-8.

[27] Euroresidentes, "Territorios Inteligentes Beyond Smart Cities," La rueda de Boyd Cohen (componentes de una Smart City).

[28] S. Zhong *et al.*, "Machine Learning: New Ideas and Tools in Environmental Science and Engineering," *Environ Sci Technol*, p. acs.est.1c01339, Aug. 2021, doi: 10.1021/acs.est.1c01339.

[29] J. Aguilar, A. Garces-Jimenez, M. D. R-Moreno, and R. García, "A systematic literature review on the use of artificial intelligence in energy self-management in smart buildings," *Renewable and Sustainable Energy Reviews*, vol. 151, p. 111530, Nov. 2021, doi: 10.1016/j.rser.2021.111530.

[30] T. van der Ploeg and R. J. J. Gobbens, "Prediction of COVID-19 Infections for Municipalities in the Netherlands: Algorithm Development and Interpretation," *JMIR Public Health Surveill*, vol. 8, no. 10, p. e38450, Oct. 2022, doi: 10.2196/38450.

[31] J. R. Saura, "Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics," *Journal of Innovation & Knowledge*, vol. 6, no. 2, pp. 92–102, Apr. 2021, doi: 10.1016/j.jik.2020.08.001.

[32] A. D'Alconzo, I. Drago, A. Morichetta, M. Mellia, and P. Casas, "A Survey on Big Data for Network Traffic Monitoring and Analysis," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 800–813, Sep. 2019, doi: 10.1109/TNSM.2019.2933358.

[33] Z. Engin and P. Treleaven, "Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies," *Comput J*, vol. 62, no. 3, pp. 448–460, Mar. 2019, doi: 10.1093/comjnl/bxy082.

[34] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, May 2020, doi: 10.1002/widm.1355.

[35] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.