NOTA

1. BREVE GUÍA PARA LA INNOVACIÓN EN EL SECTOR PÚBLICO DESDE UN ENFOQUE DATA-DRIVEN

Lic. Nicolás Gottig[1]

Brief public sector innovation guide to a data-driven approach

La administración de los recursos públicos involucra actores que tienen un doble desafío: en primer lugar y como toda organización, debe tomar decisiones con recursos limitados en un contexto de incertidumbre. Por otro lado, las prácticas en la gestión pública están condicionadas a las lógicas políticas e institucionales (normas) que rigen el comportamiento de quienes administran el Estado, incluyendo la necesidad de dar respuesta a una comunidad heterogénea a través de una estructura generalmente burocrática. Utilizar herramientas de toma de decisiones incide en la eficiencia del control de los recursos de la administración pública (Piccirilli & Farías, 2015).

El objetivo de este documento es caracterizar una serie de pasos que deben considerarse al momento de implementar un sistema de información basado en la automatización del procesamiento de la información, describiendo algunos de sus posibles usos y reconociendo la necesidad de las interdisciplinas para llevarlo a cabo.

En este sentido, la administración del estado deriva en la importancia de tener información precisa y actualizada que aporte a la toma de decisiones. No sólo para realizar lecturas rápidas referidas a la gestión, sino para estudiar el impacto de las decisiones que se toman y, si corresponde, ajustarlas. Como afirman Piccirilli & Farias (2015), no contar con información apropiada hace que no se pueda hacer control, ni interno ni externo. El enfoque donde los datos son integrados al proceso de innovación y gestión es conocido como innovación basada en datos, o data-driven innovation (Janssen et al (2017) en Salvador y Ramió, 2020).

[1] Secretario de Extensión de la Facultad de Ciencias de la Gestión (UADER). Maestrando en Estadística Aplicada (UNR). Economista (UADER). Docente Universitario (UADER-UNR).

Las políticas de gobierno abierto y transparencia tienen como una de sus acciones la difusión de datos relativos a la gestión pública. Es realmente un desafío, ya que las diferencias en los distintos sectores provocan la necesidad de realizar modificaciones en la comunicación de la información, de acuerdo al destinatario (Ozlak, 2013).

En principio, la definición clara de indicadores para su posterior recolección, el análisis y su difusión es fundamental, aunque no suficiente: La sistematización y automatización de la recolección de los datos es un trabajo técnico-informático, y debe realizarse luego de ejecutar tareas en las que deben participar todos los actores que son parte de la gestión.

Luego, se debe proceder a la construcción de indicadores o estadísticos que resuman la información y, si corresponde, la construcción de modelos[2] que permitan hacer inferencia (generalizaciones) sobre el comportamiento de esos indicadores (o su pronóstico)[3]. Por último, la difusión de la información es la meta del proceso: se debe definir a quién se va a comunicar la información y, por lo tanto, en qué formato: No es lo mismo publicar un informe técnico que una infografía sobre un determinado indicador, o una base de datos en línea y de acceso público. Tampoco es lo mismo la información generada a los efectos de difundirla en la comunidad, que la información generada para la toma interna de decisiones.

La sistematización de la recolección de los datos, su procesamiento y su reportería propone reducir la incertidumbre en la toma de decisiones diaria. Asimismo, la estadística permite cuantificar las chances de fallar y, aunque esta medida no es exacta, es intuitiva. En este sentido, reconocer las herramientas generadoras de datos (por ejemplo, los sistemas de información propios de cada institución) o diseñar métodos de recolección, es una de las tantas tareas a las que hay que enfrentarse, aunque no la única.

A continuación, se sugieren las preguntas que los actores de las instituciones públicas pueden hacerse en cada etapa del proceso de generación y difusión de la información, al que de ahora en adelante denominaremos *sistematización de la información*, poniendo especial énfasis en la etapa de diseño y análisis ya que como afirma Sosa Escudero (2020) las grandes cantidades de datos sin su tratamiento y posterior lectura son similares a Funes el memorioso, el polémico personaje de Borges que recordaba cada detalle hasta el punto de que, para relatar los eventos de un día, tardaba ni más ni menos que 24 horas: una habilidad tan precisa como inútil, ya que la abstracción es un requisito a la hora de comunicarnos.

[2] Los modelos son expresiones formales y simplificadas de la realidad.

[3] Sin caer en la premisa de que es posible hacer "futurología" los modelos matemáticos que se usan para predecir el comportamiento de un fenómeno poseen, en general, más carencias que bondades. Existen muchos debates en torno a su uso, aunque en este artículo nos vamos a limitar a lo escrito por George Box (1976) quien afirma que "todos los modelos (estadísticos) son incorrectos, pero algunos son útiles" haciendo referencia a que no tiene sentido evaluar un modelo por su posibilidad de acertar en la mayoría de los casos.

El ejercicio del análisis (y de la estadística) se pone en valor, además, cuando reconocemos el aluvión de datos como efecto del acceso a internet y los dispositivos inteligentes. Sin un diseño consistente del problema y del tratamiento, la Big Data[4] es inútil (como Funes al momento de relatar un hecho).

A su vez, las formas de trabajo de este informe están basadas en metodologías ágiles y realiza una lectura de la propuesta atravesada por elementos del análisis estadístico. El informe corresponde a la fundación Sadosky y se denomina Innovar con Ciencia de Datos en el sector Público (2022). Sin más preámbulo, comencemos:

El ciclo de análisis de la información desde una perspectiva estadística

Aunque se asocie a la estadística con cálculos extraños, o se la utilice de forma errónea en el intento de generalizar conclusiones a partir de una muestra (el incorrecto uso de herramientas estadísticas, a veces de forma deliberada, a veces no, amerita un artículo aparte) gran parte del trabajo estadístico y, por lo tanto, en la sistematización de la información, transcurre en la definición del problema.

Al momento de tomar decisiones debemos construir indicadores, o reconocer nuestros parámetros de interés. Para una empresa, un indicador puede ser la rentabilidad sobre ventas, su tasa de solvencia o su eficiencia operativa. La administración pública no persigue fines de lucro, aunque la eficiencia debería ser uno de sus pilares fundamentales. Por lo tanto, los indicadores que se utilicen para monitorear las acciones deben brindar información suficiente sobre los procesos en los que está involucrado un organismo. En este sentido, en la definición de las métricas deben participar representantes de todos los actores involucrados en la gestión de la institución que está atravesando ese proceso.

La tasa de internados condicionados a la vacuna contra el COVID-19 por departamento puede ser un indicador útil para el sector de la salud, mientras que las notas promedio por área (matemáticas, sociales, económicas, naturales…) puede ser un indicador importante para el sistema educativo. Entonces, el proceso para construir y monitorear estos indicadores se constituye de la siguiente forma:

1. Planteo del problema.
2. Planificación del estudio.
3. Adquisición de los datos.

[4] Big Data hace referencia a la adquisición de datos en grandes volúmenes, inmensurables para el cerebro humano. Como referencia, el conocido software Excel almacena 10.485.000 filas de datos aproximadamente. Esto podría ser equivalente a tener datos de 10.485.000 usuarios. Si pensamos en Big Data, deberíamos multiplicar este número por varios millones.

4. Análisis.
5. Difusión.

Este proceso se denomina PPDAC, y fue planteado por Wild y Pfannkuch (1999) como una generalización del proceso para el análisis estadístico. Si entrecerramos los ojos, observamos que, en general, es el proceso de la metodología de la investigación.

Si abordamos el problema desde un paradigma de *design thinking* (Martínez, María V., 2022) entonces la secuencia estaría dada por:

1. Diagnóstico
2. Ideación
3. Prototipado
4. Ajustes y validación
5. Difusión y escalamiento
6. Cambio sistémico

Como en el caso anterior, si entrecerramos los ojos, observamos que es el proceso planteado por la teoría fundamental de la administración: planificación, organización, dirección, integración, control y retroalimentación.

Dependiendo si necesitamos un estudio puntual (por ejemplo, un censo o relevamiento) o la construcción de sistemas de información (información resumida siempre disponible) nos será más útil uno u otro proceso.

Aunque a continuación planteamos un esquema de trabajo partiendo del último enfoque, pero por la naturaleza de los objetivos, estamos atravesados por el ciclo PPDAC.

Etapas para la construcción de sistemas de información en el sector público

El objetivo de este recorrido es sugerir posibles caminos para sistematizar la información disponible en las organizaciones, con el fin de que esté a disposición cualquiera sea el momento de la consulta. Cuando necesitamos información instantánea debemos pensar en sistemas de monitoreo, mientras que, si los datos se recolectan de fuentes secundarias que se publican bajo cierta periodicidad, o se relevan a través de un cuestionario, entonces debemos pensar en proyectos puntuales.

En el diagnóstico y la definición de indicadores nos preguntaremos qué queremos saber y para qué, mientras que en la ideación del proyecto pensaremos el cómo, además de nombrar algunos recursos gratuitos para concretar objetivos puntuales. En las etapas de prototipado, ajuste y validación se ejecutan los proyectos de forma parcial y se difunden los resultados preliminares entre los colaboradores. Luego de la retroalimentación, se realizan los ajustes necesarios y se difunde la información con la comunidad o, si corresponde, se pone a disposición de los equipos de gestión.

1. Diagnóstico y definición de indicadores

En esta etapa es importante definir el para qué. La naturaleza del problema debe ser pertinente al equipo de gestión involucrado, y deben especificarse resultados e impactos esperados luego de la aplicación de la herramienta. Es importante definir qué parámetros quieren conocerse (un promedio de costos de construcción, un porcentaje de personas satisfechas con un servicio, etc.) y cuál es la población de interés. También deben escribirse los criterios de inclusión y exclusión (qué forma parte de esa población y qué no) y, si corresponde, definir indicadores[5]. Si es posible, debe garantizarse la trazabilidad de las unidades que se estudiarán para que, en caso de que se registre la información de forma incorrecta o exista un emergente, poder consultar nuevamente la fuente.

Algunos criterios para la definición de indicadores según Martínez, María V. (2022) son:

- El sentido de un indicador es claro.
- Existe información disponible o se puede recolectar fácilmente.
- Un indicador es tangible y se puede observar.
- La tarea de recolectar datos está al alcance de la dirección del proyecto y no requiere expertos para su análisis.
- Un indicador es lo bastante representativo para el conjunto de resultados esperados.
- Un criterio que debe ser evaluado, es que los indicadores deben ser independientes. A su vez, es esperable no asignar una relación de causa-efecto entre el indicador y el objetivo que se evalúa.

Definidos y escritos los objetivos, las expectativas, la población, los criterios de inclusión y exclusión, y los indicadores, se debe definir la forma de recolectar la información:

[5] Los indicadores pueden ser simples (por ejemplo, cantidad de estudiantes que aprobaron una determinada materia) o complejos (por ejemplo, la tasa de ocurrencia de los ciudadanos a un determinado servicio, para optimizar los costos de atención).

- En caso de utilizar información secundaria provista de otros organismos, se debe evaluar la periodicidad de la publicación.
- En caso de relevar información a través de **todas** las unidades que constituyen la población, el proceso se reduce considerablemente.
- En caso de relevar información a través de una **muestra**, debe realizarse un diseño muestral que garantice que todas las unidades tengan las mismas chances de ser elegidas, además de ser independientes.

En caso de realizar muestreos, se debe garantizar la representatividad de ésta. Aunque es verdad que la cantidad de unidades que constituyen la muestra es importante (el reconocido $\sqrt{n}$[6]), las diferencias (variabilidad) entre estas unidades también son importantes. Existe un conocido dicho que establece que no es necesario probar una olla de salsa para saber si está salada, basta con una cuchara. Esta afirmación, inocente pero propositiva, hace referencia a que los diseños muestrales reducen los costos de adquisición de la información y, muchas veces, son más precisos que proponer la realización de un censo. Una práctica común para garantizar la representatividad es dividir la población de estudio en categorías o regiones, y luego tomar una muestra aleatoria de cada una. Aunque existen otros tipos de técnicas de muestreo (cada una según el problema) no es el objetivo de este documento.

Bajo esta premisa, si el objetivo de un área de desarrollo social es diagnosticar las carencias y las fortalezas de una determinada ciudad, debe garantizar la representatividad en todas las zonas. ¿Esto significa que tiene que relevar todas las zonas? la respuesta es no.

En el caso de trabajar con encuestas, debe diseñarse el instrumento considerando las unidades a relevar y como se va a proporcionar para garantizar la representatividad. Se pueden tener en cuenta los siguientes ítems:

- Siempre que sea posible, debe realizarse la encuesta de forma personal o vía telefónica. Si se realiza por correo, debe definirse y recortarse previamente la muestra y realizar el seguimiento de la respuesta. El envío masivo de correos electrónicos o las encuestas en redes sociales no garantiza representatividad y deben ser métodos para casos puntuales.
- De ser posible evitar las respuestas abiertas. Aunque hay formas de resumir cadenas de texto, las preguntas deben limitarse a obtener, sencillamente, una determinada dimensión (edad, sexo, si participa o no de una determinada actividad, etc.) de la unidad de análisis (ciudadanos, docentes, equipos, etc.).

[6] La literatura estadística identifica con la letra n a la cantidad de unidades que constituyen una muestra.

- Los encuestadores deben capacitarse y reconocer quienes serán los encuestados, qué preguntas realizarán y con qué tono y, si corresponde, qué observaciones no preguntar y marcar. Se debe, además, estudiar la reglamentación de secreto estadístico vigente en la región.

- Si a posterior se busca comparar muestras (Ej. Ciudadanos que separan residuos vs. ciudadanos que no separan sus residuos, equipos eficientes vs. equipos ineficientes, etc.) o se quiere conocer el efecto de un factor sobre un fenómeno (Ej. Causas de las bajas notas en determinadas disciplinas o causas de delincuencia en determinadas zonas) se deben incluir en el diseño muestral unidades con todas las características. En el ejemplo de los ciudadanos que reciclan versus los ciudadanos que no reciclan, deberíamos tener información de ambos. Por otro lado, en el ejemplo de los estudiantes con bajas notas en una materia, si conocemos quienes son los estudiantes con bajas notas y a los estudiantes con notas altas, debemos encuestar a ambos.

- En el caso anterior, cuando no se pueden categorizar a las unidades (por ejemplo, si no sé quiénes reciclan) una buena opción es construir una *variable indicadora*. Es decir, en vez de realizar la pregunta "¿Los residuos son separados en su hogar?" puede ser preferible preguntar: "¿Realiza acciones para el cuidado del medio ambiente?" si/no, y "¿Con qué frecuencia, siendo 0 ningún día y 7 todos los días, separa la basura en su hogar?". Así, además de tener mayor descripción de la unidad de análisis, podemos construir una variable ficticia denominada "¿El usuario recicla?" donde la respuesta será "si" si separa la basura en su hogar más de 3 días, o "no" si separa 3 días o menos, independientemente de si realiza acciones para el cuidado del medio ambiente. Esto es sólo un ejemplo y no propone la definición de separar la basura.

- De ser posible, debe probarse el formulario en unidades de análisis con características similares a las de nuestra población de interés y recibir una devolución. Por ejemplo, si en un estudio sobre la presión arterial le consultamos a las personas "¿Hace cuanto dejaste de fumar?" algunos pueden contestarnos "hace 5 años" y otros "en 2005". El hecho de que las escalas de medición sean distintas puede provocar incongruencias en análisis posteriores si no se detecta a tiempo.

En síntesis, algunas de las preguntas que debemos responder en esta etapa son:

- ¿Cuáles son nuestros objetivos de gestión y qué indicadores demuestran su cumplimiento?
- ¿Qué quiero mejorar o erradicar en los procesos institucionales que estoy llevando a cabo?
- ¿Se puede abordar la temática desde un enfoque de datos?
- ¿Poseo el equipo y los instrumentos tecnológicos para llevarlo a cabo?
- ¿Qué información necesito para construir los indicadores y cómo puedo recolectarla?
- ¿Quiénes constituyen las unidades de interés y qué criterio utilizo para definir si una unidad es o no parte de mi grupo de interés?

- ¿Es viable relevar información de toda la población o debe realizarse una muestra?
- ¿En qué escala voy a relevar los datos? ¿Serán apreciaciones subjetivas? ¿Son datos numéricos o categorías?
- ¿Qué canales utilizaré para llegar a mis unidades de interés? ¿Puedo garantizar la trazabilidad y el seguimiento de la respuesta?

2. Ideación y diseño del proyecto

Según Martínez, María V. (2022), la ideación y diseño de un proyecto de sistematización de la información puede involucrar los siguientes objetivos:

- Recolección y/o sistematización de datos.
- Análisis, visualización y monitoreo de los datos.
- Estimaciones y simulaciones basadas en los datos.
- Predicciones basadas en los datos para predecir o prevenir.

Cualquier proceso de sistematización de la información puede incluir una o varias etapas de las nombradas previamente. Aunque la recolección y la sistematización de datos es costosa en términos de tiempo y recursos humanos, no es suficiente si el objetivo es visualizar la tasa promedio de espera de un bus urbano.

En este sentido, aunque cada ítem de la lista anterior es un objetivo por sí mismo, constituyen además una serie ordenada de pasos que hay que ejecutar para alcanzar cada uno (si mi objetivo es realizar estimaciones, entonces tengo que ejecutar los dos pasos previos).

La **recolección y/o sistematización** de los datos involucra acciones para registrar, mediante tablas, las dimensiones de nuestro interés en cada una de las unidades[7] de análisis. Las tablas son similares a las siguientes:

Tabla 1 - Registro de personas ficticias. Fuente: elaboración propia.

ID	Sexo	Edad	estadoCivil	salarioPorHora
Persona1	0	25	3	4220
Persona2	1	43	2	3322

[7] En el ejemplo de la tabla 1 las unidades son personas, aunque podría ser un registro de la tasa de alfabetización por año escolar, en el tiempo, en una provincia determinada. Es por esto que se utiliza "unidades".

Persona3...	1	32	3	5401
Persona n	0	21	2	3033

Donde cada columna representa los atributos de las unidades (filas) y "Persona n" se refiere a la enésima persona entrevistada. En este caso, ID se refiere a un identificador único de cada unidad de estudio. *Sexo* se refiere al sexo declarado por la persona n (0 si es hombre, 1 si es mujer), *estadoCivil* se refiere al estado civil que manifestó la persona n (las etiquetas están más abajo) y *salarioPorHora* se refiere al salario por hora declarado por la persona n expresado en pesos. En bases de datos, esto se denomina base de datos relacional y aunque existen otras, es la de uso frecuente (y suficiente para nuestros objetivos). Por otro lado, si bien el sexo de una persona y el estado civil son factores (es decir, dimensiones cualitativas) se asigna un valor numérico para facilitar la carga de datos. Por ejemplo, para la dimensión que representa el estado civil de una persona se le pueden asignar los siguientes valores:

- 0 si *soltero/a*
- 1 si *casado/a*
- 2 si *divorciado/a*
- 3 si *viudo/a*

De acuerdo a la fuente de los datos, la estrategia será distinta: si necesitamos tasas de pobreza[8] a nivel provincial para un área de presupuesto o información sobre la ocupación hotelera para medir el impacto de una política pública, estaremos supeditados a los formatos en los que se provee la información y a la periodicidad de los mismos, aunque el costo de adquirirla será cero.

La mayoría de los organismos descentralizados nacionales y provinciales provee información sobre indicadores socioeconómicos y demográficos, financieros, entre otros. También existen *APIs*: algoritmos que conectan nuestro servidor web o local (nuestra computadora) con un servidor específico (la computadora de una organización) que contiene datos, y los intercambia a través de los protocolos correspondientes. Así, podemos acceder a información actualizada y de forma instantánea sobre, por ejemplo: el clima, las opiniones en redes sociales sobre un determinado tema, información sociodemográfica[9], entre otros.

[8] Existen muchas formas de medir la pobreza desde su aceptación común y si alteramos la definición, muchas más. Definir los indicadores con sus límites y alcances es parte de la etapa del planteo del problema

[9] La encuesta permanente de hogares posee un paquete no oficial para acceder a los datos mediante lenguaje R, aunque los datos también se encuentran en la página de INDEC.

Otra fuente de datos son los sistemas internos de las organizaciones, tales como los dispositivos de ingreso y salida, los softwares de administración interna, o los distintos sistemas de gestión. En este caso, el desafío es poder transformar las salidas que producen en información que nos sea significativa para tomar decisiones.

En relación a lo anterior, un desafío general es la transformación de los datos que carga un colaborador (por ejemplo, una planilla con proveedores y precios o información sobre la actividad curricular de un conjunto de estudiantes). En nuestro trabajo cotidiano, generalmente tenemos sistemas de control intuitivos para nosotros, pero difícil de resumir para un sistema. Al momento de elaborar las planillas para la carga y el control de procesos, es importante tener esto en cuenta para que el análisis no implique esfuerzos extraordinarios.

Por último, las fuentes de información primarias involucran entrevistas, encuestas, o distintos relevamientos que pueden hacerse en formato papel o con dispositivos digitales. Al momento de realizarlas, es importante considerar las sugerencias de los párrafos anteriores.

En caso de realizarse en formato papel, al momento de la carga pueden categorizarse las variables cualitativas (Por ejemplo, si en el relevamiento se le pregunta al usuario por su sexo, podemos registrar 0 si es hombre, 1 si es mujer y 9 si responde "otro"). Esto facilita el análisis posterior, la detección de errores y la curación (limpieza) de los datos.

Por otro lado, si el objetivo está vinculado al **Análisis, visualización y monitoreo de los datos**, al momento de plantear el problema se debe tener en cuenta que medidas se quieren describir, además de las distintas herramientas para la visualización y el monitoreo.

Si se trata de información de una población, entonces un análisis descriptivo será suficiente. Los porcentajes, o las medidas de la distribución de los datos cuantitativos (media, mediana, desvío) proveerá información precisa sobre el fenómeno o el indicador que se está evaluando.

Si se trata de datos de una muestra, entonces será necesario utilizar herramientas de inferencia estadística. De acuerdo a variabilidad y al tamaño de la muestra, puede calcularse intervalos más precisos sobre los parámetros que se desean estimar (sea un porcentaje, un promedio poblacional, etc.). Para poder generalizar las conclusiones a una población es importante garantizar que todas las unidades hayan tenido la misma chance de ser elegidas, entre otras cosas.

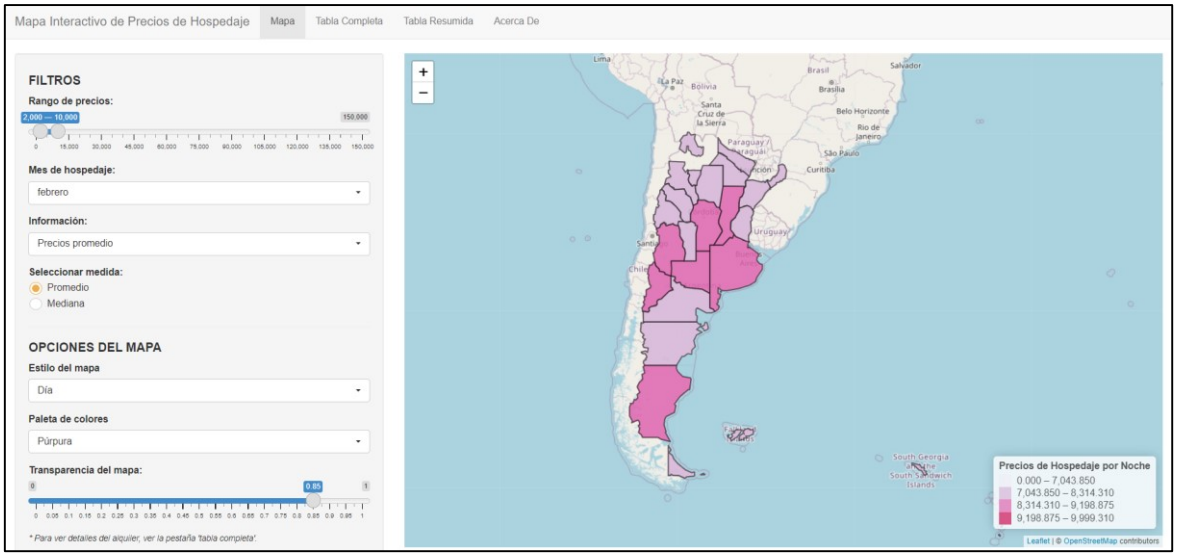
Con respecto a la visualización y al monitoreo de los datos, hay distintas herramientas que proveen, de manera gratuita, esquemas de visualización de datos conocidos en la jerga como *Dashboards* o tableros.

Los tableros contienen piezas visuales sobre los indicadores de interés. Así, los responsables de salud pueden monitorear, en vivo, los operativos de vacunación dada la densidad poblacional de una zona, o las

fuerzas de seguridad pueden observar el impacto de las políticas públicas y las estrategias de seguridad en la tasa de delincuencia.

Existen muchas formas de hacer los tableros, y todas involucran el uso de softwares o lenguajes de programación. Aunque el uso de softwares es más intuitivo, la mayoría son pagos y sus costos tienden a ser altos. Por otro lado, los lenguajes de programación limitan la construcción de tableros a los desarrolladores, aunque hay lenguajes sencillos como R y Python, con paquetes gratuitos que permiten construir los gráficos. En este último caso, la construcción de tableros no requiere costos adicionales además del programador, excepto que se desee publicar el tablero en un servidor para que toda la organización tenga acceso.

Figura 1 - Tablero de ejemplo que visualiza el precio promedio de hospedajes por provincia elaborado en R. Fuente: elaboración propia con datos del sitio Booking.



Si el objetivo es monitorear un fenómeno o un indicador, entonces los datos se deberán almacenar en un servidor al que el tablero deberá tener acceso. Cuando los datos se actualicen, se actualizarán las medidas resumen y la visualización de la información.

Por último, la mayoría de los tableros pueden almacenarse en servidores web. Esto significa que cualquier persona (bajo los protocolos de seguridad correspondientes) puede acceder a la información desde cualquier dispositivo o sólo en dispositivos específicos.

En relación a las **Estimaciones y simulaciones basadas en los datos** éstas aportan información para modelar matemáticamente un fenómeno de interés. Por ejemplo: el comportamiento de ciertos sectores económicos, las chances de que un estudiante discontinúe una trayectoria académica, o de que un turista seleccione una determinada ciudad para vacacionar habiendo invertido en publicidad una determinada cantidad de dinero. Estas estimaciones permiten construir escenarios posibles dada la información histórica, pudiendo planificar y ejecutar, por ejemplo, un proceso de distribución de energía, la asignación de recursos ante una emergencia, o las políticas de evaluación de los procesos de escolarización. También permiten

encontrar parámetros poblacionales (por ejemplo, el grado de satisfacción de todos los ciudadanos) dada una muestra, y asegurando los supuestos correspondientes.

Por último, las **predicciones basadas en los datos para predecir o prevenir** proveen información sobre los posibles futuros de un evento determinado, con un margen de error capaz de ser cuantificado. Los modelos y sus leyes se encuentran en los libros de estadística, aunque es importante no caer en las *estimaciones espurias.* Es decir, debe existir una teoría o un conjunto de teorías de alguna disciplina que explique la relación entre las dimensiones de un modelo. Por ejemplo, si mi objetivo es hacer inferencias sobre las causas que incrementan los salarios promedio de una provincia, entonces necesito un marco teórico que defina las posibles causas que incrementan los salarios. Sin esto, tal vez concluyamos que los ingresos aumentan cuando aumentan las habitaciones del hogar simplemente porque existe una asociación cuantitativa entre estas variables.

Además, debe tenerse en cuenta la presencia de *variables de confusión.* Es decir, posibles características que modifiquen el comportamiento del indicador bajo estudio y que no estamos teniendo en cuenta al momento del análisis. Por ejemplo, si le preguntamos a 300 mujeres si fuman o no, y del 50% que afirma no fumar falleció el 70% 20 años más tarde, puede caerse en la conclusión errónea que fumar alarga la vida. Sin embargo, si se tiene en cuenta la edad de las mujeres al momento de responder la encuesta, las conclusiones podrían ser otras.

Por último, tal vez desea prevenir un determinado fenómeno, como evitar algún comportamiento fraudulento o la deserción escolar. Existen técnicas específicas que permiten modelar estos comportamientos. Por ejemplo, los modelos de clasificación.

El aprendizaje supervisado y el aprendizaje no supervisado son áreas de estudio para la construcción y actualización recurrente de estos modelos. Así, si se sistematiza la extracción y la transformación de los datos, es posible actualizar de forma automática el modelo predictivo. Además, la Ciencia de Datos[10] trae consigo una batería de opciones para realizar los modelos y las simulaciones correspondientes.

Dentro de este campo disciplinar, el *Machine Learning* -aprendizaje automático- consiste en la construcción de modelos (ecuaciones) que permiten a las computadoras conocer la relación entre distintas dimensiones del universo de estudio y, en función de estos, realizar una devolución o establecer criterios de clasificación. En este sentido, al momento de construir los modelos para predecir o prevenir un fenómeno, es importante tener en cuenta quien es el receptor de la información generada: algunos modelos estadísticos no son suficientes para explicar la variabilidad de un fenómeno (por ejemplo, la inseguridad de una determinada zona puede estar dada por la presencia policial, las condiciones socioeconómicas y la luminaria,

[10] La Ciencia de Datos es un campo interdisciplinar relativamente nuevo, articula los conocimientos de la programación, la estadística y disciplinas específicas (en general Economía, pero pueden ser muchas otras) para construir modelos predictivos o de clasificación a través de algoritmos complejos.

aunque puede haber muchos factores que no conocemos y que explican, en gran parte, la inseguridad) pero su interpretación es intuitiva para las personas. En este caso, es preferible tener un modelo menos robusto, pero que sea posible de explicar. A esto se lo denomina *parsimonia*. Por otro lado, si el objetivo es realizar predicciones de forma automática, sin discutir las variables que inciden sobre el fenómeno bajo estudio, entonces puede ser preferible un modelo más complejo, aunque menos intuitivo. Ya que el objetivo en sí es predecir, no explicar. Así, podemos realizar transformaciones no lineales en las variables del modelo, establecer asociaciones (interacción) entre ellas, etc.

En consonancia, también se debe establecer si el aprendizaje será *supervisado* o *no supervisado*.

En el aprendizaje supervisado existe un indicador o variable objetivo y una o más variables explicativas. Es decir, hay una dimensión del universo que queremos explicar en función de otras. Por ejemplo, la probabilidad de que un estudiante del ciclo orientado deba repetir el año escolar dada su trayectoria en años anteriores, sus notas del año en curso y sus actividades extracurriculares. También, el costo promedio de una determinada obra pública dada la cantidad de trabajadores, los metros cuadrados, la participación relativa de los materiales y la zona.

En cambio, el aprendizaje no supervisado se basa en encontrar patrones entre las distintas unidades y agruparlas en base a sus características. Por ejemplo, agrupar los barrios de una ciudad en función de sus condiciones socioeconómicas, sanitarias, educativas y civiles, permite reorientar la política pública de forma horizontal, evitando realizar grandes ajustes en las medidas. Otro ejemplo consiste en agrupar ciudades a través de distintos indicadores (indicadores de desarrollo, indicadores de inversión pública, etc.) para evaluar qué "perfil" tienen las ciudades que mayor bienestar poseen.

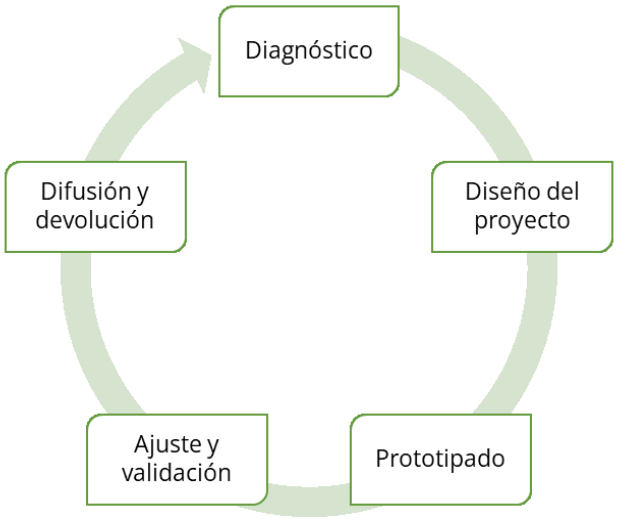
En conclusión, las preguntas que podemos respondernos en esta etapa son:

- ¿De dónde vamos a sacar los datos que necesitamos?
- ¿Qué herramientas informáticas y recursos técnicos necesitamos para realizar el registro de los datos?
- ¿Necesitamos un informe diario? ¿Semanal? ¿Mensual? ¿Trimestral? ¿Con qué herramientas lo realizamos?
- ¿Necesitamos un sistema de visualización? ¿Necesitamos monitorear los indicadores a diario?
- ¿Necesitamos realizar inferencia sobre una población? ¿Qué herramientas estadísticas utilizaremos?
- ¿Necesitamos estimar el indicador en un futuro no muy lejano?
- ¿Queremos agrupar unidades o pronosticar fenómenos?
- En caso de construir un modelo para predecir un indicador o prevenir un fenómeno ¿podemos automatizar su aprendizaje?
- ¿Qué infraestructura se debe generar para sistematizar los procedimientos?

3. Prototipado, ajuste y validación.

Una vez recolectados los datos y realizados los análisis necesarios (incluyendo el ajuste de un modelo, en caso de que corresponda) puede ponerse en funcionamiento el sistema de medición o realizarse los informes y su difusión correspondiente. Es importante contar con la devolución por parte de los actores que utilizan la información provista para tomar decisiones, y así realizar los ajustes necesarios, tal como se muestra en la siguiente figura.

Figura 2 - Etapas de un proyecto de sistematización de la información. En base a Martinez (2022) en Innovar con Ciencia de Datos en el Sector Público



El prototipado consiste en elaborar herramientas preliminares, que den respuesta de forma parcial al fenómeno que se está estudiando. Finalizada la primera versión del sistema de información, se difunde entre los integrantes del equipo y se recibe una retroalimentación.

El ajuste y la validación, por otro lado, consiste en realizar los cambios necesarios en la primera versión. Así, es importante difundir la versión definitiva con los integrantes del equipo, y documentar nuevos cambios si son necesarios.

De acuerdo al objetivo de la etapa 2, las herramientas pueden consistir en:

- El desarrollo de un sistema de recolección y sistematización de datos provenientes de fuentes propias o de terceros.
- La automatización de los reportes de los datos recolectados, exportable en formatos de lectura (Ej. PDF) o formato abierto (Ej. CSV).
- La construcción de un tablero para la visualización de los datos (Ej. Shiny, Power BI, Tableau, etc.).

- Los resultados de un estudio inductivo (inferencia) para generalizar las conclusiones a una población, bajo los criterios de validez externa.
- El reporte de un modelo estadístico que explique las causas de un determinado fenómeno.
- La construcción y puesta en marcha de un modelo de aprendizaje automático, con el fin de pronosticar un determinado escenario o prevenir un determinado fenómeno.

Por otro lado, el ajuste y la validación del modelo deberán realizarse bajo los criterios establecidos al momento del diseño del problema, y debe evitarse modificarlo a tal punto que la respuesta de la herramienta sea otra. Sí es importante evaluar la calidad y la claridad en la interpretación de los resultados de los análisis, su pertinencia (que un costo no de un valor negativo, por ejemplo).

Entonces, las preguntas a realizarse en esta etapa son:

- ¿La herramienta construida es útil para satisfacer las necesidades planteadas en la primera etapa?
- ¿Todos los integrantes del equipo son capaces de interpretar la información? (nótese que no nos preguntamos si somos capaces de llegar a las mismas conclusiones, puesto que esto depende de la herramienta).
- ¿Existe algo para mejorar en la herramienta construida?
- ¿Hay información en exceso que puede confundir la lectura de los reportes?

4. Difusión

Al momento de difundir los datos se debe definir la periodicidad de acuerdo a los costos y al comportamiento del indicador. Sosa Escudero (2020) cita a un conocido profesor de Macroeconomía, quien afirma que el Producto Bruto Interno se publica cada tres meses no sólo por los costos que se deben afrontar para su estimación, sino porque publicarlo con mayor frecuencia dada la volatilidad del indicador "alimentaría las ansiedades".

Además, se debe reconocer al público objetivo, y se debe identificar el objetivo de la difusión de los datos, además de su causa.

Los destinatarios de la información pueden ser los equipos de gestión, propios de cada institución o la comunidad a la que atiende. En el primero de los casos, y de acuerdo a la herramienta, podría ser necesario capacitar a los agentes que la utilizaran para tomar decisiones o realizar informes. En el segundo de los casos, la información debe ser resumida, consistente y debe estar clara la definición del indicador y, en grandes rasgos, los elementos para la construcción. También debe estar a disposición un manual con la información precisa sobre los procedimientos, cálculos, y la justificación de las decisiones llevadas a cabo.

También se debe garantizar el secreto estadístico y la correcta presentación de la información, evitando caer en escalas incorrectas o expresando la información de forma parcial. Si en un estudio sobre el bienestar

ciudadano se afirma que las chances de que un ciudadano tenga bienestar pleno son 2 en 10.000, mientras que poniendo canteras de agua esas chances aumentan a 4 en 10.000, la conclusión es que la cantidad de bienestar pleno en la ciudad… ¡Se duplica con las canteras! Aunque en términos absolutos, el incremento parece no ser tan significativo.

Otro ejemplo se presenta en el siguiente gráfico, donde la escala de la variable que identifica los días de marzo no está representada de forma equidistante (del día 1 pasa al 6 en la misma distancia que del 6 al 7).

Los instrumentos para comunicar las conclusiones pueden ser informes, tablas, gráficos, sitios interactivos, o tableros de información, como los desarrollados en la etapa 2.

En conclusión, las preguntas que debemos hacernos en esta etapa son:

- ¿Las publicaciones siguen las normas de confidencialidad de la información?
- ¿Quién es mi público objetivo?
- ¿En qué formato o tecnología realizaré la difusión de la información?
- ¿El manual metodológico es consistente?
- ¿La información es imparcial y entendible por todos en el público objetivo?
- ¿Con qué frecuencia actualizaré los informes?

Figura 3 - Un ejemplo de lo que no se debe hacer. Fuente: Material del Seminario "Introducción a la Estadística, Probabilidad e Inferencia" (UNR) citando a la Tapa del diario "El Cronista Comercial" – 28 de marzo de 2019.

Reflexiones finales

En el recorrido se planteó una serie de prácticas a tener en cuenta al momento de pensar en sistematizar la información de una organización, sea con el fin de obtener un diagnóstico específico, encontrar explicaciones de un evento (por ejemplo, las causas de la delincuencia) pronosticar un fenómeno (por ejemplo, las chances de fraude) o prevenir una acción (por ejemplo, la deserción). Además, se expusieron las tecnologías vigentes para automatizar este proceso, cualquiera sea el objetivo.

Por otro lado, se identificaron algunas nociones a tener en cuenta para el diseño muestral, la recolección, el tratamiento y análisis y su posterior difusión, sin la intención de ser una guía técnica. Cada situación amerita un abordaje particular, y las técnicas de muestreo, recolección, interpretación y difusión deberán ser específicas. El dominio de estas herramientas se encuentra en los textos de estadística, investigación cuantitativa de mercado, econometría, entre otros.

A su vez, como se observa en la figura 2, el ciclo del proyecto incluye una retroalimentación constante, lo que implica que el diseño de cualquier sistema de información no es lineal y mucho menos perfecto. De allí la noción de prototipo y ajuste. Además, es importante que los actores que participen para los distintos proyectos incluyan representantes de las distintas áreas que harán uso de esa información, principalmente en la etapa de diseño, donde se alinean las expectativas en el uso de las herramientas.

Por último, la recolección automática de los datos puede parecer una tarea engorrosa o costosa. Si bien es fundamental asegurar la calidad y disponibilidad de los datos, la mayoría de las instituciones públicas ofrecen sistemas de gestión administrativa, contable, de recursos humanos, etc. que proveen reportes más o menos legibles (a veces por estar codificados, otras veces por defecto). Por lo que el desafío hoy en día, excepto que se trate de datos ajenos a la organización, es la sistematización, y no su recolección.

BIBLIOGRAFÍA

- Martinez, M. V., Dumas, V. G., Sarabia, M., & Kisilevsky, I. F. (2022). *Innovar con Ciencia de Datos en el sector público* (1 ed.). Fundación Sadosky. Obtenido de https://innovacionpublicacondatos.fundacionsadosky.org.ar/descargar/HojaDeRuta.pdf
- Ozlak, O. (2013). *Gobierno abierto: hacia un nuevo paradigma de gestión política.* Red GEALC.
- Piccirilli, G. C., & Farías, C. M. (2015). Sistema de Información para la toma de decisiones y control de la administración pública. *Perspectivas de las Ciencias Económicas y Jurídicas*, 53-82.
- Salvador, M., & Ramió, C. (2020). Capacidades analíticas y gobernanza de datos en la administración pública como paso previo a la introducción de la inteligencia artifical. *Revista del CLAD Reforma y Democracia*, 7-36.
- Sosa Escudero, W. (2014). *Qué Es (y Qué No Es) la Estadística: Usos y Abusos de una Disciplina Clave en la Vida de Los Países y Las Personas.* El Cid Editor Incorporated.

- Sosa Escudero, W. (2019). *Big data: breve manual para conocer la ciencia de datos que ya invadió nuestras vidas.* Siglo Veintiuno Editores Argentina.
- Sosa Escudero, W. (2020). *Borges, big data y yo: guía nerd (y un poco rea) para perderse en el laberinto borgeano.* Siglo Veintiuno Editores.
- Ruggieri M. (2010) *Métodos estadísticos I.* Reimpresión. Rosario. UNR editora.