permitido innovar

Guías para transformar el Estado chileno

¿Cómo podemos desarrollar proyectos de ciencia de datos para innovar en el sector público?

Ministerio de Hacienda

Gobierno de Chile

Laboratorio de Gobierno

Coordinación CARLOS CARRILLO, IGNACIO PAIVA
Textos CARLOS CARRILLO, IGNACIO PAIVA, ERNA GÓMEZ, CATALINA GUTIÉRREZ, LAURA GONZÁLEZ, EDUARDO NAVARRO, MARÍA PAZ HERMOSILLA, GABRIELA DENIS, MARIANA GERMÁN **Edición** DANIELA HERRERA **Diseño gráfico y sistematización visual** MYRIAM MEYER **Fotografías** EQUIPO LABORATORIO DE GOBIERNO.

Equipo GobLab UAI
MARÍA PAZ HERMOSILLA, GABRIELA DENIS, MARIANA GERMÁN, JESÚS SANTORCUATO, VITA SALDÍAS, CLAUDIO ARACENA.

Equipo Laboratorio de Gobierno
ALEJANDRA GÓMEZ, CARLOS CARRILLO, CATALINA GUTIÉRREZ, CONSTANZA PÉREZ, DANIELA HERRERA, EDITHA FUENTES, EDUARDO NAVARRO, ELISA BREULL, ERNA GÓMEZ, FRAN GARRETÓN, FRANCISCA MOYA, FREMBERLING RAMOS, GIANCARLO SILLERICO, IGNACIO PAIVA, JAVIERA MIRANDA, LAURA GONZÁLEZ, LORENA TORRES, MYRIAM MEYER, RAÚL HENRÍQUEZ, RODRIGO ALBORNOZ, SEBASTIÁN ALTIMIRA, TOMÁS DINTRANS.

¿Cómo citar este libro?
Permitido Innovar: Guías para transformar el Estado chileno ¿Cómo podemos desarrollar proyectos de ciencia de datos para innovar en el sector público?. Laboratorio de Gobierno, Gobierno de Chile y Universidad Adolfo Ibáñez (2022).

Esta guía surge del trabajo colaborativo entre el Laboratorio de Gobierno del Ministerio de Hacienda y el GobLab UAI.

UNIVERSIDAD ADOLFO IBÁÑEZ

lab.gob.cl

Índice

Índice de herramientas

Podrás encontrar la versión digital de todas las herramientas descritas en esta Guía en el siguiente link: bit.ly/3y70knB

Prólogo

El diseño e implementación de políticas públicas requiere **no solo tener en el centro a las personas, sino además, tomar decisiones basadas en evidencia.**

Muchas veces nos vemos enfrentados a buscar soluciones sin tener la información específica disponible para adoptar una definición. **Es por ello que explorar la conexión entre los datos y la toma de decisiones es un deber para los países y los Estados.** La evidencia que nos deja el análisis de grandes volúmenes de registros, como los que se manejan en el Estado, es una responsabilidad enorme y requiere **una gestión eficaz que asegure la generación de políticas públicas que permitan construir un Estado cada vez más presente y que responda ágilmente a las demandas ciudadanas.**

En esta era de la información tenemos la oportunidad de hacer de los datos una herramienta poderosa para contribuir a **recuperar la confianza de las personas en las instituciones públicas.** Es importante incorporar este tema en el debate público y observar muchas de nuestras discusiones con esa perspectiva; los países que destacan por la confianza en sus instituciones, son justamente quienes **diseñan políticas públicas basadas en evidencia, lo que incluye usar información proveniente de datos que genera el propio Estado.**

La ciencia de datos brinda capacidades para crear políticas públicas oportunas y coordinadas entre los organismos, lo que sumado a un enfoque de innovación pública basada en datos y en el desarrollo de las capacidades en funcionarias y funcionarios públicos, nos permitirá llegar a soluciones validadas para desafíos públicos urgentes.

El Estado chileno ha comenzado a dar los pasos necesarios para que la ciencia de datos se convierta en un proyecto concreto, instalándolo con un enfoque de innovación dentro del sector público. Así, puede generar sinergias con el diseño de soluciones, centrarse en las personas usuarias y en la eficiencia en el uso de los recursos públicos, además de implementar la Ley de Transformación Digital y las iniciativas de la Agenda de Modernización del Estado.

Esta quinta Guía Permitido Innovar, elaborada por el Laboratorio de Gobierno del Ministerio de Hacienda, en colaboración con la Universidad Adolfo Ibáñez, es una invitación para aquellas funcionarias y funcionarios del Estado que deseen **explorar esta conexión entre el conocimiento invaluable que nos brindan los datos y nuestro mandato por crear más y mejores políticas públicas centradas en las personas.**

Mario Marcel Cullell
Ministro de Hacienda

¡Hola!

Les presentamos una nueva ***Guía Permitido Innovar: ¿Cómo podemos desarrollar proyectos de ciencia de datos para innovar en el sector público?*** En esta oportunidad, la quinta versión de las guías del Laboratorio de Gobierno para transformar el Estado chileno, contó con la colaboración de la Escuela de Gobierno de la Universidad Adolfo Ibáñez, específicamente el GobLab UAI, la cual fue creada a partir del modelo metodológico de innovación pública del Laboratorio de Gobierno del Ministerio de Hacienda, dándole una interpretación desde la ciencia de datos.

Esta es una disciplina nueva y amplia, que puede ir desde análisis sencillos hasta interpretaciones como el *machine learning*. En algunos casos, puede volverse complejo avanzar en proyectos que requieran esta mirada; es por esto que el documento está compuesto por pasos y herramientas, abordables para quien la utilice en el desarrollo de proyectos, con el objetivo de convertirse en una guía para funcionarias y funcionarios que deseen explorar la ciencia de datos dentro de sus instituciones.

Esta Guía está dirigida a funcionarias y funcionarios públicos que puedan ser gestores de un proyecto de innovación dentro de sus instituciones. Para ponerla en práctica se deben tener conocimientos básicos de estadística y/o análisis de datos, conocer las limitaciones en su uso, y contar con un genuino interés en resolver los problemas que afectan a las personas usuarias.

Queremos reconocer el virtuosismo de la relación entre el Estado y la Academia, donde la retroalimentación en conocimientos es clave para generar valor público para las personas y avanzar hacia servicios públicos más presentes, actualizados y de alto nivel.

Como cada Guía Permitido Innovar, esta edición fue testeada por distintos perfiles, como personas expertas de la academia y funcionarias y funcionarios, así como también desde diferentes perspectivas; en lenguaje claro, en su estructura y visualización, además de las áreas de ética e implementación de proyectos. Agradecemos su disposición en la construcción de esta Guía.

Queremos invitarles a que pongan en práctica esta Guía y comenzar a utilizar la ciencia de datos como una herramienta más para mejorar y complementar la labor que día a día ejercemos como funcionarias y funcionarios para construir un mejor Estado desde la innovación.

Catalina Gutiérrez Ricci
Directora Ejecutiva (s)
Laboratorio de Gobierno

María Paz Hermosilla
Directora
GobLab UAI

Introducción

La ciencia de datos es un campo interdisciplinario que emplea un amplio repertorio de técnicas para generar valor a partir de los datos *(Provost & Fawcett, 2013; Van Der Aalst, 2016)*. **Podríamos definir su objetivo como "la producción de creencias informadas por datos, para ser utilizadas como base para la toma de decisiones"** *(Igual & Seguí, 2017, p.2)*. Su popularidad ha aumentado rápidamente en los últimos años debido a la existencia de *softwares* libres para su implementación, sus sólidos fundamentos científicos, la gran disponibilidad de datos que pueden ser utilizados, entre otros.

El uso de la ciencia de datos no es una práctica instalada en las instituciones públicas, aunque se cuente con abundantes datos. Ésta se vuelve necesaria cuando existe el recurso humano para el tratamiento de datos, se dispone de apoyo institucional, se cuenta con soporte tecnológico, se poseen recursos financieros y, por sobre todas las cosas, se debe resolver un problema que aqueja a las personas usuarias del servicio.

En general, en los proyectos de ciencia de datos, el ejercicio que se realiza es transformar los datos a información y luego a conocimiento *(Davenport & Prusak, 1998)*. Recordar que los **datos** son un conjunto de rasgos sobre un hecho real o una observación. En ese sentido, no entregan razones sobre el por qué de las cosas ni son orientativos para la acción. **En cambio, si los datos se contextualizan, categorizan, calculan, corrigen o condensan podemos hablar de información, la que posee significado, relevancia y propósito.** Luego, si se compara, se obtienen consecuencias, se conecta y se conversa acerca de lo producido, podemos hablar de la obtención de **conocimiento**. Este se crea entre las personas, como una mezcla de experiencias, valores, información y "saber hacer" que es útil para la acción y la toma de decisiones.

Por otro lado, en los proyectos de innovación pública ocurre algo similar, dado que también buscan generar una mejora sustantiva, ya sea en términos de los procesos de trabajo, como en su relación con personas u otras instituciones usuarias o beneficiarias. **Una innovación puede tratarse de ideas nuevas que hayan sido generadas tanto por la propia como por otra institución (pública, privada o del tercer sector), pero que en cualquier caso deben haber pasado por un proceso demostrable de ajuste a las características específicas de la entidad** (Laboratorio de Gobierno, 2022).

>> TERCER SECTOR
Conjunto de instituciones privadas, ubicadas fuera del Estado y del Mercado, que se ocupan de entregar servicios y prestaciones predominantemente de carácter social.

Cabe destacar que la relación entre ciencia de datos e innovación pública ha sido ampliamente aprovechada en proyectos del Estado chileno:

- En el *Sistema de Admisión Escolar*[1] (SAE) del Ministerio de Educación que utiliza los datos para asignar estudiantes de educación escolar en los colegios de su preferencia.
- El *Portal Geomin*[2] del Servicio Nacional de Geología y Minería que permite visualizar y consultar servicios de cartografía geológica, minería y sus metadatos.

1. Disponible en sistemadeadmisionescolar.cl
2. Disponible en portalgeominbeta.sernageomin.cl

- El *Sistema Agromet*[3] del Ministerio de Agricultura que integra en un portal la información proveniente de varias redes previamente existentes, para entregarla de manera uniforme, consistente y con cobertura a lo largo de todo el país.
- El *Registro Social de Hogares (RSH)*[4], del Ministerio de Desarrollo Social y Familia, sistema de información construido con datos aportados por el hogar y bases administrativas que posee el Estado, cuyo fin es apoyar los procesos de selección de beneficiarios de un conjunto amplio de subsidios y programas sociales.

En un esfuerzo de sistematizar el uso de ciencia de datos por parte de instituciones públicas de Chile, la Universidad Adolfo Ibáñez ha construido el repositorio de *Algoritmos Públicos*[5], que a la fecha cuenta con 50 proyectos identificados donde se ha automatizado la toma de decisiones a través de la ciencia de datos, como en la selección de subsidios de arriendo del Ministerio de Vivienda y Urbanismo, la asistencia virtual de Servicio Electoral, o la gestión de licencias médicas del Fondo Nacional de Salud y la Superintendencia de Seguridad Social.

Del mismo modo, el Laboratorio de Gobierno del Ministerio de Hacienda pone a disposición de la ciudadanía el listado de *Casos de Innovación Pública*[6], donde es posible conocer experiencias que emplean ciencia de datos como *WhatsApp Mujer* con el Ministerio de la Mujer y Equidad de Género, *Un nuevo Fono Drogas* en colaboración con el Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol, y *Dame esos 5* en conjunto con el Ministerio de Educación.

En otros países también existen iniciativas que tienen proyectos documentados. Por ejemplo, la fundación *Data Science for Social Good,* que tiene como misión fomentar el uso de la ciencia de datos para obtener impactos sociales positivos, ha documentado múltiples iniciativas de este estilo en su página web[7]. Además, esta fundación creó la plataforma *Solve for Good*[8], donde cualquier persona o institución puede registrarse y publicar su proyecto de ciencia de datos, ya sea para compartirlo con otros, recibir comentarios, o para trabajar colaborativamente con voluntarios de todo el mundo.

En esta *Guía Permitido Innovar: ¿Cómo podemos desarrollar proyectos de ciencia de datos para innovar en el sector público?,* desarrollada por Laboratorio de Gobierno en colaboración con la Universidad Adolfo Ibáñez, se presenta una metodología de trabajo para resolver problemas públicos empleando ciencia de datos. **Para su construcción se recuperaron los principios más relevantes de la ciencia de datos y la innovación pública, lo que da lugar a una metodología centrada en las personas usuarias, orientada a resolver problemas prácticos, basada en evidencia, y con un fuerte componente de trabajo colaborativo y multidimensional.**

3. Disponible en agromet.cl
4. Disponible en registrosocial.gob.cl
5. Disponible en algoritmospublicos.cl
6. Disponible en lab.gob.cl/casos/list
7. Disponible en dssgfellowship.org/projects
8. Disponible en solveforgood.org

¿Qué elementos de ética y seguridad debemos considerar?

- Privacidad de los datos
- Sesgos y justicia en ciencia de datos
- Transparencia con la ciudadanía

Antes de empezar, corresponde recalcar que cada paso descrito en esta Guía para la gestión y desarrollo de un proyecto de ciencia de datos, **debe considerar y validar el cumplimiento de aspectos éticos, por lo que instamos a cuestionar y revisar constantemente.** En ciertos casos, esto será crítico para respetar aspectos normativos que involucran al proyecto en distintas etapas o para evitar la afectación de derechos fundamentales de las personas, incluyendo el de la protección de datos personales. Particularmente, debemos proteger la privacidad de los datos, velar por la justicia del algoritmo y ser lo más transparentes y claros con la información generada, dentro de los márgenes permitidos. Con respecto al último punto, el Laboratorio de Gobierno dispone de una *Guía de Lenguaje Claro*[9] que busca generar una comunicación simple, clara y efectiva entre el Estado y la ciudadanía.

>> **JUSTICIA DEL ALGORITMO**
Para efectos de esta Guía entendemos este concepto desde la definición de discriminación arbitraria de la Ley N° 20.609, que establece medidas contra ella. Por lo demás, la justicia se puede definir estadísticamente con diferentes métricas y, en otros más complejos se recomienda complementar con la opinión de personas expertas. Para mayor detalle sugerimos revisar el estudio publicado por Verma y Rubin el 2018 en el que se recopilan y explican las definiciones más destacadas de justicia algoritmica (Verma y Rubin, 2018).

>> **ALGORITMO**
Conjunto ordenado y finito de operaciones (o instrucciones) que permite hallar la solución de un problema (realizar un cómputo, procesar datos y llevar a cabo otras tareas o actividades).

Privacidad de los datos

La protección de datos personales está reconocida como derecho fundamental en la Constitución Política de la República de Chile vigente al 2022 (Art. 19 N° 4)[10] y regulado a nivel legal principalmente a través de la Ley N°19.628[11] sobre protección de la vida privada, **la que dispone expresamente que las instituciones públicas pueden utilizar datos personales siempre y cuando su uso esté dentro de materias de su competencia, resguardando la confidencialidad de la información y solicitando el consentimiento previo, expreso e informado del titular cuando corresponda.** Los *datos personales*, definidos por la citada ley, son "aquellos relativos a cualquier información concerniente a personas naturales identificadas o identificables". Bajo esta definición, son datos personales los nombres, apellidos, números telefónicos, documentos de identidad (RUN, RUT, Pasaporte), información financiera, legal, georreferenciada, información biométrica, entre otros. El tratamiento de datos personales debe respetar los derechos de las y los titulares de los datos, resguardando su confidencialidad. Dado esto, es un deber tomar todas las medidas de protección y resguardo para que esta información no se filtre, ni sea posible reidentificar a las personas y comprometerlas en alguna medida.

>> **REIDENTIFICAR**
La reidentificación es el análisis de observaciones o ficheros anonimizados con el fin de identificar a personas específicas a partir de ellos. Por ejemplo, se descubrió que el 87% (216 millones de 248 millones) de la población de Estados Unidos había declarado caracteristicas que probablemente les hacían únicos basándose únicamente en el ZIP de 5 dígitos, el sexo y la fecha de nacimiento (Sweeney, 2000).

Cabe destacar la existencia de un subconjunto de los datos personales conocidos como *datos sensibles* o especialmente protegidos "que se refieren a las características físicas o morales de las personas o a hechos o circunstancias de su vida privada o intimidad". Su tratamiento está restringido, y solo se permite cuando: (i) Otras leyes autoricen su uso, como por ejemplo a prestadores de acciones de salud, conforme a lo establecido en la Ley N° 20.584, que regula los derechos y deberes que tienen las personas en relación con acciones vinculadas a su atención de salud; (ii) si los titulares otorgan su consentimiento y (iii) para determinar u otorgar beneficios de salud a sus titulares. Estos datos son aquellos referidos a opiniones políticas, convicciones religiosas, salud mental o física, entre otros ejemplos enunciados en la norma.

9. Disponible en innovadorespublicos.cl/documentation/publication/49
10. Disponible en bcn.cl/2kdfp
11. Disponible en bit.ly/3TCCpF7

Para impedir la reidentificación, sugerimos anonimizar los datos, es decir, "transformar los datos individuales de las unidades de observación, de tal modo que no sea posible identificar sujetos o características individuales de la fuente de información, preservando las propiedades estadísticas en los resultados" *(INE, 2022)*. Es posible emplear técnicas adicionales para reducir la probabilidad de revelar información sobre individuos, empresas u otras organizaciones. Los métodos de control de divulgación estadística (en inglés *Statistical Disclosure Control*, o SDC) minimizan el riesgo de divulgación a un nivel aceptable mientras liberan tanta información como sea posible.

Para definir hasta qué punto se deben intervenir los datos, primero es necesario medir el riesgo de divulgación de datos. Un indicador que se utiliza bastante es el *k-anonimato* (*k-anonymity* en inglés), que indica cuántas observaciones comparten los mismos valores en un conjunto de variables que permiten la identificación de las personas *(INE, 2021, p.16)*. Si se decide que es necesario intervenir, se pueden emplear métodos perturbativos y/o no perturbativos. **Los primeros falsifican los datos antes de la publicación, al introducir un elemento de error a propósito por razones de confidencialidad. En cambio, los métodos no perturbativos reducen la cantidad de información liberada por supresión o agregación de datos.**

Para mayor detalle sobre control de la divulgación estadística se sugiere revisar *Guía para el Control de Divulgación Estadística en Microdatos*[12] del Instituto Nacional de Estadística (2021), la que contiene procedimientos y actividades para controlar la divulgación de conformidad a lo establecido en la referida Ley N° 19.628; en la Ley N° 17.374[13], de secreto estadístico; y en la demás normativa aplicable. Dos documentos recomendables para conocer experiencias extranjeras en la materia, son la *Guía para la Anonimización de bases de datos en el Sistema Estadístico Nacional*[14] del Departamento Administrativo Nacional de Estadística de Colombia (2018), y el *Manual de Control de Divulgación Estadística*[15] de la Unión Europea (2010).

Sesgos y justicia en ciencia de datos

En la ciencia en general es fundamental saber que pueden existir sesgos, es decir, que pueden haber errores sistemáticos que condicionan cierta información en alguna dirección. Esto podría implicar que los resultados sean injustos y, en algunos casos, la situación de grupos vulnerados se replique y se agrave. **Identificar estos sesgos y saber cómo abordarlos para que el modelo se implemente sin consecuencias negativas en la ciudadanía es una tarea especialmente compleja en política pública.**

>> MODELO
Esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja, como la evolución económica de un país, que se elabora para facilitar su comprensión y el estudio de su comportamiento.

12. Disponible en bit.ly/3SBNfdd
13. Disponible en bcn.cl/2nlkk
14. Disponible en bit.ly/3zgJrrb
15. Disponible en bit.ly/3D0yWsQ

Aquí se muestran distintos tipos de sesgos que pueden afectar un proyecto de datos y cómo pueden ser abordados para evitar sacar conclusiones erróneas:

EL SESGO MUESTRAL

Ocurre cuando la muestra seleccionada excluye sistemáticamente a un grupo de interés. Por ejemplo, el uso de una base de datos que considera mayoritariamente personas que declararon su renta voluntariamente para un proyecto que busca predecir la probabilidad de declarar en la siguiente Operación Renta. Este sesgo se puede corregir (i) ajustando el peso de los grupos menos representados de acuerdo al peso que le corresponde en la población, (ii) consiguiendo los datos de los grupos subrepresentados y no representados o (iii) cambiando los objetivos del proyecto limitándolo a quienes declaran su renta voluntariamente.

EL SESGO DE MEDICIÓN

Ocurre cuando el instrumento de medición induce a un sub o sobre reporte sistemático. Por ejemplo, un sensor para medir el oleaje de un sector marítimo que no esté calibrado y, por lo tanto, entrega un subreporte del nivel de las mareas. Este sesgo se puede corregir ajustando el instrumento de levantamiento –ya sea un sensor o las preguntas de una encuesta– o el modo de levantamiento de información.

EL SESGO POR MODELAMIENTO

Ocurre cuando el modelo favorece o perjudica a uno o más individuos de manera infundada o no apropiada *(Friedman & Nissenbaum, 1996)*. Por ejemplo, un algoritmo que entregue becas universitarias a un grupo por sobre otro cuando esto no corresponda. Este sesgo se puede corregir desplegando auditorías algorítmicas para evaluar la justicia del modelo, o ejecutando un piloto para medir posibles efectos negativos que se den en el desarrollo del modelo. No es trivial elegir las variables adecuadas para que el algoritmo genere valor y, al mismo tiempo, asegure justicia social. Se sugiere el uso de herramientas gratuitas para realizar auditorías algorítmicas, que permiten comprobar sesgos de modelamiento, evaluar paridad estadística, paridad racial, paridad en falsos positivos y en falsos negativos, tales como *Aequitas*[16], *Fairness 360*[17] o *What If*[18].

Cabe destacar que los elementos que se han expuesto sobre ética y seguridad en el uso de datos atraviesan íntegramente cualquier proyecto de ciencia de datos, por lo que se deben considerar a lo largo de todo su desarrollo.

16. Disponible en bit.ly/3Fkqibx
17. Disponible en bit.ly/3Fb5W4y
18. Disponible en bit.ly/3zgKtDz

Transparencia con la ciudadanía

Para que el modelo sea efectivamente utilizado debe poseer licencia social, es decir, que la ciudadanía "acepte la implementación de la herramienta" *(Data Futures Partnership, 2017)*. **Uno de los puntos claves para lograrlo es la transparencia tanto del impacto, esperado y alcanzado, como de las variables que lo explican.** Un modelo fácil de interpretar es un buen punto de partida para crear una narrativa y confianza en la población, pero cuando los datos y el objetivo tienen demasiada complejidad, las autoridades de las instituciones públicas generalmente deben encontrar maneras de explicarlo a la ciudadanía. Solo en algunas ocasiones no se recomienda comunicar ciertos aspectos del proyecto, por ejemplo si puede afectar la seguridad nacional o la efectividad de un modelo de fiscalización.

Una buena práctica es crear mecanismos de retroalimentación con las personas usuarias para comprender las principales dudas sobre la herramienta, de tal manera de ajustar los contenidos comunicacionales. Un buen ejemplo de participación ciudadana para favorecer la transparencia son los Consejos de la Sociedad Civil (COSOC), mecanismo consultivo y autónomo creado a través de la Ley N° 20.500[19], sobre asociaciones y participación ciudadana en la gestión pública. Los COSOC deben ser convocados por los órganos de administración del Estado, y están conformados por representantes de asociaciones sin fines de lucro que tengan relación con la competencia del órgano respectivo. Sus sesiones pueden tratar temas de diversa índole, desde las definiciones estratégicas para la institución hasta reportar el avance de un proyecto en particular.

Para conocer más en detalle sobre las consideraciones éticas y legales de su potencial proyecto de ciencia de datos, junto con ejemplos concretos y buenas prácticas, recomendamos consultar la *Guía de Formulación Ética de Proyectos de Ciencia de Datos*[20] publicada por la División de Gobierno Digital del Ministerio Secretaría General de la Presidencia y la Universidad Adolfo Ibáñez en 2022.

19. Disponible en bit.ly/3FjbzNX
20. Disponible en bit.ly/3Fg2OEo

¿Qué metodología utilizamos?

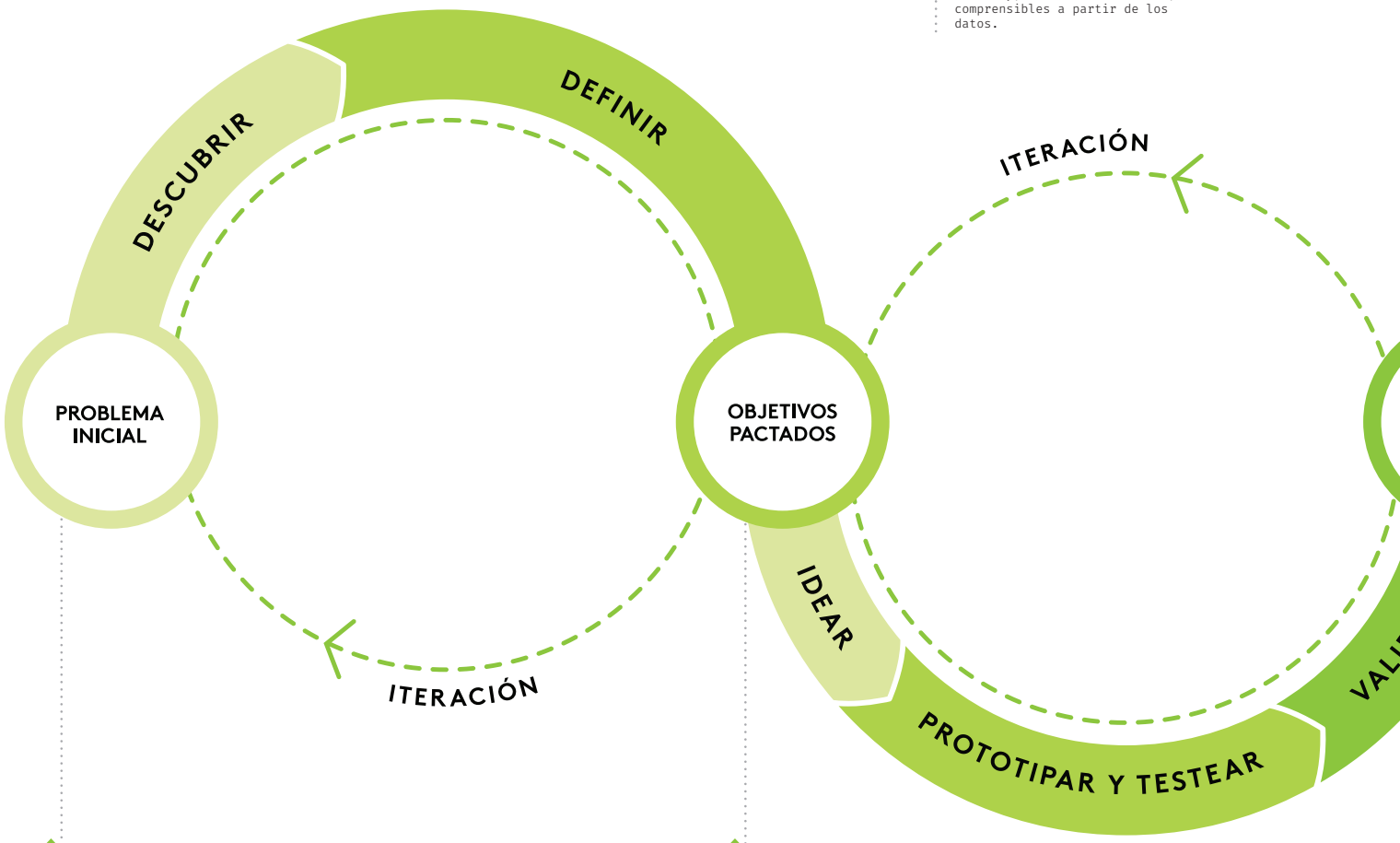
- Investigación del problema
- Diseño de propuestas de solución
- Desarollo del piloto
- Implementación de la solución

Esta Guía propone una interpretación de la metodología de innovación pública desarrollada por el Laboratorio de Gobierno desde la perspectiva de la ciencia de datos. Para enriquecer esta metodología, se han tomado aprendizajes de otras relacionadas a la ciencia de datos. Por ejemplo, de *CRISP-DM* utilizamos la importancia de un proceso iterativo en proyectos de ciencia de datos, de *KDD* el efecto final de generación de conocimiento y de *IBM* la importancia de comenzar con una pregunta que debe ser respondida, lo que en este caso se entiende como la importancia de la definición del problema. A modo general, todas las metodologías aportan también con los nombres y algunos de los pasos que utilizamos en esta Guía.

>> CRISP-DM: CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING
El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases: (i) Comprensión del negocio; (ii) Comprensión de los datos; (iii) Preparación de los datos; (iv) Fase de Modelado; (v) Evaluación; (vi) Implementación. El modelo de CRISP-DM es flexible y se puede personalizar fácilmente, lo que permite crear un modelo de minería de datos que se adapte a las necesidades concretas.

>> KDD: KNOWLEDGE DISCOVERY IN DATABASES
Apunta a procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil en ellos, de esta manera permitirá a la persona usuaria el uso de esta información valiosa para su conveniencia. Es el proceso de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos.

Figura 1: Fases de un proyecto de ciencia de datos.



FASE 1
INVESTIGACIÓN DEL PROBLEMA

DESCUBRIR
Paso 1: Conformar el equipo del proyecto
Paso 2: Describir el problema
Paso 3: Analizar la prefactibilidad del proyecto
Paso 4: Identificar actores y sus actividades
Paso 5: Mapear datos

DEFINIR
Paso 6: Definir los objetivos del proyecto

FASE 2
DISEÑO DE PROPUESTAS DE SOLUCIÓN

IDEAR
Paso 7: Seleccionar y alinear tipos de análisis de datos

PROTOTIPAR Y TESTEAR
Paso 8: Obtener y cargar los datos
Paso 9: Transformar los datos
Paso 10: Realizar análisis exploratorio de datos
Paso 11: Desarrollar y ajustar el modelo

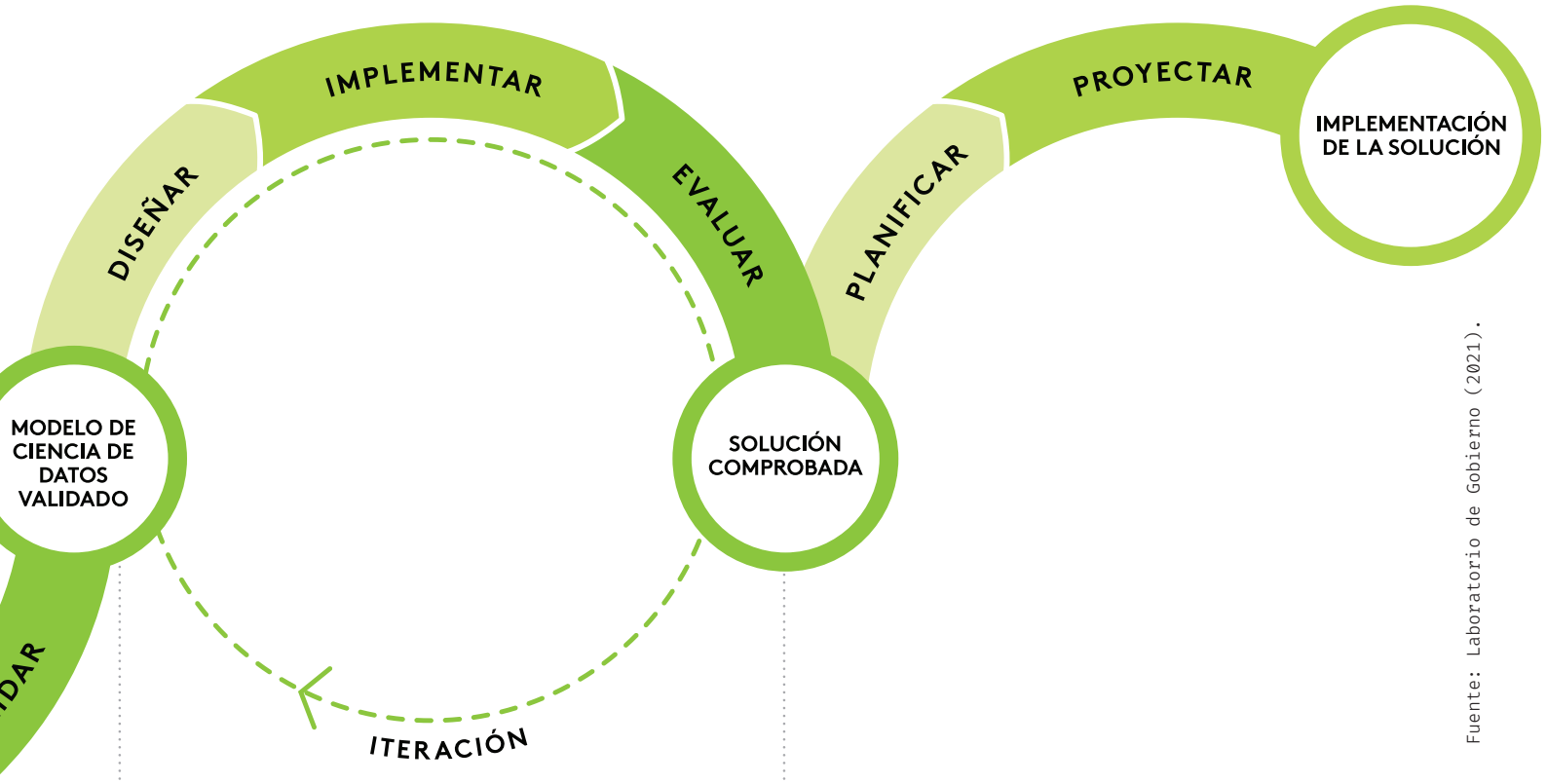
VALIDAR
Paso 12: Validar modelo
Paso 13: Generar conclusiones

>> IBM: INTERNATIONAL BUSINESS MACHINES
La Metodología Fundamental para la Ciencia de Datos consta de 10 etapas que forman un proceso iterativo para el uso de datos para descubrir nuevos conocimientos. Esta metodología tiene algunas similitudes con las metodologías reconocidas para la minería de datos, pero pone el énfasis en varias de las nuevas prácticas en la ciencia de datos, como el uso de grandes volúmenes de datos, la incorporación de la analítica de texto en el modelado predictivo y la automatización de algunos procesos.

>> SOLUCIÓN(ES)
"Acción y efecto de resolver una duda, dificultad o problema." Es diferente de un modelo puesto que abarca más aspectos, tales como la gestión del cambio y la sistematización de procesos involucrados.

Así, definimos la siguiente metodología que establece una orientación de cómo trabajar durante todas las fases de un proyecto de ciencia de datos (ver figura 1), con el objetivo de lograr la mayor eficiencia posible en resolver las problemáticas e implementar las mejores soluciones para las instituciones y sus personas usuarias.

La metodología asume la recopilación previa de algunos antecedentes que conforman un problema inicial, el cual se define como prioritario para las instituciones públicas y que tentativamente podría resolverse mediante la ciencia de datos. Estos antecedentes se consideran algo dado, por lo que no constituyen un paso propiamente tal en esta Guía.



Fuente: Laboratorio de Gobierno (2021).

FASE 3
DESARROLLO DEL PILOTO

DISEÑAR
Paso 14: Diseñar evaluación de impacto

IMPLEMENTAR
Paso 15: Implementar el piloto de la solución

EVALUAR
Paso 16: Evaluar resultados del piloto

FASE 4
IMPLEMENTACIÓN DE LA SOLUCIÓN

PLANIFICAR
Paso 17: Planificar el despliegue

PROYECTAR
Paso 18: Monitorear el desempeño
Paso 19: Robustecer el modelo

Este proceso en particular sirve para llevar a cabo proyectos de ciencia de datos en el sector público mediante 19 pasos organizados en cuatro fases. La primera fase se enfoca en comprender en profundidad el problema para identificar riesgos, aumentar las probabilidades de éxito y evaluar si la ciencia de datos es la disciplina más adecuada para resolverlo. La segunda fase se enfoca en el diseño del modelo de ciencia de datos, la que termina con un modelo de ciencia de datos validado. En tercer lugar se incorpora una fase centrada en desarrollar un piloto del modelo basado en metodologías experimentales que permita obtener evidencia de sus resultados en un contexto real. Finalmente, el proyecto se cierra con una fase de implementación y escalamiento de la iniciativa.

Dentro de cada una de las fases existe un trabajo iterativo permanente e intensivo entre pasos, con múltiples avances y retrocesos. Así mismo, cada paso puede realizarse más de una vez si fuese necesario, como podría ocurrir al desarrollar un modelo inicial que se va ajustando y validando hasta obtener el resultado deseado. Esta iteración acaba cuando se logran acuerdos significativos en el equipo de proyecto y se avanza a la siguiente fase.

Entre fases hay una relación secuencial, cuyo propósito es que el proyecto avance hacia su implementación. En este sentido, el retorno a una fase anterior debe estar acotado a casos extremadamente excepcionales, como podrían ser efectos no deseados de la implementación del modelo, promulgación de leyes que obliguen a modificar decisiones clave del proyecto, recortes presupuestarios no previstos, entre otras.

Si bien los tiempos de extensión de los proyectos son variables, esta metodología propone un enfoque ágil, que contemple fases de desarrollo rápidos, iterativos y que permitan el avance a partir de productos concretos. En este sentido, se propone que la duración de la Fase 1 sea de aproximadamente cuatro semanas. La Fase 2 pudiera considerar una extensión promedio de seis semanas, y la Fase 3 de cuatro semanas de trabajo, con un margen de flexibilidad asociado a la cantidad de iteraciones, información disponible en la institución, y las condiciones de trabajo en general. La fase de implementación no tiene un límite definido ya que depende del alcance de cada proyecto.

Sobre la base de su experiencia aplicada en proyectos en el sector público, el Laboratorio de Gobierno ha desarrollado un marco conceptual que define diversos ámbitos en los cuáles es posible desarrollar innovación pública, desde una perspectiva sostenible en el tiempo.

A partir de este modelo, se desprenden 12 tipos de innovación que son útiles como estructura de trabajo a considerar para el diseño e implementación de innovaciones.

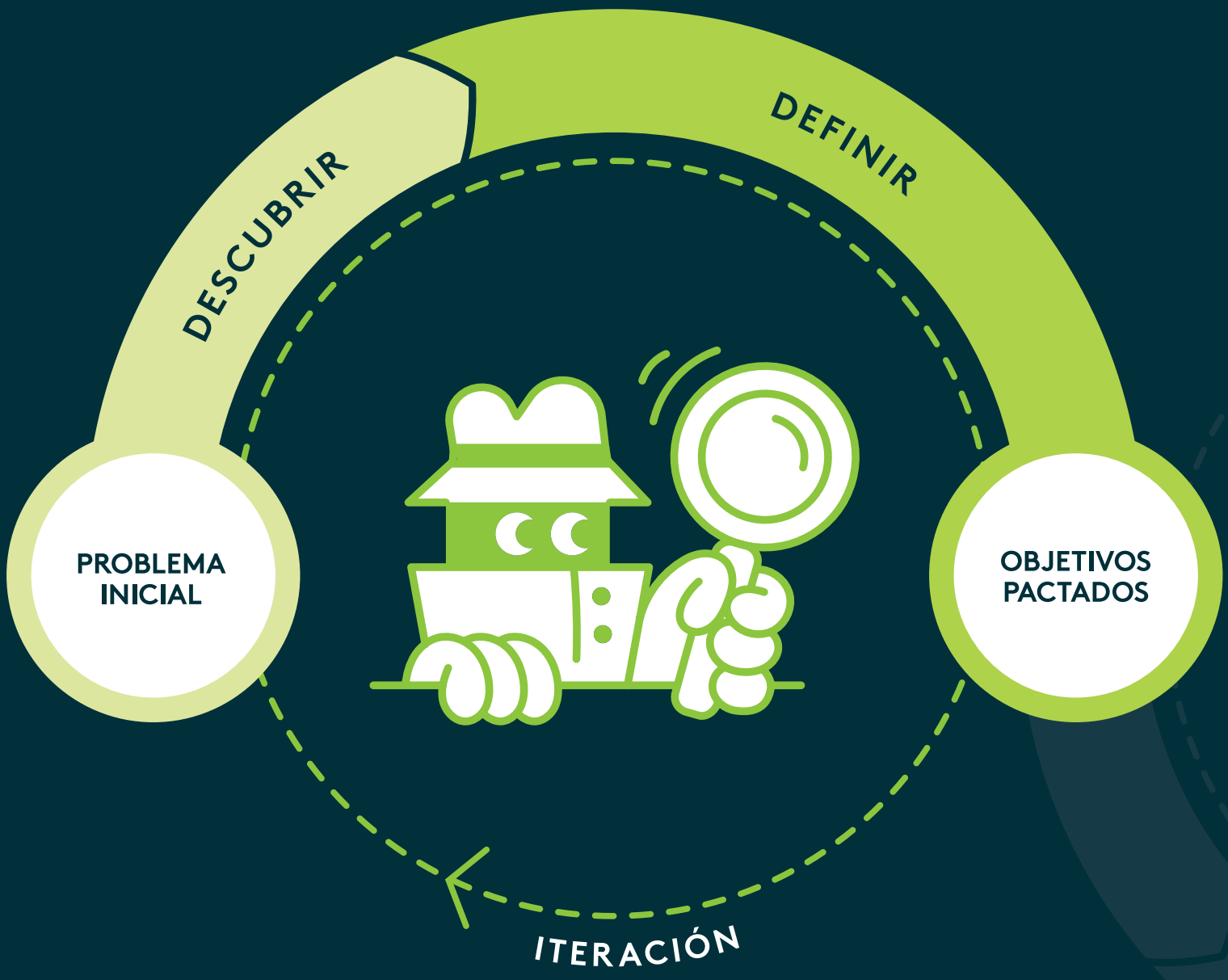
Esta Guía se concentra en el tipo de innovación “Gestión y Uso de datos”, entregando pasos simples y aplicables. Asimismo, considera aspectos relacionados con los “Procesos” y la “Tecnología”, todos pertenecientes al ámbito de Operación.

No obstante, las temáticas que puede abordar un proyecto de ciencia de datos son diversas, por lo que pudiera estar enfocado a mejorar el modelo de atención, el diseño organizacional, el comportamiento de personas usuarias, entre otras, según lo defina la institución. Para conocer más detalles de esta tipología, se sugiere consultar la publicación *Otro Ángulo*[21] y las videoclases[22] elaboradas por el Laboratorio de Gobierno.

21. Disponible en lab.gob.cl/otro-angulo
22. Disponibles en lab.gob.cl/videoclases

Figura 2: Tipos de Innovación y Ámbitos de Cambio.

Fuente: Laboratorio de Gobierno (2021).


DESCUBRIR
DEFINIR
PROBLEMA INICIAL
OBJETIVOS PACTADOS
ITERACIÓN

Fase 1
Investigación del problema

Productos

Objetivos para desarrollar un proyecto de innovación que utilice ciencia de datos, considerando equipo, actores relacionados e identificación de datos.

Esta fase comienza con un problema inicial que, aparentemente, requiere ser abordado mediante la ciencia de datos y que moviliza los pasos que vienen a continuación.

El objetivo de esta fase es obtener un conocimiento profundo del problema que afecta a las personas usuarias, y poner a prueba la idea de que la ciencia de datos es la forma de encarar las soluciones.

El cierre de esta fase ocurre cuando el problema y su contexto se conocen lo suficiente y se pactan objetivos entre todo el equipo del proyecto.

ESTA FASE CONTIENE 6 PASOS:

DESCUBRIR

Paso 1
Conformar el equipo de proyecto

Paso 2
Describir el problema

Paso 3
Analizar la prefactibilidad del proyecto

Paso 4
Identificar actores y sus actividades

Paso 5
Mapear datos

DEFINIR

Paso 6
Definir los objetivos del proyecto

Paso 1
Conformar el equipo de proyecto

Es común pensar que los proyectos de ciencia de datos son de exclusiva responsabilidad del equipo técnico de la institución, es decir de un conjunto de analistas y responsables de los datos. Esto es un error. **Estos proyectos son llevados a cabo por equipos multidisciplinarios con competencias que se complementan.** Más aún, la ciencia de datos no es posible sin un acabado conocimiento del dominio y entorno en el que se implementará el modelo. Los roles más importantes a considerar son:

JEFATURA DEL PROYECTO

Es la persona responsable del proyecto en general. Debe velar porque se logren sus objetivos, gestionar los recursos necesarios para su desarrollo y poseer las facultades para dirimir las controversias que surjan. Es quien distribuye las cargas de trabajo, organiza y convoca las diferentes instancias de colaboración, vela por el cumplimiento de los objetivos y plazos asociados al proyecto, y media el vínculo entre el equipo del proyecto y los interesados externos de la institución (*stakeholders*). También es quien tiene en consideración los aspectos éticos desde el punto de vista del impacto social.

REPRESENTANTE DEL ÁREA A INTERVENIR

Es la persona encargada de levantar y sistematizar el conocimiento de la organización para que el desarrollo del proyecto se integre al funcionamiento de la institución. Además de poseer capacidades para levantar los procesos de la institución, debe integrar habilidades de gestión del cambio para levantar los focos de resistencia desde un inicio y asegurar la adopción de la solución en la institución estratégicamente.

ANALISTA DE DATOS

Es la persona responsable de la manipulación de los datos y los análisis correspondientes. Recibe múltiples nombres dependiendo de su nivel de especialización y enfoque (analista de datos, científico/a de datos, estadístico/a, entre otros). En cualquier caso, debe ser capaz de programar y realizar los tipos de análisis necesarios para construir el modelo.

RESPONSABLE DE LOS DATOS Es la persona con el conocimiento más acabado sobre la calidad de los datos, sus procesos de recolección y almacenamiento porque estará a cargo de su constante vigilancia. Ya sea que los datos se alojen dentro o fuera de la institución, es recomendable incorporar a esta persona al equipo del proyecto. Además, tendrá la responsabilidad del anonimato de los datos.

RESPONSABLE LEGAL Es la persona responsable levantar y advertir los potenciales riesgos éticos y legales del proyecto, y proponer una o más acciones o medidas de mitigación. También debe velar por el cumplimiento de otras consideraciones normativas como la no discriminación o posibles leyes sectoriales. Además, es común que los proyectos de ciencia de datos utilicen datos personales o información de instituciones que requieren de un tratamiento especial, por lo que será la persona responsable de sugerir los términos y condiciones de tratamiento.

El equipo responsable del proyecto debe conformarse en el comienzo de la fase de investigación del problema, buscando la representación de todas las áreas involucradas y la diversidad de competencias que permitan construir la solución y resolver el problema.

Una buena práctica es favorecer una gobernanza horizontal, ya que facilita que todos los miembros del equipo puedan entregar su opinión respecto de las decisiones que se tomen, al mismo tiempo que facilita el trabajo colaborativo y la co-creación de la solución en la próxima fase.

>> **GOBERNANZA HORIZONTAL**
Una estructura funcional horizontal es aquella que otorga la misma capacidad de tomar decisiones a distintos integrantes del equipo, sin tener necesariamente la autorización de una persona con un cargo superior (Ulloa, Masacon y Rodriguez 2019).

Por último, en una institución pública puede ser difícil sumar gente para un proyecto de estas características por lo que sugerimos contemplar este capítulo desde la perspectiva de las funciones. Es decir, independiente del número de personas que integren el equipo, todas las funciones deben estar cubiertas, incluso subcontratando servicios. En este sentido, este paso es altamente iterativo, ya que las personas que integren el equipo podrían cambiar dependiendo de la definición del problema y los objetivos del proyecto.

Paso 2
Describir el problema

En innovación pública, los proyectos de ciencia de datos tienen su origen y son motivados por los problemas que enfrentan las personas usuarias en uno o más servicios entregados por el Estado. Su propósito debe estar orientado a generar valor público de cara a la ciudadanía, por lo que entender y describir detalladamente el problema que enfrentan las personas usuarias es central para definir cómo un proyecto de ciencia de datos podría ofrecer una solución. Una tentación es definir el problema

como la ausencia de un modelo de ciencia de datos, o creer que la mera existencia de datos en una institución exige la existencia de este modelo, pero no se debe olvidar que su valor yace en los problemas que resuelve y el impacto final para las personas usuarias.

Para conocer en profundidad el problema será necesario recurrir a múltiples fuentes de información que permitan fortalecer la comprensión del problema, triangulando y contrastando perspectivas. Esta información se puede levantar específicamente para enriquecer la comprensión del problema mediante entrevistas con personas usuarias, la aplicación de un cuestionario online, acompañando a una persona usuaria en una interacción real con el servicio, entre otras mencionadas en la *Guía Permitido Innovar: ¿Cómo podemos resolver problemas públicos a través de Proyectos de Innovación?*[23]. O bien, se pueden ocupar datos ya existentes como informes previos, registros de sus interacciones con las oficinas de atención ciudadana o cualquier otro dato que sirva para enriquecer la descripción del problema.

En cualquier caso, obtener el punto de vista de las personas usuarias es crucial ya que el proyecto debe responder a sus problemáticas, así como la perspectiva de las funcionarias y funcionarios que conocen el funcionamiento interno y pueden ver lo que las personas usuarias no ven. Una buena práctica es complementar este diagnóstico con experiencias locales desde donde se puedan transferir aprendizajes, experiencias internacionales que sirvan como referencia *(benchmark)* y con evidencia científica nacional e internacional.

La descripción será un paso necesario pero no suficiente para que quienes toman las decisiones consideren necesario abordar la problemática. La relevancia de la problemática será crucial para generar interés en las personas y que el proyecto se desarrolle.

Se hace indispensable conocer en detalle el problema planteado para tener certeza de la pertinencia de desarrollar un proyecto de ciencia de datos, puesto que la solución puede pertenecer a otro tipo de innovación. Por ejemplo, un proyecto puede ser sobre la creación de un *chatbot* en la web institucional para acceder de manera más rápida a la información. Sin embargo, si el problema se origina porque la página web no es lo suficientemente clara o accesible, entonces una solución puede ser simplemente reorganizar la información en la plataforma y no un proyecto de gestión y uso de datos.

Para favorecer la comprensión del problema, delimitar su alcance, y ver cómo afecta al usuario o usuaria final se sugiere usar la herramienta **Formulación del Problema** (Herramienta I). Para ejemplificar, se presenta una formulación correcta e incorrecta del proyecto *WhatsApp Mujer,* que consistió en la generación de un canal silencioso para orientar a mujeres que viven violencia de género en situación de confinamiento por la pandemia y que requieren apoyo especialista.

23. Disponible en **lab.gob.cl/guias-permitido-innovar**

HERRAMIENTA I

Formulación del Problema

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. **Convocar a todo el equipo del proyecto** en torno a un documento compartido. Invitar a personas que estén involucradas en el problema, para que aporten con su experiencia.
2. **Definir quiénes son las personas usuarias** (ciudadanía, organizaciones de la sociedad civil u otras instituciones públicas) que sufren el problema.
3. **Contextualizar el problema** en el espacio y el tiempo identificando puntos de contacto con el servicio.
4. **Levantar indicadores** que demuestren la existencia y magnitud del problema.
5. **Elaborar una hipótesis** que permita entender las causas del problema.
6. **Identificar cuáles son las medidas actuales** que abordan esta problemática, con o sin éxito.

PRODUCTO DE LA HERRAMIENTA

Formulación del problema considerando todos los elementos identificados en las preguntas anteriores.

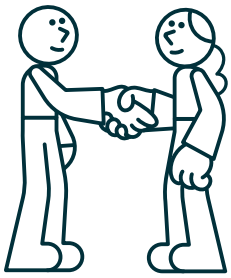
CATEGORÍA	PREGUNTA	CORRECTA ✓	INCORRECTA ✗
PERSONAS USUARIAS	**¿Quiénes se ven afectados/as por este problema?**	*Mujeres que viven violencia de género en situación de confinamiento.*	*Mujeres*
CONTEXTO DEL PROBLEMA	**¿Cuándo y dónde les afecta?**	*En Chile, especialmente durante el período de confinamiento por la pandemia COVID-19, al momento de requerir apoyo y contención de manera silenciosa por violencia de género.*	*Chile*
INDICADOR	**¿Cuánto les afecta?**	*Se observó un aumento del 70% en la cantidad de llamadas al canal telefónico 1455 del Servicio Nacional de la Mujer y la Equidad de Género (SernamEG) con respecto al mismo período del año 2019. Asimismo, se conocieron casos de llamadas simulando otro tipo de situaciones, pero con el objetivo de solicitar apoyo, ante el miedo de ser escuchadas por agresores. Esto coincide con las tendencias internacionales.*	*Existe un aumento en los casos de violencia de género en la pandemia.*
HIPÓTESIS	**¿Por qué existe este problema? ¿Cuáles son sus causas?**	*Las mujeres se contactan de manera más creciente con el canal telefónico, sin embargo se desconoce quienes no lo hacen puesto que conviven con su agresor y carecen de canales silenciosos, flexibles, anónimos, y disponibles las 24h.*	*Las mujeres que viven violencia no se contactan con los canales de apoyo.*
MEDIDAS ACTUALES	**¿Existe alguna medida que aborde este problema actualmente?**	*Los canales tradicionales de atención son telefónicos (fonos 149, 134, 1455, 600 400 0101). Su limitación es que las mujeres deben hablar en voz alta, pudiendo alertar a sus agresores y generando consecuencias que ponen en riesgo su integridad.*	*Sí, pueden llamar por teléfono.*
FORMULACIÓN DEL PROBLEMA *[Los/as usuarios/as] + "de" + [contexto] "evidencian" + [indicadores] + "porque" + [hipótesis] + "y" + [medidas actuales insuficientes] + "son insuficientes".*		*Las mujeres que viven violencia de género en Chile durante la pandemia del COVID-19 evidencian un incremento del 70% en contactos al canal telefónico de apoyo y contención y no existe referencia sobre las que no lo hacen porque requieren canales más discretos, porque la situación de confinamiento ha incrementado los niveles de violencia y los canales de contacto actuales son insuficientes para abordarlo.*	*Las mujeres chilenas no denuncian cuando viven violencia de género en el hogar.*

Paso 3
Analizar la prefactibilidad del proyecto

Una vez formulado el problema, es importante responder algunas preguntas iniciales que entregarán una primera aproximación a los potenciales riesgos del proyecto y los correspondientes mecanismos de atenuación. Esto permite adelantarse a inconvenientes que pueden surgir a lo largo del ciclo de vida de un proyecto. En esta línea, es recomendable sumar a quienes conforman el equipo indicado en el paso anterior, y también a otras personas de distintas áreas de la institución que puedan abordar el problema desde dimensiones diversas, de modo de lograr un análisis como el que se propone a continuación.

La prefactibilidad del proyecto se puede analizar considerando los **ámbitos políticos, económicos, sociales, tecnológicos y legales, conocido como PESTL por sus siglas.** A continuación, se describen las principales consideraciones que debiese tener el equipo al responder cada ámbito.

ÁMBITO POLÍTICO



En proyectos de innovación pública es fundamental que las autoridades relacionadas al proyecto estén involucradas, o al menos representadas desde el momento inicial, ya que esto permitirá mantener el alineamiento estratégico y disponer de los recursos que se necesitarán en su desarrollo. Además, si las autoridades que están relacionadas directa o indirectamente al proyecto patrocinan su desarrollo, la gestión con actores del ecosistema será más fácil. Por lo tanto es recomendable reunirse con el equipo directivo de la institución para consignar en un documento su apoyo en el proyecto y la prioridad en solucionar el problema.

También conviene el involucramiento de las áreas que verán cambios en su quehacer para que el proyecto aumente sus probabilidades de éxito, tenga las bases para que el modelo se implemente y se mantenga en el tiempo.

ÁMBITO ECONÓMICO

Anticipar proyecciones financieras de un proyecto servirá para considerar los riesgos económicos asociados. La aprobación del proyecto puede estar sujeta al balance entre los costos y los potenciales beneficios esperados. Por lo mismo, se recomienda considerar desde un inicio con los recursos necesarios (financieros y de personas) para desarrollar el proyecto de ciencia de datos, no solamente para su creación, sino también para la operación y mantención del algoritmo que resulte como solución al problema.

ÁMBITO SOCIAL

En innovación pública es primordial considerar y mejorar la experiencia de las personas usuarias de los servicios entregados por el Estado, por lo tanto se deben tener claros los beneficios potenciales y riesgos sociales del proyecto. Por ejemplo, en un proyecto que difunde herramientas educativas y consejos para controlar el estrés por mensajería instantánea a cuidadores de niños y niñas de 0 a 6 años durante la pandemia, un beneficio social sería la mejora en la salud mental de quienes integran el hogar. Por otro lado, un riesgo social asociado a este mismo proyecto es una posible reducción de la asistencia de los niños y niñas en jardines infantiles. Simultáneamente, se debe considerar la opinión pública y en la medida de lo posible, prever si el proyecto será bienvenido o no cuando se implemente.

ÁMBITO TECNOLÓGICO

Los proyectos de ciencia de datos pueden utilizar diversas fuentes de datos y es común que se utilicen datos tanto internos como externos. Los datos internos o administrativos son aquellos generados y tratados por los organismos públicos en el ejercicio de sus funciones y dentro del marco de sus competencias.

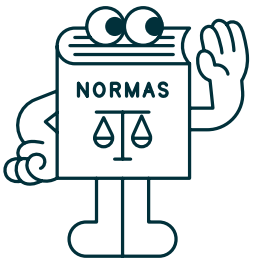
Por otro lado, los datos externos son los producidos por otras entidades. Si se trata de datos abiertos, son de fácil acceso. Si son de carácter exclusivo, se deben considerar otras interferencias que podrían impedir el desarrollo fluido y ágil de proyecto, especialmente en lo relacionado a restricciones legales.

Cualquiera sea el caso, se debe revisar la arquitectura tecnológica disponible para consultar los datos, y las factibilidades de conectarse por medio de webservices o API, según esté establecido previamente. Considerando que en materia de innovación pública es relevante mantener la agilidad, es recomendable estar muy seguro de que podamos acceder de manera oportuna a los datos actualizados de otra institución.

>> WEBSERVICES
Es un término genérico para una función de software interoperable de máquina a máquina que se aloja en una ubicación direccionable de red.
Suele proporcionar una interfaz orientada a objetos y basada en la web a un servidor de bases de datos, utilizado por ejemplo por otro servidor web, o por una aplicación móvil, que proporciona una interfaz al usuario o usuaria final.

>> API: APPLICATION PROGRAMMING INTERFACE
Son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos. Por ejemplo, el sistema de software del instituto de meteorología contiene datos meteorológicos diarios. La aplicación meteorológica de su teléfono "habla" con este sistema a través de las API y le muestra las actualizaciones meteorológicas diarias en su teléfono.

ÁMBITO LEGAL



El desarrollo de un proyecto de ciencia de datos puede implicar desafíos legales y normativos que es mejor prevenir puesto que impactarán en la planificación, sobre todo si se usarán datos de otras instituciones. Por lo demás, también es relevante hacerse preguntas de carácter jurídico con respecto a la implementación y evitar, o al menos anticipar, problemas que puedan surgir en etapas finales. En ese sentido sugerimos revisar los convenios de colaboración vigentes entre instituciones que regulan el traspaso, manejo y uso de datos entre instituciones públicas, así como las últimas novedades tanto en materia normativa como jurisprudencial.

Para analizar sistemáticamente estos ámbitos, se recomienda usar el **Análisis PESTL** (Herramienta II). Esta herramienta sirve para advertir cuáles son los puntos que hoy son críticos y se deben resolver, y para prevenir los pormenores que pueden surgir en el transcurso del proyecto. A continuación se presenta esta Herramienta completada con un ejemplo ficticio del Servicio Nacional de Aduanas, que necesita identificar los embarques más riesgosos para priorizar su fiscalización.

Es recomendable sumar a quienes conforman el equipo indicado en el paso anterior y también a otras personas de distintas áreas de la institución que puedan abordar el problema desde dimensiones diversas.

HERRAMIENTA II

Análisis PESTL

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. **Convocar** al menos a dos personas expertas en cada ámbito a un taller presencial o digital, y mejor si son de la(s) institución(es) relacionadas directamente con el problema.

2. **Explicar el problema** formulado en la Herramienta I (Formulación del problema) a quienes asisten, con todas sus características y justificaciones asociadas.

3. **Agrupar a quienes participan en mesas de trabajo (o sala virtual si la sesión es digital) por ámbito**. Por ejemplo, en la mesa del ámbito político solo deben estar las personas relacionadas a este mismo.

4. **Asignar una persona con rol de moderación por mesa o sala virtual**, quién estará a cargo de formular las preguntas asociadas al ámbito, registrar las respuestas y velar por el cumplimiento de los tiempos que se definan. Quien ejerza este rol no debe ser parte de los "expertos", sino que alguien del equipo de proyecto.

5. **Cada mesa responde colaborativamente a las preguntas en la columna *Actualidad*,** marcando una de las tres alternativas (Sí, Parcialmente, o No). Las respuestas deben estar basadas en la situación actual de cada ámbito. Cada pregunta respondida con un "No" indica un tema que deberá ser considerado por el equipo de trabajo antes de seguir avanzando en el proyecto.

6. Luego, **cada mesa completa la columna *Proyección*** con las posibles oportunidades y/o amenazas externas que se puedan presentar en su ámbito. Las respuestas deben estar basadas en la situación futura de cada ámbito. Estas pueden estar asociadas a las preguntas que se acaban de responder, o a otras que el equipo pueda identificar.

7. Finalmente se dejará tiempo para un **plenario** en el que el resto de los asistentes puedan escuchar lo que definieron los expertos de cada ámbito y complementar con su experiencia.

PRODUCTO DE LA HERRAMIENTA

Focos de riesgos y oportunidades del proyecto en 5 ámbitos relevantes.

ÁMBITO	PREGUNTA	ACTUALIDAD			PROYECCIÓN
(Para designar mesas de trabajo)	*(Para responder colaborativamente)*	SÍ	PARCIALMENTE	NO	OPORTUNIDADES Y AMENAZAS
POLÍTICO	**¿Existe claridad sobre quiénes son las autoridades internas de la institución cuyas decisiones puedan influenciar el desarrollo del proyecto?**		✓		*En dos meses más llegará una nueva jefa de servicio, por lo que no existe la certeza de que su convenio de Alta Dirección Pública esté alineado a este proyecto.*
	¿Estas autoridades están alineadas con los objetivos del proyecto?		✓		
	Si tenemos que asociarnos con otros organismos públicos, ¿existe claridad sobre quiénes son las autoridades?	✓			
	Si existen áreas que verán afectado su quehacer ¿están invitadas a participar del proyecto?	✓			

ÁMBITO	PREGUNTA	ACTUALIDAD			PROYECCIÓN
(Para designar mesas de trabajo)	*(Para responder colaborativamente)*	SÍ	PARCIALMENTE	NO	OPORTUNIDADES Y AMENAZAS
ECONÓMICO	**¿Existen los recursos financieros para el desarrollo de un modelo de ciencia de datos y su sostenibilidad en el tiempo?**		✓		*Por ahora solo se cuenta con presupuesto para el desarrollo de la solución. El resto del financiamiento estará sujeto al proceso presupuestario del próximo año.*
	¿Hay fondos disponibles de otros organismos internacionales o nacionales?			✓	
	¿Existen beneficios económicos asociados al proyecto?	✓			
SOCIAL	**¿Existe aprobación por parte de la ciudadanía sobre un proyecto así?**	✓			*La principal amenaza es que el algoritmo distribuya de manera injusta la necesidad de fiscalización de los embarques.*
	¿Existen beneficios sociales o externalidades positivas asociadas al proyecto?	✓			
	¿El proyecto se hace cargo de los riesgos sociales o externalidades negativas?	✓			
	¿El proyecto se hace cargo de los riesgos éticos (privacidad, justicia y/o transparencia)?			✓	
TECNOLÓGICO	**¿Existen los datos necesarios para el proyecto?**	✓			*La institución se encuentra migrando parte de sus servicios y capacidad de almacenamiento a la nube, por lo que se encuentra en un periodo de ajuste. Al completar la transición se contará con mucha más capacidad de cómputo y almacenamiento que en las condiciones actuales.*
	¿Se dispone de la capacidad humana para procesar y modelar los datos?		✓		
	¿Existen *softwares* gratuitos para procesar y modelar estos datos?	✓			
	¿Se cuenta con la infraestructura tecnológica para almacenar y procesar los datos?		✓		
LEGAL	**¿Está dentro de las funciones y competencias de la institución actuar sobre el problema?**	✓			*Es necesario un convenio con el Servicio de Impuestos Internos para tener la información actualizada de quienes podrían operar en los pasos fronterizos.*
	¿Existe un convenio con la(s) institución(es) responsable(s) de los datos necesarios para el proyecto?			✓	

SUGERENCIA

Incluir en esta misma instancia una segunda actividad basada en la Herramienta III: **Mapa de Actores Clave.** *(pág. 39)*

Paso 4
Identificar actores y sus actividades clave

En la primera fase de un proyecto de ciencia de datos también es importante identificar a los actores que se relacionan con el problema descrito en el Paso 2, cómo lo hacen y de qué manera se involucran. De esta forma, podremos entender la situación actual, identificar en qué medida esta se verá afectada por el modelo y, por lo tanto, considerar desde un inicio lo necesario para el cambio del quehacer institucional de acuerdo al nuevo escenario.

a. Actores clave

En primer lugar, corresponde identificar a todos los actores asociados al problema en tres niveles:

- **NIVEL MICRO,** correspondiente a las personas usuarias, interacciones y puntos de contacto donde el servicio es entregado.
- **NIVEL MESO,** que corresponde a las organizaciones o grupos de personas usuarias.
- **NIVEL MACRO,** referente a los sistemas que rigen y definen a los anteriores.

En cada nivel, los actores pueden ser personas, instituciones, agrupaciones o entidades. Para eso se puede emplear un **Mapa de Actores Clave** (Herramienta III).

HERRAMIENTA III

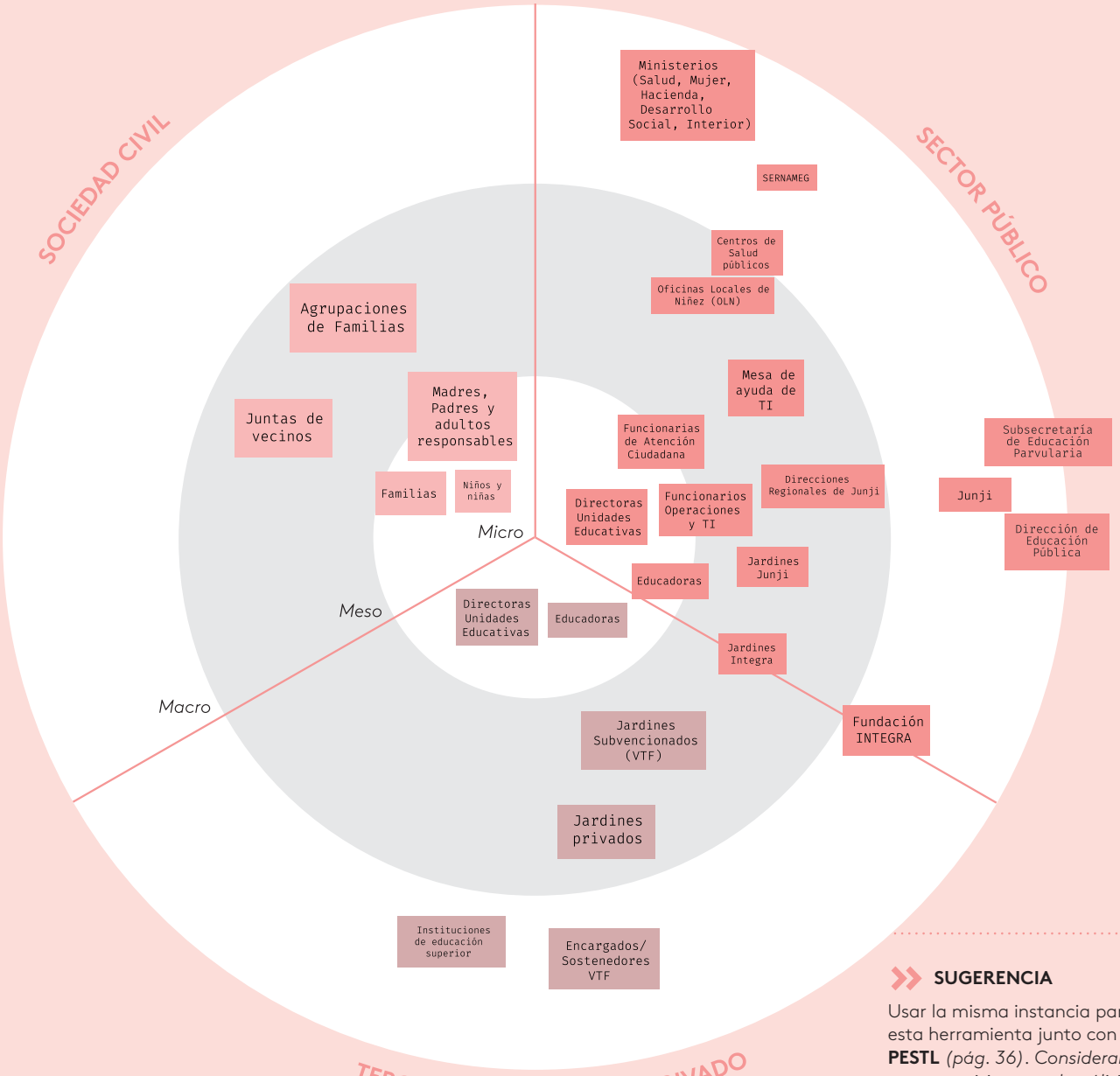
Mapa de Actores Clave

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. **Convocar** a personas de la(s) institución(es) involucrada(s) en el problema con formación y experiencia variada.
2. **Explicar a los asistentes el problema** que enfrentan las personas usuarias, con todas sus características levantadas en la Herramienta I.
3. **Agrupar a quienes asisten en mesas de trabajo (o salas virtuales si la sesión es digital)**, conformadas de tal forma que no hayan personas de una misma área por mesa. Es decir, la actividad se estructura de manera opuesta a lo indicado en la Herramienta II: Análisis PESTL.
4. Cada mesa tendrá tiempo para **levantar los actores relevantes** a nivel micro, meso y macro en el sector privado, el sector público y la sociedad civil.
5. Finalmente, desplegar un **plenario** en el que cada mesa exponga el trabajo realizado al resto del grupo.

PRODUCTO DE LA HERRAMIENTA

Levantamiento de actores por nivel y por sector.



SUGERENCIA

Usar la misma instancia para aplicar esta herramienta junto con el **Análisis PESTL** *(pág. 36)*. *Considerar que no es requisito que el análisis de la Herramienta II esté finalizado.*

b. Actividades clave

Las actividades son un conjunto de acciones o tareas específicas dentro de un proceso que ejecuta una persona responsable de un cargo. La importancia de detallar las actividades clave yace en que **es necesario contar con un entendimiento acabado de cómo se realizan actualmente y cómo el modelo de ciencia de datos las podría modificar, enriquecer, facilitar o reemplazar.** En esta etapa corresponde determinar cómo se conforma el proceso detrás del problema descrito, cuáles son sus etapas y qué actividades las componen. Un aspecto importante es **entender cómo esas actividades se encadenan para entregar el servicio a personas o instituciones usuarias.** Para una introducción al mapeo de procesos sugerimos revisar las videoclases 1 y 2 de la serie *Procesos para Innovar*[24] del Laboratorio de Gobierno.

Para trabajar con todas las actividades relacionadas al problema o desafío, se propone usar una **Ficha de Actividades Clave** (Herramienta IV). Esta permite identificar y describir la o las actividades clave, identificar sus insumos y conocer su propósito. Además, esta ficha permite evidenciar si las actividades seleccionadas se encuentran relacionadas con otras (reciben insumos desde algún actor y envían su resultado a otro actor) y si estas están relacionadas al problema al ser valoradas por las personas usuarias del servicio o se trata de una acción secundaria dentro de los procesos de la institución. Finalmente, permite entender cómo una solución basada en ciencia de datos modificaría estas actividades. El ejemplo que se presenta en la ficha muestra solo dos actividades de un mismo proceso, pero es posible modificar la tabla libremente para incluir más actividades del proceso de interés.

24. Disponible en lab.gob.cl/videoclases

HERRAMIENTA IV

Ficha de Actividades Clave

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. Una vez formulado el problema y definidos los actores relacionados, serán **quienes conformen el equipo y tengan conocimiento especialista en procesos** de la institución afectada, las personas que tengan la responsabilidad de levantar actividades clave.

2. **Convocar** a una reunión bilateral a una o más personas de cada área de cada institución relacionada directamente con la problemática definida.

3. En las reuniones bilaterales, se deben **levantar todas las actividades** vinculadas a la problemática y registrarlas en la parte superior de las columnas.

4. En las mismas sesiones, se deben **responder a las preguntas de cada actividad** asociadas al "antes", "durante" y "después".

5. Con esas interrogantes resueltas, se debe **describir cómo se quiere cambiar la actividad**. La respuesta debe ser una declaración de intenciones suficientemente general para permitir múltiples soluciones (ver ejemplos en la Herramienta), no solo una basada en ciencia de datos.

6. **Repetir los puntos 4 y 5 de esta herramienta** para cada actividad identificada hasta completarlas todas.

» PRODUCTO DE LA HERRAMIENTA

Actividades clave identificadas y caracterizadas. Además, primer indicio de cómo podrían ser modificadas.

MOMENTO	PREGUNTA	ACTIVIDADES		
(De la actividad)	*(Para responder colaborativamente)*	Actividad 1: Verificación de documentos de postulación a Junji	Actividad 2: Completar matrícula a Junji	Actividad N
ANTES	**¿Qué insumo se necesita para realizar esta actividad?**	*Los documentos que acreditan la condición de las familias.*	*El listado de postulaciones verificadas a cada unidad educativa.*	-
	¿Quién entrega el insumo para realizar esta actividad?	*Las familias postulantes que fueron seleccionadas para ingresar a una unidad educativa.*	*Cada directora de la unidad educativa.*	-
DURANTE	**¿En qué consiste la actividad?**	*Verificar la autenticidad de los documentos presentados por las familias, siguiendo un instructivo.*	*Solicitar a las familias postulantes que confirmen la matrícula que se les ofrece.*	-
	¿Quién realiza la actividad?	*Cada directora de unidad educativa.*	*Cada directora de la unidad educativa.*	-
	¿Con qué frecuencia se realiza esta actividad?	*Múltiples veces al día. Cada vez que hay una familia que podría ser matriculada.*	*Múltiples veces al día. Cada vez que se va a matricular a una familia.*	-
	¿Cuál es el resultado de la actividad?	*Familias verificadas.*	*Familias matriculadas.*	-
DESPUÉS	**¿Quién recibe el resultado de la actividad?**	*Directoras de unidades educativas.*	*Directoras de unidades educativas.*	-
	¿Qué hacen con el resultado de la actividad?	*Confirman la matrícula de las familias a Junji (columna siguiente).*	*Registran a niños o niñas en la unidad educativa.*	-
¿CÓMO QUEREMOS CAMBIAR LA ACTIVIDAD?		Eliminar la revisión manual de los documentos, liberando a las directoras de esa responsabilidad.	Que la familia confirme su matrícula de manera autónoma, sin necesidad de la intervención de las directoras.	

Paso 5
Mapear datos

Los datos han existido siempre, pero la transformación digital y adopción de nuevas tecnologías dio espacio al *Big Data*. El volumen de datos disponibles para analizar, su variedad y la velocidad a la cuál estos se obtienen, implicaron el desarrollo de nuevos y distintos mecanismos de análisis para transformar los datos en conocimiento y retroalimentar las decisiones de las organizaciones públicas. Por lo mismo, el objetivo de este paso es mapear, de forma preliminar, la existencia de datos adecuados para resolver la problemática definida.

La materia prima de un proyecto de ciencia de datos son, precisamente, los datos. Sin ellos no existe proyecto ni hay posibilidad de ofrecer una solución de estas características a las personas usuarias del servicio. Por un lado tenemos los **datos que son diseñados**, los que son generados para responder a una inquietud en particular. Ejemplos de estos datos son los obtenidos en encuestas de calidad de servicio, evaluaciones de impacto, censos, etc.

Por otro lado, están los **datos que se denominan orgánicos**, que se caracterizan por generarse a partir del quehacer institucional, ya sea de sus procesos o transacciones. Por ejemplo el registro de consultas en un motor de búsqueda, el conteo de personas en una sucursal de atención al cliente mediante un sensor, el registro de los préstamos en una biblioteca digital, la ubicación satelital de las personas usuarias al interactuar con una plataforma, etc. Para un proyecto de ciencia de datos se pueden usar ambos tipos. Más aún, es común que un proyecto utilice múltiples bases de datos.

Más importante que el tipo de datos, es lo que se puede hacer con ellos. Por un lado, tienen que servir para construir el modelo de ciencia de datos. Para esto se deben identificar todas las fuentes de datos disponibles para modelar, y evaluar si contienen lo necesario para construir un modelo. En este punto no es necesario seleccionar las variables que se emplearán en el modelamiento, ya que para aquello existe un proceso iterativo que se explicará en el Paso 11.

Por otro lado, deben permitir medir los indicadores definidos para estimar el impacto de la iniciativa, referida a los objetivos del proyecto que se establecen en el Paso 6. Por ejemplo, si el proyecto busca aumentar la satisfacción usuaria en atención presencial, entonces se debe poder medir dicha satisfacción para estimar el logro del objetivo del proyecto.

Si se trabaja con datos ya existentes se debe asegurar que el equipo ejecutor del proyecto tenga acceso a ellos. **En los casos en que el equipo no sea el responsable de los datos, se recomienda que los datos sean solicitados formalmente para asegurar el acceso, comprometer un protocolo de seguridad de datos y darle sostenibilidad al proyecto.** Si se tratase de datos obtenidos por otras instituciones, y que no sean abiertos, un paso necesario más adelante es firmar un convenio de colaboración que regule el uso de los datos o, en su defecto, algún acuerdo escrito que regule el traspaso, uso y tratamiento de los mismos.

En cambio, si fuese necesario producir datos nuevos a través de un método de registro de interacciones con una plataforma, o el levantamiento de un cuestionario para una muestra de personas usuarias de un servicio, se tiene la oportunidad de diseñar la recolección de datos para que estos sean útiles para el proyecto, ya sea porque son un insumo para el modelo que se construya o para medir la efectividad del proyecto.

En cualquiera de los casos anteriores es importante que la madurez de los datos sea suficiente, es decir, que su almacenaje, contenido, calidad, privacidad y documentación permitan realizar el proyecto. En este punto del proyecto no es necesario que los datos sean perfectos, sino que sean adecuados para construir posibles soluciones al problema anteriormente descrito. Por ejemplo, si el problema guarda relación con la entrega de beneficios sociales a nivel personal, pero solo se cuenta con información de hogares, entonces los datos no son suficientes. En casos como estos, se deberán modificar ciertas definiciones del proyecto, o bien cambiar la forma en la que se recopilan los datos.

Para evaluar esa madurez, se sugiere utilizar la **Matriz de Madurez de Datos** (Herramienta V), elaborada por la iniciativa *Data Science for Social Good*[25] que aquí se ofrece con algunos ajustes. Con esto, se tendrá una mirada amplia de los datos disponibles para que el equipo pueda definir si estos permiten abordar la problemática definida, mejorar los datos disponibles y/o privilegiar otras iniciativas. En el ejemplo de la matriz se presentan respuestas "tipo" para todos los niveles de madurez, sin embargo, estos son puramente referenciales y cada equipo podrá tener sus propias definiciones.

>> MADUREZ DE LOS DATOS
Es una medida de la capacidad de una organización para utilizar los datos, junto con lo bien que la organización aprovecha esas capacidades. Cuando una organización tiene madurez de datos, significa que puede desplegar sus recursos de datos para lograr una serie de objetivos. En muchos casos, esto no sólo significa tomar decisiones basadas en los datos, sino también hacer que los recursos de datos sean más accesibles en toda la organización.

>> CALIDAD DE LOS DATOS
Se puede considerar que los datos son de alta calidad si son aptos para su uso previsto, o si representan correctamente el constructo al que se refieren en la realidad. Por otro lado, mientras más bases de datos se disponen, la coherencia interna de estos suma relevancia. Dada la variedad de definiciones y situaciones, los puntos de vista de las personas sobre la calidad de los datos pueden fácilmente estar en desacuerdo, incluso cuando se habla del mismo conjunto de datos utilizados para el mismo propósito. La ISO 8000 describe las características y atributos a tener en cuenta con respecto a la calidad de los datos en distintos contextos que se les presentan a las organizaciones. Para tener una mejor idea vean el ejemplo de la herramienta.

Es importante que la madurez de los datos sea suficiente, es decir, que su almacenaje, contenido, calidad, privacidad y documentación permitan realizar el proyecto.

25. Disponible en bit.ly/3TIQPnu

HERRAMIENTA V

Matriz de Madurez de Datos

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura del proyecto debe **designar la responsabilidad** de completar esta matriz a quien tenga el rol de analista de datos y/o responsable de los datos.
2. **Identificar la(s) fuentes de datos** a analizar en conjunto con la(s) institución(es) relacionadas directamente a la problemática.
3. **Acceder a las bases de datos** y su documentación.

NOMBRE Y FUENTE DE LA BASE DE DATOS:

CATEGORÍA	ÁREAS	PREGUNTA	NIVEL DEFICIENTE	
¿CÓMO SE ALMACENA LA INFORMACIÓN?	ACCESIBILIDAD	¿Cuál es el nivel de accesibilidad de los datos requeridos?		
	ALMACENAMIENTO	¿Cómo es el almacenamiento de los datos?		
	INTEGRACIÓN	¿Qué tan integrados están los datos con otras fuentes de datos?		
¿QUÉ INFORMACIÓN SE RECOLECTA?	RELEVANCIA Y SUFICIENCIA	¿Qué tan relevantes y suficientes son los datos para resolver tu problema?	✓	*Los datos son irrelevantes para el problema. Por ejemplo, se quiere estimar la probabilidad de terminar la universidad pero no se tiene datos acerca de quién se gradúa.*
	CALIDAD	¿Cómo es la calidad de los datos?		
	FRECUENCIA DE RECOLECCIÓN	¿Con qué frecuencia se recolectan los datos?		
	GRANULARIDAD	¿Cuál es el nivel de granularidad de las fuentes de datos?	✓	*Agregado a nivel de ciudad.*
	HISTORIA	¿Cuánta historia está almacenada y cómo se administran sus actualizaciones?		
¿CÓMO SE ACCEDE A LOS DATOS?	PRIVACIDAD Y USO	¿En qué nivel se encuentran las políticas de privacidad y de uso de los datos en la institución?		
	DOCUMENTACIÓN	¿Cómo es la documentación de los datos?	✓	*No existe documentación digital o metadata. Los códigos de las variables no están documentados.*

4. **Marcar una casilla** que determine el nivel deficiente, básico, intermedio o avanzado con respecto a la pregunta formulada en la tercera columna.

5. Luego deben **justificar su marca respondiendo en el espacio de la casilla.**

PRODUCTO DE LA HERRAMIENTA

Un acuerdo entre el equipo de proyecto sobre los niveles suficientes en cada área para abordar la problemática formulada y acordar los objetivos en el Paso 6.

>> GRANULARIDAD

La granularidad es el detalle, o el nivel más reducido en el que pueden presentarse los datos para la realización de un análisis. Un buen ejemplo de la granularidad de los datos es cómo se subdivide un campo de nombre, si está contenido en un solo campo o subdividido en sus componentes, como el primer nombre, el segundo nombre, el apellido paterno y el apellido materno. La ventaja de los datos granulares es que se pueden moldear de la manera en que el/la científica de datos requiera. Si los datos no están granulados entonces se hace más difícil manipularlos y analizarlos.

	NIVEL BÁSICO		NIVEL INTERMEDIO		NIVEL AVANZADO	
			✓	*Los datos están en formatos accesibles como CSV, JSON, XML o una base de datos accesible de forma remota.*		
	✓	*PDF o imágenes*				
	✓	*Los datos se exportan ocasionalmente y se integran de manera ad-hoc.*				
					✓	*No hay problemas de falta de datos o errores de digitación; las bases están limpias.*
	✓	Anual de manera manual				
			✓	*Se guarda la información histórica ocupando una marca temporal.*		
					✓	*Está definido el acceso a los datos y se controla la privacidad de los mismos para preservar la privacidad de los individuos.*

SUGERENCIA Incluir en la convocatoria a alguien que posea conocimientos de la operación de la o las instituciones proveedoras de los datos.

Paso 6
Definir los objetivos del proyecto

Los objetivos del proyecto son los resultados que se buscan para impactar positivamente en la ciudadanía con el desarrollo e implementación de la solución. La definición de los objetivos permite acotar el alcance del proyecto considerando la evidencia obtenida en los pasos anteriores. Definir objetivos permite planificar el trabajo puesto que representan el impacto esperado y entregan métricas de éxito.

Existe una amplia y compleja variedad de posibles objetivos a establecer en un proyecto de ciencia de datos. Pueden estar orientados a incentivar comportamientos eco responsables, mejorar la comprensión de las normas, facilitar acceso a algún servicio, incrementar el número de personas vacunadas, entre otros. Por lo tanto, en el momento de definir los objetivos, particularmente en proyectos de ciencia de datos, recomendamos basarse en la metodología de *objetivos SMART*[26] *(Doran, 1981)* para avanzar con más certeza y con mayor control del proceso, definiendo objetivos específicos, medibles, alcanzables, realistas y que se ajusten a la temporalidad del proyecto. Se propone emplear la **Definición de Objetivos SMART** (Herramienta VI).

>> **LÍNEA BASE**
Un diagnóstico que se utiliza como punto de partida para hacer comparaciones entre mediciones diferidas en el tiempo.

26. Por sus siglas en inglés *"Specific, Mensurable, Achievable, Relevant, Timely"*.

HERRAMIENTA VI

Definición de Objetivos SMART

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. **Convocar** a un taller presencial o digital a la totalidad del equipo del proyecto.
2. La jefatura del proyecto tendrá la responsabilidad de **exponer a quienes sean parte del taller la formulación del problema** abordado en la Herramienta I.
3. Quién conozca mejor los procesos de la institución, tendrá la responsabilidad de **exponer el levantamiento de acciones y actores abordado en la Herramienta IV.**
4. El/la analista de datos tendrá la responsabilidad de **explicar la matriz de madurez de datos** abordado en la Herramienta V desde la perspectiva de las posibilidades para un proyecto de ciencia de datos y considerando un enfoque de lenguaje claro, de manera que pueda ser comprensible por un público menos capacitado en ciencia de datos.
5. Luego iniciará la actividad para definir y acotar los objetivos del proyecto basado en el lienzo de la herramienta impreso en grande. Para eso se deben **responder colaborativamente todas las preguntas** de la herramienta.
6. Al terminar de abordar cada aspecto, deben usar el formato sugerido y **acordar una redacción para el o los objetivos.** El equipo puede acordar tantos objetivos como desee, siempre y cuando cada uno mantenga la formulación SMART.

PRODUCTO DE LA HERRAMIENTA

Objetivos específicos, medibles, alcanzables, relevantes y con un límite temporal, para el proyecto.

ASPECTO DEL OBJETIVO	PREGUNTA VERIFICADORA	RESPUESTA *Ejemplo de la Tesorería General de la República, TGR.*
ESPECÍFICO	**¿Qué quiero lograr?**	*Potenciar la migración de personas usuarias de los canales presenciales hacia los no presenciales de TGR.*
	¿Por qué quiero conseguirlo?	*Hoy existe aglomeración y mala experiencia de servicio en oficinas TGR, altos costos de tiempos asociados a trámites y la sub-utilización de incentivos económicos disponibles para las personas usuarias.*
	¿Quiénes están involucrados? ¿Cómo?	*TGR con la contraparte definida para este proyecto y las funcionarias y funcionarios asociados a sus oficinas de atención.* *Además, las personas usuarias de los canales presenciales de TGR.*
MEDIBLE	**¿Es posible definir cuánto quiero mejorar y saber cuando lo he logrado?**	» *Aumentar en 20% el uso de canales digitales en un año (trámites; atención remota).* » *Aumentar en 20% el conocimiento de canales digitales (encuesta).*
	¿Cuál es la línea base?	» *10% del total de las personas usuarias utilizan los canales digitales.* » *No tenemos línea base sobre el conocimiento de nuestros canales digitales.*
ALCANZABLE	**¿La ciencia de datos serviría para alcanzar la meta?**	*Sí, con la ciencia de datos será posible personalizar la atención derivando a canales remotos a ciertos tipos de usuarias y usuarios, determinados por sus necesidades y características.*
	¿Cuáles son las fortalezas de la institución y oportunidades del entorno para alcanzar el objetivo?	*TGR tiene un sólido equipo de programación que pueden ayudar a dar forma a la solución.*
REALISTA	**¿Coincide esto con otros esfuerzos o necesidades?**	*El Estado chileno está en un intenso proceso de modernización de sus instituciones y digitalización de sus servicios, por lo que iniciativas como esta son necesarias y atingentes.*
	¿Cuáles son los obstáculos y las limitaciones para alcanzar el objetivo?	*TGR tiene múltiples oficinas de atención, las que presentan importantes diferencias entre sí que pueden resultar desafiantes. Es pertinente trabajar directamente con ellas.*
	¿Es aplicable en el entorno socioeconómico y político actual?	*Migrar hacia canales de atención remota implica una reducción de costos transaccionales permitiendo a las personas usuarias un ahorro relevante en dinero y tiempo.*
	¿Existe la infraestructura tecnológica para abordarlo?	*Las herramientas tecnológicas de hoy permiten sostener procesos digitales y los servicios del mundo están migrando hacia canales de atención remotos.*
TEMPORALIDAD	**¿Cuándo deberá estar terminado?**	*El objetivo es implementar este proyecto en un año.*

FORMULACIÓN DEL OBJETIVO

[Verbo específico] + "en" + [magnitud del cambio] + "el" + [indicador] + "en el plazo de" + [tiempo] + "mediante un modelo de ciencia de datos"

Potenciar en 20% el volumen de contribuyentes que usan canales remotos en el plazo de un año mediante un modelo de ciencia de datos.


ITERACIÓN
OBJETIVOS PACTADOS
IDEAR
PROTOTIPAR Y TESTEAR
EVALUAR
MODELO DE CIENCIA DE DATOS VALIDADO

Fase 2
Diseño de propuestas de solución

Productos

Modelo de ciencia de datos validado por el equipo de proyecto, considerando identificación de tipos de análisis, obtención y exploración de datos y registro de la elaboración y validación del modelo.

Esta fase involucra la definición de el o los tipos de análisis que se deben realizar para lograr los objetivos pactados en la fase anterior, la realización de actividades de prototipado y testeo, para cerrar con un modelo de ciencia de datos validado y listo para ser piloteado en un contexto real.

Dentro de esta fase es esperable que existan múltiples momentos iterativos, con sucesivos avances y retrocesos entre pasos, ya que **se trata de un momento de experimentación, donde se pueden probar varias ideas diferentes.** Además, se introducirán técnicas más avanzadas para el correcto análisis de datos en grandes volúmenes, con alta variabilidad y veloces en su actualización.

Para desarrollar esta fase es necesario que quienes conforman el equipo en su rol de analistas de datos cuenten con conocimiento técnico de nivel general.

ESTA FASE CONTIENE SIETE PASOS:

IDEAR

Paso 7
Seleccionar y alinear tipos de análisis de datos

PROTOTIPAR Y TESTEAR

Paso 8
Obtener y cargar los datos

Paso 9
Transformar los datos

Paso 10
Realizar análisis exploratorio de datos

Paso 11
Desarrollar y ajustar el modelo

VALIDAR

Paso 12
Validar modelo

Paso 13
Generar conclusiones

Paso 7
Seleccionar y alinear tipos de análisis de datos

El inicio de esta nueva fase está centrado en la ideación, es decir, en generar ideas de solución que puedan ser prototipadas y testeadas. Se pueden consultar múltiples técnicas de ideación en la *Guía Permitido Innovar: ¿Cómo podemos resolver problemas públicos a través de Proyectos de Innovación?*[27], las que son apropiadas para usar en este contexto.

En un proyecto de ciencia de datos, parte importante de la ideación consiste en seleccionar el o los tipos de análisis de datos que se llevarán a cabo. En este punto no es necesario seleccionar una técnica en particular, pero sí un tipo de análisis que se debería emplear para el desarrollo del modelo. Es habitual que los proyectos de ciencia de datos utilicen uno o más tipos de análisis dependiendo de la naturaleza del problema. Para seleccionar cada tipo de análisis se debe tener un problema y objetivos bien formulados. Así se podrá acordar si el modelo debe tener ser descriptivo, exploratorio, inferencial, predictivo, causal, y/o mecánico *(Peng & Matsui, 2015)*. Normalmente se tiende a pensar en modelos de carácter predictivo, sin embargo las otras alternativas también pueden ser útiles para dar solución al problema formulado *(Shmueli, 2010)*. A continuación presentamos las definiciones de cada tipo de análisis *(Leek & Peng, 2015)*.

a. **DESCRIPTIVO**
Resumir mediciones en un único conjunto de datos que se presentan de manera ordenada, sin hacer interpretaciones.

b. **EXPLORATORIO**
Analizar y descubrir tendencias, correlaciones o relaciones entre las mediciones para generar ideas o hipótesis.

c. **INFERENCIAL**
Cuantificar la probabilidad de que un patrón observado se mantenga más allá del conjunto de datos disponibles. Este es el análisis estadístico más común en la literatura científica formal.

d. **PREDICTIVO**
Predecir un resultado a partir de un subconjunto de características. Los problemas de clasificación entran en esta categoría.

e. **CAUSAL**
Averiguar qué ocurre con la media de una variable si se realiza un cambio en la media de otra. Dicho análisis identifica tanto la magnitud como la dirección de las relaciones entre las medias de las variables.

f. **MECÁNICO**
Demostrar que el cambio de una variable determina siempre y exclusivamente a un comportamiento específico en otra.

27. Disponible en bit.ly/3N6AGpl

En un proyecto de ciencia de datos, parte importante de la ideación consiste en seleccionar el o los tipos de análisis de datos que se llevarán a cabo.

Cabe destacar que en este paso basta con acordar el o los tipos análisis pertinentes para el proyecto ya que el modelado mismo viene en el Paso 11 que será realizado principalmente por quienes tengan el rol de analistas. Por lo tanto, para alinear la problemática con los objetivos propuestos y los tipos de análisis se sugiere revisar la **Pertinencia de los Tipos de Análisis** (Herramienta VII).

HERRAMIENTA VII

Pertinencia de los Tipos de Análisis

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura del proyecto debe **convocar** a una reunión al equipo del conformado.
2. **Redactar el problema formulado** que trabajaron en la Herramienta I.
3. **Redactar cada objetivo formulado** que trabajaron en la Herramienta VI.
4. **Responder a las preguntas** y asociar al menos un tipo de análisis a un objetivo.
5. **Repetir este proceso** para cada objetivo.

PRODUCTO DE LA HERRAMIENTA

Tipo(s) de análisis definidos por objetivo del proyecto.

SUGERENCIA

- Asociar más de un tipo de análisis por objetivo, considerando siempre el descriptivo.
- El equipo completo debe validar los tipos de análisis definidos.

PROBLEMA FORMULADO

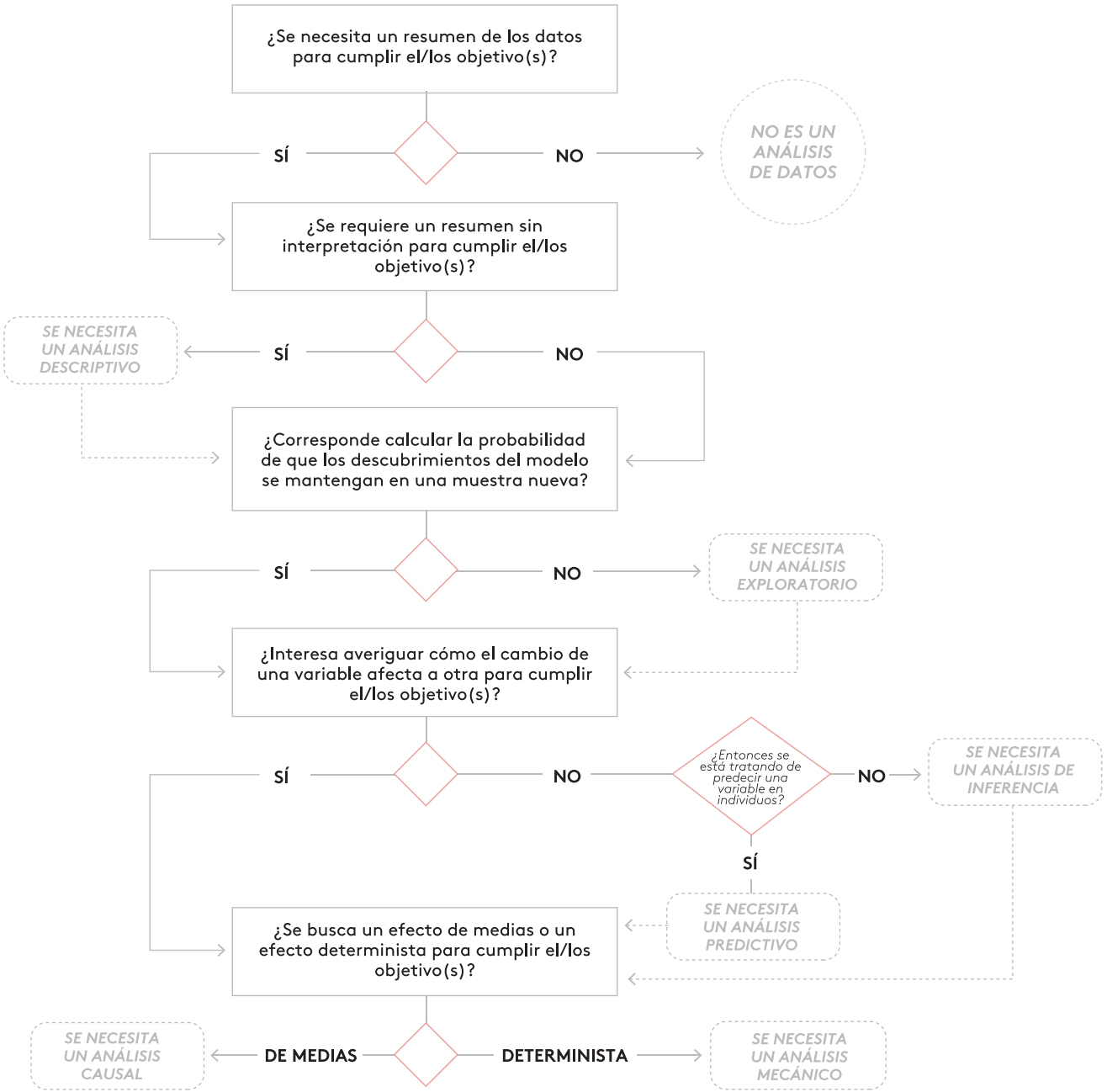
(Escribir problema)

OBJETIVO(S)

(Escribir el o los objetivos)

TIPOS DE ANÁLISIS RESULTANTES

(Escribir el o los análisis resultantes del flujo)



Fuente: Adaptado de Leek & Peng (2015).

Paso 8
Obtener y cargar los datos

El primer aspecto a considerar en este paso es la gestión para conseguir los datos necesarios para construir el modelo y cumplir con los objetivos del proyecto. Si en la etapa de planificación se consideró la utilización de datos internos, puede ser necesario establecer procedimientos claros para el traspaso veloz y eficiente de información. Por otro lado, si se consideró utilizar datos externos a la institución, será necesario desplegar otras actividades que pueden aumentar los costos del proyecto, desde la generación y acuerdo de un convenio de colaboración para el traspaso de datos hasta la definición de espacios y compra de materiales para el almacenaje de estos. En cualquier caso, el estándar es establecer protocolos de transferencia de datos para acordar el volumen de datos y la temporalidad del traspaso.

Recapitulando lo aprendido en el Paso 5 (Mapear datos), los datos pueden existir previo al proyecto o recolectarse debido a este, pueden ser diseñados acorde a sus objetivos o existir de manera orgánica. Los datos pueden extraerse con encuestas, mediante observación humana, o con inteligencia artificial. Es posible recoger bases de datos archivadas en formato *SQL* o *NoSQL*, alojadas en sistemas *CRM* y *ERP*, o incluso pueden provenir de archivos planos, de correos electrónicos, páginas web, y más.

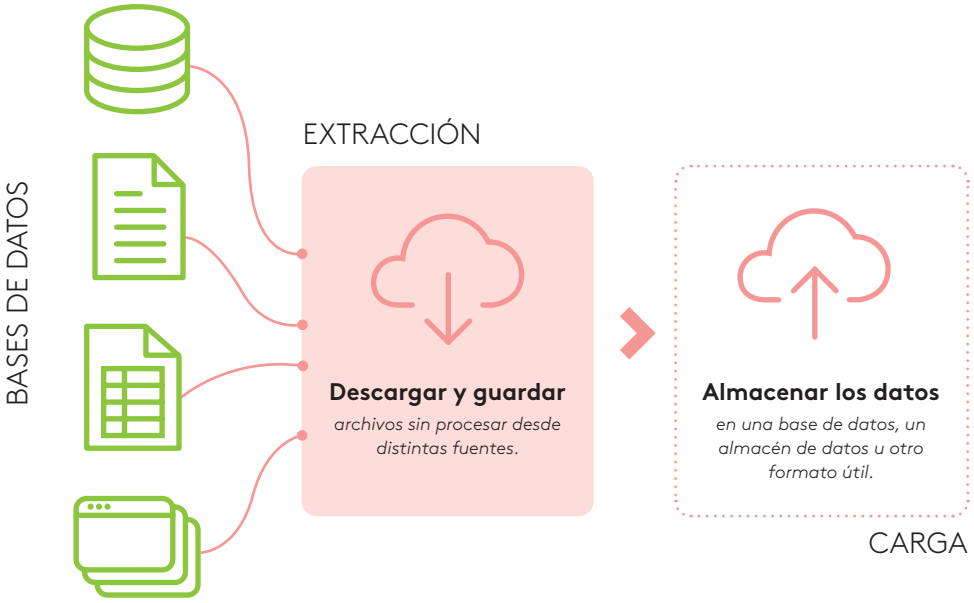
En ese sentido, las herramientas para extraer los datos necesarios también pueden ser diversas y su elección dependerá de las características del proyecto. Dependiendo de los tipos de análisis o los tipos de datos que se presenten se definirán el o los lenguajes óptimos que contengan las funciones y librerías adecuadas para su obtención. Típicamente se utiliza el lenguaje *SQL, Python, R*, entre otros, los cuales son de código abierto y tienen librerías útiles para el tratamiento de datos tales como *Spark*, o *Panda*. En el caso de *SQL*, las operaciones se realizan en distintos *softwares* (o query engine), tales como *Postgres, MySQL o PrestoDB*, que normalmente son parte de la infraestructura que ofrece *Amazon Web Services, Google Cloud, y Azure*. En esta línea, instamos a priorizar el uso de *software* de uso gratuito, ya que facilitarán transparentar la información del proyecto con la ciudadanía, ayudarán a reducir los costos del mismo y favorecerán su sostenibilidad en el tiempo.

La extracción culmina con la carga o traslado de los datos a un espacio de almacenamiento donde se guardan para su tratamiento. A menudo, los datos se asignan a objetos, que son bloques de construcción útiles para estructurar, manipular y utilizar los datos. Luego, los objetos pueden ser almacenados en algún formato idéntico al de origen o uno distinto. Una opción eficiente es el almacenamiento y tratamiento de datos en la nube con su oferta de distintos modelos de servicios de *cloud computing* que puede gestionar un tercero, tales como *IaaS, PaaS, SaaS y/o CaaS*.

>> IAAS, PAAS, SAAS Y/O CAAS
Infraestructura como Servicio es un servicio que ofrece las capas de virtualización, servidores, almacenamiento y redes para ser utilizadas de forma inmediata. En otras palabras, un tercero se encarga de la infraestructura hardware en la nube.
Plataforma como Servicio es un servicio que ofrece hardware y una plataforma de software para los aplicativos. Este servicio es ideal para Desarrolladores y Programadores.
Software como Servicio es un servicio que ofrece todas las capas de componentes del software para utilizarlo de manera inmediata y no requiere implementar infraestructura, redes, software o cualquier otro componente adicional. Se utiliza mediante navegador.
Contenedores como Servicio es un servicio que ofrece la posibilidad de gestionar e implementar aplicaciones en contenedores para facilitar el transporte, la instalación, el despliegue, entre otros.

Para mayor detalle sugerimos leer el siguiente articulo: **red.ht/3TH1XB5**

Figura 3: Proceso de extracción y carga de datos.



Fuente: Adaptado de Chen, Rubin y Cornwall (2021).

Paso 9
Transformar los datos

Este paso consiste en ejecutar distintos procedimientos que permitan convertir los datos en un formato listo para el análisis, es decir, que permiten el desarrollo del modelo propiamente tal. Para ello, es necesario considerar conjuntos de técnicas y transformaciones cuyo uso depende de los objetivos de cada proyecto.

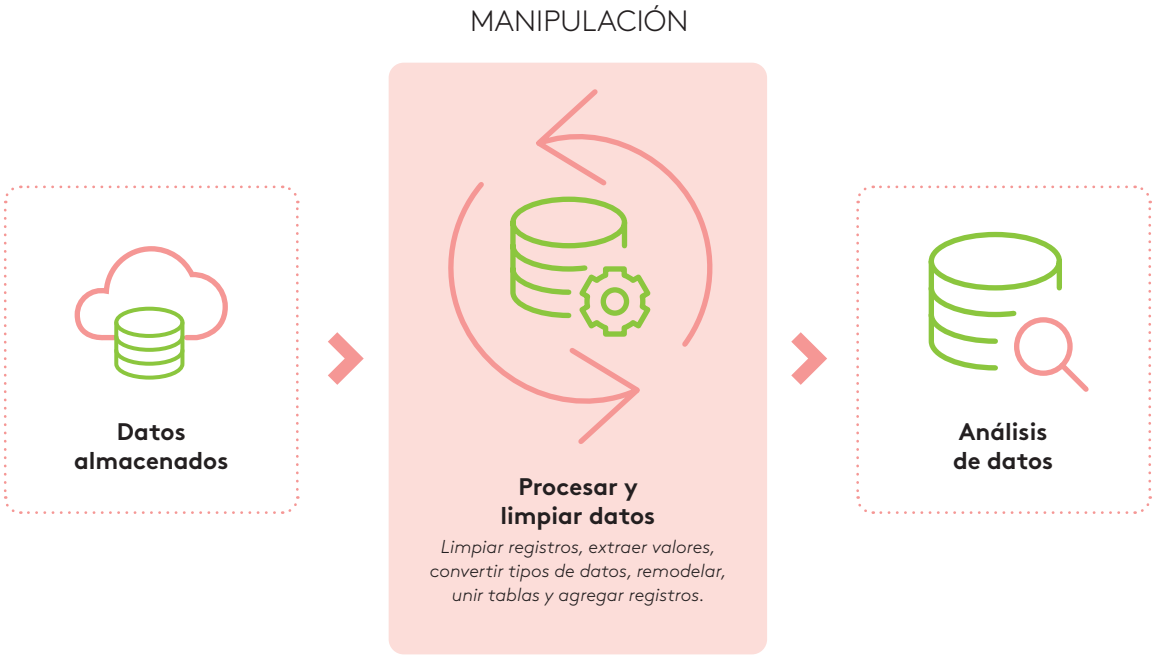
Este paso puede ocurrir entre la obtención y carga de datos, o posterior a la carga de datos, sin embargo, se separa para esclarecer la importancia de contar con datos en un formato adecuado. Además, porque esta etapa habitualmente toma tiempo.

Otro aspecto a considerar en cualquier proyecto de ciencia de datos es que el código computacional que se utilice debe incluir todas las actividades de tratamiento de los datos, inclusive lo relacionado con su obtención. Esto permitirá mejorar la replicabilidad, trazabilidad y auditar los análisis realizados. De lo contrario, se restará eficiencia a la administración de los datos. Por ejemplo, el *Índice de Innovación Pública*[28] del Laboratorio de Gobierno utiliza *Python* en *Colab*, un *software* gratuito disponible en la nube, para revisar información, unificar datos y calcular la capacidad de innovación de otras instituciones públicas.

28. Disponible en **indice.lab.gob.cl**

Dependiendo del lenguaje y *softwares* que se eligieron en el paso previo, distintos códigos serán necesarios para estructurar, modelar, extraer, unir y/o reordenar información de texto, de datos numéricos, de fechas, de imágenes, entre otros. En esta etapa, la programación debe garantizar las operaciones de limpieza y transformaciones necesarias para proceder con los cálculos y operaciones del análisis exploratorio y posteriormente, del modelo.

Figura 4: Transformación de datos.



Fuente: Adaptado de Chen, Rubin y Cornwall (2021).

Un aspecto a considerar en cualquier proyecto de ciencia de datos es que el código computacional que se utilice debe incluir todas las actividades de tratamiento de los datos, inclusive lo relacionado con su obtención.

Independiente de las elecciones que se hagan y códigos que se utilicen en esta etapa, es imperativo que la programación cumpla con tres principios básicos:

1. **EFICIENCIA**
El código debe ser eficiente y no generar funciones innecesarias. Por ejemplo, utilizando funciones y comandos que ejecuten operaciones de forma automática en reiteradas ocasiones.

2. **REPLICABILIDAD**
El código debe permitir su reutilización y el aprendizaje del ecosistema. Para esto, una buena práctica es dejar anotaciones que permitan entender el código y publicar el procedimiento final.

3. **ANONIMATO**
Es fundamental verificar la anonimización de las observaciones de la base de datos para impedir la identificación de sujetos, incluso puede llegar a ser necesario eliminar información personal y sensible (ver capítulo II referido a ética y seguridad).

Una vez finalizada la limpieza de los datos, es necesario generar el análisis exploratorio que permita conocer los datos, visualizar patrones y asociar variables.

Paso 10
Realizar análisis exploratorio de datos

Limpiados y organizados los datos, éstos se encuentran listos para comenzar con el análisis exploratorio, que no solo entregará información acerca de la calidad de los datos, sino que también permitirá dar los primeros pasos hacia la construcción de la solución.

En algunas ocasiones este análisis se transforma en una poderosa herramienta y permite hallazgos significativos que guían el resto del proyecto, o incluso puede transformarse en el proyecto en sí mismo.

El equipo técnico del proyecto es el responsable de llevar a cabo el análisis exploratorio. Sin embargo, es necesario complementar la interpretación de los hallazgos con representantes de las áreas asociadas al problema. Las respuestas a las preguntas del tipo *¿por qué existe un grupo de la población tan diferente a otro? ¿por qué existen tantas diferencias entre hombres y mujeres?*, vienen dadas por el conocimiento del problema y del quehacer institucional, y no únicamente de los datos.

Se sugiere organizar el análisis según tres propósitos: conocer a la población, visualizar patrones e identificar asociaciones entre variables. A continuación, se presenta la descripción de cada uno de estos y las técnicas de análisis que se pueden usar en cada caso.

a. Conocer a la población

Consiste en el primer análisis de datos propiamente tal, por lo que se deben resumir los datos y brindar una breve interpretación. Para esto, las técnicas de análisis descriptivo son muy útiles.

>> VARIABLES CUALITATIVAS
Son aquellas que permiten la expresión de una característica, una categoría, un atributo o una cualidad. Por ejemplo: región de residencia.

En el caso de variables cualitativas (nominales u ordinales) se puede comenzar con una tabla que indique la distribución de las variables en sus categorías. Esto se puede llevar a cabo con un conteo simple de cada categoría para conocer la magnitud del fenómeno, o con sus porcentajes para hacer comparaciones con otras variables.

>> VARIABLES CUANTITATIVAS
Son aquellas variables estadísticas que otorgan, como resultado, un valor numérico. Por ejemplo: Edad.

Si se trata de variables cuantitativas (intervalo o razón) las medidas de tendencia central (media, mediana y moda) son útiles para conocer dónde se encuentran los valores centrales. En tanto, las medidas de dispersión (rango, desviación estándar y varianza) sirven para saber cuán agrupados o dispersos están los datos en torno a dicho centro. Por su parte, las medidas de posición (cuartiles, deciles y percentiles) permiten obtener una primera aproximación a la distribución de estas variables.

Si bien se trata de técnicas sencillas, su rol en un proyecto de ciencia de datos es crucial pues son los cimientos de los modelos que se construirán más adelante y permiten tomar decisiones metodológicas antes de comenzar de lleno el desarrollo del modelo de ciencia de datos. Por ejemplo, las regresiones lineales, uno de los algoritmos más populares, usan la media como medida fundamental de sus estimaciones, por lo que se vuelve fundamental conocer sus valores y brindarles una interpretación sustantiva.

b. Visualizar patrones

Consiste en graficar los datos disponibles para identificar patrones y detectar valores atípicos.

Para graficar variables cualitativas, los gráficos circulares, de columna (vertical) y de barra (horizontal) son los más utilizados. Sin embargo, hay alternativas interesantes a estas opciones más tradicionales que se pueden revisar en los informes de resultados *del Índice de Innovación Pública*[29].

En el caso de variables cuantitativas la situación es distinta y la visualización de datos permite más alternativas. Por ejemplo, para visualizar la distribución de una variable continua, un histograma es una excelente herramienta que permite examinar el recuento de casos para cada valor de la variable. Además, permite detectar si se trata de una distribución unimodal, bimodal, asimétrica, entre otras, lo que no es fácilmente identificable con los estadísticos ya mencionados.

29. Disponibles en **indice.lab.gob.cl/#/site/resultados**

Por su parte, los gráficos de línea son muy útiles cuando se trata de identificar patrones en series de tiempo o datos longitudinales. Estos pueden tratarse de precios diarios, asistencia semanal de personas usuarias a oficinas, visitas por hora a la web institucional, entre otros. Se sugiere considerar al menos tres componentes para su análisis:

TENDENCIA
Indica si los datos crecen o decrecen en el tiempo.

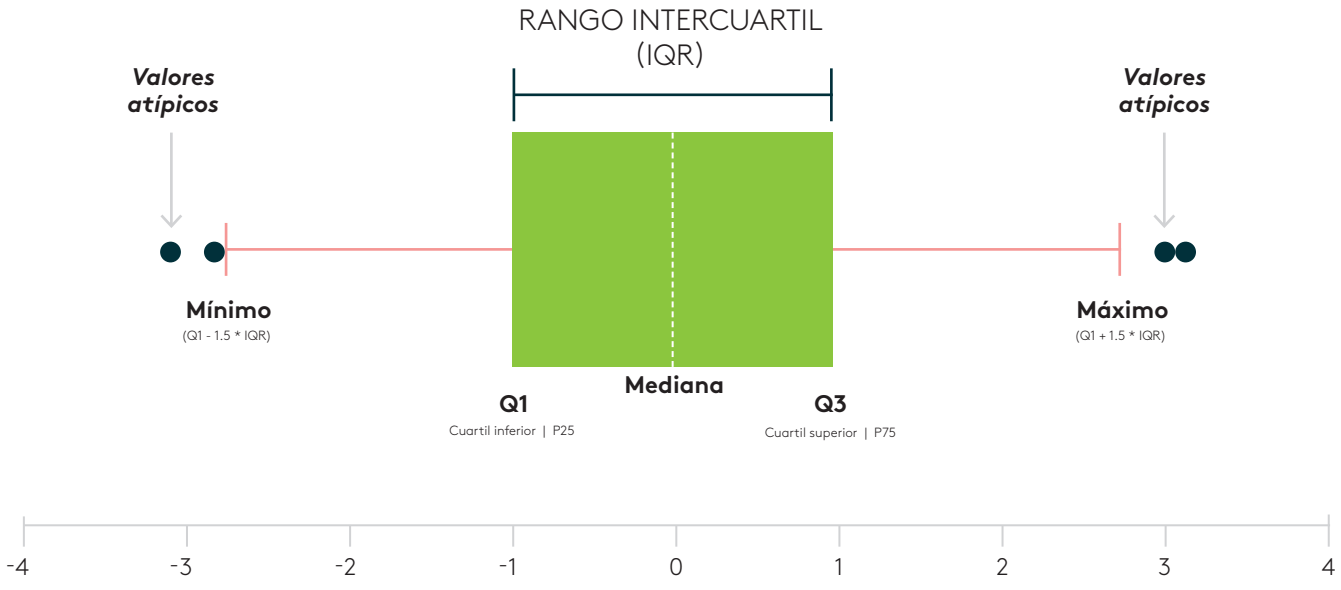
ESTACIONALIDAD
Indica si existen patrones o ciclos que se repiten cada cierto intervalo de tiempo.

RUIDO O DATOS RESIDUALES
Indica si los datos presentan fluctuaciones no predecibles o que escapan a la tendencia general.

Si se quiere complementar ese conocimiento con una primera aproximación a los datos atípicos se puede emplear un gráfico de caja y bigotes (*box plot* en inglés) como el que se muestra en la figura 5. Este gráfico contiene una caja cuyo centro está indica la mediana de los datos, y sus extremos son el cuartil inferior (25% inferior de los datos) y el cuartil superior (25% superior de los datos). Los "bigotes" o "patillas" de la caja corresponden a 1,5 veces el rango entre el cuartil inferior y el superior, también llamado rango intercuartílico (IQR). Todos los datos que se encuentren por sobre o bajo esos bigotes se consideran atípicos. Incluso, hay quienes agregan una clasificación de casos extremos para los datos que excedan en tres veces el rango intercuartílico.

>> DATOS ATÍPICOS
Que por sus caracteres se aparta de los modelos representativos o de los tipos conocidos. En el caso de los datos, es una observación que es numéricamente distante del resto.

Figura 5: Gráfico o diagrama de caja y bigotes.



Fuente: Elaboración propia.

Es importante mencionar que esta definición de datos atípicos y extremos no es la única, y se pueden emplear otros indicadores para detectarlos como la desviación estándar o la desviación absoluta mediana. En cualquier caso, su detección es relevante para poder analizarlos y tomar una decisión respecto de ellos.

c. Identificar asociaciones entre variables

Consiste en explorar las relaciones entre las variables disponibles, lo que permite identificar cuáles son las más importantes para el trabajo posterior o reducir la cantidad de variables con las que se trabaja.

En el caso de examinar la asociación entre variables cualitativas, se puede emplear una tabla de contingencia que muestre la distribución condicional de una variable sobre la otra (ver tabla 1). Un indicador que resume esa asociación es **Chi-cuadrado** (x^2), el que toma valor 0 si no detecta asociación entre las variables y un valor mayor si lo detecta.

Tabla 1: Tenencia de celular según área de residencia.

	RURAL	URBANO	TOTAL
SÍ TIENE CELULAR	37%	95%	91%
NO TIENE CELULAR	63%	5%	9%
TOTAL	100%	100%	100%

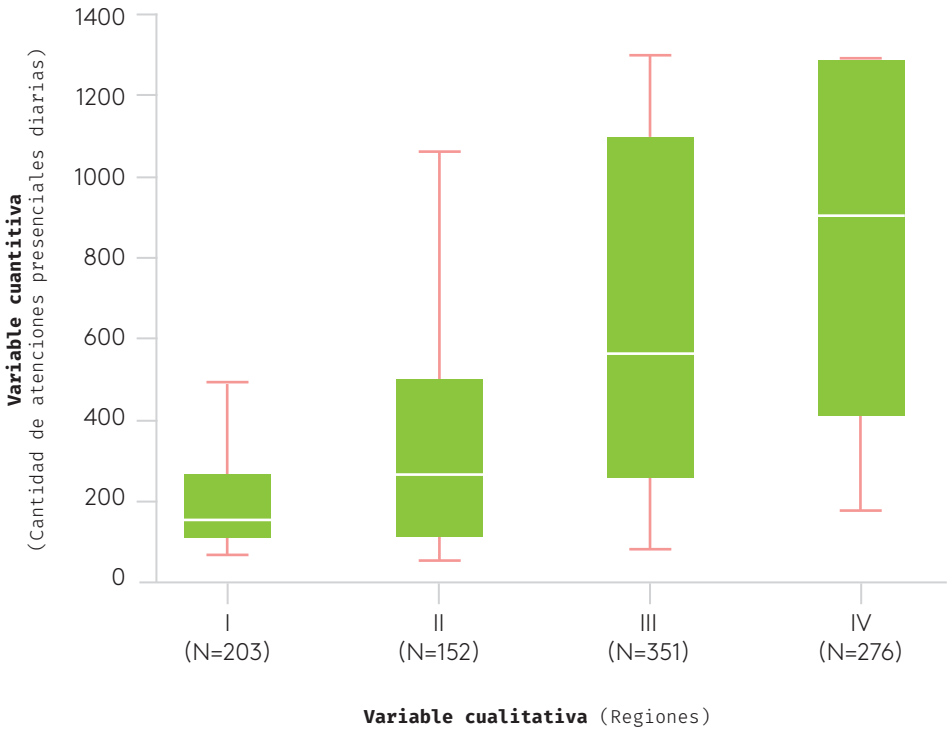
Fuente: Elaboración propia.

Cuando se quiere revisar la asociación entre variables cuantitativas y cualitativas, lo más común es revisar la distribución de las variables cuantitativas para cada categoría de las variables cualitativas. Esto se puede realizar fácilmente con un gráfico de caja que muestre la distribución de la variable cuantitativa en cada categoría de la variable cualitativa (ver figura 6).

Figura 6: Distribución condicional de variable cuantitativa según variable cualitativa.

Ejemplo: Cantidad de atenciones presenciales diarias en las oficinas de una institución pública según región.

N: puntos de atención

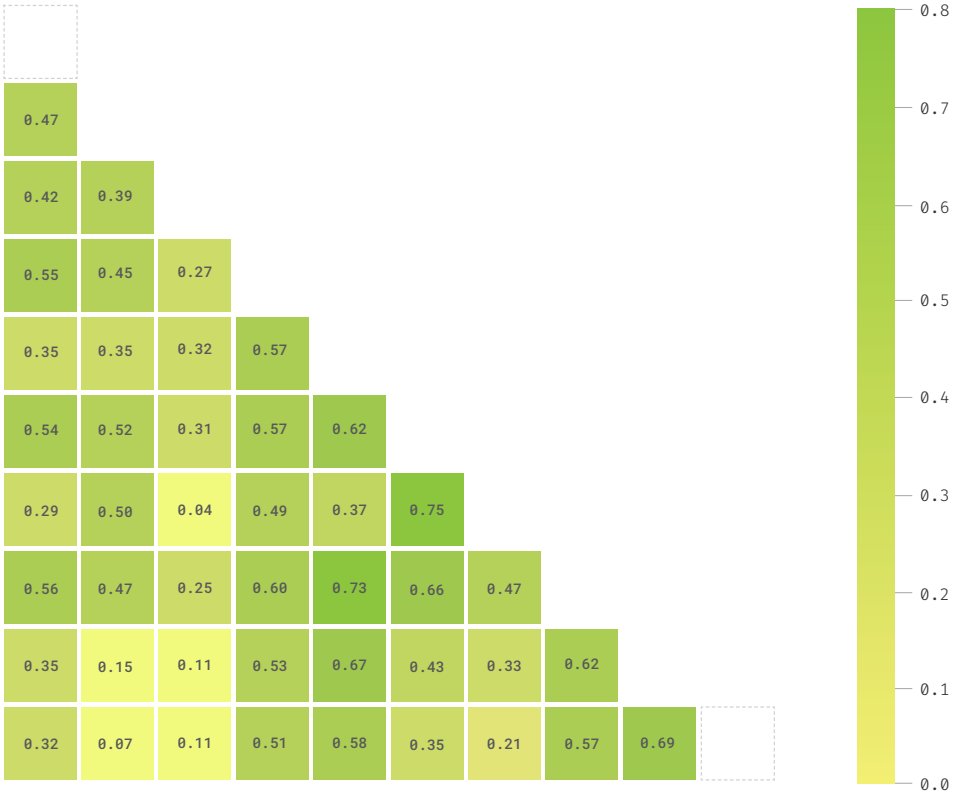


Fuente: Elaboración propia.

En cambio, si se desea examinar la asociación entre dos variables cuantitativas, se puede emplear un gráfico de dispersión, con una variable en cada eje. En los casos que se quiera revisar múltiples asociaciones bivariadas se puede lograr rápidamente mediante el cálculo del coeficiente de correlación. Uno de los más utilizados es el *coeficiente de Pearson* que permite entender la intensidad y sentido de una correlación. Dicho coeficiente puede fluctuar de -1 a 1, en donde -1 demuestra la existencia de una correlación perfectamente negativa, 0 una correlación inexistente, y 1 una correlación perfectamente positiva. Una forma de graficar esta información es mediante una matriz de correlaciones, la cual resume y permite visualizar las correlaciones para cada par de variables (ver figura 7).

Figura 7: Matriz de correlaciones.

Ejemplo: Matriz de correlaciones puntajes del Índice de Innovación Pública en las 10 capacidades clave medidas.



Fuente: Nota Técnica de Resultados 2021 del Índice de Innovación Pública, disponible en indice.lab.gob.cl

Conociendo la correlación entre variables se puede realizar un **Análisis de Componentes Principales** (*PCA* por sus siglas en inglés). Esta técnica permite reducir la *dimensionalidad*, es decir las variables o coordenadas que están a la base del fenómeno de interés, al mismo tiempo que se minimiza la pérdida de información. Si se trabaja con variables que emplean distintas escalas, es necesario estandarizar o normalizar las variables antes de llevar a cabo un análisis de componentes principales. La principal desventaja de esta transformación es que la interpretación de las variables se vuelve menos intuitiva para quienes están menos familiarizados con el análisis de datos, ya que las variables transformadas pasan a medirse en desviaciones estándar.

Una forma de graficar esta información es con un plano que da cuenta de los dos componentes o dimensiones principales que surgen del análisis, y que son las que permiten explicar de manera más adecuada la distribución de las variables.

>> ESTANDARIZAR
Es llevar un conjunto de datos con a tener media nula (=0) y desviación estándar unitaria (=1). Este método de escalado es útil cuando los datos siguen una distribución normal. La fórmula teórica para estandarizar una variable es la siguiente (Ali y Faraj, 2014):

$$x'' = \frac{(x - \mu)}{\rho}$$

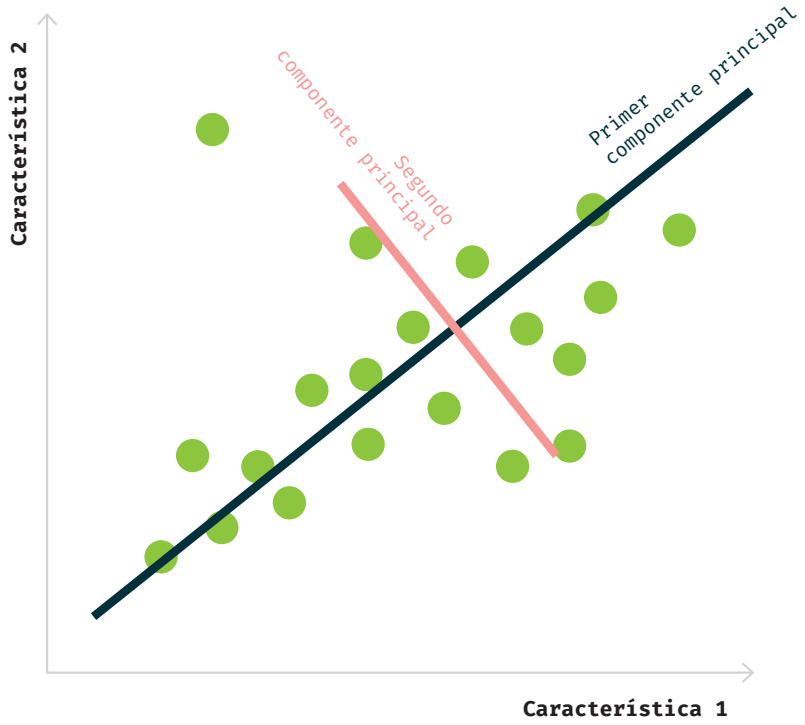
x'': variable estandarizada
μ: media
ρ:desviación estándar

>> NORMALIZAR
Es llevar los valores de un conjunto de variables a una escala comparable entre ellas entre 0 y 1, sin modificar su distribución. Este método de escalado es útil cuando los datos no presentan valores atípicos. Por ejemplo esto permitiría comparar tasas de interés, desempleo y valores bursátiles. La fórmula teórica para estandarizar una variable es la siguiente (Ali y Faraj, 2014):

$$x' = \frac{x - min}{max - min}$$

x': variable normalizada
min: valor mínimo de los datos
max:valor máximo de los datos

Figura 8: Plano de análisis de los dos componentes principales para los tipos de trámites realizados en el Estado.



Fuente: Elaboración propia, solo con fines ilustrativos.

Al presentar estos análisis es importante considerar y conocer a la audiencia objetivo. Así se podrá decidir el énfasis de los análisis y el nivel de detalle que se entregará en las visualizaciones gráficas. Respecto de la representación misma, se sugiere seguir los siete principios de la presentación efectiva de datos de *Dykes (2019)*.

Siete principios de la presentación efectiva de datos.

1

Visualizar
los datos
correctos

2

Elegir
la visualización
más adecuada

3

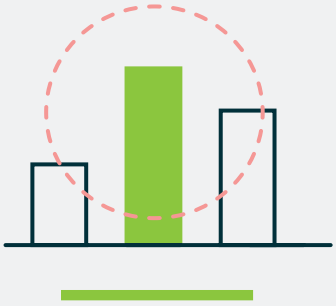
Ajustar
la visualización
a tu mensaje

4

Eliminar
el ruido
innecesario

5

Centrar la
atención
en lo importante



6

Hacer los datos
accesibles y
atractivos

7

Infundir
confianza en
las cifras

Fuente: Dykes, 2019.

Paso 11
Desarrollar y ajustar el modelo

Finalizando el análisis exploratorio, es momento de llevar a cabo análisis de mayor complejidad. Éstos dependerán del problema que afecta a las personas usuarias (Paso 2), los objetivos del proyecto (Paso 6), el tipo de análisis acordado (Paso 7) y los datos disponibles (Pasos 8 y 9). Puede ocurrir que los hallazgos del análisis exploratorio sean suficientes para solucionar el problema identificado y no sea necesario recurrir a análisis más complejos. Por ejemplo, para el manejo de los datos de COVID-19, se utilizaron técnicas de análisis descriptivo para entregar regularmente resúmenes estadísticos, como la proporción de personas hospitalizadas por la cantidad de casos de contagiados, y en base a esto se tomaron decisiones sobre las medidas sanitarias a implementar.

Para entender este paso, partiremos recordando la definición sobre qué es un modelo. **Este corresponde a una representación de la realidad que se crea a partir de los datos y que sirve para realizar distintos tipos de análisis que apoyen la toma de una decisión o que sirven para crear procesos de tomas de decisiones automatizadas.** El tipo de modelo a desarrollar dependerá de los datos disponibles, de los objetivos que se hayan establecido y de los recursos que existan para solucionar el problema o desafío.

El desarrollo, ajuste y validación de un modelo es un proceso altamente iterativo, donde, si bien el desarrollo comienza con una decisión informada, parte de la naturaleza propia de estos modelos es el ensayo y error. Por lo tanto, es muy común probar distintas técnicas y ajustarlas hasta llegar al mejor modelo para el problema que se quiere solucionar.

Para la creación del modelo, se pueden escoger una o varias técnicas. Lo importante en este punto es que el equipo técnico que esté a cargo del desarrollo del modelo justifique, basado en la literatura y experiencia comparada, cuál técnica es la más apropiada para el tipo de análisis y el problema que se quiere solucionar. Adicionalmente, se sugiere registrar los supuestos subyacentes para que luego el modelo y sus conclusiones sean válidos y sostenibles en el tiempo. A continuación, se mencionan las técnicas más utilizadas dentro del área de la ciencia de datos.

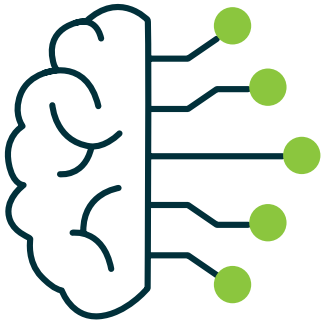
INFERENCIA ESTADÍSTICA

Es el conjunto de técnicas que permite inferir, desde los datos de una muestra, el comportamiento de la población por medio de la estimación de la distribución de probabilidad poblacional. **A partir de esto, se pueden realizar estimaciones de los parámetros de la distribución, como la media y varianza, y los consecuentes test de hipótesis sobre ellos.** Se utiliza especialmente cuando solamente se tiene datos muestrales, por ejemplo de las personas que ya han sido seleccionadas para un beneficio, o cuando los datos provienen de encuestas, pero se requieren tomar acciones en toda la población.

GEOANÁLISIS

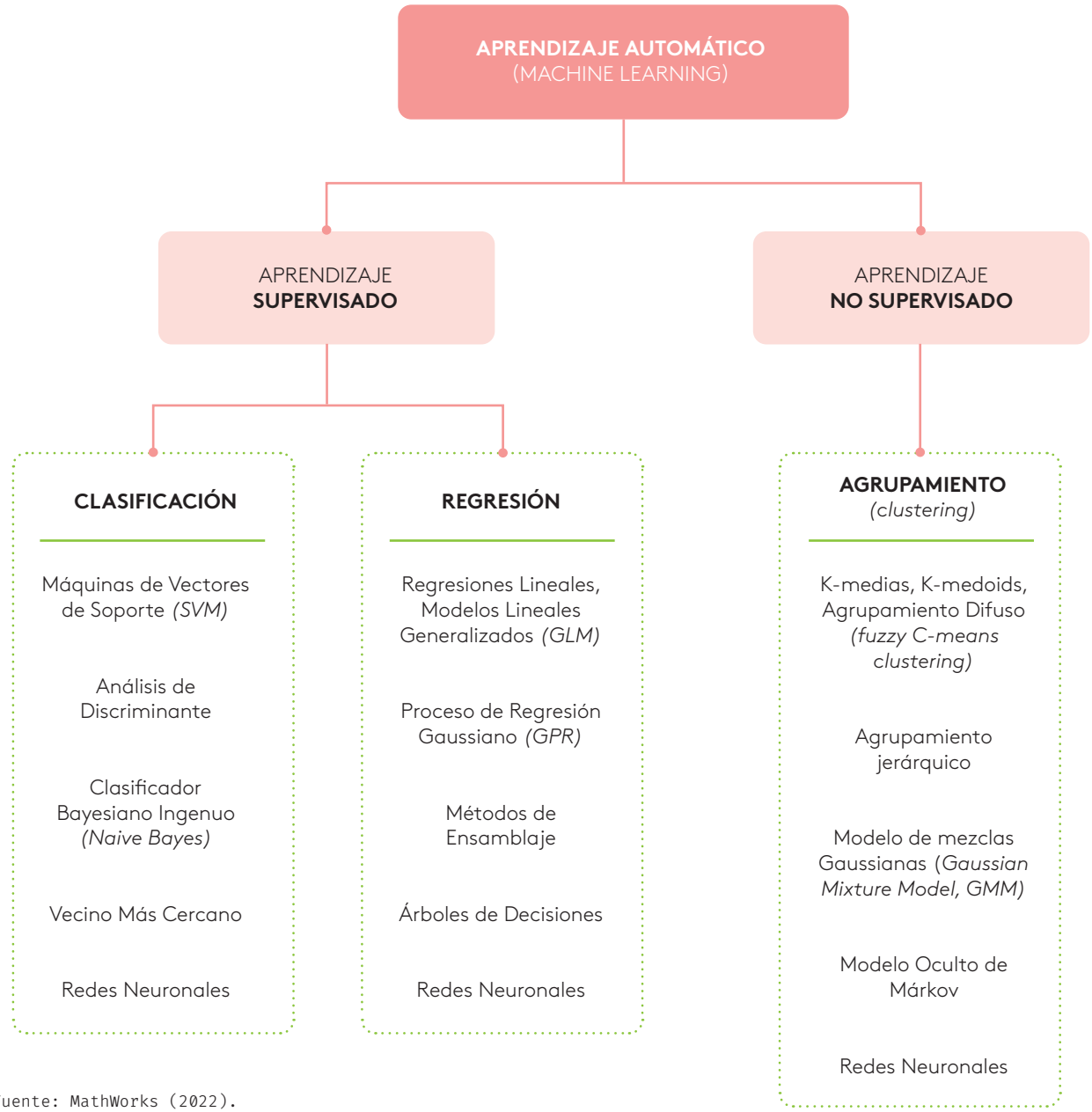
El análisis de datos geográficos comprende un conjunto de herramientas que permite estudiar los datos de un cierto fenómeno con sus localizaciones espaciales, para su almacenamiento, organización y análisis. Situar un conjunto de datos en un lugar específico del espacio puede ayudar a tener una noción más acertada de ciertos problemas, ya que permite evaluarlos tomando en cuenta su contexto geográfico y así tomar decisiones acertadas sobre estos. **El método más utilizado del geoanálisis es la *georreferenciación*, que permite integrar datos sociales, económicos o ambientales de diferentes fuentes con sus coordenadas geográficas.** Por ejemplo, para un modelo de fiscalización de contaminación ambiental, se podrían integrar datos de sensores de aire con las coordenadas de su ubicación. De esta manera, se obtendrían datos de los lugares en los que suben los niveles de gases tóxicos de acuerdo a ciertos estándares, lo que permitiría regular la situación.

APRENDIZAJE AUTOMÁTICO



El aprendizaje automático o *machine learning* se refiere a la categoría de modelos de ciencia de datos donde el objetivo es que los algoritmos aprendan a partir de los datos, identifiquen patrones y tomen decisiones a partir de estos, con poca intervención de las personas. **Existen dos tipos de modelos de machine learning: los de aprendizaje supervisado y los de aprendizaje no supervisado. La diferencia entre estos modelos es la forma en la que se presentan los datos al algoritmo.** En un modelo de aprendizaje supervisado, los datos se encuentran etiquetados como datos de entrada, que son las variables que el algoritmo usa para moldear el problema o fenómeno, y variable objetivo o datos de salida, que es la respuesta a la que se quiere llegar con la implementación del modelo. En un modelo de aprendizaje no supervisado, los datos no se encuentran etiquetados, por lo que el algoritmo debe ser más autónomo para reconocer los patrones en los datos y llegar a las conclusiones o respuestas requeridas. Existe una gran variedad de técnicas dentro de la categoría de aprendizaje automático.

Figura 9: Técnicas de aprendizaje automático.



Fuente: MathWorks (2022).

Las **técnicas de aprendizaje supervisado se basan en algoritmos que usan datos etiquetados para aprender y llegar a las conclusiones requeridas.** Estas se dividen en dos subcategorías: técnicas de clasificación y técnicas de regresión. **La clasificación se usa para problemas donde se quieren identificar categorías.** Estas pueden ser binarias (por ejemplo, recibe o no un beneficio) o multiclase (por ejemplo, zona con baja contaminación, media contaminación o alta contaminación). Por otro lado, la regresión se usa para descubrir la tendencia de los datos y, así, predecir el comportamiento de estos. Tanto la clasificación como la regresión pueden ser usadas tanto para datos discretos como para datos continuos; solo se deben adaptar y escoger la técnica correcta dependiendo del problema.

Dentro de las **técnicas no supervisadas,** una de utilidad para problemas públicos es el *clustering*, o formación de grupos según grados de similitud entre los datos. Esta técnica se utiliza cuando es necesario entender tipologías. Los algoritmos de clustering forman grupos de manera automatizada, y luego es tarea de las personas expertas en el tema interpretarlos para darles sentido. Por ejemplo, un servicio público que otorga beneficios sociales podría tener el problema de que las personas usuarias objetivo no están accediendo a tiempo a dichos beneficios. Para diseñar una estrategia de contactabilidad y difusión de los beneficios, conocer los tipos de personas que no acceden a ellos podría ser de utilidad.

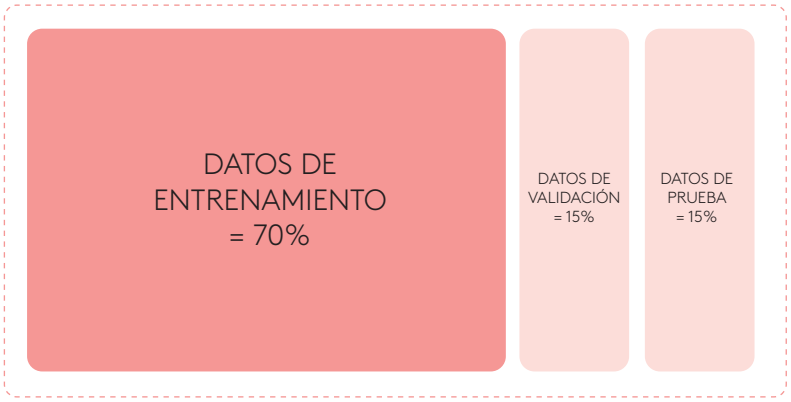
Las técnicas mencionadas tienen distintos grados de complejidad para ser comunicadas. Por ejemplo, en una regresión lineal es más fácil observar la relación y efecto de las variables, en comparación con una red neuronal. De esta forma, tal y como se mencionó en el capítulo II sobre Ética y Seguridad, la técnica debe ser lo más comprensible posible para eventuales rendiciones de cuenta, en la medida que las características del proyecto lo permitan.

>> RED NEURONAL
Una red neuronal es un método de la inteligencia artificial que enseña a las computadoras a procesar datos de una manera que está inspirada en la forma en que lo hace el cerebro humano. Se trata de un tipo de proceso de aprendizaje automático llamado aprendizaje profundo, que utiliza los nodos o las neuronas interconectados en una estructura de capas que se parece al cerebro humano. Crea un sistema adaptable que las computadoras utilizan para aprender de sus errores y mejorar continuamente. De esta forma, las redes neuronales artificiales intentan resolver problemas complicados, como la realización de resúmenes de documentos o el reconocimiento de rostros, con mayor precisión.

Después de escoger las técnicas que se van a utilizar, comienza la etapa de **entrenamiento y prueba del modelo.** Un modelo de ciencia de datos aprende de los datos y extrae información de ellos para entregar la predicción, clasificación o solución que se busca. La fase en la que el modelo está aprendiendo se llama *Entrenamiento* y la fase en la que se ven los resultados de lo que el modelo aprendió se llama *Prueba*.

Para realizar estas dos fases se deben dividir los datos disponibles en dos grupos: Datos de Entrenamiento y Datos de Prueba. Una práctica recomendable, especialmente en modelos que utilizan redes neuronales, es realizar una tercera división: datos de validación. La partición de los datos se explica a continuación:

Figura 10: División de datos.



Fuente: Elaboración propia.

DATOS DE ENTRENAMIENTO
Son la parte más grande de la partición. Con ellos, el modelo identifica patrones, calcula probabilidades y extrae la información útil.

DATOS DE VALIDACIÓN
Son una pequeña parte de los datos que se utilizan después del entrenamiento para ver cómo el modelo va aprendiendo y realizar ajustes si es necesario. La idea es utilizar datos que el modelo no ha visto antes para ver cómo se comporta frente a ellos. Con estos datos, se hace una primera evaluación del desempeño del modelo utilizando algunas de las métricas que se explican en el siguiente paso.

DATOS DE PRUEBA
Esta es la partición final de los datos, que no se utilizó ni en el entrenamiento ni en la etapa de validación, permite ver cómo se comporta el modelo entrenado con nuevos datos que no han sido procesados. Con estos datos es posible observar cómo se comportaría el algoritmo en un escenario real, en el que siempre habrá nuevos datos que procesar.

Paso 12
Validar modelo

Una vez que se tiene un modelo es importante validarlo de acuerdo a indicadores específicos antes de pilotearlo en condiciones controladas. Los modelos son simplificaciones de la realidad, por lo que no es posible tener un rendimiento perfecto en todas las métricas y por eso es importante que el equipo del proyecto entienda cuáles son las más pertinentes de acuerdo a las particularidades del modelo y los objetivos formulados en el Paso 6. Existen distintos instrumentos para validar modelos que se adecuan a las exigencias y características de cada uno.

- Por ejemplo, cuando la variable dependiente, o que resulta del modelo es categórica, se suele utilizar matriz de confusión y los indicadores que se derivan de ella.
- Por otro lado, si la variable de interés es categórica o continua y se utiliza un modelo de regresión, se utilizan instrumentos de ajuste clásicos de esos modelos, como el error cuadrático medio, el R^2, **el AIC y el BIC.**
- Si se trata de un modelo de inteligencia artificial se suele recurrir a la validación cruzada.

A continuación se describen algunas de las herramientas que nos servirán para validar el modelo, sin embargo insistimos en que las **métricas adecuadas siempre dependen del contexto de cada proyecto.**

a. Matriz de confusión para variables categoricas

Esta matriz se usa para modelos que clasifican los datos en dos categorías y muestra qué tan bien reconoce cada una. Por ejemplo, puede hacer referencia a la presencia o ausencia de una enfermedad, o a la necesidad de entregar o no un beneficio social.

Normalmente se escoge la categoría que es más importante predecir como la positiva, es decir, siguiendo el ejemplo anterior, con enfermedad, y la restante como la negativa, o sea sin enfermedad.

La matriz permite cruzar las variables que resultan de la predicción del modelo, en contraste con aquellas que son observables del conjunto de datos de prueba que se utilizaron para construirlo.

Tabla 2: Matriz de Confusión.

PREDICCIÓN DEL MODELO	VALOR REAL *Positivo*	VALOR REAL *Negativo*
POSITIVO	**Verdaderos Positivos (VP)** *Datos que el modelo detectó positivos y realmente eran positivos.*	**Falsos Positivos (FP)** *Datos que el modelo detectó positivos, pero realmente no lo eran. Por ejemplo, el modelo puede indicar que una persona tiene una enfermedad, cuando los datos observables indican que no la tiene.*
NEGATIVO	**Falsos Negativos (FN)** *Datos que el modelo detectó negativos, pero realmente eran positivos. Por ejemplo, el modelo indica que una persona no tenía una enfermedad, cuando los datos observables indican que sí la tenía.*	**Verdaderos Negativos (VN)** *Datos que el modelo detectó negativos y realmente eran negativos.*

Fuente: Adaptado de Chen, Rubin y Cornwall (2021).

Con la matriz de confusión se obtiene el desempeño del modelo. En algunos escenarios, es más importante que el modelo prediga lo mejor posible la categoría positiva, aunque no detecte tan bien la negativa; en otros casos, es necesario lo contrario.

Para algunas situaciones, es importante que el modelo prediga ambas categorías lo mejor posible, aunque no sean perfectas. Saber qué indicador es más importante para el modelo depende de los objetivos del proyecto, el problema a solucionar y en algunos casos incluso de los recursos disponibles. Según estos resultados se debe decir si el proyecto necesita mayor precisión, sensibilidad, especificidad y/o exactitud.

PRECISIÓN

Se refiere a qué tan confiable es el modelo al detectar la categoría positiva. Entrega el número de casos que realmente son positivos entre la cantidad de positivos que detectó. En el ejemplo anterior, esta medida diría cuántas personas que fueron identificadas por el algoritmo cómo que tienen una enfermedad, realmente la tenían.

$$Precisión = \frac{VP}{VP + FP}$$

SENSIBILIDAD

Es para saber qué tan bueno es el modelo detectando la categoría positiva. Entrega el número de verdaderos positivos entre todos los positivos reales. En el ejemplo, esta medida diría cuántas personas que tienen la enfermedad realmente fueron detectadas como tal.

$$Sensibilidad = \frac{VP}{VP + FN}$$

ESPECIFICIDAD

Mide qué tan bueno es el modelo detectando la categoría negativa. Entrega el número de verdaderos negativos entre todos los negativos reales. En el ejemplo, esta medida diría cuántas personas a las cuales no se les detectó la enfermedad, realmente estaban sanas.

$$Especificidad = \frac{VN}{VN + FP}$$

EXACTITUD

Muestra la cantidad de predicciones positivas y negativas que realizó el modelo. En el ejemplo, esta medida diría cuántas personas fueron clasificadas en la categoría a la que realmente pertenecían.

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VP}$$

Por lo demás, **los costos en términos económicos y sociales de un falso positivo y un falso negativo pueden ser muy distintos dependiendo del proyecto.** Se sugiere hacer un análisis de costo-beneficio en el que se evalúen las implicancias de cada uno en el contexto en el que se quiere implementar el modelo.

Por último, además de considerar que el algoritmo haga un buen trabajo desde el punto de vista técnico para toda la población, también se debe resguardar la ética asociada a su despliegue. Por ejemplo, se puede requerir medir el desempeño del modelo para subgrupos que hayan sido identificados como prioritarios según la problemática, tales como habitantes de sectores rurales o personas mayores, entre otros, de manera de asegurar que su desempeño sea equitativo.

Otros instrumentos que complementan este análisis a la hora de evaluar un modelo de clasificación son la curva *ROC* (por sus siglas en inglés, *Característica Operativa del Receptor*) y el *AUC* (por sus siglas en inglés, *Área Debajo de la Curva*). Ambas son herramientas gráficas que reflejan el desempeño del modelo y se desprenden de la matriz de confusión[30].

La ventaja principal de estas métricas es que ayudan a ajustar el modelo dependiendo de lo que es más importante para el problema, puesto que no es posible que todas las medidas tengan un valor perfecto. Por ejemplo, si se usa un modelo para entregar un beneficio social, y éste es de extrema importancia para las personas, se puede privilegiar la tasa de falsos positivos *(FPR)*, para asegurar que la mayor cantidad de verdaderos positivos pasen el filtro y reciban el beneficio. Por otro lado, si el presupuesto para el proyecto es ajustado, se puede querer privilegiar la tasa de verdaderos positivos *(TPR)*, ya que el beneficio social sería entregado a las personas que el modelo está seguro de que lo necesitan más.

b. MSE, R^2, AIC y BIC: instrumentos de ajuste para variables categóricas y continuas

Un modelo muy común en ciencia de datos es la regresión para predecir el comportamiento de un fenómeno en el futuro, ya sea con variables de interés de tipo categórica o continua. En estos casos, se estudia el comportamiento de una variable dependiente en relación a las otras variables que se encuentran en los datos. Para ello, los dos instrumentos más comunes para medir el desempeño de un modelo de este tipo son el *Error Cuadrático Medio* (*MSE* por sus siglas en inglés), el coeficiente de determinación o R^2, el *Criterio de Información de Akaike* (*AIC* por sus siglas en inglés) y el *Criterio de Información de Schwarz o Bayesiano* (*BIC* por sus siglas en inglés).

- **El MSE** mide el promedio del error entre el valor que predice el modelo y el valor real observado. Para un modelo perfecto, el *MSE* es 0, y a medida que aumenta, peor sería el desempeño del modelo.

30. Por un lado, la curva ROC se traza calculando el cociente entre los verdaderos positivos (TPR), que es igual a la Sensibilidad, y los falsos positivos (FPR), que se calcula como 1-Especificidad. Luego, el AUC se refiere al área bajo la curva ROC, que mientras más cerca esté de 1 mejor clasificador se tiene en general.

- **El R^2** representa qué tan cerca están los datos reales de la línea de predicción que generó el modelo de regresión. Un R^2 igual a 0% significa que el modelo no captura el comportamiento de los datos y, por otro lado, uno de 100% significa que lo hace a la perfección. Cabe destacar que los valores bajos no siempre implican que un modelo sea malo ni viceversa. Dadas las características de los datos, normalmente los modelos de regresión que buscan predecir el comportamiento humano suelen tener un R^2 menor al 50% aunque el modelo sea correcto. Por otro lado, un modelo con un R^2 muy cercano a 100% podría significar que el modelo está sesgado o que se está haciendo un sobreajuste de los datos. En ese sentido, se sugiere verificar con el R^2 ajustado, el cual corrige su valor por la cantidad de variables explicativas del modelo, por lo que lo castiga si se utilizan muchas.

- **Los AIC y BIC** sirven para comparar dos o más modelos que tienen la misma variable de interés pero distintas variables explicativas, penalizando por el número de variables incluidas. Como regla general, se preferirán modelos con menores valores *AIC* y *BIC*.

c. Validación cruzada para modelos de inteligencia artificial

En modelos de aprendizaje automático la validación cruzada es una técnica utilizada para validar sus resultados y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. **Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones.** Normalmente, se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica.

Paso 13
Generar conclusiones

Con un modelo prototipado y testeado, llega el momento de interpretar los resultados y obtener un modelo de ciencia de datos validado por todo el equipo, desde un punto de vista técnico y sustantivo, lo que permitirá avanzar a la fase siguiente con seguridad.

Si bien ocurren instancias de análisis e interpretación a lo largo de todo el proyecto, el propósito de este paso es distinto, ya que busca que todas las personas integrantes a nivel general evalúe y valide la solución. Asimismo se acuerda que la solución está lista para ser probada en un contexto real y no debiese tener modificaciones mayores en el corto plazo. Para ello, es relevante contar con el respaldo de las autoridades relacionadas con el problema, a fin de disponer del tiempo y todo tipo de recursos necesarios para avanzar. Esto permitirá dar inicio a la Fase 3 de desarrollo del piloto. Para orientar la reflexión, se sugieren tres principios adaptados de *Peng y Matsui (2015)*.

a. Revisitar los objetivos del proyecto

Consiste en recordar a todo el equipo los objetivos del proyecto acordados al término de la Fase 1. Aunque pueda parecer simple, este es un principio elemental ya que los objetivos se encuentran en el origen de esta fase de trabajo y son los que orientan y movilizan las acciones del proyecto.

Desde el punto de vista del análisis, los objetivos entregan un marco interpretativo de los resultados del modelo, mantienen la solución anclada en los comportamientos humanos de interés, y contienen los resultados esperados que servirán para la toma de decisiones. Solo si no se ha realizado aún, este es un excelente momento para revisitar los riesgos identificados en el análisis de prefactibilidad (Paso 3 de esta Guía), para evaluar si son apropiadamente abordados en esta iniciativa.

b. Evaluar el resultado del modelo

Consiste en evaluar, desde un punto de vista estadístico, el funcionamiento del modelo y de los resultados que entrega. La forma de realizar esto depende del tipo de análisis que se haya realizado. Dado que esta Guía no busca ser exhaustiva en estos términos específicos, a continuación se presentan indicaciones generales que buscan orientar la reflexión del equipo del proyecto.

Algunos elementos a considerar son:

- **ASOCIACIÓN ENTRE LAS VARIABLES**
  ¿La dirección, magnitud, y precisión de las estimaciones es la esperada? ¿Sirve para cumplir los objetivos del proyecto?

- **PARSIMONIA DEL MODELO**
  ¿El modelo es lo suficientemente simple para entenderlo, al mismo tiempo que cumple con su propósito?

- **CALIDAD DEL ANÁLISIS**
  ¿Cómo se compara el desempeño del modelo con la línea base? Si se trata de un modelo predictivo ¿es aceptable de acuerdo al criterio experto del equipo, cantidad de falsos negativos y falsos positivos que entrega el modelo?

- **JUSTICIA DEL MODELO**
  ¿El desempeño del modelo es equitativo entre subgrupos de la población? Si no es así ¿cuenta con medidas de mitigación?

- **AJUSTE DEL MODELO**
  ¿El modelo explica buena parte de la variabilidad de los datos?

- **SENSIBILIDAD DEL MODELO**
  ¿Los resultados se sostienen si se excluyen los casos extremos *(outliers)*? ¿Y si se introducen otras variables que puedan operar como factores de confusión *(confounders)*?

c. Considerar la evidencia externa al modelo

Este principio busca desafiar el modelo con una mirada amplia que considere lo que ya se sabe de otros proyectos en el área, así como traer aprendizajes de otras áreas para enriquecer el análisis. Esta evidencia puede venir del conocimiento del equipo sobre el tema en cuestión, otros proyectos similares, resultados de análisis afines, o información sobre la población objetivo que no ha sido incluida al modelo.

Pensemos, por ejemplo, en el *Sistema Integrado de Información*[31] para el seguimiento domiciliario de pacientes COVID-19, impulsado por el Servicio de Salud Metropolitano Sur Oriente. Este sistema busca automatizar el seguimiento de síntomas de COVID-19 mediante llamados pregrabados y alertar a la autoridad sanitaria cuando detecte casos de mayor riesgo.

Los desafíos que se presentan son múltiples. Respecto al resguardo, almacenamiento y uso de información personal y sensible de las personas usuarias, el proyecto se podría enriquecer con aprendizajes del proyecto *WhatsApp del Ministerio de la Mujer y Equidad de Género*[32], que automatizó el registro y derivación de contactos realizados por violencia de género mediante un canal silencioso de atención 24/7. De igual modo, respecto de cómo entregar las notificaciones a los responsables y las acciones que se desencadenan, el proyecto se podría beneficiar con la experiencia del *Sistema de Alerta Temprana de Deserción Escolar*[33], que mediante el análisis de datos identifica a estudiantes en riesgo de deserción y activa un procedimiento para prevenirla.

Un excelente espacio para posibilitar este tipo de intercambios es la *Red de Innovadores Públicos*[34] del Laboratorio de Gobierno. En ella, sus más de 22.000 integrantes pueden conectarse con otros servidores públicos, emprendedores, académicos, estudiantes, dirigentes sociales, profesionales y ciudadanos, compartir sus experiencias directamente y aprender de ellos.

31. Disponible en bit.ly/3FkTWgP
32. Disponible en lab.gob.cl/casos/13
33. Disponible en bit.ly/3st2B9l
34. Disponible en innovadorespublicos.cl

Tras el análisis de toda la evidencia, el equipo del proyecto podría decidir que la solución aún se podría ajustar, ya que no está orientada al logro de los objetivos del proyecto o que su desempeño es insuficiente. En este caso, se puede volver a recorrer los pasos de esta fase para hacer los ajustes necesarios que permitan cumplir los objetivos satisfactoriamente. En caso de que se concluya que no es posible cumplir los objetivos del proyecto, se podría plantear un nuevo proyecto con objetivos distintos, lo que implica recorrer nuevamente la Fase 1 de *"Investigación del problema"*.

En cambio, si el equipo concluye que la solución permite cumplir los objetivos del proyecto, muestra resultados satisfactorios, y ha sido puesta a prueba exitosamente, se llega al fin de esta fase de *"Diseño de propuestas de solución"* y se da inicio a la siguiente fase.

Para asegurarse de considerar todos estos elementos, se sugiere emplear la **Ficha de Consolidado de la Solución** (Herramienta VIII). Además esta permite acordar acciones para obtener el mejor modelo posible.

HERRAMIENTA VIII

Ficha de Consolidado de la Solución

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura del proyecto debe **convocar** al equipo del proyecto a un taller presencial o remoto. Es indispensable que en este taller participen quienes fueron parte del desarrollo del modelo de ciencia de datos y quienes han conformado el equipo del proyecto desde el comienzo.

2. La jefatura del proyecto hará una **recapitulación** de los objetivos del proyecto, las fuentes de datos, los tipos de análisis y las técnicas de análisis empleadas. Luego, **se registran en los espacios asignados** para ello en la herramienta.

3. En conjunto, los asistentes deben **identificar evidencia externa al modelo** que pueda ser considerada para desafiar su funcionamiento. Esta debe ser preparada por las personas asistentes antes de asistir al taller.

4. En conjunto, se deben **responder las preguntas que permiten evaluar al modelo** en torno a su parsimonia justicia, sensibilidad y validación.

5. **Las respuestas deben ser acordadas** por la mayoría de las personas participantes. **En caso de que no exista acuerdo** para responder una pregunta o se entregue una pregunta no satisfactoria, **se deben identificar acciones para resolverla.** En ese caso será necesario volver a pasos previos de esta fase, siguiendo un trabajo iterativo de ensayo, error y ajuste.

PRODUCTO DE LA HERRAMIENTA

Esta herramienta tiene dos posibles productos. El primero es un listado de acciones que permitan ajustar y validar el modelo. El segundo es un amplio acuerdo en el equipo de proyecto de que la solución está en condiciones de ser piloteada.

OBJETIVO(S)

Escribir los objetivos levantados en la Herramienta VI

FUENTE(S) DE DATOS

Escribir las fuentes de datos extraídas y cargadas en el Paso 8

TIPO(S) DE ANÁLISIS

Escribir los tipos de análisis definidos en la Herramienta VII

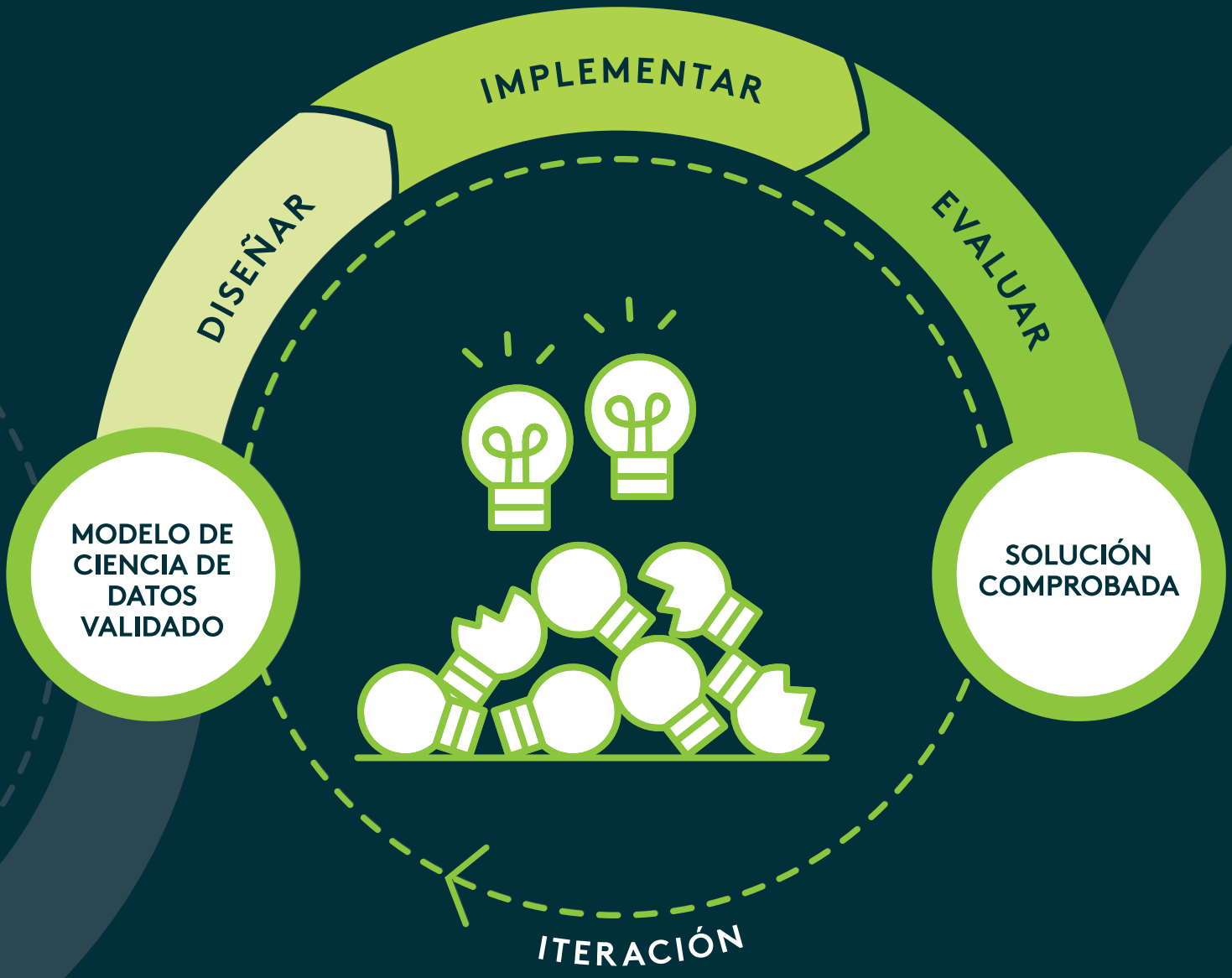
TÉCNICA(S) DE ANÁLISIS

Escribir las técnicas de análisis establecidas en el Paso 11

EVIDENCIA EXTERNA

¿Qué aprendizaje de otras áreas, modelos o soluciones podríamos necesitar en este proyecto?

CRITERIOS DE EVALUACIÓN DEL MODELO	PREGUNTA	RESPUESTA	ACCIONES REPARATORIAS
PARSIMONIA	**¿El modelo es lo suficientemente simple para entenderlo?**	*No. Cuenta con numerosas variables y no ha sido fácil de comunicar a personas externas al proyecto.*	*Revisar la cantidad de variables utilizadas y simplificar la explicación del modelo.*
JUSTICIA	**¿El modelo incluye por igual a todos los grupos de la población sin discriminar?**	*Sí. No existen grupos que sean discriminados por el modelo.*	-
SENSIBILIDAD	**¿Los resultados se sostienen si se excluyen los valores atípicos?**	*No se sabe. No se ha realizado este ejercicio.*	*Realizar un análisis diferenciado para los principales subgrupos de la población.*
	¿Los resultados se sostienen si se introducen otras variables que puedan operar como factores de confusión?	*Sí. El modelo incluye numerosas variables que permiten predecir con precisión el resultado.*	-
VALIDACIÓN	**¿El modelo fue validado con las métricas correspondientes a su tipo de análisis (Paso 12)?**	*Sí, y mostraron un buen desempeño.*	*Se deberá realizar nuevamente este análisis una vez que se lleven a cabo las acciones comprometidas en esta herramienta.*


DISEÑAR
IMPLEMENTAR
EVALUAR
MODELO DE CIENCIA DE DATOS VALIDADO
SOLUCIÓN COMPROBADA
ITERACIÓN

Fase 3
Desarrollo del piloto

Productos

Solución de ciencia de datos comprobada, considerando el impacto, adopción y funcionamiento de la solución en un contexto real.

La Fase 3 consiste en poner a prueba la solución diseñada en condiciones reales controladas.

En esta fase se podrán identificar problemas prácticos en la implementación y estimar el impacto de la solución.

El hito que marca el fin de esta fase es el acuerdo entre todo el equipo del proyecto de que los resultados del piloto son satisfactorios y suficientes para avanzar hacia la implementación.

ESTA FASE CONTIENE
3 PASOS:

DISEÑAR

Paso 14
Diseñar evaluación de impacto

IMPLEMENTAR

Paso 15
Implementar el piloto de la solución

EVALUAR

Paso 16
Evaluar resultados del piloto

Paso 14
Diseñar evaluación de impacto

En esta Guía, **el desarrollo del piloto está centrado en evaluar el impacto de la solución en un contexto real. Esto permitirá conocer con certeza si la solución permite lograr los objetivos del proyecto.** Dado que las decisiones de este paso requieren conocimiento técnico en estimación de impacto, medición y muestreo, se sugiere que el analista de datos tome mayores responsabilidades.

a. Estimación de impacto

El primer elemento a considerar es el diseño de la evaluación de impacto. Para esto se recomienda emplear diseños experimentales, es decir, donde se conformen aleatoriamente dos grupos, uno que reciba la solución (grupo de tratamiento) y otro que no (grupo de control). Estos diseños facilitan la identificación de la relación causal entre la solución y los resultados de interés, ya que eliminan los posibles factores de confusión y generan grupos cuya única diferencia estadística es haber recibido la solución. Al comparar los resultados de ambos grupos, se podrá estimar el impacto de la solución.

>> GRUPO DE CONTROL
Grupo conformado aleatoriamente que no recibe la solución. En una evaluación de impacto funciona como comparación del grupo de tratamiento que si recibe la solución.

Por ejemplo, *Dame esos 5*[35] fue un proyecto conjunto entre la Subsecretaría de Educación Parvularia y el Laboratorio de Gobierno, cuyo objetivo fue favorecer el desarrollo de niñas y niños en casa durante la pandemia. Para lograrlo, se implementó el envío de mensajes automatizados vía *WhatsApp* a personas cuidadoras de niñas y niños de 0 a 3 años, con contenido que buscaba fomentar cinco principios que los adultos pueden poner en práctica para mejorar el desarrollo integral de quienes tienen a su cuidado.

Para evaluar el impacto de la solución, se seleccionó una muestra de 27.897 personas cuidadoras, de las cuales 16.738 fueron elegidos aleatoriamente para recibir los mensajes durante cinco semanas, y 11.159 como grupo de control. Tras este periodo, se aplicó una encuesta para estimar el impacto en los comportamientos esperados entre personas cuidadoras y niños y niñas (leer, hablar de emociones, enumerar, etc.).

Los resultados mostraron un aumento de un 32,4% más de lectura y 8,5% más de actividad física en cuidadores que ingresaron al canal de *WhatsApp* en comparación con el grupo de control, lo que justificó escalar la solución a una población objetivo de más de 400.000 personas cuidadoras.

35. Disponible en bit.ly/3DBDsjj

Estos diseños, sin embargo, suelen enfrentar múltiples limitaciones para su implementación. Por ejemplo, desde un punto de vista práctico, podría resultar muy difícil tener dos formas de asignación de beneficios funcionando paralelamente, una basada en un modelo de ciencia de datos y otra más tradicional, ya que no se puede detener la entrega del servicio. Desde un punto de vista ético, tampoco sería aceptable entregar una solución de manera diferenciada (o excluir a un grupo de ella) con el único propósito de estimar su impacto, sobre todo cuando se cuenta con sólida evidencia previa de que será una intervención beneficiosa. En estas ocasiones, se puede optar por hacer uso de diseños cuasi experimentales para estimar el impacto. Los más conocidos son:

DIFERENCIAS EN DIFERENCIAS	También denominado *diff-in-diff*, permite comparar los cambios en el estado de los grupos de tratamiento y control a lo largo del tiempo. Para su ejecución se requiere medir las variables de interés de ambos grupos antes y después de la entrega de la solución.
PAREAMIENTO POR PUNTAJE DE PROPENSIÓN	En inglés *Propensity Score Matching*, permite simular la asignación aleatoria cuando se cuentan con datos posteriores a la entrega de la solución. Es particularmente útil cuando el uso de la solución es voluntario y no se cuenta con grupos de control y tratamiento asignados aleatoriamente. Para su implementación se requiere que en ambos grupos existan personas con baja y alta probabilidad de participación en variables observadas.
REGRESIÓN DISCONTINUA	Cuando el acceso a la solución depende de alcanzar un puntaje continuo (cierto nivel de ingreso, por ejemplo), se puede comparar cerca del umbral de selección los resultados de los grupos de tratamiento y control. Para su implementación se requiere el índice de clasificación y el umbral de selección.
VARIABLES INSTRUMENTALES	Permite simular la asignación aleatoria a través de una variables que afecte la participación en el programa, pero no influya directamente en sus resultados.

Una guía introductoria a la evaluación de impacto recomendable, en la que se discuten estos y otros temas con mayor profundidad, es la publicación *Evaluación de Impacto en la Práctica*[36] de Gertler, Martínez, Premand y Rawlings (2017) la que se puede descargar de manera libre.

36. Disponible en bit.ly/3Dxogn9

b. Medición

Un segundo elemento a considerar en el desarrollo del piloto es la medición. Esto incluye los indicadores clave y de funcionamiento de la solución. Los primeros permiten evaluar el impacto de la solución y los segundos permiten conocer si las personas usuarias usaron la solución y si se implementó tal como fue diseñada.

- Típicamente, **indicadores clave** son la satisfacción de las personas usuarias, tiempos de espera de las personas usuarias para obtener una respuesta, cantidad de documentos entregados en un periodo de tiempo, cantidad de postulaciones recibidas, cantidad de personas usuarias atendidas, entre otros.

- En tanto, **indicadores de funcionamiento** son la cantidad de personas usuarias que usaron la solución, cantidad de veces que las personas usuarias interactúan con la solución, tiempo de interacción de las personas usuarias con la solución, tiempo que la solución permanece disponible, entre otros.

En ambos casos, es necesario velar por obtener mediciones válidas y confiables. Desde un punto de vista metodológico, esto significa que las mediciones midan aquello que se proponen medir, y que sus resultados no varíen entre aplicaciones (o lo hagan lo menos posible). Si se trata de una encuesta, una buena práctica consiste en emplear preguntas que se hayan aplicado previamente en contextos similares o usar datos levantados por un tercero, lo que reduce los riesgos de introducir sesgos o errores. Si se trata de sensores, deben estar bien ubicados, calibrados, resguardados del ruido de medición, mantenidos regularmente, etc.

c. Muestra

Finalmente, se debe definir el tamaño muestral. En una evaluación de impacto, es clave que ésta permita detectar el efecto que se espera de la solución, lo que se conoce como el efecto mínimo detectable. No basta con estimar el resultado final con precisión en el grupo de tratamiento, sino que también se pueda comparar con el resultado del grupo de control. La diferencia entre estos grupos es lo que se busca estimar. A mayor diferencia entre los grupos, es necesario un menor tamaño muestral, pero si se espera que el efecto sea pequeño, se necesitará una muestra más grande. Este cálculo implica contar con los siguientes elementos:

- **Población objetivo**: Es el grupo de la población sobre el que se quiere inferir.
- **Marco muestral:** Es un listado de todos los elementos de la población objetivo.
- **Efecto mínimo detectable:** Corresponde al cambio mínimo que se espera tras la aplicación de la solución, lo que debería estar basado en los objetivos del proyecto.
- **Significancia estadística deseada:** Es la probabilidad de que los resultados obtenidos se deban al azar. Típicamente se utiliza un 5%.
- **Potencia estadística:** Es la probabilidad de que no se detecte un efecto de la solución cuando efectivamente no hay un efecto. Dicho de otro modo, es la probabilidad de rechazar la hipótesis nula cuando la hipótesis nula es falsa. Típicamente se usa con un 80% de probabilidad.

Para identificar qué medir, cómo medir, cuáles son los resultados de hoy (si es que los hay), y cuál es el resultado esperado, se puede emplear la herramienta de **Diseño de Evaluación** (Herramienta IX).

HERRAMIENTA IX

Diseño de Evaluación

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura debe **convocar** al equipo del proyecto a una sesión de trabajo.
2. **Redactar en la primera columna los objetivos** que fueron formulados en la herramienta VI.
3. **Registrar los indicadores de interés** en la segunda columna, alineados con cada objetivo.
4. **Registrar la medición actual** de dichos indicadores en la tercera columna, si es que existe.
5. El equipo debe **acordar y registrar cuál es el resultado esperado** del piloto en la cuarta columna, cuál es el efecto mínimo detectable. Para esto se debe basar en los objetivos del proyecto.
6. **Definir cómo se va a capturar la evidencia** para calcular los indicadores descritos previamente.

PRODUCTO DE LA HERRAMIENTA

Resumen del diseño de la evaluación.

OBJETIVOS FORMULADOS EN LA PRIMERA FASE *Escribir lo(s) objetivo(s) levantados en la Herramienta VI.*	**¿QUÉ VAMOS A MEDIR?** *Indicadores de interés.*	**¿QUÉ RESULTADOS ARROJA EL GRUPO DE CONTROL? ¿CUÁL ES LA LÍNEA BASE?** *Medición de indicadores asociados al grupo que no recibe la solución.*	**¿QUÉ RESULTADOS ESPERAMOS?** *A partir de qué nivel de impacto consideramos exitoso el piloto.*	**¿CÓMO VAMOS A CAPTURAR LA EVIDENCIA?** *Instrumentos de recogida de información.*
	% de personas cuidadoras que implementan las herramientas dispuestas. *Autoeficacia de las personas cuidadoras.*	*Indicadores del grupo de control.*	*A partir de un aumento del 10% en el uso de las herramientas entregadas.* *A partir de un aumento del 5% de la autoeficacia de las personas cuidadoras.*	*Encuesta vía WhatsApp.*

Paso 15
Implementar el piloto de la solución

La implementación del piloto involucra una gran variedad de aspectos operativos. Aquí los hemos agrupado en recursos, responsables, lugares de aplicación y comunicaciones. Vale mencionar que se han dejado fuera otros aspectos que podrían ser necesarios, con el objetivo de mantener esta Guía centrada en los más relevantes para un proyecto de ciencia de datos.

a. Recursos

Los recursos necesarios para implementar una solución de ciencia de datos pueden ser de dos tipos, principalmente.

- Por un lado, son necesarios **recursos tecnológicos físicos** en el caso que fuera necesario disponibilizar dispositivos para que las personas usuarias puedan acceder a la solución en oficinas de atención, o para que sea operada por funcionarias y funcionarios en oficinas o fuera de ellas.
- Por otro lado, es necesario identificar si son necesarios **recursos financieros** para acceder a almacenamiento en línea o escalar la capacidad de cómputo en máquinas virtuales.

b. Responsables

Junto con los recursos, se deben designar a los **responsables de mantener el modelo y de ejecutar las actividades clave.** La mantención es una actividad crítica, y la persona responsable debe contar con alta disponibilidad para solucionar rápidamente los problemas que puedan surgir durante el piloto, tales como intermitencias en su funcionamiento, entrega de información confusa, clasificaciones/predicciones erróneas, etc.

Si la solución no está completamente automatizada, también puede ser necesario asignar responsables que ejecuten las actividades clave. Por ejemplo, responder preguntas de las personas usuarias con casos complejos, registrar datos en una plataforma para que el modelo pueda trabajar con ellos, llamar a las personas usuarias para hacer seguimiento de casos, etc. Sin embargo, **en estos casos no solo se debe considerar la capacitación de los responsables en el uso de la solución o el cambio en los procesos, si no que también se debe velar por una gestión del cambio que impida la caída el bienestar y productividad de los funcionarias y funcionarios en la organización.**

c. Lugar de aplicación

El lugar de aplicación es un elemento opcional que **puede ser necesario considerar si la solución que se ofrece está asociada a un espacio físico**, como es es el caso de recintos de salud, establecimientos educacionales, oficinas de atención a la ciudadanía, entre otros. En estos casos, una acción necesaria es conseguir la autorización y el respaldo de las autoridades locales o jefes de sucursales para la ejecución del piloto en su territorio. La selección de esos lugares viene dada por la muestra definida en el paso anterior, aunque es importante mantener una escala reducida durante el piloto (pocas sucursales) para que sea fácilmente administrable. Ahora bien, si la solución se ofrece únicamente de manera online y no está asociada a un territorio particular, como en el caso de *Dame esos 5*, no es necesario considerar este elemento y la muestra para la realización del piloto puede ser mucho más extendida.

d. Comunicación

Una vez que se tiene claridad de los aspectos operativos anteriores, es necesario **definir cómo comunicar la existencia y funcionamiento de la solución.** Siguiendo las recomendaciones de la *Guía de Lenguaje Claro: ¿Cómo podemos generar una comunicación simple, clara y efectiva entre el Estado y la ciudadanía?*[37], el proceso de co-creación y testeo de productos comunicacionales debe considerar su contenido, estructura y visualización. Para los proyectos de innovación en ciencia de datos, que pueden considerar la existencia de modelos complejos, también es necesario considerar la elaboración de manuales o instructivos para el uso de la solución. La gestión de estos aspectos debe estar a cargo de la jefatura del proyecto.

Para resguardar la existencia de las tareas mencionadas para la implementación y evaluación, se puede usar la **Ficha de Implementación del Piloto** (Herramienta X). Esta consiste en un listado con preguntas que permiten asegurar que se cuente con todos los elementos necesarios para la implementación de la solución.

37. Disponible en bit.ly/3TEzMTr

HERRAMIENTA X

Ficha de Implementación del Piloto

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura debe **convocar** a una mesa de trabajo al equipo del proyecto junto con las personas responsables de los pasos "Diseñar evaluación de impacto" e "Implementar el piloto".
2. **Exponer** los pasos "*Diseñar evaluación de impacto*" e "*Implementar el piloto*" al equipo del proyecto para que lo validen.
3. **Marcar las casillas "Sí" o "No"** para responder a cada pregunta de la segunda columna. En caso de responder una pregunta con un No, completar la columna de tareas faltantes.

PRODUCTO DE LA HERRAMIENTA

Se identifican cuál son las tareas faltantes para desplegar el piloto de la solución.

DIMENSIÓN		PREGUNTA	SÍ	NO	TAREAS FALTANTES
EVALUACIÓN		**¿Se definió un diseño para la evaluación?**	✓		*No se necesitan*
		¿Se definió una forma de medir los indicadores clave?	✓		*No se necesitan*
		¿Se definió una forma de medir los datos de funcionamiento?	✓		*No se necesitan*
		¿Se definió la muestra donde se aplicará el piloto?		✓	*La jefatura del proyecto debe validar la lista de personas levantadas de forma aleatoria por el equipo de analistas de datos.*
IMPLEMENTACIÓN	RECURSOS	**¿Existen los recursos financieros para ejecutar el piloto?**	✓		*No se necesitan*
		¿Existen los recursos tecnológicos para sostener la solución?	✓		*No se necesitan*
		¿Existen los recursos humanos para ofrecer la solución?		✓	*Las jefaturas del lugar de implementación deben aprobar que un grupo de colaboradores de su organización destinen tiempo al piloto de la solución.*
	RESPONSABLES	**¿Están asignados los responsables de mantener el modelo?**		✓	*Una vez que se apruebe el recurso humano se podrán asignar las responsabilidades correspondientes*
		Si el funcionamiento de la solución depende de funcionarias y funcionarios, ¿fueron designados los responsables de ejecutar las actividades clave?		✓	
		Si el funcionamiento de la solución depende de funcionarias y funcionarios, ¿fueron capacitados los responsables de ejecutar las actividades clave?		✓	*Una vez que se asignen las responsabilidades se ejecutará el plan de capacitación que preparó el equipo del proyecto.*
	CONTEXTO	**¿Están definidos los lugares (ej: región, provincia, comuna, sucursal, etc) donde se implementará el piloto?**	✓		*No se necesitan*
		¿Está definido el período en el cuál se implementará el piloto?	✓		*No se necesitan*
		¿Las autoridades locales autorizaron la ejecución del piloto en su territorio?	✓		*No se necesitan*
	COMUNICACIONES	**¿Existe una forma de comunicar a las personas usuarias o funcionarias la existencia de la solución?**	✓		*No se necesitan*
		¿Existen capacitaciones, manuales o instrucciones para el uso de la solución?	✓		*No se necesitan*

Paso 16
Evaluar resultados del piloto

El fin de la fase de piloto ocurre cuando hay una solución comprobada en términos de impacto y adopción. Es crucial que todo el equipo de trabajo acuerde que la solución cumple con los objetivos propuestos, ya que ese es el principal indicador de logro del proyecto. De lograrlo, se deben diseñar mecanismos de difusión de la solución exitosa. Por otro lado, resultados adversos pueden implicar ajustes para intentarlo de nuevo.

a. Impacto

Al referirnos al impacto de la solución, lo primero que se debe hacer es cuantificarlo. Si se empleó un diseño experimental, el impacto se obtiene con una simple diferencia de medias de la variable de interés entre los grupos de tratamiento y control. Una formulación típica de impacto es "Para las personas usuarias en el grupo de tratamiento, en promedio, el uso de la solución se asocia a un aumento de 8% en la variable X, una reducción de 300 unidades en la variable Y, además de un aumento de 23% en la probabilidad de Z, en comparación al grupo de control". La suficiencia de estos resultados se pone en perspectiva comparándolos con los objetivos del proyecto.

En este punto, sin embargo, todavía se debe considerar si el diseño de la evaluación podría tener algún problema de validez externa. **Esto quiere decir si el impacto se mantendrá al implementarlo en condiciones reales no controladas.** Un elemento importante a tener presente es si esta solución compite con otras medidas, o si los beneficios que genera compiten entre sí a gran escala. Por ejemplo, un modelo de ciencia de datos que permita la entrega automatizada de beneficios estatales, podría aumentar la entrega de beneficios en 50.000 al mes por persona, en comparación a las personas que postulan de manera tradicional. Sin embargo, aplicado a gran escala, esta solución podría ver limitado su impacto debido a restricciones financieras.

b. Adopción

En cuanto a la **adopción de la solución**, se sugiere prestar atención a tres elementos.

- Primero, a la cantidad de personas usuarias que, estando en el grupo de tratamiento, utilizaron la solución o la prefirieron por sobre otra. Una solución con alto impacto podría verse perjudicada si pocas personas usuarias la prefieren.

- Segundo, los problemas que enfrentaron las funcionarias y funcionarios o las personas usuarias al usar la solución, lo que podría dar respuestas a la pregunta anterior.

- Finalmente, la curva de aprendizaje del uso de la solución. Se debe considerar que soluciones de uso repetido podrían necesitar evaluar cuán rápido las funcionarias y funcionarios o las personas usuarias aprenden a usarla.

Para orientar la reflexión en torno al piloto, se propone usar la ficha de **Evaluación de la solución en un contexto real** (Herramienta XI). Está organizada en dos criterios, impacto y adopción, cada uno con tres preguntas verificadoras que buscan asegurar que la decisión sobre la implementación definitiva sea lo más exhaustiva posible. La herramienta está completada con el ejemplo de un proyecto que buscaba reducir la inasistencia a citas médicas.

Si el equipo acuerda que los resultados del piloto son suficientes para la inversión, pueden ser escalables y no existen mayores inconvenientes en el uso de la solución, entonces se considerará una solución comprobada que puede avanzar hacia la fase de implementación.

De lograr los objetivos, se deben diseñar mecanismos de difusión de la solución exitosa. Por otro lado, resultados adversos pueden implicar ajustes para intentarlo de nuevo.

HERRAMIENTA XI

Ficha de Evaluación de la Solución en un Contexto Real

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura debe **convocar** al equipo del proyecto, incluyendo al área de comunicaciones.
2. **Responder las preguntas del criterio de Impacto** para analizar los resultados del piloto, y evaluar si se alcanzaron los objetivos del proyecto definidos en la Herramienta VI.
3. **Responder las preguntas del criterio de Adopción** para articular la estrategia comunicacional que permita difundir, interna y externamente, la existencia de la solución y los resultados que ofrece.

PRODUCTO DE LA HERRAMIENTA

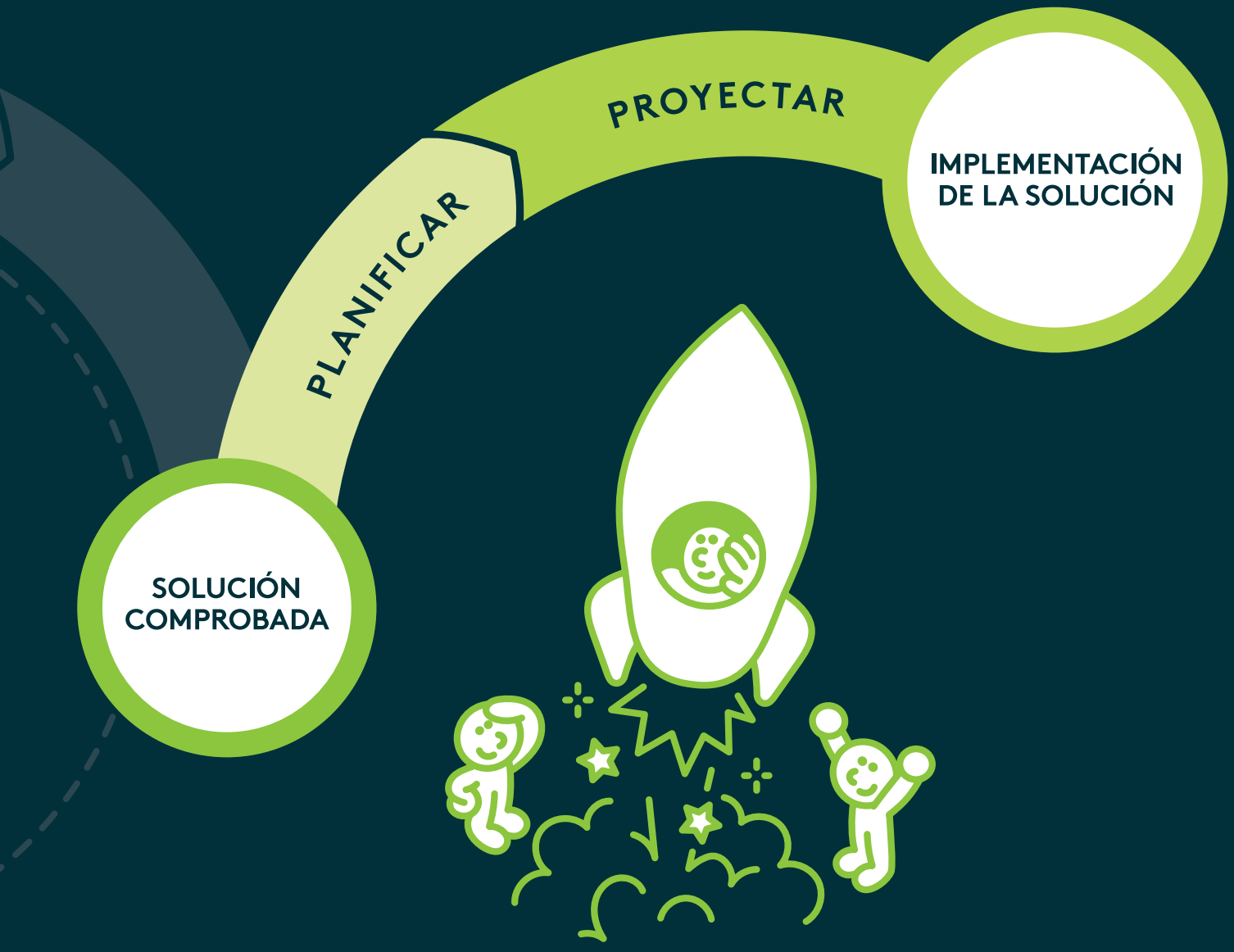
Resumen del diseño de la evaluación.

CRITERIO	PREGUNTAS VALIDADORAS	EJEMPLO: *Reducción de inasistencia a citas médicas.*
IMPACTO	**¿Cuál fue el impacto de la solución?**	*Un 25% menos de inasistencias a citas médicas.*
	¿El impacto es suficiente para justificar su implementación (ver efecto mínimo detectable en la Herramienta VIII)?	*Sí, se logra el objetivo del proyecto que era una reducción de 15% en las inasistencias.*
	De acuerdo a la validez externa del piloto, ¿es esperable que los resultados se mantengan o mejoren al implementar la solución?	*Sí, ya que las y los pacientes agendados podrían asistir. No hay competencia por un mismo servicio en este punto.*
ADOPCIÓN	**¿Cuántas personas utilizaron la solución?**	*De las y los pacientes con riesgo de inasistencia, se envió exitosamente un recordatorio vía WhatsApp al 85%, y el 40% respondió las llamadas de recordatorio.*
	¿Qué problemas enfrentaron al utilizarla?	*Personas usuarias con teléfonos válidos pero sin internet no lograban recibir los mensajes de WhatsApp.*
	Entre las personas que usaron la solución dos o más veces ¿demostraron un mejor uso?	*Las personas usuarias no usan directamente la solución (modelo de predicción de inasistencia y notificaciones), por lo que no aplica este criterio. De todos modos, las personas usuarias con riesgo de inasistencia en su primera cita, se redujeron en un 30% en su siguiente cita.*

SUGERENCIA Para responder las preguntas del criterio de Adopción sugerimos desplegar instrumentos de levantamiento de información con personas usuarias, tales como entrevistas, encuestas, grupos focales, observaciones participantes, sombras, safaris o etnografías, entre otras mencionadas en la ya citada *Guía Permitido Innovar: ¿Cómo podemos resolver problemas públicos a través de Proyectos de Innovación?*[38]

38. Disponible en bit.ly/3N6AGpl

Taller de testeo de las herramientas de la Guía con personas expertas y mienbros de la Red de Innovadores Públicos.


PLANIFICAR
PROYECTAR
SOLUCIÓN COMPROBADA
IMPLEMENTACIÓN DE LA SOLUCIÓN

Fase 4
Implementación de la solución

Productos

Solución de ciencia de datos implementada, considerando su monitoreo y robustecimiento.

Ahora corresponde pasar a una fase de despliegue de la solución en la que se implementará en el mundo real.

Esta fase requiere que se planteen objetivos en una línea de tiempo específica, que se capacite a las funcionarias y funcionarios que trabajarán directamente con el modelo y que se tome en cuenta la opinión pública.

La planificación del despliegue debe venir acompañada de un marco de gobernanza institucional, un plan de monitoreo frecuente del modelo y su consecuente robustecimiento.

ESTA FASE CONTIENE TRES PASOS:

PLANIFICAR

Paso 17
Planificar el despliegue

PROYECTAR

Paso 18
Monitorear el desempeño

Paso 19
Robustecer el modelo

Paso 17
Planificar el despliegue

El primer paso para la implementación de la solución a gran escala consiste en su planificación, la que hemos dividido en la determinación de una **gobernanza** con sus responsabilidades, la **gestión presupuestaria** y por último, la **estrategia** para lidiar con la resistencia al cambio.

Gobernanza

La gobernanza proporciona un marco de rendición de cuentas claro y específico para la iniciativa. En el plano general **considera dos grandes componentes; determinar la propiedad de la iniciativa y por otro lado, las actividades clave que se deben procedimentar para su sostenibilidad.**

Concretamente el área responsable de la iniciativa será dueña de los resultados generados por lo cuál debe velar por el rendimiento del modelo, la mantención de los sistemas asociados, la actualización del código, la rendición de cuentas a la ciudadanía y el mejoramiento de la solución.

Quienes sean responsables deben establecer mecanismos de control acordando con la institución cuáles serán las métricas de desempeño que se monitorearán para garantizar que la solución se mantenga entregando buenos resultados en el tiempo, tanto en sus fases iniciales como posteriores.

Por último, para que una gobernanza esté completa y sea eficaz en el logro de sus propósitos debe poseer las facultades para resolver conflictos, imprevistos y eventualmente poder arbitrar controversias. Asimismo, es clave que realice una nueva revisión de los aspectos jurídicos vinculados a la solución a implementar.

En este sentido, lo anterior resulta fundamental para asegurar la factibilidad normativa de la iniciativa, proponiendo cambios y/o ajustes normativos necesarios para su continuidad. Por ejemplo, ajustes a protocolos de seguridad de la información, o generación de convenios de colaboración y conectividad.

Gestión presupuestaria

Una gestión presupuestaria eficaz identifica, planifica y obtiene los recursos necesarios para la sostenibilidad de la iniciativa, evita los sobrecostos, reduce el riesgo de gastos imprevistos o estimaciones inexactas y perfecciona la planificación de próximas iniciativas. Dada la estructura presupuestaria del Estado de Chile, el presupuesto de la iniciativa también debe enmarcarse en una lógica anual e integrar las restricciones propias de los ciclos presupuestarios.

La gestión de costos es el proceso de estimar, asignar y controlar los costos de la solución, lo que permite a un equipo predecir los gastos futuros para reducir las desviaciones por sobre lo presupuestado. En ese sentido, esta planificación debe considerar tanto costos de implementación como de mantención, con sus certezas y riesgos, desde una perspectiva interna y externa.

Enfoque en las personas

La resistencia al cambio es un componente crítico dentro de los procesos de implementación de una nueva solución. **Las personas pueden sentir incertidumbre, rechazo o desconfianza frente a la solución que se está introduciendo. Es importante enfrentar este desafío presentando la información y comunicando de manera clara y oportuna**, para que tanto las funcionarias y funcionarios públicos que interactúan con el modelo, como las personas usuarias de la herramienta entiendan cómo funciona y cuáles son sus implicancias.

Con respecto a las funcionarias y funcionarios que forman parte de la institución, corresponde consolidar la relación que se ha trabajado desde el primer paso con la conformación del equipo. Es necesario diseñar una estrategia y reservar presupuesto para generar los instrumentos de comunicación, capacitación, *coaching*, y otros necesarios. Asimismo, entregar los espacios suficientes para levantar resistencias y comunicar los beneficios que tiene el cambio para ellos y su institución. Las personas funcionarias que tienen interacción directa o indirecta con el modelo deben ser capacitados adecuadamente para cumplir su rol y entender el funcionamiento del modelo. Es recomendable complementar el proceso de capacitación con un manual manual interno para el uso del modelo, al que sea posible recurrir cuando tengan dudas.

Para las personas usuarias afectadas se debe diseñar una estrategia comunicacional para entregar la información pertinente y oportunamente a la población. En primer lugar, se debe garantizar la protección de datos para dar seguridad a las personas de que su información personal y sensible está siendo resguardada y que sus datos están siendo utilizados de manera responsable. En segundo lugar, explicar el funcionamiento de la herramienta y sus implicancias en la ciudadanía. En la etapa de comunicación se debe evaluar cómo presentar los resultados más relevantes del proyecto, desde la forma en que se tomaron las decisiones, hasta los resultados encontrados.

Para diseñar esta estrategia de comunicación y definir la cantidad de información que se entregará se deben tener en cuenta los tipos de opacidad del algoritmo (Burrell, 2016).

LA OPACIDAD INTRÍNSECA

Se refiere a la dificultad de entender sistemas sumamente complejos. Por ejemplo, un árbol de decisión es más probable que presente una opacidad intrínseca baja, ya que es aparentemente más fácil de entender cómo las variables afectan al resultado. Sin embargo, un sistema que utiliza redes neuronales, ya no permite una lectura tan sencilla entre las variables y el resultado por las múltiples interacciones propias del modelo.

LA OPACIDAD INTENCIONAL

Alude a las características del algoritmo que no corresponde transparentar porque ponen en riesgo los objetivos. Por ejemplo, un sistema de predicción de infracciones al pago de impuestos, donde el potencial infractor, en caso de conocer en detalle cómo funciona el algoritmo, puede buscar otros resquicios para evitar la fiscalización y consecuente multa.

LA OPACIDAD ANALFABETA

Levanta el hecho de que el modelado de algoritmos de ciencia de datos es un proceso técnico y, por lo tanto, es probable que las personas tengan dificultades para comprenderlo. Se produce una opacidad que no viene dada por el sistema mismo sino que por una falta de conocimiento de la materia. Por ejemplo en conceptos como una capa oculta, o una semilla de aleatorización.

Cada tipo de opacidad requiere de una estrategia distinta y hay que considerar que un grupo de individuos puede no adaptarse a la solución pese a la capacitación, para lo cual deben prepararse vías de salida para estos.

Un sistema de comunicación con la ciudadanía consolidado incluye mecanismos de retroalimentación, protocolos de respuesta para los casos en que una persona usuaria solicite una explicación sobre algún aspecto de la iniciativa, instrucciones de uso y una sección de preguntas frecuentes. Cualquiera sea el caso, sugerimos siempre redactar las comunicaciones siguiendo los lineamientos de la *Guía de Lenguaje Claro: ¿Cómo podemos generar una comunicación simple, clara y efectiva entre el Estado y la ciudadanía?*.

Paso 18
Monitorear el desempeño

El monitoreo de la iniciativa parte por ser una responsabilidad clave de sus propietarios para implementar un marco sólido que detecte a tiempo los errores en el funcionamiento del modelo y, simultáneamente, vigile la entrega de un servicio responsable y justo para las personas usuarias.

Una vez la solución comienza con su despliegue en la población general, es importante crear mecanismos de monitoreo periódicos o "evaluación continua". Si existe algo del proceso con el que el equipo no esté satisfecho, se puede mejorar en esta etapa.

El propósito de esta etapa es ver cómo funciona la solución en un escenario real, detectar errores y posibles sesgos que no fueron visibles en la etapa de entrenamiento y durante el desarrollo del piloto.

Esta fase es útil para capturar nuevos aprendizajes que puedan surgir e incorporarlos a la solución para mejorar su desempeño.

Los modelos de ciencia de datos pueden degradarse y perder vigencia con el paso del tiempo. Esto ocurre por distintas razones como que el modelo supone que existe una relación estática entre los datos, pero en realidad estos interactúan y las relaciones cambian con el tiempo. También se puede producir un cambio en la calidad de los datos de entrada debido a que la manera en la que son recopilados cambia, sea un cambio en preguntas, sensores, automatización de algún proceso o modificación del preprocesamiento. Esto podría resultar en que las variables que se tomen en cuenta para generar un resultado ya no aporten la misma información y pierdan o ganen relevancia de una manera que no se había considerado.

Estos cambios pueden afectar la forma en la que el modelo de ciencia de datos toma decisiones, lo que se puede traducir en que sus resultados sean menos precisos o comiencen a presentar resultados erróneos con el tiempo. El monitoreo es la herramienta que permite detectar estos errores de degradación del modelo.

Se debe tener siempre presente cuáles son los supuestos del modelo y actualizarlos en caso de ser necesario. Además, es necesario observar y controlar en el tiempo las métricas utilizadas en la fase de validación del modelo, como la tasa de verdaderos positivos y negativos, la tasa de falsos positivos y negativos, la precisión, la sensibilidad, la especificidad, la exactitud, la curva *ROC* y *AUC*, el *MSE* y el R^2. Hay que recordar que las métricas que se escojan para monitorear el modelo dependerán del algoritmo utilizado, de la naturaleza del problema y de las prioridades que se hayan definido en los objetivos del proyecto.

Se sugiere usar la **Ficha de Implementación** (Herramienta XII) para resguardar que ocurra el seguimiento de la solución. En caso de que se detecten problemas, esta herramienta permite identificar las medidas reparatorias y asignar un responsable.

HERRAMIENTA XII

Ficha de Implementación

Para utilizar esta herramienta sugerimos seguir los siguientes pasos:

1. La jefatura debe **convocar** al equipo del proyecto para responder colaborativamente a las preguntas de la herramienta.
2. **Responder en la tercera columna** la las preguntas planteadas en la segunda columna.
3. **Registrar las medidas reparatorias y su responsable** en la cuarta columna cada vez que no se pueda dar una respuesta satisfactoria a la pregunta planteada.

PRODUCTO DE LA HERRAMIENTA

Se acuerdan los puntos para monitorear el desempeño de la iniciativa y se delegan responsabilidades para abordar los aspectos por resolver.

DIMENSIÓN	PREGUNTAS VERIFICADORAS	RESPUESTA	MEDIDAS REPARATORIAS Y RESPONSABLES
GOBERNANZA E INDICADORES	**¿Quién está a cargo de realizar el monitoreo de la solución?**	*Quien haya cumplido el rol de jefatura de proyecto deberá monitorear la solución.*	**MEDIDA** **RESPONSABLE**
	¿Qué indicadores son importantes de monitorear?	» *Indicadores de impacto, por ejemplo tiempo de espera o satisfacción de personas usuarias.* » *Métricas de validación del modelo, según las que se definieron en el Paso 12.* » *Indicadores del funcionamiento, por ejemplo porcentaje de personas que usan la solución o intermitencia del funcionamiento del modelo.*	**MEDIDA** **RESPONSABLE**
	¿Con qué herramienta se recolectan dichos indicadores?	*Los respectivos indicadores se recolectan con las mismas herramientas que se desplegaron durante el piloto, en lugares representativos de la población objetivo.*	**MEDIDA** **RESPONSABLE**
	¿Con qué frecuencia se recolectan dichos indicadores?	*Durante los primeros seis meses se hará seguimiento de los indicadores de impacto y de validación mensualmente, luego se volverán a consultar al cumplir un año desde su implementación.* *Por otro lado, los indicadores de funcionamiento se deben monitorear periódicamente cada seis meses mientras la solución esté activa.*	**MEDIDA** **RESPONSABLE**

DIMENSIÓN	PREGUNTAS VERIFICADORAS	RESPUESTA	MEDIDAS REPARATORIAS Y RESPONSABLES
SOSTENIBILIDAD DEL MODELO	**¿Bajo qué supuestos se implementó el modelo?**	*Puesto que se implementó un modelo de derivación automática, el supuesto era que se podía contar con todas las características necesarias para derivar correctamente.*	**MEDIDA** **RESPONSABLE**
	¿Cómo se han comportado dichos supuestos con el tiempo?	*Efectivamente se han recolectado todas las características definidas como necesarias para el desarrollo del modelo.*	**MEDIDA** **RESPONSABLE**
	¿Cómo podría cambiar el comportamiento de dichos supuestos en el tiempo?	*Si se decidiera recolectar otras características, estas podrían ser más difíciles de conseguir.*	**MEDIDA** **RESPONSABLE**
SOSTENIBILIDAD DE LA INICIATIVA	**¿Qué recursos financieros, humanos y/o tecnológicos se necesitan para que la iniciativa continúe funcionando?**	*Por un lado, se necesita un presupuesto anual de 20.000.000 CLP.* *Por otro lado, se necesita mantener en funcionamiento el servicio de nube contratado por la organización para el funcionamiento del modelo*	**MEDIDA** *Planificar la aprobación anual del presupuesto para la sostenibilidad de la iniciativa año a año. Considerar otras figuras jurídicas para la estabilidad financiera del proyecto en el futuro.* **RESPONSABLE** *Unidad de Análisis Jurídico Financiero.*
	¿Existen aspectos administrativos y/o regulatorios que deban modificarse o mantenerse para asegurar la sostenibilidad de la iniciativa?	*Mientras esté vigente el convenio de colaboración que se firmó al inicio, la solución puede funcionar correctamente.*	**MEDIDA** **RESPONSABLE**

Paso 19
Robustecer el modelo

En el desarrollo del proyecto podrían quedar objetivos pendientes, que hayan sido definidos inicialmente como relevantes, pero que quedaron relegados a un segundo plano puesto que dejan de alinearse con las prioridades que se tienen o no eran factibles en el momento. Es natural que los objetivos cambien o se reformulen en el transcurso del proyecto, para acoplar perspectivas o problemas no previstos y entregar una mejor solución al problema que se está enfrentando.

Por otro lado, al recibir comentarios de parte de las funcionarias y funcionarios y de las personas usuarias sobre el funcionamiento de la solución pueden surgir nuevos objetivos para el proyecto. Estos pueden estar ligados a la interacción con las persona usuarias, la forma en cómo se presenta la herramienta o la forma en la que el modelo de ciencia de datos está diseñado y entrenado. Dependiendo de los objetivos asociados, podría ser necesario volver a hacer o reiterar etapas del proyecto, para mejorar la herramienta de acuerdo a la experiencia y expectativas de las personas usuarias.

Por lo mismo, **es crucial incorporar instrumentos de retroalimentación para que el modelo mejore, logre solucionar el problema satisfactoriamente y continúe vigente en el tiempo.** De tener una opinión positiva se debe insistir con la difusión, sin embargo el equipo debe permanecer abierto a las críticas y a los malos resultados de la solución, evaluando si son necesarios cambios radicales.

Una de las vías para recibir estos comentarios sería realizar reuniones periódicas con las personas implicadas para escuchar problemas y sugerencias y generar soluciones viables. Los desafíos de implementación que se superen y las mejoras que se implementen en el camino deberían ser incorporadas al manual de personas usuarias, para mantener una comunicación clara sobre la evolución de la solución.

¿Cómo elaboramos esta Guía de ciencia de datos?

La elaboración de la *Guía Permitido Innovar: ¿Cómo podemos desarrollar proyectos de ciencia de datos para innovar en el sector público?* comenzó con una instancia de co-creación para definir los contenidos que la compondrían. El equipo del Laboratorio de Gobierno y el GobLab UAI, se reunieron para proponer y sistematizar experiencias de diversos proyectos realizados junto a con otras instituciones públicas, generando una **primera propuesta de metodología para el desarrollo de un proyecto de ciencia de datos en el sector público.** Este insumo sirvió para poner a prueba el contenido, su estructura y visualización en tres oportunidades distintas y fundamentales para la elaboración de esta Guía, cada una con objetivos específicos.

En una primera instancia, testeamos el contenido y estructura de la Guía con el equipo de la Universidad Adolfo Ibáñez y del Laboratorio de Gobierno. Utilizamos la metodología de *Levantamiento de alertas* -disponible en la *Guía Permitido Innovar de Lenguaje Claro*- para pudiesen emitir comentarios según su experiencia en el desarrollo de proyectos de innovación pública. Gracias a este proceso se realizaron sustantivos cambios a la primera versión, los que derivaron en una versión más avanzada de la Guía.

Con estos ajustes incorporados, pusimos nuevamente a prueba los contenidos, esta vez con un grupo de expertas y expertos en el desarrollo de proyectos de ciencia de datos, quiénes nos propusieron, entre otras cosas, profundizar en ciertos conceptos críticos para la coherencia del documento y dar mayor énfasis a algunos pasos de la metodología.

Por último, realizamos talleres con miembros de la *Red de Innovadores Públicos* en los que evaluamos la capacidad de esta Guía para sintetizar y visualizar su contenido en distintas herramientas que sea útiles para un proyecto de estas características. En estas instancias pudimos determinar cuán efectivas resultaban las indicaciones para el uso y aplicación de dichas herramientas, identificando cuáles facilitaban el proceso y desarrollo del proyecto y qué teníamos que mejorar en aquellas que fueron más difíciles de utilizar.

Agradecemos a las **más de 40 personas** que formaron parte de este proceso, a miembros de la *Red de Innovadores Públicos,* a funcionarias y funcionarios de instituciones públicas e investigadores de la academia que retroalimentaron valiosamente esta Guía desde su experiencia y visión experta. Algunas de las personas que colaboraron son:

- **Adriana Herrera** del Instituto Nacional de Estadísticas
- **Alejandra Neira** de de la Municipalidad de Providencia
- **Alejandra Soto** del Fondo Nacional de Salud
- **Alfredo Muñoz** del Servicio Nacional de Mejor Niñez
- **Armando García** de la Superintendencia de Pensiones
- **Braulio Neira** de la División de Gobierno Digital
- **Carla González** de la Superintendencia de Educación
- **Claudio Reyes** de División de Gobierno Digital
- **Eliana Scheihing** de la Universidad Austral
- **Enrique Simpson** del Fondo Nacional de Salud
- **Felipe Welch** del Instituto Nacional de Propiedad Industrial
- **Felipe Yavar** del Ministerio de Vivienda y Urbanismo
- **Francisca Vidal** de la Secretaría de Modernización del Ministerio de Hacienda
- **Hugo Gutiérrez** de la Corporación de Asistencia Judicial
- **Javier Carrasco** de la Municipalidad de Vitacura
- **Jennifer Álvarez** de la Tesorería General de la República
- **Jessica Carvajal** del Fondo de Solidaridad de Inversión Social
- **Jorge Avendaño** del Ministerio de Educación
- **Juan Pablo Gajardo** del Ministerio de Bienes Nacionales
- **Klaus Lehmann** del Instituto Nacional de Estadísticas
- **Laura Salazar** del Servicio Nacional de Turismo
- **Marcela Garzón** del Fondo de Solidaridad de Inversión Social
- **María José González** de la Corporación de Fomento de la Producción
- **Michael Cortés** de la Tesorería General de la República
- **Miguel Carrasco** de la Universidad Adolfo Ibáñez
- **Nicolás Soto** del Ministerio de Salud
- **Ninoska Kroff** de la División de Gobierno Digital
- **Octavio Espinoza** del Servicio Nacional del Patrimonio Cultural
- **Pablo Aguirre** de Universidad Adolfo Ibáñez y Superintendencia de Medio Ambiente
- **Reinel Tabares** de la Universidad Adolfo Ibáñez
- **Rodolfo Bravo** del Servicio de Impuestos Internos
- **Sebastián Moreno** de la Universidad Adolfo Ibáñez
- **Sergio Brito** de la Central de Abastecimiento del Sistema Nacional de Servicios de Salud

Referencias

» **Burrell, J. (2016).** How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big data & society, 3(1), 2053951715622512.
» **Chen, J. C., Rubin, E. A., & Cornwall, G. J. (2021).** Data Science for Public Policy. Springer International Publishing.
» **Data Futures Partnership (2017).** A Path to Social License: Guidelines for Trusted Data Use.
» **Davenport, T. & Prusak, L.,(1998).** Working knowledge: how organizations manage what they know.
» **Departamento Administrativo Nacional de Estadística de Colombia (2018).** Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional.
» **Doran, G. T. (1981).** There’s a SMART way to write management’s goals and objectives. Management review, 70(11), 35-36.
» **Dykes, B. (2019).** Effective data storytelling: how to drive change with data, narrative and visuals. John Wiley & Sons.
» **Friedman, B., & Nissenbaum, H. (1996).** Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.
» **Gertler, P. J., Martínez, S., Premand, P., & Rawlings, L. B. (2017).** La evaluación de impacto en la práctica. World Bank Publications.
» **Igual & Seguí (2017).** Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications
» **INE (2021).** Guía para el control de divulgación estadística en microdatos.
» **INE (2022).** Glosario de conceptos estadísticos.
» **Laboratorio de Gobierno (2021).** Otro Ángulo: Perspectivas de innovación pública.
» **Laboratorio de Gobierno (2022).** Glosario de conceptos para responder el Índice de Innovación Pública. Levantamiento 2022.
» **Leek, J. T., & Peng, R. D. (2015).** What is the question?. Science, 347(6228), 1314-1315.
» **Peng, R. D., & Matsui, E. (2015).** The art of data science. A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC.
» **Peshawa J. Muhammad Ali, Rezhna H. Faraj (2014).** Data Normalization and Standardization: A Technical Report. Machine Learning Technical Reports, 1(1), pp 1-6.
» **Provost, F., & Fawcett, T. (2013).** Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.
» **Shmueli, G. (2010).** To explain or to predict?. Statistical science, 25(3), 289-310.
» **Sweeney, L. (2000).** Simple demographics often identify people uniquely. Health (San Francisco), 671(2000), 1-34.
» **Ulloa, W. L. O., Masacon, N. H. H., & Rodríguez, A. F. C. (2019).** La estructura organizacional en función del comportamiento del capital humano en las organizaciones. RECIMUNDO, 3(4), 114-137.
» **Unión Europea (2010).** Handbook on Statistical Disclosure Control.
» **Van Der Aalst, W. (2016).** Process mining: data science in action (Vol. 2). Heidelberg: Springer.
» **Verma, S., & Rubin, J. (2018, May).** Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.

Equipo
Laboratorio de Gobierno

Alejandra Gómez
Carlos Carrillo
Catalina Gutiérrez
Constanza Pérez
Daniela Herrera
Editha Fuentes
Eduardo Navarro
Elisa Breull
Erna Gómez
Fran Garretón
Francisca Moya
Fremberling Ramos
Giancarlo Sillerico
Ignacio Paiva
Javiera Miranda
Laura González
Lorena Torres
Myriam Meyer
Raúl Herríquez
Rodrigo Albornoz
Sebastián Altimira
Tomás Dintrans

Recuerda que con las Guías Permitido Innovar del Laboratorio de Gobierno, podrás contribuir a transformar el Estado chileno a tráves de:

1. Proyectos de Innovación Pública
2. Concursos de Innovación Abierta
3. Facilitación de espacios de Innovación
4. Lenguaje claro

¡Descárgalas!
lab.gob.cl/permitido-innovar

permitido :: innovar :: Guías para transformar el Estado chileno

Ministerio de Hacienda

Gobierno de Chile

Laboratorio de Gobierno

En colaboración con:

UNIVERSIDAD ADOLFO IBÁÑEZ