# Hoja de trabajo para definir el alcance de un proyecto de ciencia de datos

**1. Nombre del proyecto:**

**2. Nombre de la organización:**

**3. Descripción del problema:**

**3.1 ¿Cuál es el problema al que te enfrentas?**

**3.2 ¿Quién/qué se ve afectado por este problema?** (Cierto tipo de personas, organizaciones, barrios, medio ambiente)

**3.3 ¿Cuántas personas/organizaciones/localidades/etc. se ven afectadas y en qué medida?** (P. ej., tiempo medio de espera para una cirugía, número de estudiantes que abandonan los estudios, costo debido a la evasión fiscal, etc.)

**3.4 ¿Por qué resolver este problema es una prioridad para tu organización?**

**4. Objetivos (en orden de prioridad)**

* La solución técnica (p. ej., modelo predictivo) no es el objetivo.
* El objetivo debe ser medible.
* Alcanzar el objetivo debería ayudar a solucionar el problema.
* Algunos objetivos habituales son mejorar/maximizar/aumentar o disminuir/mitigar/reducir un resultado o una métrica.
* Las limitaciones típicas incluyen el presupuesto, la falta de capital humano, las restricciones legales, la voluntad política y la licencia social.
* Considera las compensaciones (*trade-offs*) entre objetivos contrapuestos.

|  |  |  |
| --- | --- | --- |
|  | **Objetivo** | **Limitaciones** |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

**5. Acciones**

* Las acciones son las actividades que las instituciones hacen/pueden hacer respecto de un problema. Implicación la asignación de recursos, por ejemplo, inspeccionar instalaciones, brindar servicios preventivos, promoción, etc.
* Las acciones podrán ser realizadas de mejor forma cuando la institución tenga la información generada en el proyecto de ciencia de datos.
* Las acciones ideales deberían ayudarte a alcanzar los objetivos definidos anteriormente.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Acción 1 | Acción 2 | Acción 3 |
| **Acción:**  *P. ej., fiscalizar el cumplimiento de cuotas de pesca en desembarques industriales y artesanales en puertos* |  |  |  |
| **¿Quién ejecuta la acción?**  *P. ej., Fiscalizador (inspector) Departamento del Fiscalizador* |  |  |  |
| **¿Sobre quién/qué se está realizando la acción?**  *P. ej., barcos pesqueros artesanales e industriales* |  |  |  |
| **¿Con qué frecuencia se toma la decisión de realizar esta acción?**  *P. ej., diariamente* |  |  |  |
| **¿Qué canales se están usando/se pueden usar para realizar esta acción?**  *P. ej., presencialmente* |  |  |  |
| **Otra información útil sobre la acción** |  |  |  |

**6. Datos**

* Los datos tienen que conectarse con las acciones que respaldan, de modo que la organización pueda alcanzar su objetivo.
* Los proyectos típicos de ciencia de datos usan datos administrativos como la fuente primaria de datos y la mejoran con fuentes de datos disponibles públicamente (censo, otros datos abiertos). La asociación con el sector privado u organizaciones sin fines de lucro podría ser una forma de obtener datos que podrían faltarte a nivel interno.

**A. ¿Qué datos tienes internamente?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Fuente de datos 1** | **Fuente de datos 2** | **Fuente de datos 3** |
| **Nombre**  *P.ej., Sistema de alta hospitalaria* |  |  |  |
| **¿Qué contiene?**  *Describir los atributos con el mayor detalle posible,  P.ej., registros de admisión y alta de hospitales a nivel nacional, con datos socio-demógráficos del paciente, diagnóstico, días de estadía, tipo de seguro de salud, información doctor* |  |  |  |
| **¿Qué nivel de granularidad?**  *P. ej., transacción, persona, organización, ubicación* |  |  |  |
| **¿Con qué frecuencia se recopila/actualiza una vez que se capta?**  *P. ej., en tiempo real, a diario, semanalmente, mensualmente, anualmente, de manera excepcional* |  |  |  |
| **¿Tiene identificadores únicos y confiables que se pueden vincular con otras fuentes de datos?**  *P. ej., RUN, SSN, DNI* |  |  |  |
| **¿Quién es el propietario interno de los datos?**  *P. ej., Departamento de registros Hospital* |  |  |  |
| **¿Cómo se almacena?**  *P. ej., en una base de datos, PDF, Excel, SPSS* |  |  |  |
| **Comentarios adicionales** |  |  |  |

**B. ¿Qué datos puedes obtener de fuentes externas, privadas o públicas?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Fuente de datos 1** | **Fuente de datos 2** | **Fuente de datos 3** |
| **Nombre**  *P.ej., Registro de Calidad del aire* |  |  |  |
| **¿Qué contiene?** *Describir los atributos con el mayor detalle posible,  P .ej., Concentración de contaminantes en el aire tales como material particulado de distintos tamaños* |  |  |  |
| **¿Qué nivel de granularidad?**  *P.ej., estación de monitoreo geolocalizada por hora* |  |  |  |
| **¿Con qué frecuencia se recopila/actualiza una vez que se capta?**  *P.ej., diaria* |  |  |  |
| **¿Tiene identificadores únicos y confiables que se pueden vincular con otras fuentes de datos?**  *P.ej. Código de la estación de monitoreo* |  |  |  |
| **¿Quién es el propietario interno de los datos?**  *Ministerio del Medio Ambiente* |  |  |  |
| **¿Cómo se almacena?**  *P.ej. base de datos descargable mediante un API en un portal de datos abiertos.* |  |  |  |
| **Comentarios adicionales** |  |  |  |

**C. En un mundo ideal, ¿existen datos adicionales que te gustaría obtener/reunir que serían relevantes para este problema?** (Encuestas, circuito cerrado de televisión, registros telefónicos, ADN, diferente frecuencia o granularidad para datos disponibles actualmente, etc.)

**7. Análisis**

* Los proyectos típicos de ciencia de datos incluyen una combinación de análisis.
* El análisis no es el objetivo del proyecto.
* Elige el análisis adecuado para el problema correcto.
* Debes probar el análisis, y el proceso de validación debe coincidir con tu objetivo.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Análisis 1:** | **Análisis 2:** | **Análisis 3:** |
| **Tipo de análisis**  *P.ej., descripción, predicción, detección, cambio de comportamiento* |  |  |  |
| **Propósito del análisis**  *P. ej., entender el comportamiento histórico de las personas, estimar el riesgo de* *enfermedad de un paciente,  identificar las acciones que lograrían disminuir la sobrepesca* |  |  |  |
| **¿Qué acciones utilizarán la información generada por  este análisis?**  *P. ej., la fiscalización de barcos pesqueros industriales y artesanales* |  |  |  |
| **¿Cómo validarás este análisis usando datos existentes?**  *P. ej., usar datos históricos, ejecutar un ensayo aleatorio controlado* |  |  |  |

**8. Consideraciones éticas**

|  |  |
| --- | --- |
| **Privacidad**  ¿Estás trabajando con datos personales y/o sensibles identificables a nivel individual? ¿Cuáles? |  |
| **Transparencia**  ¿Qué partes interesadas deberían estar al tanto del proyecto?  *Las partes interesadas suelen incluir a formuladores de políticas, trabajadores de primera línea, organizaciones de la sociedad civil, organismos públicos, personas que se verán afectadas por las acciones, etc.* |  |
| **Discriminación/equidad**  ¿Existen grupos específicos para los que deseas garantizar la equidad de los resultados?  *P. ej., grupos dado su género, edad, localización, clase social, nivel educativo, urbano-rural, etnia* |  |
| **Licencia social**  Si la población completa del país se entera de tu proyecto, ¿lo aprobará? ¿Por qué? |  |
| **Rendición de cuentas**  ¿Quiénes son las personas responsables de todos los elementos anteriores? |  |
| **Otras consideraciones, como consentimiento, legales, etc.** |  |

**9. ¿Qué prueba de campo o ensayo aleatorio controlado puedes diseñar para validar los resultados del proyecto?** ¿En qué población se va a probar? ¿Por cuánto tiempo? ¿Existe una línea base contra la cual comparar? Se debe medir los impactos del modelo en diferentes grupos (ver sección 8).

**10. ¿Cuáles son las organizaciones externas y los departamentos internos que necesitarán estar involucrados?**

Generalmente, los proyectos de ciencia de datos requieren la participación de propietarios de datos, propietarios de infraestructura de TI, propietario del problema/proceso, profesionales de analítica

|  |  |  |
| --- | --- | --- |
| **Organización/departamento** | **Descripción de la participación deseada** | **Nombre/rol de la contraparte** |
| *Departamento de TI* | *Brindar infraestructura de datos* | *Jefe del Departamento de TI* |
| *Organismo de estadísticas* | *Proporcionar datos de la población* | *Jefe del Departamento de Estadísticas* |
|  |  |  |

|  |
| --- |
| Esta hoja de trabajo fue desarrollada originalmente por el *Center for Data Science and Public Policy* de la Universidad de Chicago. Para más información sobre nuestros programas y trabajo, por favor visita http://datasciencepublicpolicy.org o escríbenos a info@datascienceforsocialgood.org  Esta versión de la hoja de trabajo ha sido actualizada a través de una colaboración entre el GobLab UAI, Carnegie Mellon University e ITAM.  El GobLab UAI es el laboratorio de innovación de la Escuela de Gobierno de la Universidad Adolfo Ibáñez. Su misión es promover el uso de la ciencia de datos en el sector público con el fin de mejorar la gestión pública y tener más políticas públicas basadas en evidencia. Capacita a funcionarios públicos y realiza investigaciones y proyectos aplicados en asociación con organismos gubernamentales. Para obtener más información, visita [**https://**](https://gobierno.uai.cl/centro-investigacion/goblab-uai/)goblab.uai.cl o envía un correo electrónico a [goblab@uai.cl](mailto:goblab@uai.cl). **Attribution ShareAlike (CC BY-SA)image** |