Informe

Relevamiento de actividades de Ciencias de Datos y posibles bancos de datos de interés.

Mayo 2020

Autoras:	Lorena Etcheverry, María Inés Fariello
Fecha de creación:	20/02/2020
Fecha de última actualización:	25/05/2020
Nombre del Proyecto:	Hoja de Ruta en CD/AA
Tipo de Documento:	Informe
Referencia / Versión:	1.1
Número de páginas:	21
Destinatarios:	SNCYT, TU, MIEM

1. Introducción

El Gabinete Ministerial de Transformación Productiva y Competitividad priorizó las actividades que conforman el conjunto inicial de Hojas de Ruta, con focos específicos ya identificados al interior de algunas de ellas. Se trata de actividades de alto potencial para la transformación productiva, en línea con el objetivo principal de impulsar la expansión de actividades innovadoras con mayor valor agregado y contenido tecnológico nacionales.

Este es el caso de la Hoja de Ruta en Ciencia de Datos y Aprendizaje Automático (CD/AA)(Sistema Nacional de Transformación Productiva y Competitividad -Transforma Uruguay, 2019), presentada en mayo de 2019 y de la cual comenzaron a implementarse proyectos durante ese año. La construcción de esta HR fue liderada por el Ministerio de Industria, Energía y Minería (MIEM) en consulta a un grupo de expertos en la materia[1]. Esta Hoja de Ruta visualiza a nuestro país como un referente para el año 2030 en la aplicación de soluciones de CD/AA en sectores estratégicos, y a empresas del sector TIC de Uruguay como sus generadoras. Para alcanzar este propósito se identificaron varias líneas de trabajo, con su objetivo, líneas de acción y proyectos

[1] Participaron enl la elaboración de la hoja de Ruta en Ciencia de Datos y Aprendizaje Automático Javier Barreiro (AGESIC); Gustavo Crespi (BID); Carlos Fournier (ANCAP); Diego Garat (FING - Udelar); Sebastián García (Idatha); Ignacio Horvath (ANCAP); Matías Jackson; Federico Lecumberry (FING -UdelaR); Leonardo Loureiro (Quanam – CUTI); Benjamín Machín (Idatha); Fabrizio Scrollini (Open Data Latin American Initiative – ILDA).

concretos, agrupadas en dos grandes dimensiones: 1) aspectos que facilitarían un entorno habilitante para desarrollos vinculados a CD/AA, y 2) oportunidades para su aplicación a sectores estratégicos nacionales.

Para generar un entorno habilitante se destaca la necesidad de mejorar la educación y formación asociada a las áreas de Ciencia de Datos y Aprendizaje Automático, impulsar la atracción de talentos a nuestro país, promover en mayor medida las capacidades de investigación e innovación, actualizar la reglamentación existente para clarificar las posibilidades de actuación e impulsar los espacios de articulación internacional que permitan posicionar a Uruguay en la discusión y agenda regional y global en torno a CD/AA. Asimismo, se identifican capacidades y oportunidades para aplicar Ciencia de Datos y Aprendizaje Automático en áreas de relevancia y dinamismo clave a nivel nacional, tanto en el sector productivo y social, como en el Estado.

En el marco del proyecto “PROMOCIÓN DE LA INSERCIÓN INTERNACIONAL DE URUGUAY EN SERVICIOS Y BIENES INTENSIVOS EN CONOCIMIENTO (SBIC)” suscrito entre BID y ROU, se encuentra el Componente 1 de la Cooperación Técnica que tiene por objetivo diseñar e implementar planes estratégicos de promoción internacional sectorial conducentes a la atracción de nuevas inversiones en sectores SBIC -particularmente en materia digital y adopción de nuevas tecnologías disruptivas, como inteligencia artificial y ciencias de los datos- y así favorecer su proceso de internacionalización. En este caso se relevarán las brechas existentes que limitan el crecimiento y la extensión de la CD/AA en los sectores productivos de Uruguay y las posibles recomendaciones de políticas, instrumentos y ecosistemas que promuevan su desarrollo.

Dado que la Ciencia de Datos y el Aprendizaje Automático basan sus desarrollos en el análisis de datos, garantizar la disponibilidad de datos en áreas de interés, de manera actualizada y velando por su calidad, constituye un aspecto de fundamental importancia. Es por esto que dentro de las acciones prioritarias y proyectos específicos de la Hoja de Ruta en CD/AA se encuentra la realización de un relevamiento de posibles bancos de datos de interés a partir de líneas prioritarias de investigación con CD/AA. El presente documento da cuenta de dicho relevamiento.

El resto de este documento se organiza de la siguiente manera. La Sección 2 el marco teórico y un conjunto de definiciones que le dan contexto al presente relevamiento. Luego, en la Sección 3 se presenta la metodología utilizada para realizar el relevamiento. A continuación, en la Sección 4 se analizan las respuestas obtenidas y por último, en la Sección 5 se esbozan conclusiones. En el Anexo A1 se detallan las respuestas de algunas preguntas.

2. Marco teórico y definiciones

¿Qué entendemos por Ciencia de Datos?

La Ciencia de Datos (CD) puede definirse como la disciplina que busca extraer conocimiento, de forma sistemática y computacionalmente eficiente, a partir de los datos de un dominio. Para ello se basa en el método científico, utilizando principalmente métodos y técnicas matemáticas, estadísticas (modelos probabilistas y estadísticos, aprendizaje estadístico) y computacionales (programación, aprendizaje automático, modelado de datos). Pese a que la mayoría de los

modelos y las técnicas utilizados en CD existían previamente, la proliferación de datos disponibles, junto con la capacidad de cómputo para procesarlos y la mejora constante en los métodos utilizados, hacen necesaria una aproximación sistemática, que combine conocimientos de diferentes ramas y construya metodologías que los articule (Rose, 2016). El estándar Cross Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000), originalmente propuesto para actividades de Data Mining (o minería de datos), suele utilizarse también para organizar las actividades de la CD. Este estándar pone a los datos en el centro del proceso, y plantea una serie de etapas que se describen a continuación. La Figura 1 presenta un esquema de dicho proceso.

Toda actividad de CD se da en el contexto de algún dominio de aplicación. Es en el marco de ese dominio en el que se plantean problemas o preguntas que se busca resolver utilizando datos en forma intensiva. El conocimiento sobre el dominio en cuestión es, por lo tanto, un componente esencial de este tipo de procesos y se encuentra además fuertemente vinculado a la capacidad de comprender los datos existentes o la necesidad de incorporar datos nuevos. En muchos casos, debido a la complejidad del dominio, es necesaria una gran participación de expertos que no son necesariamente los que realizarán el procesamiento de los datos y el desarrollo de los modelos. Por otro lado, esto lleva a que en algunos casos, los expertos en el dominio busquen complementar su formación y adquirir capacidades en CD. Luego de identificar los datos que se utilizarán, estos se preparan para ser analizados o utilizados en modelos. Esta preparación suele involucrar transformaciones y cambios en su representación, de forma de permitir que su análisis y explotación se realicen en forma computacionalmente eficiente. A continuación estos datos son utilizados en la etapa de modelado, donde se aplican algoritmos para la extracción de conocimiento a partir de los datos). Luego de la evaluación de los resultados obtenidos, estos pueden ser comunicados o bien utilizados dentro de otros proyectos o productos.

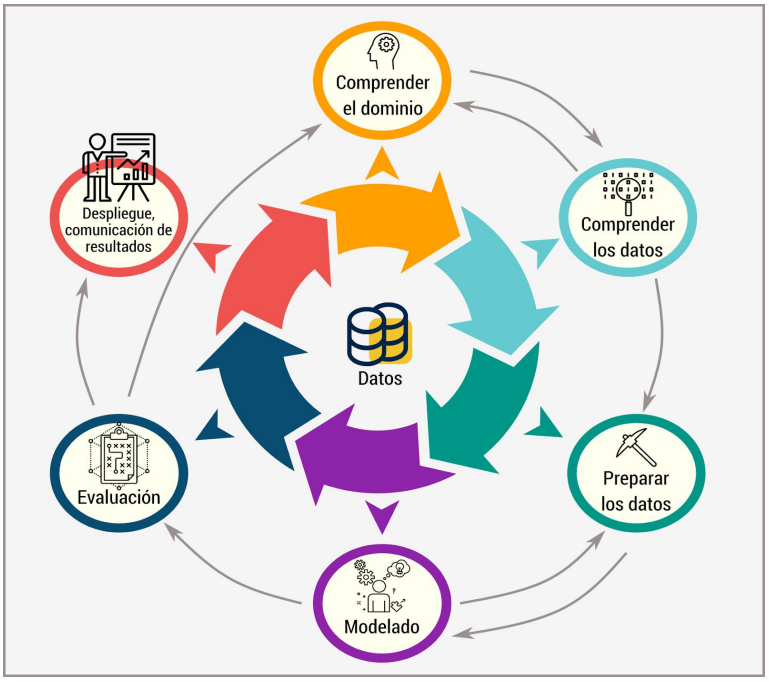

Figura 1: *Cross Industry Standard Process for Data Mining (CRISP-DM)*

Dentro de las técnicas de modelado, la subárea de la Inteligencia Artificial (IA) conocida como Aprendizaje Automático (AA) tiene una gran relevancia. Ésta área del conocimiento puede definirse como la rama que estudia los programas o agentes que aprenden, es decir que mejoran su performance en una tarea a partir de la experiencia. Esta definición incluye diferentes formas de aprendizaje (supervisado, no supervisado, por refuerzos, entre otros), así como muy diferentes modelos computacionales (redes neuronales artificiales, árboles de decisión, aprendizaje estadístico). Pese a que la Inteligencia Artificial involucra a un conjunto más amplio de disciplinas y técnicas, cuando se habla hoy de Inteligencia Artificial, prácticamente en todos los casos se está haciendo referencia al Aprendizaje Automático en una de sus variantes o modelos.

La Ciencia de Datos y el Aprendizaje Automático pueden ser aplicados en multiplicidad de contextos y actividades. En el marco de la investigación científica, hoy en día existen ejemplos de su utilización en casi todas las ramas del conocimiento. En cuanto al sector productivo y empresarial, el interés en la aplicación de este tipo de técnicas continúa en ascenso (Columbus, 2019). La Comisión Económica para América Latina y el Caribe (CEPAL) realiza un relevamiento de casos y analiza la posible interacción entre el Aprendizaje Automático y los Objetivos de Desarrollo Sustentable, identificando un amplio espectro de escenarios de uso posibles que incluyen a la telemedicina o la agricultura de precisión (CEPAL, 2018).

Sin embargo, la incorporación de CD/AA dentro de las organizaciones, los procesos productivos y los productos presenta diversos desafíos. Entre estos pueden destacarse la necesidad de la transformación organizacional, la dificultad en la formación de equipos de Ciencia de Datos, la escasez de recursos humanos formados y la dificultad de generar capacidades locales, o el hecho de que la curva de aprendizaje de estos métodos y técnicas sea bastante pronunciada (UNCTD, 2020) (Ernst & Young LLP, 2018).

Sin datos no hay Ciencia de Datos. La digitalización de las organizaciones, la gestión de datos de calidad, la disponibilización de datos y los cambios organizacionales necesarios para poner los datos en primer plano siguen siendo grandes desafíos a enfrentar para que la CD/AA tenga impacto en las organizaciones y los productos (Dhasarathy *et al.*, 2020). Los datos abiertos, y en particular los datos abiertos de gobierno, son identificados en muchos casos como un factor necesario y como impulsores de los procesos de adopción de CD/AA en las organizaciones, posibilitando el desarrollo de nuevos productos y servicios. Asimismo, la promoción de prácticas responsables de gestión de datos y el cumplimiento de los denominados principios FAIR (del inglés *Findable, Accessible, Interoperable* y *Reusable)* son identificados por la Comisión Europea como aspectos claves para generar confianza en las soluciones de IA y garantizar la reutilización de los datos (European Commission, 2020). Estos principios, donde se hace un juego de palabras entre el acrónimo y la idea de condiciones justas o adecuadas, proponen un conjunto de pautas para la disponibilización de datos que se adaptan mejor a contextos donde la apertura completa no es posible, por ejemplo para el caso de datos que puedan comprometer la privacidad de las personas (Hodson *et al.*, 2018).

3. Metodología

Se decidió realizar un relevamiento dentro de sectores clave de la investigación y la innovación nacional que podrían beneficiarse de la aplicación de técnicas de CD/AA. El objetivo general del relevamiento es el de capturar, por un lado, el nivel de conocimiento y aplicación de estas técnicas, y por otro, recopilar información sobre los conjuntos de datos que las organizaciones utilizan o podrían ser utilizados, así como el nivel de madurez de los datos y de los procesos de gestión de datos existentes. Se hizo particular énfasis en relevar cuál es la noción de ciencia de datos que las organizaciones poseen, así como en recopilar información sobre los recursos humanos que realizan este tipo de actividades, y las modalidades de vínculo de éstos con las organizaciones.

3.1. Diseño del relevamiento

El relevamiento se realizó mediante un cuestionario web, que apunta a detectar el estado de las organizaciones respecto a algunos de los aspectos que resultan desafiantes en el uso de CD/AA, en particular por un lado las capacidades y RRHH, y por otro las fuentes de datos. En particular, una de las preguntas que este relevamiento quisiera responder es si las organizaciones utilizan conjuntos de datos externos que tienen un costo de suscripción para su uso, y explorar si eso podría representar una dificultad u obstáculo para el desarrollo de este tipo de actividades en algún sector o rama de actividad.

El cuestionario tiene flujos alternativos que dependen de las respuestas a las preguntas elaboradas. La Figura 2 muestra un esquema del cuestionario. Se comienza por una auto clasificación entre las organizaciones que actualmente realizan actividades de CD/AA y aquellas en las que no. Para las que ya realizan actividades de este tipo, se incluye una serie de preguntas sobre las técnicas utilizadas y los recursos humanos que las realizan (Sección 2). Luego, se pide información acerca de los conjuntos de datos que actualmente utilizan (Sección 3). El detalle de las preguntas para este caso se presenta en la Tabla 1.

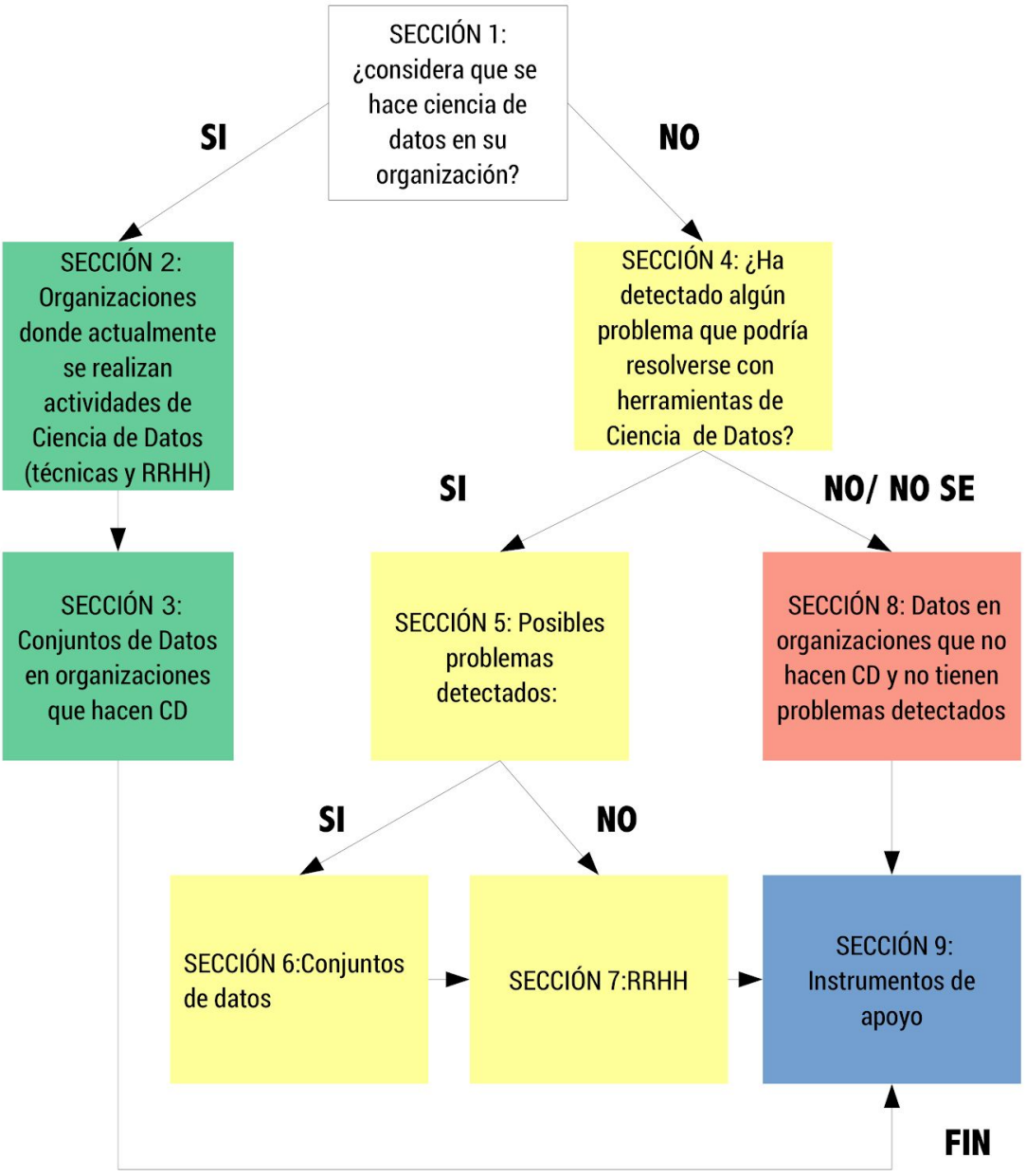

Figura 2: Esquema de las secciones del formulario diseñado y la navegación entre ellas.

Tabla 1: Preguntas realizadas a organizaciones que ya realizan actividades de CD/AA

<table>
<tr><td rowspan="7">Sección 2</td><td colspan="2">1- Describa brevemente un ejemplo de problema que esté resolviendo con Ciencia de Datos</td></tr>
<tr><td colspan="2">2 - Describa brevemente técnicas, herramientas, modelos y algoritmos que usualmente utilizan</td></tr>
<tr><td colspan="2">3- ¿Quienes realizan actividades de Ciencia de Datos en su organización?
1. personal que ya trabajaba en su organización (y eventualmente recibió formación complementaria)
2. se contrató RRHHs con la formación adecuada
3. se tercerizó la tarea, contratando el servicio de terceros para el desarrollo de la solución</td></tr>
<tr><td colspan="2">4- Describa brevemente la composición del equipo técnico que realiza ciencia de datos, indicando su formación académica</td></tr>
<tr><td colspan="2">5- ¿Tiene vínculos con alguna Universidad o equipo de investigación para realizar estas actividades? Por favor detalle su respuesta</td></tr>
<tr><td rowspan="5">Sección 3</td><td colspan="2">1- Provea información sobre los conjuntos de datos más relevantes para las actividades de Ciencia de Datos que está realizando</td></tr>
<tr><td rowspan="4">Para cada conjunto:</td><td>Descripción del conjunto de datos</td></tr>
<tr><td>El conjunto de datos ¿es generado o es responsabilidad de su organización o de un externo?
1. Por mi organización
2. Por un externo
3. Son datos internos combinados con externos</td></tr>
<tr><td>Si corresponde, indique el externo que provee estos datos</td></tr>
<tr><td>Indique todas las opciones que correspondan al conjunto de datos:
1. no está digitalizado aún
2. está digitalizado
3. está digitalizado pero sería necesario integrarlo con otros datos
4. está digitalizado pero sería necesario evaluar su calidad antes de utilizarlo
5. está disponible como datos abiertos
6. está disponible pero requiere del pago de una suscripción</td></tr>
<tr><td></td><td>En caso de que el acceso al Conjunto de datos requiera del pago de una suscripción, por favor indique su costo.</td></tr>
<tr><td colspan="2">2- ¿Ha identificado algún tipo de dificultades para aplicar Ciencia de Datos en Uruguay o para ampliar el dominio donde lo aplican?</td></tr>
</table>

A las organizaciones que indican que aún no realizan actividades de CD/AA se les pregunta si han detectado problemas que podrían resolverse con estas herramientas y en caso afirmativo se

les pide más información sobre estos problemas (Sección 5), información sobre conjuntos de datos que podrían usar, en caso de haberlos identificado (Sección 6), y detalles sobre los recursos humanos que realizarían estas actividades (Sección 7). El detalle de las preguntas para este caso se presenta en la Tabla 2.

Tabla 2: Preguntas realizadas a organizaciones que no realizan actividades de CD/AA pero que han detectado problemas en los que podrían aplicar estas técnicas

Sección 4	1- Según su criterio ¿qué tipo de problemas resuelve la ciencia de datos?
	2 -¿ha detectado algún problema que podría resolverse con herramientas de Ciencia de Datos?
Sección 5	1- De un ejemplo de problema que Ud. considere que podría resolverse aplicando Ciencia de Datos
Sección 6	1- Provea información sobre los conjuntos de datos que podrían ser relevantes para realizar actividades de Ciencia de Datos (preguntas análogas a la pregunta 6 de la Sección 3)
Sección 7	1- En caso de decidirse a utilizar Ciencia de Datos en su organización para el problema en cuestión, indique cuál de las siguientes opciones es la más probable: 1. que la tarea la realice alguien que ya trabaja en su organización (eventualmente formar a alguien) 2. contratar RRHHs con la formación adecuada 3. tercerizar la tarea, contratando el servicio de un tercero para el desarrollo de la solución
	2- ¿Tiene vínculos con alguna Universidad o equipo de investigación para asesorarse y apoyarse en estas actividades? Por favor detalle su respuesta
	3- ¿Ha identificado otras dificultades para aplicar Ciencia de Datos en su dominio?

Por último, a las organizaciones que actualmente no realizan actividades de CD/AA ni tienen identificados problemas en que podrían aplicar estas técnicas, se les consulta sobre su percepción de la ciencia de datos y sobre los conjuntos de datos que se producen en su organización (Sección 8). El detalle de las preguntas para este caso se presenta en la Tabla 3.

Tabla 3: Preguntas realizadas a organizaciones que no realizan actividades de CD/AA ni han detectado problemas donde se podrían aplicar estas técnicas.

Sección 4	1- Según su criterio ¿qué tipo de problemas resuelve la ciencia de datos?
	2 -¿ha detectado algún problema que podría resolverse con herramientas de Ciencia de Datos?
Sección 8	1- Dentro de su su organización ¿se generan datos?
	2- Si se generan datos, descríbalos brevemente e indique si están digitalizados o no.

En todos los casos, al final del cuestionario se presenta una serie de preguntas sobre los instrumentos de apoyo existentes en Uruguay para este tipo de actividades (Sección 9) que se detallan en la Tabla 4.

Tabla 4: Preguntas sobre instrumentos de apoyo realizadas en todos los casos

Sección 9	1- ¿Está al tanto de instrumentos de apoyo a la incorporación de Ciencia de Datos, por ejemplo de ANII?
	2- Si está al tanto, por favor indique algunos de estos apoyos.
	3 - ¿Quisiera recibir información complementaria sobre instrumentos de apoyo?

3.2. Alcance del relevamiento

En este relevamiento se convocaron actores del ámbito público y privado vinculados con la investigación y la innovación en diferentes sectores. Por un lado, se convocó a las instituciones y organizaciones participantes de la gobernanza de la Hoja de Ruta en CD/AA, y se optó por complementar con instituciones del sector Agropecuario (con particular énfasis en la producción de alimentos), del sector Salud (considerando tanto al sector Farmacéutico como a lo vinculado con las historias clínicas electrónicas y la medicina de precisión), del sector Energético y del sector Industrias Creativas.

En el caso de las Tecnologías de la Información y la Comunicación (TIC), la distribución del cuestionario a los actores se delegó en la cámara respectiva (en este caso la CUTI, quien participa de la gobernanza de la Hoja de Ruta). Esto también se realizó en el caso de las industrias creativas, comunicándose con asociaciones y cámaras.

La distribución del mismo se realizó por correo electrónico a través de la Secretaría de Ciencia y Tecnología en Febrero 2020. El formulario fue enviado, directamente por este medio, a más de 120 organizaciones, incluyendo la CUTI que se encargó de distribuirlo entre sus socios. Se informó a los participantes de que contaban con un mes de plazo para enviar sus contribuciones. Luego de transcurrido este plazo se obtuvieron 29 respuestas, que corresponden a 20 organizaciones. La Tabla 5 presenta la cantidad de respuestas por organización y sector.

Tabla 5: Cantidad de respuestas por organización y sector

Sector	**Organización**	**#Respuestas**
Agro	Instituto Nacional de Investigación Agropecuaria (INIA)	3
Educación/Investigación	Instituto de Investigaciones Biológicas Clemente Estable (IIBCE)	1
	Universidad Católica del Uruguay	1
	Universidad CLAEH	1

	Universidad de la República	7
	INEFOP	2
Gobernanza	Uruguay XXI	1
	Agesic (Salud.uy)	1
	Asociación Española	1
	Genia	1
Salud	Urufarma	1
	Adagio Consultores	1
	Concepto	1
	eagerWorks	1
	Globant	1
	IDATHA	1
	Ideasoft	1
	Kreilabs	1
	Rootstrap	1
TIC	Tryolabs	1
Total		**29**

4. Análisis de las respuestas

De las 29 respuestas recibidas, el 76% (22) indican que en su organización ya se realizan actividades de CD/AA, el 21% (6) indican que actualmente no realizan actividades de este tipo pero tienen problemas identificados, mientras que sólo en una respuesta (3%) se indicó que no se tienen problemas identificados, pese a que se indica que se genera una gran cantidad de datos.

Dentro de las respuestas que indican que actualmente realizan actividades de CD/AA encontramos organizaciones que corresponden a todos los sectores que respondieron el cuestionario: Agro, Educación/Investigación, Gobernanza, Salud, TIC. A continuación analizamos las respuestas obtenidas de acuerdo a cuatro ejes: los recursos humanos, los datos, las técnicas y problemas estudiados, y las dificultades reportadas.

4.1. Sobre los recursos humanos

Respecto a la integración de los equipos de CD/AA, las respuestas también son variadas, tanto en la formación de los recursos humanos como en la cantidad de personas que se dedican a este tipo de actividades en las organizaciones. Por un lado, más de la mitad de las respuestas da

cuenta de que en estas actividades participan RRHH con nivel de maestría o doctorado. Pero cuando se analiza la distribución por ámbito (público o privado) y por sector (ver Tabla 6), se observa que los RRHH con nivel de doctorado se concentran en las organizaciones del sector Educación/Investigación. En particular, parecería que ninguna de las empresas del sector TIC que completó el formulario cuenta con doctores en sus equipos de CD/AA. Esto coincide con lo relevado en el censo realizado a doctores uruguayos, donde se reporta que cerca del 80% de los encuestados trabaja en centros de investigación o universidades (Méndez *et al.*, 2019) y con lo relevado a partir de los currículos ingresados en el sistema CVUy, donde además se indica que apenas el 0,02% de los que tienen nivel de doctorado trabajan en empresas públicas, y el 1,4% en empresas privadas (*Ocho de cada diez doctorados trabajan en el sector académico, «poquitos» en el gobierno y marginales en empresas productivas*, 2018). Esta realidad, que trasciende al tema en cuestión dado que es transversal respecto a todas las áreas del conocimiento, podría tener consecuencias en la capacidad del sector productivo de incorporar y desarrollar soluciones innovadoras y originales en el área de CD/AA. Realizar alianzas con equipos de investigación locales podría ser una estrategia para compensar esta situación.

Tabla 6: Máximo nivel educativo de los integrantes de los equipos de CD/AA (distribución por sector de actividad y ámbito)

		Máximo nivel de formación alcanzado				
Ámbito	**Sector**	**Grado**	**Maestría**	**Doctorado**	**N/A**	
Privado	Educación/Investigación			1		**1**
	SALUD	1				**1**
	TIC	4	2		3	**9**
Público	AGRO			2		**2**
	Educación/Investigación	1		7		**8**
	Gobernanza	1				**1**
Total		**7**	**2**	**10**	**3**	**22**

Sobre el vínculo de los equipos de CD/AA con la organización parece interesante comparar las respuestas de las organizaciones que ya realizan actividades (Figura 3) de las que aún no (Figura 4). Mientras que en las primeras predominan los equipos formados por personas que trabajan en la organización, en las segundas se manifiesta una preferencia por contratar a terceros para este tipo de actividades. Si bien la tercerización parece ser una estrategia razonable para realizar pilotos o primeras aproximaciones a la CD/AA, el éxito de la incorporación de este tipo de actividades como parte de la dinámica de la organización en el largo plazo suele estar sustentado en la formación de equipos de trabajo. Se plantea en este escenario un dilema entre el desarrollo de soluciones locales o *in-house*, la compra de productos existentes, o la contratación de servicios externos para la realización de actividades de CD/AA. La decisión de cuál es la mejor estrategia para cada organización y cada momento depende de una serie de factores, entre los que aparecen el tamaño de la organización, la especificidad del problema, el tiempo y presupuesto disponibles, los recursos humanos, y la madurez analítica de la organización (Krensky y Linden, 2016; AAnalytics, 2017)

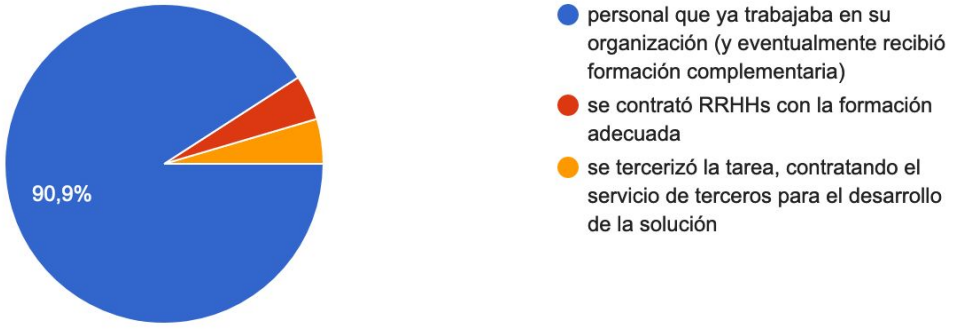

Figura 3: Tipo de vínculo de los equipos de CD/AA (en organizaciones donde se realiza CD/AA)

Figura 4: Tipo de vínculo de los equipos de CD/AA (en organizaciones donde no se realiza CD/AA)

4.2. Sobre los datos

Para hacer foco sobre los conjuntos de datos utilizados, y los costos asociados, se diseñó un conjunto de preguntas tanto para a las organizaciones que ya realizan actividades de CD/AA como para las que aún no realizan. Los detalles de estas preguntas se presentan en la Sección 3 de la Tabla 1. En cada cuestionario se podía brindar detalles de dos conjuntos de datos relevantes. De las 22 respuestas obtenidas de organizaciones que ya realizan actividades, sólo ocho contienen información de dos conjuntos de datos, el resto sólo reporta información de un conjunto. En el caso de las organizaciones que aún no realizan actividades de CD/AA se reporta un total de cinco conjuntos de datos. En total se releva información de 33 conjuntos de datos. La Figura 5 presenta la distribución de los conjuntos según quien genera o es responsable de los datos. Cabe aclarar que las empresas del sector TIC que participan del relevamiento realizan actividades de CD/AA para terceros, y por lo tanto los datos son externos a estas organizaciones

pese a que son datos del cliente y por lo tanto podrían pensarse como datos que son responsabilidad de la entidad que contrata el análisis.

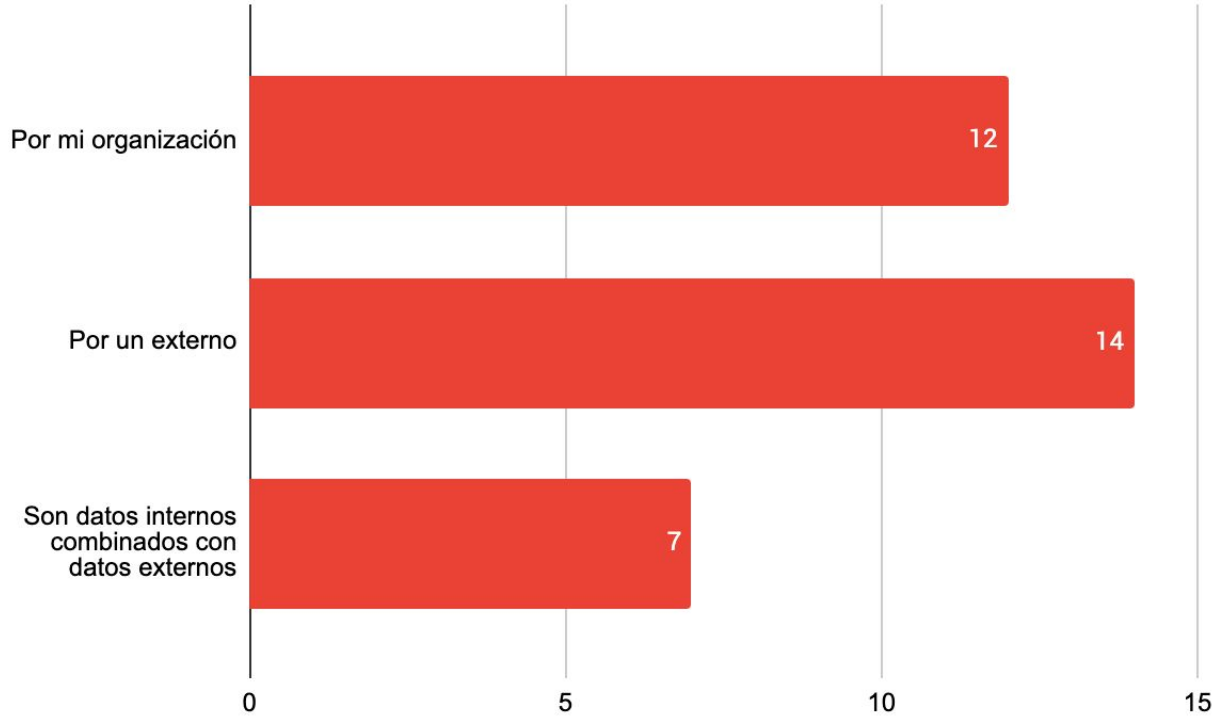

Figura 5: Distribución de conjuntos de datos por origen

Otro aspecto que interesa relevar es el nivel de madurez de estos conjuntos de datos, en el sentido de si ya están digitalizados, si tienen buena calidad, si es preciso integrarlos con otros datos para realizar análisis, y por otro lado la disponibilidad de los mismos. Ésto último es una de las motivaciones para la realización de este relevamiento, donde se buscar recabar información acerca de conjuntos de datos que pudieran ser relevantes y requieren pago de una suscripción. La Figura 6 presenta los resultados para estas características de madurez y disponibilidad, donde sólo el 9% (3) de los conjuntos de datos mencionados en el relevamiento está disponible como datos abiertos, y sólo el 3% (1) de ellos requiere suscripción. El caso reportado de datos por suscripción tiene un costo anual de aproximadamente USD 10.000 y corresponde al acceso empresarial de los servicios ofrecidos por la empresa Crunchbase[2]. Esta empresa recopila y ofrece datos de empresas e inversiones, que pueden ser utilizados por equipos de ventas, investigadores de mercado y equipos de innovación para descubrir potenciales clientes y desarrollar nuevos productos.

[2] Crunchbase (https://www.crunchbase.com/)

Parece relevante destacar que en 54% (18) de los casos es necesario realizar tareas de gestión de datos, como la integración de los datos a analizar con otros conjuntos de datos o la evaluación de la calidad de los mismos y eventualmente la realización de actividades de limpieza. En este sentido el relevamiento da cuenta de un fenómeno global. Es frecuente encontrar referencias en la literatura al costo de las tareas de preparación de datos, las cuales en muchos casos representan cerca del 80% del tiempo dedicado a los proyectos de CD/AA (Kandel *et al.*, 2012) .

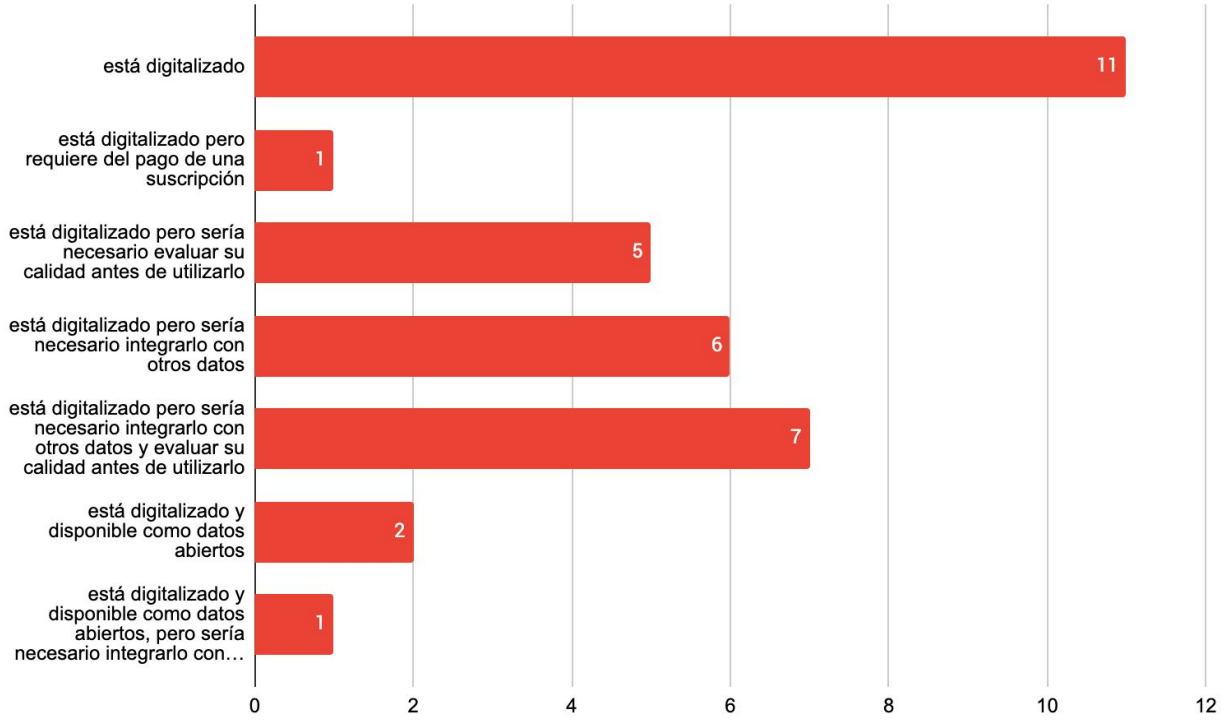

Figura 6: Madurez y disponibilidad de los conjuntos de datos

4.3. Sobre las técnicas y problemas abordados

Las preguntas 1 y 2 de la Sección 2 del cuestionario (ver Tabla 1) apuntan a obtener información sobre los problemas y las técnicas aplicadas, las cuales son sumamente variadas. En particular, respecto a las técnicas y herramientas, se observa cierta prevalencia de la aplicación de técnicas clásicas de Aprendizaje Automático. La Tabla 7 del Anexo detalla todas las respuestas obtenidas a la pregunta 2 (Describa brevemente técnicas, herramientas, modelos y algoritmos que usualmente utilizan).

De las respuestas obtenidas no es posible determinar si las organizaciones han desarrollado herramientas o productos que utilicen CD/AA (por ejemplo, sistemas recomendadores para algún problema en particular) o si lo que hacen es realizar la aplicación de técnicas a problemas específicos, eventualmente aplicando las mismas técnicas a problemas similares en diferentes

contextos o a diferentes clientes. Sería interesante realizar un estudio específico sobre estos puntos.

4.4. Sobre las dificultades de aplicar CD/AA

Tanto en la Sección 2 como en la Sección 3 se incluyeron preguntas sobre las dificultades encontradas. Las respuestas obtenidas pueden organizarse en cuatro ejes: 1) dificultades vinculadas a los datos (disponibilidad, calidad, metadatos adecuados, estructura de los datos, etc) que se mencionan en un 40% de las respuestas, 2) dificultades vinculadas a los Recursos Humanos (escasez de RRHH formados en algunos temas, dificultad de compartir experiencias con otros actores) que aparecen en un 40% de las respuestas, 3) dificultades relacionadas con las contrapartes en los proyectos de CD/AA, que aparecen en un 30% de las respuestas y que pueden resumirse en falta de experiencia por parte de los contratantes en lo que implica un proyecto de CD/AA, y por último y en un solo caso, 4) dificultades en encontrar apoyos económicos para la contratación de RRHH. La Tabla 8 del Anexo detalla todas las respuestas obtenidas.

5. Conclusiones

Pese a que podría ser un poco aventurado extraer conclusiones de una muestra tan pequeña de respuestas, creemos que algunos aspectos llaman la atención. Los recursos humanos altamente calificados parecen ser fundamentales para la aplicación exitosa de técnicas de CD/AA, ya sea en ámbitos de investigación como en proyectos de innovación y desarrollo. Llama la atención, aunque no es novedoso para la realidad de Uruguay, la constatación de la escasa participación de postgraduados y en particular doctores en iniciativas privadas, sobre todo en empresas de TIC que realizan proyectos de CD/AA. Tal como ya fue mencionado, esto podría representar una desventaja respecto a la realidad de otros países, donde los equipos de I+D+I del sector productivo contratan activamente RRHH altamente capacitados para el desarrollo de productos y soluciones. Si bien en Uruguay, y en particular a través de ANII, se han realizado esfuerzos para incentivar la contratación recursos humanos altamente calificados en las empresas, éstos han tenido escasa repercusión a juicio de las autoridades (*Ocho de cada diez doctorados trabajan en el sector académico, «poquitos» en el gobierno y marginales en empresas productivas*, 2018)

Por otro lado, y respecto a los datos, dentro de los casos relevados por este estudio el uso de datos que requieren el pago de una suscripción es muy marginal. Tampoco se manifiesta la necesidad de acceder a este tipo de datos, o se menciona el pago de suscripciones como una dificultad u obstáculo para los proyectos. Esto parece descartar, por el momento, la necesidad de invertir fondos públicos en suscripciones a datos.

Complementando esta observación, llama la atención el escaso uso que se hace de datos abiertos. Casi todas las iniciativas relevadas utilizan datos propios o de clientes, que no se encuentran publicados como datos abiertos. Aparentemente tampoco integran estos datos con

datos abiertos existentes, aunque en algunas respuestas la dificultad de conseguir datos abiertos ha sido mencionada como una dificultad. A partir del relevamiento tampoco es posible detectar conjuntos de datos que sería deseable abrir. En particular, en el caso de datos abiertos de gobierno, Uruguay tiene una larga trayectoria en promover la apertura de datos a través del Catálogo de Datos Abiertos[3], y parece llamativo que ninguna de las organizaciones que completó el relevamiento menciona el uso de datos del catálogo. En este sentido, quizás sería interesante seleccionar algunos sectores o casos de uso claves, y realizar actividades piloto focalizadas que permitan detectar conjuntos de datos y realizar las actividades preparatorias necesarias para su uso en proyectos de CD/AA.

También existe una gran oportunidad en cuanto a los datos generados en el contexto de actividades científicas y proyectos de investigación que desarrolla la comunidad local. Sin dudas existen posibilidades de generar valor a partir de la colaboración entre proyectos, y de la combinación y reuso de datos existentes en nuevos proyectos utilizando técnicas de CD/AA. Si bien muchos de estos datos están digitalizados y disponibles, dado que la mayoría de las revistas científicas solicitan el depósito de los datos utilizados en repositorios, no hay herramientas que asistan en la búsqueda y acceso a datos que se encuentran distribuidos. En este sentido, hoy en Uruguay existen iniciativas como el proyecto SILO[4], que configura un sistema nacional de repositorios de acceso abierto de ciencia y tecnología con el objetivo de promover el acceso y la visibilidad de la producción científica y tecnológica del Uruguay. Sin embargo, por el momento estos repositorios no abarcan la publicación de datos y metadatos científicos.

Referencias

AAnalytics (2017) «The Data Science Procurement Dilemma — Build, Buy, Or Outsource?» Medium. Disponible en: https://medium.com/@suomynonascitylana/the-data-science-procurement-dilemma-build-buy-or-outsource-b82259284eec (Accedido: 23 de mayo de 2020).

CEPAL (2018) «Inteligencia artificial para el desarrollo», en *Datos, algoritmos y políticas: la redefinición del mundo digital*. Publicación de las Naciones Unidas ((LC/CMSI.6/4)), pp. 167-184.

Columbus, L. (2019) «State Of AI And Machine Learning In 2019», *Forbes Magazine*, 8 septiembre. Disponible en: https://www.forbes.com/sites/louiscolumbus/2019/09/08/state-of-ai-and-machine-learning-in-2019/.

Dhasarathy, A. *et al.* (2020) *Accelerating AI impact by taming the data beast*. Disponible en: https://www.mckinsey.com/industries/public-sector/our-insights/accelerating-ai-impact-by-taming-t

[3] Catálogo de datos Abiertos, Uruguay https://catalogodatos.gub.uy/
[4] Proyecto SILO, https://silo.uy/vufind/

he-data-beast (Accedido: 23 de mayo de 2020).

Ernst & Young LLP (2018) «Artificial Intelligence in Europe. Spain. Outlook for 2019 and Beyond. How 277 major European companies benefit from AI». Disponible en: https://pulse.microsoft.com/es-es/business-leadership-es-es/na/fa1-situacion-de-la-adopcion-de-la-inteligencia-artificial-en-las-empresas-espanolas/.

European Commission (2020) «White Paper on Artificial Intelligence: a European approach to excellence and trust». Disponible en: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Hodson, S. *et al.* (2018) «Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data», *European Union: Brussels, Belgium*.

Kandel, S. *et al.* (2012) «Enterprise Data Analysis and Visualization: An Interview Study», *IEEE transactions on visualization and computer graphics*, 18(12), pp. 2917-2926.

Krensky, P. y Linden, A. (2016) *Machine-Learning and Data Science Solutions: Build, Buy or Outsource?*, *Gartner*. Disponible en: https://www.gartner.com/en/documents/3531217 (Accedido: 23 de mayo de 2020).

Méndez, L. *et al.* (2019) «Primer censo de personas uruguayas e inmigrantes con título de doctorado: informe de resultados», *Documento de Trabajo. FCS-UM. PP; 3*. Udelar. FCS-UM. Disponible en: https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/22319/1/DT%20UM-PP%2003.pdf.

Ocho de cada diez doctorados trabajan en el sector académico, «poquitos» en el gobierno y marginales en empresas productivas (2018) *Búsqueda*. Disponible en: https://www.sb.uy/nota/ocho-de-cada-diez-doctorados-trabajan-en-el-sector-academico-poquitos-en-el-gobierno-y (Accedido: 23 de mayo de 2020).

Rose, D. (2016) *Data Science: Create Teams That Ask the Right Questions and Deliver Real Value*. Apress, Berkeley, CA.

Shearer, C. (2000) «The CRISP-DM model: the new blueprint for data mining», *International Journal of Data Warehousing and Mining*, 5(4), pp. 13-22.

Sistema Nacional de Transformación Productiva y Competitividad -Transforma Uruguay (2019) «Hoja de ruta en Ciencia de Datos y Aprendizaje Automático». Sistema Nacional de Transformación Productiva y Competitividad -Transforma Uruguay. Disponible en: https://www.transformauruguay.gub.uy/es/documentos/tic.pdf.

UNCTD (2020) «Creación y captura de valor en la economía digital: Una perspectiva global», *Informe sobre la Economía Digital 2019*, pp. 89-113. doi: 10.18356/6df3198d-es.

ANEXO

Tabla 7: Respuestas obtenidas a la pregunta "Describa brevemente técnicas, herramientas, modelos y algoritmos que usualmente utilizan"

Mi área de investigación (el Procesamiento de Lenguaje Natural) es, junto con el reconocimiento de imágenes, la que más utiliza todos los modelos de aprendizaje automático, así que la lista es muy grande como para enumerar. Hoy en días, los modelos basados en redes neuronales tienen mucho peso.
Árboles de decisión, redes neuronales, análisis de sentimientos
Para ese caso particular, algoritmos de clustering, de selección de variables, de estadística en grafos.
machine learning en toda su diversidad, así como procesos de captura masiva de datos y análisis estadísticos a partir de software específicos como R, Stata, SPSS y desarrollos particulares y Phython, R y otros lenguajes
Para el manejo de los datos y construcción del DataWarehouse se utiliza una arquitectura de datos basada en PostgreSQL como motor de bases de datos relacional. Para el manejo de la infraestructura de integración de datos se utiliza una arquitectura basada en la plataforma de gestión y ejecución de procesos ETL de Pentaho Data Integration (también conocido como Kettle). Para manejo de la capa de visualización y consulta de los datos (considerando los datos operacionales, transaccionales y analíticos) se utiliza: CTools (HTML5/CSS/Javascript) sobre la plataforma de Pentaho Business Analytics para los paneles y cuadro de mando integral (CMI), Pentaho Report Designer para el módulo de reportes, Saiku y Mondrian para el armado y visualización de cubos.
Modelos de supervivencia, análisis gráficos varios, cluster análisis.
Para la imputación de las series diarias se probaron varios modelos basados en fundamentalmente dos familias: modelos aditivos no paramétricos y modelos lineales dinámicos Bayesianos. Se preparan algunas visualización estadística de las olas de extremos combinando gráficos estáticos de series temporales con histogramas bidimensionales. El reporte final y sus resultados se presenta dentro de una aplicación web que permite gráficos interactivos. Todo el procesamiento de datos, los modelos estadísticos, la visualización de datos y aplicación web es realizado con R (paquetes más relevantes: dlm, gamlss, ggplot2, shiny).
Clustering
Bases de datos genómicos, alineamiento de lecturas de secuenciacion masiva, data mining diversa, programación, compresión de datos, cuantificación de expresión genica.), machine learning, transcriptomica, traductomica, metagenomica. Análisis de imágenes

diversas (fuerza atómica, confocal, fluorescencia etc), reconstrucción tridimensionales. Análisis de datos derivados de experimentos biofísicos y neurociencias.
Redes Neuronales, NLP, Regresión Logística, Árboles de decisión.
Tenemos un amplio espectro de herramientas y modelos en nuestra práctica de Data Science. ML tradicional (optimización, clasificación, clustering, etc) Computer vision NLP/NLU Modelos recurrentes, series de tiempo y variables transaccionales Motores de recomendación
Machine Learning (clustering, clasificación, regresión, son muchos los algoritmos y dependiendo de la solución buscada se utilizan unos u otros), estadística, Business Intelligence
Clusterización, K-Means, Regresiones logísticas, redes neuronales.
Python, Tensorflow, Keras
Redes neuronales, principalmente convolucionales y recurrentes. También hemos usado XGBoost
Usamos de todo, según el caso. Trabajamos bastante con computer vision (CV), natural language processing (NLP) y predictive analytics. Usamos tanto modelos basados en deep learning (DL), como modelos más clásicos (XGBoost, SVM, etc.) y también CV clásico. Trabajamos con el stack Python (scikit learn, Pandas, PyTorch/TensorFlow, etc.).
La mayoría de trabajo fue realizada con R, para reestructurar y limpiar datos usamos herramientas del tidyverse, realizamos visualización estadística usando ggplot2, utilizamos una serie de algoritmos de machine learning implementados en H2O y para la aplicación web usamos shiny
Algoritmos y BD distribuídos, clustering y técnicas de aprendizaje no supervisado, automatización de despliegue de los modelos y APIs, GUIs de monitoreo en tiempo real.
K-means, Random Forest, Elbow method, Optimización bayesiana, Suavizado por Fourier.
Trabajos con fast.ai, tensorflow, sklearn, azure para diferentes abordajes sobre problemas de inteligencia artificial. También trabajamos con Power BI para consultoría sobre datos, Open Refine y Pandas para features engineering

Tabla 8: Respuestas a la pregunta "¿Ha identificado algún tipo de dificultades para aplicar Ciencia de Datos en Uruguay o para ampliar el dominio donde lo aplican?"

Sigue siendo difícil conseguir datos abiertos en algunas instituciones.
Si, acceso a los datos, capacidad de las organizaciones de extraer los datos necesarios. Objetivos concretos.

Es necesario tener una persona especializada en la institución exclusivamente para esta tarea y esto no es evidente para los que toman las decisiones. Para la creación y gestión de datos internos falta cultura organizacional que comprenda la importancia del correcto manejo de los datos.
Poca comunicación entre los equipos que realizan CdeD
Apoyo a la contratación de recursos humanos y adquisición de materiales. Formación integrada de recursos humanos entre disciplinas diferentes (ej, Biología e Informática o Matemática o Estadística).
Calidad de la información, Capacidad de procesamiento.
Hay poca gente que realmente conozca Predicción y Prescripción, y cuando uno la encuentra falta experiencia en el mercado más allá de lo académico.
No. Es cuestión de formar a las personas correctas.
encontrar personal calificado
En base a nuestra experiencia y que arrancamos hace unos 10 años, las dificultades que encontramos que están relacionadas a aplicar Ciencia de Datos desde Uruguay no son técnicas. A veces el no estar cara a cara con los clientes dificulta o alarga los ciclos de venta, porque muchas veces los clientes no entienden ciencia de datos. Para esto, viajamos frecuentemente.
pocos recursos humanos con formación específica para procesar datos genómicos, bioinformáticos.
Disponibilidad de información, problemas en la estructura de los datos, no estandarización de la forma de recolección de los datos que hace difícil la consistencia año a año, muchas veces no se piensa para que se recolectan los datos o el posible uso de los mismos se guarda información sin contar con metadatos que permitan saber que son, con estructuras inapropiadas sin definición de variables claras codificaciones que varían entre otros.
Las que existen en otros mercados (justificación de proyectos iniciales en organizaciones que tienen expectativas pero no han iniciado un trabajo en esta área, apreciación del valor, organizaciones inmaduras en trabajo gobierno de datos, etc.). No es una dificultad pero si una característica, muchos casos tienden a tener volúmenes de datos significativamente menores que en otros mercados, facilita algunos aspectos pero también impacta en las técnicas y enfoques.
Encontrar personas con quien intercambiar experiencias.
La falta de calidad de datos y conocimiento de sus implicancias
Puede ser la Evaluación de Impacto en referencia a la Competitividad de las Empresas, en función de los apoyos públicos recibidos.
La obtención de una buena calidad de datos de los prestadores de salud es la dificultad más grande que visualizamos hasta el momento.
La generación, almacenamiento y sistematización de los datos.