Ciencia de datos y Agilidad: una revisión de la literatura

Paula B. Lima [1]; Gabriel E. Giana[1]; Verónica Andrea Bollati [2]; Liliana Cuenca Pletsch[1]
[1] CInApTIC, UTN Facultad Regional Resistencia, Chaco, Argentina
[2] CONICET - CInApTIC, UTN Facultad Regional Resistencia, Chaco, Argentina
{limapaulabelen, gabrielgiana77, vbollati,
liliana.cuencap}@gmail.com

Abstract. La creciente producción y recolección de los datos involucrados en proyectos de Data Science generan la necesidad de un marco de trabajo que permita el procesamiento eficiente de los mismos. A pesar de los progresos logrados en esta temática, aún se desconoce una forma de trabajo que se ajuste a estos proyectos y permita obtener mayores beneficios. Frecuentemente, se opta por la utilización de prácticas ágiles, pero los equipos se ven obligados a realizar adaptaciones sobre éstas, debido a las diferencias existentes entre los proyectos de desarrollo de software y los proyectos de Data Science. En este trabajo presentamos una revisión sistemática de la literatura con el fin de obtener el estado del arte de los enfoques de trabajo utilizados en proyectos de Data Science, sobre todo aquellos que utilizan prácticas ágiles (ya sea total o parcial), y los roles que los integrantes de los equipos desempeñan a fin de determinar puntos en común de trabajo o problemas que se evidencian en sus funciones.

Keywords: Data Science, Agilidad, Framework, Técnicas.

1 Introducción

La ciencia de datos (DS, por su denominación en inglés *Data Science*) es un concepto que no sólo sintetiza y unifica el campo de la estadística, análisis de datos y sus métodos relacionados, sino también busca comprender los resultados obtenidos [1]. Surge por la gran cantidad de datos generados sobre las personas que se encuentran disponibles en distintos medios de almacenamiento [2] y por la heterogeneidad de los elementos bajo estudio (texto, imágenes o videos) [3]. Esto genera la necesidad de incorporar herramientas de las ciencias de la computación y otras disciplinas, que permitan la integración de dos o más tipos de datos distintos [4].

A partir de la necesidad de hacer un uso eficiente y eficaz de los datos [2], se implementan herramientas de Ingeniería de Software (IS), como metodologías de desarrollo, a proyectos de esta rama. Debido a la inexactitud y al retardo de la creación de valor en las metodologías tradicionales (ej. la metodología en cascada), el enfoque predominante en proyectos de DS es el propuesto por las prácticas ágiles [5], aunque, generalmente, no son utilizadas en su totalidad, puesto que los proyectos de esta índole son muy variados, por lo que se ajustan o reemplazan algunas de sus tareas.

Entre las principales dificultades en DS se pueden mencionar, la definición de un framework de trabajo [5], los roles de los integrantes de los equipos [6] y la inexactitud

de las formas de trabajo existentes de IS para este ámbito [5]. En este sentido, se considera que la utilización de prácticas ágiles, adaptadas a los proyectos de DS, pueden constituir una solución a estos problemas.

En este artículo, se presenta el análisis del estado del arte, por medio de un proceso de revisión sistemática de literatura, sobre el uso de prácticas ágiles en proyectos de DS.

El artículo se estructura de la siguiente manera: primero, se detalla el método utilizado para realizar el análisis del estado del arte (sección 2). Seguidamente, se exponen los resultados obtenidos a partir de su aplicación (sección 3). Luego, se presenta la discusión donde se da respuesta a las preguntas de investigación planteadas (sección 4). Finalmente, se presentan las conclusiones del trabajo.

2 Método

El método utilizado para la realización de este estudio es el de Revisión Sistemática de Literatura (RSL), propuesto por Kitchenham y Charters [7] y, en particular la versión propuesta por Biolchini et al. [8], cuyas fases principales son: planificación, ejecución y análisis de resultados. Paralelo a todo el proceso, se almacena la información obtenida en cada etapa, y se incorporan puntos de control para validar cada etapa. En esta sección se presentan los aspectos propuestos para la etapa de planificación.

2.1 Preguntas de investigación

En primer lugar se define el objetivo principal de la RSL, el cual es: Identificar y analizar el estado del arte de los enfoques de trabajo, procesos y equipos de DS en el contexto de la IS. A partir del mismo, se plantean las preguntas de investigación (PI):

- PI1: ¿Cuál es el enfoque de trabajo en proyectos de DS? Debido a las particularidades de los proyectos de DS (por ej., el valor entregado al cliente se relaciona con la información obtenida y no con un *Minimun Viable Product (MVP)*) y su consecuente diferencia con el desarrollo de software, interesa conocer cuál es la manera de trabajar, es decir, conocer el modo en el que se combinan las actividades científicas con las de desarrollo, identificando los diferentes frameworks que se utilizan en proyectos de DS y evaluando la efectividad de los mismos.
- PI2: ¿Cuáles son los perfiles involucrados en proyectos de DS? Debido a la naturaleza científica de los proyectos de DS participan roles propios de ciencias como estadísticas y matemáticas. Tras la necesidad de incorporar conceptos de IS para la gestión eficiente de los proyectos es importante determinar los perfiles presentes en los equipos de desarrollo, sus responsabilidades y limitaciones.
- PI3: ¿Los enfoques utilizados se aplican de manera pura o tienen adaptaciones? Es interesante conocer la forma en la que se aplican los frameworks en los proyectos de DS, para identificar las actividades que presentan mejores resultados y generan mayor beneficio, como así también identificar los problemas originados por aspectos o enfoques que no pueden aplicarse a los proyectos de DS.

2.2 Fuentes de datos y consulta

La etapa de planificación también involucra enumerar las fuentes de datos que se utilizarán para buscar estudios o trabajos previamente realizados y la definición de las consultas que se ejecutarán en dichas fuentes [7, 8].

Puesto que los frameworks aplicados a proyectos de DS son variados y en constante actualización, la búsqueda de estudios no se limitó a librerías digitales sino también se extendió a blogs técnicos. Esto no sólo permite obtener experiencias compartidas por profesionales que se desempeñan en el campo, sino también dificultades con las técnicas, a partir de vivencias de los desarrolladores. Podemos destacar las siguientes fuentes consultadas (Nombre[Acrónimo][Tipo de fuente]: website):

- IEEEXplore [IEEEX] [Librería digital]: http://ieeexplore.ieee.org/
- Google Scholar [GS] [Librería digital]: https://scholar.google.com/
- Springer Link [SPRINGER] [Librería digital]: https://link.springer.com
- Science Direct [SD] [Librería digital]: https://www.sciencedirect.com/
- ACM Digital Library [ACM] [Librería digital]: https://dl.acm.org/
- Scholar Space [SS] [Librería Digital]: https://scholarspace.manoa.hawaii.edu/
- Google [Google][Buscador generalista]: https://www.google.com

Dado que cada una de estas librerías digitales poseen su propia sintaxis, se adapto la consulta de búsqueda para ser utilizada en cada motor, como se muestra en la tabla 1.

2.3 Criterios de inclusión y exclusión

Para incrementar la precisión de la investigación, se han eliminado estudios a través de la definición de criterios de inclusión y exclusión, de acuerdo con lo recomendado por [7]. Los criterios de inclusión planteados fueron:

- El título sugiere que su contenido tiene relación con DS.
- El título sugiere que su contenido está relacionado con la gestión de proyectos de datos.
- El título posee alguna de las siguientes palabras: *Agile, Data Science, Data Teams, Data Projects, Roles*.
- El Abstract contiene alguna de las siguientes frases: *Agile Data Science, Data Teams, Data Projects, Roles*.
- El Abstract permite determinar que está relacionado con DS como así también, con su implementación en un entorno de desarrollo determinado o con la definición de perfiles asociados a este tipo de proyectos.
- Para aquellas búsquedas realizadas en los motores de búsqueda generalistas, se tendrá en cuenta si es un artículo de blog.

Tabla 1. Cadenas de texto utilizadas para las búsquedas

Fuente	Consulta utilizada	Condiciones
IEEX	(("All Metadata":data science) AND "All Metadata":agile)	Title (matches all) Filtros: 2015-2019

	((("All Metadata":data science) AND "All Metadata":software) AND "All Metadata":teams)	Filtros: 2015-2019 Conferences, Journals
Google Scholar	allintitle: agile "data science"	
	allintitle: agile "data science" teams	
Springer	agile data science	Title (matches all) AND Abstract (matches all) AND Keywords (matches all) Filtros: Entre 2016 y 2020
	agile data science teams	Title (matches all) AND Abstract (matches all) AND Keywords (matches all) Filtros: Entre 2018 y 2019
ACM	(+agile +"data science")	Any field matches all (agile “data science”
	(+agile +"data science" +teams)	Any field matches all (agile “data science” teams
SS	title:("data science")	Abstract (matches “data science” AND agile)
	title:("data science")	Abstract (matches "data science" AND agile AND teams))
Google	“agile teams OR software "data science”	
	“agile OR software OR engineer OR management "data science”	
	“software OR teams OR management "data engineer"	

Además, hemos definido criterios que permiten identificar artículos que aportan información imprecisa o irrelevante a fin de descartarlos. Es por ello que, de los artículos preseleccionados, se eliminaron aquellos que cumplían con los siguientes criterios de exclusión:

- Estudios que no estén relacionados y que no mencionen los roles asociados a proyectos de DS.
- Estudios que no proveen información detallada sobre la práctica o enfoque para abordar los proyectos de DS particularmente.
- Estudios cuyo contenido no expone información acerca de la forma de gestionar los proyectos de datos.
- Para los artículos de blog, se excluirán aquellos de los que no se pueda obtener de alguna forma, la fecha de publicación y los autores involucrados.
- Estudios que se correspondan con cursos online, foros, videos, y relacionados.
- Estudios que nos fuere imposible acceder, por motivos ajenos a los revisadores.

2.4 Extracción de datos y análisis

En la fase de extracción de datos se reunió toda la información de los estudios que permitirán dar respuesta a las preguntas de investigación [7]. Primeramente, se recolectó información fundamental para identificar cada uno de los estudios: Título y autores, Abstract, Tipo de estudio (papers, blogs, etc.), Año de publicación.

Es importante mencionar, que los artículos de blogs no tienen Abstract, por lo que se omitió este elemento. Además, para analizar con mayor profundidad los estudios seleccionados, se decidió extraer la siguiente información:

- El nombre del enfoque de trabajo adoptado.
- Las prácticas ágiles identificadas en la propuesta.
- Las adecuaciones realizadas sobre prácticas ágiles.
- Los roles involucrados en el desarrollo de proyectos de datos.
- Los problemas que surgen a partir de los roles incluidos o aquellos derivados de la ejecución de las prácticas.

Además, se conformaron Grupos de Estudio Primarios (GEP) para simplificar el análisis realizado sobre los trabajos recolectados. Se agruparon aquellos estudios que poseen líneas de investigación similares o tópicos en común. Utilizando el enfoque de GEP y las verificaciones de calidad, se pueden contestar las preguntas establecidas en la Sección 2.1.

Validaciones GEP.
En vistas de obtener un mejor análisis de los GEP con respecto al tópico de investigación, se definen 4 evaluadores de GEP (VG) y los puntajes que se tendrán en cuenta. El proceso seleccionado fue: Sí (S) = 1, Parcialmente (P) = 0,5 y No (N) = 0. Las preguntas de validación usadas en esta revisión, fueron:

- VG1: El estudio, ¿cita y describe un enfoque de trabajo para abordar los proyectos de datos? Calificación: S: Cita un enfoque de trabajo para abordar los proyectos de datos. Además, describe prácticas puntuales y/o cómo implementarlas. P: No cita un enfoque de trabajo para abordar los proyectos de datos, pero describe o indica cómo se desarrollan los proyectos. N: No cita ningún enfoque o práctica de trabajo.
- VG2: El estudio, ¿menciona los roles involucrados en la gestión del proyecto de datos? Calificación: S: Menciona todos los roles involucrados en la gestión del proyecto de datos. P: Menciona los roles asociados únicamente a la gestión de datos. N: No menciona roles asociados a la gestión de proyectos de datos.
- VG3: El estudio, ¿describe los perfiles de datos involucrados en el enfoque de trabajo adoptado? Calificación: S: Describe los perfiles de datos involucrados. P: Sólo menciona los roles relacionados a la gestión de los datos. N: No menciona roles.
- VG4: El estudio, ¿expone ventajas o beneficios de aplicar el enfoque de trabajo propuesto a partir de la experiencia? Calificación: S: Expone ventajas y beneficios de aplicar el enfoque de trabajo propuesto a los proyectos de datos a partir de casos de estudio reales. P: Expone ventajas y beneficios que podrían obtenerse del enfoque de trabajo propuesto, pero no surgen de ninguna aplicación real. N: No expone ni ventajas, ni beneficios.

2.5 Proceso que conduce la revisión

El último paso en la etapa de planificación es definir el proceso que conduce la revisión. En este caso, definimos un proceso basado en el propuesto en [9]. El primer paso en el proceso consiste en enumerar las fuentes de datos (FD) y las consultas de búsqueda usadas. Una vez identificadas las FD se debe trabajar con su motor de búsqueda. Para este propósito, el motor de búsqueda de cada FD es analizado de manera de adaptar las

consultas a la sintaxis de cada motor: por lo que cada cadena de búsqueda (CB) es usada para buscar estudios en la FD correspondiente (Tabla 1). El siguiente paso consiste en almacenar y enumerar todos los estudios obtenidos de cada FD, este paso implica la extracción de algunos datos (título, autores, etc).

A continuación, la selección de estudios primarios consiste en chequear el cumplimiento de los criterios de inclusión definidos (sección 2.3). Luego, se remueven aquellos estudios duplicados que pudiera aparecer producto de la utilización de múltiples fuentes de datos; y se evalúa cada estudio no duplicado según los criterios de exclusión definidos en la sección 2.3 de manera de eliminar aquellos que no formarán parte del conjunto final, de los cuales se extraerán los datos definidos en la sección 2.4.

3 Resultados

En esta sección se presentan los resultados obtenidos en la RSL siguiendo lo detallado previamente. Nótese que el análisis fue realizado a cada estudio de forma individual.

3.1 Búsqueda y selección primaria

Se realizó la búsqueda en librerías digitales y, posteriormente, en el motor de búsqueda "Google". Los resultados se presentan agrupados.

En la Tabla 2, se muestran los resultados obtenidos, agrupados por FD. Como se puede observar, mediante la ejecución de estas búsquedas se obtuvieron 796 resultados. La segunda columna muestra los resultados de cada FD. De ellos, se seleccionaron 70, lo que representa el 8,79%, que cumplían con algún criterio de inclusión. El porcentaje se vio afectado por estudios que exponían el uso de DS para obtener métricas sobre las prácticas ágiles implementadas por algún equipo de desarrollo de software, lo cual no forma parte del objetivo de esta investigación.

La FD Google aportó 55,71% de los Estudios Relevantes (ER), siendo la de mayor peso. Sólo se tuvieron en cuenta los resultados que se encontraban hasta la segunda página (30 resultados por página). Esto se debe a la gran cantidad de elementos que el buscador obtiene, y a partir de pruebas realizadas, se concluyó que los resultados destacados se encuentran entre la primera y la segunda página.

Seguidamente, IEEEXplore aportó 22,86% de ER y la librería digital con menor porcentaje (1,43%) fue Scholar Space, aportando un único artículo.

Tabla 2. Resumen de librerías digitales

Librería digital	Resultados de la búsqueda	Estudios relevantes [E.R.]	% de resultados relevantes	% de todos los ER
IEEX	631	16	2,54	22,86
Google Scholar	14	5	35,71	7,14
Springer Link	13	3	23,08	4,29
Science Direct	28	3	10,71	4,29
ACM	13	3	23,08	4,29
Scholar Space	7	1	14,29	1,43
Google	90	39	43,33	55,71
Total	796	70	8,79	100,00

La Tabla 3 detalla las cantidades de estudios duplicados, donde el 15,71% de los estudios relevantes encontrados en esta RSL están presentes en más de una fuente de datos. A continuación, hemos removido los estudios duplicados, resultando un total de 59 estudios únicos (no duplicados). Luego, se retiraron aquellos que cumplían con algún criterio de exclusión, constituyendo un total de 26 ER descartados. Finalmente, se obtuvieron 33 ER para realizar las validaciones de calidad.

Tabla 3. Cantidad de estudios duplicados

	Cantidad	Porcentaje (%)
Estudios relevantes	70	100,00
Estudios relevantes duplicados	11	15,71
Estudios relevantes no-duplicados	59	84,29

De acuerdo con los criterios establecidos en la sección 2.4, se conformaron 6 agrupaciones GEP, que se exponen en la Tabla 4.

Tabla 4. Grupos de Estudios Primario

Nº	Tema	Autores	Título	Año
1	Roles asociados a los datos en equipos de desarrollo	M. Kim; T. Zimmermann; R. DeLine; A. Begel	Data Scientists in Software Teams: State of the Art and Challenges	2018
		M. Kim; T. Zimmermann ; R. DeLine ; A. Begel	The Emerging Role of Data Scientists on Software Development Teams [22]	2016
		J. S. Saltz; N. W. Grady	The Ambiguity of Data Science Team Roles and the Need for a Data Science Workforce Framework	2017
		I. Hukkelberg; M. Berntzen	Exploring the Challenges of Integrating Data Science Roles in Agile Autonomous Teams	2019
		J. S. Saltz; S. Yilmazel; O. Yilmazel	Not all Software Engineers can become good Data Engineers	2016
		I. Marin	Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas	2019
		DJ Patil	Building data science teams	2011
		J. Anderson	Data engineers vs data scientists	2018
		O. Kharkovyna	Who is a data engineer & how to become a Data Engineer? [23]	2019
		H. Danish	The Misunderstood Role Of a Data Engineer	2019
		S. K. White	What is a data engineer? An analytics role in high demand	2018
		V. Borda	How Data Science R&D Teams Can Adapt Agile Processes [17]	2019
2	Prácticas de la IS aplicada al DS	J. S. Saltz; I. Shamshurin	Does pair programming work in data science context? An initial case study	2017
		J. S. Saltz; I. Shamshurin; K. Crowston	Comparing Data Science Project Management Methodologies via a Controlled Experiment [14]	2017
		M. Shcherbakov; N. Shcherbakova; A.	Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development	2014

		Brebels; T. Janovsky; V. Kamaev		
3	Agilidad en el DS	N. W. Grady	Agile big data analytics: AnalyticsOps for data science	2017
		J. Sauvala	Combining data science with agile software development: a case study [13]	2019
		R. Jurney	Agile Data Science 2.0	2017
		R. Jurney	A manifesto for Agile data science [10]	2017
		E. Yan	Data Science and Agile. Part - 1 [16]	2019
		E. Yan	Data Science and Agile. Parte – 2 [28]	2019
		O. Cohen	Data Science? Agile? Cycles? My method for managing data-science projects in the Hi-tech industry	2019
		B-M. Swanson	Taking an Agile Approach to Data Science	2017
		R. Dev	Applying agile to data science [25]	2018
		J. Akred	Getting Real World Results From Agile Data Science Teams [26]	2017
		N. Hotz	Is Agile a Fit for Data Science?	2019
		Ch. H. Lee	Agile in Data Science: Why my scrum doesn't work?	2019
		T. Petersen; G. Ericson; K. Sharkey; D. Mabee; C. Casey; M. Learned; J. Martens	Agile development of data science projects [27]	2019
		J. Humpherys	Why Data Science doesn't respond well to agile methodologies	2019
4	Análisis del futuro de la agilidad y los datos	D. Larson; V. Chang	A review and future direction of agile, business intelligence, analytics and data science [11]	2016
5	Problemáticas en el DS	R. DeLine	Research Opportunities for the Big Data Era of Software Engineering	2015
		N. Hotz	Lessons from 20 Data Science Teams	2019
6	Escalar hacia organización guiada por datos	A. Fabijan; P. Dmitriev; H. Olsson; J. Bosch	The evolution of continuous experimentation in software product development: from data to a data-driven organization at scale	2017

3.2 Resultado de los datos obtenidos

A continuación, se realizó la extracción de datos de acuerdo con lo expresado en la sección 2.4, que se expone en la Tabla 5. Es importante mencionar, que el GEP5 ha sido eliminado por no cumplir con la validación de calidad.

Tabla 5. Resumen de la extracción de datos

	Enfoque	Prácticas ágiles	Adecuaciones	Roles	Problemas
GEP1	CRISP-DM	Scrum, Kanban, Meetings, Entrega continua, equipos multifuncionales	Prototipado, se añade etapas de investigación, etc	Data Scientist Data Engineer Researcher Data Analyst	Riesgos asociados; ambigüedad de roles.

GEP2	No menciona	Scrum, Kanban, Pair programming	Rol driver-researcher para pair programming	Data Scientist	Definición de sprints. estimación, el nivel de definición de las tareas.
GEP3	Team Data Science Process; Agile Data Science	Prácticas de Scrum y Kanban; Desarrollo iterativo	Equipos de investigación; perfiles académicos; no todas las iteraciones finalizarán en un MVP	Data Scientists Project manager	Obtener mayor productividad de agilidad en DS, sin comprometer la transparencia y reproductibilidad; requerimiento no implica tener producto.
GEP4	Agile Data Warehousing Extreme Scoping	Prácticas de Scrum; Ciclos cortos	Se propone una arquitectura, visión, prototipo de los datos	Data Scientists	Proceso de convertir los datos en información; lograr la participación de stakeholders.
GEP6	No menciona	Conceptos de Lean Despliegue continuo	Experimentación continua	Data scientists Software Eng. Researcher Program manager	No menciona

Estos datos permiten obtener una aproximación de las prácticas comúnmente utilizadas en los proyectos relacionados a los datos, los roles involucrados y las problemáticas que se deben abordar.

Resultados de las validaciones de GEP

Posteriormente, se verificó la calidad de los GEP, donde se ha asignado una calificación a cada estudio dentro de un GEP por cada pregunta establecida en la sección "Validaciones GEP". Por cuestiones de espacio, el detalle de los resultados se puede consultar en frre.utn.edu.ar/cinaptic/paginas/view/item/verifcalidadgep.

La tabla 6 presenta un resumen donde se agrupan los valores obtenidos por pregunta de verificación de calidad del GEP.

La pregunta con mayor puntaje es la VG2: "El estudio, ¿menciona los roles involucrados en la gestión del proyecto de datos?". Esto sucede ya que los estudios, en su mayoría, al menos citan los roles que debería tener un proyecto relacionado al DS. Por el contrario, la que obtuvo menor puntaje es la VG3, "El estudio, ¿describe los perfiles de datos involucrados en el enfoque de trabajo adoptado?". Si bien los estudios citaban roles asociados a los proyectos de DS, no especifican para el enfoque de trabajo que describen.

Tabla 6. Puntaje de las VG

	VG1	VG2	VG3	VG4	Total
Total	2,93	3,54	2,63	2,96	12,06
%Total	24,31	29,38	21,77	24,54	100

Aunque se esperaba obtener un puntaje más alto en la VG1, los estudios no declaran un nombre para el enfoque de trabajo con el que desarrollan los proyectos de datos, sino que mencionaba las prácticas, formas de trabajar, comunicarse, herramientas utilizadas, etc., lo que disminuyó el valor de esta pregunta. Luego de las validaciones realizadas, se cuenta con 20 estudios para responder las preguntas que originan esta RSL.

4 Discusión

En esta sección se responderá a las preguntas de investigación planteadas en la sección 2.1, utilizando los principales resultados presentados en la sección anterior.

4.1 ¿Cuál es el enfoque de trabajo de proyectos de DS?

La velocidad de procesamiento es muy importante para este ambiente debido a que generalmente los datos a analizar provienen del campo de Big Data, y los modelos predictivos que se elaboran, validan y alteran en proyectos de DS, suelen desarrollarse en varios intentos. En este sentido la adopción de prácticas ágiles, y en particular el uso de iteraciones, permiten obtener resultados rápidamente, adaptarse a los cambios, eliminar pasos innecesarios en el diseño y corregir los errores en etapas tempranas [6], [10]–[13]. Esto permite disminuir los riesgos y mejorar los incrementos realizados [11].

Autores como [14]-[16] describen que el uso de Scrum en proyectos de DS permite iniciar rápidamente con las labores asociadas al desarrollo, pero presenta inconvenientes asociados a la duración de los sprints y la variabilidad de los objetivos establecidos inicialmente. Aunque menos utilizado, Kanban permite lograr mejores resultados en las tareas asociadas a la investigación [14], [15]. Además, uno de los estudios se basa en Scrumban para crear un enfoque de trabajo [17], lo cual representa una aproximación interesante.

Por otro lado, que compartir los avances frecuentemente con todo el equipo sea parte del proceso, genera un aumento en la colaboración [6]. Por lo anterior, el prototipado y la construcción de un MVP son también prácticas utilizadas en los proyectos de DS, tanto para compartir el trabajo con miembros del equipo y los Stakeholders, como para recibir feedback apropiado de los mismos [13]. Por último, existen propuestas para trabajar con agilidad en proyectos de DS en ámbitos de experimentación a gran escala [18] y en la industria [19].

Si bien en los artículos analizados se observan una gran variedad de técnicas ágiles implementadas en proyectos de DS, no están claros los beneficios que otorgan dichas técnicas.

4.2 ¿Cuáles son los perfiles en proyectos de DS?

Principalmente, se menciona el perfil del Data Scientist. Dentro de las habilidades pertinentes a este rol, se encuentran la fuerte competencia en Matemáticas, Estadísticas y Física [6], [20]–[22]. Sus actividades están relacionadas al análisis y procesamiento de los datos [5], [22], como así también, a la interacción con los clientes o miembros

del negocio para la comprensión del dominio del problema en pos de obtener una mejor visión sobre los datos [20].

Otro rol frecuentemente mencionado es el Data Engineer. A diferencia del anterior, éste posee conocimientos tanto en herramientas de programación como aquellas que son de utilidad en un proyecto de desarrollo enfocado a los datos [20], [23], [24]. Las actividades que realiza están vinculadas a la construcción de los *pipelines* que los Data Scientist emplean para construir los modelos [20] y a facilitar el formato adecuado a los datos para trabajar sobre ellos [20], [24], es decir, se concentrarán en construir y mantener la infraestructura que permitirá la accesibilidad a los datos [6]. Por otro lado, se mencionan perfiles espécificos a las necesidades de la organización, como el de "Investigador" [5], [22] y "Especialistas en Modelos" [22]. En cuanto a los relacionados con la IS, se mencionan al "Ingeniero de Software", "Project Manager" y "Stakeholders", entre otros.

En la mayoría de los trabajos analizados, se incorpora el rol de Data Scientist (del que se obtuvieron diversas descripciones en cuanto a tareas y responsabilidades), y se incluyen diferentes roles en función de las necesidades de los proyectos.

4.3 ¿ Los enfoques se aplican de manera pura o tienen adaptaciones?

A menudo, los equipos realizan adaptaciones para obtener más beneficios de las prácticas ágiles como, por ejemplo, mitigar riesgos relacionados a la incertidumbre que puede tener el cliente sobre las funcionalidades que solicita [17]. Se incluyen etapas asociadas a la investigación, donde se comprende el dominio del problema, se observan las demandas del producto y se piensan en posibles soluciones a las características requeridas [19], [25].

La mayoría de los estudios proponen el uso de Scrum o Kanban adaptándolas a proyectos de DS: modificando la duración de los sprints y las actividades asociadas [11], [26], [27], modificando el enfoque o los temas tratados en las reuniones de Scrum [11], [17], [19], [26] e incorporando columnas al tablero de Kanban [14], [15]. Otras de las adaptaciones que se pueden citar al implementar la técnica "*pair programming*", el rol del observador se convierte en "investigador" [12]. Además, en [28] se sugiere invertir tiempo en planificar el proyecto y escribirlo para esclarecer ideas que pueden parecer confusas.

De acuerdo con [15], la principal problemática de aplicar Scrum al DS surge de congelar el objetivo establecido para el Sprint durante 2 semanas ya que, con cada nueva información obtenida de los datos o el dominio del problema, el objetivo podría modificarse y hasta tornarse inalcanzable, y se debe interrumpir el Sprint con frecuencia.

Asimismo, existe dificultad para definir los requerimientos de acuerdo con [16], ya que no existe un camino claro hacia el objetivo del proyecto, por lo que es dificil comprometerse con necesidades al principio del sprint, ya que las mismas pueden cambiar día tras día. Esto, además, genera dificultades al realizar estimaciones. Por último, destaca la diferencia de resultados entre proyectos de desarrollo de software y proyectos de DS, donde no siempre se obtendrá un entregable para el cliente (por ejemplo, el resultado de un experimento, muy común en el ámbito de DS). Cabe

destacar que en [5] se menciona la necesidad de obtener descripciones de los frameworks de trabajo, de forma tal que se pueda identificar, reclutar, entrenar, desarrollar y mantener aquellas habilidades pertinentes a través de la provisión de un lenguaje para categorizar y describir el tipo de DS que debe realizarse.

Por último, en [25] se menciona la dificultad relacionada a obtener la productividad, principal beneficio del DS Ágil, sin comprometer la transparencia y la reproductibilidad en los flujos de trabajo.

Se ha podido observar que se realizan ajustes en las técnicas ágiles de acuerdo a las particularidades de cada proyecto maximizando los beneficios obtenidos. Por lo que no se cuenta con un listado de mejoras sugeridas, sino que las mismas dependen de los proyectos y la experiencia de las personas que participan.

5 Conclusiones

Este estudio presenta un análisis del estado del arte de los enfoques de trabajo, procesos y equipos de Data Science en el contexto de la Ingeniería de Software, fundamentalmente, de la aplicación de prácticas ágiles para la gestión de proyectos y la definición de los roles involucrados.

Las conclusiones que se obtienen a partir del objetivo planteado son las que se nombran a continuación:

Se encontró una pluralidad de equipos aplican agilidad al Data Science. Sin embargo, dadas las diferencias existentes entre los proyectos de desarrollo de software y los de Data Science, las prácticas implementadas se modifican (al menos ligeramente) para obtener beneficios de su implementación y no incrementar los riesgos. Es por ello que se descubrieron propuestas de workflows para proyectos de Data Science elaboradas por profesionales a partir de sus experiencias.

A pesar de la creciente tendencia de aplicar agilidad en los proyectos de Data Science, no se dispone de descripciones de frameworks homogéneos que se implementen de forma similar en las empresas dedicadas a este tipo de trabajo, con lo que el workflow presenta considerables variaciones de organización en organización. Esto puede ser un inconveniente ya que los Data Scientists frecuentemente no poseen conocimientos en gestión de proyectos de desarrollo, con lo cual, deberán aprender cuál es el enfoque de trabajo en la organización que se insertan, pudiendo demorar más tiempo de lo usual que comiencen a producir valor.

Las responsabilidades de los distintos roles asociados a los proyectos no se encuentran claramente definidas. Esto deriva en la confusión de qué roles son los que posee la organización, las tareas que debe realizar cada miembro y qué habilidades se requieren al incorporar recursos humanos a las empresas.

Cabe destacar que la mayor cantidad de estudios se obtuvieron del motor de búsqueda “Google”. Desde nuestro punto de vista, esto puede deberse a la emergente aplicación del DS, y la facilidad de publicar información, experiencias o técnicas de gestión implementadas en los proyectos de esta índole, a través de *blogs.*

Como trabajo futuro, se pretende extender está revisión para profundizar en la problemática de no disponer de descripciones de enfoques de trabajo de forma tal que

se pueda corroborar si realmente los Data Scientists perciben el inconveniente de demorar tiempo entendiendo el workflow antes de comenzar a entregar resultados.

Por otro lado, se plantea realizar un análisis sobre el estado actual de los enfoques de trabajo implementados en equipos de Data Science en la República Argentina para identificar si se aplica algún framework de trabajo, cuáles son los más habituales y cuál es el nivel de implementación de agilidad en estas temáticas, con el objetivo de proponer un enfoque ágil que se ajuste a las necesidades de dichos proyectos.

Agradecimientos

Este trabajo ha sido financiado en forma conjunta por CONICET y la UTN. Se agradece el apoyo brindado por estas instituciones. Además, ha sido parcialmente financiado por el Ministerio de Economía, Industria y Competitividad del Gobierno de España, bajo el proyecto MADRID (TIN2017-88557-R)

Referencias

1. C. Hayashi, “What is Data Science ? Fundamental Concepts and a Heuristic Example,” pp. 40–51, 1998.
2. M. Loukides, *What ’ s Data Science ?* 2011.
3. A. Liu, “Data Science and Data Scientist,” p. 11, 2015.
4. B. Dhar, “Data Science and Prediction,” *Commun. ACM*, vol.56, no.12, pp.64–73, 2013.
5. J. S. Saltz and N. W. Grady, “The ambiguity of data science team roles and the need for a data science workforce framework,” *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 2355–2361, 2017.
6. I. Hukkelberg and M. Berntzen, “Exploring the Challenges of Integrating Data Science Roles in Agile Autonomous Teams,” *2019 Agil. Process. Softw. Eng. Extrem. Program. - Work.*, vol. 364, pp. 37–45, 2019.
7. B. Kitchenham, S. Charters, “*Guidelines for performing Systematic Literature Reviews in SE*,” pp. 1–44, 2007.
8. P. Mian, T. Conte, A. Natali, J. Biolchini, and G. Travassos, “A Systematic Review Process for Software Engineering,” *Empir. Softw. Eng.*, vol. 32, no. 3, pp. 1–6, 2007.
9. F. Pino, F. García, M. Piattini, “SW process improvement in small and medium software enterprises: A systematic review,” *Softw. Qual. J.*, vol.16, no.2, pp.237–261, 2008.
10. R. Jurney, “A manifesto for Agile data science,” 2017. Avail: https:// www.oreilly.com/radar/a-manifesto-for-agile-data-science/. [Acces: 27-Nov-2019].
11. D. Larson and V. Chang, “A review and future direction of agile, business intelligence, analytics and data science,” *Int. J. Inf. Manage.*, vol. 36, no. 5, pp. 700–710, 2016.
12. J. S. Saltz and I. Shamshurin, “Does pair programming work in a data science context? An initial case study,” *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 2348–2354, 2017.
13. J. Sauvala, “Combining Data Science with Agile Software Development: A Case Study,” Aalto University, 2019.
14. J. S. Saltz, I. Shamshurin, and K. Crowston, “Comparing Data Science Project Management

Methodologies via a Controlled Experiment," *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, pp. 1013–1022, 2017.
15. H. L. Chang, "Agile in Data Science: Why My Scrum Doesn't Work?," 2019. Avail: https://changhsinlee.com/agile-ds-scrum-kanban/. [Acces: 27-Nov-2019].
16. E. Yan, "Data Science and Agile - Part 1," 2019. Avail: https://towardsdatascience.com/agiledatascience-3b7ca65278a4. [Acces: 27-Nov-2019].
17. V. Borda, "How Data Science R&D Teams Can Adapt Agile Processes," 2019. Avail: https://blog.invoca.com/inside-engineering-adapting-agile-for-data-science-rd-teams/. [Acces: 27-Nov-2019].
18. A. Fabijan, P. Dmitriev, H. Olsson, J. Bosch, "The Evol of Continuous Experimentation in SW Product Development: From Data to a Data-Driven Organization at Scale," *Proc. - 2017 IEEE/ACM 39th Int. Conf. Softw. Eng.*, pp. 770–780, 2017.
19. O. Cohen, "Data-science? Agile? Cycles? My method for managing data-science projects in the Hi-tech industry.," 2019. Avail: https:// towardsdatascience.com/data-science-agile-cycles-my-method-for-managing-data-science-projects-in-the-hi-tech-industry-b289e8a72818. [Acces: 27-Nov-2019].
20. J. Anderson, "Data engineers vs. data scientists," 2018. Avail: https://www.oreilly.com /radar/data-engineers-vs-data-scientists/. [Acces: 27-Nov-2019].
21. M. Kim, T. Zimmermann, R. Deline, A. Begel, "Data scientists in sw teams: State of the art and challenges," *IEEE Trans. Softw. Eng.*, vol. 44, no. 11, pp. 1024–1038, 2018.
22. M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "The emerging role of data scientists on software development teams," *Proc. - Int. Conf. Softw. Eng.*, vol. 14-22-May-, pp. 96–107, 2016.
23. O. Kharkovyna, "Who Is a Data Engineer & How to Become a Data Engineer?," 2019. Avail: https://towardsdatascience.com/who-is-a-data-engineer-how-to-become-a-data-engineer-1167ddc12811. [Acces: 27-Nov-2019].
24. J. S. Saltz and S. Yilmazel, "Not All Software Engineers Can Become Good Data Engineers," *2016 IEEE Int. Conf. Big Data (Big Data)*, pp. 2896–2901, 2016.
25. R. Dev, "Applying agile to data science," 2018. Avail: https://medium.com /@rahul.dev/applying-agile-to-data-science-63d482149206. [Acces: 27-Nov-2019].
26. J. Akred, "Getting Real World Results From Agile Data Science Teams," 2017. Avail: https://www.kdnuggets.com/2017/02/real-world-results-agile-data-science-teams.html. [Acces: 27-Nov-2019].
27. T. Petersen *et al.*, "Agile development of data science projects - Team Data Science Process," 2019. Avail: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/agile-development. [Acces: 27-Nov-2019].
28. E. Yan, "Data Science and Agile - Part 2," 2019. Avail: https://towards datascience.com/data-science-and-agile-1cdfb1667789. [Acces: 27-Nov-2019].