PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Marco de Trabajo para el desarrollo de proyectos de analítica de datos

Trabajo de investigación para obtener el grado académico de Maestro en Informática con mención en Ciencias de la Computación que presenta:

César Alberto Olivera Cokan

Asesores:

Mg. Alejandro Toribio Bello Ruiz

Dr. José Antonio Pow Sang Portillo

Lima, 2024

Informe de Similitud

Yo, Alejandro Toribio BELLO RUIZ, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor del trabajo de investigación titulado "Marco de Trabajo para el desarrollo de proyectos de analítica de datos" del autor César Alberto OLIVERA COKAN**,** dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 18%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 3/07/2024.
- He revisado con detalle dicho reporte y el trabajo de investigación, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

San Miguel, 3 de Julio de 2024.

<table>
<tr><td colspan="2">Apellidos y nombres del asesor / de la asesora:
BELLO RUIZ, Alejandro Toribio</td></tr>
<tr><td>DNI:
16656624</td><td rowspan="2">Firma
  </td></tr>
<tr><td>ORCID:
https://orcid.org/0000-0002-8608-6364</td></tr>
</table>

Resumen

El desarrollo de proyectos de analítica de datos en las organizaciones requiere de procesos bien definidos para su éxito. Existen procesos estándar de analítica de datos, como CRISP-DM, que han tenido una amplia adopción en las últimas décadas. Sin embargo, mediante una búsqueda sistemática de la literatura se ha podido evidenciar que muchas de las organizaciones a menudo no aplican CRISP-DM o procesos similares, como SEMMA y KDD, tal como están, sino que muchos de ellas adaptan estos marcos de trabajo para abordar requerimientos específicos en diversos contextos de la industria. Además, según estos estudios se evidencia que un grupo considerable de empresas emplea Scrum u otros marcos de trabajo para el desarrollo de software con el fin de llevar a cabo sus proyectos de analítica de datos, lo cual no es correcto pues estos marcos de trabajo no abordan las particularidades de un ciclo de vida de una solución analítica. Si bien CRISP-DM es el marco de trabajo para analítica de datos más empleado, este mismo posee un conjunto de falencias enfocadas en diversos casos de uso o procesos de negocio que ha llevado a muchas organizaciones a adaptar este marco a sus necesidades. Hasta ahora no se ha sugerido ninguna adaptación que permita abordar las falencias que los diferentes dominios en la industria poseen. Este artículo aborda la propuesta del diseño de un marco de trabajo para proyectos de analítica de datos general denominado GEN-DA (*Generic Data Analytics framework* por sus siglas en inglés). GEN-DA extiende y modifica CRISP-DM para solucionar las diferentes falencias encontradas en la literatura y lograr un ciclo de vida del proyecto de analítica de datos que pueda ser empleado en todos los contextos de la industria. Este marco de trabajo ha sido diseñado y evaluado de forma iterativa empleando una metodología en ciencias del diseño gracias a la participación de expertos en analítica de datos mediante el método de validación por Juicio Experto. Los resultados obtenidos son alentadores y habilita la factibilidad de emplear este marco propuesto en un entorno real, cuyos resultados, se presume, que serán satisfactorios.

Palabras clave

Analítica de datos; Minería de datos; Ciencia del diseño; Marco de trabajo; Juicio experto

Tabla de contenidos

Índice de imágenes

Índice de tablas

1. Introducción

Durante las últimas décadas, los proyectos de analítica de datos han aumentado su adopción entre las organizaciones que buscan mantener y mejorar su nivel de competitividad y valor comercial en la industria (Davenport & Harris, 2017). Esta tendencia ha llevado a varias organizaciones grandes a administrar una rica cartera de proyectos de analítica de datos (Davenport & Harris, 2017). El éxito de este tipo de proyectos en las empresas se basa en tener un enfoque estructurado, claro y repetible. En un estudio realizado en el año 2021 por McKinsey & Company (McKinsey & Company, 2021) se evidencia que, si bien la adopción de soluciones analíticas por parte de las organizaciones ha ido en aumento, la adopción aún no sigue siendo amplia, lo cual nos indica que no hay un buen índice de éxito de proyectos de analítica de datos.

Con el fin de mejorar la adopción de soluciones analíticas, surgen los marcos de trabajo para proyectos de analítica de datos, siendo el más utilizado el CRISP-DM, según estudios realizados por la comunidad KDnuggets (KDnuggets, 2014) y Data Science Process Alliance (Data Science Process Alliance, 2020). También existen otros marcos de trabajo descritos en la literatura que han sido adoptados en las empresas como SEMMA y KDD. Sin embargo, podemos observar, en los estudios mencionados, que la adopción de estos marcos no es total, lo cual se puede evidenciar por la considerable cantidad de empresas que han decidido utilizar marcos de trabajo de desarrollo de software e incluso optaron por desarrollar sus propios marcos según sus necesidades debido a que los marcos de trabajo existentes cuentan con falencias acorde a sus procesos de desarrollo.

Este artículo apunta a cubrir estos vacíos que presentan los marcos de trabajo existentes estableciendo un nuevo marco de trabajo para proyectos de analítica de datos llamado **GEN-DA (Generic Data Analytics Framework)**. GEN-DA adapta y extiende los marcos de trabajo existentes descritos en el estado del arte y la búsqueda sistemática de la literatura. Para la construcción del nuevo marco de trabajo, se analizó marcos de trabajo reportados en la literatura realizando una búsqueda sistemática y en base a ello un catálogo de vacíos o puntos no cubiertos por los marcos de trabajo. A continuación, se definió un nuevo ciclo de vida de proyectos de analítica de datos con lo cual fue posible elaborar una nueva propuesta de marco de trabajo para proyectos de analítica de datos. Finalmente, esta nueva propuesta fue validada mediante un método de juicio experto teniendo la participación de 4 expertos en el campo de analítica de datos.

En las siguientes secciones del artículo la información está estructurada de la siguiente manera: En la sección 2, el estado del arte introduce conceptos principales de la analítica de datos y marcos de trabajo existentes. A continuación, en la sección 3, el desarrollo, se describe la búsqueda sistemática realizada, el catálogo de falencias (*gaps*), la identificación del ciclo de vida, la elaboración del nuevo marco de trabajo para proyectos de analítica de datos y la validación de la propuesta de marco de trabajo, en donde se detallan los resultados de haber realizado el método de juicio de expertos. Así mismo, se describen los ajustes que fueron necesarios realizar al marco en base a las observaciones del jurado de expertos. Finalmente, en la sección 4, las conclusiones realizando un análisis de los resultados de la validación y recomendaciones sobre trabajos futuros que podrían mejorar la propuesta de marco de marco de trabajo.

2. Estado del arte

2.1. Marco conceptual

En esta sección se aborda los conceptos más importantes asociadas a marcos de trabajo y analítica de datos. Estos conceptos permitirán tener un mayor entendimiento del alcance de la propuesta de marco de trabajo.

2.1.1. Marcos de trabajo y Metodologías

Como sabemos existe una diferencia entre la definición de marco de trabajo y metodología. Por un lado, marco de trabajo es un grupo de prácticas y conceptos que permiten dar solución a un problema o a varios problemas de la misma índole utilizando lo que sea más útil para su solución. Por otro lado, según la RAE, una metodología es la manera de hacer las cosas ordenadamente, es decir, debemos seguir una serie de pasos para lograr un propósito, sin olvidarse de cada uno de ellos. En otras palabras, una metodología hace referencia a un modelo aplicable, el cual debe seguir necesariamente, una selección de técnicas concretas (o métodos), aun cuando estas resultan cuestionables. Por lo tanto, un marco de trabajo es más flexible y así mismo se adapta mejor a los cambios que una metodología, mientras que una metodología tiene un alcance más cerrado y brinda una serie de pasos que deben seguirse de manera más rígida. Sin embargo, si bien hablamos de que existe una gran diferencia conceptual entre estos dos términos, a veces es casi inevitable usar el término metodología, para referirnos a "*frameworks*" tal y como lo hacen CRISP-DM, SEMMA y KDD, las cuales en muchos artículos son considerados marcos de trabajo y otros como metodologías.

En el presente artículo abordaremos el establecimiento de una nueva propuesta de marco de trabajo por permitirnos mayor flexibilidad y adaptación a los cambios, lo cual nos permitirá una mayor generalización a nivel de todas las industrias.

2.1.2. Analítica de datos y minería de datos

La analítica de datos es la gestión de integrar datos heterogéneos provenientes de diferentes fuentes, realizar inferencias a través del análisis de estas, y tomar decisiones para habilitar la innovación con ello lograr ventajas competitivas y la ayuda en la toma de decisiones estratégica. Existen diferentes tipos de analítica de datos como la analítica descriptiva, analítica diagnóstica, analítica predictiva y analítica prescriptiva (Gudivada, 2017). La minería de datos, en específico, se refiere a la técnica de extraer información útil de grandes conjuntos de datos crudos estructurados y heterogéneos, para encontrar patrones e identificar relaciones ocultas entre ellos. Se enfoca, sobre todo, en el campo de la analítica predictiva, no obstante, la minería de datos es una piedra angular de la ciencia de datos. Por lo tanto, podemos concluir que la analítica de datos consta de un conjunto de procesos para la gestión de los datos y la minería de datos se encuentra dentro del proceso enfocado en analítica predictiva.

Muchas de los marcos de trabajo comerciales como CRISP-DM, SEMMA y KDD nacieron para abordar problemáticas acerca de la minería de datos. Sin embargo, podemos percatarnos que en la actualidad son empleadas en diferentes procesos de la analítica de datos.

2.1.3. Ciclo de vida de proyectos de analítica de datos

Debido a su importancia, existen diversas definiciones del ciclo de vida de proyectos de AA que constan de varias etapas, pero las más representativas son cuatro. En primer lugar, el análisis exploratorio de datos (Rolllins, 2015) consiste en investigar conjuntos de datos y resumir sus principales características y correlaciones de variables, a menudo con métodos de visualización de datos. En segundo lugar, el preprocesamiento de datos y limpieza de datos (Rolllins, 2015) consiste en tratar los datos no válidos,

valores faltantes, duplicidad de datos y brindar un formato adecuado de los datos. Con ello combinar los datos de múltiples fuentes y transformarlo en variables más útiles. En tercer lugar, el modelado (Rolllins, 2015) consiste en desarrollar modelos predictivos o descriptivos usando algoritmos de aprendizaje automático o de aprendizaje profundo que son capaces de brindar descubrimientos sobre la data para la toma de decisiones. Finalmente, se realiza la evaluación de modelos (Rolllins, 2015), la cual consiste en evaluar el actual rendimiento de los algoritmos elaborados. Con esta evaluación se podrá saber si el modelo analítico sufre de problemas de bajo y sobre ajustes, así mismo se podrá mejorar la selección de características dependiendo del rendimiento computacional que este conlleve.

2.2. Metodología para la construcción del marco de trabajo

Existen diversas metodologías de investigación que pueden ser empleados para la construcción de un marco de trabajo. El autor J. R. Venable et al.(Pries-Heje et al., 2017) recomienda que se utilice la Metodología de Investigación de Desarrollo de Sistemas (SDRM) si el resultado del artefacto de la investigación debe ser un sistema de TI, la metodología de Investigación en Ciencias del Diseño (DSRM) si se necesita una amplia adaptación del artefacto al uso diario, o se usa el Modelo de Proceso DSR (DSRPM) si el objetivo de la investigación es desarrollar la teoría del diseño. Como esta investigación tiene como objetivo desarrollar un nuevo artefacto; es decir, un nuevo marco de trabajo, así como asegurar su utilidad, aplicabilidad y relevancia en el dominio de la aplicación, se realizará una amplia adaptación del artefacto al uso diario en entornos prácticos. Por lo tanto, DSRM (Peffers et al., 2008) ha sido seleccionado como el método de investigación más adecuado.

Metodología de Investigación en Ciencias del Diseño (DSRM)

Esta metodología consta de 6 pasos que constituyen un proceso iterativo (Peffers et al., 2008). El primer paso consiste en la Identificación del Problema y Motivación. Este paso apunta a definir el problema de investigación y la motivación respecto a la significancia de la solución. En el segundo paso se definen los objetivos de la solución sean cuantitativos o cualitativos. El tercer paso consiste en el Diseño y Desarrollo del marco de trabajo. Tanto las funcionalidades deseadas como la arquitectura del artefacto son definidas y el prototipo es creado en este paso. El cuarto paso es el paso de la Demostración, el cual consiste en realizar un conjunto determinado de experimentos, simulaciones, casos de estudio y/u otros métodos aplicables. El uso del artefacto y como el problema de investigación es resuelto son presentados en este paso. El quinto paso consiste en la Evaluación. En este paso una evaluación formal es ejecutada para averiguar qué tan bien el artefacto asiste en resolver el problema. Al igual que en el paso de la Demostración, en este paso se suele ejecutar un conjunto de métodos de validación y, de acuerdo con los resultados, el artefacto es mejorado. Finalmente, el sexto y último paso es el de la Comunicación, el cual consiste en comunicar potenciales mejoras que pueden ser implementadas en futuros proyectos de investigación.

En el presente proyecto de investigación se consideró solo 4 pasos y se seleccionó métodos específicos en algunos de ellos, lo cual es detallado a continuación:

- **Identificación de los problemas y Motivación:** En este paso se definió la problemática mediante un árbol de problemas.
- **Formulación de los Objetivos:** En este paso se definió los objetivos de investigación soportada con una revisión sistemática de literatura y la elaboración de un catálogo de falencias (gaps).
- **Diseño y desarrollo de artefactos:** En este ciclo se define un catálogo de meta-requerimientos y principios de diseño. Así mismo, se elabora un ciclo de vida para proyectos de analítica y con ello un primer prototipo del marco de trabajo
- **Evaluación:** En este paso se realizar el método de juicio de expertos (Escobar-Pérez & Cuervo-Martínez, 2008). Con la ayuda de un jurado se evalúa el marco de trabajo iterativamente

obteniendo recomendaciones de mejora hasta llegar un estado mínimo aceptable por los expertos.

2.3. Antecedentes y trabajos relacionados

Existen diversos marcos de trabajo y metodologías de desarrollo de proyectos de Analítica de datos, pero los más usados son CRISP-DM, SEMMA y KDD según un estudio publicado por la comunidad KDnuggets (Data Mining Community's Top Resource)(KDnuggets, 2014). Según el mismo estudio, también existen marcos de trabajo que han sido creados y/o adaptados por los mismos equipos de analítica de datos. Por tal motivo, se empleó una búsqueda sistemática en los principales motores de búsqueda de artículos científicos con la finalidad de encontrar marcos de trabajo para proyectos de analítica de datos creados y/o adaptados.

2.3.1. Marcos de trabajo populares

En primer lugar, la metodología CRISP-DM es la más usada y consta de cuatro niveles de abstracción organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. De forma general, esta metodología consiste en un proceso que está organizado en seis fases las cuales van desde analizar el problema, analizar los datos, preparar los datos, modelar, evaluar y explotar. En segundo lugar, la metodología SEMMA (Azevedo & Santos, 2008) fue desarrollada por SAS Institute y es una de las más usadas por los equipos de AA. Consiste en un proceso de selección, exploración y modelado de datos para descubrir patrones desconocidos en estos datos. El proceso tiene cinco fases básicas las cuales van desde el muestreo de datos, exploración de datos, manipulación de datos, modelado y evaluación del modelo. Finalmente, el Proceso KDD (Knowleadge Discovery in Databases) (Azevedo & Santos, 2008) fue el primer marco que fue aceptado por la comunidad científica y estableció etapas principales para un proyecto de explotación de datos, enfocado principalmente en la minería de datos. El proceso KDD es interactivo e iterativo y consta de cinco fases, las cuales van desde la selección de datos, preprocesamiento y/o limpieza de datos, transformación de datos, minería de datos y evaluación.

2.3.2. Marcos de trabajo creados y/o adaptados

Búsqueda Sistemática

La búsqueda sistemática fue realizada de acuerdo con las directrices establecidas por B. A. Kitchenham y S. Charters (Charters & Kitchenham, 2007) para la revisión sistemática de la literatura. La elaboración de la cadena de búsqueda se basó en el método PICo (Población, Intervención y Contexto). La definición de conceptos generales mediante el uso de PICo se detalla a continuación:

- **Población**: Marco de trabajo, Metodología
- **Intervención**: Proyectos de analítica de datos, Proyectos de data y analítica, Proyectos de minería de datos
- **Contexto**: -

La cadena de búsqueda resultante considera algunos sinónimos para lograr una búsqueda más completa. Así mismo, se consideraron un conjunto de filtros además de considerar solo las publicaciones que datan del 2013 en adelante con el fin de analizar el estado actual de los marcos de trabajo.

C1: (methodology OR framework) AND
C2: ("advanced analytics project" OR "data analytics project" OR "data mining project") AND
C3: (publication year > 2012)

Los resultados de la búsqueda sistemática presentaron un total de 67 artículos de los cuales se aplicaron un conjunto de filtros y se seleccionaron finalmente 5 artículos que incluyeron marcos de trabajo para proyectos de analítica de datos, los cuales se detallan a continuación:

En un reciente artículo, (Plotnikova et al., 2022) propone un nuevo marco de trabajo para proyectos de analítica de datos enfocado el dominio del entorno financiero. Este nuevo marco nace a partir de un conjunto de principios, requerimientos y la identificación de un conjunto de falencias (gaps) que posee CRISP-DM frente al ámbito del proceso financiero. En base a estos principios, requerimientos y lista de falencias propone mejoras en las fases de entendimiento del negocio, evaluación y despliegue modificando e incorporando nuevos elementos al marco de trabajo. Así mismo, incorpora una nueva fase denominada post-despliegue que contempla actividades relevantes del ciclo de vida de un proyecto de analítica de datos, pues la solución de análisis de datos deberá tener un proceso de mejora claro a lo largo del tiempo, como, por ejemplo, definir planes de ingesta de datos y reentrenamiento de modelos analíticos.

Los autores (Qadadeh & Abdallah, 2020) proponen un nuevo marco de trabajo enfocado en el entorno del gobierno. Este nuevo marco de trabajo nace a partir de un contexto gubernamental que requiere sustentar y documentar adecuadamente los proyectos de software. Por tal motivo, se basa en CRISP-DM pero modifica las dos primeras fases que comprende el entendimiento del negocio y el entendimiento de los datos.

Enfocado en integrar procedimientos de analítica de datos en el ámbito de sistemas de manufactura, (Kozjek et al., 2020) propone un nuevo marco de trabajo conceptual que modifica y ajusta CRISP-DM en base a un conjunto de hechos y requerimientos de procesos de manufactura complejos. Si bien el marco de trabajo propuesto modifica y extiende las fases de CRISP-DM, este también incorpora nuevos elementos como el desarrollo y evaluación iterativa de prototipos de la solución de análisis de datos y refuerza el hecho de que es posible retornar a fases anteriores con el hecho de mejorar continuamente la solución analítica. Así mismo, la reorganización de los elementos y las fases del marco de trabajo da una mayor capacidad de adaptación para entornos con procesos de manufactura reales.

3. Desarrollo

En esta sección se describe el proceso de construcción del marco de trabajo que se llevó a cabo. El proceso se basa en la metodología DSRM la cual se dividió en 3 fases y 4 iteraciones. En primer lugar, la primera fase consiste en la formulación de los problemas y objetivos de la investigación. En segundo lugar, la segunda fase consiste en el diseño y desarrollo del marco de trabajo. Finalmente, la tercera fase consiste en la evaluación del marco de trabajo mediante un juicio de expertos con lo cual se obtendrá una retroalimentación. Las fases fueron distribuidas en cada iteración, por lo que la iteración 0 consiste en la formulación de problemas y objetivos. La iteración 1 al 3 consiste en repetir las fases 2 y 3 teniendo como entrada a la retroalimentación obtenida por los expertos.

3.1. Formulación de problemas y objetivos de la investigación

La hipótesis inicial plantea que los marcos de trabajo existentes no pueden abordar todos los problemas y casos de uso en la industria. Para apoyar esta hipótesis se realizó una búsqueda sistemática sobre marcos de trabajo en la industria. Los resultados revelaron que existen muchas falencias en los marcos de trabajo existentes e incluso se plantean nuevos marcos de trabajo extendiendo en muchos casos a CRISP-DM, marco de trabajo más utilizado. Por lo tanto, se plantea como objetivo el diseño de un nuevo marco de trabajo para el desarrollo de proyectos de analítica de datos que pueda cubrir las falencias detectadas. Así mismo, como objetivos específicos se plantea una revisión sistemática de la literatura de marcos de trabajo existentes, la elaboración de un catálogo de falencias (gaps), el desarrollo de un

nuevo marco de trabajo para el desarrollo de proyectos de analítica de datos y la evaluación del nuevo marco mediante el método de juicio de expertos. Esta fase concluye con la presentación del catálogo de falencias (ver tabla 1). En este catálogo de falencias (*gaps*) se muestra un consolidado de falencias recabadas en la búsqueda sistemática y anotaciones brindadas por expertos. Finalmente, se definen los principios de diseño y meta-requerimientos del marco de trabajo.

Falencia	Definición	Fuente de datos
G1 - Gestión de requisitos & elicitación	Falta de tareas para la validación y modificación de los requisitos existentes y obtención de nuevos	FIN-DM (Plotnikova et al., 2022)
G2 - Interdependencias	Falta de iteraciones entre las diferentes fases	FIN-DM (Plotnikova et al., 2022), MAN-DM (Kozjek et al., 2020)
G3 - Universalidad	Falta de soporte para varios resultados analíticos, técnicas especializadas y no supervisadas, y formatos de implementación	FIN-DM (Plotnikova et al., 2022)
G4 - Cumplimiento normativo	Falta de tareas para abordar el cumplimiento normativo	FIN-DM (Plotnikova et al., 2022), GOV-DM (Qadadeh & Abdallah, 2020)
G5 - Validación	Falta de apoyo para probar modelos en entornos de la vida real	FIN-DM (Plotnikova et al., 2022)
G6 - Accionabilidad	Falta de soporte para pilotar modelos en entornos de la vida real	FIN-DM (Plotnikova et al., 2022), MAN-DM (Kozjek et al., 2020)
G7 - Proceso	Controles de procesos de analítica de datos, garantía de calidad y habilitadores de procesos críticos (no se toman en cuenta datos, códigos, herramientas, infraestructura y factores organizacionales) necesarios para la ejecución efectiva de proyectos de analítica de datos	FIN-DM (Plotnikova et al., 2022), GOV-DM (Qadadeh & Abdallah, 2020), MAN-DM (Kozjek et al., 2020)
G8 - Identificación de problemas y negocio	Falta de soporte para el empleo de técnicas y metodologías para la identificación de problemas	GOV-DM (Qadadeh & Abdallah, 2020)
G9 - Esquema de ejecución de proceso	Falta de soporte para brindar técnicas de cómo segregar y paralelizar esfuerzos en tareas de analítica de datos	GOV-DM (Qadadeh & Abdallah, 2020)
G10 - Arquitectura de solución	Falta de soporte para abordar consideraciones sobre la solución analítica, prototipado y su construcción	MAN-DM (Kozjek et al., 2020)
G11 - Manejo del Know-How	Falta de soporte para la gestión del know-how de elementos clave del negocio y problema	MAN-DM (Kozjek et al., 2020)
G12 - Gestión de riesgos	Falta de soporte para la gestión de riesgos del proyecto.	Elaboración propia.

Tabla 1. Catálogo de falencias (Gaps) consolidado. Elaboración propia.

3.2. Desarrollo y validación del marco de trabajo propuesto

3.2.1. Desarrollo del marco de trabajo

En la fase de Desarrollo, es construido un nuevo ciclo de vida para proyectos de analítica de datos, los elementos lógicos que componen las fases del marco de trabajo y las relaciones que estos poseen. Como punto de partida en el diseño del nuevo marco de trabajo, CRISP-DM es tomado como base, ya que es el marco de trabajo más utilizado y difundido para proyectos de analítica de datos, lo cual permite a los usuarios GEN-DA poder familiarizarse con mayor facilidad al nuevo modelo lógico. Para poder expresar el ciclo de vida, este se representa mediante un modelo lógico empleando un diagrama de actividades siguiendo la notación UML 2.5.1, con el fin de poder representar las fases, elementos lógicos y relaciones entre elementos del marco de trabajo propuesto (véase la figura 1).

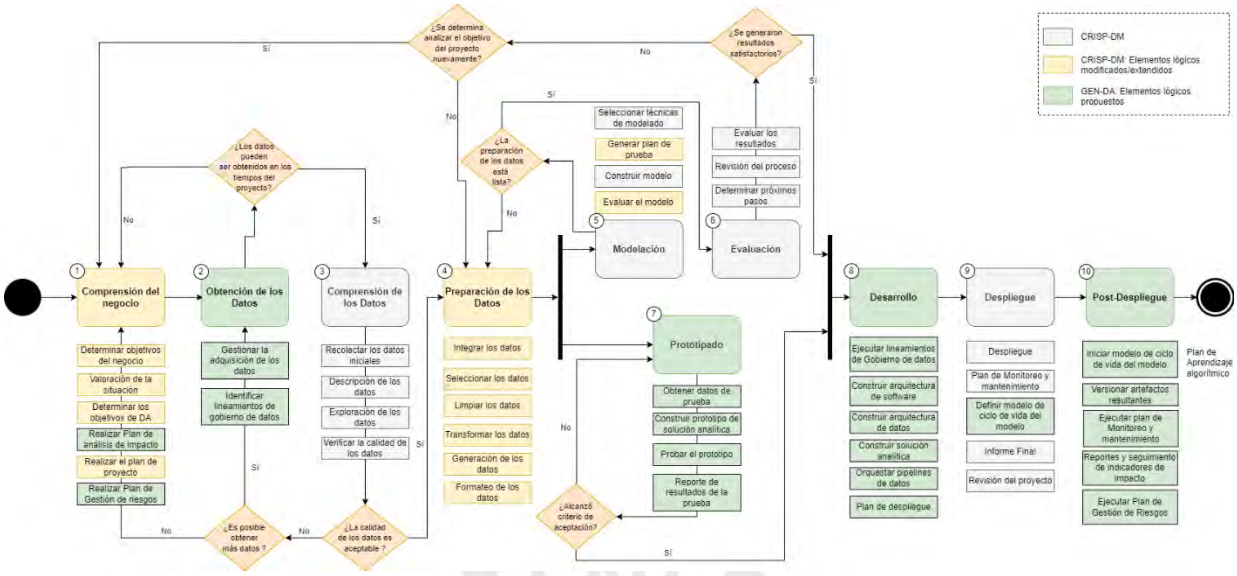

Figura 1. Diagrama de actividades de Marco de trabajo GEN-DA (Generic Data Analytics Framework) con notación UML 2.5.1. Elaboración propia.

El marco de trabajo GEN-DA tiene el objetivo de cubrir las necesidades evidenciadas en el catálogo de falencias (véase tabla 1). Así mismo, se llevó a cabo el proceso de definir un conjunto de principios de diseño a partir de meta-requerimientos que fueron el resultado de la traducción de la lista de falencias al dominio del diseño. Por ejemplo, según la tabla 1, la falencia G3 respecto a la universalidad puede traducirse en el meta-requerimiento no funcional de ser independiente a la plataforma, método y resultado, lo cual está relacionado al principio de diseño de Excluir especificaciones de tecnologías, métodos o resultados relacionados. Este último principio de diseño nos permitió tener una mayor claridad para la redacción de la descripción de componentes lógicos del marco de trabajo GEN-DA, ya que se evitó incluir información muy específica a la tecnología, método o resultado de algún dominio.

El catálogo de falencias (Gaps) y los principios de diseño sirvieron para diseñar la primera versión del marco de trabajo propuesto. La primera versión de GEN-DA sufrió diversos ajustes debido al proceso de validación, que constó de 3 evaluaciones empleando el método de juicio experto. En cada evaluación se finalizó con material para realizar ajustes al diseño del marco de trabajo. Una vez finalizado el proceso de validación con la satisfacción de los expertos, se concluye con una versión final del marco de trabajo propuesto denominado GEN-DA.

El marco de trabajo propuesto GEN-DA posee 10 fases (véase figura 2), las cuales pueden ser adaptadas, extendidas o acotadas según las necesidades del equipo de analítica de datos.

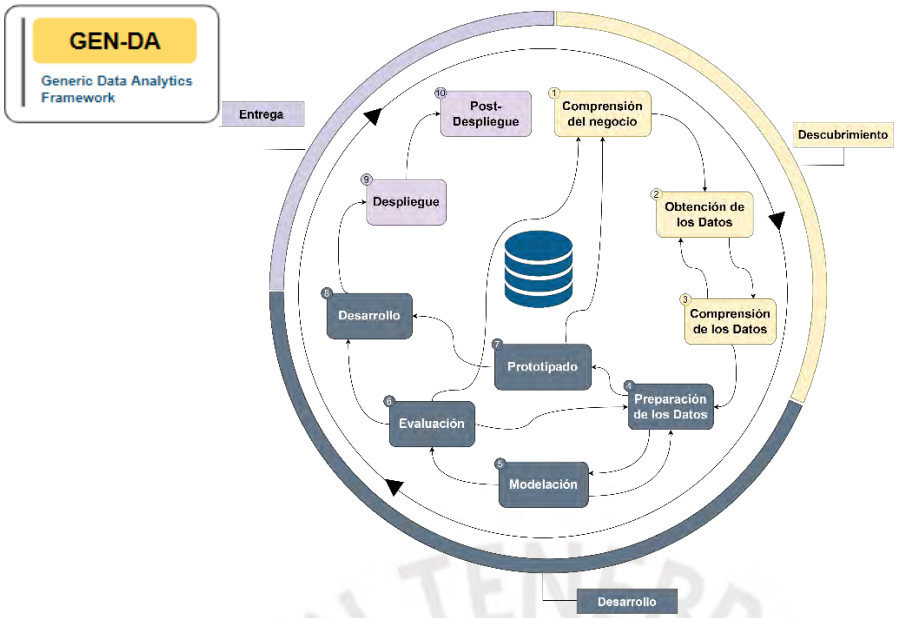

Figura 2. Diagrama de Fases del Marco de trabajo GEN-DA (*Generic Data Analytics Framework*). Elaboración propia.

La fase 1, conocida como ***comprensión del negocio***, es una extensión de la fase 1 de CRISP-DM tomado de la revisión sistemática de la literatura. Se incorporaron elementos lógicos enfocados en encontrar la factibilidad del proyecto, un plan de proyecto claro y en realizar un plan de gestión de riesgos adecuado debido al grado de incertidumbre que poseen este tipo de proyectos. Posterior a encontrar los objetivos de analítica de datos y previo a realizar el plan de proyecto, se debe definir métricas de impacto claras y entendibles. La métrica de impacto depende de la naturaleza del negocio, ya que algunas empresas, por ejemplo, buscarán tener un impacto financiero mientras que otras empresas sin fines de lucro buscarán obtener un impacto social o ambiental. Con la métrica de impacto definida se realiza un plan de análisis de impacto para dar seguimiento a esta métrica posteriormente. Luego de haber definido el plan de análisis de impacto, se realiza el plan de proyecto, para lo cual se recomienda brindar mayor detalle respecto a los recursos, responsables y tiempos. Finalmente, se realiza un plan de gestión de riesgos el cual deberá contener todos los riesgos detectados en el proyecto y un plan claro de qué acción tomar ante la materialización de alguno de los riesgos. Es posible que existan riesgos no detectados por lo que deberá definirse un plan de contingencia claro para dichos casos.

La fase 2, denominada como ***obtención de datos***, es una nueva fase añadida al modelo lógico que contempla todos los pasos que se deben seguir para lograr la adquisición de los datos para el proyecto. La primera actividad consiste en iniciar la gestión para adquisición de los datos. En esta actividad se solicita al área o equipo encargado del gobierno de datos el acceso o la adquisición de uno o más conjuntos de datos. En el proceso de adquisición se suele intercambiar información sobre estos orígenes de datos y se definen finalmente los esquemas y datos que son necesarios para el desarrollo del proyecto. Finalmente, en caso de que la organización en cuestión posea procesos de gobierno de datos, se suele solicitar los lineamientos aplicables al proyecto y los orígenes de datos solicitados. Es posible además recibir políticas de retención y/o destrucción de datos. Estas políticas pueden provenir por decisión del equipo de gobierno de datos o por orden de los clientes o usuarios finales debido a cláusulas contractuales.

La fase 3, denominada como ***comprensión de los datos***, está basada en el modelo de ciclo de vida base CRISP-DM tomado de la revisión sistemática de la literatura. La cual contempla las mismas actividades que el modelo de ciclo de vida base. Es decir, recolectar los datos iniciales, realizar una

descripción de los datos recolectados, realizar la actividad de exploración de los datos y finalmente realizar la validación de la calidad de los datos.

La fase 4, denominada ***preparación de los datos,*** es una extensión y modificación del modelo lógico de CRISP-DM tomado de la revisión sistemática de la literatura. Es decir, se modificó el orden de las tareas, se modificó la descripción de las tareas y se añadieron tareas nuevas. En primer lugar, se lleva a cabo la integración de los datos de diversos orígenes para obtener una sola tabla consolidada. En segundo lugar, se selecciona los datos más relevantes para el modelo analítico, a esta tarea también se le conoce como selección de características. Esta puede ser en base a una técnica como, por ejemplo, arboles de decisión. En tercer lugar, se realiza la limpieza de datos, la cual es una de las tareas que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc. En cuarto lugar, se realiza la transformación de los datos, la cual incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros a partir de agregaciones o transformación de valores para atributos existentes. En quinto lugar, se realiza la generación de datos, la cual involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes y creación de nuevos registros para el balanceo del conjunto de datos. Finalmente, se realiza la tarea de formateo de los datos, la cual consiste, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de Analítica de datos en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.). La fase de preparación de datos está relacionada a la fase de modelación, ya que los datos serán preprocesados de acuerdo con la técnica de modelado elegida. Es por ello que la fase de preparación de datos y la de modelación interactúan constantemente. También está relacionada a la fase de evaluación debido a que puede solicitarse mejorar el modelo analítico como siguientes pasos producto de la evaluación del modelo.

La fase 5, denominada ***modelación***, es también una adaptación y extensión del modelo lógico de CRISP-DM tomado de la revisión sistemática de la literatura y consiste en seleccionar las técnicas de modelación, generar plan de pruebas, construir los modelos analíticos y evaluar los modelos construidos. La cual contempla las mismas tareas que el modelo de ciclo de vida base, con la consideración de que se debe contemplar la construcción de uno o más modelos. Adicionalmente, se modifica una actividad a consecuencia del juicio de expertos y del catálogo de falencias. Ambas actividades están muy relacionadas a la tarea de plan de pruebas. Ya que se pone especial atención a la particularidad de que en ocasiones es importante definir una métrica de calidad ad-hoc al negocio, que puedan ser la combinación de una o más métricas de calidad, debido a que las métricas de calidad conocidas en la literatura o por el negocio por sí solas no llegan a ser suficientes para una correcta validación del modelo.

La fase 6, denominada ***evaluación***, está basada en la fase de evaluación del modelo lógico de CRISP-DM tomado de la revisión sistemática de la literatura. Se proponen cambios en la definición de las tareas respecto a los componentes originales de CRISP-DM. Por ejemplo, en el proceso de revisión se pone bastante énfasis en la preparación de una presentación que resuma el proceso de analítica de datos llevado y los resultados del modelo, con ello se suele emplear una línea base que ha sido definida en la fase de Comprensión del negocio para tener una referencia de la eficiencia del modelo. Así mismo, se recomienda emplear técnicas para la presentación de alto impacto como la técnica de Story Telling.

La fase 7, denominada ***prototipado*** es una nueva fase propuesta en base al catálogo de falencias. Esta fase consiste en realizar un prototipo de la solución analítica; es decir, una simulación que no es el producto de software final pero que será proporcionado al usuario final con el fin de obtener conclusiones relevantes para el proyecto. En esta fase el equipo de analítica de datos puede concluir si el proyecto es importante y fácil de usar por los usuarios finales. Así mismo, es posible determinar, en base a un criterio de aceptación, si el proyecto debe ser evaluado nuevamente con los objetivos de la organización; es decir, debe ser enviado a la fase inicial o fase de comprensión del negocio. Para llevar a cabo esta fase, se realizan 4 actividades. En primer lugar, es importante obtener datos de prueba, si es posible, obtener datos reales con lo cual se define una prueba con usuarios en la cual se puede emplear una técnica o metodología de evaluación de usabilidad. En segundo lugar, se diseña y construye un prototipo de la solución analítica para desarrollar las pruebas. En tercer lugar, con el prototipo construido y el plan de pruebas y se ejecuta la prueba. Finalmente, se realiza un reporte de resultados de la prueba. En esta instancia, se suele evaluar estos resultados para tomar decisiones sobre los siguientes pasos. Finalmente, en caso de que los resultados sean satisfactorios y se haya alcanzado el criterio de aceptación, se aprovecha la fase de prototipado para diseñar una arquitectura de software que pueda soportar la solución analítica en un ambiente real.

La fase 8, denominada ***desarrollo***, es una nueva fase propuesta en base al catálogo de falencias. Esta fase consiste en construir la solución analítica como producto de software. Una solución analítica está compuesta por el modelo analítico, el software que empaqueta el modelo y su interfaz gráfica con la que el usuario final interactúa (ver figura 3). Para llevar a cabo esta fase, se realizan 4 actividades.

A. Ejecutar lineamiento de gobierno de datos. En caso de que su organización posea procesos de gobernanza de datos, es importante ejecutar y respetar los lineamientos de gobierno de datos aplicables al proyecto de analítica de datos. Es posible que parte de la aplicación de estos lineamientos influya en la construcción de la solución analítica, ya que puede haber políticas de confidencialidad de los datos y de accesos que deben ser implementados a la hora del diseño de la arquitectura de software y la arquitectura de datos.
B. *Construir arquitectura de software*. En el caso de que se tome la decisión de desarrollar una aplicación web o móvil se requiere diseñar y construir una arquitectura de software. Esta arquitectura de software se diseña en base al prototipo, al modelo analítico y a los datos obtenidos en las fases anteriores. Una tarea importante es la elección y despliegue de los componentes de arquitectura y de la plataforma de donde se aprovisionan estos componentes. Por ejemplo, una arquitectura aprovisionada en la nube o una arquitectura aprovisionada en un ambiente on-premise por políticas internas de la organización. En caso de seleccionar una herramienta BI como visualizador de datos, de igual forma es importante elegir el motor de procesamiento del modelo analítico y el sistema de almacenamiento donde se gestionará los datos para el modelo y los reportes o *dashboards*. Esta actividad está muy relacionada a la construcción de la arquitectura de datos, especialmente en la elección del sistema de almacenamiento, ya que se debe analizar los datos a almacenar y sus metadatos.
C. *Construir arquitectura de datos.* Esta actividad consiste en diseñar y construir la arquitectura de datos de la solución analítica. A diferencia de una arquitectura de software, una arquitectura de datos describe cómo se gestionan los datos, desde la recopilación hasta la transformación, la distribución y el consumo. Por lo tanto, establece el plan para los datos y la forma en que fluyen a través de los sistemas de almacenamiento de datos. Existen diversos marcos de trabajo para el desarrollo de una arquitectura de datos. Entre los marcos de trabajo más conocidos se encuentran: DAMA-DMBOK 2 (Technics Publications, 2017), TOGAF (Bhupesh et al., s. f.) y Zachman Framework for Enterprise Architecture (Inmon et al., 1997). Estos marcos de trabajo proporcionan una guía de cómo se elabora una correcta arquitectura de datos para la solución analítica en base a diversos criterios como el tipo de datos, volumen, rapidez, seguridad, entre otros.

D. *Construir la solución analítica*. Esta actividad consiste en construir la solución analítica. La solución analítica está compuesta por el modelo analítico que fue creado en la fase de modelación, los pipelines de datos y la interfaz gráfica de usuario donde se interactúa con el modelo analítico y se visualizan los resultados.
Los pipelines de datos representan un conjunto de pasos o fases involucradas en un proceso de movimiento o procesamiento de datos. Orientado a soluciones analíticas, estos contienen las tareas para alimentar el modelo analítico, transformar los datos y almacenar sus resultados de forma persistente en sistemas de almacenamiento. Los pipelines de datos pueden ser clasificados según el orden de procesamiento y el tipo de procesamiento. Por ejemplo, según el orden de procesamiento, un pipeline de datos puede ser un ETL, ELT o un EL. Siendo la diferencia entre ellos el orden en que la extracción, transformación y carga de datos se da. Por otro lado, según el tipo de procesamiento, los pipelines de datos pueden ser de procesamiento en tiempo real o streaming y procesamiento por lotes o batch. Siendo la diferencia entre estos tipos de pipelines, la frecuencia en la que se procesan los datos y la latencia en la que los datos están disponibles en los sistemas de almacenamiento. El modelo analítico, creado previamente en la fase de Modelación, es empaquetado y empleado dentro de los pipelines de datos para poder obtener los resultados deseados según el proyecto. Los modelos analíticos entrenados en entornos de experimentación suelen guardarse en archivos binarios que contienen la información de los pesos e hiperparámetros para poder ser reutilizados en el entorno real. Los formatos más populares en los que los modelos analíticos son guardados son el formato pickle y h5, ya que librerías conocidas de Python emplean estos formatos para la reutilización de modelos. La aplicación o interfaz gráfica con la que el usuario final interactúa suele ser un reporte (dashboard) diseñado en una herramienta BI o una aplicación de software a la medida o un software como servicio (SaaS). La decisión de cómo se construirá esta componente de visualización se toma en base a las necesidades del negocio, el presupuesto del proyecto, los plazos de entrega y del conocimiento del equipo de DA. Por ejemplo, las herramientas de visualización BI permiten visualizar los datos en un formato de reporte ejecutivo empresarial para poder entender de una forma resumida y profesional los resultados de la solución analítica.

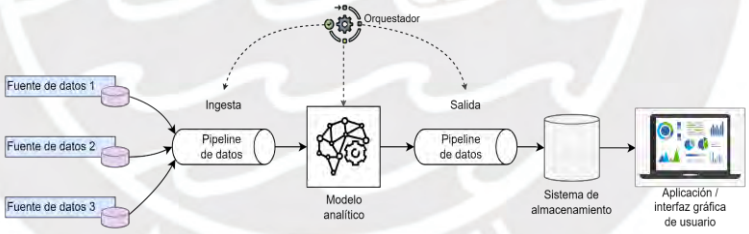

Figura 3. Diagrama de una solución analítica. Elaboración propia.

E. *Orquestar pipelines de datos*. Esta actividad consiste en realizar el orquestamiento de los pipelines de datos de la solución analítica, con el fin de que los datos puedan fluir con una frecuencia y rapidez según las necesidades del negocio. El orquestamiento de los datos permite dar una secuencialidad, paralelismo y/o dependencia a los pipelines y elementos que componen el pipeline. Así mismo, definir las eventualidades y/o dependencias para que los diferentes pipelines se ejecuten (triggers).

F. *Plan de despliegue*. Para desplegar la solución analítica en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su despliegue. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior ejecución. En el plan de despliegue es recomendable definir un periodo de marcha blanca para poder probar y monitorear el funcionamiento de la solución analítica con los usuarios finales.

La fase 9, denominada ***despliegue***, es también una extensión del modelo lógico de CRISP-DM tomado de la revisión sistemática de la literatura. La cual contempla las mismas actividades que el modelo de ciclo de vida base y se le añade una actividad a raíz de la revisión sistemática de la literatura. Esta fase consiste en 5 tareas:

A. *Despliegue*. En esta tarea se despliega la solución analítica siguiendo el plan de despliegue definido en la fase anterior.
B. *Plan de Monitoreo y Mantenimiento*. Si la solución analítica es desplegada en el dominio del problema como parte de la rutina del proceso de negocio, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre la solución analítica. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si la solución analítica está siendo utilizada apropiadamente.
C. *Definir modelo de ciclo de vida de la solución analítica.* Esta actividad consiste en definir un modelo de ciclo de vida de la solución analítica a ser contemplada después de haberla liberado a los usuarios finales. Este modelo de ciclo de vida contiene información de las fases que atravesará la solución analítica una vez haya sido desplegada. Así mismo, deberá contener información sobre:
    - ***Plan de aprendizaje algorítmico***: Se debe realizar un plan de cómo el modelo aprenderá a lo largo del tiempo. Es posible definir criterios por los cuales tomar la decisión de reentrenar el modelo analítico. Por ejemplo, se decide reentrenar debido a que se ha obtenido una cierta cantidad de datos, por haber alcanzado un periodo de vigencia, por haber logrado superar la métrica de calidad del modelo con los nuevos datos e hiperparámetros, entre otros.
    - ***Plan de Gestión de los datos del modelo***: Los datos del modelo son importantes ya que en base a ellos el modelo ha podido aprender y puede ser ejecutado para obtener resultados. Por lo tanto, es importante que los equipos de analítica de datos definan un plan de Gestión de los datos para el uso del modelo y sobre los datos que genera el modelo por intermedio del usuario final.
    - ***Mantenimiento del modelo***: El modelo analítico puede tener parámetros de configuración que deben ser ajustados a lo largo del ciclo de vida para poder obtener mejores resultados. Por ello, es importante definir un plan de cómo llevar a cabo este mantenimiento con el fin de asegurar la calidad de los resultados.
    - ***Responsables de llevar a cabo las acciones***: Se debe realizar un plan de acción claro ante cada una de las fases del ciclo de vida y los responsables por actividad.
D. *Informe Final.* Es la conclusión del proyecto de DA realizado. Dependiendo del plan de despliegue, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.
E. *Revisión del proyecto*. En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar

La fase 10, denominada ***post-despliegue***, es la última fase de GEN-DA. Esta es una nueva fase propuesta en base a los resultados de la revisión sistemática de la literatura. Esta fase consiste en ejecutar todos los planes de acción definidos para soportar la implantación de la solución analítica en un entorno real con usuarios finales. La fase de post-despliegue consta de 4 actividades las cuales son:

A. ***Iniciar el modelo de ciclo de vida del modelo***. Consiste en poner en marcha el ciclo de vida y se considera el plan de aprendizaje algorítmico definido en la fase de despliegue. En esta tarea

se suele aplicar los criterios para monitorear, mantener y reentrenar los modelos analíticos en producción.

B. ***Versionar artefactos resultantes.*** Esta tarea está muy relacionada al plan de aprendizaje algorítmico y consiste en versionar los artefactos resultantes del reentrenamiento de modelos, ya que en un proyecto real se pueden realizar un conjunto de reentrenamientos del modelo y es importante poder llevar un control de la evolución de los modelos por técnica, por grupo de hiperparámetros y por conjunto de datos pre-procesados que fueron empleados. Esto brinda una mayor claridad y soporte a la selección del modelo óptimo y los criterios de su elección.

C. ***Ejecutar plan de monitoreo y mantenimiento***. Consiste en ejecutar el plan de monitoreo y mantenimiento definido en la fase de despliegue. Es importante dar un monitoreo de la solución analítica con respecto a su funcionamiento y uso por parte de los usuarios. Mayormente se emplean algunas herramientas de observabilidad para este fin. Así mismo, es importante seguir el esquema definido para dar mantenimiento y soporte a la solución analítica con esfuerzos, fechas y tiempos claros. Dependiendo de las necesidades se puede haber definido acuerdos de servicio según la capacidad del equipo de soporte.

D. ***Reportes de seguimiento de indicadores de impacto***. Esta actividad consiste en realizar un seguimiento a los indicadores de impacto del negocio que viene teniendo la solución analítica en la realidad en comparación a lo planificado en la fase de entendimiento del negocio. Por ejemplo, en caso de medir un impacto financiero es importante obtener un estado real de la monetización de la solución analítica, ya que, en casos críticos, en los que los resultados reales estuvieron muy por debajo de los esperados, es posible, incluso, tomar la decisión de dar de baja a la solución analítica o empezar un análisis más exhaustivo de la razón de impacto negativo para lograr recuperar los niveles esperados en el impacto financiero.

E. ***Ejecutar plan de gestión de riesgos***. Esta actividad consiste en ejecutar los mecanismos ante la materialización de uno o más riesgos que han sido detectados. Es posible que existan incidentes o acontecimientos que no hayan sido previamente detectados por lo que también será necesario ejecutar un plan de contingencia para dichos casos.

3.2.2. Validación del marco de trabajo

El método empleado para la validación de GEN-DA es el juicio de expertos. Para tal fin, se consideró la participación de 4 expertos (ver tabla 2.0). Según la metodología empleada, basada en DSRM, se ejecutan 3 iteraciones para las fases de desarrollo y validación. Por lo tanto, se validó el marco de trabajo con los expertos en 3 ocasiones como máximo o hasta conseguir su aprobación.

Nº	Perfil	Área de expertise	Rol Actual | Empresarial
1	Data Scientist	Ciencias de datos y proyectos de analítica de datos	Senior Data Scientist
2	Project Manager	Ciencias de datos y proyectos de analítica de datos. Estadística	Docente Universitario | Data Lead
3	Experto	Ciencias de datos y proyectos de analítica de datos.	Docente Universitario | Coordinador de Planeamiento y Análisis de Datos
4	Experto	Ciencias de datos y proyectos de analítica de datos	Docente Universitario | Líder del Grupo de Inteligencia Artificial PUCP - IA-PUCP

Tabla 2. Expertos participantes en la evaluación por Juicio experto – Perfiles y principales características.

Método de Juicio Experto

Los pasos para llevar a cabo la evaluación de **GEN-DA (Generic Data Analytics Framework)** mediante el método de juicio de expertos, se basaron en la Guía para la realización de juicio de expertos de P. Escobar et al. (McGarland et al., 2003). A continuación, se detalla los pasos realizados:

En primer lugar, se definió el objetivo del juicio de expertos con el fin de tener clara la finalidad de llevar a cabo esta validación, en este caso el objetivo es validar la suficiencia y la usabilidad de **GEN-DA**. En segundo lugar, se seleccionó y contactó a 4 expertos con conocimientos en el área de analítica de datos, específicamente con experiencia en la participación o gestión de proyectos de analítica de datos empleando marcos de trabajo. Esta cantidad de expertos se consideró adecuada porque supera la cantidad mínima requerida, el cual es de dos expertos según el autor D. McGarland et al. . En tercer lugar, se realizó una guía a manera de manual de usuario de GEN-DA (ver figura 4) con la descripción de las fases, tareas y relaciones que estas paseen. El manual de usuario de GEN-DA incluye más información sobre los pasos a seguir para ejecutar el ciclo de vida de un proyecto de analítica de datos. Así mismo, se apoya de diagramas con notación UML 2.5.1. En cuarto lugar, se elaboró una rúbrica de evaluación en formato de cuestionario (ver Figura 5) que contiene una lista de ítems separándolos en dos categorías (suficiencia y usabilidad) con respuestas basadas en la escala de Likert y se definió el objetivo de la validación, el cual consiste en usar los resultados obtenidos para el rediseño del marco de trabajo y la obtención de conclusiones. En quinto lugar, se elaboró una plantilla con los elementos antes descritos, lo cual sirve como instrumento de validación para recoger los resultados y observaciones de los jueces. Finalmente, se ejecutó 3 evaluaciones como máximo con cada uno de los jueces. En cada evaluación se le brindó el Manual de usuario de GEN-DA y la plantilla de evaluación por juicio experto como instrumentos de validación. Las observaciones de los jueces sirvieron para realizar ajustes al diseño original del marco de trabajo GEN-DA, con lo cual el marco de trabajo se optimizó hasta llegar a la aprobación de todos los jueces.

Fase 1. Comprensión del negocio o problema

La primera fase de la guía de referencia GEN-DA, denominada fase de comprensión del negocio o problema (véase figura 3.0), consiste en todas las tareas para poder comprender los objetivos y requisitos del proyecto desde la perspectiva de la organización, con el fin de detectar y traducir los problemas y el conocimiento del negocio a objetivos técnicos y un plan de proyecto en el ámbito de la Analítica de datos. Esto debido a que para poder obtener mayor provecho de la Analítica de datos, es importante comprender de manera más profunda el problema que se desea resolver desde una perspectiva empresarial. Una vez logrado este conocimiento, se consigue establecer objetivos en el ámbito de la analítica de datos que estén alineados a los objetivos de la organización.

Figura 3.0. Fase de comprensión del negocio o problema de GEN-DA. Elaboración propia.

La descripción de las principales tareas a desarrollar en esta fase del proceso son:

Determinar los objetivos del negocio. Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Analítica de datos y definir los criterios de éxito.

Los problemas pueden ser diversos como, por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc.

Figura 4. Previsualización del manual de usuario de GEN-DA. Fase 1 – Comprensión del negocio o problema. Elaboración propia.

RÚBRICA DE LA VALIDACIÓN POR JUICIO DE EXPERTOS

Objetivo del Juicio de expertos: Validar la suficiencia y la usabilidad de GEN-DA, Marco de trabajo para Desarrollo de proyectos de Analítica de Datos.

Objetivo de la validación: Usar los resultados obtenidos de la validación para la obtención de conclusiones. Se usarán valores promedio o representativos obtenidos a partir de la calificación de los revisores.

Tenga en cuenta al momento de responder las siguientes afirmaciones:

1=Totalmente en desacuerdo	2= En desacuerdo	3= Ni de acuerdo ni en desacuerdo	4= De acuerdo	5= Totalmente de acuerdo

Suficiencia

En esta sección se evalúa la **suficiencia** de **GEN-DA**, abordaje a las falencias de otros marcos de trabajo existentes y su completitud frente a problemas de analítica de datos en las empresas.

N°	Pregunta	1	2	3	4	5
1	Cubre y soluciona falencias (véase Anexo 2) del proceso de ciclo de vida de la analítica de datos.					X
2	Cubre y soluciona los diferentes tipos de proyectos de analítica de datos que usted ha afrontado en su experiencia laboral.					X
...	...					

Observaciones:

_____
_____
_____

Usabilidad

En esta sección se evalúa la **usabilidad** de **GEN-DA**, abordaje a la facilidad de uso y factibilidad de aplicación en el entorno laboral.

N°	Pregunta	1	2	3	4	5
1	Excluye **especificaciones** de tecnologías, plataformas, metodologías, métodos o resultados relacionados, pero sí brinda recomendaciones.					X
2	El marco de trabajo permite añadir y remover elementos. Así mismo, permite extender y transformar elementos.					X
...	....					

Observaciones:

_____
_____
_____

Figura 5. Cuestionario plantilla para la validación por juicio de expertos. El cuestionario completo cuenta con 5 ítems en la categoría de Suficiencia y Usabilidad Elaboración propia.

Resultados del Juicio Experto

Se recabaron un total de 12 evaluaciones; es decir, 4 evaluaciones en cada ronda. Donde se recogió los puntajes de 5 ítems sobre la suficiencia y 5 ítems sobre la usabilidad de GEN-DA basado en la escala de Likert del 1 al 5, donde 1 es la percepción más negativa y 5 es la más positiva. Las percepciones promedio por evaluación y su desviación estándar se muestran en la figura 6.

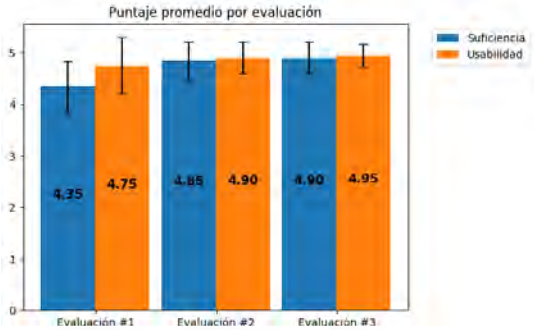

Figura 6. Resultados de la validación por juicio de expertos. Todas las respuestas están dadas en la escala de Likert del 1 al 5, que se basaban entre fuertemente en desacuerdo (1) y totalmente de acuerdo (5). Las barras indican la media (m) y la desviación estándar (SD) para cada escala. Elaboración propia.

Así mismo, se recabó las puntuaciones promedio por ítem por evaluación y categoría. Los resultados se muestran en la figura 7 y figura 8 para las categorías de Suficiencia y Usabilidad de GEN-DA respectivamente. Podemos ver que los puntajes son altos y han ido de forma ascendente debido a que en cada evaluación se absolvió las observaciones de la evaluación anterior.

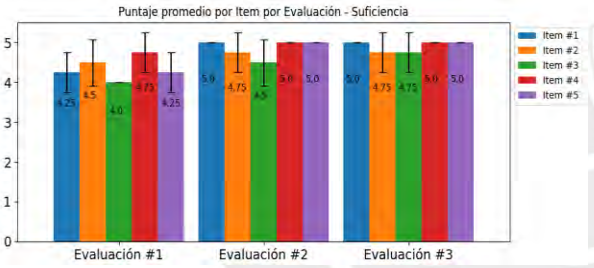

Figura 7. Resultados de la validación por juicio de expertos por ítem por Evaluación para la **categoría de Suficiencia**. Todas las respuestas están dadas en la escala de Likert del 1 al 5, que se basaban entre fuertemente en desacuerdo (1) y totalmente de acuerdo (5). Las barras indican la media (m) y la desviación estándar (SD) para cada escala. Elaboración propia.

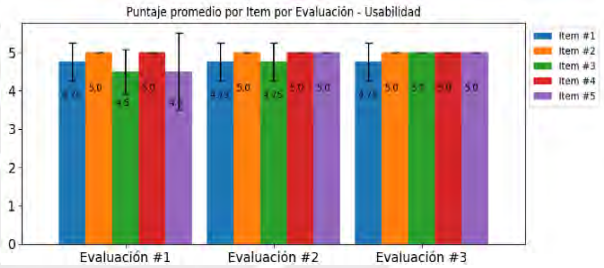

Figura 8. Resultados de la validación por juicio de expertos por ítem por Evaluación para la **categoría de Usabilidad**. Todas las respuestas están dadas en la escala de Likert del 1 al 5, que se basaban entre fuertemente en desacuerdo (1) y totalmente de acuerdo (5). Las barras indican la media (m) y la desviación estándar (SD) para cada escala. Elaboración propia.

Adicionalmente a las puntuaciones según el cuestionario, se solicitó a los expertos dejar observaciones lo más detalladas posibles. Esto con la finalidad de apoyar a los ajustes en el diseño del marco de trabajo después de cada evaluación. Las cantidades de observaciones fueron en descenso debido a que en cada evaluación se absolvió las observaciones de la evaluación anterior (ver figura 9).

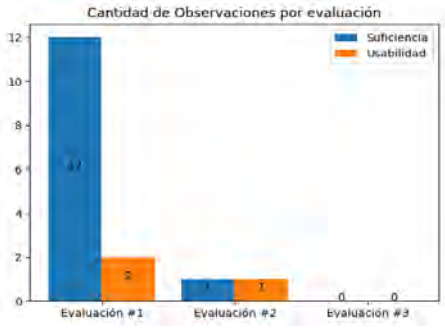

Figura 9. Cantidad de observaciones por evaluación y categoría. Elaboración propia.

4. Conclusiones y trabajos futuros

Esta articulo ha presentado el diseño y evaluación de un marco de trabajo para el desarrollo de proyectos de analítica de datos en el dominio empresarial. Con la aplicación de la metodología de investigación en ciencias del diseño (DSRM) y el apoyo de 4 expertos con experiencia en el área de ciencias de datos y proyectos de analítica de datos, se desarrolló y propuso GEN-DA (Generic Data Analytics Framework) o marco de trabajo genérico para el desarrollo de proyectos de analítica de datos. GEN-DA fue diseñado para abordar las falencias detectadas en la literatura y bajo la experiencia de los participantes de este proyecto.

Se desarrolló un prototipo y se llevaron a cabo evaluaciones con los expertos para realizar un conjunto de ajustes al diseño de GEN-DA. Los resultados obtenidos de las evaluaciones realizadas a GEN-DA fueron muy alentadores, ya que concluyeron con la aprobación unánime de los expertos evidenciado en los resultados de la validación por el método de juicio experto. Por un lado, como se puede observar en la figura 6, los promedios de las percepciones de los expertos cerraron en la última evaluación con 4.90 para la categoría de Suficiencia, lo cual nos permite concluir cualitativamente que el marco de trabajo propuesto GEN-DA aborda las falencias enumeradas en el catálogo de falencias (ver tabla 1) y es completo frente a diversos problemas de analítica de datos experimentados en las empresas. Por otro lado, se puede apreciar que los promedios de las percepciones de los expertos cerraron con 4.95 para la categoría de Usabilidad, lo cual nos permite concluir cualitativamente que el marco de trabajo propuesto GEN-DA es fácil de usar y factible de aplicar en diversos entornos laborales.

Las observaciones recabadas fueron diversas y ayudaron para realizar ajustes al marco de trabajo propuesto con lo que se consiguió la alta aceptación de los expertos. Estas observaciones se dividieron en observaciones asociadas a la suficiencia y a la usabilidad. Las observaciones más destacadas se presentan a continuación:

Entre las observaciones más destacadas, en términos de suficiencia del marco propuesto, se menciona, para la fase de comprensión del negocio, la importancia de considerar un buen plan de proyecto que detalle los recursos, responsables y tiempos en cada fase. Así mismo, una estimación de costos de personas y de tecnología. Por otro lado, los expertos mencionan que, en la fase de comprensión del negocio o problema, es importante identificar si ya existe una lógica que resuelve el caso de uso o problema en cuestión. Esto permite definir una línea base contra qué comparar la solución analítica deseada y diseñar indicadores de seguimiento más adecuados. Finalmente, con respecto a la fase de modelación, se afirma la importancia de incluir más ejemplos de indicadores de calidad, ya que los diferentes tipos de proyectos podrían no verse reflejados. Con lo cual se procedió a brindar más ejemplos de indicadores para poder dar un mayor soporte a la definición de métricas de calidad. Se puede concluir que estas observaciones permitieron dar mayor abordaje a las falencias encontradas en la literatura y a evidencias completitud al momento de abordar diversos casos de uso en diversos modelos de negocio,

incluso permite la adaptación de GEN-DA con tecnologías disruptivas y la aparición de eventuales nuevos modelos de negocio.

Las observaciones más destacadas en términos de usabilidad del marco propuesto se basan en ajustes al diseño del diagrama de actividades inicial y al evitar el empleo de conceptos muy específicos en la guía de GEN-DA. Es decir, se resaltó que, en fases más sofisticadas como la fase de desarrollo, específicamente en la tarea de construir una arquitectura de datos y arquitectura de software, no es recomendable usar conceptos muy específicos a las tecnologías empleadas en la arquitectura de datos y de software, por lo que en los ajustes de diseño se generalizó los conceptos tratados. Así mismo, se recomendó nuevos órdenes en las tareas para la fase de preparación de datos, nuevas conexiones entre fases y una agrupación para las 10 fases en tres categorías más simples que involucra el marco de trabajo propuesto, con el fin de facilitar la familiarización y el entendimiento. Se puede concluir que estas observaciones permitieron que el marco sea más fácil de usar y se pueda aplicar en un entorno real, según las percepciones de los expertos.

Se recomienda para futuras investigaciones reforzar la evaluación de la suficiencia y usabilidad mediante la elaboración de un diseño experimental con y sin grupo de control para poder evaluar a GEN-DA en un proyecto de analítica de datos real. Es posible comparar resultados entre GEN-DA y otros marcos de trabajo como CRISP-DM. Con ello poder evidenciar con mayor exactitud el éxito del marco de trabajo propuesto GEN-DA en un entorno real. Finalmente, otra propuesta de trabajo futuro podría estar asociada con tener un alcance más amplio que abarque temas no descritos en la lista de falencias detectadas en la literatura como, por ejemplo, la Ética, la protección y privacidad de los datos. Así mismo, con abarcar conceptos más tecnológicos como MLOps (Machine Learning Operations) dentro de la fase de Desarrollo (Kreuzberger et al., 2022). Que, si bien el marco de trabajo brinda soporte a los conceptos de integración, despliegue y entrega continua, este no lo desarrolla a manera más técnica otras buenas prácticas de este paradigma.

5. Referencias

Azevedo, A., & Santos, M. F. (2008). *KDD, semma and CRISP-DM: A parallel overview Business Intelligence-Implantation on Federal Institute of Triângulo Mineiro (IFTM) System View project Book Editing: Handbook of Research on E-Assessment in Higher Education View project KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. https://www.researchgate.net/publication/220969845

Bhupesh, M., Smolander Supervisor, K., & Erja Mustonen-Ollila, A. (s. f.). *BUILDING IT ARCHITECTURE FOR DEVELOPING NATIONS USING TOGAF*.

Charters, S., & Kitchenham, B. (2007). Guidelines for performing Systematic Literature Reviews in software engineering. *Int. Conf. Soft. Engin.*, *45*(4ve), 1051. https://doi.org/10.1145/1134285.1134500

Data Science Process Alliance. (2020). *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects - Data Science Process Alliance*. https://www.datascience-pm.com/crisp-dm-still-most-popular/

KDnuggets. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects - KDnuggets*. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Technics Publications. (2017). *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*.

Davenport, T. H., & Harris, J. G. (2017). *Competing on Analytics : The new science of winning*.

Escobar-Pérez, J., & Cuervo-Martínez, Á. (2008). Validez De Contenido Y Juicio De Expertos: Una Aproximación a Su Utilización. *Avances en Medición*, *6*, 27-36.

McKinsey & Company. (2021). *Global survey: The state of AI in 2021 | McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

Gudivada, V. N. (2017). Data Analytics: Fundamentals. En *Data Analytics for Intelligent Transportation Systems* (pp. 31-67). Elsevier Inc. https://doi.org/10.1016/B978-0-12-809715-1.00002-X

Inmon, W., Zachman, J., & Geiger, J. (1997). *Data Stores, Data Warehousing and the Zachman Framework: Managing Enterprise Knowledge*.

Kozjek, D., Vrabič, R., Rihtaršič, B., Lavrač, N., & Butala, P. (2020). Advancing manufacturing systems with big-data analytics: A conceptual framework. *International Journal of Computer Integrated Manufacturing*, *33*(2), 169-188. https://doi.org/10.1080/0951192X.2020.1718765

Kreuzberger, D., Kühl, N., & Hirschl, S. (2022). *Machine Learning Operations (MLOps): Overview, Definition, and Architecture*.

McGarland, D., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77. https://doi.org/10.2753/MIS0742-1222240302

Plotnikova, V., Dumas, M., Nolte, A., & Milani, F. (2022). Designing a data mining process for the financial services domain. *Journal of Business Analytics*. https://doi.org/10.1080/2573234X.2022.2088412

Pries-Heje, J. R., Baskerville ; Venable, J. ;, Pries-Heje, J. R., & Baskerville, J. (2017). Choosing a Design Science Research Methodology. En *Australia Choosing a Design Science Research Methodology* (Vol. 2). APA.

Qadadeh, W., & Abdallah, S. (2020). *An Improved Agile Framework For Implementing Data Science Initiatives in the Government*.

Rolllins, J. B. (2015). Metodología Fundamental para la Ciencia de Datos. *IBM Analytics*. https://www.ibm.com/downloads/cas/6RZMKDN8