

Prof. **Edgar Serna M.** (Ed.)

INVESTIGACIÓN FORMATIVA EN INGENIERÍA

Cuarta Edición

Medellín - Antioquia
2020



Editorial IAI

Evaluación comparativa de las metodologías Team Data Science Process TDSP y Analytics Solutions Unified Method for Data Mining ASUM-DM desde la perspectiva de la ciencia de datos

Giuliana Fois
Gustavo Andrés Agüero Crovella
Paola Verónica Britos
Universidad Nacional de Río Negro
Argentina

En proyectos de ciencia de datos es importante planificar para abordar el problema y llegar a una solución. Para ello se utilizan metodologías que organizan las actividades. En este trabajo de investigación se realizará un estudio comparativo de las metodologías Team Data Science Process TDSP, o Proceso del Equipo de Ciencia de Datos, y Analytics Solutions Unified Method for Data Mining ASUM-DM, o Método unificado de soluciones de análisis para la minería de datos, donde se evaluarán las ventajas y desventajas de su utilización.

1. INTRODUCCIÓN

La ciencia de datos es una disciplina que en los últimos años ha tenido un gran crecimiento. Tanto las personas como los sensores y dispositivos de comunicación utilizados, han ido generando grandes volúmenes de datos. La Ciencia de Datos (o Data Science) surge a partir de la insuficiencia de métodos y análisis clásicos para procesar los grandes volúmenes de datos. Si bien el término mencionado es relativamente nuevo, los orígenes se remontan a la década de 1960 - 1970 donde Turkey [1, 2] ya se preocupaba por los grandes conjuntos de datos y proponía poner énfasis en el análisis de datos para sugerir hipótesis a probar. Por otra parte, la ciencia de datos se establece como multidisciplinar ya que involucra la estadística, la matemática y la ingeniería de datos entre otros campos, para desarrollar procesos, técnicas y sistemas que extraigan conocimiento de grandes volúmenes de fuentes de información diversas y complejas. Fayad, et al. [3] en 1996 definieron el término Descubrimiento de Conocimiento en Bases de Datos (o Knowledge Discovery in Databases, KDD) como un proceso que identifica patrones válidos, novedosos, potencialmente útiles y comprensibles de una manera no trivial.

La Minería de Datos (o *Data Mining*) es una parte central dentro del proceso de descubrimiento en bases de datos, lo cual requiere que se ejecuten pasos que definan procesos y tecnologías que transformen en conocimiento la información. A partir de ello, surge la explotación de información que es una sub-disciplina que aporta a la inteligencia de negocio las herramientas para la transformación de información en conocimiento y define la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información [4], para eso, define un grupo de tareas relacionadas lógicamente que, a partir de un conjunto de información con un cierto grado de valor para la organización, se ejecuta para lograr otro, con un grado de valor mayor que el inicial [5].

Britos, en 2008 [6] propone como procesos de explotación de la información lo siguiente: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos (atributos importantes para el entorno de negocio que se aplica), descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos. Cada proceso tiene asociadas técnicas de minería de datos para descubrir patrones a través de su aplicación. En la ciencia de datos, planificar cómo abordar el problema para llegar a una solución es de suma importancia. Para ello se utilizan metodologías que nos ayudan a organizar y planificar las actividades. En este trabajo de investigación se realizará un estudio comparativo de las metodologías TDSP [7] y ASUM-DM [8], donde se evaluarán las ventajas y desventajas de su utilización.

Estas metodologías fueron propuestas por dos de las empresas más renombradas en la industria del software durante décadas alrededor del mundo: Microsoft e IBM. Ambas han tenido protagonismo en distintos hitos de la historia de la informática y en el desarrollo de las grandes áreas de estudio que devinieron de la creciente capacidad de cómputo.

En los últimos tiempos, han surgido metodologías para la Ciencia de Datos que siguen la filosofía ágil que ha adoptado la Ingeniería de Software. En este caso, se trata de un conjunto de técnicas y métodos iterativos e incrementales, apropiados para proyectos que tienen requerimientos cambiantes, que generan poca documentación y que se propone lograr un desarrollo a corto y mediano plazo [9]. Las metodologías ágiles persiguen principios claves [10]: satisfacer al cliente, aprovechar el cambio para obtener una ventaja competitiva con el cliente, hacer entregas de calidad con frecuencia, involucrar al cliente en todo el proceso, organizar el equipo de trabajo, reflexionar cómo ser más eficaz, promover el desarrollo sostenible, mantener conversaciones fluidas entre el equipo de trabajo y la simplicidad son esenciales.

Resulta de interés entender cómo se traslada este concepto a la Ciencia de Datos y cómo se aplica en las metodologías que vamos a analizar. Sánchez y Félix Serrano [11], señalan que, ante la disrupción de tecnologías, el contexto social dinámico, los nuevos modelos de negocio y gestión, debemos ampararnos en el agilísimo, para ello, es vital trabajar colaborativamente, y los proyectos de ciencia de datos no están excluidos.

TDSP [7] fue propuesta por Microsoft en 2017 y es una metodología de ciencia de datos ágil e iterativa para ofrecer soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente, que pretende mejorar el aprendizaje y colaboración del equipo sugiriendo roles que funcionan mejor juntos. Dentro de la misma hay un apartado dedicado al desarrollo ágil de proyectos en Ciencias de Datos. El mismo explica los pasos a seguir para planificar un sprint, agregar ítems de trabajo a uno (características, historias de usuario, tareas o bugs) y crear una plantilla de ítem de trabajo dentro de las etapas del ciclo de vida.

La metodología desarrollada en 2015 por IBM Analytics ASUM-DM [8] es un proceso iterativo para implementar un proyecto de minería de datos o análisis predictivo. ASUM [12] utiliza la implementación de metodologías ágiles y los principios tradicionales de la Ingeniería del Software para lograr sus objetivos y proporcionar un resultado óptimo. Estos principios son:

- El proyecto es evaluado para la aplicación de principios ágiles.
- El alcance del proyecto concuerda con los objetivos de la organización.
- Tanto el personal de negocios como el de tecnología de la información forman parte integral del equipo de implementación del proyecto.
- Los requisitos se aclaran y ajustan a través de un número de sprints iterativos de prototipos.
- Según el número y la prioridad de los requisitos, el cronograma y recursos disponibles, se adopta una implementación por etapas para alcanzar los objetivos.
- Los resultados de la creación de prototipos se comparan con los requisitos totales para evaluar logros y determinar nuevas iteraciones.
- El desarrollo iterativo e incremental se utiliza para finalizar la configuración y construcción.
- Después de pruebas realizadas a lo largo del ciclo de vida del proyecto, se pone en marcha la primera etapa de la solución.
- Las etapas restantes del proyecto siguen la misma dinámica de prototipos de sprints (iterativos e incrementales desarrollados en la primera etapa.
- Los resultados de las etapas implementadas son monitoreados continuamente en la fase Operar y optimizar y si se producen desviaciones, se tiene lugar a nuevas rondas de desarrollo.

2. MÉTODO

Para hacer un estudio comparativo de las metodologías, como primera instancia observamos una serie de atributos que destacan sus rasgos más pronunciados en cuanto a sus características. Según ASUM [8] su comienzo inicia como una extensión del estándar abierto de proceso CRISP-DM (del acrónimo en inglés *Cross Industry Standard Process for Data Mining*) [13], en cambio TDSP [7] no da cuenta de su origen.

Por otro lado, analizamos la rigurosidad de la metodología de IBM, en contraposición de Microsoft que propone una flexibilidad que permite personalizaciones ad hoc para adaptarse a herramientas y técnicas requeridas. La estructura de cada metodología tiene elementos en común y elementos dispares, para los cuales haremos un análisis específico que se pueden ver en las Tablas 1 y 2.

En la Tabla 1 se pueden ver los ejes conceptuales definidos para comparar las metodologías. Con estos se tiene una visión global de las similitudes y las diferencias que hay entre ellas.

Tabla 1. Ejes de las metodologías

| Ítem de comparación | ASUM | TDSP |
|---|-----------------------|--------------------------|
| Origen | Extensión de CRISP-DM | Independiente |
| Flexibilidad | No expresada | Expresada explícitamente |
| Componentes | Fases | Etapas de ciclo de vida |
| | Actividades | Tareas principales |
| | Tareas | Roles |
| | Subtareas | Tareas de roles |
| | Pasos | |
| | Roles | |
| Interacción entre fases | Secuencial | Multidireccional |
| Modalidad | Ágil | Ágil |
| ¿Recomienda herramientas para llevar adelante la metodología? | Sí, SPSS | Sí, Azure |

Se seleccionan dos de los componentes con mayor nivel de abstracción, indicando los elementos que los integran, estos se pueden observar en la Tabla 2.

Tabla 2. Componentes analizados de las metodologías

| Ítem de comparación | ASUM | TDSP |
|---------------------|---|---|
| Fases / Etapas | Analizar. Diseñar. Configurar y Construir. Implementar y/o Desplegar. Operar y Optimizar. | Entendimiento del Negocio. Adquisición de datos y comprensión. Modelado. Despliegue. Aceptación del cliente. |
| Roles | Administrador de aplicaciones del cliente. Cliente patrocinador comercial. Analista de datos del cliente. Administrador de base de datos del cliente. Usuarios del sistema de clave de cliente. Administrador de red del cliente. Gerente de proyecto de cliente. Administrador de seguridad del cliente. Grupos de interés del cliente. Experto en la materia del cliente. Gerente de soporte al cliente. Administrador de herramientas de cliente. Minero de datos / Científico de datos. Arquitecto empresarial. Gerente de proyecto. Gerente de proyectos de SPSS. | Administrador de grupo. Responsable de equipo. Responsable de proyecto. Colaboradores individuales del proyecto. |

La secuencia en que ocurre cada fase o etapa, tiene una disposición diferente: si bien se trata de fases iterativas, en general, ASUM establece un orden secuencial (Figura 1), mientras que TDSP dispone estadios en paralelo con múltiples posibilidades de inter-comunicaciones (Figura 2). Como se expuso anteriormente, ambas metodologías persiguen la modalidad ágil, a pesar de sus diferentes formas de aplicación.

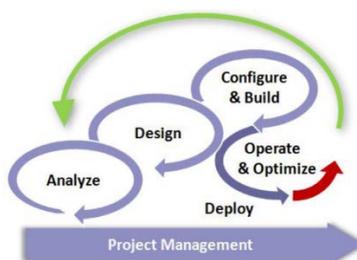


Figura 1. Analytics Solutions Unified Method ASUM

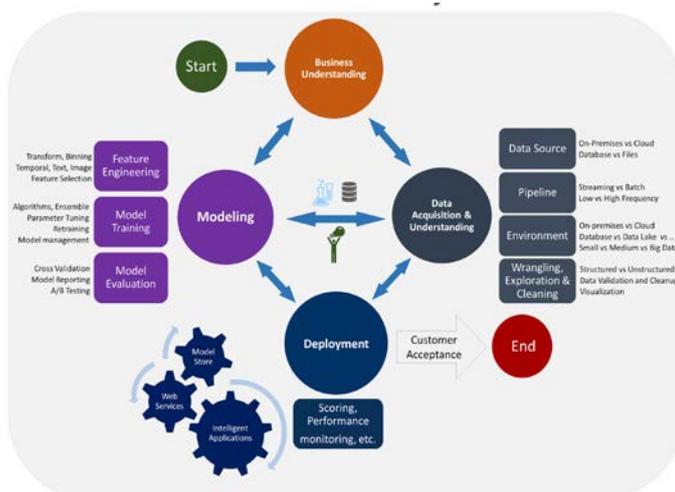


Figura 2. Team Data Science Process TDSP

3. RESULTADOS Y ANÁLISIS

Para hacer un análisis más profundo se comparan diferentes estratos que se corresponden en Fases/Etapas, Actividades o Tareas en la Tabla 3. Al tratarse de componentes de distinta dimensión o alcance, no se trata de una comparación en relación de elementos equivalentes, sino una relación en que un elemento está contenido en el otro. Cabe aclarar que los componentes de menor nivel en ASUM (actividades y tareas) no han sido considerados para este análisis, se prevé a hacerlo posteriormente. Para poder identificar el nivel de desagregación de los distintos componentes se realiza la siguiente notación: 1) en **negrita** las Fases/Etapas, 2) las Actividades de ASUM en correspondencia con las Tareas de TDSP se señalan en *cursiva*, y 3) las tareas que presentan mayor desagregación de la primera metodología mencionada en este párrafo, con letra normal.

Tabla 3. Comparación entre fases/etapas, actividades y tareas

| ASUM | TDSP |
|--|--|
| Fase 1: Analizar | Etapa 1: Entendimiento del Negocio |
| <i>Actividad 1.1 - Prepárese para la implementación.</i> | NO IDENTIFICADA. |
| Tarea 1.1.1 - Enlace con ventas y revisión de detalles del proyecto. | NO IDENTIFICADA. |
| Tarea 1.1.2 - Identificar recursos. | NO IDENTIFICADA. |
| <i>Actividad 1.2 - Realizar evaluación de preparación.</i> | NO IDENTIFICADA. |
| Tarea 1.2.1 - Evaluar la preparación del cliente para la implementación. | NO IDENTIFICADA. |
| <i>Actividad 1.3 - Realizar el lanzamiento del proyecto.</i> | NO IDENTIFICADA. |
| Tarea 1.3.1 - Preparar la plataforma de lanzamiento del proyecto. | NO IDENTIFICADA. |
| Tarea 1.3.2 - Oriente los recursos del proyecto. | NO IDENTIFICADA. |
| Tarea 1.3.3 - Ejecutar lanzamiento de proyecto. | NO IDENTIFICADA. |
| <i>Actividad 1.4 - Comprender el negocio.</i> | <i>Tarea 1.1 - Definición de objetivos.</i> |
| Tarea 1.4.1 - Determinar los Objetivos. | <i>Tarea 1.1 - Definición de objetivos.</i> |
| Tarea 1.4.2 - Evaluar la situación. | NO IDENTIFICADA. |
| Tarea 1.4.3 - Determinar objetivos de minería de datos. | <i>Tarea 1.1 - Definición de objetivos.</i> |
| Tarea 1.4.4 - Crear un plan de proyecto. | NO IDENTIFICADA. |
| Tarea 1.4.5 - Crear un informe de comprensión empresarial. | NO IDENTIFICADA. |
| <i>Actividad 1.5 - Comprender los datos.</i> | Etapa 2: Adquisición de datos y comprensión. |
| Tarea 1.5.1 - Recopilar datos Iniciales. | Etapa 2: Adquisición de datos y comprensión. |
| Tarea 1.5.2 - Describir los datos. | NO IDENTIFICADA. |
| Tarea 1.5.3 - Explorar los datos. | NO IDENTIFICADA. |
| Tarea 1.5.4 - Verificar la calidad los datos. | NO IDENTIFICADA. |
| Tarea 1.5.5 - Crear informe de comprensión de los datos. | NO IDENTIFICADA. |
| Fase 2: Diseñar. | NO IDENTIFICADA. |
| <i>Actividad 2.1 - Diseñar y validar la infraestructura.</i> | NO IDENTIFICADA. |
| Tarea 2.1.1 - Diseño de infraestructura técnica. | NO IDENTIFICADA. |
| Tarea 2.1.2 - Diseño de infraestructura de seguridad. | NO IDENTIFICADA. |
| Fase 3: Configurar y Construir. | NO IDENTIFICADA. |
| <i>Actividad 3.1 - Configurar entornos.</i> | <i>Tarea 2.1: Introducción de los datos en el entorno de análisis de destino.</i> |
| Tarea 3.1.1 - Configurar los entornos en el sitio. | Etapa 3: Modelado. |
| Tarea 3.1.2 - Configurar los entornos en la nube. | <i>Tarea 3.1: Diseño de características.</i> |
| <i>Actividad 3.2 - Preparar los datos.</i> | NO IDENTIFICADA. |
| Tarea 3.2.1 - Seleccionar los datos. | <i>Tarea 2.2 - Exploración de los datos para determinar si su calidad es suficiente para responder a la pregunta.</i> |
| Tarea 3.2.2 - Limpiar los datos. | <i>Tarea 2.3 - Configuración de una canalización de datos para puntuar los datos nuevos o que se actualizan con regularidad.</i> |
| Tarea 3.2.3 - Construir los datos. | <i>Tarea 1.2 - Identifique los orígenes de datos.</i> |
| Tarea 3.2.4 - Integrar los datos. | NO IDENTIFICADA. |
| Tarea 3.2.5 - Formatear los datos. | NO IDENTIFICADA. |
| Tarea 3.2.6 - Crear informe de preparación de datos. | NO IDENTIFICADA. |
| <i>Actividad 3.3 - Construir el modelo.</i> | <i>Tarea 2.1: Introducción de los datos en el entorno de análisis de destino.</i> |
| Tarea 3.3.1 - Seleccionar técnicas de modelado. | Etapa 3: Modelado. |
| Tarea 3.3.2 - Generar diseño de prueba. | <i>Tarea 3.1: Diseño de características.</i> |
| Tarea 3.3.3 - Construir el modelo. | <i>Tarea 3.2: Entrenamiento del modelo.</i> |
| Tarea 3.3.4 - Evaluar el modelo. | <i>Tarea 3.2: Entrenamiento del modelo.</i> |
| <i>Actividad 3.4 - Evaluar el modelo.</i> | NO IDENTIFICADA. |
| Tarea 3.4.1 - Evaluar resultados. | <i>Tarea 3.3 - Determine si el modelo es adecuado para su uso en producción.</i> |
| Tarea 3.4.2 - Proceso de revisión. | NO IDENTIFICADA. |
| Tarea 3.4.3 - Determinar los próximos pasos. | NO IDENTIFICADA. |
| <i>Actividad 3.5 - Realizar transferencia de conocimiento analítico.</i> | <i>Tarea 3.3 - Determine si el modelo es adecuado para su uso en producción.</i> |
| Tarea 3.5.1 - Orientar y transferir conocimiento al equipo analítico del proyecto del cliente. | NO IDENTIFICADA. |
| <i>Actividad 3.6 - Definir enfoque de implementación.</i> | Etapa 4: Despliegue. |
| Tarea 3.6.1 - Crear plan de implementación. | NO IDENTIFICADA. |
| <i>Actividad 3.7 - Diseño de estrategia de prueba operacional.</i> | <i>Tarea 5.1: Validación del sistema.</i> |
| Tarea 3.7.1 - Identificar y acordar planes de prueba. | NO IDENTIFICADA. |
| Tarea 3.7.2 - Crear planes de prueba. | NO IDENTIFICADA. |
| <i>Actividad 3.8 - Validar y probar en entorno de control de calidad.</i> | <i>Tarea 5.1: Validación del sistema.</i> |
| Tarea 3.8.1 - Crear archivos de datos de control de calidad. | NO IDENTIFICADA. |
| Tarea 3.8.2 - Asegúrese de que el entorno de control de calidad esté listo. | NO IDENTIFICADA. |
| Tarea 3.8.3 - Migrar / Restaurar modelo analítico en QA. | NO IDENTIFICADA. |
| Tarea 3.8.4 - Sistema de conducta y pruebas de rendimiento. | NO IDENTIFICADA. |
| Tarea 3.8.5 - Realizar prueba de aceptación del usuario. | Etapa 5: Aceptación del cliente. |

| Fase 4: Implementar/Desplegar. | Etapa 4: Despliegue. |
|---|---|
| <i>Actividad 4.1 - Realizar transferencia de conocimiento operacional.</i> | <i>Tarea 5.2 -Entrega del proyecto.</i> |
| Tarea 4.1.1 Orientar y transferir conocimiento al equipo operativo del proyecto del cliente. | NO IDENTIFICADA. |
| <i>Actividad 4.2 - Prepararse para el mantenimiento continuo.</i> | NO IDENTIFICADA. |
| Tarea 4.2.1 - Establecer un cronograma para el soporte de guardia. | NO IDENTIFICADA. |
| Tarea 4.2.2 - Programar actividades de mantenimiento para el entorno de producción. | NO IDENTIFICADA. |
| Tarea 4.2.3 - Programar actividades de monitoreo para el entorno de producción. | NO IDENTIFICADA. |
| <i>Actividad 4.3 - Implementar la solución.</i> | NO IDENTIFICADA. |
| Tarea 4.3.1 - Crear archivos de datos de producción. | NO IDENTIFICADA. |
| Tarea 4.3.2 - Crear y realizar pruebas de preparación operacional. | NO IDENTIFICADA. |
| Tarea 4.3.3 - Migrar / restaurar el modelo de control de calidad a producción. | NO IDENTIFICADA. |
| <i>Actividad 4.4 - Tránsito al soporte de IBM.</i> | NO IDENTIFICADA. |
| Tarea 4.4.1 - Orientar a los administradores a los recursos de soporte al cliente de IBM. | NO IDENTIFICADA. |
| Tarea 4.4.2 - Entregar para apoyar. | NO IDENTIFICADA. |
| <i>Actividad 4.5 - Lanzamiento.</i> | <i>Tarea 5.2 -Entrega del proyecto.</i> |
| Tarea 4.5.1 - Ir a vivir. | NO IDENTIFICADA. |
| Tarea 4.5.2 - Implementar el plan de lanzamiento de comunicación. | NO IDENTIFICADA. |
| Tarea 4.5.3 - Lanzamiento de revisión. | NO IDENTIFICADA. |
| <i>Actividad 4.6. -Prepararse para el cierre del proyecto.</i> | <i>Tarea 5.2 - Entrega del proyecto.</i> |
| Tarea 4.6.1 - Crear informe final. | NO IDENTIFICADA. |
| Tarea 4.6.2 - Reunión de revisión de conducta. | NO IDENTIFICADA. |
| Tarea 4.6.3 - Seguimiento de las acciones de la reunión de revisión. | NO IDENTIFICADA. |
| Tarea 4.6.4 - Obtenga el consentimiento para cerrar el proyecto. | NO IDENTIFICADA. |
| Tarea 4.6.5 - Obtenga el consentimiento para la iniciación de referencia. | NO IDENTIFICADA. |
| Tarea 4.6.6 - Realizar el evento del equipo de cierre del proyecto. | NO IDENTIFICADA. |
| Tarea 4.6.7 - Realizar una revisión interna del proyecto y las lecciones aprendidas. | NO IDENTIFICADA. |
| Fase 5: Operar y Optimizar. | NO IDENTIFICADA. |
| <i>Actividad 5.1 - Modelo de monitor.</i> | <i>Tarea 3.2: Entrenamiento del modelo.</i> |
| Tarea 5.1.1 - Monitorear continuamente la precisión del modelo y la actualización del modelo. | NO IDENTIFICADA. |
| <i>Actividad 5.2 - Operar, optimizar y mejorar el Sistema.</i> | NO IDENTIFICADA. |
| Tarea 5.2.1 - Programar y operar el sistema. | NO IDENTIFICADA. |
| Tarea 5.2.2 - Monitorear procesos de producción. | NO IDENTIFICADA. |
| Tarea 5.2.3 - Gestionar procesos de mantenimiento. | NO IDENTIFICADA. |
| Tarea 5.2.4 - Gestionar defectos. | NO IDENTIFICADA. |
| Tarea 5.2.5 - Administrar mejoras. | NO IDENTIFICADA. |
| Tarea 5.2.6- Optimizar el sistema. | NO IDENTIFICADA. |
| Tarea 5.2.7 - Mantener la documentación de la aplicación. | NO IDENTIFICADA. |
| Tarea 5.2.8 - Gestionar competencia. | NO IDENTIFICADA. |
| <i>Actividad 5.3 - Comunidad de usuarios de soporte.</i> | NO IDENTIFICADA. |
| Tarea 5.3.1 - Administrar operaciones de la mesa de ayuda. | NO IDENTIFICADA. |
| Tarea 5.3.2 - Administrar problemas. | NO IDENTIFICADA. |
| Tarea 5.3.3 - Administrar recursos de la mesa de ayuda. | NO IDENTIFICADA. |
| Tarea 5.3.4 - Gestionar comunicación. | NO IDENTIFICADA. |
| <i>Actividad 5.4 - Administrar infraestructura,</i> | NO IDENTIFICADA. |
| Tarea 5.4.1 - Administrar la configuración del sistema. | NO IDENTIFICADA. |
| Tarea 5.4.2 - Gestionar capacidad. | NO IDENTIFICADA. |
| Tarea 5.4.3 - Gestionar disponibilidad del sistema. | NO IDENTIFICADA. |
| Tarea 5.4.4 - Administrar lanzamientos. | NO IDENTIFICADA. |
| Tarea 5.4.5 - Administrar seguridad. | NO IDENTIFICADA. |
| <i>Actividad 5.5 - Programa de ciclo de vida del sistema de gobierno.</i> | <i>Tarea 5.2 - Entrega del proyecto.</i> |
| Tarea 5.5.1 - Verificar la realización de beneficios. | NO IDENTIFICADA. |
| Tarea 5.5.2 - Revisar el rendimiento del sistema. | NO IDENTIFICADA. |
| Tarea 5.5.3 - Revisión del plan del ciclo de vida del sistema. | NO IDENTIFICADA. |
| Tarea 5.5.4 - Gestionar calidad. | NO IDENTIFICADA. |

Como se puede observar en la tabla anterior, ASUM tiene componentes que se nombran de la misma manera, pero se diferencian en el nivel de profundidad. Por ejemplo, la *Actividad 3.3 - Construir el modelo* (hace referencia a las múltiples iteraciones que ejecutan los modelos utilizando los parámetros predeterminados y luego ajustan los parámetros o vuelven a la actividad) y Tarea 3.3.3 - Construir el modelo (Ejecutar la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos).

Lo mismo ocurre con la Tarea 3.3.4 - Evaluar el modelo (Determina qué modelo(s) son lo suficientemente precisos o efectivos para ser definitivos) y *Actividad 3.4 - Evaluar el modelo.* (Evaluar los modelos utilizando los criterios de éxito empresarial).

Los componentes citados permiten encontrar puntos comunes dentro de cada metodología, y aquellos en donde las tareas están asignadas en distinto orden o momento de avance del proyecto. Además, resulta visible el nivel de detalle

y desglose de componentes de ASUM, frente a los de TDSP que se enuncian y describen con mayor abstracción. Además de los elementos que determinamos como comparables y semejantes, se identifican estratos que no presentan correspondencia entre las metodologías.

Finalmente, se realizó un análisis de roles de cada metodología, cabe aclarar que las mismas no establecen la cantidad de personas que participan dentro del proyecto, y que un integrante puede ocupar más de un rol, como así también un rol podría ocuparlo un grupo de integrantes. En la Tabla 4 se resumen las diferencias y similitudes de roles mencionadas.

Tabla 4. Comparación entre Roles

| ASUM | TDSP |
|---|--------------------------------------|
| Administrador de aplicaciones del cliente. | |
| Cliente Patrocinador Comercial. | |
| Analista de datos del cliente. | |
| Administrador de base de datos del cliente. | |
| Usuarios del sistema de clave de cliente. | |
| Administrador de red del cliente. | |
| Gerente de proyecto de cliente. | NO TIENEN CORRESPONDENCIA. |
| Administrador de seguridad del cliente. | |
| Grupos de interés del cliente. | |
| Experto en la materia del cliente. | |
| Gerente de soporte al cliente. | |
| Administrador de herramientas de cliente. | |
| Minero de Datos / Científico de Datos. | Colaborador Individual del Proyecto. |
| Arquitecto Empresarial. | NO TIENE CORRESPONDENCIA. |
| Gerente de proyecto. | Responsable de Proyecto. |
| Gerente de proyectos de SPSS. | Responsable de Equipo. |

4. CONCLUSIONES

Como se vió en la Tabla 1, se identifican distintos elementos presentes en las metodologías con características comunes y dispares. En ambos casos se reconoce el carácter ágil con etapas iterativas e incrementales y propósitos semejantes. Sin embargo, estas fases/etapas transcurren y se retroalimentan de distinto modo. También, cada empresa propone herramientas de software para abordar proyectos de Ciencias de Datos.

De acuerdo a la comparación realizada en la Tabla 3, se observa que ASUM es más estructurada y detallista que TDSP, mientras que esta última se presenta flexible y abierta. Además, la fase de diseño de la primera metodología, no se encuentra reflejada en la segunda. En otro orden, todas las tareas de TDSP tienen una correspondencia con ASUM en mayor o menor medida.

En cuanto a la comparación de roles (Tabla 4), se observa que el rol de Arquitecto Empresarial no tiene un paralelo en la metodología TDSP. Podemos interpretar entonces que ASUM identifica los roles según las responsabilidades distribuidas a los fines del proyecto, mientras que TDSP organiza los roles en función de la gestión del equipo de trabajo en Ciencia de Datos, a través de distintos proyectos. En ASUM la figura del cliente tiene mayor presencia, ya que se le atribuyen un conjunto de roles que TDSP no reconoce dentro de la metodología.

Finalmente, podemos concluir que ambas metodologías, tienen una impronta ágil como respuesta a la vorágine de proyectos de todas las áreas de sistemas con necesidades tan inmediatas como cambiantes. La elección de una metodología u otra está fundada en cada caso por la experiencia del equipo de trabajo en proyectos.

REFERENCIAS

- [1] Tukey, J. (1962). The future of data analysis. *Ann. Math. Stat.*, 33(1), 1e67.
- [2] Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley.
- [3] Fayyad, U. et al. (1996). *Advances in knowledge discovery and data mining*. AAAI Press.
- [4] Martins, S. (2014). Derivación del proceso de explotación de información desde el modelado del negocio. *Revista Latinoamericana de Ingeniería de Software*, 2(1), 53-76.
- [5] García, R., Britos, P. y Rodríguez, D. (2013). Information mining processes based on intelligent systems. *Lecture Notes on Artificial Intelligence*, 7906, 402-410.
- [6] Britos, P. (2008). *Procesos de explotación de información basados en sistemas inteligentes*. Disertación doctoral. Universidad Nacional de La Plata.
- [7] Microsoft. (2020). *Team data science process*. Recuperado: <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process>.
- [8] IBM. (2015). *Analytics Solution Unified Method (ASUM)*. Recuperado: http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html.
- [9] Beck, K. et al. (2001) *Manifiesto Ágil*. Recuperado: <https://agilemanifesto.org/>.
- [10] Beck, K. et al. (2001) *Principios Ágiles*. Recuperado: <http://agilemanifesto.org/principles.html>.

- [11] Sánchez, C. y Félix Serrano, M. (2020). Ciencia de datos, producto y agilidad. Recuperado: <https://jeronimopalacios.com/product-delivery/ciencia-de-datos-producto-y-agilidad/>.
- [12] IBM. (2016). Analytics solutions unified method implementations with agile principles. Recuperado: <ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf>
- [13] Chapman, P. et al. (1999). CRISP-DM 1.0 Step by step Blguide. Edited by SPSS. Recuperado: <http://www.crispdm.org/CRISPWP-0800.pdf>.