MODELO BASADO EN CRISP-DM EXTENDIDO MEDIANTE PRÁCTICAS DE METODOLOGÍAS ÁGILES PARA PROYECTOS MEDIANOS DE ANALÍTICA DE DATOS

CRISTIAN DANIEL MAVESOY MURCIA

UNIVERSIDAD DE MEDELLÍN
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SOFTWARE
MEDELLÍN
2018

MODELO BASADO EN CRISP-DM EXTENDIDO MEDIANTE PRÁCTICAS DE METODOLOGÍAS ÁGILES PARA PROYECTOS MEDIANOS DE ANALÍTICA DE DATOS

CRISTIAN DANIEL MAVESOY MURCIA

Trabajo de grado para optar al título de
Maestría en Ingeniería de Software

Director
Juan Bernardo Quintero, PhD
Doctor en Ingeniería Electrónica

Codirector
Bell Manrique Losada, PhD
Doctora en Ingeniería

UNIVERSIDAD DE MEDELLÍN
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SOFTWARE
MEDELLÍN
2018

Nota de aceptación:

_____

_____

_____

_____

_____

_____

_____

Director

_____

Jurado

_____

Jurado

Medellín, Febrero 1 de 2019.

Dedicado a ...

Mi familia, que está ahí en todo momento sin importar las decisiones que haya tomado en la vida; mi Padre, incondicional, motivador y un gran apoyo para buscar solución a las adversidades de la vida; mi Mamá, el amor de mi vida, la que me aconseja y acompaña todo el tiempo a pesar de la distancia; mi hermano, el que me recuerda siempre de dónde soy y lo que quiero ser; mi hijo, sencillamente el motor de mi vida, mi luz e inspiración para salir adelante.

AGRADECIMIENTOS

Primero agradezco a DIOS por permitirme participar en este proyecto de vida, nuevamente a mi familia, también a todos los docentes de la Universidad de Medellín que aportaron en mi crecimiento profesional en este proyecto de Maestría, en especial a los directores de este trabajo de grado (Juan Bernardo y Bell) y a Gloria Gasca que me recibió con mucha disposición desde el primer día en este programa y a esas personas con las que me involucré laboralmente que aportaron con su conocimiento en la construcción de este proyecto.

CONTENIDO

pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

pág.

LISTA DE ANEXOS

pág.

ACRÓNIMOS

AUP:	Agile Unified Process
BI:	Business Intelligence
CRISP-DM:	Cross Industry Standard Process for Data Mining
KDD:	Knowledge Discovery in Databases
XP:	Extreme Programming
SPEM:	Software Process Engineering Metamodel
ASD:	Adaptative Software Development
DAO:	Data Access Object
ME:	Method Engineering
AD:	Analítica de Datos

RESUMEN

Los proyectos de analítica de datos adquieren un papel importante en las organizaciones, apoyando la toma de decisiones, identificando oportunidades de mejora, mercados potenciales o para predecir el comportamiento de los clientes bajo algunas variables de entorno. Por lo tanto, se debe hacer énfasis en los procesos usados para este tipo de proyectos y adaptarlo a entornos ágiles, los cuales permitan realizar entregas tempranas, de valor e identificar amenazas del proyecto de forma oportuna lo antes posible.

Este documento propone una metodología de analítica de datos en entornos ágiles llamada CRISP-DM Ágil para proyectos medianos, la cual se basa en CRISP-DM donde se detallan actividades específicas de proyectos e incluye prácticas de metodologías ágiles para realizar tableros de control o *Dashboard* y/o *Reporting*.

El caso de estudio basado en un diseño experimental donde se comparan dos proyectos construidos en dos (2) universidades de la misma naturaleza, independiente de su tecnología, realizados usando el estándar CRISP-DM y CRISP-DM ÁGIL.

Por último, se compara mediante un marco evaluativo donde se determina que al incluir prácticas ágiles en proyectos medianos de analítica de datos se mejora la dinámica del equipo de trabajo y se disminuyen los tiempos de entrega según la priorización de las necesidades de la organización del cliente.

Palabras clave: CRISP-DM, Analítica de datos, SPEM, Ágil, SCRUM.

PARTE I

INTRODUCCIÓN

“Acepta la responsabilidad de tu vida. Date cuenta que tú eres quien va a llegar a donde quiere ir, nadie más” -- *Les Brown*

CAPÍTULO 1

Introducción

Durante los últimos años se ha tenido un incremento exponencial de todo tipo de información que las organizaciones han empezado a aprovechar para buscar patrones, tendencias e identificar oportunidades en su entorno para poder entender esas necesidades que no se han suplido o requieren un acompañamiento especial. Como lo indica (Gandomi & Haider, 2015), la Analítica de datos ofrece un sin número de técnicas, herramientas y metodologías para desarrollar proyectos que se apropian de la información y contribuyen a la toma de decisiones para mejorar su competitividad.

Por lo tanto, esta tesis contribuye en los proyectos de analítica de datos con una propuesta metodológica basada en CRISP-DM que incluye prácticas de metodologías ágiles. Así, los profesionales de datos cuentan con actividades definidas en un proyecto donde los requisitos sufran cambios en el tiempo y se requiera de una constante comunicación con los usuarios finales para dar respuesta a las necesidades de las organizaciones en el menor tiempo posible a sus clientes, proveedores o competencia.

La propuesta denominada **CRISP-DM Ágil** es el resultado de una convergencia de algunas actividades realizadas en la metodología de analítica CRISP-DM, usando prácticas de metodologías ágiles como SCRUM y eXtreme Programming (XP), donde se describe la interacción entre elementos como fases, actividades y artefactos generados por roles específicos.

Al hacer uso de esta metodología propuesta, se muestra que es posible disminuir el esfuerzo en las entregas de valor al cliente y mejorar la comunicación entre los actores involucrados en un proyecto de analítica de datos, ofreciendo una reacción temprana y oportuna a los impedimentos presentados.

Este documento se divide en 5 partes, la Parte 1 **Introducción** que brinda un contexto de la propuesta; luego la Parte 2 **Exploración,** define la teoría usada que sirve como base para conocer los aspectos de esta investigación. La parte 3 **Propuesta,** describe todos los elementos necesarios y de qué manera se ha construido la metodología presentada en este documento. Se continúa con la Parte 4 **Validación,** que describe un caso de estudio y los resultados obtenidos al hacer uso de CRISP-DM Ágil en un proyecto de analítica de datos en entornos ágiles y por último la Parte 5 **Conclusiones,** donde se muestran los resultados obtenidos y la experiencia ganada al utilizar este tipo de metodologías en este campo.

1.1. Justificación

El mundo de la información es un entorno de constantes cambios en los que se ven afectados todos los actores que de una u otra forma tienen interacción con éste, como ingenieros, contadores, médicos, administradores y docentes. Los profesionales de datos no son ajenos a estos cambios. Por esto se ven en la necesidad de evolucionar con su entorno no solo mejorando sus técnicas, métodos o herramientas, sino adaptando los procesos a su entorno cambiante y a las necesidades de su organización.

CRISP-DM es una metodología de trabajo completa para el desarrollo de cualquier tipo de proyecto de analítica de datos, lo que la conduce a ser genérica, debido a que carece de un nivel de detalle que permita a un profesional de datos con conocimientos moderados en el campo de analítica de datos saber cómo llevar a

cabo un proyecto en cada una de las fases que ofrece (Sharma, Stranieri, Vamplew, & Martin, 2017).

Debido a que CRISP-DM es el modelo de referencia estándar más usado en este campo, el cual divide el ciclo de vida de su desarrollo en seis fases (Sharma *et al*., 2017), se cuenta con una base importante que brinda una gran oportunidad para adaptar CRISP-DM a las nuevas tendencias de desarrollo como lo son las metodologías ágiles y contribuir a que los profesionales de datos realicen proyectos de analítica donde involucren a los usuarios en todas las fases de desarrollo. Se puede convertir en una potencialidad el aprovechar en entornos ágiles los beneficios de CRISP-DM, como modelo de referencia más usado en proyectos de analítica de datos, basado en procesos de las metodologías ágiles para el desarrollo de proyectos de software (Muntean & Surcel, 2013).

Las ventajas de usar metodologías ágiles se pueden evidenciar en los tiempos de entrega y desarrollos más flexibles, para identificar fallos o mejoras en etapas tempranas (Golfarelli, Rizzi, & Turricchia, 2012), lo cual es un escenario ideal para los proyectos de analítica de datos donde la comunicación con el cliente es un factor clave de éxito en los proyectos. Además, en la literatura se evidencian investigaciones que aproximan las metodologías ágiles para el desarrollo de software hacia el ámbito del desarrollo de proyectos de analítica de datos (Rehani, 2011), (Muntean & Surcel, 2013), (Muntean, 2014), quienes presentan modelos de desarrollo de proyectos ágiles como SCRUM, ASD y XP, pero que no incluyen CRISP-DM como modelo de referencia en sus marcos de trabajo propuestos.

Por lo anterior, es importante adoptar CRISP-DM para aplicarlo a proyectos medianos de analítica de datos en entornos ágiles y aprovechar esas buenas prácticas que son conocidas por los profesionales de datos a través de la incorporación de metodologías ágiles para lograr disminuir los esfuerzos en la entrega de proyectos de analítica de datos y así responder al comportamiento y las

necesidades de las organizaciones, donde las entregas tempranas, de valor y una constante comunicación con el cliente son fundamentales para el éxito de los proyectos.

1.2. Problema de Investigación

Los profesionales de datos, son personas encargadas de dar a conocer información importante u oculta en los datos de una organización y/o su contexto a través de un proceso de recolección, modelado, procesamiento y presentación a sus clientes a través de diferentes elementos como reportes, **KPI** (*Key Performance Indicator*) y *Dashboard*. Por lo tanto, necesitan de una metodología donde exista una constante comunicación con todas las personas involucradas en el proyecto durante el proceso de desarrollo de un producto de analítica de datos y permita la adaptación a los cambios de los requisitos, diseño y otros factores (Ostadzadeh & Shams, 2013).

En la literatura y en la industria existen diversos métodos de desarrollo de proyectos de analítica de datos para diferentes entornos como el estándar SEMMA por sus siglas en inglés para *Sample, Explore, Modify, Model* y *Assess,* el cual define el proceso de realizar un proyecto de minería de datos; luego, KDD por sus siglas en inglés *Knowledge Discovery in Databases*, el cual sirve para extraer lo que se considera conocimiento de acuerdo con la especificación de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación de la base de datos; y por último el estándar CRISP-DM por sus siglas CR*oss-Industry Standard Process for Data Mining* (Azevedo & Santos, 2008).

La metodología CRISP-DM se ha convertido en el estándar para procesos de Minería de Datos más utilizada en la industria (Nadali, Kakhky, & Nosratabadi, 2011), pero su proceso se ve enmarcado en unos pasos secuenciales afectando la

comunicación con los usuarios finales y el constante seguimiento de los objetivos de negocio (Sharma et al., 2017). Lo anterior debido a que sólo incluye validación de resultados en la penúltima fase llamada ‘Evaluación’, donde se compara el resultado del modelo generado, con los objetivos de negocio identificados en la primera fase conocida como ‘entendimiento del negocio’.

Una comunicación constante entre todos los integrantes de un equipo de trabajo es esencial para el éxito de los proyectos en cualquier campo. Por esta razón, las metodologías ágiles de desarrollo de software más utilizadas como ASD (*Adaptative Software Development*), XP (*eXtreme Programming*) y SCRUM se están usando en proyectos de analítica de datos para definir entornos ágiles, considerando que son los más usados por la industria del software (Krawatzeck, Dinter, & Thi, 2015).

Cuando los profesionales de datos usan la metodología tradicional CRISP-DM para el desarrollo de proyectos se encuentran con diversas debilidades, tales como: un ciclo de desarrollo largo que se vuelve un proceso incierto para el usuario final durante su ejecución, los usuarios no están involucrados en el desarrollo del proyecto, los requerimientos no pueden cambiar en el tiempo de ejecución del proyecto porque se verían seriamente afectados los tiempos y costos del mismo y no existen pruebas parciales que permitan ver el avance del proyecto, ya que existe sólo una fase de evaluación y despliegue (Muntean & Surcel, 2013). Por lo tanto, se presentan grandes esfuerzos relacionados con tiempos o costos en la entrega de proyectos de analítica de datos usando CRISP-DM (Muntean & Surcel, 2013).

1.3. Campo de acción y enfoque de la investigación

1.3.1. Analítica de Datos

Los datos son la fuente de la que se obtienen las variables, las relaciones entre ellas, el conocimiento inducido o los patrones de comportamiento identificados, convirtiéndose en un elemento vital de todo análisis predictivo (Espino & Martinez, 2017). La analítica de datos se encarga de darle un tratamiento a los datos generados dependiendo de las necesidades de un negocio particular desde diferentes líneas de profundización como lo son la Analítica Descriptiva, Analítica Predictiva y Analítica Prescriptiva (Delen & Demirkan, 2013). El grupo de investigación de Microsoft ha querido involucrar una nueva línea llamada Analítica Diagnóstica (Garzón, 2018).

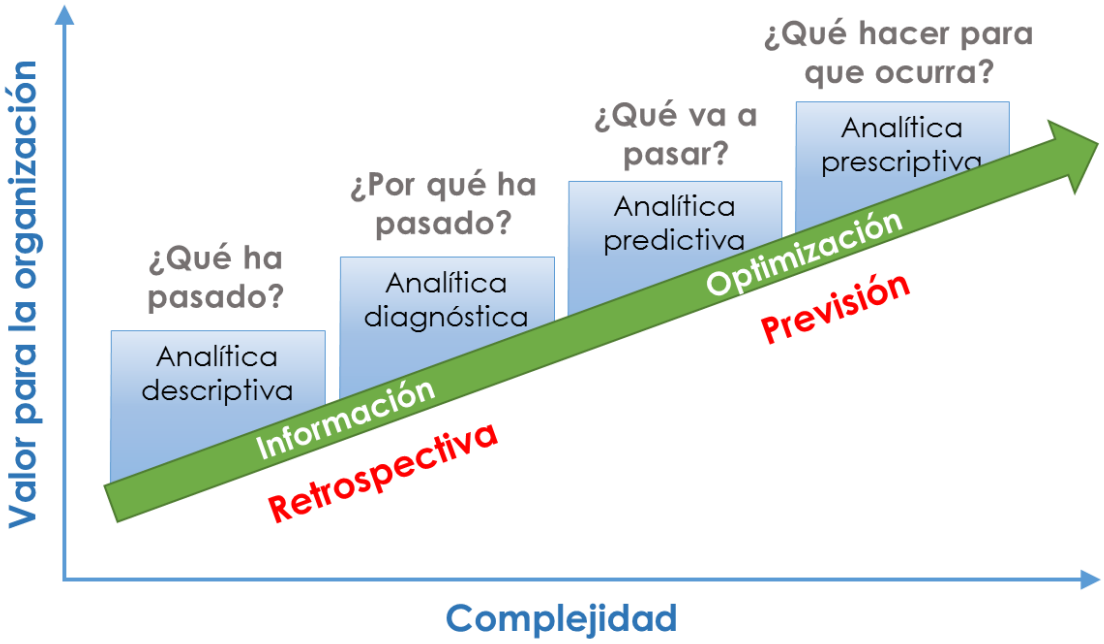


Imagen 1 Tipos de Analítica de Datos

La gráfica anterior representa los tipos de analítica de datos, donde el tamaño de un proyecto de analítica depende de dos factores, su valor para la organización y su complejidad, aunque vale la pena resaltar que dependiendo del tamaño del

conjunto de datos que se cuente y los objetivos de analítica definidos pueden incidir en el tamaño del proyecto de analítica de datos.

Tener un proceso repetible y bien definido puede ayudar a los equipos de ciencia de datos a través de un rango de desafíos, que incluye entender quién necesita estar incluido como stakeholder en el proceso, seleccionar una arquitectura/infraestructura técnica de datos apropiada, determinar las técnicas analíticas apropiadas y validar los resultados (Saltz, Shamshurin, & Crowston, 2017). Además de que podría generar un comportamiento proactivo y aumentar la comunicación efectiva entre los integrantes del proyecto.

1.3.2. Minería de Datos

La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más (SAS, 2019).

Además, satisface su objetivo principal al identificar correlaciones y patrones válidos, potencialmente útiles y fácilmente comprensibles presentes en los datos existentes. Este objetivo de la minería de datos puede satisfacerse modelizándolo como de naturaleza predictiva o descriptiva. El modelo predictivo funciona haciendo una predicción sobre los valores de los datos, que utiliza resultados conocidos encontrados en diferentes conjuntos de datos. Las tareas incluidas en el modelo de minería de datos predictivo incluyen clasificación, predicción, regresión y análisis de series de tiempo. Los modelos descriptivos identifican principalmente patrones o relaciones en conjuntos de datos. Sirve como una manera fácil de explorar las propiedades de los datos examinados anteriormente y no predecir nuevas propiedades (Agyapong, Hayfron-Acquah, & Asante, 2016).

1.3.3. Metodologías ágiles

En la industria del software, las metodologías de desarrollo han sido un marco de trabajo que ha permitido elaborar proyectos con una estructura ordenada.

Las metodologías ágiles nacen a partir de las necesidades presentadas durante el desarrollo de proyectos usando metodologías tradicionales, donde un grupo de personas en 2001 crean la organización sin ánimo de lucro "*The Agile Alliance*", y nace el manifiesto ágil que se basa en 4 valores, los individuos e interacciones por encima de los procesos y las herramientas, software funcionando por encima de la documentación, la colaboración del cliente por encima de la negociación del contrato y la respuesta al cambio por encima del seguimiento de un plan, además de 12 principios para realizar entregas tempranas y de valor para la organización (Uribe & Ayala, 2007).

1.3.4. Ingeniería del método

La ingeniería del método es definida como la disciplina de la ingeniería para diseñar, construir y adaptar métodos, técnicas y herramientas para el desarrollo de sistemas (Henderson-Sellers & Ralyté, 2010).

Tradicionalmente, la ingeniería de métodos (ME) se ocupa de los procesos de diseño, construcción y adapta métodos dirigidos al desarrollo de sistemas de información (Bucher, Klesse, Kurpjuweit, & Winter, 2007).

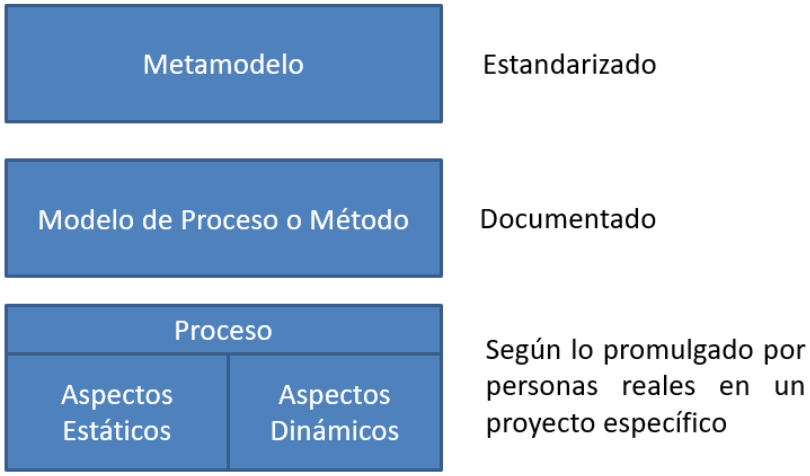


Imagen 2 Las "capas" de proceso y terminología de método

Una vez identificados los campos de acción requeridos para el cumplimiento de los objetivos propuestos a continuación, a partir de los lineamientos de la ingeniería del método, se propone un modelo para proyectos medianos de analítica de datos basado en metodologías de analítica de datos y las buenas prácticas de las metodologías ágiles. En la Imagen 3 se muestra el marco para diseñar la metodología.

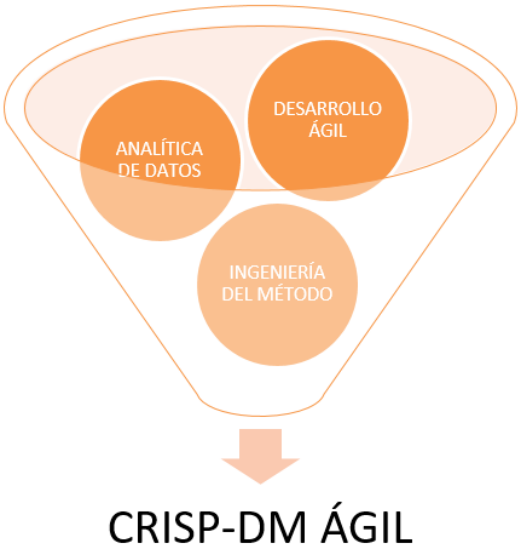


Imagen 3 Marco de desarrollo de la propuesta

Elaboración Propia

1.4. Hipótesis

Aplicando CRISP-DM e integrando prácticas de metodologías ágiles en proyectos medianos de analítica de datos se disminuye el esfuerzo en la entrega al cliente.

1.5. Objetivos de la investigación

El desarrollo de la propuesta metodológica plantea los siguientes objetivos:

Objetivo General: Proponer una metodología de trabajo basada en CRISP-DM extendida con prácticas de Metodologías Ágiles para disminuir el esfuerzo en el desarrollo de proyectos medianos de Analítica de Datos.

Objetivos específicos: El objetivo general se alcanza mediante los siguientes objetivos específicos:

1. Identificar fortalezas y debilidades de la metodología CRISP-DM para todas sus fases en proyectos de analítica de datos.
2. Realizar una caracterización de los procesos de metodologías de analítica de datos en entornos ágiles.
3. Diseñar una metodología basada en CRISP-DM incluyendo prácticas de metodologías ágiles para proyectos de analítica de datos
4. Validar la propuesta metodológica con un caso de estudio en un proyecto de analítica de datos de tamaño mediano.

El proceso de investigación definido para alcanzar estos objetivos se ilustra en la Imagen 4, mostrando las diferentes fases adelantadas y el principal resultado de cada una.

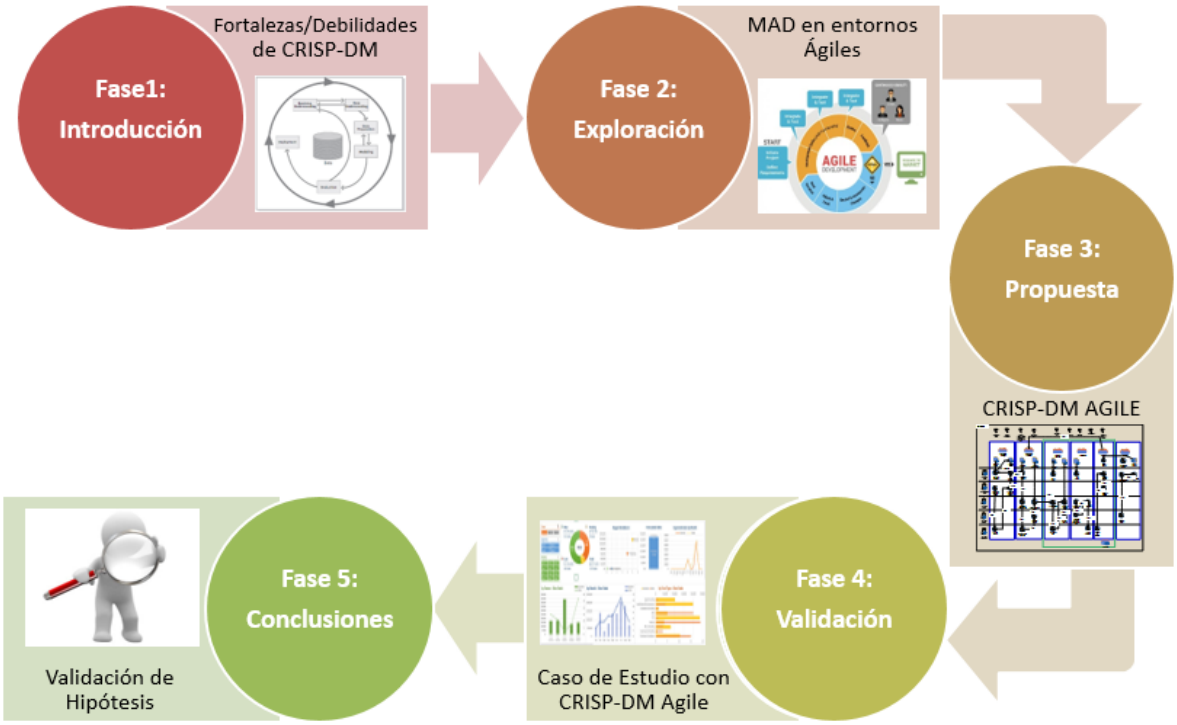


Imagen 4 Proceso de Investigación para alcanzar los objetivos Elaboración Propia

A continuación, se describen brevemente las anteriores fases con el principal resultado:

Fase 1 – Introducción: Contiene las generalidades de este documento, definiciones, el porqué de la importancia de esta propuesta metodológica y especifica los objetivos logrados.

Fase 2 - Exploración: Durante la fase de exploración se describen los conceptos generales de los componentes necesarios para la realización de la propuesta y los antecedentes encontrados en diferentes bases de datos científicas que sirvieron para orientar el camino de esta investigación.

Fase 3 - Propuesta: Se define la caracterización de aspectos estáticos y dinámicos para realizar el Modelo basado en CRISP-DM extendido mediante prácticas de metodologías ágiles para proyectos medianos de analítica de datos definiendo roles, procesos, actividades y artefactos llamado CRISP-DM AGILE.

Fase 4 - Validación: Desarrollo de casos de aplicación en proyectos medianos de analítica de datos del modelo propuesto y usando solo la metodología tradicional CRISP-DM.

Fase 5 - Conclusiones: Contiene las recomendaciones al usar la propuesta metodológica, las conclusiones de este documento y trabajos futuros.

Este documento consigna el trabajo adelantado para cubrir estas cinco fases, con una presentación estructurada en partes para dar mayor claridad con respecto a su contenido.

1.6. Estructura de la tesis

La realización de un trabajo de esta índole, deja una amplia gama de testimonios y experiencias que necesitan ser documentadas de forma organizada para facilitar su estudio y comprensión. Por tal motivo este trabajo está organizado en 10 capítulos agrupados en 6 partes, como se ilustra en la Imagen 5.

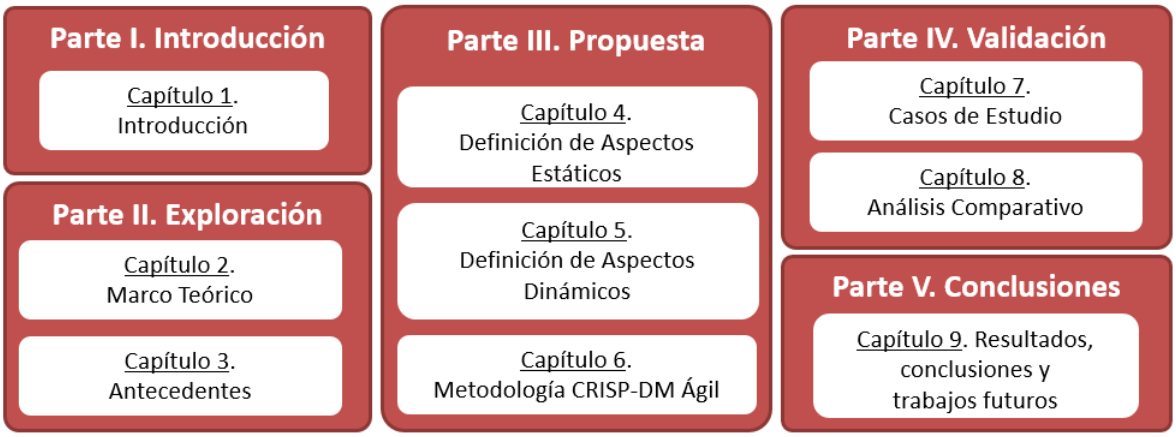


Imagen 5 Estructura de la tesis.

Elaboración Propia

A continuación, se describen brevemente los capítulos de este trabajo con su respectivo contenido:

PARTE I - INTRODUCCIÓN

CAPÍTULO 1. Introducción: Realiza una contextualización, presentando su campo de acción y explicando el proceso investigativo que sigue. Contiene una breve descripción de los diferentes tipos de Analítica de Datos, los objetivos, la estructura de la tesis y las rutas de lectura para abordarla.

¡Error! No se encuentra el origen de la referencia.**PARTE II - EXPLORACIÓN**

CAPÍTULO 2. Marco teórico: Identifica, define y compara las diferentes metodologías de analítica de datos existentes, con una breve descripción de sus procesos y fases donde luego son comparadas bajo ciertos criterios que permiten

dar importancia a la selección de una metodología que sea base para el modelo propuesto.

CAPÍTULO 3. Antecedentes: Identifica, define y compara las diferentes metodologías de desarrollo ágil existentes, con una breve descripción de sus procesos, fases, herramientas y técnicas que luego son comparadas bajo ciertos criterios donde la interacción con el cliente es el factor más importante para seleccionar diferentes prácticas que aporten a la disminución de esfuerzos en las entregas del producto final al cliente.

PARTE III - PROPUESTA

CAPÍTULO 4. Definición de aspectos estáticos: Como parte de los elementos de los que se forma la metodología presentada en esta investigación, se discriminan componentes estáticos como artefactos o roles.

CAPÍTULO 5. Definición de aspectos dinámicos: Como parte de los elementos de los que se forma la metodología presentada en esta investigación, se discriminan componentes dinámicos como fases o actividades.

CAPÍTULO 6. Metodología CRISP-DM Ágil: Se describe cómo están relacionados los elementos estáticos y dinámicos de los capítulos 4 y 5 de este documento haciendo uso de un diagrama SPEM-.

PARTE IV. VALIDACIÓN

CAPÍTULO 7. Casos de Estudio: Presenta los momentos llevados a cabo en un proyecto mediano de analítica de datos, el alcance presentado, los actores involucrados y el reporte de la ejecución del proyecto usando la metodología CRISP-DM Ágil.

CAPÍTULO 8. Análisis Comparativo: Se analizan algunos criterios de comparación para metodologías de analítica de datos usando CRISP-DM Ágil y CRISP-DM puro.

PARTE V - CONCLUSIONES

CAPÍTULO 9. Conclusiones, recomendaciones y trabajos futuros: inicia con un análisis del cumplimiento de los objetivos específicos que se plantearon en la investigación. Muestra los aportes de este trabajo al igual que sus beneficiarios.

Presenta cómo los proyectos medianos de analítica de datos pueden ser tratados usando prácticas de metodologías ágiles basados en un modelo que permite mapear las actividades específicas para un proyecto de esta naturaleza.

CAPÍTULO 2 Marco Teórico

2.1. Metodologías de analítica de datos

En el desarrollo de cualquier tipo de proyectos se requiere de marcos o metodologías que brinden una manera organizada de realizar el trabajo y asignar responsables a las actividades que se van a elaborar. Es importante contar con un conjunto de técnicas que permitan materializar una idea, pero si no existe un orden y unos compromisos que se adquieran cuando participan varias personas, el cumplimiento del proyecto sería indeterminado.

2.1.1. CRISP-DM

El modelo de referencia CRISP-DM es el estándar de minería de datos más ampliamente utilizado que divide el ciclo de vida de un ejercicio de minería de datos en seis fases diferentes. También define las tareas que se llevarán a cabo en cada fase y el resultado esperado. El siguiente paso depende del resultado del paso anterior, le permite volver a él y es menos rígido que un modelo de cascada tradicional utilizado en el desarrollo de software (Sharma *et al*., 2017). Como lo definen Daihani y Feblian (2016) es la estandarización del proceso de minería de datos como una estrategia general de resolución de problemas del negocio o unidad de estudio.

CRISP-DM es flexible y se puede personalizar fácilmente como se muestra en la Imagen 6. Por ejemplo, si una organización intenta detectar actividades de blanqueo de dinero, es probable que necesite realizar una criba de grandes cantidades de datos sin un objetivo de modelado específico. En lugar de realizar el modelado, el trabajo se debe centrar en explorar y visualizar datos para descubrir patrones

sospechosos en datos financieros. CRISP-DM permite crear un modelo de minería de datos que se adapte a necesidades concretas (IBM, 2012).

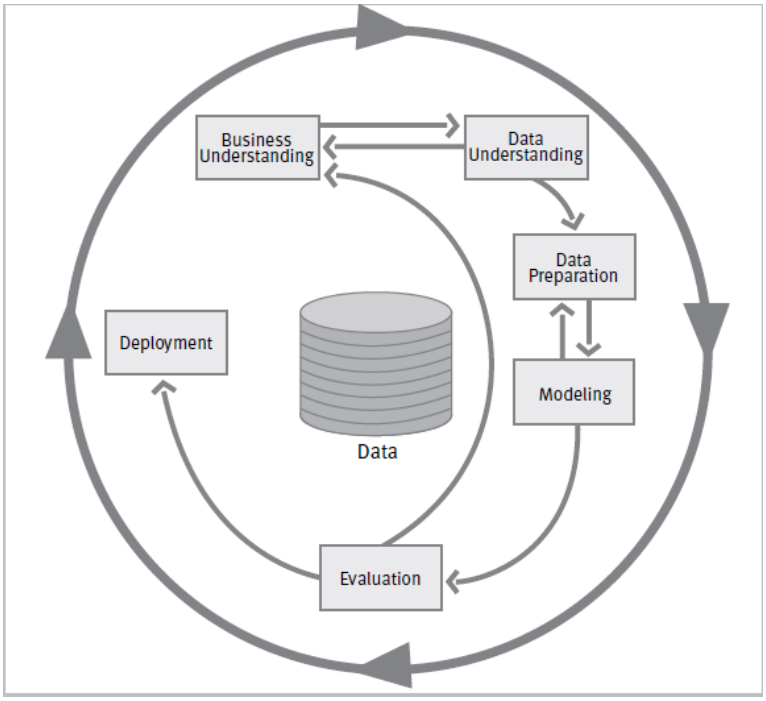


Imagen 6 Fases de CRISP-DM

Fuente (Chapman *et al.*, 2000)

2.1.2. KDD

Knowledge Discovery Databases (KDD) es el proceso de extraer el conocimiento oculto según las bases de datos. KDD requiere conocimiento previo relevante y una comprensión breve del dominio y los objetivos de la aplicación. El modelo de proceso KDD es de naturaleza iterativa e interactiva (Shafique & Qaiser, 2014).

El proceso KDD, como se presenta en (Fayyad, 1996) es el proceso de usar métodos DM para extraer lo que se considera conocimiento de acuerdo con la especificación de medidas y umbrales, utilizando una base de datos junto con

cualquier pre procesamiento, submuestreo y transformación requeridos de la base de datos (Azevedo & Santos, 2008).

Imagen 7 Proceso KDD

Fuente (Fayyad, 1996)

2.1.3. SEMMA

La metodología SEMMA toma su nombre de las diferentes etapas que lideran el proceso de explotación de información, muestreo (*Sample*), exploración (*Explore*), modificación (*Modify*), modelamiento (*Model*) y evaluación (*Assess*). La metodología es en sí misma un ciclo cuyos pasos internos se pueden realizar de forma iterativa de acuerdo con las necesidades. SEMMA proporciona un proceso fácil de entender que permite el desarrollo y mantenimiento de proyectos de explotación de información (Jair *et al*., 2017).

El estándar SEMMA (*Sample, Explore, Modify, Model, and Assess*) es un método de minería de datos desarrollado por el instituto SAS. Ofrece y permite la comprensión, organización, desarrollo y mantenimiento de proyectos de minería de datos. Ayuda a proporcionar soluciones para los problemas y los objetivos del negocio. SEMMA está vinculado a *SAS Enterprise Miner* y es básicamente una organización lógica de las herramientas funcionales para proyectos de minería de

datos. Tiene un ciclo de cinco etapas o pasos como lo muestra la Imagen 8 (Shafique & Qaiser, 2014).

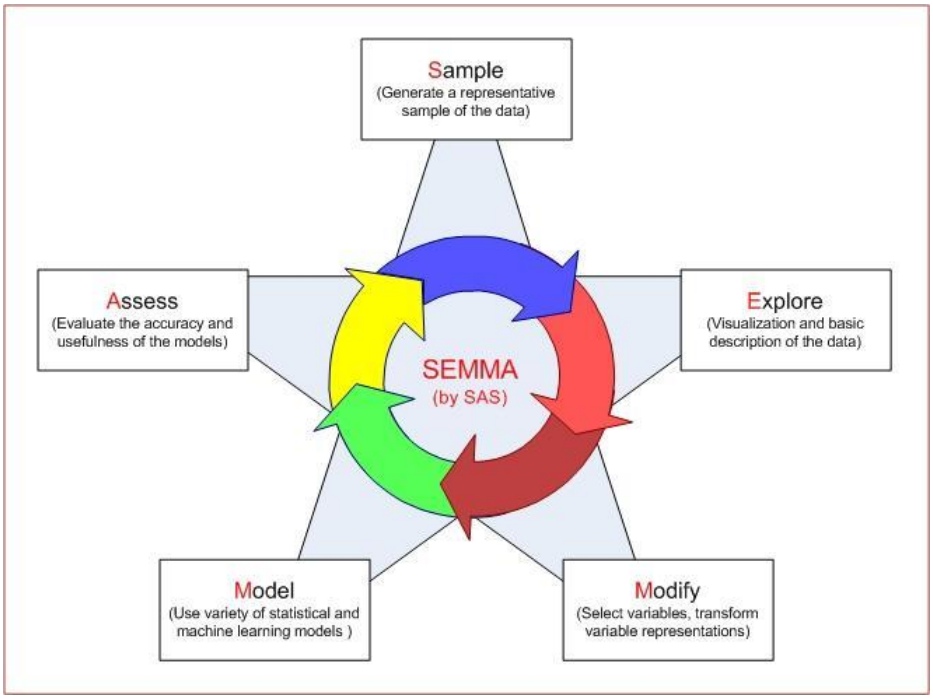


Imagen 8 Metodología SEMMA. Fuente (Shafique & Qaiser, 2014)

2.1.4. TDSP

El *Team Data Science Process* (TDSP) es una metodología ágil e iterativa de ciencia de datos para ofrecer soluciones analíticas predictivas y aplicaciones inteligentes de manera eficiente. TDSP ayuda a mejorar la colaboración y el aprendizaje en equipo. Contiene una destilación de las mejores prácticas y estructuras de Microsoft y otros en la industria que facilitan la implementación exitosa de iniciativas de ciencia de datos. El objetivo es ayudar a las empresas a darse cuenta de los beneficios de su programa de análisis (GuhaThakurta, Ericson, Martens, & Bradley, 2017).

La Imagen 9 describe las etapas principales que normalmente se ejecutan en los proyectos, a menudo de manera iterativa:

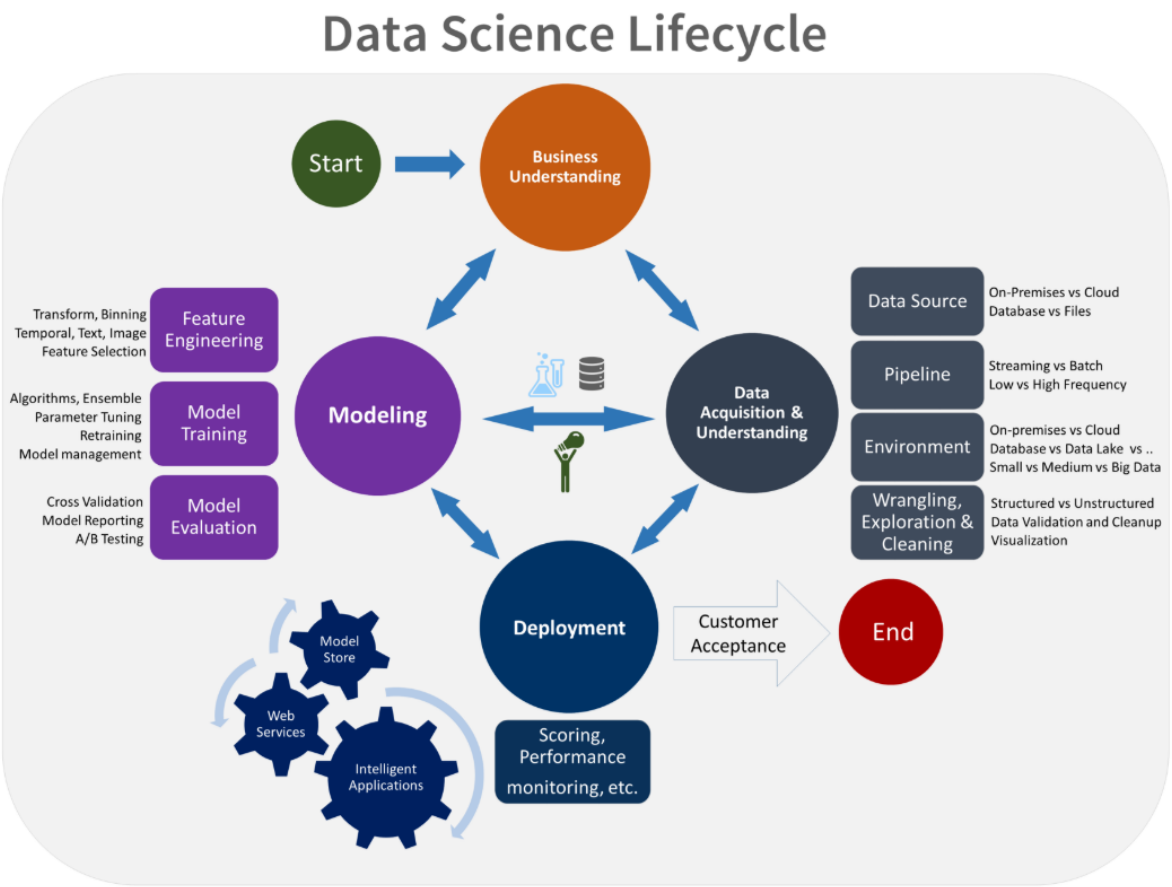


Imagen 9. Ciclo de Vida de la Ciencia de Datos. Fuente (GuhaThakurta, Ericson, Martens, & Bradley, 2017)

2.2. Desarrollo Ágil

2.2.1. Adaptative Software Development (ASD)

Creado por Jim Highsmith, es un modelo iterativo, orientado a los componentes de software más que a las tareas y es tolerante a los cambios. Este modelo propone 3 fases: especulación, colaboración y aprendizaje. Con la primera inicia y planifica el proyecto, en la segunda se desarrollan las características y finalmente en la tercera se revisa la calidad y se entrega al cliente. Las revisiones frecuentes sirven para

aprender de los errores y volver a iniciar el ciclo de desarrollo (Garcés Uquillas, 2015). La Imagen 10 permite una representación gráfica del comportamiento de ASD:

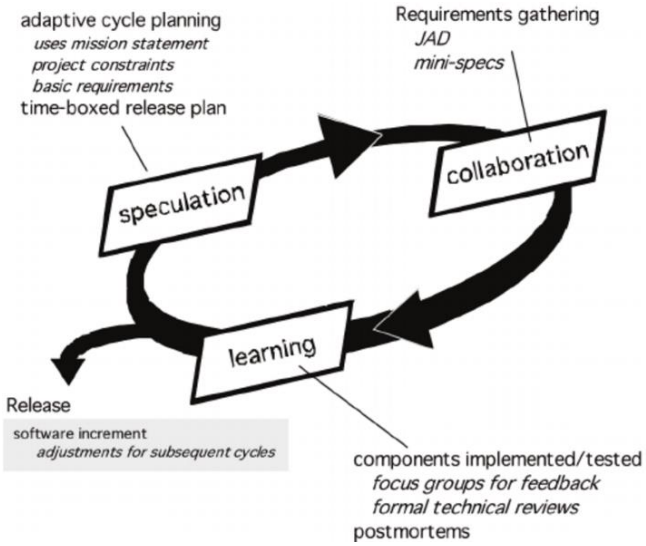


Imagen 10 Proceso ASD. Fuente (Garcés Uquillas, 2015)

2.2.2. Lean Software Development (LSD)

Lean Software Development es una metodología iterativa ágil desarrollada originalmente por Mary y Tom Poppendieck. LSD le debe gran parte de sus principios y prácticas al movimiento Lean Enterprise y las prácticas de compañías como Toyota. LSD enfoca al equipo en entregar valor al cliente, en la eficiencia del "flujo de valor" y los mecanismos que entregan ese valor. Los principios principales de esta metodología según (Shcherbakov, Shcherbakova, Brebels, Janovsky, & Kamaev, 2014) incluyen:

- Eliminando residuos
- Aprendizaje amplificador
- Decidiendo lo más tarde posible
- Entregando tan rápido como sea posible
- Empoderando al equipo
- Integridad en lo construido

- Ver el todo

En la Imagen 11 se representa el ciclo de vida de un proyecto basado en esta metodología pasando por las fases de Identificación de flujo de valor, mapear el flujo de valor, crear el flujo, establecer salidas y buscar la perfección.

Imagen 11 Proceso LEAN. Fuente. (Shcherbakov et al., 2014)

2.2.3. Programación Extrema (XP)

XP, originalmente descrito por Kent Beck, se ha convertido en una de las metodologías ágiles más populares y controvertidas. XP es un enfoque disciplinado para entregar software de alta calidad de forma rápida y continua. Promueve una alta participación del cliente, ciclos rápidos de retroalimentación, pruebas continuas, planificación continua y trabajo en equipo para entregar software en funcionamiento a intervalos muy frecuentes, por lo general cada 1-3 semanas. La receta original de XP se basa en cuatro valores simples: "simplicidad, comunicación, retroalimentación y coraje" y doce prácticas de apoyo (McLaughlin, 2018), todo esto se puede evidenciar en la Imagen 12 que representa su comportamiento:

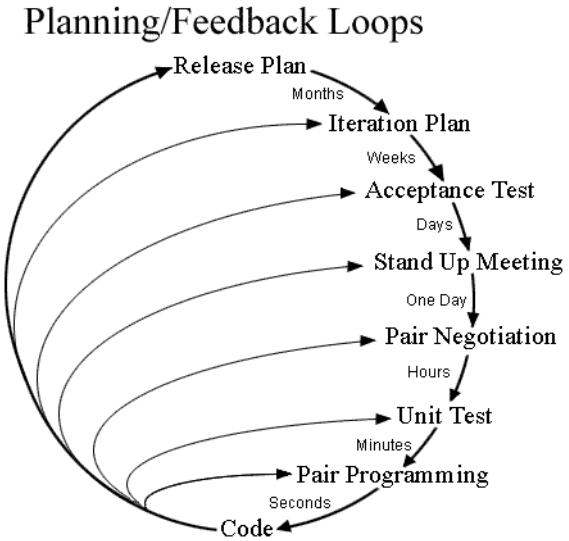


Imagen 12 Marco eXtreme Programming. Fuente. (McLaughlin, 2018)

2.2.4. SCRUM

Es un marco de trabajo dentro del cual las personas pueden abordar complejos problemas de adaptación, mientras que ofrecen productiva y creativamente productos del más alto valor posible, además es ligero, simple de entender y difícil de dominar (Schwaber & Sutherland, 2017).

Scrum ha sido utilizado para gestionar el trabajo en productos complejos desde principios de los años noventa. Scrum no es un proceso, técnica o método definitivo. Más bien, es un marco dentro del cual se puede emplear varios procesos y técnicas. Scrum deja en claro la eficacia relativa de su gestión de productos y técnicas de trabajo para que pueda mejorar continuamente el producto, el equipo y el entorno de trabajo (Schwaber & Sutherland, 2017).

Por otro lado, es una de las metodologías ágiles más populares. Es una metodología de adaptación, iterativa, rápida, flexible y eficaz, diseñada para ofrecer un valor significativo de forma rápida en todo el proyecto. Scrum garantiza transparencia en la comunicación y crea un ambiente de responsabilidad colectiva y de progreso continuo. Como se define en la Guía SBOK™, está estructurado de tal manera que

es compatible con los productos y el desarrollo de servicios en todo tipo de industrias y en cualquier tipo de proyecto, independientemente de su complejidad (SCRUMstudy™, 2016).

Así mismo, tiene una amplia aplicabilidad para administrar y controlar proyectos iterativos e incrementales de todo tipo. Ken Schwaber, Mike Beedle, Jeff Sutherland y otros han contribuido significativamente a la evolución de Scrum en la última década. Scrum ha ganado cada vez más popularidad en la comunidad de desarrollo de software ágil debido a su simplicidad, productividad comprobada y su capacidad para actuar como envoltorio de varias prácticas de ingeniería promovidas por otras metodologías ágiles.

En la Imagen 13 se representa el flujo de trabajo dentro del marco de trabajo Scrum:

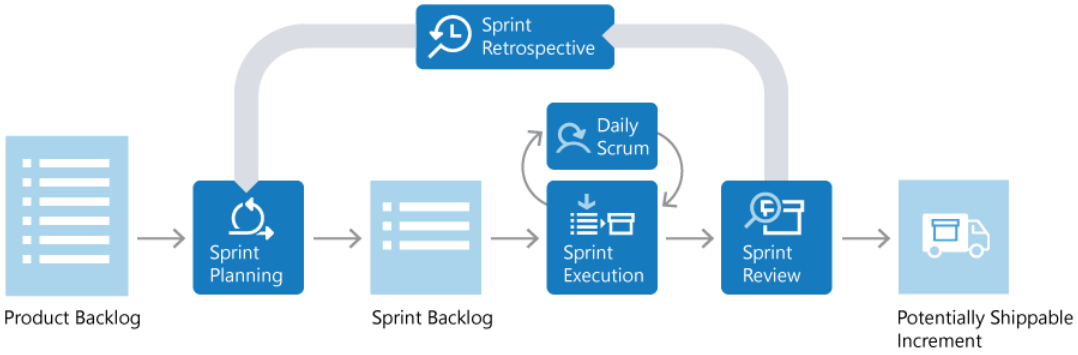


Imagen 13 Ciclo de Vida de Scrum.

(Boer, 2017)

2.3. Ingeniería del Método Situacional

La ingeniería del método tiene como objetivo aportar soluciones efectivas a la construcción, mejora y modificación de los métodos utilizados para desarrollar sistemas de información y software (Nehan & Deneckere, 2007).

Es la disciplina para construir métodos específicos del proyecto, llamados métodos situacionales, de partes de los métodos existentes, llamados fragmentos de métodos (Henderson-Sellers & Ralyté, 2010).

PARTE II
EXPLORACIÓN

"Si deseas tener éxito, debes emprender nuevos caminos, en lugar de recorrer los caminos tradicionales y trillados del éxito aceptado." -- John D. Rockefeller

CAPÍTULO 3

Antecedentes

3.1. Metodologías de Analítica

Los proyectos de analítica de datos requieren ser implementados en entornos ágiles para responder a las necesidades de las organizaciones en la actualidad, por lo que es importante resaltar que este tipo de proyectos están sufriendo la misma transformación que han pasado los proyectos de software, por lo tanto, cabe mencionar la diferencia entre metodologías tradicionales de las ágiles y por qué es importante dar el salto a estos entornos. En la Tabla 1 se muestran unos criterios de comparación que permiten demostrar por qué usar metodologías ágiles.

Comparativo de Metodologías

Criterio	Metodologías ágiles	Metodologías tradicionales
Adaptabilidad a cambios	SI	NO
Metodología Rígida	NO	SI
Cliente participa en el proyecto	SI	NO
Políticas y Normas Rigurosas	NO	SI
Interacción del Cliente en Reuniones	SI	NO
Pocos roles	SI	NO

Tabla 1 Comparativo Metodologías.

(García Aguilar, 2016)

Las metodologías tradicionales se caracterizan por: resistencia al cambio; planes de proyecto que deben respetarse durante todo proceso; poco involucramiento de los clientes; el cliente o los encargados de usar el producto pueden participar para brindar una retroalimentación sólo hasta que se tiene el producto para su revisión y

pruebas; y seguir fases al pie de la letra cuando se define una metodología de trabajo, ya que se vuelven políticas de proyecto.

Las metodologías ágiles se preocupan por involucrar al cliente en el desarrollo del proyecto, las políticas no son rigurosas permitiendo un gran campo de trabajo para tomar acciones que satisfagan las necesidades del cliente en una actividad cualquiera del proyecto.

3.1.1. Criterios de Comparación de las Metodologías

Existen propuestas que permiten identificar criterios relevantes para seleccionar qué metodología se puede usar en un proyecto específico. Por un lado, *Krawatzeck et al.* (2015) presenta un catálogo de identificación de principios, modelos de procesos, técnicas y tecnologías particulares para un proyecto de Analítica de Datos y según su entorno se puede decidir sobre el uso de una metodología particular. Luego, Moine y otros (2015) proponen un marco comparativo que propone, mediante ciertas preguntas agrupadas en cuatro categorías que se muestran en la Imagen 14, conocer la metodología más apropiada para proyectos de Analítica de datos.

Imagen 14 Aspectos del marco comparativo.

Fuente (Moine et al., 2015)

Este marco comparativo que propone Moine y otros, entrega un detalle de los aspectos más importantes en un proyecto de analítica de datos a nivel de proceso

y gestión de proyecto, por lo que ha sido implementado para apoyar la decisión del uso de CRISP-DM como metodología base para la presente propuesta metodológica.

Teniendo en cuenta el marco comparativo anterior, se realizó un estudio para determinar la importancia de las metodologías evaluadas, sobre un grupo de profesionales de analítica (2 docentes de la Universidad de Medellín, 4 Analistas de Datos, 2 Estudiantes de Maestría en Ingeniería de Software) para comprobar cuál de esas metodologías se debería usar como base para esta propuesta. Mediante el diseño de un instrumento en google forms, compuesto por las preguntas sugeridas en el marco comparativo, se encontraron los resultados de la Tabla 2.

Criterio de Evaluación	CRISP-DM	SEMMA	KDD
Nivel de detalle en la descripción de las actividades de cada fase	4	2	3
Escenarios de aplicación	3	2	2
Actividades específicas que componen cada fase.	6	3	2
Actividades destinadas a la dirección del proyecto	5,5	0,5	0,25

Tabla 2. Resultados Criterios de Evaluación. Fuente. Elaboración propia

De los criterios de comparación podemos decir que CRISP-DM es la mejor como base para la propuesta metodológica por las siguientes razones:

- Llega a un nivel de detalle sobre cómo realizar las actividades para generar las diferentes salidas o artefactos que soportan el desarrollo del proyecto,
- Es el modelo de proyectos de analítica más adoptado en muchos proyectos de minería de datos, y uno de los primeros modelos hacia la estandarización (Alnoukari, 2016).

- Cuenta con un nivel de detalle sobre las actividades a nivel de pasos para completar la actividad, lo que permitiría realizar los ajustes necesarios a los proyectos de Analítica de Datos en entornos ágiles.

CRISP-DM presenta algunas dificultades para proyectos en entornos ágiles, como: carecer del uso explícito de *data warehouse/data marts*; la etapa de despliegue es un punto muerto que hace que sea difícil adaptarse a los cambios del entorno (Alnoukari, 2016). Por otro lado, no tiene roles definidos para las actividades durante el desarrollo del proyecto y por último genera un gran número de artefactos que ponen en riesgo la velocidad de un proyecto de analítica en entornos ágiles. Estas dificultades son una oportunidad que esta propuesta metodológica busca suplir.

3.2. Principales metodologías de analítica

La dinámica de las organizaciones hoy en día demanda proyectos que generen valor de negocio en el menor tiempo posible y la Analítica de Datos no es ajena a estas necesidades que requieren de respuestas tempranas para ser competitivos en una industria globalizada.

Las metodologías de analítica, según (Rogalewicz & Sika, 2016), (Saltz, Shamshurin, & Connors, 2017), (Hochsztain & Tasistro, 2015) y otros, hacen gran énfasis en CRISP-DM, SEMMA y KDD siendo las más usadas y aplicadas por los profesionales de analítica. Estos son los estándares de metodologías de analítica tradicionales más seleccionados para proyectos. Piatetsky y KDnuggets (2014) han realizado encuestas a lo largo de los años demostrando que son las más importantes para el desarrollo de proyectos de analítica.

Por lo tanto, es válido tener un detalle más profundo sobre las 3 metodologías mencionadas anteriormente sobre sus fases para comparar cómo se lleva a cabo

un proyecto de Analítica de Datos. En la tabla 2 se muestran las fases de las que se componen las principales metodologías durante toda su ejecución.

Resumen de Procesos KDD, CRISP-DM y SEMMA

Modelos de procesos de minería de datos	KDD	CRISP-DM	SEMMA
No. De pasos	9	6	5
Nombre de los Pasos	Desarrollo y comprensión de la aplicación	Entendimiento de Negocio	----------
	Crear un conjunto de datos de destino	Entendimiento de los Datos	Muestra
	Limpieza y preprocesamiento de datos		Explorar
	Transformación de Datos	Preparación de los Datos	Modificar
	Elegir la Tarea de Minería de Datos adecuada	Modelamiento	Modelar
	Elegir el algoritmo de minería de datos adecuado		
	Empleando Algoritmo de Minería de Datos		
	Interpretando patrones minados	Evaluación	Evaluar
	Usando el conocimiento descubierto	Despliegue	----------

Tabla 3 Resumen de Procesos KDD, CRISP-DM y SEMMA.

Fuente: (Shafique & Qaiser, 2014)

Como se muestra en la Tabla 3, se intentan homologar los pasos o fases existentes en cada metodología para comparar el camino que recorrería un proyecto de analítica de datos en KDD, CRISP-DM y SEMMA. Éste último se ha descartado por la dependencia que tiene con las herramientas de SAS y CRISP-DM al cubrir todas las necesidades en un menor número de fases, abarca las actividades necesarias para el desarrollo de un proyecto de analítica de datos.

3.3. Revisión de literatura de metodologías de analítica en entornos ágiles

Durante los últimos años, con el crecimiento de la información generada por un sin número de elementos como redes sociales, sistemas de información, prácticas empresariales, estudios demográficos, entre otras, por lo tanto, el campo de la ciencia de datos ha tomado un papel importante para interpretar esa información brindando nuevas tecnologías, técnicas y herramientas capaces de soportar las exigencias que demanda la información día a día.

Es por esto que la propuesta ha iniciado con un proceso de revisión de tipo exploratorio, basado en la revisión sistemática de literatura (RSL) con el fin de identificar propuestas existentes en el desarrollo de proyectos de analítica de datos en entornos ágiles. Luego, se definieron palabras clave relacionadas con el problema de investigación para conformar unas cadenas de búsqueda como “crisp+dm+scrum” y “analítica+de+datos+entornos+agiles” para utilizarlas en las bases de datos científicas IEEE, Science Direct y ACM. También se realizaron búsquedas en Google Scholar, además se realizó un proceso de clasificación de artículos primarios publicados entre los años 2012 y 2018.

Los resultados arrojados por las bases de datos científicas con las cadenas de búsqueda aplicadas fueron 57 artículos para IEEE, 46 artículos para Science Direct y 360 artículos para ACM, de los cuales se filtraron a un total de 29 artículos mediante criterios de inclusión y exclusión centrados en si los artículos usaban metodologías de Analítica de Datos basadas en CRISP-DM, SEMMA o KDD.

Debido a la necesidad planteada en este proyecto, la revisión de literatura se divide en dos grupos: Metodologías basadas en CRISP-DM y Metodologías basadas en otros modelos de referencia como SEMMA y KDD.

3.3.1. Metodologías basadas en CRISP-DM

La primera propuesta es la de Zhu (2017), basada en CRISP-DM y Agile, y creada con 2 etapas principales: la Etapa de Reconfirmación de Requisitos (RR) y la Etapa de Refinado de Rendimiento (PR), como se ve en la Imagen 15 y se describe a continuación.

Proceso de Minería Ágil

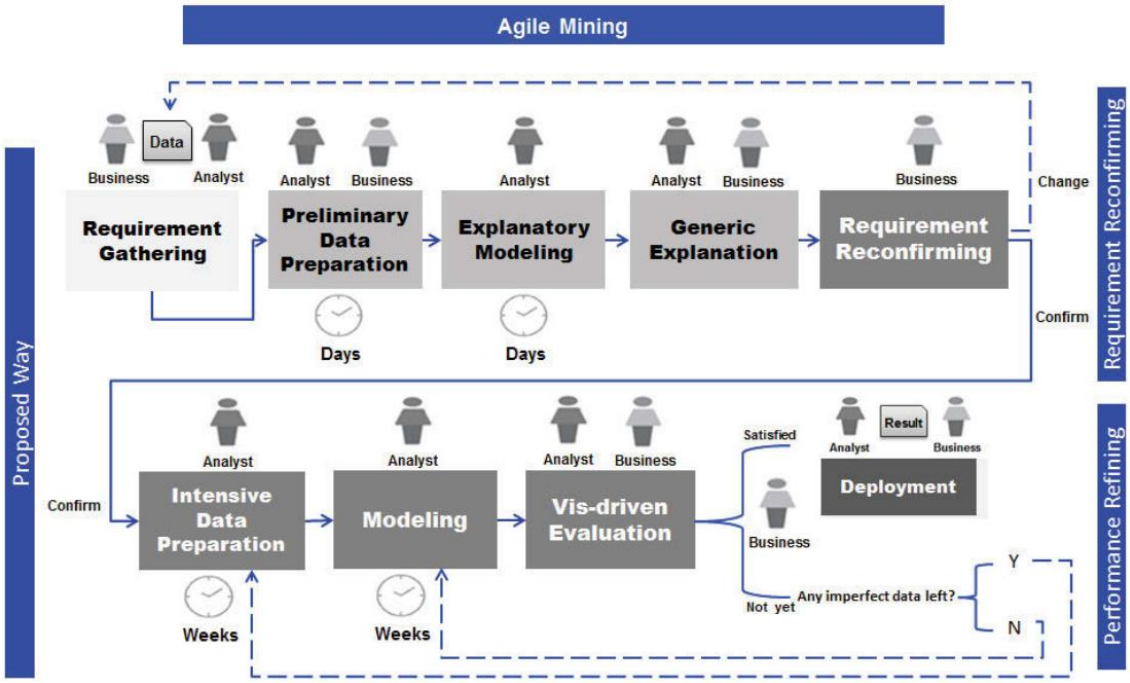


Imagen 15 Proceso de Minería Ágil.

Fuente (Zhu, 2017)

Recolección de requisitos: Recopila la descripción de las partes interesadas sobre el propósito del proyecto. La descripción podrían ser historias desde la perspectiva empresarial, pero deberían ser convertibles a objetivos analíticos individuales. Después de eso, estos objetivos serán priorizados por el gerente del proyecto y un representante de negocios de la compañía colaboradora para formar un retraso inicial del proyecto. Este representante comercial debe ser el usuario final de este proyecto entregable, de modo que las preocupaciones esenciales de la situación práctica se tengan en cuenta y atraigan una mayor prioridad. El analista luego tomará los elementos de la cartera de proyectos (PBI) con la máxima prioridad y los

dividirá en varios pasos lógicos como la preparación de la cartera de pedidos en Scrum.

Requisito de reconfirmación: los interesados y el representante comercial deciden si se desea el objetivo proporcionado para el propósito del proyecto. Según la explicación del paso anterior, las partes interesadas podrían obtener una comprensión más clara sobre cómo se lograrán los requisitos de su negocio, o qué tipo de información se puede generar a partir del objetivo analítico actual. Sobre la base de esta conciencia evolucionada, las empresas pueden juzgar mejor si el requisito inicial propuesto es lo suficientemente competente para adaptarse al propósito del proyecto.

Por otro lado se tiene la propuesta *Agile Mining Process* (Sharma *et al.*, 2017), que implica un refinamiento progresivo de los objetivos en lugar de confiar en objetivos estrechamente especificados en hitos predeterminados que no cambian una vez que se ha alcanzado un hito. Estos autores presentan un estudio de caso en un hospital que reveló tres categorías principales de decisiones que fueron tomadas por los grupos; decisiones sobre en qué problema enfocarse, cómo interpretar los datos y cómo aprovechar el ejercicio de minería de datos.

El estudio de caso involucró tres Sprints; un primer Sprint que involucró al Analista Líder y al Gerente de RI en discusiones con respecto a una especificación de problema adecuada para la minería. El segundo Sprint involucró a los tres Analistas de manera independiente, generando ideas potenciales relevantes para el problema. El tercer Sprint se produjo algún tiempo después e involucró al analista de una segunda universidad implementando un sistema predictivo. Cada Sprint fue puntuado con una presentación y reflexión sobre el progreso que involucró al personal de la organización y a los analistas, interpretando datos y deliberando sobre los próximos pasos.

ASD-DM (Alnoukari, Alzoabi, & Hanna, 2008) combina características de la metodología ASD con los pasos de solución de minería de datos de predicción. La fase de especulación incluye la comprensión comercial y de datos, y los preparativos de datos, incluidas las operaciones ETL (Extraer / Transformar / Cargar). Esta fase es la más importante, ya que requiere un tiempo y recursos considerables. Esta fase de preparación finalizará al crear el almacén de datos de la empresa y los mercados y cubos de datos requeridos. La fase de colaboración asegura la alta comunicación en una diversidad de partes interesadas experimentadas con el fin de utilizar el mejor algoritmo de modelado para el proceso de minería de datos predictivos. Las pruebas y evaluación de dichos algoritmos ocurren en la fase de "Aprendizaje", luego, los resultados son discutidos entre los miembros del equipo del proyecto, si los resultados son aceptables, se libera una nueva versión; de lo contrario, se realizará de nuevo la fase de colaboración para elegir un mejor algoritmo de minería de datos.

ASD-BI es un enfoque basado en el modelo adaptativo que involucra las fases de Especulación, Colaboración y Aprendizaje, como se muestra en la Imagen 16.

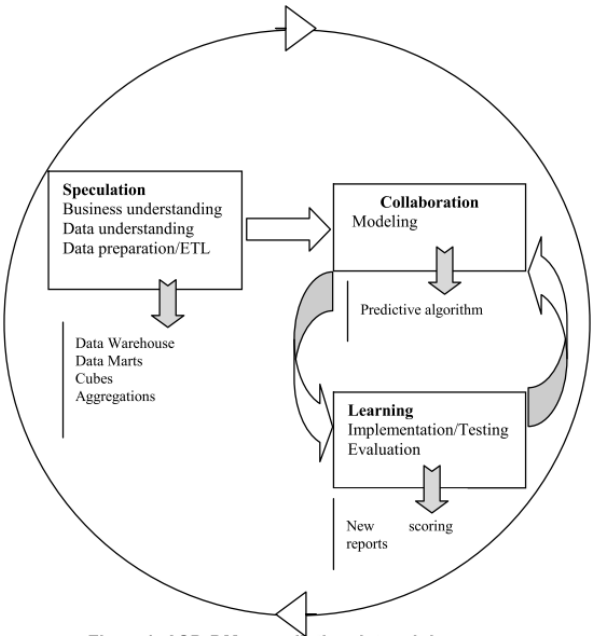


Imagen 16 ASD-DM: un marco de proceso de minería de datos predictivo basado en la metodología ASD.

Fuente (Alnoukari et al., 2008)

Alnoukari y otros (2016) anuncia una nueva versión llamada ASD-BI (Alnoukari, 2016), el cual usa la misma lista de procesos en CRISP-DM además de un proceso adicional para la configuración de objetivos e hipótesis. La lista de procesos de ASD-BI incluye: comprensión empresarial, comprensión de datos, fijación de objetivos/hipótesis, preparación de datos/ETL, modelado/minería de datos, evaluación e implementación. Categorizar los procesos en fases agrega más comprensibilidad y organización de tareas, cooperación y aprendizaje. La categorización de procesos ayudaría a analizar fases individuales por separado, establecer hitos y asignar los recursos necesarios para cada tarea dividida en el ciclo de vida de la minería de datos (cuyas etapas principales son pre procesamiento, extracción de datos y pos procesamiento) representados en la Imagen 17.

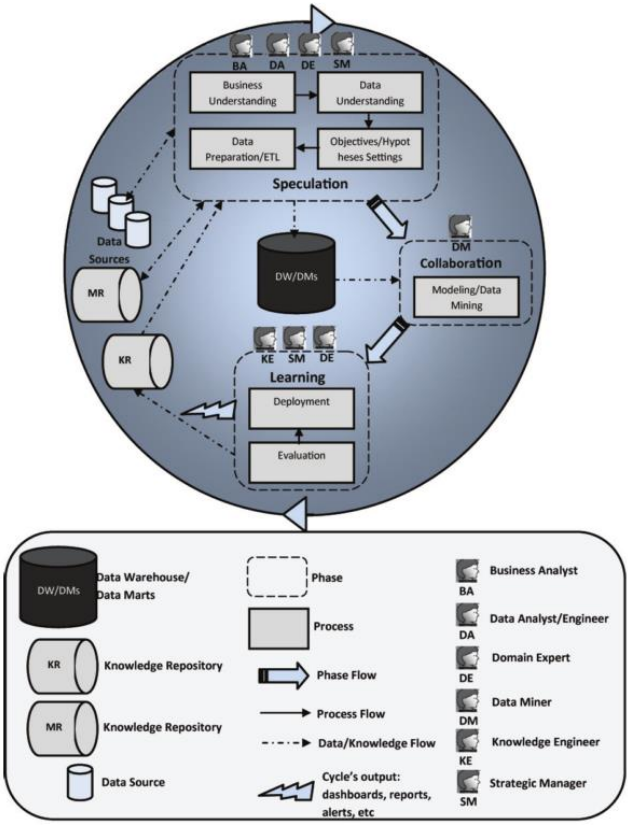


Imagen 17 Modelo de proceso ASD-BI, una descripción detallada

Fuente (Alnoukari, 2016)

Según Alnoukari (2016), ASD-BI es un marco que se ajusta a las necesidades de proyectos BI por su adaptación, trabajo de equipo colaborativo, de constante aprendizaje y dirigido bajo objetivos.

Al usar como base la metodología CRISP-DM, requiere que se incluyan roles específicos para una actividad en un contexto, por otro lado es importante mencionar que dicha propuesta aporta un gran punto de partida debido a que no necesitó eliminar ninguna fase de CRISP-DM resaltando la necesidad de cada una dentro de un proyecto de analítica y los ha embebido dentro del proceso de la metodología ágil ASD.

Por otro lado, Sharma et al. (2017) en su propuesta ágil propponen separar por sprint el desarrollo de un proyecto de analítica como si fuera un desarrollo realizado puramente con CRISP-DM. Debido a que uno de los principios de las metodologías ágiles es entregar productos de valor para el cliente en cada iteración realizada por el equipo de trabajo, divide las actividades de CRISP-DM en sprint, no me parece una manera adecuada de llamarla una propuesta ágil.

3.3.2. Metodologías basadas en otros modelos de referencia

AABA (*Architecture –centric Agile Big Data Analytics*), se centra en la arquitectura de la solución utilizada para generar valor al cliente donde los actores principales son los científicos de datos y el arquitecto de software basado en los objetivos de negocio (Chen, Kazman, & Haziyev, 2016) como se puede ver en la Imagen 18.

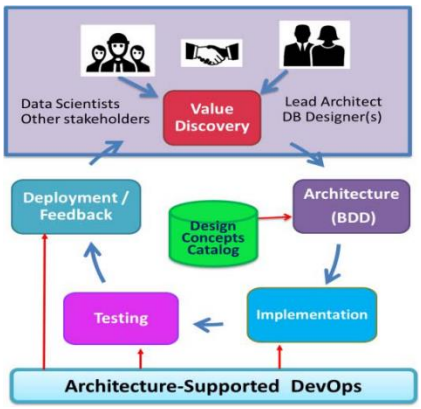


Imagen 18 Metodología AABA

Fuente (Chen et al., 2016)

AABA se diferencia de otras metodologías de desarrollo de sistemas en que comienza con una estrecha colaboración entre científicos de datos (y otras partes interesadas) y un arquitecto de software (y otros ingenieros de software clave, como un diseñador de bases de datos). Estos dos grupos deben unirse para determinar la propuesta de valor para el sistema que se está construyendo, en función de los objetivos comerciales que el sistema busca mejorar.

Agile Way of BI (Rehani, 2011) se basa en el enfoque tradicional de proyectos ágiles de software donde el *backlog* de requisitos se recibe de los usuarios. Estos requisitos se dividen en *sprints* según su prioridad y complejidad. Un sprint es un ciclo de vida completo de comprensión del requisito, análisis, diseño, construcción y prueba del usuario. Un *Sprint* tiene una duración de 1 a 2 semanas. La demostración del usuario (generalmente llamada Mostrar y Contar) se realiza después de cada *Sprint* para obtener los comentarios del usuario.

Agile Data Warehouse (Golfarelli *et al.*, 2012) propone una planificación de sprints que tiene en cuenta las principales variables que afectan la priorización de historias de usuario y la composición de sprints. En la Imagen 19 se muestra el flujo de la planificación del proyecto teniendo en cuenta los siguientes objetivos para un plan óptimo.

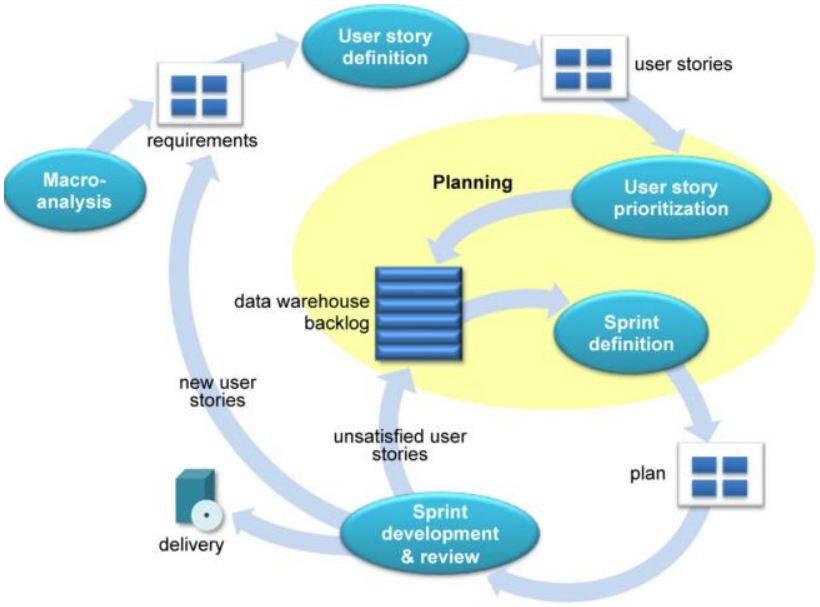


Imagen 19 Agile Framework DW

Fuente (Golfarelli et al., 2012)

1. Satisfacción del cliente. Se puede obtener entregando historias de usuarios con mayor utilidad primero. En la filosofía ágil, esto también aumenta la conciencia y la confianza del usuario.
2. Gestión de afinidad. Historias similares deben llevarse a cabo en el mismo sprint para aumentar su valor para los usuarios.
3. Gestión de riesgos. Se puede lograr (i) avanzando historias de usuarios críticas para evitar efectos secundarios tardíos, por un lado; (ii) distribuir historias inciertas en diferentes sprints y por otro lado posponerlas para reducir el riesgo de que la entrega del sprint se retrase.

(Analuisa Barona, 2016) propone una metodología que consiste en descomponer el proceso de desarrollo de *Data Warehouse* (DWH) en sub procesos o fases, identificar las tareas a ejecutar en cada fase y establecer los procedimientos para su ejecución, para lo cual se cuenta con herramientas y técnicas pre establecidas.

El proceso de desarrollo de DWH, según Kimball y Ross (2013), se descompone en las siguientes fases las cuales son consideradas en el marco metodológico presentado en la sección de **Normas de Operación** de la Imagen 20Imagen 20:

- Definición de Requerimientos del Negocio.
- Diseño de la Arquitectura Técnica.
- Selección de Productos.
- Modelo Dimensional.
- Diseño Físico.
- Diseño y Construcción del ETL.
- Especificaciones y Desarrollo de Aplicaciones BI.
- Implementación de la Solución.

Imagen 20 Estructura Procedimiento de desarrollo de productos Data WareHouse. Fuente (Analuisa Barona, 2016)

Agile BI *By Using In-Memory Analytics* (Muntean, 2014) tiene como objetivo principal reemplazar las soluciones de BI tradicionales basadas en disco. Las diferencias importantes entre ellos son: velocidad, volumen, persistencia y precio.

Las tecnologías de BI en memoria cargan todo el conjunto de datos en la RAM antes de que los usuarios puedan ejecutar una consulta. Además, la mayoría de ellos puede ahorrar un tiempo de desarrollo significativo al eliminar la necesidad de agregar y diseñar cubos y esquemas de estrella. Basados en el flujo de la Imagen 21 se pueden tomar decisiones que aporten mejor entendimiento de los problemas de negocio en la organización.

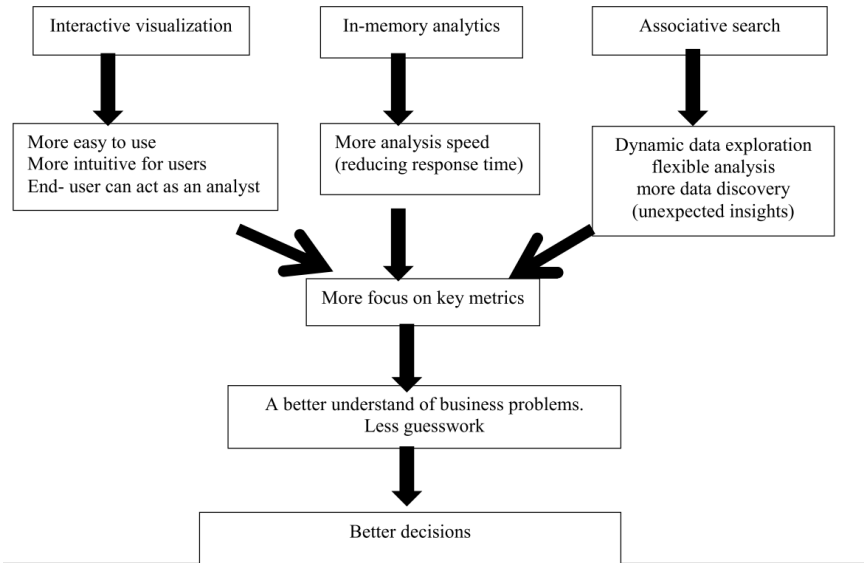


Imagen 21 Cómo la analítica en memoria, la visualización interactiva y la búsqueda asociativa afectan a las empresas

Fuente (Muntean, 2014)

Diversas propuestas se han desarrollado para implementar metodologías ágiles en proyectos de analítica de datos en entornos ágiles, con elementos importantes que muestran que CRISP-DM es la metodología con mayor número de actividades en las fases que se propone, lo cual es óptimo para modificar y adaptar a proyectos más específicos sin perder control de los recursos dentro de un equipo de trabajo.

En la Tabla 4 se muestra una comparación de las metodologías más relevantes encontradas en esta investigación:

Comparación de Metodologías Encontradas

Título	Año	Tiene caso de estudio	Metodología de Analítica Base	Metodología Ágil
(Zhu, 2017)	2017	SI	CRISP-DM	NO
(Krawatzeck & Dinter, 2015)	2015	NO	NO	ASD
(Rehani, 2011)	2011	NO	NO	Scrum
(Sharma et al., 2017)	2017	SI	CRISP-DM	NO
(Golfarelli et al., 2012)	2012	SI	NO	Scrum
(Analuisa Barona, 2016)	2016	SI	NO	SCRUM
(Muntean, 2014)	2014	NO	NO	SCRUM
(Alnoukari, 2016)	2016	NO	CRISP-DM	ASD

Tabla 4 Comparación de Metodologías Encontradas Elaboración Propia

Golfarelli et al. (2012) tiene una interesante propuesta en la fase de planificación basada en Scrum, por lo que ha sido un importante punto de partida para identificar la manera de organizar un proyecto de analítica de datos en entornos ágiles, resaltando la importancia de las ceremonias de Scrum y una definición adecuada para las historias de usuario.

Por su parte (Alnoukari, 2016) en su propuesta de incluir dentro de un marco de desarrollo ágil como ASD, al tener como referencia SCRUM en esta propuesta, impone una manera óptima de extender CRISP-DM hacia entornos ágiles en proyectos medianos de analítica de datos.

En consecuencia, aunque no se cuenta con suficientes casos de estudio realizados incorporando SCRUM como metodología de desarrollo ágil en los proyectos de Analítica de Datos, es pertinente demostrar el gran aporte que se puede realizar en

los equipos de desarrollo donde existe la necesidad de un trabajo colaborativo y de constante comunicación (Rehani, 2011).

Según el nivel de analítica los métodos de minería de datos que se pueden aplicar están agrupados por descripción, clasificación, regresión o clustering, abarcando todos los niveles de proyectos de analítica (descriptiva, diagnóstica, predictiva y prescriptiva), por lo tanto, usar CRISP-DM como modelo de referencia aplica tanto para proyectos de analítica descriptiva como para proyectos de analítica predictiva.

El modelo descriptivo se puede definir para descubrir regularidades interesantes en los datos, para descubrir patrones y encontrar subgrupos interesantes en la mayor parte de los datos. Agrupación, Métodos Estadísticos, pruebas de hipótesis, correlación de datos, son algunos de los métodos de minería de datos usados en este nivel y son representados mediante reportes o gráficos (Agyapong *et al.,* 2016).

El propósito del modelo predictivo es determinar el resultado futuro en lugar del comportamiento actual. Su salida puede ser categórica o de valor numérico. Árboles de Regresión, Redes Neuronales, Clasificación, Series de Tiempo, son algunos de los métodos usados y son representados mediante modelos de predicción (Agyapong *et al*., 2016).

Debido a que la minería de datos contiene métodos que pueden ser aplicados en analítica descriptiva, es factible usar CRISP-DM como metodología de referencia para este tipo de proyectos.

PARTE III
CONSTRUCCIÓN

“Permanencia, perseverancia y persistencia a pesar de todos los obstáculos, desalientos e imposibilidades: Es esto, que en todas las cosas se distingue el alma fuerte de la débil.” -- Thomas Carlyle

CAPÍTULO 4

Definición de Aspectos Estáticos

Para describir con mayor rigor una metodología basados en SPEM por sus siglas en inglés *Software Process Engineering Methods*, (Henderson-Sellers & Ralyté, 2010) propone definir aspectos estáticos como actores y artefactos que se van a mantener estables en el desarrollo de un proyecto.

4.1. Actores

Los actores son elementos que representan el rol de una persona que se encarga de realizar actividades del proceso. CRISP-DM no tiene roles definidos, pero se consideran aquí los roles actuales desempeñados en el campo de analítica de datos, como: Científico de Datos, Analista de Datos, Arquitecto de Datos, Ingeniero de Datos, Estadístico, Administrador de Base de Datos, Analista de Negocios, Administrador de Analítica y Datos los cuales son responsables de desarrollar modelos, explicaciones, probar y proponer hipótesis utilizando métodos analíticos. A continuación, se describe cada uno.

Scrum Data Product Owner representa los intereses de la comunidad de socios para el equipo Scrum. Este rol es responsable de garantizar una comunicación clara sobre el producto y los requisitos de funcionalidad del servicio con el equipo Scrum, al igual que definir los criterios de aceptación, y asegurar que se cumplan dichos criterios. En otras palabras, el propietario del producto es responsable de asegurar que el equipo Scrum ofrezca valor.

Scrum Data Master es el líder servicial del equipo Scrum, y es quien modera y facilita las interacciones del equipo como entrenador y motivador del mismo. Este rol es responsable de asegurarse que el equipo tenga un ambiente de trabajo productivo al protegerlo de influencias externas, eliminando todos los obstáculos, y confirmando que se cumplan los principios, aspectos y procesos de Scrum.

Scrum Data Analyst es el responsable del desarrollo del producto, servicio o de cualquier otro resultado. Consiste en una persona o grupo de personas que trabajan en las historias de usuario y en la lista de pendientes del sprint para crear los entregables del proyecto.

Scrum Data Stakeholders son aquellas personas que no intervienen directamente en el producto, pero están involucradas de diferentes maneras como los socios del proyecto, los usuarios finales o los patrocinadores.

Scrum Data Architect es un ingeniero de datos con una amplia experiencia en varias tecnologías distribuidas, y que también tiene una buena comprensión de conceptos de arquitectura orientados a servicios y aplicaciones web (conceptos SOA y marcos REST) además de los conjuntos de habilidades del desarrollador (Biguru, 2015).

4.2. Artefactos

Son los elementos que representan el resultado de una actividad realizada por un actor, la cual se puede evidenciar mediante algunos reportes a nivel de gráficas o documentos.

4.2.1. Artefactos de gestión

- **Kanban**. Para maximizar la capacidad de un equipo para entregar software de alta calidad de manera consistente, Kanban se enfatiza en dos prácticas

principales. La primera, visualizar el flujo de trabajo, requiere que mapee las etapas del flujo de trabajo de su equipo y configure su tablero Kanban para que coincida con todas las actividades del equipo. El segundo, restringir la cantidad de trabajo en progreso, requiere que establezca límites de trabajo en curso (KathrynEE, Danielson Steve, & Erickson Doug, 2018). En la Imagen 22 se muestra un tablero Kanban mínimo usado en la industria del software para diferentes proyectos.

Imagen 22 Kanban Básico

Fuente (Kanbantool, n.d.)

Un tablero Kanban convierte el *backlog* en un letrero interactivo, proporcionando un flujo visual de trabajo. A medida que el trabajo avanza desde la idea hasta su finalización, actualiza los elementos en la pizarra. Cada columna representa una etapa de trabajo, y cada tarjeta representa una historia de usuario (tarjetas azules) o un error (tarjetas rojas) en esa etapa de trabajo (KathrynEE et al., 2018). Como se puede ver en la Imagen 23, el Kanban lo proporcionan diferentes aplicaciones de gestión de proyectos para llevar seguimiento de las actividades designadas para terminar una historia de usuario.

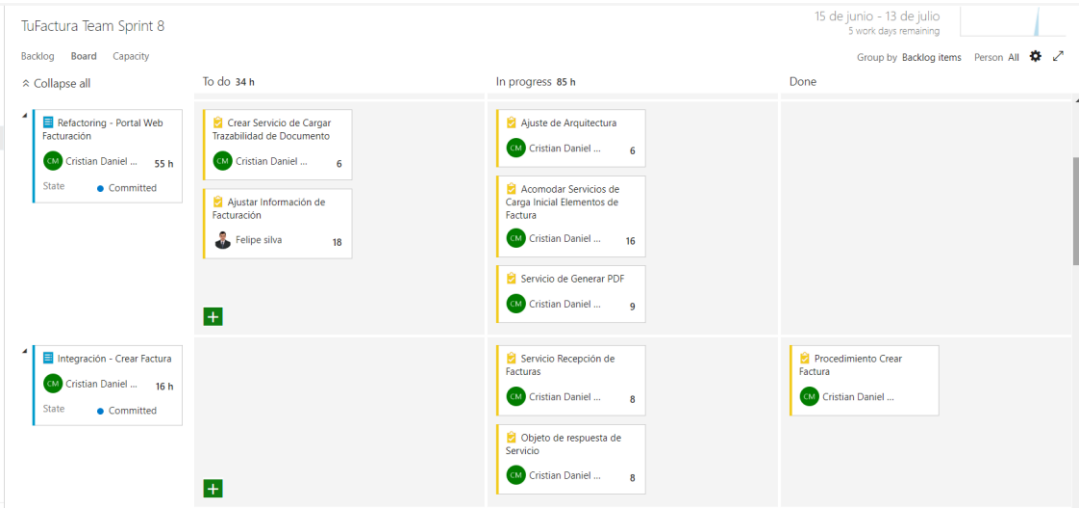


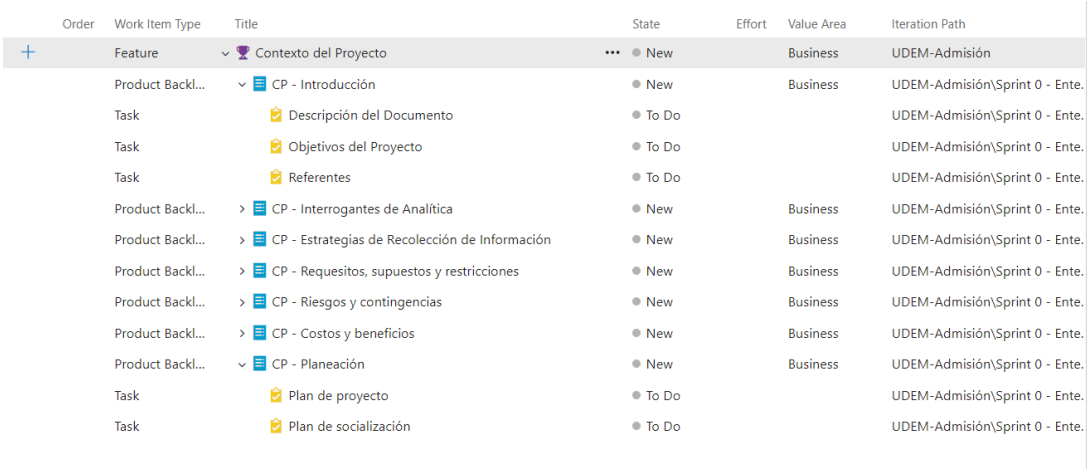
Imagen 23 Kanban en TFS

Fuente Elaboración Propia

- ***Product Backlog*.** Como se describe en la Guía de Scrum (Schwaber, 2017), es una lista ordenada de todo lo que se sabe que se necesita en el producto. Es la única fuente de requisitos para cualquier cambio que se realice en el producto. El propietario del producto es responsable de la acumulación de historias de usuario, incluido su contenido, disponibilidad y pedido.

  Un *product backlog* nunca está completo. Su desarrollo más temprano establece los requisitos inicialmente conocidos y mejor comprendidos. El *Product Backlog* evoluciona a medida que evoluciona el producto y el entorno en el que se utilizará, es dinámico, cambia constantemente para identificar que el producto sea apropiado, competitivo y útil. Si existe un producto, también existe su *Product Backlog*.

  En la Imagen 24 semuestra un ejemplo de un *product backlog*, el cual muestra el listado de las historias de usuario creadas para un proyecto



Order	Work Item Type	Title	State	Effort	Value Area	Iteration Path
	Feature	Contexto del Proyecto	New		Business	UDEM-Admisión
	Product Backl...	CP - Introducción	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Task	Descripción del Documento	To Do			UDEM-Admisión\Sprint 0 - Ente.
	Task	Objetivos del Proyecto	To Do			UDEM-Admisión\Sprint 0 - Ente.
	Task	Referentes	To Do			UDEM-Admisión\Sprint 0 - Ente.
	Product Backl...	CP - Interrogantes de Analítica	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Product Backl...	CP - Estrategias de Recolección de Información	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Product Backl...	CP - Requesitos, supuestos y restricciones	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Product Backl...	CP - Riesgos y contingencias	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Product Backl...	CP - Costos y beneficios	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Product Backl...	CP - Planeación	New		Business	UDEM-Admisión\Sprint 0 - Ente.
	Task	Plan de proyecto	To Do			UDEM-Admisión\Sprint 0 - Ente.
	Task	Plan de socialización	To Do			UDEM-Admisión\Sprint 0 - Ente.

Imagen 24 Product Backlog en TFS

Fuente Elaboración Propia

- **Backlog de Sprint,** es el conjunto de elementos del *product backlog* seleccionados para el Sprint, más un plan para entregar el Incremento del producto y lograr el Objetivo del Sprint. El *Backlog* de *Sprint* es una previsión del Equipo de Desarrollo sobre qué funcionalidad habrá en la próxima entrega y el trabajo necesario para entregar esa funcionalidad totalmente terminada para que esté en estado “Hecho”.

- **Gráfica de quemado (*Burn Down Chart*),** es un gran gráfico que relaciona la cantidad de trabajo restante (en el eje vertical) y el tiempo transcurrido desde el inicio del proyecto o un sprint (en la horizontal, mostrando el futuro y el pasado).

  Muestra cuánto trabajo quedaba al final de los intervalos especificados durante un sprint. La fuente de los datos en bruto es la acumulación de sprint. El eje horizontal muestra los días en un sprint, y el eje vertical mide la cantidad de trabajo que queda para completar las tareas en el sprint. El trabajo que queda se muestra en horas.

- **Historia de usuario.** Las historias de usuario se apegan a una estructura específica predefinida y son una forma simple de documentar los requerimientos y funcionalidades que desea el usuario final (SCRUMstudy™, 2016). Generalmente responden preguntas como ¿Qué valor genera al usuario de esta solución? Para Analítica debe identificar ¿qué pregunta de analítica la solución puede ayudar a alguien a responder? o ¿Qué decisión de negocios podría soportar la solución?

  En la Imagen 25 se presenta un modelo de formulario para diligenciar una historia de usuario en TFS.

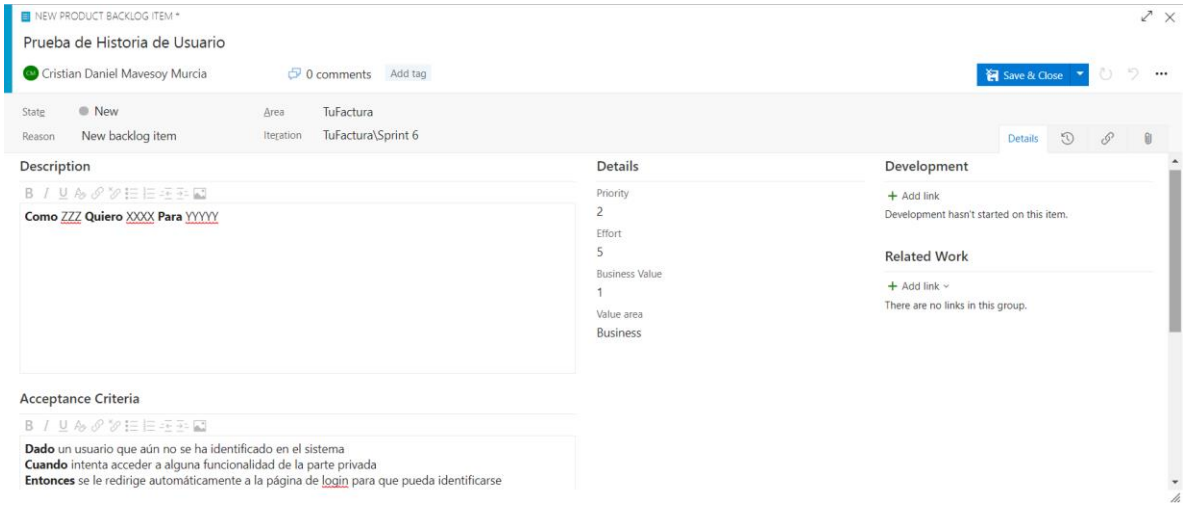


Imagen 25 Historia de Usuario TFS

Fuente Elaboración Propia

4.2.2. Artefactos de analítica

- Contexto del proyecto. Reúne un conjunto de entregables propuestos para CRISP-DM de la fase de Entendimiento de Negocio usados para la metodología propuesta.

- Modelo de analítica. Consigna el resultado de las actividades realizadas en la construcción de los tableros de control o Dashboards y/o reportes del proyecto.
- Diccionario de datos. Contiene un listado de todas las conexiones, tablas, columnas y/o objetos de datos usados para el proyecto.
- Implementación. Especifica los elementos usados para que el proyecto de analítica funcione como servidores, aplicaciones, interfaces de comunicación, puede ser representado mediante un Diagrama de Despliegue de UML, además de información sobre el acceso a las fuentes de datos dispuestas para el proyecto.

CAPÍTULO 5

Definición de Aspectos Dinámicos

Para describir con mayor rigor una metodología basados en SPEM, Henderson-Sellers & Ralyté (2010) propone definir aspectos dinámicos para el desarrollo de un proyecto, como fases y actividades. Estos especifican el comportamiento de un proyecto realizado por unos individuos para la generación de entregables.

5.1. Fases del proceso

A continuación, se describen las fases propuestas con Chapman y otros (2000):

- **Entendimiento del Negocio.** Esta fase inicial se encarga de entender los objetivos del proyecto y requisitos desde una perspectiva de negocio, luego se convierte este conocimiento en una definición de problema de minería de datos, para finalmente diseñar un plan preliminar para lograr estos objetivos.

- **Entendimiento de los Datos**. Inicia con la recolección de datos iniciales y continua con actividades que ayudan a familiarizarse con los datos, identificar problemas de calidad de datos, descubrir primeras ideas sobre los datos y/o detectar subconjuntos interesantes para formar hipótesis con respecto a la información oculta.

- **Preparación de los Datos**. La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (datos que se incorporarán a la(s) herramienta(s) de modelado) a partir de los datos brutos iniciales. Es probable que las tareas de preparación de datos se realicen

varias veces y no en ningún orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para herramientas de modelado.

- **Modelamiento**. En esta fase se seleccionan y aplican diversas técnicas de modelado, y sus parámetros se calibran a valores óptimos. Normalmente, hay varias técnicas para el mismo tipo de problema de minería de datos, algunas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, volver a la fase de preparación de datos a menudo es necesario.

- **Evaluación**. En esta etapa se ha creado un modelo (o modelos) que parece tener alta calidad desde una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, para asegurarse de que el modelo logre los objetivos de negocio correctamente. Un objetivo clave es determinar si hay algún problema de negocio importante que no se haya tenido suficientemente en cuenta. Al final de esta fase, se debe alcanzar una decisión sobre el uso de los resultados de la minería de datos.

- **Despliegue**. La creación del modelo generalmente no es el final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de forma que el cliente pueda utilizarlo. A menudo implica la aplicación de modelos "en vivo" dentro de los procesos de toma de decisiones de una organización, por ejemplo, la personalización en tiempo real de las páginas web o la puntuación repetida de las bases de datos de marketing. Según los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso repetible de minería de datos en toda la empresa.

5.2. Actividades por fases del proceso

CRISP-DM describe ciertas actividades sin importar el tipo de proyecto de analítica que se lleve a cabo, pero en este caso, se requiere mostrar las actividades específicas para un proyecto mediano de analítica de datos en entornos ágiles para el desarrollo de tableros de control o *Dashboards*.

Existen dos tipos de actividades que se realizan en el desarrollo de un entregable: de producto y de proceso. Las actividades de producto describen lo que se debe realizar para la construcción de los entregables dependiendo de la fase donde se encuentre. Las actividades de proceso se realizan de manera paralela a la construcción del producto para tener control de los recursos asignados.

Durante los siguientes apartados, se presenta una imagen que define las actividades por fase tomadas del documento oficial de CRISP-DM (Chapman *et al.* 2000), para luego hacer una selección ajustando la necesidad particular de proyectos medianos de tableros de control, las cuales son actividades de producto y algunas actividades puntuales de SCRUM o XP (*eXtreme Programming*) correspondientes a las actividades de proceso, por último, al final de cada punto se mostrarán las actividades finales para CRISP-DM Ágil.

5.2.1. Actividades para entendimiento de negocio

Este grupo de actividades especifica lo que debe realizarse al inicio de un proyecto para tener un punto de partida de lo que se quiere realizar, con qué recursos se cuenta y cuál será el alcance en función de las necesidades de la organización para gestionar su información. En la Imagen 26 se representan las actividades genéricas de CRISP-DM que son el punto inicial para sacar las más importantes en la propuesta metodológica.

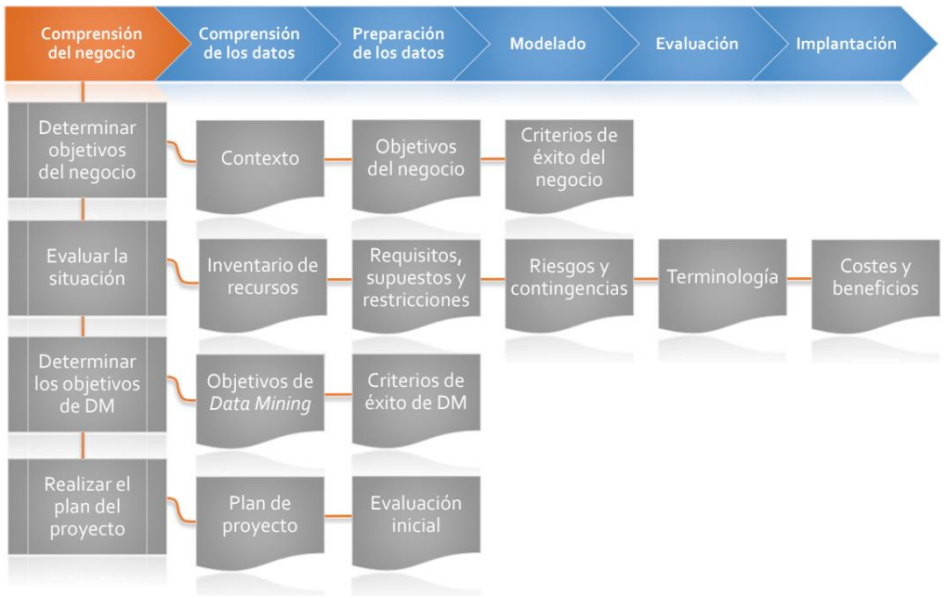


Imagen 26 Actividades CRISP-DM Fase 1

Fuente (Chapman *et al*., 2000)

Actividades de CRISP-DM:

- Determinar objetivos del Negocio. Se definen los objetivos generales y/o específicos para el proyecto de analítica.
- Evaluar la situación. Implica una búsqueda de datos más detallada sobre todos los recursos, restricciones, suposiciones y otros factores que deben considerarse al determinar el objetivo del análisis de datos y el plan del proyecto (Chapman *et al*., 2000).
- Determinar objetivos de Minería de Datos. Un objetivo de negocio establece objetivos en la terminología de negocios. Un objetivo de minería de datos establece los objetivos del proyecto en términos técnicos. Por ejemplo, el objetivo de negocio podría ser "Aumentar las ventas por catálogo a los clientes actuales". Un objetivo de minería de datos podría ser "Predecir cuántos widgets comprará un cliente, dadas sus compras en los últimos tres años, información demográfica (edad, salario, ciudad")., etc.), y precio del artículo" (Chapman *et al*., 2000).
- Realizar Plan de Proyecto. Incluye las siguientes actividades:

- Crear Historias de Usuario

- Plan de Proyecto. Definir un *backlog* de producto con todas las actividades a realizar de forma organizada y priorizada.

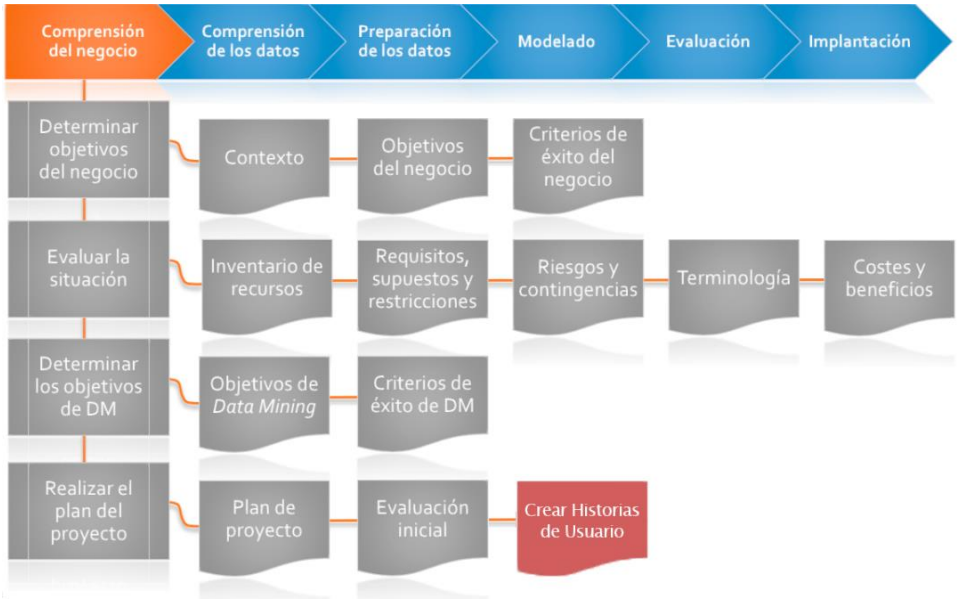


Imagen 27 Actividades CRISP-DM Ágil Fase 1. Fuente. Elaboración Propia

5.2.2. Actividades para entendimiento de los datos

Esta fase describe las actividades requeridas para identificar las fuentes de datos que se usarán para generar los entregables en cada sprint del proyecto, la Imagen 28 muestra las actividades base para la construcción de CRISP-DM ÁGIL.

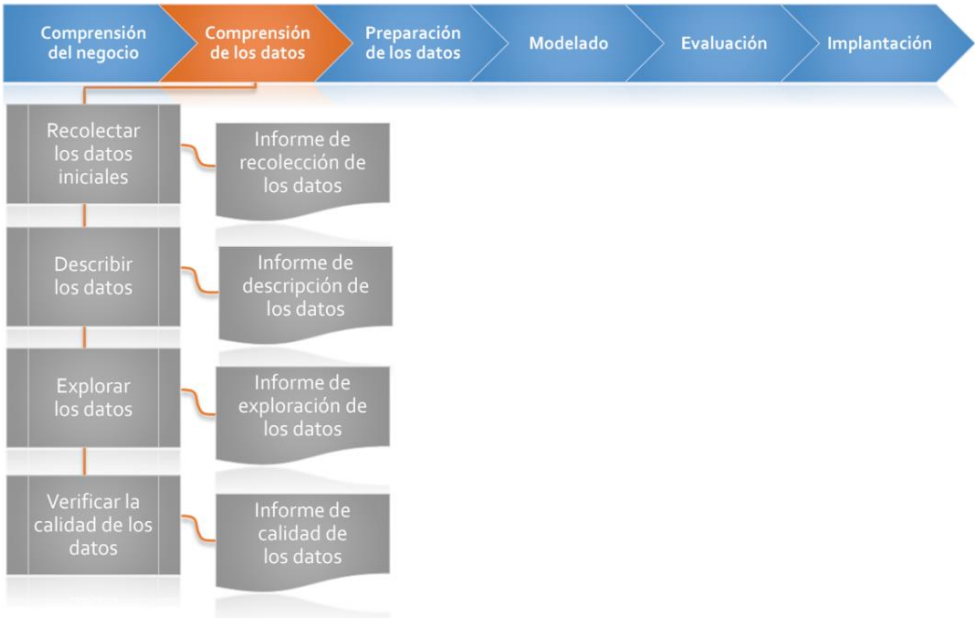


Imagen 28 Actividades CRISP-DM Fase 2

Fuente (Chapman *et al.*, 2000)

Actividades ágiles:

- Crear *Backlog* de Sprint. Según la capacidad del equipo de trabajo, se listan las historias de usuario priorizadas que serán ejecutadas en la próxima iteración o sprint.
- Recolectar y Describir los Datos. Se adquieren los datos (o acceso a los datos) enumerados en los recursos del proyecto, se examinan las propiedades "brutas" o "superficiales" de los datos adquiridos y se informa sobre los resultados (Chapman *et al*., 2000).
- Verificar la Calidad de los datos. Se examina la calidad de los datos, abordando preguntas como: ¿Están completos los datos (cubren todos los casos requeridos)? ¿Es correcto o contiene errores y, si hay errores, qué tan comunes son? ¿Hay valores faltantes en los datos? Si es así, ¿cómo se representan, ¿dónde se producen y qué tan comunes son? (Chapman *et al.,* 2000).
- Seleccionar los datos. Decidir sobre los datos que se utilizarán para el análisis. Los criterios incluyen relevancia para los objetivos de minería de

datos, calidad y restricciones técnicas, como los límites en el volumen de datos o los tipos de datos (Chapman *et al.,* 2000).

- Listar Indicadores de Negocio. Basado en los objetivos de minería de datos, se relacionan los indicadores de negocio (KPI) a utilizar para la visualización en los diferentes tableros de control creados.

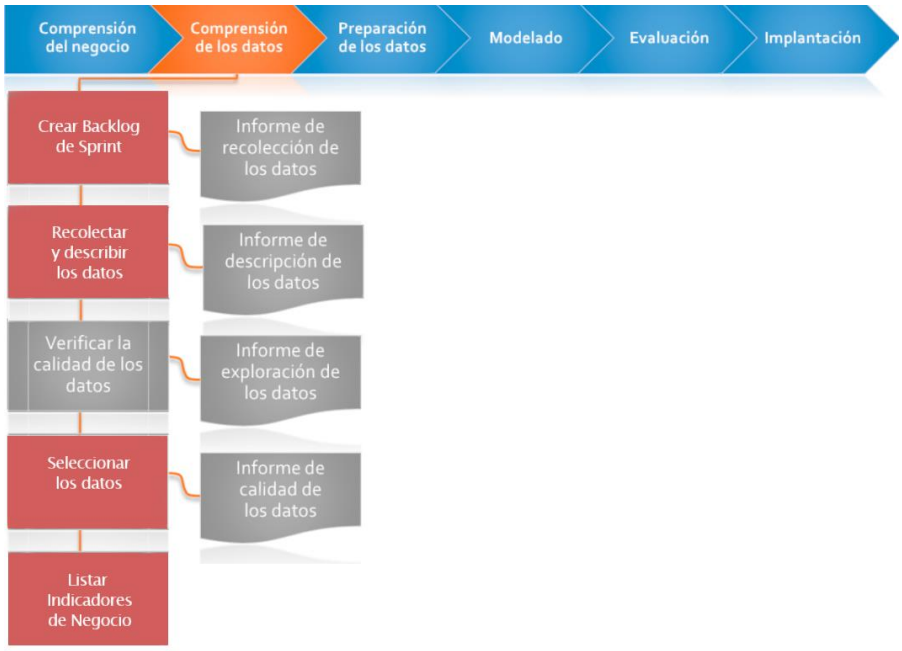


Imagen 29 Actividades CRISP-DM Ágil Fase 2. Fuente. Elaboración Propia

5.2.3. Actividades para preparación de los datos

En la Imagen 30 se muestran las actividades propuestas por Chapman para realizar una transformación a los datos de acuerdo a las necesidades de negocio identificadas en el *sprint planning*.

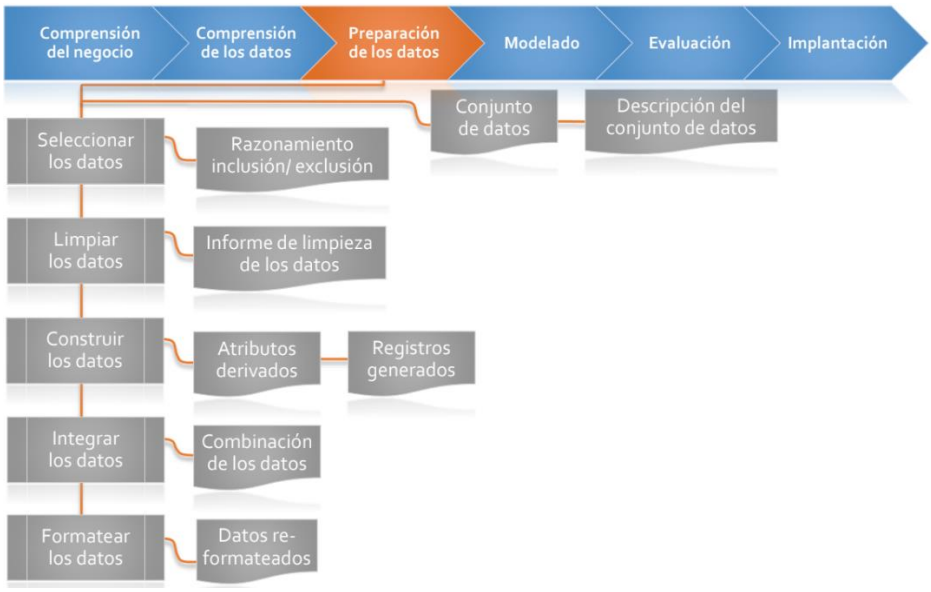


Imagen 30 Actividades CRISP-DM Fase 3

Fuente (Chapman et al., 2000)

- Limpiar los datos. Se aumenta la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de los datos, la inserción de valores predeterminados adecuados o técnicas más ambiciosas, como la estimación de los datos faltantes mediante el modelado (Chapman *et al*., 2000).
- Construir e integrar los datos. Estos son métodos mediante los cuales la información se combina de varias tablas o registros para crear nuevos registros o valores.
- Construir *Mock-ups*. Se realiza una vista previa del reporte o *dashboard* para ser evaluado por el usuario final antes de su implementación.
- Realizar Sprint Actual. El Equipo de trabajo se encarga de realizar las tareas asignadas en el *Sprint backlog*.

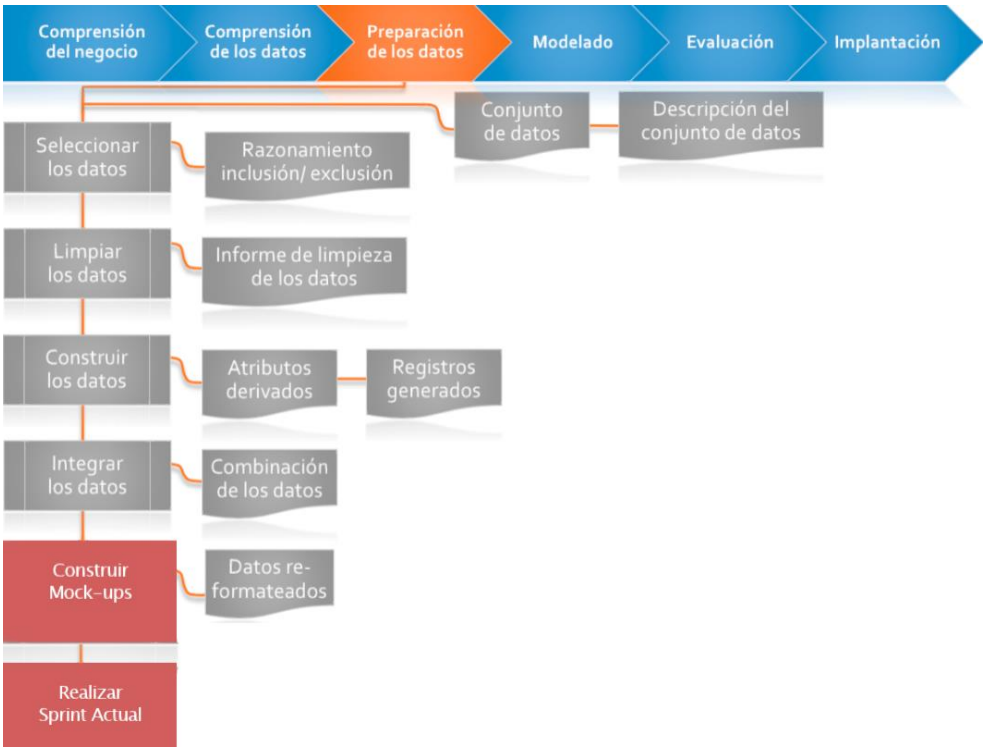


Imagen 31 Actividades CRISP-DM Ágil Fase 3. Fuente. Elaboración Propia

5.2.4. Actividades para la fase de modelamiento

Las actividades consignadas en esta fase proponen la forma en que se va a generar el modelo de datos que alimentará las visualizaciones diseñadas como los Dashboards o reportes según se ha definido en las historias de usuario del proyecto desarrollado. Chapman propone usar las actividades que se presentan en la Imagen 32.

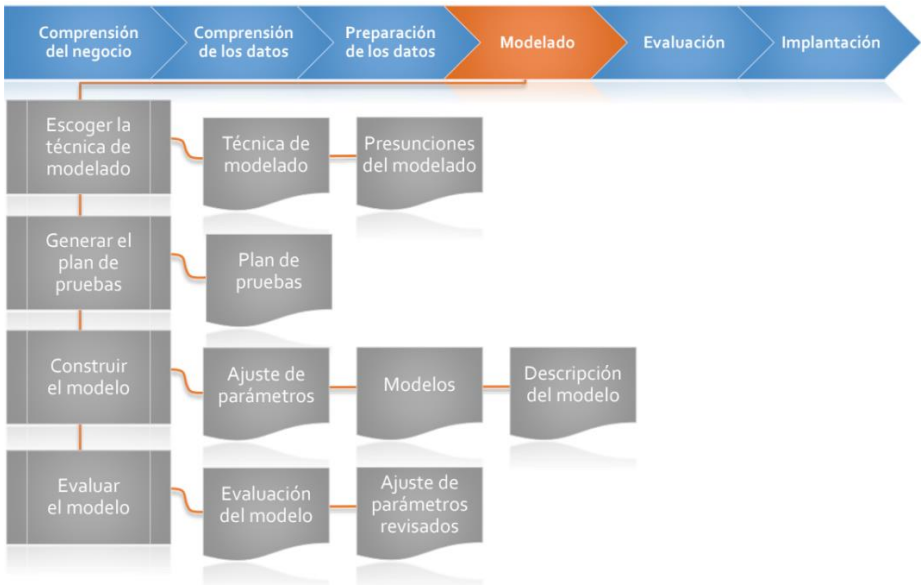


Imagen 32 Actividades CRISP-DM Fase 4

Fuente (Chapman et al., 2000)

- Crear Esquema de Visualización. Basado en los datos y las técnicas seleccionadas, se procede a especificar los indicadores a usar, los reportes y/o los tableros de control, según los entregables definidos por el sprint (Chapman et al., 2000).
    - Análisis de Indicadores. Lista de los indicadores clave de desempeño (KPI: *Key Performance Indicator*), con la explicación de cada uno. Es importante destacar la naturaleza del indicador, el origen de los datos necesarios para calcularlo, la forma en la que se realizará el cálculo y la meta que se pretende alcanzar inicialmente para el indicador.
    - Análisis de reportes. Lista de los reportes a realizar con la explicación de cada uno. Es importante destacar el tipo de reporte (listado, matriz, rompimiento, etc.), el origen de los datos necesarios para generarlo y cómo se puede realizar el *drill down,* técnica usada para ampliar el detalle de los elementos presentados en el reporte.
    - Diseño de Reportes (*Reporting*). Imágenes de los prototipos de los reportes ilustrando en lo posible datos reales o aproximados de la

situación actual y la forma de ampliar el detalle de algunos elementos del reporte.

- Diseño de Tableros (*Dashboard*). Imagen del prototipo del tablero de control ilustrando en lo posible datos reales o aproximados de la situación actual y la forma de ampliar el detalle de algunos elementos del tablero de control.

- Seleccionar técnicas de presentación de datos. Como primer paso en el modelado, se selecciona la técnica de presentación real que se va a utilizar. Aunque es posible que ya haya seleccionada una herramienta durante la fase de 'Entendimiento de negocio', esta tarea se refiere a la técnica de presentación específica, por ejemplo, usar un diagrama de barras o definir los niveles de los KPI. Si se aplican múltiples técnicas, se realiza esta tarea por separado para cada técnica (Chapman *et al*., 2000).
- Construir los tableros de control o Dashboards. Se construye el esquema de visualización basado en el *mock-up* aprobado por el usuario, para mostrar los datos creados previamente.
- Evaluar los tableros de control o *Dashboards*. Se socializa el producto con el cliente para que tenga interacción con el entregable y pueda retroalimentar según lo vea necesario de acuerdo a su juicio de expertos como dueños del negocio.
- Realizar *Daily Stand-Up*. Reunión diaria para el seguimiento del equipo de trabajo y actualización del estado de las actividades en el Kanban.

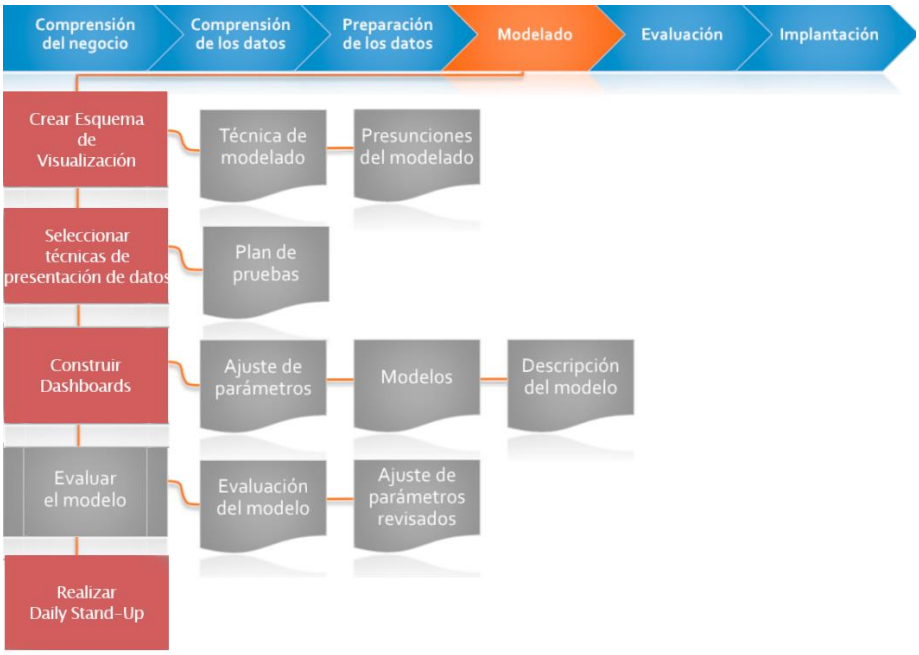


Imagen 33 Actividades CRISP-DM Ágil Fase 4. Fuente. Elaboración Propia.

5.2.5. Actividades para la fase de evaluación

Además de evaluar el producto, también se requiere evaluar el proceso realizado por parte de los integrantes del proyecto para identificar oportunidades de mejora en cada iteración de tiempo, Chapman propone las actividades de la Imagen 34:

Imagen 34 Actividades CRISP-DM Fase 5

Fuente (Chapman et al., 2000)

- Evaluar los resultados. Se socializa el entregable con todos los stakeholders para determinar la veracidad de la información suministrada y adaptarse a los objetivos definidos en la fase 'Entendimiento de los Datos'.
- Revisión de Sprint. Se realiza a nivel de proceso una entrega formal del entregable mediante una ceremonia de socialización.
- Determinar próximos pasos. Dependiendo de los resultados de la evaluación y la revisión del proceso, el equipo del proyecto decide cómo proceder. El equipo decide si terminar el proyecto y continuar con la implementación, iniciar nuevas iteraciones o configurar nuevos proyectos de minería de datos (Chapman *et al*., 2000).

Imagen 35 Actividades CRISP-DM Ágil Fase 5. Fuente. Elaboración Propia.

5.2.6. Actividades para la fase de implantación

En la última fase de CRISP-DM proponen actividades para el cierre del proyecto. En la Imagen 36 se muestran las actividades que se realizan para finalizar formalmente el proyecto.

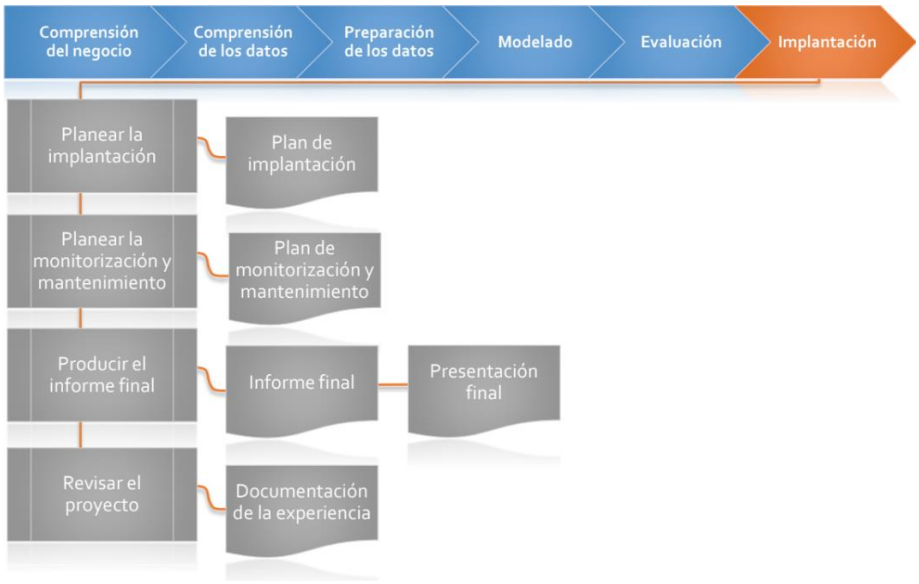


Imagen 36 Actividades CRISP-DM Fase 6

Fuente (Chapman et al., 2000)

- Despliegue del plan. Esta tarea toma los resultados de la evaluación y determina una estrategia para el despliegue. Si se identificó un procedimiento general para crear los modelos relevantes, este procedimiento se documenta en el *backlog* de producto para su posterior implementación (Chapman et al., 2000).
- Retrospectiva de Sprint. Es una oportunidad para que el equipo se inspeccione a sí mismo y cree un plan para que se realicen mejoras durante el próximo Sprint. El propósito de la Retrospectiva de Sprint es:

  • Inspeccionar cómo fue el último Sprint con respecto a las personas, las relaciones, el proceso y las herramientas;

  • Identificar y ordenar los elementos principales que salieron bien y las mejoras potenciales; y,

  • Crear un plan para implementar mejoras en la forma en que el equipo hace su trabajo (Schwaber, 2017).

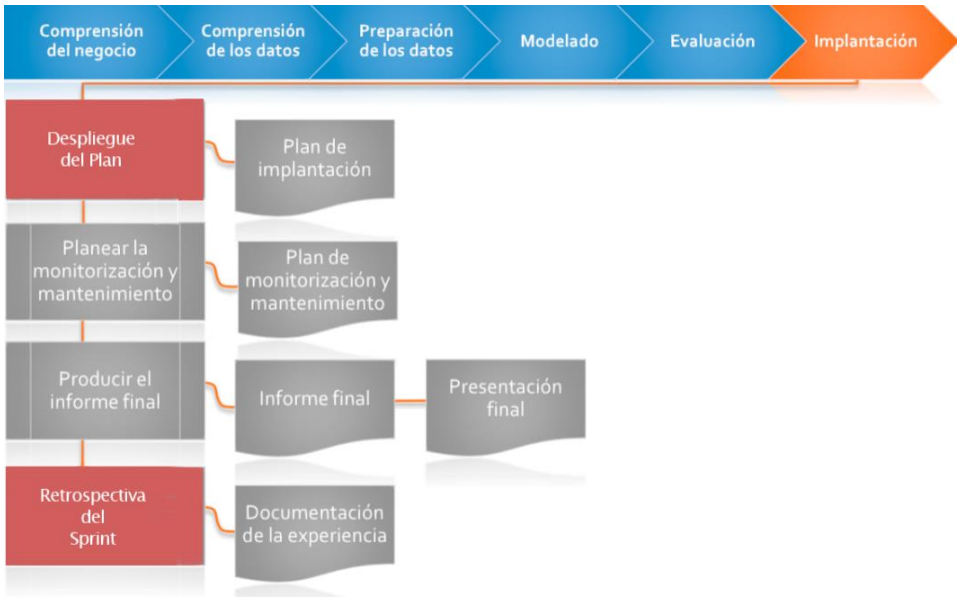


Imagen 37 Actividades CRISP-DM Ágil Fase 6. Fuente. Elaboración Propia

CAPÍTULO 6

Elementos de la Metodología CRISP-DM Ágil

6.1. Consideraciones para articular los aspectos estáticos y dinámicos

De acuerdo con Menéndez Domínguez & Castellanos Bolaños (2008), SPEM propone unos elementos mínimos para definir una metodología de trabajo los cuales se mencionan en la Tabla XX.

Elementos de CRISP-DM ÁGIL

Elemento	SPEM	DESCRIPCIÓN
Fases		Son las etapas por las que va a pasar el equipo de trabajo durante el desarrollo de un proyecto de analítica de datos.
Roles		Un Rol (*Role Definition*) define un conjunto de habilidades, competencias y responsabilidades relacionadas, de un individuo o de un grupo. No se deben confundir roles con personas, ya que la vinculación entre personas y roles se realiza durante la planificación del proyecto y puede ocurrir que un individuo desempeñe varios roles o que un rol sea desempeñado por varios individuos. Un rol es un Elemento de Método usado en las Definiciones de Tareas para señalar quiénes las realizan (Ruiz & Verdugo, 2008).
Actividades		Describen una parte del trabajo desarrollado por un rol, las tareas, operaciones y acciones que son desempeñadas por un rol o las que el rol puede asistir. Una Actividad se puede componer de elementos atómicos llamados pasos (Menéndez Domínguez & Castellanos Bolaños, 2008).
Artefactos o productos de trabajo		Corresponden a las evidencias o los entregables realizados durante la actividad o actividades asociadas.
Líneas de Flujo		Permite representar la relación entre dos elementos, ya sea una asociación de acción, de producir un elemento o de hacer uso de algo.

Tabla 5 Elementos de CRISP-DM ÁGIL

Fuente Elaboración Propia

6.2. Esquema SPEM de la metodología CRISP-DM Ágil

Para representar de manera gráfica la metodología CRISP-DM Ágil, SPEM en su versión 2.0 propone una serie de componentes gráficos que permiten interactuar entre sí para ilustrar el comportamiento que debe llevar un proyecto en ejecución los cuales se especifican en la Tabla 5.

Haciendo uso de estos elementos mínimos, se construye la propuesta metodológica bajo elementos de SPEM llamada CRISP-DM ÁGIL, cuya representación está en la Imagen 38.

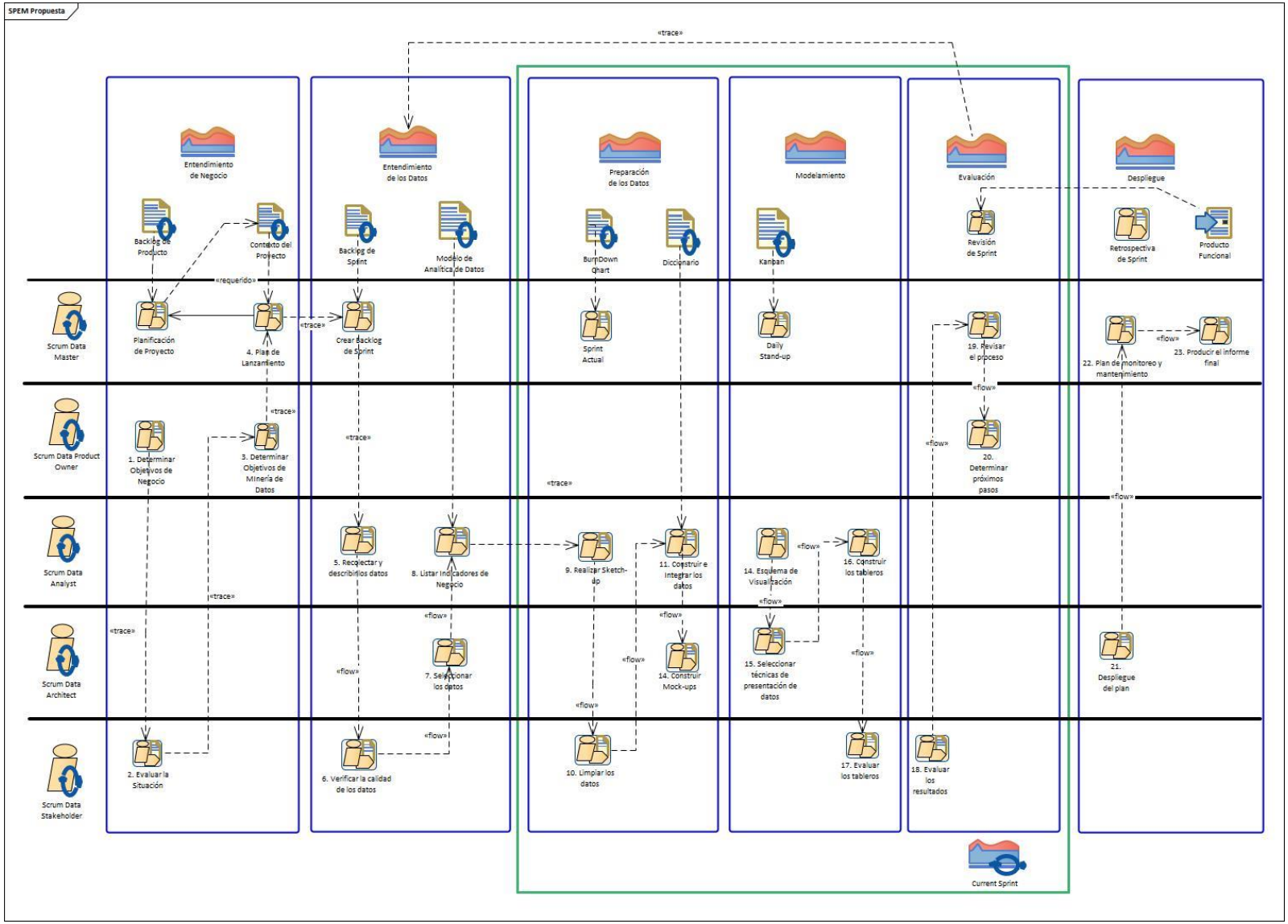


Imagen 38 Diagrama SPEM CRISP-DM ÁGIL

Fuente Elaboración Propia

Se han utilizado carriles en diferentes colores para permitir interpretar en qué fase va una actividad y quién es el responsable de que esa actividad se realice de manera satisfactoria.

Primero se crearon las fases ubicadas en la parte superior basados en CRISP-DM encerrados en carriles verticales de color azul, se agrega un carril adicional de color verde que encierra diferentes fases de analítica que hacen parte de la fase de *Current Sprint*, donde representa el grupo de actividades realizada por el equipo de trabajo para la construcción de los entregables por cada sprint.

Luego, se agregaron de manera vertical los roles separados por carriles negros de izquierda a derecha para ubicar en cada carril las actividades correspondientes al rol responsable y en la fase donde se realiza dicha actividad.

Por último, se asocian los productos de trabajo a las actividades encargadas de generar como salida un *backlog* de producto, un diccionario de datos o un modelo de analítica de datos.

PARTE IV
VALIDACIÓN

“Tu tiempo es limitado, de modo que no lo malgastes viviendo la vida de alguien distinto. No quedes atrapado en el dogma, que es vivir como otros piensan que deberías vivir. No dejes que los ruidos de las opiniones de los demás acallen tu propia voz interior. Y, lo que es más importante, ten el coraje para hacer lo que te dice tu corazón y tu intuición.” -- Steve Jobs

CAPÍTULO 7

Caso de Estudio

La validación de la hipótesis planteada en este proyecto de maestría, se ha realizado el caso de estudio bajo un diseño metodológico de estudio de casos en dos (2) proyectos institucionales: el primero es **preparatorio** que busca entender el comportamiento de un proyecto de analítica de datos usando CRISP-DM estándar, el cual se realizó en la Universidad de Medellín en un proyecto llamado "Esfuerzo Institucional" que hace parte de un macroproyecto llamado ODISEA. El segundo proyecto es **experimental** y usa la metodología CRISP-DM ÁGIL para implementarla en el desarrollo de "Admisión BI" en la Fundación Universitaria Autónoma de las Américas.

Teniendo en cuenta los diagramas de componentes UML donde se especifican las piezas de software, controladores embebidos y/o servicios externos necesarios para que un sistema pueda funcionar, se definieron 4 componentes principales:

- **Repositorio de Datos**: Describe las bases de datos usadas en los sistemas transaccionales.
- **Servidor de Analítica**: Almacena la aplicación usada para poder exponer los elementos de analítica creados en el proyecto como los *Dashboards* y reportes.
- **Herramienta de Visualización**: Usada para poder acceder a la aplicación de analítica alojada en el servidor de analítica, como Tableau Desktop o Power BI.

- **Herramienta de Preparación de Datos**: Describe el componente encargado de ajustar la información extraída para que responda a las necesidades del negocio de la aplicación de analítica.

La tecnología utilizada en cada proyecto es diferente, lo cual permite corroborar que la metodología CRISP-DM Ágil es independiente de la tecnología del proyecto. E la Tabla 6 se muestran las diferencias en las tecnologías utilizadas para cada caso de estudio en los componentes de los proyectos desarrollados.

Tecnología Utilizada en Caso de Estudio

Componente	Esfuerzo Institucional	Admisiones
Repositorio de Datos	ORACLE 11g	SQL Server 2012
Servidor de Analítica	Tableau Server	Analysis Services
Herramienta de Visualización	Tableau Desktop	Power BI
Herramienta de Preparación de Datos	Spyder / SQLDev / Excel	SSMS

Tabla 6 Tecnología Utilizada en Caso de Estudio

Fuente Elaboración Propia

La herramienta usada en los proyectos realizados en el caso de estudio para llevar seguimiento y control del proyecto fue *Team Foundation Services* (TFS) de Microsoft, la cual implementa plantillas donde se puede agregar el backlog del producto, las historias de usuario y las actividades, al igual que asignar responsabilidades a las personas involucradas y generar reportes del avance a nivel de proceso.

7.1. Proyecto de Esfuerzo Institucional

7.1.1. Planteamiento del Problema

Analizar los elementos críticos asociados a la formación académica de los estudiantes de la Universidad para medir el desempeño de la institución por medio de la exposición de indicadores, comparables y analizables en el tiempo. El objetivo final de la implementación es buscar mediante los tableros de visualización cómo se da la transformación y avance de conocimiento de un estudiante dentro de la Universidad de Medellín.

7.1.2. Desarrollo

Basado en las necesidades del proyecto y la metodología CRISP-DM, se asignaron los roles a los encargados de la ejecución y seguimiento del proyecto Esfuerzo Institucional. Durante las actividades de la fase de entendimiento de negocio, se define el alcance basado en los ítems de los artefactos que se generarían en el desarrollo del proyecto.

El seguimiento del proyecto se realizó cada 15 días debido a la carga administrativa para diferentes actividades de la institución por parte del equipo de trabajo; sin embargo, los compromisos directos o de último momento con la Institución obligaban a la cancelación de las reuniones en diferentes ocasiones.

Finalmente, el proyecto tuvo que ser realizado por un proveedor externo en consecuencia a los retrasos que se estaban presentando en la entrega al cliente, por lo que el registro de tiempos y demás se siguió llevando rigurosamente.

7.1.3. Resultados

Los colaboradores del proyecto no tenían un 100% de asignación de tiempo por la carga administrativa y otras funciones que desempeñaban durante el día en la

institución; por lo tanto, se generó un atraso en el desarrollo de la recopilación de las fuentes de datos para llevar a cabo la definición de los objetivos de analítica basados en los objetivos de negocio, las evidencias del proyecto implementado quedan consignadas en el anexo 1.

En la Tabla 7 se muestra cómo se empleó gran cantidad de tiempo en las fases de entendimiento de negocio y entendimiento de los datos, generado por los trámites de adquisición de las fuentes de datos exigidos por la institución. Las reuniones de seguimiento se programaron de forma espaciada, lo que generó inconvenientes de entendimiento de datos.

Las fases de evaluación y despliegue tuvieron conflictos para poder finalizar el proyecto por parte del equipo inicial, lo cual llevó a las directivas de la Universidad a finalizar el proceso con un proveedor externo, el cual conocía ampliamente los métodos y plataformas utilizados, razón por la cual en los esfuerzos no se registran los de evaluación y despliegue.

Resumen de tiempos por fases de CRISP-DM

Fase	Horas
Entendimiento de Negocio	186
Entendimiento de los Datos	186
Preparación	172
Modelamiento	52
Evaluación	0
Despliegue	0

Tabla 7 Resumen de tiempos por fases de CRISP-DM

Fuente Elaboración Propia

7.2. Proyecto BI-Admission

7.2.1. Planteamiento del Problema

Cubrir todos los informes necesarios para el departamento de admisión (módulo de inscripciones, gestión del solicitante, reconocimientos, módulo de entrevistas sin registro).

7.2.2. Desarrollo

Una vez se ha identificado la necesidad por parte de la institución se procede a conformar el equipo de trabajo que se encargó de la ejecución tanto a nivel de proceso como de producto.

Para describir la metodología a implementar con el equipo de trabajo se usó el ciclo de vida del proyecto bajo CRISP-DM ÁGIL que se muestra en la Imagen 39.

Ciclo de Vida CRISP-DM ÁGIL

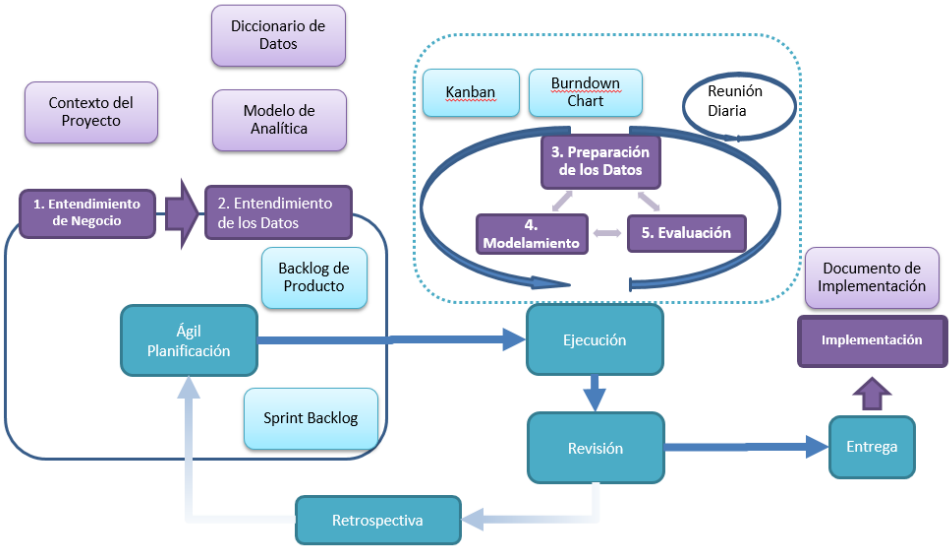


Imagen 39 Ciclo de Vida CRISP-DM ÁGIL

Fuente Elaboración Propia

Durante la primera reunión, se asignaron las personas que asumirían el rol requerido en la propuesta metodológica para el desarrollo del proyecto, y se describió la necesidad inicial, el alcance estimado y la duración de cada *Sprint* que para este proyecto fue de 1 semana.

El *Scrum Data Product Owner* se encargó de realizar las historias de usuario y las actividades asociadas para realizar los tableros de control, apoyado por el *Scrum Data Architect* quien conoce las herramientas disponibles en la organización y la arquitectura de analítica a implementar.

Antes de iniciar un *Sprint*, se reunía el equipo de trabajo para agregar, modificar o quitar actividades definidas para luego asignar responsables y estimar el tiempo de ejecución para cada actividad durante ese *Sprint*.

En la ejecución del primer *Sprint*, la actividad de Recolectar y Describir los Datos donde el responsable eran los *Scrum Data Analyst*, implicó un acompañamiento relevante del *Scrum Data Architect* para poder identificar las tablas que se iban a usar en las consultas que hacen parte del modelo tabular.

Durante los *Dailys* realizados se despejaban las inquietudes de entendimiento de negocio y se ajustaba el modelo relacional de acuerdo con las historias de usuario e información obtenida en cada reunión con el equipo de trabajo y el usuario final.

Al final del *sprint* 1, durante la actividad de Revisión del *Sprint*, se realiza la entrega formal del primer *Dashobard* funcional, el cual muestra las generalidades de la información del proceso de inscripciones y la validación de los criterios de aceptación consignados en las historias de usuario, logrando una aceptación importante entre los usuarios potenciales del *Dashboard*.

El *Sprint Planning* para el sprint 2 tuvo en cuenta la necesidad de excluir la actividad de Realizar Sketch-Up, dado que los analistas manifestaron que la herramienta de visualización permitía crear el *Dashboard* de manera sencilla y se podría ir mostrando al usuario final mientras se trabaja el modelo de datos. El modelo de datos fue creciendo a la medida que se iban cumpliendo las actividades de las historias de usuario, lo que permitía ir liberando de manera controlada los entregables propuestos en el sprint. Al final quedó un modelo tabular (Ver Imagen 47 de anexo 2) que abarcó todas las necesidades de información para el proceso de admisión de estudiantes.

El *Sprint* 3 tuvo problemas con la realización de los *Dailys* por motivos administrativos por parte de los analistas del proveedor, lo cual generó un retraso en las actividades de implementación del modelo para cubrir las necesidades de las historias de usuario consignadas en esta iteración. En la Imagen 51 que se encuentra en el anexo 2 se puede ver el retardo que se tuvo en el desarrollo de las actividades mediante el *Burndown Chart*. De manera general, en el desarrollo de cada sprint, se buscó respetar la ceremonia del *Daily Meeting*, la cual permitió identificar y abordar impedimentos de entendimiento de los datos, técnicas de visualización o despliegue del producto. Esta práctica permitió que las actividades de validación de datos se realizaran en tiempos más cortos de lo estimado. En la Tabla 8 se mencionan los hallazgos más relevantes en las ceremonias realizadas.

Hallazgos Relevantes en los Dailys

Detalle
Se despejan inquietudes de entendimiento de negocio (a)
Se aclaran dudas del modelo relacional (a)
Avanzan en modelo tabular (b)
Se define que el scketch-up no es necesario debido a que el mock-up se puede montar directamente en la herramienta final para tener una pre visualización (b)
Se entrega el modelo tabular en la herramienta de Analysis Services (c)
Inician con la construcción de los tableros de control (c)
Se entrega la primera versión del dashboard integrado con los datos (d)
El equipo de desarrollo solicita más datos para cargar visualización con más detalle (d)

El equipo solicita datos para publicación del dashboard para una primera revisión en ambiente de pruebas por parte del usuario. (d)
Se responden solicitudes de dudas de consultas (e)
Se incluye participante pasivo de operaciones (e)
Cargue de información (ver tablero por el usuario) (e)
Métricas a las alertas, medidas grandes (e)
Comparar datos de power bi con datos de usuario (f)
Visualizar con usuario el resultado final antes de publicación. (f)

Tabla 8. Hallazgos Relevantes en los Dailys

Fuente Elaboración Propia

La participación de los *Scrum Data Stakeholders* en algunas ceremonias del *Daily Meeting* permitió una mejor destreza en el dominio de negocio por parte de los *Scrum Data Analyst* y posibilitó la realización de algunos cambios en la visualización como mejora, buscando no alejarse de los criterios de aceptación de las historias de usuario. Por último, las retrospectivas de cada *Sprint* mejoraron las asignaciones de las actividades de acuerdo a las habilidades de los *Scrum Data Analyst* al interior del equipo.

7.2.3. Resultados

Durante los *Daily Meeting* donde se incluyó al usuario final antes de terminar el *Sprint Review* se identificó de manera proactiva que el modelo tabular presentaba inconsistencias en los indicadores mostrados y un estimado de la meta que se debería alcanzar para cada indicador. En la Imagen 40 se muestra el *Dashboard* principal con los indicadores más relevantes del proyecto, el resto de evidencias se encuentran en el Anexo 2.

Dashboard Implementado

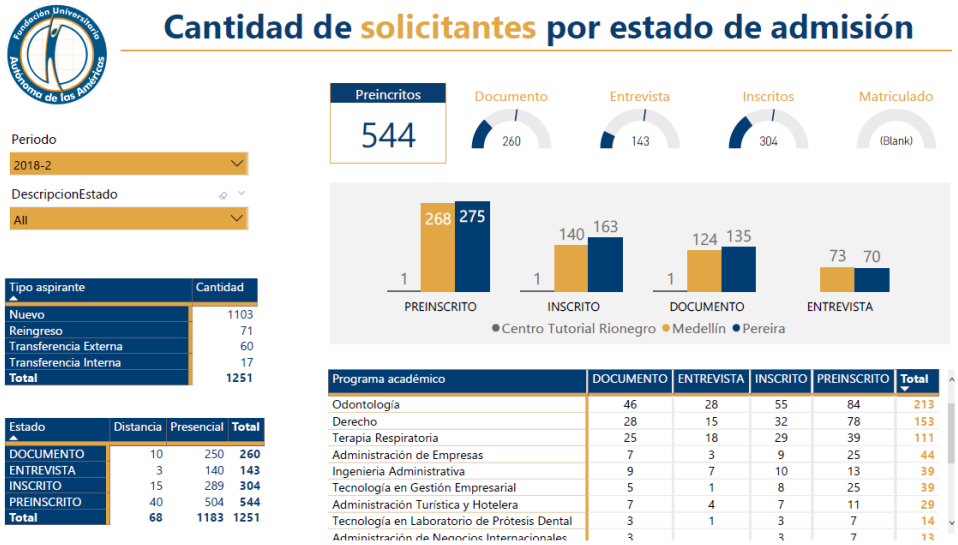


Imagen 40 Dashboard Implementado

Fuente Elaboración Propia

Se realizó un conteo de todos los tiempos de cada actividad para cada fase de los *Sprints* realizados para obtener el resumen de la Tabla 9.

Fase	Horas
Entendimiento de Negocio	38
Entendimiento de los Datos	103
Preparación	67
Modelamiento	68
Evaluación	8
Despliegue	14

Tabla 9. Resumen de Tiempos por Fase de CRISP-DM ÁGIL

Los tiempos de las primeras fases fueron relativamente cortos porque se contaba con la base de un documento con requerimientos definidos, el modelo relacional donde estaba la información y algunos scripts que cargaban ciertas reglas de negocio para el sistema transaccional; adicionalmente, se entregaron unos diseños de los reportes y algunas visualizaciones para representar los indicadores de negocio más relevantes.

Los *Dailys* realizados durante el desarrollo del proyecto disminuyeron el tiempo de evaluación porque se eliminaban las incidencias en 1 o 2 días.

CAPÍTULO 8

Análisis Comparativo

De las estrategias para realizar la comparación de metodologías de analítica de datos, se usó el marco comparativo de (Moine et al., 2015) el cual es mencionado en el apartado **3.1.1 Criterios de Comparación de las Metodologías**, para determinar la pertinencia de CRISP-DM ÁGIL para los proyectos medianos de analítica de datos en entornos ágiles, según la experiencia de las personas involucradas en el caso de estudio de la Fundación Universitaria Autónoma de las Américas.

El ejercicio anterior no se realizó en el caso de estudio de la Universidad de Medellín porque la metodología usada fue CRISP-DM para realizar comparación de tiempos pero no de evaluación metodológica, dado que se cuenta con una valoración de CRISP-DM en la metodología de comparación utilizada, como se puede ver en la Tabla 10.

Comparar CRISP-DM ÁGIL con Estándar

Criterio de Evaluación	CRISP-DM	CRISP-DM ÁGIL
Nivel de detalle en la descripción de las actividades de cada fase	**4**	**4**
Escenarios de aplicación	**3**	**3**
Actividades específicas que componen cada fase.	**9**	**10**
Actividades destinadas a la dirección del proyecto	**6**	**8**
Total	**5,5**	**6,25**

Tabla 10. Comparar CRISP-DM ÁGIL con Estándar.

Fuente Elaboración Propia

En la tabla anterior se muestra el ponderado de todas las preguntas requeridas en el marco de comparación de metodologías de analítica realizadas al equipo de trabajo al final de proyecto, utilizando la metodología CRISP-DM ÁGIL.

Se realizó el formulario en google docs y se emitió a seis (6) personas involucradas en el proyecto de BI – Admisiones, los resultados se promediaron para tener como resultado los datos de la columna 3 de la Tabla 10.

CRISP-DM ÁGIL, a través de las prácticas ágiles implementadas, apoya las actividades destinadas a la dirección del proyecto, lo que hace un factor diferenciador en proyectos de analítica de datos en entornos ágiles y garantiza una dinámica en los requisitos de acuerdo con las necesidades que surjan a lo largo de la construcción del proyecto.

Tiempos Proyectos de Caso de Estudio

Fase	Esf. Inst.	Admisiones
	Horas	Horas
Entendimiento de Negocio	186	38
Entendimiento de los Datos	186	103
Preparación	172	67
Modelamiento	52	68
Evaluación	0	8
Despliegue	0	14

Tabla 11. Tiempos Proyectos de Caso de Estudio.

Fuente Elaboración Propia

Para la fase de entendimiento de negocio, existe una gran diferencia entre los proyectos porque en el alcance de **BI- Admisiones** se contaba con un documento previo donde estaban los requisitos definidos.

En la fase de entendimiento de los datos con el proyecto de **Esfuerzo Institucional** fue más costoso definir y acceder a las fuentes de datos por el rigor administrativo que implicó, mientras que en **BI-Admisiones** solo era una fuente de datos, lo que facilitó su acceso para el desarrollo de esta fase.

Desde el primer *Sprint Review*, el usuario final pudo tener retroalimentación de las observaciones realizadas durante los *Dailys* donde tuvo participación, lo cual logró un impacto muy positivo y una aceptación de la metodología al eliminar impedimentos tempranos de datos y diseño de los tableros de control en las fases de Preparación, Modelamiento y Evaluación, esto permite validar que, el aplicar CRISP-DM integrando prácticas de metodologías ágiles en proyectos medianos de analítica de datos, se disminuye el esfuerzo en la entrega al cliente.

Al hacer seguimiento a los dos proyectos del caso de estudio se puede evidenciar la diferencia de tiempos de ejecución en cada fase que, aplicando prácticas ágiles, disminuye el esfuerzo en el desarrollo de las actividades de construcción de los entregables en los proyectos de analítica de datos.

PARTE V
CONCLUSIONES

“Mi padre me explicó que la educación y el conocimiento es lo que le permitirá a los niños mejorar el mundo.” -- Steve Wozniak

CAPÍTULO 9

Conclusiones, Recomendaciones y Trabajo futuro

9.1 Conclusiones

La metodología de analítica de datos CRISP-DM es robusta y brinda diversidad de actividades en todas las fases de un proyecto. A pesar de esto, no cuenta con unos roles específicos para asignar dichas actividades por el tipo de versatilidad que propone.

Las metodologías ágiles están siendo utilizadas en proyectos de analítica de datos aprovechando la experiencia y el éxito que se ha tenido en proyectos de software, como lo demuestran diferentes casos aplicados como (Zhu, 2017), (Alnoukari, 2016), (Chen et al., 2016), (Rehani, 2011), (Golfarelli et al., 2012), (Analuisa Barona, 2016), (Muntean, 2014), indicando la necesidad aplicar todas esas prácticas ágiles que permiten realizar entregas continuas, en periodos de tiempos cortos y de valor para la organización.

Hacer uso de SPEM permite al equipo de trabajo tener una perspectiva de cómo se van a ejecutar las actividades de un proyecto de analítica de datos, por qué fases va a pasar el proyecto y dónde se va a ver involucrado de acuerdo con el rol designado.

El uso de CRISP-DM ÁGIL permitió involucrar activamente a los integrantes del proyecto, utilizando prácticas ágiles tanto de SCRUM como de XP, mitigando los incidentes antes liberar el entregable en ambiente productivo.

Las herramientas utilizadas para la construcción del proyecto son independientes de la metodología CRISP-DM Ágil.

Incluir los *Daily Meeting* a CRISP-DM permite identificar de manera temprana los impedimentos que van surgiendo en la construcción del proyecto para corregir o evaluar una solución, lo cual genera menos reproceso y aumenta la confianza en los usuarios finales ya que están al tanto del avance del proyecto.

La dinámica de CRISP-DM ÁGIL permitió una comunicación efectiva en el equipo de trabajo.

Aplicando CRIS-DM ÁGIL en proyectos de analítica de datos permite disminuir el esfuerzo.

9.2 Cumplimiento de objetivos

A continuación, se describen los apartados/secciones del documento donde se evidencia el cumplimiento de los objetivos propuestos en este proyecto.

Objetivo Específico	Capítulo o Sección
Identificar fortalezas y debilidades de la metodología CRISP-DM para todas sus fases en proyectos de analítica de datos.	Antecedentes
Realizar una caracterización de los procesos de metodologías de analítica de datos en entornos ágiles	Metodologías basadas en otros modelos de referencia

Diseñar una metodología basada en CRISP-DM incluyendo prácticas de metodologías ágiles para proyectos de analítica de datos	Representación SPEM de la metodología CRISP-DM ÁGIL
Validar la propuesta metodológica con un caso de estudio en un proyecto de analítica de datos de tamaño mediano	Análisis Comparativo

9.3 Recomendaciones

CRISP-DM ÁGIL requiere contar con recursos disponibles durante la ejecución de cada sprint, respetar las ceremonias como el *Daily Meeting* y el *Sprint Retrospective*, ayuda a que todo el equipo de trabajo tenga sincronía a la hora de participar en sus actividades asignadas y ser un equipo auto gestionable en torno a los incidentes generados.

La metodología CRISP-DM ÁGIL está diseñada para proyectos de analítica de datos descriptiva de tamaño mediano que requiera el desarrollo de tableros de control y/o reportes.

Se define que la actividad de REALIZAR SKECTCH-UP no es necesario realizarla debido a que la actividad de REALIZAR MOCK-UP se diseña directamente en la herramienta final de visualización con gráficos finales.

En la fase de entendimiento de negocio, la participación del *Scrum Data Architect* fue esencial para apoyar la construcción del *Product Backlog*, debido a que el dominio en la tecnología usada permitió delimitar mejor las historias de usuario.

El uso de una herramienta de gestión de proceso como el TFS te permite generar artefactos de prácticas ágiles para visualizar el progreso del equipo como el *Burn Down Chart,* el *Kanban*, el flujo acumulativo, entre otros.

9.4 Trabajos Futuros

CRISP-DM ÁGIL ha sido diseñado para realizar proyectos de analítica descriptiva como *Dashboards* y/o *Reporting*, y puede ser modificado para realizar proyectos de Analítica predictiva en entornos ágiles donde las actividades de la fase de modelamiento correspondan a la necesidad puntual.

Evaluar un equipo de analítica de datos realizando varios proyectos con la misma metodología para poder evaluar el nivel de madurez trabajando analítica de datos en entornos ágiles.

Apoyar el entendimiento de la metodología CRISP-DM ÁGIL mediante artefactos didácticos que soporten un programa de inducción para involucrar a personas de la organización externas a la dinámica propuesta.

9.5 Contribuciones

Al finalizar el proceso de desarrollo de este proyecto, se logró realizar las siguientes actividades, que sustentan las contribuciones realizadas:

- Participación en evento académico SICC 2018 con el artículo APROXIMACIÓN METODOLÓGICA PARA EL DESARROLLO DE PROYECTOS DE ANALÍTICA USANDO METODOLOGÍAS ÁGILES, bajo la participación en modalidad póster, realizado en la Universidad de Medellín.

- Realización de talleres de formación en las áreas relacionadas con el proyecto, en el Departamento de Informática de la Universidad de Medellín y en la Fundación Universitaria de las Américas.
- Sometimiento de trabajo “MODELO BASADO EN CRISP-DM EXTENDIDO MEDIANTE PRÁCTICAS DE METODOLOGÍAS ÁGILES PARA PROYECTOS MEDIANOS DE ANALÍTICA DE DATOS” al II Simposio Regional de Maestrías y Doctorados, a realizarse en la Universidad de Medellín en Octubre del 2019.

10 Bibliografía

Agyapong, K. B., Hayfron-Acquah, J. B., & Asante, M. (2016). *An Overview of Data Mining Models (Descriptive and Predictive)* (Vol. 4). Retrieved from www.ijournals.in

Alnoukari, M. (2016). ASD-BI : A Knowledge Discovery Process Modeling Based on Adaptive Software Business Intelligence and Agile Methodologies for Knowledge-Based Organizations : Cross-Disciplinary Applications, (January), 28.

Alnoukari, M., Alzoabi, Z., & Hanna, S. (2008). Applying Adaptive Software Development (ASD) agile modeling on predictive data mining applications: ASD-DM methodology. *Proceedings - International Symposium on Information Technology 2008, ITSim*, *2*. https://doi.org/10.1109/ITSIM.2008.4631695

Analuisa Barona, J. F. (2016). Formulación de un marco metodológico para el desarrollo de soluciones de inteligencia de negocios, empleando metodologías agiles. Caso: área de datawarehouse del servicio de rentas internas. Retrieved from http://dspace.udla.edu.ec/handle/33000/5689

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS - DM*. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Biguru. (2015). What roles do you need in your data science team? | The Business Intelligence Blog. Retrieved May 21, 2018, from https://biguru.wordpress.com/2015/06/25/what-roles-do-you-need-in-your-data-science-team/

Boer, G. (2017). What is Scrum? - Azure DevOps | Microsoft Docs. Retrieved May 16, 2018, from https://docs.microsoft.com/es-es/azure/devops/agile/what-is-scrum

Bucher, T., Klesse, M., Kurpjuweit, S., & Winter, R. (2007). Situational Method Engineering. Retrieved from http://www.iwi.unisg.ch

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76. https://doi.org/10.1109/ICETET.2008.239

Chen, H. M., Kazman, R., & Haziyev, S. (2016). Agile big data analytics development: An architecture-centric approach. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (Vol. 2016-March, pp. 5378–5387). IEEE.

https://doi.org/10.1109/HICSS.2016.665

Daihani, D. U., & Feblian, D. (2016). IMPLEMENTATION OF CRISP-DM MODEL IN ORDER TO DEFINE THE SALES PIPE LINES OF PT X. Retrieved from https://isiem.net/wp-content/uploads/2016/10/9th_ISIEM_2016_paper_59_dss.pdf

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, *55*(1), 359–363. https://doi.org/10.1016/j.dss.2012.05.044

Espino, T. C., & Martinez, X. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. Retrieved from http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117memòria.pdf

Fayyad, U. (1996). KDD Fayyad_Piatetsky-1996-TheKDD, *39*(11), 27–34. https://doi.org/10.1145/240455.240464

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Garcés Uquillas, M. B. (2015). Estudio comparativo de metodologías e implementación de alternativas business intelligence opensource vs. propietarias en entornos tradicionales; caso prototipo en las pymes en el sector agroindustrial. Retrieved from http://200.24.220.94/handle/33000/2660

García Aguilar, J. (2016). CRONOS-ANALYZER: Herramienta para el análisis y extracción de conocimiento a partir de datos sobre el uso del tiempo personal. Retrieved from https://ruidera.uclm.es/xmlui/handle/10578/8195

Garzón, A. (2018). Introducción a Data Science - Microsoft Virtual Academy. Retrieved June 13, 2018, from https://mva.microsoft.com/es-es/training-courses/introduccin-a-data-science-18330?l=e2izowPcE_5411982333

Golfarelli, M., Rizzi, S., & Turricchia, E. (2012). Sprint Planning Optimization in Agile Data Warehouse Design BT - Data Warehousing and Knowledge Discovery. *Data Warehousing and Knowledge Discovery*, *7448*(Chapter 3), 30–41. https://doi.org/10.1007/978-3-642-32584-7_3

GuhaThakurta, D., Ericson, G., Martens, J., & Bradley, R. (2017). Business understanding stage of the Team Data Science Process lifecycle - Azure | Microsoft Docs. Retrieved May 13, 2018, from https://docs.microsoft.com/en-us/azure/machine-learning/team-

data-science-process/lifecycle-business-understanding

Henderson-Sellers, B., & Ralyté, J. (2010). Situational Method Engineering: State-of-the-Art Review. Retrieved from https://pdfs.semanticscholar.org/1a8f/2fd4c185a18017a2802697b0176e6d41c786.pdf

Hochsztain, E., & Tasistro, A. (2015). Teaching an Learning Business Intelligence: Business Evaluation Last but Not Least. In *2015 International Workshop on Data Mining with Industrial Applications (DMIA)* (pp. 66–71). IEEE. https://doi.org/10.1109/DMIA.2015.19

IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler*. (C. I. C. 1994, Ed.). Copyright IBM Corporation 1994. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Jair, H., Palacios, G., Andrés, R., Toledo, J., Albeiro, G., Pantoja, H., … Navarro, M. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. Retrieved from www.astesj.com

Kanbantool. (n.d.). Metodología Kanban | Kanban Tool. Retrieved January 20, 2019, from https://kanbantool.com/es/metodologia-kanban

KathrynEE, Danielson Steve, & Erickson Doug. (2018). Understand and configure your Kanban board - Azure Boards | Microsoft Docs. Retrieved December 20, 2018, from https://docs.microsoft.com/en-us/azure/devops/boards/boards/kanban-basics?view=vsts&tabs=new-nav

Krawatzeck, R., & Dinter, B. (2015). Agile Business Intelligence: Collection and Classification of Agile Business Intelligence Actions by Means of a Catalog and a Selection Guide. *Information Systems Management*, *32*(3), 177–191. https://doi.org/10.1080/10580530.2015.1044336

Krawatzeck, R., Dinter, B., & Thi, D. A. P. (2015). How to Make Business Intelligence Agile: The Agile BI Actions Catalog. In *2015 48th Hawaii International Conference on System Sciences* (pp. 4762–4771). IEEE. https://doi.org/10.1109/HICSS.2015.566

McLaughlin, M. (2018). Agile Methodologies for Software Development. Retrieved May 16, 2018, from https://www.versionone.com/agile-101/agile-methodologies/

Menéndez Domínguez, V. H., & Castellanos Bolaños, M. E. (2008). Software Process Engineering Metamodel (SPEM). *Revista Latinoamericana de Ingenieria de Software*, *3*(2), 92–100. https://doi.org/10.18294/relais.2015.92-100

Moine, J. M., Haedo, A. S., Nacional, U. T., & Rosario, F. R. (2015). Una herramienta para la evaluación y comparación de metodologías de minería de datos.

Muntean, M. (2014). Toward Agile BI By Using In-Memory Analytics. *Informatica Economică*, *18*(3). https://doi.org/10.12948/issn14531305/18.3.2014.03

Muntean, M., & Surcel, T. (2013). Agile BI – The Future of BI. *Informatica Economică*, *17*(3). https://doi.org/10.12948/issn14531305/17.3.2013.10

Nadali, A., Kakhky, E. N., & Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology*, *6*, 161–165. https://doi.org/10.1109/ICECTECH.2011.5942073

Nehan, Y.-R., & Deneckere, R. (2007). Component-based Situational Methods BT - Situational Method Engineering: Fundamentals and Experiences. In J. Ralyté, S. Brinkkemper, & B. Henderson-Sellers (Eds.) (pp. 161–175). Boston, MA: Springer US.

Ostadzadeh, S., & Shams, F. (2013). Towards a Software Architecture Maturity Model for Improving Ultra-Large-Scale Systems Interoperability. *The International Journal of Soft Computing and Software Engineering*, *3*(3), 69–74. https://doi.org/10.7321/jscse.v3.n3.13

Piatetsky, G., & KDnuggets. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved June 12, 2018, from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Rehani, B. (2011). Agile way of BI implementation. In *2011 Annual IEEE India Conference* (pp. 1–6). IEEE. https://doi.org/10.1109/INDCON.2011.6139618

Rogalewicz, M., & Sika, R. (2016). Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering. *Management and Production Engineering Review*, *7*(4), 97–108. https://doi.org/10.1515/mper-2016-0040

Saltz, J., Shamshurin, I., & Connors, C. (2017). A Framework for Describing Big Data Projects (pp. 183–195). Springer, Cham. https://doi.org/10.1007/978-3-319-52464-1_17

Saltz, J., Shamshurin, I., & Crowston, K. (2017). Comparing Data Science Project Management Methodologies via a Controlled Experiment. Retrieved from http://hl-128-171-57-22.library.manoa.hawaii.edu/handle/10125/41273

SAS. (2019). ¿Qué es la minería de datos? | SAS. Retrieved August 13, 2019, from

https://www.sas.com/es_co/insights/analytics/data-mining.html

Schwaber, K. (2017). The Scrum Guide ™, (November).

Schwaber, K., & Sutherland, J. (2017). Scrum Guide | Scrum Guides. Retrieved May 16, 2018, from http://www.scrumguides.org/scrum-guide.html

SCRUMstudy™. (2016). *SCRUM*.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research ISSN*, *12*(1), 2351–8014. Retrieved from http://www.ijisr.issr-journals.org/

Sharma, V., Stranieri, A., Vamplew, P., & Martin, L. (2017). An Agile Group Aware Process beyond CRISP-DM : A Hospital Data Mining Case Study, 109–113.

Shcherbakov, M., Shcherbakova, N., Brebels, A., Janovsky, T., & Kamaev, V. (2014). Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development (pp. 708–716). Springer, Cham. https://doi.org/10.1007/978-3-319-11854-3_61

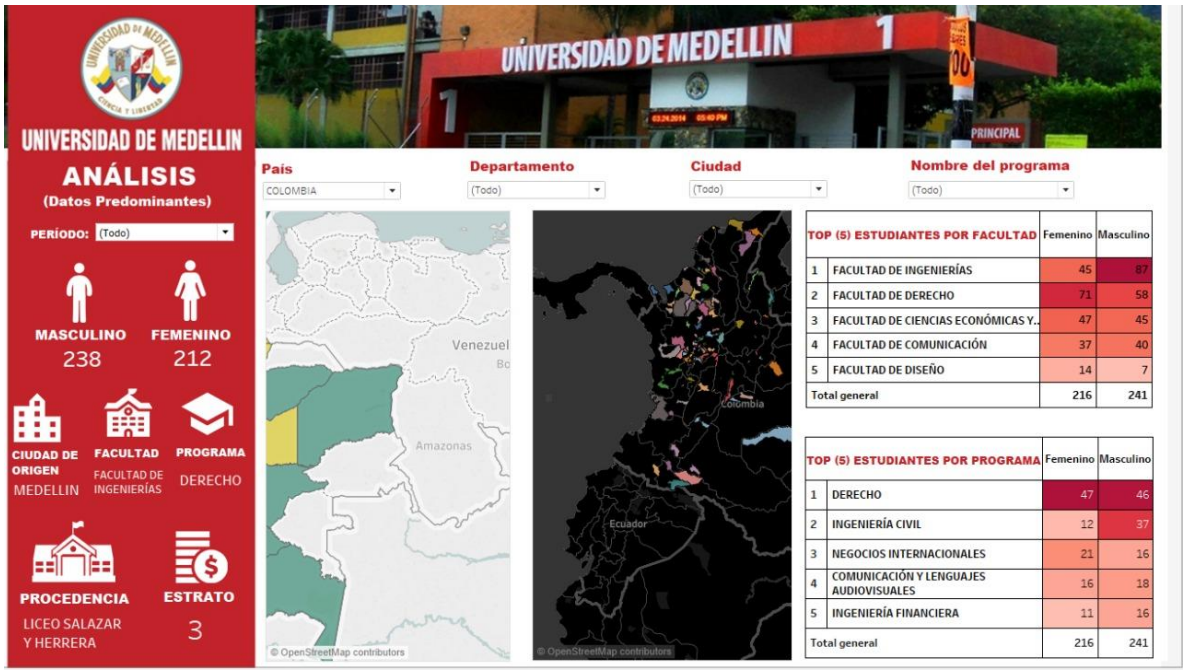
Uribe, E. H., & Ayala, L. E. V. (2007). Del manifiesto ágil sus valores y principios. *Scientia Et Technica*, *XIII*(34), 381–386. Retrieved from http://www.redalyc.org/resumen.oa?id=84934064%5Cnhttp://www.redalyc.org/articulo.oa?id=84934064

Zhu, X. (2017). Agile mining : a novel data mining process for industry practice based on Agile Methods and visualization. Retrieved from https://opus.lib.uts.edu.au/handle/10453/123178

ANEXOS

ANEXO 1

Desarrollo de Tableros de Control para Esfuerzo Institucional de la Universidad de Medellín.



TOP (5) ESTUDIANTES POR FACULTAD		Femenino	Masculino
1	FACULTAD DE INGENIERÍAS	45	87
2	FACULTAD DE DERECHO	71	58
3	FACULTAD DE CIENCIAS ECONÓMICAS Y...	47	45
4	FACULTAD DE COMUNICACIÓN	37	40
5	FACULTAD DE DISEÑO	14	7
Total general		216	241

TOP (5) ESTUDIANTES POR PROGRAMA		Femenino	Masculino
1	DERECHO	47	46
2	INGENIERÍA CIVIL	12	37
3	NEGOCIOS INTERNACIONALES	21	16
4	COMUNICACIÓN Y LENGUAJES AUDIOVISUALES	16	18
5	INGENIERÍA FINANCIERA	11	16
Total general		216	241

Imagen 41. Dashboard Estuerzo Institucional

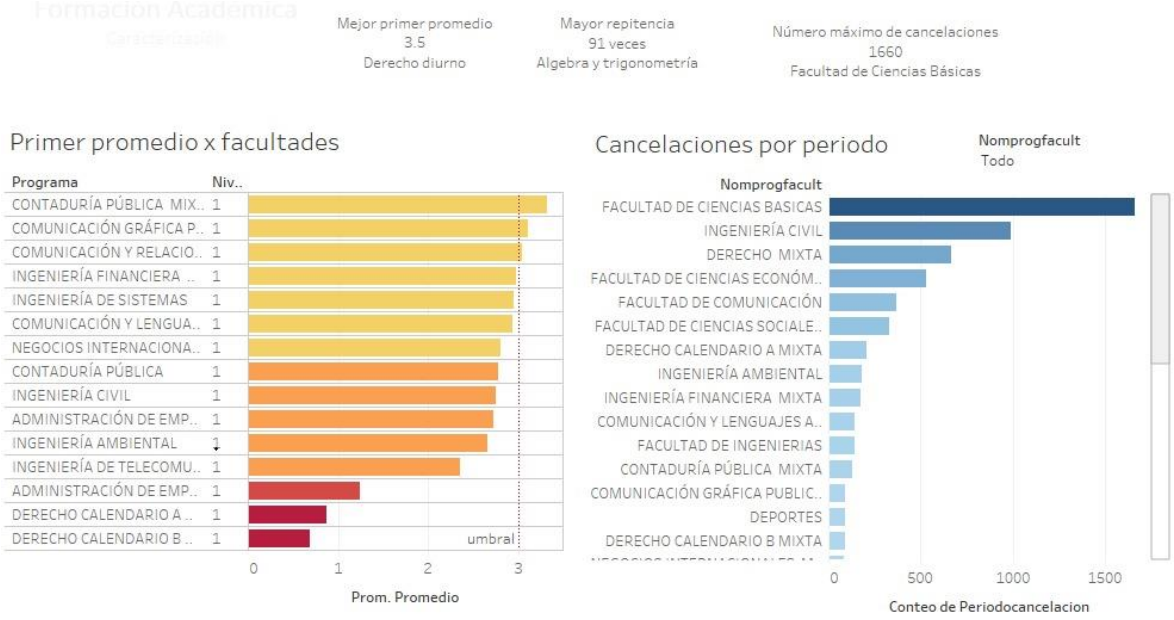


Imagen 42. Detalle de Facultades

Porcentaje de estudiantes por facultad

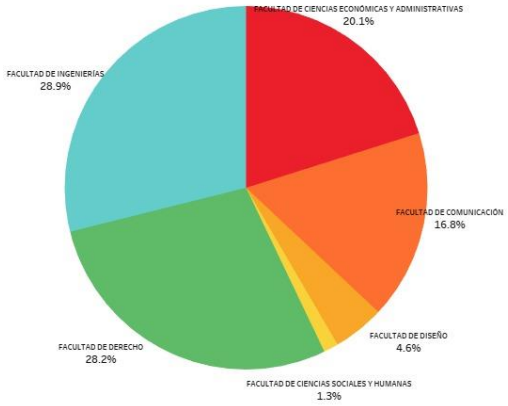


Imagen 43. DrillDown Estudiantes por Facultad

ANEXO 2

Desarrollo de un Tablero de Control para realizar seguimiento al proceso de Admisiones de la FUNDACIÓN UNIVERSITARIA AUTÓNOMA DE LAS AMÉRICAS

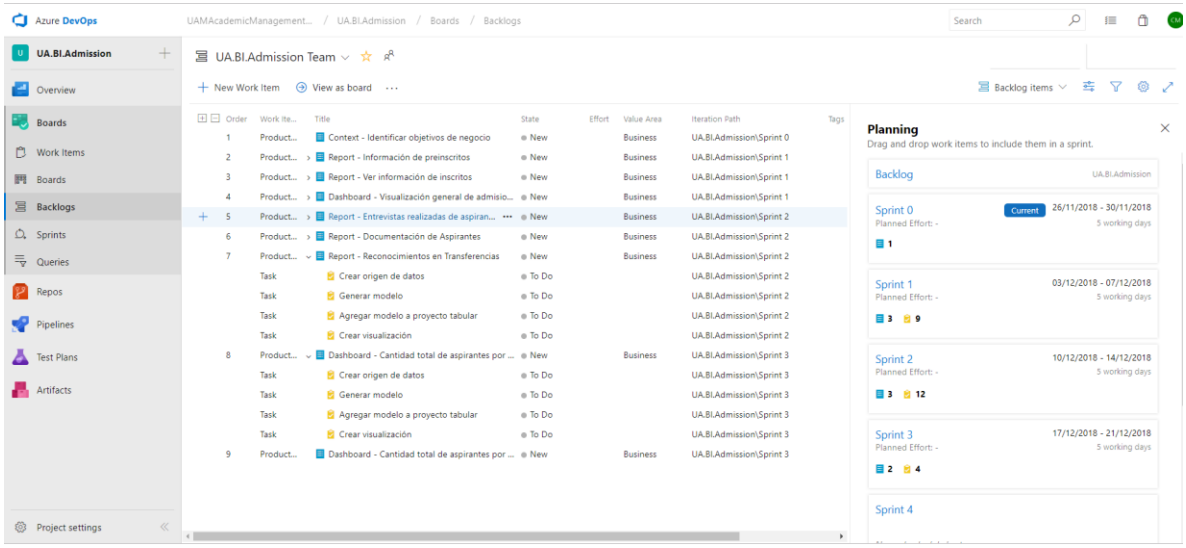


Imagen 44 Product Backlog

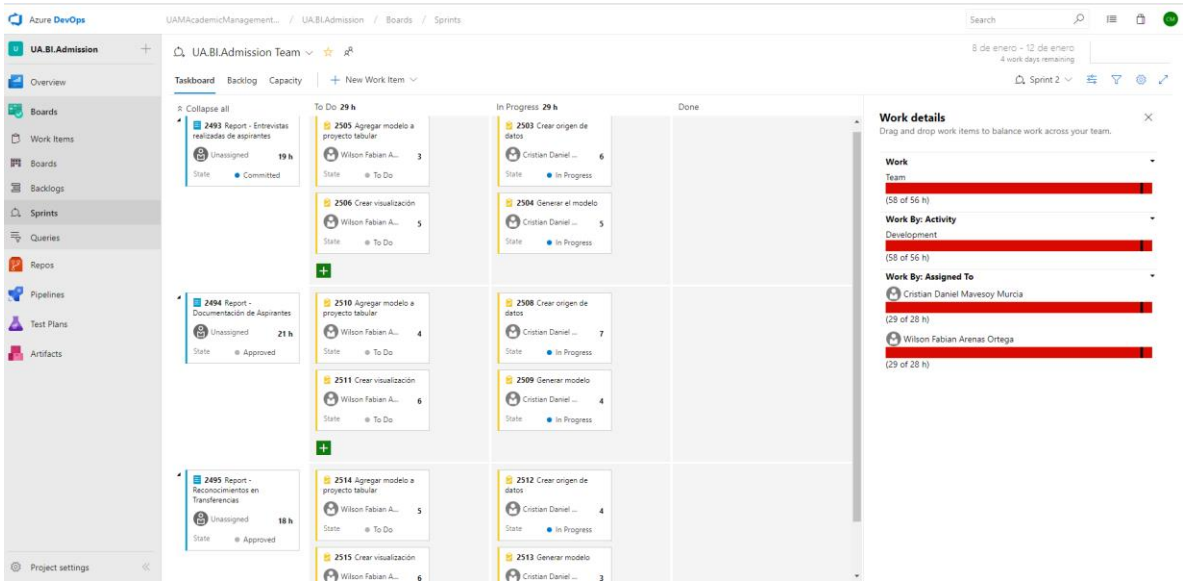


Imagen 45 Ejecución de Sprint.

Fuente Elaboración Propia

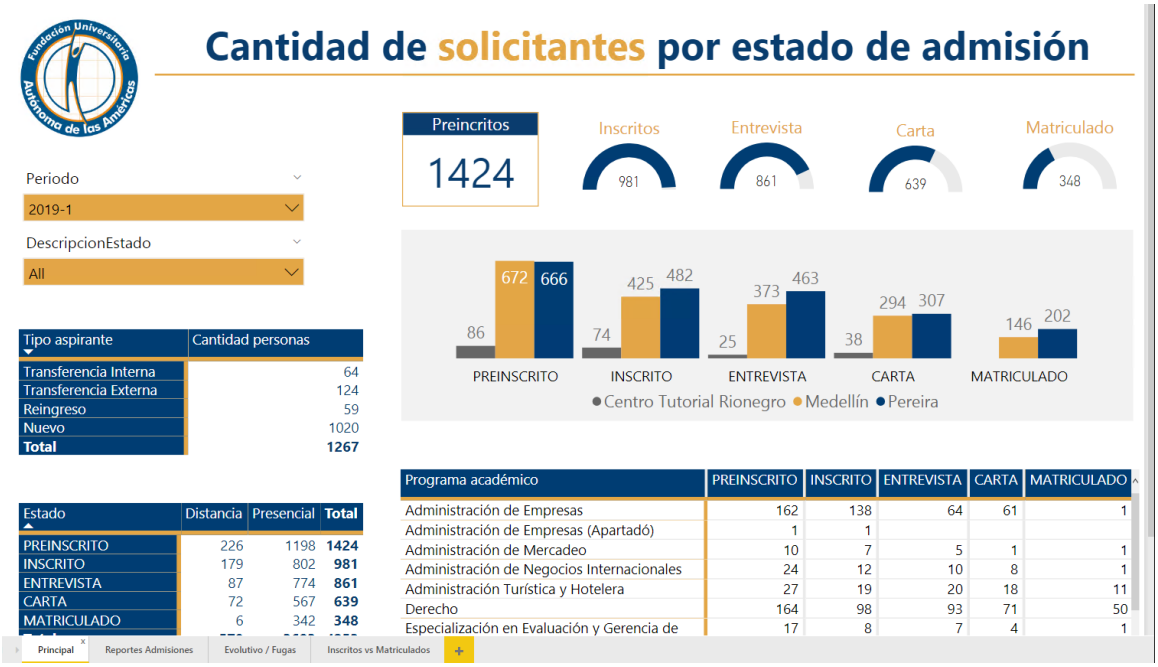


Imagen 46 Dashboard Principal.

Fuente Elaboración Propia

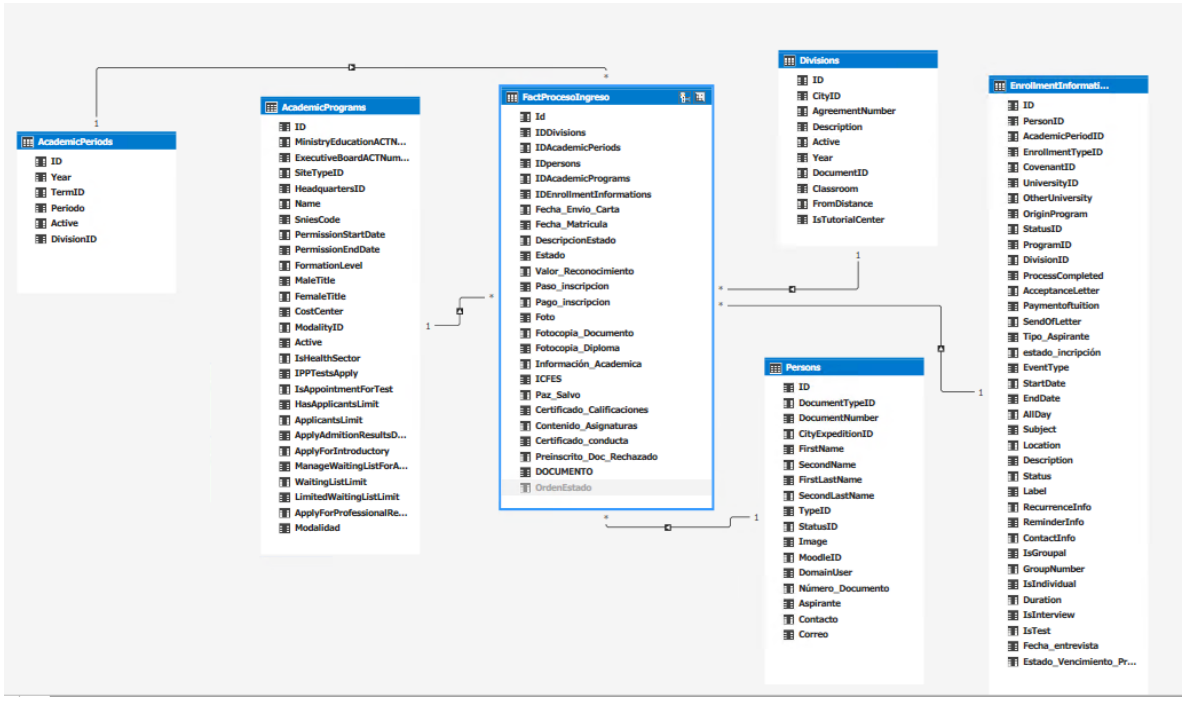


Imagen 47 Modelo Tabular Final.

Fuente Elaboración Propia

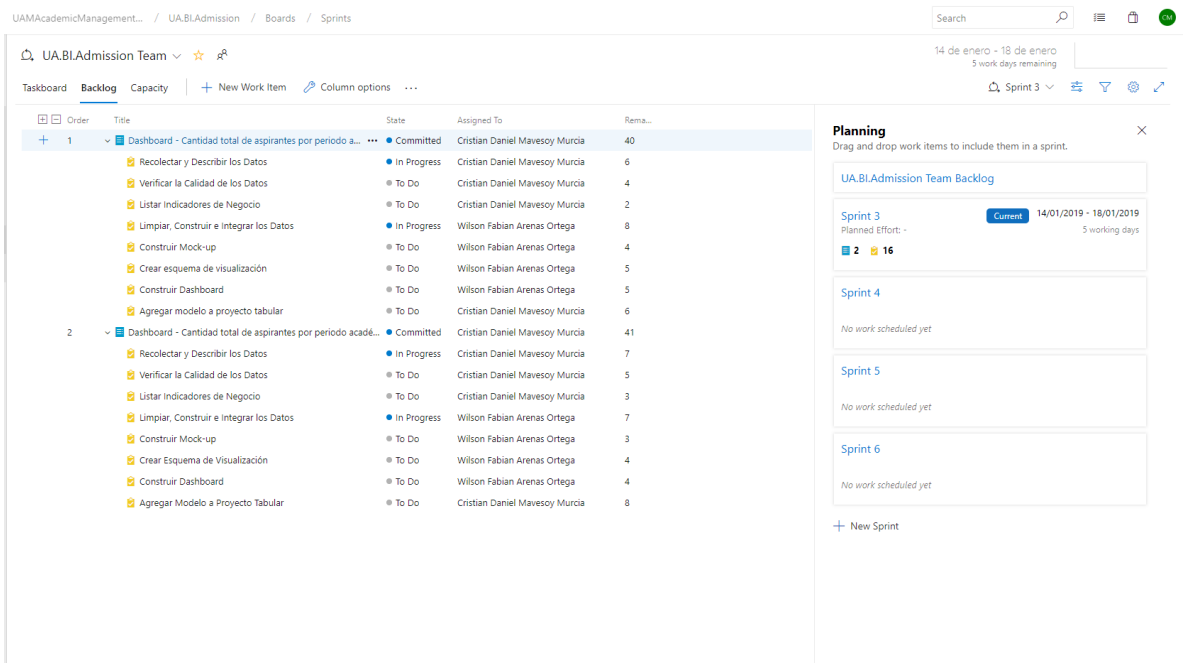


Imagen 48 Sprint Planning.

Fuente Elaboración Propia

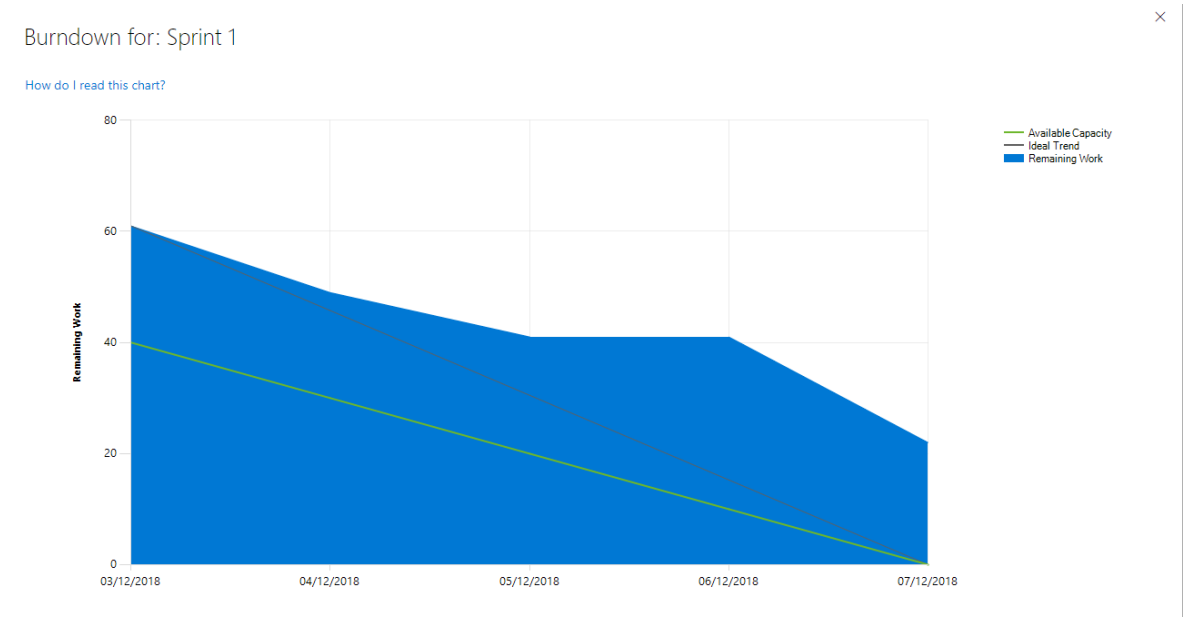


Imagen 49. Burndown Chart Sprint 1.

Fuente Elaboración Propia

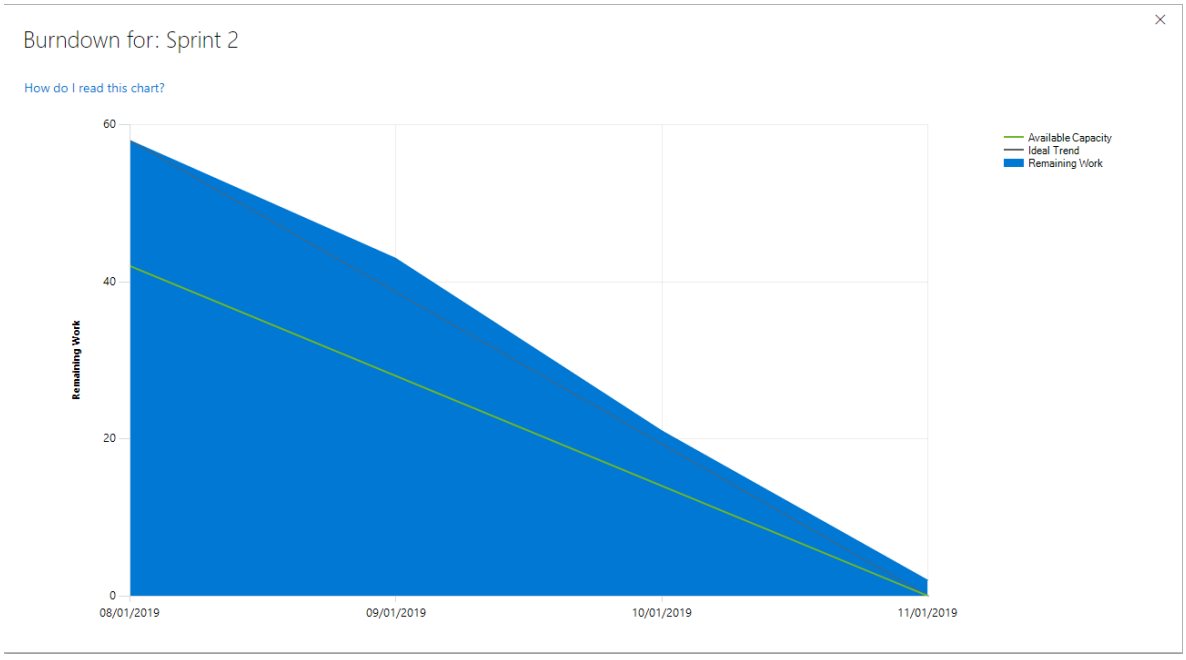


Imagen 50. Burndown Chart Sprint 2.

Fuente Elaboración Propia

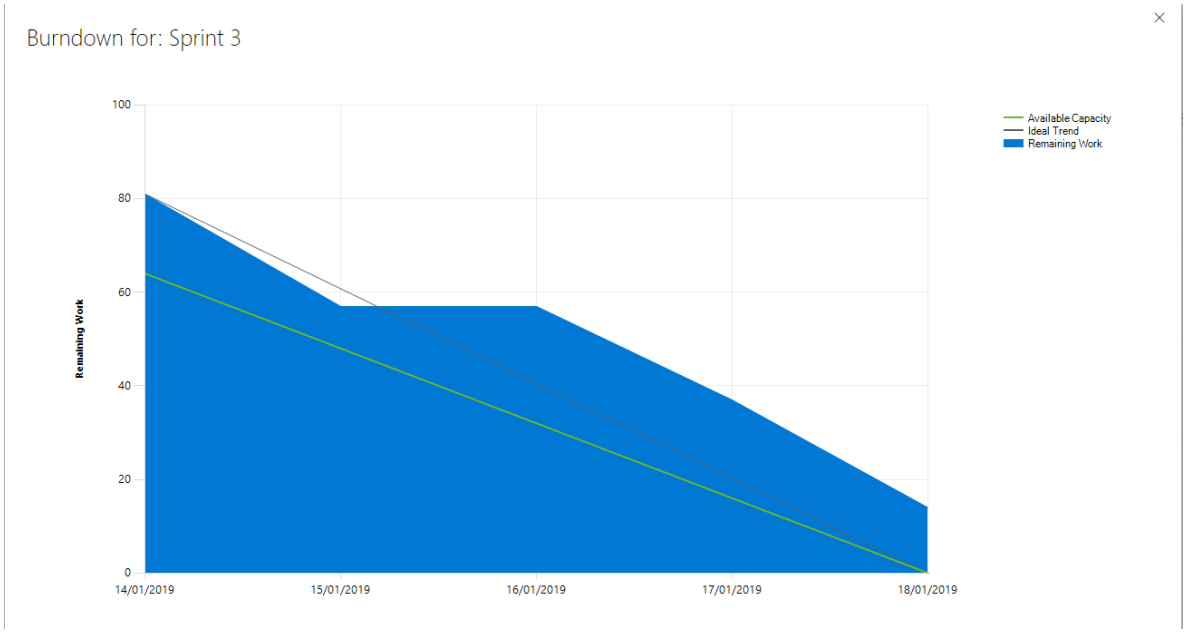


Imagen 51. Burndown Chart Sprint 3.

Fuente Elaboración Propia

Imagen 52. Dashboard. Generar Reportes.

Fuente Elaboración Propia

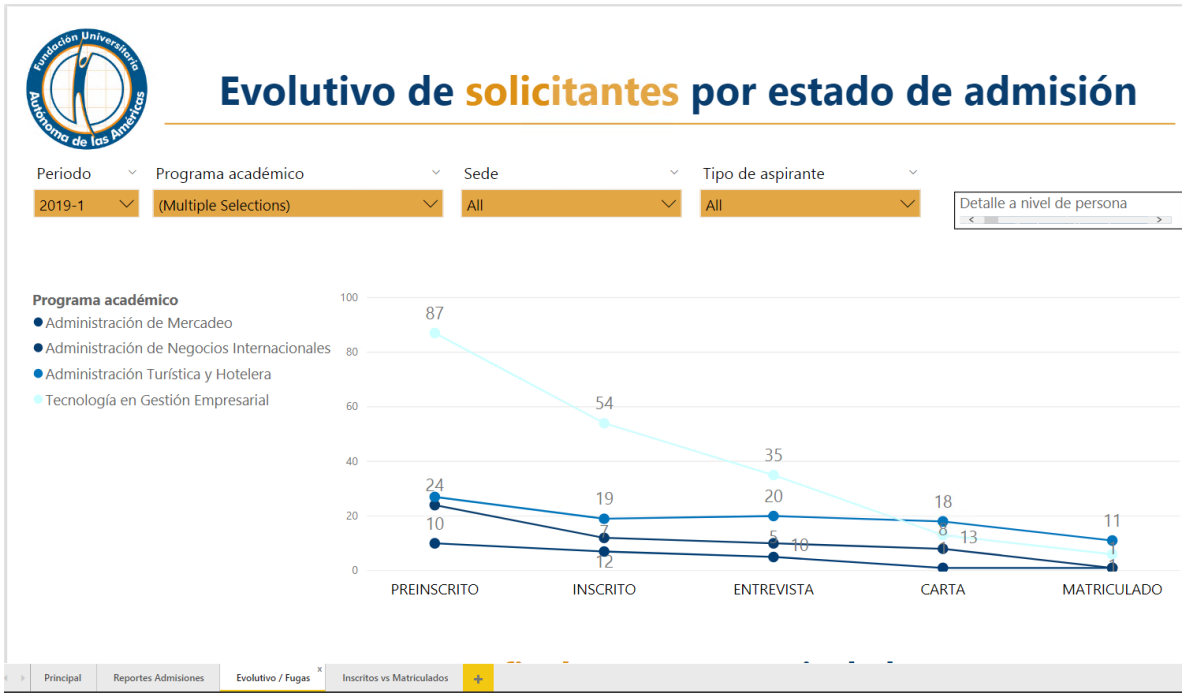


Imagen 53 Dashboard Estado de Admisión.

Fuente Elaboración Propia

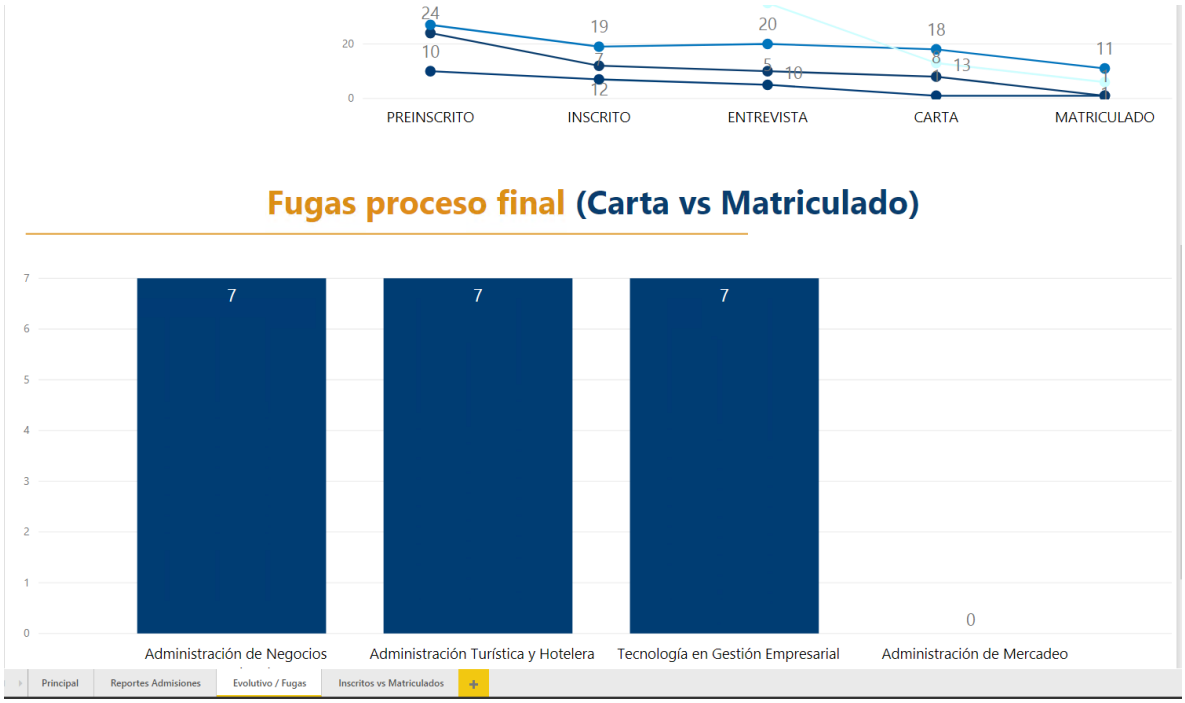


Imagen 54. Dashboard Indicador de Fugas.

Fuente Elaboración Propia