



CRISP-DM/SMEs

Metodología de Analítica de Datos para PYME

sin ánimo de lucro - ESAL

Autores

Jhon Montalvo-Garcia ¹ [0000-0002-5783-6373]

Juan Bernardo Quintero ² [0000-0002-0237-2276]

Bell Manrique-Losada ³ [0000-0001-9848-5162]

Afiliación de autores

¹ Seventh-day Adventist Church of Colombia, Medellín. Cra 84 #33aa-169, Colombia
jhonmontalvo@gmail.com

² Universidad de Medellín, Medellín. Cra 87 #30-65, Colombia
juquintero@udem.edu.co

³ Universidad de Medellín, Medellín. Cra 87 #30-65, Colombia
bmanrique@udem.edu.co

Universidad de Medellín, Colombia - 2019



Prefacio

CRISP-DM/SMEs es una metodología creada para realizar proyectos de analítica de datos en Pequeñas y Medianas Empresas (**PYME**) como Entidades Sin Ánimo de Lucro (**ESAL**), basada en las actividades más importantes y necesarias para este tipo de organizaciones. Sin embargo, esta metodología también puede ser aplicada a cualquier otro tipo de PYME con fines lucrativos, por lo tanto, solo se mencionará el concepto de PYME de ahora en adelante para un mayor alcance.

CRISP-DM/SMEs guía a las PYME en todo el proceso de desarrollo de sus proyectos de analítica de datos con el menor esfuerzo posible, reduciendo los costos de las distintas fases del proyecto.

CRISP-DM/SMEs fue desarrollada a partir de CRISP-DM y simplificada para el desarrollo de analítica de datos en PYME y enriquecida con tableros de control (*Dashboard*) y roles de la ciencia de datos.

Contenido

INTRODUCCIÓN A CRISP-DM/SMEs	5
PALABRAS CLAVES	5
CONCEPTOS DE LA METODOLOGÍA	6
<i>Roles</i>	6
<i>Fases</i>	7
<i>Actividades</i>	7
<i>Productos de trabajo</i>	7
<i>Guía</i>	7
<i>Herramientas</i>	7
REPRESENTACIÓN GRÁFICA	8
FASE 1: DEFINICIÓN DEL PROYECTO DE ANALÍTICA	9
HERRAMIENTA DE ENTORNO DE TRABAJO COLABORATIVO	9
ACTIVIDAD 1: SELECCIONAR OBJETIVOS EMPRESARIALES	9
ACTIVIDAD 2: DEFINIR OBJETIVOS DEL PROYECTO	10
CRITERIOS DE ACEPTACIÓN	11
ACTIVIDAD 3: ASIGNAR RECURSOS	12
ACTIVIDAD 4: DETERMINAR EL ALCANCE Y RIESGOS	14
INTERROGANTES DE ANALÍTICA	14
CONTEXTO DEL PROYECTO	15
FASE 2: GESTIÓN DE DATOS	16
HERRAMIENTA PARA GESTIÓN DE DATOS	16
ACTIVIDAD 5: RECOLECTAR DATOS	16
ACTIVIDAD 6: EXPLORAR DATOS	17
ACTIVIDAD 7: INTEGRAR DATOS	18
ACTIVIDAD 8: FORMATEAR DATOS	18
INTERROGANTES DE ANALÍTICA	19
DATOS FORMATEADOS	19
FASE 3: MODELADO	20
ACTIVIDAD 9: SELECCIONAR MODELO	20
ACTIVIDAD 10: SELECCIONAR HERRAMIENTA	21
ACTIVIDAD 11: CONSTRUIR TABLERO DE CONTROL	22
INTERROGANTES DE ANALÍTICA	22
TABLERO DE CONTROL	23
FASE 4: EVALUACIÓN	24
ACTIVIDAD 12: EVALUAR TABLERO DE CONTROL	24
ACTIVIDAD 13: ANALIZAR RESULTADOS	25
INTERROGANTES DE ANALÍTICA	25
ANÁLISIS DE RESULTADOS	25

FASE 5: DESPLIEGUE	26
ACTIVIDAD 14: AUTOMATIZAR PROCESO.....	26
ACTIVIDAD 15: DISTRIBUIR RESULTADOS.....	27
INTERROGANTES DE ANALÍTICA.....	27
INFORME FINAL.....	28
RECOMENDACIÓN FINAL.....	28
REFERENCIAS	29

INTRODUCCIÓN A CRISP-DM/SMEs

La Metodología CRISP-DM/SMEs fue desarrollada para proyectos de analítica de datos (DA) en PYME sin ánimo de lucro (ESAL) que involucran tableros de control (*dashboard*), y está basada en las principales actividades de la metodología CRISP-DM. Esta Metodología pretende ser una solución para superar las barreras que tienen las PYME en cuanto a la complejidad, los costos y la falta de capacitación de personal en la analítica de datos. Por lo tanto, el uso de CRISP-DM/SME permite la definición del proyecto, la gestión de los datos, el modelado, la evaluación y el despliegue rápido en cualquier PYME que incursione en la analítica de datos.

PALABRAS CLAVES

Esta guía metodológica hace uso de las siguientes palabras claves en todo el documento y para una mejor comprensión de estas, se brinda una breve descripción a continuación:

- PYME:** Pequeña y Mediana Empresa o **SMEs** por su sigla en inglés (*Small and Medium-sized Enterprises*). Las PYME se clasifican de acuerdo a número de trabajadores y sus activos totales según el país. En Colombia, la pequeña empresa comprende entre 11 y 50 trabajadores y activos totales mayores a 501 y menores a 5.001 salarios mínimos mensuales legales vigentes (SMMLV), mientras que la mediana empresa es aquella que tiene entre 51 y 200 trabajadores con unos activos totales entre 5.001 y 15.000 SMMLV.
- ESAL:** Entidades Sin Ánimo de Lucro (*Non-profit*). Son empresas jurídicas o sociales creadas para producir bienes y servicios, cuyo estatuto jurídico no les permite ser fuente de ingreso, beneficios u otras ganancias financieras para las unidades que las establecen, controlan o financian [1].
- DA:** Analítica de Datos (*Data Analytics*). Es la disciplina que se encarga de explorar, descubrir e inclusive interpretar patrones en los datos, con el propósito de sacar conclusiones [2].
- CRISP-DM:** (*Cross Industry Standard Process for Data Mining*). Es una metodología de minería de datos creada por el grupo de empresas SPSS, NCR y Daimler Chrysler en el año 2000 [3]. Actualmente es una de las metodologías más usadas por la industria en la analítica de datos [4].

CONCEPTOS DE LA METODOLOGÍA

Para representar la Metodología CRISP-DM/SMEs se usa un diagrama construido en SPEM (Software & Systems Process Engineering Meta-Model), ya que es un lenguaje estándar de modelado de procesos de desarrollo de software orientado a productos de trabajo. SPEM permite representar una familia de procesos de desarrollo de software y sus componentes, y proporciona una sintaxis y estructura para cada aspecto, incluyendo roles, fases, actividades, productos de trabajo, guías y herramientas, entre otros [5]. La propuesta metodológica usa los elementos representados en la figura 1.







Estereotipo	SPEM 2.0
Rol	
Fase	
Actividad	
Producto de trabajo	
Guía	
Definición de Herramienta	

Figura 1. Elementos usados de SPEM



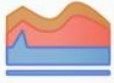
ROLES

La metodología CRISP-DM/SMEs define los roles desde la perspectiva de la industria de ciencia datos, como lo sugiere DataCamp [6]. Los roles seleccionados para esta metodología se describen a continuación:

Ingeniero de datos o Data Engineer: Es el principal rol dentro de los proyectos DA en las PYME. Es un profesional con conocimientos en matemáticas, estadística, programación, comunicación, y tiene dominio del sector. Este rol es quien se encarga de realizar la ejecución de la metodología, así que está presente en todas las fases del proyecto y puede ser interno o externo a la organización. Sus competencias mínimas son manejo de bases de datos, hojas de cálculo y diseño de tableros de control. Requiere, además, conocimientos asociados al uso de lenguajes java, R, SQL, C++, SPSS, HTML y Javascript.

Analista de Negocio o Business Analyst: Es el rol encargado de mejorar los procesos del negocio a través de la analítica, y es el intermediario entre el ingeniero de datos y el administrador de datos. En las PYME este rol lo suele asumir el director de área. Dentro de las competencias básicas están el manejo de herramientas de visualización como Power BI o Tableau, herramientas de office y modelado de datos. También se requiere un conocimiento básico en el lenguaje SQL.

Administrador de Datos o Data Manager: Es el rol encargado de liderar el equipo DA y suministrar los recursos financieros, humanos y de software para el proyecto. Sus competencias son el liderazgo, la comunicación interpersonal y análisis de información. Debido a estas competencias, este rol suele ser asumido por el gerente de la PYME, así que las competencias tecnológicas son mínimas para este caso.



FASES

Una fase es un conjunto de actividades desarrolladas por uno o varios roles y tiene como objetivo un producto de trabajo. La metodología CRISP-DM/SMEs comprende 5 fases partiendo de la *Definición del Proyecto* para luego realizar la *Gestión de los Datos*, *Modelado*, *Evaluación* y *Despliegue*. Estas fases pueden ser secuenciales o cíclicas, dependiendo de la evaluación de los interesados del proyecto (stakeholders).



ACTIVIDADES

Las actividades de son un conjunto de tareas del Proyecto de DA que se desarrollan para cumplir el objetivo de la fase a la cual pertenece. Una actividad o conjunto de actividades pueden originar un producto de trabajo que sirve de entrada para una siguiente fase y es desarrollada por uno o varios roles según las tareas asignadas.



PRODUCTOS DE TRABAJO

Un producto de trabajo es un elemento tangible del Proyecto DA que se produce y utiliza por en una fase. En la Metodología CRISP-DM/SMEs, cada fase del Proyecto DA tiene como resultado un producto de trabajo, que puede ser usado en la siguiente fase o en la interrelación de estas. Un producto de trabajo puede ser un documento, un conjunto de datos o un reporte de visualización. Los artefactos de la metodología son: i) contexto del proyecto, ii) los datos formateados, iii) el tablero de control, iv) el análisis de resultados y v) el Informe final.



GUÍA

Una guía es una serie de instrucciones o recomendaciones a seguir. En la fase de modelado se requiere de una guía para realizar eficazmente el tablero de control puesto que se debe realizar una gráfica correcta, ya que no todas representan bien los datos, y además está sujeta a la herramienta de visualización.



HERRAMIENTAS

Las herramientas son aplicaciones de software que ayudan a desarrollar las actividades de las fases. Una ayuda a la gestión del Proyecto DA, otra permite la gestión o formateo de los datos y finalmente la que ayuda construir el tablero de control. Las herramientas son escogidas por el Ingeniero de datos, según los recursos asignados para el Proyecto DA.

REPRESENTACIÓN GRÁFICA

La aplicación de la propuesta metodológica se representa en el diagrama SPEM, en la figura 2. La secuencia de las fases no es estricta, sino que se puede interactuar entre cada una de las fases; en la ejecución del Proyecto DA se puede avanzar o retroceder de ser necesario. Dichas fases contienen unas series de actividades mínimas necesarias para garantizar la calidad del proyecto, y al finalizar una o varias actividades se obtiene como resultado un producto de trabajo. Para ejecutar las actividades, los roles pueden usar una guía y varias herramientas.

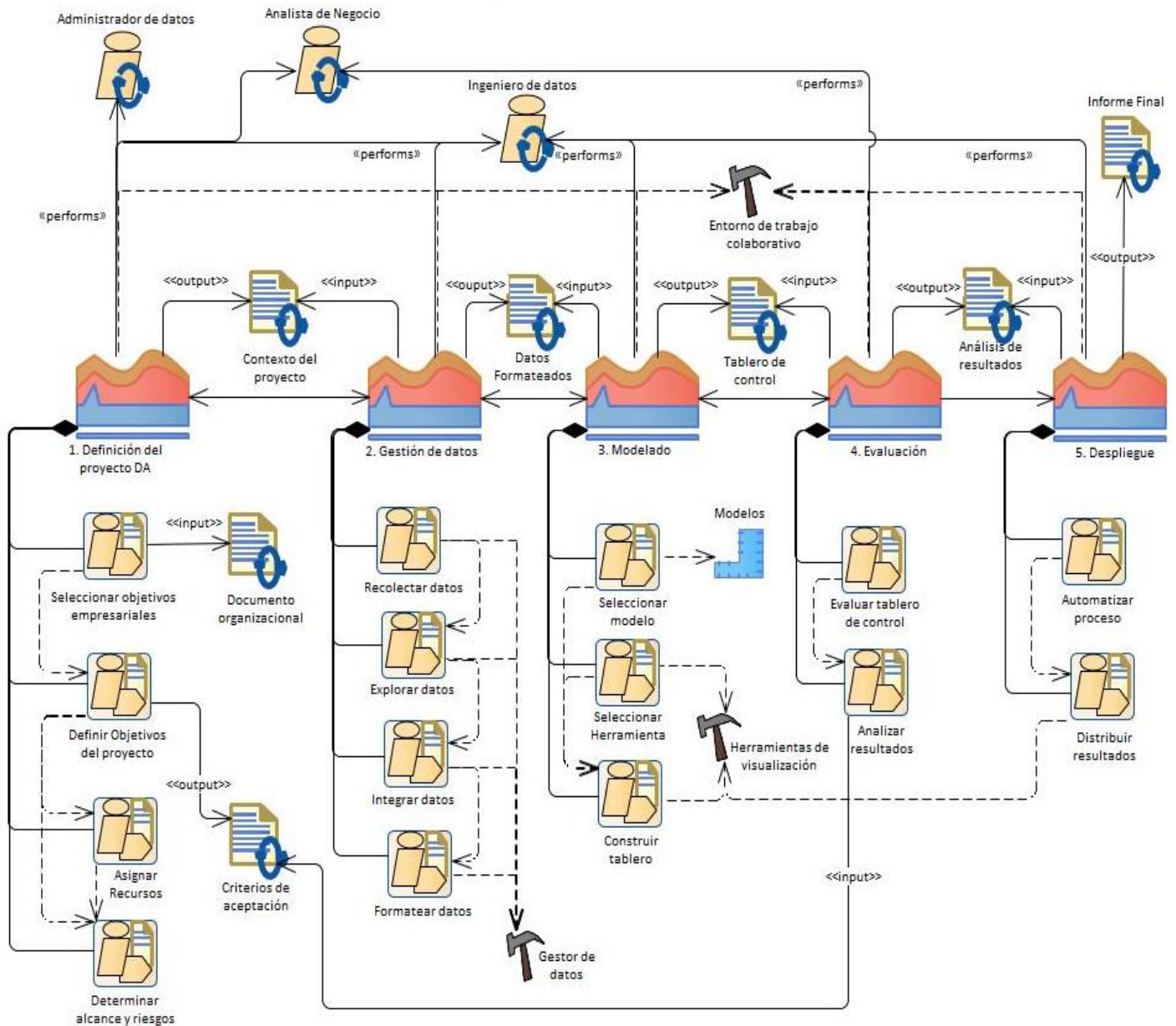
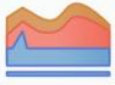


Figura 2. Representación gráfica de CRISP'DM-SMEs



FASE 1: DEFINICIÓN DEL PROYECTO DE ANALÍTICA

Para definir el Proyecto DA, es necesario identificar las personas que participarán activamente en el Proyecto DA, y el respectivo rol que asumirán. También se debe definir el problema o la necesidad a abordar en la PYME previamente. Estos dos componentes son esenciales para empezar el proyecto de Analítica de Datos ya que la Metodología CRISP-DM/SMEs excluye cómo descubrir una problemática o necesidad en una PYME. Para completar esta fase, se deben desarrollar las siguientes actividades: “Seleccionar objetivos empresariales”, “Definir objetivos del proyecto”, “Asignar recursos” y “Determinar el alcance y riesgos”. Con estas actividades y usando una “Herramienta de trabajo colaborativo”, se construye el “*Contexto del Proyecto*” como un producto de trabajo necesario para la siguiente fase.



HERRAMIENTA DE ENTORNO DE TRABAJO COLABORATIVO

La herramienta de entorno colaborativo será escogida en esta fase de la metodología. El equipo de trabajo debe estar de acuerdo en el uso de esta herramienta para dar seguimiento a todo el Proyecto DA. Cada uno de los integrantes debe tener acceso en línea para desarrollar las actividades que le corresponden e informar a todo el equipo del avance de estas. Dicha herramienta debe ser identificada en la asignación de recursos y puede ser una herramienta que use actualmente la PYME o puede elegirse entre una de acceso público o comercial, según el presupuesto disponible.



ACTIVIDAD 1: SELECCIONAR OBJETIVOS EMPRESARIALES

DESCRIPCIÓN DE LA ACTIVIDAD

Esta actividad consiste en seleccionar los objetivos empresariales de la PYME relacionados con el Proyecto de Analítica, ya que los proyectos en una organización deben estar orientados a alcanzar los objetivos organizacionales.

ENTRADAS

- Definición por escrito del problema o necesidad de la PYME.
- Documentar los requisitos de los usuarios que darán solución a la problemática planteada.
- Definición del equipo de trabajo de analítica y su respectivo rol.

TAREAS

- Solicitar los objetivos empresariales de la PYME.
- Discutir con el equipo cuáles son los objetivos empresariales que aborda el Proyecto de DA
- Describir cómo se obtendría un acercamiento a los objetivos con la implementación del Proyecto DA.

BUENAS PRÁCTICAS

- Hable con el equipo de analítica sobre la situación actual del problema o necesidad y trate de descubrir procesos relacionados al proyecto.
- Analice la Misión y Visión de la PYME para encontrar elementos organizaciones que puedan aportar al proyecto.
- Indague con el equipo o con personas relacionadas sobre los proyectos similares realizados previamente que puedan brindarle un panorama más amplio.
- Identificar personas indirectamente relacionadas que puedan facilitar información o algún tipo de soporte durante las actividades futuras, por ejemplo, el encargado de redes, servidores, seguridad o el administrador de bases de datos.

SALIDAS

- Documentación de los objetivos empresariales seleccionados para construir el producto de trabajo Contexto del Proyecto.



ACTIVIDAD 2: DEFINIR OBJETIVOS DEL PROYECTO

DESCRIPCIÓN DE LA ACTIVIDAD

Esta actividad consiste en definir los objetivos del Proyecto de DA de la PYME, alineados con los objetivos empresariales establecidos previamente. Estos objetivos se elaboran a partir de los requisitos de los stakeholders del proyecto. Como resultado de esta tarea se obtienen los criterios de aceptación del Proyecto DA.

ENTRADAS

- Documentación de los objetivos empresariales seleccionados para construir el producto de trabajo Contexto del Proyecto.

TAREAS

- Definir con el equipo de trabajo los objetivos del Proyecto DA según los requisitos de los usuarios.
- Definir con el equipo de trabajo los criterios de aceptación acorde los objetivos del Proyecto DA.

BUENAS PRÁCTICAS

- Clasifique los objetivos en generales y específicos teniendo en cuenta los objetivos empresariales.
- Tenga en cuenta que los objetivos se redactan empezando con un verbo en infinitivo y deben ser claros sin que se mal interpreten.
- Los objetivos deben ser alcanzables y medibles.
- Por cada objetivo del proyecto debe existir al menos un criterio de aceptación.

Los criterios de aceptación son un producto de trabajo como resultado de esta actividad, y será usado en la fase de evaluación, por lo tanto, se puede trabajar como un artefacto independiente.

SALIDAS

- Definición de los objetivos del Proyecto DA para construir el producto de trabajo *Contexto del Proyecto*.
- Criterios de aceptación como un producto de trabajo.



CRITERIOS DE ACEPTACIÓN

DESCRIPCIÓN DEL PRODUCTO DE TRABAJO

Los criterios de aceptación son las condiciones bajo las cuales se aceptan un requisito y se definen en la fase de la definición del proyecto. El objetivo principal de los criterios de aceptación es cumplir con los requisitos y expectativas de los stakeholders. Debe existir por lo menos un criterio de aceptación por cada objetivo del Proyecto DA definidos en la actividad 2. Estos criterios son usados en la fase de evaluación y determinarán si es necesario retroceder a una fase previa o avanzar a la fase de despliegue.

ENTRADAS

- Objetivos del Proyecto DA.

TAREAS

- Definir con el equipo de trabajo cada uno de los criterios de aceptación de acuerdo a los objetivos del Proyecto DA.

BUENAS PRÁCTICAS

- Los criterios de aceptación deben tener contexto, evento, y resultado. La tabla 1 presenta una guía con los elementos básicos para realizar el producto de trabajo criterios de aceptación.

SALIDAS

- Criterios de aceptación para futura evaluación.

Tabla 1: Criterios de Aceptación, basado en [7]

		Criterios de Aceptación				
	Objetivo	Contexto	Evento	Resultado	Evaluación	
1	[Objetivo del Proyecto relacionado con el criterio de aceptación]	1.1	[Proporciona una descripción sobre las condiciones que desencadenan el escenario.]	[Representa la acción que el usuario ejecuta, en el contexto definido]	[Representa el resultado de la acción que ejecuta el usuario dado un contexto]	[Resultado obtenido en la evaluación]
		1.2				
2		2.1				
		2.2				
		2.3				
3		3.1				
		3.2				



ACTIVIDAD 3: ASIGNAR RECURSOS

DESCRIPCIÓN DE LA ACTIVIDAD

Se estiman las horas de cada rol involucrado según las actividades para medir el esfuerzo del Proyecto DA, también se lista el software y los recursos físicos empleados en dicho proyecto. Es muy importante establecer las personas involucradas y el tiempo estimado que se dedicará al desarrollo del proyecto según las fases de la metodología CRISP-DM/SMEs

ENTRADAS

Participación de roles en las actividades. En la tabla 2 se presenta las actividades marcadas con una X en la columna del rol que participa en la ejecución de esta.

Tabla 2: Participación de roles en las actividades

PARTICIPACION DE ROLES EN LAS ACTIVIDADES			
METODOLOGIA CRISP-DM/SMEs	ROL		
	Administrador de datos	Analista de Negocio	Ingeniero de Datos
Fase 1: Definición del Proyecto DA			
1	Seleccionar objetivos empresariales	X	X
2	Definir objetivos del proyecto	X	X
3	Asignar recursos	X	X
4	Determinar alcance y riesgos	X	X
Fase 2: Comprensión de los datos			
5	Recolectar datos		X
6	Explorar datos		X
7	Integrar los datos		X
8	Formatear los datos		X
Fase 3: Modelado			
9	Seleccionar modelo		X
10	Seleccionar herramienta de visualización		X
11	Construir tablero de control		X
Fase 4: Evaluación			
12	Evaluar tablero de control	X	
13	Analizar resultados	X	X
Fase 5: Despliegue			
14	Automatizar proceso		X
15	Distribuir resultados		X

TAREAS

- Estimar el total de horas por cada actividad presentada en la tabla 2 y totalice por fase y por rol para completar la tabla 3.
- Establecer el software a utilizar en cada fase para la construcción de los productos de trabajo.
- Seleccionar las herramientas necesarias para llevar a cabo las actividades del Proyecto DA.
- Determinar los recursos físicos o tecnológicos en cada fase del proyecto.
- Asignar los costos financieros ponderados del Proyecto DA según los recursos identificados.
- Completar la tabla 3 con la información correspondiente.

Tabla 3: Asignación de recursos

ASIGNACIÓN DE RECURSOS						
TIPO DE RECURSO/FASES	FASE 1	FASE 2	FASE 3	FASE 4	FASE 5	Total horas
Recurso Humano	Cant. Horas	Cant. Horas	Cant. Horas	Cant. Horas	Cant. Horas	
Administrador de datos						
Analista de negocio						
Ingeniero de datos						
Recursos Software a utilizar						
1						
2						
Recursos Físicos/Tecnológicos						
1						
2						
TOTALES						

BUENAS PRÁCTICAS

- Tener en cuenta si es necesario la intervención de una persona externa al proyecto como un consultor o asesor o un empleado con conocimientos específicos sobre el área.
- Determine la cantidad de horas al día o a la semana que cada rol empleará en el proyecto. Las horas en reuniones también deben sumarse al proyecto como también el tiempo incurrido en la realización de consultas o informes.
- Establezcan fechas de obligaciones.
- Realice un inventario de software y hardware que tiene la PYME disponibles para la ejecución del Proyecto DA y establezca que software es necesario adquirir para la realización del Proyecto DA.
- Una vez identificado el recurso humano, de software y físicos o tecnológicos, establezcan el presupuesto financiero con el cual trabajará el proyecto.
- Opcionalmente se puede emplear un diagrama de Gantt para la asignación de los recursos y tener un mayor control sobre la evolución del proyecto.

SALIDAS

- Tabla con la asignación de recursos a utilizar en la ejecución del Proyecto DA.



ACTIVIDAD 4: DETERMINAR EL ALCANCE Y RIESGOS

DESCRIPCIÓN DE LA ACTIVIDAD

Esta actividad determina el alcance que debe realizarse para entregar el proyecto con las características y funciones especificadas. También se establecen los posibles riesgos que se puedan presentar, clasificándolos en bajo, medio o alto y cuáles serían las estrategias de contingencia para detectar los problemas.

ENTRADAS

- Objetivos del Proyecto DA.

TAREAS

Alcance: Describir con el equipo de trabajo el alcance del Proyecto DA teniendo en cuenta:

- Un cronograma con todas las actividades a desarrollar con su respectivo responsable.
- Estrategia de recolección de información
- Plan de socialización para la implementación del proyecto

Riesgos: Identificar con el equipo de trabajo los riesgos, sus respectivos niveles y estrategias de contingencias del Proyecto DA, completando la tabla 4.

Tabla 4: Riesgos y contingencias

Riesgo	Nivel			Contingencia
	Bajo	Medio	Alto	

BUENAS PRÁCTICAS

- Cuando se describe el alcance de un proyecto mediante palabras es importante que la descripción sea lo más concisa y directa posible.

SALIDAS

- Alcance del Proyecto DA
- Riesgos y contingencias del Proyecto DA.



INTERROGANTES DE ANALÍTICA

- ¿Cuáles son las preguntas que se van a responder a través del proyecto?
- ¿Para qué sirven los reportes e indicadores que se generan en el proyecto?
- ¿Qué condiciones o restricciones se deben tener en cuenta con respecto a la generación o el uso de los reportes e indicadores del proyecto?



CONTEXTO DEL PROYECTO

INTRODUCCION

Finalidad del documento y aspectos más resaltantes sobre el Proyecto DA.

✓ **Descripción del Problema**

Describe detalladamente el problema o necesidad de la PYME.

✓ **Responsables del Proyecto**

Nombre de los responsables del Proyecto DA y su respectivo rol dentro de la metodología de trabajo.

OBJETIVOS EMPRESARIALES

Liste los objetivos empresariales de la PYME que fueron tratados en la actividad 1 “Seleccionar los objetivos empresariales”.

OBJETIVOS DEL PROYECTO DA

Liste los objetivos para el Proyecto DA que fueron definidos en la actividad 2 “Definir objetivos del proyecto”.

CRITERIOS DE ACEPTACIÓN

Escriba los criterios de aceptación que fueron definidos en la actividad 2.

ASIGNACIÓN DE RECURSOS

Escriba los recursos asignados en la actividad 3.

DETERMINACION DEL ALCANCE Y LOS RIESGOS

Escriba el alcance y los riesgos identificados en la actividad 4.

COSTOS

Ponderado de los costos del proyecto teniendo en cuenta la asignación de recursos.

PLANEACION

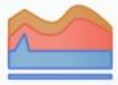
Construya el cronograma con todas las actividades a realizar de forma organizada y cronológica.

ESTRATEGIAS DE RECOLECCIÓN DE INFORMACIÓN

Literales con métodos, herramientas y personal utilizados para reunir la información necesaria.

PLAN DE SOCIALIZACIÓN

Estrategia para la recepción, adaptación y adopción del sistema por parte de los usuarios.



FASE 2: GESTIÓN DE DATOS

La gestión de datos es la fase que más tiempo requiere ya que se debe identificar las fuentes u orígenes y los respectivos accesos para la exploración de los datos. Además, el proceso de formatear los datos y dejarlos preparados para cualquier tipo de reporte analítico es un proceso complejo ya que debe garantizar la calidad y la integración de estos. La fase de gestión de datos comprende cuatro actividades, a saber; recolectar los datos, explorar los datos, integrar los datos y formatear los datos. Al completar estas actividades se obtendrá el producto de trabajo “Datos Formateados”.



HERRAMIENTA PARA GESTIÓN DE DATOS

Cada una de las actividades de esta fase puede usar una herramienta para la gestión de los datos y está sujeta a la experticia que tenga el Ingeniero de datos. Dicha herramienta debe estar identificada en la asignación de recursos. El objetivo es crear una base de datos estructurada con las distintas fuentes u orígenes de datos cuyo resultado final sea un producto de trabajo con los “Datos Formateados”



ACTIVIDAD 5: RECOLECTAR DATOS

DESCRIPCIÓN DE LA ACTIVIDAD

Según el contexto del Proyecto DA y el área de implementación, se recogen los datos de las distintas fuentes, sea a través de respaldos, bases de datos estructuradas o no estructuradas, registros web, archivos de Excel o cualquier otra fuente de datos.

ENTRADAS

- Contexto del Proyecto DA.

TAREAS

- Pedir el acceso a las distintas fuentes de datos a las personas que las custodian.
- Identificar si se requiere de información externa de la compañía y solicitar o buscar las fuentes.
- Firmar el documento de confidencialidad de la información según la política de tratamiento de datos de la PYME.
- Extraer un respaldo de los datos para ser explorados.

BUENAS PRÁCTICAS

- Realice un respaldo de la información suministrada en caso de daños o pérdida de datos.

- Almacene la información teniendo en cuenta la seguridad de los datos. Recuerde que solo las personas que han firmado el documento de confidencialidad pueden acceder a la información.

SALIDAS

- Base de datos para el Proyecto DA.



ACTIVIDAD 6: EXPLORAR DATOS

DESCRIPCIÓN DE LA ACTIVIDAD

Se explora, analiza y limpia los datos de las distintas fuentes de datos. Esta tarea es de gran utilidad para encontrar errores ya sea de captura, almacenamiento o calidad de los datos. Luego se identifica la relación entre las distintas fuentes y se documentan las propiedades de los datos.

ENTRADAS

- Base de datos para el Proyecto DA.

TAREAS

- Seleccionar los datos a utilizar en el Proyecto DA.
- Determinar el tamaño del origen de datos.
- Explorar posibles errores de captura en los datos recolectados.
- Analizar si existen atributos perdidos o campos nulos en las fuentes.
- Excluir los errores o en caso de que haya datos que generen ruido.
- Identificar y documentar la relación entre las distintas fuentes de datos.

BUENAS PRÁCTICAS

- Realice cada una de las tareas por cada fuente u origen de datos que se necesitan para el Proyecto DA.
- Si el tamaño de la fuente de datos es muy grande, entonces empiece con una selección de datos de muestra para realizar una prueba de exploración. Luego de identificar patrones de errores, o atributos perdidos y datos ruidosos, entonces tome la totalidad de los datos y realice la respectiva limpieza.

HERRMIENTA

- La herramienta por usar está bajo criterio del Ingeniero de datos. Puede crear consultas de selección, eliminación y actualización para realizar las tareas de limpieza.

SALIDAS

- Base de datos limpia



ACTIVIDAD 7: INTEGRAR DATOS

DESCRIPCIÓN DE LA ACTIVIDAD

Se integran las distintas fuentes de datos que tiene la empresa y que han sido suministrados para el Proyecto DA. De ser necesario, se adicionan datos externos de acceso público como información demográfica, de geolocalización, o macroeconómica.

ENTRADAS

- Fuentes de datos limpios
- Fuentes de datos externos

TAREAS

- Fusionar los datos de distintas fuentes ya sean internas o externas a la PYME.
- Agregar registros o columnas a la fuente de datos en caso de necesitar información para indicadores claves de desempeño (KPI).
- Asignar nombres a las columnas en las fuentes de datos acorde al Proyecto DA.
- Realizar pruebas de integración y calidad de datos.

BUENAS PRÁCTICAS

- De ser necesario, vuelva a la actividad “Exploración de datos” para analizar si se generaron nuevos errores debido a la integración con otras fuentes.
- Tenga en cuenta los cuidados necesarios al fusionar las fuentes ya que esto genera nuevas columnas o características en los datos.
- Si se va a adicionar información de una fuente a otra, tenga en cuenta mantener el mismo tipo de datos para cada columna.

SALIDAS

- Fuente de datos integrados.



ACTIVIDAD 8: FORMATEAR DATOS

DESCRIPCIÓN DE LA ACTIVIDAD

Esta es la última actividad antes de entrar a la fase del modelado, y muy importante saber que tipos de datos requieren de un formato especial o renombrar los campos para que el usuario final pueda entender fácilmente los datos. Se crea un nuevo conjunto de datos con los campos requeridos sólo para el análisis.

ENTRADAS

- Fuente de datos integrados.

TAREAS

- Seleccionar solo los datos que se van a usar en el Proyecto DA
- Crear un modelo de datos o un almacén de datos (Datamart) para el Proyecto DA.

- Crear dimensiones (metadatos) para los hechos (datos de estudio) que brinden mayor claridad para el análisis de la información.
- Formatear, relacionar, ordenar, clasificar o agrupar los datos según la necesidad del modelado.
- Configurar los datos para que sean accesibles a la herramienta del modelado.

BUENAS PRÁCTICAS

- La creación de un almacén de datos o Datamart multidimensional en estrella o copo de nieve facilitará la gestión de los datos durante todo el proceso de esta fase y la siguiente.

SALIDAS

- Datos formateados como un producto de trabajo.



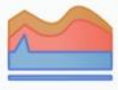
INTERROGANTES DE ANALÍTICA

- ¿Están accesibles los datos para la herramienta del modelado?
- ¿Existen datos externos que varían el formato cada vez que son actualizados?
- ¿Se realizaron las pruebas de calidad a los datos después de integrar las distintas fuentes?
- ¿Qué medidas de seguridad se ha tomado para proteger los datos donde están almacenados?



DATOS FORMATEADOS

Los datos formateados son el subconjunto de datos, o almacén de datos o Datamart creados específicamente para el Proyecto DA y deben estar listos para la fase del modelado. Esto quiere decir que todos los datos deben estar accesibles para la herramienta de construcción de los tableros de control o dashboard, y tener los elementos o atributos necesarios para la minería de datos.



FASE 3: MODELADO

En la base de Modelado se representa gráficamente los datos a través de elementos visuales como diagramas, gráficos de barras, mapas y tablas para comprender tendencias, patrones, indicadores, entre otros. El modelado desarrolla 3 actividades que son “Seleccionar el modelo”, “Seleccionar la herramienta de visualización” y la “Construcción del tablero de control”. Para llevar a cabo cada una de las actividades de esta fase se requiere de definir el modelo de la representación gráfica y la herramienta de visualización que puede crear dicho modelo.



GUÍA DE MODELADO

La guía para el modelado está directamente relacionada con la herramienta de visualización y dependerá de los tipos de modelos que tiene incorporado para su respectivo uso. Sin embargo, la gran mayoría de proveedores de herramientas de visualización tienen en común los modelos más usados en la analítica de datos.



HERRAMIENTA DE VISUALIZACIÓN

La herramienta de visualización es uno de los recursos de software más importantes en el desarrollo del Proyecto DA ya que esta convierte los datos en información de manera visual para la organización. Además, debe presentar el resultado final de manera efectiva sin dar lugar a información engañosa y su distribución debe ser práctico y entendible para el usuario final. La herramienta debe ser seleccionada por el Ingeniero de Datos teniendo en cuenta las características del proyecto, los recursos disponibles, los modelos definidos por el equipo de trabajo y los usuarios finales.



ACTIVIDAD 9: SELECCIONAR MODELO

DESCRIPCIÓN DE LA ACTIVIDAD

La selección del modelo esta estrechamente relacionada con la herramienta de visualización. Dicha herramienta realiza el análisis de los datos aplicando modelos de representación gráfica que satisfaga los objetivos del Proyecto DA.

ENTRADAS

- Producto de trabajo “Datos Formateados”.

TAREAS

- Analizar cómo representar visualmente los objetivos del Proyecto DA.
- Seleccionar los modelos gráficos de visualización que mejor representen a cada objetivo del Proyecto DA.

BUENAS PRÁCTICAS

- Seleccione distintos modelos de visualización, al menos 3 de estas clasificaciones: 1) gráfico de barras, 2) gráfico circular o pastel, 3) gráfico lineal, 4) diagrama de dispersión, 5) Mapas de calor, 6) gráfico de bala o indicadores y 7) tablas.
- Los modelos deben representar efectivamente el objetivo sin dar lugar a información engañosa.
- Tenga en cuenta las siguientes preguntas antes de seleccionar un modelo: ¿Qué es? ¿Cuándo usarlo? ¿Por qué elegir este tipo de gráfico?

SALIDAS

- Modelos de representación visual para cada objetivo.

**ACTIVIDAD 10: SELECCIONAR HERRAMIENTA****DESCRIPCIÓN DE LA ACTIVIDAD**

La selección de la herramienta de visualización es muy importante en el desarrollo del Proyecto DA, ya que ésta influirá en la evaluación de los resultados. Una herramienta dinámica, amigable, escalable y económica es lo que se busca en esta fase.

ENTRADAS

- Modelos de representación visual para cada objetivo.

TAREAS

- Analizar las herramientas que ofrece el mercado para construir tableros de control basado en los modelos definidos para cada objetivo.
- Identificar la herramienta que se acomode al presupuesto de la organización.
- Seleccionar la herramienta de visualización y configúrela.

BUENAS PRÁCTICAS

- Seleccione una herramienta de visualización que pueda usarse online y a través de dispositivos móviles.
- Seleccione una herramienta que pueda ser fácilmente mantenida por el usuario final sin mayor esfuerzo.
- Si es una herramienta de visualización de uso gratis, analice que no exponga la información de la organización.
- Revise detalladamente las medidas de seguridad de la información que tiene dicha herramienta.

SALIDAS

- Herramienta de visualización.



ACTIVIDAD 11: CONSTRUIR TABLERO DE CONTROL

DESCRIPCIÓN DE LA ACTIVIDAD

Los tableros de control se construyen conforme a los objetivos definidos en la primera fase y contemplados en los criterios de aceptación, usando la herramienta de visualización previamente seleccionada. Como resultado de esta actividad se obtendrá el tablero de control como un producto de trabajo para ser analizado en la siguiente fase de evaluación. En esta actividad se puede realizar muestreos tomando un grupo de datos pequeños en el caso de que la fuente de datos sea muy grande, con el fin de validar resultados rápidamente.

ENTRADAS

- Datos formateados
- Herramienta de visualización
- Modelos de representación gráfica de cada objetivo.

TAREAS

- Cargar los datos formateados a la herramienta de visualización.
- Construya el tablero de control según los requisitos del usuario cumpliendo cada objetivo y criterio de aceptación.

BUENAS PRÁCTICAS

- Todos los elementos del tablero de visualización deben verse en una sola pantalla.
- El tablero de control debe explicarse por sí solo sin necesidad de ir a textos adicionales.
- Las unidades de medidas deben ser claras para el usuario final.
- El tablero de control debe contar una historia sobre la PYME.
- Ubique los indicadores claves en la parte superior del tablero, luego los indicadores de contexto y por último los indicadores de detalles.
- Use moderadamente los colores sin saturar el tablero de control.

SALIDAS

- Tablero de control como un producto de trabajo.



INTERROGANTES DE ANALÍTICA

- ¿El tablero presenta problemas de calidad de datos?
- ¿Los modelos construidos son los más efectivos para representar la información?
- ¿El tablero de control cumple con los criterios de aceptación?



TABLERO DE CONTROL

El tablero de control como producto de trabajo debe elaborarse teniendo en cuenta en primer lugar los criterios de aceptación, y luego incluyendo las características de un buen tablero de control, para que sea un artefacto visualmente entendible y responsivo para la ejecución desde cualquier dispositivo, como portátil o computador de escritorio, tablet o iPad, y dispositivos móviles, como se muestra en la figura 3.

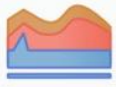


Figura 3. Tablero Power BI Responsivo [8]

Los indicadores claves de gestión o KPIs por sus siglas en ingles pueden ser elaborados según los niveles de interés que tiene la PYME y las áreas de análisis las cuales pueden ser de:

- Finanzas
- Producción
- Logísticos
- Calidad
- Recursos humanos

También pueden ser KPIs externos que suelen ser de gran importancia a la hora de elaborar un tablero de control, como es el análisis de la competencia o de clientes, e indicadores sociales, geográficos, demográficos, entre otros.



FASE 4: EVALUACIÓN

El tablero de control debe ser presentado a los interesados del proyecto para por su respectiva evaluación y aceptación. Los resultados obtenidos se evalúan de acuerdo con los criterios de aceptación y objetivos del proyecto DA. Si todos los elementos cumplen con los criterios de aceptación y objetivos de Proyecto DA, entonces se procede a la fase final de despliegue, sino se debe determinar cuáles son los criterios que no se alcanzaron y volver a una actividad anterior para corregir los problemas detectados.



ACTIVIDAD 12: EVALUAR TABLERO DE CONTROL

DESCRIPCIÓN DE LA ACTIVIDAD

El tablero de control con los respectivos modelos de representación gráfica es sujeto a evaluación por parte del analista del negocio.

ENTRADAS

- Tablero de control

TAREAS

- Comparar los resultados con uno o varios indicadores hechos manualmente.
- Verificar la calidad de los datos.
- Analizar la pertinencia de los modelos.
- Consultar la información según los filtros del tablero de control y analizar la información resultante.
- Evaluar el tablero de control según los criterios de aceptación construidos en la tabla 1.

BUENAS PRÁCTICAS

- Tenga en cuenta los criterios de aceptación durante todo el proceso de evaluación.
- Analice cada uno de los filtros de datos y los resultados obtenidos cuando se seleccionan periodos, áreas o tipos de datos.

SALIDAS

- Tablero evaluado



ACTIVIDAD 13: ANALIZAR RESULTADOS

DESCRIPCIÓN DE LA ACTIVIDAD

Esta actividad analiza los resultados obtenidos de la evaluación del tablero de control y documenta el alcance de cada criterio de aceptación con el fin de proceder a la fase de despliegue o retomar una actividad previa para mejorar los datos, el modelo o el tablero de control, o incluso si es necesario ajustar los criterios de aceptación u objetivos.

ENTRADAS

- Criterios de aceptación.
- Tablero de control evaluado

TAREAS

- Completar la tabla 1 de los criterios de aceptación.
- Decidir que actividad debe seguir para la continuación del Proyecto DA.

BUENAS PRÁCTICAS

- La evaluación es realizada por el Analista de Negocio junto con el Ingeniero de Datos, pero el Administrador de Datos puede participar activamente de ser necesario.
- Revise los interrogantes de esta fase e inclúyalos en el análisis de resultados.

SALIDAS

- Análisis de resultados.



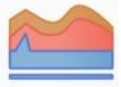
INTERROGANTES DE ANALÍTICA

- ¿Cuáles fueron los errores encontrados en la evaluación del tablero de control?, y si los hubo,
- ¿Cómo se pueden corregir los errores encontrados en la evaluación?
- ¿Cuánto tiempo se necesitará para realizar las correcciones al modelado?
- ¿Cuáles son los pasos siguientes a la evaluación?



ANÁLISIS DE RESULTADOS

El análisis de resultados es un documento formal como producto de trabajo con los hallazgos encontrados en la evaluación del tablero de control con respecto a los criterios de aceptación y las decisiones tomadas para seguir con la fase de despliegue. Este documento tendrá todas las evaluaciones realizadas junto con las interacciones con otras actividades que fueron necesarias para mejorar el modelo y el tablero de control.



FASE 5: DESPLIEGUE

La fase de despliegue consiste en automatizar la fuente de los datos para que sea formateados y que sirvan de entrada para la herramienta de visualización sin necesidad de realizar mayor esfuerzo por parte de los usuarios finales, y que los objetivos puedan ser representados gráficamente por el tablero de control. Esta fase está compuesta por las actividades de Automatizar el proceso y “Distribuir los resultados”. Como producto de trabajo se realiza un informe final al concluir la fase para informar a la stakeholders de la culminación del proyecto.



ACTIVIDAD 14: AUTOMATIZAR PROCESO

DESCRIPCIÓN DE LA ACTIVIDAD

Esta tarea consiste en automatizar las fuentes de datos para los tableros de control, donde se construye una solución que pueda tomar automáticamente los datos actualizados de la empresa, integrarlos y formatearlos para su posterior uso en el modelo y tablero de control.

ENTRADAS

- Fuente de datos
- Herramienta de gestión de datos.
- Herramienta de visualización.

TAREAS

- Construir un proceso automatizado para la integración de las fuentes de datos y el formateo de los datos.
- Establecer una conexión permanente y flexible para que la herramienta de visualización acceda a los datos formateados.
- Identificar los procesos necesarios para que los datos capturados por los usuarios tengan la calidad e integridad requerida para el Proyecto DA.
- Determinar cuales son las acciones a seguir si hay un cambio en el origen de datos debido a una decisión organizacional.
- Capacitar a los usuarios que capturan información para que los datos cumplan con la calidad e integridad necesaria para el Proyecto DA, o proponer requisitos de desarrollo para el sistema de información si fuere el caso.

BUENAS PRÁCTICAS

- Desarrolle un plan de capacitación para los usuarios cuando se requiera que algún proceso cambie en la captura de los datos.

- Identifique con qué frecuencia se actualizarán los datos que se usarán en el tablero de control.

SALIDAS

- Tablero de control automatizado.



ACTIVIDAD 15: DISTRIBUIR RESULTADOS

DESCRIPCIÓN DE LA ACTIVIDAD

Esta actividad consiste en publicar el tablero de control con los stakeholders del proyecto, mediante aplicaciones, enlaces o accesos.

ENTRADAS

- Herramienta de visualización.
- Accesos de usuarios.

TAREAS

- Compartir el tablero de control con los usuarios que tienen acceso a la herramienta de visualización
- Realizar capacitación sobre el uso de la herramienta de visualización.

BUENAS PRÁCTICAS

- Tenga en cuenta los criterios de aceptación durante todo el proceso de evaluación.
- Establezca un periodo de frecuencia para realizar mantenimiento al tablero de control.

SALIDAS

- Tablero de control distribuido.



INTERROGANTES DE ANALÍTICA

- ¿El modelo del tablero de control cumplió con las expectativas del Proyecto DA?
- ¿La organización quedó conforme con el Proyecto DA y le gustaría continuar con otro proyecto similar?
- ¿El Proyecto DA se ejecutó dentro del presupuesto asignado?
- ¿Los tiempos de desarrollo fueron acordes a los establecidos en el plan de trabajo?



INFORME FINAL

El informe final es un documento escrito como producto de trabajo con el detalle de la ejecución del Proyecto DA. Este informe está dirigido a los miembros del equipo de trabajo. Este informe debe contener el problema presentado por la empresa y cómo se dio solución a la problemática desde la analítica de datos.

Los principales apartados del informe son [9]:

- Nombre de la empresa
- Descripción del problema
- Objetivos Organizacionales relacionados con el Proyecto DA
- Objetivos del Proyecto DA.
- Costos del proyecto.
- Herramientas utilizadas.
- Principales hallazgos en la ejecución de las actividades.
- Desviaciones tenidas con respecto al Contexto del Proyecto.
- Recomendaciones
- Firma de aceptación final por parte de los stakeholders.

RECOMENDACIÓN FINAL

Si la PYME autoriza, cree un documento con información que se pueda compartir para publicar en la web como un caso de éxito de PYME en proyectos de analítica de datos.

REFERENCIAS

- [1] L. Rodríguez, "Entidades sin ánimo de lucro," *Confed. Colomb. ONG*, no. 019, pp. 2012–2013, 2011.
- [2] L. J. K. Shmueli G, Bruce PC, Yahav I, Patel NR, *Data mining for business analytics: Concepts, techniques, and applications with XLMiner*. John Wiley & Sons, 2016.
- [3] J. Mi. Moine, A. Haedo, and S. Gordillo, "Estudio comparativo de metodologías para minería de datos," *XIII Work. Investig. en Ciencias la Comput.*, pp. 278–281, 2011.
- [4] KDnuggets, "What main methodology are you using for your analytics, data mining, or data science projects?," *KDnuggets*, 2014. [Online]. Available: <https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>. [Accessed: 31-Oct-2017].
- [5] V. . Menendez and M. . Castellanos, "Software Process Engineering Metamodel (SPEM)," *Rev. Latinoam. Ing. Softw.*, vol. 3, no. 2, pp. 92–100, 2008.
- [6] DataCamp, "The Data Science Industry: Who Does What (Infographic) (article) - DataCamp," *2015-09-10*, 2015. [Online]. Available: <https://www.datacamp.com/community/tutorials/data-science-industry-infographic>. [Accessed: 18-Oct-2018].
- [7] PMOinformatica.com, "Plantillas Scrum: historias de usuario y criterios de aceptaciónNo Title," *La oficina de proyectos de informática*, 2012. [Online]. Available: <http://www.pmoinformatica.com/2012/10/plantillas-scrum-historias-de-usuario.html>. [Accessed: 28-Feb-2019].
- [8] Microsoft, "Business intelligence like never before," 2019. [Online]. Available: <https://powerbi.microsoft.com/en-us/>. [Accessed: 01-Mar-2019].
- [9] I. B. M. IBM, "Manual CRISP-DM de IBM SPSS Modeler," *IBM Corp.*, p. 56, 2012.