CRISP-DM/SMES;

UNA METODOLOGÍA DE PROYECTOS DE ANALÍTICA DE DATOS

PARA LAS PYME

JHON FREDY MONTALVO GARCIA

UNIVERSIDAD DE MEDELLIN

MAESTRIA EN INGENIERIA DE SOFTWARE

FACULTAD DE INGENIERIAS

MEDELLIN

2019

CRISP-DM/SMES;
UNA METODOLOGÍA DE PROYECTOS DE ANALÍTICA DE DATOS
PARA LAS PYME

JHON FREDY MONTALVO GARCIA

Trabajo de grado presentado como requisito para optar al título de

MAGISTER EN INGENIERIA DE SOFTWARE

Director: JUAN BERNARDO QUINTERO

PhD en Ingeniería Electrónica

Codirector: BELL MANRIQUE LOSADA

PhD en Ingeniería

UNIVERSIDAD DE MEDELLIN

FACULTAD DE INGENIERIAS

MAESTRIA EN INGENIERIA DE SOFTWARE

MEDELLIN

2019

Nota de aceptación:

Director

Jurado

Jurado

Medellín, junio 6 de 2019.

Dedicado a Dios y a mi familia.

Agradecimientos

A Dios, fuente de vida, por darme la salud, sabiduría y demás bendiciones, para crecer académicamente dentro del área que me asignando para cumplir su labor.

A la Iglesia Adventista del Séptimo Día de Colombia – Unión Colombiana del Norte por apoyarme financieramente en los estudios realizados a nivel de postgrado.

A la Universidad de Medellín por proveer los recursos académicos para el desarrollo de la Maestría.

Al Doctor Juan Bernardo Quintero y la Doctora Bell Manrique, por asumir la dirección de mi proyecto de grado y brindarme todas las asesorías para el desarrollo de la investigación.

A Álvaro Daniel Castillo, por su disposición y la evaluación de la propuesta en el caso de estudio.

A la Cooperativa Unión Colombiana – COOMUNION por permitir evaluar la propuesta en su organización y sus aportes a la investigación.

Tabla de contenido

Resumen 13
Summary 14
1. Introducción 16
2. Contexto de la Investigación 19
2.1 Planteamiento del problema 19
2.1.1 Justificación. 19
2.1.2 Pregunta de investigación. 20
2.1.3 Hipótesis. 21
2.1.4 Definición de variables 21
2.2 Objetivos 21
2.2.1 Objetivo general. 21
2.2.2 Objetivos específicos. 22
2.3 Alcances y limitaciones de la investigación 22
2.3.1 Alcances 22
2.3.2 Limitaciones. 23
3. Marco teórico 25
3.1 Principales metodologías de analítica de datos 25
3.1.1 KDD 26
3.1.2 SEMMA. 27
3.1.3 CRISP-DM. 29
3.2 Pequeña y Mediana Empresa 30
3.2.1 Clasificación de las PYME. 30
3.2.2 PYME ESAL. 31
3.2.3 Panorama de analítica en PYME. 33
3.3 Requerimientos de las PYME para implementar Proyectos de analítica de datos. 34
4. Diseño metodológico 36
4.1 Tipo de investigación 36
4.2 Diseño de la investigación 36
4.2.1 Ciclo de análisis del entorno o relevancia. 37
4.2.2 Ciclo de análisis de la base de conocimiento o rigor 38
4.2.3 Ciclo del diseño y evaluación del modelo. 38
5. Revisión de literatura 41
5.1 Selección de fuentes 41
5.2 Cadenas de búsqueda. 41
5.3 Selección de estudios 42
5.3.1 Criterios de inclusión y exclusión. 42
5.3.2 Estudios seleccionados. 43

5.4 Resultados de los estudios 44
5.5 Comparacion de KDD, SEMMA y CRISP-DM 45
5.6 Modelo de referencia para proyectos de analítica....46
5.7 Elementos de la metodología CRISP-DM adaptados. 48
6. Diseño de la propuesta metodológica CRISP-DM-SMEs....51
6.1 Introducción a CRISP-DM/SMEs....51
6.2 Conceptos de la Metodología....51
6.2.1 Roles. 52
6.2.2 Fases. 53
6.2.3 Actividades. 54
6.2.4 Productos de trabajo. 54
6.2.5 Guía. 54
6.2.6 Herramientas. 54
6.2.7 Representación gráfica. 55
6.3 Fase 1: Definición del Proyecto de Analítica 57
6.3.1 Herramienta de entorno de trabajo colaborativo. 57
6.3.2 Actividad 1: Seleccionar objetivos empresariales. 58
6.3.3 Actividad 2: Definir objetivos del proyecto. 59
6.3.4 Criterios de aceptación. 60
6.3.5 Actividad 3: Asignar recursos. 63
6.3.6 Actividad 4: Determinar el alcance y riesgos. 66
6.3.7 Interrogantes de analítica. 67
6.3.8 Contexto del Proyecto. 68
6.4 Fase 2: Gestión de datos 69
6.4.1 Herramienta para gestión de datos....70
6.4.2 Actividad 5: Recolectar datos. 70
6.4.3 Actividad 6: Explorar datos. 71
6.4.4 Actividad 7: Integrar datos. 72
6.4.5 Actividad 8: Formatear datos. 73
6.4.6 Interrogantes de analítica. 74
6.4.7 Datos Formateados. 75
6.5 Fase 3: Modelado 75
6.5.1 Guía de Modelado. 75
6.5.2 Herramienta de visualización. 75
6.5.3 Actividad 9: Seleccionar modelo. 76
6.5.4 Actividad 10: Seleccionar herramienta. 77
6.5.5 Actividad 11: Construir tablero de Control. 78
6.5.6 Interrogantes de analítica. 79
6.5.7 Tablero de Control 80
6.6 Fase 4: Evaluación 81
6.6.1 Actividad 12: Evaluar tablero de Control. 81
6.6.2 Actividad 13: Analizar resultados. 82

6.6.3 Interrogantes de analítica. 83
6.6.4 Análisis de resultados. 84
6.7 Fase 5: Despliegue 84
6.7.1 Actividad 14: Automatizar proceso 84
6.7.2 Actividad 15: Distribuir resultados 86
6.7.3 Interrogantes de analítica 86
6.7.4 Informe Final 87
6.7.5 Recomendación final 87
7. Evaluación comparativa de la propuesta metodológica 89
7.1 Evaluación de los aspectos del marco comparativo 90
7.1.1 Aspecto 1: Descripción de las actividades de cada fase. 90
7.1.2 Aspecto 2: Escenarios de aplicación. 90
7.1.3 Aspecto 3: Actividades específicas que componen cada fase. 91
7.1.4 Aspecto 4: Actividades destinadas a la dirección del proyecto. 92
7.2 Evaluación final del marco comparativo 93
8. Caso de estudio 94
8.1 Definición del Caso de Estudio 94
8.1.1 Selección de la PYME. 94
8.1.2 Descripción del Proyecto DA. 94
8.1.3 Esfuerzo actual de la PYME. 96
8.2 Aplicación de la Propuesta Metodológica 97
8.2.1 Roles en el caso de estudio. 97
8.2.2 Fuentes de datos 97
8.2.3 Herramientas en el caso de estudio 98
8.2.4 Producto de trabajo: Tablero de control. 99
8.2.5 Esfuerzo según la propuesta metodológica. 99
8.2.6 Resultados del Caso de Estudio 101
8.3 Definición de la línea base 103
8.3.1 Juicio de Expertos. 103
8.3.2 Esfuerzo según la línea base. 107
9. Análisis y discusión de resultados 108
10. Conclusiones 112
11. Recomendaciones, trabajos futuros y aportes de la investigación 114
11.1 Recomendaciones 114
11.2 Trabajos futuros 114
11.3 Aportes de la investigación 115
11.3.1 Artículo científico 115
11.3.2 Publicación de la Metodología 115
REFERENCIAS BIBLIOGRÁFICAS 116

ANEXOS 121
Anexo 1: Clasificación de las ESAL 121
Anexo 2: Evaluación comparativa de la propuesta metodológica 122
Anexo 3: Organigrama Cooperativa Unión Colombiana 128
Anexo 4: Niveles de riesgo de liquidez 129
Anexo 5: Cuadrante mágico de Gartner para plataformas de analítica 132
Anexo 6: Certificado de participación ICICT 2019 133

Lista de tablas

Tabla 1. Criterios de búsqueda en las bases de dato.....42

Tabla 2. Resultados de la búsqueda en las bases de datos.....42

Tabla 3. Selección de estudios.....43

Tabla 4. Criterios de análisis 45

Tabla 5. Resumen de las correspondencias entre KDD, SEMMA y CRISP-DM.....46

Tabla 6. Comparación CRISP-DM con CRISP-DM/SME's.....49

Tabla 7. Criterios de Aceptación 62

Tabla 8. Participación de roles en las actividades 63

Tabla 9. Asignación de recursos.....65

Tabla 10. Riesgos y contingencias 67

Tabla 11. Evaluación del nivel de detalle en las actividades que componen cada fase 90

Tabla 12. Evaluación de los escenarios de aplicación.....91

Tabla 13. Evaluación general de las actividades específicas.....92

Tabla 14. Evaluación general para las actividades de dirección del proyecto 92

Tabla 15. Evaluación final de todos los aspectos del marco comparativo 93

Tabla 16. Esfuerzo actual de la PYME.....96

Tabla 17. Roles en el caso de estudio.....97

Tabla 18. Herramientas en el caso de estudio. 98

Tabla 19. Esfuerzo según las actividades de la propuesta metodológica. 101

Tabla 20. Evaluación de juicios de expertos 105

Lista de figuras

Figura 1. Estructura de la tesis. Elaboración propia. 18

Figura 2. Principales metodologías para analítica. KDnuggets (2014) 25

Figura 3. Etapas de KDD. Fayyad et al. (1996).....27

Figura 4. Proceso de desarrollo del modelo SEMMA. Azevedo y Santos (2008).....28

Figura 5. Fases del modelo de referencia CRISP-DM. Chapman et al. (2000).....29

Figura 6. Actividades y Tareas CRISP-DM. Chapman et al. (2000).....30

Figura 7. Factores que preocupan a las PYME. Sinnetic (2017).....34

Figura 8. Requerimientos de las PYME para implementar Proyectos DA. Elaboración propia. 35

Figura 9. Ciencia basada en el diseño. Gonzalez y Pomares (2012). 37

Figura 10. CRISP-DM/SME's basado en CRISP-DM. Elaboración propia.....47

Figura 11. Elementos usados de SPEM. Menendez y Castellanos (2008) 52

Figura 12. Representación gráfica de CRISP'DM-SMEs. Elaboración propia 56

Figura 13. Tablero Power BI Responsivo. Tomado de https://powerbi.microsoft.com/en-us/ ...80

Figura 14. Aspectos del Marco Comparativo. Moina y Haedo (2015).....89

Figura 15. Tablero de control en Power BI. Elaboración propia. 100

Figura 16. Porcentaje de esfuerzo por fases. Elaboración propia 102

Figura 17 . Esfuerzo según la metodología en el caso de estudio. Elaboración propia. 109

Glosario

PYME: Pequeña y Mediana Empresa.

SME: PYME en inglés (Small and Medium-sized Enterprises).

ESAL: Entidades Sin Ánimo de Lucro (Non-profit).

DA: Analítica de Datos (Data Analytics).

CRISP-DM: Metodología de minería de datos (Cross Industry Standard Process for Data Mining).

SPEM: Lenguaje estándar de modelado de procesos de desarrollo de software orientado a productos de trabajo (Software & Systems Process Engineering Meta-Model).

Resumen

El aumento exponencial de la información debido al avance tecnológico y el desarrollo de las comunicaciones ha creado la necesidad de tomar decisiones basadas en el análisis de los datos. Por un lado, las empresas se ven en la necesidad de seguir metodologías de minería de datos – DA, para gestionar los grandes volúmenes de información con herramientas Big Data; tendencia que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones. Por otro lado, existen las pequeñas y medianas empresas – PYME Sin Ánimo de Lucro – ESAL que realizan esfuerzos para abordar la analítica de datos según sus diversas fuentes y formas, encontrando desafíos como la falta de conocimiento en herramientas metodológicas y de software, que le permitan un despliegue oportuno para la toma de decisiones. En este trabajo se propone CRISP-DM/SMEs, una metodología de analítica de datos para PYME ESAL. El diseño de esta metodología está basado en CRISP-DM como marco de referencia, se representa con SPEM y se caracteriza por ser sencilla, flexible, y con bajo costo de implementación. La evaluación de la metodología se realizó bajo un marco comparativo y en la aplicación de un caso de estudio en una PYME ESAL donde los resultados fueron positivos en cada uno de los aspectos evaluados y demostró una disminución del esfuerzo durante la ejecución del proyecto DA en comparación con el proceso actual de la PYME y la línea base de referencia.

Palabras claves: ANALÍTICA DE DATOS, CRISP-DM, PYME, ESAL.

Summary

The exponential increase in information due to technological progress and the development of communications has created the need to make decisions based on data analysis. On the one hand, companies are in need of following data mining methodologies - DA, to manage large volumes of information with Big Data tools; trend that has opened the doors to a new approach to understanding and decision making. On the other hand, there are small and medium-sized enterprises - Non-Profit SMEs, that make efforts to address data analytics according to their various sources and forms, finding challenges such as lack of knowledge in methodological tools and software, which allow for timely deployment for decision making. This paper proposes CRISP-DM / SMEs, a data analytics methodology for Non-Profit SMEs. The design of this methodology is based on CRISP-DM as a frame of reference, it is represented with SPEM and it is characterized by being simple, flexible, and with low cost of implementation. The evaluation of the methodology was carried out under a comparative framework and in the application of a case study in an Non-Profit SMEs where the results were positive in each of the aspects evaluated and demonstrated a decrease in effort during the execution of the DA project in comparison with the current process of the SME and the baseline of reference.

Keywords: DATA ANALYTICS, CRISP-DM, SME, NON-PROFIT

PARTE I

INTRUCCIÓN

1. Introducción

En esta época donde los productos, el mercado y la competencia son tan dinámicos, es complejo mantener un nivel de competitividad sólido y sostenible en las empresas para garantizar una posición en el mercado. Autores como Nenzhelele y Pellissier (2014) afirman que las empresas se enfrentan a un entorno cada vez más competitivo en el que es difícil mantener una ventaja sostenida. Davenport (2013) asegura que bajo el contexto de Analytics 3.0, la analítica impulsa la economía de datos en la empresa, y para ello se necesitarán de nuevos enfoques para la toma de decisiones. Los administradores deben sentirse cómodos con la experimentación basada en datos y repensar fundamentalmente cómo el análisis de los datos puede crear valor para ellos y para sus clientes.

Los métodos y técnicas de análisis inteligente de datos, también conocidos como algoritmos de minería de datos, se han convertido en un área de investigación importante, ya que su aplicación conjunta con análisis de datos tradicionales puede revelar relaciones ocultas de conocimiento, patrones de comportamiento, perfiles de entidad y regularidades similares en datos almacenados en grandes bases de datos o almacenes (Olivera, Sasa, y Zita, 2009). Pero a pesar de que las herramientas de DA fueron concebidas para el tratamiento de la información sin importar el tipo de organización (Florez, 2012), los estudios se centran en las grandes empresas por su capacidad económica. Por lo anterior se infiere que las PYMES no son el foco de estudio para el descubrimiento del conocimiento a través de la DA.

Dicho descubrimiento es un proceso muy intrincado, incierto y que consume mucho tiempo (Olivera et al., 2009). Por lo tanto, es de suma importancia seguir la orientación metodológica existente, entre las cuales el CRISP-DM es la más favorecida (Oztekin, Best,

y Delen, 2014). Aunque CRISP-DM fue concebido para cualquier tipo de organización, sus tareas pueden convertirse en un proceso largo y complejo para las PYME, y es por eso por lo que se pretende analizar cada una de sus fases para determinar cuáles son las más relevantes en la implementación de Proyectos DA en las PYMES.

Como consecuencia, se propone una metodología para Proyectos DA en PYME, que disminuye el esfuerzo con respecto a CRISP-DM. Y para evaluar la propuesta se desarrolla en un caso de estudio, donde se selecciona una PYME sin ánimo de lucro (ESAL) con interés en un Proyecto DA que incluye tableros de control.

Para el desarrollo de la propuesta metodológica se emplea una metodología de investigación científica basada en el diseño (Gonzalez y Pomares, 2012), cuyo objetivo es contribuir a la solución de un problema relevante en la industria, al mismo tiempo que se hace un aporte en el área del conocimiento de la analítica de datos, de una manera novedosa y rigurosa a través del diseño de productos de trabajo. Sus respectivos ciclos de relevancia, diseño y rigor se enmarcan en las partes de exploración, propuesta y evaluación del contenido de este trabajo. En total son seis partes las que contiene el desarrollo de esta investigación, como se muestra en la figura 1.

Figura 1. Estructura de la tesis. Elaboración propia.

2. Contexto de la Investigación

2.1 Planteamiento del problema

2.1.1 Justificación.

El procesamiento de información, según Palmer y Hartley (Guarda, Santos, Pinto, Augusto, & Silva, 2013), se ha convertido gradualmente en la base para lograr una ventaja competitiva y por lo tanto las organizaciones tienen que creer que tienen la información correcta en el momento adecuado y para las personas indicadas. Para Soto (2004), se debe proporcionar a los directivos de las empresas las herramientas apropiadas para la explotación y análisis de los datos que les permitan obtener el conocimiento necesario en el proceso de toma de decisiones estratégicas. Es así como en la última década los almacenes de datos (en inglés *Data Warehouse*, DW) se han convertido en un componente esencial para lograr competitividad con los modernos sistemas de apoyo a la toma de decisiones en la mayoría de las empresas.

Sin embargo, las empresas que han logrado alcanzar niveles de maduración en la toma de decisiones basada en la analítica, son las que manejan grandes volúmenes de datos y estas sólo representan una mínima parte del mundo de los negocios. Mientras que "Las PYME constituyen la forma dominante de organización empresarial en todos los países del mundo, ya que representan más del 95% y hasta el 99% de la población empresarial según el país" (Pytel, Hossian, Britos, y Garcia-Martinez, 2015, p. 2); se están quedando atrás de las grandes empresas (Gudfinnsson y Strand, 2018), a pesar que se consideran importantes en la economía nacional (Mullins et al., 2007).

En el marco de las PYME y del sector social, están las entidades sin ánimo de lucro (ESAL), quienes persiguen los fines sociales y comunitarios. Aunque no generan reparto de

utilidades ni enriquecimiento a los socios, sí se toman decisiones estratégicas y financieras que involucran el análisis de grandes volúmenes de datos. No obstante, las PYME ESAL, recopilan información de distintas fuentes y están interesadas en los sistemas de inteligencia de negocios (*Business Intelligent*, BI) (Lawton, 2009), y en la tendencia hacia la analítica de datos (*Data Analytics*, DA), la cual va en aumento gracias al avance tecnológico.

Pero a pesar de este interés, el desarrollo de proyectos de analítica de datos se ve frustrado, ya que su implementación suele ser una tarea compleja debido a los costos que genera. Según Flórez (2012), estos costos están asociados a la infraestructura tecnológica, el costo administrativo, la capacitación de personal y las herramientas de software. Es así como la implementación de proyectos de analítica en PYME ESAL tiene pocas alternativas, ya que la analítica se enfoca en las grandes empresas que tienen mayor capacidad financiera (Guarda et al., 2013) para implementar metodologías de analítica de datos.

Las metodologías que usan las empresas para el desarrollo de proyectos de analítica son KDD, SEMMA y CRISP-DM, por ser las más referenciadas, pero ninguna de ellas tiene un enfoque hacia las PYME. En consecuencia, surge la necesidad de desarrollar un modelo de referencia que pueda ayudar a disminuir el esfuerzo en los proyectos de analítica de datos en las PYME ESAL.

2.1.2 Pregunta de investigación.

¿Cómo se puede disminuir el esfuerzo en proyectos de analítica de datos adaptando la metodología CRISP-DM para PYME ESAL?

2.1.3 Hipótesis.

La implementación de una metodología modificada de CRISP-DM disminuye el esfuerzo en los proyectos de analítica de datos en las PYMES ESAL, en comparación con los proyectos de analítica del entorno empresarial que utilizan otras metodologías o ninguna en específico.

2.1.4 Definición de variables

Las variables de la investigación están explícitas en la pregunta de investigación, las cuales son:

a) Variables independientes:

- Metodología de analítica de datos modificada de CRISP-DM
- Otra o ninguna metodología de analítica de datos.

b) Variable dependiente:

Esfuerzo en los proyectos de analítica de datos en las PYME.

El esfuerzo se medirá según las horas de trabajo hombre invertidas en el Proyecto de analítica de datos según las metodologías implementadas.

2.2 Objetivos

2.2.1 Objetivo general.

Proponer una metodología para proyectos de analítica de datos en PYMES ESAL a partir de la metodología CRISP-DM como modelo de referencia.

2.2.2 Objetivos específicos.

Para responder la pregunta de investigación y validar las hipótesis se plantean los siguientes objetivos específicos (OE) identificados con un código para su posterior referencia:

OE1. Identificar los requerimientos que tienen las PYME ESAL para implementar metodologías de proyectos de analítica de datos.

OE2. Definir los elementos de las metodologías de analítica de datos que deberían adaptarse para la implementación de proyectos de analítica de datos en las PYME ESAL.

OE3. Diseñar una metodología a partir de los requerimientos identificados y los elementos adaptados para disminuir el esfuerzo en los proyectos de analítica de datos en las PYME ESAL.

OE4. Evaluar la metodología propuesta con la aplicación de un caso de estudio, en un Proyecto de analítica de datos en una PYME ESAL de la ciudad de Medellín.

2.3 Alcances y limitaciones de la investigación

2.3.1 Alcances.

Para cumplir con el objetivo general, el alcance de la investigación estará dado con la creación de una propuesta metodológica como modelo de referencia para proyectos de analítica de datos en PYME ESAL. Para la divulgación del trabajo se presenta un artículo científico para ser sujeto a revisión por un proceedings en evento científico. Y finalmente, para cumplir con los objetivos específicos, se entregará el documento de investigación con el desarrollo de cada una de las partes descritas en la introducción.

Para el caso de estudio, se tomará una PYME ESAL de la ciudad de Medellín con la intención de implementar un Proyecto de analítica de datos que involucre la construcción de tableros de control.

2.3.2 Limitaciones.

La presente investigación tiene como limitaciones lo siguiente:

- Cuando se refiere a las metodologías de analítica de datos, solo se incluyen las más referenciadas en el campo empresarial, a saber; KDD, SEMMA y CRISP-DM, las cuales tienen un paralelismo según Azevedo y Santo (2008).
- Los tipos de PYME objeto de estudio, son las que se clasifiquen dentro de las Entidades Sin Ánimo de Lucro (ESAL), como las instituciones de educación superior, instituciones religiosas, fundaciones, cooperativas, entre otras.
- Los proyectos de analítica a los que está orientado este estudio son aquellos que involucren la construcción de tableros de control o *dashboard* en inglés.

PARTE II

FUNDAMENTOS TEÓRICOS

3. Marco teórico

3.1 Principales metodologías de analítica de datos

Los modelos más referenciados que se encuentra en la comunidad científica y que han sido propuestos para el desarrollo de este tipo de proyectos son: KDD (*Knowledge Discovery in Databases),* SEMMA (*Sample, Explore, Modify, Model, Assess*) y CRISP-DM (*Cross Industry Standard Process for Data Mining*), siendo CRISM-DM la más usada en los últimos años (KDnuggets, 2014), como se evidencia en la figura 2.

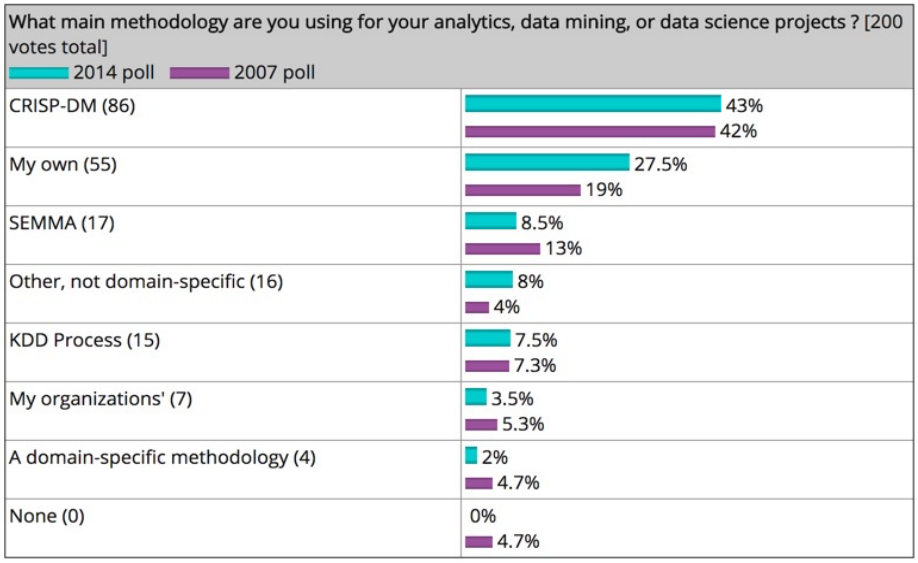


Figura 2. Principales metodologías para analítica. KDnuggets (2014)

Estos modelos no son recientes. A inicios del año 1996, el modelo KDD se constituyó en el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información. Luego, a partir del año 2000, con el gran crecimiento que surgió en el área de la minería de datos, se desarrollaron dos nuevos modelos que plantean

un enfoque sistemático para llevar a cabo el proceso: SEMMA y CRISP-DM. A continuación, se describe cada una de ellas.

3.1.1 KDD.

KDD fue presentado en 1996 como el proceso de DM para extraer lo que se considera conocimiento según la especificación de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación requeridos de la base de datos. Se consideran cinco etapas: (i) Selección, que consiste en crear un conjunto de datos objetivo, o centrarse en un subconjunto de variables o muestras de datos, en el que se debe realizar el descubrimiento; (ii) Preprocesamiento, la cual consiste en la limpieza y preprocesamiento de datos de destino para obtener datos coherentes; (iii) Transformación, que consiste en la transformación de los datos utilizando métodos de reducción o transformación de dimensionalidad; (iv) Minería de datos, que consiste en buscar patrones de interés en una forma de representación particular, según el objetivo de DM (generalmente, predicción); y (v) Interpretación/Evaluación, que consiste en la interpretación y evaluación de los patrones minados (Fayyad, Piatetsky-Shapiro, y Smyth, 1996). En la figura 3 se representan las etapas anteriormente mencionadas.

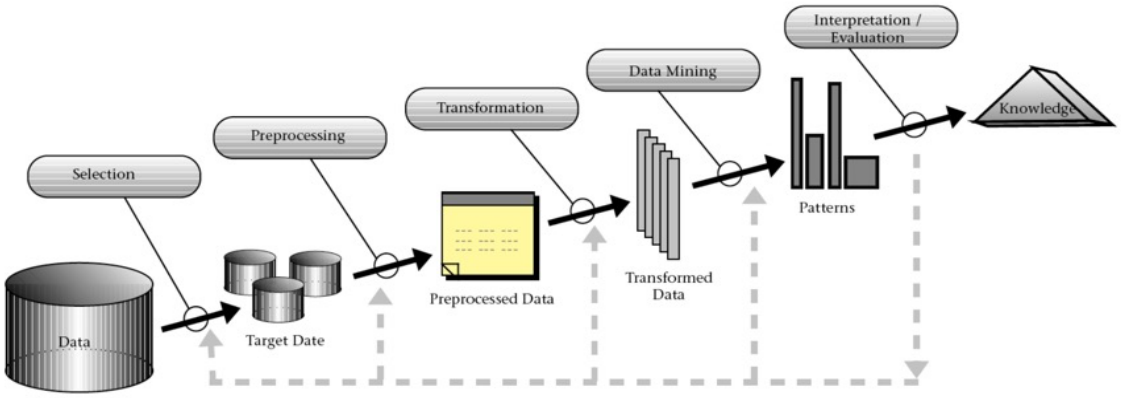


Figura 3. Etapas de KDD. Fayyad et al. (1996)

3.1.2 SEMMA.

El acrónimo SEMMA significa Muestrear, Explorar, Modificar, Modelar, Evaluar (*Sample, Explore, Modify, Model, Assess*), asociado a las fases del proceso de conducir un proyecto de DM. El SAS Institute (citado en Azevedo y Santos, 2008) considera un ciclo con 5 etapas:

- Muestrear: consiste en muestrear los datos extrayendo una porción de un gran conjunto de datos lo suficientemente grande como para contener la información significativa, pero lo suficientemente pequeña como para manipularla rápidamente.
- Explorar: consiste en la exploración de los datos mediante la búsqueda de tendencias y anomalías imprevistas con el fin de obtener comprensión e ideas.
- Modificar: consiste en la modificación de los datos creando, seleccionando y transformando las variables para enfocar el proceso de selección del modelo. Modelar: consiste en modelar los datos al permitir que el software busque automáticamente una combinación de datos que prediga de manera confiable un resultado deseado.

- Evaluar: consiste en evaluar los datos mediante la evaluación de la utilidad y fiabilidad de los hallazgos del proceso de DM y estimar qué tan bien funciona.

La metodología SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS. Este producto organiza sus herramientas con base en las distintas fases que componen la metodología. Es decir, el software proporciona un conjunto de herramientas especiales para la etapa de muestreo, otras para la etapa de exploración, y así sucesivamente. Sin embargo, el usuario podría usarlo siguiendo cualquier otra metodología de minería de datos como CRISP-DM (Moine, Haedo, y Gordillo, 2011). En la figura 4 se representa el proceso del modelo SEMMA.

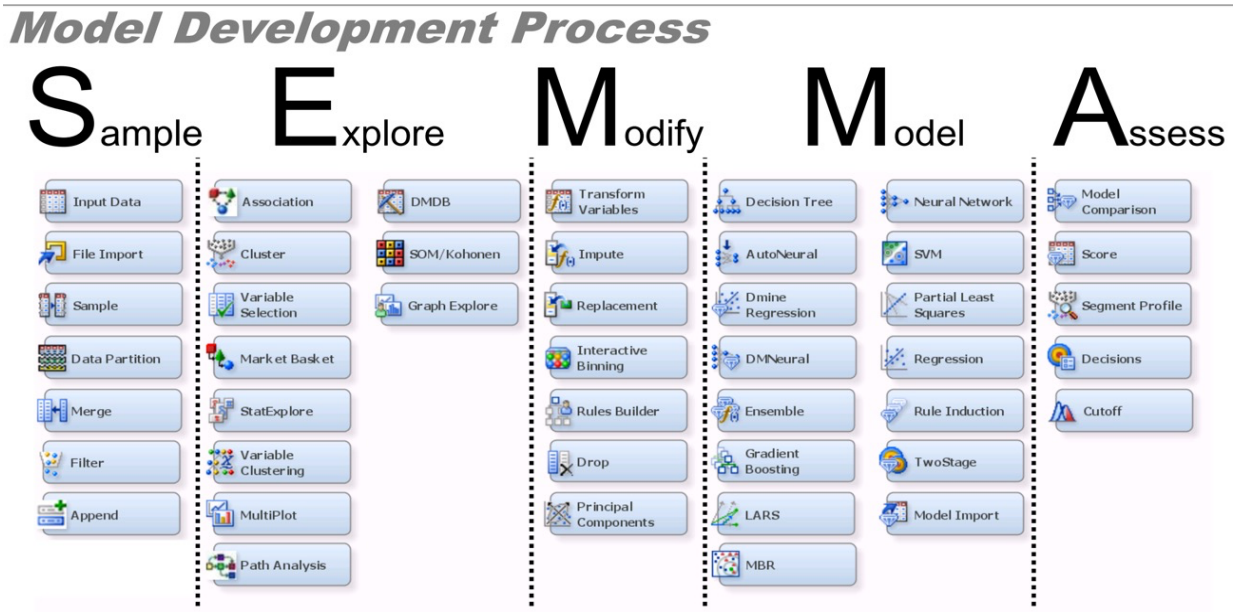


Figura 4. Proceso de desarrollo del modelo SEMMA. Azevedo y Santos (2008)

3.1.3 CRISP-DM.

CRISP–DM fue creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000 (Moine et al., 2011) y consta de seis fases, que se muestran en la figura 5. La secuencia de las fases no es rígida. Siempre se requiere avanzar y retroceder entre ellas. El resultado de cada fase determina qué fase, o tarea particular de una fase debe realizarse a continuación. Las flechas indican las dependencias más importantes y frecuentes entre las fases. El círculo exterior en la figura 5 simboliza la naturaleza cíclica de la minería de datos en sí misma. La extracción de datos no termina una vez que se implementa una solución. Las lecciones aprendidas durante el proceso y a partir de la solución implementada pueden generar nuevas preguntas empresariales, a menudo más enfocadas. Los procesos posteriores de minería de datos se beneficiarán de las experiencias de los anteriores (Chapman et al., 2000).

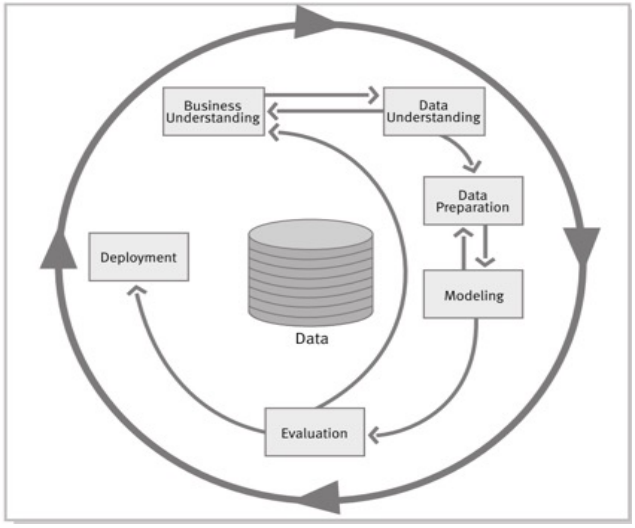


Figura 5. Fases del modelo de referencia CRISP-DM. Chapman et al. (2000)

Cada fase de CRISP-DM contiene tareas genéricas (en negrilla) y salidas (en cursiva), como se evidencia en la figura 6.

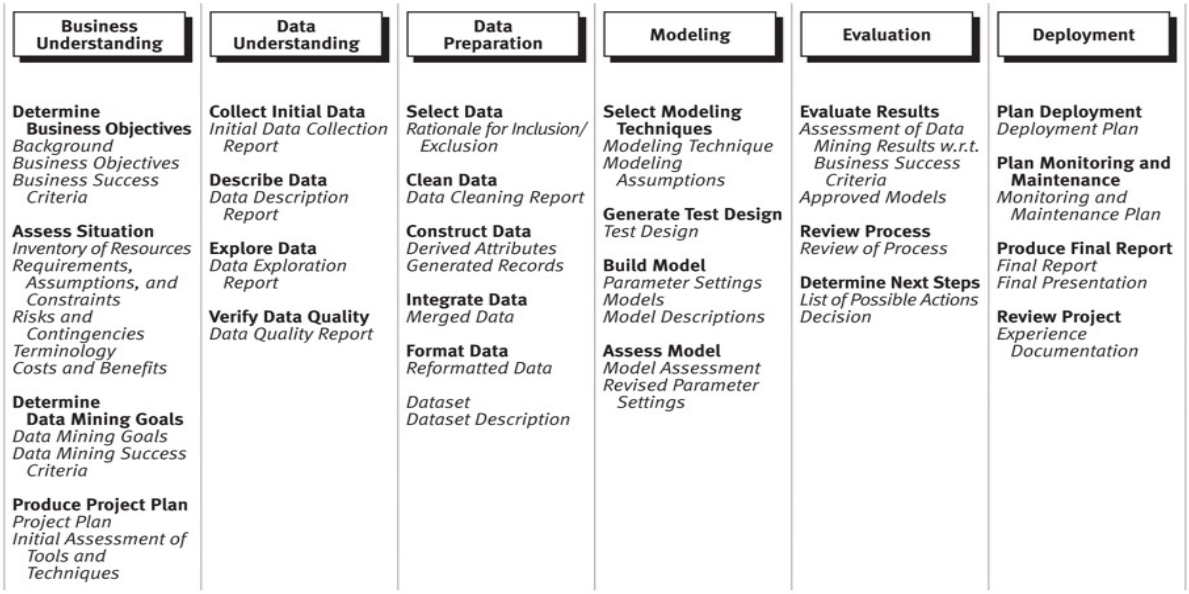


Figura 6. Actividades y Tareas CRISP-DM. Chapman et al. (2000)

3.2 Pequeña y Mediana Empresa

3.2.1 Clasificación de las PYME.

En Colombia, según la Ley para el Fomento de la Micro, Pequeña y Mediana Empresa, Ley 905, las PYME se clasifican de acuerdo a número de trabajadores y sus activos totales, a saber; 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.

De acuerdo a esta clasificación, las PYME representan el 80% del empleo en el país (Revista Dinero, 2016). Y según el Ministerio de Comercio, Industria y Turismo

(MinCIT) hasta el 8 de agosto 2016 estaban registradas en las Cámaras de Comercio 2’518.120 pequeñas y medianas empresas, de las cuales 39,9% corresponden a sociedades y el restante 60,1% son personas naturales.

3.2.2 PYME ESAL.

Las ESAL son entidades 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 (Rodríguez, 2011). Por lo tanto, pueden beneficiar a los asociados, a terceros o al público en general. Aunque no persiguen distribuir utilidades, en algunos casos sí hay reparto de excedentes, como sucede con las entidades del sector solidario.

Las ESAL han tomado una gran importancia en el mundo, no solamente como organizaciones que prestan servicios de tipo social, sino como generadoras de empleo e impulsoras de la actividad económica. A causa de esto, las ESAL son agrupadas en 12 grupos a nivel internacional (Rodríguez, 2011), las cuales son inspeccionadas, vigiladas o controladas por distintas entidades del estado como alcaldías, gobernaciones, superintendencias y ministerios del gobierno. según la Confederación Colombiana de ONG (CCONG, 2016), las ESAL se clasifican en los siguientes grupos:

- Grupo 1: Organizaciones no gubernamentales ONG
- Grupo 2: Movimiento comunal.
- Grupo 3: Instituciones políticas.
- Grupo 4: Instituciones educativas y culturales
- Grupo 5: Instituciones de economía solidaria y social

- Grupo 6: Instituciones de Educación
- Grupo 7: Iglesias
- Grupo 8: Bienestar familiar
- Grupo 9: Instituciones del sector privado – empresarial
- Grupo 10: Convivencia
- Grupo 11: Medios de comunicación comunitarios
- Grupo 12: Instituciones de representación de trabajadores
- Grupo 13: Indígenas
- Grupo 14: Campesinos
- Grupo 15: Protección de animales
- Grupo 16: Defensa de los consumidores

El anexo 1 amplía la información sobre los tipos de organizaciones que conforman cada grupo, la cantidad de empresas por grupos, y las entidades supervisoras de dichas entidades.

Las ESAL tienen como principal fuente de ingreso los dineros recibidos por parte de personas naturales, jurídicas o entidades públicas en representación del Estado, por medio de las donaciones o subvenciones. En cuanto a la estructura organizacional, las ESAL tienen un grado de complejidad, por la amplia gama de posibilidades de organización (central, divisional, funcional o geográfica). Dentro del grupo de las ESAL más reconocidas están las Fundaciones, Universidades, Corporaciones, Asociaciones, Cooperativas, e Iglesias.

3.2.3 Panorama de analítica en PYME.

La capacidad de las PYME para tener éxito frente a competidores más grandes se centra en la intuición personal y la capacidad de proporcionar un servicio superior. Dado que los grandes datos están cambiando el panorama empresarial, algunos competidores grandes y pequeños están utilizando Big Data para mejorar la calidad del producto, las operaciones de marketing y las relaciones con los clientes. Esta nueva eficiencia de los competidores más grandes puede ser una amenaza real para la sostenibilidad del negocio de las PYME (Ogbuokiri, Udanor, y Agu, 2015).

Según la Fundación Unipymes (Unipymes, 2014), de cada 10 empresas que en el país están accediendo a la analítica de datos, hay dos o tres que son de tamaño mediano. Es decir que entre el 20% y 30% de las PYME están empezando a abordar la analítica para mejorar sus capacidades y conseguir sus objetivos, ya que esto conlleva a organizar el volumen de datos, descubrir y visualizar información que les permita tomar decisiones más acertadas y ejecutar estrategias más efectivas.

Sin embargo, hay otras preocupaciones en las PYME que se interponen al crecimiento de la analítica de datos, y son las nuevas regulaciones que impone el gobierno cada año. Un claro ejemplo de ello es la implementación de las Normas Internacionales de Información Financiera-NIIF, políticas de protección de datos, facturación electrónica y regulaciones de la Unidad de Gestión Pensional y Parafiscales -UGPP, que han incrementado en gran porcentaje del año 2016 al 2017 como se puede evidenciar en la figura 7. Esto ha dejado a un lado el avance en la transformación digital, aspecto clave en la carrera de Big Data.

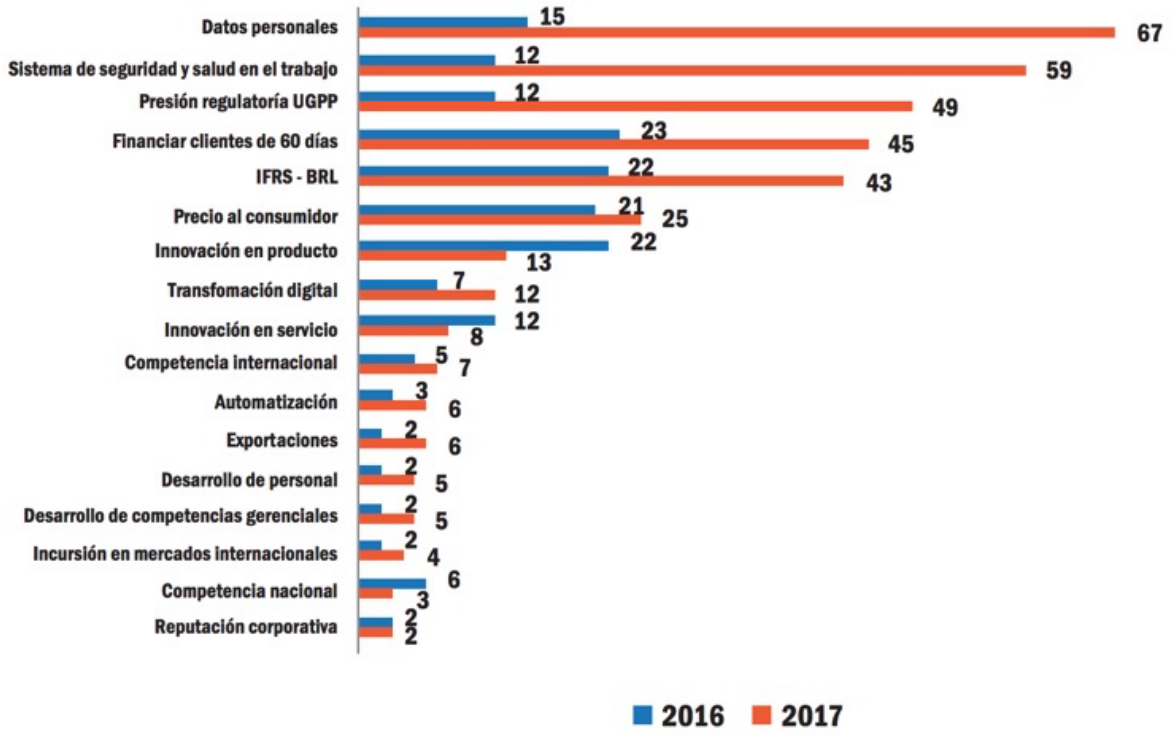


Figura 7. Factores que preocupan a las PYME. Sinnetic (2017)

3.3 Requerimientos de las PYME para implementar Proyectos de analítica de datos.

A pesar de las preocupaciones empresariales a nivel de PYME, un estudio de EMC revela que el 58% de las organizaciones colombianas tienen planes actuales para implementar tecnologías de Big Data. El otro 42% asegura que la falta de interés corresponde a que la cultura empresarial aún no está lista (46%); pero también porque es costoso implementarlo con respecto a la situación económica actual (28%), y existe una falta de entendimiento con respecto a esta tendencia (25%) (Gonzalez, 2014).

Es así como, para minimizar la complejidad, los costos y la falta de capacitación de personal, las PYME requieren que un proyecto de analítica de datos pueda ser

desplegado rápidamente y que pueda ser fácil de modelar y replicar a otras áreas de la organización. También demanda que los modelos desarrollados puedan ser fáciles de mejorar ante cualquier cambio externo que afecte a la organización y, además, que sea flexible para la integración de distintas fuentes de datos. En la figura 8 se muestra lo que las PYME requieren al momento de implementar un proyecto de analítica de datos.

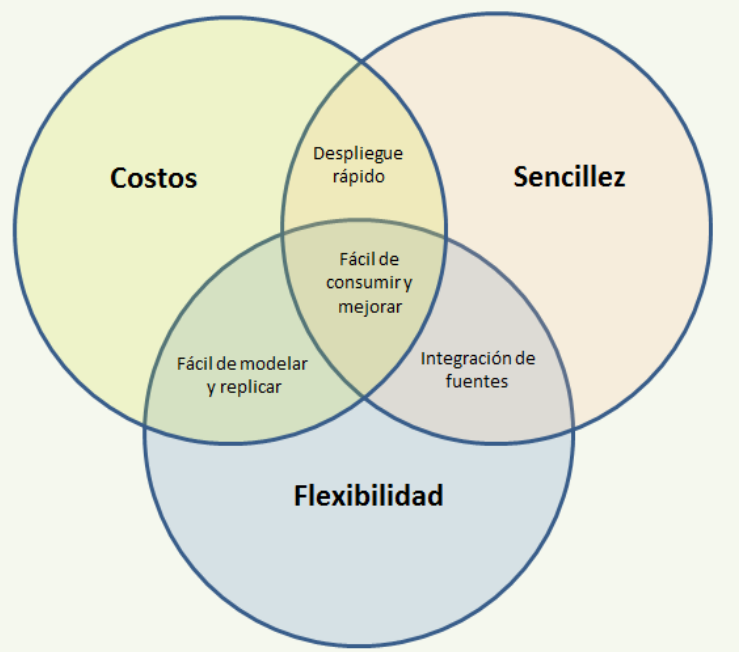


Figura 8. Requerimientos de las PYME para implementar Proyectos de analítica de datos. Elaboración propia.

4. Diseño metodológico

4.1 Tipo de investigación

El presente trabajo se desarrolla bajo un estudio experimental, donde se tiene el control de la variable independiente, es decir, las metodologías de analítica de datos, en la aplicación de un caso de estudio para evaluar el efecto que ésta tiene en la variable dependiente, que es el esfuerzo medido en horas/hombre.

4.2 Diseño de la investigación

De acuerdo a los objetivos específicos de la investigación, se utiliza la metodología de investigación ciencia basada en el diseño (D*esign Science in Information Systems Research*). La ciencia basada en el diseño es un paradigma de resolución de problemas en investigaciones de informática y ciencias de la computación, y tiene como objetivo contribuir en la solución de problemas relevantes al mismo tiempo que hacer aportes significativos a un área del conocimiento, mediante el análisis de problemas aún no resueltos en un ambiente del mundo real y su resolución de una manera novedosa (Hevner, March, Park, y Ram, 2004). En la figura 9 se representa la metodología de investigación.

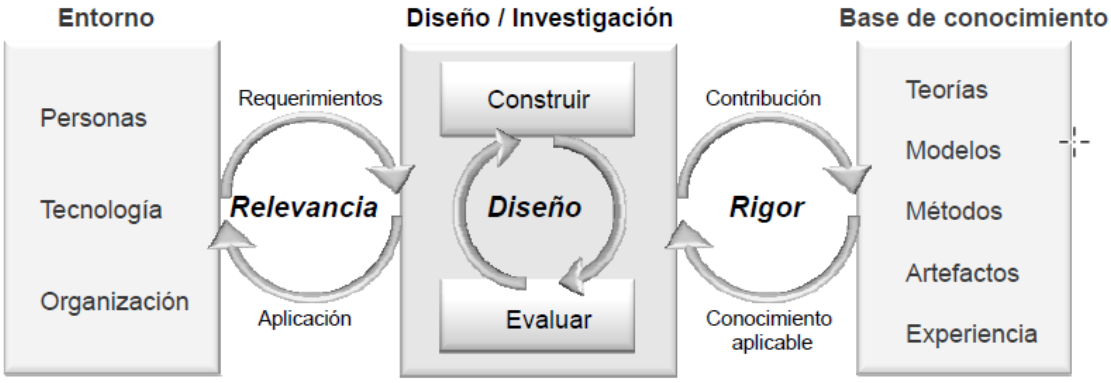


Figura 9. Ciencia basada en el diseño. Gonzalez y Pomares (2012).

En consecuencia, el desarrollo de este estudio pretende solucionar el problema difuso de saber cuál metodología es más adecuada para los proyectos de analítica de datos en las PYMES ESAL como un problema real en el ambiente empresarial. Para ello se establece el desarrollo de un proceso de investigación basado en tres ciclos, a saber; relevancia, establece un problema para aplicar una propuesta de solución; rigor, revisa la literatura científica como soporte teórico para una solución que pueda ser aprovechada; y, por último, diseño, que permite generar un nuevo conocimiento con la propuesta de una metodología que pueda ser evaluada con el problema real. A continuación, se describen las actividades asociadas a los objetivos específicos dentro de cada uno de los ciclos.

4.2.1 Ciclo de análisis del entorno o relevancia.

En este ciclo se realiza una revisión de literatura científica para identificar los requerimientos que tienen las PYME para implementar metodologías de proyectos de analítica de datos, como lo dice el objetivo específico OE1. Las actividades por desarrollar bajo este objetivo son:

- A1. Definición de las fuentes de estudio
- A2. Selección de literatura científica de acuerdos a los criterios de inclusión y exclusión que responda a la pregunta de investigación.
- A3. Selección de las metodologías de proyectos de analítica de datos más usadas.
- A5. Identificación de requerimientos de los proyectos de analítica de datos en PYME a través de la búsqueda bibliográficas.

4.2.2 Ciclo de análisis de la base de conocimiento o rigor

Este ciclo se orienta a definir los elementos de las metodologías de analítica de datos que deberían adaptarse para la implementación de proyectos de analítica de datos en las PYME, de acuerdo al objetivo específico OE2, cuyas actividades se presentan a continuación:

- Selección del modelo de referencia para la propuesta metodológica.
- Definición de los elementos de mayor relevancia en los proyectos de analítica de datos en las PYME.

4.2.3 Ciclo del diseño y evaluación del modelo.

En este ciclo se diseñará una metodología a partir de los requerimientos identificados y los elementos definidos para realizar proyectos de analítica de datos en las PYME según el objetivo específico OE3, y se hará la evaluación a través de la aplicación de un caso de estudio en un Proyecto de analítica de datos en una PYME de la ciudad de Medellín como se presenta en el objetivo específico OE4. Las actividades de esta fase son:

- Diseño: Documentación de cada una de las actividades de la propuesta metodológica de proyectos de analítica de datos para PYME ESAL.
- Comparación: Evaluación comparativa de las características de la propuesta metodológica con respecto al modelo de referencia.
- Evaluación: Aplicación de la propuesta metodológica de analítica de datos en una PYME ESAL comparando el esfuerzo con respecto a otra metodología de analítica de datos.

PARTE III

EXPLORACIÓN

5. Revisión de literatura

Para alcanzar el objetivo específico OE2 de esta investigación, se realiza una revisión de literatura con el fin de definir los elementos que deberían adaptarse en los proyectos de analítica de las PYME. La revisión de literatura que se llevó a cabo se presenta a continuación.

5.1 Selección de fuentes

Las bases de datos seleccionadas para la realización de este estudio fueron:

- IEEE
- Science Direct y
- Google Scholar.

5.2 Cadenas de búsqueda.

Con las palabras claves CRISP-DM, *Analytics, Small and Medium-sized Enterprise* o SME, se formaron las siguientes cadenas de búsqueda:

- **Cadena A:** CRISP-DM, *Small and Medium-sized Enterprise*
- **Cadena B:** *Analytics, Small and Medium-sized Enterprise*

En la tabla 1 se muestra los criterios de búsquedas en las bases de datos definidas previamente.

Tabla 1. Criterios de búsqueda en las bases de dato.

Base de Datos	A	B
IEEE	Búsqueda en metadatos y el texto completo 2009-presente.	Búsqueda metadatos y el texto completo 2009-presente
ScienceDirect	Búsqueda en texto complete 2009-presente.	Búsqueda en texto complete 2009-presente.
GoogleScholar	2009-presente.	Se puso cada cadena en "" 2009-presente.

Los resultados de la búsqueda en cada base de datos de acuerdo a las cadenas de búsqueda se muestran en la tabla 2.

Tabla 2. Resultados de la búsqueda en las bases de datos.

Base de Datos	**A**	**B**
IEEE	5	9
ScienceDirect	91	29
GoogleScholar	162	195

5.3 Selección de estudios

5.3.1 Criterios de inclusión y exclusión.

Para realizar las búsquedas en las diferentes bases de datos, se tuvieron encuentra los estudios desde el 2009 en adelante, y aquellos trabajos que evidenciaran en sus títulos una propuesta de analítica de datos para las PYME o SME.

5.3.2 Estudios seleccionados.

De dicha búsqueda se seleccionaron aquellos artículos que se acercaron a la pregunta de investigación, los cuales se listan en la tabla 3.

Tabla 3. Selección de estudios.

#	Título	Autores	Año
1	Intelligent Data Analysis - Support for Development of SMEs Sector	Olivera et al	2009
2	CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data	Bošnjak, Grljević, & Bošnjak	2009
3	Feasibility and effort estimation models for medium and small size information mining projects	Pytel et al.	2015
4	Implementing Big Data analytics for small and medium enterprise (SME) regional growth	Ogbuokiri et al.	2015
5	SMART: An Application Framework for Real Time Big Data Analysis on Heterogeneous Cloud Environments	Dos Anjos et al.	2015
6	Leveraging big data technology for small and medium-sized enterprises (SMEs)	Kalan & Ünalir.	2016
7	Big data driven customer insights for SMEs in redistributed manufacturing	Soroka, Liu, Han, & Salman	2017
8	A Data Analytics Framework for Business in Small and Medium-Sized Organizations	Dittert, Härting, Reichstein, & Bayer	2018

5.4 Resultados de los estudios

Para analizar los estudios seleccionados, se plantearon 7 criterios de análisis que se presentan en la tabla 4. A pesar de que los artículos tratan el tema de analítica de datos en las PYME, los artículos 1, 3, 6 y 7 no usan un modelo de analítica de datos de referencia, ni proponen una nueva metodología ni tampoco mencionan como disminuir el esfuerzo en los proyectos de las PYME. En cuanto al artículo 2, que usa CRISP-DM como modelo de referencia, herramientas y modelos de analítica de datos, le falta considerar la disminución del esfuerzo. Por otro lado, el artículo 4 que se centra en aprovechar el Big Data para el crecimiento de las PYME con la propuesta de una serie de herramientas como IBMs Watson Analytics y Google Analytics, entre otros, para reducir el esfuerzo, se olvida de usar una metodología para alcanzar dicho fin. Lo mismo ocurre con el artículo 5 que trata de disminuir el esfuerzo en un ambiente de nube heterogéneo, pero no desarrolla el uso de una metodología de analítica de datos para alcanzar los resultados que propone.

Finalmente, el artículo 8 presenta una propuesta para disminuir el esfuerzo en los proyectos de analítica de datos en las PYMES. El artículo fue publicado en el 2018 y es el que más se acerca a la pregunta de investigación. Este plantea una metodología basada en CRISP-DM como modelo de referencia, sin embargo, deja por fuera aspectos importantes como el entendimiento del negocio y el despliegue.

Tabla 4. Criterios de análisis

Criterio de análisis	Art #1	Art #2	Art #3	Art #4	Art #5	Art #6	Art #7	Art #8
Usa CRISP-DM como modelo de referencia?	No	Sí	No	No	No	No	No	Sí
Propone una nueva metodologia DA para PYME?	No	No	No	No	No	No	No	Sí
Disminuye el esfuerzo de la analitica en PYME?	No	No	No	Si	Si	No	No	Sí
Aplica caso de estudio?	No	Sí	No	No	Sí	No	No	Sí
Cuál es el marco de aplicación?		PYME		PYME	PYME	PYME		PYME
Qué Herramientas usa para DA?		MS Access - DataEngine		IBM's Watson Analytics	MApReduce			RapidMiner
Qué Modelos usa para DA?		Neural networks and fuzzy						ANN Red reuronal Artificial

Del análisis de los estudios, se encontró que son pocas las publicaciones que proponen una metodología de analítica de datos que busque la disminución del esfuerzo en los proyectos de las PYME basándose en una metodología ya existente como CRISP-DM o en una nueva propuesta. A falta literatura sobre cuál modelo de referencia seguir para crear una metodología ajustada para las PYME, se tiene que realizar una comparación con las metodologías más referenciadas y que fueron descritas en el marco teórico, que permita adaptarse para las PYME buscando dar solución a la problemática del tema de investigación.

5.5 Comparacion de KDD, SEMMA y CRISP-DM

Según Azevedo y Santos (2008), al hacer una comparación de las etapas KDD y SEMMA, afirma que son equivalentes, como se muestra en la tabla 5. Aquí se plantea que las cinco etapas del proceso SEMMA se pueden ver como una implementación práctica de las cinco

etapas del proceso KDD, ya que están directamente relacionadas con el software SAS Enterprise Miner.

Por otro lado, CRISP-DM no es tan fácil de relacionar con las otras dos metodologías. Sin embargo, en el mismo estudio se hace una comparación con KDD, donde la fase de comprensión del negocio es comparada con la fase Pre-DDD la cual estudia la comprensión del dominio de la aplicación, el conocimiento previo relevante y los objetivos del usuario final. La fase de comprensión de datos se identifica con la combinación de selección y pre-procesamiento. La fase de preparación de datos se identifica con transformación. La fase de modelado se puede identificar con minería de datos. La fase de evaluación se puede identificar con Interpretación/Evaluación. Por último, la fase de despliegue se identifica con la fase Post KDD o consolidación del conocimiento.

Tabla 5. Resumen de las correspondencias entre KDD, SEMMA y CRISP-DM

KDD	SEMMA	CRISP-DM
Pre-KDD	---------	Comprensión del Negocio
Selección	Ejemplificar	Comprensión de los datos
Pre procesamiento	Explorar	
Transformación	Modificar	Preparación de los Datos
Minería de datos	Modelar	Modelado
Interpretación/Evaluación	Valoración	Evaluación
Post KDD	---------	Despliegue

Nota: Tomado de Azevedo y Santos (2008).

5.6 Modelo de referencia para proyectos de analítica

Según Achmad, Sabur, Pritasari y Reinaldo (2016), CRISP-DM es considerado el modelo estándar para desarrollar proyectos de minería de datos y descubrimiento de

conocimiento. Y para Dittert, Härting, Reichstein y Bayer (2018) es un estándar completo pero flexible y se puede adaptar fácilmente a cada tarea de analítica en términos de procesos. Además, CRISP-DM profundiza en mayor detalle sobre las tareas y actividades a ejecutar en cada fase de minería de datos, mientras KDD y SEMMA proveen sólo una guía general del trabajo a realizar. Esta última se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando y fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS (Moine, Haedo, y Gordillo, 2011).

En la figura 10 se presenta cómo CRISP-DM es la metodología más importante para analítica de datos, existiendo otras alrededor menos comunes, e incluso de dominios específicos y propias en cada organización; otras en cambio sólo toman algunos elementos de CRISP-DM para la ejecución de los proyectos, pero sin enfocarse en el proceso y el esfuerzo. También se muestra que la metodología propuesta se basará en dicha metodología, pero con un enfoque más pequeño para disminuir el esfuerzo en el proceso de implementación de un proyecto de analítica de datos en una pequeña o mediana organización.

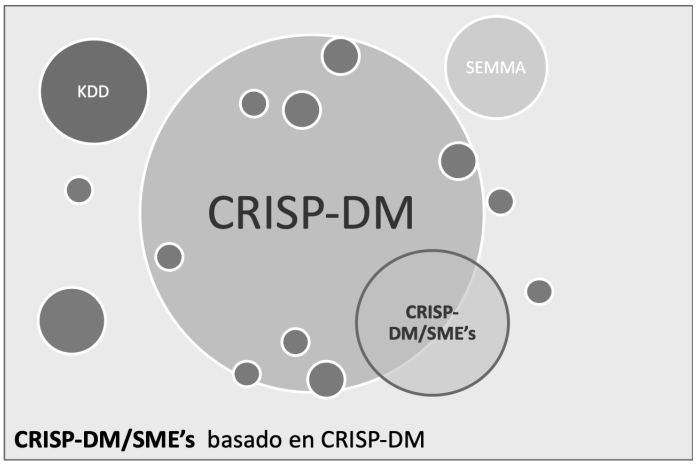


Figura 10. CRISP-DM/SME's basado en CRISP-DM. Elaboración propia.

Para la propuesta metodológica se tomará CRISP-DM como modelo de referencia para la metodología de proyectos de analítica de datos en PYME ESAL, justificado en los siguientes aspectos:

a. Es el modelo más referenciado por su amplia aceptación
b. Todas sus etapas están debidamente organizadas, estructuradas y definidas.
c. Facilita la comprensión y revisión de un proyecto.

Al modelo de referencia se le excluirán algunas tareas y se propondrán otras más resumidas o adaptadas para que sea sencillo de implementar, mejorar y replicar; flexible de adaptar y, que pueda ser costeado por la organización. Todo ello con el fin de disminuir el esfuerzo de personal, tiempo y los costos en la ejecución de los proyectos de analítica de datos.

5.7 Elementos de la metodología CRISP-DM adaptados.

En la tabla 6 se presenta la metodología CRISP-DM con sus respectivas fases y actividades y al frente en una segunda columna, las fases que han sido adaptadas para la propuesta metodológica para PYME. Se han seleccionado las actividades (aparecen en cursiva debajo de cada fase) más relevantes en los proyectos de analítica de datos en PYME con el fin de disminuir el esfuerzo en la organización y se han agrupado en fases guardando la relación con CRISP-DM. Cada fase de esta nueva propuesta da como resultado un artefacto o producto de trabajo que puede ser usado en la siguiente fase.

Tabla 6. Comparación CRISP-DM con CRISP-DM/SME's.

CRISP-DM	CRISP-DM/SME	Artefacto
Fase 1: Comprensión del negocio	**Fase 1: Definición del proyecto DA**	**Contexto del proyecto**
Determinar objetivos del negocio	*Seleccionar objetivos empresariales*	
Evaluar la situación	*Definir objetivos del proyecto*	
Determinar los objetivos DM	*Asignar recursos*	
Realizar el plan del proyecto	*Determinar alcance y riesgos*	
Fase 2: Comprensión de los datos	**Fase 2: Gestión de datos**	**Datos formateados**
Recolectar datos iniciales	*Recolectar datos*	
Describir los datos	*Explorar datos*	
Explorar los datos	*Integrar los datos*	
Verificar la calidad de los datos	*Formatear los datos*	
Fase 3: Preparación de los datos		
Seleccionar los datos		
Limpiar los datos		
Construir los datos		
Integrar los datos		
Formatear los datos		
Fase 4: Modelado	**Fase 3: Modelado**	**Tablero de control**
Escoger la técnica de modelado	*Seleccionar modelo*	
Generar el plan de pruebas	*Seleccionar herramienta de visualización*	
Construir el modelo	*Construir tablero de control*	
Evaluar el modelo		
Fase 5: Evaluación	**Fase 4: Evaluación**	**Análisis de resultados**
Evaluar los resultados	*Evaluar tablero de control*	
Revisar el proceso	*Analizar resultados*	
Determinar los próximos pasos		
Fase 6: Despliegue	**Fase 5: Despliegue**	**Informe final**
Planear el despliegue	*Automatizar proceso*	
Planear la monitorización y mantenimiento	*Distribuir resultados*	
Producir el informe final		
Revisar el proyecto		

PARTE IV
PROPUESTA

6. Diseño de la propuesta metodológica CRISP-DM-SMEs

Con los elementos de CRISP-DM definidos en la revisión de literatura y según la identificación de los requerimientos que necesitan las PYME para la implementación de proyectos de analítica de datos, se presenta la propuesta metodológica a continuación.

6.1 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.

6.2 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 (Menendez y Castellanos, 2008). La propuesta metodológica usa los elementos representados en la figura 11.

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

Figura 11. Elementos usados de SPEM. Menendez y Castellanos (2008)

6.2.1 Roles.

La metodología CRISP-DM/SMEs define los roles desde la perspectiva de la industria de ciencia datos, como lo sugiere (DataCamp, 2015), Los roles seleccionados para esta metodología se describen a continuación:

6.2.1.1 Ingeniero de datos o Data Engineer.

Es el principal rol dentro de los proyectos de analítica de datos 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.

6.2.1.2 *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 puede asumir el director de área o cargos similares. 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.

6.2.1.3 *Administrador de Datos o Data Manager.*

Es el rol encargado de liderar el equipo de analítica de datos 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 puede ser asumido por el gerente de la PYME, así que las competencias tecnológicas son mínimas para este caso.

6.2.2 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*).

6.2.3 Actividades.

Las actividades son un conjunto de tareas del Proyecto de analítica de datos 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.

6.2.4 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 analítica de datos 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.

6.2.5 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.

6.2.6 Herramientas.

Las herramientas son aplicaciones de software que ayudan a desarrollar las actividades de las fases. Una ayuda a la gestión del Proyecto de analítica de datos, otra permite la gestión o formateo de los datos y finalmente la que ayuda a construir el tablero

de control. Las herramientas son escogidas por el Ingeniero de datos, según los recursos asignados para el Proyecto de analítica de datos.

6.2.7 Representación gráfica.

La aplicación de la propuesta metodológica se representa en el diagrama SPEM, en la figura 12. La secuencia de las fases no es estricta, sino que se puede interactuar entre cada una de las fases; en la ejecución del Proyecto de analítica de datos 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.

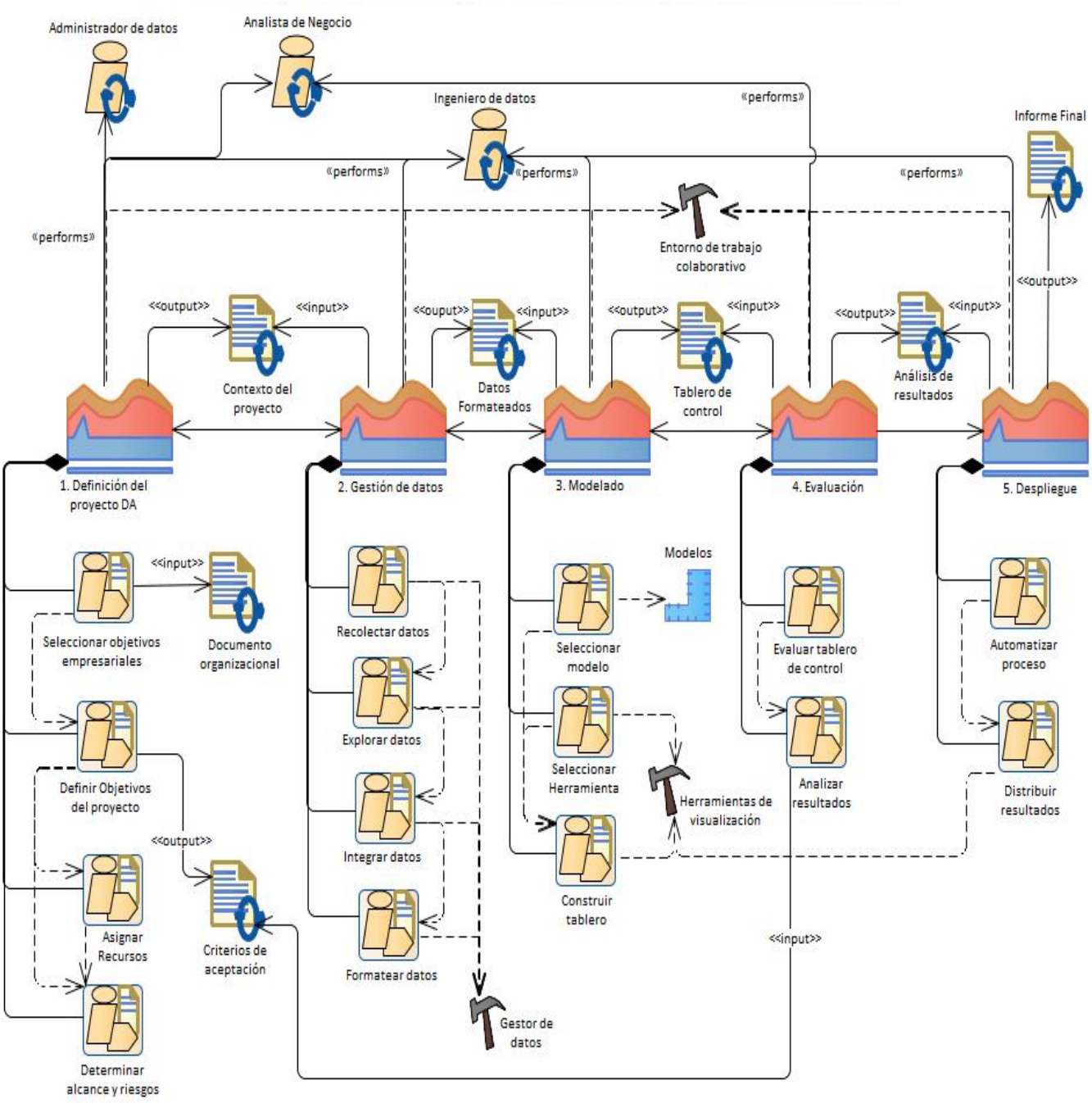


Figura 12. Representación gráfica de CRISP'DM-SMEs. Elaboración propia

6.3 Fase 1: Definición del Proyecto de Analítica

Para definir el Proyecto DA, es necesario identificar las personas que participarán activamente, 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 DA 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.

6.3.1 Herramienta de entorno de trabajo colaborativo.

La herramienta de entorno colaborativo se escoge 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.

6.3.2 Actividad 1: Seleccionar objetivos empresariales.

6.3.2.1 Descripción de la actividad.

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

6.3.2.2 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.

6.3.2.3 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.

6.3.2.4 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 organizacionales 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.

6.3.2.5 Salidas.

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

6.3.3 Actividad 2: Definir objetivos del proyecto.

6.3.3.1 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.

6.3.3.2 Entradas.

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

6.3.3.3 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.

6.3.3.4 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.

6.3.3.5 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.

6.3.4 Criterios de aceptación.

6.3.4.1 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.

6.3.4.2 Entradas.

Objetivos del Proyecto DA.

6.3.4.3 Tareas.

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

6.3.4.4 Buenas prácticas.

Los criterios de aceptación deben tener contexto, evento, y resultado. En la tabla 7 se presenta una guía con los elementos básicos para realizar el producto de trabajo ‘criterios de aceptación’.

6.3.4.5 Salidas.

Criterios de aceptación para futura evaluación.

Tabla 7. Criterios de Aceptación

		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				

Nota: Tomado de http://www.pmoinformatica.com/2012/10/plantillas-scrum-historias-de-usuario.html.

6.3.5 Actividad 3: Asignar recursos.

6.3.5.1 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 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

6.3.5.2 Entradas.

Participación de roles en las actividades. En la tabla 8 se presenta las actividades en la primera columna y marcadas con una X, en la columna del rol, los roles que participan en la ejecución de ésta.

Tabla 8. 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	X
2	*Definir objetivos del proyecto*	X	X	X
3	*Asignar recursos*	X	X	X
4	*Determinar alcance y riesgos*	X	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

6.3.5.3 Tareas.

- Estimar el total de horas por cada actividad presentada en la tabla 8 y totalizar por fase y por rol para completar la tabla 9.
- 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 9 con la información correspondiente.

Tabla 9. 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							

6.3.5.4 Buenas prácticas.

- Considere 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.
- Establezca 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 físico o tecnológico, establezca 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.

6.3.5.5 Salidas.

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

6.3.6 Actividad 4: Determinar el alcance y riesgos.

6.3.6.1 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.

6.3.6.2 Entradas.

Objetivos del Proyecto DA.

6.3.6.3 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 10.

Tabla 10. Riesgos y contingencias

Riesgo	Nivel			Contingencia
	Bajo	Medio	Alto	

6.3.6.4 Buenas prácticas.

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

6.3.6.5 Salidas.

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

6.3.7 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?

6.3.8 Contexto del Proyecto.

A continuación, realice el contexto del Proyecto como un producto de trabajo y entregable para cada uno de los miembros del equipo. El documento debe contener las siguientes partes:

6.3.8.1 Introducción.

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

6.3.8.2 Descripción del Problema.

Describe detalladamente el problema o necesidad de la PYME.

6.3.8.3 Responsables del Proyecto.

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

6.3.8.4 Objetivos empresariales.

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

6.3.8.5 Objetivos del Proyecto DA.

Lista de los objetivos para el Proyecto DA que fueron definidos en la actividad 2 ‘Definir objetivos del proyecto’.

6.3.8.6 Criterios de Aceptación.

Criterios de aceptación que fueron definidos en la actividad 2.

6.3.8.7 Asignación De Recursos.

Recursos asignados en la actividad 3.

6.3.8.8 Determinación del alcance y los riesgos.

Alcance y riesgos identificados en la actividad 4.

6.3.8.9 Costos.

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

6.3.8.10 Planeación.

Cronograma con las actividades a realizar de forma organizada y cronológica.

6.3.8.11 Estrategias de recolección de información.

Descripción de métodos, herramientas y personal utilizados para reunir la información necesaria.

6.3.8.12 Plan de socialización.

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

6.4 Fase 2: Gestión de datos

La gestión de datos es la fase que más tiempo requiere ya que se deben 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".

6.4.1 Herramienta para gestión de datos.

En cada una de las actividades de esta fase se 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"

6.4.2 Actividad 5: Recolectar datos.

6.4.2.1 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.

6.4.2.2 Entradas.

- Contexto del Proyecto DA.

6.4.2.3 Tareas.

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

6.4.2.4 Buenas prácticas.

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

6.4.2.5 Salidas.

Base de datos para el Proyecto DA.

6.4.3 Actividad 6: Explorar datos.

6.4.3.1 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.

6.4.3.2 Entradas.

Base de datos para el Proyecto DA.

6.4.3.3 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.

6.4.3.4 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 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.

6.4.3.5 Salidas.

Base de datos limpia.

6.4.4 Actividad 7: Integrar datos.

6.4.4.1 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.

6.4.4.2 Entradas.

- Fuentes de datos limpios
- Fuentes de datos externos

6.4.4.3 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.

6.4.4.4 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.

6.4.4.5 Salidas.

Fuente de datos integrados.

6.4.5 Actividad 8: Formatear datos.

6.4.5.1 Descripción de la actividad.

Esta es la última actividad antes de entrar a la fase del modelado, se necesita conocer los tipos de datos que requieren 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.

6.4.5.2 Entradas.

Fuente de datos integrados.

6.4.5.3 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 briden 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.

6.4.5.4 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.
- Verifique con el usuario la completitud del conjunto de datos final.

6.4.5.5 Salidas.

Datos formateados como un producto de trabajo.

6.4.6 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 ha tomado para proteger los datos donde están almacenados?

6.4.7 Datos Formateados.

Los datos formateados son el subconjunto de datos, 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.

6.5 Fase 3: Modelado

En la fase de Modelado se representan gráficamente los datos a través de elementos visuales como diagramas, gráficos de barras, mapas y tablas para comprender tendencias, patrones, indicadores, entre otros. En el modelado se desarrollan 3 actividades que son "Seleccionar el modelo", "Seleccionar la herramienta de visualización" y "Construcción del tablero de control". Para llevarlas a cabo se requiere definir el modelo de representación gráfica y la herramienta de visualización que puede crear dicho modelo.

6.5.1 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.

6.5.2 Herramienta de visualización.

La herramienta de visualización es uno de los recursos de software más importantes en el desarrollo del Proyecto DA ya que ésta 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 práctica 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.

6.5.3 Actividad 9: Seleccionar modelo.

6.5.3.1 Descripción de la actividad.

La selección del modelo está 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.

6.5.3.2 Entradas.

Producto de trabajo “Datos Formateados”.

6.5.3.3 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.

6.5.3.4 Buenas prácticas.

- Seleccione los 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?

6.5.3.5 Salidas.

Modelos de representación visual para cada objetivo.

6.5.4 Actividad 10: Seleccionar herramienta.

6.5.4.1 Descripción de la actividad.

La selección de la herramienta de visualización es 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.

6.5.4.2 Entradas.

Modelos de representación visual para cada objetivo.

6.5.4.3 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 configurarla en el computador donde se realizará la fase de modelado.

6.5.4.4 Buenas prácticas.

- Investigue las herramientas de visualización líderes del mercado. Puede guiarse por el cuadrante mágico de Gartner.
- 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.

6.5.4.5 Salidas.

Herramienta de visualización.

6.5.5 Actividad 11: Construir tablero de Control.

6.5.5.1 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 grande, con el fin de validar resultados rápidamente.

6.5.5.2 Entradas.

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

6.5.5.3 Tareas.

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

6.5.5.4 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.

6.5.5.5 Salidas.

Tablero de control como un producto de trabajo.

6.5.6 Interrogantes de analítica.

- ¿El tablero presenta problemas de calidad de datos?
- ¿Los modelos construidos son los más efectivo para representar la información?

- ¿El tablero de control cumple con los criterios de aceptación?

6.5.7 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 13.

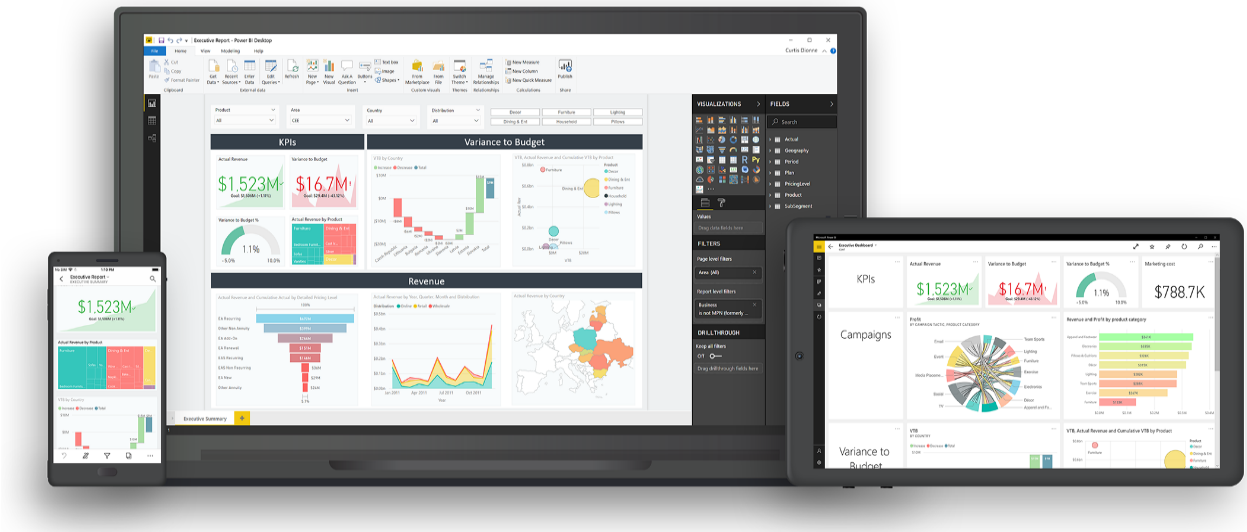


Figura 13. Tablero Power BI Responsivo. Tomado de https://powerbi.microsoft.com/en-us/

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.

6.6 Fase 4: Evaluación

El tablero de control debe ser presentado a los interesados del proyecto para 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 deben determinar cuáles son los criterios que no se alcanzaron y volver a una actividad anterior para corregir los problemas detectados.

6.6.1 Actividad 12: Evaluar tablero de Control.

6.6.1.1 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.

6.6.1.2 Entradas.

Tablero de control.

6.6.1.3 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.

6.6.1.4 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.

6.6.1.5 Salidas.

Tablero evaluado

6.6.2 Actividad 13: Analizar resultados.

6.6.2.1 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.

6.6.2.2 Entradas.

- Criterios de aceptación.
- Tablero de control evaluado

6.6.2.3 Tareas.

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

6.6.2.4 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.

6.6.2.5 Salidas.

- Análisis de resultados.

6.6.3 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?

6.6.4 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.

6.7 Fase 5: Despliegue

La fase de despliegue consiste en automatizar la fuente de los datos para que sean 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 la culminación del proyecto.

6.7.1 Actividad 14: Automatizar proceso

6.7.1.1 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.

6.7.1.2 Entradas

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

6.7.1.3 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 cuáles 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.

6.7.1.4 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.

6.7.1.5 Salidas

Tablero de control automatizado.

6.7.2 Actividad 15: Distribuir resultados

6.7.2.1 Descripción de la actividad

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

6.7.2.2 Entradas

- Herramienta de visualización.
- Accesos de usuarios.

6.7.2.3 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.

6.7.2.4 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.

6.7.2.5 Salidas

Tablero de control distribuido.

6.7.3 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?

6.7.4 Informe Final

El informe final es un documento escrito como producto de trabajo con el detalle de la ejecución del Proyecto DA y la experiencia adquirida por el 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 (IBM Corp, 2012):

- 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.
- Experiencia adquirida por el equipo.
- Recomendaciones
- Firma de aceptación final por parte de los stakeholders.

6.7.5 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

PARTE V
EVALUACIÓN

7. Evaluación comparativa de la propuesta metodológica

Para evaluar la propuesta como una metodología de minería de datos, se emplea el marco comparativo propuesto por Moina y Haedo (2015), que incluye 52 características divididas en 4 aspectos a evaluar, como se muestra en la figura 14.

Figura 14. Aspectos del Marco Comparativo. Moina y Haedo (2015)

Cada uno de estos aspectos con sus respectivas características, deben estar presentes en una metodología de minería de datos bien definida. Por lo tanto, la propuesta CRISP-DM/SME's se comparará con CRISP-DM bajo este marco comparativo con el fin de evaluar la propuesta si cumple con las condiciones para ser una metodología propiamente dicha.

7.1 Evaluación de los aspectos del marco comparativo

7.1.1 Aspecto 1: Descripción de las actividades de cada fase.

El primer aspecto evaluado fue el nivel de detalle en la descripción de las actividades de cada fase. La propuesta arrojó un resultado del 100% de valoración positiva en las características, frente a un 80% con CRISP-DM como se evidencia en la tabla 11.

Tabla 11. Evaluación del nivel de detalle en las actividades que componen cada fase

#	Característica	CRISP-DM	CRISP-DM/ SMEs
1.1	¿Se definen actividades específicas para cada fase del proceso?	SI	SI
1.2	¿Se explicitan los pasos a seguir para llevar a cabo cada actividad?	SI	SI
1.3	¿Se definen las entradas de cada actividad?	NO	SI
1.4	¿Se definen las salidas de cada actividad?	SI	SI
1.5	¿Se provee una guía de buenas prácticas para cada una de las actividades específicas?	SI	SI
	Valoraciones positivas	4/5 = 80%	5/5 = 100%

7.1.2 Aspecto 2: Escenarios de aplicación.

Debido a que la propuesta no contempla actividades para identificar los problemas de la PYME, sino que su punto de partida es un problema ya definido por la organización y tampoco hay partidas alternativas, la valoración positiva es de un 50% frente a un 75% de CRISP-DM, como se muestra en la tabla 12.

Tabla 12. Evaluación de los escenarios de aplicación

#	Característica	CRISP-DM	CRISP-DM/SMEs
2.1	¿Se especifican actividades para la definición y el análisis del problema u oportunidad con el cual colaborará la minería de datos?	SI	NO
2.2	¿Se consideran puntos de partida alternativos donde el usuario no refiere un problema, sino que sólo desea explorar sus datos?	NO	NO
2.3	¿La metodología es independiente del dominio de aplicación?	SI	SI
2.4	¿La metodología es aplicable a proyectos de diferente tamaño?	SI	SI
	Valoraciones positivas	3/4 =75%	2/4=50%

7.1.3 Aspecto 3: Actividades específicas que componen cada fase.

La evaluación detallada de las características de este aspecto están en el anexo 2, y el resumen se presenta en la tabla 13, donde se muestra una valoración positiva de la propuesta del 69% frente a un 77% de CRISP-DM. La fase de evaluación fue la que influyó en este resultado debido a que no se cumplen las características de ponderar modelos, y tener una vía alternativa en caso de que el modelo no resulte viable. Cabe aclarar que la fase del modelado en ambos casos es distinta, ya que los modelos planteados por CRISP-DM contemplan todas las técnicas de modelado mientras que la propuesta se enfoca en la técnica de visualización de los datos.

Tabla 13. Evaluación general de las actividades específicas

Fase	CRISP-DM	CRISP-DM/SMEs
Análisis del problema	6/9 = 66%	5 /9 = 55%
Selección y preparación de los datos	3/5 = 60%	4 /5 = 75%
Modelado	3/4 = 75%	3 /4 = 75%
Evaluación	4/4 = 100%	2 /4 = 50%
Implementación	4/4 = 100%	4/4= 100%
Total	20/26 = 77%	18/26 = 69%

7.1.4 Aspecto 4: Actividades destinadas a la dirección del proyecto.

El aspecto 4 contiene las características a evaluar para las actividades destinadas a la dirección del proyecto, como lo son la gestión del alcance, del tiempo, del costo, del equipo de trabajo y del riesgo. La evaluación de las características de estas áreas se evidencia en el anexo 2.

La evaluación general del aspecto 4 presentado en la tabla 14 muestra que la propuesta alcanza un 82% de valoración positiva, esto debido a la herramienta de colaboración que ayuda a la gestión del Proyecto DA, en comparación con CRISP-DM que no propone cómo gestionar los recursos y el equipo de trabajo, por lo cual, valoración positiva disminuye a un 47%.

Tabla 14. Evaluación general para las actividades de dirección del proyecto

Área	CRISP-DM	CRISP-DM/SMEs
Gestión del alcance	1/2 = 50%	2 /2 = 100%
Gestión del tiempo	3/4 = 75%	4 /4 = 100%
Gestión del costo	1/4 = 25%	4 /4 = 100%
Gestión del equipo de trabajo	1/3 = 33%	2 /3 = 66%
Gestión del riesgo	2/4 = 100%	2/4 = 50%
Total	8/17 = 47%	14/17 = 82%

7.2 Evaluación final del marco comparativo

Después de la evaluación de las 52 características envueltas en los 4 aspectos del marco comparativo, y resumidas en la tabla 15, se concluye que la propuesta metodológica CRISP-DM-SMEs tiene una calificación positiva mayor que el modelo de referencia CRISP-DM. La propuesta obtuvo una calificación del 75% frente a su referente de comparación que obtuvo un 67% de valoración. La mayor diferencia se dio en las actividades de dirección de proyecto, ya que la propuesta incluye el uso de herramientas de colaboración para la gestión de Proyecto.

Tabla 15. Evaluación final de todos los aspectos del marco comparativo

Aspecto	CRISP-DM	CRISP-DM/SMEs
Nivel de detalle en la descripción de las actividades	4/5 = 80%	5 /5 = 100%
Escenarios de aplicación	3/4 = 75%	2 /4 = 50%
Actividades específicas de cada fase	20/26 = 77%	18 /26 = 69%
Actividades de dirección del proyecto	8/17 = 47%	14 /17 = 82%
Total	35/52 = 67%	39/52 = 75%

A pesar de que el marco comparativo no tiene una característica que evalúen el uso de tableros de control, la propuesta metodológica logró obtener una calificación positiva alta.

Finalmente, se concluye que la propuesta metodológica CRISP-DM/SMEs cumple con todos los requisitos de una metodología de analítica de datos superando a su modelo de referencia.

8. Caso de estudio

8.1 Definición del Caso de Estudio

Para dar cumplimiento al OE4, de la investigación, se evalúa la propuesta metodológica con un caso de estudio en un Proyecto DA en una PYME ESAL de la ciudad de Medellín.

8.1.1 Selección de la PYME.

La Cooperativa de Ahorro y Crédito Unión Colombiana (COMUNION) es una entidad que forma parte del sector solidario en Colombia y está ubicada en Medellín. Es Vigilada y regulada por la Superintendencia de la Economía Solidaria y por el Fondo de Garantías de Entidades Cooperativas (FOGACOOP). El propósito de Comunión es brindar mediante el servicio de ahorro y crédito bienestar y cobertura a las necesidades de la Iglesia Adventista en Colombia y a los de sus empleados y familiares. Inicia operaciones desde el año 2006 y cuenta con 12 empleados, 1.128 asociados ubicados en todo el país, y unos activos de $ 30.000’ 000.000. Para mayor información sobre su estructura organización ver el anexo 3.

8.1.2 Descripción del Proyecto DA.

COMUNION necesita desarrollar un Proyecto DA para la construcción de un tablero de control o dashboard el cual pueda ser consultado en línea por el Director Financiero, el Director de Riesgo y el Gerente. Además, que sea fácilmente compartido con todos los miembros del Comité de Riesgos, una vez al mes. El tablero de control debe contener el análisis financiero, de riesgo, de proyección social, y de la competencia, resumidos así:

8.1.2.1 Indicadores de cobertura GAP

- 6 indicadores de concentración
- 4 indicadores de liquidez
- 3 indicadores de exceso de liquidez
- Nivel de riesgo total

En el anexo 4 se presentan los indicadores detalladamente.

8.1.2.2 Indicadores de captación.

- Ahorros
- Ahorro contractual
- CDAT
- Intereses de ahorros

8.1.2.3 Indicadores de Colocación.

- Crédito libre inversión
- Crédito hipotecario
- Crédito vehículo
- Crédito educativo
- Otros

8.1.2.4 Análisis de la competencia.

- Tasas de interés en CDT

- Máxima tasa de interés en CDAT
- Tasa de interés de CDAT

8.1.2.5 Impacto social.

- Total inversión por año
- Cantidad de personas beneficiadas

8.1.2.6 Asociados.

- Total Asociados
- Asociados por ciudad
- Crecimiento de asociados

8.1.3 Esfuerzo actual de la PYME.

El Director Financiero de la Cooperativa prepara los indicadores de Riesgo de Liquidez para el Comité de Riesgo que se reúne cada mes. Para preparar dichos indicadores, el Director Financiero exporta la información a un archivo plano .txt del Sistema Contable, para posteriormente digitarlo o copiarlo en una plantilla en Excel. Este proceso tarda aproximadamente 36 horas. Luego, la información es enviada al Asesor Financiero para que termine de realizar los ajustes, quien tarda 3 horas en la terminación de los indicadores.Para generar los indicadores de Riesgos de liquidez se requieren de 39 horas hombre mensual como se muestra en la tabla 16.

Tabla 16. Esfuerzo actual de la PYME

Personal	Tiempo
Director Financiero	36 horas
Asesor Financiero	3 horas
Totales	**39 horas/mes**

8.2 Aplicación de la Propuesta Metodológica

Después de la evaluación de la metodología CRISP-DM/SMEs en el caso de estudio para el desarrollo del Proyecto DA, se identificaron los siguientes elementos relevantes para el estudio de esta investigación.

8.2.1 Roles en el caso de estudio.

Los roles que participaron en la implementación de la metodología de analítica y sus respectivas competencias se presentan en la tabla 17.

Tabla 17. Roles en el caso de estudio.

Rol en ciencia de datos	Rol en la PYME	Descripción	Competencias
Ingeniero de datos	Ingeniero de Sistemas	Es un profesional en Informática, administración e ingeniería. Es una persona externa a la PYME y es quien implementa la metodología de analítica de datos en el proyecto.	Sus competencias son manejo de bases de datos, hojas de cálculo y diseño de tableros de control. Con conocimientos asociados al uso de SQL y herramientas de visualización.
Analista de negocio	Director Financiero	Contador Público el encargado de mejorar los procesos del negocio a través de la analítica, y quien prepara los indicadores de riesgo de liquidez para presentarlo al Comité de Riesgos.	Amplio conocimiento sobre los riesgos financieros. Dentro de las competencias básicas están el manejo herramientas de office y modelado de datos
Administrador de datos	Gerente	Contador Público y especialista en cooperativismos. Apoya el Proyecto DA y suministrar los recursos financieros, humanos y de software para el desarrollo de las actividades.	Sus competencias son el liderazgo, la comunicación interpersonal y análisis de información.

8.2.2 Fuentes de datos

- Base de datos en SQL 2012 Express con más de 300.000 registros contables.
- Base de datos en SQL 2012 Express con información de los asociados.

- Base de datos en Excel con las tasas de interés de todas las entidades financieras de Colombia. Esta información se descarga cada mes de la superfinanciera.
- Base de datos en Excel con información de personas impactadas en los proyectos sociales.

8.2.3 Herramientas en el caso de estudio.

Las herramientas utilizadas en la metodología fueron escogidas por la accesibilidad que tenía la PYME para adquirirlas, y la experticia del Ingeniero de datos. En la tabla 18 se presentan las herramientas utilizadas en el caso de estudio con una breve descripción.

Tabla 18. Herramientas en el caso de estudio.

Tipo de Herramienta	Herramienta	Descripción
Colaboración	Microsoft Teams	Con esta plataforma de Microsoft se gestionaron las actividades del proyecto relacionadas con la planificación y control.
Gestor de datos	SQLServer, Excel.	En SQL se construyó un Datamart, scripts y Jobs para la gestión de los datos de tal forma que quedaran formateados para la herramienta de visualización.
Visualización	Power BI	Se seleccionó la herramienta de Microsoft Power BI para la construcción del tablero de control por su liderazgo en el mercado, su integridad con las demás herramientas, costos y la experiencia del ingeniero de datos.

8.2.4 Producto de trabajo: Tablero de control.

El tablero de control fue evaluado según los criterios de aceptación que fueron definidos en el proyecto y cumplió satisfactoriamente con los objetivos del Proyecto DA. En la primera evaluación del tablero se realizaron algunos ajustes a la presentación del modelo según las observaciones del analista de negocio, hasta que finamente se aceptaron los modelos de visualización, como se evidencia en la figura 15.

El tablero de control accede a los datos automáticamente del Datamart construido en SQL, el cual se actualiza diariamente a través de Jobs que ejecutan los scripts para la limpieza de los datos. Otras fuentes para los indicadores externos se almacenan en OneDrive y son actualizados por el usuario periódicamente, cuando los datos de la competencia cambian. Los datos de la información social son también almacenados en OneDrive para ser accedidos por Power BI para la creación del tablero de control.

8.2.5 Esfuerzo según la propuesta metodológica.

Cada una de las actividades planteadas en la propuesta metodológica se llevaron a cabo para la implementación del Proyecto DA en el caso de estudio y los resultados del esfuerzo obtenido se presenta en la tabla 19.

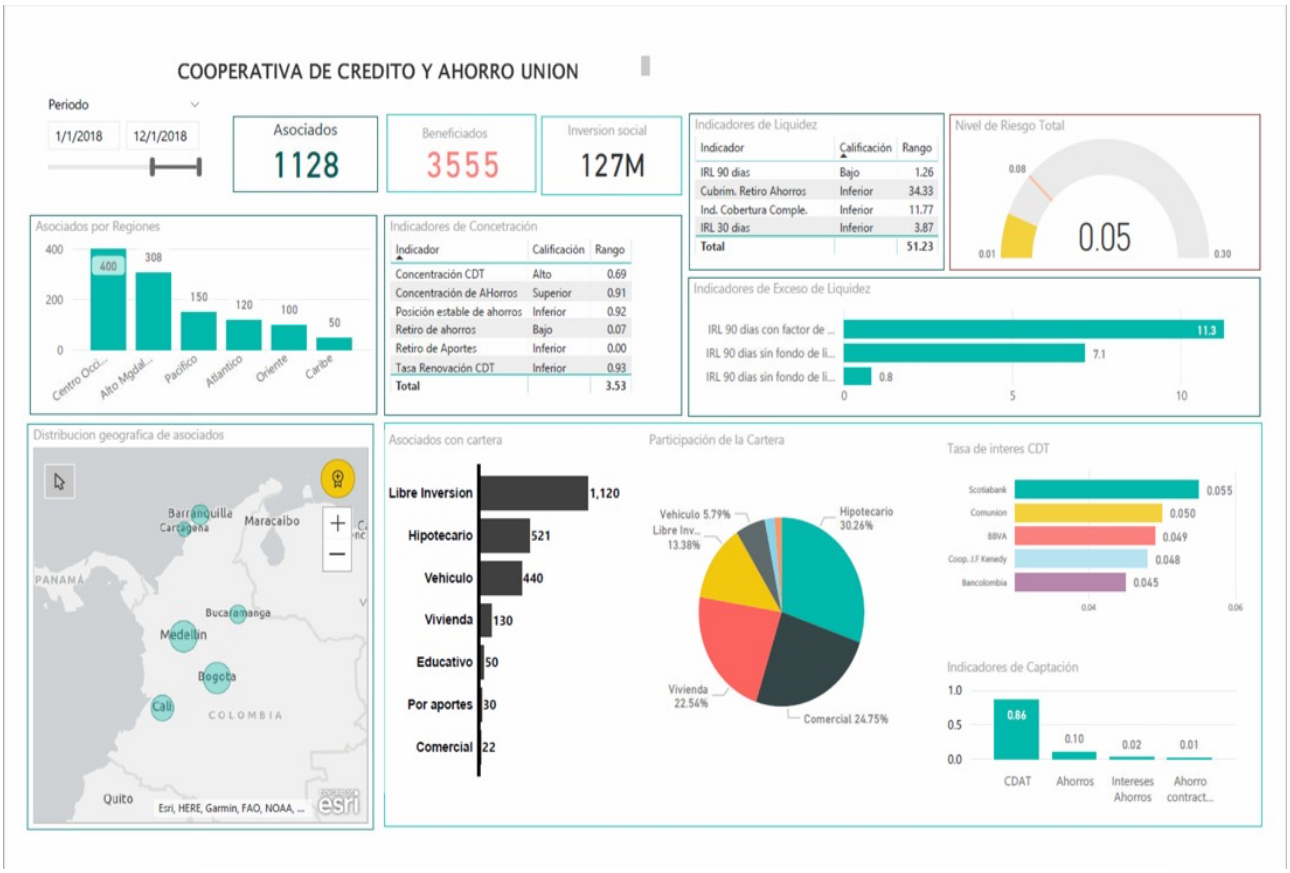


Figura 15. Tablero de control en Power BI. Elaboración propia.

Tabla 19. Esfuerzo según las actividades de la propuesta metodológica.

Fase	Actividad	Horas	Total Horas
Fase 1	Selección de los Objetivos Empresariales	3	**20**
	Definición de los Objetivos del proyecto	3	
	Asignación de recursos	3	
	Determinación del alcance y Riesgos	5	
	Creación Documento Contexto del proyecto	6	
Fase 2	Recolección de los Datos	32	**230**
	Exploración de los Datos	40	
	Integración de los Datos	64	
	Formateo de Datos	94	
Fase 3	Selección del modelo	2	**45**
	Selección de la Herramienta	3	
	Construcción del Tablero	40	
Fase 4	Evaluación del Tablero	3	**5**
	Análisis Resultados	2	
Fase 5	Automatización de los procesos	16	32
	Distribución de los Resultados	16	
Total esfuerzo en horas hombre			**332**

8.2.6 Resultados del Caso de Estudio

Los resultados obtenidos en la evaluación de la metodología en el caso de estudio muestran que el esfuerzo total empleado para el Proyecto DA fue de 332 horas hombre, teniendo en cuenta todos los roles involucrados en el equipo de trabajo.

La fase que mayor esfuerzo tuvo fue gestión de datos, que se consumió el 69% del tiempo total del proyecto. Dicho tiempo se justifica debido a que en esta fase están las actividades que mayor esfuerzo requieren, especialmente el formateo de los datos para la creación de los KPI de riesgos de liquidez que se consumió el 28% del total del esfuerzo del proyecto. La segunda actividad que mayor esfuerzo requirió fue la integración de los datos con un 19% del esfuerzo total. Las otras actividades que jugaron un papel importante en la medición del esfuerzo fueron la exploración de los datos con 12% y la recolección de los datos con un 10% del total de horas.

A pesar de que las demás fases incluían varias actividades, solo llegaron hasta un 14% de esfuerzo. En el caso del modelado, la construcción del tablero de control fue la que más esfuerzo empleó con un 12% del total de horas. En la figura 16 se presenta la participación total de cada una de las fases de la propuesta metodológica.

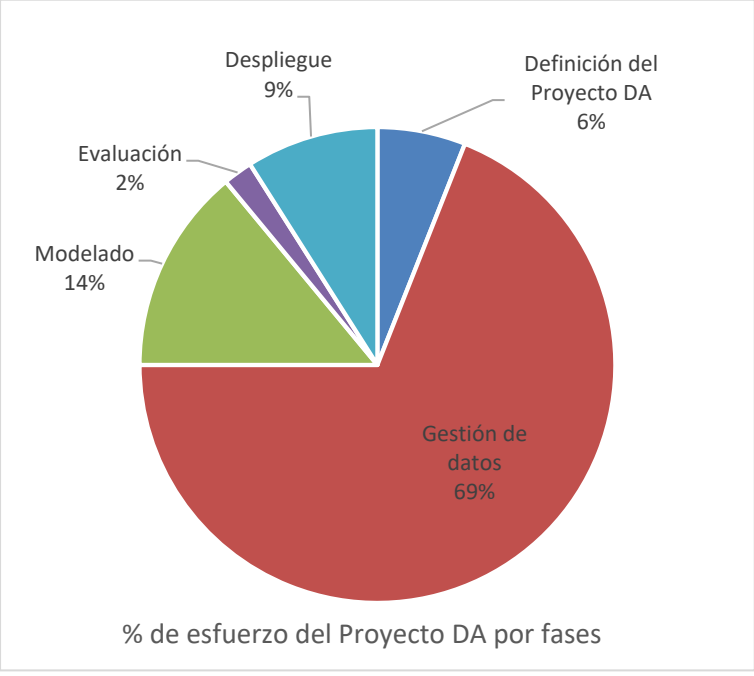


Figura 16. Porcentaje de esfuerzo por fases. Elaboración propia

8.3 Definición de la línea base

Para estimar el esfuerzo que tendría el Proyecto DA con otra metodología y en ausencia del uso de metodologías de analítica de datos en la PYME, se optó por estimar el tiempo que toma desarrollar el caso de estudio usando el método Wideband Delphi (Wiegers, 2000); la cual combina la técnica Delphi y el método probabilístico para obtener una línea base de comparación para la propuesta CRISP-DM/SMEs.

8.3.1 Juicio de Expertos.

Para realizar la estimación se reunió en un mismo lugar a un grupo de 5 expertos, a quienes se les presentó la descripción del proyecto y se les pidió que realizaran 3 estimaciones; la optimista (Do), la pesimista (Dp) y la media (Dm), teniendo en cuenta la aplicación de la metodología CRISP-DM. Luego se discutieron los resultados obtenidos y cada uno presentó sus argumentos sobre la estimación realizada en cada fase de dicha metodología. Luego se les pidió a los expertos que ajustaran sus estimaciones teniendo en cuenta los nuevos elementos que detectaron en la discusión. De esta forma cada uno de los expertos pudo volver a estimar los tiempos totales realizando nuevamente las 3 estimaciones y cuyos resultados se presentan en la tabla 20.

Para hallar la duración estimada $\mathbf{D_e}$, se utilizó la siguiente ecuación:

$$\mathbf{D_e} = (D_o + 4D_m + D_p) / 6$$

donde,

D_o es la duración optimista

D_m es la duración media

D_p es la duración pesimista

La varianza de dicha estimación se calcula con la siguiente manera:

$$V^2 = ((D_p - D_o) / 6)^2$$

Y la desviación estándar se calcula con la raíz cuadrada de la varianza, entonces se tiene:

$$\text{Desviación estándar: } \mathbf{S} = \sqrt{V^2}$$

Tabla 20. Evaluación de juicios de expertos

METODOLOGIA CRISP-DM	Experto #1			Experto #2			Experto #3			Experto #4			Experto #5		
	Dp1	Dm1	Do1	Dp2	Dm2	Do2	Dp3	Dm3	Do3	Dp4	Dm4	Do4	Dp5	Dm5	Do5
Fase 1: Comprensión del negocio															
Determinar objetivos del negocio															
Evaluar la situación	50	48	40	100	64	32	45	36	27	128	64	32	100	80	50
Determinar los objetivos DM															
Realizar el plan del proyecto															
Fase 2: Comprensión de los datos															
Recolectar datos iniciales															
Describir los datos	70	60	55	100	80	64	80	63	45	90	60	32	110	90	60
Explorar los datos															
Verificar la calidad de los datos															
Fase 3: Preparación de los datos															
Seleccionar los datos															
Limpiar los datos	200	160	130	480	240	180	250	180	135	100	80	60	80	60	50
Construir los datos															
Integrar los datos															
Formatear los datos															
Fase 4: Modelado															
Escoger la técnica de modelado															
Generar el plan de pruebas	100	90	70	270	120	32	135	90	63	40	30	20	50	40	30
Construir el modelo (Tablero de control)															
Evaluar el modelo															
Fase 5: Evaluación															
Evaluar los resultados	45	40	30	120	32	16	45	27	18	25	20	16	60	50	40
Revisar el proceso															
Fase 6: Despliegue															
Planear el despliegue	30	28	25	16	12	8	36	27	18	20	16	8	20	15	10
Planear la monitorización y mantenimiento															
Producir el informe final															
ESTIMACION TOTAL DEL ESFUERZO EN HORAS	495	426	350	1086	548	332	591	423	306	403	270	168	420	335	240

A partir de los tiempos totales del proyecto se calcularon las duraciones promedio para cada tipo de estimación:

$D_{o(proyecto)}$=279 horas/hombre

$D_{m(proyecto)}$=400 horas/hombre y

$D_{p(proyecto)}$=599 horas/hombre.

Al aplicar la ecuación, se obtiene la duración estimada, así:

$$\mathbf{D_e}_{(proyecto)} = (279 + 4 * 400 + 599) / 6 = 413 \text{ horas/hombre}$$

Y la varianza del proyecto obtenida es

$$V^2_{(proyecto)} = ((599 – 279) / 6)^2 = 2841$$

La desviación estándar se calcula con la raíz cuadrada de la varianza, entonces se tiene:

Desviación estándar: $\mathbf{S} = \sqrt{2841} = (\pm)$ 53 horas

La desviación estándar pudo ser más pequeña si se hubiese ajustado los resultados en una tercera y cuarta estimación por parte de los expertos, ya que en la segunda evaluación aún se encontraron estimaciones muy distantes entre un experto y otro, como

es el caso del experto 1 que estimó un promedio de 81 horas en un escenario medio, mientras que el experto 2 lo hizo con 796 horas para el mismo escenario.

8.3.2 Esfuerzo según la línea base.

Según los resultados de la línea base, aplicando el método Wideband Delphi, se estima que el esfuerzo del proyecto aplicando la metodología CRISP-DM en el Proyecto DA en el caso de estudio de esta investigación es de 413 horas hombre, con una desviación estándar de 53 horas. Esto quiere decir que el esfuerzo del proyecto puede variar entre 53 horas menos o 53 horas más de lo estimado.

9. Análisis y discusión de resultados

La propuesta metodológica fue diseñada a partir CRISP-DM como modelo de referencia para crear una metodología que ayude a disminuir el esfuerzo en los proyectos de analítica de datos en las PYME ESAL. Por lo tanto, en el proceso de construcción de dicha propuesta, se extrajeron las actividades más importantes de la analítica para ser modificadas según las necesidades de las PYME. Además, se incluyeron algunos elementos importantes como son los roles, las herramientas, los productos de trabajo y los tableros de control para darle solidez a la metodología.

CRISP-DM/SMEs fue evaluada con respecto a su modelo de referencia bajo un marco comparativo y obtuvo una calificación positiva del 75%, un poco mayor que la de su referente CRISP-DM que tiene un 67%, siendo esta una de las más referenciadas en la literatura y en la industria. Por lo tanto, se puede decir que la propuesta metodológica cumple satisfactoriamente con las características de una metodología de analítica de datos.

En cuento a la hipótesis que planteaba que la implementación de una metodología modificada de CRISP-DM disminuye el esfuerzo en los proyectos de analítica de datos en las PYME ESAL, en comparación con los proyectos de analítica del entorno empresarial que utilizan otras metodologías o ninguna en específico, se puede concluir que los resultados fueron prometedores, ya que en la aplicación del caso de estudio, tuvo un resultado satisfactorio, como se muestra en la figura 17.

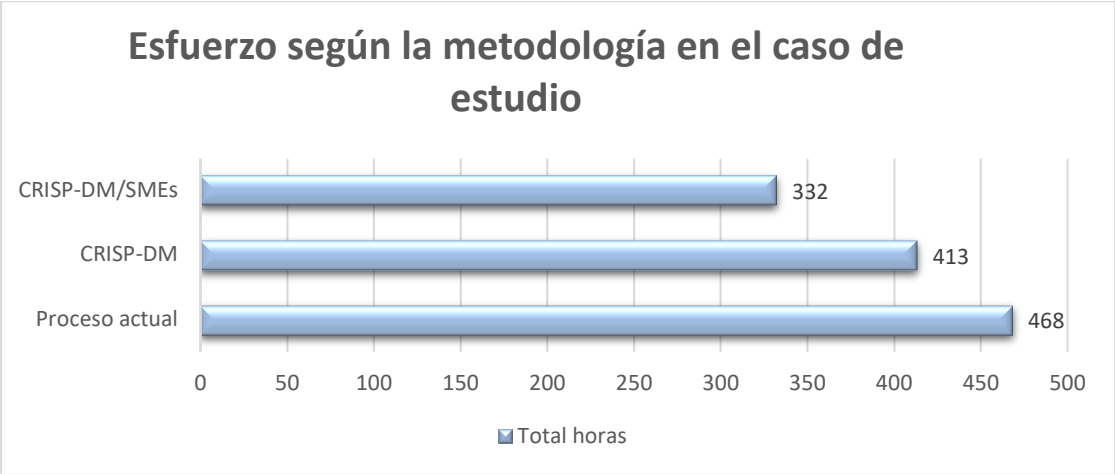


Figura 17 . Esfuerzo según la metodología en el caso de estudio. Elaboración propia.

El esfuerzo total siguiendo la propuesta metodológica fue de 332 horas hombre, mientras que con la metodología CRISP-DM, según el juicio de expertos, el esfuerzo promedio fue de 413 horas hombre, en un escenario medio. Por otro lado, el esfuerzo actual de la PYME para construir los indicadores, propuesto en el Proyecto DA, tarda 468 horas hombre por año; es decir que con respecto a la metodología de referencia hubo una disminución del esfuerzo del 20% aproximadamente, y con respecto al proceso que llevaba la empresa, disminuyó en un 30% aproximadamente.

Estos resultados indican que seguir la propuesta metodológica CRISP-DM/SMEs disminuyó el esfuerzo en la implementación del Proyecto DA en la PYME en comparación con CRISP-DM o con el proceso manual que llevaba la organización para obtener unos resultados similares.

Sin embargo, cabe resaltar que el conocimiento de las herramientas usadas en la metodología de analítica de datos por parte del ingeniero de datos influye significativamente en el resultado, ya que las actividades de las fases gestión y modelado son las que mayor esfuerzo

requieren y son desarrolladas por dicho rol. Además hay que tener en cuenta que el esfuerzo medido en la metodología es de horas hombre, mas no lo que cuesta la hora de cada rol, así que los resultados pueden variar si se representan las horas en dinero.

Finalmente, si en la definición del Proyecto DA se proponen herramientas sobre las cuales el ingeniero de datos tiene poca experiencia, entonces se recomienda una medición de la curva de aprendizaje y recalcular el esfuerzo total.

PARTE VI
CONCLUSIONES

10. Conclusiones

Se propone la metodología CRISP-DM/SME's para disminuir el esfuerzo de la implementación de proyectos de analítica de datos en las PYME, mejorar la recolección, almacenamiento, procesamiento y análisis de los datos, e inducir la creación de conciencia en la toma de decisiones basada en la exploración de la información. Ya que al contar con una metodología de analítica para PYME permite minimizar la complejidad, los costos y ayuda a tener personal capacitado para la implementación de proyectos DA. Además, permite un despliegue rápido, fácil de mejorar e integrar en otros proyectos.

La integración de roles en la metodología ayudó a la asignación de las tareas a cada uno de los involucrados en el proyecto DA dentro de la PYME, empleando las competencias y habilidades de los profesionales de la industria de la ciencia de los datos. Además, la inclusión de tableros de control como un producto de trabajo fue determinante para la analítica de datos en la PYME, pues representó gráficamente el resultado del proyecto y se constituyó en una herramienta útil para la gerencia y toma de decisiones.

El uso del lenguaje SPEM para la representación de la propuesta metodológica permitió una descripción abstracta de los elementos fundamentales del proceso de analítica de datos en la PYME, y también la descripción de cómo ellos estaban relacionados entre sí, especialmente los roles con sus respectivas actividades, que a su vez consumían y producían productos de trabajo, facilitando de este modo, el uso de la metodología.

La evaluación de la metodología en el caso de estudio fue satisfactoria, pero se requiere de más casos de éxito para dar mayores evidencias de la disminución del esfuerzo con la propuesta metodológica y confirmar los resultados obtenidos en la

primera evaluación. No obstante, la metodología puede ser evaluada en PYME que no necesariamente sean de tipo ESAL, para ampliar el campo de aplicabilidad de la propuesta.

11. Recomendaciones, trabajos futuros y aportes de la investigación

11.1 Recomendaciones

Se recomienda adoptar la propuesta metodológica para los proyectos de analítica de datos en PYME ESAL para seguir evaluando y documentando los resultados del esfuerzo obtenido en comparación con otra metodología. También se recomienda el uso de herramientas que puedan ser accesibles a las PYME, ya sea por el costo o porque ya las ha adquirido previamente.

Debido a que las PYME apenas están incursionando en la analítica de datos y la mayoría no cuentan con Ingenieros de datos, se recomienda que este rol sea de una persona externa por servicios con las competencias básicas en el área para que implemente el Proyecto DA.

11.2 Trabajos futuros

Como trabajo futuro se propone la validación de la propuesta metodológica en diversos Proyectos DA en PYME ESAL para la identificación de mejoras que conlleve a la disminución del esfuerzo en comparación con otras metodologías de analítica de datos.

Además, se propone la realización de una plataforma web alcance de las PYME que permita el repositorio de cada uno de los Proyectos DA, con las respectivas fases y tareas, roles, productos de trabajo y las guías de cómo realizar cada tarea. Y como aporte a la investigación, dicha plataforma comparta los resultados de los proyectos a la comunidad científica y empresarial.

11.3 Aportes de la investigación

Como resultado de esta investigación se construyeron los siguiente entregables los cuales quedan a disposición de la comunidad académica y empresarial.

11.3.1 Artículo científico

Se envió el artículo ***"CRISP-DM / SMEs: A Data Analytics Methodology for Non-profit SMEs"*** al ICICT 2019 *"International Congress on Information and Communication Technology"*, celebrado el 25-26 de febrero de 2019 en Londres, Inglaterra, el cual fue evaluado por pares investigadores y fue aprobado para ser publicado en Springer para que sea referenciado en futuras investigaciones. El anexo 6 contiene el certificado del aporte realizado en el evento.

11.3.2 Publicación de la Metodología

Se construyó un documento con la propuesta, llamado **"CRISP-DM/SMEs Metodología de Analítica de Datos para PYME sin ánimo de lucro – ESAL"** la cual estará al alcance de las PYME, académicos y científicos de datos.

REFERENCIAS BIBLIOGRÁFICAS

Achmad, H., Sabur, V. F., Pritasari, A., & Reinaldo, H. (2016). Sains Humanika Data Mining and Sharing to Create Usable Knowledge , Implementation in Small Business in Indonesia. *Sains Humanika*, *2*(2016), 69–75.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP:DM: A parallel overview. *IADIS European Conference Data Mining 2008 2.*, 182–185.

Bošnjak, Z., Grljević, O., & Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. *Proceedings - 2009 5th International Symposium on Applied Computational Intelligence and Informatics, SACI 2009*, *xx*(1), 509–514. https://doi.org/10.1109/SACI.2009.5136302

CCONG. (2016). Quiénes conforman el sector de las Entidades Sin Ánimo de Lucro- ESAL en Colombia | CCONG :: Confederación Colombiana de ONG. Retrieved August 30, 2018, from http://ccong.org.co/ccong/documentos/grafico-ii:-quienes-conforman-el-sector-de-las-entidades-sin-animo-de-lucro--esal-en-colombia_617

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

DataCamp. (2015). The Data Science Industry: Who Does What (Infographic) (article) - DataCamp. Retrieved October 18, 2018, Recuperado de https://www.datacamp.com/community/tutorials/data-science-industry-infographic

Davenport, T. H. (2013). Analytics 3.0. Recuperado de https://hbr.org/2013/12/analytics-30

Dinero. (2016). Pymes contribuyen con más del 80% del empleo en Colombia. Recuperado de http://www.dinero.com/edicion-impresa/caratula/articulo/porcentaje-y-contribucion-de-las-

pymes-en-colombia/231854

Dittert, M., Härting, R., Reichstein, C., & Bayer, C. (2018). A Data Analytics Framework for Business in Small and Medium-Sized Organizations. *Springer International Publishing AG 2018*, *73*, 13. https://doi.org/10.1007/978-3-319-59424-8

Dos Anjos, J. C. S., Assuncao, M. D., Bez, J., Geyer, C., de Freitas, E. P., Carissimi, A., … Pereira, R. (2015). SMART: An Application Framework for Real Time Big Data Analysis on Heterogeneous Cloud Environments. *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 199–206. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.29

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–53. https://doi.org/10.1609/aimag.v17i3.1230

Florez, H. (2012). Inteligencia de negocios como apoyo a la toma de decisiones en la gerencia. *Revista Vinculos*, *9*(2), 11–23.

Gonzalez, A. (2014). Big data y analítica en Colombia: A un paso de despegar. Recuperado de https://searchdatacenter.techtarget.com/es/cronica/Big-data-y-analitica-en-Co

Gonzalez, R.A. & Pomares, A. (2012) "La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería". Reunión Nacional ACOFI, Sep. 12-14, Medellín.

Guarda, T., Santos, M., Pinto, F., Augusto, M., & Silva, C. (2013). Business Intelligence as a Competitive Advantage for SMEs. *International Journal of Trade, Economics and Finance*, *4*(4), 187–190. https://doi.org/10.7763/IJTEF.2013.V4.283

Gudfinnsson, K., & Strand, M. (2018). Challenges with BI adoption in SMEs. *2017 8th International Conference on Information, Intelligence, Systems and Applications, IISA 2017*, *2018–Janua*(August 2017), 1–6. https://doi.org/10.1109/IISA.2017.8316407

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems*, *28*(1).

IBM (2012). Manual CRISP-DM de IBM SPSS Modeler. IBM Corp., p. 56.

Kalan, R., & Ünalir, M. (2016). Leveraging big data technology for small and medium-sized enterprises (SMEs). *Computer and Knowledge Engineering (*, (Iccke), 1–6. https://doi.org/10.1109/ICCKE.2016.7802106

KDnuggets. (2014). What main methodology are you using for your analytics, data mining, or data science projects? Recuperado de https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Lawton, G. (2009). Users Take a Close Look at Visual Analytics. *Computer*, *42*(2), 19–22. https://doi.org/10.1109/MC.2009.61

Menendez, V. ., & Castellanos, M. . (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. Mi., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores En Ciencias de La Computación*, 278–281. Recuperado de http://sedici.unlp.edu.ar/handle/10915/20034

Mullins, R., Duan, Y., Hamblin, D., Burrell, P., Jin, H., Ewa, Z., & Aleksander, B. (2007). A Web Based Intelligent Training System for SMEs. *The Electronic Journal of E-Learning*, *5*(1), 39–48.

Nenzhelele, T. E., & Pellissier, R. (2014). Competitive Intelligence Implementation Challenges of Small and Medium-Sized Enterprises. *Mediterranean Journal of Social Sciences*, *5*(16), 92–99. https://doi.org/10.5901/mjss.2014.v5n16p92

Ogbuokiri, B. ., Udanor, C. ., & Agu, M. . (2015). Implementing bigdata analytics for small and medium enterprise (SME) regional growth. *IOSR Journal of Computer Engineering Ver. IV*, *17*(6), 2278–2661. https://doi.org/10.9790/0661-17643543

Olivera, G., Sasa, B., & Zita, B. (2009). Intelligent Data Anallysis - Support for Development of SMEs Sector. *Perspectives of Innovations, Economics & Business*, *3*(114), 57–58.

Oztekin, A., Best, K., & Delen, D. (2014). Analyzing the predictability of exchange traded funds characteristics in the mutual fund market on the flow of shares using a data mining approach. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 779–788. https://doi.org/10.1109/HICSS.2014.104

Pytel, P., Hossian, A., Britos, P., & Garcia-Martinez, R. (2015). Feasibility and effort estimation models for medium and small size information mining projects. *Information Systems*, *47*, 1–14. https://doi.org/10.1016/j.is.2014.06.004

Rodríguez, L. (2011). Entidades sin ánimo de lucro. *Confederación Colombiana de ONG*, (019), 2012–2013. Recuperado de http://exitojuridico.blogspot.com.co/2011/05/entidades-sin-animo-de-lucro.html

Sinnetic. (2017). PYMES se desaceleran en transformación digital e innovación por responder a múltiples requerimientos estatales, *18*, 1–3. Recuperado de http://www.sinnetic.com/noticias/SINNETIC-NEWS-18.pdf

Soroka, A., Liu, Y., Han, L., & Salman, M. (2017). Big data driven customer insights for SMEs in redistributed manufacturing. *Procedia CIRP*, *63*, 692–697.

https://doi.org/10.1016/j.procir.2017.03.319

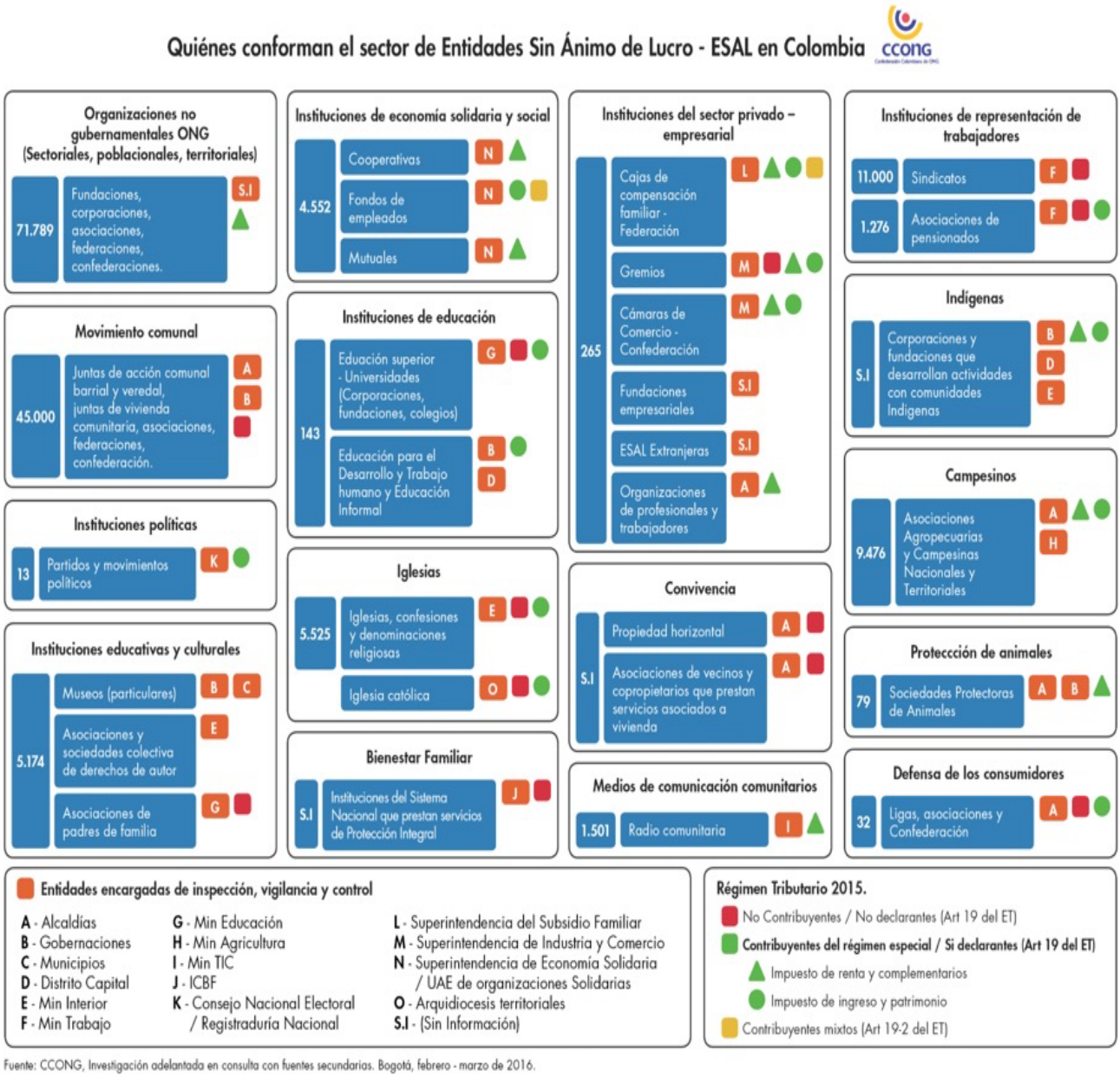
Soto, E. (2004). La información como recurso estratégico generador de conocimientos. Un enfoque de recursos y capacidades. (Tesis doctoral) Universidad de la Laguna. España.

Unipymes. (2014). Empresas Colombianas en crecimiento empiezan a acceder a la analitica de negocios. Recuperado de https://www.unipymes.com/empresas-colombianas-en-crecimiento-empiezan-acceder-la-analitica-de-negocios/

Wiegers, K. (2000). Stop promising miracles, Software Development. Vol 8, no 2 pp 49-53 Recuperado de https://www.processimpact.com/articles/reqtraps.pdf

ANEXOS

Anexo 1: Clasificación de las ESAL



Fuente: CCONG, Investigación adelantada en consulta con fuentes secundarias. Bogotá, febrero - marzo de 2016.

Anexo 2: Evaluación comparativa de la propuesta metodológica

Detalle del aspecto 3: Actividades específicas que componen cada fase.

Evaluación de las actividades en la fase de análisis del problema

#	Característica	CRISP-DM	CRISP-DM/SMEs
3.1	¿Se propone una evaluación general de la organización?	SI	NO
3.2	¿Se identifica al personal involucrado en el proyecto (stackeholders)?	SI	SI
3.3	¿Se define el problema u oportunidad de negocio?	SI	SI
3.4	¿Se propone una evaluación de las fuentes de datos?	NO	SI
3.5	¿Se analizan todas las soluciones posibles al problema?	NO	NO
3.6	¿Se especifican los objetivos del proyecto?	SI	SI
3.7	¿Se define un criterio de éxito para el proyecto?	SI	SI
3.8	¿Se realiza una evaluación general de las técnicas de minería que podrían utilizarse?	SI	NO
3.9	¿Se especifica de qué forma el usuario utilizará el nuevo conocimiento?	NO	NO
	Valoraciones positivas	6/9 = 66%	5/9 = 55%

Evaluación de las actividades en la fase de selección y preparación de los datos

#	Característica	CRISP-DM	CRISP-DM/SMEs
3.10	¿Se propone un análisis exploratorio inicial de los datos?	SI	SI
3.11	¿Se sugieren actividades para la limpieza de los datos?	SI	SI
3.12	¿Se contemplan actividades para la transformación de variables y la creación de atributos derivados?	SI	SI
3.13	¿Se realiza un análisis descriptivo final sobre los datos depurados?	NO	NO
3.14	¿Se verifica con el usuario la completitud del conjunto de datos final?	NO	SI
	Valoraciones positivas	3/5 = 60%	4/5 = 75%

Evaluación de las actividades en la fase de modelado

#	Característica	CRISP-DM	CRISP-DM/SMEs
3.15	¿Se efectúa una selección de las técnicas que se utilizarán?	SI	SI
3.16	¿Se planifica la forma en la que se evaluarán los resultados?	SI	SI
3.17	¿Se efectúa una evaluación inicial de los modelos obtenidos?	SI	SI
3.18	¿Se proveen directivas para el caso donde se dificulta el descubrimiento de patrones?	NO	NO
	Valoraciones positivas	3/4 = 75%	3/4 = 75%

Confrontación de las actividades en la fase de evaluación

#	Característica	CRISP-DM	CRISP-DM/SMEs
3.19	¿Se interpretan los modelos en función de los objetivos organizacionales?	SI	SI
3.20	¿Se comparan y ponderan los modelos obtenidos?	SI	NO
3.21	¿Se propone una revisión general del proceso?	SI	SI
3.22	¿Se proveen directivas para el caso donde ninguno de los modelos obtenidos resulta viable?	SI	NO
	Valoraciones positivas	4/4 = 100%	2/4 = 50%

Evaluación de las actividades para la fase de implementación

#	Característica	CRISP-DM	CRISP-DM/SMEs
3.23	¿Se planifica la implementación del nuevo conocimiento?	SI	SI
3.24	¿Se propone la creación de un programa de mantenimiento?	SI	SI
3.25	¿Se entrega al usuario un resumen del proyecto?	SI	SI
3.26	¿Se documenta la experiencia adquirida por el equipo de trabajo?	SI	SI
	Valoraciones positivas	4/4 = 100%	4/4 = 100%

Detalle del aspecto 4: Actividades destinadas a la dirección del proyecto.

Evaluación de la gestión del alcance

#	Característica	CRISP-DM	CRISP-DM/SMEs
4.1	¿Se propone la selección de los entregables que se generarán durante el proyecto?	SI	SI
4.2	¿Se **especifican** actividades de control del alcance?	NO	SI
	Valoraciones positivas	1/2 = 50%	2/2 = 100%

Evaluación de la gestión del tiempo

#	Característica	CRISP-DM	CRISP-DM/SMEs
4.3	¿Se realiza una definición y secuenciación de las actividades que se ejecutarán durante el proyecto?	SI	Si
4.4	¿Se realiza una estimación de la duración de cada actividad?	SI	SI
4.5	¿Se construye un cronograma para el proyecto?	SI	SI
4.6	¿Existen actividades de control del cronograma?	NO	SI
	Valoraciones positivas	3/4 = 75%	4/4 = 100%

Evaluación de la gestión del costo

#	Característica	CRISP-DM	CRISP-DM/SMEs
4.7	¿Se efectúa una estimación de los recursos afectados por cada actividad?	SI	SI
4.8	¿Se realiza una estimación de los costos del proyecto?	NO	SI
4.9	¿Se construye un presupuesto de costos?	NO	SI
4.10	¿Existen actividades de control del presupuesto a medida que avanza el proyecto?	NO	SI
	Valoraciones positivas	1/4 = 25%	4/4 = 100%

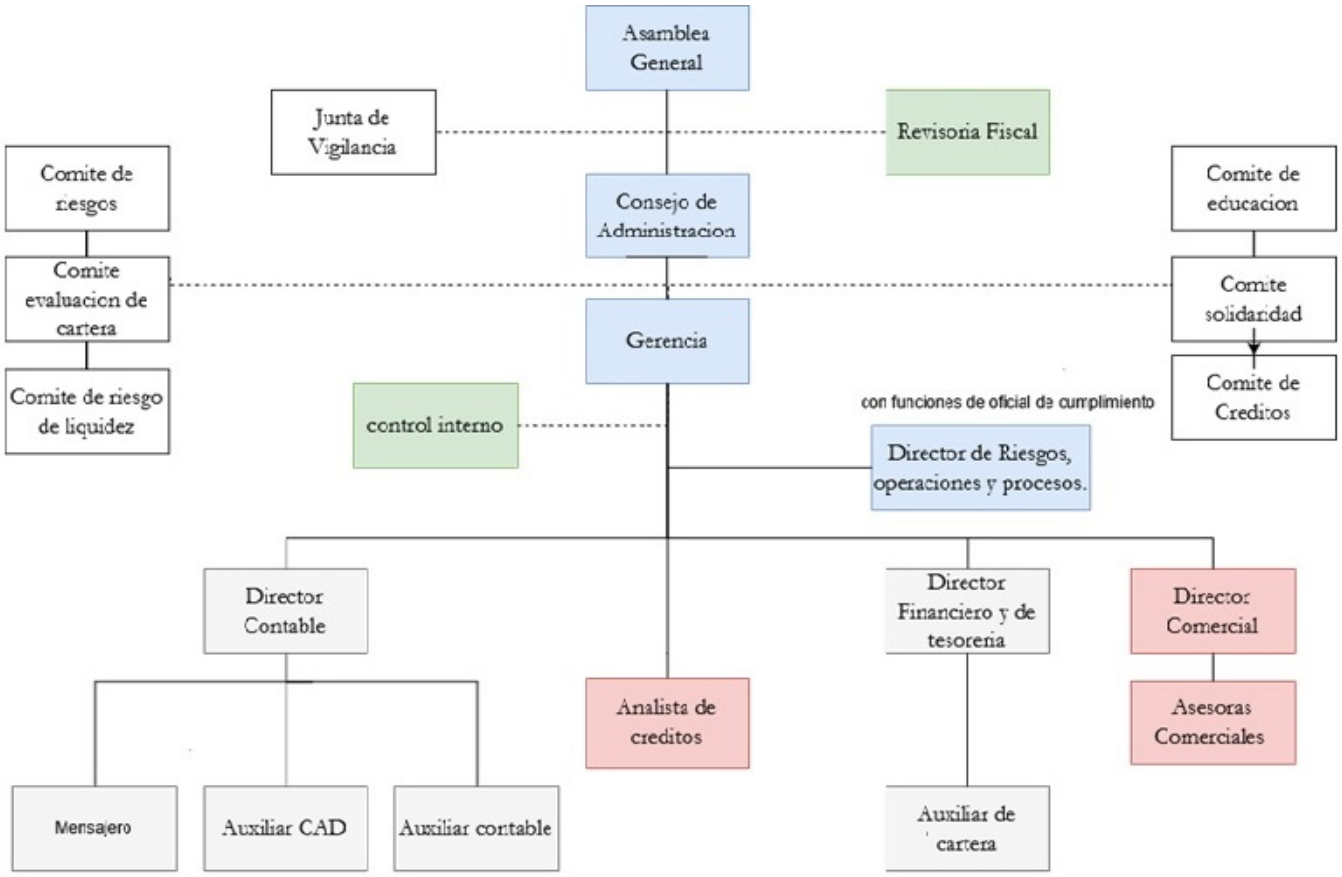
Evaluación de la gestión del equipo de trabajo.

#	Característica	CRISP-DM	CRISP-DM/SMEs
4.11	¿Se efectúa una planificación de los recursos humanos?	SI	SI
4.12	¿Se proponen actividades para motivar la interacción entre los miembros del equipo?	NO	NO
4.13	¿Se efectúa un seguimiento del rendimiento de los recursos humanos?	NO	SI
	Valoraciones positivas	1/3 = 33%	2/3 = 66%

Evaluación de la gestión del riesgo

#	Característica	CRISP-DM	CRISP-DM/SMEs
4.14	¿Se efectúa una identificación de los riesgos del proyecto?	SI	SI
4.15	¿Se realiza una cuantificación de los riesgos?	NO	NO
4.16	¿Se planifican acciones de respuesta ante cada riesgo?	SI	SI
4.17	¿Existen actividades de supervisión y control de los riesgos?	NO	NO
	Valoraciones positivas	2/4 = 50%	2/4 = 50%

Anexo 3: Organigrama Cooperativa Unión Colombiana



Anexo 4: Niveles de riesgo de liquidez

A continuación, se muestran los indicadores de Riesgo de Liquidez que se requieren presentar en un tablero de control como un proyecto de analítica.

INDICADORES DE COBERTURA GAP

	Apr-18	May-18	Jun-18
NIVEL RIESGO TOTAL	5.83%	5.83%	5.83%
NIVEL RIESGO TOTAL	MINIMO	MINIMO	MINIMO

INDICADORES DE CONCENTRACION

INDICADOR	Apr-18	May-18	Jun-18	W PONDERACION	RANGO	CALIFICACIÓN (D)	CALIFICACIÓN RIESGO
Retiro máximo aportes - mensual	0.29%	0.12%	0.22%	17%	<0,03	0%	INFERIOR
					>=0,03 <0,05	2%	
					>=0,05 <0,09	8%	
					>=0,09 <0,12	15%	
					>=0,12 <0,15	20%	
					>=0,15 <0,25	25%	
					>=0,25 <1	30%	
Retiros de los ahorros sobre activos liquidos	6.60%	7.02%	6.51%	17%	<0,03	0%	BAJO
					>=0,03 <0,05	2%	
					>=0,05 <0,09	8%	
					>=0,09 <0,12	15%	
					>=0,12 <0,15	20%	
					>=0,15 <0,25	25%	
					>=0,25	30%	
Concentración de ahorros (a la vista) - Gini	90.28%	91.30%	91.38%	17%	<0,09	2%	SUPERIOR
					>=0,1 <0,19	8%	
					>=0,2 <0,29	15%	
					>=0,3 <0,39	25%	
					>=0,4 <1	**30%**	
Posicion estable de ahorros (a la vista)	93.26%	92.86%	92.42%	17%	<0,15	30%	INFERIOR
					>=0,16 <0,3	25%	
					>=0,31 <0,45	8%	
					>=0,46 <0,6	15%	
					>=0,61 <0,75	8%	
					>=0,76 <0,9	2%	
					>=0,91 <1	**0%**	
Concentración depósitos a término - Gini	69.85%	69.27%	69.29%	17%	<0,2	2%	ALTO
					>=0,2 <0,3	8%	
					>=0,3 <0,5	15%	
					>=0,5 <0,7	20%	
					>=0,7 <1	30%	
Tasa de renovación certificados a término PJ	93.80%	92.87%	93.23%	17%	<0,15	30%	INFERIOR
					>=0,16 <0,3	25%	
					>=0,31 <0,45	8%	
					>=0,46 <0,6	15%	
					>=0,61 <0,75	8%	
					>=0,76 <0,9	2%	
					>=0,91 <1	**0%**	
NIVEL DE RIESGO	9.7%	9.7%	9.7%				
NIVEL DE RIESGO	BAJO	BAJO	BAJO				

INDICADORES DE LIQUIDEZ

INDICADOR	Apr-18	May-18	Jun-18	W PONDERACION	RANGO	CALIFICACIÓN (D)	CALIFICACIÓN RIESGO
IRL 90 DIAS	130.23%	115.62%	126.10%	25%	<90%	30%	BAJO
					>=90 , <100%	20%	
					>=100% , <150%	8%	
					>=150% , < 200%	2%	
					>=200%	0%	
Indicador de cobertura complementaria	1374.29%	1159.41%	1176.92%	25%	>=0 <1	30%	INFERIOR
					>=1 <1,5	20%	
					>=2 <2,5	8%	
					>=3 <10	0%	
IRL 30 DIAS	264.59%	472.61%	387.37%	25%	<90%	30%	INFERIOR
					>=90 , <100%	20%	
					>=100% , <150%	8%	
					>=150% , < 200%	2%	
					>=200%	0%	
Cubrimiento de retiro máximo ahorros	4089.39%	3381.81%	3432.87%	25%	<1	30%	INFERIOR
					>=1 <2	20%	
					>=2 <2,5	8%	
					>=3 <10	0%	
NIVEL DE RIESGO	2.00%	2.00%	2.00%				
NIVEL DE RIESGO	MINIMO	MINIMO	MINIMO				

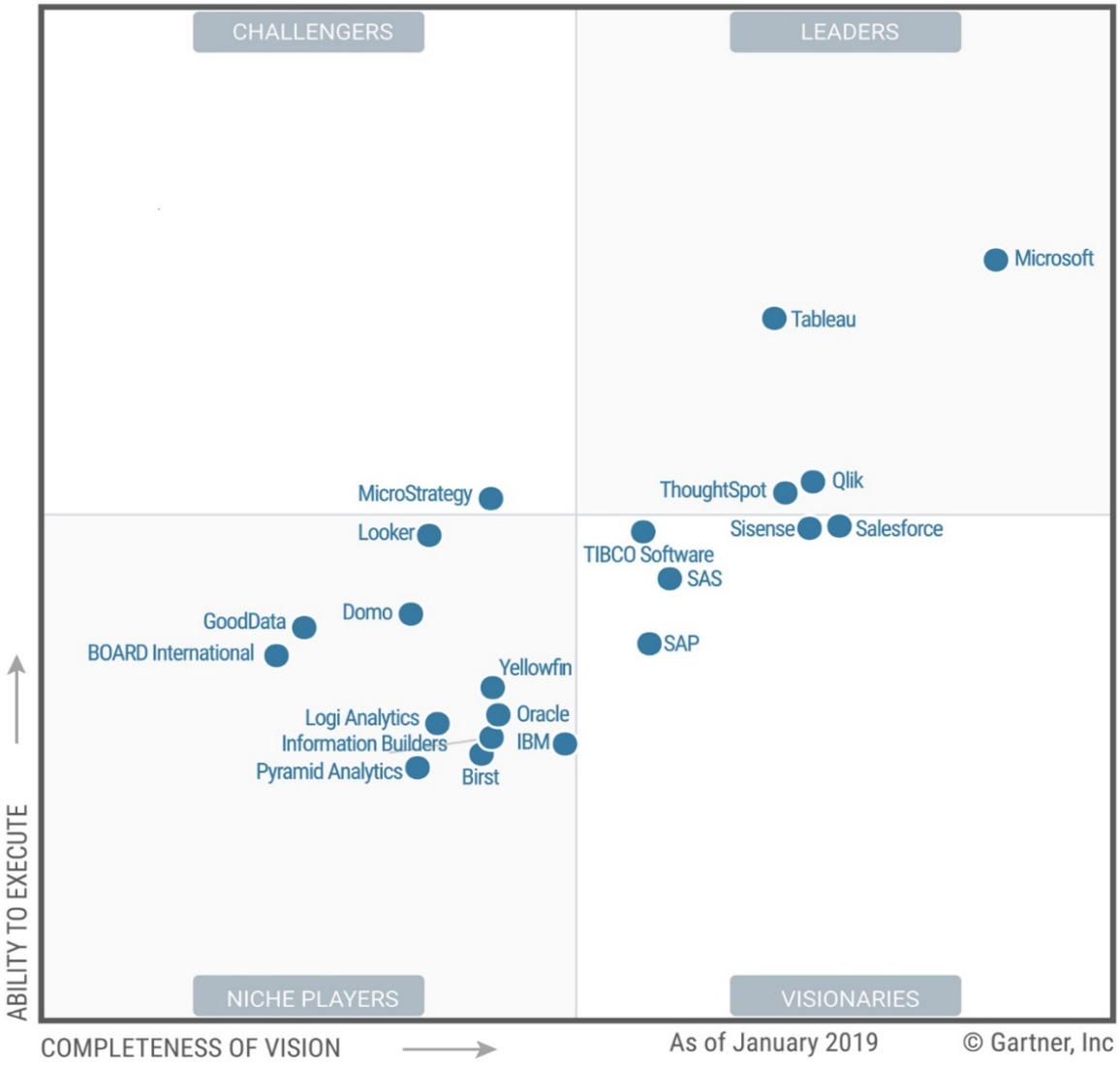
CRITERIOS DE CALIFICACION	Nivel de Riesgo
0%	INFERIOR
2%	MINIMO
8%	BAJO
15%	MEDIO
20%	ALTO
25%	MUY ALTO
30%	SUPERIOR

INDICADORES DE EXCESO DE LIQUIDEZ

INDICADOR	Apr-18	May-18	Jun-18	W PONDERACION	RANGO	CALIFICACIÓN (D)	CALIFICACIÓN RIESGO
IRL 90 DIAS SIN FONDO DE LIQUIDEZ CON FACTOR	914.38%	808.54%	714.80%	33%	<90%	30%	INFERIOR
					>=90 , <100%	20%	
					>=100% , <150%	8%	
					>=150% , < 200%	2%	
					>=200%	0%	
IRL 90 DIAS SIN FONDO DE LIQUIDEZ SIN FACTOR DE RENOVACION	81.70%	71.39%	81.84%	33%	<90%	30%	SUPERIOR
					>=90 , <100%	20%	
					>=100% , <150%	8%	
					>=150% , < 200%	2%	
					>=200%	0%	
IRL 90 DIAS CON FACTOR DE RENOVACION	1457.55%	1309.48%	1126.06%	33%	<90%	30%	INFERIOR
					>=90 , <100%	20%	
					>=100% , <150%	8%	
					>=150% , < 200%	2%	
					>=200%	0%	

NIVEL DE RIESGO	10%	10%	10%
NIVEL DE RIESGO	BAJO	BAJO	BAJO

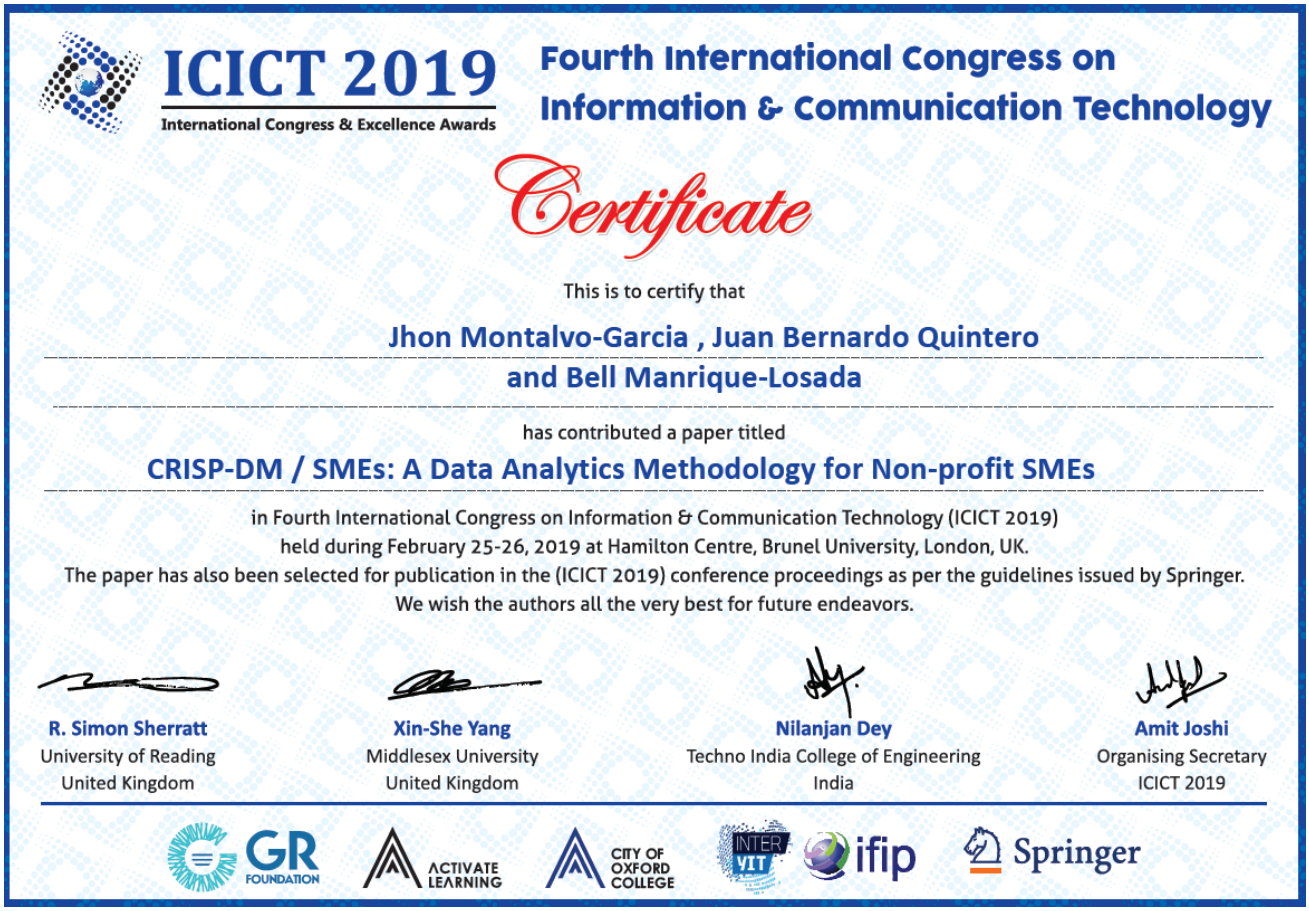
Anexo 5: Cuadrante mágico de Gartner para plataformas de analítica



Source: Gartner (February 2019)

Fuente: https://info.microsoft.com/rs/157-GQE-382/images/EN-CNTNT-GartnerMQ-BI2019.jpg

Anexo 6: Certificado de participación ICICT 2019



ICICT 2019
International Congress & Excellence Awards

Fourth International Congress on Information & Communication Technology

Certificate

This is to certify that

Jhon Montalvo-Garcia , Juan Bernardo Quintero and Bell Manrique-Losada

has contributed a paper titled

CRISP-DM / SMEs: A Data Analytics Methodology for Non-profit SMEs

in Fourth International Congress on Information & Communication Technology (ICICT 2019) held during February 25-26, 2019 at Hamilton Centre, Brunel University, London, UK.
The paper has also been selected for publication in the (ICICT 2019) conference proceedings as per the guidelines issued by Springer.
We wish the authors all the very best for future endeavors.

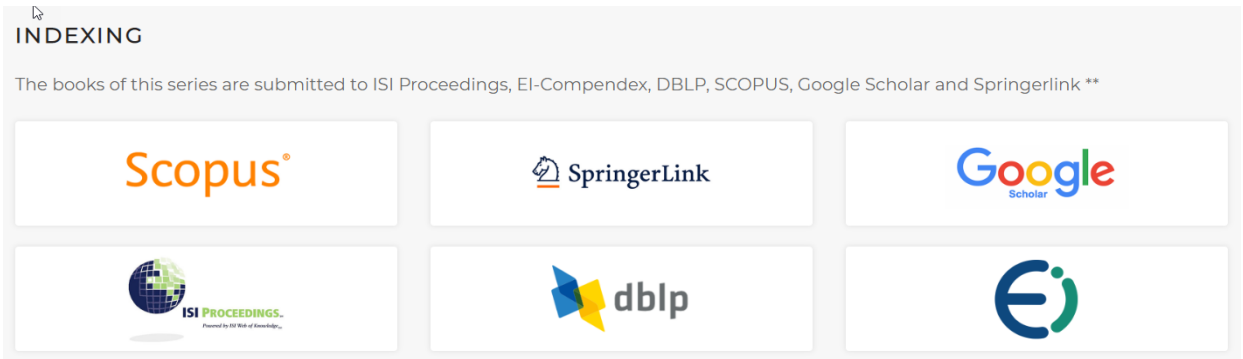
R. Simon Sherratt
University of Reading
United Kingdom

Xin-She Yang
Middlesex University
United Kingdom

Nilanjan Dey
Techno India College of Engineering
India

Amit Joshi
Organising Secretary
ICICT 2019

GR FOUNDATION · ACTIVATE LEARNING · CITY OF OXFORD COLLEGE · INTER VIT · ifip · Springer



INDEXING

The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink **

Scopus · SpringerLink · Google Scholar · ISI Proceedings · dblp · Ei