Modelo de Implementación, estrategia de Analítica de Datos como soporte de las funciones de IVC de la SDEGC

Propuesta de trabajo de grado presentado como requisito parcial para optar al título de Magíster en Gestión Estratégica de la Información

Soluciones Empresariales

Grupo de Investigación y Desarrollo en Informática y Telecomunicaciones

Director:

Magíster en Ingeniería. Luis Carlos Correa Ortiz

Codirector:

Doctor en Ciencias de la Educación. Carlos Betancourt Correa

Universidad de Manizales

Facultad de Ciencias e Ingeniería

Maestría en Gestión Estratégica de la Información

Manizales, 2023

Resumen

Este proyecto de investigación aborda la problemática institucional de la Superintendencia Delegada para Energía y Gas Combustible de la Superintendencia de Servicios Públicos Domiciliarios (SSPD), en el fortalecimiento metodológico de los procesos de analítica de datos orientados a obtener el mayor valor agregado del Sistema Único de Información de los prestadores de servicios públicos domiciliarios (SUI). Se realiza la construcción de un modelo de implementación soportado por un modelo documental inspirado en la metodología CRISP-DM y la aplicación de metodologías de desarrollo ágil de proyectos. Una vez construidos modelos de implementación y soporte documental se realiza un proceso de socialización y armonización con las dependencias administrativas de la SSPD, de cara a un proceso de adopción institucional y la identificación de escenarios de validación práctica del proyecto. Finalmente, se da aplicación al modelo de implementación y estructura documental en el desarrollo de tres procesos de analítica de datos que atienden necesidades institucionales.

PALABRAS CLAVE: Analítica de Datos, procesos, armonización, CRISP-DM y metodologías de desarrollo ágil.

Abstract

This research project addresses the institutional problems of the Delegated Superintendence for Energy and Fuel Gas of the Superintendence of Domiciliary Public Utilities (SSPD), in the methodological strengthening of data analytics processes aimed at obtaining the greatest added value of the Single Information System of the providers of domiciliary public utilities (SUI). The construction of an implementation model supported by a documentary model inspired by the CRISP-DM methodology and the application of agile project development methodologies is carried out. Once the implementation and documentary support models have been built, a socialization and harmonization process is carried out with the SSPD's administrative departments, with a view to an institutional adoption process and the identification of practical validation scenarios for the project. Finally, the implementation model and document structure are applied in the development of three data analytics processes that meet institutional needs.

Keywords: Data analytics, processes, harmonization, CRISP-DM and agile development methodologies

Tabla de contenido

Índice de Ilustraciones

Índice de Tablas

Introducción

La presente investigación está motivada en la necesidad de la Superintendencia de Servicios Públicos Domiciliarios (SSPD) de fortalecer los procesos institucionales de Inspección, Vigilancia y Control propios de su misión, a través del aprovechamiento intensivo de la información disponible en el Sistema Único de Información (SUI). Los procesos institucionales desarrollados actualmente están soportados por equipos interdisciplinarios pertenecientes a cada una de las dependencias con ausencia de unidad conceptual y metodológica que permita la interoperabilidad y la transferencia de conocimientos entre las diferentes instancias de la organización.

El Estado Colombiano establece un marco técnico-legal para el proceso de transformación digital de las diferentes instituciones que hacen parte del gobierno de la república de Colombia. (Mintic, 2020) A la luz de este marco se plantea para las diferentes organizaciones que componen el estado colombiano una inmersión en el enfoque productivo de la cuarta revolución industrial. Se plantea que las diferentes organizaciones que componen el gobierno colombiano deben consolidar productos informáticos orientados a la generación de valor a partir de la información disponible. Dentro de este enfoque del estado como creador de valor a partir de la información las tecnologías de la analítica de datos y la inteligencia artificial toman relevancia, y se hace necesario incorporarla a la visión y los procesos institucionales.

El proyecto tiene como objetivo unificar la adopción técnica y metodológica en el contexto de la transformación digital y las demandas institucionales para aumentar la productividad, disminuir los tiempos de reacción y ofrecer valor a los múltiples actores involucrados en el proceso de suministro de servicios públicos.

1. Planteamiento del problema de investigación

1.1 Descripción del área problemática

Para entender el contexto del problema a tratar es necesario entender que tipo de funciones cumple la Superintendencia de Servicios Públicos Domiciliarios. Esta entidad del gobierno colombiano es la responsable de vigilar a los prestadores de servicios domiciliarios del país, función que fue asignada a través del artículo 53 de la Ley 142 de 1994. La ley 142 designa la responsabilidad de crear los sistemas de información que permitan capturar la información necesaria de los prestadores de servicios públicos, como insumo para la vigilancia del desarrollo de su función (Congreso de Colombia, 1994).

El gobierno colombiano mejoró estos roles en respuesta a las obligaciones descritas en la Ley 142 al modificar la Ley 142 de 1994 con la Ley 689 de 2001. El establecimiento, administración, mantenimiento y operación del Sistema (SUI) son tareas dadas por los cambios aprobados. La plataforma tecnológica referida tiene como objetivo recopilar y compilar los datos presentados por los proveedores de acuerdo con los estándares especificados por las acciones administrativas realizadas por la institución. (Congreso de Colombia, 2001).

Estos actos administrativos, obedecen a Circulares y Resoluciones que recopilan las disposiciones emitidas por la Comisión Reguladora de Energía y Gas – CREG – como órgano regulador, así como las diferentes condiciones y modelos de negocio para cada actividad de la cadena energética (energía eléctrica y gas combustible), como necesidades del órgano vigilante.

La tarea destinada a la Superintendencia fomenta sus bases en el documento CONPES 3168 de 2002, donde se establece la Estrategia para la puesta en marcha del Sistema Único de Información de los Servicios Públicos Domiciliarios, en este documento se puede evidenciar la

planificación de la estrategia en el diseño y adopción, cronograma y financiación, al igual que los actores gubernamentales favorecidos con su desarrollo.

El documento CONPES 3168 de 2002, define la estrategia para el desarrollo del Sistema Único de Información de Servicios Públicos Domiciliarios, y sirve de base para el trabajo encomendado a la Superintendencia. El documento describe la preparación para la concepción e implementación de la estrategia, el calendario, la financiación, así como las figuras políticas que apoyaron su creación.

Como desarrollo de esta función desde el año 2010, se realiza el acopio de la información en un sistema centralizado denominado Sistema Único de Información SUI, donde los prestadores reportan información de índole Comercial, Técnica y Financiera, en cumplimiento a los actos administrativos expedidos por este ente de vigilancia.

Por otra parte, a medida que se ha desarrollado la plataforma tecnológica, se han realizado una serie de cambios para implementar diversas metodologías regulatorias, incluidas implementaciones para patrimonio técnico transaccional, pérdidas y distribución, entre otros aspectos. La aplicación de una nueva metodología no es inmediatamente aplicable a todos los proveedores, por lo que se genera una transición de un esquema a otro, definiendo una vigilancia diferencial y complejizando las reglas de carga de información al Sistema Unificado de Información -SUI-. Esto, es uno de los primeros elementos que genera cambios significativos en la plataforma tecnológica de la Superintendencia.

Lo anterior, en un sistema poco flexible a estos cambios genera un ciclo de desarrollo lento y complejo, que consume demasiado tiempo en el análisis de la reingeniería en la arquitectura para el nuevo cargue de información, más aún, cuando se trata de unificar por medio de un acto todas las características propias de cada uno de los vigilados, es claro que desde la Superintendencia se regula de forma única, sin perder de vista para el análisis estas características.

Ahora bien, que se logra el objetivo inicial de almacenar, centralizar y poner a disposición la información certificada por los proveedores en los plazos establecidos en los actos administrativos, los enfoques antes mencionados se dirigen a la recopilación de información, que, en el contexto de la operación actual, parece ser un mal menor.

Sin embargo, es en el momento de procesar y divulgar resultados de la información, donde se encuentran dificultades por los tiempos que se destinan en el proceso de extracción, limpieza y transformación, para en muchas ocasiones no finalizar con un resultado favorable en lo planificado y/o esperado, o peor aún, perder la oportunidad en la posibilidad de tomar medidas correctivas por medio de requerimientos a las empresas que incumplen con las reglas normativas o medidas sancionatorias materializadas en multas, las cuales representan incurrir en gastos a nivel económico y reputacional en el sector.

Analizando los diagnósticos realizados a la plataforma en términos del proceso de divulgación de información, se evidenciaron debilidades en la plataforma tecnológica que limitaban los procesos de intercambio de información para promover la investigación y la publicación de informes a los proveedores de servicios. (Urdaneta, 2019).

Considerando esta circunstancia, existe un deseo creciente de mejorar el paradigma de la vigilancia mediante la valoración de la información y el uso del análisis de datos para mejorar la toma de decisiones informada.

La problemática expresada respecto al procesamiento de información podría reducirse a través de la aplicación de metodologías específicas, tal como lo manifiesta Hernández en su texto: “La aplicación de algunas técnicas de preprocesamiento permite que los algoritmos de aprendizaje sean más eficientes, por ejemplo, al reducir la dimensión, los algoritmos de aprendizaje podrían funcionar más rápido y se mejora su eficiencia”. (Hernández, 2018)

Otro ejemplo de lo expresado es presentado por Aristizábal en la siguiente afirmación: “La transformación de datos en información y la información en conocimiento, no es una tarea trivial. Se requieren ciertas habilidades y conocimientos, los cuales, aunque no son necesariamente complicados, sí platean cierta disposición y competencia. Una mala presentación de información, a pesar de que haya una buena fuente de datos, puede conducir a una mala interpretación o viceversa, lo que puede acarrear problemas al momento de tomar decisiones”. (Aristizábal, 2016)

Otras consideraciones respecto a la inteligencia de negocios evidencian lo siguiente: La inteligencia de negocios se define como la capacidad de una empresa para tomar decisiones. Esto se logra mediante el uso de métodos, aplicaciones y tecnologías que permiten la recolección, limpieza, transformación y aplicación de técnicas de minería de datos analíticos. Rosado. (Parr, 2020).

En conclusión, se deben establecer líneas base, acompañadas de documentación técnica para ser implementadas con propósito en la vigilancia inteligente por medio de la analítica de datos, donde los profesionales dediquen su tiempo al análisis de las alertas generadas por los sistemas implementados y no a generar sus propias líneas bases en el procesamiento de la información.

1.2 Formulación del problema

Con base en la información establecida en las resoluciones SSPD 20155 de 2019 y 12515 de 2021, se requiere establecer un mecanismo que permita por medio de indicadores normativo y de vigilancia, transformar estos datos en información que establezca el comportamiento comercial y técnico de las empresas prestadoras del servicio de energía eléctrica del país.

¿Es posible mejorar el desempeño de los procesos de Inspección, Vigilancia y Control adelantados por la Superintendencia delegada de Energía y Gas Combustible, a través del desarrollo de un modelo de implementación para la estrategia de analítica de datos del SUI?

1.3 Justificación

Si bien, la Superintendencia de Servicios Públicos Domiciliario, ha estado a cargo de administrar, mantener y operar el Sistema único de Información (SUI), su avance tecnológico no ha ido a la par con el de sus supervisados, por tanto, se han requerido esfuerzos significativos en el procesamiento de una gran cantidad de información de una manera estática por sus profesionales. De igual forma, con el transcurrir del tiempo se han realizado varios actos administrativos modificando los diferentes cargues de información, realizando ajustes a nivel de validaciones y reglas de negocio que han sido necesarios implementar. Es así, que esta condición se convierte en otro factor negativo en el procesamiento ágil y oportuno en la toma de decisión.

En consecuencia, es necesario construir la línea base que sirva de guía para el proceso de aprovechamiento de la información capturada, dando especial importancia a que este proceso se realice en forma ágil y oportuna. Se plantea la necesidad de unificar los diferentes esquemas de información que hoy existen en las bases de datos del Sistema Único de Información (SUI), con el fin de dar equilibrio a la brecha que se ha generado entre vigilante y vigilados a través del uso de intensivo de la analítica de datos.

Estas salidas de información deben estar fundamentadas en analítica de datos, la cual se considera la disciplina responsable de examinar e incluso interpretar patrones en los datos para sacar conclusiones. Cuando estas actividades se realizan para respaldar decisiones comerciales o crear valor para la organización, hablamos de "analítica de negocio". Quintero (como se citó en Shmueli, Patel, & Bruc, 2016).

Es por medio de esta analítica de datos, que desde el modelo de vigilancia se puede establecer el perfilamiento de riesgo de las empresas en modelos predictivos con base en la lectura al comportamiento periódico de reporte de información, identificando patrones que permitan establecer conductas no reguladas pero que pongan en riesgo la estabilidad energética del país.

En conclusión, incrementar la automatización de los procesos de analítica de datos permite liberar a los profesionales para que en lugar de realizar procedimientos técnicos complejos para preparar la información, se dediquen a actividad de análisis e identificación de oportunidades para el fortalecimiento de la Superintendencia de Servicios Públicos Domiciliarios.

2. Objetivos

2.1 Objetivo General

Diseñar un modelo de implementación de Analítica de Datos que permita construir un marco general para el desarrollo de proyectos basados en analítica al interior de la Superintendencia Delegada de Energía y Gas Combustible.

2.2 Objetivos Específicos

- Diseñar la arquitectura del modelo de implementación de Analítica de Datos en la SSPD.
- Construcción del modelo documental que soporte la arquitectura propuesta para el modelo de implementación de Analítica de datos en la SSPD inspirado en la metodología CRISP - DM.
- Diseñar la arquitectura tecnológica que de soporte a la operación del proceso de analítica de datos en la Superintendencia Delegada de Energía y Gas Combustible.
- Validar el modelo de implementación propuesto con las dependencias institucionales responsables de la estructuración y control administrativo y operativo de la SSPD.
- Aplicar el modelo propuesto en un caso de uso al interior de la Superintendencia Delegada de Energía y Gas Combustible.

3. Antecedentes

La estructura de la Superintendencia de Servicios Públicos Domiciliarios fue modificada por el Decreto 1369 de 2020 a la luz de la recomendación del documento CONPES 3985 de 2020 para fortalecer la base institucional de los servicios públicos domiciliarios.

En este sentido, en el artículo séptimo del Decreto referido, se transformó la estructura institucional de la superintendencia identificando en ella la Oficina de Administración de Riesgos y Estrategia de Supervisión (OARES). (Presidencia de la República de Colombia [PRC], 2020).

Por su parte, el artículo 1 establece entre otras las siguientes responsabilidades asignadas a OARES:

1. Proponer al Superintendente los lineamientos estratégicos respecto de información; gobierno de los datos; estándares prudenciales y de gestión de riesgos; y prácticas de supervisión

4. Desarrollar los productos de analítica para la Superintendencia y el suministro de información de interés del sector

8. Coordinar el desarrollo de investigaciones, estudios, indicadores y reportes de analítica sobre aspectos financieros, técnicos, administrativos y tarifarios, y análisis de riesgos de los prestadores de servicios públicos domiciliarios.

9. Suministrar los lineamientos para la elaboración de los reportes estadísticos de la Superintendencia.

10. Proponer al Superintendente para su aprobación, las políticas de gobernabilidad de los datos en la Superintendencia, en coordinación con la Oficina Asesora de Planeación e Innovación Institucional.

11. Diseñar la metodología para evaluar la consistencia, homogeneidad y calidad de la información de los prestadores de servicios públicos domiciliarios, recibida o capturada por la Superintendencia (Presidencia de la República de Colombia [PRC], 2020).

Otra área identificada en el artículo séptimo del Decreto 1369 de 2020, es la Oficina de Tecnologías de la Información y las Comunicaciones, a la cual se le asignan entre otras las siguientes funciones:

4. Definir lineamientos tecnológicos para el cumplimiento de estándares y buenas prácticas de seguridad y privacidad de la información y en especial la interoperabilidad de los sistemas que la soportan.

5. Aplicar los lineamientos y procesos de arquitectura tecnológica de la Superintendencia en materia de software, hardware, redes y telecomunicaciones, acorde con los parámetros gubernamentales para su adquisición, operación, soporte especializado y mantenimiento.

8. Administrar, mantener actualizado y operar tecnológicamente, el Sistema Único de Información - SUI de que tratan los artículos 53 de la Ley 142 de 1994 y 14 de la Ley 689 de 2001, de acuerdo con lo definido por las Superintendencias delegadas y sus Direcciones Técnicas de Gestión.

11. Definir la arquitectura de información y datos necesaria para el desarrollo de las funciones de la Superintendencia.

12. Dirigir y orientar el desarrollo de los contenidos y ambientes virtuales requeridos para el cumplimiento de las funciones y objetivos de la Superintendencia.

14. Apoyar los procesos de transformación digital, arquitectura empresarial y continuidad del negocio, en lo referente al componente tecnológico (Presidencia de la República de Colombia [PRC], 2020).

Ante lo anteriormente expuesto, queda evidenciado que la responsabilidad de la creación de la política de datos institucional y la armonización de su futura implementación y su seguimiento es la oficina denominada OARES. Esta dependencia ha venido adelantando análisis sobre diferentes propuestas metodológicas para abordar el proceso de análisis institucional, de estos estudios ha identificado la metodología CRISP-DM como la herramienta más prometedora de cara a consolidar un proceso unificado para la implementación de proyectos de analítica institucional. Sin embargo, pese a los procesos adelantados, la situación institucional establece una serie de desafíos relacionados con el campo de acción institucional y la interoperabilidad con los demás actores involucrados en los procesos de planificación, control y operación de los servicios públicos domiciliarios.

Un diagnóstico institucional realizado a inicios de la década de 2010, evidenció diferentes inconvenientes operacionales entre la SSPD y CREG (Miranda, 2012).

Los inconvenientes se presentan cuando estas dos instituciones (comisiones de regulación y superintendencias) interactúan o materializan sus funciones, ya que muy a pesar de que en el “papel” (ley, decretos, sentencias) esta medianamente claro que es “regulación” y que es “vigilancia y control”, lo cierto es que estos organismos no parecen tener claro cuáles son sus límites y hasta dónde pueden llegar.

Lo anterior, se materializa en las necesidades de cada ente en establecer modelos tanto de regulación como de vigilancia, sin unir esfuerzos para que exista la sinergia en este actuar. Es así como desde el ente vigilante, se deben establecer Resoluciones para los diferentes cargues en obediencia estricta al cumplimiento de una norma del Regulador. En principio, esto no debería ser un inconveniente puesto que, desde sus creaciones, los roles estaban establecidos y claro.

Ya en un plano de ejecución de funciones, se denota las diferencias en las necesidades de contar con información para la vigilancia de los prestadores, es decir, en la regulación se establecen generalidades que, para un modelo de vigilancia basado en el dato, requiere en ocasiones de mucha especificidad, la cual se consigue en actos administrativos que en algunos momentos parecieran regular.

Los actos administrativos creados desde la Superintendencia se materializan en el reporte de información del Sistema Único de Información (SUI), el cual alberga información de todos los servicios Públicos Domiciliarios. Entonces, es gracias al SUI que se realiza la vigilancia a los prestadores, agrupando la información en tópicos técnicos, comerciales, administrativos y financieros, sin dejar de lado el tema de los auditores externos y las Peticiones Quejas y Reclamos. Según (Castro 2020), hay quienes consideran que la Ley 142 de 1994, amerita una nueva regulación, actualizada con las diferentes normas de procedimiento general que han sido emitidas desde 1994, al estimar también que esta Ley aún es restrictiva en cuanto a los postulados del debido proceso y garantías al usuario, entre otros aspectos.

El reporte de información realizado por los prestadores de servicios públicos a la Superintendencia se realiza en diferentes intervalos de tiempo: mensual, trimestral, semestral, anual y por demanda, siendo esta última una declaración expresa del prestador indicando que cuenta con la información correspondiente a un formato o formulario de un periodo especifico y desea reportarlo. Estos periodos constan de una fecha de inicio de reporte y una fecha límite de reporte de la información, que en muchos casos atiende a una disposición regulatoria y/o modelo de negocio.

Para el servicio de energía eléctrica existe una resolución que captura el marco normativo expedido por la CREG, denominado Metodología de Distribución 097, conformada por un total de 27 formatos y formularios y, se denomina Resolución 20102400008055 de 2010.

Una vez expedido el nuevo marco normativo por parte de la Comisión Reguladora, bajo la Resolución CREG 015 de 2018, que hace referencia a la nueva metodología de distribución, se estructuró la Resolución SSPD 20192200020155 del 25 de junio de 2019, conformada por un total de 87 formatos y formularios.

De esta forma se estructura los actos administrativos de la Superservicios, en aplicación a un acto administrativo de la Comisión reguladora. Este ejercicio se plantea para el entendimiento de la aplicación normativa y cadena de valor que finaliza con el reporte de los prestadores y de igual forma, es de aplicación al servicio de gas combustible con sus actos administrativos particulares. Como ejemplo se citan los siguientes:

Para la información correspondiente al esquema de reporte de la Circular Conjunta SSPD - CREG 001 del 2005, y la Circular Conjunta SSPD-CREG No. 20091000000044 de 2009 por la cual se regula el reporte de información comercial básica del sector del gas licuado de petróleo. (SSPD, 2005 - 2009).

Para la información correspondiente al esquema de reporte de información de las Resolución SSPD 20141300040755 de 2014, Circular Conjunta SSPD – CREG 006 de 2003, por la cual se establece el reporte de información de facturación. (SSPD, 2014 -2003)

Los comercializadores mayoristas, transportadores, distribuidores y comercializadores minoristas de GLP, están obligados a reportar la información correspondiente a las actividades que realizan de acuerdo con el marco regulatorio vigente y las respectivas metodologías tarifarias, según se describe en la Circular conjunta SSPD -CREG 002 de 2016, modificada por la 004 de 2016. (SSPD, 2016).

De lo anterior, es evidente la necesidad de crear un modelo de implementación de una estrategia de Analítica de Datos, que permita establecer las condiciones mínimas para dar inicio a cualquier ejercicio de este tipo dentro de la SSPD, para desde allí, procesar toda la información acopiada dentro de las resoluciones realizadas y que sean acordes a las fases del marco del trabajo de la metodología CRISP-DM.

3.1 Antecedentes institucionales

En el desarrollo del ejercicio de investigación institucional, se evidenció el Plan Estratégico De Tecnologías de la Información (PETI) para el período 2021-2022, documento público dispuesto en la página de la Superintendencia de Servicios Públicos Domiciliarios, el cual relaciona los ejercicios que se han realizado a nivel de arquitectura empresarial y transformación digital, en busca de identificar necesidades y fortalecer las iniciativas tecnológicas que han sido recopiladas por la Oficina de Tecnologías de la Información y las Comunicaciones (OTIC), mediante procedimiento interno TI-P-004 denominado Gestión de los Sistemas de Información.

Como resultado de este ejercicio, se pueden identificar ocho macro iniciativas: 1) Fortalecimiento del Sistema de Información Único (SUI), 2) Implementación Política de Gobierno Digital, 3) Modernización de Sistemas de apoyo, 4) Actualización de Infraestructura Tecnológica, 5) Seguridad de la Información, 6) Impulsar la seguridad digital, 7) Explotación de Información y 8) Mejorar la gestión de la Oficina de TIC. (PETI, 2021).

De lo anterior, se puede evidenciar dentro estas ocho macro iniciativas, la asociación a los planes de acción institucionales y proyectos de inversión, agrupados al desarrollo de cada una de las iniciativas y es allí, donde para la macro iniciativa de Explotación de Información, se enmarca el objetivo de Implementar modelos de gestión de información a través de analítica de datos para la toma de decisiones, por medio del Proyecto de Inversión BPIN 2018011000323.

Para lograr este objetivo, desde la Oficina de Tecnologías de la Información y las Comunicaciones (OTIC), se incorporó en uno de los productos estratégicos definido en su plan de acción, un producto del proyecto de Inversión denominado Servicio de información implementado, cuyas actividades son: realizar los procesos necesarios para evidenciar las necesidades y, definir y construir las adecuaciones del SUI, implementar las nuevas soluciones tecnológicas del SUI, establecer los modelos de gestión de información a través de analítica de datos que permitan apoyar la toma de decisiones institucionales.

Como estrategia de implementación para la explotación de información, se han adelantado actividades encaminadas a establecer una estrategia de publicación de la información, para ello se cuenta con la herramienta O3 para los cubos y las bodegas (ETL's manuales) y para los reportes (fábrica de reportes), para la exposición información con la ciudadanía en general.

De igual forma, desde la OTIC se adelanta la actualización de convenios de acuerdo de nivel de servicio para interoperabilidad de información con instituciones como el DANE, Ministerio de Minas y Energía, Unidad de Planeación Minero Energética (UPME), entre otros.

Como estrategia de implementación de ejercicios de analítica, se establecieron lineamientos para el uso de tecnologías tales como: R, Python, ARGIS y la creación de un Data Lake institucional que atienda las necesidades de almacenamiento de los diferentes modelos de datos de los proyectos a realizar por parte de las áreas de la Superintendencia (este ejercicio se encuentra en prueba de concepto).

Todas estas estrategias están encaminadas a implementar dentro de la superintendencia de servicios públicos domiciliarios, una cultura de toma de decisiones basada en datos, a través del inicio de la aplicación de la analítica de datos como herramienta de apoyo a los procesos de vigilancia e inspección.

A partir de la implementación de estos ejercicios, desde la Delegada de Energía y Gas Combustible de la Superintendencia, se está trabajando en la ejecución del proyecto de inversión BPIN 2017011000304 que tiene como objetivo de desarrollar Innovación en el monitoreo de los prestadores de los servicios de energía eléctrica y gas combustible a nivel nacional, para desde allí, construir herramientas que permitan identificar el comportamiento de los prestadores de los servicios de energía y gas combustible, a través de la información reportada en el SUI.

3.2 Antecedentes Académicos externos

Dentro de la producción académica disponible en el contexto colombiano, se evidencia bajo nivel de profundización en temáticas relacionadas con la aplicación de metodologías para la implementación exitosa de procesos de analítica de datos dentro del sector de servicios públicos y las empresas vinculadas a él, dando gran importancia a los procesos de modelado, codificación y análisis de resultados. El análisis de ejercicios similares desarrollados en otros escenarios obtuvo como resultado dos documentos académicos enfocados a la aplicación de la metodología CRISP-DM mediante el enfoque ágil en la analítica de datos, y la aplicación de la metodología referida a un proyecto de minería de datos en el entorno Universitario.

En el primero de ellos, titulado: Modelo Basado en CRISP-DM extendido mediante prácticas de metodologías ágiles para proyectos medianos de analítica de datos, el investigador detalla las tareas a seguir e incluye prácticas de métodos ágiles para el desarrollo de los tableros de control (Mavesoy, 2018).

Es así, como el autor basó su estudio de caso en un diseño experimental comparando dos (2) proyectos del mismo tipo producidos por dos (2) universidades utilizando las técnicas CRISP - DM y CRISP - DM ÁGIL.

Al respecto, Mavesoy concluyó que, la metodología de análisis de datos CRISP-DM es resiliente y brinda diversidad de actividades en todas las fases del proyecto, sin embargo, no asigna responsabilidades particulares a las diferentes actividades.

Respecto a la utilización de la metodología ágil, en este caso CRISP – DM ÁGIL, consideró involucrar activamente a los participantes del proyecto, empleando los principios ágiles de SCRUM y XP, abordando los problemas antes de desplegar el entregable en una atmósfera productiva. Esta aproximación permitió la detección temprana de barreras que se desarrollan durante la construcción del proyecto para corregir o evaluar una solución, lo que resulta en reducción de reprocesamiento y más confianza en los usuarios finales porque están al tanto del progreso del proyecto.

Cabe resaltar que, teniendo en cuenta las bondades ofrecidas por el enfoque ágil, en las recomendaciones, el autor fue puntual en recalcar la importancia de que, en la fase del entendimiento del negocio, la presencia de Scrum Data Architect fue una característica crucial para ayudar a la construcción del Product Backlog.

En un segundo estudio identificado, orientado a la Aplicación de la Metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario, el investigador indica que: la minería trata de extraer la mayor cantidad de información posible de los almacenes de datos, no se conforma con simplemente visualizar los datos a través de consultas simples, sino que trata de obtener resultados en cuanto a la relación que existe entre ellos y cómo se pueden beneficiar el negocio de alguna manera (Galán, 2015).

Adicionalmente, dio a conocer que, a través de la metodología elegida, en este caso CRISP- DM, la ejecución del objetivo general fue precisamente, aplicar estrictamente cada uno de los pasos definidos en la metodología sobre los datos académicos obtenidos en la Universidad tomada como caso de estudio y alojados en los sistemas de información misionales. Adicionalmente, el autor precisó que, el objetivo de la metodología en sí es sacar conclusiones y predicciones lo más fiables posibles a partir de una serie de datos, afirmó enfáticamente que no cumplir los objetivos del negocio no implica necesariamente que no se vayan a cumplir los objetivos del proyecto, ya que el objetivo estará cubierto en cualquier caso siempre y cuando hayamos aplicado completamente la metodología a nuestro problema. (Galán, 2015).

En este sentido, se identificó que, el autor se centró en el enfoque de negocio para los resultados, precisando en su documento que, se había establecido como principal criterio de éxito poder hacer predicciones con un porcentaje de confiabilidad " aceptable “, criterio algo subjetivo, por lo que es inevitable confiar principalmente en los criterios de éxito desde el punto de vista de la extracción de datos que son mucho más específicos y precisos.

Finalmente, el investigador analizó el modelo con base a los resultados de los indicadores adquiridos, de lo cual, se autorizaron los modelos 1 y 2 porque cumplían con los criterios de éxito, pero se descartó el modelo 3, en razón a que no cumplía con los requisitos tanto de negocio como de minería de datos.

3.3 Antecedentes en el uso de métodos predictivos en empresas de servicios públicos domiciliarios

Dentro de la gran variedad de estudios realizados a nivel mundial, cuyo objetivo principal es mejorar la calidad y confiabilidad de la prestación de servicios públicos se toman como referencia para la aplicación, dos estudios específicos que permiten hacer un análisis de la demanda y de la predicción de las interrupciones utilizando el método de Holtz-Winters y redes bayesianas respectivamente.

3.3.1 Método de Holtz-Winters en la predicción de demanda

El modelo de Holtz-Winters desarrollado a finales de la década del 50 como una parametrización de los procesos de suavizamiento exponencial (exponential-smoothing), ha demostrado ser eficientemente aplicado en el análisis de series de tiempo donde se evidencia marcado efecto de la estacionalidad. El modelo puede resumirse como la combinación lineal de diferentes componentes de análisis donde se evidencian: un componente inmutable o no des-

estacionalizado, un componente de tendencia y el comportamiento que describe la estacionalidad, de allí que su uso se haya extendido al análisis de señales y otros fenómenos.

En el estudio adelantado (Tuland, Bello. 2022) aplicado a empresas de energía establece que el análisis de la demanda de energía como insumo para la predicción encaminada al diseño, mantenimiento y confiabilidad del sistema eléctrico es en sí mismo un proceso estocástico que representa un desafío para operadores y entes de control. Tomando como fuente de información la demanda mensual de energía de los usuarios de la empresa CEPALCO, ubicada en Centroamérica y considerando un periodo de análisis de 8 años, los investigadores desarrollan un modelo predictivo para la demanda mensual de energía, partiendo de la aplicación del modelo de Holtz-Winters.

Si bien, el análisis previamente descrito evidenció la existencia de estacionalidad en el consumo de energía eléctrica para países ubicados en la zona tórrida, región a la cual pertenece Colombia, el estudio se limitó a verificar la idoneidad del método para dicho proceso sin explorar en forma individual los componentes resultantes del análisis y sus diferentes ventajas como indicadores claves de la operación de la red eléctrica.

3.3.2 Redes Bayesianas en la predicción de fallos

La investigación realizada (Yue et al, 2017) busca integrar información climatológica a un proceso de predicción de ocurrencia y duración de interrupciones de servicio eléctrico considerando un enfoque novedoso que se aleja de los modelos regresivos tradicionales y los sistemas basados en conjuntos de lógica difusa. La novedad propuesta en este análisis parte del uso de información climatológica georreferenciada consistente con las zonas de influencia de las interrupciones. La información utilizada en el este estudio parte de la identificación de cuatro variables clave dentro de la información extraída de la plataforma NEXRAD. Las variables utilizadas en el estudio provienen del modelo de análisis de señales climatológicas tal como puede verse a continuación:

Tabla 1: Variables del escáner climatológico incluidas en la predicción de interrupciones

Symbol	Long Name	Description	Units
NCZ	Composite reflectivity	Maximum radar reflectivity in each vertical column; describes storm structure and intensity.	dBZ
N0S	Storm relative velocity	Produced by subtracting storm motions from the general wind field; helps identify rotation in storms, which can be particularly damaging.	Knots
N1P	One-hour accumulated precipitation	Hourly accumulated precipitation	Inches
*NVL, DVL	Vertically integrated liquid	Water content of a 2.2 x 2.2 nm column of air; describes storm intensity.	kg m^{-2}

Fuente: Yue et al 2017

Estas variables se utilizan para predecir diferentes niveles de fallos en transformadores y los comportamientos de las solicitudes de atención de contingencias realizados a la empresa. El método utilizado en los predictores corresponde con la regresión logística cuyos resultados serán incorporados a un modelo bayesiano que los agregue considerando criterios espaciales. Como resultado de la investigación se diseñó un modelo bayesiano agregado por áreas donde se identifican zonas de impacto de interrupciones y se evalúa su predictibilidad haciendo uso de una regresión probabilística. El modelo fue puesto a prueba teniendo en cuenta la información climatológica del huracán Irene y a partir del análisis de confiabilidad del modelo se identifican diez grupos de transformadores dentro de la red eléctrica para cada uno de los cuales se cuenta de una confiabilidad de desempeño del modelo. La implementación demuestra que mientras más acotada y reducida es el área de análisis pueden lograrse predicciones más certeras, sin embargo, la complejidad computacional de dicha aproximación hace necesario conciliar la precisión con la velocidad de predicción especialmente en escenarios climáticos cambiantes como el que vivimos ahora.

4. Referente Normativo y Legal

4.1 Normativa Externa

La Constitución Política de Colombia del año 1991 en el artículo 370: otorga al Presidente de la República de Colombia la responsabilidad de ejercer por medio de la Superintendencia de Servicios Públicos Domiciliarios como ente vigilante, el control la inspección y vigilancia de las entidades que lo presten. (Constitución Política de Colombia, 1991).

La Ley 142 de 1994, en el artículo 69 Numeral 2, establece la creación de la Comisión de Regulación de Energía y Gas Combustible como ente regulador, adscrito al Ministerio de Minas y Energía. (Congreso de Colombia, 1994).

4.2 Normativa interna de la Superintendencia de Servicios Públicos Domiciliarios

Para el servicio de energía eléctrica, existe una Resolución que establece el marco normativo expedido por la CREG, denominado Metodología de Distribución 097, conformada por un total de 27 formatos y formularios agrupados bajo la Resolución 20102400008055 de 2010 (SSPD, 2010).

Una vez expedido un nuevo marco normativo por parte de la Comisión Reguladora, bajo el nombre de Resolución CREG 015 de 2018, que hace referencia a la nueva metodología de distribución que toma forma bajo la Resolución SSPD 20192200020155 del 25 de junio de 2019, la cual, está conformada por un total de 87 formatos y formularios (SSPD, 2019).

De esta forma se estructura los actos administrativos de la Superservicios, en aplicación a un acto administrativo de la Comisión reguladora. Este ejercicio se plantea para el entendimiento de la aplicación normativa y cadena de valor que finaliza con el reporte de los prestadores y de igual forma, es de aplicación al servicio de gas combustible con sus actos administrativos particulares. Como ejemplo se citan los siguientes:

1. Para la información correspondiente al esquema de reporte de la Circular Conjunta SSPD - CREG 001 del 2005, y Circular Conjunta SSPD-CREG No. 20091000000044 de 2009 por la cual se reporta la información comercial básica del sector del gas licuado de petróleo. (SSPD, 2005 - 2009)

2. Para la información correspondiente al esquema de reporte de las Resolución SSPD 20141300040755 de 2014, la Circular Conjunta SSPD – CREG 006 de 2003, establece el reporte de información de facturación. (SSPD, 2014 -2003)

3. Para la información correspondiente al esquema de reporte contenido en la Circular Conjunta SSPD – CREG 002 de 2016, modificada por la circular 004 del 2016, establece en que forma los comercializadores mayoristas, los transportadores, los distribuidores y los comercializadores minoristas de GLP, generan el reporte de la información correspondiente a actividades que se desarrollen respecto al marco regulatorio vigente y sus metodologías tarifarias respectivas. (SSPD, 2016)

5. Referente Teórico

5.1 Transformación digital

El rápido crecimiento de los sistemas informáticos da lugar al en la década de los años 80 a un nuevo concepto denominado Transformación digital. Hasta ese momento el alto costo de la tecnología sugería que su apropiación estaba limitada a segmentos del mercado muy específicos. Con la llegada del computador personal y las tendencias posteriores de personalización de la interacción con las fuentes informáticas que aun en nuestros días ponen en evidencia que las transformaciones tecnológicas se están realizando a tal velocidad que su apropiación generalizada no se ha alcanzado cuando una nueva tecnología que reemplace a la anterior entra en funcionamiento.

El caso de la industria que debe enfrentar un mundo tecnológicamente cambiante genera una necesidad de conceptualizar estratégicamente los procesos empresariales de identificación, apropiación y uso de las diferentes tecnologías en con el ánimo de fortalecer los procesos institucionales y elevar la productividad. Por las razones expuestas cada organización debe contar con una estrategia orientada a soportar los procesos de transformación digital, entendiendo esta como el proceso metodológico que contiene etapas estructuradas para garantizar el éxito empresarial. Todo proceso de transformación digital debe apuntar a dar solución a una problemática institucional.

Con la identificación de la problemática se hacen evidentes los objetivos del proceso de transformación digital integrando tres dimensiones estratégicas para todo negocio: la generación de rentabilidad, el fortalecimiento de los procesos institucionales y el aprovechamiento del conocimiento e información generados por la organización. Bajo estas consideraciones: “Todo proceso de transformación digital requiere un ejercicio consolidado de gestión del conocimiento como fuente de la cual se nutren las futuras actividades estratégicas relacionadas con el proceso” (Trejo, 2021)

Ilustración 1: Construcción de los objetivos de un proceso de transformación digital

Fuente: (Trejo, 2021)

A grandes rasgos los componentes principales del proceso de transformación digital son:

ANÁLISIS CONTEXTUAL: Diagnóstico de la realidad institucional, considerando el contexto productivo, y los diferentes involucrados que participan en cada una de las etapas de negocio.

DEFINICIÓN DE OBJETIVOS: Es el diseño estratégico de metas que aparecen como respuesta a las problemáticas identificadas en el análisis contextual.

DISEÑO DE FUTUROS: Es un proceso prospectivo que tiene en cuenta los objetivos definidos sin realizar juicios de valor en términos de confiabilidad y/o factibilidad. En este proceso se permite plantear escenarios idealizados sin ningún tipo de restricciones.

VALORACIÓN DE LOS OBJETIVOS: En esta etapa se consideran las diferentes restricciones del negocio, normativas, legales, financieras, humanas, entre otras. Como resultado de la inclusión de dichas restricciones se evalúa si los futuros diseñados son alcanzables y en qué medida posibles. Si el resultado del proceso de valoración es negativo deberá iterarse sobre los objetivos y escenarios de futuro diseñados.

IMPLEMENTACIÓN DEL PLAN: Con la valoración realizada se hace necesario definir un proceso ordenado de ejecución que parte del proceso empresarial, las personas involucradas (usuario, cliente, empleado, socio, ciudadano, entre otros), la tecnología o tecnologías apropiadas y la gestión del conocimiento. (Trejo 2021)

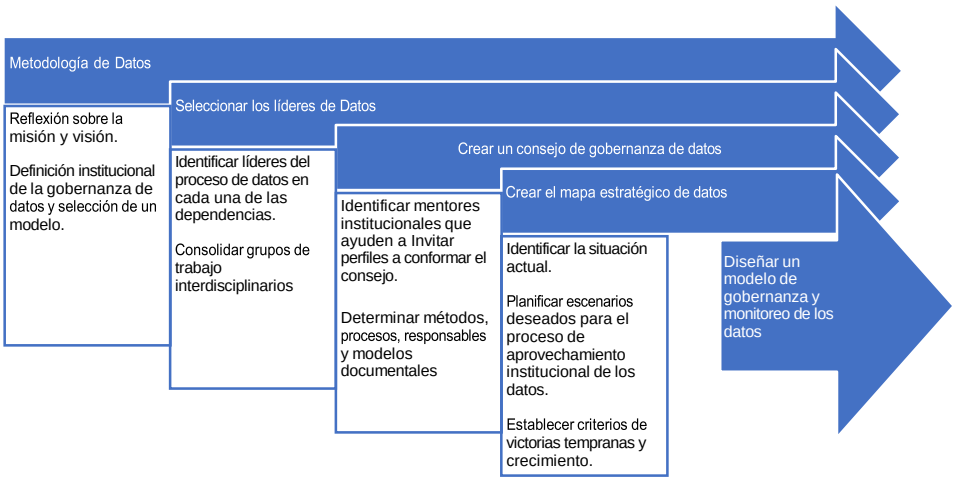
Como puede verse los procesos de transformación digital van mucho más allá de la selección de una tecnología y su apropiación dentro de un proceso productivo. El resultado de estas actividades da lugar a nueva información en forma de datos, indicadores y conocimiento en general, en esta etapa se hace evidente la necesidad de gestionar y administrar toda la nueva información generada, dando origen a un proceso posterior de transformación digital para dar uso eficiente de dichos datos. Una vez una compañía inicia una transición hacia la transformación digital este proceso se realizará indefinidamente a lo largo de todo su ciclo de vida.

5.2 Data-Driven Organizations

El concepto de Data-Driven va más allá de su denominación por directa traducción, una organización Data-Driven no solo utiliza datos para soportar cada una de sus decisiones en los diferentes procesos que componen su cadena de valor, en realidad una organización Data-driven considera sus datos como la mayor fuente de riqueza y ha construido en torno a esta visión toda una cultura organizacional. (Maffeo, 2023)

Existen dentro de la teoría de las organizaciones, diferentes aproximaciones al proceso de transformación institucional. Considerando la última versión propuesta por Maffeo, la versión sintetizada de una metodología para el desarrollo de organizaciones Data-driven está formada por la estructura de seis componentes propuesta a continuación.

Ilustración 2: Modelo de implementación Data-Driven



Fuente: Maffeo 2021.

5.3 La metodología CRISP-DM

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining DM) es considerada una de las guías más utilizadas para el desarrollo de proyectos de minería de datos en la actualidad, y está conformada por seis etapas, a saber: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación e implementación. Por su parte, la minería de datos o data mining es el proceso de extraer información útil, comprensible y nueva a partir de grandes cantidades de datos, tiene por objetivo principal encontrar información oculta o indirecta que no se puede obtener mediante métodos estadísticos convencionales. Con el paso de los años el uso de la metodología CRISP-DM ha probado ser útil en otras formas de análisis de información y se ha generalizado para su aplicación en procesos complejos de inteligencia artificial.

La metodología plantea que la secuencia de pasos puede no ser rígida y define un conjunto de tareas y actividades para cada fase de un proyecto, pero no establece cómo llevarlas a cabo, por lo que se considera un aspecto positivo para el equipo de analítica o científico de datos, dado que puede incorporar su criterio en el desarrollo de las mismas (Moine et al., 2011). A continuación, se presentan de manera general las fases definidas:

- Compresión del Negocio: Esta fase se enfoca a la necesidad del científico de datos de dedicar tiempo representativo para investigar las expectativas que la empresa tiene respecto al proceso de análisis de datos. Lo anterior se considera un momento estratégico para familiarizarse con los datos de la empresa, de manera que se logre que todos los involucrados en el aprovechamiento de la información estén de acuerdo antes de compartir los recursos necesarios. Uno de los objetivos principales de esta fase es la de recolectar información de las ambiciones de la entidad y las necesidades que dan origen a la implementación de procesos de análisis de información. Lo anterior, permite tener muy claros los objetivos de la entidad y los aspectos necesarios a tener en cuenta. Asimismo, se definen los objetivos técnicos del proceso de análisis, se establece el plan de proyecto y se efectúa una conceptualización básica sobre la compresión de los datos, es decir la forma de acceder a éstos y explorarlos.

- Comprensión de los datos: Suele ser la fase más larga del proyecto, dado que se deben estudiar los datos disponibles dentro de la organización y alinearlos con el proyecto de análisis a desarrollar. Es importante tener presente el origen de los datos, determinar la estructura de los datos existentes, los datos adquiridos y los datos adicionales. Es allí donde resulta importante hacer uso de preguntas directoras, tal como lo sugiere el manual de IBM : 1) ¿qué columnas de la base de datos no parecen relevantes para el análisis y se pueden excluir?, 2) ¿Se cuenta con datos suficientes para realizar conclusiones generales del tema a abordar y/o predicciones acertadas? y 3) ¿Se tiene claridad respecto a la forma de gestionar los valores que se consideren perdidos en cada origen de los datos analizados? (IBM, 2021)

Respecto a la descripción de los datos, que se encuentra incorporada en esta fase, se considera pertinente tener en cuenta la cantidad y la calidad de los datos. En este sentido, la Corporación IBM establece como premisa los siguientes elementos: 1) Cantidad de los datos (lo cual impacta el procesamiento), 2) Tipos de valores: como lo son numéricos, categóricos (cadena), booleano (verdadero/falso), lo cual es importante a fin de no tener posibles situaciones problemáticas en la fase del modelado, y 3) Esquema de codificación:

Para describir este aspecto, la Corporación IBM ejemplariza un caso de la siguiente forma: “un conjunto de datos puede utilizar H y M para representar hombre y mujer, mientras que otro puede utilizar los valores numéricos 1 y 2.”, y finalmente respecto a la fase de exploración de los datos, se determina como relevante el apoyo de las gráficas y demás herramientas de visualización para su desarrollo. (IBM, 2021)

- Preparación de los datos: En esta fase se realiza la preparación de los datos lo cual, según la literatura relacionada, se lleva en tiempo entre un 50 y 70% del proyecto. Así, tal como lo indica la Corporación IBM en su sitio web, conforme a la entidad se tienen en cuenta las siguientes tareas: 1) Combinación de datos y/o registros, 2) Selección de una muestra de un subconjunto de datos, 3) Agregación de registros, 4) Creación de nuevos atributos, 5) Clasificación de datos para modelado, 6) Eliminación o sustitución de valores vacíos o faltantes y, 7) División de secciones de datos de prueba y entrenamiento. (IBM ,2021)

- Modelado: Corresponde a la incorporación del trabajo realizado en las anteriores fases y a la utilización de las técnicas y/o herramientas analíticas seleccionadas, a través de los parámetros predeterminados. De acuerdo con la consulta teórica efectuada, se consideró pertinente mencionar lo definido por la Corporación IBM, respecto al proceso de generación de los modelos por parte del científico de datos, en donde según lo indicado se dispondrá de tres tipos de información que podrán ser utilizados en la toma de decisiones por parte de las partes interesadas: 1). Configuración de parámetros, 2) Los modelos reales producidos y 3) Las descripciones de los resultados de los modelos (Problemas de los datos y el rendimiento). (IBM,2021)

- Evaluación: El propósito de esta fase es que se logre determinar si los modelos técnicamente se pueden considerar correctos y efectivos, teniendo en cuenta los criterios de éxito definidos en la etapa inicial del proceso de minería de datos. Así, de acuerdo a lo expuesto por IBM, la minería de datos puede dar como resultado lo siguiente:

1) La materialización de las fases anteriores establecidas en la metodología CRISP-DM y 2) “las conclusiones o interferencias obtenidas de los modelos y del proceso de minería de datos, que recibe el nombre de descubrimientos”. (IBM, 2021)

- Despliegue: En esta fase se implementa el modelo en la entidad a fin de obtener mejoras en la situación inicial. En este punto, se logra obtener beneficio por parte de la entidad tanto en la aplicación de la metodología dada por la minería de datos como por nuevos conocimientos adquiridos y tomados como oportunidades para ser aplicados. Lo anterior, es medido tanto en el aporte a la planificación del proceso u otros procesos a analizar como por la toma de decisiones por las partes interesadas. Es importante resaltar que, esta fase incluye como actividades la planificación y el control del despliegue, y la producción de un informe final y la revisión del proyecto enfocado a la mejora continua. (IBM, 2021)

5.4 Inteligencia de Negocios

Tal como lo indica Lund et al. (2021), la inteligencia de negocios se refiere de manera general a las metodologías, procesos y tecnologías que hacen posible la generación de evidencias para sustentar la toma de decisiones a fin de alcanzar objetivos y metas. Los citados autores mencionan tres elementos necesarios para la BI, estos son el dato, información y la información sintetizada, analizada e interpretada, a lo cual agregan:

Dato es la representación simbólica, no tiene contenido semántico; mientras que la información refiere a un conjunto de datos procesados que tiene un significado; a su vez la información sintetizada, analizada e interpretada da origen al conocimiento. Estos dos últimos constituyen las piezas claves para una acertada toma de decisiones (Lund et al, 2021, p.2)

Cabe resaltar que, BI permite analizar los datos actuales, realizar proyecciones con base en datos históricos y proporcionar a las partes interesadas la información adecuada para asegurar la mejor ruta para la acertada toma de decisiones.

Por lo anterior, los referidos autores consideran un DataWareHouse como: mucho más que un repositorio o una base de datos organizacional, involucra un conjunto de tecnologías y procesos articulados para lograr el objetivo final, que es asistir a la toma de decisiones. (Lund et al, 2021)

De otra parte, las autoras del documento Inteligencia de Negocios aplicada a los procesos de autoevaluación de la Universidad de Manizales, reconocen como principales ventajas de BI:

Permitir integrar datos de diferentes fuentes o áreas de la empresa, y acceder a esta información a través de un formato único.

Aportar la información basada en tiempo y en hechos reales, distribuyéndola en toda la organización y para los diferentes actores de esta.

Las herramientas ofrecidas por Business Intelligence, permiten una fácil y rápida interacción con los usuarios, además de mostrar la información a gran velocidad.

Permitir que la empresa tenga un continuo seguimiento de los procesos, para tener mejores y acordes visiones de la empresa a largo plazo. (Arenas y Gómez, 2017)

5.5 Herramientas y técnicas para la Inteligencia de Negocios

Las herramientas y técnicas para la inteligencia de negocios, se enfocan de manera específica en un aporte significativo en a la toma de decisiones y a dar a entender los procesos a las partes interesadas.

De acuerdo a lo indicado por los investigadores en el documento académico Estrategias para fomentar el emprendimiento y desarrollo empresarial, se define que el instrumento

conocido como inteligencia de negocios, abarca secuencias de habilidades y recursos destinados a administrar la importante base de datos recopilada por las empresas, produciendo estudios que emplean y aprovechan al máximo los datos obtenidos, efectuando un mayor uso y beneficio de los datos recopilados. De igual forma, lograr proporcionar a las personas claves de la organización, una comprensión completa del curso de la misma, es uno de sus muchos beneficios. (Tapia, et al. 2020).

Lo anterior, se complementa de manera clara, con la capacidad de entender los datos a través de la visualización que pueda permitir analizar y evaluar a quien los observe.

Cabe resaltar lo mencionado por los autores del documento Inteligencia de Negocios para las Organizaciones, en donde se resalta que las empresas realmente reducen el margen de error al utilizar la inteligencia de negocios como su principal instrumento de investigación, dado que permite minimizar el margen de error en la toma de decisiones, aprovechar al máximo los recursos internos y dar información de manera efectiva para contextualizar los modelos de gestión de la organización. (Cevallos et al., 2021)

Finalmente, respecto a la reporteo y análisis, los autores referidos en el párrafo anterior, indican que, con el uso de datos recopilados, el Power BI para inteligencia de negocios (Herramienta de visualización utilizada) proporciona visualizaciones históricas, actuales y predictivas de las actividades comerciales de una entidad, en un estilo que es fácil de entender visualmente, lo cual puede ser utilizado para para crear informes, realizar análisis y extraer datos. (Cevallos et al., 2021)

5.6 El proceso de análisis descriptivo de información

El análisis descriptivo de información está compuesto por los procesos de análisis de información derivados de la estadística descriptiva clásica considerando distribuciones de población, medidas de comportamiento central, histogramas, segmentación en cuartiles, y en general las diferentes herramientas de la estadística tradicional. El objetivo de aplicar procesos de análisis descriptivo tiene tres objetivos principales: la identificación de valores por fuera de

rango, la segmentación y categorización de la información y los análisis de distribución probabilística que permitan comprender la naturaleza de los fenómenos que se analizan.

5.7 Data visualization

El proceso de visualización de datos es una particularización de los procesos comunicativos que hacen uso de representaciones gráficas para evidenciar resultados propios del proceso de análisis de información. Contrario a lo comúnmente se cree la visualización de datos tiene más elementos de la teoría de la comunicación y del arte que de complicados procesos de análisis de datos, de allí que sea tan valioso dentro de los ejercicios de analítica de datos.

5.8 Análisis geoestadísticos

El análisis geoestadístico corresponde a la integración tecnológica de los sistemas de bases de datos relaciones y los sistemas de información geográfica en el proceso de aplicación de principios del análisis de información. A grandes rasgos la geoestadística es la integración de análisis de información compleja que busca una correspondencia en unidades geográficas: puntos, líneas y polígonos; cada una de estas estructuras geográficas y sus correspondientes indicadores asignados se transforman en modelos de visualización espacial, útiles en la comunicación de datos relacionados con dimensiones humanas y sociales, entre otros.

5.9 Modelos de Predicción y Simulación Predictiva

El análisis predictivo es un tipo de análisis de datos que se enfoca en realizar proyecciones sobre el comportamiento futuro del fenómeno o problema investigado, por ejemplo, qué podría pasar con los indicadores de gestión del sector público, calidad y desempeño. De esta forma podemos anticiparnos a posibles escenarios favorables o desfavorables y tomar decisiones en consecuencia (MINTIC et al., 2022)

Al respecto, se indica que en un análisis predictivo de datos se obtiene información histórica del tema a predecir y se utiliza machine learning, así como algoritmos matemáticos

para detectar de manera específica posibles relaciones entre algunas variables, a fin de que logre predecir diferentes escenarios a futuro.

En este sentido, la simulación predictiva, es una herramienta que utiliza la información recopilada del tema sobre el comportamiento en tiempo real del tema objeto de análisis para lograr plantear escenarios hipotéticos, en los cuales pudiera llegar a ser afectado de manera negativa o positiva la situación analizada y, tiene como fin la mejora de la gestión ya que tiene como aspecto relevante anticiparse a la ocurrencia de los hechos y así lograr tener planeado un control que minimice los posibles impactos y/o consecuencias.

Es pertinente mencionar que, para lograr realizar simulaciones predictivas cuyos resultados sirvan de apoyo para el mejoramiento de la gestión, es imperativo que el modelo virtual cuente con adquisición de datos en tiempo real o repositorios de información histórica sobre el fenómeno a analizar. Este conjunto de datos iniciales se denomina el escenario inicial o base para comparar de manera acertada el comportamiento que se tenía antes de la simulación del tema objeto de análisis. (Salvador, 2020).

En conclusión, este tipo de análisis se enfoca en lograr que con los resultados obtenidos se pueda definir acciones preventivas, estableciendo estrategias para evitar situaciones de riesgo sino también para que las partes interesadas puedan tomar decisiones o influir en ellas.

5.10 Método Holt-Winters

El método Holt-Winters es un método de predicción perteneciente a los modelos de descomposición exponencial. En este método se obtienen resultados segmentados en cuatro componentes principales: tendencia, el factor cíclico, la estacionalidad, y el componente irregular, los cuales al integrarse aditivamente replican la información inicial. Es posible identificar si las fluctuaciones en los valores de las series de tiempo varían en forma independiente o integrada con la tendencia con el fin de establecer criterios de análisis complementarios que sean de utilidad a los tomadores de decisiones.

Cuando una serie sigue un esquema multiplicativo y presenta estacionalidad, es el método estacional. Una vez desestacionalizada la serie podremos realizar predicciones para periodos futuros. En las series temporales que siguen una tendencia aproximadamente lineal, y además están sometidas a la incidencia del factor estacional, el método de predicción más adecuado resulta ser el método de Holt-Winter. (Guerrero, 2006).de la razón a la media móvil el más apropiado, por su consistencia y uso, para eliminar el factor

5.11 Redes Bayesianas

Una red bayesiana es un modelo de probabilidad estático que representa un conjunto de variables aleatorias y sus dependencias condicionales mediante análisis dirigido. (DAFP, 2021)

Una Red Bayesiana es un modelo gráfico probabilístico, un gráfico acíclico dirigido que representa un conjunto de variables (nodos) y su independencia probabilística. Los nodos pueden representar cualquier variable: un parámetro medido como Ra, una variable latente o una hipótesis. Existen algoritmos eficientes que realizan inferencia y aprendizaje en Redes Bayesianas. Un gráfico acíclico dirigido es Red Bayesiana con respecto a un conjunto de variables si el conjunto de distribuciones de probabilidad de variables nodales se puede escribir como el producto de las distribuciones locales de cada nodo y sus padres. (Correa, 2008)

5.12 Simulación de MonteCarlo

Conforme a lo indicado por Zapata y demás autores en su artículo, el método de simulación de Montecarlo y Estudios de Confiabilidad de Sistemas de Distribución de Energía Eléctrica, la Simulación de Montecarlo Se basa en la generación de números aleatorios y la finalidad del procedimiento es simular el comportamiento aleatorio del sistema para obtener artificialmente los índices de fiabilidad de los puntos de carga. (Zapata et al, 2004)

De igual forma, los autores indican que es el método más versátil dado que:

1). Permite que cualquier distribución modele la salida y el retorno de los componentes, 2) Permite resolver sistemas que no tienen solución analítica, 3) Es posible obtener las distribuciones de probabilidad de los índices de confiabilidad del punto de carga lo cual es muy útil para estimar el riesgo de ocurrencia de diferentes valores del índice, y 4) Los cambios del sistema se realizan en la base de datos obtenida sin necesidad de realizar ningún cambio en el software.

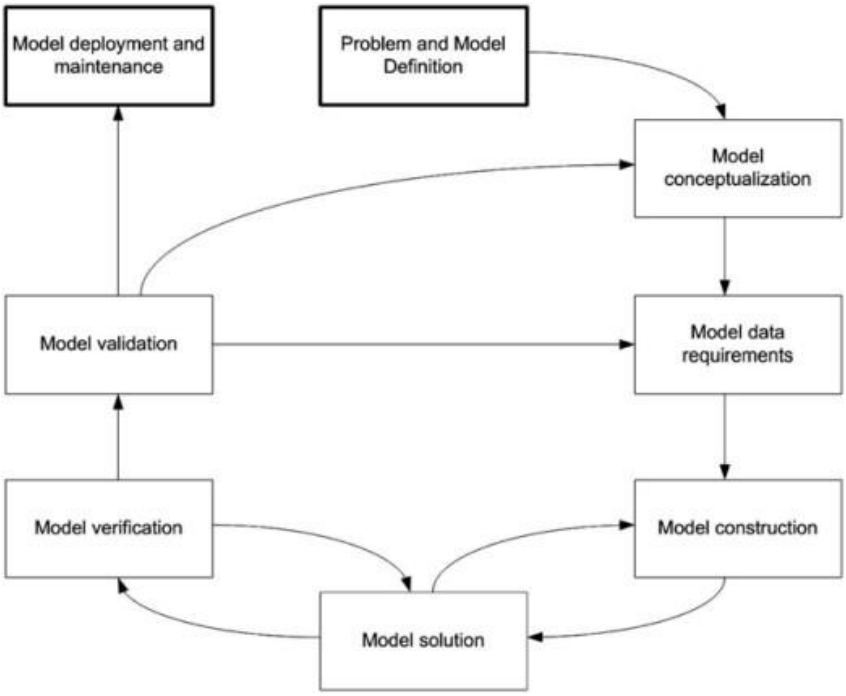
La desventaja definida por los autores referidos es que este método requiere una gran cantidad de tiempo de computación. Esto se debe en parte a que, la cadena de distribución primaria tiene una gran cantidad de componentes y puntos de carga para los cuales el software debe procesar una gran cantidad de datos. (Zapata et al, 2004)

5.13 Diseño de modelos

En términos generales Modelar es el proceso humano de representar la realidad percibida o imaginada con ayuda de estructuras simples que permitan la síntesis del conocimiento, la replicabilidad y la comunicación eficiente de las ideas. El acto humano de modelar parte siempre de tres elementos conductores: la identificación de un sistema, el propósito y la forma de representación de un objetivo específico, partiendo por lo general de una problemática a resolver. (Cameron, Giani. 2021)

La técnica de modelado parte de la necesidad planteada se desarrolla a través de un ejercicio ciclico con realimentaciones paulatinas siguiendo el modelo propuesto en la siguiente imagen:

Ilustración 3: Método para el diseño de modelos



Fuente: (Cameron, Giani. 2021)

Las etapas y bucles de realimentación están diseñadas para permitir el fortalecimiento de la conceptualización y representación del conocimiento y volver constantemente sobre la conceptualización del modelo, pues la representación del conocimiento es mutable en función de las dinámicas del fenómeno que se analiza y las necesidades que lo motivan.

Teniendo en cuenta la producción teórica de Cameron y Giani, conceptualizar un modelo es una actividad altamente compleja que parte de un conjunto de interrogantes clave para representar la fenomenología que se desea representar, comprender, optimizar o simular. Las preguntas clave más utilizadas en el proceso de modelamiento son:

¿Cuál es el Sistema estudiado y qué tipo de restricciones o límites operan sobre él?. Cuando se trata de procesos administrativos o estratégicos el sistema se refiere al contexto organizacional y las restricciones los marcos normativo, administrativo y estratégico.

¿Es posible identificar subsistemas dentro del gran sistema estudiado, estos subsistemas tienen reglas especificas y han definido criterios de interacción con otros

subsistemas?

¿Cuáles son las condiciones iniciales del sistema?, considerando su estructura, entradas y salidas, esta situación inicial es determinante en la conceptualización que viene desarrollándose.

Después de aplicar las preguntas claves, el diseño del modelo se enfrenta a un conjunto de criterios complementarios para poder dar forma a la representación del conocimiento:

Permitir la “tormenta de ideas”, dando la posibilidad de contemplar posibilidades que están por fuera de los criterios de factibilidad.

Los modelos preliminares en la tormenta de ideas deben pasar por el principio de parsimonia o navaja de Occam: “La mejor solución a cualquier problema suele ser la más sencilla”.

Surtidas las etapas descritas se realiza la transición a la definición los bloques conceptuales del modelo que son en su orden:

Entornos: subsistemas que hacen parte de la estructura inicial del sistema donde desea implementarse el modelo.

Instancias: Categorías que agrupan las diferentes acciones y procedimientos que componen el modelo, pueden representarse con la representación de caja negra haciendo claridad sobre las entradas y salidas.

Actividades independientes: Son acciones que no pertenecen a ninguna instancia en específico, o que por el contrario pretenden conectar dos instancias del modelo.

Flujos: modelo de representación del conocimiento llevado por lo general a representación de línea que pone en evidencia la forma como las entradas y salidas de las diferentes instancias interactúan y se articulan con los entornos.

6. Metodología

6.1 Enfoque metodológico

El proyecto de investigación tiene un enfoque metodológico mixto desarrollado en tres etapas. En la primera etapa se realiza una investigación descriptiva orientada a la identificación de la base teórica conceptual que permita definir una propuesta de modelo de implementación armónica con procesos institucionales y soportada en estructuras documentales concretas. En la segunda etapa se realiza un análisis cualitativo de la pertinencia del modelo propuesto con la dinámica organizacional de la institución motivo de estudio y una tercera etapa donde se realiza un análisis cuantitativo de información institucional para identificar patrones de comportamiento.

La complejidad del proyecto de investigación y la diversidad de temáticas se armonizan en función de su aplicabilidad concreta dentro de la institución motivo de estudio para dar un norte a la determinación de los tipos de estudio y la estructura de ejecución requerida.

6.2 Tipo de estudio

Para cada uno de los enfoques de investigación incluidos en este proyecto y que determinan las etapas del proceso de investigación se establecen los diferentes tipos de estudio a aplicar:

Fase 1: Estudio descriptivo que parte de la recopilación de documentación y estudios de caso sobre procesos de gestión de proyectos de analítica de datos para realizar una propuesta metodológica afín a la estructura organizacional de la SSPD. En este estudio se identificarán propuestas documentales armónicas con los procesos de implementación ágil de dinámicas empresariales basados por lo general en modelos soportados en la experiencia.

Fase 2: Estudio cualitativo donde se recopilan las percepciones, visiones y realimentaciones del modelo propuesto y construido en la fase 1. Las dependencias incluidas

en este análisis serán las directamente responsables dentro del gobierno institucional para dar trámite y gestión a los procesos de analítica. Resultado de esta interacción directa se dará paso a un análisis de pertinencia y afinidad de la propuesta.

Fase 3: Estudio cuantitativo sobre fuentes de datos disponibles en el SUI, donde se identifiquen patrones de comportamiento de las diferentes empresas prestadoras del servicio de energía eléctrica. Los estudios incluirán elementos de la estadística descriptiva clásica, modelos regresivos basados en estacionalidades y modelos predictivos estocásticos.

7. Resultados

7.1 Modelo de implementación de proyectos de analítica propuesto a la Superintendencia delegada de Energía y Gas Combustible

Resultado Objetivo específico número 1:

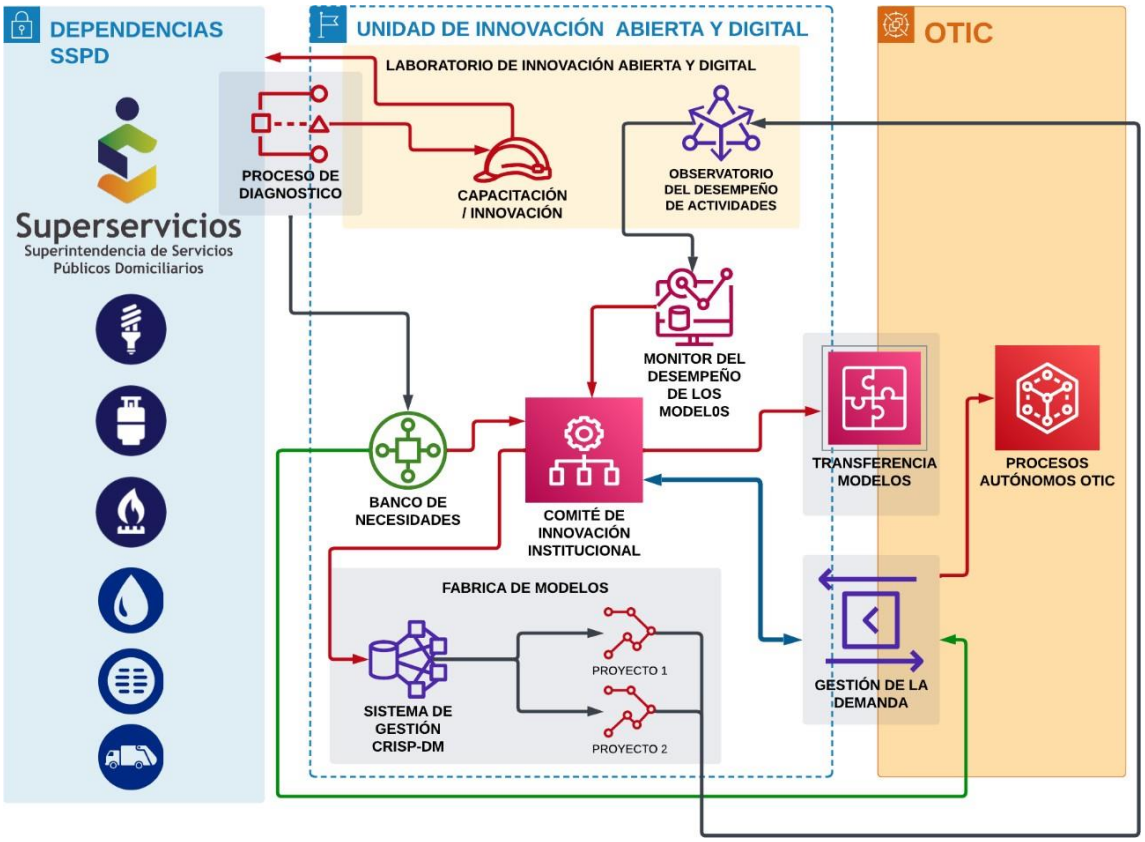
Metodología de Investigación: Enfoque metodológico mixto desarrollado en tres etapas. En la primera etapa se realiza una investigación descriptiva orientada a la identificación de la base teórica conceptual que permita definir una propuesta de modelo de implementación armónica con procesos institucionales y soportada en estructuras documentales concretas. En la segunda etapa se realiza un análisis cualitativo de la pertinencia del modelo propuesto con la dinámica organizacional de la institución motivo de estudio y una tercera etapa donde se realiza un análisis cuantitativo de información institucional para identificar patrones de comportamiento.

El desarrollo metodológico del presente Modelo fue abordado en el Anexo 2, en donde se evidencia el uso de la metodología propuesta en el Modelo, alineado con la metodología de minería de datos (CRISP-DM).

Considerando los diferentes criterios del modelamiento de instituciones data-driven se diseña un modelo que incluye tres componentes fundamentales: Entornos, instancias y

actividades. Los entornos describen cada una de las dependencias institucionales involucradas directamente en los procesos de análisis, uso y gobernanza de los datos. Las instancias son nuevas unidades administrativas del negocio de duración temporal o duradera que articulan la comunicación entre instancias y cumplen funciones estratégicas dentro del proceso de análisis, uso y gobernanza de los datos. Las actividades corresponden con cada uno de los procesos realizados directamente por las instancias y que dan como resultados los entregables que alimentan futuras etapas del proceso de análisis, uso y gobernanza de los datos.

Ilustración 4: Modelo SSPD Data-Driven



Fuente: Elaboración propia

Dentro del modelo planteado se identifican claramente tres entornos: La Oficina de Tecnologías de Información y las Comunicaciones (OTIC), las diferentes dependencias administrativas y operativas de componen la SSPD y la Unidad de Innovación Abierta y Digital (UIAD). De los entornos descritos solo la UIAD no existe explícitamente al interior de la organización y hace parte de las recomendaciones de reestructuración administrativa que

deben ser socializadas y discutidas con la SSPD.

Las instancias diseñadas para hacer posible una SSPD Data-Driven son respectivamente:

El laboratorio de innovación abierta y digital: instancia que debe contar con un espacio físico e infraestructura tecnológica para soportar las actividades de capacitación, análisis del desempeño de los modelos e involucrados en el proceso de análisis de datos y diagnóstico de necesidades de las diferentes dependencias.

Fábrica de modelos de analítica: instancia donde se desarrollan todos los modelos de análisis requeridos para dar respuesta a las necesidades institucionales identificadas.

Banco de proyectos: Instancia donde se priorizan y analizan las diferentes necesidades identificadas en la actividad de diagnóstico, haciendo uso para ello de un proceso metodológico de valoración de la data y la viabilidad tecnológica.

Comité de innovación institucional: instancia donde se evalúa el banco de proyectos para determinar en forma óptima la priorización de proyectos y la asignación de recursos en las etapas posteriores.

Adicional a las actividades propias de cada una de las instancias diseñadas existen actividades que actúan como interfaces entre diferentes instancias y procesos tales como:

Monitor del desempeño de modelos: que evalúa el uso de los diferentes modelos y herramientas de analítica puestos en servicio para identificar su importancia, nivel de apropiación y necesidades de intervención. El reporte de esta información estará consignado en una plataforma automatizada que permitirá su evaluación por parte del comité de innovación institucional.

Gestión de la demanda: Actividad que evalúa la capacidad técnica y disponibilidad tecnológica y humana con la cual se cuenta para dar paso a los ejercicios de puesta en producción y despliegue. Esta actividad permite el dialogo entre el comité de innovación institucional y las actividades realizadas por OTIC para el despliegue y transferencia de

modelos.

Como puede verse el modelo propuesto incluye las diferentes etapas del modelo CRISP-DM y las materializa a través de una estructura administrativa y operativa que se alinea con los principios de la transformación digital y los conceptos de desarrollo ágil. A partir de este modelamiento se abre paso a un dialogo con las diferentes dependencias responsables de la planificación estratégica de la SSPD con el fin de armonizar el resultado.

Existen una correspondencia directa entre cada una de las etapas del modelo CRISP-DM y el modelo de implementación para proyectos de analítica propuesto que se evidencia en la siguiente tabla donde se describen las acciones concretas que unen ambas construcciones:

Tabla 2: Correspondencia entre el modelo institucional y CRISP-DM

	LABORATORIO DE INNOVACION	BANCO DE NECESIDADES	COMITÉ DE INNOVACIÓN	FABRICA DE MODELOS	MONITOR DEL DESEMPEÑO	TRANSFERENCIA DE MODELOS
COMPRENSIÓN DEL NEGOCIO	Diagnóstico y capacitación	Informe de Diagnóstico	Informe de diagnóstico			
COMPRESIÓN DE LOS DATOS		Análisis de estructura y calidad de los datos	Informe de calidad de los datos			
PREPARACIÓN DE LOS DATOS		Extracción, transformación y carga de los datos	Informe de preparación de los datos			
MODELADO			Priorización de proyectos de analítica de datos	Procesos de Modelado, visualización y desarrollo	Sandbox del producto de analítica	
EVALUACIÓN					Proceso de valoración de modelos	
DESPLIEGUE			Priorización del despliegue			Despliegue a producción

7.2 Modelo documental para procesos de Analítica Institucional inspirado en la metodología CRISP-DM.

Resultado Objetivo específico número 2

Metodología de Investigación: Enfoque metodológico mixto desarrollado en tres etapas. En la primera etapa se realiza una investigación descriptiva orientada a la identificación de la base teórica conceptual que permita definir una propuesta de modelo de implementación armónica con procesos institucionales y soportada en estructuras documentales concretas. En la segunda etapa se realiza un análisis cualitativo de la pertinencia del modelo propuesto con la dinámica organizacional de la institución motivo de estudio y una tercera etapa donde se realiza un análisis cuantitativo de información institucional para identificar patrones de comportamiento.

El desarrollo metodológico del presente Modelo fue abordado en el Anexo 2, en donde se evidencia el uso de la metodología propuesta en el Modelo, alineado con la metodología de minería de datos (CRISP-DM).

A partir del intercambio de conocimientos derivado del modelo propuesto se realiza una construcción de las estructuras documentales que darán soporte a su implementación sirviendo de guía institucional a los proyectos institucionales. En primer lugar, se realiza una adaptación de componentes CRISP-DM con la realidad institucional y la visión estratégica liderada por OARES, de tal forma que las estructuras documentales responsables den fe del conocimiento instruccional y realicen su transición hacia formatos documentales adoptados legalmente dentro de los procesos administrativos de la organización. El resumen de estas actividades es mostrado en la siguiente imagen:

Ilustración 5: Componentes CRISP-DM apropiados por la SSPD

FASE	PASOS	TAREAS	HERRAMIENTAS DE APOYO DOCUMENTAL
FASE 1. COMPRENSIÓN DEL NEGOCIO	Metas del negocio	Contexto	Esquema diagnóstico Técnico contextual
		Objetivos del Negocio	
		Criterios de éxito del negocio	
	Diagnóstico de la situación	Inventario de recursos	Esquema de pre-factibilidad técnica institucional
		Requisitos, supuestos y restricciones	
		Riesgos y contingencias	
		Terminología	Diccionario de términos de la ciencia de Datos
		Costes y beneficios	Esquema Técnico presupuestal
	Objetivos del análisis	Metas del análisis	Esquema de Preguntas inteligentes
		Criterios de éxito	Esquema técnico de dimensiones de éxito
	Realizar el plan del proyecto	Plan de actividades	Esquema técnico factibilidad
		Valoración	
FASE 2. ESTUDIO Y COMPRENSIÓN DE DATOS	Recolectar los datos iniciales	Reporte de la recolección de los datos	Esquema de Inventario de Fuentes de información
	Descripción de los datos	Reporte de la descripción de los datos	

	Exploración de los datos	Reporte de la exploración de los datos	Esquema de sumarización estadística y valoración de dimensiones de calidad
	Verificar la calidad de los datos	Reporte para verificar la calidad de los datos	
FASE 3. PREPARACIÓN DE LOS DATOS	Datasets	Descripción del Dataset	Esquema de Procedimientos de Data Staging
	Selección de Datos	inclusión o exclusión de los datos	
	Limpieza de Datos	Reporte de la calidad de los datos	
	Construcción de nuevos datos y estructura	Derivación de atributos	Esquema de Creación de nuevos datos
		Generación de registros	
	Integración de datos	Unificación de los datos	Esquema de creación de Datamarts
	Formato de datos	Reporte de calidad de datos	
FASE 4. MODELADO	Selección de técnicas de modelado	Descripción de técnicas seleccionados	Esquema Matriz de valoración de modelos
		Supuestos del modelo	Esquema Consideraciones de modelado
	Construcción del plan de pruebas	Plan de pruebas	Esquema pruebas de desempeño del modelo
	Construcción del modelo	Análisis de parámetros	Esquema descripción del modelo
		Modelo	
		Descripción del modelo	
	Evaluación del modelo	Evaluar la calidad del modelo	Esquema Pruebas de desempeño del modelo
		Revisión de parámetros	
FASE 5, EVALUACION	Evaluación de resultados	Valoración de los resultados	Esquema pruebas de desempeño del modelo
		Modelos aprobados	
	Proceso de revisión	Revisión del proceso	
	Actividades futuras	Técnicas consideradas	
		Listado de posibles acciones	
FASE 6. DESPLIEGUE	Plan de implantación	Arquitectura propuesta	Esquema Diseño de Arquitectura
		Requerimientos de infraestructura	
	Plan de monitoreo y mantenimiento	Plan de monitoreo y mantenimiento	Esquema modelo de operación y soporte
	Informe final	Documento memoria del proceso	Esquema Ejecutivo de informe de resultados
		Recomendaciones y trabajos futuros	
	Revisión del proyecto	Documentación de experiencias	

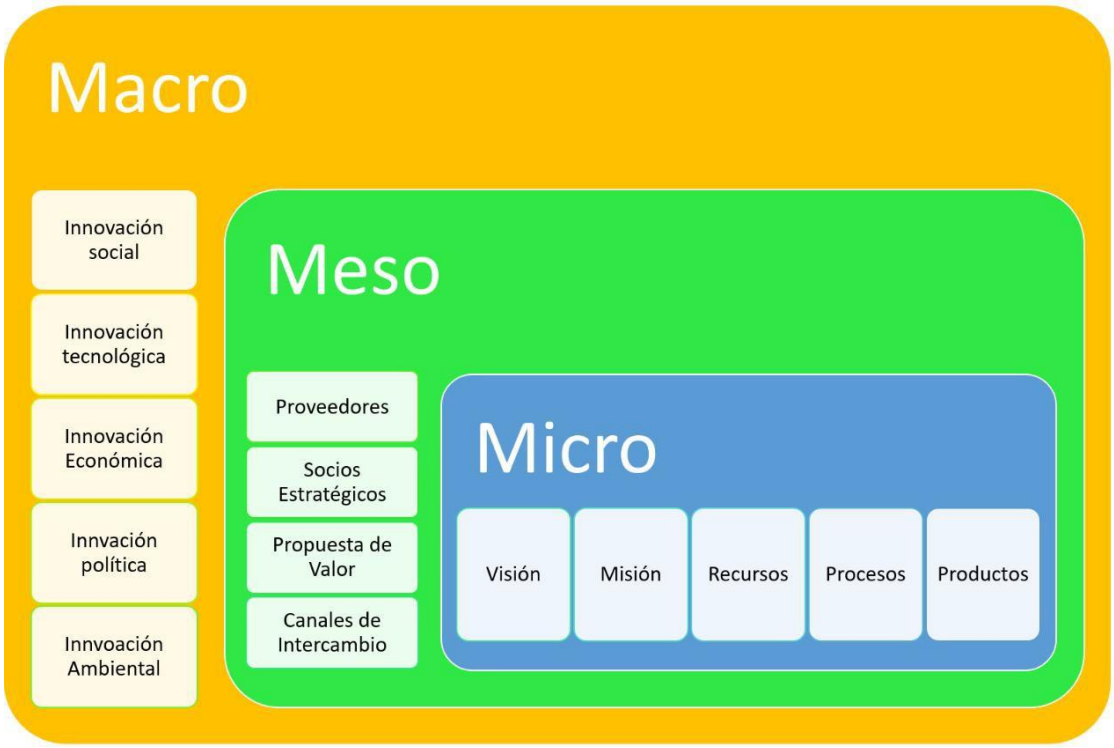
Fuente: (IBM, 2021)

Cada uno de los esquemas propuestos para la gestión documental es producto de un ejercicio metodológico que incluye elementos provenientes de diferentes áreas de conocimiento afines a la Ciencia de Datos, a continuación, hacemos un resumen de los bloques constructores y la lógica subyacente en cada uno de ellos.

7.2.1 Diagnostico Técnico Contextual

Para contextualizar el estado del negocio se hace importante establecer un análisis del contexto general partiendo de los modelos de diagnóstico institucional de tres dimensiones: micro, meso y macro. Estos análisis pretenden establecer una visión integral de la organización, sus objetivos, sus restricciones y las iniciativas que está dispuesta a asumir. El diagrama resumen de este formato es mostrado a continuación.

Ilustración 6: Estructura del diagnóstico técnico contextual



Fuente: Aguilar 2022

En términos generales el análisis incluye una estructura de diagnóstico PEST, propuesto en HARDVARD por Francis Aguilar como la herramienta más efectiva para ayudar a las organizaciones a planificar estratégicamente los proyectos e iniciativas, ofreciendo herramientas que identifiquen las oportunidades de los cambios que ocurren en el contexto convirtiéndolas en oportunidades estratégicas.

7.2.2 Prefactibilidad Técnica institucional

El análisis de la viabilidad Técnica institucional parte del modelo de madurez de los datos en la organización, una forma eficiente de lograrlo es a través de la matriz de diagnóstico interno de la institución a partir de la cual se da respuesta a los tres interrogantes clave del análisis de prefactibilidad técnica: recursos tecnológicos disponibles, características del contexto interno tecnológico y el análisis de riesgos asociados al desarrollo de proyectos de analítica de datos.

Tabla 3: Modelo de madurez institucional en el uso de Datos

	EXPLORACION	USO	LIDERAZGO	INNOVACIÓN
ESTRATEGIA DE NEGOCIO	Datos utilizados únicamente para reportes	Hallazgos en los datos apoyan las decisiones del negocio	La estrategia del negocio se construye a partir de la data	La data es utilizada para identificar la continua evolución de la estrategia de negocio
DATOS	La organización solo utiliza data interna	La organización utiliza proveedores de datospara enriquecer y complementar su propia data	Se utiliza información de terceros como elemento diferenciador	La organización busca continuamente integrar nuevas fuentes de datos provenientes de fuentes no evidentes

	EXPLORACION	USO	LIDERAZGO	INNOVACIÓN
CULTURA	El uso de datos y el análisis depende del individuo	Los datos son utilizados para medir resultados, pero no para la planificación	Los tomadores de decisiones se basan en resultados de análisis de datos para maximizar los resultados del negocio	La organización ha desarrollado herramientas de INTELGENCIA ARTIFICIAL / MACHINE LEARNING que adaptan y mejoran los objetivos del negocio
ARQUITECTURA	El negocio carece de una arquitectura de datos coherente y cohesiva	Existe alguna arquitectura para automatizar y analizar flujos de datos	La arquitectura posibilita a todos los miembros de la organización tomar decisiones basadas en datos	La arquitectura está hecha para garantizar la velocidad, la distribución y la manipulación de grandes volúmenes de datos
GOBERNANZA	La gobernanza de datos se hace manual y de forma inconsistente	El proceso se realiza en el mismo lugar para proteger la calidad de los datos a lo largo de la organización	Hay plena confianza en los datos y los resultados derivados de los análisis	La gobernanza de datos hace parte de los procesos empresariales
ADQUISICIÓN	No hay carga de fuentes de datos externas	Los equipos son responsables individualmente de la carga de datos para el análisis	Hay una línea de alimentación para los datos, sin embargo, la carga no es universal	La organización tiene un equipo de provisión de datos encargado de la carga de nuevas fuentes

Fuente: Gartner 2020

La metodología de uso de este formato parte por marcar cada recuadro de color verde en caso de que la etapa ya haya sido cumplida por la organización de acuerdo con los criterios descritos, de color amarillo si es una etapa que viene adelantándose o ya se ha surtido en forma parcial con dichos criterios y de color rojo si definitivamente dichos criterios no se cumplen. Este modelo semaforizado es el soporte conceptual de las preguntas clave previamente descritas, a la vez que establece en cada dimensión en que role de madurez se encuentra la organización. Teniendo en cuenta estos resultados en un documento no

superior a una página se resumirán los componentes del estado actual, que posibilidades para proyectos de analítica de datos representa y que situaciones de riesgo se identifican.

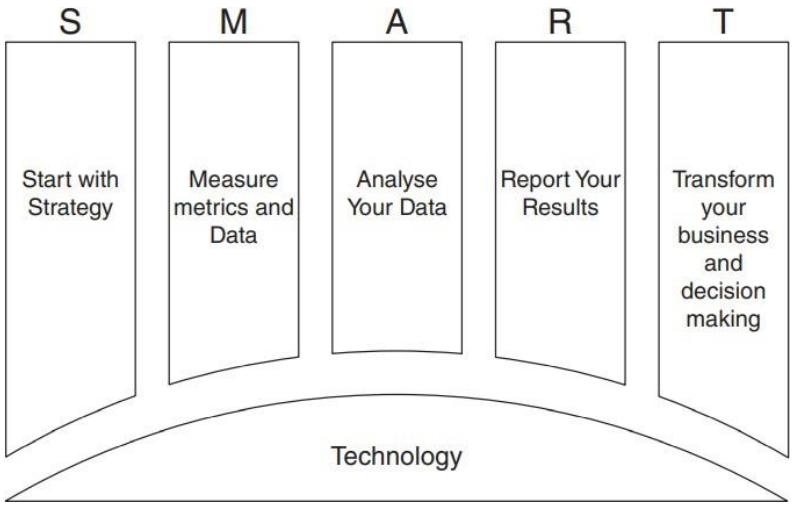
7.2.3 Diccionario de términos relacionados con el negocio y la ciencia de datos

El diccionario mínimo común a la gran mayoría de los proyectos de analítica de datos puede verse evidenciado en la propuesta disponible en el anexo 1 de este documento.

7.2.4 Preguntas inteligentes

El desarrollo de preguntas inteligentes parte del uso de la metodología SMART Strategy Board (Bernard Marr, 2015). Desarrollado dentro de un conjunto de herramientas orientadas a apoyar el desarrollo de actividades de consultoría, representa el eje central del análisis intuitivo de la institución de cara a establecer objetivos específicos para la aplicación de la ciencia de datos. La estructura general de los componentes filosóficos y del modelo de aplicación son mostrados a continuación.

Ilustración 7: Elementos constructores de la metodología SMART



Fuente: Marr, 2018

En la implementación práctica de proyectos de analítica solo el 10% de las ocasiones la información institucional disponible está preparada para iniciar procesos de análisis de datos, en el resto de ocasiones es necesario desarrollar nuevos procesos de recolección y preparación de datos de cara a dar solución a las necesidades institucionales, a este tipo de actividades se refiere el componente de “ESTRATEGIA” mostrado en el diagrama anterior. Sin embargo, esta validación preliminar no sería posible si no se parte de un conjunto de criterios específicos que definan claramente los objetivos institucionales que se buscan, es allí cuando se hace necesario desarrollar un diagnóstico rápido y contundente que incluya las diferentes dimensiones del negocio en la conceptualización de una pregunta única, de forma análoga a los procesos formales de investigación científica y direccionado a dar respuesta a necesidades específicas del mercado. A este conjunto de interrogantes se les denomina “Preguntas Inteligentes” (Smart Questions) y para construirlas se hace uso del “TABLERO DE ESTRATEGIA SMART” (Smart Strategy Board)

Tabla 4: Tablero de estrategia

<table>
<tr><td colspan="3"><u>PROPÓSITOS / AMBICIONES:</u>
Misión, Visión, Objetivos particulares de la Institución y sus miembros</td></tr>
<tr><td><u>CLIENTES</u>

Mercado Objetivo
Propuestas de Valor</td><td><u>FINANZAS</u>

Objetivos financieros, procesos de generación de efectivo y rentabilidad</td><td><u>INVOLUCRADOS EXTERNOS</u>

Identificación de involucrados externos que inciden en el negocio: competidores, legislación, otras organizaciones….
Análisis de los riesgos asociados y de los desafíos que estos implican</td></tr>
<tr><td><u>OPERACIONES</u>
Análisis de Aliados
Competencias institucionales Claves para el negocio</td><td></td><td></td></tr>
<tr><td colspan="3"><u>RECURSOS</u>

Infraestructura tecnológica, herramientas de tecnologías de la información, talento humano, valores institucionales</td></tr>
</table>

Fuente: Marr, 2018

Al listar cada uno de los criterios presentes en el tablero de estrategia se evidencian las dimensiones que dan origen a la pregunta inteligente dentro del modelo de negocio. Cada organización dependiendo de su momento, desea apoyar sus objetivos de negocio o gestionar mejor a sus clientes, optimizar sus operaciones y recursos, fortalecer su estructura financiera o gestionar sus competidores y riesgos. En ocasiones enfocarse en un aspecto especifico es la mejor forma de dar respuesta al mayor número de dimensiones posible, sea cual fuere el caso formular una pregunta inteligente es un proceso iterativo que debe poder

construirse en forma ágil para luego pasar a la validación y aprobación de la misma, Las mejores recomendaciones para apoyar el diseño de la pregunta inteligente son:

1. Diligenciar el tablero de estrategia.
2. Seleccionar una o dos dimensiones.
3. Elaborar una interrogante iniciando con la expresión: ¿ES POSIBLE?
4. Describir un criterio específico dentro de cada dimensión, en caso de aplicar concatenar los criterios pertenecientes a dimensiones diferentes.
5. Preguntarse sin contemplar restricciones tecnológicas (Estas hacen parte del modelo de valoración de la pregunta inteligente).
6. Una vez cuente con varias preguntas realice el proceso opuesto identificando las dificultades para dar respuesta a cada pregunta.
7. Por último, seleccione el camino más simple, la pregunta con menos restricciones técnicas asociadas es una pregunta inteligente que vale la pena evaluar.

NOTA: En caso de fallar la evaluación técnica de la implementación de una pregunta inteligente, diligenciar nuevamente el tablero de estrategia y comenzar el proceso de construcción de preguntas inteligentes.

7.2.5 Dimensiones de éxito para los proyectos de analítica de datos

El formato técnico de dimensiones de éxito es un instrumento que contempla las tres líneas fundamentales de valoración de un proceso de analítica: Calidad de la solución analítica, Impacto en los procesos institucionales y fortalecimiento de la base de conocimiento. Cada una de estas dimensiones cuenta con métricas específicas que son resultados de procesos complementarios al ejercicio de la analítica en sí mismo, tal como se evidencia en el siguiente diagrama.

Ilustración 8: Dimensiones para evaluación del éxito de proyectos de analítica en la SSPD

Corto Plazo: Calidad de la Solución	Mediano y Largo Plazo: Mejoramiento de Procesos	Incremento de la base de conocimiento institucional
• Aplicación de métricas de desempeño para los modelos predictivos con evaluaciones superiores a 90% de Precisión. • Valoración de la usabilidad de las herramientas de análisis descriptivo por parte de los analistas de la SSPD (Aplicar instrumentos de recolección de datos).	• Análisis de Tráfico de usabilidad del producto de analítica y sus componentes. • Validación de la pertinencia del producto y su aporte al desempeño de las diferentes dependencias del negocio (Aplicar instrumentos de recolección de datos).	• Numero de Insights de calidad de la información encontrados durante el proceso de analítica. • Numero de insights de requerimiento infraestructura tecnológica encontrados en el proceso. • Numero de procesos de transformación tecnológica institucional derivados de los insigths encontrados a lo largo de la experiencia.

Fuente: Elaboración propia.

En cualquier caso, el éxito de un proyecto de analítica requiere ventanas de tiempo para su análisis, el análisis inmediato de la usabilidad y las métricas propias de los modelos solamente indicará el estado actual del proyecto desde un enfoque subjetivo. Por tal motivo todo proyecto de analítica debe establecer cuáles de las métricas aquí presentes deberán implementarse a corto, mediano y largo plazo, y cuáles de ellas deben hacerse recurrentemente como ocurre en procesos de aprendizaje continuo a través de metodologías de inteligencia artificial. Por lo general las métricas de calidad de la solución e incremento de la base de conocimiento se realizan recurrentemente, mientras las métricas asociadas al mejoramiento de procesos se realizan a mediano y largo plazo teniendo en cuenta los ciclos de adopción y mejora continua de las soluciones de analítica.

Las preguntas clave a ser incluidas en los formatos de recolección de información para validación de impactos incluyen:

¿Encuentra útiles los resultados mostrados por la solución?

¿Dichos resultados hacen parte de los cálculos o procedimientos que realiza día a día en el desarrollo de sus actividades?

¿Si pudiera calcular el impacto en tiempos, ¿cuál sería el ahorro de tiempos que el uso de esta herramienta?

¿Cree que es posible mejorar la herramienta para que pueda ser utilizada con mayor frecuencia? Por favor compartir las sugerencias y propuestas de mejoramiento.

7.2.6 Inventario de fuentes de información

Las fuentes de información disponibles en la organización se identifican a través de un proceso itinerante con las dependencias responsables de la administración y mantenimiento de los recursos de información institucional, en cada caso deben tenerse en cuenta los siguientes criterios técnicos generales para describir las fuentes:

COMPONENTE DEL NEGOCIO AL QUE PERTENECE: Es una descripción sencilla del origen de cada fuente de información haciendo énfasis en que procesos soporta y que dependencias de la organización hacen uso de ella.

PERIODICIDAD DE ACTUALIZACION: Se refiere al periodo de tiempo con el cual es actualizada la información perteneciente a la fuente.

DESCRIPCIÓN DE LOS DATOS: Listado de los campos que conforman la fuente de información y su contenido.

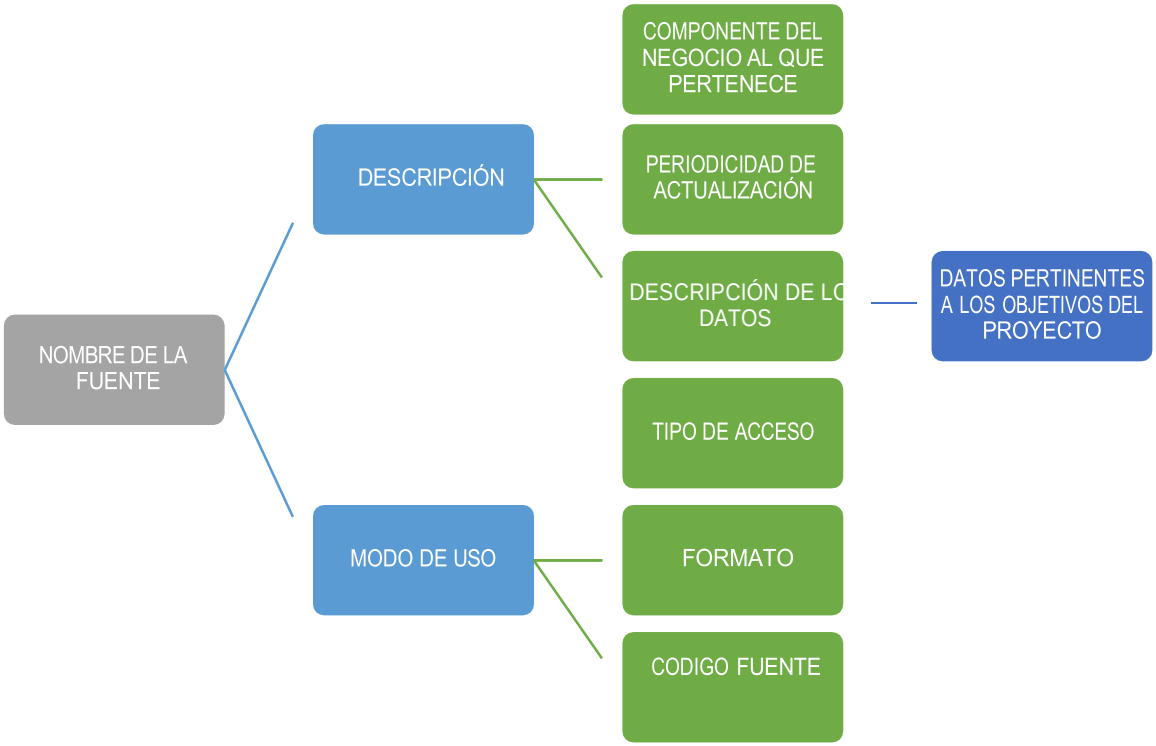
DATOS PERTINENTES A LOS OBJETIVOS DEL PROYECTO: Conjunto de datos que por su naturaleza están alineados con las actividades estratégicas definidas para el proyecto:

TIPO DE ACCESO: Descripción simple de la forma como se accede a la información.

FORMATO: Se refiere al formato que tienen los datos para leerse

CODIGO FUENTE: En caso de ser requerido, deben mostrarse las líneas de código que hacen posible acceder a los datos.

Ilustración 9: Componentes del inventario de fuentes de datos



Fuente: elaboración propia

7.2.7 Resumen Estadístico

El esquema de sumarización estadística se aplica se a cada uno de los datos pertinentes a los objetivos del proyecto, dando como resultado una matriz de análisis que incluye los siguientes valores e indicadores:

Tabla 5 Componentes de sumarización estadística a aplicar en cada variable

CRITERIO	DESCRIPCIÓN
Numero de filas	Hace referencia al número de registros de la fuente de información
Número de Columnas	Hace referencia al número de campos de la fuente
Tamaño	Tamaño en MegaBytes
Valor Mínimo	**El valor numérico mínimo de toda la muestra**

CRITERIO	DESCRIPCIÓN
Valor Máximo	El valor numérico máximo de toda la muestra
Media	Promedio de valor para todos los valores de la muestra
Moda	En datos nominales o categóricos el valor que más se repite
Desviación Standard	Medida utilizada para calcular la dispersión de los datos en una muestra estadística
1 cuartil	se refiere al valor máximo bajo el cual se encuentra el 25% de los datos de la muestra, sirve como indicador del apiñamiento de los datos en una distribución estadística.
2 cuartil	se refiere al valor máximo bajo el cual se encuentra el 50% de los datos de la muestra, sirve como indicador del apiñamiento de los datos en una distribución estadística.
3 cuartil	se refiere al valor máximo bajo el cual se encuentra el 75% de los datos de la muestra, sirve como indicador del apiñamiento de los datos en una distribución estadística.
Tipo de Variable	Las variables pueden ser numéricas, nominales, categóricas, fechas, coordenadas geográficas o texto no estructurado
Numero de Datos	Número total de datos presentes en la muestra
Número de Datos Vacíos	Numero de datos vacíos (sin valor) dentro de la muestra
Numero de Datos NULOS	Numero de datos que presenta datos nulos o con valores que representan dicho estado

Fuente: Elaboración propia

7.2.8 Análisis de Calidad de los datos

Este análisis debe aplicarse y presentarse en una matriz unificada por cada fuente de datos, de cara a fortalecer el proceso de calidad de la información. Las métricas más utilizadas para evaluar la calidad de los datos son en su conjunto las siguientes:

Tabla 6: Indicadores de calidad de los datos

ATRIBUTO	SIGNIFICADO	MÉTRICAS
CONSISTENCIA	No importa en qué lugar de la data te encuentres no encontrarás incongruencias en la naturaleza y contenido de los datos.	Número de inconsistencias y análisis porcentual de las mismas. *[Ejemplo: En dos tablas distintas unas usuarias presentan diferente número de acciones relativas a un mismo proceso]*
EXACTITUD	Los datos contenidos en la base de datos tienen sentido y se ajustan a la realidad	Número de datos inexactos y análisis porcentual de los mismos. *[Ejemplo: Las coordenadas geográficas no coinciden con un valor conocido]*
COMPLETITUD	Todos los elementos presentes en la base de datos tienen valores con trazabilidad	Numero de valores vacíos *[Ejemplo: No es posible encontrar un valor para la fecha de un proceso]*
AUDITABILIDAD	La información es accesible y es posible realizar una trazabilidad de los cambios realizados	Porcentaje de los datos donde no es posible hacer trazabilidad de los cambios realizados *[Ejemplo: Ante nuevos ingresos de actividad de un cliente, no es posible identificar cuando o que artículos adquirió]*

ATRIBUTO	SIGNIFICADO	MÉTRICAS
ORDINALIDAD	La información ingresada tiene sentido	Numero de datos sin sentido [Ejemplo: Información de fechas con cambio de formato que alteran la forma de interpretación de los datos]
UNICIDAD	Existen identificadores únicos que permiten identificar la	Numero de valores duplicados [Ejemplo: dos clientes tienen el mismo ID en el sistema]
TEMPORABILIDAD	La data presenta la realidad de la información en un período de tiempo	Numero de datos que están por fuera de la sincronía temporal [Ejemplo: La dirección de un usuario ha variado desde la última compra, pero este valor no se ha actualizado en los registros de compra anteriores]

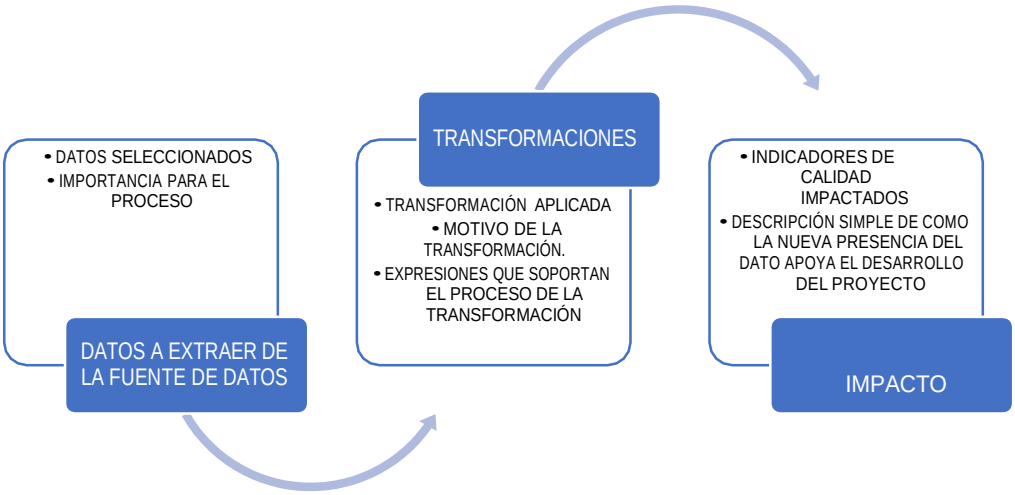
Fuente: Data Quality, 2017

7.2.9 Procedimientos de Area-Stagging:

El procedimiento de área Staging parte del análisis de calidad de los datos y el inventario de datos para realizar una selección de los mismos y establecer los procesos de corrección y ajuste requeridos para el proceso, en esta etapa se realizan los procedimientos de EXTRACCION y TRANSFORMACIÓN PRIMARIA DE LOS DATOS.

Para dar cumplimiento a este procedimiento se parte del ejercicio de identificación de los datos requeridos, la transformación que debe realizarse, los motivos de dicha transformación y si lo tiene el impacto sobre los indicadores de calidad de la organización.

Ilustración 10: Estructura de elementos de la documentación de ETL´s

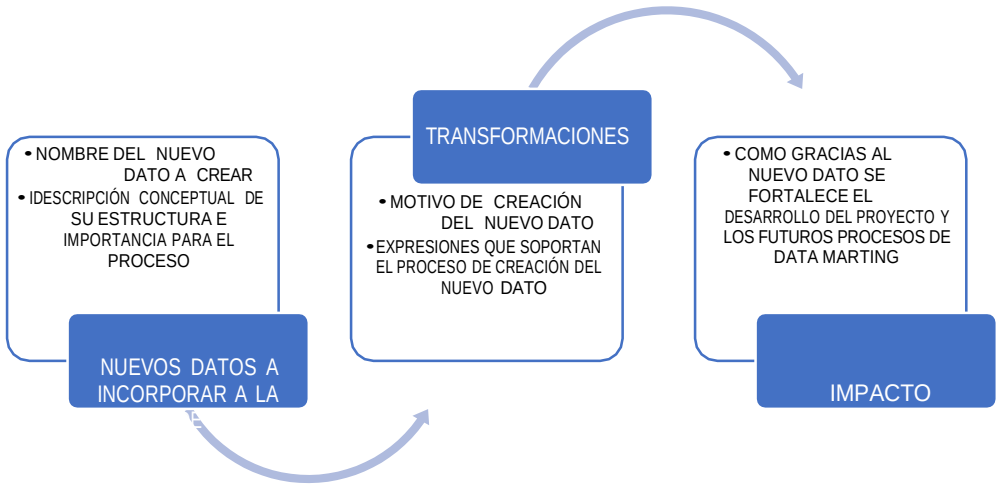


Fuente: Elaboración propia

7.2.10 Creación de nuevos datos

La creación de nuevos datos es un procedimiento de TRANSFORMACIÓN que es necesario de cara a la preparación de la información para el consumo, de forma análoga al proceso de data Staging se identifica la fuente que afectan, los procedimientos utilizados y el impacto sobre la fuente de datos. La estructura de dichas actividades es mostrada a continuación:

Ilustración 11: Estructura de elementos de la documentación de datos creados



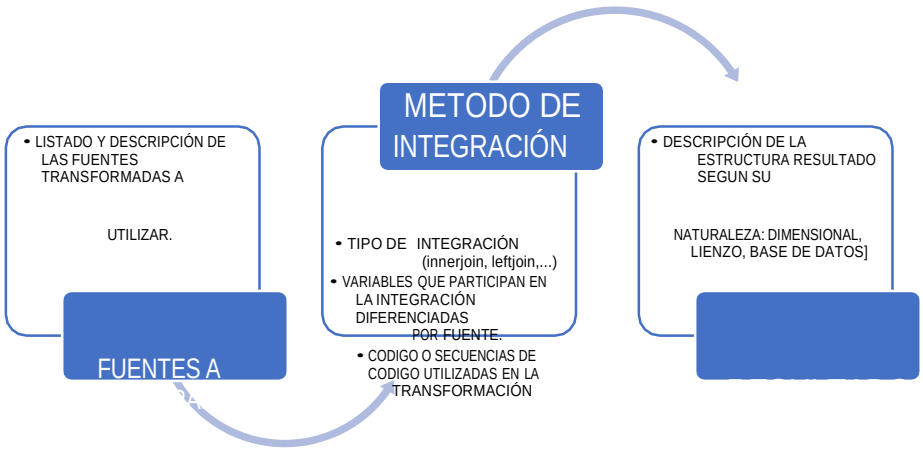
Fuente: Elaboración propia

7.2.11 Datamarts

Este proceso corresponde al proceso de CARGA DE INFORMACIÓN o DESPLIEGUE DE INFORMACIÓN, es decir la determinación de la estructura de datos consolidada que puede ser utilizada dentro de la estructura de una DATA WARE HOUSE para el consumo de los diferentes usuarios y herramientas desarrolladas.

Para apoyar este proceso se parten de dos esquemas, el esquema conceptual de creación de lienzos de data y el esquema conceptual de la estructura de los data warehouses, bases de datos o estructuras NOSQL a partir de las cuales se dará la opción de consumo.

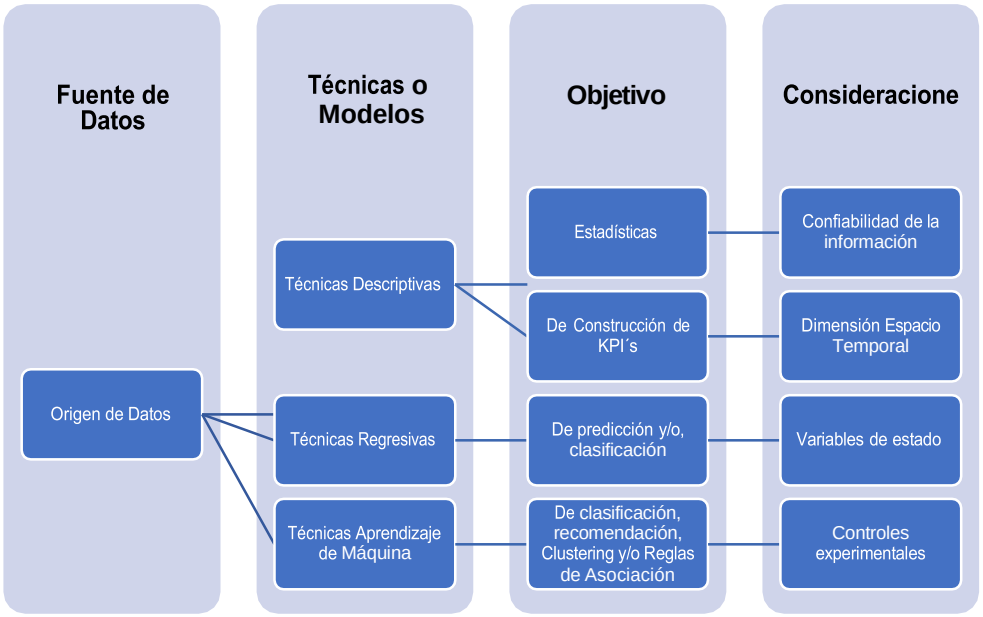
Ilustración 12: Estructura de los elementos de la documentación de DATAMARTS



Fuente: Elaboración propia

7.2.12 Selección y evaluación de modelos

Ilustración 13: Clasificación de los modelos aplicables teniendo en cuenta los objetivos



Fuente: Elaboración propia

Al conceptualizar un proceso de modelado dentro del ejercicio metodológico, los componentes mínimos a tener en cuenta tienen la siguiente estructura.

Confiabilidad de la información: parte de un análisis de la calidad de los datos y su relevancia dentro de la toma de decisiones institucionales

Dimensión espacio temporal: para el análisis tener en cuenta el intervalo de tiempo aplicado y la segmentación geográfica de la información.

Variables de estado: Incluir siempre del proceso de modelado la identificación de las variables e indicadores clave para la administración del negocio y concertar los indicadores nuevos con expertos de cada dependencia.

Controles experimentales: Pruebas de validación de la calidad de los modelos, caracterización de los tamaños muestrales y planificación de los procesos de reentrenamiento de los modelos.

7.2.13 Desempeño de modelos

Para realizar las pruebas de desempeño del modelo deben partirse de las consideraciones y objetivos planteados para el modelo. Según dichos objetivos y sus consideraciones es posible utilizar según el caso las siguientes métricas para la evaluación del desempeño. Si bien existen métricas nuevas producto de la evolución de estos procedimientos, las aquí contenidas son las más usadas y consideradas dentro del FRAMEWORK de testing que empresas como MICROSOFT e IBM. A continuación, se muestran los criterios de análisis aplicables a los diferentes tipos de modelado y aprobados por OARES.

Tabla 7: Métricas de desempeño para modelos de clasificación binaria

Métricas de evaluación para la clasificación binaria		
Métricas	**Descripción**	**Valor Esperado**
Precisión	El porcentaje de predicciones hechas correctamente utilizando datos de prueba constituye una mayor precisión, proporcionando predicciones precisas para todas las muestras de entrada. Su funcionamiento es correcto con un mayor número de similitudes en la muestra de igual clase.	Un valor de exactamente 1,00 indica un problema (a menudo, fuga de etiqueta/objetivo, sobreajuste o prueba con datos de entrenamiento). La eficacia de un clasificador no se refleja completamente en la precisión cuando los datos de prueba están desequilibrados (la mayoría de los ejemplos pertenecen a una clase), el conjunto de datos es pequeño o las puntuaciones están cerca de 0,00 o 1,00. Por lo tanto, es necesario examinar otras métricas**.**

Métricas de evaluación para la clasificación binaria		
Métricas	**Descripción**	**Valor Esperado**
AUC	AucROC o Área bajo la curva mide el área bajo la curva que se creó limpiando la tasa de positivos verdaderos frente a la tasa de falsos positivos.	**Cuanto más cerca de 1,00, mejor**. Debe ser mayor que 0,50 para que un modelo sea aceptable. Un modelo con AUC de 0,50 o menos no tiene ningún valor.
AUCPR	aucPR o *Área bajo la curva de una curva de precisión-recuperación*: Medida útil de éxito de predicción cuando las clases están poco equilibradas (conjuntos de datos muy sesgados).	**Cuanto más cerca de 1,00, mejor**. Las puntuaciones altas cercanas a 1,00 muestran que el clasificador devuelve resultados precisos (alta precisión), así como una mayoría de todos los resultados positivos (recuperación alta).
Puntuación F1	La puntuación F1 también se denomina puntuación F equilibrada o F medida F. Es la media armónica de la precisión y la recuperación. La puntuación F1 resulta útil cuando desea buscar un equilibrio entre la precisión y la recuperación.	**Cuanto más cerca de 1,00, mejor**. Una puntuación F1 alcanza el mejor valor en 1,00 y la peor puntuación en 0,00. Indica cuán preciso es el clasificador.

Fuente: IBM, 2020

Tabla 8: Métricas de desempeño para modelos de clasificación multiclase

Métricas para la clasificación multiclase		
Métricas	**Descripción**	**Buscar**
Microprecisión	La precisión de micro promedio agrega las contribuciones de todas las clases para calcular la métrica promedio. Es la fracción de instancias que se predijeron correctamente. El micro promedio no tiene en cuenta la pertenencia a una clase. Básicamente, todos los pares de ejemplo y clase contribuyen del mismo modo a la métrica de precisión.	**Cuanto más cerca de 1,00, mejor**. En una tarea de clasificación de varias clases, la microequi exactitud es preferible a la precisión de macros si sospecha que puede haber un desequilibrio de clases (es decir, puede tener muchos más ejemplos de una clase que de otras clases).

Métricas para la clasificación multiclase		
Métricas	**Descripción**	**Buscar**
Macro precisión	La precisión de macro promedio es la precisión promedio en el nivel de clase. La precisión de cada clase se calcula y la macro precisión es el promedio de estas precisiones. Básicamente, todas las clases contribuyen del mismo modo a la métrica de precisión. Las clases minoritarias tienen el mismo peso que las clases más grandes. La métrica de macro promedio proporciona el mismo peso a cada clase, independientemente de cuántas instancias de esa clase contiene el conjunto de datos.	**Cuanto más cerca de 1,00, mejor**. Calcula la métrica de forma independiente para cada clase y, a continuación, toma la media (por lo tanto, se consideran todas las clases de igual forma)
Pérdida de registro	La pérdida logarítmica mide el rendimiento de un modelo de clasificación donde la entrada de predicción es un valor de probabilidad de entre 0,00 y 1,00. La pérdida de registro aumenta a medida que la probabilidad de predicción difiere de la etiqueta real.	**Cuanto más cerca de 0,00, mejor**. Un modelo perfecto tendría una pérdida de registro de 0,00. El objetivo de nuestros modelos de Machine Learning es minimizar este valor.

Métricas para la clasificación multiclase		
Métricas	**Descripción**	**Buscar**
Reducción de pérdida logarítmica	La reducción de pérdida logarítmica se puede interpretar como la ventaja del clasificador sobre una predicción aleatoria.	**Parte de -inf y 1,00, donde 1,00 equivale a una predicción perfecta, y 0,00 indica una predicción aproximada**. Por ejemplo, si el valor equivale a 0,20, se puede interpretar como "la probabilidad de que una predicción correcta sea 20 % mejor que el cálculo aleatorio"

Tabla 9: Métricas de desempeño para modelos de regresión y recomendación

Métricas para modelos de regresión y recomendación		
Métrica	**Descripción**	**Buscar**
R cuadrado	R cuadrado (R2), o el coeficiente de determinación representan la eficacia predictiva del modelo como un valor comprendido entre -inf y 1,00. 1,00 significa que hay un ajuste perfecto y, dado que el ajuste puede ser arbitrariamente deficiente, las puntuaciones pueden ser negativas. Una puntuación de 0,00 significa que el modelo consiste en adivinar el valor esperado para la etiqueta. Un valor R2 negativo indica que el ajuste no sigue la tendencia de los datos y el modelo funciona peor que el cálculo aleatorio. Esto solo es posible con modelos de regresión no lineal o con regresión lineal restringida. R2 mide la proximidad de los valores de datos de prueba reales a los valores de predicción.	**Cuanto más cerca de 1,00, es mejor la calidad**. Sin embargo, a veces valores bajos de R cuadrado (por ejemplo, 0,50) pueden ser completamente normales o lo suficientemente buenos en un escenario, y los valores altos de R cuadrado no siempre son buenos y pueden ser sospechosos.

Pérdida absoluta	La pérdida absoluta o la desviación media (MAE) mide la proximidad de las predicciones a los resultados reales. Se trata de la media de todos los errores del modelo, donde el error del modelo es la distancia absoluta entre el valor de la etiqueta predicho y el valor de la etiqueta correcto. Este error de predicción se calcula para cada registro del conjunto de datos de prueba. Por último, el valor medio se calcula para todas las desviaciones medias registradas.	**Cuanto más cerca de 0,00, es mejor la calidad.** La desviación media utiliza la misma escala que los datos que se van a medir (no se normaliza en un intervalo específico). La pérdida absoluta, la pérdida cuadrática y la pérdida de RMS solo pueden usarse para realizar comparaciones entre los modelos del mismo conjunto de datos o el conjunto de datos y una distribución de valores de etiqueta similar.

Pérdida cuadrática	Squared-loss o Mean Squared Error (MSE), también denominado Mean Squared Deviation (MSD) (Desviación cuadrada media [MSD]), le indica cómo se cierra una línea de regresión a un conjunto de valores de datos de prueba tomando las distancias desde los puntos hasta la línea de regresión (estas distancias son los errores E) y los iguala. El cuadrado proporciona más peso a las grandes diferencias.	Siempre es un valor no negativo y los **valores próximos a 0,00 son mejores**. En función de los datos, puede resultar imposible obtener un valor muy pequeño para el error cuadrático medio.
Pérdida de RMS	La pérdida de RMS o el error de raíz cuadrada media (RMSE) (también denominado desviación de raíz cuadrada media, RMSD), mide la diferencia entre los valores predichos por un modelo y los valores que se observan en el entorno que se está modelando. La pérdida de RMS es la raíz cuadrada de la pérdida cuadrática y tiene las mismas unidades de la etiqueta, similar a la pérdida absoluta, aunque proporciona más peso a las grandes diferencias. El error de raíz cuadrada media se usa habitualmente en la climatología, la previsión y el análisis de regresión para comprobar resultados experimentales.	Siempre es un valor no negativo y los **valores próximos a 0,00 son mejores**. RMSD es una medida de precisión para comparar errores de previsión de diferentes modelos en determinado conjunto de datos y no entre conjuntos de datos, ya que es dependiente de la escala.

Tabla 10: Métricas de desempeño para clústeres

Métricas para evaluación de clusteres		
Métrica	**Descripción**	**Buscar**
Distancia media	Promedio de la distancia entre los puntos de datos y el centro de su clúster asignado. La distancia media es una medida de proximidad de los puntos de datos a los centroides de clúster. Es una medida del grado de "ajuste" del clúster.	Los valores más próximos a **0** son mejores. Cuanto más se acerque a cero la distancia media, más agrupados estarán los datos. Tenga en cuenta, sin embargo, que esta métrica disminuirá si se aumenta el número de clústeres y, en el caso extremo (en el que cada uno de los distintos puntos de datos es su propio clúster) será igual a cero.
Índice de Davies Bouldin	La relación media entre las distancias dentro del clúster y las distancias entre clústeres. Cuanto más ajustado sea el clúster y más separados estén los clústeres, más bajo será este valor.	Los valores más próximos a **0** son mejores. Los clústeres que estén más separados y menos dispersos generarán una mejor puntuación.
Información mutua normalizada	Se puede usar si los datos de entrenamiento usados para entrenar el modelo de agrupación en clústeres se incluyen también con etiquetas verdaderas (es decir, agrupación en clústeres supervisada). La métrica de información mutua normalizada mide si se asignan puntos de datos similares al mismo clúster y puntos de datos dispares a clústeres distintos. La información mutua normalizada es un valor entre 0 y 1	Los valores más próximos a **1** son mejores

Tabla 11: Métricas de desempeño para clasificador

Métricas para evaluar la calidad de las clasificaciones		
Métrica	**Descripción**	**Buscar**
Ganancias acumuladas descontadas	Las ganancias acumuladas descontadas (DCG) son una medida de calidad de la clasificación. Se derivan de dos suposiciones. Una: los elementos altamente pertinentes resultan más útiles si aparecen más arriba en orden de clasificación. Dos: la utilidad realiza un seguimiento de la pertinencia, es decir, cuanto mayor es la pertinencia, más útil es un artículo. Las ganancias acumuladas descontadas se calculan para conseguir una posición determinada en el orden de clasificación. Suma la calificación de pertinencia dividida por el logaritmo del índice de clasificación hasta la posición de interés. Se calcula mediante $\sum_{i=0}^{p} \frac{rel_i}{\log_e i+1}$ Las calificaciones de pertinencia se proporcionan a un algoritmo de entrenamiento de clasificación como etiquetas verdaderas. Un valor de DCG se proporciona para cada posición de la tabla de clasificación, de ahí el nombre de **ganancias** acumuladas descontadas.	**Los valores más altos son mejores**
Ganancias acumuladas descontadas normalizadas	La normalización de DCG permite la comparación de la métrica para las listas de clasificación de diferentes longitudes	**Los valores más próximos a 1 son mejores**

Tabla 12: Métricas para la identificación de anomalías en modelos

Métricas para la evaluación en detección de anomalías		
Métrica	**Descripción**	**Buscar**
Área bajo la curva de ROC	El área bajo la curva receptor-operador mide el grado de eficacia del modelo al separar los puntos de datos anómalos y los habituales.	**Los valores más próximos a 1 son mejores**. Solo los valores mayores que 0,5 muestran la eficacia del modelo. Los valores de 0,5 o menos indican que el modelo no es mejor que la asignación aleatoria de las entradas a categorías anómalas y habituales
Tasa de detección en el recuento de falsos positivos	La tasa de detección en el recuento de falsos positivos es la relación entre el número de anomalías identificadas correctamente y el número total de anomalías de un conjunto de prueba, indexada por cada falso positivo. Es decir, hay un valor de la tasa de detección en el recuento de falsos positivos para cada elemento de falsos positivos.	**Los valores más próximos a 1 son mejores**. Si no hay ningún falso positivo, este valor será 1

7.2.14 Arquitectura para la implementación de los procesos de analítica

La arquitectura para procesos de analítica de datos automatizada se propone una particularización de la arquitectura general “DATA LAKES” que se ajuste a las condiciones técnicas y administrativas de la oficina T.I que gestiona y coordina los procesos tecnológicos institucionales.

Bajo la propuesta metodológica mostrada en la imagen la infraestructura óptima para el desarrollo de procesos de analítica de datos institucional cuenta con cuatro capas fundamentales:

INGESTA (IngestionLayer): Capa donde se realiza la lectura de la información disponible y se extrae la información útil a ser almacenada para futuros procesamientos.

ALMACENAMIENTO (Storage or Staging Area): Capa de almacenamiento donde todas las fuentes de información seleccionadas se alojan y preparan para su posterior consumo.

TRANSFORMACIÓN (Transformation Layer): Capa donde tienen lugar los procesos de análisis de calidad, limpieza y transformación de los datos contenidos en las fuentes de información, el resultado de esta capa son los diferentes componentes de información de utilidad para procesos de visualización y parametrización.

INTERACCION (Interaction Layer): Capa donde ocurren los procesos de administración efectiva de la información y se disponen los procesos de consumo de datos cualificados para las diferentes actividades de analítica de datos.

El análisis de las tecnologías disponibles en el mercado para la implementación de lagos de datos pone en evidencia que la disponibilidad de dichas arquitecturas para proyectos de mediana y baja envergadura se encuentra restringida entre otras características por: requisitos de pago de servicios de almacenamiento en la nube para obtener la mayor potencialidad de los servicios, altos costos y curvas de aprendizaje muy pronunciadas debido

a la complejidad de uso haciendo necesario contratar personal especializado.

Tabla 13: Comparativo de servicios de DATA-LAKES disponibles en el mercado

PROVEEDOR	HADOOP	IBM	MICROSOFT	GOOGLE	DREMIO
PRECIO	FREE O PAGO POR DEMANDA	LICENCIA PAGO POR CONSUMO	LICENCIA	LICENCIA PAGO POR CONSUMO	FREE O PAGO POR LICENCIA
MODELO	CLUSTERES LOCALES O DEPENDENCIAS AWZ entre otras	CLOUD SERVICES	CLOUD SERVICES	CLOUD SERVICES	CLUSTERES LOCALES O DEPENDENCIAS AWZ entre otras
ARQUITECTURA	DISTRIBUIDA PERSONALIZABLE	DISTRIBUIDA NO PERSONALIZABLE	DISTRIBUIDA NO PERSONALIZABLE	DISTRIBUIDA NO PERSONALIZABLE	TOTEMICA DEDICADA AL ALMACENAMIENTO
INTERACCION CON OTRAS TECNOLOGIAS	PERMITIDA, PERSONALIZABLE	RESTRINGIDA	RESTRINGIDA	RESTRINGIDA	PERMITIDA, PERSONALIZABLE
INCLUYEN TECNOLOGIAS ETL / ELT	SI	SI	SI	SI	NO
INCLUYEN HERRAMIENTAS DE VISUALIZACION	NO	SI	SI	SI	NO
INCLUYEN HERRAMIENTAS DE ANALITICA AVANZADA	SI	SI	SI	SI	NO
INCLUYEN HERRAMIENTAS AI	PROYECTO EN DESARROLLO	SI	SI	SI	NO
ESQUEMA DE SALIDA	DATA WAREHOUSE, DATAMARTS, JSON	SAND BOXES & DATA MARTS	APPLICATIONS, DATASETS, VISULIZATIONS	DATAMARTS, MODELS, DATA STORES	DATA WARE HOUSE

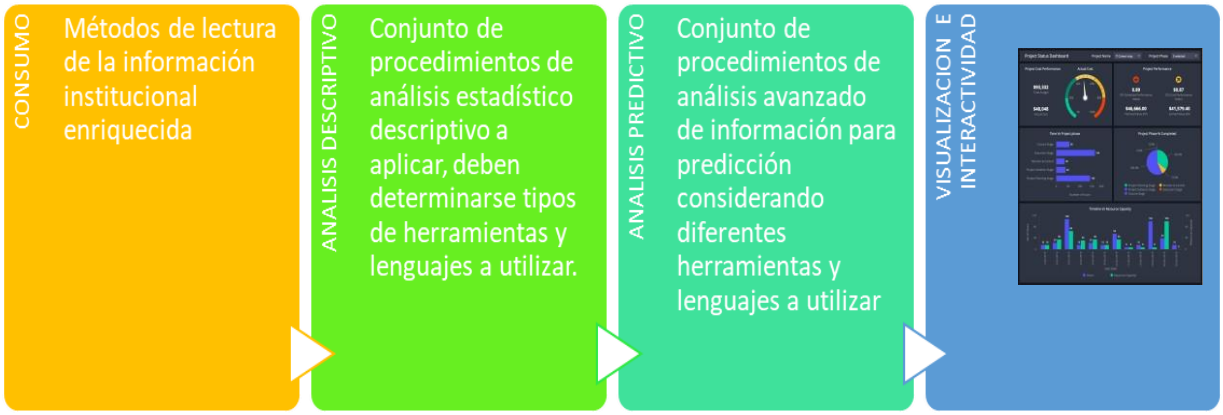
Fuente elaboración propia

En este proceso documental es necesario aplicar la matriz comparativa de servicios y tecnologías para la arquitectura tecnológica que soporta el desarrollo de la analítica de datos institucional. En este documento debe aclararse el conjunto de criterios que motivó la selección de tecnología y el proveedor de la misma.

7.2.15 Modelo de operación y soporte

El modelo de operación es resultado del proceso posterior al diseño de la arquitectura, las aplicaciones o desarrollos realizados por la organización deben tener en cuenta la arquitectura tecnológica particular y respetar dichos criterios en sus etapas de diseño e implementación. Los componentes generales del modelo de operación se muestran en el siguiente diagrama y deben ser ampliamente explicados en la estructura documental.

Ilustración 14: Modelo de operación y soporte de proyectos de analítica institucional SSPD



7.2.16 Resumen ejecutivo de Resultados

El esquema ejecutivo de informe de resultados, es un documento estructurado a partir del cual se narra la experiencia del proyecto y se identifican las principales lecciones aprendidas en el proceso, debe contener principalmente los siguientes elementos:

Introducción a la metodología utilizada

Descripción de los objetivos del proyecto y sus motivaciones Exposición de los resultados y hallazgos realizados por el proceso.

Descripción de contingencias y dificultades encontradas a lo largo del proceso y la forma como se abordaron con el fin de dar cumplimiento a los objetivos

Recomendaciones en tres niveles de profundización: administrativas, tecnológicas y de la lógica del negocio.

Lecciones aprendidas y propuestas de nuevas experiencias.

7.3 Arquitectura tecnológica

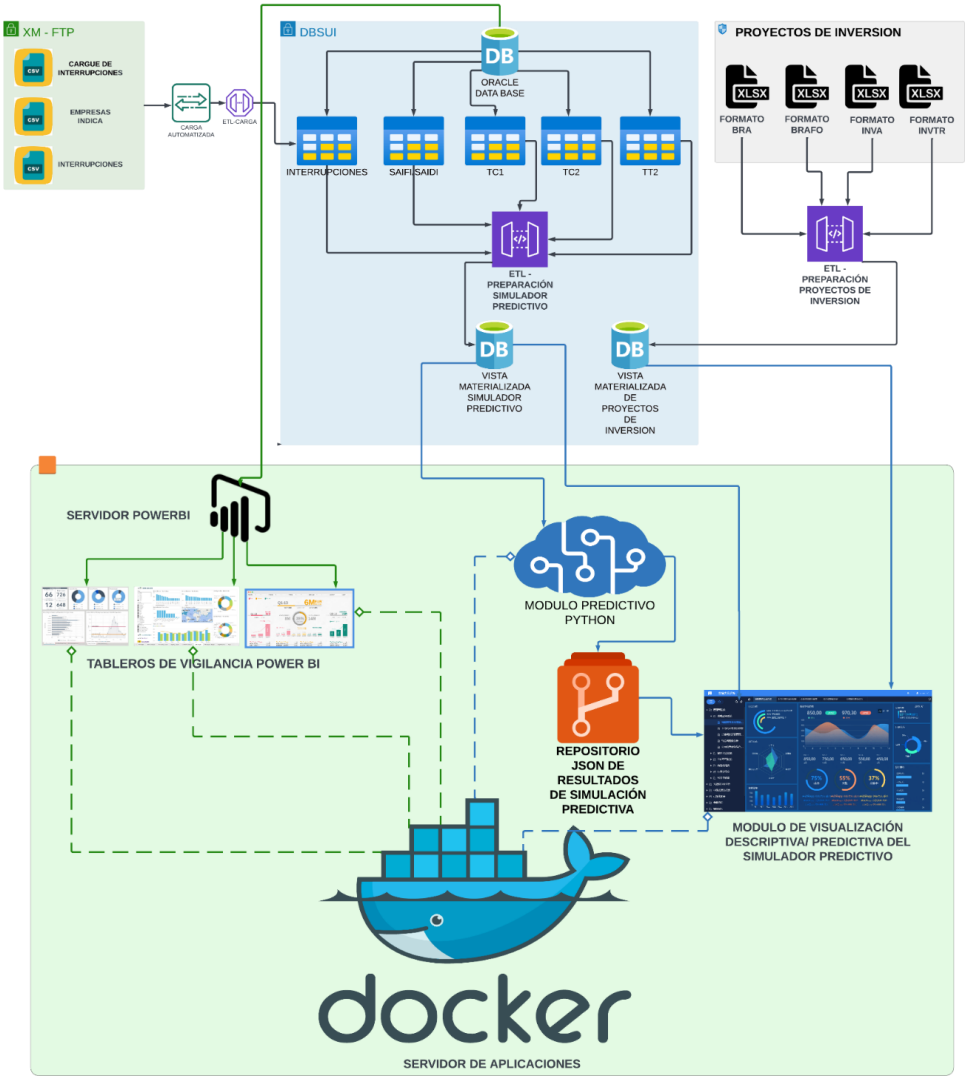
Resultado Objetivo específico número 3

Metodología de Investigación: Enfoque metodológico mixto desarrollado en tres etapas. En la primera etapa se realiza una investigación descriptiva orientada a la identificación de la base teórica conceptual que permita definir una propuesta de modelo de implementación armónica con procesos institucionales y soportada en estructuras documentales concretas. En la segunda etapa se realiza un análisis cualitativo de la pertinencia del modelo propuesto con la dinámica organizacional de la institución motivo de estudio y una tercera etapa donde se realiza un análisis cuantitativo de información institucional para identificar patrones de comportamiento.

El desarrollo metodológico del presente Modelo fue abordado en el Anexo 2, en donde se evidencia el uso de la metodología propuesta en el Modelo, alineado con la metodología de minería de datos (CRISP-DM).

La arquitectura propuesta para la implementación del SIMULADOR

PREDICTIVO es descrita en el siguiente diagrama



El análisis de los componentes que describen esta arquitectura puede dividirse en 3 grandes entornos tecnológicos:

ENTORNO XM: Compuesto por el servidor FTP dispuesto por XM para compartir la información respectiva a las interrupciones reportadas por los operadores.

ENTORNO SUI: Compuesto por la herramienta de carga y ETL de la información proveniente de XM, la infraestructura de bases de datos del SUI y el

conjunto de procedimientos almacenados que la componen:

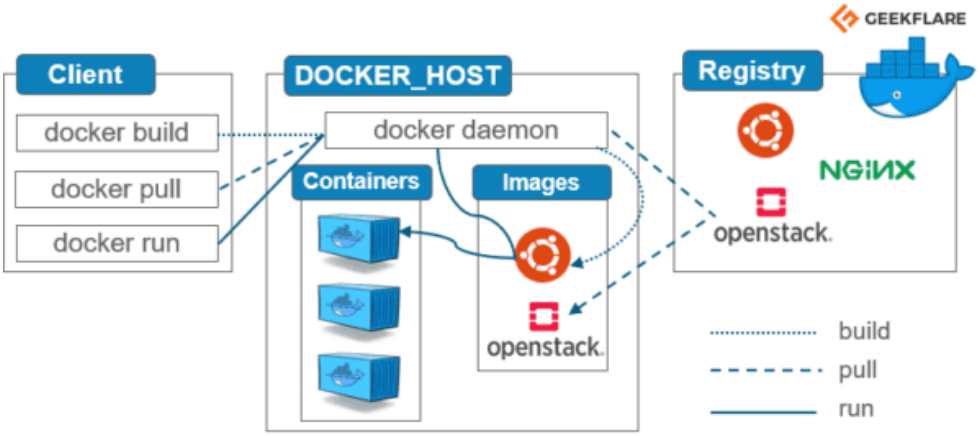
ETL-CARGA XM: Conjunto de procedimientos de cruce y transformación de datos orientados a consolidar la información de las interrupciones provistas por XM. Este proceso incluye la transformación de las fechas a un único formato y el cruce de las tablas CARGUE INTERRUPCIONES, EMPRESAS INDICA e INTERRUPCIONES para generar una única tabla que contenga las interrupciones discriminadas por usuario y activo de transformación tal como puede evidenciarse en los numerales 2.3.2 y 2.3.3.

ETL-PREPARACION SIMULADOR PREDICTIVO: Proceso de integración de las 5 fuentes de datos primordiales para el análisis predictivo (INTERRUPCIONES, TC1, TC2, TT2, SAIFI Y SAIDI), en una única sábana que corresponda a la estructura relacional requerida para alimentar los procesos de simulación y análisis descriptivos. Ver numerales 2.3.4 y 2.3.5

ETL-PREPARACION PROYECTOS DE INVERSIÓN: Proceso de cruce de la información proveniente de los 4 formatos que conforman la carga de información de los prestadores del servicio de energía eléctrica a la SSPD, este proceso está descrito en los numerales 2.3.2 y 2.3.3

ENTORNO SERVIDOR DE APLICACIONES: El servidor de aplicaciones provisto por la oficina OTIC para el despliegue del tablero de vigilancia inteligente y los componentes del simulador predictivo cuenta con una arquitectura de contenedores que operan conforme a la arquitectura DOCKER, mostrada a continuación:

Ilustración 15: Arquitectura para despliegue de soluciones con DOCKER



Fuente: (Goffinet, 2021)

Docker Engine: Aplicación basada en la arquitectura cliente-servidor. La aplicación se instala en una máquina que hace las veces de hosting y tiene 3 componentes:

Servidor: El servidor llamado Dockerd tiene la capacidad de crear y administrar imágenes, contenedores, redes y todos los demás componentes de la arquitectura.

API - RESTFUL: Es un servicio configurable que da al servidor direccionalidad para hacer actividades.

Interfaz de línea de comandos (CLI): Es un prompt para ingresar comandos a Docker de modo manual.

Cliente Docker: Aplicación que permite la conexión de los usuarios a Docker. Cada vez que un usuario envía un comando el api-restful redirige las acciones para indicarle al servidor que debe hacer. Un mismo cliente Docker puede comunicarse con diferentes servidores y api´s.

Registros de Docker: Ubicación de memoria donde se almacenan las imágenes de Docker. Pueden configurarse registros de acceso público y/ o privado

dependiendo de los requerimientos.

Imágenes: Plantillas de solo lectura, que contienen las instrucciones para crear contendores, pueden configurarse si es el deseo o utilizarse directamente para replicar una configuración exitosa. La imagen de Docker tiene una capa base que es de solo lectura y la capa superior se puede escribir. Cuando edita un archivo docker y lo reconstruye, solo la parte modificada se reconstruye en la capa superior. (Goffinet, 2019)

Contenedores: Es el resultado de ejecutar una imagen. Es la unidad de trabajo base de todas las aplicaciones y su entorno se ejecutan dentro de un contenedor.

Volúmenes: Son datos recurrentes que se almacenan para ser utilizados en forma inmediata por los contendores de Docker de tal forma que el tiempo de inicio de las aplicaciones o funcionalidades desarrolladas será mínimo. Dentro de un volumen pueden incluirse información de ejecución de sistemas operativos como Windows y Linux.

Redes: Configuración de las formas de acceso a la información entre contenedores, existen 5 controladores de red:

Puente: Configuración de red canónica de Docker.

Anfitrión: Este controlador elimina el aislamiento de red entre los contenedores de la ventana acoplable y el host de la ventana acoplable. Se utiliza cuando no necesita ningún aislamiento de red entre el host y el contenedor.

Superposición: Esta red permite que los servicios de enjambre se comuniquen entre sí. Se utiliza cuando los contenedores se ejecutan en diferentes hosts de Docker o cuando los servicios de enjambre están formados por múltiples

aplicaciones.

Ninguna: Este controlador desactiva todas las redes.

Macvlan: Controlador que asigna direcciones MAC estáticas para los contenedores de tal forma que operen como servidores físicos.

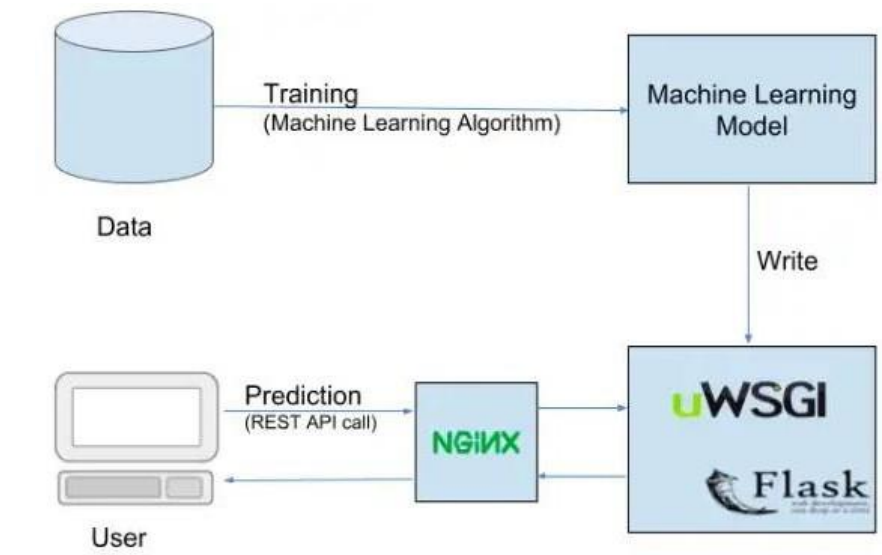
Dentro de la arquitectura DOCKER cada tablero de vigilancia, el módulo predictivo Python y el módulo de visualización descriptiva/predictiva se desplegarán en forma independiente permitiendo realizar un control individualizado de cada solución de cara a su puesta en servicio, mantenimiento y actualización.

Como resultado del análisis del proceso de despliegue se identifican productos emergentes del proceso de desarrollo del simulador predictivo, estos productos son respectivamente API´s de consumo para los modelos predictivos de demanda e interrupciones discriminadas por tipo para cada una de las dimensiones de precisión geográfica establecidas: TRANSFORMADORES, MUNICIPIO, EMPRESAS, DEPARTAMENTOS.

La estructura conceptual de las API´s y sus ventajas como servicios de consumo son descritas a continuación: Una API es una posibilidad para ofrecer servicios online permitiendo el acceso diferenciado a la información a través de métodos estructurados, elevando la seguridad y la velocidad de desempeño.

La estructura de despliegue para estos productos emergentes es mostrada en la siguiente imagen:

Ilustración 16: Despliegue de modelos de machine learning con javascript



Fuente: (Goffinet, 2019)

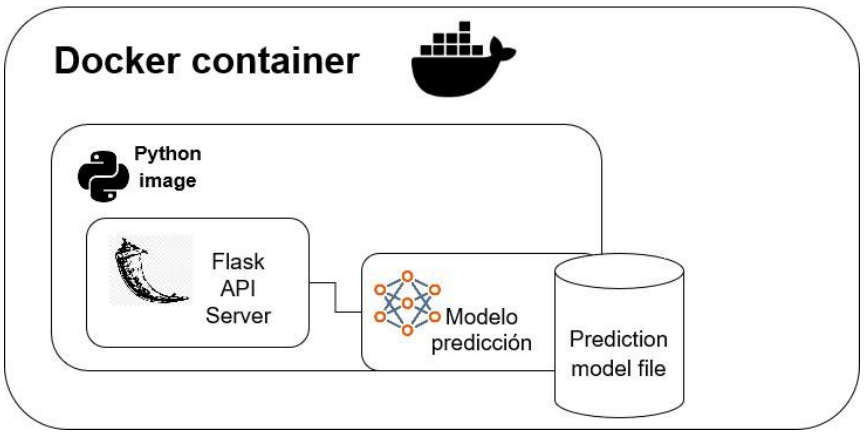
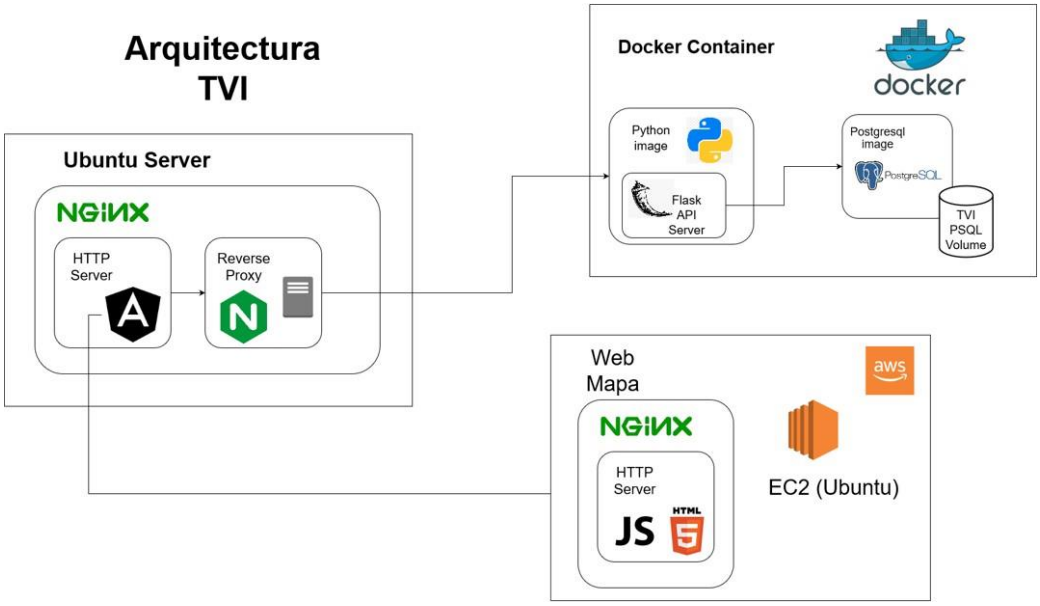


Ilustración 17: Arquitectura propuesta para el TVI utilizando DOCKER



Fuente: Elaboración propia.

7.4 Validación y apropiación del modelo con las dependencias de la SSPD.

Resultado Objetivo específico número 4

Medología de Investigación: Enfoque cualitativo.

Se realiza socialización y validación del modelo de implementación para analítica de datos, la estructura documental que soporta el modelo y la arquitectura tecnológica requerida para procesos de analítica de la delegada de Energía y Gas Combustible, con cada una de las dependencias responsables de procesos estratégicos en la SSPD: OARES, OTIC, PLANEACIÓN. De la realimentación obtenida como resultado del proceso de socialización y validación se da lugar a un proceso de apropiación y escalamiento del modelo dentro de los ejercicios de analítica institucional y se genera un procedimiento de gestión y desarrollo de productos de analítica que se encuentra registrado en el sistema institucional de

calidad.

Como parte del proceso de apropiación descrito se plantea la priorización de la aplicación del modelo de implementación al interior de la delegada de energía y gas combustible en tres productos específicos:

Tablero de vigilancia inteligente considerando las dimensiones financiera, comercial y técnica para las empresas prestadoras del servicio de energía eléctrica.

Análisis descriptivo profundo de la información histórica de interrupciones de energía eléctrica.

Simulación predictiva del impacto de proyectos de inversión considerando escenarios probabilísticos de interrupción.

7.5 Aplicación del Modelo propuesto en un ejercicio de la Superintendencia delegada de Energía y Gas combustible

Resultado Objetivo específico número 5

Corresponde a la aplicación del modelo propuesto, los cuales se pueden visualizar en cada uno de los siguientes enlaces:

7.5.1 Tablero de vigilancia inteligente

Dimensión Financiera
https://app.powerbi.com/view?r=eyJrIjoiNTA2MzUxNjItYzQxOS00OGZhLWJlMTQtYTdiZDdiMjVlM2VjIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection6c8d4782db67cc8d466d

imensión Financiera / Ranking
https://app.powerbi.com/view?r=eyJrIjoiZjY1MzQ5YTctZjhiMS00ZDBhLTkwNDYtYjg3OTgyYTYyNDA5IiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection21b1419e4d4db4a6b7c1

Dimensión Comercial / Tarifas
https://app.powerbi.com/view?r=eyJrIjoiODI3Y2ZmYzQtZWYzYS00N2VlLTg5NDQtNDg1Y

2Q1MTJ
lYWNlIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsIm
MiOjR9& pageName=ReportSection23edcd34d86c3ec43114

Dimensión Comercial / Costo Unitario
https://app.powerbi.com/view?r=eyJrIjoiMjI0ZDFmNGYtYzBjMC00ZmVmLWJlZTItNmFiM2Q4OW
ZiN2RhIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9
&pageName=ReportSection4117270c11492224d287

Dimensión Comercial / Estrategia Compra de Energía
https://app.powerbi.com/view?r=eyJrIjoiNjUyMjZiNzQtZWUwMC00NWM5LTg2MGMtNjI5ODV
mZmQ2Y2JiIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMi
OjR9&pageName=ReportSection8a6efc91c1ea7c068385

Dimensión Técnica / SAIDI_SAIFI
https://app.powerbi.com/view?r=eyJrIjoiZTkwNDAzMjYtNDkxZS00YzQ0LTg4MDAtNjgwYzM5Nzc
4YzQxIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&
pageName=ReportSection

Dimensión Técnica / DIU_FIU
https://app.powerbi.com/view?r=eyJrIjoiNTEwY2NkYWMtNWNiYS00NmViLWFhYjUtYjljNzAwYm
M3YzIzIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9
&pageName=ReportSectiond9f12db04075c988288b

7.5.2 Análisis Descriptivo de Interrupciones

Tablero Interrupciones
https://app.powerbi.com/view?r=eyJrIjoiZTlhNjVhMDAtMWY1ZC00NDkyLWE1Y2ItZjg2ZmQyYzZ
kNzU2IiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&
pageName=ReportSection

7.5.3 Simulador Predictivo de proyectos de inversión

https://servermaegei.umanizales.edu.co/dashboard-desc

8. Impactos

El desarrollo de este proyecto de investigación ha generado impactos de diferente índole en la dinámica organizacional de la SSPD:

Dinamización de los procesos organizacionales para la construcción de una visión institucional hacia una Superintendencia DATA-DRIVEN.

Fortalecimiento del modelo documental para la adopción de la metodología CRISP-DM bajo la dirección y coordinación de la oficina OARES.

Sensibilización de las demás dependencias institucionales en la identificación de necesidades particulares de analítica de datos.

Inicio de un tránsito armónico desde la transformación digital soportada en la analítica de datos hacia una trasformación cultural.

Fortalecimiento de los equipos de trabajo hacia la especialización de actividades derivadas de la aplicación del modelo propuesto.

Identificación de oportunidades de fortalecimiento de la carga de información realizada por los proveedores de servicios de energía y gas combustible, como estrategia de fortalecimiento de la analítica de datos para la inspección, vigilancia y control.

9. Conclusiones

La construcción de este trabajo tiene como propósito documentar la implementación de un ejercicio basado en un modelo de analítica de datos a nivel de la Superintendencia Delegada Para Energía y Gas Combustible - SDEGC, entidad del estado, con fines de fortalecer el ejercicio de vigilancia e inspección a través de tres herramientas funcionales que apoyen a los diferentes equipos técnicos en la toma de decisiones basados en datos. Estos ejercicios soportados metodológicamente se identifican como: i). Descriptivo TVI - Tablero de Vigilancia Inteligente -, ii) Descriptivo Interrupciones y iii). Simulador Predictivo, herramientas basadas en el desarrollo de indicadores de medición en los tópicos financieros, técnicos y comerciales que determinan el comportamiento de los prestadores de los servicios de energía y gas combustible a nivel nacional, ejercicio fundamental para identificar a los prestadores que serán objeto de vigilancia.

A partir de la implementación de este modelo y su armonización al interior de los equipos técnicos, se evidencia un cambio cultural frente al análisis de las necesidades y el tratamiento de las misma, potenciando así las decisiones basadas en ejercicios analíticos con la información reportada por los prestadores.

Con base en la arquitectura definida y socializada a las demás áreas institucionales de interés, se ha identificado la necesidad de replicar este ejercicio bajo las reglas establecidas para cada servicio público domiciliario, generando de esta forma una instancia única para el desarrollo de ejercicios analíticos al interior de la Superintendencia.

Con el uso institucional de estas herramientas, se genera una cultura de análisis de la información, dejando de lado el ejercicio de procesamiento de información, componente que consumía un recurso importante de tiempo y no permitía llegar con oportunidad a la prevención de las medidas correctivas hacía los vigilados.

Frente a la materialización de los productos que fueron expuestos al modelo de analítica de la SDEGC, en cada una de sus instancias, se evidenció que, a partir de los resultados de los diferentes indicadores, facilitaron la clasificación de empresas para su vigilancia y de esta forma ser mas eficientes en la toma de decisión de aquellas empresas que producto de estos resultados deberían ser objeto de inspección.

Por último, empoderar a los ciudadanos con estas herramientas, permite generar un espacio de academia para contar con usuarios informados y aliados en el ejercicio de vigilancia de la Superintendencia de Servicios Públicos Domiciliarios.

10. Recomendaciones

A partir de los resultados obtenidos en el contexto institucional y las evidencias de los resultados de los modelos de analítica implementados, emergen las siguientes recomendaciones:

Con el fin de apoyar el proceso de adopción institucional del modelo de implementación para proyectos de analítica de datos es necesario fortalecer el modelo documental propuesto bajo la estructura del modelo de calidad institucional y elevarlos al status de formato institucional, debidamente registrado y aprobado mediante acto administrativo.

Fomentar la discusión institucional en la vía de la transformación digital y en el enfoque Data-Driven de la institución de tal forma que la adopción del modelo propuesto y su manifestación en la estructura organizacional y de procesos de la SSPD se materialice en el mediano plazo.

Establecer mesas de trabajo técnicas institucionales con propósito de establecer las instancias y responsables de la unidad de analítica abierta y digital propuesta en el modelo, para establecer los requerimientos tecnológicos y procedimentales que se deban adelantar para la implementación en la Superservicios, a partir de la experiencia desarrollada en la Delegada de Energía y Gas Combustible.

Los modelos descriptivos y predictivos desarrollados requieren de actividades periódicas de actualización de la información y reentrenamiento que requieren transitar el camino hacia la automatización en la medida que la arquitectura tecnológica propuesta sea implementada a cabalidad.

Consolidar los diferentes modelos a nivel del sector, academia y usuarios, para de esta forma convertir a la superintendencia en un referente nacional de consumo de información para la toma de decisiones frente a los servicios públicos domiciliarios.

11. Referencias

Congreso de Colombia (1994) Ley 142 de 1994 por la cual se establece el régimen de los servicios públicos domiciliarios y se dictan otras disposiciones. Bogotá, D.C: Congreso de Colombia.

Congreso de Colombia (2001) Ley 689 de 2001 por la cual se modifica parcialmente la Ley 142 de 1994. Bogotá, D.C: Congreso de Colombia.

Departamento Nacional de Planeación (2002) Documento CONPES 3168 Estrategia para la puesta en marcha del Sistema Único de Información de los Servicios Públicos Domiciliarios. Bogotá D.C: Republica de Colombia

Quintero Bernardo Juan (2018) Analítica de datos para sistemas de costos basados en actividades en la era de big data. Dialnet. N. Extra 1. 6

Constitución Política de Colombia (1991) Articulo 370 otorga al Presidente de la República ejercer por medio de la Superintendencia de Servicios Públicos Domiciliarios el control, inspección y vigilancia de las entidades que presten el servicio público domiciliario. Bogotá D.C.: Congreso de Colombia

Congreso de Colombia (1994) Ley 142 de 1994 Articulo 69 numeral 2 por la cual se crea la Comisión Reguladora de Energía y Gas Combustible. Bogotá, D.C: Congreso de Colombia.

Miranda Ortega Lucy Maria. (2012) Consideraciones sobre el modelo de las autoridades administrativas de regulación, inspección, vigilancia y control de los servicios públicos domiciliarios de energía y gas en Colombia (Trabajo de Grado). Universidad del Norte, Barranquilla, Colombia

Castro Dulcey Fanny Gineth (2020) El procedimiento administrativo sancionatorio en el marco de las empresas de servicios públicos domiciliarios: aproximación teórica a sus postulados básicos. Revista IUSTA. Número 53. 8-9

Superintendencia de Servicios Públicos Domiciliarios (2010) La Superintendencia Delegada de Energía y Gas, publica el Acto administrativo unificado para el reporte de información al SUI de las empresas del sector de Energía Eléctrica, Bogotá D.C: Superintendencia de Servicios Públicos Domiciliarios.

Comisión Reguladora de Energía y Gas (2008) Por la cual se aprueban los principios generales y la metodología para el establecimiento de los cargos por uso de los Sistemas de Transmisión Regional y Distribución Local. Bogotá D.C: CREG

Superintendencia de Servicios Públicos Domiciliarios (2019) Por la cual se expiden los lineamientos para el cargue de información al Sistema Único de Información - SUI aplicable a los prestadores del servicio público de energía eléctrica del Sistema Interconectado Nacional - SIN. Bogotá D.C: Superintendencia de Servicios Públicos Domiciliarios.

Comisión Reguladora de Energía y Gas (2018) Por la cual se establece la metodología para la remuneración de la actividad de distribución de energía eléctrica en el Sistema Interconectado Nacional. Bogotá D.C: CREG

Claudia Hernández, Jorge Enrique Rodríguez Rodríguez (2018) Procesamiento de datos estructurados. Revista vínculos. Volumen 4 Número 2. doi: https://doi.org/10.14483/2322939X.4123

Aristizabal. F. Jorge Alexander (2016) Analítica de datos de aprendizaje (ADA) y gestión educativa. Revista Gestión Educación Escuela de Administración Educativa. Volumen 6 Número 2. Doi: DOI 10.15517/RGE.V1I2.25499

Alveiro Alonso Rosado Gómez, Dewar Willmer Rico Bautista (2010) Inteligencia de negocios: Estado del arte. Revista Scientia et Technica. Volumen 1 Número 44. Doi: https://doi.org/10.22517/23447214.1803

Superintendencia de servicios públicos Domiciliarios (2005 – 2009) por la cual se reporta de información comercial básica del sector del gas licuado de petróleo. Bogotá D.C: Superintendencia de Servicios Públicos Domiciliarios

Superintendencia de servicios públicos Domiciliarios (2014 -2003) por la cual se establece el reporte de información de facturación. Bogotá D.C: Superintendencia de Servicios Públicos Domiciliarios

Superintendencia de servicios públicos Domiciliarios (2016) por la cual los comercializadores mayoristas, transportadores, distribuidores y comercializadores minoristas de GLP, deben reportar la información correspondiente a las actividades que desarrollan de

conformidad con el marco regulatorio vigente y las metodologías tarifarias respectivas. Bogotá D.C.

Moine, J. M., Gordillo, S. E., & Haedo, A. S. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. In Congreso Argentino de Ciencias de la Computación (Vol. 17).

© Copyright IBM Corporation, (2021). SPSS Modeler. IBM Documentación. Recuperado de: https://www.ibm.com/docs/es/spss-modeler/saas?topic=deployment-overview [16 de noviembre de 2022].

Ministerio de Tecnologías y Comunicaciones de la República de Colombia, Departamento Administrativo de la Función Pública, Departamento Nacional de Planeación, Escuela Superior de Administración Pública (2022), Diplomado Servidor Público 4.0, El poder de los datos en el sector Público. Bogotá, Colombia.

Salvador Shinji, L. S. (2020). Implementación de un sistema de monitoreo en tiempo real con simulación predictiva para sistemas de potencia.

Guerrero, J. F. J., Fernández, R. S., & Abad, J. C. G. (2006). La capacidad predictiva en los métodos Box-Jenkins y Holt-Winters: una aplicación al sector turístico. Revista Europea de Dirección y Economía de la Empresa, 15(3), 185-198.

Departamento Administrativo de la Función Pública (2021) Guía para la analítica de datos y su uso en la planificación y ejecución de auditorías internas basadas en riesgos versión julio. Bogotá, Colombia.

Correa, M., Bielza, C., Pamies-Texeira, J., & Alique López, J. R. (2008). Redes Bayesianas vs redes neuronales en modelos para la predicción del acabado superficial.

Zapata, Cj, Piñeros, Lc, & Castaño, Da (2004). El Método De Simulación De Montecarlo En Estudios De Confiabilidad De Sistemas De Distribución De Energía Eléctrica. Scientia et Technica , X (24), 55-60.

Lund, M. I., Migani, S. I., Vera, C., Orellana Vasallo, A., Gómez, A. M., Pinto, S. E., ... & Molinari, M. L. (2021). Inteligencia y analítica de negocios para la toma de decisiones en diferentes

contextos. In XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja).

Arenas López, M. C., & Gómez Montes, A. M. (2017). Inteligencia de negocios aplicada a los procesos de autoevaluación de la Universidad de Manizales.

Mavesoy Murcia, C. D. (2019). Modelo basado en CRISP-DM extendido mediante prácticas de metodologías ágiles para proyectos medianos de analítica de datos.

Galán Cortina, V. (2016). *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario* (Bachelor's thesis).

Presidencia de la República de Colombia (2020) Decreto 1369 de 2020 Por el cual se modifica la estructura de la Superintendencia de Servicios Públicos Domiciliarios. Bogotá, D.C: Presidencia de la República.

Presidencia de la República de Colombia (2020) Decreto 1369 de 2020 Por el cual se modifica la estructura de la Superintendencia de Servicios Públicos Domiciliarios. Bogotá, D.C: Presidencia de la República.

Tulang Alfeo, Bello Alwielland. 2022. Forecasting Power Load Demand using Holt-Winters Model. International Journal of Educational Research for Higher Learning Volume 24 Number 2 October 2018. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4107011

Yue M, Toto T. , Jensen M, Giangrande S, Lofaro R. 2017. A Bayesian Approach Based Outage Prediction in Electric Utility Systems Using Radar Measurement Data. IEEE Transactions on Smart Grid July 2017.

Trejo Daniel. 2021. Transformación digital y administración del conocimiento para directores. Editorial DIDAC, Mexico.

Maffeo Lauren. 2023. Designing Data Governance from the Ground Up. Pragmatic Programming Series.

Mintic, Gobierno de la República de Colombia. 2020. Marco de la Transformación Digital para el Estado Colombiano. Retrieved from: https://www.mintic.gov.co/portal/715/articles-149186_recurso_1.pdf

Parr, Olivia. 2000. Data mining cookbook. Obtenido de http://books.google.com.co/books?id=L3w0loZrcU0C&printsec=frontcover&dq=Data+Mining+Cookbook#v=onepage&q=&f=false.

Cameron Ian, Gani Rafiqul. 2011. Product and Process Modelling A Case Study Approach. Editorial Elsevier. Paises Bajos.

Goffinet Francois. 2019. Orchestration de conteneurs, Docker – Swarm – k8s. Obtenido de: https://goffinet.gitlab.io/docker-k8s-training/containers.goffinet.org.pdf

Cevallos, C. J. V., & Párraga, D. M. (2021). Inteligencia de Negocios para las Organizaciones. Revista Arbitrada Interdisciplinaria Koinonía, 6(12), 304-333.

Tapia, H. A., Erazo, J. C., Narváez, C. I., & Matovelle, M. M. (2020). Estrategias para fomentar el emprendimiento y desarrollo empresarial [Strategies to promote entrepreneurship and business development]. Revista Arbitrada Interdisciplinaria Koinonía, (5),10, 833-861. http://dx.doi.org/10.35381/r.k.v5i10.837

Anexos:

A. Tesauro Institucional Ciencia de datos

Anexo1_Tesaurio_CRG.pdf

B. Aplicación del modelo documental al TVI

Anexo2_Documentacion_TVI_CRG.pdf

C. Aplicación del modelo de implementación al análisis descriptivo de interrupciones y simulador predictivo

Anexo3_Implementacion_Modelo_CRG.pdf

D. Anexos Visualizaciones, Power BI

Dimensión Financiera

https://app.powerbi.com/view?r=eyJrIjoiNTA2MzUxNjItYzQxOS00OGZhLWJlMTQtYTdiZDdiMjVlM2VjIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection6c8d4782db67cc8d466d

Dimensión Financiera / Ranking
https://app.powerbi.com/view?r=eyJrIjoiZjY1MzQ5YTctZjhiMS00ZDBhLTkwNDYtYjg3OTgyYTYyNDA5IiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection21b1419e4d4db4a6b7c1

Dimensión Comercial / Tarifas
https://app.powerbi.com/view?r=eyJrIjoiODI3Y2ZmYzQtZWYzYS00N2VlLTg5NDQtNDg1Y2Q1MTJlYWNlIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection23edcd34d86c3ec43114

Dimensión Comercial / Costo Unitario
https://app.powerbi.com/view?r=eyJrIjoiMjI0ZDFmNGYtYzBjMC00ZmVmLWJlZTItNmFiM2Q4OWZiN2RhIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection4117270c11492224d287

Dimensión Comercial / Estrategia Compra de Energía
https://app.powerbi.com/view?r=eyJrIjoiNjUyMjZiNzQtZWUwMC00NWM5LTg2MGMtNjI5ODVmZmQ2Y2JiIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection8a6efc91c1ea7c068385

Dimensión Técnica / SAIDI_SAIFI
https://app.powerbi.com/view?r=eyJrIjoiZTkwNDAzMjYtNDkxZS00YzQ0LTg4MDAtNjgwYzM5Nzc4YzQxIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection

Dimensión Técnica / DIU_FIU
https://app.powerbi.com/view?r=eyJrIjoiNTEwY2NkYWMtNWNiYS00NmViLWFhYjUtYjljNzAwYmM3YzIzIiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSectiond9f12db04075c988288b

Tablero Interrupciones
https://app.powerbi.com/view?r=eyJrIjoiZTlhNjVhMDAtMWY1ZC00NDkyLWE1Y2ItZjg2ZmQyYzZkNzU2IiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9&pageName=ReportSection

Tablero proyectos de inversión
https://servermaegei.umanizales.edu.co/dashboard-desc