INSTITUTO COLOMBIANO DE BIENESTAR FAMILIAR

GUÍA METODOLÓGICA PARA LA IMPLEMENTACIÓN DE PROYECTOS DE ANALÍTICA

TABLA DE CONTENIDO

1. Introducción

El Plan de Analítica Institucional del Instituto Colombiano de Bienestar Familiar – ICBF propone la implementación de ejercicios de analítica o minería de datos, cuyo planteamiento y ejecución constante requiere de la aplicación de una metodología de trabajo que abarque no sólo los procesos de modelado, optimización y análisis estadístico, sino que además se integre con las etapas de obtención, limpieza y comprensión de los datos. La presente guía pretende establecer los lineamientos necesarios, fases y actividades mínimas para desarrollar proyectos, análisis o ejercicios de analítica al interior del Instituto, a través de la combinación de metodologías desarrolladas por IBM y Microsoft como la Cross-Industry Standard Process for Data Mining (CRISP-DM), -así como su versión más actualizada, Analytics Solutions Unified Method (ASUM)- y la Team Data Science Process (TDSP).

El contenido se estructura de la siguiente manera: en la primera parte se presentan los objetivos y el alcance, en la segunda parte se aborda el marco de referencia que comprende la revisión de los antecedentes y en la última parte, se describe el desarrollo de un ejercicio de analítica en sus diferentes fases.

2. Objetivo

Establecer los lineamientos metodológicos, fases y las actividades necesarias para desarrollar ejercicios de analítica, incluyendo su formulación, su implementación, su seguimiento y su evaluación en el Instituto Colombiano de Bienestar Familiar.

3. Alcance

La presente guía aplica a los procesos de gestión y procesamiento para el desarrollo de ejercicios de analítica en el Instituto Colombiano de Bienestar Familiar a nivel nacional.

4. Definiciones

Metodología: Estrategia general que sirve de guía para los procesos y actividades que están dentro de un dominio determinado. No depende de tecnologías ni herramientas específicas, ni es un conjunto de técnicas o recetas. La metodología proporciona un marco sobre cómo proceder con los métodos, procesos y argumentos que se utilizarán para obtener respuestas o resultados.

Analítica de datos: Disciplina orientada a analizar datos mediante técnicas científicas y herramientas automatizadas con énfasis en identificar hechos, relaciones, patrones ocultos de comportamiento de variables, correlaciones y tendencias, que brindan conocimiento respecto de los fenómenos de la realidad que antes permanecían ocultos debido a la complejidad de su medición y análisis por otros medios. De acuerdo con el objetivo que busca, la analítica puede ser descriptiva, diagnóstica, predictiva o prescriptiva.

Ciencia de datos (Data Science): Un campo de investigación y práctica que se enfoca en resolver problemas del mundo real, a menudo usando grandes cantidades de datos y combinando habilidades de diferentes áreas del conocimiento: matemáticas, ciencias de computación, estadísticas, ciencias sociales e incluso periodismo de datos o arte.

Big data: Conjunto de aplicaciones, técnicas, tecnologías e iniciativas que dan lugar a la generación de valor con base en los datos con gigantescas cantidades de información.

5. Antecedentes

El Plan de Analítica Institucional del ICBF tiene como propósito fortalecer al Instituto en las dimensiones organizacional, tecnológica y de cultura de datos, de forma que mejoren la calidad de los procesos, tratamiento y difusión de las estadísticas oficiales generadas. Además, estableció como uno de sus grandes objetivos sentar las bases metodológicas, éticas y culturales del desarrollo de buenas prácticas de analítica institucional. Este plan contempló las bases para la política del “aprovechamiento de datos, mediante el desarrollo de las condiciones para que sean gestionados como activos para generar valor social y económico.” (Conpes 3920 de 2018). De igual forma, adoptó las recomendaciones de la consultoría que realizó el DNP a través de Economía Urbana y Galileo para la construcción del índice de la capacidad del Instituto en la explotación de datos (Imbigda, índice de madurez de Big Data, 2019).

El índice de madurez midió la profundidad de la implementación de aspectos relacionados con el ámbito organizacional, cultural, tecnológico, jurídico y ético. Estas dimensiones comprenden el entendimiento y aplicación por parte de la entidad tanto de los preceptos normativos relacionados con la protección de datos personales, la transparencia y el gobierno abierto, como de las consideraciones éticas que pueden traer consigo la aplicación de algoritmos o de otras técnicas de análisis.

La evaluación realizada a través de este índice le permitió al ICBF plantearse metas hacia el 2022 en materia de explotación de datos, que involucran el compromiso de diversas áreas misionales, de grupos de profesionales de diferentes áreas del conocimiento, y que igualmente plantean la necesidad de una metodología para el desarrollo de buenas prácticas que permita trazar un plan de trabajo apropiado con metas fijas y claras en el proceso.

6. Metodologías para la implementación de ejercicios analíticos.

La presente guía de analítica se realizó tomando como base las metodologías Cross-Industry Standard Process for Data Mining (CRISP-DM), Analytics Solutions Unified Method (ASUM) desarrolladas por IBM y Team Data Science Process (TDSP) desarrollada por Microsoft. También tuvo en cuenta algunas consideraciones éticas y legales que se hacen cada vez más indispensables para la aplicación de técnicas de analítica en el estudio de fenómenos sociales, tal y como es el caso del ICBF.

6.1. Consideraciones éticas y legales previas

Todos los ejercicios de analítica desarrollados al interior del ICBF deberán ceñirse a las disposiciones reglamentarias y normativas existentes en el país en materia de protección de datos personales y uso ético de la información. También deberá ajustarse a las políticas operacionales de los procedimientos de procesamiento, gestión y análisis de información del Instituto Colombiano de Bienestar Familiar (P16.GTI, P17.GTI, P18.GTI).

Adicionalmente, el uso de los algoritmos de modelación utilizados deberá hacerse de manera responsable y considerando los posibles aspectos discriminatorios que podría traer consigo su aplicación. En este sentido, deberá prevalecer para su uso los principios éticos de justicia y beneficio hacia la población objetivo.

El desarrollo o implementación de los ejercicios analíticos no deberá desencadenar situaciones de discriminación y odio. En caso de que se identifiquen riesgos de este tipo, el equipo que desarrolla el ejercicio analítico deberá acudir a técnicas para su mitigación a

través de la entrega agregada de resultados o de la modificación parcial o total del método o artefacto de implementación de los resultados del modelo. Para esto, el ICBF también adoptará la Guía ética para la implementación de algoritmos analíticos (G24.GTI) y otros documentos de Gobierno y de buenas prácticas internacionales en materia.

6.2. Etapas de la Metodología

La metodología de proyectos de analítica planteada se encuentra compuesta por cinco (5) etapas que son: i) Comprender el problema; ii) Gestionar los datos; iii) Modelar; iv) Implementar; v) Retroalimentar.

6.2.1. Etapa I – Comprender el problema

Esta es quizá la etapa más importante, por cuanto implica establecer con claridad los objetivos del ejercicio analítico a partir de la comprensión profunda de la necesidad del usuario final de los resultados. Para atacar con éxito esta etapa es necesario resolver las siguientes preguntas:

¿Cuál es el principal objetivo del ejercicio?
¿Este objetivo implica categorizar, detectar anomalías, pronosticar o llegar a alguna recomendación?
¿Cuánto tiempo y recursos se tienen para desarrollar el ejercicio?
¿Cuál sería el artefacto más adecuado para consumir los resultados del ejercicio analítico? ¿Web service en tiempo real de un modelo, un front-end, etc.?

Como resultado de esta primera etapa el equipo deberá contar con los siguientes productos:

a) Un documento guía situacional del proyecto que contenga en tiempo real el avance del ejercicio. Inicialmente contendrá los objetivos y las preguntas que se pretende resolver a través del ejercicio analítico, así como una revisión inicial de la literatura que permita inferir posibles variables o tipo de información que podría utilizarse para el problema o cuestionamiento que intenta resolverse. También contendrá los puntos c), d) y e) planteados en este listado.
b) Una herramienta de planeación del proyecto (tipo Microsoft Planner, por ejemplo) que establezca: i) los miembros del equipo que adelantará el proyecto y su rol; ii) las fases generales del proyecto con el plazo de ejecución pactado por el equipo; iii) actividades preliminares que son requeridas para cumplir con cada fase.
c) Un repositorio compartido entre los miembros del equipo, en el que se puedan consolidar versiones de código, documentación soporte de la problemática, entre otros. Pueden usarse herramientas libres como Git, GitHub, u otros como DevOps de Azure.

d) Una identificación inicial de datos disponibles (raw data) para el ejercicio junto con sus respectivos metadatos y diccionario de datos.
e) Una identificación del enfoque analítico que se implementará, así como de la(s) variable(s) objetivo del ejercicio: detección de anomalías, predicción, prospección, descripción, clasificación. Esto permitirá seleccionar más fácilmente el tipo de modelado que se implementará durante la etapa III.
f) Una identificación inicial del diseño de la solución a través de la cual se consumirán los resultados del modelo/ejercicio analítico propuesto.

6.2.2. Etapa II – Gestionar los datos

En esta etapa el equipo dispone de los datos en un medio apropiado para la solución planteada, los explora en busca de anomalías o situaciones que deben ser corregidas antes de seguir con la modelación, y establece un nivel de comprensión profunda de la información con la que se cuenta. De este modo, comprende:

i) La ingestión de datos en un medio apropiado para el desarrollo de la solución de consumo y gestión de resultados del ejercicio analítico: Involucra disponer los datos en un servicio tecnológico que permita implementar la solución diseñada para consumo de los datos. Por ejemplo, disponerlos en un servidor de Amazon, en una estructura de consumo local, entre otros.
ii) Comprender los datos: Implica realizar un proceso de auditoría de datos en busca de datos perdidos, outliers o datos anómalos, así como un proceso estadístico de análisis de distribución, relación entre variables y la variable objetivo, entre otros.
iii) Feature-engineering previa: Aunque este paso está ampliamente ligado al tipo de modelo, algunos hallazgos de la actividad de comprensión de los datos pueden llevar a la conclusión de que los datos deben ser limpiados o procesados previamente antes de la etapa del modelaje. En esta actividad se incluyen los procesos de merge o fusión y agregación de bases de datos, así como la imputación de algunas variables.

Como resultado de esta etapa el equipo deberá contar con los siguientes productos:

a) Un reporte de calidad de datos que establezca de forma concisa los principales hallazgos del punto ii)
b) Arquitectura de datos de la solución, que corresponde a la descripción (bien usando un diagrama u otro medio que se considere apropiado) de las fuentes de datos y disposición de ellos en ambiente que permita el desarrollo de la solución, así como las instrucciones o recomendaciones que faciliten el reentrenamiento del modelo usado.
c) Un notebook con código que contenga descripción inicial de los datos que permita establecer de forma gráfica y usando pruebas estadísticas, las relaciones entre las variables (correlación, por ejemplo), la significancia estadística individual con relación a la (s) variable (s) objeto de estudio (a través de pruebas como t-student, chi-

cuadrado, F-Fischer, entre otros), así como la distribución estadística de las variables más importantes.

6.2.3. Etapa III – Modelar

En la etapa de modelado el equipo se enfrenta a la primera versión del conjunto de datos preparados en la etapa II, e intenta hacer que estos ofrezcan información relevante y útil para resolver el problema planteado. Modelar no necesariamente implica predecir. El modelado puede tener como objeto la clasificación de una población en grupos, la detección de alguna anomalía entre los datos con el ánimo de establecer características comunes en ciertas personas o fenómenos que se quieran describir, el establecimiento de estrategias de acción para el futuro a través de construcción artificial de escenarios, o simplemente la descripción y consulta de datos relevantes que no se habían detectado por el gran volumen de información disponible. Cualquiera que sea el objetivo planteado, el modelaje requiere del desarrollo de las siguientes actividades:

i) Feature engineering final: En esta actividad se crean, adicionan o transforman variables con el objetivo de que se ajusten a los requerimientos temáticos y técnicos de los métodos de modelación disponibles. Se trata de encontrar un balance entre la experiencia o conocimiento temático de la problemática, la información disponible y la utilidad estadística de las variables, de modo que se minimice el ruido introducido a través de demasiadas variables estadísticamente inútiles que sobreentrenan a los modelos. Se recomienda usar técnicas de suavizamiento para la selección de variables no solo de tipo filtro como correlación, chi-cuadrado, etc., sino de wrapper (como Sequential Forward Selection, SBS, Recursive Feature Elimination) o embedded (como L1, L2, elastic net).
ii) Entrenamiento, selección y evaluación del modelo: El proceso de entrenamiento, selección y evaluación del modelo comprende una serie de actividades iterativas incrementales que dependen de lo encontrado en la etapa II: dividir los datos entre entrenamiento y evaluación de forma aleatoria; construir el modelo usando los datos de entrenamiento; evaluar los resultados y supuestos del modelo tanto en la partición de entrenamiento como en la de evaluación; determinar la mejor alternativa o el mejor modelo a partir de la comparación de métricas de los distintos modelos.

Existen distintos tipos de modelos que dependen de los objetivos del ejercicio analítico. Los principales tipos de modelos son los siguientes:

- Aprendizaje supervisado: cada dato se encuentra asociado con una categoría o valor de interés para el modelo. Un ejemplo de asignación de etiquetas es el precio asociado a un carro determinado. El objetivo de este tipo de modelos es estudiar las diferentes variables y realizar predicciones o construir escenarios a partir del compartimiento futuro.
- Aprendizaje no supervisado: En estos modelos se asume que no existen etiquetas o valores predeterminados en los datos y el objetivo es justamente organizarlos,

clasificarlos, encontrar su estructura. El principal grupo de modelos que clasifican dentro de esta categoría son los modelos de clústeres como k-medias (*k-means*).

- Aprendizaje reforzado: En el aprendizaje reforzado el algoritmo trata de tomar una decisión distinta en respuesta al dato específico al que se enfrenta. Es utilizado principalmente en robótica y en inteligencia artificial detrás de botchats, entre otros.

Los productos de esta etapa son los siguientes:

a) Códigos implementados compartidos en el repositorio (usando herramientas como Git, GitHub, DevOps, etc.), así como descripción del feature-engineering realizado.
b) Reporte del modelo: Cada uno de los modelos entrenados o probados debe tener una tabla informativa que contenga sus principales métricas de evaluación, de forma que permita en el futuro sustentar la decisión tomada por el equipo.

6.2.4. Etapa IV – Implementar

Esta etapa implica darles vida a los resultados obtenidos en la etapa III – Modelar, de forma que no se queden en un repositorio, sino que sean utilizados por el equipo de la forma más fácil y útil posible a sus propósitos. Para exponer los resultados para el consumo por otras aplicaciones, se considera útil el desarrollo de una interfaz API abierta. Las aplicaciones que podrían consumir los resultados a partir de este desarrollo podrían ser:

- Sitios web
- Hojas de cálculo
- Dashboard
- Aplicaciones front-end
- Aplicaciones back-end

Cualquiera que sea el artefacto de implementación seleccionado, es imperativo recurrir a las disposiciones y reglamentaciones establecidas al interior del Instituto Colombiano de Bienestar Familiar para el desarrollo de este tipo de alternativas. Para tales efectos, es necesario consultar lo dispuesto en:

- Procedimiento para desarrollo y mantenimiento de Sistemas de Información, P6.GTI
- Guía de arquitectura de referencia para proyectos de analítica, G17.GTI

En caso de que los recursos y/o el alcance del ejercicio de analítica no impliquen la implementación a través de consumo activo o en tiempo real de los resultados, el equipo deberá asegurarse de que se dispongan de la forma más expedita y amigable posible al equipo o área misional que utilizará los resultados para el alcance de sus objetivos.

Los principales productos de esta etapa (aunque no los únicos, por cuanto algunas soluciones de implementación pueden ser más complejas que otras) son los siguientes:

a) Un reporte del modelo final seleccionado con documentación detallada de su implementación, dispuesto en un repositorio de acceso libre a quienes van a usar/evaluar los resultados.
b) Un documento de solución arquitectónica final de la implementación: este deberá considerar los aspectos normativos y reglamentarios del ICBF en cuanto al desarrollo de software o aplicativos nuevos se refiere.

6.2.5. Etapa V – Retroalimentar

En esta última etapa del ejercicio, el equipo deberá asegurarse de comunicar los resultados del modelo y sus principales parámetros al mayor grupo de expertos posibles, en busca de validación y crítica del enfoque utilizado, las variables, entre otros. También deberá tener una comunicación constante con el equipo o área misional que usa los resultados para el alcance de sus objetivos, de modo que pueda contar de primera mano con una validación “en campo” del desempeño del modelo.

Todo lo anterior permitirá al equipo mejorar el modelo y obtener versiones mejoradas o robustecidas que incrementen la utilidad misional de sus resultados.

El producto de esta etapa es el siguiente:

a) Retroalimentación constante del documento guía situacional del proyecto, el cual como se mencionó en la etapa I, deberá contener en tiempo real el avance del ejercicio.

7 DOCUMENTOS DE REFERENCIA

- DANE (2018). Metodología para el desarrollo de planes estadísticos.
- ICBF (2020). Plan de Analítica Institucional - Instituto Colombiano De Bienestar Familiar.
- P17.GTI Procedimiento para el procesamiento y generación de información estadística
- P18.GTI Procedimiento para generar análisis de información estadística oficial

- IBM (2015). Metodología Fundamental para la Ciencia de Datos. https://www.ibm.com/downloads/cas/6RZMKDN8
- IBM (2016). Analytics Solutions Unified Method (ASUM). ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf
- MICROSOFT (2020). Teams Data Science Process. https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

8 CONTROL DE CAMBIOS

Fecha	Versión	Descripción del Cambio
N/A	N/A	N/A