Criterios de explotación de datos en entidades públicas para su transparencia

JUNIO DE 2019

Departamento Nacional de Planeación (DNP)

Dirección General
Luis Alberto Rodríguez Ospino (2019, actual)
Gloria Amparo Alonso Másmela (2018-2019)

Subdirección General Sectorial
Daniel Gómez Gaviria (2019, actual)
Rafael Puyana Martínez

Subdirección General Territorial
Amparo García Montaña (2019, actual)
Diego Rafael Dorado Hernández (2018-2019)

Secretaria General
Diana Patricia Ríos García (2019, actual)
Jenny Fabiola Páez Vargas, (2018-2019)

Dirección de Desarrollo Digital
Iván Mauricio Durán Pabón (2020, actual
Paola Andrea Bonilla Castaño (2018-2019)

Subdirección Prospectiva Digital
Iván Mauricio Durán Pabón (2018-2020)
Viviana Rocio Vanegas Barrero (2020 actual)

Este estudio ha contado con el apoyo de los siguientes asesores
Carlos Andrés Rocha Ruiz
Eduardo Escobar Gutierrez

Criterios de explotación de datos en entidades públicas para su transparencia
Documento final

Grupo de Comunicaciones y Relaciones Públicas
Luis Segundo Gamez Daza
Coordinador

Calle 26 núm. 13-19
Bogotá, D.C.
PBX: 3815000
www.dnp.gov.co

Contenido

Glosario de términos

Big Data: Término usualmente utilizado para la descripción de una serie de conceptos que engloba la capacidad tecnológica de almacenar, agregar y procesar grandes volúmenes de datos de manera eficiente[1].

Caja blanca: Es el elemento detallado de la lógica interna y la estructura de un código, modelo o algoritmo del cual se tiene total conocimiento de su funcionamiento[2].

Caja negra: Es el elemento en el cual no se tiene ningún conocimiento del funcionamiento interno de la aplicación, modelo o algoritmo. En este se examina los aspectos fundamentales del sistema sin dar relevancia a la estructura lógica interna del sistema[2].

Algoritmo: Es un conjunto de pasos e instrucciones lógicas predefinidas y finitas para llevar a cabo un proceso matemático, estadístico, computacional, etc.[3].

Modelo: Es una formulación matemática de la relación entre variables, parámetros, atributos, etc. Siendo una extracción sucinta de la realidad con el objetivo de estudiar el comportamiento o los factores clave de un fenómeno económico, natural, social, etc[3].

Entrenamiento del modelo: El proceso por el cual se cuantifica el nivel de ajuste de las predicciones de un modelo dadas las observaciones y se obtiene la memoria de largo plazo del modelo[4].

Pruebas del modelo: Proceso en el cual se cuantifica la exactitud que tiene el modelo a las predicciones dadas nuevas observaciones de modo que el modelo no fue entrenado[4].

[1] (Mauro, 2015)
[2] (Khan & Khan, 2013)
[3] (RAE, 2018)
[4] (James, Witten, Hastie, & Tibshirani, 2013)

1 Introducción

La implementación y ejecución de un proyecto de analítica de datos suele tener diversos enfoques y estos dependen en gran medida de la naturaleza del proyecto y de la disponibilidad de los datos para el desarrollo de este. Garantizar su transparencia en el sector público no solo es necesario, sino deseable debido a la naturaleza de las políticas públicas que puedan diseñarse e implementarse a partir de los resultados de un proyecto.

Así mismo, gran cantidad de entidades públicas y privadas concentran una parte de sus recursos en el desarrollo y el fortalecimiento de áreas de analítica y en mantener una cultura de datos, por tal motivo, los incentivos existentes en generar, implementar y ejecutar proyectos de analítica en diversas entidades están latente, lo cual conlleva a un alto crecimiento en este sector.

La variedad de opciones por la cual se puede abordar un proyecto de analítica de datos debe estar enfocado y canalizado hacia el objetivo final del proyecto, evitando caer en una ambigüedad y subjetividad dentro de la elección de un algoritmo, modelo, esquema, etc. Para ello se requiere, en un principio, definir aquello que se considere como la opción preferible en la elección de los algoritmos y evitar posibles dificultades futuras en el proceso.

El criterio para la elección destacable de los algoritmos depende en una medida del objetivo final del proyecto y de los resultados que el investigador busca encontrar. Por un lado, es necesario tener en cuenta diferentes escenarios y posibilidades que puedan llevar a dar solución al proyecto de una manera satisfactoria y efectiva. De tal modo que, con la implementación de una solución de analítica para una problemática, esta sea beneficioso.

Dentro del marco de la mejor elección de la variedad de opciones posibles por la cuales abordar un proyecto de analítica de datos, **la transparencia de los algoritmos** resulta ser un factor de gran relevancia en su desarrollo por motivos relacionados al objetivo final de los proyectos, como lo puede ser factores de **interpretabilidad de los algoritmos** o la **reproducibilidad de los algoritmos**. Para ello, es necesario tener en cuenta la diferenciación entre los diferentes algoritmos existentes, el nivel de

transparencia de acuerdo con su interpretabilidad y como estos pueden ser utilizados en un proyecto de analítica en entidades públicas.

En este documento busca dar las herramientas necesarias para la implementación y ejecución de proyectos de analítica que garantice la transparencia y responsabilidad de los algoritmos, así como las buenas prácticas utilizadas por el investigador en el ejercicio de la implementación de modelos de aprendizaje automático para la solución de un problema o la evolución de un proyecto e identificar unos criterios a tener en cuenta por parte de las entidades públicas que implementen proyectos de explotación de datos y así, lograr un estándar de ejecución de dichos proyectos que garanticen transparencia y responsabilidad en los mismos.

Las necesidades de tener una guía de ruta y unos parámetros claros en la implementación de proyectos de analítica son conclusiones obtenidas a partir de la "Mesa de trabajo transparencia de los proyectos de explotación de datos y criterios que deberían incorporarse en un marco ético", realizada en la Universidad Externado de Colombia en junio de 2018 (Anexo 1).

2 Identificación de la necesidad de transparencia y responsabilidad en los algoritmos

La penetración de los algoritmos de inteligencia artificial en la sociedad lleva consigo una serie de retos por los cuales esta debe ser canalizada y ejecutada de tal modo que oriente su potencial beneficio a un bien común y minimice los potenciales riesgos que lleve consigo su implementación. El alcance de la Inteligencia Artificial (IA) y de todos los algoritmos de aprendizaje automático se centra en la capacidad de describir un fenómeno, predecirlo o verlo desde un punto de vista de analítica prescriptiva, descriptiva o predictiva, por tal motivo, el alcance de la IA y la aplicabilidad de esta en la sociedad tiene diversos factores a considerar para garantizar la responsabilidad y transparencia, mitigando los posibles riesgos de su implementación.

2.1 Responsabilidad

El creciente impacto de los algoritmos de IA en la sociedad de la información, las operaciones y las “decisiones” que se toman día a día con base o en su totalidad por IA va en incremento. Más frecuentemente, los algoritmos de IA están presentes en las decisiones e intermediando en la operación diaria de la sociedad y de las economías productivas de los países (Brent Daniel Mittelstadt, 2016).

De igual manera, dentro de un marco individual en la sociedad, los algoritmos de IA determinan la información, la publicidad y los bienes y servicios que cada persona consume (Martin, 2018). Por tal motivo, la necesidad de mantener un margen de error mínimo en las clasificaciones que genera un algoritmo de IA (y de acuerdo con la sensibilidad en la cual este sea propuesto) es de vital importancia en su ejecución.

La responsabilidad de los algoritmos, de igual manera, va orientada al mecanismo de verificación y rigurosidad de los algoritmos utilizados, así como su fuente de datos y sus posibles resultados. Esta parte está relacionada en gran medida con la

trasparencia de los algoritmos en cuanto a su propuesta general, que termina incentivando prácticas adecuadas para la construcción de algoritmos verificables, auditables y reproducibles. Para ello, es vital el completo entendimiento del mecanismo interno de los modelos y algoritmos, junto con sus respectivas virtudes y falencias para resolver cada problema en particular.

En ese orden de ideas, la responsabilidad en los algoritmos se puede ver desde dos frentes dentro de un marco social beneficioso. El primero es la responsabilidad social de los algoritmos de IA en cuanto a las garantías sociales relacionadas con las oportunidades y la equidad dentro del entorno de mercado laboral (Wisskirchen, 2017) y el segundo es el relacionado al posible sesgo en los resultados de los algoritmos a raíz de prácticas no adecuadas.

Por tal motivo, dentro del contexto de los potenciales riesgos que puede llevar consigo la implementación de diferentes algoritmos de IA se requiere identificar y mitigar dichos riesgos en su ejecución:

1. **Responsabilidad social de los algoritmos:** El desarrollo e implementación de un sistema de IA puede aumentar la productividad de un sector y la competitividad de este y, por consiguiente, generar incentivos para su implementación, ya sea un sistema de recomendación de producto a clientes, la clasificación de sectores de la sociedad que requieren de un subsidio o no, de un seguro, etc. Por tal motivo, desde el punto de vista de productividad, la implementación de un algoritmo de aprendizaje automático y la automatización de un proceso repetitivo que requiera datos, genera mayores beneficios en las organizaciones y del país frente a la no utilización de dicha automatización (Iain M. Cockburn, 2017), sin embargo puede tener efectos negativos en los niveles de empleabilidad en algunos sectores donde la automatización logre reemplazar una labor que este siendo realizada por un ser humano, lo cual tiene el potencial de generar mayor inequidad social y económica (James Manyika, 2017).

   Los algoritmos de IA suplen una necesidad para el desarrollo de labores rutinarias y repetitivas que involucren información y datos para la toma de decisiones. Por tal motivo, un sistema de IA logra minimizar los costos y aumentar la operabilidad de labores que se han desarrollado hasta el momento

por seres humanos, desde reconocimiento de fraudes, análisis contables, financieros, etc. Sin embargo, para el desarrollo de dichos sistemas de IA, se requiere de nueva mano de obra que supla la emergente demanda, lo que conlleva a la generación de nuevos empleos, a la transformación de empleos actuales, pero también a la destrucción de otros empleos reemplazados por la IA.

En un principio, los nuevos empleos creados para el desarrollo de los sistemas de IA requieren de conocimientos y competencias muy específicas (Wisskirchen, 2017), lo cual implica una barrera de entrada a una parte de la población, para desempeñarse en este tipo de ocupaciones. Por otro lado, la destrucción de los empleos, en los cuales se reemplazan por un sistema de IA, suelen ser empleos que requieren labores repetitivas característicos por ser de ingresos medios. (Ekkehard Ernst, 2018). De acuerdo con lo anterior, la productividad puede aumentar, pero consigo podría también aumentar la inequidad en el mercado laboral y en el nivel de ingresos, por el hecho de que los empleos que se pueden estar reemplazando, pueden ser empleos sin altas restricciones ni barreras de entrada, con salarios medios que requieran la utilización de datos y repetitivos, aumentando así las brechas salariales entre trabajadores. Por tal motivo, al momento de desarrollar un sistema de IA, se requiere evaluar los costos y beneficios sociales que implican su implementación.

2. **Responsabilidad en el desarrollo de los algoritmos:** De igual manera, la implementación de algoritmos de aprendizaje automático dentro de un sistema de IA requiere de un desarrollo bajo ciertos parámetros y criterios que permiten regular la implementación del sistema en un proyecto. La necesidad de mantener un desarrollo responsable radica en la necesidad misma de mantener los resultados sin falencias, evitando cualquier tipo de posible sesgo. A pesar de ello, factores asociados al sesgo en los resultados de un algoritmo de aprendizaje automático pueden estar relacionados a diversos motivos. En un principio, puede existir factores controlables y no controlables que pueden generar sesgo en los resultados. El objetivo es evitar aquellos aspectos que sean controlables que conlleven a sesgos en los resultados del algoritmo.

Los **factores no controlables** que pueden ocasionar sesgo en los resultados van ligados a la naturaleza misma de los datos, esto quiere decir que la información recolectada para su análisis puede encontrarse con errores de medición o la muestra elegida para su estudio, puede no ser representativa de la población y estar ignorando una parte de esta que sea relevante. Dichos escenarios no son totalmente controlables, ya sea por la infraestructura de recolección de información, problemas en los sistemas de información o simplemente la disponibilidad de la información es limitada. Por tal motivo, las medidas que se pueden tomar para mitigar estos factores son limitadas.

Por otro lado, los **factores controlables** que pueden ocasionar sesgos en los resultados están relacionados con la implementación de los algoritmos de aprendizaje automático o la elección de un segmento no representativo de la información. Dichos factores, pueden ser generados por incentivos de terceros para presentar información espuria en los resultados del algoritmo, por tal motivo, se requiere de mecanismos que garanticen desincentivar el uso de estas prácticas que pueden llegar a ser perjudiciales.

En este orden de ideas, surge la necesidad de mantener un nivel de responsabilidad en los algoritmos que permita un desarrollo sostenible y beneficioso para la sociedad, y mantener los niveles óptimos de productividad y eficiencia en los procesos donde un algoritmo de aprendizaje automático interviene.

2.2 Transparencia

La transparencia de los algoritmos se refiere a la capacidad de un algoritmo de ser entendible y explicable de acuerdo con sus finalidades. Como se mencionó anteriormente, este tema va relacionado con la interpretabilidad de los modelos. Si un modelo es interpretable, esto implica que es transparente al público en las decisiones que se tomen a partir de esa información adquirida. Sin embargo, si la interpretabilidad interna del modelo no es uno de los objetivos del proyecto, la transparencia de los algoritmos va orientada a los resultados que este genere y la facilidad con que estos sean comprendidos y explicados.

En ambos casos es importante que los modelos realizados sean reproducibles en cualquier entorno bajo los mismos parámetros, supuestos, datos y condiciones

utilizadas originalmente. Esto tiene el objetivo de tener un proceso auditable y expuesto a mejoras, además de incentivar la rigurosidad en el modelo.

1. **Modelos de caja negra:** Los modelos de caja negra son aquellos que, en su estructura interna de modelamiento matemático de parámetros, pesos, distancias, etc., no cuentan con interpretabilidad directa de las variables de entrada. Dependiendo de la naturaleza del proyecto, la interpretabilidad puede ser o no ser necesaria, sin embargo, dentro del marco de transparencia de los algoritmos, la importancia de conocer el mecanismo por el cual un sistema de IA clasifica o predice bajo un escenario de política pública, toma gran relevancia debido a la necesidad de conocer las razones intrínsecas de la predicción. Sin embargo, la necesidad de interpretabilidad de los modelos también está sujeta a la naturaleza misma del proyecto. En ese sentido, si el proyecto en un principio no requiere interpretación intrínseca interna del modelo y la ganancia es sustancial (como lo puede ser modelos de pronóstico del clima), la implementación de modelos de caja negra puede ser beneficioso.
2. **Reproducibilidad:** Así mismo, la importancia de reproducibilidad de un algoritmo es beneficioso en el sentido de que los resultados obtenidos por cualquier algoritmo de aprendizaje automático pueden ser replicables para su validación, bajo las mismas condiciones. La posibilidad de reproducibilidad del algoritmo permite llevar un control general de los caminos y posibles oportunidades de mejora que un algoritmo de aprendizaje automático pueda tener. Así mismo, los incentivos que se pueden generar y que, de algún modo, pueden afectar negativamente los factores controlables en la responsabilidad del desarrollo de algoritmos, pueden ser mitigados o controlados debido a que la reproducibilidad del algoritmo permite que entidades, grupos o personas externas (y de acuerdo con la sensibilidad de la información) auditen y encuentren dichos inconvenientes

3 Buenas prácticas y mecanismos de autorregulación en los algoritmos

En términos generales, no existe una única rama o camino por el cual se debe desarrollar un proyecto de analítica. Esto depende de la organización, recursos

disponibles e inclusive de la persona encargada. Sin embargo, existen ciertos criterios recomendados para tener en cuenta en la implementación de un modelo de aprendizaje automático y en la autorregulación en los modelos.

Este documento da una visión general de los procesos y pasos estándar por los cuales se pasa al momento de implementar un modelo o algoritmo de analítica para el desarrollo de un proyecto. Varios de los posibles riesgos a los cuales un proyecto se puede enfrentar, son identificados, junto con algunas sugerencias para enfrentarlos.

Los pasos definidos en la metodología CRISP DM son marcos conceptuales que dan una guía para un resultado exitoso y satisfactorio que cumpla con todos los parámetros del proyecto y que permita generar valor dentro del área al cual fue concebido. Por tal motivo, las buenas prácticas en el desarrollo de un proyecto de analítica de datos están en el marco de esta metodología.

Para el desarrollo de un sistema de IA, los mecanismos de autorregulación de los algoritmos pueden ser vistos desde dos perspectivas distintas sin ser mutuamente excluyentes sino complementarias, la primera es la autorregulación en el desarrollo por parte del investigador o desarrollador del algoritmo y la segunda es el mecanismo que tiene el algoritmo por sí mismo de sobre o sub entrenarse, utilizando el mejor modelo posible y de ajustarse satisfactoriamente a las nuevas necesidades.

Para el primer caso, dentro del marco de política pública, en la identificación e implementación de mecanismos de autorregulación en los algoritmos y en la necesidad de definir los criterios de explotación de datos, se logran identificar buenas prácticas en el desarrollo de un proyecto de aprendizaje automático. Dichos mecanismos de autorregulación de los algoritmos están orientados hacia la generación de incentivos para evitar malas prácticas en el desarrollo de este, de tal modo que, las posibles anomalías que puedan surgir en el desarrollo del sistema de IA puedan ser detectadas con facilidad, sin la necesidad de auditar frecuentemente al sistema. Por tal motivo, dentro de los criterios de explotación de datos en entidades públicas, la necesidad de manejar sistemas responsables y transparentes anteriormente mencionados. En otras palabras, el desarrollo transparente y responsable de los algoritmos, bajo un marco de criterios de explotación de datos, genera incentivos para un desarrollo del proyecto de analítica de datos con buenas prácticas, siendo un mecanismo de autorregulación de los algoritmos.

En el segundo caso, el mecanismo de autorregulación de los algoritmos, es la flexibilidad que tiene este de trabajar con el mejor modelo posible y evitar los potenciales riesgos asociados a la utilización equivocada de un modelo (como puede ser sub o sobre entrenamiento). Lo anterior quiere decir a que el algoritmo de inteligencia artificial debe ser flexible en cuanto a trabajar con un modelo u otro, esto debido a que, en un principio, puede que un modelo sea el óptimo, pero debido a un choque externo que genere algún cambio en los datos de entrada, puede que otros modelos o inclusive el mismo tipo de modelo, pero con otros hiperparametros, tengan mejores resultados que el propuesto en un principio.

Por último, un sistema de IA debe estar orientado a cumplir las necesidades de transparencia anteriormente mencionadas, de tal modo que garantice la **reproducibilidad del algoritmo** dentro de su validación. Sin embargo, no existe un único criterio factible que garantice la reproducibilidad de los algoritmos, a pesar de ello, se pueden dar algunas recomendaciones que puedan dirigir el proyecto a dicho propósito de reproducibilidad.

1. **Utilizar un sistema de control de versiones**: Dentro del desarrollo de un algoritmo de aprendizaje automático, utilizar un sistema de control de versiones es beneficioso dado que, por un lado, sirve para generar "puntos de control" en el desarrollo del código que permita devolverse y corregir posibles errores en cualquier versión que se encuentre el código. Por otro lado, se genera una trazabilidad al código para un tercero que requiera replicarlo y validar su desarrollo.
2. **Tener un código lo más automático posible**: La automatización de los procesos es uno de los puntos más importantes en los beneficios que implica la implementación de un sistema de IA. Por tal motivo, aquellos procesos donde se realice una modificación externa desde que se adquieren nuevos datos y se dan los resultados (como una modificación externa en la base de datos con el apoyo humano) no es idealmente beneficioso debido a que el mismo sistema de IA debería contener el procesamiento previo de los datos en su totalidad para luego realizar los respectivos cálculos de la nueva información.

3. **Establecer un punto de partida común en caso de generación aleatoria de información (establecer semillas)**: Para la reproducibilidad del algoritmo, pueden existir puntos en los cuales se requiera una generación aleatoria de información (cómo la segmentación de datos en *cross validation* o algoritmos que lo requieren por sí mismo). Por tal motivo, por cuestiones de reproducibilidad, y para garantizar los mismos resultados cada vez que se ejecute un código, es recomendable utilizar "semillas" en los algoritmos. Dichas semillas son para garantizar que el algoritmo siempre ejecute la misma generación aleatoria de información, ya sea para dividir la base, cálculo de parámetros, optimización de funciones, etc.
4. **Guardar registros de modelos utilizados y trazabilidad del mismo**: Dentro del desarrollo de un sistema de IA, la comparación entre diferentes modelos y metodologías para encontrar el que mejor satisfaga la necesidad es un factor rutinario y clave. Por tal motivo, tener registro de la trazabilidad de los algoritmos implementados es deseable tanto para la validación de los algoritmos como para el desarrollo mismo de los algoritmos.

4 Criterios de explotación de datos en entidades públicas

Para la ejecución de proyectos de explotación de datos en entidades públicas, se deben tener en cuenta una serie de criterios que sirven de guía de ruta para promover la transparencia en los algoritmos para entidades públicas. Sin embargo, la optimización de los recursos utilizados y el óptimo manejo de los datos depende en un principio del origen de estos y deben estar orientados al objetivo final del proyecto.

Existen diversos principios por los cuales un algoritmo puede o no ser valioso en un proyecto de explotación de datos. Sin embargo, existen pautas específicas para identificar cuando se puede incurrir en diferentes problemas futuros en la implementación de una política pública con base en información obtenida a partir de los resultados de un algoritmo de explotación de datos. Dichos problemas surgen a partir de no mantener parámetros de transparencia en la ejecución de una política con recursos públicos. Por tal motivo, se busca evitar caer en una falta de claridad en el mecanismo por el cual se toman las decisiones de una política pública con base en resultados de algoritmos analítica y explotación de datos.

Por un lado, un algoritmo de explotación de datos debe tener, y dependiendo del proyecto, un nivel de **interpretabilidad** en el mecanismo por el cual este se desarrolla con el fin de mantener un camino coherente y entendible en sus resultados. Por otro lado, se deben seguir unas pautas que garanticen que este sea **reproducible** para poder ser **auditable** desde cualquier usuario bajo los parámetros, supuestos, datos, códigos, etc., propuestos por el desarrollador del proyecto.

Así mismo, los diferentes esquemas de los problemas que se mencionaron anteriormente se desarrollarán más adelante en el documento, basados en un marco de unos criterios que promuevan la transparencia y orienten a las entidades públicas en la ejecución de un proyecto de explotación de datos bajo dichos parámetros. La importancia de plantear dichos criterios se remonta en la necesidad de parametrizar un estándar de ejecución en las entidades públicas que garanticen la transparencia y responsabilidad en su ejecución y que consigo se pueda lograr un desarrollo satisfactorio y completo en cualquier proyecto en entidades públicas que involucre diferentes metodologías de explotación de datos.

Basados en la metodología CRISP DM que proporciona una descripción normalizada y estándar del ciclo de vida de los proyectos de análisis de datos (Wirth & Hipp, 2000), y basados en los diferentes esquemas de interpretabilidad y sesgo en los resultados explorados por (Angwin, Larson, Mattu, & Kirchner, 2016) y (Khan & Khan, 2013) y basados en las fases de un modelo propuestas por (Matthew Mayo, 2017), se exponen los criterios para la ejecución de un proyecto de explotación de datos para promover la transparencia en los algoritmos así como identificar los potenciales riesgos asociados al proyecto:

1. **Entendimiento de la necesidad y del proyecto**: Este criterio se centra en entender las necesidades del proyecto y el objetivo final por lo cual este ha sido creado. De igual manera se busca los requisitos y necesidades de este para una posterior conversión a un problema de análisis de datos. Dicho entendimiento se basa en mantener claridad en el objetivo, los insumos y los resultados del proyecto.
2. **Entendimiento de los datos**: Este proceso empieza con la recolección de los datos y en las actividades posteriores como la visualización de estos y las estadísticas descriptivas para tener una visión general. Esta fase se encuentra altamente relacionada con el entendimiento de la necesidad y del proyecto (punto anterior). Para el entendimiento de los datos, también se deben identificar las posibles falencias que estos tengan y el **posible sesgo que estos tengan** ya sea por sus problemas de medición, recolección, etc. Dicho sesgo es importante tenerlo en cuenta, porque puede influir en la interpretación de un algoritmo y en su transparencia.
3. **Preparación de los datos**: Este criterio se centra en todas las actividades alrededor de construir el *dataset* final para su óptima explotación y que serán utilizados en el posterior periodo de modelamiento e interpretación. De igual manera, estos deben estar estructurados de tal manera que sea eficiente el modelamiento y deben contener toda la información (y en lo posible) relevante que expliquen el fenómeno para evitar posibles sesgos en los resultados. Por tal motivo, existe una necesidad de estructurar los datos que usualmente no están en la estructura adecuada para la implementación de un modelo de analítica. Estructurar los datos se refiere a organizarlos, dividirlos y

distribuirlos de acuerdo con las necesidades que se tengan para que posteriormente sea procesada y visualizada la información.

4. **Visualización de los datos:** El uso de diferentes técnicas de visualización de datos previo a la construcción del modelo es muy importante pues permite tener una visión clara y general de la composición de los datos. Este paso ayuda a despejar dudas y a tener mayor intuición de cómo utilizar los datos del proyecto y a cómo sacar el mayor provecho de ello. La utilización de tablas, gráficas, mapas, diagramas son comúnmente utilizados.
5. **Modelamiento**: El criterio de modelamiento es un conjunto de pasos y criterios que son necesarios tener en cuenta para un óptimo aprovechamiento de los datos. Para ello es necesario conocer el funcionamiento y las fases intrínsecas en los diferentes modelos, sus diferentes tipos, los diferentes riesgos y beneficios que tiene el uso de uno u otro, etc. De igual manera, existen diferentes criterios a tener en cuenta para garantizar su **transparencia como lo son su interpretabilidad, el posible sesgo que este tenga y su reproducibilidad**.
6. **Evaluación**: En esta fase, se tiene en cuenta las diferentes metodologías de prueba que existan para evaluar los diferentes modelos identificados para elegir el que mejor se ajusta a las necesidades del proyecto. Para ello, se implementan diferentes pruebas que dan sustento empírico a la utilización de un modelo y no otro. El criterio de evaluación debe cumplir con los parámetros que sean más eficientes con las necesidades de transparencia del proyecto.
7. **Desarrollo**: Una vez construido el modelo y haber obtenido el resultado de este. El conocimiento obtenido debe ser organizado y presentado para que logre tener algún impacto y genere valor. Este debe cumplir con todos los parámetros clave que permitan satisfacer la necesidad del proyecto. De igual manera, los proyectos deben ser **reproducibles** para poder ser auditables y de ser necesarios, ajustables, así como interpretables al nivel que este lo requiera.

5 Modelos de aprendizaje automático y analítica

El entendimiento del funcionamiento interno de un modelo de aprendizaje automático es clave para dar una solución coherente y satisfactoria al problema de cada proyecto de analítica. Por su parte, el entendimiento del funcionamiento interno los modelos dan las bases para su óptimo uso y que cumplan con los parámetros de transparencia y responsabilidad que ellos suponen.

Los modelos de aprendizaje automático buscan en términos sucintos, la mejor capacidad de generalizar. A esto se refiere en la capacidad que tiene el modelo de ajustar los resultados a datos por los cuales este no ha sido entrenado o, en otras palabras, la capacidad de un modelo de ajustarse a datos nuevos.

La estructura general de un modelo y su implementación se basa en la estructura de los datos y se representa en la Figura 1:

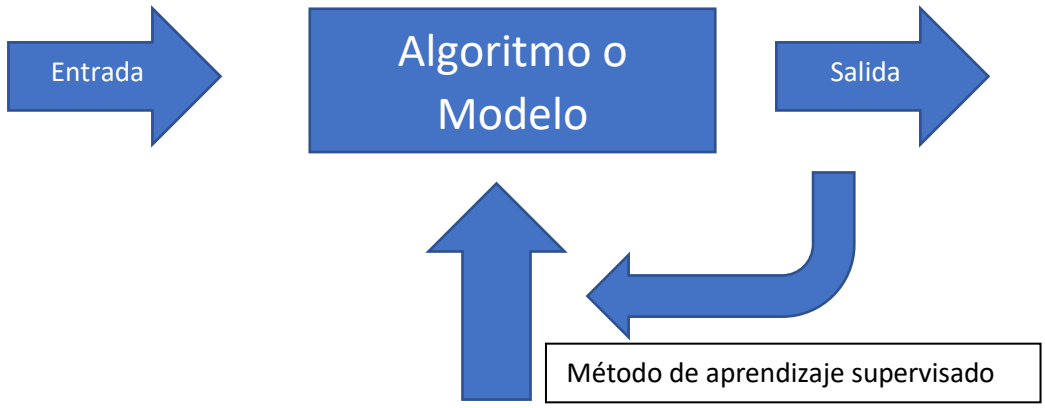

Figura 1: Algoritmos o modelos de aprendizaje automático

Todos los modelos están compuestos por unas entradas (*inputs*), se hace una transformación numérica (algoritmo o modelo) y expulsa una salida (*outputs*). Por su parte, en el proceso de aprendizaje, el modelo genera una memoria de largo plazo y otra memora de corto plazo, el primero son los resultados obtenidos en el entrenamiento del modelo después de lograr el periodo de aprendizaje de este y el segundo son las salidas y resultados parciales del modelo dentro del proceso de aprendizaje.

Para saber cuál es el objetivo que debe tener un modelo de analítica para la solución de un problema, lo primero que es necesario tener claro es ¿qué necesito que haga el modelo? Para ello, es necesario saber qué tipo de modelos existen y cual es óptimo para el proyecto. La Figura 2 muestra un esquema general de los modelos de aprendizaje de máquina y analítica:

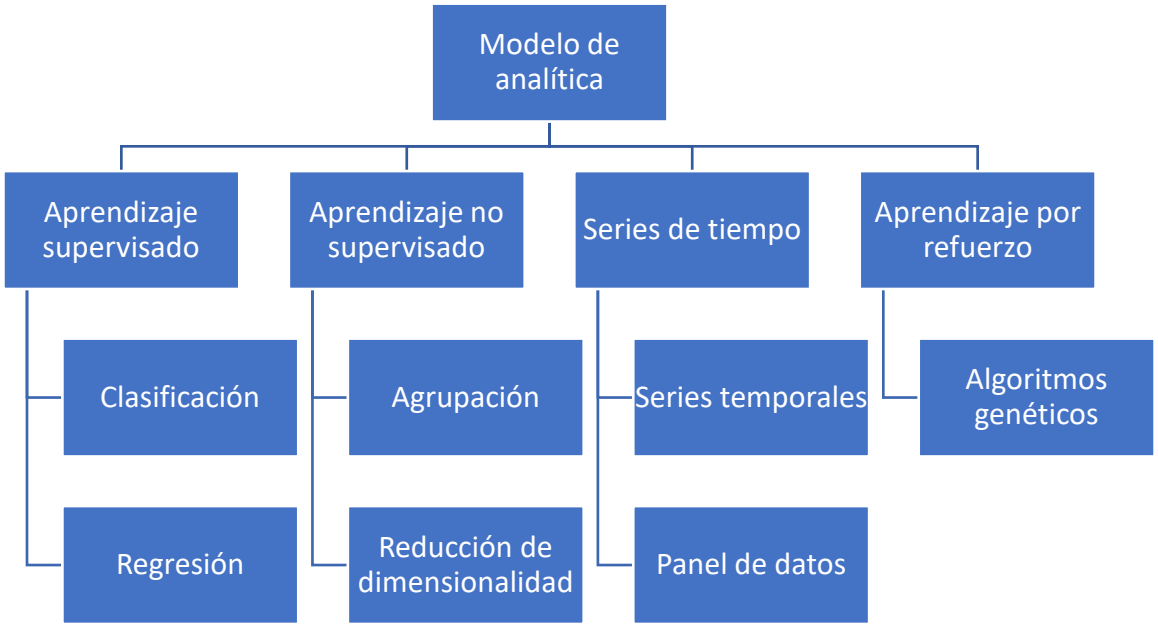

Figura 2: Tipos de modelos de analítica

Los diferentes tipos de modelos están orientados a distintos tipos de objetivos y diferentes estructuras de datos. Dependiendo del objetivo del investigador y la disponibilidad de datos, se elige el tipo de modelo a utilizar.

- **Aprendizaje supervisado:** Los modelos de aprendizaje supervisado se basan en que es conocida la salida dentro del set de datos, un ejemplo sencillo es un set de datos con las calificaciones de estudiantes y las horas dedicadas al estudio. Cada registro de tiempo dedicado al estudio tiene una calificación asociada, por lo que el output del modelo estará relacionado con la calificación obtenida por el estudiante y el input será el tiempo dedicado al estudio, los modelos comunes son los de regresión y clasificación.

- **Aprendizaje no supervisado:** Estos modelos se basan en que NO es conocida la salida explicita dentro del set de datos. Volviendo al ejemplo anterior, si solo se conoce el tiempo que el estudiante dedica al estudio, se puede implementar un modelo no supervisado para agrupar aquellos estudiantes que estudian más y aquellos que estudian menos.
- **Series de tiempo:** Estos son modelos comúnmente utilizados en la econometría y se basan en el valor de una variable a través del tiempo en uno o más escenarios. Un ejemplo de esto es la producción anual de uno o varios países en un periodo de tiempo determinado.
- **Aprendizaje por refuerzo:** Son modelos que "aprenden de sus errores", los cuales son asociados un "valor de ganancia" cuando va por buen camino y un "valor de perdida" cuando comete un error. Estos algoritmos comúnmente se basan en conceptos genéticos y de selección natural.

5.1 Identificación de la necesidad de interpretabilidad del modelo

La identificación de la necesidad y su conexión con la transparencia del proyecto y de los algoritmos está relacionada con la interpretabilidad de estos de cómo se desarrollan e implementan. Si este es o no interpretable tiene repercusiones posteriores en cuanto a la transparencia en la toma de decisiones que surgen a partir de los resultados del modelo utilizado.

Teniendo claro los distintos modelos y algoritmos a utilizar, la necesidad de utilizar uno u otro depende estrictamente del objetivo final del proyecto y de la estructura de datos disponibles. Por un lado, ¿el modelo debe priorizar la interpretabilidad (transparencia) o la capacidad de generalizar (desempeño)? Ambos componentes no son mutuamente excluyentes; sin embargo, existe gran variedad de modelos y en varios casos puede suceder que los modelos no interpretables sean los que alcanzan un mejor desempeño.

Para la implementación de modelos de analítica para el apoyo a políticas públicas, muchas veces es indispensable que haya una interpretabilidad en el modelo utilizado. La implementación o uso de un modelo que carezca de este atributo puede ser problemática. A continuación, se presentan algunos ejemplos:

1. La implementación de un modelo judicial para ayudar a los jueces al momento de juzgar casos y dictar sentencias. En este caso, las partes deben tener claridad de cómo se llegó a la decisión final, para garantizar el debido proceso y las garantías legales a las que haya lugar
2. Un modelo que determine cuáles familias recibe o no un subsidio del Estado puede llegar a tener serios inconvenientes si no es capaz de explicar o dar sustento técnico a su proceso de decisión.
3. Para el cuidado médico, si un modelo define a qué pacientes dar prioridad para la aplicación de un determinado medicamento (por ejemplo, por cuestiones de costo/beneficio), es de vital importancia que quien toma la decisión esté en capacidad de justificar por qué eligió a una persona y a otra no.

5.1.1 Modelos interpretables

Los modelos interpretables son aquellos para los cuales, dentro del proceso de predicción, es posible entender el proceso por el cual el algoritmo predice, optimiza, clasifica, etc. Este proceso es comprensible y se tiene total entendimiento y una explicación intuitiva (transparencia del algoritmo) del mismo (Lipton, 2017; Weng, 2017). En ejemplo de este tipo de modelos es la regresión lineal, que comprende una serie de entradas (regresores o variables independientes) y entrega una salida (variable dependiente). En el caso de la regresión lineal se sabe cuál es la importancia de cada variable independiente frente a la variable dependiente (por ejemplo, las pruebas de significancia o el análisis de intervalos de confianza) e inclusive se puede saber cuánto afecta dicha variable a la variable objetivo (coeficiente). Estos modelos se denominan modelos de caja blanca. Gran cantidad de los "modelos clásicos" tienen un método de interpretabilidad ya establecido además de herramientas que ayudan en dicho proceso (Kim, Khanna, & Koyejo, 2016).

5.1.2 Modelos no interpretables

Los modelos no interpretables son aquellos que no ofrecen una forma de conocer explícitamente el proceso al interior del modelo para tomar una serie determinada de entradas y producir una salida en particular. Este tipo de modelos se puede ver como una caja negra, en la que solo interesa la relación entre entradas y salidas. Un

ejemplo de este tipo de modelos son las redes neuronales profundas, que tienen una complejidad (cantidad de parámetros y relaciones lineales y no lineales entre ellos) tan alta que es extremadamente difícil exactamente cómo afectan las variables de entrada a la variable de salida.

Sin embargo, existen diferentes metodologías que permiten dar cierto grado de interpretabilidad a estos modelos de caja negra. Algunas de estas técnicas estudian el cambio de la salida del modelo al cambiar individualmente cada una de las entradas de este, para estimar la relación entre cada variable de entrada y la salida. Este enfoque es utilizado por (Robnik-Sikonja & Kononenko, 2008) y consiste en descomponer la predicción hecha por un modelo en "instancias" o "momentos". Otras técnicas utilizadas para dar interpretabilidad a modelos de caja negra incluyen las metodologías de análisis de sensibilidad y las técnicas LIME (*Local Interpretable Model-Agnostic Explanations*), BETA (*Black Box Explanation through Transparent Approximations*) y LOCO (*Leave One Covariate Out*), entre otras.

5.2 Elección del modelo o algoritmo

Dentro de todo el proceso de un proyecto de analítica, pueden existir varios caminos para llegar al resultado esperado. Sin embargo, no todos los caminos son iguales y estos pueden llevar a diferentes resultados. La elección del modelo es el mecanismo por el cual, se selecciona el mejor modelo, dentro de todos los tipos, para darle solución al problema. Sin embargo, la ambigüedad del concepto de "mejor modelo" puede generar confusión al hacer esta elección.

Una vez se tiene un modelo que sea apropiado para la naturaleza del problema, la disponibilidad de datos y la necesidad que se desea resolver, y que cumple con los requerimientos puntuales de interpretabilidad del proyecto, este modelo puede ser utilizado para el desarrollo del proyecto de analítica de datos.

Si existen dos o más modelos/algoritmos que cumplan con estas condiciones, el siguiente criterio para elegir el modelo, y los parámetros de este, es su desempeño, que a la larga se traduce en la habilidad del modelo para generar salidas confiables para datos nuevos. Para encontrar y comparar el desempeño de distintas alternativas, es necesario seguir un proceso de validación de los diferentes modelos.

6 Validación de modelos de aprendizaje de máquina

En esta Sección se presentan recomendaciones para validar el desempeño de un modelo de analítica de datos, y posibles riesgos que se deberían evitar para desarrollar modelos sin sesgos y que funcionen bien con nuevos datos.

6.1 Periodo de entrenamiento y prueba

Los modelos de aprendizaje de máquina normalmente son sometidos a un proceso de entrenamiento a partir de datos, para ajustar sus parámetros y mejorar su poder desempeño. Probar los modelos con datos no considerados en la etapa de entrenamiento es tan importante como el entrenamiento mismo. La etapa de prueba es la que verdaderamente permite evaluar cómo se va a comportar un modelo en producción, cuando trabaje con nuevos datos.

Para ejecutar ambas etapas en un proyecto de analítica, usualmente los datos disponibles se dividen en dos conjuntos: datos de entrenamiento y datos de prueba (Figura 3).

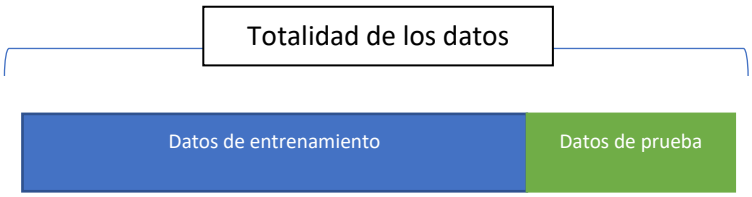

Figura 3: División de los datos para entrenamiento y prueba

La finalidad de dividir los datos en entrenamiento y prueba es en un principio, poder diferenciar y validar los resultados y también evitar la sobre estimación del modelo y de tal modo, encontrar el mejor camino para el desarrollo del proyecto.

6.2 Pruebas del modelo o algoritmo

Dependiendo del objetivo del modelo y la naturaleza del problema, se pueden utilizar diferentes indicadores y herramientas para medir el desempeño del modelo. Estas medidas permiten cuantificar qué tan bien está funcionando un algoritmo con un determinado set de parámetros.

Adicionalmente, si se calculan dichos indicadores para el modelo sobre el conjunto de entrenamiento y el conjunto de prueba por separado, es posible comparar dichas medidas. Esta comparación permite identificar si el modelo está sub ajustado o sobre ajustado a los datos de entrenamiento. Ambos fenómenos son no deseados, por lo que se quiere evitarlos siempre que sea posible. La Figura 4 muestra un cuadro con estos conceptos.

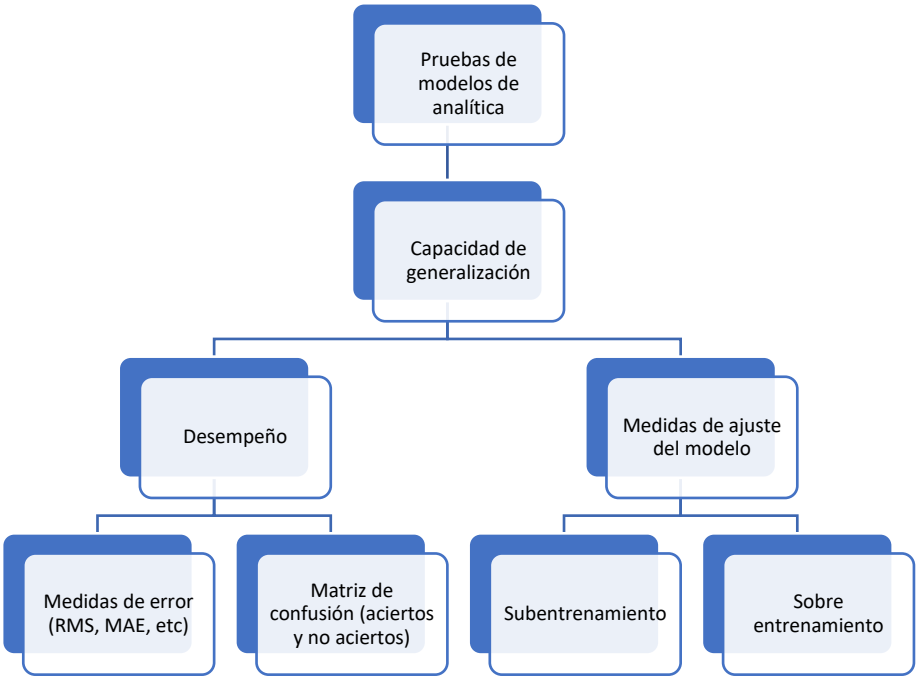

Figura 4: Apreciación para la implementación de pruebas a los modelos de analítica

6.2.1 Capacidad de generalización

Como se mencionó anteriormente, la capacidad de generalización del modelo es la habilidad de mantener su desempeño frente a datos nuevos, que nunca ha visto antes. Una vez se obtienen los parámetros del modelo durante el proceso de entrenamiento del modelo, estos deben tener la capacidad de enfrentarse satisfactoriamente a nueva información.

6.2.2 Desempeño

La capacidad de predicción/clasificación/segmentación/planeación de un modelo está asociada a la naturaleza del problema. Dependiendo del objetivo del modelo y

de la naturaleza de los datos, diferentes indicadores pueden ser utilizados para cuantificar el desempeño alcanzado. Elegir el indicador de desempeño adecuado para un modelo es de vital importancia para realizar una correcta elección del modelo y sus parámetros.

Dependiendo de la naturaleza del modelo, si este es de regresión o de clasificación, existen diferentes medidas en el error de resultados que se van a explorar a continuación:

1. Modelos de clasificación:

   Para modelos de clasificación, usualmente las medidas de desempeño más utilizadas son provenientes de la **matriz de confusión** el cual representa el número de predicciones de cada clase frente el número real de la clase, quiere decir que muestra el número de predicciones acertadas y no acertadas que hace el modelo en comparación con las clases reales. La matriz de confusión contiene la probabilidad de cometer un error de tipo 1 o **falso positivo** que es un pronóstico afirmativo del modelo, cuando la clase real es negativo y la probabilidad de cometer un error de tipo 2 o **falso negativo** que es un pronóstico negativo del modelo cuando en realidad la clase es positiva.

   A partir de la matriz de confusión, usualmente se calculan métricas como *precosion, recall* y *accuracy* que, en términos generales, la primera mira la proporción de todos los pronósticos que son verdaderos, el segundo mira la proporción de todos los verdaderos, cuantos se pronosticaron como tal y el tercero mira del total de pronósticos, la proporción de cuantos fueron acertados.

2. Modelos de regresión:

   Por otro lado, los modelos de regresión manejan otro tipo de medidas de error, usualmente basadas en el **error de pronóstico** que es la diferencia entre el valor real y el valor pronosticado. Por tal motivo, se utilizan usualmente medidas como la suma acumulada de errores de pronóstico (CFE), desviación media absoluta (MAD), error cuadrático medio (MSE), error porcentual medio absoluto (MAPE) entre otros.

6.2.3 Medidas de ajuste del modelo

Este punto está relacionado con que tan bien se ajusta el modelo a los datos con los cuales este fue entrenado, y se puede dividir en dos tipos.

- Sub entrenamiento del modelo (*Under-fitting*): Un modelo sub entrenado es aquel que no se ajusta lo suficiente a los datos con los cuales fue entrenado. Esto quiere decir que los valores arrojados por el modelo muestran resultados atípicos y poco coherentes, ocasionando frecuentes errores en la salida del modelo.
- Sobre entrenamiento del modelo (*Over-fitting*): El sobre entrenamiento del modelo está asociado a la dificultad de generalizar para nuevos datos. El modelo puede tener muy buenos resultados para los datos a los cuales fue entrenado, pero al salirse de esta zona, los resultados son muy variables y hasta erróneos. Dentro del ejemplo, es la imposibilidad de que el vehículo conduzca por fuera de Sillicon Valley pero que sí lo haga bien en esta área.

6.3 Sesgo en los resultados

El sesgo en los resultados de un modelo se debe principalmente a que sus parámetros o resultados (*outputs*) están orientados o tienden hacia un sector, valor, grupo, etc. Existen varias naturalezas en el sesgo, algunas más técnicas que otras y que es necesario tener en cuenta. A continuación, se presentará en el siguiente ejemplo la relevancia de mapear y mitigar los sesgos:

> En un juicio, se requiere determinar si el acusado es culpable o no. Para ello se implementó un modelo de aprendizaje automático que ayude al juez en su decisión y el modelo dice que es culpable. Otros 9 acusados son sometidos ante el mismo procedimiento, resultando 6 culpables y 3 inocentes (incluyendo al primero). Por otro lado, se someten otros 10 acusados y los resultados dan que 4 son culpables y 6 no lo son. ¿Por qué el modelo da diferentes resultados? En un principio se puede pensar en las diferencias sustanciales entre los dos grupos de acusados y que son aleatoriamente elegidos. Sin embargo, el primer grupo y el segundo grupo tienen un factor que los diferencia uno del otro, el primero es conformado por afrodescendientes y el segundo es conformado por caucásicos. Estos resultados pueden llevar a la siguiente pregunta ¿es la

etnia determinante en este proceso? Al momento de ver los datos con los cuales se entrenó el modelo, se encontró que en su mayoría son datos de delitos cometidos por personas afrodescendientes. El modelo tiene un sesgo porque los datos en un principio están sesgados; este caso es mayormente explorado en (Skeem & Lowenkamp, 2016).

El sesgo en los resultados puede estar también relacionado a la implementación de algoritmos que, de uno u otro modo, amplifican este hecho. En el procesamiento de lenguaje natural, existe un sesgo en el proceso de texto conocido como *word embedding*, que representa el texto como vectores. Sin embargo, este método muestra un sesgo en estereotipos y crece aún más debido a que se suele amplificar este sesgo (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016).

Los riesgos a los cuales se incurre en cuanto a los sesgos que estos se puedan ocasionar, tiene repercusiones en sus resultados y en las decisiones que se tomen a partir de la interpretación de estos resultados, por tal motivo, surge la importancia de tenerlo en cuenta al momento de la implementación de un modelo. Al validar los resultados de un modelo, es recomendable observar cómo se distribuye la salida de este con respecto a otras, y tratar de identificar posibles sesgos.

Por tal motivo, en orden para mitigar la posibilidad de sesgos en los resultados de los algoritmos de aprendizaje automático, se sugieren las siguientes recomendaciones:

1. Validar la calidad de los datos por medio de herramientas estadísticas como los momentos muestrales, la distribución de los datos y su comparación con resultados obtenidos previamente en la literatura sobre problemas y temáticas iguales o parecidas.
2. Validar la metodología de recolección de datos e identificar los puntos críticos o puntos en la metodología que puedan ocasionar algún tipo de sesgo (como la implementación de un tipo de muestreo que no sea adecuado para ese estudio).
3. Validar la correlación entre variables de entrada y sus dependencias lineales entre sí y con otras variables externas para evitar multicolinealidad y cointegración entre las variables.

4. Validar las metodologías previamente desarrolladas en la literatura e identificar los resultados, fortalezas y debilidades de la implementación de un algoritmo u otro.
5. Validar el sobre o sub entrenamiento del modelo y su comportamiento frente a otros modelos.

7 Conclusiones y recomendaciones finales

La necesidad de definir y esquematizar la trazabilidad en el desarrollo de un proyecto de analítica y explotación de datos con base en criterios definidos en el contexto de entidades públicas toma relevancia en un entorno donde la producción y el potencial de valor agregado de la información en la solución de problemas es abundante. Sin embargo, es necesario identificar los potenciales riesgos que puede conllevar la implementación de un sistema de IA en un problema donde antes no se utilizaba, de tal modo que los criterios y la hoja de ruta en el cual se aborde el proyecto, logre mitigar dichos riesgos lo mayor posible.

Por tal motivo, la importancia de conocer e implementar una hoja de ruta que oriente el desarrollo de un proyecto de analítica de datos a los objetivos finales de las entidades públicas bajo el marco en el cual ellas se rigen es de vital importancia, de tal modo que se garantice responsabilidad y transparencia en su proceso. Para ello, los sistemas de IA deben cumplir con ciertos requisitos explorados en este documento como lo son la reproducibilidad y la interpretabilidad (y de acuerdo con la necesidad) que garanticen transparencia y responsabilidad en los algoritmos.

Así mismo, la definición de los criterios por los cuales se puede abordar un proyecto de analítica de datos busca consolidar y satisfacer dichos requisitos para lograr un desarrollo satisfactorio y eficiente del proyecto y que pueda generar valor en el contexto al cual es expuesto. Por esta razón, este documento busca sentar las bases de explotación de datos en entidades públicas con base en las necesidades particulares del proyecto y generales de la entidad y política pública.

8 Bibliografía

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country.*

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? *Advances in neural information processing systems*, 4349-4357.

Brent Daniel Mittelstadt, P. A. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.

Ekkehard Ernst, R. M. (2018). *The economics of artificial intelligence: Implications for the future of work.* Genova: International Labour Office.

Iain M. Cockburn, R. H. (2017). The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis. *The National Bureau of Economic Research*.

James Manyika, M. C. (2017). *A future that works: Automation, Employment, and Productivity.* Mckinsey global institute.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Stadistical Learning with Applications in R.* New York: springer.

Khan, M. E., & Khan, F. (2013). A Comparative Study of White Box, Black Box and Grey Box Testng Techniques. *International Journal of Advanced Computer Science and Applications*.

Kim, b., Khanna, R., & Koyejo, O. (2016). Examples are not Enough, Learn to Criticize! *Advances in Neural Information Processing Systems*, 2280-2288.

Lipton, Z. C. (2017). The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490*.

Martin, K. (2018). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, pp 1-16.

Matthew Mayo. (2017). Obtenido de KDnuggets: https://www.kdnuggets.com/2017/11/interpreting-machine-learning-models-overview.html

Mauro, A. D. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings 1644, 97.*

RAE. (2018). *Real Academia de la Lengua Española*. Obtenido de Defiición de Algoritmo: https://dle.rae.es/?id=1nmLTsh

RAE. (2018). *Real Academia de la Lengua Española*. Obtenido de Definición de Modelo: https://dle.rae.es/?id=PTk5Wk1

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 589-600.

Skeem, J. L., & Lowenkamp, C. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 680-712.

Weng, L. (2017). *Lil'Log.* Obtenido de https://lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model.html#interpreting-black-box-models

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 29-39.

Wisskirchen, G. B. (2017). *Artificial Intelligence and Robotics and Their Impact on the Workplace.* London: IBA Global Employment Institute.

Anexo 1

Mesa de trabajo transparencia de los proyectos de explotación de datos y criterios que deberían incorporarse en un marco ético

Primer Foro: Política Nacional de Explotación de Datos (Big Data) y diseño de territorios y ciudades inteligentes

25 de junio de 2018

Contexto

De conformidad con las acciones contempladas dentro del Documento CONPES 3920 *Política Nacional de Explotación de Datos (Big Data)*, luego de presentada la política y sus alcances, se conformó una mesa de trabajo para recopilar insumos para definir los elementos que deberían tener los proyectos de explotación de datos para garantizar su transparencia y aportar a la construcción de un marco ético sobre el uso de los datos. Para ello, se presentaron a las siguientes consideraciones a los asistentes:

- La materialización del valor potencial de los datos, mediante la generación de productos de información útiles para la toma de decisiones se obtiene empleando técnicas de analítica (OCDE, 2015). Esta es la 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.

- Esto requiere medidas para evitar errores, sesgos y discriminaciones en las decisiones que se toman a partir de la explotación de datos. De acuerdo con los referentes internacionales, dentro de los riesgos que deben ser tenidos en cuenta y examinados cuidadosamente, se encuentran, entre otros, el hecho de que el análisis de datos puede llevar a la toma de decisiones erróneas, debido a sesgos que no hayan sido contemplados o eliminados de los datos empleados. Por

ejemplo, una herramienta usada para determinar la probabilidad de comisión de delitos y reincidencia puede presentar resultados errados porque los datos contenían un sesgo racial que no fue identificado y eliminado.

- Los límites éticos generan confianza en el aprovechamiento de los datos y permiten resolver cuestionamientos frente a los cuales el ordenamiento jurídico resulta inadecuado o insuficiente (Data & Society Research Institute, 2014) porque estos temores se asocian con las aplicaciones imprevisibles e imperceptibles de la explotación de datos, que exigen el compromiso de los actores para respetar la dignidad humana y aumentar el bienestar social, mediante la definición de pautas o modelos de autorregulación. Son las pautas de comportamiento que orientan a los actores del ecosistema hacia los fines socialmente deseables, de modo que se tomen las decisiones que maximicen el bienestar social en cada circunstancia particular.

- Por ejemplo, los datos pueden ser empleados de manera indebida en contextos de transacciones privadas, por ejemplo, para imponer tasas más altas en la venta de seguros a partir de la revisión del historial de navegación en Internet, limitar la capacidad de decisión mediante la exposición limitada de contenido, o el aumento de precios de acuerdo con los patrones de consumo, entre otros.

- En este contexto, donde los ciudadanos aceptan de manera rutinaria, y casi automática, los términos y condiciones de los bienes y servicios aparentemente gratuitos y que se lucran del uso de esos datos, libre y autónomamente entregados (Greenwood, Stopczynski, Sweatt, Hardjono, & Pentland, 2014), pone de presente la necesidad de evitar la entrega voluntaria de datos personales, pero sin plena conciencia de las consecuencias por parte de los ciudadanos.

Preguntas orientadoras

De acuerdo con el contexto expuesto, surgen los siguientes interrogantes, que son parte de una discusión en curso en el mundo sobre los límites que deben diseñarse para minimizar los riesgos y garantizar los derechos de los titulares frente al aprovechamiento de datos y la automatización creciente de las decisiones:

- Teniendo en cuenta su complejidad y nivel de conocimiento técnico requerido ¿Cómo podría lograrse la *explicabilidad* y auditabilidad de los proyectos de explotación de datos?
- ¿Cómo lograr un balance entre la transparencia de los proyectos y el deber de proteger datos privados, semiprivados, sensibles, reservados o clasificados, que puedan estar involucrados?
- ¿Qué mecanismos pueden mitigar los riesgos de que existan sesgos o errores en los proyectos?
- ¿Qué prácticas o experiencias de autorregulación en materia de explotación de datos conoce?
- ¿Cómo podrían adoptarse estas prácticas en el país?
- ¿Cómo lograr compromisos de todos los actores sociales para la implementación de estas prácticas?
- ¿Cómo verificar el cumplimiento de estos compromisos de autorregulación?
- El actual enfoque normativo se centra en la obtención de la autorización por parte del titular de los datos ¿Debería este enfoque complementarse con exigencias concretas respecto del uso y tratamiento, aun cuando haya sido consentido por el titular?

Resultados de la mesa de trabajo

- **Acceso y salvaguarda de los datos:** Existe una preocupación relacionada con las fuentes de datos que sean empleados para los análisis. Concretamente, respecto de su procedencia y sensibilidad. Se reconoce que no todos los datos pueden estar expuestos al público (por ejemplo, información sobre víctimas del conflicto armado en el país). En general, los asistentes debatieron sobre cómo trabajar en mecanismos que velen por la vigilancia y control sobre el uso de los datos para prevenir la vulneración de la reserva y la detección de perfiles en bases de datos, en casos en que puede revertirse el proceso de anonimización.

- **Vulnerabilidad de los algoritmos:** Además de la vulneración de los datos (acceso no autorizado o indebido) los algoritmos son susceptibles de ser "*hackeadas*" o contener "*bugs*" que, inclusive si se protegiesen los datos, pueden revelar más información de la permitida. Esto va de la mano con el "*software*" utilizado para

crear estos algoritmos. Se hace énfasis en que los programas con que son construidos estos algoritmos son paquetes de lenguajes de programación, por ello, en control debe centrarse en la construcción del algoritmo, más que en que pueda ser "*hackeable*". Lo anterior, se engloba en el término "auditoría".

- **Instancia de vigilancia y control al uso de los datos:** Siguiendo con lo anterior, se discute si debe existir una entidad con competencias para hacer la vigilancia, control, inspección y validación del buen uso de los datos, los algoritmos y los resultados. Por ejemplo, si los datos se usan para tomar mejores decisiones en política pública, es necesario garantizar que dichos resultados no estén sesgados o corruptos por las personas que participen en todo el ciclo de un proyecto en ciencia de datos. En este sentido, se sugiere evaluar si una entidad pública actualmente tiene competencia para este control y vigilancia o si se requiere una Superintendencia que realice esta función.

- **Auditabilidad de los algoritmos:** Se resalta la necesidad de permitir el control social con los datos. Al respecto, debe tenerse en cuenta que el capital humano con las herramientas para validar de qué manera se utilizan los datos requiere de una calificación significativa. Por ello, es necesario tener herramientas que permitan de manera muy ilustrativa entender lo que cada algoritmo hace y cuál es la función o papel que cumplen en la toma de decisiones.

- **Incentivo a la participación:** Se propone la realización de "*hackatones*" y actividades similares para incentivar la participación y el conocimiento respecto de los proyectos adelantados por las entidades públicas. Esto, además de fortalecer la participación ciudadana, promueve la cultura de datos.

- **Replicabilidad de los algoritmos:** La capacidad para replicar los algoritmos es una forma de validar que los resultados obtenidos sean correctos y no se encuentren produciendo resultados sesgados o manipulados por los autores de estos. Esto, garantizando en todo caso que no se permita el acceso a datos reservados, privados o semiprivados. La barrera en este aspecto es que algunos algoritmos no pueden ser replicables, porque no basta con una muestra

representativa de los datos, situación que generaría vulnerabilidades a la no divulgación de datos.

- Como alternativa a lo anterior, un esquema de clasificación de los datos puede servir para saber cuáles proyectos pueden tener replicabilidad y cuáles no. Dicho esquema de clasificación permitiría saber con qué información los ciudadanos pueden acceder de manera fácil y gratuita y cuales requieren de previa autorización de la entidad emisora para controlar el acceso de la información ahí depositada.

- Adicionalmente, los asistentes plantean como mecanismo para mitigar los riesgos éticos y morales de los algoritmos y los datos, permitir a los ciudadanos asistir personalmente a las entidades y que estas cuenten con salas de computo donde puedan correr algoritmos, acceder a la información, trabajar con los datos y manipularlos todo bajo un sistema de protección y control que monitorearían estas mismas entidades. Esto podría articularse con las acciones de atención al ciudadano que se han implementado para el acceso y consulta de documentos públicos.