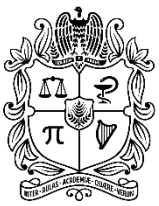


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Aplicación de técnicas de analítica de datos para identificar los factores que afectan la participación ciudadana en Medellín

Juan Pablo López Buitrago

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2022

Aplicación de técnicas de analítica de datos para identificar los factores que afectan la participación ciudadana en Medellín

Juan Pablo López Buitrago

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Ingeniería - Analítica

Director (a):

Ph.D. Olaya Morales

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2022

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Juan Pablo López Buitrago

02/08/2022

Resumen

Título en español: Aplicación de técnicas de analítica de datos para identificar los factores que afectan la participación ciudadana en Medellín

Teniendo en marcha la tercera medición del Índice de Participación Ciudadana, y aprovechando la experiencia consolidada a lo largo de los años en el cálculo e interpretación de los indicadores, una de las premisas obtenidas para la presente medición era la necesidad de dar un paso más en esta dirección. Teniendo esto como punto de partida, se propuso la exploración de herramientas de Analítica y Ciencia de Datos para garantizar un mejor aprovechamiento de los datos almacenados producto de las diferentes mediciones, y a la vez generar valor y conocimiento a partir de los datos que faciliten el ejercicio de toma de decisiones.

Como resultado, se construyó un modelo de Aprendizaje Automatizado a partir del algoritmo Random Forest Classifier. Con el objetivo de identificar las variables que más influyen en el puntaje final del IPCM se utilizaron herramientas de Feature Importance, dando lugar a inferencias, conclusiones y recomendaciones que brindarán una base sólida e informada para la elaboración de programas, proyectos y políticas públicas orientadas a mejorar el ejercicio de la participación en Medellín.

Palabras clave: Participación Ciudadana, IPCM, Machine Learning, Feature Importance, Medellín**.**

Abstract

Título en inglés: Application of data analytics techniques to identify the factors that affect citizen participation in Medellin.

Having underway the third measurement of the Citizen Participation Index, and taking advantage of the experience consolidated over the years in the calculation and interpretation of the indicators, one of the premises obtained for the present measurement was the need to take a step further in this direction. With this as a starting point, the exploration of Analytics and Data Science tools was proposed to guarantee a better use of the data stored as a result of the different measurements, and at the same time generate value and knowledge from the data to facilitate the decision-making exercise.

As a result, a Machine Learning model was built based on the Random Forest Classifier algorithm. In order to identify the variables that most influence the final IPCM score, Feature Importance tools were used, leading to inferences, conclusions and recommendations that will provide a solid and informed basis for the development of programs, projects and public policies aimed at improving the exercise of participation in Medellin.

Keywords: Innovative Citizen Participation, IPCM, Machine Learning, Feature Importance, Medellín.

Contenido

Pág.

Lista de imágenes

Pág.

Lista de gráficos

Pág.

Lista de tablas

Pág.

Introducción

Desde la génesis misma del Estado moderno, los asuntos de gobierno estuvieron supeditados a su capacidad para la recolección y manejo de datos de sus ciudadanos y del territorio que controlan. A diferencia de las distintas formas político-legales que han determinado los Estados a lo largo de la historia, la principal característica del Estado moderno es su composición legal-burocrática, que supuso una organización más eficiente y especializada de su aparato institucional, y una necesidad cada vez mayor de emular el funcionamiento de la industria moderna, lo que deriva en una búsqueda de mayor eficiencia y eficacia en la toma de decisiones (Weber, 1947). Bajo esta lógica, la construcción de políticas públicas informadas y transparentes es una necesidad imperativa para el correcto funcionamiento del Estado, no sólo por el criterio técnico de eficiencia mencionado previamente, sino además por su contribución en términos de Gobernanza urbana.

En la segunda mitad del siglo XX, la irrupción del Internet como una red de computadoras interconectadas a nivel mundial con el fin de compartir información entre sí, transformaría para siempre la forma como se producen, se almacenan y se manejan los datos (Lemus-Delgado & Pérez Navarro, 2020). De este modo, ante la rápida evolución de las tecnologías de la información (TICs) y el desarrollo de la capacidad de computación y sus facilidades de acceso para el ciudadano común, nos encontramos ante un mar de información al que antes sólo se tenía acceso a través del Estado.

En la actualidad, cualquier ciudadano a través de sus computadoras, celulares, relojes, automóviles, y en general múltiples dispositivos electrónicos, están constantemente generando datos sobre sus tendencias de consumo, movilidad, intereses e, incluso, posturas políticas. Bajo esta perspectiva, conceptos como *Gobernanza* (González, 2019; Danaher, Hogan, Noone, et al., 2017; Boudjelida & Mellouli, 2016) y *Territorio* (Lemus-Delgado & Pérez Navarro, 2020) sufren una profunda transformación, el primero por la capacidad cada vez mayor de generar discusiones políticas e incidencia en las decisiones de gobierno a nivel local, nacional y supranacional, a causa del impacto de las redes

sociales como “ágora digital” en el marco de la conceptualización de las ciudadanías digitales; y el segundo, por su parte, como consecuencia de unos procesos de territorialización en constante cambio debido a los fenómenos de globalización e interconexión social, política y económica. De este modo, cuando se habla de Gobernanza se hace referencia al proceso de construcción de las decisiones políticas (Mayorga y Córdova, 2007). Dicho en otros términos, mientras la Gobernabilidad se preocupa por las cuestiones de orden y estabilidad política-estatal, la Gobernanza se preocupa por el cómo se gobierna, por la acción o el ejercicio de gobierno.

En definitiva, el presente siglo y sus avances tecnológicos no sólo traen consigo transformaciones sociales cada vez más evidentes, sino además nuevas herramientas para el análisis de la realidad debido al crecimiento exponencial de los datos y la capacidad de su procesamiento.

Todo este proceso se condensa en el término de Big Data, que se refiere a aquellos conjuntos de datos cuyas características se definen a partir de las tres V: volumen, variedad y velocidad (Meneses Rocha, 2018). El *volumen* se determina a partir de la cantidad de registros contenidos en una base de datos, que, si bien se suele asumir a partir del millón de registros, existe un leve consenso en determinar Big Data a toda base de datos cuyo peso supere los 1114 terabytes; aunque, dado el constante crecimiento y desarrollo de los datos y su procesamiento, este peso mínimo es susceptible de ser aumentado. La *variedad*, por su parte, se define a partir de los tipos de datos, ya sean texto, imágenes, audio y/o video. La *velocidad*, por último, hace referencia a la velocidad de crecimiento de la base de datos, por lo que a menudo se consideran aquellas que contienen un flujo de datos constante e, incluso, los está capturando en tiempo real. Además, esta característica está determinada por la capacidad de procesamiento; de este modo, se considera Big Data a aquellos conjuntos de datos que no es posible ser procesados a través de técnicas y herramientas convencionales (bases de datos relacionales, paquetes estadísticos o herramientas de visualización).

En este orden de ideas, la analítica -o Analytics- surge como resultado de este proceso de desarrollo tecnológico. El término analítica de grandes datos (Big Data Analytics) se refiere a una serie de técnicas y metodologías, desarrolladas para el manejo de conjuntos de datos de gran volumen, variedad y velocidad de actualización, y es en la actualidad una de las bases fundamentales de la Ciencia de Datos.

La analítica de grandes datos no difiere mucho de la estadística clásica; por el contrario, es una evolución de ella y combina la estadística con la informática, las ciencias de la computación, el desarrollo de software y la comunicación (Lemus-Delgado & Pérez Navarro, 2020). Por otro lado, la Ciencia de Datos es una disciplina que aborda tanto el diseño, limpieza y estructuración de bases de datos, la formulación de algoritmos y modelos estadísticos para el análisis de datos, la transformación de datos en información para la toma de decisiones, y la correcta comunicación y visualización de los mismos.

Existe también la categoría de Small Data -o microdatos-, entendiendo esta como aquellos conjuntos de datos cuyo volumen es limitado, la captura de la información se hace a partir de un proceso lento y controlado (como el de un proceso de encuestaje, por ejemplo) y su variedad suele ser escasa o simplemente estar compuesta por un sólo tipo de datos (Meneses Rocha, 2018).

Esto último no supone una limitación, todo lo contrario. A la luz de la Analítica descriptiva, no es necesario un vasto volumen de datos para poder realizar análisis profundo de los mismos; además, la combinación de distintos conjuntos de datos, como lo es este caso, ya hace necesario el uso de herramientas de analítica para convertirlos en información valiosa y en insumos para la generación de nuevos campos de investigación y conocimiento.

Por último, es necesario reconocer que, dada la exposición mediática a la que se ha visto acogido el Big Data como el eje y soporte en la toma de decisiones en el mundo empresarial, y a su vez, como presente y futuro de las formas de gobierno - apareciendo incluso conceptos muy interesantes como smart city, e-governance y Algorithmic Governance (Islam Sarker, Khatun, Alam & Islam, 2020; Danaher, Hogan, Noone, et al., 2017; Ju, Liu & Feng, 2018; Meijer & Bolívar, 2016)-, ha habido una suerte de fetichización en torno a este concepto y sus técnicas; de este modo, no son pocas las consideraciones en las que se cree que este método, por sí solo, puede resolver problemas de orden sociopolítico como la criminalidad, la movilidad o la formulación de políticas públicas. Tampoco se trata de despreciar todo su potencial en el análisis científico y en la solución de problemas de esta índole, pero es necesario reconocer que, en este nuevo giro positivista, el privilegio del método no puede derivar en un detrimento la teoría; la analítica de datos no es una bola de cristal.

En definitiva, la Analítica de datos se presenta más como un medio que como un fin en sí mismo. Y dado su alto potencial y su carácter novedoso en el manejo de conjuntos de datos susceptibles de ser convertidos en información, esta metodología supone un insumo muy poderoso, mas no suficiente, para la generación de conocimiento que cualifique la participación ciudadana y la gobernanza urbana, entendiendo esta última a partir del proceso de construcción de las decisiones políticas (Mayorga y Córdova, 2007); con ayuda, por supuesto, de los métodos etnográficos. Por tanto, en el presente ejercicio se propone revisitar los datos de las mediciones anteriores del IPCM (2017 y 2019), los cuales sumados a los resultados de la más reciente medición (2021) proporcionarán la capacidad de realizar análisis comparativos, explorar tendencias y fenómenos, y construir reflexiones a partir del análisis de la dispersión de los datos a través de técnicas de clasificación.

Planteamiento del problema[1]

Hablar de un Enfoque Territorial para la Participación Ciudadana en Medellín implica ubicar este fenómeno histórica y territorialmente, de ahí que se reconozca toda la experiencia y el antecedente participativo con el que cuenta la ciudad y se estudie de acuerdo a su configuración histórica particular, de modo que pueda constituirse las condiciones para generar un conocimiento más preciso de las dinámicas y ejercicios de participación ciudadana en Medellín, así como su calidad.

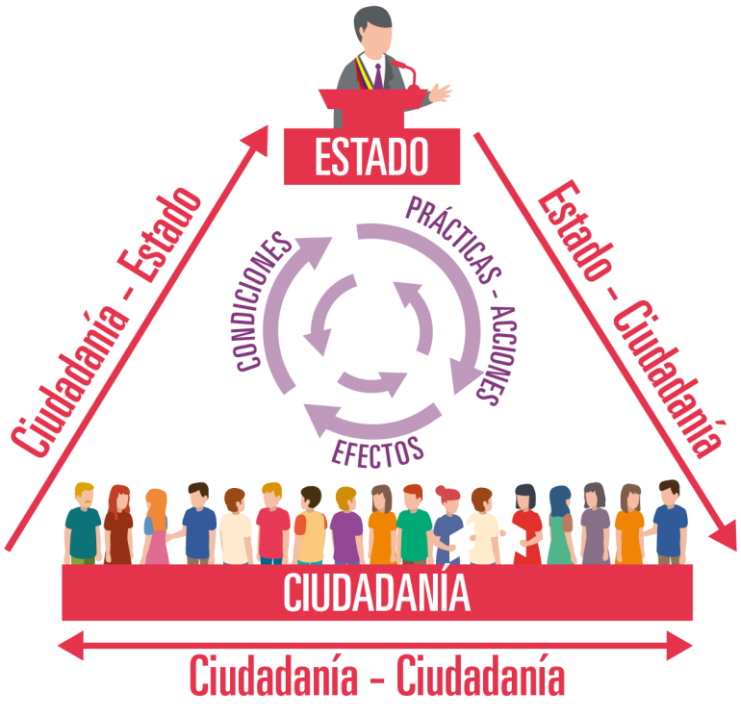
De este modo, desde el Enfoque Territorial, la **participación ciudadana** se entiende como una práctica política situada, individual o colectiva, de involucramiento con los asuntos públicos de interés en una o varias escalas territoriales (local, municipal, nacional), cuyos efectos o resultados dependen de las condiciones político-territoriales pre-existentes en las que se agencia, y del tipo de acción que establecen los actores. Esto lleva a trascender la mirada jurídico-administrativa del territorio para dar lugar a su comprensión como un espacio producido por diversos actores sociales que a su vez despliegan en él prácticas y relacionamientos de tipo político. Así pues, en este escenario se descubren tres ámbitos de la práctica política, en cuya relación dialéctica se constituye la arena política democrática:

Los primeros dos de ellos se ubican en el plano de la relación directa entre la Ciudadanía (Sociedad Civil) y el Estado, cada uno abordando un direccionamiento distinto. Así pues, el primero de ellos hace referencia a la relación Estado-Ciudadanía (Top Down), referida a aquellos escenarios institucionales en los que el Estado interviene en la Ciudadanía, ya sea por medio de la ejecución de actos legislativos, la redacción de la normatividad y sus límites, o la puesta en escena de planes y proyectos encaminados a la consolidación (o restricción) de espacios de apertura democrática. El segundo de ellos, al contrario, está direccionado a la relación Ciudadanía-Estado (Bottom-up), refiriéndose así al afianzamiento de vías institucionalizadas o no institucionalizadas para intervenir, exigir o controlar las acciones estatales; en este ámbito caben prácticas políticas como la

[1] El contenido de este subtítulo hace parte de la construcción conceptual denominada "Enfoque Territorial de la Participación Ciudadana (ET-PC): referente teórico-práctico para la comprensión situada de la participación en Medellín", elaborada en el seno del grupo de investigación *Nuevas Metodologías para la Participación Ciudadana de Medellín*

movilización social o la protesta ciudadana, los organismos de control o veeduría ciudadana, el plebiscito, la tutela y demás mecanismos de participación ciudadana instituidos en la Constitución Política de Colombia de 1991, la Ley 134 de 1994 y la Ley Estatutaria 1757 de 2015. El último ámbito, por su parte, se ubica en la relación horizontal Ciudadanía-Ciudadanía, haciendo referencia a aquellos espacios de interacción y deliberación generados entre diversos colectivos y organizaciones sociales sin la necesaria intervención del Estado; este ámbito ha caracterizado la participación ciudadana en la ciudad de Medellín desde la década de 1960, en donde las juntas cívicas, los convites, las natilleras, y demás formas de organización social construyeron los barrios de la ciudad con muy poca intervención estatal. Además, la consolidación de este ámbito es fundamental para la constitución de una cultura política y democrática más fuerte dentro de la ciudadanía, contribuyendo a la construcción de escenarios de apropiación política y territorial autónomos. En síntesis, los tres ámbitos del Enfoque Territorial se ven representados gráficamente de la siguiente manera:

Gráfico 1. Esquema analítico ET-PC



A su vez, al interior del esquema de relacionamiento dialéctico entre la Ciudadanía y el Estado a través de los tres ámbitos reseñados, es posible encontrar tres vectores que movilizan el proceso de significación y dotación de sentido generado en el

entrecruzamiento de los diversos intereses y subjetividades que entran en disputa en la arena política. De este modo, uno de los vectores (1) hace referencia a las condiciones territoriales en las cuales se inscribe la participación ciudadana, entendiendo esta como la configuración territorial y social pre-existente que determina, posibilita, facilita o restringe el ejercicio de la participación; este vector, pues, está ligado al grado de las condiciones materiales de existencia a las que está sujeto el actor social, y engloba en sí mismo desde el nivel de acceso a derechos fundamentales como la salud, la educación o el buen vivir, hasta entornos de amenaza o coerción y elementos de infraestructura física y tecnológica. En definitiva, este vector se expresa en el nivel de las garantías y oportunidades para la participación ciudadana.

Otro de los vectores (2) se enmarca en el plano de las acciones y prácticas desplegadas en el ejercicio de la participación, evocando así aquellos procesos de agencia a través del cual los actores sociales despliegan estrategias y recursos para hacer efectivo su derecho a participar e incidir en su territorio. Es a través de este vector que se proyectan las manifestaciones de movilización social y uso de los mecanismos de participación ya referenciados, y se expresa en el nivel de la gobernanza.

El último vector (3) está adscrito a los efectos que produce el ejercicio de la participación ciudadana, entendiendo estos como la materialización de la capacidad de agencia de los actores políticos orientados a la resolución de problemas concretos, ya sea en temas de inclusión política, democratización de la gestión pública o mejoramiento de la calidad de vida en los territorios.

En síntesis, estos tres vectores (condiciones territoriales, prácticas y actores, y efectos) se traducen en dimensiones de análisis que adquieren relevancia en el estudio de la calidad de la Participación Ciudadana en la ciudad de Medellín. Por tanto, se entiende como calidad de la participación el equilibrio entre el triple proceso de relacionamiento dialéctico entre la Ciudadanía y el Estado, atravesadas por los vectores ya descritos, dando lugar a su entendimiento como un proceso dinámico que oscila entre dos puntos contrarios entre sí, en el que, por un lado, se ubica la ausencia de un ejercicio participativo, en tanto no existe correspondencia entre las variables descritas, mientras que por el otro se ubica un escenario utópico, en el que la participación alcanza su mayor expresión.

Es este el fundamento analítico que justifica la formulación en el año 2017, de un Índice que mida la calidad de la Participación Ciudadana en la ciudad de Medellín, siendo su

primera edición en el segundo semestre de ese mismo año. Cada una de las tres dimensiones previamente descritas supuso la construcción de indicadores que fueron medidos a partir de los datos recolectados vía encuesta a Individuos que participan y Organizaciones sociales y colectivos de las diferentes comunas y corregimientos de Medellín. Obtenida esta información, se computa un promedio geométrico cuyo resultado determina un Índice por dimensión de la Participación (Condiciones Territoriales, Prácticas y Actores, y Efectos), y un Índice general, que en suma representa el IPCM de Medellín.

Siguiendo una escala temporal de aplicación bianual, en la actualidad se ha llevado a cabo la medición del IPCM en los años 2017, 2019 y 2021. Sin embargo, a pesar de que esta información ha sido de suma relevancia a nivel de ciudad, en especial para la formulación de políticas públicas por parte de la Administración Municipal, es mucho el potencial que estos datos contienen para un mejor entendimiento de los fenómenos sociales asociados a la participación en la ciudad; para ellos, como complemento de las herramientas tradicionales que se usan en las ciencias sociales como la etnografía y los estudios de caso, surge la Analítica de datos. En este sentido, se espera generar procesos que permitan convertir en información y conocimiento los datos obtenidos y almacenados en las distintas mediciones, de modo que pueda construirse un plan de trabajo que dé respuesta a esas necesidades y requerimientos.

En suma, estamos ante la necesidad de optimizar el aprovechamiento de los datos generados por las mediciones del Índice de Participación Ciudadana de Medellín - IPCM en sus ediciones de 2017, 2019 y 2021. Si bien a través de la estadística descriptiva se han generado productos y procesos muy importantes como consecuencia de la interpretación de los resultados de las dos ediciones anteriores, la Analítica permitiría un mayor aprovechamiento de unos conjuntos de datos que no han sido explotados a su máximo potencial; en concordancia, se considera que a partir de esta optimización de los datos, y la adopción de las técnicas requeridas para tal fin, la información obtenida podría arrojar inferencias que aumente los campos de acción y la capacidad analítica para la Academia, Alcaldía y Sociedad Civil.

En esencia, el principal requerimiento para este proyecto implica una agregación de los conjuntos de datos que contienen los resultados globales del IPCM, de modo que pueda realizarse un análisis exploratorio que permita una observación del comportamiento de las variables que determinan un resultado bajo y alto de la calidad de la Participación

Ciudadana, haciendo énfasis en una serie de elementos que orienten la revisión de los conjuntos de datos, siendo estos: el sexo, grupo etario, territorio y organizaciones sociales. Estos ejes de análisis cumplen el papel de movilizadores para las interpretaciones generadas posterior a la implementación de los modelos de clasificación, y serán prioridad a la hora de revisar el comportamiento de los datos. Es importante hacer este ejercicio tanto para el conjunto de datos de los Individuos que participan como para el de Organizaciones sociales.

Por otro lado, y con un grado de prioridad similar al anterior, surge la necesidad de realizar dos análisis complementarios, en donde se presta especial atención al comportamiento de los datos clasificados de acuerdo a (i) las variables asociadas a la percepción positiva y negativa de la dimensión Efectos, tanto para el conjunto de datos de Individuos que participan como para el de Organizaciones, y (ii) las variables asociadas a una valoración fuerte de los liderazgos sociales, en específico para el conjunto de datos de Organizaciones sociales.

Dicha información se convertiría en un insumo fundamental para la toma de decisiones y el diseño de políticas públicas orientadas a atacar y promover aquellos fenómenos sociales asociados a las variables que empujan hacia abajo o hacia arriba los resultados del índice, de modo que no sólo se propicie el mejoramiento de los resultados en las mediciones futuras, sino que además se contribuya al mejoramiento de la calidad de la participación en la ciudad.

Por último, es necesario hacer claridad frente a los riesgos que podrían surgir de un ejercicio de este tipo. Ello conlleva el seguimiento de protocolos de seguridad adecuados para el manejo de bases de datos con información pública y sensible, dando lugar a un correcto proceso de anonimización, legitimidad y gobernanza del dato. Por otro lado, es necesario el reconocimiento de los sesgos presentes en los conjuntos de datos para no caer en errores técnicos y conceptuales; teniendo en cuenta la cantidad de población encuestada, no es posible hablar de un nivel de representatividad a nivel de ciudad. En definitiva, es necesario dejar en claro que, debido a los datos disponibles, no pueden generarse conclusiones o inferencias a nivel de ciudad y comuna; para evitar los sesgos, estas inferencias deben hacerse a nivel del conjunto de datos mismo.

Objetivos

Objetivo general

Construir un modelo de Analítica de datos que permita la generación de conocimiento a partir de los datos generados para la medición de las diferentes ediciones del Índice de Participación Ciudadana de Medellín.

Objetivos específicos

- Identificar las variables que más influencia tienen en la calidad de la participación ciudadana en Medellín.
- Comparar el impacto de las variables asociadas a la percepción positiva y negativa de la dimensión Efectos.

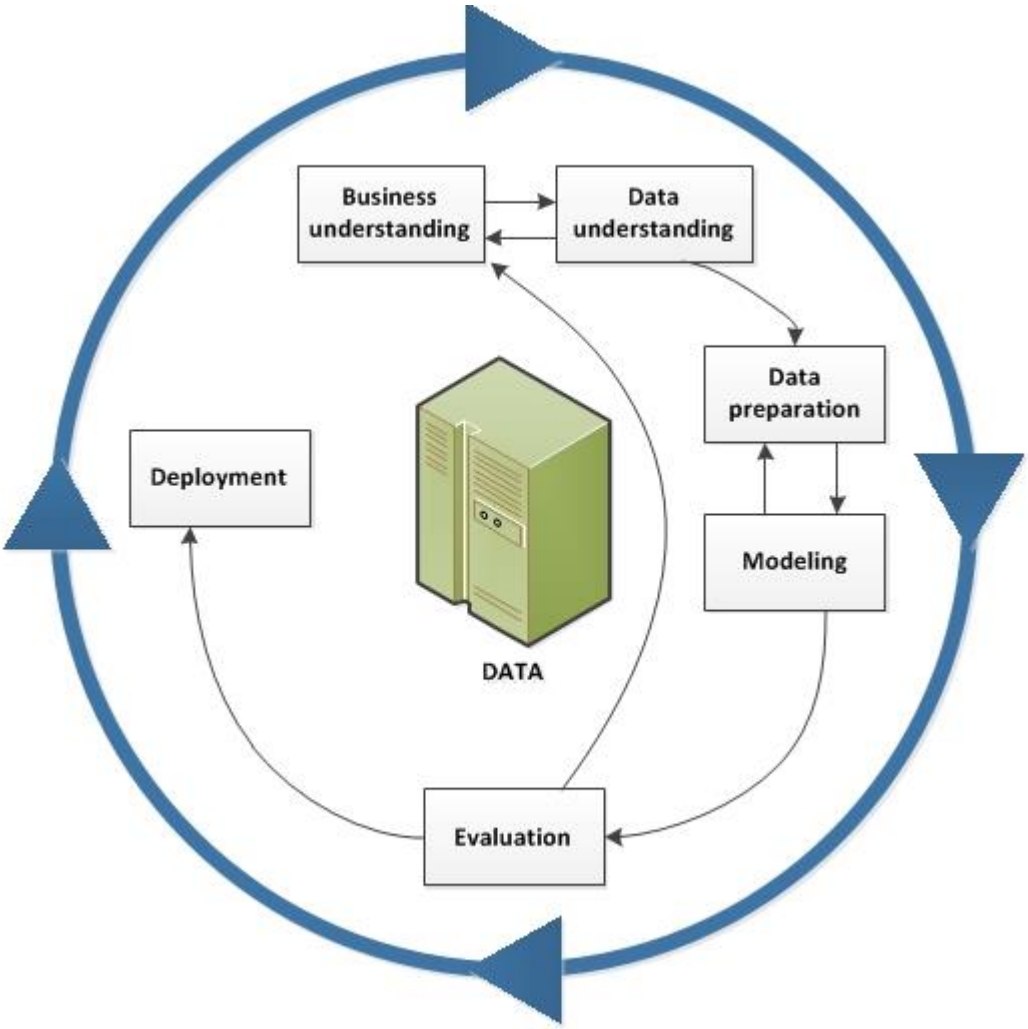
Plan de Análisis

El método estándar para los proyectos de Analítica a nivel empresarial, y en general para cualquier proyecto de este tipo, es el denominado CRISP-DM (*Cross-Industry Standard Process for Data Mining*), propuesto en el año 1999 por la empresa norteamericana IBM[2]. Si bien en el año 2015 la misma empresa propuso un modelo complementario denominado ASUM-DM (*Analytics Solutions Unified Method for Data Mining*), que en esencia recoge la metodología de CRISP-DM y la ajusta a nuevas prácticas desarrolladas en la Ciencia de Datos producto del desarrollo en procesamiento y capacidad de las tecnologías de la computación y la información, tales como el uso de volúmenes de datos enormes, el análisis de texto y el modelado predictivo; sin embargo, CRISP-DM sigue siendo la metodología más usada para este tipo de proyectos, sobre todo casos donde los conjuntos de datos a analizar no se corresponden con la caracterización de Big Data (Angée, Lozano-Argel, Montoya-Munera, Ospina-Arango & Tabares-Betancur 2018; IDECA, 2019).

En este orden de ideas, la metodología propuesta CRISP-DM consta de seis fases: Entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación, y despliegue.

Imagen 1. Ciclo de vida del análisis de datos

[2] https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview



Fuente:https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

1. Entendimiento del negocio

En esta fase se le da relevancia al reconocimiento de los objetivos del proyecto, así como las motivaciones, necesidades y requerimientos que se esperan solventar con el ejercicio de Analítica. Del proyecto, por tanto, se espera convertir en información y conocimiento los datos obtenidos y almacenados por la organización, de modo que pueda construirse un plan de trabajo que dé respuesta a esas necesidades y requerimientos.

Para la Secretaría de Participación Ciudadana existe la necesidad de optimizar el aprovechamiento de los datos generados por las mediciones del Índice de Participación Ciudadana de Medellín - IPCM en sus ediciones de 2017, 2019 y 2021. De este modo, se reconoce que a través de la estadística descriptiva se han generado productos y procesos

muy importantes como consecuencia de la interpretación de los resultados de las dos ediciones anteriores. En esencia, el principal requerimiento implica la agregación de los conjuntos de datos generados en estas mediciones, de modo que pueda explorarse las variables que determinan un resultado bajo y alto en la calidad de la participación ciudadana; dicha información se convertiría en un insumo fundamental para la toma de decisiones y el diseño de políticas públicas.

2. Entendimiento de los datos

Para esta fase ya se conocen las necesidades y requerimientos del proyecto, aunque como puede percibirse en la *Imagen 1*, esto no implica que el proceso se de en una sola dirección de forma ascendente, por el contrario, el esquema metodológico se presenta de manera dialógica, que posibilite regresar, replantear y perfeccionar algunos elementos de las distintas fases. Ahora bien, el entendimiento de los datos requiere, en primer lugar, garantizar el acceso a los conjuntos de datos que van a ser trabajados, y en caso de que estos no existan, recolectarlos.

Esta es una fase exploratoria, donde la intención es entender de la forma más completa el contenido de los conjuntos de datos, proceso que conlleva una descripción detallada a partir de criterios como el tipo de datos, la fuente y origen de los mismos, confiabilidad, volumen o cantidad de datos, velocidad de actualización y método de recolección. Es en esta fase, además, que se realiza la verificación de la calidad de los datos, se evalúa la completitud de la base de datos, los errores y datos nulos, registros duplicados y demás.

En este sentido, el primer paso (1) de este proceso es verificar la existencia de los datos y garantizar su acceso y almacenamiento en medios físicos. Ya con los archivos en el poder (2), se procede a la identificación de los datos: identificar el tipo de archivo para así determinar su carga y procesamiento; reconocer los tipos de datos que contiene el archivo, el número y tipo de variables, y el volumen del conjunto de datos. Superada esta revisión, (3) se procede a ejercer control de calidad sobre los datos; para ello, se evalúa la veracidad y confiabilidad de la fuente de información, el proceso de recolección y el contenido de las bases de datos. Por último, (4) se identifican datos nulos o faltantes, errores de tipeo y/o ortográficos.

Para este ejercicio, se reciben 6 conjuntos de datos generados a partir de las mediciones del IPCM 2017, 2019 y 2021, discriminados por los cuestionarios aplicados a Individuos que participan y Organizaciones sociales y colectivos. La recolección de los datos fue

efectuada vía encuesta, siguiendo las metodologías de muestreo aleatorio simple y muestreo por bola de nieve, cuya aplicación estuvo a cargo del Centro de Estudios de Opinión de la Universidad de Antioquia. Como resultado, se obtuvieron conjuntos de datos contenidos en archivos de tipo Excel (con formato .xlsx) y archivo plano separado por coma (con formato .csv), cuya responsabilidad y custodia recae en la Secretaría de Participación Ciudadana de la Alcaldía de Medellín.

Para la edición del IPCM del año 2017, se recibieron tres archivos: *"BD_Organizaciones20180415.xlsx"*, con los datos recogidos por la encuesta a Organizaciones sociales y colectivos, "*BD_ParticipantesV20180404.xlsx"*, con los datos recogidos por la encuesta a Individuos que participan, y "*Maestro de Preguntas 20180403.xlsx"*, con la traducción de cada una de las variables y los tipos de respuesta para ambas bases de datos; estos conjuntos de datos se componen, en su mayoría, por tipos de datos categóricos que van desde escalas "1a3", "1a5" y "SíyNo", contiene además datos de identificación y variables abiertas; ambos conjuntos de datos, además, cuentan con una cantidad importante de registros nulos, en donde sólo los datos de identificación son válidos: para la BD de Organizaciones sociales y colectivos son 904 los registros nulos, mientras para la BD de Individuos que participan son 51. En los Anexos 1 y 2, se encuentra la información detallada de la Descripción de los datos de Individuos y Organizaciones con un nivel de granularidad por variable.

Tabla 1. IPCM 2017

Nombre archivo	*"BD_Organizaciones 20180415.xlsx"*	"*BD_ParticipantesV2 0180404.xlsx"*	*Maestro de Preguntas 20180403.xlsx"*,
Contenido	Datos recogidos de la encuesta Organizaciones sociales	Datos recogidos de la encuesta Individuos que participan	Traducción de las claves utilizadas en cada una de las variables de las encuestas.
Tipos de datos	En su mayoría,	En su mayoría,	

	categóricos.	categóricos.	
Registros	1506	3510	
Registros nulos	904	51	

Para la edición 2019, fueron dos los archivos recibidos: *"dataPublica_individuos.xlsx"*, con los datos de Individuos que participan, y *"dataPublica_organizaciones.xlsx"*, con los datos de Organizaciones sociales y colectivos; a diferencia de la edición anterior, en este caso se prescinde de un Maestro de Preguntas, de modo que los datos no requieren una traducción, sino que, por el contrario, están expuestos en cada uno de los campos; además, no se encuentran registros nulos. Tal como en la edición anterior, los conjuntos de datos están compuestos, en su mayoría, por datos categóricos, con escalas de "1a3", "1a5" y "SíyNo". Pero sufrió una modificación en el orden de las variables, así como la supresión de algunas de ellas; esto último supone una diferenciación importante respecto a los conjuntos de datos del 2017, que podría ser problemático en aras de una normalización y consolidación de una base de datos general, algo que será asumido en la siguiente fase. Los Anexos 3 y 4 presentan una descripción más detallada de cada una de las variables.

Tabla 2. IPCM 2019

Nombre archivo	*"dataPublica_organizaciones.xlsx"*	"*dataPublica_individuos.xlsx"*
Contenido	Datos recogidos de la encuesta Organizaciones sociales	Datos recogidos de la encuesta Individuos que participan
Tipos de datos	En su mayoría, categóricos.	En su mayoría, categóricos.
Registros	1112	3639

Registros nulos	0	0

Para la edición 2021, se recibieron los archivos *"organizaciones_final.csv"*, que contiene los datos de Organizaciones, y *"participantes_final.csv"*, con los datos de Individuos que participan. A diferencia de los anteriores, estos conjuntos de datos contaban con una buena limpieza de los datos, lo que permitió que no fuesen intervenidos estructuralmente, más allá de pequeñas modificaciones para homogeneizar con las ediciones anteriores; tampoco cuentan con registros nulos. Los Anexos 5 y 6 contienen la descripción al detalle de cada una de las variables.

Tabla 3. IPCM 2021

Nombre archivo	*"organizaciones_final.csv"*	"*participantes_final.csv"*
Contenido	Datos recogidos de la encuesta Organizaciones sociales	Datos recogidos de la encuesta Individuos que participan
Tipos de datos	En su mayoría, categóricos.	En su mayoría, categóricos.
Registros	1125	4263
Registros nulos	0	0

2.1 Base de Datos Individuos que participan – versión 2019

Iniciar por la revisión de este conjunto de datos, en vez de hacerlo por orden histórico se explica, en primer lugar, porque este contiene mayor calidad en los datos y, segundo, porque se advirtieron modificaciones a algunas variables y opciones de respuesta. Por tanto, se intuye que para la segunda medición se efectuaron discusiones y consensos que, por obvias razones, modifican en cierto grado el cuestionario en comparación con la primera medición.

Así pues, tenemos una base de datos con 3639 registros, que se corresponde con la misma cantidad de personas encuestadas; cuenta a su vez con 181 columnas, de las cuales 6 presentan datos de identificación de la encuesta, como los datos de recolección, datos de envío, duración de la encuesta y datos de encuestador (Expresados en las siguientes variables: Date.Collected; Date.Sent Duration.(seconds); User; Latitude; Longitude), el resto contiene las variables de la encuesta. En general, este conjunto de datos es muy completo, pues no registra datos nulos, y sólo se presentan datos vacíos en variables que no son consideradas para la medición del Índice, sino que son complementarias a él, por lo que no representan un impacto en la calidad de la base de datos.

En suma, se realiza una descripción de los datos por variable de forma más detallada en el Anexo 3. Este anexo presenta datos informativos como el Nombre del archivo que contiene la base de datos, los datos que contiene la misma, la velocidad de los datos y el volumen. Las variables, por su parte, se presentan por el Código de la variable, su Tipo de dato, y un complemento del mismo. Los Tipos de datos pueden ser:

- **Geográfico:** Hace referencia a datos que contienen información sobre una ubicación geográfica. Para el IPCM, estos datos proveen información a nivel de comuna y zona. Estos datos son de tipo String, es decir, una cadena de caracteres; por tanto, aun cuando se esté hablando de la Comuna 1, por poner un ejemplo, el dato suministrado no se interpreta como número entero, sino como cadena de caracteres.

- **Fecha:** Este tipo de datos contiene información sobre el tiempo. Por tanto, al asumirse el complemento Date, se entiende este como un formato de medición y presentación del tiempo, ya sea a escala de minutos, horas, días, meses o años. El complemento Year, en específico, se utiliza en la variable que contiene los años de nacimiento de la persona encuestada.

- **Identificación:** Contienen información personal y sensible de la persona encuestada. Se expresa en cadenas de caracteres, por lo que sus entradas se asumen como String.

- **Booleano:** Los datos booleanos se expresan de forma dicotómica, en clave de Verdadero o Falso. En este sentido, las variables con opciones de respuesta “Sí” y “No” hacen parte de este tipo de dato.
- **Categórico:** Son datos cuyas entradas están determinadas a una o varias categorías. Pueden ser Nominales, cuyas opciones de respuesta no tienen un orden o importancia, y Ordinales, cuyo orden sí es importante.
- **Numérico:** Sus entradas son números enteros, descritas en el anexo como Integer.

2.2 Base de datos Individuos que participan – versión 2017

La revisión de este conjunto de datos está determinada por su comparación con la edición 2019, por las razones que fueron esgrimidas previamente. Por tanto, entendiendo que esta fue la primera medición realizada, su aplicación fue en cierto grado experimental, y muchos de los elementos del proceso, incluido el cuestionario, fueron revisados en la edición siguiente. En consecuencia, la revisión de este conjunto de datos está condicionada por la búsqueda de elementos comunes y diferenciales que podrían ser determinantes en la normalización, y su tratamiento está marcado por algunas dificultades de calidad que serán expuestas a continuación.

En términos generales, tenemos una base de datos con 3510 registros de personas que participan, y 169 columnas que en su mayoría contienen las variables que componen el IPCM, ya sea para su medición o complementarias. En su contenido, este conjunto de datos se caracteriza por presentar importantes problemas en la calidad de los datos. En primer lugar, se encuentran 51 registros que están compuestos enteramente por datos nulos, los únicos datos que fueron bien diligenciados fueron los de identificación; ello requiere, por tanto, su eliminación. Además, al igual que el anterior, este también contiene múltiples datos con entradas vacías, aunque en este caso se utiliza el código “-88” como indicador de que no existe registro, o en algunos casos para indicar la respuesta “No sabe, no responde” (estos últimos también están determinados por el código “-98” en algunas variables); para poder diferenciar a qué hace referencia este código, hay que recurrir al Maestro de Preguntas, lo que a simple vista se torna muy confuso. Este problema se soluciona en el proceso de normalización, donde se determina si la entrada está vacía o hace parte de las opciones de respuesta.

Por otro lado, a diferencia de la edición 2019, las respuestas a este conjunto de datos están determinadas por códigos numéricos a los que, para conocer su significado, es obligatorio recurrir al Maestro de Preguntas. Si bien esto es beneficioso en términos operativos, pues los algoritmos implementados tanto para medición como para modelado procesan los datos en clave numérica, su lectura se hace muy poco intuitiva para usuarios ajenos a la medición, e incluso para los que tienen conocimiento sobre ella; por eso es necesario traducir cada entrada para que represente en una cadena de caracteres cada opción de respuesta. En los casos donde la entrada es numérica, o en las variables ordinales que van de I_40_1 a I_40_5, e I_48_1 a I_48_9, las opciones de respuesta numéricas se mantienen, para mantener el formato utilizado en la edición 2019. Se observa, además, que a diferencia de la edición 2019, en esta edición no existen las variables I_17_5c, I_95, I_27a, I_57, ni la serie de variables de I_94_1 a I_94_6. Además, la variable I_56 tiene opciones de respuesta que difieren con las de 2019, pues mientras aquellas eran booleanas, las del 2017 fueron categóricas de tipo nominal; por suerte, a excepción de la I_57, ninguna de estas variables se considera para la medición del IPCM y son complementarias, por lo que no se da una afectación muy grande en términos de calidad.

2.3 Base de datos Individuos que participan – versión 2021

Para la edición 2021, la base de datos recibida ya contaba con una calidad y limpieza de los datos inmejorable. Una muestra de ello es la inexistencia de datos nulos entre los 4263 registros que la componen. Además, las columnas representan exclusivamente a las variables que son tenidas en cuenta en la medición del Índice, por lo que las variables complementarias que se encontraban en las bases de datos de las ediciones anteriores, ya habían sido eliminadas desde la fuente. En consecuencia, los datos están listos para su procesamiento.

2.4 Base de datos Organizaciones sociales y colectivos – versión 2019

Tal como ocurrió en los conjuntos de datos de Individuos, se salta el orden histórico y se le da prioridad al conjunto de datos más reciente, tanto porque en su interior alberga registros con mejor calidad en los datos, como porque se intuye que en la nueva medición se presentaron discusiones y modificaciones a algunas variables y opciones de respuesta. Del mismo modo, al considerarse como el conjunto de datos más actualizado hasta ese

momento, se toma como referencia para efectuar la normalización en la fase de Preparación de los datos.

Cuenta con 1112 registros con datos de organizaciones encuestadas, y a su vez 232 columnas de las cuales 6 representan datos de identificación de la encuesta, como los datos de recolección, datos de envío, duración de la encuesta y datos de encuestador (Expresados en las siguientes variables: Date.Collected; Date.Sent Duration. (seconds); User; Latitude; Longitude), y el resto contiene las variables de la encuesta. En general, este conjunto de datos es muy completo, pues no registra datos nulos, la cantidad de datos vacíos es muy pequeña, y en la mayoría de los casos se dan en variables que solicitan información adicional, es decir, variables de respuesta abierta o variables que están encadenadas a otras variables.

En su mayoría, este conjunto de datos está compuesto por variables categóricas, cuyas opciones de respuesta son cadenas de caracteres que representan cada categoría de las mismas, ya sean de tipo ordinal o nominal. Existen, además, variables cuya pregunta hace referencia a cantidades, por lo que sus opciones de respuesta son de tipo numérico. Las variables de tipo booleano, por tanto, si presentan una variación importante, que requiere ser corregida en la fase de Preparación: Así pues, obviando aquellas cuyas opciones de respuesta son las adecuadas (“Si” y “No”), tenemos en la O_89 respuestas como “Yes” y “No”; y tenemos 0 y 1 como opciones de respuesta en las series de O_10_1 a O_10_7, O_11_1 a O_11_9, O_17_1 a O_17_6, O_18, O_19_1 a O_19_10, O_20_1 a O_20_9, O_27, O_28_1 a O_28_9, O_37_1 a O_37_5, O_38_1 a O_38_9, O_57, O_66_0 a O_66_7, O_70_1 a O_70_6, O_71_1 a O_71_6, y O_72_1 a O_72_9.

2.5 Base de datos Organizaciones sociales y colectivos – versión 2017

La revisión del conjunto de organizaciones está caracterizada por la identificación de problemas serios de calidad en sus datos. Con 1506 registros de organizaciones sociales de la ciudad que fueron encuestadas, encontramos que 904 de ellos son registros nulos; esto es una cifra preocupante, pues estamos hablando de un 60% del total de la base de datos. Por otro lado, y al igual que el conjunto de Individuos de ese año, contiene muchos datos vacíos que fueron representados con el código -88. Además, las respuestas a este conjunto de datos están determinadas por códigos numéricos a los que, para conocer su significado, es obligatorio recurrir al Maestro de Preguntas.

En comparación al conjunto de datos de 2019, encontramos que las variables O_92, O_44_8, O_93, O_70_5, O_70_6, O_71_5, O_71_6, O_72_8 y O_72_9 no existen en esta base de datos; y las variables O_58 y O_59 cambian sus opciones de respuesta, pues mientras para el 2017 estas eran un booleano, para el 2019 se convirtieron en categóricas de tipo nominal. Por desgracia, a excepción de las variables O_92, O_44_8 y O_93, todas estas variables son consideradas para la medición del IPCM, por lo que estamos ante una variación importante en los componentes que perjudica la comparativa entre los resultados.

2.6 Base de datos Organizaciones sociales y colectivos – versión 2021

Para la edición 2021, al igual que la de Individuos que participan, la base de datos recibida ya contaba con una calidad y limpieza de los datos inmejorable. Una muestra de ello es la inexistencia de datos nulos entre los 1125 registros que la componen. Además, las columnas representan exclusivamente a las variables que son tenidas en cuenta en la medición del Índice, por lo que las variables complementarias que se encontraban en las bases de datos de las ediciones anteriores, ya habían sido eliminadas desde la fuente. En consecuencia, los datos están listos para su procesamiento.

3. Preparación de los datos

Posterior a la verificación de la calidad de los datos, se procede a su normalización. En este punto se definen los datos que serán útiles para el análisis, se realiza una limpieza general y se construyen las tablas derivadas de la integración de datos de múltiples fuentes. Para este caso en particular, fue necesario consolidar una base de datos general que agregara todas las mediciones existentes hasta la fecha. Este proceso se llevó a cabo de acuerdo a los siguientes pasos:

- Conversión de los archivos a un formato C*omma Separated Values* (o formato .csv). De este modo, tenemos seis archivos derivados de los conjuntos de datos anteriores: *BD_ParticipantesV20180404.csv* (2017), *BD_Organizaciones20180415.csv* (2017), *dataPublica_individuos.csv* (2019), *dataPublica_organizaciones.csv* (2019), *participantes_final.csv* (2021) y *organizaciones_final.csv* (2021). Estos son los archivos de entrada (inputs) que alimentarán el proceso.

- Unificación de convenciones y tipos de respuesta. Los conjuntos de datos de la edición 2017 contenían un Maestro de Preguntas independiente que traducía los registros introducidos en cada variable. La normalización se hizo con base en los conjuntos de datos de 2019 y se consolidó una Base de datos general, dando como resultado los conjuntos de datos: *BDIndividuosTotal.csv* (Anexo 7) y *BDIndividuosTotal.xlsx* (Anexo 8) para la encuesta de Individuos que participan, *BDOrganizacionesTotal.csv* (Anexo 9) y *BDOrganizacionesTotal.xlsx* (Anexo 10) para la encuesta de Organizaciones sociales y colectivos. Este proceso quedó documentado en el script *ETL.ipynb* (Anexo 11), escrito en Python.

- Anonimización de las bases de datos suprimiendo variables que contienen datos sensibles, como direcciones residenciales y números telefónicos. Además, se suprimieron también algunas de las variables territoriales que causan redundancia, debido a que el código de la comuna presente en la variable S_1 es suficiente para la identificación geográfica. También se suprimieron las variables que fueron descartadas para la medición de 2019, y a su vez, se le aplicó el orden de las variables de esta última edición. Todo ello dio como resultado la eliminación de las siguientes columnas de acuerdo a conjunto de datos y edición del IPCM:

Tabla 4. Columnas eliminadas

Conjunto de Datos	**Columnas eliminadas**	**Total columnas eliminadas**
Individuos 2017	'Unnamed: 0', 'S_00', 'I_3', 'I_4', 'I_5', 'I_6c','I_9_2','I_12','I_17','I_17_5c','I_18','I_19c','I_22c','I_23c','I_26','I_39_2c'	16
Organizaciones 2017	'form','S_00','O_12c','O_13c','O_14','O_15','O_19','O_19_11','O_29','O_33_1','O_33_2','O_35','O_37_6','O_39','O_42','O_42_1','O_42_2','O_42_3','O_42_4','O_43_1','O_43_2','O_43_3','O_43_4','O_43_5','O_43_6','O_43_7','O_43_8','O_43_9','O_44_1','O_56','O_57c','O_63','O_66_7c','O_73','O_86','O_88c'	36

Individuos 2019	'Date.Collected', 'Date.Sent', 'Duration.(seconds)', 'User', 'Latitude', 'Longitude', 'Comuna', 'cm', 'I_3', 'S_1','I_28','I_35'	12
Organizaciones 2019	'Date.Collected', 'Date.Sent', 'Duration.(seconds)', 'User', 'Latitude', 'Longitude', 'Comuna', 'X5', 'S_1'	9
Individuos 2021	Ninguna columna fue eliminada	0
Organizaciones 2021	Ninguna columna fue eliminada	0

Por último, es necesario mencionar la presencia de datos faltantes dentro de los conjuntos de datos, como consecuencia de espacios vacíos o ausencia de respuestas para algunas opciones de respuesta. Como fue advertido en la fase de Entendimiento de los datos, los conjuntos de datos de la edición de 2017 contenían una cantidad importante de registros nulos, en casos donde, salvo los datos de identificación, el resto de datos eran corruptos: esto es, 51 registros para el conjunto de datos de Individuos y 904 para el de organizaciones; estos registros fueron eliminados, dado que no aportan nada a la medición. Sin embargo, también se advierten datos faltantes entre los registros que sí contienen información relevante, configurando así vacíos dentro de dicha información. Esto se debe, además de la ausencia en las respuestas, a la adición de variables para la edición 2019 que no existían en 2017, y que representan para esta última, por tanto, espacios vacíos. En total, encontramos 11632 datos nulos para la base de datos agregada de Individuos que participan, y 4730 para la de Organizaciones sociales y colectivos. Cuando estos datos nulos se encuentran dentro de una variable que es obligatoria para la medición (ver Tabla 5), es necesario su intervención para que no represente un problema de procesamiento en el despliegue de los modelos; de este modo, los datos nulos y vacíos en las variables de tipo string fueron rellenadas con la convención “Sin registro”, mientras en las numéricas se hizo con la convención “0”, ello con el fin de no entrar en conflicto con el modelo y evitar columnas con dos tipos de datos distintos.

En general, las bases de datos no tuvieron grandes reformas en tanto contenido, salvo la corrección de las entradas que presentaban conflicto (como se ha descrito previamente). Sin embargo, se presentó una adición de tres columnas por registro que representan el puntaje del IPCM por dimensión (Condiciones, Prácticas y Efectos) así como el puntaje global. Estas cuatro columnas adicionales servirán en la construcción de modelos como variables independientes (Y), como se verá en la fase de modelado.

Además, previo a la configuración de los modelos (y dentro de los mismos), los datos fueron modificados para procesar sólo las variables que son consideradas en la medición del Índice, es decir, aquellas que tienen peso en la medición. Estas variables son:

Tabla 5. *Variables medición IPCM Individuos que participan*

Individuos	
Datos de Identificación	S_1, I_6, I_7, I_8, I_9, I_10.
Condiciones Territoriales	I_19, I_21_1, I_21_2, I_21_3, I_22 , I_23, I_24, I_26_1, I_26_2, I_26_3, I_27, I_28_1, I_29, I_32, I_33, I_40_1, I_40_2, I_40_3, I_40_4, I_40_5, I_90_1 , I_90_2 , I_90_3 , I_90_4 , I_90_5 , I_90_6 , I_90_7 , I_90_8, I_90_9, I_90_10.
Prácticas y Actores	I_31, I_35_1, I_38, I_51_2, I_51_3, I_51_4, I_52, I_54, I_55, I_48_1, I_48_2, I_48_3, I_48_4, I_48_5, I_48_6, I_48_7, I_48_8, I_48_9, I_91_1, I_91_2, I_91_3, I_91_4, I_91_5, I_91_6, I_91_7, I_91_8, I_91_9, I_91_10.
Efectos	I_30, I_41, I_44, I_45, I_47.

Tabla 6. *Variables medición IPCM Organizaciones sociales y colectivos*

Organizaciones	
Datos de	S_1, O_88, O_89.

Identificación	
Condiciones Territoriales	O_12, O_13, O_21, O_22, O_67, O_68, O_8, O_19_1, O_19_2, O_19_3, O_19_4, O_19_5, O_19_6, O_19_7, O_19_8, O_19_9, O_19_10, O_20_1, O_20_2, O_20_3, O_20_4, O_20_5, O_20_6, O_20_7, O_20_8, O_20_9, O_70_1, O_70_2, O_70_3, O_70_4, O_70_5, O_70_6, O_71_1, O_71_2, O_71_3, O_71_4, O_71_5, O_71_6, O_72_1, O_72_2, O_72_3, O_72_4, O_72_5, O_72_6, O_72_7, O_72_8, O_72_9.
Prácticas y Actores	O_24, O_27, O_48, O_52, O_54, O_57, O_58, O_59, O_90, O_28_1, O_28_2, O_28_3, O_28_4, O_28_5, O_28_6, O_28_7, O_28_8, O_28_9, O_37_1, O_37_2, O_37_3, O_37_4, O_37_5, O_38_1, O_38_2, O_38_3, O_38_4, O_38_5, O_38_6, O_38_7, O_38_8, O_38_9, O_66_0, O_66_1, O_66_2, O_66_3, O_66_4, O_66_5, O_66_6, O_66_7, O_81_1, O_81_2, O_81_3, O_81_4, O_81_5, O_81_6, O_81_7, O_82_1, O_82_2, O_82_3, O_82_4, O_82_5, O_82_6, O_82_7.
Efectos	O_40, O_41, O_45, O_49, O_55, O_61_1, O_61_2, O_61_3, O_61_4, O_61_5, O_62_1, O_62_2, O_62_3, O_62_4, O_62_5, O_62_6.

4. Modelado

Para el análisis de datos multivariados o multidimensionales es conveniente encontrar un método para detectar patrones o regularidades que permitan presentar la información de manera simplificada, para que de tal forma sea más fácil de comprender. En este esfuerzo existen una multitud de técnicas estadísticas de resumen de información, que en su mayoría han tenido como punto de partida el análisis de información de tipo cuantitativo. Sin embargo, la aplicación de estas técnicas a datos de tipo categórico usualmente enfrenta dificultades. La razón de esto se explica en que estos métodos, para detectar relaciones entre variables de tipo cuantitativo, se basan en medidas de tendencia central -usualmente el promedio o la mediana- o en medidas de distancia tales como la distancia euclidiana.

Sin embargo, tanto estas medidas de distancia como de tendencia central que usan los métodos mencionados anteriormente no son directamente aplicables a datos de tipo categórico. Por ejemplo, el criterio de distancia pierde su significado sustantivo al comparar la disimilitud entre categorías de variables. Por ejemplo, ¿cuál es la distancia que se asigna a un individuo que aporta una respuesta afirmativa (un *sí*) cuando la mayoría de respuestas son negativas (un *no*)? ¿Tiene sentido hablar de distancias entre categorías?

Evidentemente, las anteriores interrogantes dejan claro que desde un punto de vista conceptual es posible juzgar como erróneo trasladar de manera directa métodos desarrollados para datos cuantitativos al contexto de datos categóricos o cualitativos. Para los métodos de agrupamiento, podemos encontrar como una alternativa a los clusters K-means o K-medias, los llamados clusters K-modas. Reconociendo las particularidades de los datos de tipo categórico o cualitativo, los cluster K-modas utilizan como medida de tendencia central la *moda* (es decir, el dato más frecuente), y como medida de disimilitud la distancia de Hamming (Huang, 1998). Luego, cada individuo quedará asignado a un grupo cuyos patrones de respuesta sean los más parecidos posible a los suyos.

4.1 Recategorización de los datos usando segmentación

Imagen 2. Escala de referencia para la interpretación sociopolítica y territorial de la calidad de la participación ciudadana en Medellín.

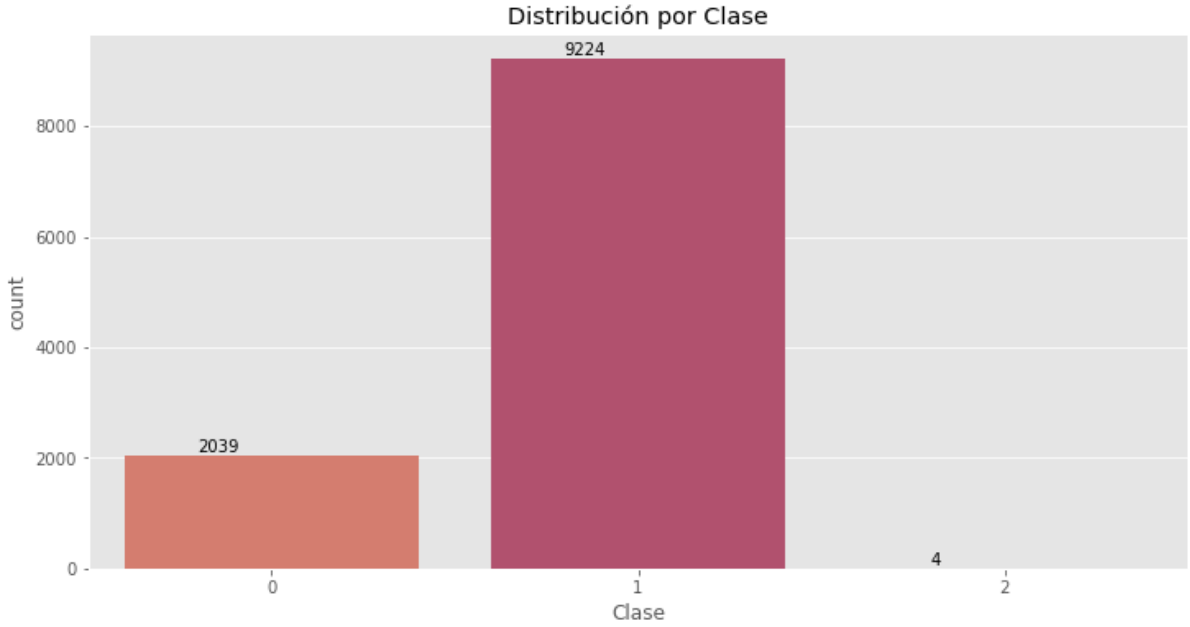
Escala ordinal	BAJA		MEDIA			ALTA	
Escala de Intervalos	**Muy baja**	**Baja**	**Media baja**	**Media**	**Media alta**	**Alta**	**Muy alta**
Valores	0 a 0,19	0,20 a 0,29	0,3 a 0,39	0,4 a 0,59	0,6 a 0,69	0,7 a 0,79	0,8 a 1

Fuente: Alcaldía de Medellín (2019). *Calidad de la participación ciudadana de Medellín. Resultados de la medición del Índice de Participación Ciudadana de Medellín 2019*.

Teniendo como propósito construir un modelo de clasificación para los datos del IPCM, se hace necesario definir las categorías para la predicción del modelo. En primera instancia, estas categorías de clasificación se definieron a partir de la Escala de referencia para la interpretación sociopolítica y territorial de la calidad de la participación ciudadana en

Medellín (ver **Imagen 2**). Estas categorías tienen relación directa con el puntaje del IPCM por cada registro, y se traducen en una escala ordinal que comprende una calidad Baja (de 0 a 0.29), Media (0.3 a 0.69) y Alta (0.7 a 1) de la Participación Ciudadana. En consecuencia, tras etiquetar los datos bajo estas condiciones, se obtiene una distribución de los datos tal como se observa en la siguiente gráfica:

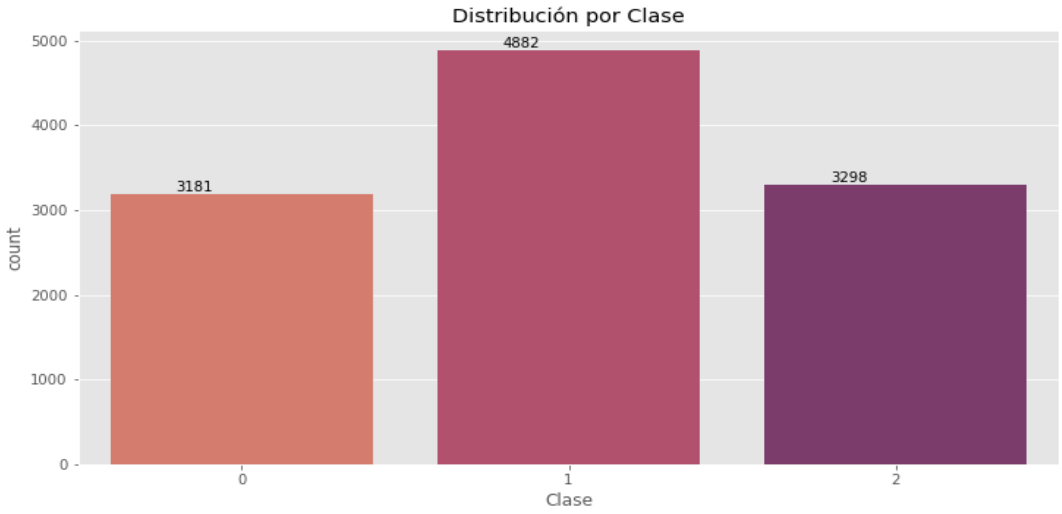
Gráfico 2. *Distribución por Clase*



No obstante, este tipo de categorización presenta serios problemas en el balance de los datos, pues el 81.8% del total de los datos está concentrado en la clase Media y sólo 4 registros (0.00035%) fueron catalogados con una calidad Alta de la Participación. Ello implica una necesidad de replantearse esta categorización, pues una concentración tan alta de los resultados en una sola clase genera un sesgo de aprendizaje que va a afectar el desempeño de los modelos y su predicción.

Por tanto, se tomó la decisión de utilizar herramientas exploratorias sobre los datos a partir de algoritmos de segmentación (clustering) y además, reconociendo que existe una prevalencia de datos categóricos, se opta por el algoritmo K-modas para esta nueva categorización, dando como resultado la siguiente distribución:

Gráfico 3. *Distribución por Clase*



Como consecuencia, el algoritmo K-modas permitió una categorización más balanceada sin comprometer los objetivos del proyecto, pues si bien en un principio este proceso de etiquetado estaba directamente anclado al puntaje del IPCM, este nuevo proceso se relaciona en mayor medida con los patrones de respuesta de cada individuo, configurando las clases bajo este criterio. Por lo tanto, se adoptan estas clases para la ejecución de los algoritmos. A continuación, se ofrece una breve descripción teórico-conceptual de cada modelo y herramienta usada en este proceso:

4.2 Modelos de clasificación

Los modelos de clasificación se caracterizan porque el análisis producto de sus resultados corresponde a un problema de clasificación; por tanto, el modelo intenta asignar una categoría de manera exitosa de acuerdo a la información obtenida por las clases asignadas previamente, en el proceso de etiquetado. A continuación, se dará una breve descripción de los modelos considerados para el proyecto, dejando su análisis para la fase de Evaluación.

I. Regresión Logística

La Regresión logística es un método estadístico cuya función es la modelación de la probabilidad de una variable dependiente binaria de acuerdo a una serie de variables independientes; una de sus principales particularidades es que es un modelo linear generalizado (GLM en sus siglas en inglés) cuya salida es una variable dependiente de tipo categórico (Diez, Çetinkaya-Rundel & Barr, 2019; García, Molina, Berlanga, Patricio,

Bustamante & Padilla, 2018), los predictores -o variables independientes- pueden ser tanto continuos como categóricos.

En principio, el que este modelo contenga una variable de salida de tipo binario puede suponer una limitación, en especial para conjuntos de datos con múltiples clases. Para evitar ello, es utilizada la Regresión Logística multinomial, muy útil para modelos con una variable dependiente de tipo nominal con más de dos categorías, y mantiene la característica de que sus predictores, o variables independientes, sean continuas o categóricas (Pando & San Martín, 2004). Su función se expresa de la siguiente manera:

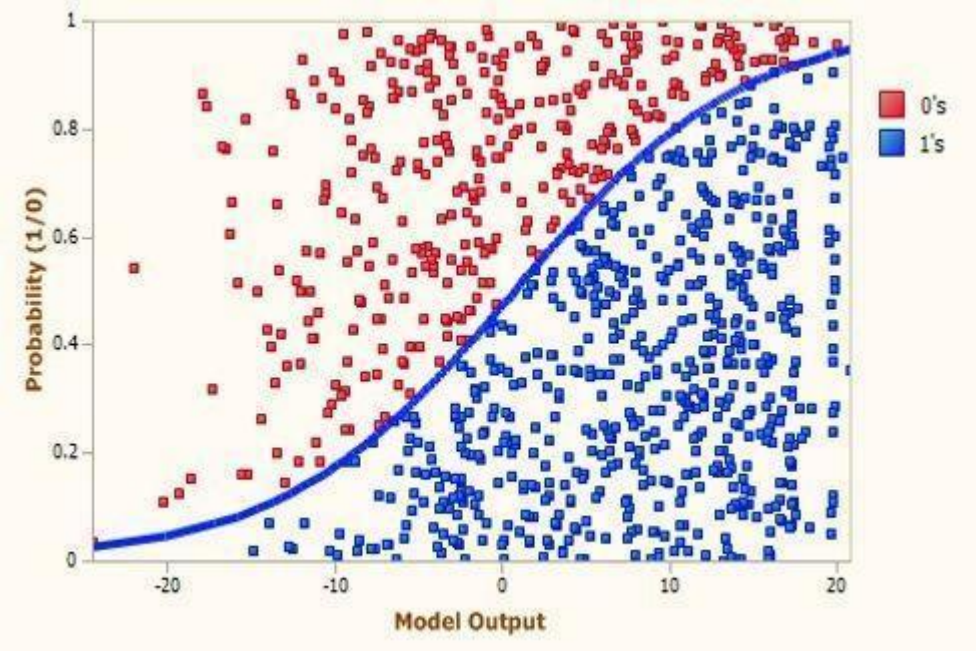
$$log(\frac{P}{1 - P}) = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_i\chi_i$$

La frontera de decisión, por su parte, está determinada por la función logística (también conocida como función sigmoide), que constituye el valor de la probabilidad Y, y se obtiene con la inversa del logaritmo natural, de este modo:

$$p(Y) = \frac{e^{\beta_0+\beta_1\chi_1+\beta_2\chi_2+\ldots+\beta_i\chi_i}}{1 + e^{\beta_0+\beta_1\chi_1+\beta_2\chi_2+\ldots+\beta_i\chi_i}}$$

Esta función se expresa gráficamente de la siguiente forma:

Imagen 3. Logistic Regression



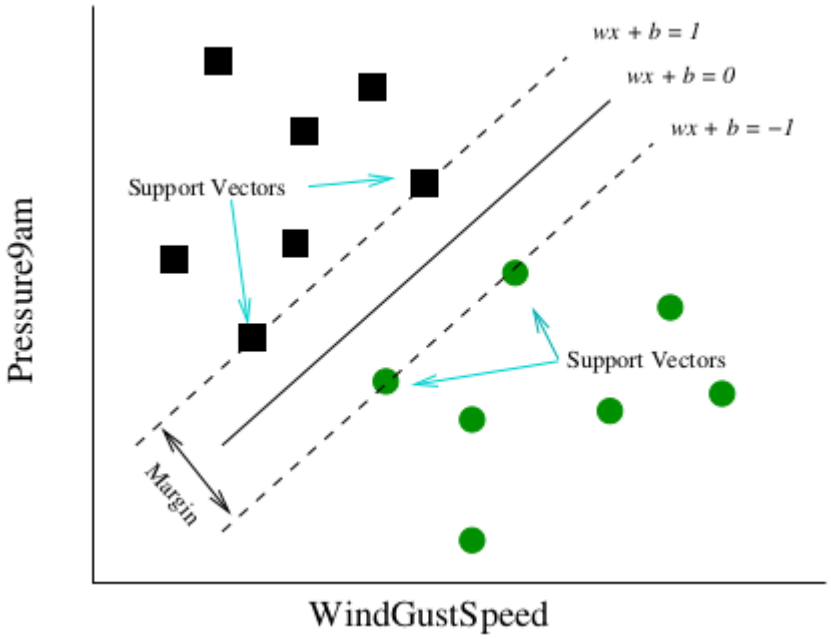
Fuente: https://www.statdeveloper.com/regresion-logistica-en-python/

II. Máquinas de Vectores de Soporte

Este algoritmo se usa tanto para clasificación como para regresión. Las máquinas de soporte vectorial (support vector machine) son uno de los métodos en los que se produce un pequeño número de combinaciones lineales de las variables de entrada originales. Estas combinaciones lineales se emplean luego como variables independientes en una regresión (Hastie et al., 2009).

Este algoritmo, conocido de forma más frecuente por su nombre original (Support Vector Machine o SVM), fue desarrollado en la década de 1990 por Vladimir Vapnik y se destaca por su aplicación en problemas de clasificación de texto, reconocimiento de voz, análisis de imágenes y series de tiempo (Deng, Tian & Zhang, 2012). Si bien en un principio fue desarrollado para clasificación binaria, su desarrollo lo ha llevado a su aplicación en clasificación múltiple y regresión. En su funcionamiento interno, construye sus resultados a partir de construir subconjuntos de los datos de entrenamiento, denominados Vectores de Soporte, y su clasificación está definida por la construcción de un Hiperplano que determina una región o margen que separa las clases (Williams, 2011; Kowalczyk, 2017; Müller & Guido, 2016), tal como se observa en el siguiente esquema:

Imagen 4. Support Vector Machine



Fuente: Williams, G. J. (2011). Data mining with Rattle and R: The art of excavating data for knowledge discovery. Springer.

III. Vecinos más cercanos (KNN)

Este algoritmo se usa tanto para clasificación como para regresión. A diferencia del SVM, en donde la clasificación se determina por la distancia o margen al hiperplano que separa cada clase, el algoritmo KNN (K-Nearest Neighbors) identifica la clase de acuerdo a la determinación de la clase mayoritaria entres sus vecinos más próximos.

En este sentido, este algoritmo es considerado dentro de la categoría de no-paramétrico (García, Molina, Berlanga, Patricio, Bustamante & Padilla, 2018; Müller & Guido, 2016).

De las métricas de distancia disponibles para este algoritmo, fueron probadas durante la fase de modelado la distancia Euclidiana, Manhattan, Chebyshev y Minkowski. A continuación, se presentan las fórmulas a partir de las cuales se calcula cada una de las distancias (Zhu & Zhang, 2020):

Distancia Euclidiana

$$d(x,y) = \sqrt{\sum_{i=1}^{k} (X_i - Y_i)^2}$$

Distancia Manhattan

$$d(x,y) = \sum_{i=1}^{k} |X_i - Y_i|$$

Distancia Chebyshev

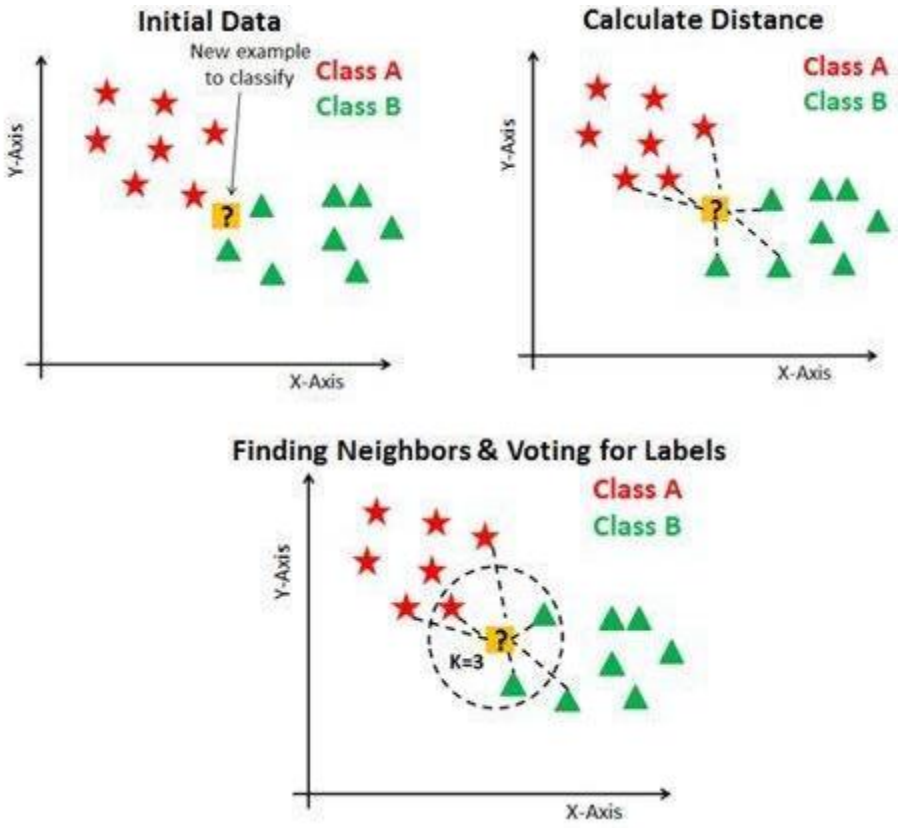
$$d(x,y) = max_i|X_i - Y_i|$$

Distancia Minkowski

$$d(x,y) = (\sum_{i} |X_i - Y_i|^p)^{\frac{1}{p}} = (\sum_{i} |d_i|^p)^{\frac{1}{p}}$$

En consecuencia, el algoritmo KNN se representa de la siguiente manera:

Imagen 5. K-Nearest Neighbors



Fuente: https://github.com/artifabrian/dynamic-knn-gpu

IV. Bosques aleatorios

Los árboles aleatorios, o Random Forest, es un algoritmo que puede ser utilizado tanto para clasificación como para regresión: por tanto, su variable dependiente puede ser categórica o continua. Es un algoritmo de ensamble, que se conforma por un conjunto determinado de árboles de decisión, cada uno modelado y entrenado a partir de un subconjunto distinto y aleatorio de datos de entrenamiento. La predicción, por tanto, es el resultado del agregado de todos los modelos entrenados, tomando la media de las predicciones. Este método es conocido como **Bagging -**o bootstrap aggregation-, y tiene como fin la reducción de la varianza en el algoritmo (Williams, 2011; García, Molina, Berlanga, Patricio, Bustamante & Padilla, 2018).

Imagen 6.

Algorithm 15.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Fuente: Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

La fórmula, para este modelo, varía de acuerdo a su uso[3]. En este sentido, si el modelo es ajustado para problemas de regresión, la fórmula está determinada por defecto de acuerdo al Error cuadrático medio (MSE), y se representa de la siguiente forma:

$$H(Q_m) = \frac{1}{n_m}\sum_{y \in Q_m}^{N} (y - \underline{y}_m)^2$$

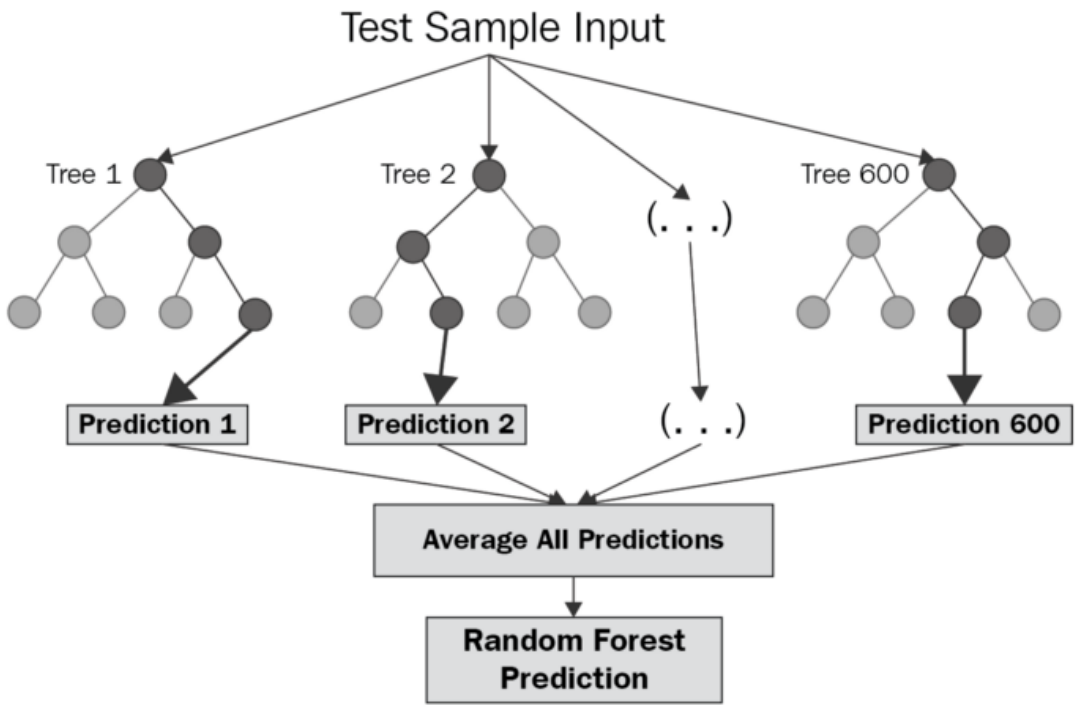
Para problemas de clasificación, se utiliza por defecto la técnica del Índice Gini, representado de la siguiente manera:

[3] En este caso se usa la formulación matemática usada por la librería SciKit-Learn, dado que son sus modelos los que se utilizan en este ejercicio. Para más información, consulte: https://scikit-learn.org/stable/modules/tree.html#tree

$$H(Q_m) = \sum_k \quad p_{mk}(1 - p_{mk})$$

El esquema general del algoritmo se presenta de la siguiente manera:

Imagen 7. Random Forest



Fuente: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

4.3 Ejecución de los modelos

En este punto, es necesario recordar que uno de los objetivos planteados inicialmente es verificar si existen variables que determinan un resultado bajo o alto en la calidad de la Participación Ciudadana, y es importante observar cuáles son estas variables.

Para este fin, se construyeron dos modelos en cuadernos Jupyter, escritos en lenguaje Python, que corresponden a los modelos de clasificación (Anexo 12) y al análisis de las variables más importantes de acuerdo al algoritmo Feature Importance (Anexos 13 y 14), se explicará cada uno de ellos de forma detallada más adelante. En ambos modelos, se dio uso de las herramientas contenidas en la librería Scikit-Learn, y se tomó como punto

de partida los conjuntos de datos obtenidos de la fase de Preparación de los datos, denominados *BDIndividuosTotal.csv* (Anexo 7) y *BDOrganizacionesTotal.csv* (Anexo 9).

La experimentación de cada modelo se realizó con los datos de Individuos, de modo que, al encontrarse el modelo con mejores resultados, se replicó el proceso en cada subconjunto de datos. Por último, al ser conjuntos de datos que están compuestos principalmente por datos categóricos, se utilizaron las herramientas de preprocesamiento Ordinal Encoder y Standard Scaler: la primera realiza una transformación de los datos de entrada a números ordinales, y evita errores de procesamiento; Standard Scaler, por su parte, estandariza todos los datos de entrada llevando su media a 0 y su desviación estándar a 1, a través de la siguiente fórmula:

$$X_* = \frac{X - mean(X)}{std(X)}$$

En la construcción de un modelo de aprendizaje automatizado, es necesario recordar que en esencia cada modelo ejecuta operaciones matemáticas y estadísticas a través de la construcción de algoritmos. Por tanto, para un modelo de clasificación, los datos de entrada del modelo deben dividirse entre variables independientes (X) y una variable dependiente (Y), que va a asumir el rol de variable objetivo.

En el caso del IPCM, para las variables independientes, que se asumen en el conjunto de datos X, se tomaron las variables que fueron utilizadas para la medición del IPCM y que ya fueron referenciadas en la *Tabla 5* y *Tabla 6*. Para los modelos de clasificación, por su parte, la variable predictora (Y) se compone de clases o etiquetas categóricas, donde el algoritmo predice, precisamente, a qué clase pertenece el dato de entrada. Para estos modelos, la clasificación puede ser binaria (0 o 1) o multiclase, que permite asignar múltiples categorías; el caso del IPCM se ubica en esta última, debido a que existen tres niveles de clasificación que indica la calidad de la Participación Ciudadana de Bajo a Alto.

I. Partición de datos

En relación con la distribución de los datos, previo a la entrada del modelo se realiza una partición del conjunto de datos inicial. Esta partición se hace en dos sentidos: en primer lugar, se dividen los datos entre las variables predictoras (X), que consta de las variables que configuran las preguntas de la encuesta, y la variable de respuesta (Y), que está compuesta por la categorización realizada en el paso anterior. La segunda partición deriva

en un conjunto de entrenamiento (train) y uno de testeo (test), tanto para X como para Y. Generalmente, esta partición se realiza de acuerdo a una distribución 80/20, quedando así los conjuntos X_train con el 80% de los datos y sus etiquetas (y_train). Con este conjunto de entrenamiento, el modelo se entrena para intentar predecir de manera correcta las etiquetas (y_test) del 20% de los datos restantes (X_test). Así pues, teniendo lista esta partición, se realiza el entrenamiento de los cuatro modelos propuestos.

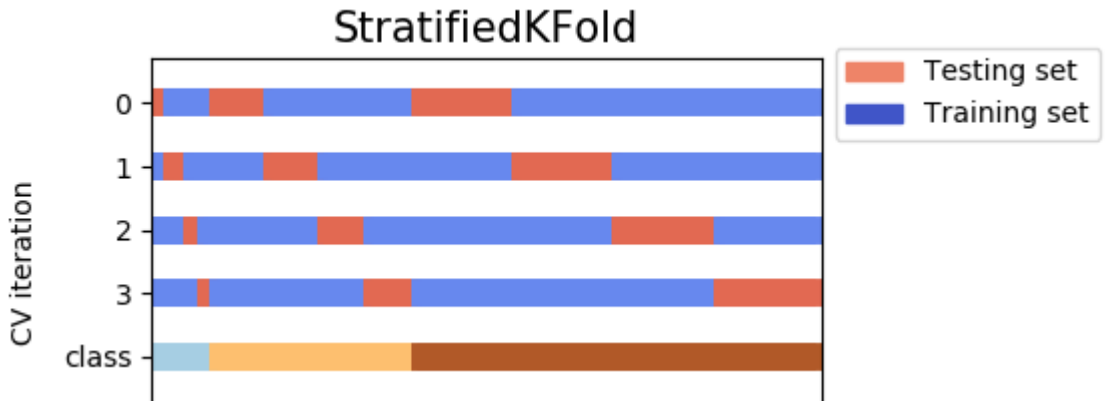
Para el entrenamiento, se utiliza la herramienta GridSearchCV[4], que recibe, además de los datos de entrada, una malla de hiperparámetros con los que entrena el modelo, así como una estrategia de Validación Cruzada; este método funciona a partir del procesamiento de múltiples modelos al tiempo, cada uno combinando las opciones de hiperparámetros de modo que evalúe el óptimo para el caso específico. Si bien ello implica un aumento en el gasto de recursos computacionales, el retorno de este ejercicio es construir el mejor modelo posible, es decir, con los hiperpatrámetros óptimos, lo que redunda en mayor precisión y al mismo tiempo en evitar un sobreajuste (overfitting)[5] del modelo.

La validación cruzada (o cross-validation) se configura como una técnica que también se utiliza para no caer en un sobreajuste o sobreentrenamiento del modelo. Para el presente proyecto, se utilizó como técnica de validación cruzada la herramienta StratifiedKFold, que busca evitar la superposición de los datos de prueba, barajándolos y respetando la proporción de las clases o etiquetas (de ahí que siga una lógica de estratificación). En otras palabras, si en un conjunto de datos existe una concentración de una clase en algunos registros de forma sucesiva, el algoritmo tradicional (K-Fold) puede sobreaprender este patrón e interpretar que otros registros sucesivos pertenezcan a dicha clase sin serlo; esta herramienta busca evitar que esto suceda (Kononenko & Kukar, 2007). La siguiente gráfica presenta un mejor entendimiento de su funcionamiento.

Imagen 8. Stratified K-Fold

[4] Esta herramienta hace parte de la librería SciKit Learn. Para mayor información, se remite a la documentación oficial contenida en https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[5] El sobreajuste (u overfitting) hace referencia al hecho en donde un modelo se ajusta tanto a los datos de entrenamiento que obtiene resultados casi perfectos, sin embargo, sólo funciona bien con esos datos, volviéndose inservible en el procesado de otro conjunto de datos distinto.



Fuente:https://www.it-swarm-es.com/es/python/diferencia-entre-stratifiedkfold-y-stratifiedshufflesplit-en-sklearn/834810622/

5. Evaluación

En esta fase, se determina el, o los modelos que se comportaron de forma más eficiente en el procesamiento de los datos, y que a su vez arrojaron resultados que son válidos, confiables, y se corresponden con los objetivos del proyecto. Esto último determina lo que se denomina *criterios de rendimiento*, que no es más que la coincidencia entre los resultados obtenidos en la fase de modelado y las necesidades y requerimientos de la organización determinados en la primera fase del proyecto. De este modo, como salida de esta evaluación se generan las inferencias y conclusiones obtenidas por el proceso de Analítica.

5.1 Validación

Ya con los modelos entrenados, se valida su eficiencia en la predicción y clasificación. Para ello se utiliza la matriz de confusión, que compara en dos ejes las etiquetas reales y las predicciones del algoritmo, corroborando así su exactitud y rendimiento. De este modo, establece una relación entre cuatro parámetros: Verdaderos Positivos (True Positive o TP), Falsos Positivos (False Positive o FP), Verdaderos Negativos (True Negative o TN) y Falsos Negativos (False Negative o FN). De acuerdo con la relación entre estos parámetros, se determinan las métricas de evaluación del modelo, que validan su rendimiento. Estas son:

- **Precision (Precisión):** Evalúa la ratio entre las predicciones positivas acertadas que dio el modelo sobre el total de predicciones positivas. Se construye a partir de la siguiente fórmula:

$$Precision = \frac{TP}{TP + FP}$$

- **Accuracy (Exactitud):** Mide el porcentaje de aciertos del modelo, es decir, las predicciones correctas sobre el total de las predicciones. Si bien este es una de las métricas más utilizadas en la industria, para conjuntos de datos con clases muy desbalanceadas puede resultar engañosa, pues el acierto de la clase mayoritaria no implica, necesariamente, un ajuste correcto del modelo a los datos; por el contrario, indica sobreajuste. Se construye a partir de la siguiente fórmula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall (Sensibilidad):** Indica la probabilidad de clasificar correctamente una clase. Se construye a partir de la siguiente fórmula:

$$Recall = \frac{TP}{TP + FN}$$

- **F1 score:** Se calcula combinando las métricas de Precisión y Sensibilidad. Para conjuntos con clases muy desbalanceadas, como el nuestro, esta métrica es una de las más útiles. Se construye a partir de la siguiente fórmula:

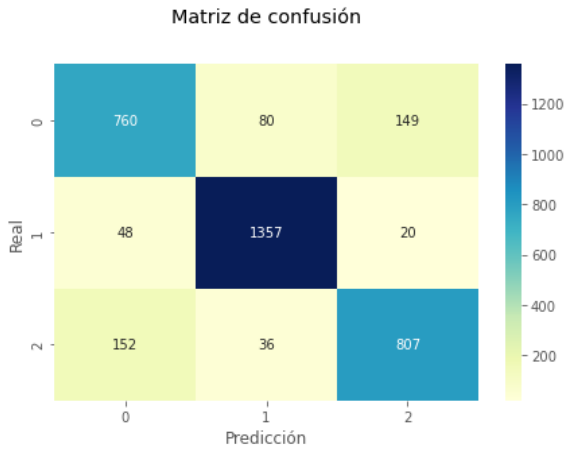
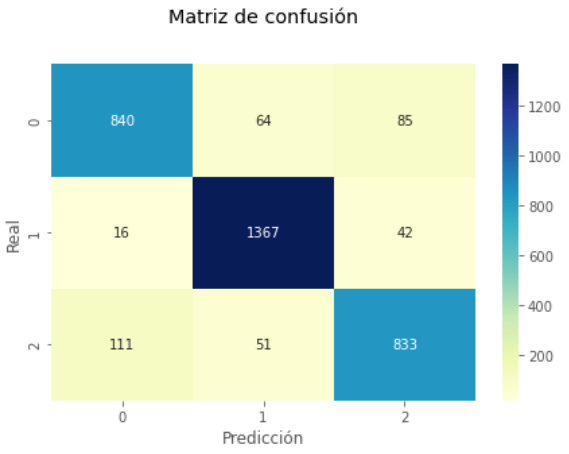
$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

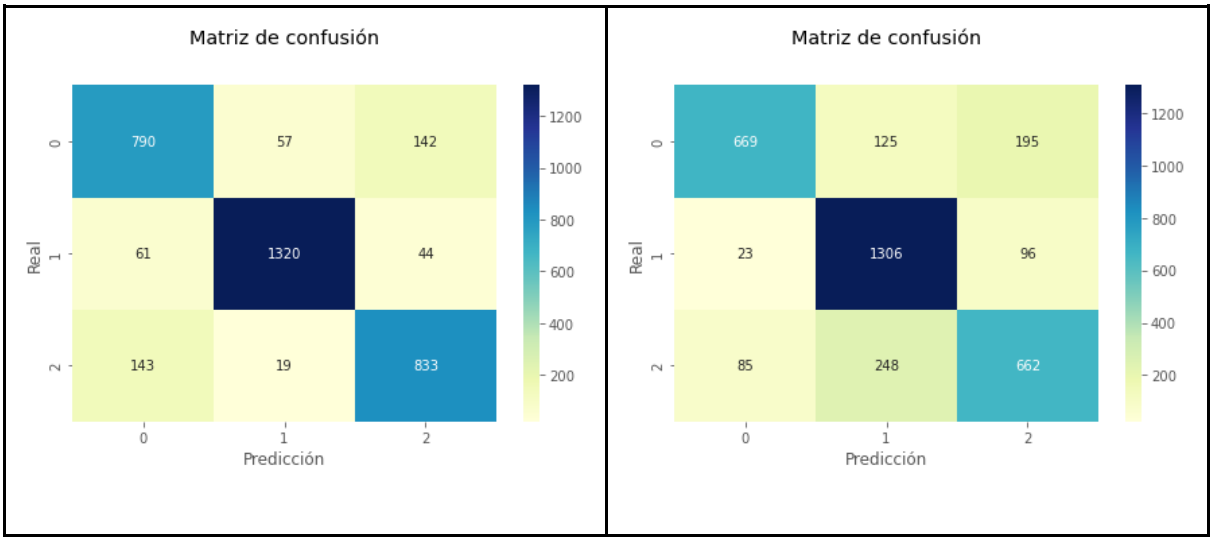
Por último, cada métrica de evaluación es calculada a partir de dos parámetros: macro avg y weighted avg. La diferencia entre ambos es su enfoque, pues mientras la primera se realiza sobre cada clase sin tener en cuenta su peso, la segunda sí lo considera; es decir, para aquellos conjuntos de datos con clases desbalanceadas, el weighted avg considera la proporción de cada clase respecto al total, reconociendo así aquellas clases que tienen más representación (y por tanto peso) sobre el total. Por tanto, el parámetro macro avg va a castigar mucho más a los casos en los que se presentan clases desbalanceadas. En definitiva, entre más alto sea el parámetro, mejor es el resultado.

6. Resultados

Ya descritos los criterios de evaluación, se utiliza la Matriz de Confusión y las Tablas de desempeño que sintetizan las métricas de evaluación; es así como se determina el modelo con el mejor rendimiento y calificación. La Matriz de Confusión consta de un gráfico de distribución de frecuencia que cruza los datos reales con la predicción del modelo, es a partir de este cruce que se determinan los Verdaderos Positivos, Verdaderos Negativos, Falsos Positivos y Falsos Negativos reseñados en la sección anterior, las métricas de evaluación son producto del cálculo realizado con estos criterios, y se presenta en las Tablas de Desempeño. Después de entrenados los modelos, estos fueron sus resultados:

Tabla 7. *Matriz de Confusión por modelo*

Regresión Logística	**Random Forest Classifier**
	
Support Vector Machine	**K_Nearest Neighbors**



A continuación, se presentan las tablas de desempeño de cada modelo, discretizadas por métrica de evaluación y por clase:

Tabla 8. Regresión logística

	Precision	Recall	F1-score	Cantidad
0	0.79	0.77	0.78	989
1	0.92	0.95	0.94	1425
2	0.83	0.81	0.82	995
Accuracy			0.86	3409
Macro avg	0.85	0.84	0.85	3409
Weighted avg	0.86	0.86	0.86	3409
Hiperparámetros: C = 0.1010280808080808, class_weight = 'balanced', penalty = 'l1', solver = 'liblinear'.				

Tabla 9. Random Forest

	Precision	Recall	F1-score	Cantidad
0	0.87	0.85	0.86	989
1	0.92	0.96	0.94	1425
2	0.87	0.84	0.85	995
Accuracy			0.89	3409
Macro avg	0.89	0.88	0.88	3409
Weighted avg	0.89	0.89	0.89	3409
Hiperparámetros: class_weight = 'balanced', max_features = 'log2', n_estimators = 500.				

Tabla 10. Support Vector Machine

	Precision	Recall	F1-score	Cantidad
0	0.79	0.80	0.80	989
1	0.95	0.93	0.94	1425
2	0.82	0.84	0.83	995
Accuracy			0.86	3409
Macro avg	0.85	0.85	0.85	3409
Weighted avg	0.86	0.86	0.86	3409

Hiperparámetros: C = 1000, class_weight = 'balanced', gamma = 0.0001, kernel = 'rbf'.

Tabla 11. K-Nearest Neighbors

	Precision	Recall	F1-score	Cantidad
0	0.86	0.68	0.76	989
1	0.78	0.92	0.84	1425
2	0.69	0.67	0.68	995
Accuracy			0.77	3409
Macro avg	0.78	0.75	0.76	3409
Weighted avg	0.78	0.77	0.77	3409
Hiperparámetros: leaf_size = 20, metric = minkowski, n_neighbors = 10, p = 1, weights = 'distance'				

Aunque los métodos dan valores similares en las métricas Recall y F1-Score (a excepción del método K-Nearest Neighbors, que presentó los resultados más pobres), el que mejor salió calificado fue el **Random Forest**, pues obtuvo el mejor puntaje de los cuatro en todas las métricas de desempeño, y mantiene buen equilibrio entre los puntajes de las clases. Esto lo posiciona como el modelo de clasificación óptimo para nuestro conjunto de datos de entre los modelos entrenados. Ahora bien, aun cuando el balance entre las clases es satisfactorio después de la aplicación del algoritmo K-modas, sigue existiendo un leve desbalance en favor de la clase 1, es así como buscando minimizar su impacto en el análisis final y en la salida del modelo, se ajustó el hiperparámetro “class_weight='balanced'”.

6.1 Feature Importance

Ya evaluados los modelos, y asumiendo el Random Forest Classifier como el modelo óptimo para el procesamiento de los datos, se procede a la aplicación de las funciones de Feature Importance. Estos métodos están orientados para calcular las variables más importantes, o dicho de otro modo, que más influyen en el resultado final, de modo que se refiere a la medida de la contribución individual de cada característica para el clasificador (Saarela & Jauhiainen, 2021). En consecuencia, evalúa cuáles son los fenómenos que más contribuyen en la calidad de la participación ciudadana en cada una de las mediciones.

Para ello tenemos tres herramientas orientadas a tal fin. La primera de ellas, (i) Feature Importance, es la más básica de todas y su aplicación se fundamenta en la varianza. Esta función toma cada una de las variables y evalúa su promedio en la reducción de varianza a nivel general (Saarela & Jauhiainen, 2021). Sin embargo, el principal defecto de este método es que tiende a mostrar preferencia frente a variables que contienen registros numéricos o respuestas categóricas de alta cardinalidad (es decir, de muchas opciones de respuesta), en consiguiente, estamos ante un sesgo que podría interpretar como muy importantes variables que simplemente cumplen una de estas dos características.

La segunda herramienta, (ii) denominada Permutation Feature Importance (Permutación de Variables Importantes) busca superar las dificultades y sesgos que presenta la herramienta por defecto (Feature Importance), y para ello toma los valores de una variable, los modifica, y realiza una nueva predicción de forma interna con esta variación; de este modo, después de repetir el proceso con cada variable determina cuáles de ellas influyen en mayor medida en la predicción final y así establece su importancia. Si bien este método exige más recursos computacionales, es recomendable debido a que tiene resultados más precisos. Las primeras dos herramientas están contenidas en la librería Scikit-Learn. Por último, la tercera herramienta (iii) es denominada SHAP (SHapley Additive exPlanations), y se basa en la teoría de juegos para calcular cómo contribuye cada variable en la predicción (Lundberg & Lee, 2017)[6].

[6] Para mayor información sobre este método y su desarrollo, puede remitirse al repositorio de Github del autor: https://github.com/slundberg/shap

Así pues, ya habiendo ejecutado el Random Forest para todo el conjunto de datos buscando evaluar su rendimiento, se repite la operación para cada medición (Individuos que participan y Organizaciones sociales y colectivos; ediciones 2017, 2019 y 2021 en conjunto; y por dimensión) para evaluar cuáles fueron las variables que más influyeron en los resultados obtenidos. Además, a cada una se le aplicó las tres herramientas mencionadas, buscando comparar entre sí y encontrar variables que se repitan en importancia.

Los resultados de este ejercicio están alojados en el Anexo 13 - Clas_Feature_Importance.ipynb. El análisis será realizado de la siguiente forma: En primer lugar(i), el ejercicio se aplicará con los datos agregados de las ediciones 2017, 2019 y 2021, tanto para el conjunto de Individuos como el de Organizaciones sociales y colectivos; con ello se espera responder al primer objetivo. En segundo lugar (ii) el análisis se replica a nivel de dimensiones, buscando con ello responder el segundo objetivo y además, generar inferencias del comportamiento de los datos por cada dimensión.

I. Individuos que participan

Buscando obtener resultados más específicos, este ejercicio de validación de las variables más importantes fue realizado a cada subconjunto de datos. Sin embargo, se tomó la decisión de realizar el análisis con los datos agregados de las ediciones 2017-2019-2021 procurando entrenar un modelo más robusto y, por tanto, obtener resultados más confiables. Por tanto, agregar los resultados en un sólo modelo entrenado permitió mejorar sus resultados, asumiendo todas las respuestas a los cuestionarios como una sola base de datos y obteniendo así inferencias más precisas y confiables.

Tras probar cada una de las herramientas de Feature Importance previamente mencionadas, se observó que la herramienta SHAP presentó los resultados más satisfactorios, pues nos trae un ranking de importancia de variables (Gráfico 4) en el que presenta una síntesis de los resultados de las dos herramientas anteriores. En ese sentido, encontramos que repiten la mayoría de variables que aparecieron en el Feature Importance y en el Permutation Importance, a excepción de **I_55** (¿Está usted está dispuesto a liderar procesos de alguna organización o colectivo, distinto a la JAC o JAL?), que no aparece en estas. Una de las grandes potencialidades que tiene esta herramienta, es también poder observar el grado de importancia de las variables seleccionadas, y si estas aportan de forma positiva o negativa a la salida del modelo (Gráfico 5).

Tabla 12. Variables más importantes - *Individuos que participan*

Código	Descripción de variable	Dimensión
I_90_10	¿Ud conoce alguno de los siguientes mecanismos de control social?: i. Audiencias públicas	Condiciones Territoriales
I_90_1	¿Ud conoce alguno de los siguientes mecanismos de control social?: a. Veedurías ciudadanas	Condiciones Territoriales
I_90_9	¿Ud conoce alguno de los siguientes mecanismos de control social?: h. Acciones de Tutela	Condiciones Territoriales
I_90_6	¿Ud conoce alguno de los siguientes mecanismos de control social?: e. Peticiones, quejas, reclamos y sugerencias (PQRS) ante autoridades competentes	Condiciones Territoriales
I_90_4	¿Ud conoce alguno de los siguientes mecanismos de control social?: c. Derechos de Petición	Condiciones Territoriales
I_90_3	¿Ud conoce alguno de los siguientes mecanismos de control social?: b. Juntas de Vigilancia	Condiciones Territoriales
I_41	Durante los últimos 24 meses, ¿Usted ha participado en iniciativas, propuestas o proyectos que beneficien su comuna o corregimiento?	Efectos
I_90_7	¿Ud conoce alguno de los siguientes mecanismos de control social?: f. Denuncias y demandas	Condiciones Territoriales
I_52	¿Participa usted en la toma de decisiones para mejorar su comuna o corregimiento?	Prácticas

I_90_2	¿Ud conoce alguno de los siguientes mecanismos de control social?: j. Acciones populares	Condiciones Territoriales

Gráfico 4. *SHAP - Individuos que participan*

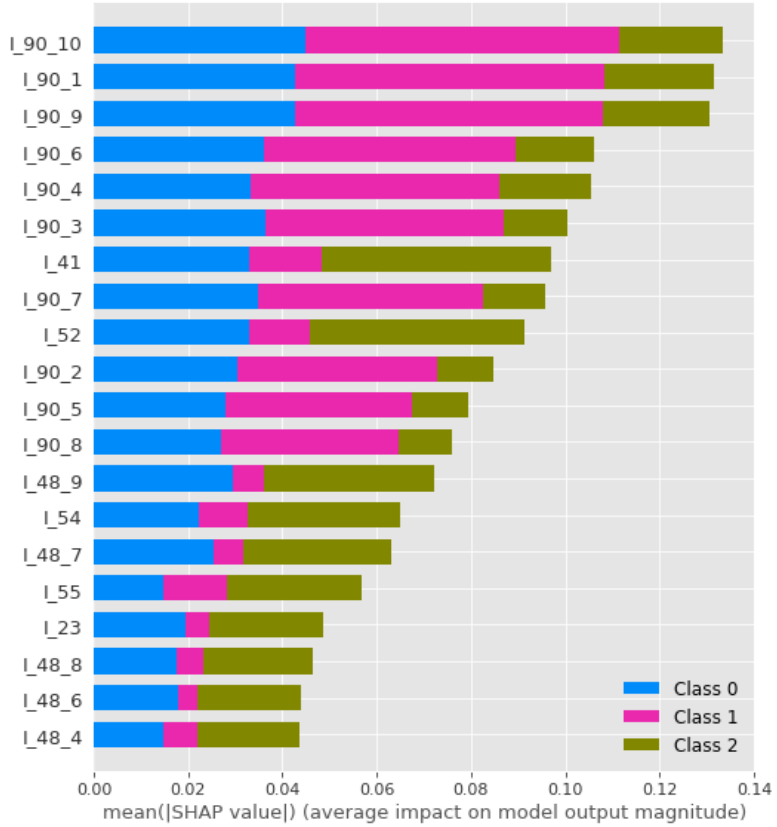
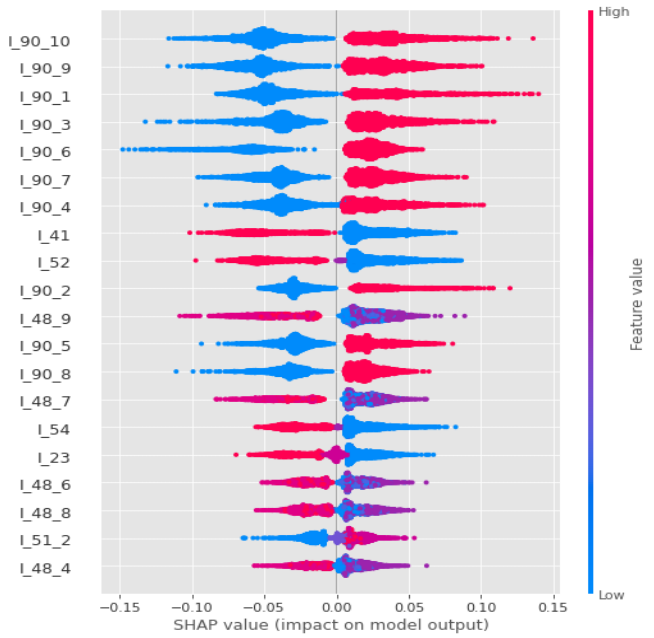
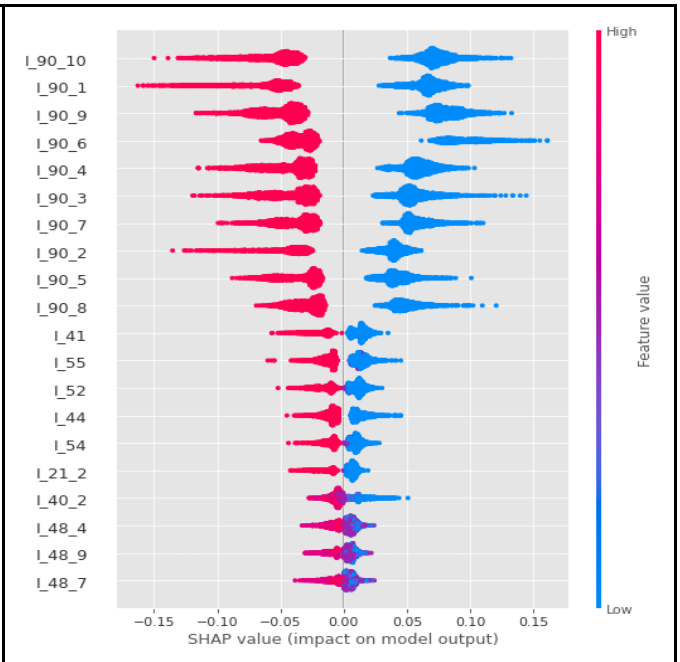


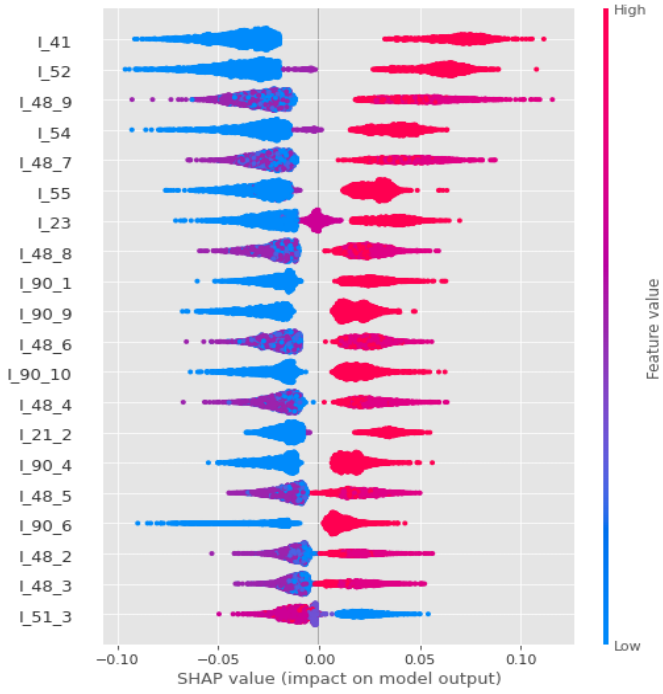
Gráfico 5. *SHAP value impact por clase. Individuos que participan*

Clase 0	***Clase 1***





Clase 2



De los resultados obtenidos para el conjunto de datos de Individuos que participan, resalta la mayoría de variables asociadas a las Condiciones Territoriales, pues 8 de las 10

variables más importantes hacen parte de esta dimensión. Los resultados cobran mayor relevancia al reconocer que estas 8 variables hacen parte del grupo **I_90**, que hace referencia al conocimiento que los individuos tienen de los mecanismos de control social. Claramente, su conocimiento y manejo tienen una fuerte influencia en la calidad de la participación ciudadana para los participantes.

Completan el listado la variable I_41 de la dimensión Efectos, que se relaciona con la participación del individuo en iniciativas, propuestas o proyectos de su territorio (comuna o corregimiento) en las que percibe que ha significado alguna transformación efectiva de su realidad, y la variable I_52 de la dimensión Prácticas y Actores, que está relacionada con la percepción de la participación del individuo en la toma de decisiones de su comuna o corregimiento. A pesar de pertenecer a dimensiones de análisis distintas, ambas variables, están fuertemente relacionadas en la dinámica política local y cobran sentido al reconocer que, en tanto el individuo sienta que su participación es efectiva y genera incidencia, tanto en la toma de decisiones como en la consecuencia de las mismas sobre su territorio, la calidad de la participación se verá profundamente afectada, ya sea de manera positiva o negativa.

II. Análisis por dimensiones - Individuos que participan

Ya realizado el análisis general con todos los registros agregados de las ediciones 2017-2019-2021, se propone replicar el proceso con cada una de las dimensiones. Esto no sólo contribuye a la respuesta del segundo objetivo, sino además permite generar mayor conocimiento sobre cómo se comportan los datos por cada una de estas esferas de la realidad social que se denominan dimensiones de la participación ciudadana, observando cuáles son las variables más importantes en cada una de ellas.

a. Condiciones Territoriales

En la medición del Índice de Participación Ciudadana de Medellín (IPCM), es necesario recordar que además del puntaje general, también se construyen indicadores por cada dimensión de la participación ciudadana. Por tanto, si se desea realizar un análisis para la dimensión Condiciones Territoriales, la variable dependiente (Y) debe estar construida a partir de ese indicador. Por tanto, se construyen nuevos modelos con el mismo proceso reseñado a lo largo del presente informe, pero ahora los datos de entrada están compuestos por las variables independientes (X) constituidas por las variables específicas

de la dimensión Condiciones Territoriales. De este modo, para esta dimensión, obtenemos los siguientes resultados:

Gráfico 6. *Matriz de confusión. Condiciones Territoriales - Individuos*

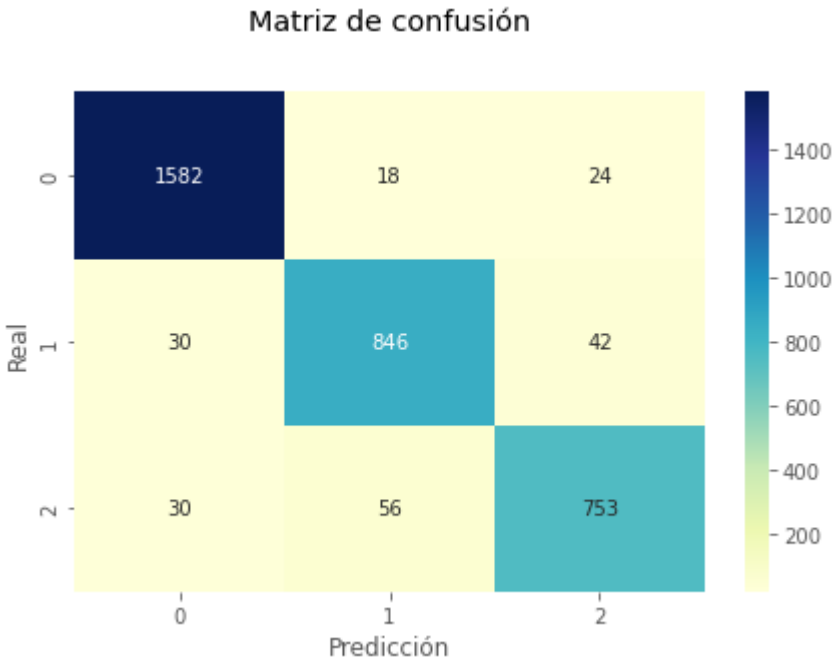


Tabla 13. *Métricas de evaluación. Condiciones Territoriales - Individuos*

	Precision	Recall	F1-score	Cantidad
0	0.96	0.97	0.97	1624
1	0.92	0.92	0.92	918
2	0.92	0.90	0.91	839
Accuracy			0.94	3381
Macro avg	0.93	0.93	0.93	3381

Weighted avg	0.94	0.94	0.94	3381
Hiperparámetros: class_weight = 'balanced', max_features = 'log2', n_estimators = 700.				

Estamos, por tanto, ante un modelo muy robusto que se ve beneficiado por la reducción de variables a considerar a comparación del modelo general. Para este análisis por dimensiones, se mantiene la estrategia de realizar el etiquetado de los datos por medio de la herramienta exploratoria de segmentación K-modas. De modo que, a pesar del pequeño desbalance en favor de la clase 0, el comportamiento de la predicción en todas las variables es satisfactorio. Por tanto, reconociendo este desbalance, el criterio Weighted avg es el mejor para definir la métrica, por lo que observamos un muy buen rendimiento al reconocer el puntaje de la Sensibilidad del modelo -o Recall- de 0.94, y el de F1 también de 0.94.

Gráfico 7. *SHAP. Condiciones Territoriales - Individuos*

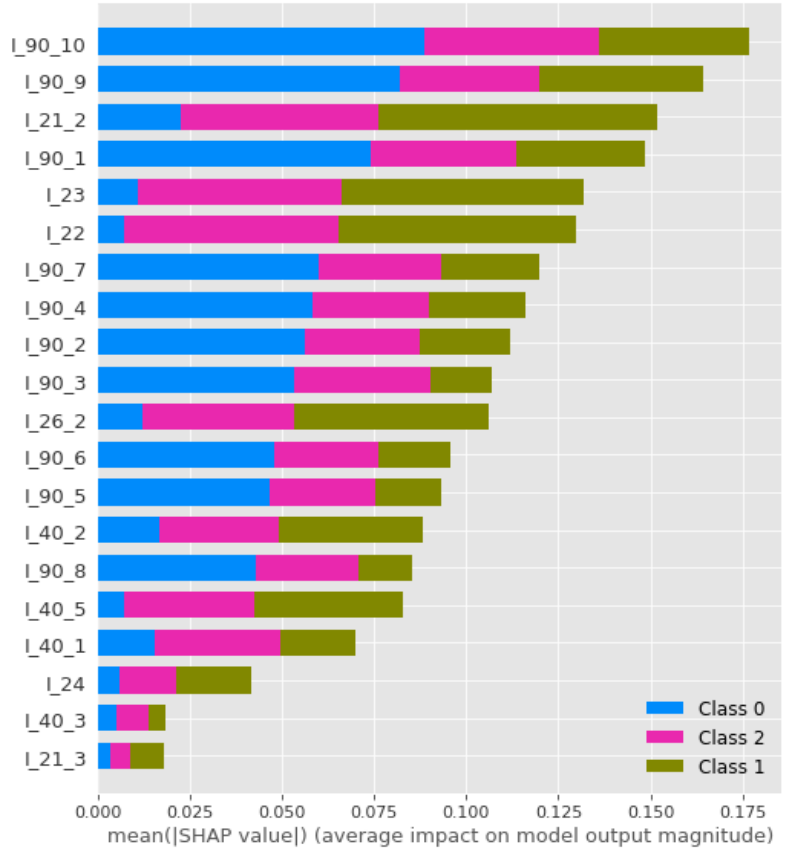


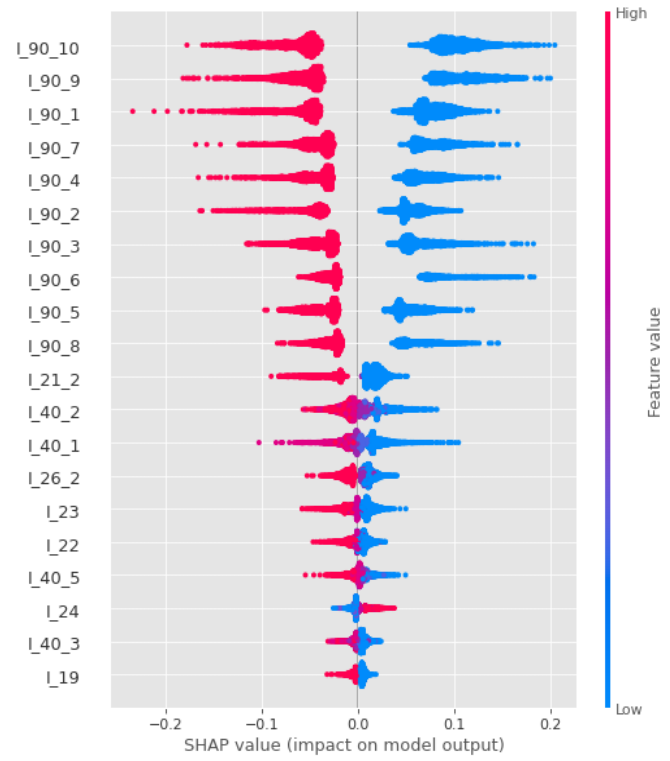
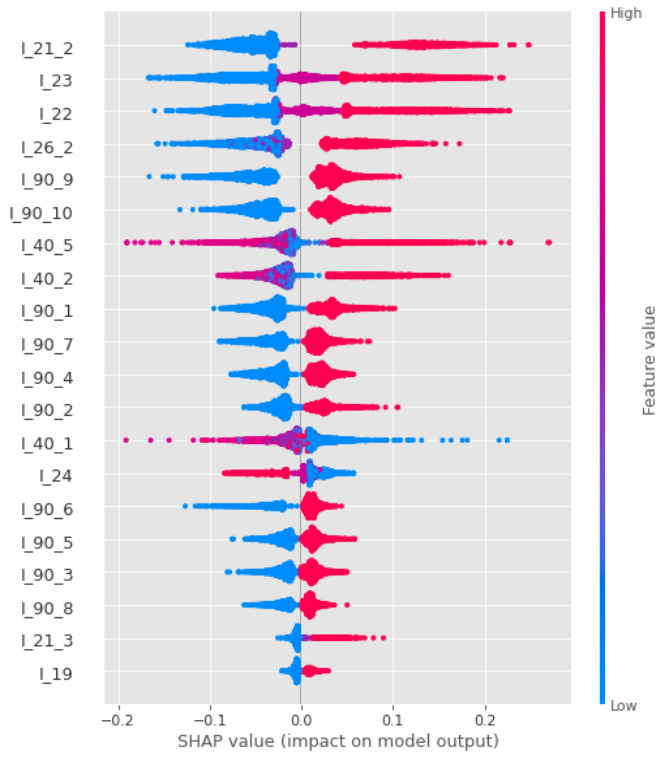
Tabla 14. Variables más importantes. Condiciones Territoriales - Individuos

I_90_10	¿Ud conoce alguno de los siguientes mecanismos de control social?: i. Audiencias públicas
I_90_9	¿Ud conoce alguno de los siguientes mecanismos de control social?: h. Acciones de Tutela
I_21_2	En el último año ha usado internet para convocar o participar de reuniones entre vecinos u otros grupos de ciudadanos
I_90_1	¿Ud conoce alguno de los siguientes mecanismos de control social?: a. Veedurías ciudadanas
I_23	En caso de una reunión o evento relacionados con actividades comunitarias o ciudadanas, tiene acceso al préstamo de sedes comunales (Por ejemplo, sedes de las Juntas de Acción Comunal, sedes sociales de edificios o unidades residenciales).
I_22	Le han prestado espacios en instituciones públicas (por ejemplo,escuelas, colegios , casas de gobierno de justicia, uvas, casa de la cultura, parques biblioteca) para realizar reuniones y eventos relacionados con actividades comunitarias o ciudadanas.
I_90_7	¿Ud conoce alguno de los siguientes mecanismos de control social?: f. Denuncias y demandas
I_90_4	¿Ud conoce alguno de los siguientes mecanismos de control social?: c. Derechos de Petición
I_90_2	¿Ud conoce alguno de los siguientes mecanismos de control social?: j. Acciones populares

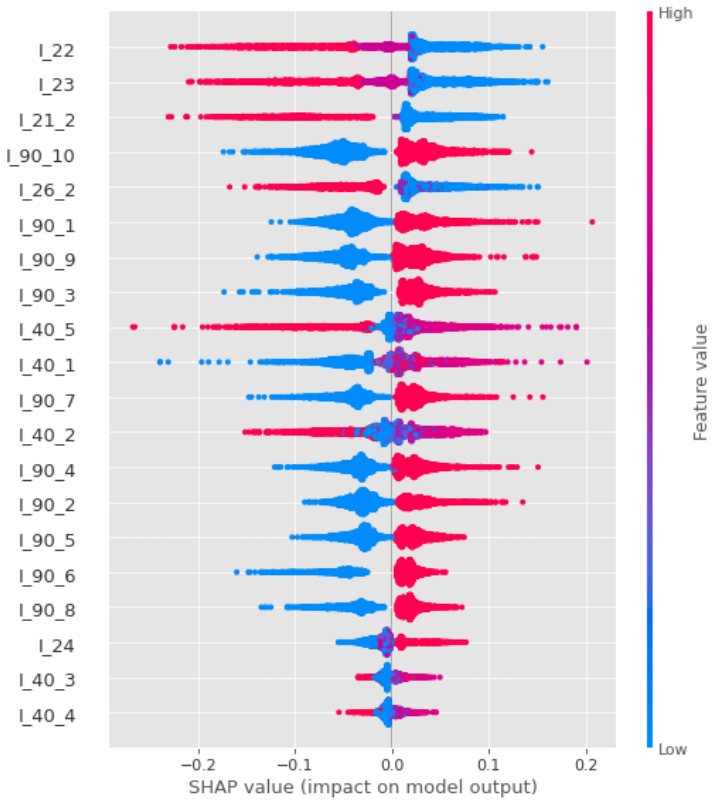
I_90_3	¿Ud conoce alguno de los siguientes mecanismos de control social?: b. Juntas de Vigilancia

Para la dimensión Condiciones Territoriales prevalecen las variables que hacen referencia al conocimiento que tienen los individuos que participan sobre los mecanismos de participación ciudadana. Para poner en perspectiva, 7 de las 10 variables más importantes hacen parte este grupo. Complementan la lista dos variables asociadas al acceso a espacios públicos para el desarrollo de actividades asociadas a la participación (sean estos espacios comunales o instituciones públicas como escuelas, colegios y/o casas de la cultura) y una variable asociada al uso de internet para la convocatoria y participación en escenarios de participación locales.

Gráfico 8. *SHAP value impact por clase. Condiciones Territoriales - Individuos*

Clase 0	***Clase 1***
	

Clase 2



El impacto de cada variable varía de acuerdo con la clase. Un ejemplo de ello es la variable **I_22** (Préstamo de espacios en instituciones públicas para reuniones y eventos relacionados con actividades comunitarias), la cual presenta muchos registros que impactan de manera fuerte y negativa a la clase 2 pero a su vez impacta de manera fuerte y positiva a la clase 1; si bien a simple vista ello podría reconocerse como una contradicción, lo que indica el impacto precisamente es qué tanto contribuye o no contribuye los registros de una variable a la clase. Es así como se evidencia que no todas las clases tienen la misma relación de impacto con las mismas variables, y que si bien a nivel general las variables del grupo I_90 (Conocimiento de los mecanismos de participación) son las de mayor impacto, esta relación no se refleja en todas las clases. Ello no implica, no obstante, que exista una alta variación entre las clases y las variables más importantes a nivel global (Gráfico 7); de hecho, salvo la variable I_26_2 (En los últimos 24 meses, dejó de asistir a reuniones: Por falta de tiempo) que aparece en las clases 1 y 2, las variables más importantes suelen ser las mismas.

b. Prácticas y Actores

Tal como en la dimensión anterior, para la dimensión Prácticas se genera un mini-modelo en donde las variables independientes (X) están compuestas por aquellas que conforman esta dimensión, y la variable dependiente (Y). Este fue su rendimiento:

Gráfico 9. *Matriz de confusión. Prácticas - Individuos*

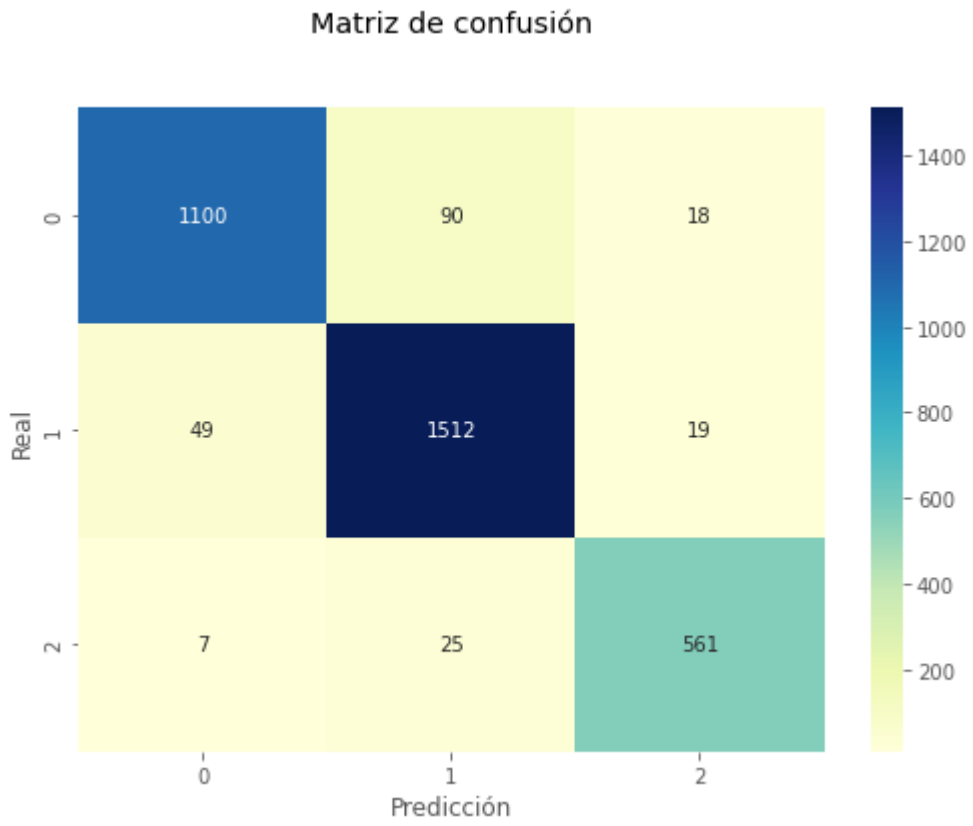


Tabla 15. *Métricas de evaluación. Prácticas - Individuos*

	Precision	Recall	F1-score	Cantidad
0	0.95	0.91	0.93	1208
1	0.93	0.96	0.94	1580
2	0.94	0.95	0.94	593

Accuracy			0.94	3381
Macro avg	0.94	0.94	0.94	3381
Weighted avg	0.94	0.94	0.94	3381
Hiperparámetros: class_weight = 'balanced', max_features = 'log2', n_estimators = 500.				

El mini-modelo demuestra tener buen rendimiento, pues sus métricas de evaluación Recall y F1 le dieron un puntaje de 0.94 en ambos casos, en su parámetro Weighted avg.

Gráfico 10. *SHAP. Prácticas - Individuos*

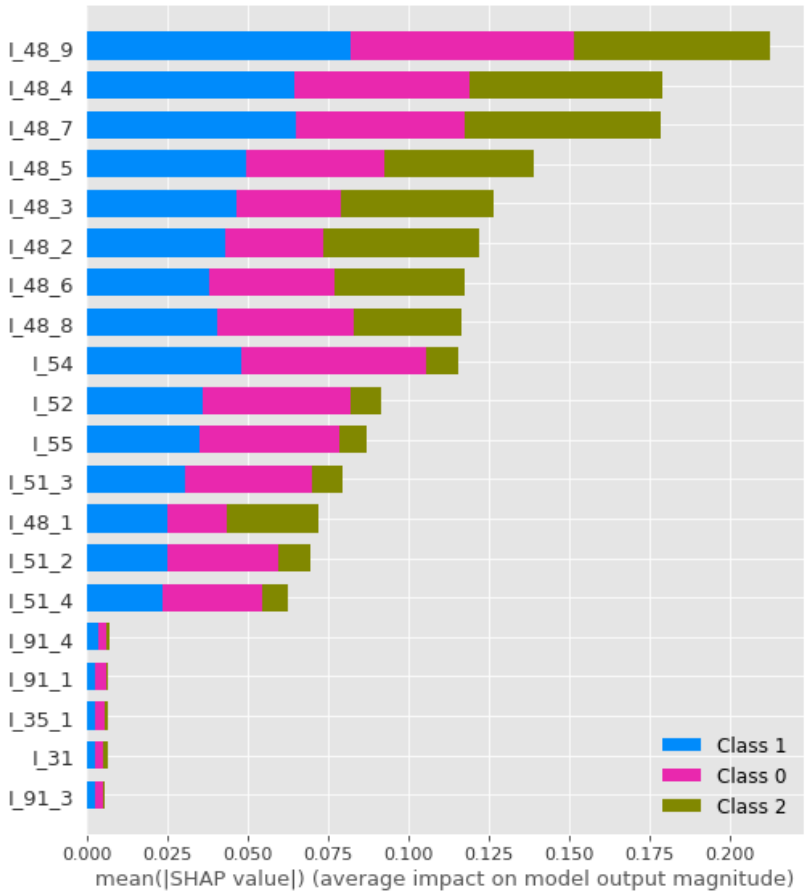
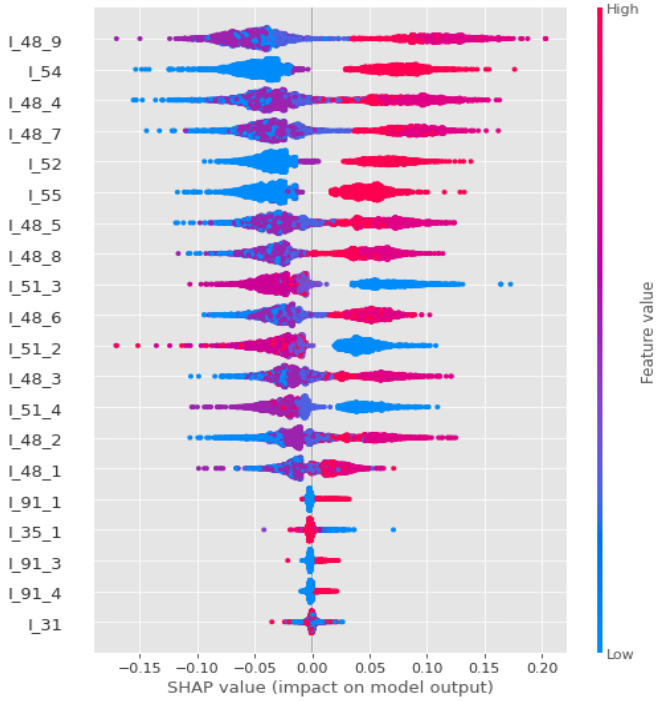


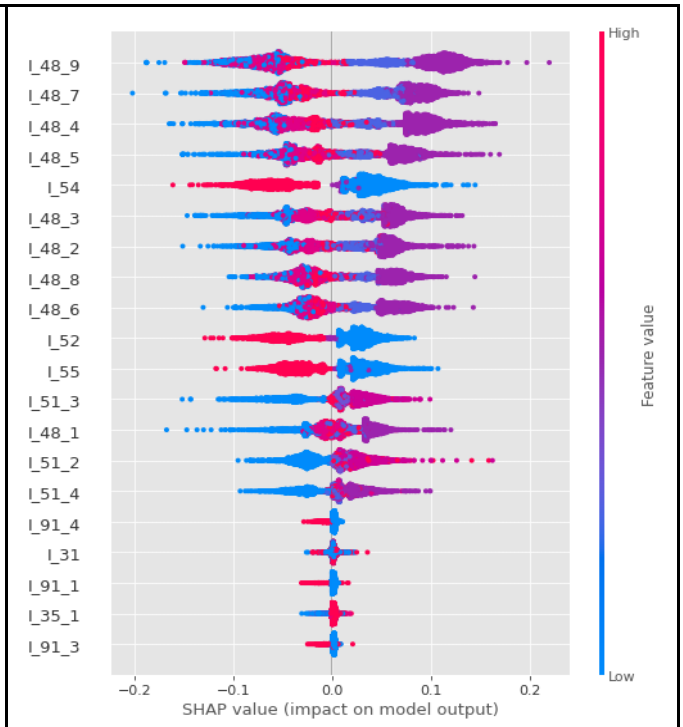
Tabla 16. Variables más importantes. Prácticas - Individuos

I_48_9	Califique siendo 1 poco y 5 mucho, qué tanto confía en: El Consejo Comunal o Corregimental
I_48_4	Califique siendo 1 poco y 5 mucho, qué tanto confía en: La Personería Municipal
I_48_7	Califique siendo 1 poco y 5 mucho, qué tanto confía en: Los integrantes de las Juntas Administradoras Locales JAL
I_48_5	Califique siendo 1 poco y 5 mucho, qué tanto confía en: La defensoría del Pueblo
I_48_3	Califique siendo 1 poco y 5 mucho, qué tanto confía en: Los funcionarios del municipio (servidores públicos)
I_48_2	Califique siendo 1 poco y 5 mucho, qué tanto confía en: Concejo municipal
I_48_6	Califique siendo 1 poco y 5 mucho, qué tanto confía en: Los integrantes de la Juntas de Acción Comunal - JAC
I_48_8	Califique siendo 1 poco y 5 mucho, qué tanto confía en: Las organizaciones de la comuna o corregimiento
I_54	¿Está usted dispuesto a liderar procesos de la JAL, JAC?
I_52	¿Participa usted en la toma de decisiones para mejorar su comuna o corregimiento?

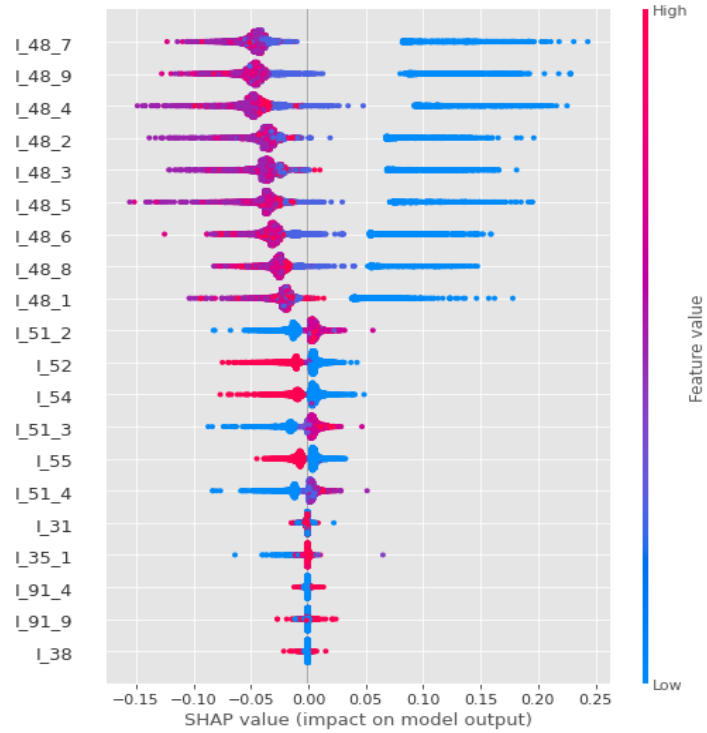
Gráfico 11. *SHAP value impact por clase. Prácticas - Individuos*

Clase 0	***Clase 1***


I_48_9
I_54
I_48_4
I_48_7
I_52
I_55
I_48_5
I_48_8
I_51_3
I_48_6
I_51_2
I_48_3
I_51_4
I_48_2
I_48_1
I_91_1
I_35_1
I_91_3
I_91_4
I_31
−0.15 −0.10 −0.05 0.00 0.05 0.10 0.15 0.20
SHAP value (impact on model output)
High
Feature value
Low


I_48_9
I_48_7
I_48_4
I_48_5
I_54
I_48_3
I_48_2
I_48_8
I_48_6
I_52
I_55
I_51_3
I_48_1
I_51_2
I_51_4
I_91_4
I_31
I_91_1
I_35_1
I_91_3
−0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low

Clase 2


I_48_7
I_48_9
I_48_4
I_48_2
I_48_3
I_48_5
I_48_6
I_48_8
I_48_1
I_51_2
I_52
I_54
I_51_3
I_55
I_51_4
I_31
I_35_1
I_91_4
I_91_9
I_38
−0.15 −0.10 −0.05 0.00 0.05 0.10 0.15 0.20 0.25
SHAP value (impact on model output)
High
Feature value
Low

Para la dimensión Prácticas prevalecen las variables del grupo I_48 (Califique siendo 1 poco y 5 mucho, qué tanto confía en:), a tal punto que las primeras 8 variables de las 10 más importantes hacen parte de este grupo (ver tabla 16). Por tanto, la conclusión es que el nivel de confianza de los individuos encuestados frente a diversos actores de la participación en la ciudad es fundamental para evaluar la calidad de la participación ciudadana en Medellín, o al menos tiene un alto impacto.

Además, resulta interesante que la primer y tercer variable más importantes (I_48_9 e I_48_7) están asociadas a actores locales (Consejo Comunal o Corregimental e integrantes de Juntas Administradoras Locales - JAL, respectivamente), dando lugar a un nivel de relevancia que se configura desde el círculo más cercano del participante. En esencia, le da preponderancia al escenario participativo local, ya sea a nivel barrial o comunal, en detrimento del nivel de ciudad, que puede generar menos impacto en el individuo. Esto último no implica que los actores a nivel municipal carezcan de preponderancia; sin embargo, resalta que en el segundo y cuarto lugar aparecen actores de ciudad que están directamente relacionados a la defensa de derechos, como lo son la Personería Municipal (I_48_4) y la Defensoría del Pueblo (I_48_5).

Por último, completan ranking dos variables: La primera (I_54) asociada a la intención de asumir liderazgos dentro del escenario participativo a nivel comunal y zonal; y la segunda (I_52) asociada a la percepción del participante frente a su participación e incidencia en la toma de decisiones a nivel comunal o corregimental. Esta última variable es de suma importancia dado que sirve de puente con la dimensión Efectos y guarda en su seno la esencia del ejercicio participativo, que se manifiesta en la incidencia de la participación individual y colectiva en la planificación y ejecución de políticas públicas que afectan en alguna forma la vida en comunidad.

c. Efectos

Para esta dimensión se repite el proceso. Se construye un pequeño modelo donde las variables independientes (X) constituyen las variables específicas que hacen parte de esta dimensión, y la variable dependiente (Y) constituye el indicador de la dimensión Efectos. Es necesario tener en cuenta que esta dimensión es la de menor cantidad de variables que la componen, por lo que la calidad del modelo se va a ver impactada en algún nivel; en este sentido, al sólo haber 5 variables para esta dimensión, el valor está en determinar el orden de importancia de ellas. Estos fueron sus resultados:

Gráfico 12. *Matriz de confusión. Efectos - Individuos*

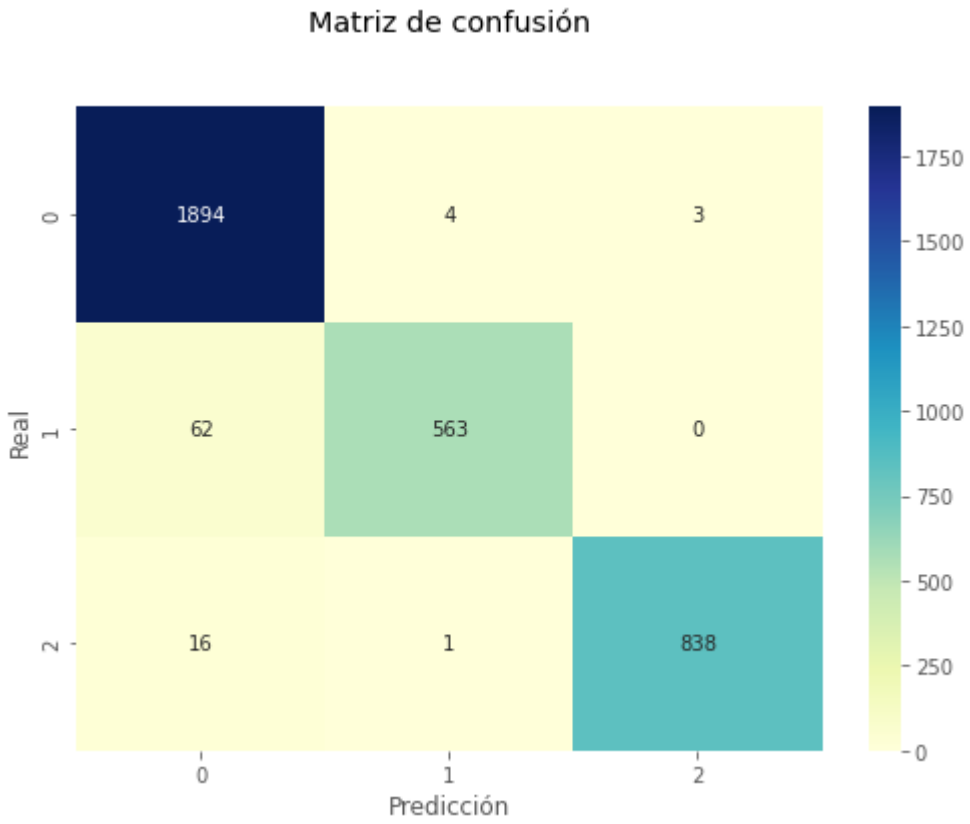


Tabla 17. *Métricas de evaluación. Efectos - Individuos*

	Precision	Recall	F1-score	Cantidad
0	0.96	1.00	0.98	1901
1	0.99	0.90	0.94	625
2	1.00	0.98	0.99	855
Accuracy			0.97	3381
Macro avg	0.98	0.96	0.97	3381

Weighted avg	0.98	0.97	0.97	3381
Hiperparámetros: class_weight = 'balanced', max_features = 'auto', n_estimators = 100				

Comparado con los mini modelos de las otras dos dimensiones, este modelo presenta un mayor índice de desbalance entre las clases, donde la Clase 0 es significativamente mayor que las otras dos; no obstante, esto no afecta la predicción, pues sus métricas de Recall y F1 obtuvieron un puntaje de 0.96 y 0.97 respectivamente.

Gráfico 13. *SHAP. Efectos - Individuos*

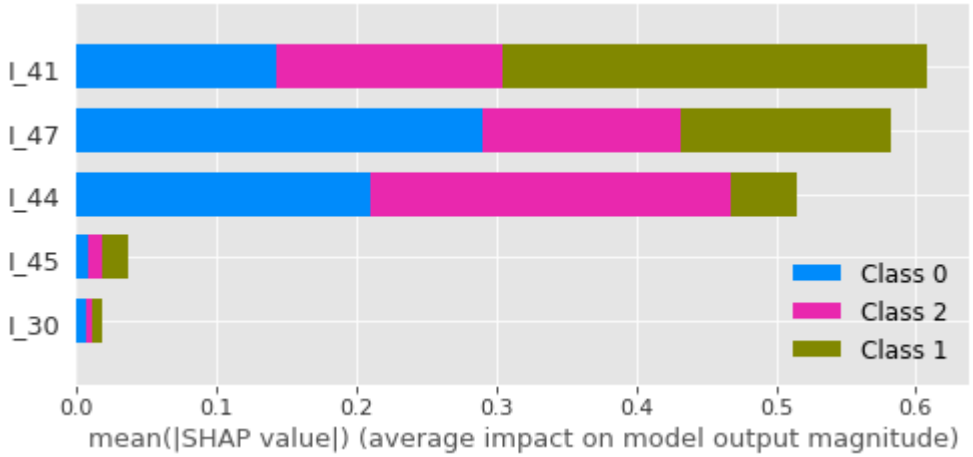


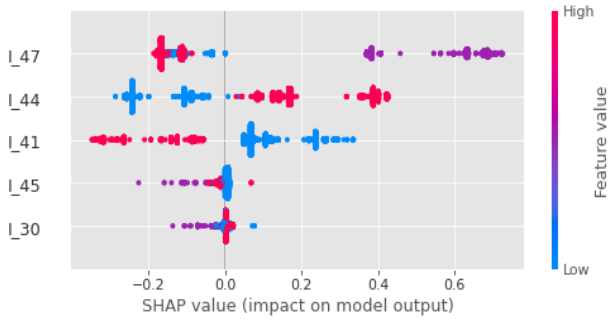
Tabla 18. Variables más importantes. Efectos - Individuos

I_41	Durante los últimos 24 meses, ¿Usted ha participado en iniciativas, propuestas o proyectos que beneficien su comuna o corregimiento?
I_47	Durante los últimos 24 meses, como resultado de las participación o movilizaciones ciudadana, considera que las condiciones de vida de su comunidad han: Mejorado - Empeorado - Permanecido Iguales.
I_44	Durante los últimos 24 meses y a nivel de ciudad, ¿La participación o movilización ciudadana ha logrado algún impacto en términos de acciones o políticas adoptadas por la administración municipal?

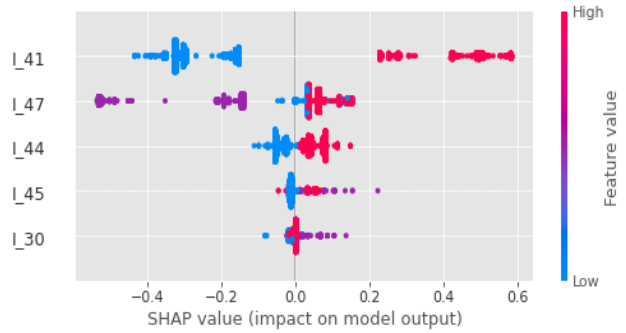
I_45	¿Ha hecho parte de la formulación de propuestas, proyectos, que han incidido sobre las acciones del gobierno local?
I_30	¿Ha observado en su comuna o corregimiento que los grupos o comunidades que participan tienen mayores beneficios que aquellas que no lo hacen?

Gráfico 14. *SHAP value impact por clase. Efectos - Individuos*

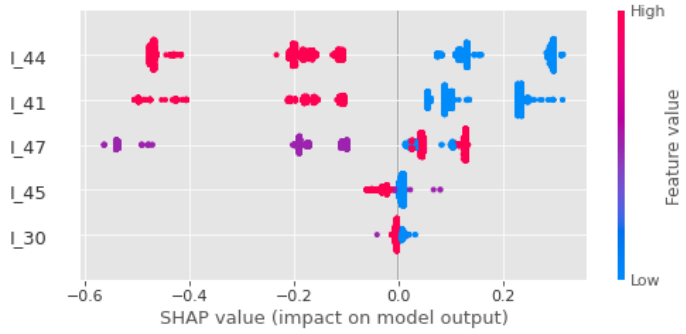
Clase 0: Baja



Clase 1: Media



Clase 2: Alta



Para esta dimensión, dada la baja cantidad de variables, resalta que I_41 sea la más importante, dado que también es la única de Efectos que aparece en el ranking general de Individuos que participan (ver Gráfico 4). Ello cobra mayor relevancia al reconocer el contenido de la variable, pues esta hace referencia a la participación efectiva del individuo

en iniciativas, propuestas y proyectos que beneficien a la comunidad a nivel comunal o corregimental. Su importancia radica en que la incidencia del participante en la toma de decisiones políticas colectivas es la esencia del ejercicio participativo. Además, resulta interesante la discrepancia de las variables en términos de impacto, pues si se observa el Gráfico 13, las últimas dos variables tienen muchísimo menos importancia que las tres restantes. Estás variables, además de la I_41 que ya fue reseñada, son la I_47 (que hace referencia a la percepción en la transformación de las condiciones de vida en la comunidad, ya sea de forma positiva, negativa o neutral) y la I_44 (que hace referencia al impacto que se percibe ha generado la participación en acciones y políticas adoptadas por la administración municipal). En definitiva, ambas variables giran en torno a la misma idea: el impacto que estas variables tienen sobre el puntaje de la dimensión es directamente proporcional al impacto que los individuos perciben de sus acciones frente a su ejercicio participativo a nivel local y a nivel de ciudad.

III. Organizaciones sociales y colectivos

Este conjunto de datos supone un gran reto para el análisis pues para la edición del año 2017, más de la mitad de los registros fueron eliminados debido a que representaban datos nulos y al final sólo una porción del total pudo ser sometido al proceso de modelado. Ello afecta enormemente tanto la calidad de los datos como el análisis sobre ellos. El buen trabajo realizado en las ediciones 2019 y 2021 mitiga el impacto y mejora la calidad del ejercicio. En comparación al conjunto de datos de Individuos que participan (ver Gráfico 3), la clasificación por medio del algoritmo K-modas dio como resultado un mayor desbalance en la distribución entre los clústers (ver Gráfico 15); sin embargo, este desbalance no afectó el desempeño del modelo, pues incluso sus métricas de desempeño son mejores que las del modelo de los datos de Individuos (ver Tabla 19).

Gráfico 15. *Distribución por Clase - Organizaciones sociales*

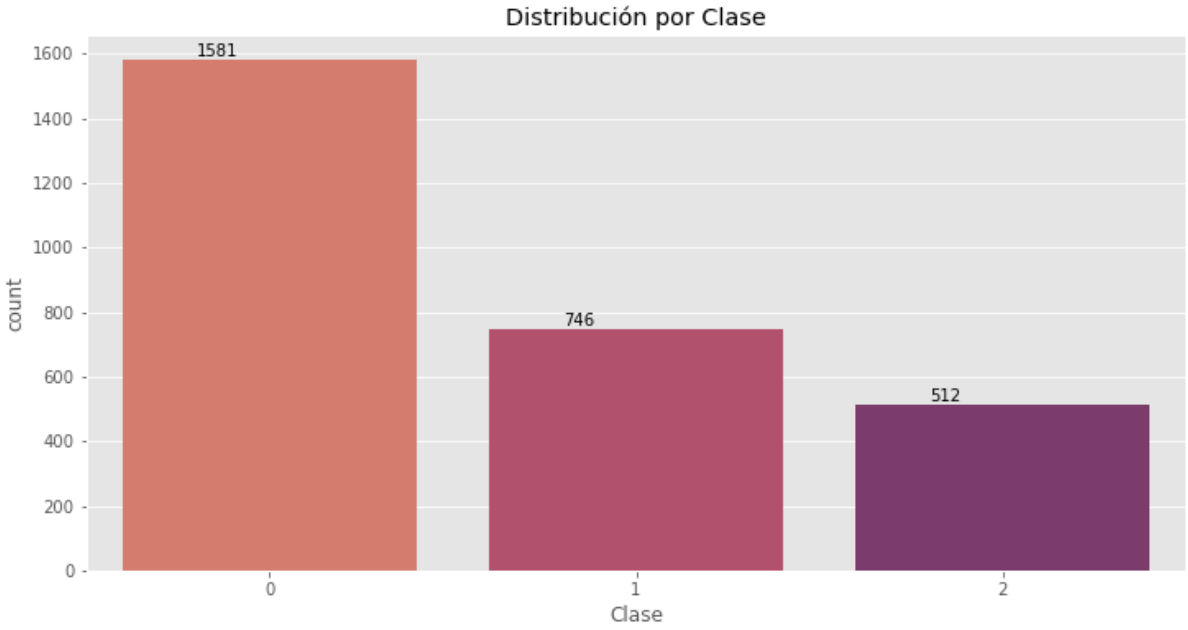


Gráfico 16. *Matriz de confusión – Organizaciones sociales*

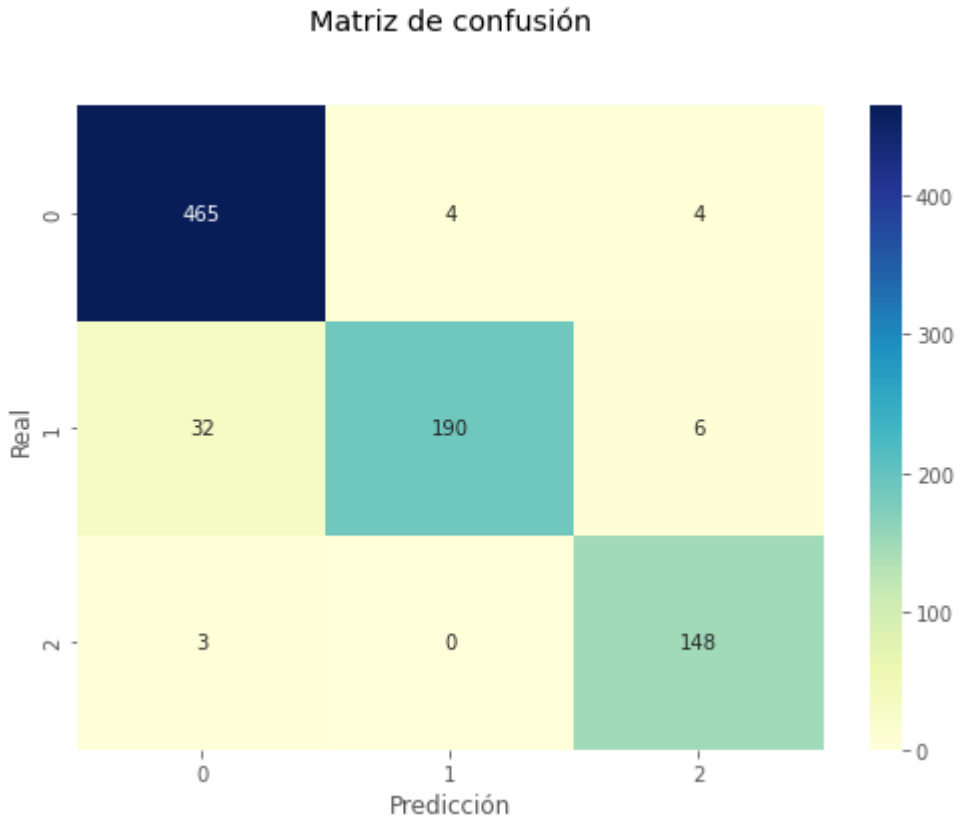


Tabla 19. Métricas de evaluación - Organizaciones sociales

	Precision	Recall	F1-score	Cantidad
0	0.93	0.98	0.96	473
1	0.98	0.83	0.90	228
2	0.94	0.98	0.96	151
Accuracy			0.94	852
Macro avg	0.95	0.93	0.94	852
Weighted avg	0.94	0.94	0.94	852
Hiperparámetros: class_weight = 'balanced', max_features = 'sqrt', n_estimators = 700.				

Por tanto, a continuación, se replica el proceso realizado con el conjunto de datos de Individuos, en donde se calculan las variables más importantes a través de Feature Importance, Permutation Importance y SHAP.

Gráfico 17. *SHAP - Organizaciones sociales*

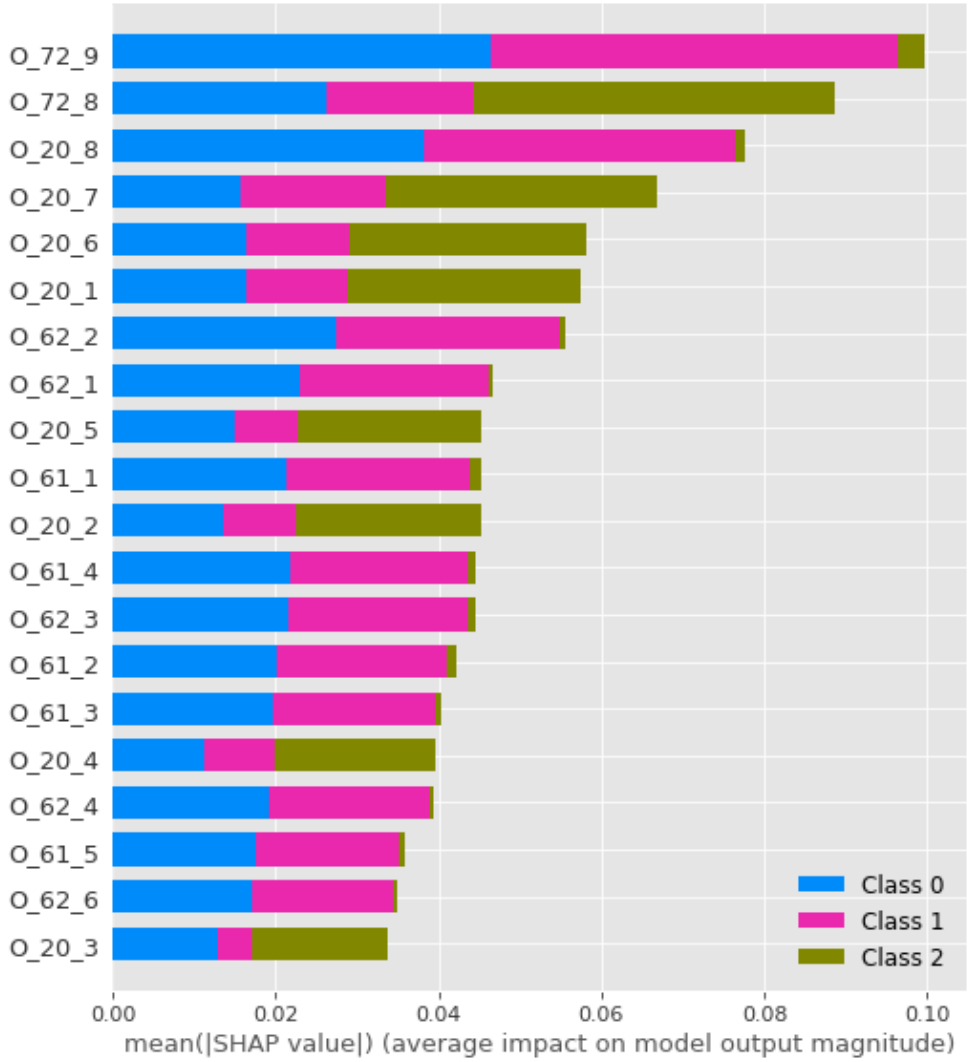


Tabla 20. Variables más importantes - *Organizaciones sociales*

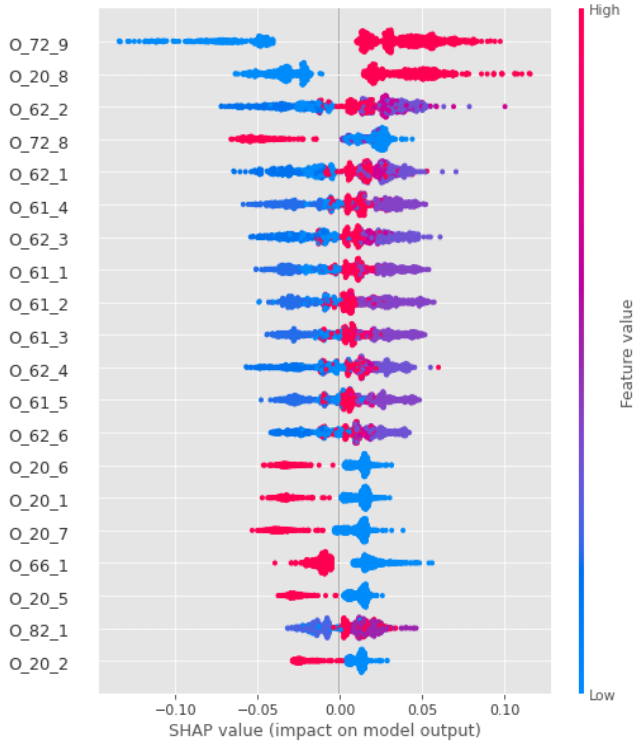
Código	**Descripción de variable**	**Dimensión**
O_72_9	Cuando su organización o colectivo social busca información sobre los programas y proyectos de la Alcaldía de Medellín, puede observar: j. no busca	Condiciones Territoriales
O_72_8	Cuando su organización o colectivo social busca información sobre los programas y proyectos de la Alcaldía de Medellín, puede observar: i. No encuentra	Condiciones Territoriales

O_20_8	Su organización y/o colectivo social ha recibido alguna intimidación o han experimentado situaciones que impliquen o le obliguen a: j. Otro ¿Cuál?	Condiciones Territoriales
O_20_7	Su organización y/o colectivo social ha recibido alguna intimidación o han experimentado situaciones que impliquen o le obliguen a: g. Pagar extorsión o vacuna	Condiciones Territoriales
O_20_6	Su organización y/o colectivo social ha recibido alguna intimidación o han experimentado situaciones que impliquen o le obliguen a: f. Expulsión de algún miembro	Condiciones Territoriales
O_20_1	Su organización y/o colectivo social ha recibido alguna intimidación o han experimentado situaciones que impliquen o le obliguen a: a. Cerrar el grupo	Condiciones Territoriales
O_62_2	Califique de 1 a 5, siendo 1 totalmente inequitativo y 5 totalmente equitativo. Cree usted que las acciones realizadas por las organizaciones y/o los colectivos de la comuna o corregimiento, en términos de recursos, obras y garantía de derechos, son equitativas entre: b. Veredas del corregimiento vs centralidad del corregimiento	Efectos
O_62_1	Califique de 1 a 5, siendo 1 totalmente inequitativo y 5 totalmente equitativo. Cree usted que las acciones realizadas por las organizaciones y/o los colectivos de la comuna o corregimiento, en términos de recursos, obras y garantía de derechos, son equitativas entre: a. Barrios de la comuna o corregimiento	Efectos
O_20_5	Su organización y/o colectivo social ha recibido alguna intimidación o han experimentado situaciones que impliquen o le obliguen a: e. Irse del barrio, de la comuna o corregimiento	Condiciones Territoriales

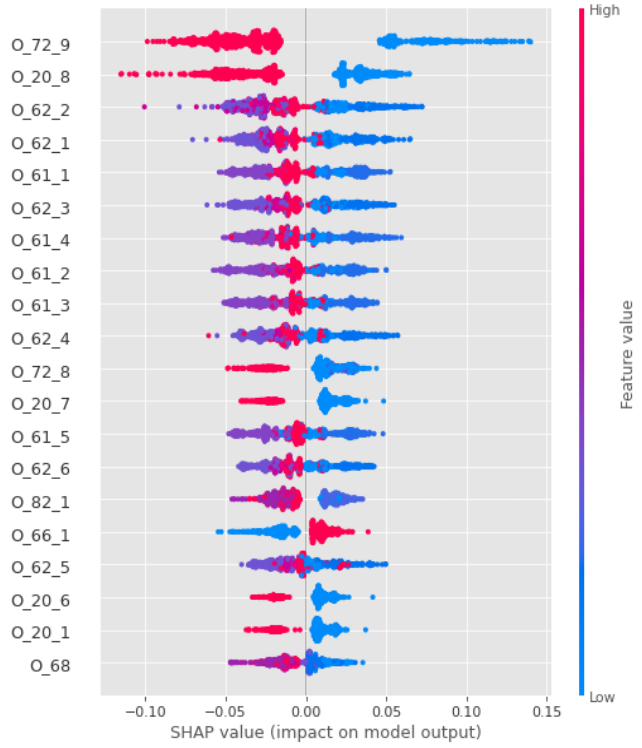
O_61_1	Califique de 1 a 5, siendo 1 totalmente inequitativo y 5 totalmente equitativo. Cree usted que las acciones realizadas por la Administración Municipal, en términos de recursos, obras y garantía de derechos, son equitativas entre: a. Comunas o corregimientos	Efectos

Gráfico 18. *SHAP value impact por clase. Organizaciones sociales*

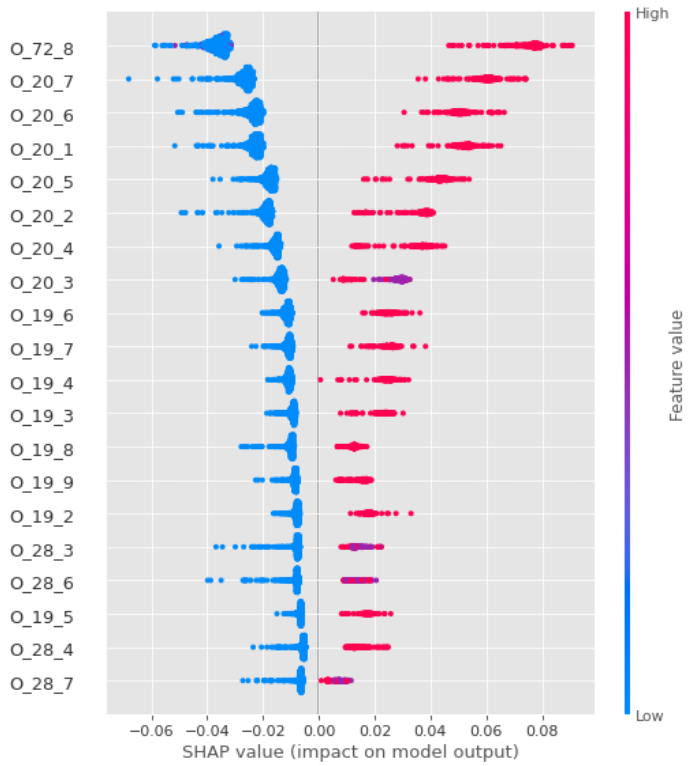
Clase 0



Clase 1



Clase 2



El primer elemento que se destaca al observar las 10 variables más importantes del conjunto de datos de Organizaciones sociales, es la ausencia de aquellas que hacen parte de la dimensión Prácticas y actores dentro del ranking. Al igual que en el conjunto de datos de Individuos que participan, existe una prevalencia de la dimensión Condiciones Territoriales con 7 de 10 variables, siendo los 3 restantes pertenecientes a la dimensión Efectos.

Las dos variables más importantes, **O_72_9** y **O_72_8**, están asociadas al acceso a la información sobre programas y proyectos de la Alcaldía de Medellín. Resalta que ambas variables están asociadas a un escenario negativo, lo que evidencia la importancia que tiene para las organizaciones sociales de la ciudad un acceso transparente, fácil y efectivo a información de carácter pública. En tanto esta condición no se cumpla, la calidad de la participación se va a ver profundamente afectada.

Por otra parte, cobran protagonismo las variables del grupo **O_20**, que están asociadas a situaciones intimidatorias y sus consecuencias en el ejercicio participativo. Su relevancia no sólo se refleja en la magnitud de esta situación, sino además en el hecho de que 5 de las variables de este ranking hacen parte de este grupo. Es preocupante que las

consecuencias que se destacan en el listado estén relacionadas al pago de extorsiones o vacunas, expulsión de miembros, cierre del grupo o incluso exilio y desplazamiento forzado intraurbano. Sin duda, dada la historia reciente de la ciudad en términos de seguridad y orden público, y el peligro que históricamente ha conllevado el ejercicio de la participación política en Colombia, este es un tema a tratar con especial atención. El lineamiento es claro: en tanto no existan garantías básicas de seguridad y la participación no implique una exposición de la integridad de las organizaciones sociales y los individuos que la componen, la calidad se va a ver profundamente afectada.

Por último, destacan las tres variables de la dimensión Efectos, específicamente la **O_62_2**, **O_62_1** y **O_61_1**. Con diferencias de escala, ya sea a nivel municipal por parte de la Alcaldía de Medellín, a nivel comunal o barrial y a nivel de centralidad y periferia en los corregimientos, estas tres variables evalúan la equidad de las acciones y proyectos llevados a cabo en el territorio en cuestión. En consecuencia, estas variables están asociadas al fenómeno de corporativización de lo público y la ejecución de recursos y proyectos a partir de criterios definidos bajo una lógica clientelar. Es así como la inequidad producida por estos eventos afecta de manera significativa la calidad de la participación ciudadana en la ciudad.

7. Análisis de Frecuencia

Tras destacar el listado de variables más importantes, identificando de este modo la influencia que estos tienen en el resultado final, el ejercicio queda inconcluso. Como se recordará, cada una de las variables corresponde a una pregunta del cuestionario que se le hizo en su momento a un Individuo que participa o a una Organización social o colectivo, por tanto, si lo que se quiere es identificar el verdadero impacto de dicha variable, es necesario recurrir a un análisis de frecuencia que nos dé una respuesta más específica. Es a partir de esta revisión que se construyen las conclusiones y recomendaciones finales.

A continuación, se presentan las gráficas con las principales variables y su distribución de Frecuencia, la interpretación de las mismas será abordada en la sección siguiente junto con sus conclusiones. Estas gráficas pueden ser manipuladas de forma interactiva en el Anexo 15.

Gráfico 19. Distribución de frecuencia variable I_41

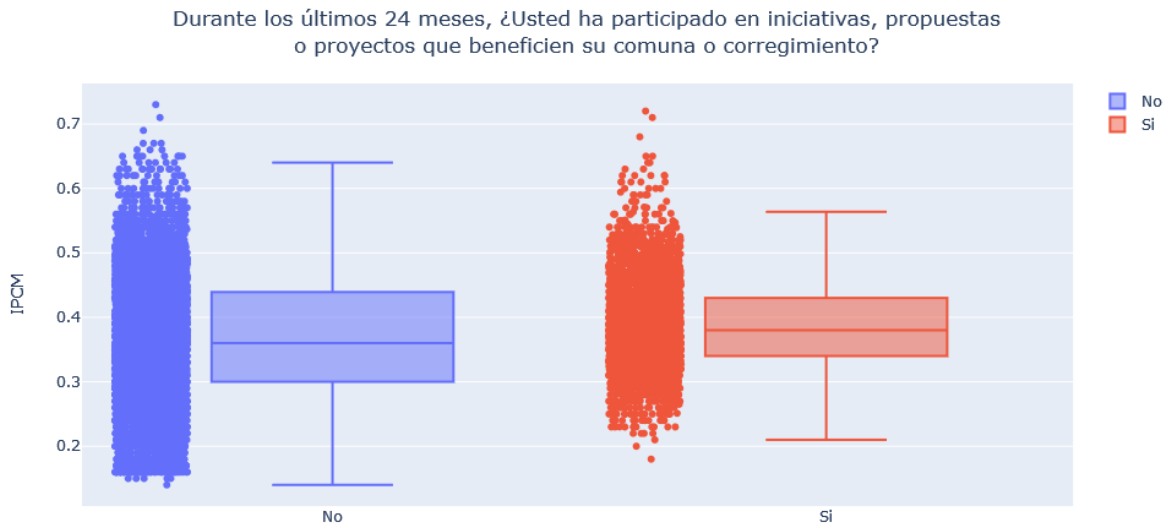


Gráfico 20. Distribución de frecuencia variable I_52

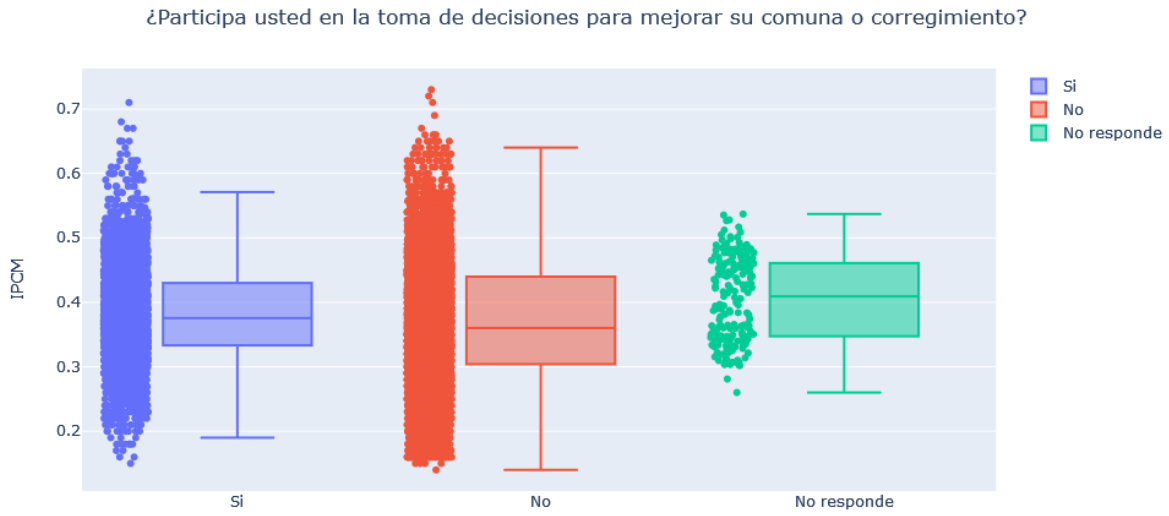


Gráfico 21. Distribución de frecuencia variable O_20_7

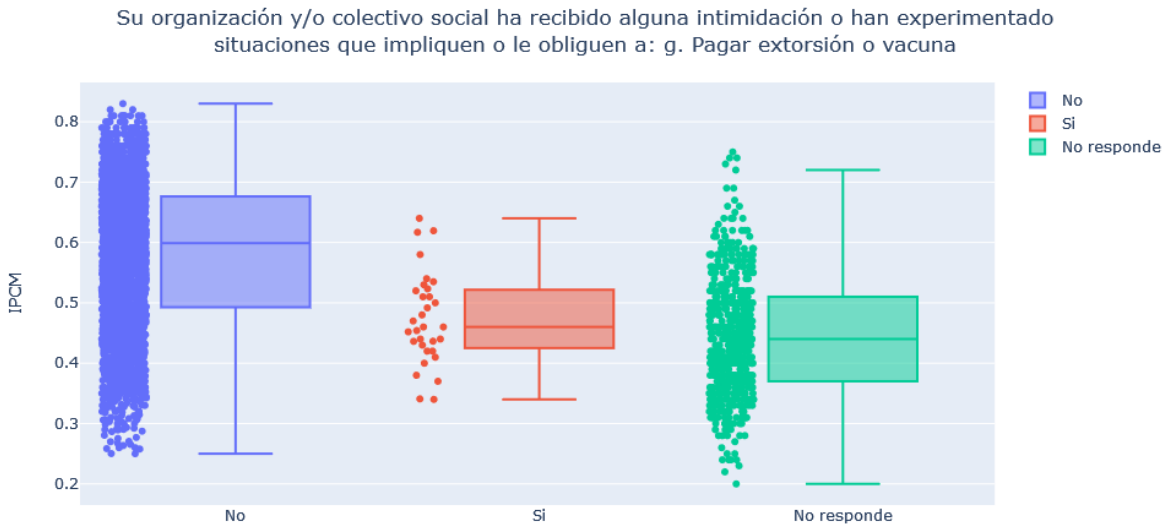


Gráfico 22. Distribución de frecuencia variable O_20_6

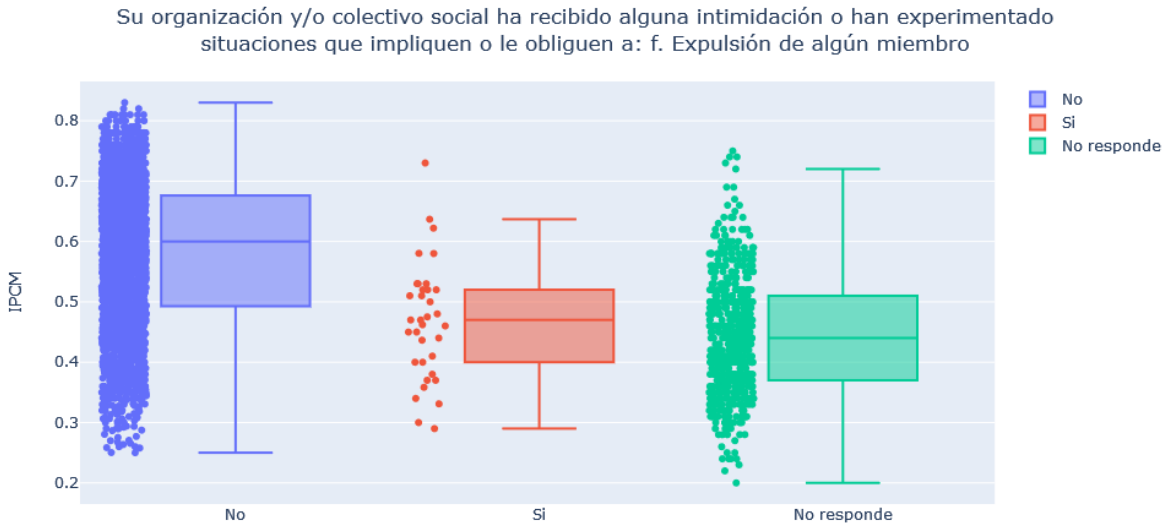


Gráfico 23. Distribución de frecuencia variable O_20_1

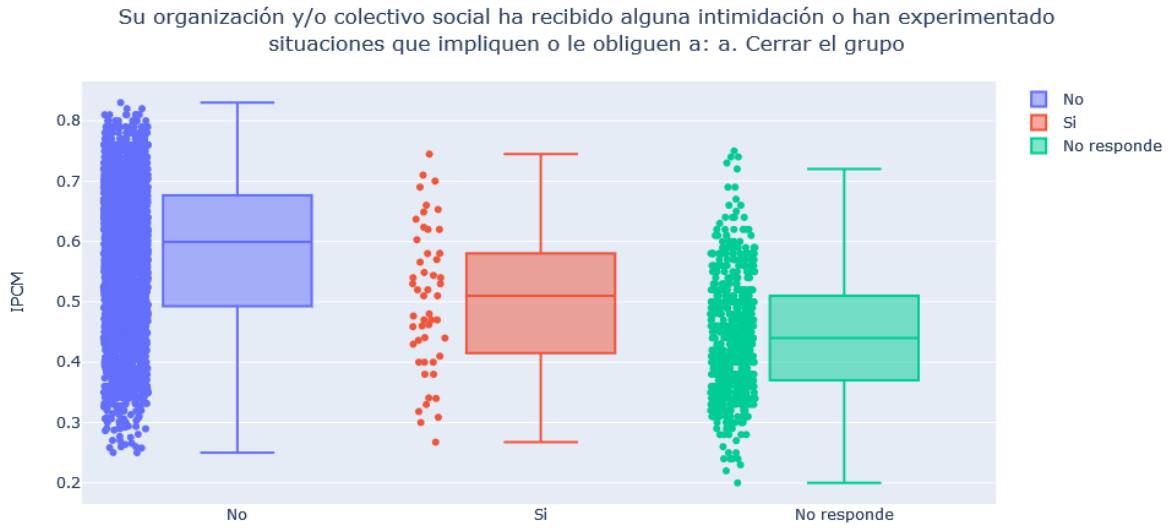
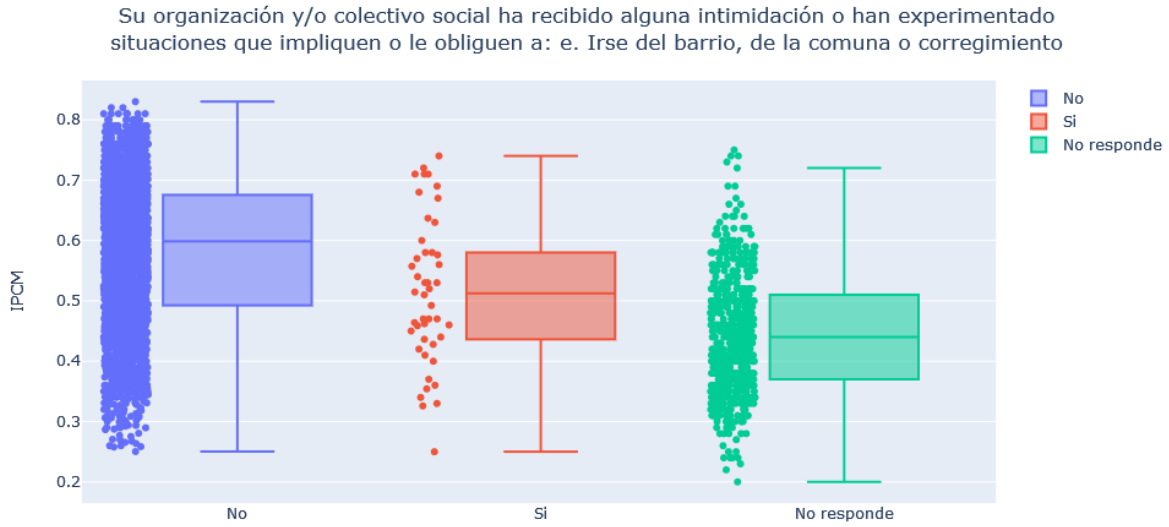


Gráfico 24. Distribución de frecuencia variable O_20_5



Conclusiones y recomendaciones

Conclusiones

Construidos los modelos y evaluados sus resultados, existe una respuesta satisfactoria a los objetivos planteados al inicio del proyecto. Así como en la sección anterior se descubrieron las variables que más influencia tienen en el puntaje del IPCM y en consecuencia, en la calidad de la participación ciudadana, el siguiente paso es evaluar el impacto de las variables en clave de sus opciones de respuesta. De este modo, la distribución de frecuencia nos dará respuestas más específicas frente a la forma como estas variables impactan sobre el puntaje final.

Como ya se vio en la sección anterior, para el conjunto de datos de individuos destacó la prevalencia de variables del grupo I_90, asociadas al conocimiento existente frente a mecanismos de control social (ver Tabla 12). De este modo, siendo evidente la influencia en el puntaje final, es recomendable revisar las estrategias de formación ciudadana llevadas a cabo por la Secretaría de Participación Ciudadana de Medellín, en específico en lo relativo al componente de control social y político, el conocimiento de estas herramientas y su uso.

De entre las variables también destaca la I_41 por dos motivos: por ser una de las 10 variables más importantes del conjunto de datos de individuos y por ser la más importante entre la dimensión Efectos. Como puede observarse en el Gráfico 19, aquellos que responden positivamente a la pregunta ¿Usted ha participado en iniciativas, propuestas o proyectos que beneficien su comuna o corregimiento? con tendencia suelen obtener un mejor puntaje del IPCM. En definitiva, es recomendable que la Secretaría de Participación Ciudadana enfoque gran parte de sus esfuerzos y trabajo en propiciar procesos que fortalezcan e impulsen la incidencia de la participación y su agencia; en tanto exista la percepción de que los actores políticos y sociales de la ciudad no inciden en los asuntos comunales, barriales y municipales, tanto el IPCM como la participación en sí se van a ver

profundamente afectadas. La variable I_52, asociada a la participación del individuo en la toma de decisiones de su comuna o corregimiento, refuerza esta reflexión.

En lo referente a los resultados del conjunto de datos de organizaciones, es realmente preocupante la prevalencia entre las variables más importantes de aquellas relativas al grupo O_20, dado que estas hacen referencia a situaciones intimidatorias que han afectado en algún grado el ejercicio participativo. Ello supone un desafío enorme que sobrepasa las funciones y capacidad de la Secretaría de Participación Ciudadana y requiere, por tanto, de acciones conjuntas que involucren la Secretaría de Seguridad y Convivencia y la Secretaría de Inclusión Social, Familia y Derechos Humanos, así como entes de carácter nacional como la Fiscalía General de la Nación y la Policía Nacional. En definitiva, teniendo en cuenta los antecedentes de violencia política, que involucra persecución, acoso y vulneración de derechos humanos de actores políticos y líderes sociales, es de carácter urgente. La secuencia de Gráficos que va desde el 21 hasta el 24 corrobora el fuerte impacto que tiene la respuesta afirmativa ante algún tipo de intimidación en el puntaje final del IPCM; destaca además que para la opción de respuesta “No responde”, la tendencia es que el puntaje sea aún más bajo, ello podría indicar un silencio deliberado por parte del representante encuestado, dada la intensidad de la intimidación y el miedo producto de esta.

Consideraciones finales

Desde su formulación y primera publicación en el año 2017, el Índice de Participación Ciudadana de Medellín ha alcanzado un grado de madurez que le ha permitido consolidarse como un proyecto de ciudad, que trasciende la agenda de un periodo de gobierno específico y que busca una mayor apropiación por parte de la sociedad civil. En consecuencia, aún con sus dificultades técnicas y socio-políticas inevitables para una ciudad en constante evolución y disputa, el IPCM ha servido como herramienta diagnóstica para la participación en la ciudad y como insumo fundamental para la elaboración de políticas públicas informadas por parte de la Administración Municipal.

La arquitectura interna del proceso de elaboración *del Índice de Participación Ciudadana de Medellín*, está compuesta por tres componentes: 1. Medición, 2. Analítico y 3. Educomunicativo. En este sentido, el componente Analítico -en el que se asume el ejercicio realizado en el presente informe- cumple un papel de doble entrada: En primer lugar, al

asumir los resultados de la medición como insumo esencial para la realización de análisis y construcción de productos de interpretación socio-política de los resultados del Índice; y en segundo lugar, al proponer ejercicios y discusiones que dinamizan y complejizan el ejercicio de la participación a nivel de ciudad. Su trabajo representa una síntesis del proceso general, dado que en su seno se articulan los demás procesos con el objetivo final de generar conocimiento sobre la participación ciudadana, en concordancia con el *Enfoque Territorial de Participación Ciudadana*.

De este modo, la razón de ser y eje misional de este componente es la producción de información cualificada a partir de los datos, ya sean estos de tipo cuantitativo o cualitativo, y que a su vez sirvan de entrada para el *Sistema de Información y Gestión del Conocimiento para la Participación Ciudadana - SIGC-PC*. En definitiva, después del recorrido por los distintos conjuntos de datos, sus modelos y resultados, estamos ante la generación de conocimiento nuevo producto del aprovechamiento de los datos generados por las distintas ediciones del IPCM que brindarán una base sólida e informada para la elaboración de programas, proyectos y políticas públicas orientadas a mejorar el ejercicio de la participación en Medellín. Por otro lado, aunque mejorable, es satisfactorio reconocer en este trabajo la introducción de metodologías que comúnmente son ajenas a las ciencias sociales, y que suponen un excelente complemento a las metodologías clásicas.

En suma, estamos ante un primer acercamiento a un ejercicio inédito a nivel de ciudad, que potencia el trabajo ya realizado por el IPCM y se pone a disposición para mediciones futuras. Sin duda, para el año 2023 en el que se realice la cuarta edición del Índice, se contará con una herramienta funcional y efectiva para el procesamiento de los datos recolectados, además de la experiencia previa que permite la elaboración de modelos más refinados. La interacción de la población civil y los distintos actores de la ciudad con los resultados obtenidos dará lugar a nuevas preguntas y necesidades a resolver. En cualquiera de los casos, esta herramienta nos da la posibilidad de generar programas, proyectos y políticas públicas informadas a partir de los datos, y esa es su principal ganancia.

Lista de Anexos[7]

1. Anexo1_DescripcionIndividuos2017
2. Anexo2_DescripcionOrganizaciones2017
3. Anexo3_DescripcionIndividuos2019
4. Anexo4_DescripcionOrganizaciones2019
5. Anexo5_DescripcionIndividuos2021
6. Anexo6_DescripcionOrganizaciones2021
7. Anexo7_BDIndividuosTotal
8. Anexo8_BDIndividuosTotal
9. Anexo9_BDOrganizacionesTotal
10. Anexo10_BDOrganizacionesTotal
11. Anexo11_ETL
12. Anexo12_Modelos_Clasificacion
13. Anexo13_Clas_Feature_Importance
14. Anexo14_Dimension_Feature_Importance
15. Anexo15_Distribucion_Frecuencia

[7] Los Anexos serán entregados junto con el documento. Además, se podrá tener acceso a través del siguiente enlace: https://github.com/jlopezbu/IPCM2021/tree/main/Anexos

Referencias

Boudjelida, A. & Mellouli, S. (2016). A Multidimensional Analysis Approach For Electronic Citizens Participation. In Proceedings of the 17th International Digital Government Research Conference on Digital Government Research (dg.o '16). Association for Computing Machinery, New York, NY, USA, 49–57. https://doi.org/10.1145/2912160.2912195

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S.-M., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., & Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. Big Data & Society. https://doi.org/10.1177/2053951717726554

Deng, N., Tian, Y., & Zhang, C. (2012). Support vector machines: optimization based theory, algorithms, and extensions. CRC press.

Diez, D. M., Barr, C. D., & Çetinkaya-Rundel, M. (2019). OpenIntro Statistics: Fourth Edition. OpenIntro.

García, J., Molina, J. M., Berlanga, A., Patricio, M. A., Bustamante, A. L., & Padilla, W. R. (2018). Ciencia de datos. Técnicas analíticas y aprendizaje estadístico. Alfaomega Colombiana S.A, Publicaciones Altaria.

González, F. (2019). Big data, algoritmos y política: las ciencias sociales en la era de las redes digitales. Cinta de moebio, (65), 267-280. https://dx.doi.org/10.4067/s0717-554x2019000200267

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3), 283–304. https://doi.org/10.1023/A:1009769707641

Islam Sarker, N., Khatun, M., Alam, M. & Islam, S. (2020). Big Data Driven Smart City: Way to Smart City Governance. 2020 International Conference on Computing and Information Technology (ICCIT-1441), 2020, 1-8. doi: 10.1109/ICCIT-144147971.2020.9213795

Ju, J., Liu, L., & Feng, Y. (2018). Citizen-centered big data analysis-driven governance intelligence framework for smart cities. Telecommunications Policy, 42(10), 881–896. https://doi.org/https://doi.org/10.1016/j.telpol.2018.01.003

Kononenko, I., & Kukar, M. (2007). Machine Learning Basics. Machine Learning and Data Mining, 59–105. doi:10.1533/9780857099440.59

Kowalczyk, A. (2017). Support Vector Machines Succintly. Syncfusion.

Lemus-Delgado, D. & Pérez Navarro, R. (2020). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. Colombia Internacional, (102), 41-62. https://doi.org/10.7440/colombiaint102.2020.03

Lundberg, S. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. ArXiv. https://doi.org/10.48550/arXiv.1705.07874

Mayorga, F. & Córdova, E., (2007), “Gobernabilidad y Gobernanza en América latina”, Working Paper NCCR Norte-Sur IP8, Ginebra. No publicado. http://www.institut-gouvernance.org/docs/ficha-gobernabilida.pdf

Meijer, A., & Bolívar, M. P. R. (2016). Governing the smart city: a review of the literature on smart urban governance. International Review of Administrative Sciences, 82(2), 392–408. https://doi.org/10.1177/0020852314564308

Meneses Rocha, M. (2018). Grandes datos, grandes desafíos para las ciencias sociales. Revista mexicana de sociología, 80(2), 415-444. https://doi.org/10.22201/iis.01882503p.2018.2.57723

Müller, A.C., & Guido, S. (2016). Introduction to Machine Learning with Python. A guide for Data Scientist. O’Reilly.

Pando, V. & San Martín, R. (2004). Regresión logística multinomial. En: Cuadernos de la Sociedad Española de Ciencias Forestales, Nº. 18, 2004, págs. 323-327.

Saarela, M., Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. **3,** 272 (2021). https://doi.org/10.1007/s42452-021-04148-9.

Weber, M. (1947). The Theory of Social and Economic Organization. New York: Oxford University Press.

Williams, G. J. (2011). Data mining with Rattle and R: The art of excavating data for knowledge discovery. Springer.

Zhu, Z. & Zhang. M. (2020). K-Nearest Neighbors(KNN) Classification with Different Distance Metrics. Shanghai Jiao Tong University.