PONTIFICIA UNIVERSIDAD JAVERIANA

Maestría en Analítica para la Inteligencia de Negocios

Facultad de Ingeniería

Tesis de Maestría

Daniel Alejandro Calambás Marín

Jaime Andrés Mendoza Mendoza

Angélica Pacheco Mejía

Leidy Andrea Ruiz Rodríguez

Bogotá, 2019

“ADACOP- Analytics”

Autores:

Daniel Alejandro Calambás Marín

Jaime Andrés Mendoza Mendoza

Angélica Pacheco Mejía

Leidy Andrea Ruiz Rodríguez

Bogotá

2019

CONTENIDO

LISTA DE TABLAS

LISTA DE ILUSTRACIONES

1. Introducción

Alianza CAOBA es el centro de Excelencia e Investigación en Big Data y Data Analytics, creado en el año 2015 como una iniciativa de Colciencias y el Ministerio de las Tecnologías de la Información y las Telecomunicaciones (MinTIC) y conglomera empresas del sector público, privado y la academia. CAOBA tiene como misión la generación de nuevo conocimiento relacionado con el Big Data y Data Analytics por medio de la investigación y desarrollo de soluciones tecnológicas. Al ser patrocinado por el gobierno nacional, CAOBA está comprometido en apoyar las iniciativas tecnológicas estatales y tiene definido un portafolio de proyectos estratégicos alineados a las necesidades del estado y la industria.

El MinTIC tiene definidos cuatro pilares, sobre los que se soporta la política "El futuro digital es de todos", con lo que se busca cerrar la brecha digital en Colombia. Los cuatro pilares son: "Entorno TIC para el desarrollo digital", con el que se busca modernizar el sector TIC; "Ciudadanos y hogares empoderados del entorno digital", cuyo objetivo es empoderar a las personas para que interactúen más de los servicios TIC; "Inclusión social digital", centrado en lograr una cobertura total de internet a lo largo del territorio nacional; "Transformación digital sectorial y territorial", para que los sectores público y privado exploten al máximo las tecnologías en sus procesos. ('El futuro digital es de todos': la nueva política TIC - Ministerio de Tecnologías de la Información y las Comunicaciones s. f.)

Teniendo en cuenta lo mencionado anteriormente, Alianza CAOBA pretende desarrollar el proyecto aquí presentado, de acuerdo con el pilar "Ciudadanos y hogares empoderados del entorno digital" con el fin de hacer uso de la información publicada en el portal de datos abiertos. Asimismo, desea apoyar la toma de decisiones de los alcaldes y gobernantes de las diferentes regiones del país, mediante la creación de una herramienta por medio de la cual se ejecuten modelos analíticos que logren generar conocimiento a partir de datos abiertos, soportando los procesos de transparencia en la contratación pública estatal.

Cabe mencionar que por medio de herramientas como dashboards que buscan resumir e integrar diversas métricas de tal manera que se puedan monitorear y planear de manera efectiva las principales variables, respaldando así los procesos de transparencia que se mencionaban anteriormente (Pauwels et al. 2009). Una de las gráficas que se pueden implementar en estos tableros de control por ejemplo pueden ser drill-down que permiten ir de un nivel más general a uno más detallado de la información permitiendo realizar agregaciones y desagregaciones en el análisis realizado (Pauwels et al. 2009). En términos

generales, los dashboards son instrumentos útiles y productivos con los que los funcionarios del gobierno pueden visualizar toda la información de manera más sencilla y eficiente (Vila, Estevez, y Fillottrani 2018).

Además, se espera poder lograr un impacto social mediante el empoderamiento de los ciudadanos, para éstos puedan conocer de una forma clara, veraz y pertinente cómo se está gestionando el presupuesto de la nación en lo concerniente a temas de contratación pública.

2. Entendimiento del negocio

2.1. Objetivos de negocio

La Alianza CAOBA cuenta con cuatro ejes estratégicos en los que fundamenta sus productos y servicios: investigación aplicada, transferencia del conocimiento, consultoría tecnológica y apoyo al emprendimiento. El presente proyecto pretende apoyar dos de ellos: 1) la investigación aplicada, debido a que se pretende crear valor mediante una plataforma que incluye un componente analítico que brindará soporte y fiabilidad al ciudadano o parte interesada y un mayor entendimiento de los datos abiertos; y 2) la transferencia de conocimiento, ya que se establece un trabajo conjunto con la maestría en ingeniería de sistemas y el Observatorio Fiscal; identificando modelos que apoyen la transparencia en los procesos de contratación pública colombiana. Adicionalmente, el software y modelos analíticos a ser desarrollados en este proyecto serán parte del portafolio de productos y servicios que la Alianza CAOBA está desarrollando con el fin de ofrecerlos a la comunidad como parte de su estrategia de sostenibilidad.

2.2. Evaluación de la situación

A nivel internacional, de acuerdo con la definición dada por *Open Data Charter*, los datos abiertos son datos digitales que son dispuestos públicamente, con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar, ver tabla 1.

Principio	Descripción
Abiertos por defecto	Los datos gubernamentales son aquellos datos en poder de: gobiernos nacionales, regionales, locales y municipales, organismos gubernamentales internacionales, y otros tipos de instituciones del sector público, pero también son aquellos pertenecientes a

	entidades eternas que están relacionadas con programas y servicios del gobierno. En este principio, se reconoce el libre acceso.
Oportunos y exhaustivos	Para que el dato genere valor al gobierno, a la ciudadanía, organizaciones sociales y demás entidades, estos deben contar con calidad y ser precisos.
Accesibles y utilizables	Deben ser visibles y accesibles y no contar con barreras burocráticas.
Comparables e interoperables	Los datos deben ser fáciles de comparar dentro de los sectores, teniendo en cuenta los factores geográficos y del tiempo, escritos en un formato entendible para los usuarios
Para mejorar la gobernanza y la participación ciudadana	La información contenida en los datos refuerza la confianza en las entidades públicas y promueve la transparencia en las obligaciones de los gobiernos, demostrando que están alineados con el desarrollo de programas públicos que le beneficien al ciudadano.
Desarrollo incluyente e innovación	Cuanto más los gobiernos, la ciudadanía y los entes en generar usen los datos abiertos, mayores serán los beneficios sociales y económicos, por ejemplo, retos globales: el hambre, cambios climáticos, etc.

Tabla 1. Seis Principios de datos abiertos.

Anualmente, el *Open Data Barometer* (ODB) realiza un estudio entre 115 países para calificar la calidad y cantidad de la información disponible en los diferentes portales de datos abiertos. El último año arrojo a Canadá, Reino Unido y Australia como los países con mayor cantidad y calidad de datos en sus portales. Este estudio encontró que solo el 7% de los datos es totalmente abierto. También se evidenció que, los datos más completos y de mayor calidad se encuentran publicados en páginas de otras agencias gubernamentales, en un 61% de los casos. (Informe Global | Open Data Barometer s. f.)

En el contexto colombiano, de acuerdo con la guía de datos abiertos, se establecieron ocho principios que rigen los datos, contenidos en la tabla 2:

Principio	Descripción
Primario	Obtener los datos de origen con un alto nivel de detalle.
Accesibles	Estar disponible al usuario

Completos	Reflejar la totalidad del tema, garantizando la totalidad de la información suministrada.
Procesables por maquina	Contenidos en formatos que permitan el procesamiento
No propietarios	Disponibles en un formato en el que ninguna entidad tenga control.
Oportunos y actualizados	Garantizar su valor y mantener frecuencia de actualización
No discriminados	Estar disponibles para cualquier persona sin requerir registro de autenticación
Licenciados de forma abierta	Deben contar con términos de uso y licenciamiento abierto

Tabla 2. Principios en el contexto colombiano.

En Colombia, la regulación de los temas de transparencia y gobierno abierto se encuentra contenida principalmente en la ley 1712 de 2014: la Ley de transparencia y derecho de acceso a la información pública. El objeto de esta es regular el derecho de acceso a la información pública que tienen todas las personas, los procedimientos para el ejercicio y la garantía del derecho fundamental, así como las excepciones a la publicidad de la información pública. Según la ley 1712 (Literal J, artículo 6. Definiciones), los datos abiertos deben estar a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos. El acceso a la información solamente podrá ser restringido excepcionalmente. Las excepciones serán limitadas y proporcionales, deberán estar contempladas en la ley o en la Constitución y ser acordes con los principios de una sociedad democrática. Por su parte, la Resolución 3564 del 31 de diciembre de 2015 establece los lineamientos respecto de los estándares para publicación y divulgación de la información, accesibilidad en medios electrónicos para población en situación de discapacidad, formulario electrónico para la recepción de solicitudes de acceso a información pública, condiciones técnicas para la publicación de datos abiertos y condiciones de seguridad de los medios electrónicos, que se establecen en los artículos 2.1.1.2.1.1, 2.1.1.2.1.11, y el parágrafo 2o del artículo 2.1.1.3.1.1 del Decreto número 1081 de 2015.

Además de la legislación en materia de transparencia de la información pública, el gobierno colombiano implementó la Política de Gobierno Digital (Decreto 1078 de 2015 libro 2, parte 2, título 9. Cap. 1), mediante el cual se regula el uso y aprovechamiento de las TIC para mejorar la provisión de servicios digitales, el desarrollo de procesos internos eficientes, la toma de decisiones basadas en datos, el empoderamiento de los ciudadanos y el impulso en el desarrollo de territorios y ciudades inteligentes. Lo anterior, logrado a partir de la consolidación de un estado y ciudadanos competitivos, proactivos, e innovadores, que generan valor público en un entorno de confianza digital. La política de Gobierno Digital señala que todo proyecto que incorpore el uso de las TIC y genere información, debe contar con sistemas de información que permitan la generación de datos abiertos de manera automática para su publicación, uso y reutilización. Los datos abiertos no son contrarios a la protección de datos personales, pues si existe información asociada a datos personales que puede ser valiosa como dato abierto, es posible desarrollar acciones de anonimización para eliminar la información sensible que afecta a personas u organizaciones y cuya identidad debe protegerse legalmente.

Es importante resaltar, que actualmente existen algunas iniciativas para el fomento de el empoderamiento ciudadano y la transparencia. Tal es el caso del Portal de transparencia económica (http://www.pte.gov.co/). También, cabe mencionar al Observatorio de Transparencia y Anticorrupción (http://www.anticorrupcion.gov.co/Paginas/index.aspx), una herramienta para la medición y análisis del fenómeno de la corrupción; a partir de la interacción entre entidades, ciudadanos, y organizaciones públicas y privadas del orden nacional y territorial, para contribuir a elevar el nivel de transparencia en la gestión pública.

Con el objetivo de complementar el presente proyecto y de obtener una visión desde un organismo especializado en el manejo de temas de transparencia, se contactó al doctor Luis Carlos Reyes, director y cofundador del Observatorio Fiscal y profesor investigador de la Universidad Javeriana. El Observatorio Fiscal es un emprendimiento conjunto de la Facultad de Ciencias Económicas y Administrativas y de la Escuela Javeriana de Gobierno y Ética Pública, en el marco de la Planeación Universitaria Javeriana. La misión del Observatorio Fiscal es democratizar la información sobre las finanzas públicas en Colombia. Con lo anterior se busca que la ciudadanía pueda

combatir la corrupción y el desperdicio de los recursos públicos a través del voto informado y el activismo civil.

2.2.1. Antecedentes

A nivel mundial, se han realizado una gran cantidad de esfuerzos para promover la publicación y uso de datos abiertos gubernamentales. El primer país en implementar una estrategia de datos abiertos fue el gobierno de los Estados Unidos en el año 2009 basado en tres pilares: colaboración, partición y transparencia (S. S. Dawes y N. Helbig , 2010). De igual manera, la Unión Europea ha introducido el concepto de gobernanza inteligente, que tiene como eje central el uso de las grandes cantidades de datos que la administración pública genera en el desarrollo de sus actividades y en sus relaciones con la ciudadanía y las empresas, para apoyar los procesos de toma de decisiones estatales y aumentar la transparencia (C. Martínez, A, 2018) . También resalta que el rol de las agencias del gobierno no debe basarse únicamente en liberar los datos, sino en crear estrategias para atraer entidades externas y *stakeholders* que generen innovación sobre los datos abiertos (Z. Yang y A. Kankanhalli, 2013). Según una encuesta realizada por Deloitte: *Shaping the Future of Open Data An assessment of the open dat*a, en 2016, ¿Qué tipo de información del sector público o conjuntos de datos específicos le interesaría o cree que debería ponerse a disposición de los ciudadanos? (European Commission, 2014), el 38% de los encuestados desean conocer más sobre decisiones de gobierno y un 6% sobre contratación pública.

En la investigación sobre datos abiertos: *Budget and Procurement Analytics using Open Government Data in Thailand* (N. Surasvadi, C. Saiprasert, S. Thajchayapong, 2017), se realizó un proceso de análisis de compras y gastos presupuestales basado en datos abiertos gubernamentales. Allí, se utilizó un proceso propio de selección y procesamiento de la información, destacando la calidad y disponibilidad de los datos. Este estudio identificó patrones de gastos de presupuestos y adquisiciones mapeando cada departamento del país y utilizando reglas de asociación como técnica analítica principal.

2.2.2. Problemática

El Gobierno de Colombia ha reconocido que la compra y contratación pública es un asunto estratégico. Por este motivo, se creó la organización “Colombia Compra

Eficiente" en el año 2011. Esta entidad crea políticas unificadas para que sirvan de guía a los administradores de compras del estado y que permitan monitorear y evaluar el desempeño del sistema generando mayor transparencia *(*Colombia Compra Eficiente, 2015). Con la creación de esta organización, se adoptó el sistema de información SECOP (Sistema Electrónico para la Contratación Pública) promoviendo la contratación abierta y el uso de la información de los procesos de compra para fomentar la colaboración, la innovación y la transformación de la entrega de bienes, obras y servicios a los ciudadanos. Un proceso de contratación comprende una serie de pasos para que las entidades estatales puedan adquirir los bienes y servicios, mediante las licitaciones en las que participan los proveedores. Los pasos se resumen de la siguiente forma:

1. Una entidad estatal abre un proceso de contratación para adquirir bienes o servicio.
2. Los proveedores que deseen contratar con el estado se registran en la plataforma y son clasificados de acuerdo con el tipo o bien que ofrecen.
3. Una vez se publica el proceso de contratación, los proveedores reciben una invitación de acuerdo con el servicio/bien que ofrecen.
4. Tras la publicación del proceso, los proveedores manifiestan el interés por participar.
5. Finalmente, los proveedores que presentaron sus ofertas son tenidos en cuenta por la entidad quien selecciona uno de ellos.

De acuerdo con la información registrada en la página oficial del sistema de compra pública colombiana, en la ilustración 1 se observan los registros relacionados con la cantidad de procesos publicados en la plataforma de Colombia compra eficiente:

Ilustración 1 - Cantidad de procesos publicados SECOP II. Imagen de autoría propia

Se evidencia un crecimiento importante al comparar los años anteriores a 2017 contra el año 2018, se observa que la contratación estatal ha aumentado en el país, suponiendo que las normativas implementadas por el gobierno están siendo más rigurosas con respecto a los contratos.

Revisando en detalle, en la ilustración 2 se muestran algunos datos importantes para el año 2018, identificando que 1135 entidades crearon procesos de contratación, equivalentes a 158.006 contratos, que benefician a 334 ciudades o municipios colombianos y para los cuales participaron 1942 proveedores.

Ilustración 2 - Detalles de contratación año 2018. Imagen de autoría propia

En niveles de proporción de costos adjudicados, se evidencia siete tipos de contratos (ilustración 3). El tipo de contrato de servicios de aprovisionamiento tiene 49.77% del total del valor adjudicado para 2018, siendo éste el mayoritario, mientras que para servicios de consultoría hay 1.27%.

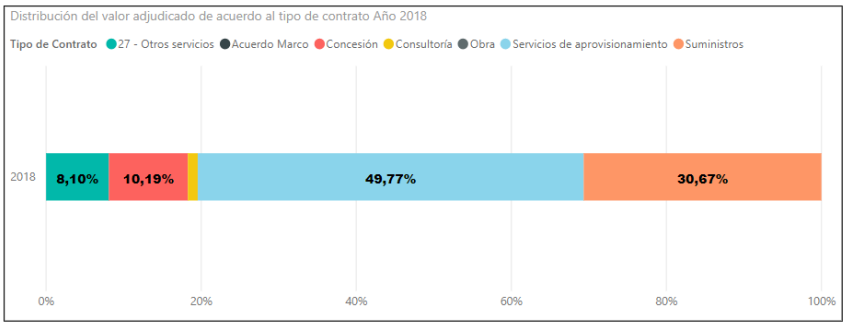

Ilustración 3 - Distribución valor total adjudicado año 2018. Imagen de autoría propia

A pesar de las iniciativas relacionadas con la contratación abierta y la publicación legible de los datos relacionados, aún hay una carencia de herramientas para facilitar las compras y promover la eficiencia a partir del insumo de la información.

2.2.2.1 Restricciones de acceso.

Los consumidores de la información expuesta dentro del portal de datos abiertos, como se mencionó anteriormente, no disponen de herramientas de consulta asertivas en cuanto a análisis de datos y visualizaciones efectivas que expongan los datos. El portal de datos abiertos utiliza Socrata (Interfaz de Socrata (2018). Plataforma de datos abiertos del gobierno colombiano) como herramienta de acceso y Plotly para visualización y análisis de datos en tiempo real; al realizar la extracción de la información para realizar gráficas descriptivas, en el portal existe una limitante de visualización en cuanto al

tamaño de la información, explorando la herramienta se obtiene que, para abrir el conjunto de datos en Plotly, se dispone máximo de 5MB.

Teniendo en cuenta el contexto expuesto previamente, el presente proyecto surge como una respuesta a la carencia de mecanismos suficientes y propone la aplicación de herramientas analíticas, enriquecidas con fuentes externas como redes sociales y portales de noticias, para que entidades públicas y ciudadanos generen capacidades de entendimiento y evaluación en las dinámicas de los procesos de contratación pública con el fin de soportar los procesos de transparencia y brindar una mayor cercanía hacia el ciudadano. Este proyecto se encuentra alineado con uno de los pilares del MinTIC que parte de los datos abiertos de entidades públicas como elementos clave para promover la transparencia, la competitividad, el desarrollo económico y la generación del impacto social en el contexto de apropiación de las TIC (Ministerio de Tecnologías de la Información y las Comunicaciones, Datos Abiertos 2018). (Wirth & Hipp, 2000)

2.3. Cliente

A pesar de que el presente proyecto tiene efecto sobre varios interesados en Colombia, como ya se ha mencionado a lo largo del documento, es importante resaltar que el cliente directo del mismo es Alianza Caoba. El papel de validación del trabajo estará a cargo del Observatorio Fiscal de la Pontificia Universidad Javeriana.

2.4. Objetivos de analítica

2.4.1. Objetivo general

A partir de la arquitectura diseñada para el proyecto "ADACOP - Sistemas", aplicar técnicas analíticas, que ayuden a promover la transparencia, la competitividad, el empoderamiento ciudadano y el impacto social en el contexto de apropiación de las TIC, mediante el análisis de los datos abiertos de contratación de entidades públicas en Colombia.

2.4.2. Objetivos Específicos

- Identificar fuentes en el portal de datos abiertos (datos.gov.co) que contengan información relacionada con contratación pública.
- Identificar fuentes complementarias, provenientes de fuentes externas que permitan enriquecer los datos abiertos relacionados con contratación pública estatal.

- Realizar un análisis descriptivo que permita entender los procesos de contratación pública en Colombia, de acuerdo a las fuentes identificadas.
- Calcular estadísticos descriptivos, con el fin de validar la calidad de los datos en las fuentes utilizadas.
- Desarrollar un modelo analítico de clasificación, que permita identificar patrones y tendencias en las demoras y sobrecostos de los contratos.
- Desarrollar un modelo analítico, que permita identificar y analizar la percepción de la ciudadanía con respecto a la contratación estatal, con base en información no estructurada.
- Desarrollar un modelo analítico no supervisado que permita agrupar los contratos por características comunes con el fin de analizar demoras y sobrecostos en dichos grupos.
- Evaluar los modelos analíticos identificados para el análisis de los datos.

2.5. Plan de proyecto

Para el desarrollo de este proyecto se cuenta con una restricción de tiempo la cual está directamente relacionada con el calendario académico definido por la Maestría en Analítica para la Inteligencia de Negocios. En el Anexo I - Cronograma se encuentra descrito el cronograma de desarrollo del proyecto el cual se divide de acuerdo a las etapas correspondientes de la metodología CRISP-DM y que de igual manera están alineadas con las fechas de entrega del trabajo de grado correspondiente.

3. Entendimiento de Datos

3.1. Recolección de datos iniciales

Se cuenta, con disponibilidad inmediata, de cinco fuentes de datos básicas para su análisis, que son relacionadas en la tabla 3. Actualmente, estas fuentes se encuentran disponibles en el portal de Datos Abiertos (datos.gov.co) de donde fueron extraídas para su análisis en este proyecto.

Fuente	Descripción y características de la fuente
SECOP I	

	Información de los procesos de compra pública registrados en la plataforma SECOP I a partir del año 2011 [5]. Archivo en formato CSV
SECOP II - Contratos	Información de los Contratos correspondientes a procesos de compra pública registrados en la plataforma SECOP II desde su existencia [6]. Archivo en formato CSV
SECOP II - Procesos	Información de los procesos de compra pública registrados en la plataforma SECOP II desde su existencia, incluyendo aquellos que cuentan con contrato y los que no, y aquellos que aún están en desarrollo [7]. Archivo en formato CSV
Multas y Sanciones SECOP I:	Registro de las Multas y Sanciones generadas en la plataforma SECOP I [8]. Archivo en formato CSV
SECOP II - Proveedores Registrados:	Información Básica de los proveedores registrados en SECOP II [9]. Archivo en formato CSV

Tabla 3. Descripción de bases de datos SECOP

3.2. Descripción de datos

En el Anexo III – Diccionario de Datos se encuentras descritas las distintas variables que conforman las bases de datos mencionadas anteriormente. Este archivo contiene una hoja por cada una de las bases y sus correspondientes descripciones.

3.3. Exploración de datos

Para efectos de un mejor entendimiento de las bases de datos identificadas, se realizó una exploración inicial con el fin de conocer la información con la que se cuenta y poder, en la siguiente etapa, definir los mejores modelos analíticos que pueden ser implementados.

La descripción de los estadísticos básicos de las bases de datos se puede encontrar en el Anexo II - Estadisticas Descriptivas. A continuación, se presentan las variables más representativas de las bases de datos con por medio de diferentes gráficos para facilitar su entendimiento.

3.3.1. SECOP I

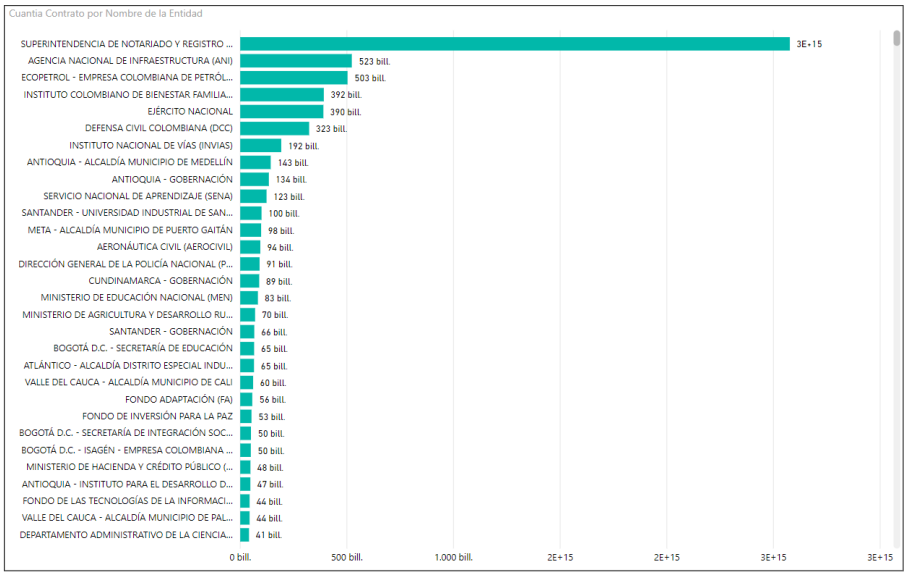

Ilustración 4- Cuantía Por Entidad. Imagen de autoría propia

En la ilustración 4 se pueden observar las entidades estatales que realizan gran cantidad de contrataciones en términos monetarios. Los valores están dados en pesos colombianos. Claramente se puede observar que la entidad que más ha gastado en contratación es la Superintendencia de Notariado y Registro y es intrigante el hecho de que esta entidad gaste más en contrataciones que la Agencia Nacional de Infraestructura (ANI).

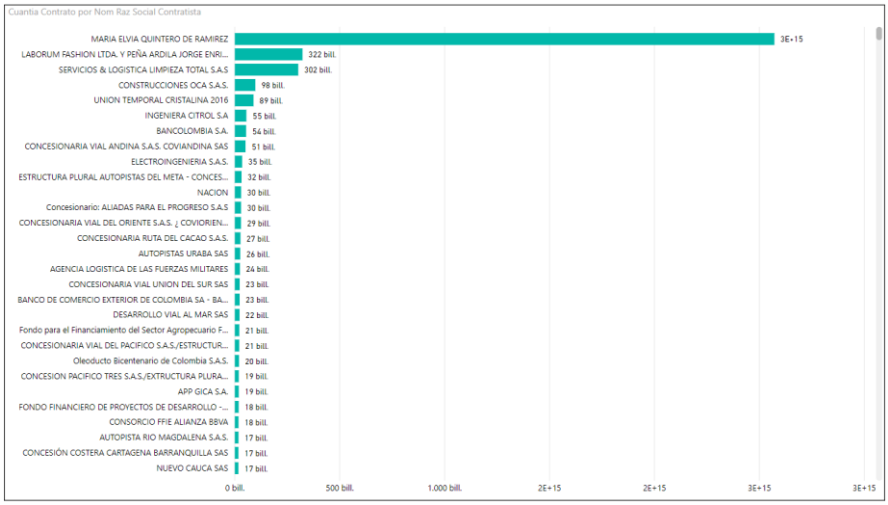

Ilustración 5 - Cuantía por Contratista. Imagen de autoría propia

En la ilustración 5 se observan los contratistas que más han recibido dinero por parte del estado. Como se observa, la entidad que más ha recibido dineros públicos es una persona natural, esto se podría deber a un error de calidad o a alguna definición de negocio que es esta etapa del proyecto se desconoce por parte del equipo de analítica. Se espera obtener más claridad al respecto durante las siguientes etapas.

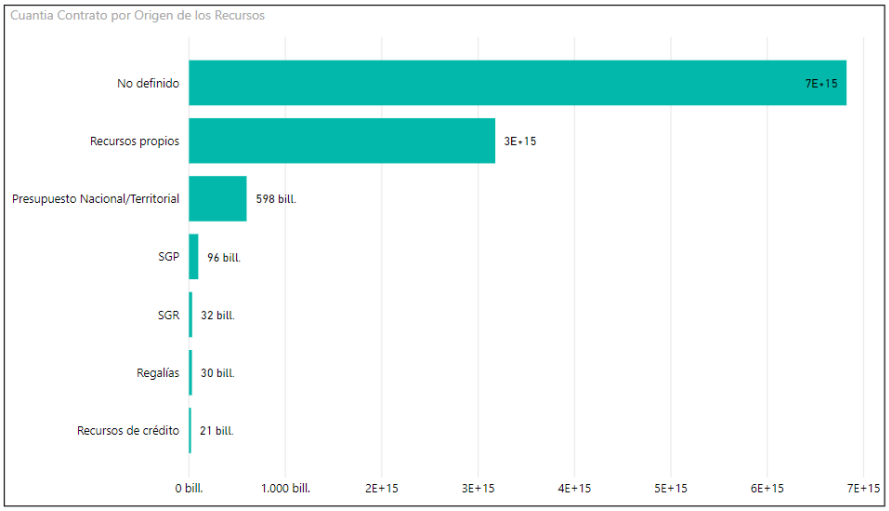

Ilustración 6 - Cuantía por Origen de Recursos. Imagen de autoría propia

La ilustración 6 presenta los orígenes de recursos que son gastados por las entidades estatales en contratación pública de bienes y servicios. Es de resaltar que la gran mayoría de los recursos no tienen definido su origen. Esto se puede deber principalmente a un error relacionado a la calidad de los datos.

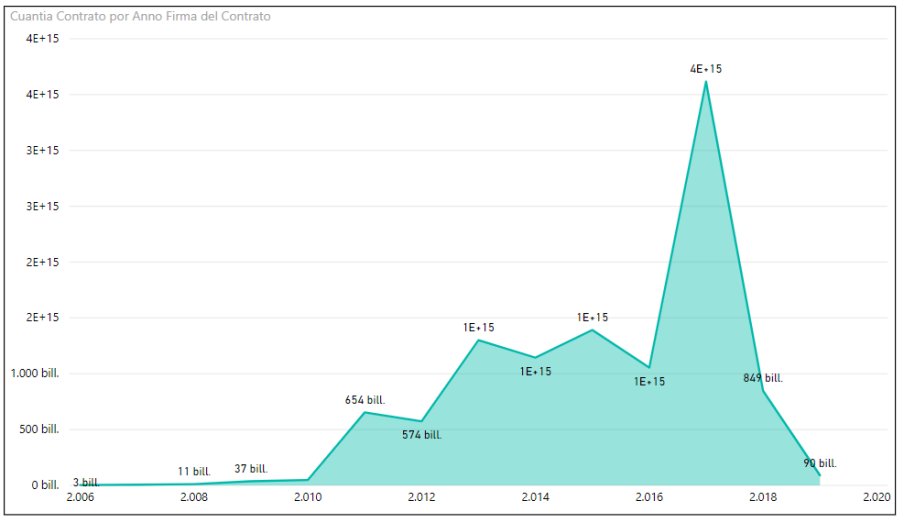

Ilustración 7 - Cuantía por Año. Imagen de autoría propia

En el gráfico de líneas, contenido en la ilustración 7, se puede apreciar la cantidad de dinero que ha sido invertido en contratación pública a lo largo de los años durante desde que el sistema SECOP fue implementado en el estado. Es de resaltar que en el año 2017 hubo un incremento significativo que en gran parte puede estar relacionado con el contexto político durante ese periodo de tiempo.

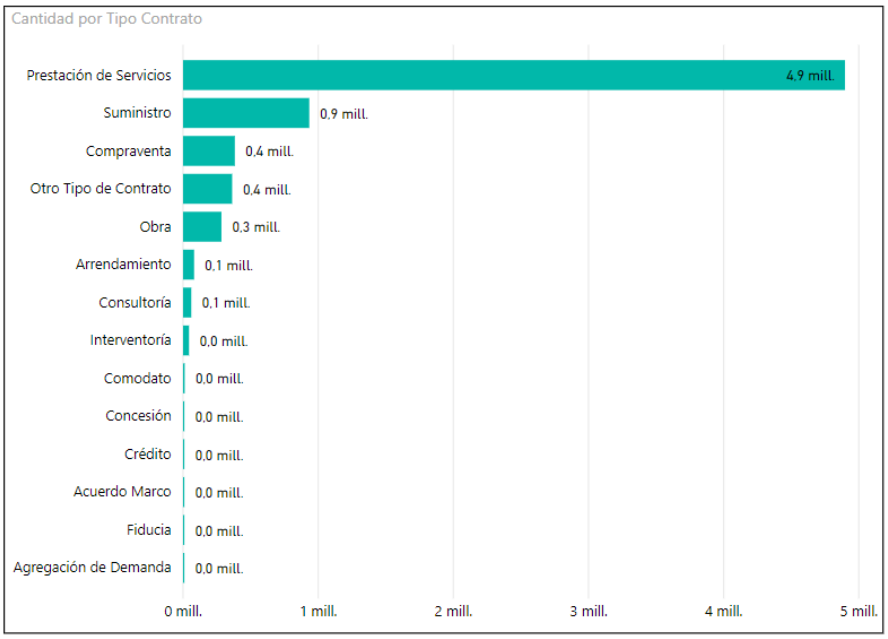

Ilustración 8 - Cantidad por Tipo de Contrato. Imagen de autoría propia

La ilustración 8 da a conocer los diferentes tipos de contrato que son celebrados por las diferentes entidades del estado. Se puede observar que el contrato que predomina es el de prestación de servicios seguido de los contratos de suministros los cuales están relacionados a diferentes insumos.

3.3.2. SECOP II- Contratos

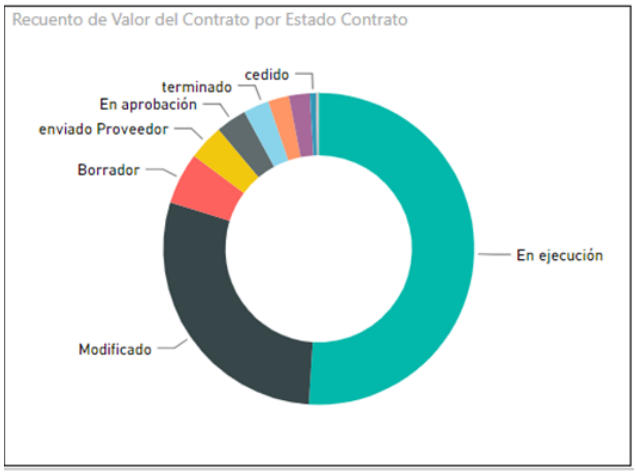

Ilustración 9 - Contrato por Estado. Imagen de autoría propia

En la gráfica de anillos de la ilustración 9, se puede apreciar como la mayoría del total de los contratos se encuentran actualmente en ejecución. Pero llama la atención que el caso de los contratos terminados sea tan pequeño en comparación con los demás, pues podría ser un indicador de incumplimiento de fechas de terminación de contratos.

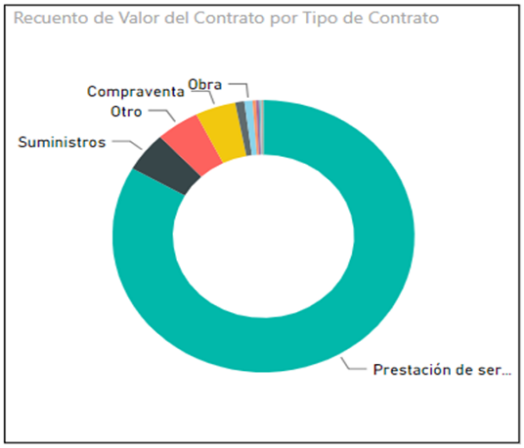

Ilustración 10 - Valor por Tipo Contrato. Imagen de autoría propia.

Por otro lado, se puede observar en la ilustración 10, que la mayor inversión en los contratos por prestación de servicios, con claramente más de un 75% de los valores de los contratos.

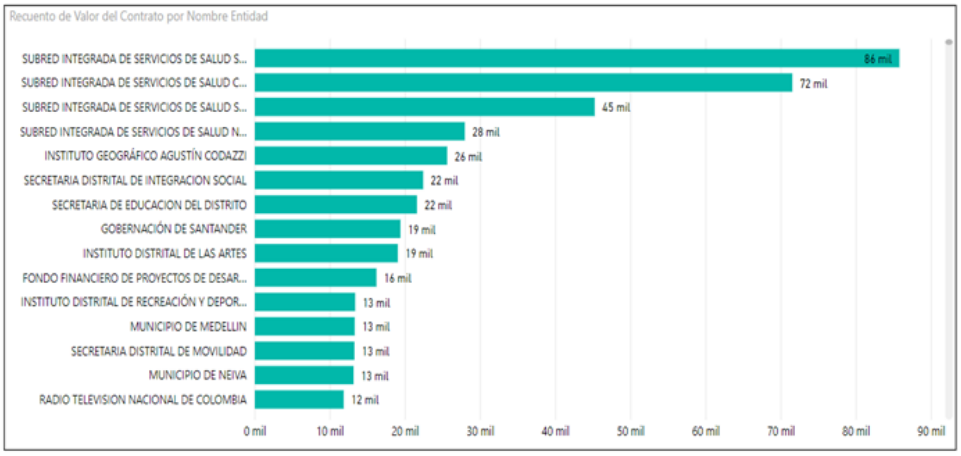

Ilustración 11 - Valor por Entidad. Imagen de autoría propia

Adicionalmente se encuentra que la mayor inversión es al sector salud (ilustración 11), pues en primeros lugares se encuentran:

- SUBRED INTEGRADA DE SERVICIOS DE SALUD SUR OCCIDENTE E.S.E
- SUBRED INTEGRADA DE SERVICIOS DE SALUD CENTRO ORIENTE E.S.E
- SUBRED INTEGRADA DE SERVICIOS DE SALUD SUR E.S.E
- SUBRED INTEGRADA DE SERVICIOS DE SALUD NORTE E.S.E

Esto llama la atención dado que superan por mucho la inversión de municipios enteros como el de Medellín y Neiva o la Gobernación de Santander.

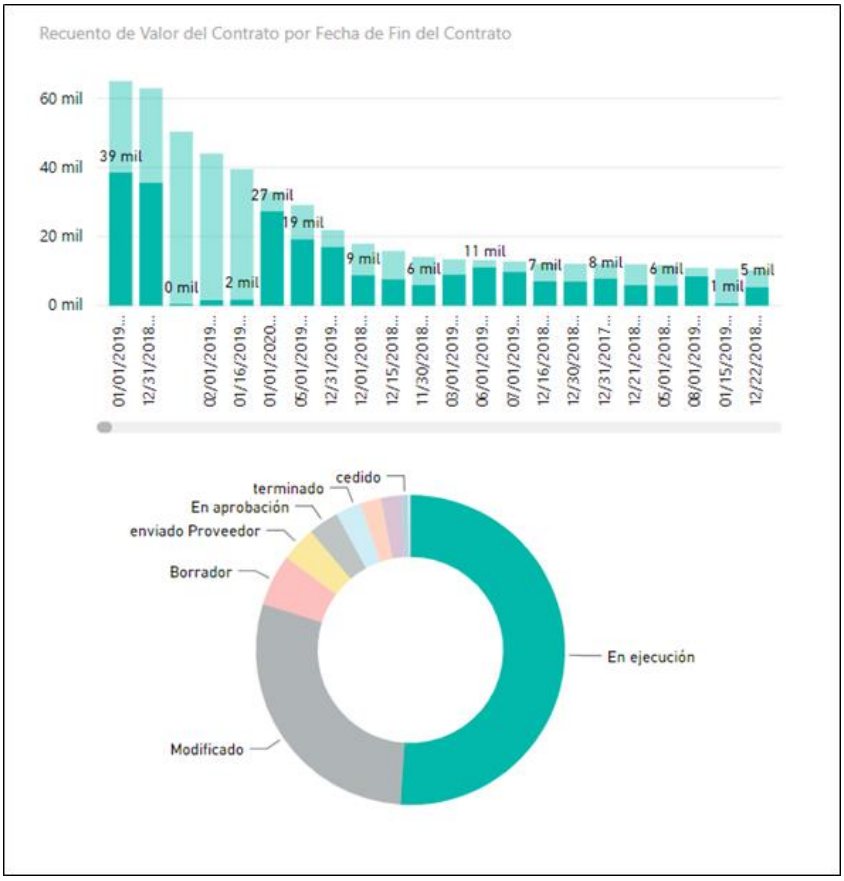

Ilustración 12 - Valor Contrato por Fecha Fin. Imagen de autoría propia

En la ilustración 12, se puede ver el cruce de los valores de los contratos contra las fechas de finalización de los mismos, pero adicionalmente se realiza un cruce con aquellos que se encuentran aún en ejecución y se puede observar como hay contratos que debían terminar a principios del 2019 y aún están en ejecución. Incluso hay casos del 2018 que no han concluido, pero lo más preocupante es que hay algunos del 2017.

3.3.3. SECOP II – Procesos

De acuerdo con los datos relacionados con los procesos de contratación pública, se tienen siete tipos de contratos; como se mencionó anteriormente, Servicios de aprovisionamiento tiene más participación en la adjudicación de los contratos; tomando los costos en miles de millones, el gráfico de la ilustración 13 refleja el valor total adjudicado para los años 2016, 2017, 2018 y 2019.

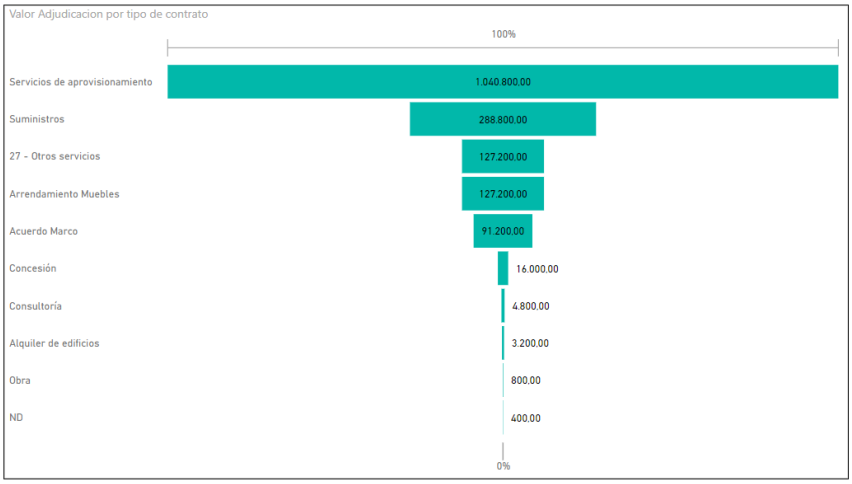

Ilustración 13 - Adjudicación por Tipo Contrato. Imagen de autoría propia

El total de contratos adjudicados es de 153.649, realizando el ranking de las ciudades con más procesos adjudicados, se evidencia que Bogotá cuenta con el 57% de la total de registros encontrados en SECOP II, como se ve en la ilustración 14.

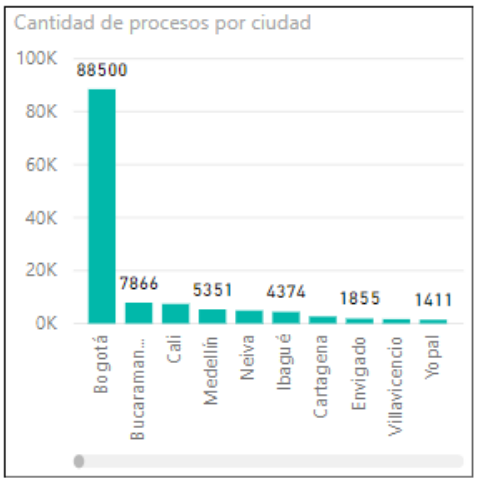

Ilustración 14 - Procesos por Ciudad. Imagen de autoría propia

Al momento de crear un proceso de contratación se define el tiempo de duración de éste (en días meses o años). Si el proceso se realiza en una ciudad, la referencia de duración se dispone en meses, mientras que, si se realiza en un municipio, se dispone en días. De acuerdo con este contexto, para un municipio, el promedio de duración de un contrato es de 186 días, mientras que, para una ciudad, como Bogotá el promedio 7,6 meses.

3.3.4. Multas y Sanciones SECOP I

Los datos revelan que, la entidad con mayor valor de sanciones es la Agencia Nacional de Infraestructura (ANI); el valor de sus multas asciende a los 118 mil millones de pesos. El

contratista cuya sanción ha sido la más alta, es la Sociedad Tren de Occidente, con una multa de 72 mil millones de pesos, aplicada a esta entidad en el año 2014. Esto se puede observar en las ilustraciones 15 y 16.

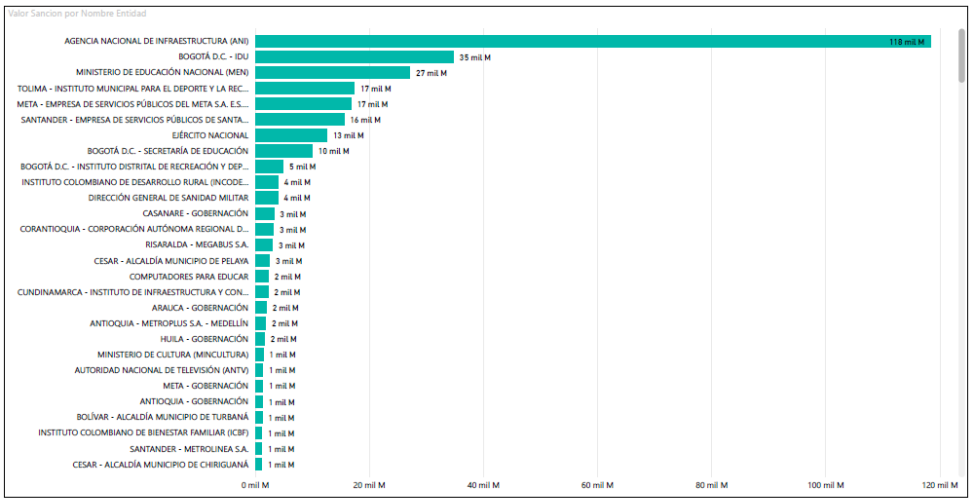

Ilustración 15 -Valor Sanción por Entidad. Imagen de autoría propia

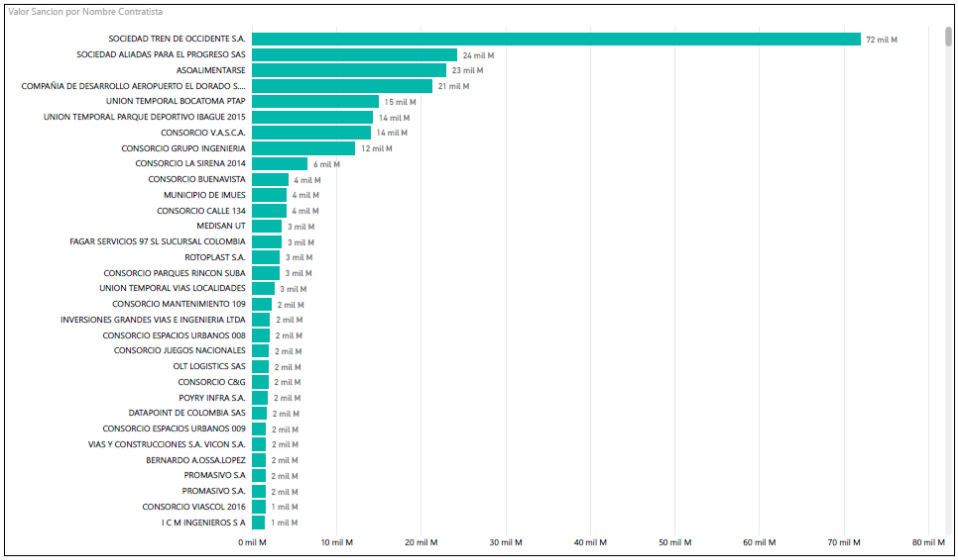

Ilustración 16 - Valor Sanción por Contratista. Imagen de autoría propia

La única variable cuantitativa del data set, "Valor Sanción", muestra una muy alta dispersión con valores de las multas que se mueven en un rango desde los cero pesos hasta los 72 mil millones de pesos, razón por la cual, la desviación estándar de la variable era tan alta (del orden de 2 mil millones de pesos).

Finalmente, resulta interesante observar en la ilustración 17, que la mayoría de las multas se otorgan en obras de orden nacional, seguido por departamental territorial y por el Distrito Capital de Bogotá.

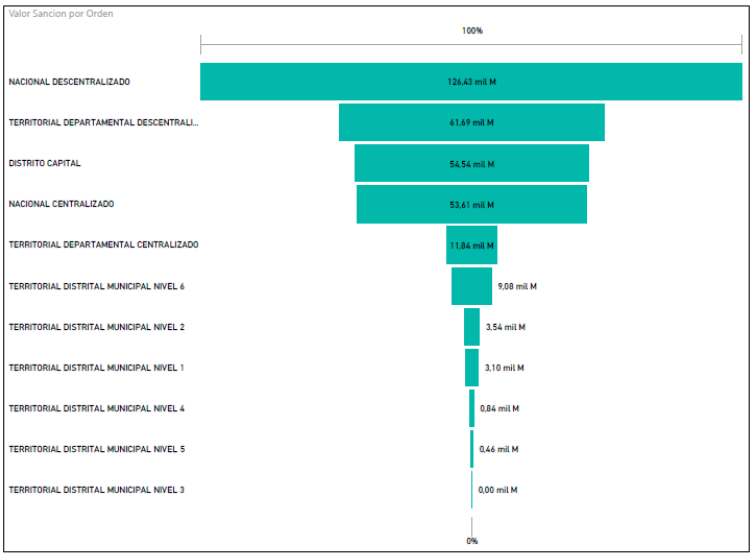

Ilustración 17 - Sanción por Orden. Imagen de autoría propia

3.3.5. SECOP II - Proveedores Registrados

Con respecto a los proveedores registrados, como se observa en la ilustración 18, en varios casos, los nombres de los proveedores registrados son inconsistentes e.g., "C", "O", o están registrados sin los apellidos correspondientes. La entidad bajo la cual figuran más registros es el IDIPRON (36).

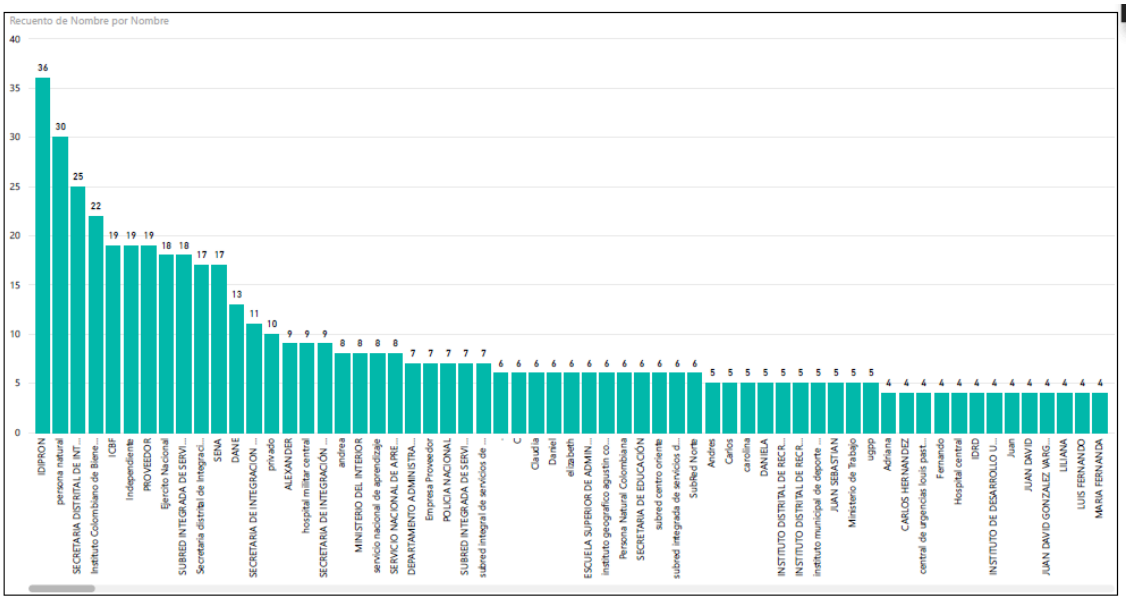

Ilustración 18- Nombre Entidades. Imagen de autoría propia

La mayoría de las empresas registradas para procesos de contratación son de tamaño grande, ya que 228 mil de éstas son no Pymes. En contraste, las empresas Pymes ascienden a 47 mil, como se observa en la ilustración 19.

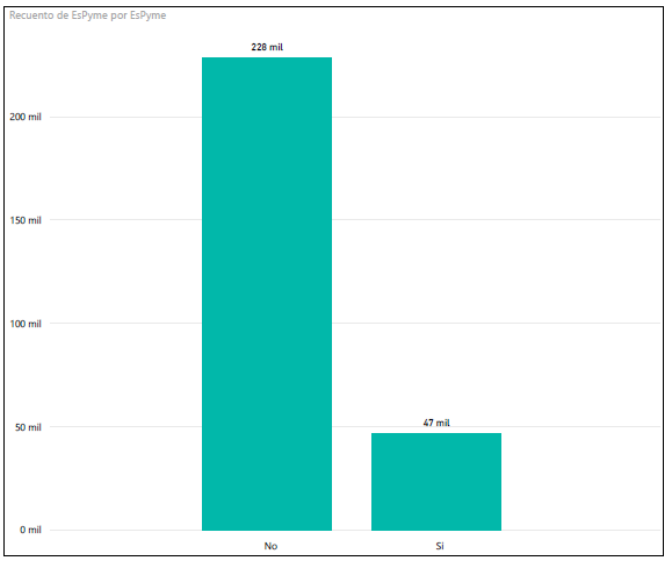

Ilustración 19- Recuento EsPyme. Imagen de autoría propia

La mayor concentración de proveedores se encuentra en el Distrito Capital de Bogotá, con un total de registros superior a 19 mil que corresponde al 76, 84%; asimismo, en su mayoría son personas naturales (221 mil), como se observa en las ilustraciones 20 y 21, respectivamente.

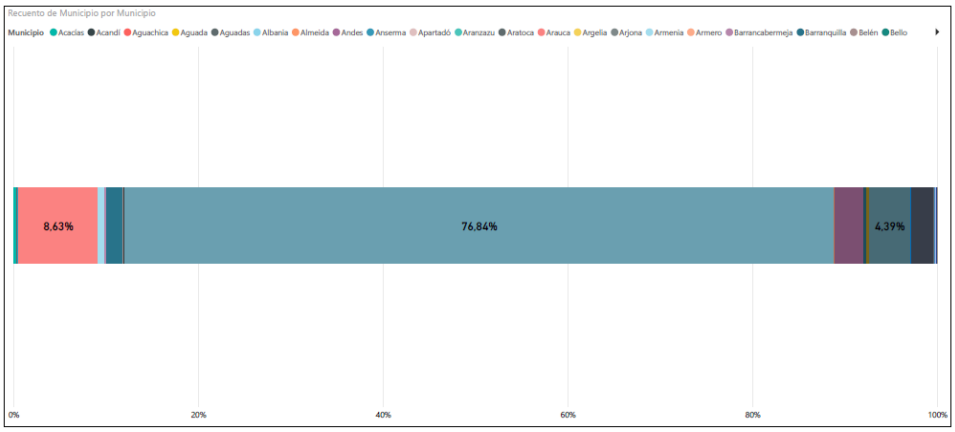

Ilustración 20 - Recuento Municipios. Imagen de autoría propia

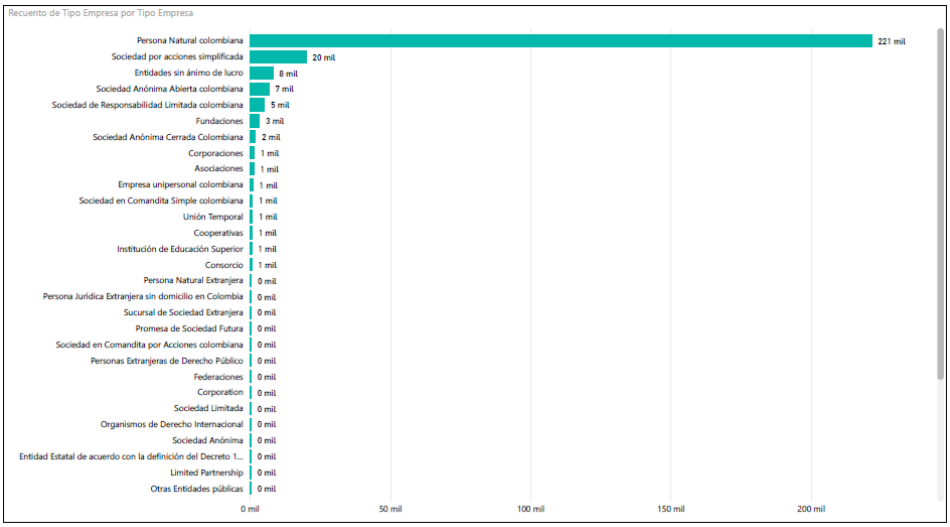

Ilustración 21- Recuento Tipo Empresa. Imagen de autoría propia

Aunque la moda de la variable "Descripción de categoría principal" corresponde a No Definido, se observa en la ilustración 22, que mayoría de los proveedores en cuyos registros sí está definido el valor de esta variable, la mayoría de los proveedores pertenecen a las categorías "Servicio de asesoría de gestión" y "Servicios profesionales de ingeniería."

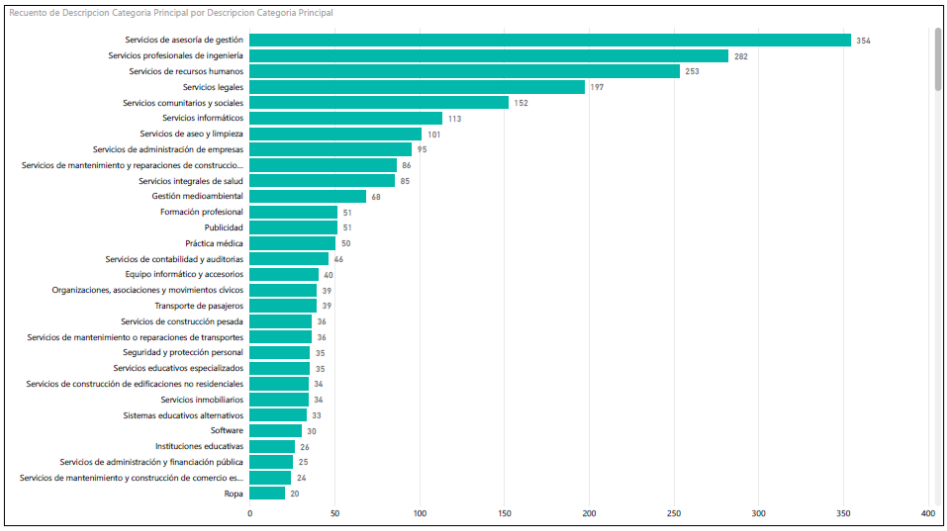

Ilustración 22- Recuento Categoría Principal. Imagen de autoría propia

3.4. Calidad de datos

3.4.1. SECOP I

De las bases de datos analizadas, es la que tiene el mayor número de registros dado que contiene el histórico de contratos desde los inicios de la plataforma. En términos

generales la calidad de los datos es aceptable. Sin embargo, es necesario analizar con más detalle las distribuciones de los datos ya que la mayoría de ellas son distribuciones de cola larga y los datos atípicos pueden llegar a ocultarse fácilmente. Adicionalmente, la cantidad de datos nulos es consistente con ciertas reglas de negocio que fueron detectadas las cuales serán abordadas en la etapa de modelado.

3.4.2. SECOP II - Contratos

En cuanto a la calidad de los datos de la base SECOP II – Contratos, se identifican algunas falencias, dado que hay varios campos a lo largo de todo el data set que se encuentran nulos o con espacios en blanco. Como es el caso de las columnas “Anno BPIN”, “Código BPIN” y “Es Post Conflicto”, en las que cada una tiene más de 700.000 registros en blanco, lo que es más del 50% del total de registros.

3.4.3. SECOP II – Procesos

En cuanto a la calidad de datos, el data set contiene valores únicos no repetidos iguales a 244.656, para este ejercicio solamente se usarán dichos valores. La variable categórica Precio_Base tiene valores negativos en 3381 registros dentro del data set. A su vez, la variable numérica Duración, refleja 13 outliers superiores a 10000 meses de duración. Dichos casos son eliminados del dataset, ya que alteran los cuartiles de la variable. En la variable Precio_Base se realiza la siguiente transformación: los números negativos se reemplazan por 0.

3.4.4. Multas y Sanciones

La calidad de los datos de la base de datos Multas y Sanciones SECOP I es buena. No existen valores inconsistentes en ninguna de las variables. Tampoco se encontraban valores faltantes para las variables, con excepción del NIT de algunas entidades, que fue un dato faltante para 415 de los 1248 registros totales.

3.4.5. Proveedores registrados

Todas los variables dentro de este conjunto de datos, son de tipo categóricas. En total, se encontraban 275069 registros. La base de datos presenta una calidad media-baja, ya que, hay bastantes datos faltantes (>23000) en todas las variables de ubicación geográfica de los proveedores. (Departamento, Municipio, Ubicación).

4. Preparación de los datos

4.1. Selección de datos

Después de haber realizado la etapa inicial de entendimiento de los datos y teniendo en cuenta las especificaciones dictadas por Alianza Caoba, se decidió que los datos con los

cuales se va a trabajar en el resto del proyecto corresponden a las bases de datos SECOP I, SECOP II - Contratos y SECOP II – Procesos. A su vez, teniendo en cuenta las métricas de calidad de los datos que se encuentran en la siguiente sección; se concluyó que para cada base existían algunas variables que no generaban ningún valor para los procesos de minería y, por lo tanto, fueron descartadas. A continuación, se relacionan dichas variables:

4.1.1. SECOP I

- Ruta Proceso en SECOP I
- Nombre Sub Unidad Ejecutora
- Proponentes Seleccionados
- Código BPIN
- ID Origen de los Recursos
- ID Adjudicación
- Nombre Clase
- ID Clase
- ID Familia
- Nombre Grupo
- ID Grupo
- Número del Contrato
- Número del Proceso
- Número del Constancia

4.1.2. SECOP II- Contratos

- Condiciones de Entrega
- Es Grupo
- Habilita Pago Adelantado
- Obligaciones Ambientales
- Obligaciones Postconsumo
- Estado BPIN
- Anno BPIN
- Código BPIN
- URL Proceso
- Documento Proveedor

- Id Contrato

4.1.3. SECOP II- Procesos

- Nit Entidad
- Nombre del Procedimiento
- Fase
- Precio Base
- Modalidad de Contratación
- Proveedores con Invitación Directa
- Visualizaciones del Procedimiento
- Proveedores que Manifestaron Interés
- Respuestas al Procedimiento
- Conteo de Respuestas a Ofertas
- Proveedores Únicos con Respuestas
- Codigo Proveedor
- NIT del Proveedor Adjudicado

4.2. Limpieza de datos

Para cada data set, se realizaron las siguientes tareas de preparación y limpieza.

4.2.1. Eliminación de outliers

- Teniendo en cuenta que las variables de las cuantías no tienen una distribución normal, para identificación de los valores, se realiza un proceso de identificación de dichos valores y por regla de negocio, se toma la decisión de eliminarlos. El proceso de identificación se realizada de la siguiente manera:

  Identificación por desviación estándar:

  ```
  df = SECOP
  data_mean, data_std = mean(df), std(df)
  cut_off = data_std * 3
  lower, upper = data_mean - cut_off, data_mean + cut_off
  ```

  En la anterior función, se tienen en cuenta el parámetro de cut_off, que da el lineamiento superior e inferior de la distribución, para identificar los valores atípicos. De los 327.530 registros, se trabaja con 291.355

4.2.2. Eliminación de datos perdidos

- Datos de variables que cuyo formato no es posible estandarizar.

  Se crea un función que recorre cada registro por cada variable, esta identifica si existe por lo menos un valor cuyo formato no corresponde al identificado por la variable, se identificaron por ejemplo caracteres especiales de tipo ¨{}¨, ¨[]¨,¨*¨, los cuales fueron eliminados; un caso específico que se identificó, corresponde a la variable *Origen de los recursos*, si bien la variable es de tipo categórico, contenía valores de ciudades que no generaba ruido en la agrupación, esto se corrige generando un conteo de valores distintos a las categorías registradas en la variable, así , si existe un valor diferente a los demás, es eliminado.

4.2.3. Selección de variables

Para realizar el ejercicio de selección de datos (sección 4.1), por cada data set, se realizó un proceso de *data profiling*, por medio del cual se identificaron los criterios de calidad contenidos en la tabla 5. Estos criterios son referenciados en mediciones realizadas a datos abiertos gubernamentales (A.Vetrò, L. Canova, M. Torchiano, C. Orozco, R. Iemma y F. Morando, 2016), ver tabla 4.

Traceability	Define el ciclo de actualización e indica el control sobre el tratamiento que se realiza en el data set (versionamiento).
Currentness	Es la actualidad de los datos, identificación de datos nuevos por registros.
Completeness	Define la coherencia entre los registros y las variables.
Compliance	Identifica valores nulos o defectos de formatos (Normalización).
Understandability	Identifica valores no legibles en las variables
Accuracy	Cálculo de la precisión de datos a nivel de columna y fila

Tabla 4. Criterios de calidad

Con base en las definiciones anteriores, se crea la matriz de indicadores y métricas de calidad, en donde se evalúan tres dimensiones: por fila (variable), por columna (registro) y por metadatos (ver tabla 5).

Categoría		Descripción
Traceability	Metadatos	Actualización = Fecha actual de actualización - última actualización
		Variación en actualizaciones =(# registros en la última actualización - # registros en la fecha de creación)/(# registros en la fecha de creación)
	Fila	Cantidad de datos actualizados por fechas = por Id de contrato realizar la comparación de fechas en la versión (Fecha inicio, fecha firma, fecha adjudicación)
	Fila	Cantidad de datos actualizados por cuantías = por Id de contrato realizar la comparación de cuantías en la versión
	Fila y Columna	Cantidad de datos = # columnas * # de filas
Currentness	Columna	Columnas nuevas = cantidad de columnas (variables) nuevas en el dataset
	Fila	Datos nuevos = compración de # campos vacíos con # campos nuevos por Id de contrato identificar la variación de los campos
Completeness	Columna	Datos vacíos por variable = Cantidad de registros no vacíos / total de registros del dataset (#contratos)
	Columna	Valores coherentes por variables= Cantidad de celdas con valores coherentes (formato de columna) / total de registros de la variable
	Fila	Registros incompletas = (1-(cantidad de filas incompletas (con datos vacíos) / cantidad total de filas registradas en el dataset))
	Fila	Registros completas = (1-(cantidad de filas completas (sin datos vacíos) / cantidad total de filas registradas en el dataset))
Compliance	Columna	Estandarización = # de columnas con información estandar (la columna de ciudad)
	Columna	Outliers = cantidad de outilers por variable / total de registros de la variable
Understandability	Columna	Valores incomprensibles por variable= # de datos no legibles por máquina / cantidad total de datos de la variable
Accuracy	Columna	Exactitud de formato Variable = # registros conformato correcto / total de registros por variable
	Fila y Columna	Exactitud de datos por contrato = registros del contratos en Secop II contratos es igual a la registrada en
	Fila y Columna	Exactitud en contratos = el id de contratos en Secop II contratos existe en Secop II procesos

Tabla 5. Criterios de calidad

En cuanto a indicadores de calidad, ser realiza un reporte (tablero de control) que muestra el cálculo de la métrica descritas anteriormente, en este contexto, se realiza la medición de la trazabilidad de los metadatos de SECOP II Procesos, obteniendo como resultados los contenidos en la tabla 6. En esta medición se evidencia la variación de datos entre las tres últimas actualizaciones. Del primer análisis al segundo, la variación en la totalidad de los datos es 83% dado que el data set contenía el 90% de datos duplicados. Ahora, sin contar los duplicados, la variación es de 18% en 16 días de diferencia entre actualizaciones. Una vez se corrigió la duplicidad, la variación de datos entre el segundo y tercer análisis, con respecto a registros nuevos, disminuye considerablemente, dando un 1% idóneo en el data set.

	Registros no duplicados	Totalidad registros	Registros duplicados
Primer Análisis	20-feb-19	20-feb-19	20-feb-19
	244.656	1.696.393	1.451.737
Segundo Análisis	8-mar-19	8-mar-19	
	288.656	288.656	
Dias de diferencia	16	**Var. Totalidad de reg.**	83%
Registros nuevos	44.000	**Var. Reg. no duplicados**	18%
Tercer Análisis	26-mar-19	26-mar-19	
	292.000	292.000	
Dias de diferencia	18	**Var. Totalidad de reg.**	1%
Registros nuevos	3.344		

Tabla 6. Medición de trazabilidad

Para el indicador de completitud por columna, se realiza el cálculo de dos métricas, la cantidad de datos completos y el porcentaje de datos faltantes de acuerdo con el total

de registros en la variable, en este caso, por ejemplo, la variable Código_bpin, es la que contiene mayor porcentaje de datos faltantes, con un 98,27 %.

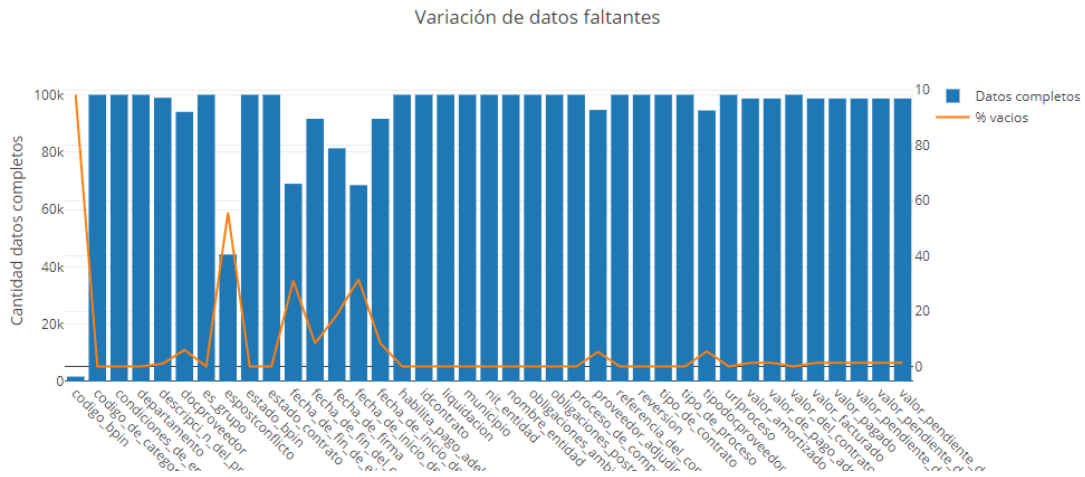

Ilustración 23 – Cantidad de datos completos con porcentaje de registros incompletos.
Imagen de autoría propia

El resultado del cálculo de completitud por filas se muestra en la tabla 7.

#_con_completos v1	#_con_completos v2	Variacion
3033	3097	2.11

Tabla 7. Medición de trazabilidad

En este indicador se realiza la comparación entre los dos últimos datasets de SECOP II, calculando la cantidad de contratos que tienen la totalidad de columnas completas, para la última versión existen 3097 registros completos, lo que representa el 1,01 % del dataset; en comparación con la versión anterior de SECOP II, se calcula la variación de dichos registros, dando como resultado que solo el 2.11% vienen completos en la última actualización.

El indicador de entendimiento se mide a nivel de lenguaje, en este se identifica el tipo de variable, para el caso de SECOP II, el ejercicio dio como resultado el contenido de la tabla 8:

Cantidad	Tipo	Lista de variables
26	Object String	['Nombre Entidad', 'Proceso de Compra', 'Descripción del Proceso','Tipo de Contrato', 'Referencia del Contrato','Fecha de Inicio del Contrato', 'Fecha de Fin del Contrato','Fecha de Inicio de Ejecucion', 'Fecha de Fin de Ejecucion','Condiciones de Entrega', 'Proveedor Adjudicado', 'Estado Contrato','Habilita Pago Adelantado', 'Liquidacion', 'Obligaciones Ambientales','Obligaciones Postconsumo', 'Reversion','Codigo de Categoria Principal', 'Tipo de Proceso', 'Fecha de Firma','Estado BPIN', 'Codigo BPIN',

		'Tipo Documento Proveedor','Documento Proveedor', 'URL Proceso', 'ID Contrato']
9	Float	['Valor de pago adelantado', 'Valor Facturado', 'Valor Pendiente de Pago', 'Valor Pagado', 'Valor Amortizado', 'Valor Pendiente de Amortizacion', 'Valor Pendiente de Ejecucion', 'Es Post Conflicto']
2	Int	[Nit Entidad, Valor del Contrato]
0	Boolean	

Tabla 8. Medición variable legibles por máquina.

De lo anterior, se identifica qué variables se deben normalizar / estandarizar y en el caso de las fechas, realizar el proceso de cambio de formato.

En cuanto al indicador actualizaciones de columnas, se evidencia que en la versión inicial de SECOP II, la variable *Municipio*, contenía la descripción del departamento / municipio en donde se ejecuta el contrato; en las dos últimas versiones de SECOP II, se separó el contenido de dicha variable, ahora en el dataset se encuentra una columna adicional, únicamente con la descripción del *Departamento,* dando como resultado para este indicador 1, que representa la columna adicional.

La actualización de filas calcula para cada registro la variación en las columnas de fechas y montos, como resultado se obtiene que en la última versión del dataset 78.448 contratos cambiaron de *Valor facturado y* en 89.964 se ajustaron las fechas de inicio de contrato y ejecución.

Finalmente, se realiza el cálculo de precisión (Accuracy) de registros, este mide si los datos de descripción de cada contrato existentes en la primera versión corresponden a los datos registrados en la última versión carga.

Como se mencionó anteriormente, las variables con más cambios son las correspondientes a fechas y valores monetarios, por esta razón estas no se tienen en cuenta, otro factor importante a discriminar, es que la última versión tendrá un delta de contratos nuevos, estos no se tienen en cuenta en el cálculo de este indicador:

N1	N2	join	acc
327530	331991	327522	99.9976

Tabla 9. Medición precisión de contratos.

Para el entendimiento del contenido de la tabla 9 se tiene que: N1: carga anterior, N2: última actualización, Join: Contratos existentes en N1 y N2; para el caso de SECOP II la

precisión de los datos descriptivos equivale al 99.99%, de lo anterior se interpreta que los valores categóricos no son cambiantes con las actualizaciones.

Vale la pena resaltar que los indicadores de calidad se miden con el dataset completo y sin modificaciones, esto ya que los indicadores son un insumo para la selección de variables y posterior justificación para la creación de los modelos y por otro lado, a nivel de negocio, para el Observatorio Fiscal de la Universidad Javeriana, es relevante identificar las mayor cantidad de variables que generen valor a nivel de contratos.

4.2.4. Estandarización / Normalización

- Para todas las variables numéricas, que definan montos o valores en dinero, se define una escala en millones.
- Las variables de fechas que definen datos tipo DATE, tienen diferentes formatos, en este caso se realiza un CAST y se estandariza un único formato dd/mmm/yyyy.
- La variable de duración del contrato está definida en años, meses y días. se estandariza para trabajar a nivel de años.
- El campo de ciudad contiene registro de municipios, este se estandariza de acuerdo con la base de municipios del DANE.

4.3. Construcción de datos

Una vez se realice el proceso de limpieza sobre las bases de datos seleccionadas correspondientes a SECOP I y SECOP II, es necesario generar nuevas variables de interés y además realizar agrupaciones con el fin de tener más claridad de los datos. De igual manera, se obtienen datos de la red social Twitter para enriquecer de cierta manera la información de SECOP.

4.3.1. Obtención datos de Twitter

La base SECOP II- Contratos fue utilizada para obtener las cuentas de los contratistas que posteriormente se usarían para extraer información relevante de Twitter. Dada la magnitud de este data set, para poder hallar las cuentas (screen_names), primero se organizó el mismo de forma descendente tomando como base la variable “valor facturado”. Una vez estuvo organizado de este modo, se procedió a dividirlo en cuatro partes iguales. De cada una de estas partes, que según cuantías se denominaron *top, high, medium* y *low*; se tomaron 25 cuentas para un total de 100. Usando la API de Twitter, los nombres/razones sociales de estos cien contratistas se buscaron en Twitter con el fin de obtener sus correspondientes screen_names. Como es obvio, no todos los

contratistas son usuarios activos de esta red social, y más aún, no todos los nombres de usuarios aparecen tal y como se encuentran consignados en la base de SECOP. Por lo anterior, se implementó un enfoque en el cual, si al hacer la primera consulta no se encontraba la cuenta con todo el nombre del contratista se le fueran eliminado palabras a dicho nombre y se continuara buscando. En algunos casos, el resultado de la búsqueda arrojaba varias cuentas de Twitter, así que en este caso se procedía a validar la ubicación geográfica de los usuarios y se conservaban las cuentas cuya localización estuviera dentro del territorio nacional.

4.3.2. Generación de variables agregadas

4.3.2.1. Fechas

Con el fin de analizar más detalladamente el comportamiento de las fechas de los contratos, las variables que contienen esta información fueron seleccionadas para hallar los tiempos transcurridos entre cada una de ellas. Aunque los dos conjuntos de datos seleccionados tienen el mismo tipo de información relacionado con los contratos ejecutados por el estado, si existen diferencias entre estas.

4.3.2.1.1. Variables de tipo fecha en SECOP I

- Fecha de Cargue en el SECOP: fecha en la que la información de los contrataos fue cargada en la plataforma de SECOP.
- Fecha de Firma del Contrato: fecha en la que el contrato fue firmado.
- Fecha Inicio Ejecución Contrato: fecha de inicio de ejecución del contrato.
- Fecha Fin Ejecución Contrato: fecha en la que finalizó la ejecución del contrato

4.3.2.1.2. Variables de tipo fecha en SECOP II

- Fecha de Firma: corresponde a la fecha de firma del contrato.
- Fecha de Inicio del Contrato: corresponde a la fecha en la que el contrato inicia.
- Fecha de Fin del Contrato: corresponde a la fecha en la que el contrato finaliza.
- Fecha de Inicio de Ejecución: corresponde a la fecha en la que inicia la ejecución de la actividades o actividades estipuladas en el contrato.
- Fecha de Fin de Ejecución: corresponde a la fecha en la que el contratista termina la ejecución de contrato

Las variables que son de interés para analizar posteriormente son las diferencias de tiempo, en días, entre las diferentes fechas, especialmente aquellas que permiten determinar el tiempo extra que se demoró un contrato a partir de lo pactado inicialmente. En el caso de SECOP II la nueva columna agregada será igual a la diferencia entre **"Fecha de Fin de Ejecución"** y **"Fecha de Fin del Contrato"** teniendo en cuenta que los valores negativos corresponderán a contratos que finalizaron a tiempo. Para SECOP I solo se cuenta con la **"Fecha Fin Ejecución Contrato"** y por ello no es posible calcular este valor. Sin embargo, esta base tiene una variable llamada **"Tiempo Adiciones en Días"** la cual brinda la información de interés.

En la tabla 10 se describen las variables agregadas generadas en las dos bases de datos con el fin de hallar las duraciones o deltas correspondientes.

Variable Generada	Operación Aplicada	Base de Datos
firma_x_inicio_c	Fecha de Firma – Fecha de Inicio contrato	SECOP II
firma_x_inicio_e	Fecha de Firma – Fecha de Inicio ejecución	SECOP I SECOP II
firma_x_fin_c	Fecha de Firma – Fecha de Fin Contrato	SECOP II
firma_x_fin_e	Fecha de Firma – Fecha de Fin Ejecución	SECOP I SECOP II
inicio_c_x_inicio_e	Fecha inicio contrato – Fecha de Inicio ejecución	SECOP II
inicio_c_x_fin_c	Fecha inicio contrato – Fecha de Fin Contrato	SECOP II
inicio_c_x_fin_e	Fecha inicio contrato – Fecha de Fin Ejecución	SECOP II
inicio_e_x_fin_c	Fecha inicio ejecución – Fecha de Fin Contrato	SECOP II
inicio_e_x_fin_e	Fecha inicio ejecución – Fecha de Fin Ejecución	SECOP I SECOP II
fin_c_x_fin_e	Fecha de Fin Contrato – Fecha de Fin Ejecución	SECOP II

Tabla 10. Variables generadas a partir de datos de tipo fecha

4.3.2.2. Cuantías

Las variables correspondientes a las cuantías proporcionan información relacionada con los valores monetarios del proceso de contratación. Para su tratamiento es importante tener en cuenta en la base SECOP I existe una variable llamada "Moneda" la cual indica la moneda en la cual se encuentran dados los valores de las cuantías. En la tabla 8 se muestra la distribución de esta variable, y como se observa son pocos los contratos dados en dólares y en otras monedas. Los contratos dados en otras monedas no se utilizarán ya que se desconoce la moneda y no es posible hacer la conversión a pesos colombianos y, en cuanto a los contratos dados en dólares, sí es necesario realizar la debida conversión ya que pueden llegar a ser significativos. Teniendo esto en cuenta, se crea una nueva columna en la cual se estandarizan los valores de todos los contratos a la misma moneda, en este caso pesos colombianos. Las nuevas variables generadas se describen en la tabla 11.

Nombre Variable	Valor	Cantidad
Moneda	Pesos (COP)	6'258.224
	No Definida	7.442
	Dólares	7.223

Tabla 11. Distribución de la variable Moneda dentro de la base de SECOP II

En la base de SECOP II no hay una variable asociada a la moneda de los contratos así que se asume que todos están dados en pesos colombianos y por lo tanto no es necesario aplicar generación de una nueva variable.

Variable Original	Variable Generada (COP)
Cuantía Proceso	Cuantía Proceso COP
Cuantía Contrato	Cuantía Contrato COP
Valor Total de Adiciones	Valor Total de Adiciones COP
Valor Contrato con Adiciones	Valor Contrato con Adiciones COP

Tabla 12. Variables agregadas normalizadas en pesos colombianos

4.3.2.3. Columnas binarias

Para el análisis de los tiempos que proporcionan las fechas se asumió que existe un orden entre éstas, el cual se describe a partir de las siguientes reglas:

1. La fecha de inicio del contrato debe ser mayor o igual a la fecha de firma de este. En caso contrario, esto podría indicar un posible error en los campos de las fechas.
2. La fecha de inicio de ejecución del contrato debe ser mayor o igual a la fecha de inicio de este. En caso contrario significaría que el contrato se empezó a ejecutar antes de la fecha pautada.
3. La fecha de finalización del contrato debe ser mayor o igual a la fecha de inicio de ejecución de este. En caso contrario puede significarse un error en los campos de fechas.
4. La fecha de fin de ejecución del contrato debe ser debe ser mayor o igual a la fecha de inicio de este. En caso contrario, se podría tratar de un posible error en los campos.

A partir de estas reglas, se crearon diferentes columnas binarias las cuales indican si la diferencia de las fechas es o no correcta. Las columnas creadas se describen en la tabla 13.

Variable Generada	Operación aplicada	Base de Datos
firma_x_inicio_c_flag	1 cuando la Fecha de Firma es mayor a la Fecha de Inicio contrato, 0 de lo contrario	SECOP II
firma_x_inicio_e_flag	1 cuando la Fecha de Firma es mayor a la Fecha de Inicio ejecución, 0 de lo contrario	SECOP I SECOP II
firma_x_fin_c_flag	1 cuando la Fecha de Firma es mayor a la Fecha de Fin Contrato, 0 de lo contrario	SECOP II
firma_x_fin_e_flag	1 cuando la Fecha de Firma es mayor a la Fecha de Fin Ejecución, 0 de lo contrario	SECOP I SECOP II
inicio_c_x_inicio_e_flag	1 cuando la Fecha inicio contrato es mayor a la Fecha de Inicio ejecución, 0 de lo contrario	SECOP II

inicio_c_x_fin_c_flag	1 cuando la Fecha inicio contrato es mayor a la Fecha de Fin Contrato, 0 de lo contrario	SECOP II
inicio_c_x_fin_e_flag	1 cuando la Fecha inicio contrato es mayor a la Fecha de Fin Ejecución, 0 de lo contrario	SECOP II
inicio_e_x_fin_c_flag	1 cuando la Fecha inicio ejecución es mayor a la Fecha de Fin Contrato, 0 de lo contrario	SECOP II
inicio_e_x_fin_e_flag	1 cuando la Fecha inicio ejecución es mayor a la Fecha de Fin Ejecución, 0 de lo contrario	SECOP I SECOP II
fin_c_x_fin_e_flag	1 cuando la Fecha de Fin Contrato es mayor a la Fecha de Fin Ejecución, 0 de lo contrario	SECOP II

Tabla 13. Variables generadas a partir de las diferencias entre las diferentes fechas

Por el lado de las cuantías, con el fin de identificar sobrecostos en los contratos, en SECOP II, se genera la variable "diferencia_valor" de tipo binaria, esta variable contiene la diferencia entre el valor pagado y el valor de contrato, calculada de tal manera que si el contrato tiene diferencia, se guarda el valor 1 , en caso contrario se guarda el valor 0.

Es importante aclarar que en las variables de fechas y valores existen campos nulos que no se pueden operar, los contratos cuyos campos están vacíos, no se tienen en cuenta para la generación del modelo de clasificación, es decir, se eliminan 73.986 registros, trabajando con 217.369.

4.3.3. Agrupaciones de datos

Con el fin de desarrollar los dashboards para visualizar la información de manera más comprensible para los usuarios, se realizaron una serie de agrupaciones sobre los datos para así generar conjuntos de datos temporales que puedan ser utilizado para las visualizaciones.

En términos del desarrollo del componente tecnológico de ADACOP, se desarrolló una función en PySpark sobre la infraestructura de CAOBA, la cual permite la agregación de datos apta para la visualización de histogramas. En esencia, esta función recibe como parámetros el nombre de la base de datos, ya

sea SECOP I o SECOP II y la variable sobre la cual se quiere realizar la tarea de agregación. Como resultado se obtiene la distribución de esta variable, es decir, todos los diferentes valores existentes dentro de la columna y su respectivo conteo de apariciones. De esta manera, se cuenta con una herramienta para analizar las distribuciones de las variables de interés requeridas para el desarrollo de modelos analíticos.

5. Modelado

5.1. Selección de técnicas de modelado

Posterior a la limpieza y preparación de las bases de datos se seleccionan las técnicas de modelado más adecuadas para lograr los objetivos definidos por el negocio. Para la selección de

Para los interesados es de gran valor analizar el comportamiento y características de aquellos contratos los cuales incurren en sobre costos en las cuantías de estos y en demoras en los tiempos de entrega.

5.1.1. Arboles de decisión

Los árboles binarios, son herramientas de clasificación que permiten el análisis de conjuntos de decisiones, organizadas en una estructura jerárquica, siendo de fácil visualización y entendimiento (SALTOS, Giger. COCEA, Mihalea,2017). Los modelos de árboles de clasificación contienen una estructura de decisión y un algoritmo que lo resuelve. El objetivo de este modelo es realizar agrupaciones por características comunes, en la base de contratos y procesos contenidos en SECOP II con el fin de analizar si el contrato va a tener demoras (variable objetivo) y si el contrato va a tener sobrecostos (variable objetivo). En este contexto, además de las variables mencionadas en la sección 4.3.2, se seleccionan las siguientes variables que generan mayor ganancia:

- Es Post Conflicto: variable binaria que identifica si el contrato tiene relación con temas de post conflicto. 1 para Si, 0 para No.
- Tipo de contrato: Modalidad a través de la cual se desarrolló el proceso de compra.
    - Prestación de servicios
    - Compraventa
    - Obra

- Suministros
- Arrendamiento de inmuebles
- Asociación Público Privada
- Otro
- Interventoría
- No Especificado
- Servicios financieros
- Arrendamiento de muebles
- Consultoría
- Comisión
- Seguros
- Acuerdo de cooperación
- Concesión
- Negocio fiduciario
- Acuerdo Marco de Precios
- Venta muebles

- Liquidación: 1 para Si, 0 para No y No definido
- Obligaciones Ambientales: 1 para Si, 0 para No.
- Estado Contrato:
    - Borrado
    - Cancelado
    - Cerrado
    - En aprobación
    - En ejecución
    - Modificado
    - Terminado
    - Activo
    - Prorrogado
    - Suspendido
    - Cedido
    - Enviado a proveedor

Para este modelo, se evaluarán los algoritmos ID3 y C 4.5, como técnicas de particionamiento, esto, teniendo en cuenta los valores resultantes de ganancia, que selecciona en cada nodo el atributo con mayor ratio de ganancia de

información, ignora datos perdidos y genera reglas de clasificación y los valores de entropía.

5.1.1.1 Entrenamiento del modelo

Antes de iniciar con el entrenamiento, se realiza una revisión de la distribución de las clases de acuerdo con las dos variables objetivo, obteniendo el siguiente resultado por clase, identificando las demoras:

diferencia_fecha	Cantidad
1	18.050
0	204.859

Tabla 14. Distribución de la variable diferencia_fecha-DEMORA total

En cuanto a los sobrecostos, se obtuvo el siguiente resultado por la clase:

diferencia_valor	Cantidad
1	14.854
0	208.055

Tabla 15. Distribución de la variable diferencia_valor -SOBRECOSTOS total

De lo anterior, es notorio que se debe aplicar una técnica de balanceo entre las clases, tanto para demoras como para sobrecostos.

5.1.2. Análisis de Sentimientos

El procesamiento de lenguaje natural es una rama de la analítica descriptiva que busca generar conocimiento adicional, con base a las interacciones humanas. Parte de este es el análisis de sentimientos que busca entender cómo se sentía una persona, tratando de determinar la polaridad de su discurso y las emociones asociadas. Para el caso especifico de ADACOP, lo interesante sería detectar la polaridad de la población civil sobre algunos de los contratistas, mediante su interacción en twitter.

Teniendo en cuenta esto, se desarrollarán 3 modelos diferentes, el primero de ellos bag of words, que toma como base una serie de términos clasificados como positivos o negativos previamente y buscándolos en los tweets para tener por cada uno una suma que cargas polares que de como resultado la polaridad del tweet. También se desarrollará un modelo Naive con tweets previamente

clasificados que se utilizarán para entrenarlo. El tercer y último modelo a desarrollar es un modelo basado en SenticNet.

SenticNet es un desarrollo en lo que respecta a procesamiento de lenguaje natural (NLP) que trabaja en la identificación de sentimientos y afectivos, teniendo en cuenta la polaridad de las palabras, la estructura gramatical y una serie de reglas gramaticales que dan como resultado el valor sentimental y afectivo de una frase. Aunque los resultados en español no son los mejores, en inglés el desarrollo está mucho más avanzado y por ende los resultados son mucho mejores y se acercan a un 89% de precisión (Cambria 2015). Con esto en mente, se piensa en la oportunidad de desarrollar un módulo de análisis de sentimientos relacionados a la interacción de los usuarios de la red social Twitter, sobre los contratos públicos en Colombia. De esta manera se podría conocer el sentimiento que genera en los ciudadanos uno u otro contrato.

6. Evaluación

6.1. Resultados y análisis

6.1.1. Árboles de predicción

Para llevar a cabo la implementación del modelo con árboles se utilizó Spark, un sistema de computación en clúster, rápido y de uso general para Big Data. Proporciona API de alto nivel en Scala, Java, Python y R, y un motor optimizado que admite gráficos de computación generales para el análisis de datos. También es compatible con un amplio conjunto de herramientas de alto nivel que incluyen Spark SQL y DataFrames, MLlib para aprendizaje automático, GraphX para procesamiento de gráficos y Spark Streaming para procesamiento de flujos. Por lo anterior, y a petición del product owner, se utilizó únicamente la librería mllib de Pyspark para la implementación.

El primer paso que se realizó, después de importar las librerías necesarias, fue cargar la base de datos Secop- II como un dataframe de Spark. Posteriormente se analizó la cantidad de valores no nulos y de valores únicos presentes en cada columna del dataset y se eliminaron de acuerdo con lo descrito en la sección 5.1.

De acuerdo a las variables a predecir: 1) diferencia de fecha y 2) diferencia de valor (sobrecosto), se crearon dos dataset dividiendo el original con respecto a las dos variables independientes Estado Contrato Cerrado y Estado Contrato Ejecución. Partiendo de los dos dataset creados (cerrados y en ejecución) se crean dos tablas nuevas independientes para cada uno de los modelos eliminando las columnas de estado de contrato, y las variables con las que se construyeron las columnas de sobrecosto y demora. Las variables categóricas se transformaron a índices como lo exigen las implementaciones de machine learning y, asimismo, se vectorizaron las variables independientes.

Se crearon dos modelos de árboles de clasificación, con un parámetro máxDepth igual a 3 y se calcularon las predicciones sobre el dataset de contratos en ejecución. A continuación, se presentan los resultados de las predicciones para las dos variables:

6.1.1.1.1. Demoras

- **Ejecución 1:** resultados contenidos en la Tabla 16.

diferencia_fecha	prediction	count
1	0	479
0	0	3986

Tabla 16. Predicción clase DEMORAS – modelo 1

Accuracy: 0.89,

Precision: None,

Recall: 0.0,

f1: None,

True_positive: 0,

True_negative: 3986,

False_positive: 0,

False_negative: 479.

- **Ejecución 2:** resultados contenidos en la Tabla 17

diferencia_fecha	prediction	count
1	0	7208
0	0	133975

Tabla 17. Predicción clase DEMORAS – modelo 2

Accuracy: 0.94,

Precision: None,

Recall: 0.0,

f1: None,

True_positive: 0,

True_negative: 133975,

False_positive: 0,

False_negative: 7208.

De acuerdo a los dos modelos anteriores, en la predicción de demoras, teniendo en cuenta los resultados de la matriz de confusión y a pesar de que el valor del accuracy es superior al 88 % en ambos casos; los datos muestran que el porcentaje de instancias bien clasificadas (Precisión) es 0, lo que significa que en la predicción no hay contratos con demoras, a pesar de que la variable objetivo contaba con casos que justificaban que el contrato si estaba retrasado.

6.1.1.1.2. Sobrecostos

- **Ejecución 1:** resultados contenidos en la Tabla 18.

diferencia_valor	prediction	count
1	0	33
0	0	3203
1	1	1062
0	1	167

Tabla 18. Predicción clase SOBRECOSTOS – modelo 1

Accuracy: 0.95,

Precision: 086,

Recall: 0.96,

f1: 0,91,

True_positive: 1062,

True_negative: 3203,

False_positive: 167,

False_negative: 33.

De acuerdo al modelo anterior, en la predicción, el valor del accuracy es superior al 95 %, con una precisión del 86% en las instancias bien clasificadas y con un F1-Score (que tiene en cuenta la precisión y la exhaustividad) del 91%, lo que indica que la evaluación del modelo y el rendimiento del clasificador son buenos, es decir, se tiene una maximiza tasa de verdaderos positivos y una minimiza tasa de falsos positivos.

- Ejecución 2: resultados contenidos en la Tabla 19

diferencia_valor	prediction	count
1	0	2707
0	0	117226
1	1	6008
0	1	15242

Tabla 19. Predicción clase SOBRECOSTOS – modelo 2

Accuracy: 0.87,

Precision: 0.28,

Recall: 0.69,

f1: 0.4,

True_positive: 6008,

True_negative: 117226,

False_positive: 15242,

False_negative: 2707

De acuerdo con el modelo anterior, en la predicción, el valor del accuracy es superior al 87 %, con una precisión del 28% en las instancias bien clasificadas y con un F1-Score (que tiene en cuenta la precisión y la exhaustividad) del 40%, indica que la evaluación del modelo y el rendimiento del clasificador no son buenos.

De acuerdo con los modelos anteriores, se identifica:

Hallazgos a nivel de negocio

- De los casos que se predijeron correctamente y que generan sobrecostos, aproximadamente el 70% de los contratos se ejecutaron en la ciudad de Bogotá, el resto se distribuye en Medellín, Manizales e Ibagué.

Hallazgos a nivel de modelo

- En la predicción de demoras, los dos modelos generados presentan sobreajuste, dado que la clase dominante es 0, se pensó en solucionarlo realizando un balanceo de clases, pero calidad de datos de SECOP en las variables seleccionadas no permitía generar un set de datos de entrenamiento y validación, nivelado.
- Dado que los árboles generados tenían pocos niveles de profundidad, pero muchas ramas (amplitud), no fue fácil realizar el gráfico de visualización.

6.1.2. Análisis de sentimientos

Para el análisis de sentimientos, se querían realizar tres enfoques diferentes uno con bag of words, otro con SenticNet y un clasificador Naive Bayes entrenado con un set de tweets ya clasificados, pero esta base de tweets no tenía nada que ver con temas políticos o de contratación pública. Los tres modelos se implementaron, pero no había una base ya clasificada de tweets con los cuales se pudiera evaluar el rendimiento de cada uno de los modelos. Para poder validar estos resultados y definir adecuadamente el mejor modelo para este caso específico, es necesario validarlos con tweets previamente clasificados manualmente, en el rango de los miles. Por este motivo, esto se salía del alcance del proyecto.

6.1.3. Extracción de tópicos y Nube de palabras

Para desarrollar la extracción de tópicos se utilizó el conjunto de los 21506 tweets de contratación estatal, con el objetivo de observar los temas más relevantes de los cuales se estaba opinando en dicha materia. Estos tweets se extrajeron utilizando hashtags y cuentas relacionados con contratación. Es importante resaltar, que el enfoque que se había pensado utilizar en un primer momento, donde se extraían los tweets de las supuestas cuentas de los contratistas, fue descartado dado que los datos recopilados no eran relevantes para el objetivo de este estudio.

La cantidad total de palabras en ese conjunto era de 83892 con un total de 14519 y una diversidad léxica: 0.173. El modelo de tópicos de construyó usando el enfoque Latent Dirichlet allocation- LDA. Para poder analizar únicamente los resultados de los tópicos

más relevantes, se corrió la implementación con un total de 5 tópicos y se analizaron 15 términos por cada uno de estos. El resultado total se muestra a continuación:

[(Grupo 0, '0.029*"presidente" + 0.009*"historia" + 0.006*"señor" + 0.005*"continuidad800temporalessena" + 0.005*"cara" + 0.005*"vía" + 0.005*"realidad" + 0.005*"vergüenza" + 0.005*"acuerdo" + 0.005*"familias" + 0.004*"problema" + 0.004*"palabras" + 0.004*"cuenta" + 0.004*"gobierno" + 0.004*"paso"')
(Grupo 1, '0.017*"negocio" + 0.014*"cosas" + 0.014*"empresas" + 0.013*"éxito" + 0.010*"corrupción" + 0.009*"rendicióndecuentas" + 0.008*"método" + 0.008*"podrás" + 0.007*"discurso" + 0.007*"transparencia" + 0.006*"comienzos" + 0.005*"caso" + 0.005*"favor" + 0.005*"años" + 0.005*"alcalde"')
(Grupo 2, '0.011*"información" + 0.010*"transparencia" + 0.009*"sector" + 0.008*"país" + 0.008*"verdad" + 0.007*"vía" + 0.007*"años" + 0.007*"vez" + 0.006*"ley" + 0.006*"servicio" + 0.006*"año" + 0.006*"proyecto" + 0.006*"campo" + 0.006*"marzo" + 0.005*"gestión"')
(Grupo 3, '0.010*"pueblo" + 0.009*"día" + 0.008*"país" + 0.008*"temporalessena" + 0.008*"guerra" + 0.007*"vida" + 0.006*"cuentas" + 0.006*"gobierno" + 0.006*"paz" + 0.006*"poder" + 0.006*"señor" + 0.006*"normalidad" + 0.006*"frontera" + 0.005*"millones" + 0.005*"pobreza"')
(Grupo 4, '0.059*"emprendimiento" + 0.015*"gracias" + 0.014*"emprendedores" + 0.011*"metro" + 0.009*"país" + 0.008*"negocios" + 0.008*"datos" + 0.007*"días" + 0.007*"gente" + 0.006*"fin" + 0.006*"ayuda" + 0.006*"mundo" + 0.006*"atención" + 0.006*"innovación" + 0.005*"estudios"')]

Algunos de los tópicos más importantes obtenidos de los tweets y de relevancia para temas de contratación estatal son:

- Rendicióndecuentas
- Emprendimiento
- Metro
- Transparencia
- Cuentas
- Pobreza
- Vergüenza

- Corrupción

Como se observa, varios de los tópicos de los que se habla en contratación estatal en Colombia presentan una polaridad de tipo negativo e.g., vergüenza, corrupción. En su mayoría, el resto de las palabras son sustantivos de tipo neutro. Lo anterior se puede interpretar como un reflejo de la imagen que tienen los ciudadanos sobre temas de contratación en el país.

Con respecto a la nube de palabras, se usó un enfoque de Bag of Words para poder obtener los términos más representativos de los tweets de contratación estatal. El objetivo de esto fue poder incluir en el dashboard de SECOP, de manera gráfica, una herramienta para brindar a los usuarios una perspectiva de la imagen de los temas de contratación pública, obtenida de la red social Twitter. La ilustración 17 muestra el resultado descrito.

Ilustración 24. Nube de palabras Imagen de autoría propia

7. Conclusiones

- La campaña de datos abiertos a nivel mundial es muy importante y tiene mucho potencial, pero, aunque trata de generar un gobierno más transparente, se queda corta teniendo en cuenta que no se tienen en cuenta barreras técnicas y tecnológicas que

vuelven prácticamente inservible toda información ahí publicada. Para cualquier persona que quiera enfrentarse a la labor de revisar los datos, requerirá de una infraestructura tecnológica grande con la que la mayoría no tiene al alcance y además debe tener algunos conocimientos medios para poder analizar de manera adecuada y eficiente todo el volumen de la información. Estas primeras barreras ya descartan a la gran mayoría de la población y adicionalmente hay que tener en cuenta que se debe tener un conocimiento de negocio para poder tener claridad sobre el significado de uno u otro valor para cada variable del dataset.

- Iniciativas como ADACOP son muy útiles y son un gran aporte para dar claridad y mayor entendimiento de la información en el portal de datos abiertos. Pero aún quedan muchas bases por trabajar, además de que queda por implementar estrategias para empoderar realmente al ciudadano dándole a conocer cuáles son y que significan al menos las variables más relevantes de cada data set y como se pueden entender dentro de los *dashboards*.
- Los árboles obtenidos dieron resultados buenos con respecto a predicciones de sobrecostos. Sin embargo, para la predicción de demoras los resultados no fueron los esperados ya que se obtuvo que ningún contrato se demoraría. Lo anterior está directamente ligado a la calidad de los datos contenidos en las bases de datos de SECOP; a pesar de los esfuerzos de limpieza y preparación de datos, se esperaban resultados mejores.
- La extracción de tópicos y nube de palabras elaborados con la información extraída de la red social Twitter refleja la opinión de los ciudadanos con respecto a temas de contratación estatal. El manejo de datos no estructurados supone un reto grande si se quiere obtener información que sea realmente relevante y pertinente para los temas bajo estudio.
- Al trabajar con datos abiertos, un factor importante y que requiere de bastante trabajo es la estandarización en la calidad de los datos, no solo en los seis indicadores que se generaron en este proyecto, que evalúan la conformidad y legibilidad de los datos para realizar posteriores cálculos y modelos, sino también la calidad a nivel de negocio, es decir, para un ciudadano interesado (con experiencia contractual) no es posible tomar decisiones de manera sencilla, justamente porque hay carencia de información.
- Aunque el problema de calidad en ciertas variables de las bases de datos de SECOP es un problema evidente, principalmente por la completitud de la información, no significa que los datos no sean aptos para desarrollar procesos analíticos, de hecho, el proceso llevado

a cabo en este trabajo de grado permitió al observatorio fiscal determinar problemas de negocio relacionados directamente a la calidad de los datos, como por ejemplo la inexistencia de procesos de verificación al momento de ingresar la información al sistema (SECOP).

- Las plataformas de SECOP implementadas por la política de Colombia Compra Eficiente integradas con la política de datos abiertos es un primer acercamiento por parte del gobierno colombiano para aumentar la transparencia y acceso a la información estatal para la ciudadanía. Sin embargo, aún se requieren políticas estatales que incentiven proyectos como ADACOP con el fin de extraer valor de los datos que realmente apalanquen el mejoramiento de los procesos de tomas de decisiones en las entidades del estado y que, adicionalmente, entreguen a la ciudadanía información fácil de entender con el fin de empoderarlo en dichos procesos.

8. Referencias

Cambria, Erik. 2015. "Sentic Computing. Springer International Publishing."

"'El futuro digital es de todos': la nueva política TIC - Ministerio de Tecnologías de la Información y las Comunicaciones". https://www.mintic.gov.co/portal/604/w3-article-79186.html (19 de febrero de 2019).

"Informe Global | Open Data Barometer". https://opendatabarometer.org/4thedition/report/?lang=es (28 de enero de 2019).

Pauwels, Koen et al. 2009. "Dashboards as a Service: Why, What, How, and What Research Is Needed?" Journal of Service Research 12(2): 175-89.

Vila, R.A., E. Estevez, y P.R. Fillottrani. 2018. "The design and use of dashboards for driving decision-making in the public sector". En , 382-88.

Ministerio de las TIC, (2018). 'El futuro digital es de todos': la nueva política TIC. Recuperado de https://www.mintic.gov.co/portal/604/w3-article-79186.html

Open data charter, (2015). Carta internacional de datos abiertos. Recuperado de: https://opendatacharter.net/principles-es/

Ministerio de las TIC, (2016). Guia de datos abiertos en Colombia: Recuperado de http://estrategia.gobiernoenlinea.gov.co/623/articles-8248_Guia_Apertura_Datos.pdf.

Datos abiertos - ministerio de tecnologías de la información y las comunicaciones. Recuperado https://www.mintic.gov.co/portal/604/w3-article-62310.html.

Ministerio de las TIC, (2014). Ley 1712 de 2014. Recuperado de https://www.mintic.gov.co/portal/604/articles-7147_documento.pdf

Ministerio de las TIC, (2015) Resolución 3564. Recuperado de https://www.mintic.gov.co/portal/604/w3-article-14476.html

S. S. Dawes y N. Helbig, (2010). Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency, en *Electronic Government* , pp. 50-60.

C Martínez, A, (2018) .Big data and open data for an intelligent governance. Profesional de la Información 27, n.º 5, Recuperado de https://doi.org/10.3145/epi.2018.sep.16 , pp. 1128-35

Z. Yang y A. Kankanhalli,(2013) Innovation in Government Services: The Case of Open Data, en *Grand Successes and Failures in IT. Public and Private Sectors*, pp. 644-651.

N. Surasvadi, C. Saiprasert, S. Thajchayapong, (2017). 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), *Budget and Procurement Analytics using Open Government Data in Thailand* .pp. 1-4. Recuperado de https://ieeexplore.ieee.org/document/8074079

European Commission. (2014) Report on high-value datasets from EU institutions value of a dataset. https://ec.europa.eu/isa2/sites/isa/files/publications/report-on-high-value-datasets-from-eu-institutions_en.pdf

Colombia Compra Eficiente, (2015). Colombia Compra Eficiente. Recuperado de https://www.colombiacompra.gov.co/colombia-compra/colombia-compra-eficiente.

Colombia Compra Eficiente, (2015). Datos abiertos. Recuperado de: https://www.colombiacompra.gov.co/transparencia/gestion-documental/datos-abiertos

Interfaz de Socrata (2018). Plataforma de datos abiertos del gobierno colombiano. Recuperado de: https://dev.socrata.com/foundry/www.datos.gov.co/aimg-uskh

Colombia Compra Eficiente, (2015). Manual para el uso de datos abiertos del SECO. Recuperado de https://www.colombiacompra.gov.co/sites/cce_public/files/cce_documentos/cce_manual_datos_abiertos_0.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones, Datos Abiertos (2018). Disponible en: https://www.mintic.gov.co/portal/604/w3-article-62310.html.

A.Vetrò, L. Canova, M. Torchiano, C. Orozco, R. Iemma y F. Morando, (2016). "Open data quality measurement framework: Definition and applicationto Open Government Data", Government Information Quarterly, pp.325–337.

HUANG, Zhexue. A fast clustering algorithm to cluster very large categorical data sets in data mining. DMKD, 1997, vol. 3, no 8, p. 34-39.

LABIOD, Lazhar; BENNANI, Younes. A Spectral Based Clustering Algorithm for Categorical Data with Maximum Modularity. En ESANN. 2011.

SALTOS, Giger. COCEA, Mihalea, (2017). An Exploration of Crime Prediction Using Data Mining on Open Data . International Journal of Information Technology & Decision Making. vol. 16, No. 05, pp. 1155-1181.

CHANDGUDE, Amar S., Kumar, Vijay. International journal of emerging technology and advanced engineering. Clustering Methods for Categorical and Numerical Dataset: A Recent Survey, 2015, vol. 5, no 6.