Entrega Final Proyecto de Grado

Resultados finales de los Requerimientos y Manual Técnico

Mariana Arias

Alejandro Murcia Pinilla

Juan Camilo Pérez

Johan Santacruz

Mayo 2023

Contenido

Tabla de Requerimientos y su diligenciamiento con Pruebas

Tabla de Requerimientos Generales

Aspecto	Requerimiento	Prueba prevista	Criterio o métrica de evaluación y rangos deseados	Estado
Negocio				
R1	Aumentar en un 20% la familiaridad sobre la información de obras públicas en el país al público general	Encuesta a usuarios antes de usar el tablero y después de usarlo	Familiaridad 2da encuesta >= Familiaridad 1er encuesta + 20%	Alcanzado
R2	Cargar la totalidad de la información de contratos de tipo "Obra"	Conteo de fuente de datos contra el datamart	El 100% de los contratos deben estar cargados en el datamart	Alcanzado
R3	Relacionar las multas y sanciones con sus contratos y proveedores correspondientes	Contar el de número de multas y sanciones contra el total de relaciones realizadas	Número de multas y sanciones disponibles = Número de relaciones realizadas con contratos y proveedores	Alcanzado
Desempeño				
R4	Desarrollar modelo predictivo que clasifique proyectos futuros como	Cross Validation	ROC >=0.65	Alcanzado con Modificaciones*

	"Exitoso" o "No exitoso"			
Funcional				
R5	Construir una visualización amigable y fácil de usar en desktop para los usuarios	Evaluación en escala de 0 a 5	Nota promedio >= 3.5	Alcanzado
R6	Procesar el dashboard rápidamente en todos sus tabs	Pruebas de performance	Tiempo de carga de todos los tabs <= 8 segundos	Alcanzado

- **Evidencias del cumplimiento del requerimiento "Aumentar en un 20% la familiaridad sobre la información de obras públicas en el país al público general"**

Después de realizar las encuestas a 10 usuarios diferentes se obtuvieron los siguientes resultados:

Usuario	Calificación antes de ver el dashboard	Calificación después de ver el dashboard	% Mejora
1	4/7	7/7	43%
2	4/7	5/7	15%
3	4/7	7/7	43%
4	1/7	7/7	85%
5	1/7	7/7	85%
6	1/7	7/7	85%
7	0/7	7/7	100%
8	3/7	7/7	43%
9	3/7	6/7	43%
10	3/7	6/7	43%

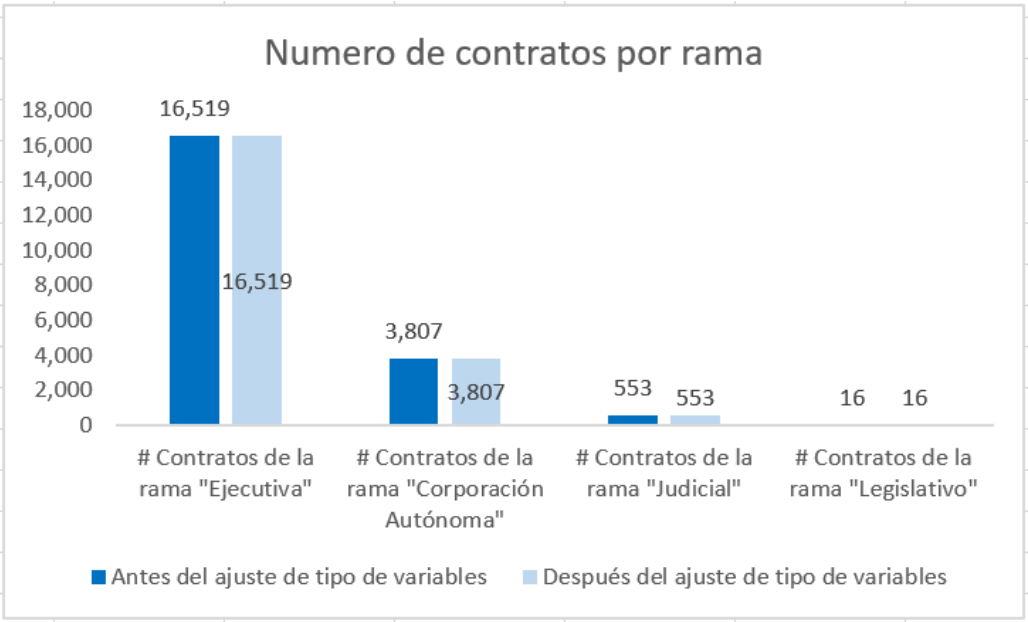
Con base en lo anterior, se puede confirmar que el público aumento en más del 20% su familiaridad sobre la información de la contratación de obras públicas en Colombia. En el siguiente link se encuentra el soporte de cada una de las respuestas de las encuestas realizadas:

- **Evidencias del cumplimiento de los requerimientos “Cargar la totalidad de la información de contratos de tipo "Obra" y “Relacionar las multas y sanciones con sus contratos y proveedores correspondientes “**

1. Contratos electrónicos: Al descargar los datos desde la página del SECOP, se pudo observar que las variables no tenían el tipo de variable correcto. Todas las variables eran de tipo texto a pesar de que algunas variables eran numéricas, fechas, etc. Por lo tanto, el primer paso fue convertir cada variable a su tipo de datos respectivo sin perder información. Para esto, se hicieron las siguientes validaciones:

Criterio	Antes de la conversión de las variables a su tipo de datos respectivo	Después de la conversión de las variables a su tipo de datos respectivo
# de filas	20.895	20.895
Valores únicos de la variable “Rama”	Ejecutivo, corporación autónoma, judicial, legislativo	Ejecutivo, corporación autónoma, judicial, legislativo
Cantidad de valores nulos	63.419	63.419

Adicionalmente, se contaron la cantidad de contratos que pertenecían a cada rama. En el siguiente gráfico se pueden ver los resultados:



2. Multas y sanciones**:** Al descargar los datos desde la página del SECOP, se pudo observar que las variables no tenían el tipo de variable correcto. Todas las variables tenían el formato de texto a pesar de tener variables numéricas y fechas. Por lo tanto, el primer paso fue convertir cada variable a su tipo de datos respectivo sin perder información. Para esto, se hicieron las siguientes validaciones:

Criterio	Antes de la conversión de las variables a su tipo de datos respectivo	Después de la conversión de las variables a su tipo de datos respectivo
# de filas	486	486
Valores únicos de la variable “Versión"	1,2,3,4,5,6,7,8,9,10,11,14,25	1,2,3,4,5,6,7,8,9,10,11,14,25
Cantidad de valores nulos	734	746

La cantidad de valores nulos es diferente ya que la variable “Aplico Garantías” tenía las opciones de “True”, “False” o “None”. Sin embargo, al ajustar los tipos de variables se agruparon los “None” y los “False”. Por lo tanto, antes del ajuste de los tipos de variables habían 10 nulos adicionales. Sin embargo, es importante mencionar que esta variable “Aplico Garantías” no se usa en nuestros análisis.

3. Procesos de contratación**:** Al descargar los datos desde la página del SECOP, se pudo observar que las variables no tenían el tipo de variable correcto. Todas las variables tenían el formato de texto a pesar de tener variables numéricas y fechas. Por lo tanto, el primer paso fue convertir cada variable a su tipo de datos respectivo sin perder información. Es importante mencionar, que nuestros análisis serán relacionados únicamente con los contratos de tipo obra, por lo que se debía verificar que solo teníamos contratos tipo “Obra”. Para esto, se hicieron las siguientes validaciones:

Criterio	Antes de la conversión de las variables a su tipo de datos respectivo	Después de la conversión de las variables a su tipo de datos respectivo
# de filas	40.917	40.917
Valores únicos de la variable “Tipo Contrato”	Obra	Obra
Valores únicos de la variable “Orden Entidad”	Nacional, Territorial, Corporación Autónoma	Nacional, Territorial, Corporación Autónoma
Cantidad de valores nulos	135.811	135.811

Después de tener todos los datos con el tipo de dato correcto, se construyeron 2 tablas fundamentales para nuestros análisis posteriores. La primera tabla es “MultaProcesoContrato” en donde se encuentran todos los contratos y procesos de tipo Obra que fueron multados. Por lo tanto, se unieron las 3 fuentes de datos mencionadas anteriormente. Para validar que no se haya perdido información en la unión de las diferentes fuentes de datos se tomaron los siguientes pasos:

1. Se encontraron los contratos que fueron multados (unión entre “Multas y sanciones” y “Contratos Electrónicos” a través de la variable “ID Contrato”)
2. Se encontraron los procesos que fueron multados (unión entre “Multas y sanciones” y “Procesos” a través de la variable “Referencia del proceso”)
3. Se encontraron los valores únicos de los contratos y procesos que fueron multados en los pasos anteriores.

4. Se verifica que la tabla “MultaProcesoContrato” tenga la misma cantidad de procesos y contratos únicos multados encontrados en el paso 3.

Criterio	**Fuente de datos Contratos Electrónicos**	**Fuente de datos Proceso**	**Nueva tabla llamada “MultaProcesoContrato”**
# Contratos multados	27	N/A	27
# Contratos multados	N/A	39	39

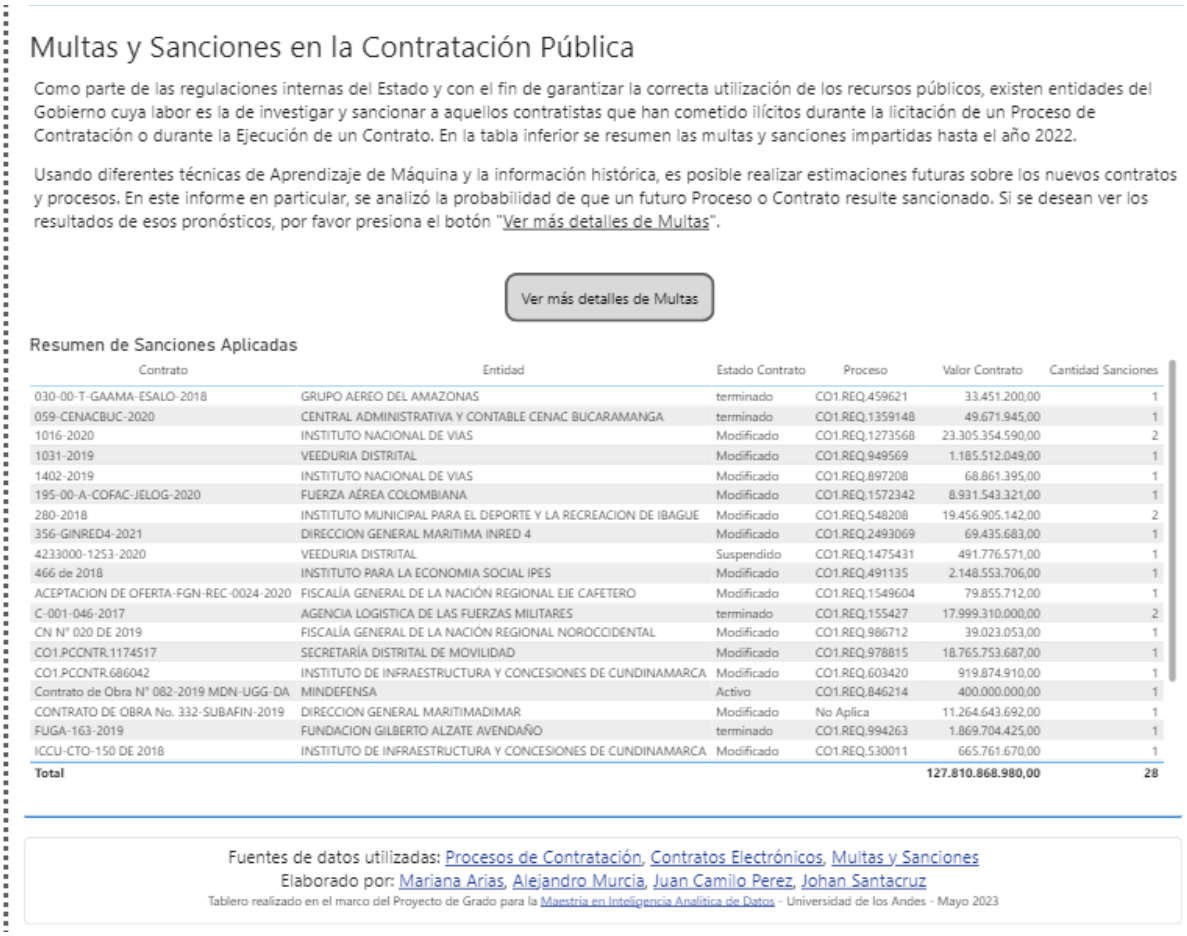
La segunda tabla fundamental para nuestros análisis es “ProcesoContrato” que es la unión entre las fuentes de datos “Procesos” y “Contratos electrónicos”. Para la construcción de esta tabla se tomaron todos los procesos de la fuente de datos “Procesos” y se buscó su respectivo contrato en la fuente de “Contratos electrónicos”. Esta tabla contiene incluso los procesos que no tienen un contrato respectivo o contratos que no tienen un proceso asociado. Por lo tanto, el número de procesos encontrados en la fuente de datos “Procesos” debe ser igual al número de procesos de la nueva tabla “ProcesoContrato”

Criterio	**Fuente de datos “Procesos”**	**Nueva tabla llamada “ProcesoContrato”**
# Procesos únicos	24.559	24.559

Finalmente, se construyó una sábana de datos uniendo las últimas dos tablas mencionadas anteriormente (MultaProcesoContrato y ProcesoContrato). Es decir que se construyó una tabla final llamada “SabanaAnalitica” que sería la unión de todas las fuentes de datos. El principal objetivo de esta tabla final es servir como entrada para el análisis de ML realizado con el fin de pronosticar futuros proyectos que potencialmente podrían ser sancionados, así como servir de fuente para las dimensiones *ProcesoContrato* y *HechoProcesoContrato* descritos anteriormente. Se hicieron las siguientes validaciones:

Criterio	**Tabla “ProcesoContrato”**	**Nueva tabla llamada “SabanaAnalitica”**
# Contratos únicos	20.892	20.892
# Procesos únicos	36.986	36.986

Adicionalmente, se puede observar que se relacionaron las multas y sanciones con sus contratos y proveedores correspondientes en el dashboard de Power BI:



Multas y Sanciones en la Contratación Pública

Como parte de las regulaciones internas del Estado y con el fin de garantizar la correcta utilización de los recursos públicos, existen entidades del Gobierno cuya labor es la de investigar y sancionar a aquellos contratistas que han cometido ilícitos durante la licitación de un Proceso de Contratación o durante la Ejecución de un Contrato. En la tabla inferior se resumen las multas y sanciones impartidas hasta el año 2022.

Usando diferentes técnicas de Aprendizaje de Máquina y la información histórica, es posible realizar estimaciones futuras sobre los nuevos contratos y procesos. En este informe en particular, se analizó la probabilidad de que un futuro Proceso o Contrato resulte sancionado. Si se desean ver los resultados de esos pronósticos, por favor presiona el botón "Ver más detalles de Multas".

Ver más detalles de Multas

Resumen de Sanciones Aplicadas

Contrato	Entidad	Estado Contrato	Proceso	Valor Contrato	Cantidad Sanciones
030-00-T-GAAMA-ESALO-2018	GRUPO AEREO DEL AMAZONAS	terminado	CO1.REQ.459621	33.451.200,00	1
059-CENACBUC-2020	CENTRAL ADMINISTRATIVA Y CONTABLE CENAC BUCARAMANGA	terminado	CO1.REQ.1359148	49.671.945,00	1
1016-2020	INSTITUTO NACIONAL DE VIAS	Modificado	CO1.REQ.1273568	23.305.354.590,00	2
1031-2019	VEEDURIA DISTRITAL	Modificado	CO1.REQ.949569	1.185.512.049,00	1
1402-2019	INSTITUTO NACIONAL DE VIAS	Modificado	CO1.REQ.897208	68.861.395,00	1
195-00-A-COFAC-JELOG-2020	FUERZA AÉREA COLOMBIANA	Modificado	CO1.REQ.1572342	8.931.543.321,00	1
280-2018	INSTITUTO MUNICIPAL PARA EL DEPORTE Y LA RECREACION DE IBAGUE	Modificado	CO1.REQ.548208	19.456.905.142,00	2
356-GINRED4-2021	DIRECCION GENERAL MARITIMA INRED 4	Modificado	CO1.REQ.2493069	69.435.683,00	1
4233000-1253-2020	VEEDURIA DISTRITAL	Suspendido	CO1.REQ.1475431	491.776.571,00	1
466 de 2018	INSTITUTO PARA LA ECONOMIA SOCIAL IPES	Modificado	CO1.REQ.491135	2.148.553.706,00	1
ACEPTACION DE OFERTA-FGN-REC-0024-2020	FISCALÍA GENERAL DE LA NACIÓN REGIONAL EJE CAFETERO	Modificado	CO1.REQ.1549604	79.855.712,00	1
C-001-046-2017	AGENCIA LOGISTICA DE LAS FUERZAS MILITARES	terminado	CO1.REQ.155427	17.999.310.000,00	2
CN N° 020 DE 2019	FISCALÍA GENERAL DE LA NACIÓN REGIONAL NOROCCIDENTAL	Modificado	CO1.REQ.986712	39.023.053,00	1
CO1.PCCNTR.1174517	SECRETARÍA DISTRITAL DE MOVILIDAD	Modificado	CO1.REQ.978815	18.765.753.687,00	1
CO1.PCCNTR.686042	INSTITUTO DE INFRAESTRUCTURA Y CONCESIONES DE CUNDINAMARCA	Modificado	CO1.REQ.603420	919.874.910,00	1
Contrato de Obra N° 082-2019 MDN-UGG-DA	MINDEFENSA	Activo	CO1.REQ.846214	400.000.000,00	1
CONTRATO DE OBRA No. 332-SUBAFIN-2019	DIRECCION GENERAL MARITIMADIMAR	Modificado	No Aplica	11.264.643.692,00	1
FUGA-163-2019	FUNDACION GILBERTO ALZATE AVENDAÑO	terminado	CO1.REQ.994263	1.869.704.425,00	1
ICCU-CTO-150 DE 2018	INSTITUTO DE INFRAESTRUCTURA Y CONCESIONES DE CUNDINAMARCA	Modificado	CO1.REQ.530011	665.761.670,00	1
Total				**127.810.868.980,00**	**28**

Fuentes de datos utilizadas: Procesos de Contratación, Contratos Electrónicos, Multas y Sanciones
Elaborado por: Mariana Arias, Alejandro Murcia, Juan Camilo Perez, Johan Santacruz
Tablero realizado en el marco del Proyecto de Grado para la Maestria en Inteligencia Analitica de Datos - Universidad de los Andes - Mayo 2023

- **Evidencias del cumplimiento del requerimiento "Desarrollar modelo predictivo que clasifique proyectos futuros como "Exitoso" o "No exitoso"**

Con base en el modelo predictivo de Random Forest que se desarrolló para clasificar los proyectos futuros como exitosos o no exitosos, se obtuvieron los siguientes resultados:

Predicciones de Sanciones para Procesos y Contratos (Desde 2023)

ReferenciaContrato	Cod Contratista	Entidad	Departamento	Estado Contrato	Referencia Proceso	Estimación
	No Adjudicado	ARMADA NACIONAL BASE NAVAL ARC MALAGA	VALLE DEL CAUCA		CO1.REQ.4317610	Riesgo Bajo
	No Adjudicado	BN4	San Andrés, Providencia y Santa Catalina		CO1.REQ.4355760	Riesgo Bajo
	No Adjudicado	GOBERNACION DEL MAGDALENA	MAGDALENA		CO1.REQ.4358980	Riesgo Bajo
	No Adjudicado	POLICIA METROPOLITANA SANTIAGO DE CALI	No Definido		CO1.REQ.4360706	Riesgo Bajo
	No Adjudicado	DIRECCIÓN DE INTELIGENCIA POLICIAL	DISTRITO CAPITAL DE BOGOTÁ		CO1.REQ.4380836	Riesgo Bajo
	No Adjudicado	GOBERNACION DEL DEPARTAMENTO DEL CESAR	No Definido		CO1.REQ.4403366	Riesgo Medio
191SO2023	713220051	MUNICIPIO DE MARINILLA	No Definido	En ejecución	CO1.REQ.4166272	Riesgo Bajo
192SO2023	713220051	MUNICIPIO DE MARINILLA	No Definido	En ejecución	CO1.REQ.4166386	Riesgo Bajo

En donde se puede observar que algunos proyectos tienen una probabilidad baja, media o alta de ser multados (es decir, ser No exitosos). Además, las métricas de desempeño para elegir entre diferentes modelos seleccionada fue el *recall* en donde nuestro modelo de Random Forest elegido tiene una media de 0.848. Escogimos esa métrica ya que es la que permite seleccionar la mayor cantidad de verdaderos positivos, sin importar tanto que sean seleccionados algunos falsos positivos. En resumen, usando como input de entrenamiento los 50 contratos/procesos con multas más la totalidad de los que no recibieron ninguna sanción hasta 2022, el algoritmo fue capaz de clasificar 8 contratos/procesos futuros (2023 en adelante) con probabilidades Medias y Bajas de recibir alguna sanción.

Indice Modelo	mean_test_accuracy	std_test_accuracy	mean_test_precision	std_test_precision	mean_test_recall	std_test_recall
RFC_OS_8	0.894	0.065	0.654	0.458	0.848	0.143
RFC_OS_6	0.992	0.008	0.684	0.442	0.843	0.221
RFC_OS_12	0.992	0.008	0.684	0.443	0.843	0.221
RFC_OS_3	0.889	0.066	0.654	0.458	0.843	0.141
RFC_OS_5	0.889	0.066	0.654	0.458	0.843	0.141

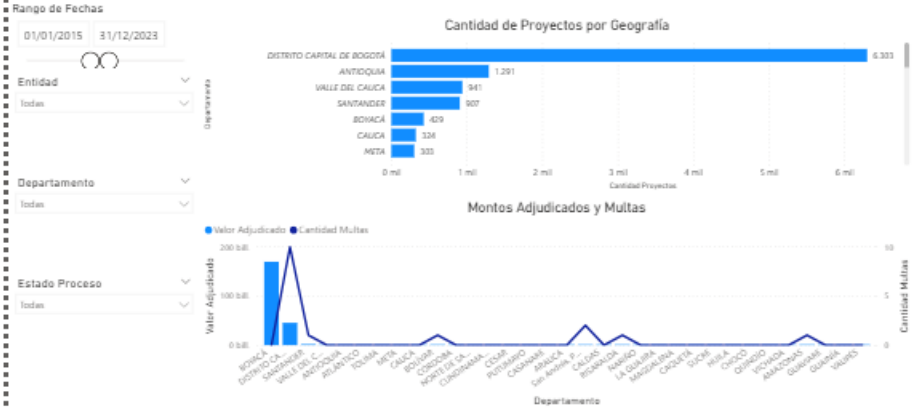
- **Evidencias del cumplimiento del requerimiento "Construir una visualización amigable y fácil de usar en desktop para los usuarios"**

Se construyó una visualización en Power BI para mostrar los resultados de las estimaciones de nuestro modelo predictivo y también, para mostrar las características de los proyectos antiguos (es decir la base de nuestro modelo predictivo)

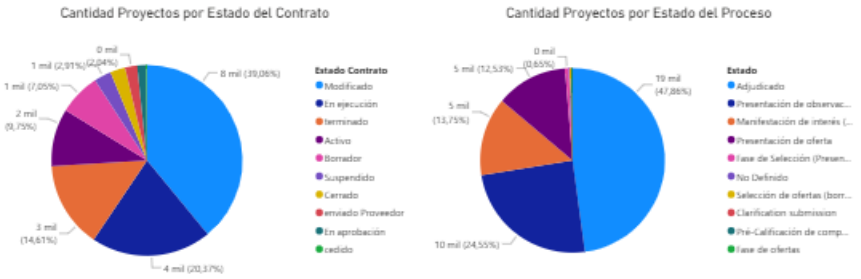
Contratación Pública en Colombia - Del Proceso al Contrato

Esta es una página interactiva que utiliza datos abiertos del estado Colombiano relacionados con el Secop II. Contiene información de Procesos de Contratación, Contratos Electrónicos y Multas y Sanciones. Es una herramienta construida con el propósito de facilitar el acceso a esta información pensando en el público general.

Este documento es de fácil acceso y uso tanto para los ciudadanos como otras partes interesadas. Con este pueden obtener una mejor comprensión de cómo se están utilizando los recursos públicos y cómo se están aplicando las leyes y regulaciones. Comenzaremos con una validación de la cantidad de proyectos distribuidos geográficamente, así como de los montos que han sido adjudicados y la cantidad de sanciones que se han creado según los filtros seleccionados:



Mediante un Proceso de Contratación, una Entidad Estatal puede realizar la selección de sus contratistas para ejecutar obras. En algunos casos, estos Procesos no son necesarios y las Entidades pueden contratar directamente. De cualquier forma, para entablar una relación entre una Entidad y un Contratista, se requiere de un Contrato. Es por ello que se pueden analizar de manera independiente tanto Procesos como Contratos:



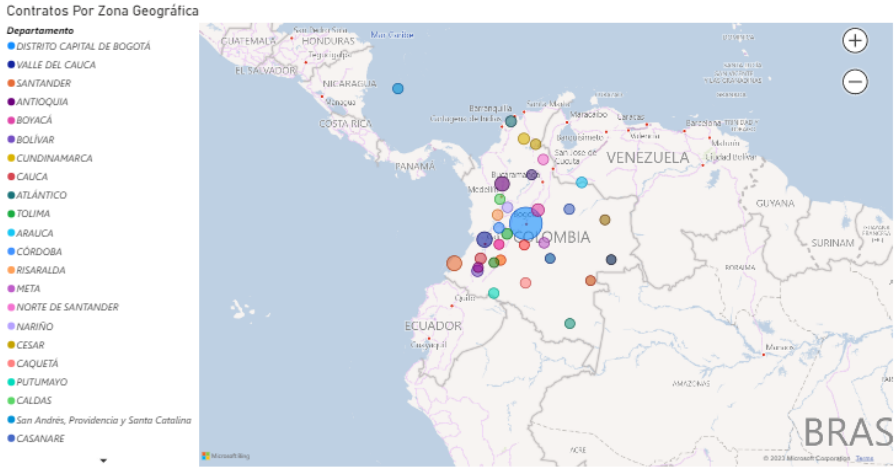
Si filtramos *Proyectos por Estado del Contrato*, notaremos que en la gráfica de *Proyectos por Estado del Proceso* solo se resalta la sección Adjudicado. Esto refleja el flujo natural de la selección de un Contratista. Sin embargo, al hacer lo opuesto (filtrar por *Estado Proceso* como Adjudicado) notaremos que no se seleccionan la totalidad de Contratos. Esto se debe a que no todos los contratos nacen desde un Proceso.

Contratación Pública en Colombia - Un repaso general

Utilizando los filtros superiores como base, podemos obtener algunas cifras relevantes en el análisis de la información de contratación de obras públicas del país. Por ejemplo, sabemos que en el rango de tiempo por defecto de este informe, han sido impactados los 32 Departamentos del país (con la adición de Bogotá). Tambien, se han ejecutado obras en más de 270 municipios y estas han sido ejecutadas por casi 10 mil contratistas. En el mapa inferior podemos observar como ha sido esta distribución a nivel de Departamento y Ciudad, en donde el radio de cada círculo es proporcional a la cantidad de proyectos ejecutados:

9.910	274	33
Cantidad de Contratistas	Ciudades Impactadas por Obras	Departamentos impactados por Obras

Contratos Por Zona Geográfica



Multas y Sanciones en la Contratación Pública

Como parte de las regulaciones internas del Estado y con el fin de garantizar la correcta utilización de los recursos públicos, existen entidades del Gobierno cuya labor es la de investigar y sancionar a aquellos contratistas que han cometido ilícitos durante la licitación de un Proceso de Contratación o durante la Ejecución de un Contrato. En la tabla inferior se resumen las multas y sanciones impartidas hasta el año 2022.

Usando diferentes técnicas de Aprendizaje de Máquina y la información histórica, es posible realizar estimaciones futuras sobre los nuevos contratos y procesos. En este informe en particular, se analizó la probabilidad de que un futuro Proceso o Contrato resulte sancionado. Si se desean ver los resultados de esos pronósticos, por favor presiona el botón "Ver más detalles de Multas".

Ver más detalles de Multas

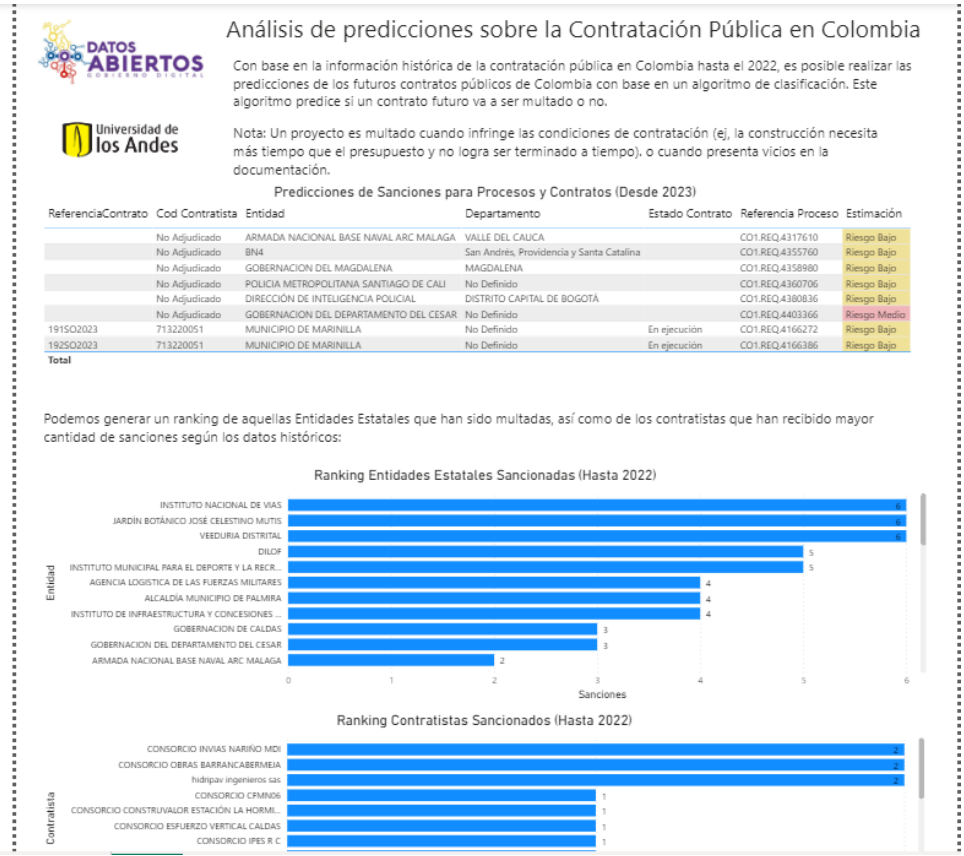
Resumen de Sanciones Aplicadas

Contrato	Entidad	Estado Contrato	Proceso	Valor Contrato	Cantidad Sanciones
030-00-T-GAAMA-ESALO-2018	GRUPO AEREO DEL AMAZONAS	terminado	CO1.REQ.459621	33.451.200,00	1
059-CENACBUC-2020	CENTRAL ADMINISTRATIVA Y CONTABLE CENAC BUCARAMANGA	terminado	CO1.REQ.1359148	49.671.945,00	1
1016-2020	INSTITUTO NACIONAL DE VIAS	Modificado	CO1.REQ.1273568	23.305.354.590,00	2
1031-2019	VEEDURIA DISTRITAL	Modificado	CO1.REQ.949569	1.185.512.049,00	1
1402-2019	INSTITUTO NACIONAL DE VIAS	Modificado	CO1.REQ.897208	68.861.395,00	1
195-00-A-COFAC-JELOG-2020	FUERZA AÉREA COLOMBIANA	Modificado	CO1.REQ.1572342	8.931.543.321,00	1
280-2018	INSTITUTO MUNICIPAL PARA EL DEPORTE Y LA RECREACION DE IBAGUE	Modificado	CO1.REQ.548208	19.456.905.142,00	2
356-GINRED4-2021	DIRECCION GENERAL MARITIMA INRED 4	Modificado	CO1.REQ.2493069	69.435.683,00	1
4233000-1253-2020	VEEDURIA DISTRITAL	Suspendido	CO1.REQ.1475431	491.776.571,00	1
466 de 2018	INSTITUTO PARA LA ECONOMIA SOCIAL IPES	Modificado	CO1.REQ.491135	2.148.553.706,00	1
ACEPTACION DE OFERTA-FGN-REC-0024-2020	FISCALÍA GENERAL DE LA NACIÓN REGIONAL EJE CAFETERO	Modificado	CO1.REQ.1549604	79.855.712,00	1
C-001-046-2017	AGENCIA LOGISTICA DE LAS FUERZAS MILITARES	terminado	CO1.REQ.155427	17.999.310.000,00	2
CN N° 020 DE 2019	FISCALÍA GENERAL DE LA NACIÓN REGIONAL NOROCCIDENTAL	Modificado	CO1.REQ.986712	39.023.053,00	1
CO1.PCCNTR.1174517	SECRETARÍA DISTRITAL DE MOVILIDAD	Modificado	CO1.REQ.978815	18.765.753.687,00	1
CO1.PCCNTR.686042	INSTITUTO DE INFRAESTRUCTURA Y CONCESIONES DE CUNDINAMARCA	Modificado	CO1.REQ.603420	919.874.910,00	1
Contrato de Obra N° 082-2019 MDN-UGG-DA	MINDEFENSA	Activo	CO1.REQ.846214	400.000.000,00	1
CONTRATO DE OBRA No. 332-SUBAFIN-2019	DIRECCION GENERAL MARITIMADIMAR	Modificado	No Aplica	11.264.643.692,00	1
FUGA-163-2019	FUNDACION GILBERTO ALZATE AVENDAÑO	terminado	CO1.REQ.994263	1.869.704.425,00	1
ICCU-CTO-150 DE 2018	INSTITUTO DE INFRAESTRUCTURA Y CONCESIONES DE CUNDINAMARCA	Modificado	CO1.REQ.530011	665.761.670,00	1
Total				**127.810.868.980,00**	**28**

Fuentes de datos utilizadas: Procesos de Contratación, Contratos Electrónicos, Multas y Sanciones

Elaborado por: Mariana Arias, Alejandro Murcia, Juan Camilo Perez, Johan Santacruz

Tablero realizado en el marco del Proyecto de Grado para la Maestría en Inteligencia Analítica de Datos - Universidad de los Andes - Mayo 2023

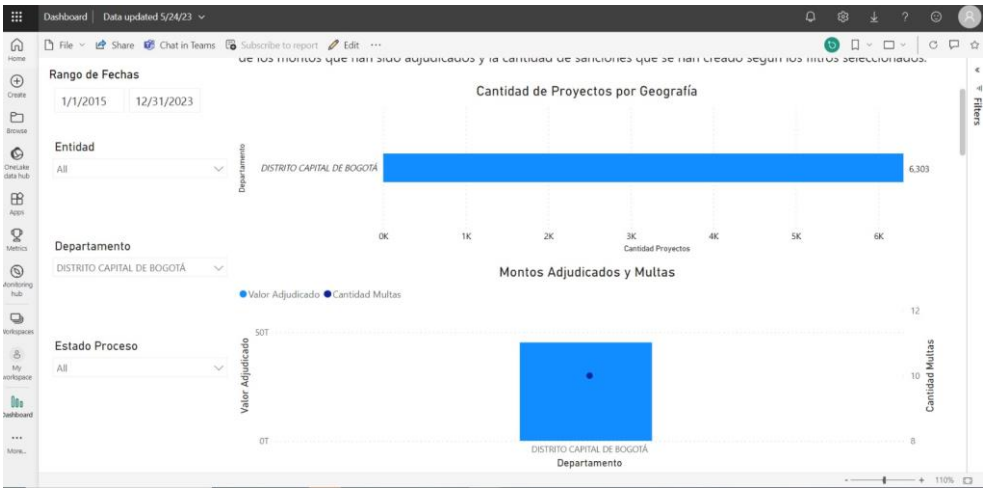


Por otro lado, también le pedimos al público que utilizó el dashboard que calificara de 1 a 5 la facilidad de uso del tablero (en donde el 1 representaba que el tablero fue muy difícil de usar, y 5 que el tablero fue muy fácil de usar). De las 10 personas encuestadas, 7 de ellas le dieron una nota de 5/5 y las otras 3 personas restantes le dieron una nota de 4/5. Por lo tanto, el promedio de la facilidad de uso del tablero de Power BI es de 4.7/5.

- **Evidencias del cumplimiento del requerimiento "Procesar el dashboard rápidamente en todos sus tabs"**

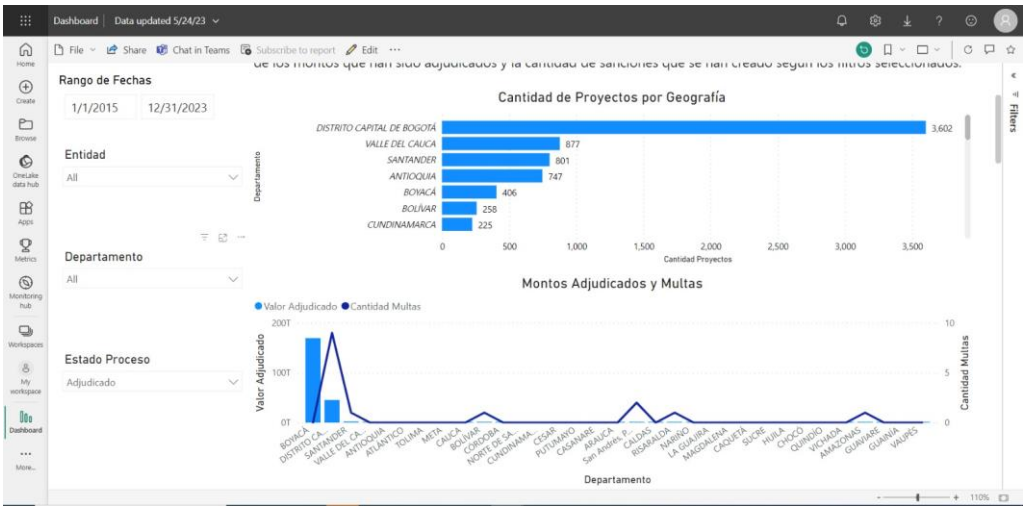
Tiempos de espera al utilizar el tab "Departamento"

Intento	Tiempo de espera (seg)
1	0.14
2	0.24
3	0.11



Tiempos de espera al utilizar el tab “Estado Proceso”

Intento	Tiempo de espera (seg)
1	0.08
2	0.17
3	0.19



Código Fuente del proyecto

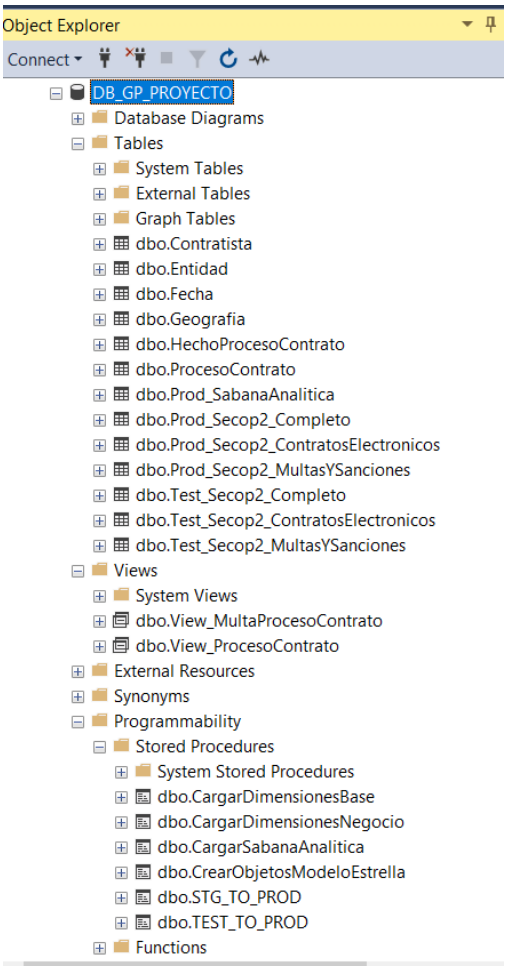
Todos los archivos utilizados para la elaboración de este proyecto están referenciados en el siguiente repositorio: https://github.com/jsantacruzl/ProyectoDeGradoMIAD_o3b

El lector se puede apoyar del readme del repositorio, así como de los manuales del presente documento para obtener una comprensión detallada del proyecto mismo.

Manuales Técnicos y de Usuario General

a. ETL (Técnico)

- Creación de la base de datos por primera vez:
  - Creación de la base de datos en SQL Server. Se puede realizar usando Azure. Tener en cuenta la configuración de permisos, el usuario y la contraseña.
  - Creación de las tablas físicas en la base de datos (Objetos en el repositorio, folder SQL). Las siguientes tablas deben ser creadas usando los scripts del repositorio. Estas son:
    - Test_Secop2_Completo
    - Test_Secop2_ContratosElectronicos
    - Test_Secop2_MultasYSanciones
    - Prod_SabanaAnalitica
    - Prod_Secop2_Completo
    - Prod_Secop2_ContratosElectronicos
    - Prod_Secop2_MultasYSanciones
  - Creación de los StoredProcedures que gestionan la carga de datos de STG a PROD, así como crean los objetos del modelo en estrella:
    - STG_TO_PROD
    - TEST_TO_PROD
    - CrearObjetosModeloEstrella
    - CargarDimensionesBase
    - CargarSabanaAnalitica
    - CargarDimensionesNegocio
  - Creación de las vistas de negocio, que son usadas para unir los datos del negocio en entidades con la granularidad deseada en el modelo en estrella
    - View_MultaProcesoContrato
    - View_ProcesoContrato
  - Ejecutar el StoredProcedure STG_TO_PROD con el parámetro @TipoCarga = 1 creará el resto de los objetos restantes, así como realizará la carga de datos tanto de datos de negocio como de dimensiones fijas (como la dimensión de Fecha que carga todos los registros diarios desde 1985 hasta 2057 y la de Geografia que carga los más de 1100 municipios de Colombia).
  - Al final, si se han seguido los pasos correctamente, las tablas, vistas y StoredProcedures dentro de la base de datos deben lucir de la siguiente forma:



- Carga de los datos desde SECOP usando Socrata. Carga incremental.
    - Se deben ejecutar manualmente los Notebooks de cada una de las entidades de negocio (Repositorio, folder ETL). Tener cuidado de actualizar el server, el usuario, el password y la DB de forma adecuada. **Hint de trabajo futuro**: Encapsular estos notebooks en scripts de Python y programarlos con el administrador de tareas de Windows (a manera de ejemplo).

ETL	Links
Contratos Electrónicos	Link a repositorio
Procesos de Contratación	Link a repositorio
Multas y Sanciones	Link a repositorio

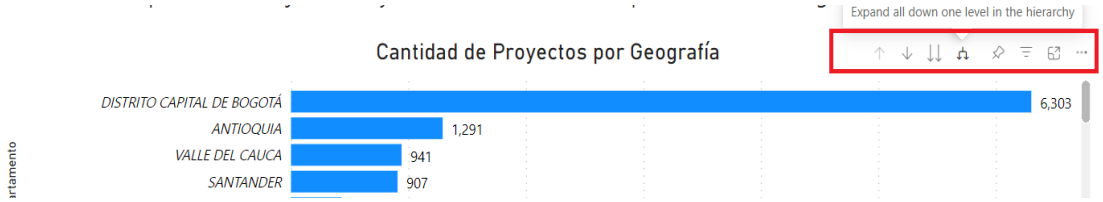
    - Ejecutar el StoredProcedure STG_TO_PROD con el parámetro @TipoCarga = 2. Esto hará que solamente se recarguen los datos que vienen directamente desde el SECOP II en el modelo en estrella, utilizando todas las tablas ya existentes. **Hint de trabajo futuro:** Enlazar el job descrito en el punto anterior con un Job de SQL Server que realice este segundo paso. De esta forma el pipeline quedará completamente automatizado.

b. Power BI (Usuario General)

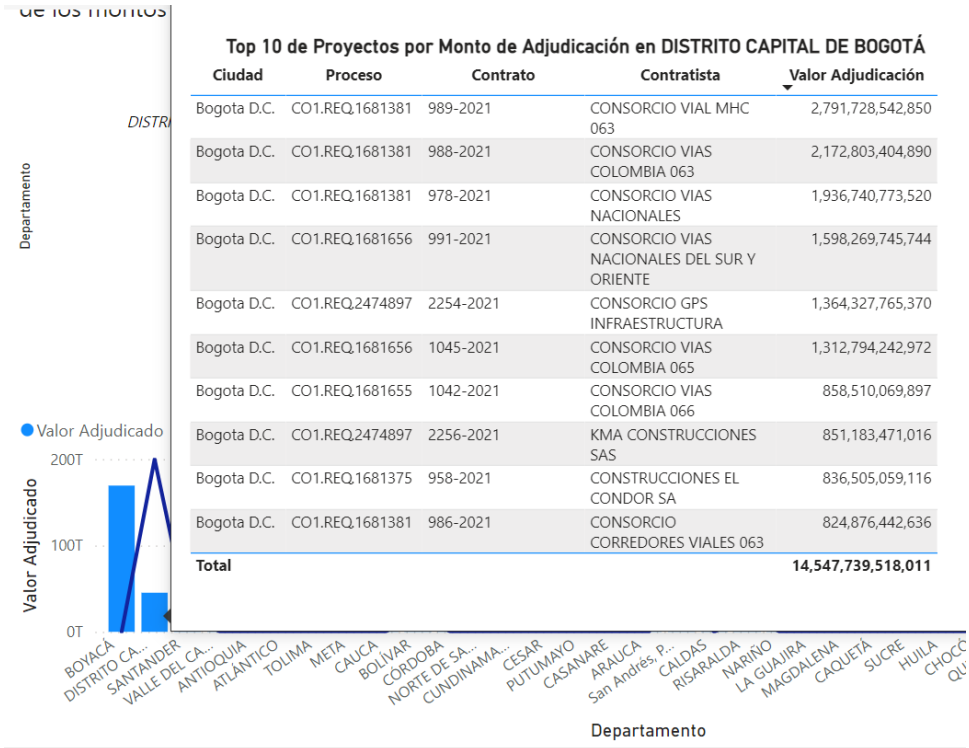
La última versión del tablero puede encontrarse en el folder Dashboard del respositorio. Desde ahí, un usuario puede descargarlo y realizar modificaciones usando Power BI Desktop (va a necesitar tener acceso a la base de datos de Azure, ya que esa es la fuente de datos). Desde ahí, puede desplegarlo en el servicio de https://app.powerbi.com/.

Asumiendo que un usuario final esté haciendo uso del tablero desde el servicio web, se adjuntan algunos comentarios sobre su funcionamiento:

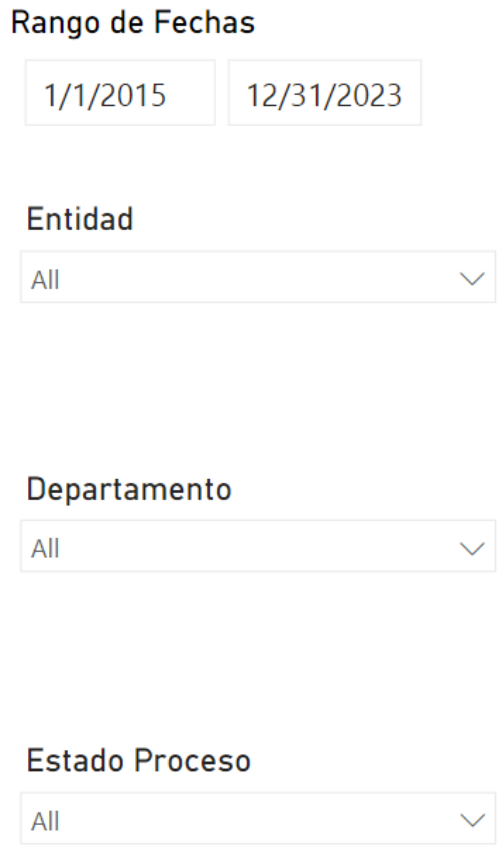
- En la mayoría de los gráficos, se tiene la funcionalidad de hacer drill up/drill down para obtener más detalles. Es de especial utilidad en donde se utilicen datos Geográficos, ya que permite obtener detalles por Departamento y Ciudad.



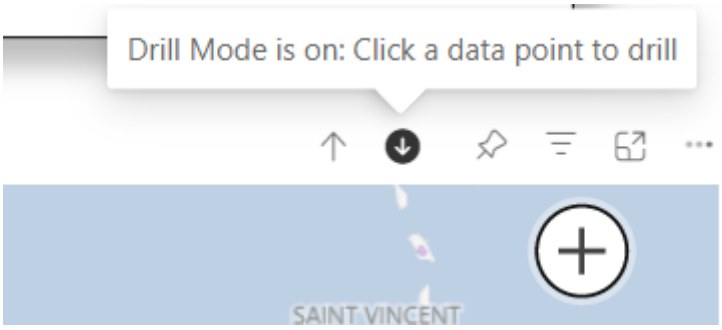
- En el gráfico de Montos Adjudicados y Multas, se cuenta con la opción de Tooltip para ver más detalles:



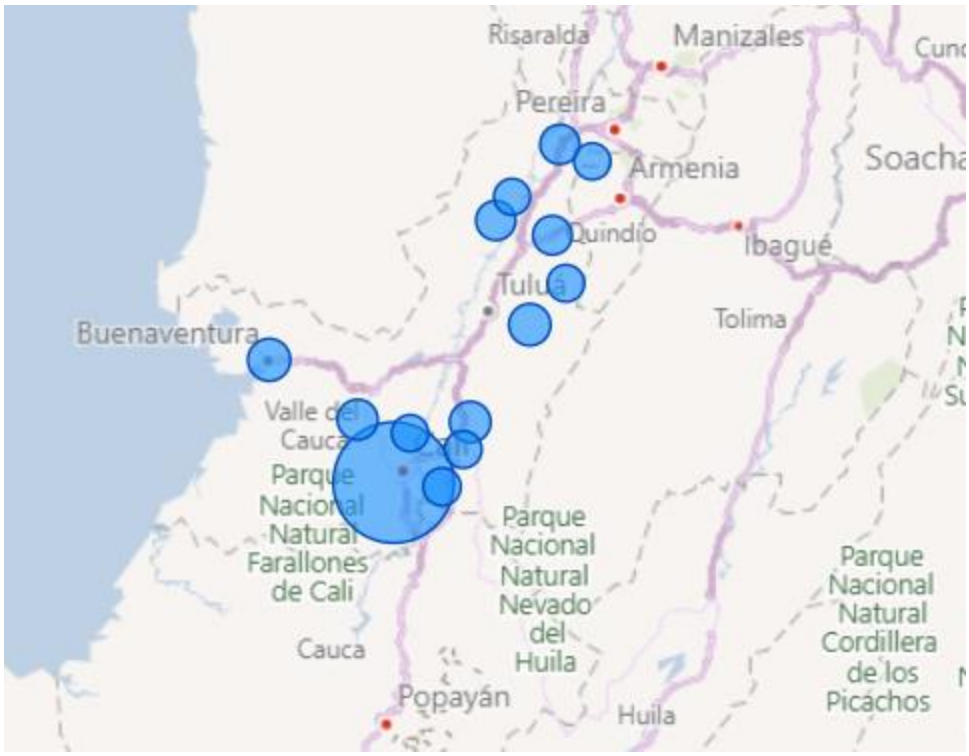
- Los filtros ubicados en la parte superior izquierda del tablero filtrarán la totalidad de los gráficos:



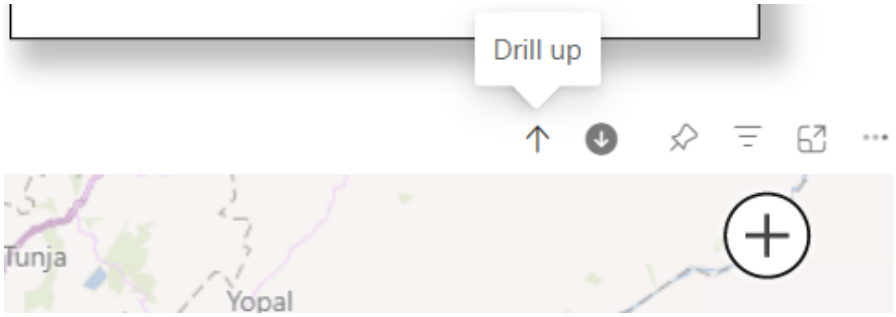
- Algunos gráficos tienen la habilidad de aplicar filtros sobre otros gráficos. El ejemplo más claro de ello es los dos pie-chart que muestran la relación entre Procesos de Contratación y Contratos.
- Para analizar los datos con el mapa, se recomienda el siguiente proceso:
    - Se da clic en la segunda flecha de tal forma que quede resaltada como se muestra en la imagen:



    - Se da clic en el círculo del Departamento del cual se quieren ver sus detalles por municipio:

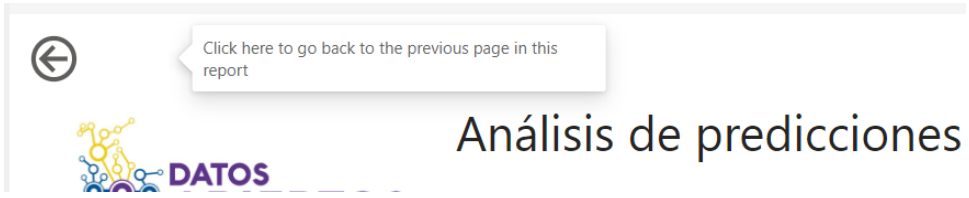


- Si se quiere volver a la vista inicial o nacional, se da clic en la primera flecha o drill up:



- Se puede hacer zoom in o zoom out con la rueda del mouse

- Si se da clic en el botón de “Ver más detalles de Multas” se irá a una página alternativa en donde se podrán ver, entre otras cosas, los pronósticos realizados a futuros contratos/procesos. Para volver a la página principal, se debe subir hasta el tope de la página secundaria y darle clic a la flecha que lo indica:



- Se pueden obtener más detalles de las fuentes usadas, así como del equipo de trabajo siguiendo los datos del footer de la página principal.

c. Modelos de ML (Técnico)

Todos los archivos referentes al modelo y su despliegue se encuentran en el Git del proyecto, el cual se puede encontrar accediendo en este link. Lo primero a mencionar es el Script completo del modelo a ejecutar, el cual se encuentra con el nombre de *GP_Proyecto_Despliegue_Modelo_Entrega.ipynb,* y en este se encuentra todo el proceso desde la lectura de los datos directamente en la base de datos dispuesta para el proyecto, la transformación de la data, la predicción y cargue final de dichas predicciones nuevamente sobre la base de datos final a usar en el reporte para el usuario.

A continuación, se presenta el paso enumerado del ejecutable que se encuentra dentro del Script mencionado;

1. Lo primero es el cargue de la sabana de datos completa, lo cual se hace por medio de Python y la conexión directa a la BD de Azure por medio de la librería de pyodbc. Toda esta data es almacenada sobre un DataFrame de pandas y dispuesta para ser procesada y desplegada por el modelo de Random Forest.
2. Se hacen unas pequeñas transformaciones finales de la data y la selección de determinadas columnas para ser implementadas en el modelo analítico.
3. Se parte la data tomando los proyectos de 2023 en adelante, los cuales van a ser los usados para predecir la probabilidad de que estos sean multados o no en el futuro.
4. Se hace uso del One Hot Encoder almacenado como *encod_oh.joblib* dentro del Git dispuesto para el proyecto. Este encoder se carga dentro del Script de Python y se utiliza para transformar las variables categóricas de los proyectos sobre los cuales se interesa generar las predicciones con el modelo de ML.
5. Se usa un KNNImputer para generar las imputaciones de los datos faltantes, el cual fue entrenado con los datos de los proyectos previstos a 2023 y con parámetros de 5 vecinos y pesos uniformes. Este modelo se encuentra almacenado en el git con el nombre de *imputer_ok.joblib* y se usa para imputar los datos faltantes de los proyectos del 2023 en adelante sean uniformes. Este modelo se encuentra almacenado en el git con el nombre
6. Estos datos ya completos se re-escalan haciendo uso de un modelo de escalado estándar, el cual fue previamente entrenado y almacenado dentro del Git del proyecto con el nombre de *scaler_ok.joblib.* Este escalador se carga en el Python y se utiliza para transformar los datos de los proyectos del 2023 en adelante.
7. Se había entrenado un modelo de PCA, con el cuál se buscaba obtener un 95% de la variabilidad de todos los datos, y luego fue almacenado dentro del Git del proyecto con el nombre de *pca_ok.joblib.* Este modelo fue cargado en el Python y fue usado para hacer las transformaciones de los proyectos de 2023 en adelante logrando obtener una reducción de dimensionalidad considerable y que quede de la misma forma en que recibe los datos el modelo predictor de Random Forest.
8. Como se había estado mencionando en entregas anteriores, el mejor modelo seleccionado luego de la experimentación fue el de Random Forest con parámetros de 100 estimadores, máxima profundidad de 5 y mínimo de 2 proyectos para cada partición de una rama. Este modelo entrenado con los proyectos previos a 2023 fue almacenado y dispuesto dentro del Git que se ha mencionado constantemente y se encuentra con el nombre de rfc_2022_ok.joblib, con el cual se le pasan los datos que

veníamos transformando hasta ahora y a la salida nos entrega las probabilidades de que cada uno de los proyectos de 2023 en adelante fueran multados en el futuro.

9. Estas probabilidades son clasificadas haciendo uso de las reglas de semáforo que se han definido (sin riesgo de ser multado, riesgo bajo y riesgo medio), pero se agrega la funcionalidad dentro del código de poder variar estos límites de acuerdo a la necesidad del momento.
10. Estos datos son almacenados dentro de la base de datos de Azure, los cuales ya quedan disponibles para ser implementados dentro del panel funcional.

Es importante mencionar que no se necesita realizar ninguna modificación sobre el script de ejecución para las predicciones. De esta forma, también se menciona que la actualización de las predicciones de los proyectos sólo requiere de hacer uso de la función de “Ejecutar todo” en el Script final.