De Problema de Negocio a Problema de Analytics

Parte I – Del Qué de problema de negocio al Cómo del problema de Analytics

Fase A

1. ¿Qué resultados esperan obtener tras la implementación del proyecto?

A nivel de resultado general, esperamos contarcon un tablero interactivo que permita al usuario conocer cuáles han sido las características más relevantes de la contratación en infraestructura del país y que por medio la aplicación de algoritmos de ML a unos atributos dados permita estimar que tan exitoso puede ser un hipotético proyecto futuro, entendido como la probabilidad de que el proyecto llegue a finalizarse.

Con esta herramienta, esperamos los siguientes resultados específicos:

- Facilitar a los entes de control el seguimiento y vigilancia de las contrataciones gubernamentales de proyectos de infraestructura pública.
- Permitir a cualquier usuario potencial de la herramienta poder realizar revisiones sobre los procesos de contratación y validar la transparencia de estos.
- Disminuir el malestar social causado por el desconocimiento de las inversiones en infraestructura estatal.
- Disminuir la sensación de corrupción en el establecimiento de nuevos contratos de infraestructura, fomentando la investigación y evaluación de la ciudadanía de estos.
- Poder detectar de manera gráfica, fácil e intuitiva, comportamientos como concentración de contratos a un mismo proveedor, modalidades de selección, valores adjudicados, contratos finalizados exitosamente, entre otros.
- 2. ¿Quiénes participaran en la solución analítica?

Los integrantes identificados hasta el momento de la solución analítica son:

- Equipo de desarrollo técnico y de gestión del proyecto, conformado por los miembros del equipo que elaboran el presente documento.
- Equipo de supervisión directo, conformado por el equipo docente de la Universidad de los Andes
- Stakeholders primarios, conformado por expertos que apoyan al equipo docente en la viabilidad/factibilidad de las propuestas presentadas por el equipo de desarrollo técnico.
- 3. ¿Qué cambios se generarían en la organización/unidad cliente como resultado del proyecto?

Se esperan dos grandes cambios dentro de la organización del cliente, que hemos definido como la sociedad colombiana y parte del gobierno:

- La ciudadanía contará con una herramienta para realizar seguimiento y control de las inversiones de infraestructura que será fácil de utilizar. Aunque actualmente existen iniciativas como *InfraestructuraVisible*, la idea del presente proyecto es ir más allá de una interfaz consolidada de datos y mostrar de una manera amigable, cifras concretas y filtros específicos que le permitan a la ciudadanía ejercer un control más eficiente de los recursos públicos.
- Los entes de control gubernamentales tendrán una herramienta adicional para realizar validaciones, especialmente para apoyarse en la evaluación de hipotéticos proyectos futuros. Sin embargo, la forma en la que los procedimientos internos de dichas entidades van a ser modificados es un área gris para el equipo del proyecto. No es seguro siquiera que esto ocurra. Aunque la posibilidad de que ocurra es latente.

Fase B (Lo que el cliente quiere)

Luego de realizar el análisis de Kano sobre el conjunto de features definido y al validar internamente en el equipo de trabajo (únicos SH alcanzables al momento de elaborar este documento), el resultado final obtenido fue el siguiente:

Feature	Must-be	Performance	Attractive	Indifferent	Reverse	Questionable	Category
Feature 1: Datos descriptivos actualizados diariamente	0%	0%	75%	25%	0%	0%	Attractive
Feature 2: Mobile Responsive	0%	0%	25%	75%	0%	0%	Indifferent
Feature 3: Filtros predefinidos	0%	75%	0%	25%	0%	0%	Performance
Feature 4: Tiempo de carga inicial menor a 10 segundos	75%	0%	0%	25%	0%	0%	Must-be
Feature 5: Explorar datos usando mapas	25%	50%	25%	0%	0%	0%	Performance
Feature 6: Sugerir si un nuevo proyecto será o no exitoso dadas algunos atributos del mismo	0%	50%	0%	25%	0%	25%	Performance

- Los requerimientos básicos u obligatorios que se atienden con el proyecto están relacionados con los tiempos de carga inicial menor a 10 segundos. Los clientes desean realizar análisis sobre la información existente de una manera rápida y efectiva.
- El requerimiento indiferente está relacionado con el diseño *mobile responsive*. La versión desktop es suficiente para el general de los SH y la versión móvil no aporta gran valor al proyecto.
- Este proyecto no cuenta, de momento, con requerimientos inversos.
- Se cuenta con 3 requerimientos unidimensionales o de performance: Filtros predefinidos en el tablero, exploración de los datos usando mapas y sugerir si un nuevo proyecto será o no exitoso dados algunos atributos de este.
- La periodicidad diaria en la actualización de los datos descriptivos es considerada como atractiva para los SH.

Fase C (Requerimientos Finales)

Los requerimientos/features finales, de acuerdo con su clasificación (importancia) son los siguientes:

ID/Orde	Requerimiento	Categoría
n		
1	Tiempo de carga inicial menor a 10 segundos	Must-be
2	Filtros predefinidos	Performance
3	Explorar datos usando mapas	Performance
4	Sugerir si un nuevo proyecto será o no exitoso dadas algunos atributos de este	Performance
5	Datos descriptivos actualizados diariamente	Atractive
6	Mobile Responsive	Indiferente

Para solucionar estos requerimientos, se debe involucrar a las siguientes áreas/SH dentro del proyecto:

Interesados	Importancia
Ciudadanos	Se debe validar con la ciudadanía en general, por medio de encuestas
	representativas, cuestiones como su estado actual de conocimiento sobre la contratación pública y deseos más relevantes de características o preguntas que
	deseen resolver por medio de la herramienta computacional.
ANCP/Entes de	Confirmar que los resultados de este proyecto faciliten su labor de control.
Control	

Equipo del	Encargados de transformar el problema de negocio en un problema de analytics
proyecto	e implementar la solución.
Equipo Docente	Validar que desde el punto de vista formativo, el proyecto cuente con los
UniAndes	fundamentos adecuados de proceso y de implementación tecnológica, acordes
	con las expectativas de producto.

Recordando los objetivos planteados en este proyecto:

- 1. Disminuir la sensación general de corrupción de nuestra sociedad.
- 2. Impactar en la reducción que existe en el desconocimiento de la ciudadanía so bre la gestión y uso de los recursos públicos
- 3. Mejorar la explotación de los recursos públicos, ya que la gente va a tener un conocimiento más claro de lo que pasa con estos.
- 4. Realizar métodos y aplicar metodologías de ML para leer documentos de texto y estandarizar la información para que sea fácilmente procesable y replicable. Esto permite que haya un desarrollo de conocimiento que pueda ser extendido en los académicos, corporativo, gubernamental.

Se puede realizar la siguiente tabla de requerimientos y su relación con dichos objetivos:

Requerimientos Para Incrementar				
Requerimiento (ID)	Escala de viabilidad (1-10)	Objetivo(s) Asociado (ID)	Escala de Impacto en objetivo(s) (1- 10)	Score Total
1	7	1,2,3	7	49
2	10	1,2	5	50
3	10	1,2,3	8	80
4	5	2,3,4	10	50
5	8	1,2,3,4	8	64

Requerimientos Para Disminuir				
Requerimiento (ID)	Escala de viabilidad (1-10)	Objetivo(s) Asociado (ID)	Escala de Impacto en objetivo(s) (1- 10)	Score Total
6	9	1,2	3	27

Basándonos en los Score obtenidos de las matrices anteriores, la siguiente es la tabla de priorización de requerimientos final del proyecto (De momento, se decide excluir el requerimiento ID 6, Mobile Responsive, de las prioridades de este dado su bajo score):

Requerimiento (ID)	Descripción	Score
3	Explorar datos usando mapas	80
5	Datos descriptivos actualizados diariamente	64
2	Filtros predefinidos	50
4	Sugerir si un nuevo proyecto será o no exitoso dadas algunos atributos de este	50
1	Tiempo de carga inicial menor a 10 segundos	49

Relación de priorización de objetivos con métricas de negocio del proyecto:

ID Métrica	Nombre	Valor Actual	Valor Objetivo
1	Encuesta de conocimiento en escala de 1 a 10 con preguntas muy cortas a una población aleatoria antes y después de presentarles el tablero final sobre su desconocimiento en temas de gestión y uso de los recursos públicos en infraestructura. (Ej. ¿Qué tanto sabe sobre los proyectos de inversión en su departamento?)	Se plantea realizar la aplicación de la encuesta a una población aleatoria, por lo cual en este momento se desconoce el valor exacto de esta métrica.	Reducir el puntaje de desconocimiento al menos en un 20% de la población.
2	Calificación del tablero medido en una escala de 0 a 5 estrellas.	0	4 estrellas en promedio.
3	Limpieza, integración y uso de la mayor cantidad posible de información. (Ej. Dado que es necesario en algunos casos trabajar con OCR para los contratos, se espera lograr el mayor porcentaje de información posible para utilizar y analizar en el tablero)	0%	85% de la información capturada.
4	Fuentes adicionales de información. (Ej. Se busca enriquecer lo máximo posible la información resultante, por lo cual se busca poder realizar la integración de otras fuentes de datos al tablero final.)	0 Fuentes	1 Fuente adicional.

ID Métrica	Submétrica	Objetivo Asociado
1	Elaborar encuesta de control para evaluar	1
	conocimientos sobre la contratación de	
	infraestructura de Colombia	
2	Definir 5 filtros por defecto para el tablero	2
2	Definir la historia a ser contada por el tablero. El	3,5,2
	tablero debe poder contestar las preguntas que	
	tengan los usuarios.	
3	Establecer cuáles son las visualizaciones que	2,3
	mejor representan los datos	
4	Poder identificar de los proyectos exitosos cuáles	4
	son los features que hicieron tener un buen	
	desempeño a este.	

Fase D (Recursos)

- Personal y perfiles requeridos: Para la realización de este proyecto son necesarios los siguientes roles y número de recursos (algunos miembros del equipo realizarán 2 roles simultáneamente):

Rol	Cantidad Mínima de personas por rol
Ingeniero de Datos	2
Ingeniero de ML/MLOps	2
UI/UX	1
Scrum Master/Product Owner	1

- Tiempo de implementación: Se tienen 8 semanas para la preparación de este proyecto y 8 semanas para la implementación de este, en donde cada integrante del equipo trabajará en este proyecto una media de 12 horas semanales, aunque se usarán las semanas fina les de preparación que queden disponibles para realizar las actividades iniciales de desarrollo.
- Recursos Tecnológicos: Inicialmente, se desea realizar la configuración del siguiente ambiente de desarrollo/producción:
 - o Repositorio de GIT
 - Máquina Virtual en GCP o AWS con las siguientes características: 32 GB de RAM,
 256 GB de DD, 8 Cores en CPU.
- Presupuesto para la inversión. Asumiendo una tarifa de \$80.000 COP por hora de trabajo de cada integrante y estimando unos 500 USD de valor final de crédito necesario para la MV, los costos estimados serían:

Rubro	Total \$
Mano de Obra	\$ 61.440.000
Infraestructura (Costos durante el desarrollo)	\$2.500.000
Total Estimado	\$ 63.940.000

Fase E. Planteamiento inicial Problema de Analytics

Se requiere construir un tablero interactivo utilizando como fuente primaria los datos de contratación pública del gobierno de Colombia, con el fin de facilitar la exploración y análisis descriptivo de estos, orientando la experiencia de usuario principalmente a un público general. Dicha exploración debe ser guiada por medio de filtros predefinidos y su visualización debe enfocarse en dispositivos de escritorio. Los tiempos de carga de los tabs de dicho reporte no pueden tomar más de 10 segundos de espera. Se requiere, además, integrar un análisis prescriptivo adicional que permita a los usuarios la evaluación de nuevos y potenciales proyectos de infraestructura, en la cual el sistema de evaluar que tan beneficiosos puedan ser estos tomando como referencia la historia de los datos ya existentes. También se requiere que la actualización de los datos sea periódica (diaria).

Para soportar este proyecto, se realizará una configuración totalmente nueva de infraestructura en una MV de GCP en la cual se amanecerán los datos de todas las fuentes involucradas, las ETL y la seguridad necesaria con el fin de garantizar la disponibilidad del servicio. Estos datos serán la fuente directa para un tablero utilizando Power BI, en la cual se construirá todos los elementos de narrativa/storytelling requeridos por los usuarios.

Parte II – Establecimiento de las fuentes de datos

Fase A. (Establecer fuentes)

1. ¿Consideran que los datos potenciales a emplear existen o se pueden generar? ¿Dónde se encuentran estos datos o como se podrían generar? ¿Se identifica que se requieren más fuentes de datos que las planteadas en la guia1?

Dentro de los objetivos del proyecto se encuentra integrar varias fuentes de datos en donde se pueda no solo conocer sobre los proyectos pasados, actuales y futuros, sino que predecir si serán exitosos o no. Los datos potenciales del proyecto se encuentran en la página web de SECOP, la cual ofrece varias bases de datos para el uso público. Allí reposa información del plan anual de adquisiciones, las multas y sanciones entre otros. Dado que en la página del SECOP reposa información de diferentes sectores, regiones, entidades, departamentos, municipios, consideramos que la información es suficiente para el alcance del proyecto.

2. ¿Los datos identificados son coherentes para abordar el problema de negocio planteado o se tiene que replantear?

Los datos son coherentes y suficientes para abordar el problema de negocio, en la página web resaltan que "Los datos abiertos cuando la información es de libre acceso y no tiene restriccio nes de uso, está estructurada, puede ser reutilizada y está en formatos de lectura fácil". Por tanto, consideramos que el problema de negocio puede dar respuesta con la información allí contenida.

3. ¿Las fuentes adicionales de información requieren el desarrollo y aplicación de algún instrumento? Nota: considerar si se requiere tiempo para obtener los datos

Las fuentes adicionales de información se encuentran dentro de la misma página del SECOP, son de uso público y sin restricción por lo que no se necesita ningún instrumento o tiempo adicional para obtener los datos. Quizá mucha de esta información, aunque pudiera venir en formatos estructurados, requerirá de procesos de estandarización de formatos de datos, o de algunas transformaciones esenciales para el manejo de esta y que facilite su uso en el proyecto.

4. Para los datos que existen y los que se tengan que generar, indica que tipo de datos son (ej., series de tiempo, datos no estructurados, entre otros)

Dado que ambas bases de datos (Plan de anual de adquisición, Multas y sanciones) pertenecen a la misma fuente de información (SECOP), las dos fuentes de información tienen la misma estructura. EL tipo de datos de dichas bases son datos estructurados, pues cuentan con una estructura definida.

Para el caso de "Plan de anual de adquisición" cada fila representa un registro o proyecto. Para la base de datos "Multas y sanciones" cada fila representa una multa o sanción. En cuanto a los datos, las variables son del tipo texto, numéricas y fechas.

5. ¿Se requiere alguna transformación que genere datos útiles?

Aunque los datos vienen estructurados, se debe hacer un proceso de limpieza de datos para validar el nivel de completitud de la base de datos, la calidad de los mismo, la unicidad y la integridad. Adicionalmente, es posible que toque hacer transformaciones de datos o creación de nuevas variables que faciliten su manipulación en el proyecto. Por ejemplo, en la base de datos de "Plan de anual de adquisición", viene una variable "Fecha estimada de inicio de proceso de selección (mes)" que es el mes del año en que se iniciará el proceso, con esta variable se puede construir una variable tipo fecha que complemente la información y permita una mejor interpretabilidad.

6. Identifica que departamentos o áreas de la organización están involucradas con las fuentes de datos

El departamento encargado del manejo de las fuentes del proyecto es la Oficina de Datos, quienes son los encargados de la administración de la información. La Oficina de Datos permite la gestión integral de los datos y tiene como principal objetivo el procesamiento de toda la información que se genera dentro de la organización, posibilitando el análisis posterior de los datos y la elaboración de informes que se utilizarían para el proyecto.

Fase B. (Determinar los objetivos del análisis de los datos)

Objetivo

- Realizar una exploración, transformación y visualización de datos que permitan identificar tendencias y/o patrones que revelen elementos importantes para tomar acciones que respalden la toma de decisiones en la contratación de infraestructura de Colombia.

Actividades

- Realizar una limpieza de los datos lo suficientemente robusta que permita un análisis claro y conciso de los datos.
- De ser necesario, realizar transformaciones o creación de nuevas variables (calculadas o dependientes) que faciliten el entendimiento, manipulación o procesamiento de los datos.
- Realizar visualizaciones de datos a partir de gráficos, tablas o imágenes que permitan concluir sobre cada una de las variables o dimensiones de la información reportada. Esto es útil para encontrar, distribuciones, tendencias, frecuencias o concentraciones de datos.
- Construir un tablero interactivo que permita al usuario conocer cuáles han sido las características más relevantes de la contratación en infraestructura del país.
- Aplicar modelos de ML que, por medio de unos atributos dados, permita estimar que tan exitoso (probabilidad de llegar a feliz término) puede ser un hipotético proyecto futuro.

Fase C. (Determinar criterios de éxito en el análisis)

1. Describe las métricas para la evaluación de las técnicas identificadas

Para medir el éxito del análisis de los datos se medirán diferentes métricas. Por lo tanto, se promediarán las multas o sanciones de los proyectos en cada año, y dichos promedios no deberían estar tan alejados entre sí.

2. Define el benchmarking para evaluar el éxito de las métricas

Estas sanciones promedio se compararán con multas o sanciones de proyectos de infraestructura contratados por el gobierno nacional de otros países de Suramérica que tengan una economía similar. De esta forma, no solo evaluaremos que nuestras métricas tengan sentido, sino que mediremos que tan alejadas están las multas o sanciones de Colombia en comparación con otros países.

3. Define si hay algunas métricas subjetivas que impacte en la decisión de si el modelo se considera exitoso o no exitoso

Nuestro modelo también será capaz de calcular si un proyecto será exitoso o no. Sin embargo, esta definición de éxito puede ser subjetiva al usuario de la herramienta.

Fase D. (Tipos de Problemas de Analytics)

- Enfoques y técnicas candidatos: Entender actividad pasada. Descubrir insights en el contenido utilizando aprendizaje supervisado y no supervisado. Procesamiento de lenguaje natural para interpretar texto de contratos.
- La idea es utilizar el descubrimiento de insights para encontrar las características comunes que tienen aquellos proyectos que han sido exitosos anteriormente.
- La forma de alinear estos resultados podría basarse en un ejercicio de clustering. El plan inicial es agrupar los proyectos en diversos clusters y luego identificar el significado de negocio de

- estos. Esperamos encontrar uno o más clúster con información asociada a proyectos mayoritariamente fallidos.
- Una vez construidos los clusters, si al momento de intentar identificar el cluster de un nuevo proyecto futuro este pertenece a algún clúster con mayoría marcada a fracaso, daremos eso como respuesta al usuario.

Fase E. (Problema de Analytics)

Con el fin de que tanto la ciudadanía como los entes de control realicen un control efectivo de las inversiones públicas en Colombia, así como también puedan aportar al análisis de la viabilidad técnica de un proyecto de infraestructura, es necesaria la construcción de una herramienta computacional que les facilite realizar dichas acciones.

Así, se requiere construir un tablero interactivo utilizando Power BI, utilizando como fuente primaria los datos de contratación pública del gobierno de Colombia, con el fin de facilitar la exploración y análisis descriptivo de estos, orientando la experiencia de usuario principalmente a un público general. Dicha exploración debe ser guiada por medio de filtros predefinidos y su visualización debe enfocarse en dispositivos de escritorio. Los tiempos de carga de los tabs de dicho reporte no pueden tomar más de 10 segundos de espera. Se requiere, además, integrar un análisis prescriptivo a dicional, utilizando técnicas de clustering, que permita a los usuarios la evaluación de nuevos y potenciales proyectos de infraestructura, en la cual el sistema de evaluar que tan beneficiosos puedan ser estos tomando como referencia la historia de los datos ya existentes. También se requiere que la actualización de los datos sea de las diversas fuentes de datos sea en batch y con una periodicidad de 24 horas.

Los datos se encuentran en el sistema SECOP. Se componen de múltiples archivos planos que deben ser limpiados y procesados de manera individual. En ellos, la existencia de valores nulos y duplicados o datos que no tienen relación entre los diferentes archivos son un reto a ser resulto. El plan es convertir todos estos archivos en un modelo en estrella que permita responder a diferentes preguntas de negocio de los usuarios.

Para evaluar la validez del resultado en términos de métricas, esperamos contar con el 85% de la información capturada depurada en nuestro modelo en estrella. Los clústers generados en el proceso de aprendizaje no supervisado deben tener un nivel de 60% o más de pureza con relación a proyectos exitosos o fallidos para ser concluyentes de cara a cliente (si un clúster no supera este umbral, no podremos dar una conclusión sobre los proyectos que pertenezcan a dicho clúster).

Para soportar este proyecto, se realizará una configuración totalmente nueva de infraestructura en una MV de GCP en la cual se amanecerán los datos de todas las fuentes involucradas, las ETL y la seguridad necesaria con el fin de garantizar la disponibilidad del servicio.

El plan de desarrollo del proyecto se compone de 16 semanas de trabajo y un equipo de desarrollo de 4 integrantes. El principal beneficio esperado de este proyecto es la de proveer a la ciudadanía en general una herramienta que les permita realizar control efectivo sobre las inversiones públicas, con ello se busca, además, disminuir el sinsabor generalizado de corrupción que sienten los ciudadanos acerca de este tipo de proyectos públicos.

Parte III – Estado del arte: referencias de investigaciones, aplicaciones en el área

Nombre del	Similitudes	Diferencias	Bibliografía
proyecto			
Contratación	- Desarrolla un	- Mide los niveles de	Espinoza, S (2017).
estatal óptima	modelo que	corrupción que generan	Contratación estatal óptima
para proyectos	determina si una	los mecanismos	para proyectos de
de	futura contratación	tradicionales de	infraestructura en entornos
infraestructura	es conveniente dado	contratación directa	corruptos: empaquetamiento
en entornos	los niveles de	como herramienta para	por medio de asociaciones
corruptos:	bienestar social	_	publico privadas vs

empaquetamient o por medio de asociaciones público-privadas vs contratación directa por etapas		desarrollar proyectos de infraestructura - Compara la conveniencia de un contrato basándose en dos modelos (APP y contratación directa por etapas)	contratación directa por etapas [Tesis, Universidad de los Andes]. https://repositorio.uniandes.edu. co/flexpaper/handle/1992/1351 1/u722660.pdf?sequence=1&is Allowed=y#page=1
Contratación pública y debilidad institucional en infraestructura en Colombia	- Disminuye el malestar causado por el desconocimiento de las inversiones en infraestructura estatal Se presiona indirectamente a los contratistas a mejorar sus procedimientos internos para mejorar su imagen al público.	- Discute medidas para fortalecer la capacidad institucional y la estructuración de los contratos, usando adecuadamente la técnica de Project financing Discute sobre cómo debe financiarse la infraestructura, proyectos actuales de infraestructura y como se compara con otros sectores económicos en Colombia	Benavides, J. (2020). Contratación pública y debilidad institucional en infraestructura en Colombia. Revisa De Ingeniería. https://web-s-ebscohost-com.ezproxy.uniandes.edu.co/e host/pdfviewer/pdfviewer?vid= host/pdfviewer?vid=

		participar en las contrataciones públicas.	tab=Todo&query=any,contains, corrupcion%20en%20contrataci on%20de%20infraestructura%2 0colombia
Las asociaciones publico privadas en Colombia: ¿Instrumento para el desarrollo?	- Analiza el desempeño de las asociaciones publico privadas en Colombia y su incidencia en el desarrollo Propone cuales serían el tipo de contrataciones que serían beneficiosas para el desarrollo del país en el futuro, y da ejemplos de contratos de infraestructura que fueron fracasos.	- Es una investigación/crítica - Justifica porque las APP han tenido una incidencia reducida en el desarrollo en una dimensión integral y como podrían reformularse para que cumplan con su objetivo	Chala, J. (2021). Las asociaciones publico privadas en Colombia: ¿Instrumento para el desarrollo? [Tesis, Universidad de los Andes]. https://uniandes.primo.exlibrisg roup.com/discovery/fulldisplay?docid=alma991005380839007681&context=L&vid=57U_UDLA:UDLA⟨=es&search_s cope=MyInst_and_CI&adaptor=Local%20Search%20Engine&tab=Todo&query=any,contains, corrupcion%20en%20contratacion%20de%20infraestructura%20colombia