Aprendizaje no Supervisado para la Detección de Corrupción en la Contratación Pública de Colombia

Propuesta Ganadora de la *Convocatoria para Jóvenes Investigadores Álvaro Reyes* del Centro de Estudios Manuel Ramírez

-Kevin Steven Mojica Muñoz-

Motivación

- Los algoritmos de Inteligencia Artificial tienen potencial como herramienta en la lucha contra la corrupción, con algunos antecedentes de: Mojica, 2021; Gallego, Prem y Vargas, 2020; y Castiblanco, 2018.

- Hasta el momento solo se han implementado algoritmos de aprendizaje supervisado como herramienta para detectar riesgo de corrupción. Estos algoritmos requieren de una variable objetivo difícil de conseguir en estos contextos.

- Castiblanco (2018) hace un análisis exploratorio con aprendizaje no supervisado en Bogotá y Antioquia, pero tenía dos dificultades: i) no lograba generar observaciones comparables entre sí, y ii) no disponía de una base de datos con variables que permitieran identificar riesgo.

Contribución: Utilizando un enfoque novedoso de aprendizaje no supervisado en dos etapas iterativo es posible generar indicadores de riesgo de corrupción en la contratación pública de Colombia utilizando los datos del SECOP II.

Aprendizaje no Supervisado para la Detección de Corrupción en la Contratación Pública de Colombia

Pregunta de investigación:

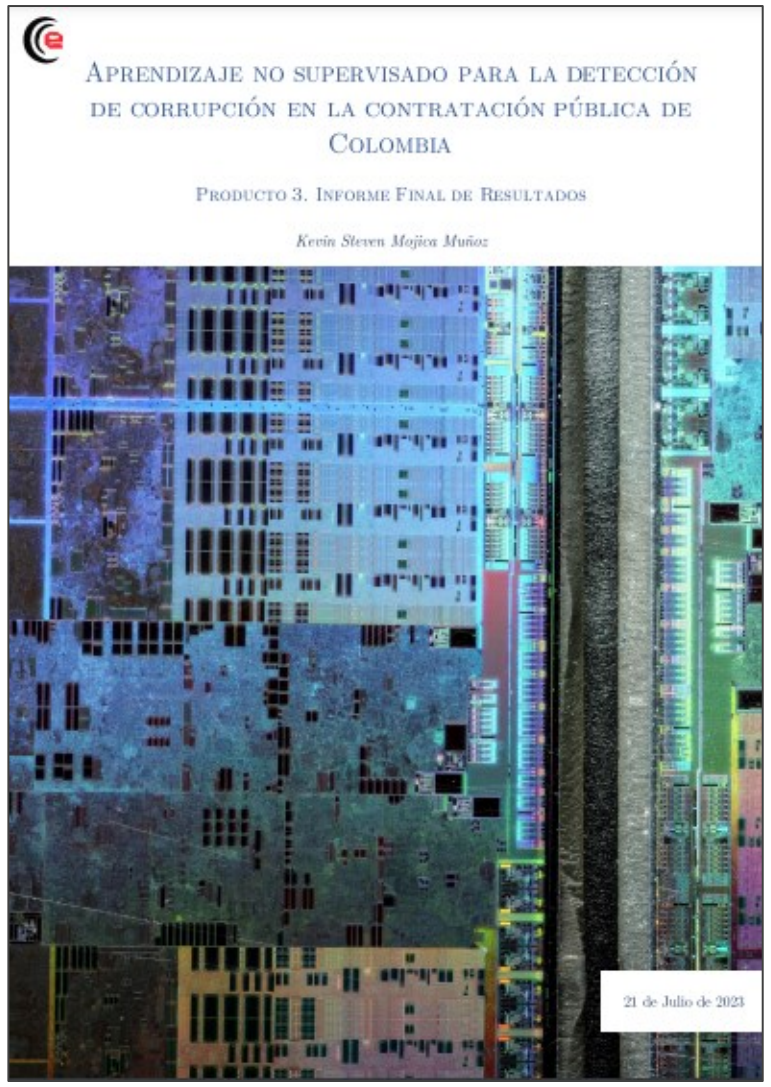
¿Pueden los algoritmos de aprendizaje de máquinas (Inteligencia Artificial) resultar ser herramientas efectivas para identificar riesgo de corrupción en la contratación pública de Colombia?

Objetivo de Investigación:

Implementar y evaluar algoritmos de aprendizaje no supervisado para detectar de manera temprana presuntas irregularidades en la contratación pública del país tomando como caso de estudio el PAE.

Impacto en la sociedad:

Desarrollar un esquema de alertas tempranas para facilitar las labores de veeduría y prevención de los organismos de control, la sociedad en general, y las agencias especializadas del gobierno.



El desarrollo de este proyecto se hizo en el marco de cuatro fases consecutivas de trabajo

---

1. **Detalle de las bases metodológicas**

    - **Resultado:** Entregable 1. Plan de Trabajo y Enfoque metodológico.
    - **Duración:** Nov 2022 – Mar 2023

2. **Procesamiento y limpieza de datos**

    - **Resultado:** Entregable 2. Informe de avance sobre la fase de procesamiento y limpieza de datos.
    - **Duración:** Dic 2022 – May 2023

3. **Implementación y evaluación de los algoritmos**

    - **Resultado:** Script de R utilizado para la implementación y evaluación de los algoritmos
    - **Duración:** May 2023 – Jul 2023

4. **Presentación de resultados**

    - **Resultado:** Entregable 3. Informa final de resultados
    - **Duración:** Jul 2023 – Dic 2023

Marco Conceptual Corrupción en el PAE

- Utilizo la aproximación de Zuleta et al. (2018), quienes hacen un estudio completo de los riesgos de corrupción existentes en el sistema de compra pública en Colombia.

1. Persistencia de modalidades de contratación no competitivas
2. Concentración de contratistas
3. Modalidades de contratación excepcionales
4. Adiciones / modificaciones

Estos factores de riesgo servirán como base para interpretar los resultados de los algoritmos de aprendizaje no supervisado y asignar de manera consecuente el riesgo de corrupción.

Supuestos

1. Las entidades que están legalmente obligadas a publicar sus contratos en SECOP II efectivamente lo hacen para la mayor parte de su contratación. Esto implica que los datos en SECOP II contienen el universo de contratos elegibles para las diferentes aplicaciones que se desarrollen con Inteligencia Artificial en el conjunto de entidades legalmente obligadas.

2. Las entidades digitan la información de los contratos de manera rigurosa para la mayor parte de su contratación y no manipulan los datos, lo que implica que la información en las bases de datos del SECOP coincide con la información real de la contratación y la documentación del proceso.

3. Las bases de datos del SECOP II se encuentran estructuradas y son de libre acceso.

4. Las bases de datos del SECOP II contienen información suficiente para entrenar modelos de aprendizaje de máquinas con alto grado de confiabilidad.

Metodología

1. **Preparación de la muestra a utilizar**

    - Pegar, filtrar y procesar las bases de datos para obtener la submuestra de contratos ejecutorios del PAE (Compilación y estandarización de los datos, tratamiento de los valores faltantes, y el filtrado de las variables y observaciones que se utilizaron en el análisis)

2. **Aplicación de los algoritmos en la submuestra de trabajo**

    - *clustering* en dos etapas con modelo base de *kmeans* y con criterio de distancia euclidiana.

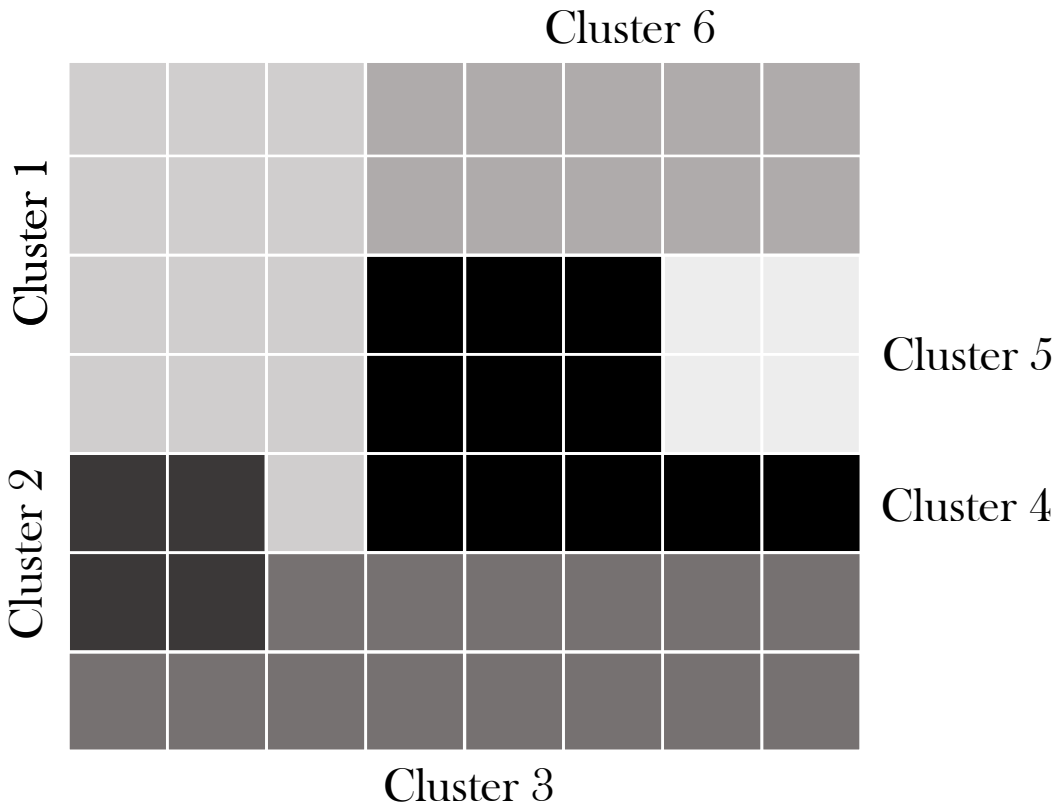
    - La primera etapa busca segmentar la muestra en conjuntos de contratos que sean comparables entre sí, considerando las variables de valor del contrato y duración.

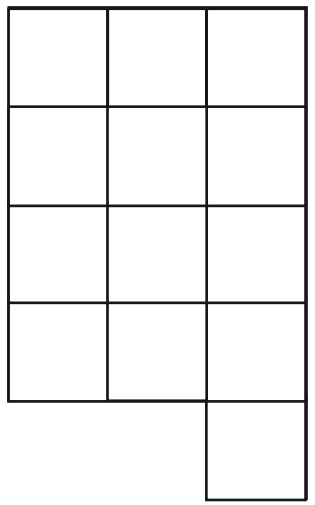
    - La segunda etapa de esta metodología busca estimar el nivel de riesgo de corrupción a través de un *clustering* multivariado (k=5) en cada uno de los subgrupos de la primera etapa y utilizando seis variables (0 - 1):

        1. Si el contrato tuvo un proceso competitivo,
        2. La concentración de la contratación en la entidad - contratista,
        3. Si se trata de una modalidad de contratación excepcional,
        4. Si el contrato tuvo adiciones o modificaciones,
        5. Si el proceso de contratación tuvo alguna garantía rechazada,
        6. El riesgo de contratación de la entidad.

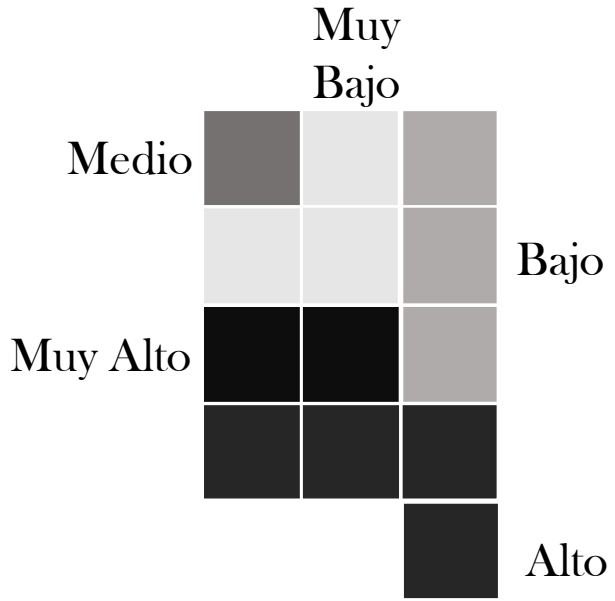
Representación gráfica del proceso metodológico

El clustering es una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como Clusters. **Lo que se busca es segmentar la información para identificar aquellos contratos con anomalías.**





Cluster 1



Metodología

2. Aplicación de los algoritmos en la submuestra de trabajo

- Las variables determinan el nivel de riesgo del contrato. Los contratos serán más riesgosos cuando la mayor cantidad de componentes estén activos.

- Los grupos que se conformen en la primera etapa deben cumplir con un requisito de observaciones mínimas.

    - Se plantea una aproximación iterativa en la conformación de los grupos de la primera etapa en la que cada iteración se hace sobre el subconjunto de observaciones que quedaron en un cluster cuyo número de observaciones es menor a 60 observaciones.

    - Esto se hace hasta que el total de observaciones esté clasificado en un cluster.

- También se compara el desempeño entre un *clustering* jerárquico (*criterio ward*) y no jerárquico (*kmeans*).

3. Evaluación de los algoritmos

- La validación de los datos de origen (proporción de contratos con inconsistencias en la información y el promedio de inconsistencias en el subconjunto de contratos con inconsistencias)

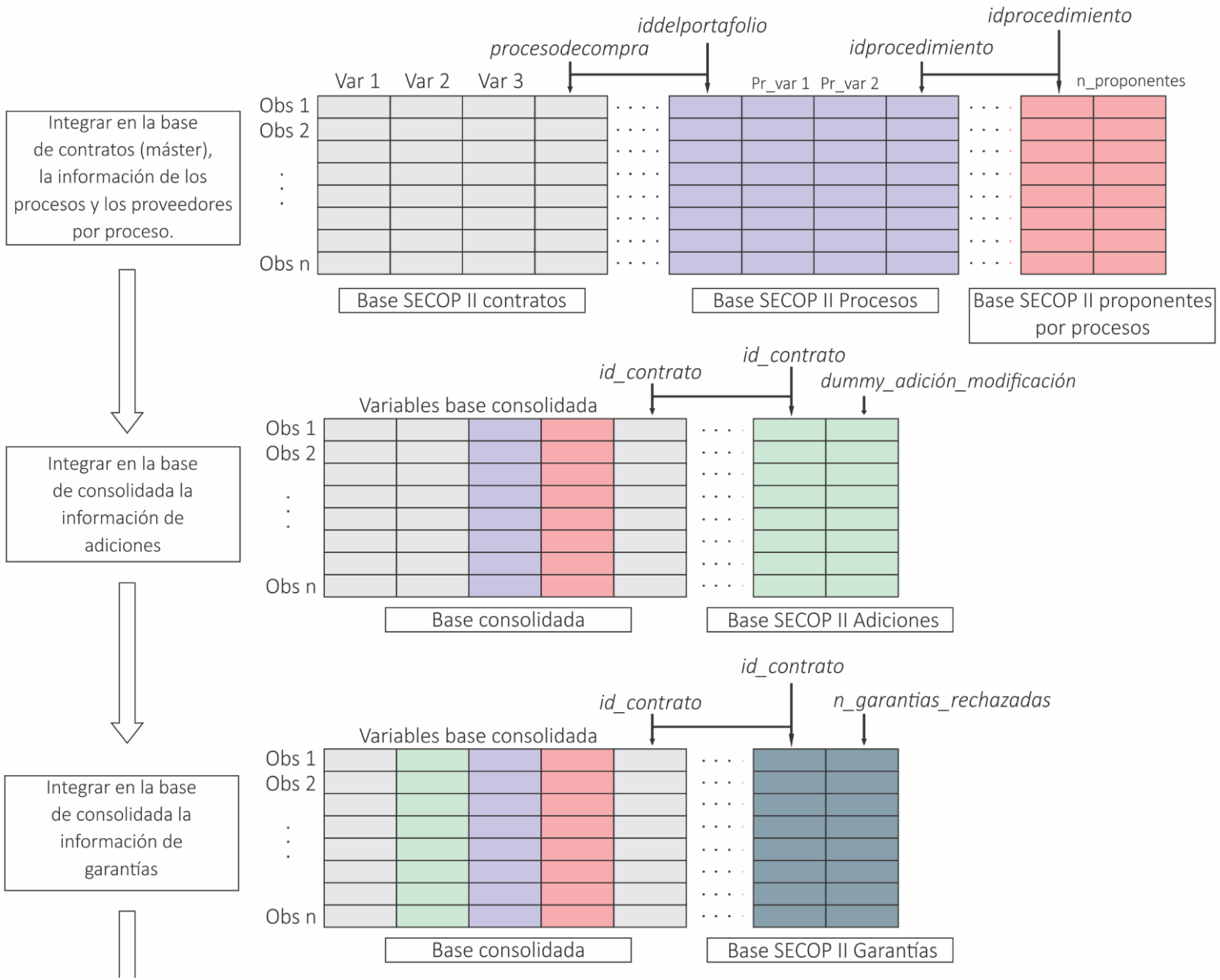
- Evaluación de la calidad del agrupamiento utilizando el coeficiente de Silhouette. $$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Datos

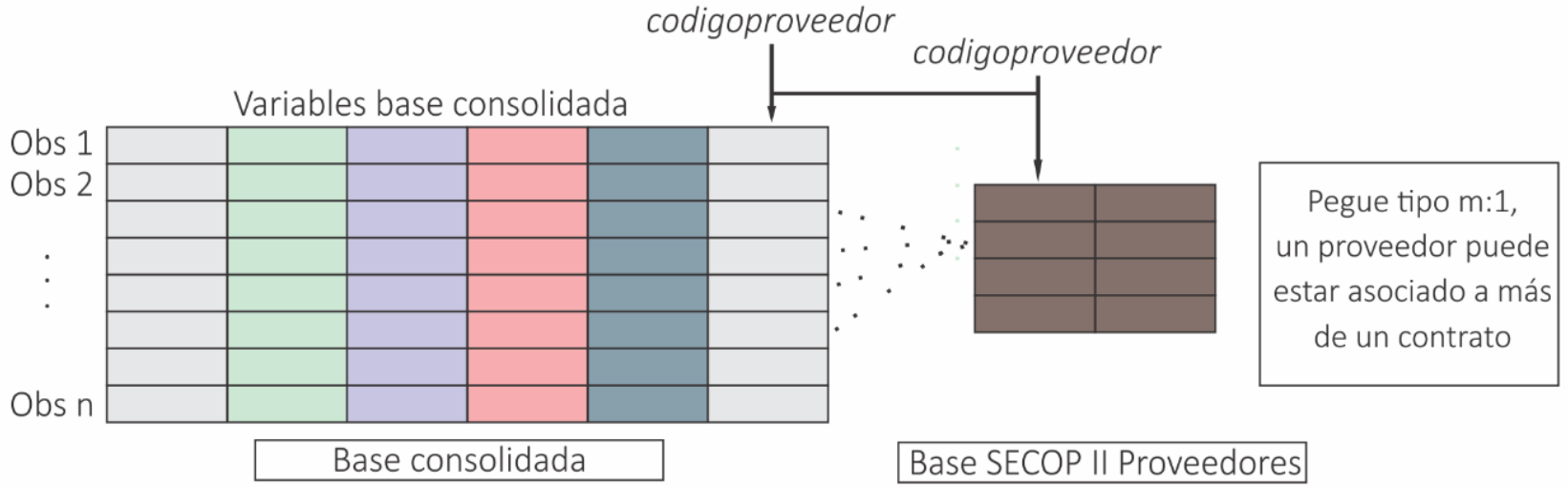
Las bases de datos del SECOP [Nov 2022] que se utilizaron para efectos de esta investigación son las siguientes:

- SECOP II Procesos de Contratación: Compila 2.24 millones de datos sobre procesos de contratación.
- SECOP II Contratos electrónicos: Compila 2.02 millones de datos sobre contratos electrónicos.
- SECOP II Adiciones: Compila 2.33 millones de datos sobre adiciones contractuales.
- SECOP II Garantías: Compila 5.01 millones de datos sobre garantías asociadas a contratos.
- SECOP II Proponentes por procesos: Compila 1.38 millones de datos sobre proponentes por procesos contractuales.
- SECOP II Proveedores registrados: 1.03 millones de datos sobre proveedores.

Adicionalmente se utilizaron los indicadores de riesgo en la contratación para entidades del Instituto Anticorrupción.

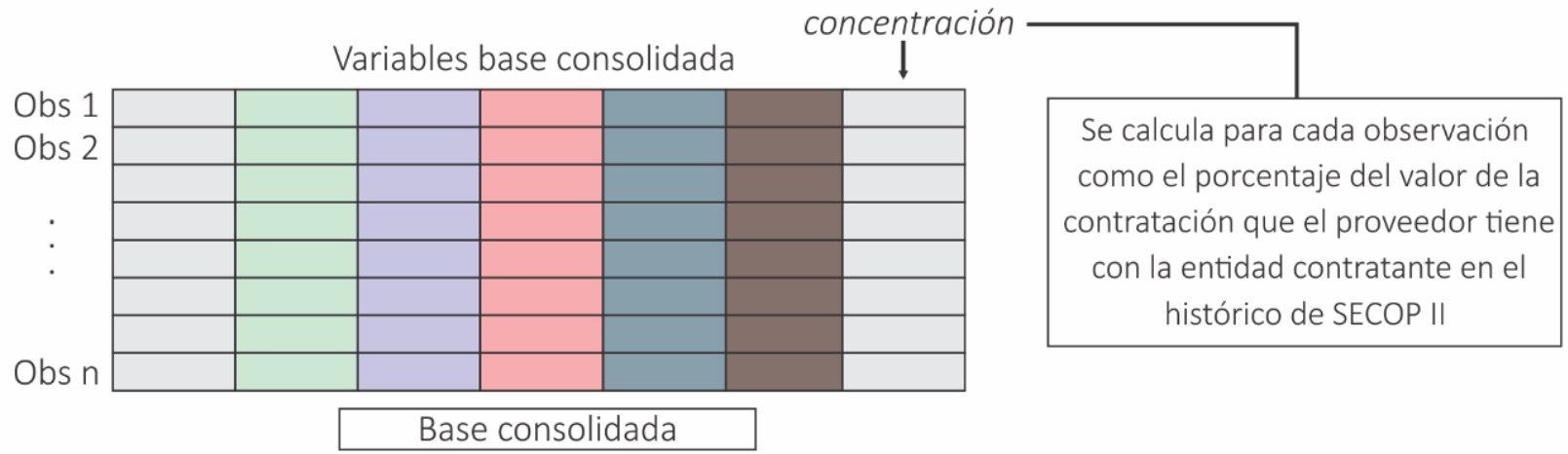

Integrar en la base de contratos (máster), la información de los procesos y los proveedores por proceso.
Integrar en la base de consolidada la información de adiciones
Integrar en la base de consolidada la información de garantías
procesodecompra
iddelportafolio
idprocedimiento
idprocedimiento
Var 1
Var 2
Var 3
Pr_var 1
Pr_var 2
n_proponentes
Obs 1
Obs 2
Obs n
Base SECOP II contratos
Base SECOP II Procesos
Base SECOP II proponentes por procesos
id_contrato
id_contrato
dummy_adición_modificación
Variables base consolidada
Obs 1
Obs 2
Obs n
Base consolidada
Base SECOP II Adiciones
id_contrato
id_contrato
n_garantias_rechazadas
Variables base consolidada
Obs 1
Obs 2
Obs n
Base consolidada
Base SECOP II Garantías

Integrar en la base de consolidada la información de proveedores



Pegue tipo m:1, un proveedor puede estar asociado a más de un contrato

Calcular en la base consolidada el porcentaje de concentración en el valor de la contratación para cada pareja proveedor- entidad



Se calcula para cada observación como el porcentaje del valor de la contratación que el proveedor tiene con la entidad contratante en el histórico de SECOP II

Datos

- Se realizó un proceso de filtrado de información desde la base consolidada hasta la muestra utilizada para los algoritmos de aprendizaje de máquinas.
    - El proceso involucró siete filtros distintos: i) entidad contratante, ii) tema, iii) tipo de contrato, iv) valor del contrato, v) duplicados, vi) código de categoría, vii) estado de contrato. En cada uno de los filtros se hicieron pruebas para verificar la calidad del procesamiento y asegurar la pertinencia de la muestra final.
    - Se evidenció que hay gobernaciones que están sobrerrepresentadas en los datos: Antioquia y Cauca.
    - Para la variable valor del contrato se hizo una transformación de tipo logaritmo natural.
    - Todos los datos tuvieron una transformación de tipo MinMax para asegurar su estandarización en escala 0 – 1.

Dificultades en los datos

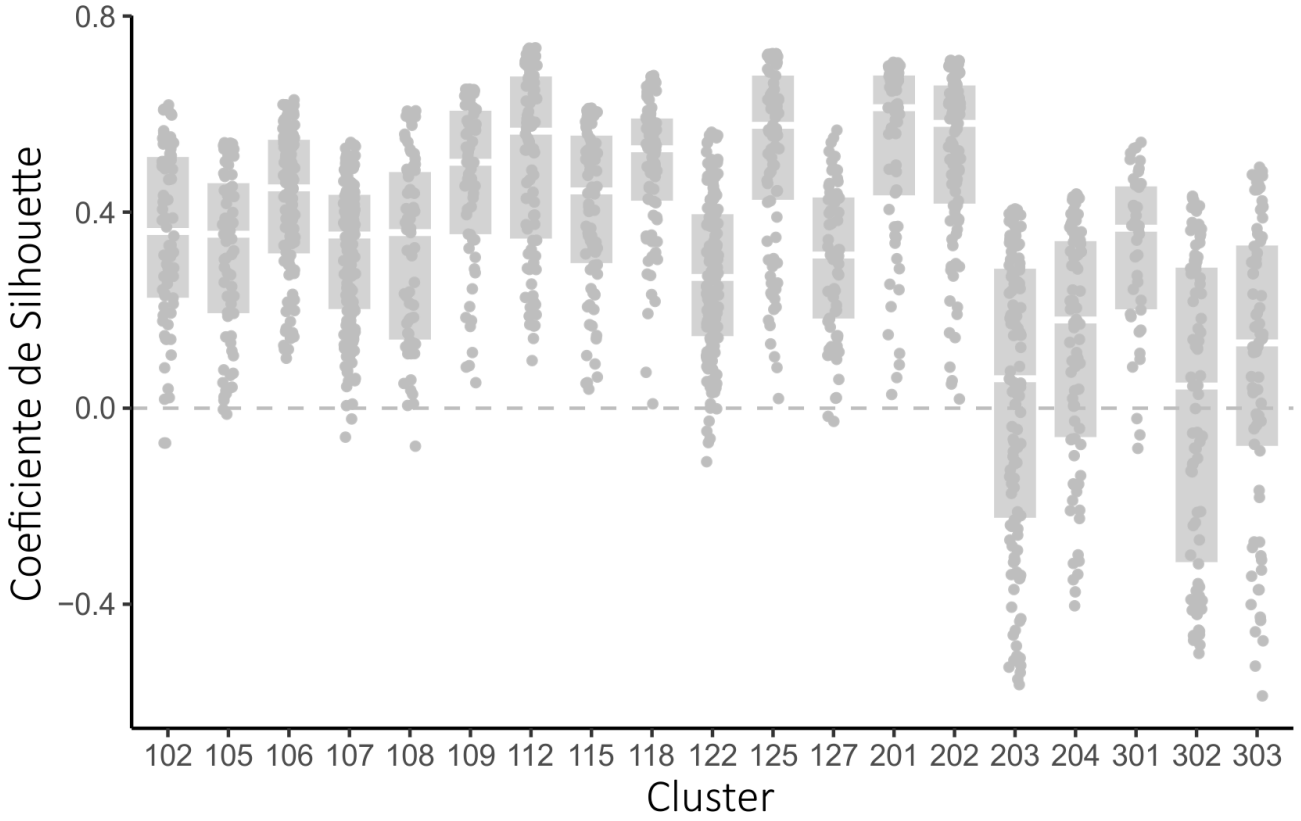
Problema	Magnitud	Tratamiento
En teoría todos los contratos electrónicos del SECOP II deben tener asociado un proceso de contratación en la base de procesos, sin embargo, existen contratos electrónicos que no tienen asociado un proceso de contratación en esta base, lo que implica que tienen información incompleta.	1.2% de la muestra de la base de contratos.	Para los casos en los que el filtrado de la base requiera de la variable 'descripcióndeproceso' de la base de Procesos, se utilizará también la variable 'objetodelcontrato' de la base Contratos.
Contratos con duración igual a cero o negativa.	1.29% de la muestra final.	Se decide eliminar estos contratos al no presentar datos confiables.
Inconsistencias en la atribución de variables. Existen contratos donde el valor o categoría que reporta la variable no corresponde con el dato real del contrato.	Afecta las variables: "orden", "rama" , "estadocontrato", "tipodecontrato", "codigodecategoriaprincipal", entre otras.	Se categorizan estas variables como 'No confiables', lo que implica que su uso como mecanismo para el filtrado de la base se limita a lo estrictamente necesario.
Contratos con procesos de contratación duplicados.	5.64% de los contratos de la muestra	Esto ocurre cuando un contrato se modifica en alguna de sus categorías principales, se decide mantener la última actualización teniendo en cuenta la fecha de modificación.
Valores faltantes.	Afecta principalmente las variables de fechas de ejecución, liquidación, y adjudicación.	Se decide no utilizar las variables de fechas de ejecución. En cambio, se utilizan las fechas de inicio y final del contrato que no presentan este problema.

Dificultades en los datos

Problema	Magnitud	Tratamient
Valores extremos en algunas variables	Afecta exclusivamente a las variables de valor del contrato y semejantes	Se decide aplicar el logaritmo natural para reducir la escala de los datos extremos.
Falta de información estratégica sobre los contratos. No se reporta el número de beneficiarios, el alcance geográfico, o las condiciones sobre las que se acuerda la entrega de los servicios. Tampoco se reporta los detalles de la modificación o ajuste al contrato una vez se implementa.	Afecta a todas las bases del SECOP II	No hay forma de resolver este problema en la actualidad, pero en el futuro se pueden utilizar técnicas avanzadas de scraping para recolectar la información de los contratos directamente sobre los documentos que los soportan.
Ausencia de identificadores que faciliten la interoperabilidad de bases de datos	Afecta a todas las bases del SECOP II	Se decide utilizar el NIT de la entidad para poder integrar en los datos los identificadores necesarios para garantizar la interoperabilidad de las bases de datos.
Los nombres de las entidades contratantes no se encuentran estandarizados bajo criterios únicos. En algunos casos la administración municipal se describe como alcaldía, en otros como municipio, en otros se pone la secretaría correspondiente. Tampoco existen variables que indiquen de manera confiable el orden de la entidad.	Afecta las variables "nombredeentidad" y "orden"	Se crea una variable que categorice el orden de las entidades al identificar todas las posibles formas en las que se denominan a las administraciones del orden territorial utilizando las raíces de los nombres en la variable "nombredeentidad"
Datos desactualizados	Afecta las variables "estadodecontrato" y las variables asociadas a los valores pagados, ejecutados o amortizados.	Se categorizan estas variables como 'No confiables', lo que implica que su uso como mecanismo para el filtrado de la base se limita a lo estrictamente necesario.

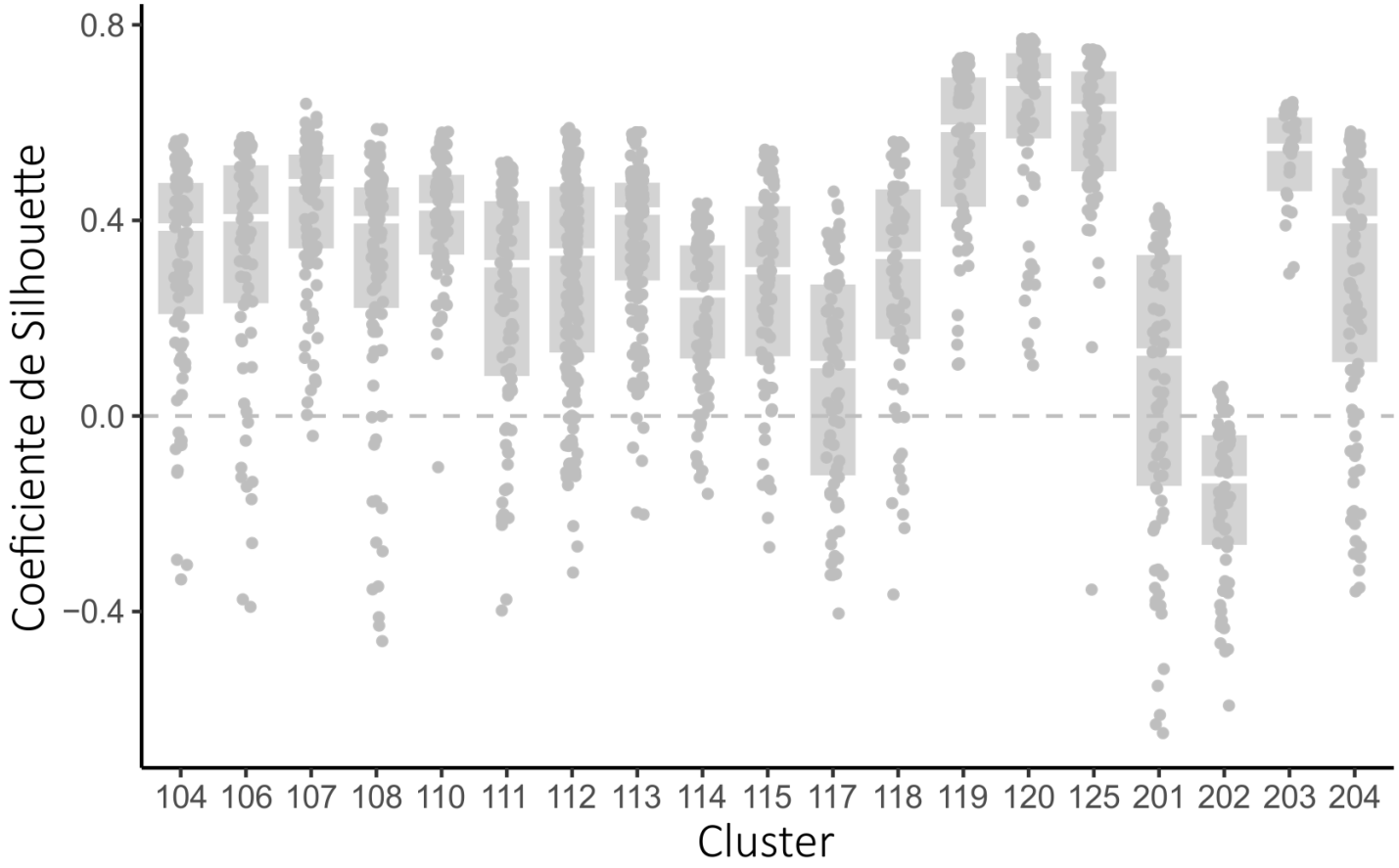
Resultados primera etapa

- El promedio del Coeficiente de Silhouette en el caso del ***clustering* no jerárquico** (kmeans) es de 0.38 para las observaciones clasificadas en la primera iteración, 0.27 para las *observaciones* clasificadas en la segunda iteración, y 0,09 para las observaciones clasificadas en la tercera iteración, para un resultado global de 0.32.

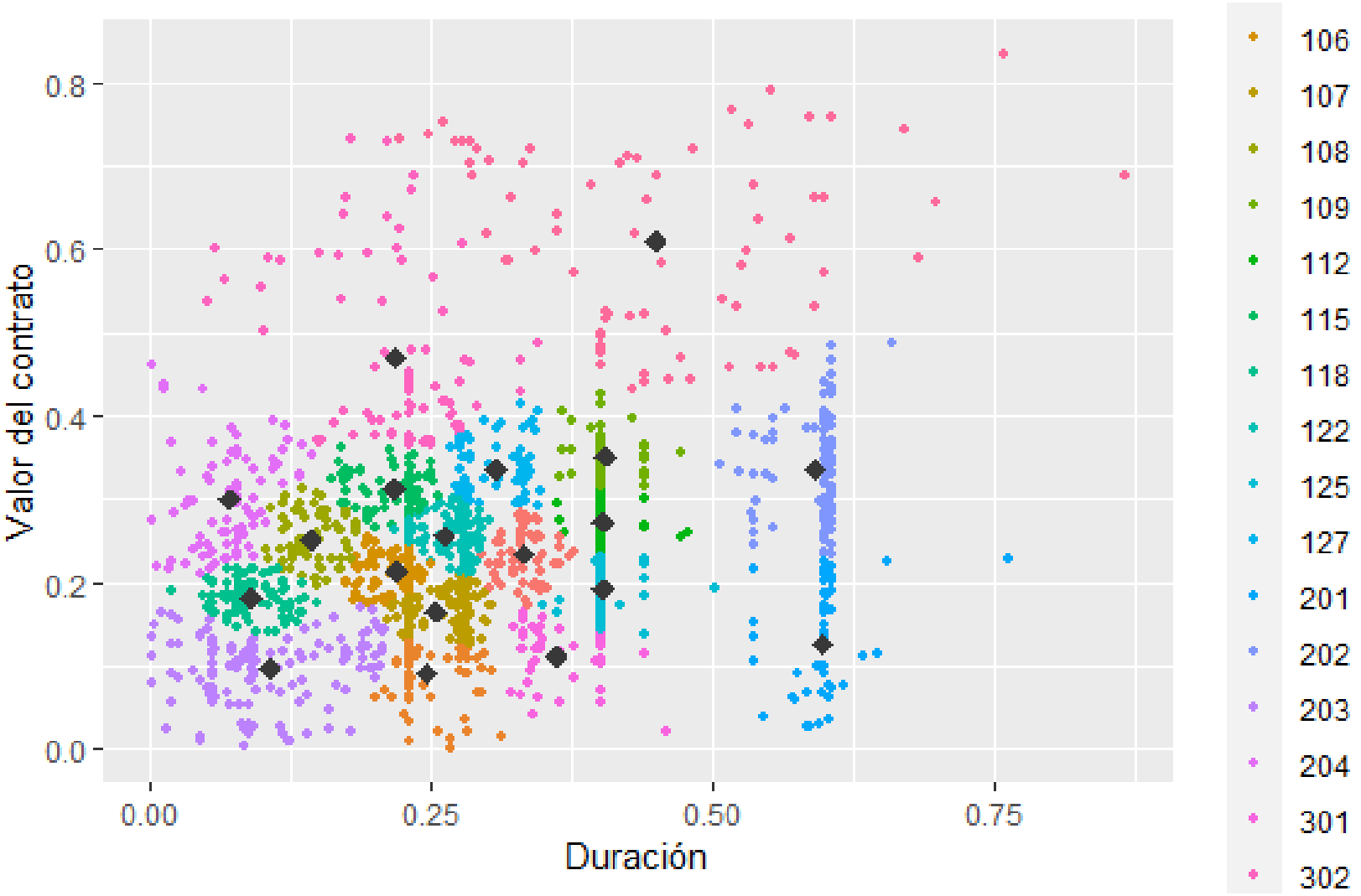


Resultados primera etapa

- Para el ***clustering*** **jerarquizado**, el promedio del Coeficiente de Silhouette es de 0.34 para las observaciones clasificadas en la primera iteración, y de 0.13 para las observaciones clasificadas en la segunda iteración, con un resultado global de 0.30.



Resultados primera etapa



Resultados segunda etapa

Clúster de la primera etapa	Coeficiente de Silhouette segunda etapa
102	0.906
105	0.959
106	0.957
107	0.958
108	0.913
109	0.936
112	0.736
115	0.946
118	0.942
122	0.927
125	0.973
127	0.869
201	0.960
202	0.905
203	0.938
204	0.879
301	0.878
302	0.674

Resultados segunda etapa

Clasificación de riesgo	Número de obs.	Riesgo	Dummy proponentes	Concen-tracion	Modalidad excepcional	Dummy adiciones	INAC Contratación	Garantías rechazadas
Muy alto	192	3.23	0.95	0.0038	0.74	0.56	0.78	0.19
Alto	286	2.28	1.00	0.0052	0.04	0.97	0.22	0.04
Medio	551	2.13	0.98	0.0019	0.12	0.50	0.52	0.02
Bajo	363	1.71	0.95	0.0029	0.00	0.11	0.63	0.01
Muy bajo	259	1.07	0.85	0.0034	0.00	0.00	0.19	0.03

Resultados segunda etapa

La validación de los datos de origen se hizo sobre este subconjunto de observaciones categorizadas en el nivel más alto de riesgo.

- Se tomó una muestra aleatoria correspondiente al 15% de las observaciones, donde se revisó para cada observación las variables utilizadas en la primera y segunda etapa, a excepción del Indicador INAC que provenía de información externa.

- Los resultados indican que, del total de observaciones analizadas**, 58.62% presentaba alguna inconsistencia entre la información de la base de datos y la información documental del contrato**, o no había podido verificarse dada la ausencia completa de información.

- En algunos casos, la información del contrato era diferente según si se revisaba la página del proceso contractual, la página del contrato electrónico, o el mismo documento.

- De entre los contratos que presentaban alguna inconsistencia, el promedio de inconsistencias fue de 1.23 inconsistencias.

Conclusiones

- De este ejercicio se puede extraer que los algoritmos de aprendizaje no supervisado pueden ser una herramienta poderosa para detectar riesgo de corrupción en la contratación pública del país a futuro.

- Por el momento los resultados tienen un alcance limitado dadas las enormes limitaciones en la calidad y disponibilidad de la información [El SECOP II].

- No se cumplen los supuestos mencionados previamente para implementar con confianza los algoritmos de aprendizaje de máquinas en el contexto colombiano.

- Hay dudas sobre la aplicabilidad de las herramientas desarrolladas previamente por Gallego, Prem y Vargas, 2020; y Castiblanco, 2018.

Recomendaciones de política

- En primer lugar, establecer mecanismos legales para obligar a las entidades públicas a subir toda la contratación que la ley les exige con los criterios más altos de calidad y consistencia.

- En segundo lugar, definir mecanismos para detectar las irregularidades en los datos de los contratos y llevar estos casos a las instancias pertinentes para que se proceda con las sanciones que haya lugar.

- En tercer lugar, corregir los datos ya existentes en las bases de datos para que puedan ser de utilidad en el futuro.

    - Puede ser necesario utilizar herramientas de scraping avanzadas que incorporen en bases de datos la información que está en los documentos anexos a los contratos.
    - Con esto se pueden establecer medidas sobre la diferencia entre la información presente en las bases de datos y los contratos.
    - Esto permitiría estimar el sesgo de la información y corregir las bases de datos para que sean aplicables a futuros desarrollos en esta materia.

Image by Alan Warburton

Muchas gracias