Universidad de Costa Rica
Sistema de Estudios de Posgrado

COMPARACIÓN DE TÉCNICAS DE ANÁLISIS MULTIVARIADO PARA DETECCIÓN DE VALORES EXTREMOS CON TÉCNICAS NO SUPERVISADAS, UNA APLICACIÓN AL CASO DE ANOMALÍAS EN CONTRATACIÓN DE BIENES EN COSTA RICA.

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Académica en Estadística

JOSE PABLO ARROYO CASTRO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2024

DEDICATORIA:

Para mi esposa, con quien después de casi 20 años prácticamente hemos fusionado nuestras identidades: esta tesis también es tuya. Gracias por compartir tantos momentos juntos, incluso cuando todo sentido común y racionalidad dictaran lo contrario.

AGRADECIMIENTOS:

Al Dr. Shu Wei Chou-Chen quien fue el mejor impulso para sacar adelante esta tesis, su tenacidad, paciencia y buenos oficios le convirtieron en la pareja perfecta para hacer frente a este reto, gracias por estar siempre presente, aún cuando yo no quería estar.

A los comentarios del Dr. Luis Barboza Chinchilla y del PhD. Gilbert Brenes Camacho, quienes hicieron observaciones vitales para mejorar el proceso del culminación de esta tesis, gracias por compartir sus conocimientos de una forma tan asertiva.

A Israel Ortiz y Alejandro Alonso Salas Vargas, quienes han sido compañeros, amigos e inspiración. También quiero agradecer a Juan Alejandro Herrera López, de quien tengo una admiración profunda, espero algún día poder contar con el nivel de compromiso, pasión y eclecticismo en sus pasiones. Gracias por ser todo un monstruo en tantos temas.

A mis padres, hermanos quienes hicieron más de lo que pudieron para asegurarse que nunca me faltara nada en la vida, son mi más grande fuente de inspiración, y mis grandes roles a seguir, gracias por las herramientas que me han permitido soñar, gracias por ser y estar.

Finalmente, a mi esposa, quien tuvo la paciencia y aceptación de permitirme crecer, gracias por nunca detenerme.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Estadística"

LUIS ALBERTO BARBOZA CHINCHILLA (FIRMA)
PERSONA FISICA, CPF-01-1109-0348.
Fecha declarada: 26/10/2024 01:15:18 PM

Dr. Luis Barboza Chinchilla
Representante de la Decana
Sistema de Estudios de Posgrado

SHU WEI CHOU CHEN (FIRMA)
Digitally signed by SHU WEI CHOU CHEN (FIRMA)
Date: 2024.10.23 20:22:38 -06'00'

Dr. Shu Wei Chou-Chen
Director de Tesis

M.Sc. Juan Alejandro Herrera López
Asesor

ALEJANDRO ALONSO SALAS VARGAS (FIRMA)
Firmado digitalmente por ALEJANDRO ALONSO SALAS VARGAS (FIRMA)
Fecha: 2024.10.24 16:26:57 -06'00'

M.Sc. Alejandro Alonso Salas Vargas
Asesor

GILBERT BRENES CAMACHO (FIRMA)
Firmado digitalmente por GILBERT BRENES CAMACHO (FIRMA)
Fecha: 2024.10.25 17:49:00 -06'00'

PhD. Gilbert Brenes Camacho
Director
Programa de Posgrado en Estadística

Jose Pablo Arroyo Castro
Candidato

TABLA DE CONTENIDO

Página

RESUMEN

Este trabajo consiste en comparar técnicas no supervisadas del análisis multivariado para la detección de anomalías en el contexto de la contratación pública de bienes en Costa Rica durante el período 2020-2022 y proponer el enfoque más adecuado. Después de una exhaustiva investigación, se seleccionaron enfoques relacionados con la aplicación de modelos lineales, proximidad, densidad, cortes paralelos al eje y proyección de Manifolds Uniformes.

Se comparó la eficiencia de las técnicas empleadas mediante 490 escenarios controlados, esto permitió evaluar la calidad predictiva de las técnicas estadísticas empleadas en diferentes contextos, considerando aspectos como la asimetría de los datos, la correlación entre las variables, el tamaño de la población y el porcentaje de anomalías presentes.

Los resultados revelaron que los enfoques basados en proximidad y densidad son los que requieren mayores recursos de procesamiento de información en comparación con los enfoques con cortes paralelos al eje o modelos lineales. En cuanto a la calidad de predicción, se observó que, en la mayoría de los escenarios, esta se ve afectada por el aumento en el tamaño de la población y la cantidad de anomalías presentes.

Finalmente, en la aplicación de estas técnicas con datos de Contratación Pública, no se identificó un único modelo óptimo que maximice la utilización de la información en todos los escenarios analizados. Sin embargo, se encontraron ventajas en varios enfoques y se establecieron las condiciones en las cuales algunos modelos presentan un comportamiento más estable a lo largo del tiempo. Se concluye que es necesario delimitar la población para realizar comparaciones adecuadas. Asimismo, se recomienda el análisis mediante una técnica de consenso que pondera las diferencias metodológicas de cada una de las técnicas utilizadas.

ABSTRACT

This work involves comparing unsupervised multivariate analysis techniques for anomaly detection in the context of public procurement of goods in Costa Rica during the period 2020-2022 and proposing the most suitable approach. After exhaustive research, approaches related to the application of linear models, proximity, density, axis-parallel cuts, and Uniform Manifold Projection were selected.

The efficiency of the employed techniques was compared using 490 controlled scenarios, allowing for the evaluation of the predictive quality of the statistical techniques used in different contexts, considering aspects such as data skewness, correlation between variables, population size, and the percentage of anomalies present.

The results revealed that approaches based on proximity and density require more information processing resources compared to those with axis-parallel cuts or linear models. Regarding prediction quality, it was observed that, in most scenarios, it is affected by the increase in population size and the amount of anomalies present.

Finally, in the application of these techniques with Public Procurement data, no single optimal model was identified that maximizes the use of information in all analyzed scenarios. However, advantages were found in several approaches, and conditions were established under which some models exhibit more stable behavior over time. It is concluded that it is necessary to define the population to make appropriate comparisons. Additionally, an analysis using a consensus technique that weighs the methodological differences of each of the techniques used is recommended.

LISTA DE CUADROS

LISTA DE FIGURAS

CAPÍTULO 1
INTRODUCCIÓN

En términos de la Administración Pública, la adquisición de bienes y servicios representa uno de los factores más relevantes para lograr el cumplimiento de los objetivos institucionales, bajo esta premisa el actuar bajo principios de responsabilidad financiera, así como bajo principios de transparencia permite una toma de decisiones razonable, basado a su vez en el control cruzado de los recursos públicos (CAF, 2021).

En Costa Rica, la unificación de las compras públicas mediante una plataforma única tiene su origen en el año 2015 con el Decreto Ejecutivo número 38830-H-MICITT, con esta unificación de la plataforma de compras públicas se logra una estandarización de los mecanismos para ejecutar los procesos de contratación administrativa, logrando a su vez una mayor facilidad para el análisis masivo de la información (DFOE, 2021).

La presente investigación se centra en la adquisición de bienes mediante el Sistema Integrado de Compras Públicas (SICOP). Empleando definiciones teóricas se establecen condiciones que permiten definir lo que se podría indicar como contrataciones de bienes exitosas, dicha valoración se basa en principios estimables con la información disponible, tal como el plazo del proceso de contratación, el costo asociado al bien contratado, o el alcance establecido como la necesidad a satisfacer por el ejecutor del procedimiento. Con este establecimiento de condiciones, se busca una identificación de aquellos procesos cuyo comportamiento se salga del esperable bajo condiciones estandarizadas de análisis.

Rosenblatt (1957) fue el pionero que dio origen a los métodos de aprendizaje de máquinas, esto mediante su propuesta de máquina llamada Perceptron, la cual podía reconocer letras del alfabeto, empleando un sistema basado en el comportamiento del sistema nervioso humano, este tipo de sistema se convirtió en el prototipo del los métodos de Redes Neuronales actuales, y el punto de partidas para los métodos de aprendizajes de máquinas en general. Estos métodos tuvieron su auge en la primera década del siglo XXI, donde la convergencia de 3 factores permitieron su mayor utilización, por una parte el uso de grandes datos, el empleo del procesamiento en

paralelo, y la creación de nuevos algoritmos que hacían uso de este tipo de técnicas (Fradkov, 2020).

En términos generales, las técnicas de aprendizaje de máquinas pueden clasificarse como supervisadas y no supervisadas, en las primeras existe una serie de datos previamente observados que permiten la identificación de patrones en la información, mientras que en la segunda no se cuenta con dichos patrones conocidos, con lo cual suelen ser útiles para descubrir aquellos comportamientos que se podrían considerar atípicos en un conjunto de datos.

Este trabajo considera diferentes modelos para detección de valores anómalos bajo técnicas no supervisadas, considerando su nivel de predicción con diferentes tamaños de población, asimetrías, porcentajes de anomalías (este concepto se abordará en el siguiente capítulo), y nivel de correlación entre el conjunto de datos analizados. Por medio de este tipo de simulación, se logró conocer el ajuste de los modelos empleados que mejor identifican valores anómalos, ayudando a predefinir factores que podrían considerarse al momento de hacer nuevas detecciones de anomalías.

La importancia de este estudio sobre contratación de bienes, se debe a que la cantidad de recursos transados por medio de la plataforma unificada del Estado, según datos de OCDE (2019) para el año 2015 la contratación pública representó un 15% del Producto Interno Bruto (PIB), y un 30% del gasto público. A su vez, el ahorro potencial que puede derivarse del uso de un sistema único alcanza un 1.5% del PIB (CGR, 2019). Considerando lo expuesto anteriormente, resulta fundamental asegurarse de identificar los procedimientos que presenten desviaciones en su comportamiento, ya que estos podrían brindar información relevante para la toma de decisiones. Dichos hallazgos están estrechamente ligados a comprender las causas y consecuencias de los resultados obtenidos.

En la literatura consultada el enfoque de detección de valores anómalos ha tenido como principal enfoque la detección de corrupción, en este caso el cambio de enfoque pretende determinar aquellos factores con un impacto potencial mayor a nivel de todas las instituciones del Sector Público (Martínez *et al.*, 2019; Zuleta *et al.*, 2019).

En cuanto al período de estudio empleado, se utilizaron datos que van desde el año 2018 hasta el año 2021, debido al volumen de la información disponible, así como la calidad de la misma durante ese período. Como resultado de esta investigación se

espera lograr una detección de valores de interés según las características definidas.

1.1 Justificación

Los grandes avances computacionales y analíticos permiten que muchos procesos que involucran datos sean monitoreados continuamente y su recolección sea más efectiva. Asimismo, la ausencia de factores que permitan identificar patrones anómalos y que puedan ser considerados como indicios de actos de corrupción, requiere la aplicación de diversas técnicas de identificación, las mismas usualmente no supervisadas debido a la ausencia de identificación previa. Derivado de la aplicación de estas técnicas es posible analizar los valores extremos en una posterior fiscalización (Ferwerda y Deleanu, 2013). Este tipo de identificación debe de hacerse considerando diversos factores, entre ellos las características particulares que enfrentan las instituciones que realizan una contratación mediante las plataformas disponibles.

Gutman (2014) determinó que las adquisiciones del sector público representan en promedio el 13% del Producto Interno Bruto (PIB) y un tercio del gasto público en los países de la Organización para la Cooperación y el Desarrollo Económico (OCDE), por lo tanto la adecuada utilización de recursos, ayudan a determinar la calidad del gasto público, así como la credibilidad fiduciaria en dicha gestión. En la actualidad se han generado diversos estudios a nivel internacional, enfocándose principalmente en el combate de la corrupción (Martínez *et al.*, 2019; Zuleta *et al.*, 2019).

Usualmente en el ámbito de la fiscalización de recursos públicos, se analizan datos mediante muestreos con el fin de analizar algunas contrataciones de bienes o servicios puntuales (Crous *et al.*, 2012). Evidentemente existe un riesgo debido a que el tamaño de la muestra no suele ser extensa debido a las capacidades de los equipos de fiscalización involucrados.

En la actualidad, con la presencia de bases de datos de mayor dimensión, es posible realizar un análisis más profundo, principalmente en el caso de aquellos sistemas que permitan una alimentación constante y basada en principios de transparencia, definidos así por la legislación actual, tal como el Reglamento para la utilización del sistema integrado de compras públicas "SICOP", N^{o}41438-H.

A su vez, existen diversos tipos de anomalías, las cuales se identifican según diver-

sos enfoques, determinados por la existencia o no de valores anómalos previamente identificados (Aggarwal y Yu, 2001). En caso de que existan dichos valores previos, la detección se podría enfocar en un aprendizaje supervisado o semi-supervisado, con lo cual un estudio previo daría mayor facilidad para la identificación.
Ahora bien, la previa identificación de valores atípicos existe bajo una serie de condiciones, pero requiere un entrenamiento para su posterior detección con los modelos respectivos, adicionalmente, existe un enfoque no supervisado el cual no requiere de una identificación previa, lo cual es más consistente con la realidad de los datos que serán analizados en la presente investigación.
Además, en el ámbito de la aplicación de técnicas estadísticas, es crucial llevar a cabo comparaciones entre los distintos métodos utilizados para determinar cuáles ofrecen las mejores cualidades en cada contexto particular. De igual manera, es esencial considerar el proceso propuesto para la calibración de técnicas no supervisadas.

1.2 Problema de investigación

El problema a investigar es: ¿cuál es la mejor técnica para la detección de anomalías en un enfoque no supervisado con poblaciones simétricas y asimétricas?

1.3 Objetivos

1.3.1 Objetivo General

Analizar la capacidad de detección de anomalías en un enfoque no supervisado mediante simulaciones bajo escenarios de diversas características, con el fin de realizar una identificación de valores anómalos en un contexto de contratación pública de bienes.

1.3.2 Objetivos Específicos

1. Identificar las técnicas estadísticas más apropiadas para la detección de anomalías en un enfoque no supervisado.

2. Analizar diferentes técnicas para la detección de anomalías en un enfoque no supervisado mediante simulaciones de escenarios controlados.

3. Aplicar la técnica más apropiada en el caso de contratación de bienes realizadas mediante la plataforma de SICOP.

Para alcanzar los objetivos planteados, en el Capítulo 2 (Marco Teórico) se presentan los fundamentos teóricos de los análisis sobre los valores extremos, la detección de anomalías y estudios previos sobre la detección de anomalías en contratación pública. En el Capítulo 3 (Marco Metodológico) se exponen las herramientas metodológicas para llevar a cabo los objetivos, tales como la simulación mediante la creación de escenarios controlados. En el Capítulo 4 (Resultados) se presentan los resultados obtenidos de las simulaciones y el análisis de la detección de valores anómalos en las contrataciones de bienes en Costa Rica. Finalmente, en el Capítulo 5 (Conclusiones) se presenta la discusión y conclusiones de la investigación realizada.

CAPÍTULO 2
MARCO TEÓRICO

2.1 Contextualización del concepto anomalía

El concepto de anomalía ha sido ampliamente discutido, pero la definición dada por Hawkins (1980) es la más frecuente en este ámbito de investigación, dicha definición es, *"una observación que se desvía tanto de otras observaciones que despierta sospechas de que fue generada por un mecanismo diferente"*. De esta manera, en términos generales se pueden entender las anomalías como aquellas desviaciones del comportamiento *"normal"*[1] en los datos, usualmente, este comportamiento normal se entiende como una secuencia de múltiples observaciones, y no tanto como un escalar [2].

Las técnicas de detección de anomalías pueden usarse para dos propósitos específicos, el primer propósito se centra en la detección de anomalías para la limpieza de los datos, permitiendo remover los patrones que generan ruido[3] en la información analizada (Amer y Abdennadher, 2011; Ghafoori, 2018).

Otro propósito es generar una identificación de las anomalías, orientada a la aplicación de técnicas de predicción. Específicamente Ghafoori (2018) explica como la detección de anomalías es un campo de Aprendizaje de Máquinas y Minería de datos, enfocado en la detección de valores anómalos en conjuntos de información.

Aggarwal (2017) explica como los algoritmos usuales utilizados de la detección de anomalías se dividen en dos enfoques. El primero basado en dar un puntaje al nivel de desviación o de anomalía de cada observación, lo cual permite una fácil identi-

[1] El comportamiento normal a nivel de esta investigación será entendido como el patrón manifestado con más frecuencia según las características propias de los individuos observados. Otra definición que puede emplearse en caso de conocer la distribución de los datos, es dada por Hautamaki *et al.* (2004), quien explica que un dato normal se define como una observación que se explica por la función de densidad de probabilidad subyacente.

[2] En el contexto de Minería de datos, autores como Aggarwal y Yu (2001) se refieren a las anomalías como anormalidades, discordancias o desviaciones atípicas. Usualmente la definición de anomalía se puede explicar según la desviación de la observación con respecto a su distribución. Por ejemplo, Hautamaki *et al.* (2004) indican que una anomalía en una distribución normal es una observación que se desvía del promedio de las observaciones en tres veces la varianza, dicha situación no es fácilmente alcanzable ya que en ocasiones no es posible determinar la distribución de los datos sin la presencia misma de las anomalías.

[3] El ruido puede ser considerado como una anomalía débil.

ficación de los elementos que poseen un comportamiento más desviado del resto de las observaciones, pero imposibilita el establecer un patrón que brinde resultados concisos sobre si la observación es anormal o no en un contexto general. Esto potencialmente podría llegar a presentar un problema ya que el comportamiento anómalo en ocasiones es impreciso, lo cual dificulta el establecimiento de umbrales, o bien, puede generarse una evolución de los datos en el tiempo, haciendo que puedan cometerse errores en la identificación de los valores atípicos (Amer y Abdennadher, 2011).
Otro tipo de enfoque utiliza la identificación de anomalías mediante etiquetas binarias, ya sea mediante el establecimiento de límites teóricos de tolerancia, o bien mediante modelos que brinden directamente un resultado binario en el proceso de identificación de anomalías. Este tipo de enfoque da una menor cantidad de información sobre las anomalías, pero en el caso de que sea el principal interés del analista, es posible que permita simplificar la toma de decisiones.
Ahora bien, existen diversos tipos de anomalías las cuales pueden ser anomalías débiles o fuertes, depende de qué tanto se alejen del comportamiento usual de los datos. En el caso de las anomalías débiles, las mismas son algún tipo de ruido en la información, representando un umbral entre los datos normales y las anomalías fuertes[4] (Aggarwal, 2017).
Complementando lo anterior Breunig *et al.* (2000) explican que desde el punto de vista de análisis de clúster, el ruido puede incluso representar un cluster completo dentro del set de datos analizados. Dicha observación puede verse gráficamente en la Figura 2.1.

Tal como se puede observar, las observaciones más aisladas se detallan en las zonas externas a la región amarilla, estas corresponde a anomalías fuertes, asimismo, la región amarilla podría denominarse como el umbral con observaciones anómalas débiles (o ruido).
Además de lo anterior, las anomalías se pueden clasificar en anomalías puntuales, anomalías contextuales o anomalías colectivas (Gogoi *et al.*, 2010). Las puntuales son aquellas cuyos registros se desvían del resto de los conjuntos de datos; las anomalías colectivas se presentan cuando un grupo de datos similares se desvía del conjunto de datos; Finalmente, Song *et al.* (2007) explican que las anomalías con-

[4]El interés de la presente investigación se centrará justamente en este rubro, siendo las anomalías fuertes consideradas como anomalías verdaderas.

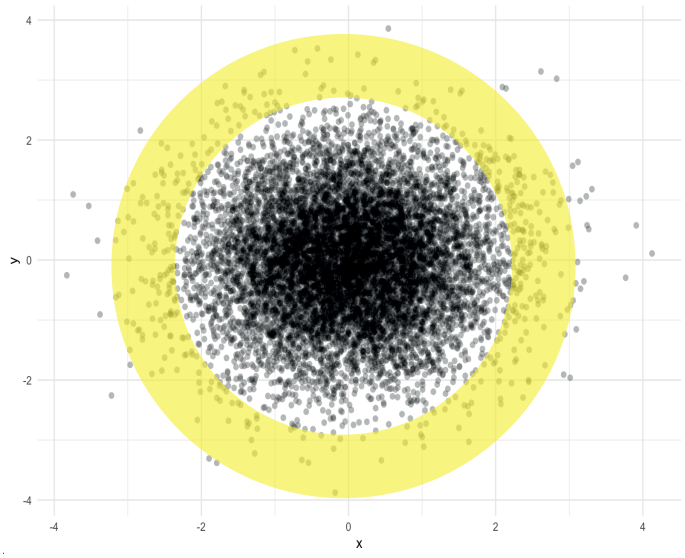

Figura 2.1. Identificación de ruido o anomalías débiles en un conjunto de datos, 2022.

textuales pueden determinarse cuando las observaciones poseen un comportamiento que podrían considerarse normales en un contexto pero anómalos en otro.
Ahora bien, Aggarwal y Yu (2001); Gogoi *et al.* (2010), indican que existen diversas maneras de detectar anomalías, y depende principalmente del tipo entrenamiento así como de sus resultados esperados. A su vez, Ghafoori (2018) indica que el objetivo común de los enfoques para detectar anomalías, se centra en identificar los patrones que no conforman un comportamiento *normal* o *esperado*. Estos tipos de enfoques se enlistan a continuación:

- Las técnicas supervisadas, son definidas por Swersky (2018) como un caso especial de clasificación binaria, ya que requiere de una cantidad de observaciones previamente identificadas como anómalas para poder clasificar los datos como anómalos o normales, mediante un entrenamiento que permita en primera instancia analizar los datos, y en una segunda instancia realizar una prueba con las observaciones restantes. En esta línea Ghafoori (2018) identifica que estas técnicas permiten abordar el problema del desequilibrio de clases dado que las observaciones anómalas suele ser mucho menores que aquellas normales.

- Otro enfoque es generado cuando la cantidad de observaciones disponibles

con la clasificación binaria de anomalía, no son suficientes para poder asegurar una clara identificación del comportamiento anormal, en este caso se puede emplear un enfoque de aprendizaje semi-supervisado, con el cual se tiene un registro de datos normales, lo cual requiere de una fase de entrenamiento y una segunda etapa donde cualquier registro que se desvíe del modelo normal se etiqueta como atípico. Para Amer y Abdennadher (2011) una de las principales dificultades de este tipo de técnicas parte del factor de requerir identificar en el entrenamiento el comportamiento normal en los datos, si los datos de entrenamiento están mal producidos se podrían generar falsos positivos en la identificación. Asimismo, existen dos puntos claves dentro de este análisis, el primero de ellos se centra en el adecuado balance de la generalización para detectar anomalías, y como segundo factor es necesario hacer la técnica de la manera más robusta para evitar que anomalías dentro del conjunto de datos contaminen las observaciones normales (Ghafoori, 2018).

- Finalmente, se tiene el enfoque sin supervisión, en el cual no existe una fase de entrenamiento y se asume que las anomalías son mucho menos comunes que los datos normales en el conjunto de datos, este tipo de identificación se realiza según diversos métodos disponibles. Para Swersky (2018) la selección del método más adecuado puede ser basado según la densidad de las observaciones o bien por otros factores que permitan identificar un aislamiento de las observaciones anómalas. Amer y Abdennadher (2011) explican a su vez que las técnicas no supervisadas en ocasiones no son las más deseables, pero si las más aplicables, porque el establecimiento de etiquetadas binarias y entrenamiento de los datos puede llegar a ser muy demandante en términos de recursos y tiempo.

Una de las principales dificultades para la identificación adecuada de valores anómalos es la dimensionalidad de los datos, en un caso de valores extremos en una dimensión o dos, es posible la identificación de los mismos con técnicas estadísticas más sencillas tal como herramientas visuales. Ahora bien, en el caso de múltiples dimensiones podría generarse una concentración de las distancias en donde una anomalía puede no ser fácilmente identificable, porque podría pasar por ser un conjunto de puntos en una región específica, lo cual hace necesario determinar el tipo

de plano en el cual se pretenden buscar los valores extremos (Aggarwal, 2017).
Dado el contexto de las contrataciones públicas, y la ausencia de un registro oficial de líneas de contratación anómalas, el uso del enfoque no supervisado representa la solución obvia. En la presente investigación se hará una selección de enfoques con el cual abordar el problema, profundizando el análisis en los siguientes apartados.

2.2 Detección de valores por modelos lineales

La detección de valores por modelos lineales puede realizarse, mediante el clasificador de Componentes Principales. Shyu *et al.* (2003) indican que esta técnica posee sus ventajas ya que no posee ningún supuesto respecto a la distribución de los datos, asimismo, en un contexto de alta dimensión la reducción de dimensiones permite llegar a un clasificador simple, permitiendo a su vez que el modelo pueda ser calculado en poco tiempo durante la etapa de detección. Esta técnica busca encontrar hiperplanos de representación óptima de cualquier dimensionalidad. En otras palabras, el Análisis de Componentes Principales (en adelante ACP) puede determinar el hiperplano $k - dimensional$ (para cualquier valor de $k < n$) que minimiza el error de proyección al cuadrado sobre las dimensiones restantes $(n - k)$ (Aggarwal y Yu, 2001).

2.2.1 Componentes Principales

El ACP fue propuesto por Pearson (1901), para encontrar las líneas y los planos que mejor se ajustan a un sistema de puntos en el espacio. Wold *et al.* (1987) explican que esta aproximación se realiza al tomar una matriz X la cual se recompone en términos del producto de dos matrices más pequeñas μ y W', con estas se puede capturar el patrón principal de la matriz X.
Matemáticamente, es posible realizar una proyección de una matriz X en un subespacio de una dimensión A, la cual se da por medio de la matriz de proyección W'. Esta matriz a su vez brinda las coordenadas del objeto en el plano μ (Wold *et al.*, 1987).
Además, Suzuki (2021) indica que necesariamente $(1 \leq d \leq p)$, lo cual permite observar que mientras más pequeño sea el valor de d, más comprimida está la infor-

mación en X.

Adicionalmente, si se considera un entorno con N instancias correspondientes a los puntos de datos $d - dimensionales$, denotados por $\overline{X}_1 \dots \overline{X}_N$. Si el número de puntos N es mayor que la dimensionalidad d, el sistema de ecuaciones está sobredeterminado y es posible que todos los puntos en los datos no satisfagan la condición de independencia lineal $w_1 \cdot x_1 + \ldots + w_d \cdot x_d = 0$.

Ahora bien, sin importar el vector de coeficientes[5] $\overline{W} = (w_1, \ldots, w_d)$, es conocido que se obtendrá un valor de error ϵ_j, ya que la existencia de residuos es esperable:

$$\overline{W} \cdot \overline{X_j} = \epsilon_j \quad \forall \quad \epsilon \{1 \dots N\} \tag{1}$$

Según lo observado en (1), se evidencia el objetivo del modelo, el cual es determinar un vector de coeficientes $\overline{W} = (w_1, \ldots, w_d)$, con el cual la suma de errores al cuadrado $\sum_{i=1}^{N} \epsilon_j^2$ es minimizado, asimismo, es importante evitar una solución trivial de $\overline{W} = 0$, por lo tanto se asume lo siguiente:

$$\left\| \overline{W} \right\|^2 = \sum_{i=1}^{d} w_i^2 = 1 \tag{2}$$

Podríamos indicar que X es una matriz $N \times d$, la cual contiene información de d dimensiones en sus filas, y están denotadas por $\overline{X}_1 \dots \overline{X}_N$. Por lo tanto, se podría escribir el vector de columnas de sus N-dimensiones de los diferentes puntos del plano de regresión como $\overline{\epsilon} = X\overline{W}^T$ el cual según lo indicado en (1) contiene los puntajes de los errores que servirán de base para determinar los valores anómalos.

Con el objetivo de minimizar el vector de distancias $\left\| X\overline{W}^T \right\|^2$ se buscan los coeficientes óptimos de $w_1, \ldots, w_d$, mediante el multiplicador de Lagrange

$$\mathcal{L} := \left\| X\overline{W}^T \right\|^2 - \lambda \left(\left\| \overline{W} \right\|^2 - 1 \right) \tag{3}$$

Con lo anterior se puede obtener el valor óptimo del vector $\overline{W}$, el cual es un eigenvector:

[5] Este factor se conoce también como cargas, lo cual da el peso o la importancia que tiene cada variable en cada componente, y por lo tanto ayuda a conocer que tipo de información recoge cada uno de ellos.

$$\left[\mathrm{X}^{\mathrm{T}} X\right] \overline{W}^{T} = \lambda \overline{W}^{T} \tag{4}$$

Suzuki (2021) indica que el procedimiento de ACP busca obtener d vectores, que cumplan con la maximización correspondiente[6], considerando una matriz de información $X \in \mathbb{R}^{N \times d}$ con $(d \leq N)$, en este caso se cumplirá lo siguiente:

$w_1 \in \mathbb{R}^N$ Con $\|w_1\| = 1$ lo cual maximiza $\|Xw_1\|$

$w_2 \in \mathbb{R}^N$ Con $\|w_2\| = 1$ lo cual es ortogonal a w_1 y maximiza $\|Xw_2\|$

$\vdots$

$w_{\mathrm{d}} \in \mathbb{R}^N$ Con $\|w_{\mathrm{d}}\| = 1$ lo cual es ortogonal a $w_1, \ldots, w_{\mathrm{d}-1}$ y maximiza $\|Xw_2\|, \ldots,$

Amat Rodrigo (2017) indica que una vez que se ha calculado el primer componente, se calcula el segundo, pero añadiendo la condición de que la combinación lineal no puede estar correlacionada con el primer componente, este proceso se repite hasta calcular todos los posibles componentes, donde el orden de importancia de cada componente viene dado por la magnitud del eigenvalue asociado a cada eigenvector. Retomando lo indicado en (4) es posible obtener más de un valor de λ que permita satisfacer la ecuación indicada

$$\|Xw_1\|^2 = w_1^{\mathrm{T}} \mathrm{X}^{\mathrm{T}} X w_1 = \lambda_1 \|w_1\|^2 = \lambda_1$$

En ese caso se tiene que elegir el valor más grande de λ_1 Adicionalmente, para w_2,

$$\mathrm{X}^{\mathrm{T}} X w_2 = \lambda_2 w_2$$

Dado que w_2 es un eigenvector de $X^T X$, debe cumplir la condición tal que $\lambda_1 \geq \lambda_2$, asimismo, w_1 es ortogonal a w_2, dejando abierta la posibilidad que $\lambda_1 = \lambda_2$ por lo tanto se podría decir que w_1 está en el mismo espacio propio que w_2.

Cabe recordar que los eigenvalues son definidos como una matriz no negativa. Para Aggarwal y Yu (2001) se denota que la matriz $\mathrm{X}^{\mathrm{T}} X$ es una versión escalada de la matriz de covariancia Σ de los datos centrados en la media de la matriz X, lo cual

[6] A su vez Amat Rodrigo (2017) indica que el problema de optimización de las cargas permite encontrar el punto donde se maximiza la varianza, mediante el cálculo de eigenvector-eigenvalue de la matriz de covariancias

se representa como sigue:

$$\sum := \frac{1}{\mathrm{N}} \mathrm{X}^{\mathrm{T}} X \tag{5}$$

Ahora bien, si consideramos (5) y reemplazamos $\frac{\lambda_1}{N}, \ldots, \frac{\lambda_d}{N}$ con $\mu_1, \ldots, \mu_N$, se puede reexpresar la ecuación de la siguiente manera:

$$\Sigma w_1 = \mu_1 w_1 \tag{6}$$

Donde, Σ, es la matriz de covarianza basada en muestras y $\mu_1 \geq \ldots \geq \mu_p \geq 0$ son los eigenvalues [7].

Es posible por lo tanto seleccionar los d principales componentes $w_1, \ldots, w_d$ con las mayores variancias $\mu_1 \geq \ldots \geq \mu_{\mathrm{d}} \geq 0$, permite tener por lo tanto, una porción de un $K - esimo$ componente principal, $\frac{\mu_k}{\sum_{\mathrm{i}=1}^{\mathrm{p}} \mu_{\mathrm{i}}}$, y un $k - esimo$ del componente principal acumulado, $\frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mu_{\mathrm{i}}}{\sum_{\mathrm{i}=1}^{\mathrm{p}} \mu_{\mathrm{i}}}$

En este caso se debe considerar algo importante, y es el hecho de que es necesario estandarizar las unidades de medida de las variables, ya que podría presentarse el caso de un desbalance ocasionado por este factor. Cuando la matriz X es generada de manera aleatoria, es improbable que exista más de un eigenvalue que coincida, por lo tanto es seguro asumir que

$$\mu_1 > \ldots > \mu_{\mathrm{m}}$$

Dado que Σ es simétrico, entonces se tiene que

$$\mu_1 \neq \mu_j \quad \Rightarrow w_i^{\mathrm{T}} w\mathrm{j} = 0$$

Con lo anterior, se puede replantear (6), estableciendo lo siguiente,

$$\mu_{\mathrm{j}} w\mathrm{i}^{\mathrm{T}} w_{\mathrm{j}} = {w_{\mathrm{i}}}^{\mathrm{T}} \Sigma w_{\mathrm{j}} = w_{\mathrm{j}}^{\mathrm{T}} \Sigma w_{\mathrm{i}} = \mu_{\mathrm{i}} w\mathrm{i}^{\mathrm{T}} w_{\mathrm{j}}$$

Donde $\left(\mu_{\mathrm{i}} - \mu_{\mathrm{j}}\right) w_{\mathrm{i}}^{\mathrm{T}} w_{\mathrm{j}} = 0$, lo cual significa que $w_{\mathrm{i}}^{\mathrm{T}} w_{\mathrm{j}} = 0$, por lo tanto si se localizan los m eigenvalues mayores y sus eigenvectors, no es necesario verificar que sean

[7] Es conocido que el error cuadrático acumulado junto con un particular eigenvector es igual al eigenvalue.

ortogonales.
Otra manera de indicar lo anterior, es considerar que existe un subespacio en el cual existe una combinación lineal de $(d-1)$ eigenvectors, los cuales dan un sistema de ejes con el cual los datos pueden ser representados aproximadamente con una mínima pérdida de información. Es por lo tanto, de interés conocer los puntos en los cuales el error de esta aproximación es alto, ya que esos puntos serían considerados como anómalos. Aggarwal y Yu (2001) indican que el set de eigenvectors deben satisfacer las siguientes propiedades:

- Si los datos se transforman en el sistema de ejes correspondiente a los eigenvectors ortogonales, la varianza de los datos transformados a lo largo de cada eje (eigenvector) es igual al eigenvalue correspondiente. Las covarianzas de los datos transformados en esta nueva representación son 0.

- Dado que las varianzas de los datos transformados a lo largo de los eigenvectors con pequeños eigenvalues son bajos, las desviaciones significativas de los datos transformados de los valores medios a lo largo de estas direcciones pueden representar valores anómalos.

Es importante indicar que la solución dada por ACP potencialmente es capaz de brindar todas las soluciones a la Ecuación 4, y no únicamente una solución con el eigenvector más bajo.

Para el caso del análisis de detección de anomalías, Aggarwal (2017); Amat Rodrigo (2020) indican que es posible tener una reconstrucción total de las observaciones originales si se realiza la multiplicación de los eigenvectors con los eigevalues, ahora bien, es posible tener un error cuadrático medio de reconstrucción, calculado como el promedio de las diferencias al cuadrado entre el valor original de sus variables y el valor reconstruido, lo cual es utilizado para identificar anomalías, dado que se espera que los datos normales se representen de forma correcta, mientras que aquellas observaciones con mayor error de reconstrucción se considerarían como atípicas.

2.2.2 Componentes Principales Robustos

El método de ACP es muy sensible a la presencia de valores anómalos, por lo tanto, la presencia de un valor anómalo en una variable, hace que la varianza se denote elevada en esa dirección, con lo cual el subespacio creado se moverá en exceso hacia la dirección de la anomalía[8] (Amat Rodrigo, 2020).

Es por lo anterior, que Aggarwal y Yu (2001), exponen el método de Análisis de Componentes Principales Robusto (ACPR) siento este el enfocado a encontrar valores atípicos orientados a la independencia de los valores extremos. Esta solución puede estar definida por dos enfoques, el primero destinado a utilizar una matriz de covariancia robusta, o bien el cálculo de cada componente mediante un método robusto.

En términos generales, Oyeyemi e Ipinyomi (2010) indican que se pueden considerar dos métodos para obtener la matriz de covariancia robusta, ambos desarrollados por Rousseeuw (1984):

- Estimador de elipsoide de volumen mínimo (MVE[9]), con el cual se busca encontrar un elipsoide de volumen mínimo que cubra un subconjunto de al menos h puntos de datos, este subconjunto se llama medio conjunto porque a menudo se elige que h sea más de la mitad de los m puntos de datos. El estimador de ubicación es el centro geométrico del elipsoide y a su vez el estimador de la matriz de varianza-covarianza definida como el elipsoide mismo multiplicado por una constante apropiada para asegurar la consistencia.

- Determinante de Covarianza Mínima (MCD[10]), el cual se obtiene al encontrar el conjunto medio de puntos multivariados que brinda un valor mínimo al determinante de la matriz de covarianza. El estimador de ubicación resultante es el vector medio muestral de los puntos. Con este enfoque un pequeño eigenvelue sugiere una dependencia casi lineal de los datos en el espacio, lo cual indica que hay un grupo de puntos similares entre sí.

[8]Es por este motivo que diversos autores como Aggarwal y Yu, Suzuki y Amat Rodrigo indican la necesidad de normalizar las observaciones, ya que datos no normalizados podrían direccionar los resultados de manera equivocada, no tanto por un factor de identificación de anomalías, sino por unidad de medida de la variable utilizada.

[9]MVE por sus siglas en inglés, Minimum Volume Ellipsoid.

[10]MCD por sus siglas en inglés, Minimum Covariance Determinant.

En un estudio llevado a cabo por Carvalho Rocha *et al.* (2013), se realizó una comparación de métodos para determinar la homogeneidad de un nuevo material, en este escenario el modelo ACPR demostró un mejor rendimiento a la hora de identificar anomalías. Asimismo en la Figura 2.2 es posible visibilizar las diferencias en los resultados entre los modelos de ACP y ACPR.

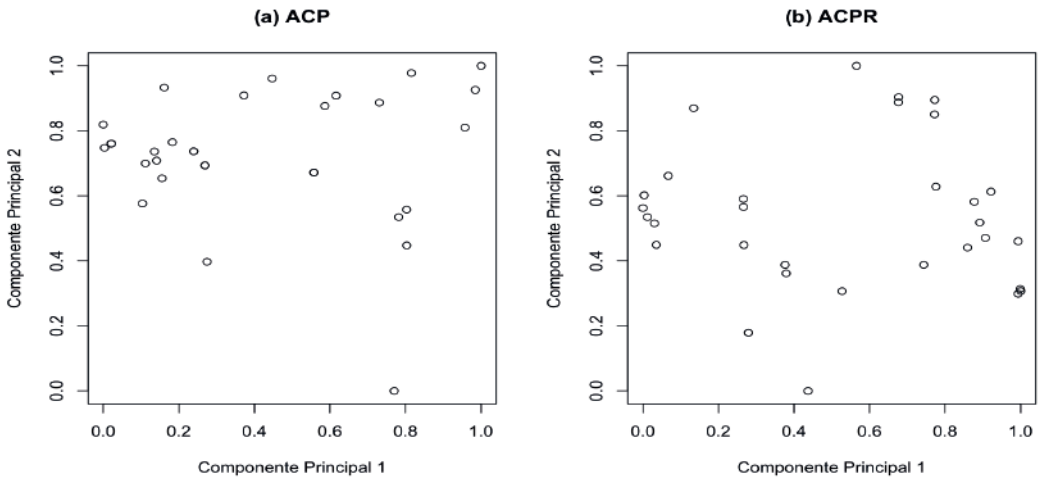

Figura 2.2. Resultados de proyección según metodologías ACP y ACPR.

2.3 Detección de valores extremos basado en la proximidad

La detección de valores extremos basado en la proximidad, se trata de un enfoque comúnmente utilizado mediante el cual se determina la proximidad entre las observaciones, si la distancia es muy corta se considera que están representando similitudes entre las observaciones. Por lo tanto, la detección del valor atípico se alcanza cuando su localidad está escasamente poblada (Su, 2011).

El caso del clúster de K medias funciona con la suposición natural de que las k distancias vecinas más cercanas de los puntos de datos atípicos son mucho mayores que las de los puntos de datos normales. Este tipo de enfoque permite una capacidad alta de distinguir entre valores atípicos débiles y fuertes en conjuntos de datos ruidosos (Aggarwal y Yu, 2001).

2.3.1 Clúster de K medias

Ye *et al.* (2006) indican que los clúster de K medias es uno de los métodos más utilizados en minería de datos, debido a su eficiencia y escalabilidad para realizar clústers en grandes conjuntos de datos. Asimismo, esta técnica considera un conjunto de clústers divididos en N muestras $X_1, \ldots, X_N$ con p variables entre K sets disconjuntos. Esta técnica requiere que el valor de K sea predeterminado de antemano y el valor determinado cambiará la composición del resultado final tal como se observa en la Figura 2.3 (Suzuki, 2021).
Una vez se ha determinado el valor de K, se realizan dos pasos específicos:

P.1 Para cada clúster $k = 1, \ldots, K$ se localiza el centro (siendo este el promedio del sector).

P.2 Para cada muestra $i = 1, \ldots, N$, se asigna el clúster para el cual cada centro es el más cercano entre los K clúster disponibles.

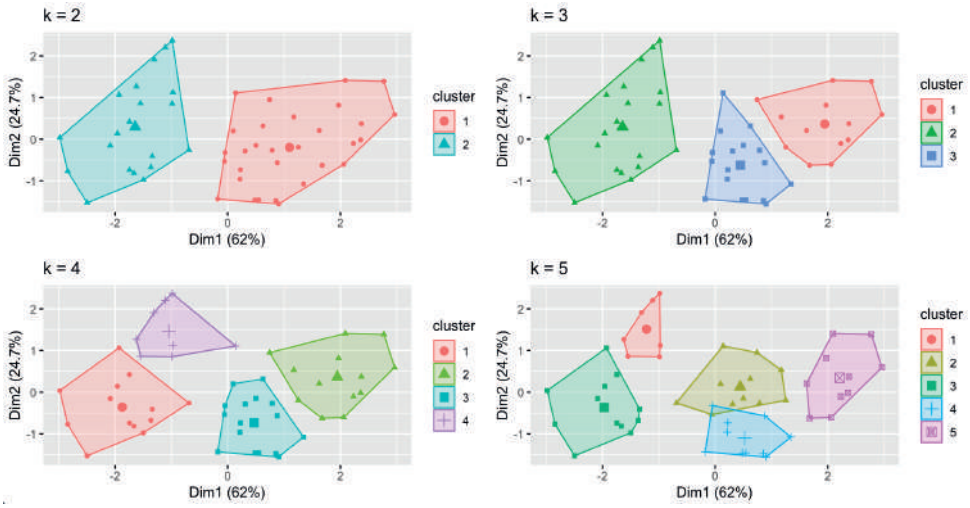

Figura 2.3. Visualización Clúster de K Medias con diferentes niveles de K.

Cada clúster es un set de vectores $p - dimensionales$ y su media aritmética corresponde al centro, una vez que se obtiene dicho valor, se evalúa la distancia euclidiana:

$$\|a - b\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_p - b_p)^2},$$

donde $a = \left[a_1, \ldots, a_p\right]^T, b = \left[b_1, \ldots, b_p\right]^T \ \epsilon \ \mathbb{R}^p$.

Además, es posible determinar el puntaje que permite minimizar la distancia intra-cluster,

$$S := \sum_{k=1}^{K} min \ z_k \ \epsilon \ \mathbb{R}^p \sum_{i \ \epsilon \ C_k} \|x_i - z_k\|^2 \tag{7}$$

donde, C_k es un set de índices i de muestras en los k-ésimo clúster y Z_k es el centroide de cada clúster k.

Es importante indicar que dicho puntaje no incrementa con cada ejecución de los clúster de K-medias. De hecho, Suzuki (2021) indica que el valor de

$$\sum_{i \ \epsilon \ C_k} \|x_i - z_k\|^2$$

observado en (7) es la suma al cuadrado de la distancia entre x y z_k, de esta manera los puntos son minimizados cuando x es elegido para ser el centro $\bar{x}_k$ del clúster k. Concretamente para cada $k = 1, \ldots, K$ el centro de cada clúster se obtiene de la siguiente forma:

$$\sum_{i \ \epsilon \ C_k} \|x_i - z_k\|^2 = \sum_{i \ \epsilon \ C_k} \left\|(x_i - \bar{z}_k) - (x_i - \bar{z}_k)\right\|^2$$

$$= \sum_{i \ \epsilon \ C_k} \|x_i - \bar{z}_k\|^2 + \sum_{i \ \epsilon \ C_k} \|x_i - \bar{z}_k\|^2 - 2\,(x - \bar{z}_k)^T \sum_{i \ \epsilon \ C_k} (x_i - \bar{z}_k)$$

$$= \sum_{i \ \epsilon \ C_k} \|x_i - \bar{x}_k\|^2 + \sum_{i \ \epsilon \ C_k} \|x_i - \bar{z}_k\|^2 \geq \sum_{i \ \epsilon \ C_k} \|x_i - \bar{z}_k\|^2,$$

donde se tiene que

$$\frac{1}{|C_k|} \sum_{i \ \epsilon \ C_k} x_i,$$

lo cual indica que el puntaje no incrementa luego de aplicar **P.1**. Asimismo, al realizar la acción de **P.2** se observa que si el clúster al cual pertenecen las muestras se cambia al clúster más cercano, el puntaje de S tampoco incrementa.

En línea con lo anterior, el resultado de los clúster de K medias depende de la selección inicial del clúster, lo cual no necesariamente permite garantizar que se obtendrá una solución óptima, por lo tanto, es necesario realizar este análisis con

diferentes valores iniciales, y seleccionar el clúster con el mejor resultado (Suzuki, 2021).

Ahora bien, si tomamos N muestras como sigue:

$$x_1 = \left[x_{1,1},\ \dots\ , x_{1,p}\right]^T,\ \dots\ ,\quad x_N = \left[x_{N,1},\ \dots\ , x_{N,p}\right]^T,$$

y se escribe como un set de muestras de índices que pertenecen a un clúster k (un subset de $\{1,\dots,N\}$), a su vez, se describe el centro como C_k y $\bar{x}_k = \left[\bar{x}_{k,1},\ \dots\ , \bar{x}_{k,p}\right]^T$, obtendríamos que para cada $k = 1,\dots,K$, se tiene la siguiente relación:

$$\frac{1}{|C_k|}\sum_{i,i'\ \epsilon\ C_k}\sum_{j=1}^{p}\left(x_{i,j} - x_{i',j}\right)^2 = 2\sum_{i\ \epsilon\ C_k}\sum_{j=1}^{p}\left(x_{i,j} - \bar{x}_{k,j}\right)^2. \tag{8}$$

Respecto a la Ecuación 8, Suzuki (2021) plantea que la misma puede expresarse como:

$$\frac{1}{|C_k|}\sum_{i\epsilon C_k}\sum_{i'\epsilon C_k}\left(x_{i,j} - x_{i',j}\right)^2 = 2\sum_{i\epsilon C_k}\left(x_{i,j} - \bar{x}_{k,j}\right)^2,$$

para $j = 1,\dots,p$. En particular se puede transformar la parte izquierda de la ecuación a lo siguiente,

$$\begin{aligned}\frac{1}{|C_k|}\sum_{i\epsilon C_k}\sum_{i'\epsilon C_k}\left\{\left(x_{i,j} - \bar{x}_{k,j}\right) - \left(x_{i',j} - \bar{x}_{k,j}\right)\right\}^2 = \\ \frac{1}{|C_k|}\sum_{i\epsilon C_k}\sum_{i'\epsilon C_k}\left(x_{i,j} - \bar{x}_{k,j}\right)^2 - \frac{2}{|C_k|}\sum_{i\epsilon C_k}\left(x_{i,j} - \bar{x}_{k,j}\right)\sum_{i'\epsilon C_k}\left(x_{i',j} - x_{k,j}\right) \\ +\frac{1}{|C_k|}\sum_{i\epsilon C_k}\sum_{i'\epsilon C_k}\left(x_{i',j} - \bar{x}_{k,j}\right)^2\end{aligned} \tag{9}$$

donde el segundo término de (9) es cero, dado que $\bar{x}_{k,j} = \frac{1}{|C_k|}\sum_{i'\epsilon C_k} x_{i',j}$, el primer el y tercer término en (9) comparte el mismo valor que $\sum_{i\epsilon C_k}\left(x_{i,j} - \bar{x}_{k,j}\right)^2$, y la suma coincide con el valor derecho de la Ecuación 8. Desde (9), se observa que los clúster de k medias busca la configuración que minimiza la suma al cuadrado de las distancias de los pares de muestras en los conglomerados.

Con base en todo lo observado, la búsqueda de espacios poco poblados es el factor determinante a la hora de detectar valores anómalos, partiendo del principio de que la atipicidad de los datos se puede ejemplificar con los clúster que tienen poca

presencia de observaciones, o bien, como lo indica Hautamaki *et al.* (2004) se define este proceso como la definición de una observación que no calza en el patrón general establecido por los clústers.

2.3.2 Agregación de K Medias

Angiulli y Pizzuti (2002) indican que las observaciones con puntajes mayores tienen vecindarios más dispersos y por lo tanto suelen ser valores atípicos más fuertes, lo cual no siempre es posible afirmarlo, ya que es posible tener observaciones con puntajes iguales pero resultados disímiles como se ve en la Figura 2.4

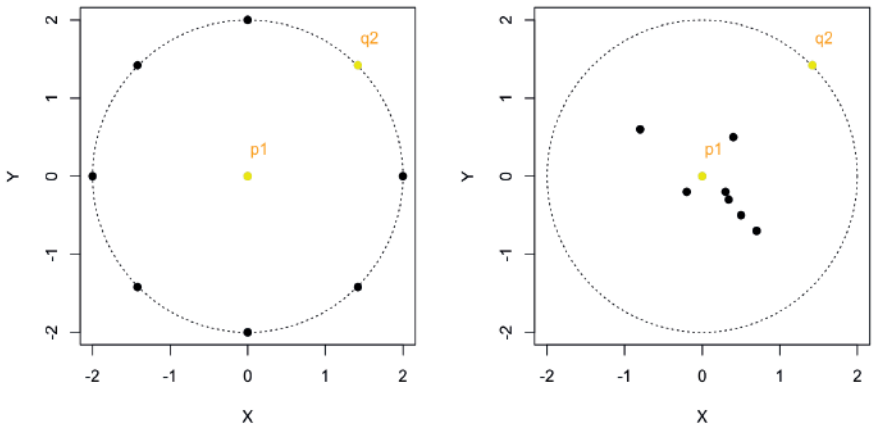

Figura 2.4. Caso con igual puntaje pero diferente identificación de anomalía.

.

Cada observación p posee una suma de distancia de todos sus vecinos más cercanos, en este caso se puede nombrar dicha suma como el peso $w_k(p)$ para cada observación, lo cual permite ranquear todas las observaciones de cada set de datos, a su vez permitiendo identificar los posibles valores anómalos.

Ahora bien, un set de datos con d dimensiones (DB), podría ajustarse a un hipercubo $D = [0,1]^d$ [11], lo cual permite mapear el comportamiento de D dentro de un intervalo $I = [0,1]$ (Angiulli y Pizzuti, 2002).

Con el objetivo de identificar los k vecinos más cercanos, se emplean dos conceptos básicos, el primero es el propuesto por Peano (1890), y mediante el cual se descubre la existencia de una curva continua que pasa a través de cada punto dentro de un

[11] Harary *et al.* (1988) definen un hipercubo (cubo n dimensional), como el producto cartesiano de dos planos. Por ejemplo, si un cubo se representa como $Q_1 = K_2$, un hipercubo se representaría como $Q_n = K_2 \times Q_{n-1}$

cuadrado cerrado, esta curva de Peano de un intervalo $I = [0, 1]$ puede verse a su vez como parte de un cuadrado cerrado $S = [0, 1]^2$.

Con base en lo anterior, Hilbert (1935), identificó el procedimiento geométrico general que permite la construcción de toda una clase de curvas que llenan espacios, con la premisa de que si existe un intervalo I que se puede mapear continuamente en el cuadrado S, es posible dividir I en cuatro subintervalos congruentes y S en cuatro subcuadrados congruentes, dando como resultados que cada subintervalo se puede mapear continuamente en uno de los subcuadrados, hasta un nivel de particionamiento infinito tal como se observa en la Figura 2.5.

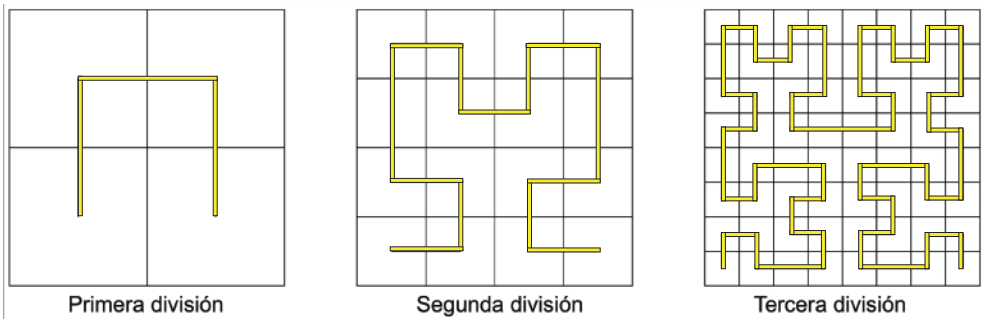

Figura 2.5. Primeros 3 niveles del llenado espacial de Hilbert, tomado de Moon *et al.* (2001)

.

Moon *et al.* (2001) explican como el llenado de espacio de Hilbert, permite obtener los K vecinos más cercanos en cada punto según el análisis de sus predecesores y sucesores en I, en este caso si se observa que hay dos observaciones cercanas en I es posible que también posean cercanía en D. A su vez, Angiulli y Pizzuti (2002) define L_t como la distancia entre dos puntos $p = (p_1 \dots p_d)$ y $q = (q_1 \dots q_d)$, esta distancia conocida como la norma-t, se define de la siguiente manera,

$$d_t(p, q) = \left(\sum_{i=1}^{d} |p_i - q_i|^t \right)^{1/t}, \tag{10}$$

donde $1 \leq t < \infty$ y además $max_{1 \leq i < d} |p_i - q_i|$ para $t = \infty$.

Con la Ecuación 10 se puede definir el peso como $w_k(p) = \left(\sum_{i=1}^{k} d_t\left(p, nn_i(p)\right) \right)$ donde $nn_i(p)$ indica el i-ésimo vecino más cercano de p en DB.

Dado un set de datos y parámetros k y n, existe un punto $p \, \epsilon \, DB$ es la n-ésima anomalía con respecto a k, y se define como $anomalia_k^n$, si hay exactamente $n - 1$

puntos q en DB tal como $w_k(q) < w_k(p)$. Ahora bien, se define $ANOM_k^n$ como el set de observaciones de anomalías máximas en DB con respecto a k.

Por lo tanto, considerando $ANOM^*$ como un conjunto de n observaciones de DB y ε como un número real positivo, se define a $ANOM^*$ como una ε-aproximación de $ANOM_k^n$ si $w^*\varepsilon \geq w^n$, donde w^* es $Min\left\{w_k(p) \mid p \ \epsilon \ ANOM^*\right\}$ y w^n es el peso de $ANOM_k^n$.

Los n puntos $ANOM_k^n$ con mayores valores de w_k son considerados como anomalías, para calcular los pesos, los k vecinos más cercanos son obtenidos usando el llenado del espacio como fue definido previamente. Es importante indicar que en las particiones de Hilbert a nivel práctico no se repiten hasta el infinito sino que tiene un punto de finalización luego de h pasos para llenar el espacio. En el caso de $h \geq 1$ y $d \geq 2$, se podría asumir un valor $\mathcal{H}_h^d$ el cual denota la aproximación de orden h de una curva de llenado de espacio de Hilbert d-dimensional que mapea 2^{hd} subintervalos de longitud $1/2^{hd}$ en 2^{hd} subhipercubos cuyos puntos centrales son considerados como puntos en un espacio de granularidad finita.

La curva de Hilbert pasa por cada observación en un espacio d-dimensional, solamente una única vez y con un mismo orden, genera un mapa entre los valores en el intervalo I y las coordinadas de cada d-dimensión. En este caso podemos decir que dado D como un conjunto donde $\left\{p \,\epsilon\, \mathbb{R}^d : 0 \leq p_i,\ 1 \leq i \leq d\right\}$ y p como un punto d-dimensional en D. En este caso la imagen inversa de p bajo el mapeo indicado es llamado el Valor de Hilbert y se denota como $\mathcal{H}(p)$, y dado que se había definido a DB como un conjunto de puntos en D, se pueden ordenar todos los puntos en función del orden por el que pasan las curvas por ellos. Se puede denotar $\mathcal{H}(DB)$ como el set $\left\{\mathcal{H}(DB) \mid p \,\epsilon\, DB\right\}$, el cual está ordenado en relación con lo inducido por la curva Hilbert.

Dado un punto cualquiera, se tendrá tanto un predecesor como un sucesor de dicha observación, el cual podríamos denotarlo como $\mathcal{H}_{pred}(p)$ y $\mathcal{H}_{suc}(p)$ en $\mathcal{H}(DB)$ son los puntos más cercanos con respecto al orden inducido por la curva de Hilbert. Finalmente, el m-ésimo predecesor y sucesor de p son definidos como $\mathcal{H}_{pred}(p, m)$ y $\mathcal{H}_{suc}(p, m)$.

En línea con lo anterior, Angiulli y Pizzuti (2002) indican que una región r es un hipercubo abierto en $[0, 2]^d$ con una longitud de lado $r = 2^{1-l}$ teniendo la forma $\prod_{i=0}^{d-1}\left[a_i r, (a_i + 1)\, r\right]$ donde cada a_i se ubica en un intervalo tal que, $0 \leq i \ \leq d$,

asimismo, l se ubica en $\mathbb{N}$.

En caso de tener dos puntos, p y q, se denota por $MinReg\,(p,q)$ el lado más pequeño de la región r que contiene tanto a p como a q, asimismo, $MaxReg\,(p,q)$ será el lado más grande que contiene a p pero no a q, por lo tanto se obtiene lo siguiente,

$$MinReg\,(p,r) = min_{i=1}^{d}\left\{min\left\{p_i,\ mod\, r, r - p_i\, mod\, r\right\}\right\} \tag{11}$$

$$MaxDist\begin{cases} \sum_{i=1}^{d} max\left\{p_i, mod\, r, r - p_i\, mod\, r\right\}^{\,t1/t} & para\ 1 \leq t < \infty \\ max_{i=1}^{d}\left\{max\left\{p_i,\ mod\, r, r - p_i\, mod\, r\right\}\right\} & para\ t = \infty \end{cases} \tag{12}$$

Donde $x\ mod\ r\ =\ x\ -\ \left[x/r\right]\ r$ y p_i denota el valor de p a lo largo de la i-ésima coordenada, son respectivamente la distancia perpendicular desde p a la cara más cercana de la r-región de lado r que contiene p, es decir, un límite inferior para la distancia entre p y un punto que se encuentra fuera de la región r anterior, y la distancia desde p hasta el vértice más lejano de la r-región del lado r que contiene p, es decir, un límite superior para la distancia entre p y un punto que se encuentra en la región r anterior.

Puede verse ejemplificado en la Figura 2.6 donde se observa como el proceso iterativo aprovecha cada característica del punto para determinar de manera rápida si ya se han encontrado los k vecinos más cercanos exactos.

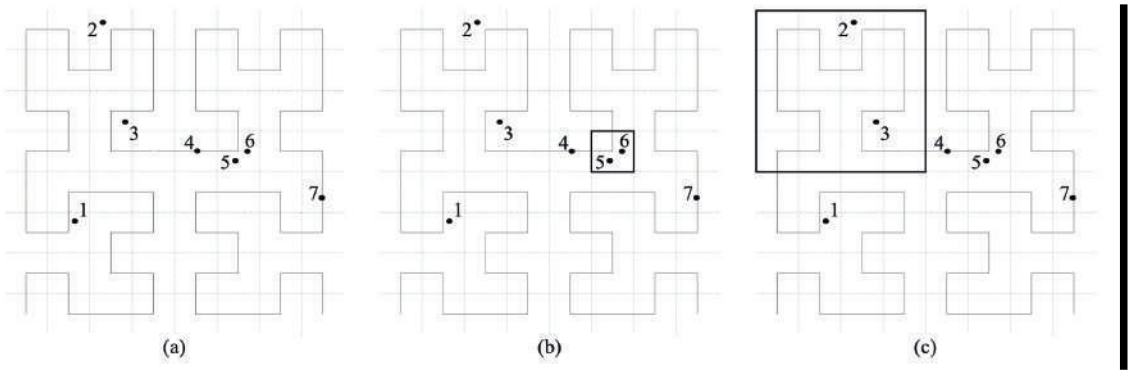

Figura 2.6. Identificación de anomalías bajo principio del llenado espacial de Hilbert, tomado de Angiulli y Pizzuti (2002)

.

2.4 Detección de valores extremos basado en densidad

La detección de valores extremos basado en densidad, se realiza cuando en una localidad a pesar de existir diversas observaciones, todas estas pueden ser atípicas. Bajo esta premisa, la técnica del valor atípico local (en adelante LOF[12]), identifica cuán aislado está el objeto con respecto al vecindario circundante.

Breunig *et al.* (2000) identificaron que esta técnica permite encontrar valores atípicos significativos, pero que de otra manera no pueden identificarse. Asimismo, Ramaswamy *et al.* (2000) explican como la noción de valores atípicos basados en la distancia se amplía y puede ser usada utilizando la distancia al vecino k más cercano para clasificar los valores atípicos, este tipo de clasificación se centra principalmente en el agrupamiento y no necesariamente está optimizado para la detección de valores atípicos. Podría evitar que las anomalías débiles sean correctamente clasificadas, ya que usualmente son ignoradas, o bien pueden identificarse con un resultado dicotómico, no permitiendo ver el nivel de desviación que posee la observación.

Intuitivamente la técnica de LOF, busca que las observaciones puntúen según el grado en que su densidad se desvía con respecto a sus vecinos, es por lo tanto más probable que una observación con una puntuación más alta que otras en el conjunto de datos de prueba sea una anomalía (Ghafoori, 2018).

2.4.1 Valor Atípico Local

Breunig *et al.* (2000) introdujo el concepto de LOF, para la identificación de valores atípicos en un conjunto de datos multivariados, una de las ventajas que da este tipo de técnica es que permite otorgar un puntaje con el cual se cuantifica el grado de atipicidad, en este caso se denomina como un valor atípico local ya que está restringido a la vecindad existente en el contexto de cada observación.

Se propone una distancia entre dos objetos p y q[13], asimismo, C representa un clúster o set de datos, lo cual podría simplificarse indicando que $d\,(p, C)$ denota la distancia mínima entre dos objetos p y q,

[12]Se utilizarán las siglas de Local Outlier Factor (LOF) para la identificación de esta técnica.

[13]$d\,(p, q)$ representa la notación a utilizar

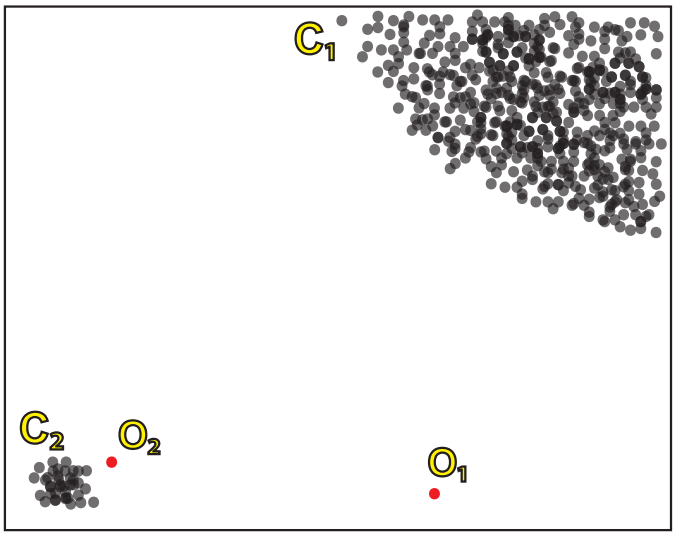

Figura 2.7. Composición de clúster y anomalías, según Breunig *et al.* (2000)

.

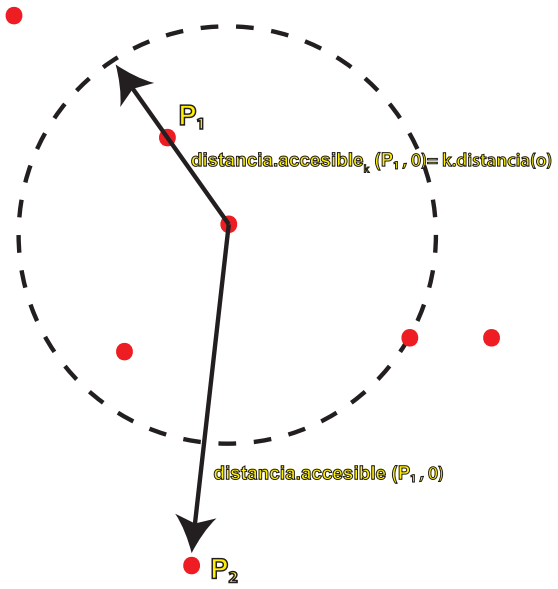

Figura 2.8. Distancia mínima y máxima accesible para K=4

.

$$d\left(p,q\right) = min\left\{d\left(p,q\right) \mid q \,\epsilon\, C\right\}. \tag{13}$$

Lo observado en la Ecuación (13) no permite asegurar que la identificación de anomalías sea completa o correcta, ya que existen complejidades derivadas de los datos a analizar, y la múltiple existencia de clústeres dificultaría una identificación única

del clúster C, como se observa en la Figura 2.7

En consecuencia con lo anterior, se identifica un factor de anomalía, como una $k\,distancia$ de un objeto p, siendo esta la distancia $d\,(p,o)$ entre p y un objeto o el cual pertenece a D, cumpliendo dos condiciones básicas:

- Para al menos k objetos $o' \,\epsilon\, D \setminus \{p\}$ se mantiene que $d\left(p,o'\right) \leq d\,(p,o)$
- Para un máxinmo de $k-1$ objetos $o' \epsilon\, D \setminus \{p\}$ se mantiene que $d\left(p,o'\right) < d\,(p,o)$

Con lo anterior, se podría indicar que dada una $k - distancia$ de p, la $k\,distancia$ del vecindario de p contiene cada objeto cuya distancia desde p no es mayor que la $k\,distancia$, por lo tanto todos los objetos q son k vecinos cercanos de p,

$$N_{k.distancia(p)}\,(p) = \left\{q \,\epsilon\, D \setminus \{p\} \mid \; d\,(p,q) \leq k.distancia(p)\right\} \tag{14}$$

Además se identifica la distancia accesible de un objeto p a un objeto o, lo cual se define como,

$$distancia.accesible_k(p,o) = max\left\{k\,distancia(o),\; d\,(p,q)\right\} \tag{15}$$

Con las definiciones propuestas en (14) y en (15), se puede profundizar en dos nociones finales para la aplicación de la técnica de LOF, primero la definición del parámetro $MinPts$ el cual define un mínimo número de objetos; y segundo la especificación del volumen mediante un parámetro. Es justamente con la combinación de estos dos elementos que es posible generar un umbral con el cual el algoritmo opera (Breunig *et al.*, 2000).

Ahora bien, con el objetivo de detectar anomalías basadas en la densidad, se hace necesario comparar las densidades de diferentes conjuntos de objetivos, tomando lo indicado en le Ecuación (14) se puede definir $MinPts$ como el único parámetro a usar, permitiendo tener una medida de volumen de la densidad del vecindario de un objeto p.

La definición propuesta por Breunig *et al.* (2000) para hablar de la densidad de accesibilidad local (en adelante IRD [14]), es la siguiente:

[14]Siglas de Local Reachability Density

$$IRD_{MinPts}(p) = 1/\left[\frac{\sum\left(o\ \epsilon\ N_{MinPts}(p) distancia\ alcanzable_{MinPts}(p,o)\right)}{\left|N_{MinPts}(p)\right|}\right] \quad (16)$$

El IRD de un objeto p es la inversa del promedio de la distancia de accesibilidad local basada en los $MinPts$ vecinos más cercanos de p, es importante notar que en el caso de que existan valores duplicados en el conjunto de datos, las distancias accesibles se volverían 0, con lo cual el valor del IRD sería infinito.
Con la Ecuación (16), es posible definir el Valor Atípico Local:

$$LOF_{MinPts}(p) = \frac{\sum\left(o\ \epsilon\ N_{MinPts}(p)\frac{IRD_{MinPts}(o)}{IRD_{MinPts}(p)}\right)}{\left|N_{MinPts}(p)\right|} \quad (17)$$

El factor de anomalía del objeto p captura el grado de anormalidad, además, es posible identificar el hecho de que mientras mayor sea el valor de las densidades de accesibilidad local de los $MinPts$ vecinos más cercanos de p o bien más bajo el valor de IRD de p, mayor será el resultado de LOF para el caso de dicha observación.
Partiendo del principio expresado en la Ecuación (13), se podría relacionar la distancia mínima accesible de una serie de datos en un clúster

$$distancia.minima.accesible = min\left\{distancia.accesible(p,q) \mid p,q\ \epsilon\ C\right\}$$

Asimismo, se podría identificar la distancia máxima alcanzable, con lo cual es posible definir un parámetro tal que, $\varepsilon = \frac{distancia.maxima.accesible}{distancia.minima.accesible} - 1$, lo cual podría llevarnos a dos conclusiones puntuales

- todos los q $MinPts$ *vecinos cercanos* de p están en C
- todos los o $MinPts$ *vecinos cercanos* de q están también en C

Usando lo anterior, tendríamos que $1/(1+\varepsilon) \leq LOF(p) \leq (1+\varepsilon)$, con lo cual intuitivamente si la observación que está siendo analizada está dentro de un clúster en el centro de todos sus vecinos, el valor de ε tenderá a ser 0, con lo cual el resultado de LOF será cercano a 1.

En el caso de los clúster de K medias, la detección de anomalías pasa por la detección de distancias poco pobladas según las distancias mínimas accesibles, partiendo nuevamente del principio de que la atipicidad de los datos pasaría por ser un aislante de dichas observaciones en los planos (Breunig *et al.*, 2000).

2.5 Detección de valores extremos según subespacios paralelos al eje

Aggarwal (2017) define la detección de valores extremos según subespacios paralelos al eje, como aquella metodología donde un valor atípico puede ser aislado debido a su ubicación. En el caso de la técnica de bosques de aislamiento propuesta por Liu *et al.* (2008), su premisa nace del concepto de que las anomalías son pocas y diferentes, lo cual las hace más susceptibles al aislamiento que aquellas observaciones consideradas como normales.

Utilizando una estructura de árbol de clasificación[15] se considera que las anomalías se aíslan más cerca de la raíz, mientras que los puntos normales están aislados en el extremo más profundo del árbol, lo cual requiere de un mayor trabajo para lograr su separación, a su vez, si se realiza una múltiple aplicación de árboles de aislamiento se obtiene un bosque de aislamiento que ha demostrado buenos resultados en un contexto de alta dimensionalidad.

2.5.1 Bosques de aislamiento

Los valores atípicos a menudo están incrustados en subespacios con el potencial de ser fácilmente identificables, pero por diversos factores podrían tener enmascaramiento que dificulten su interpretabilidad, lo cual hace que sea necesario la búsqueda de nuevos subespacios. Es por lo tanto evidente que a mayor dimensionalidad de los datos, mayor es el número de posibles proyecciones necesarias para identificar valores de interés (Aggarwal, 2017).

Liu *et al.* (2008) explican como el modelos de Bosques de Aislamiento [16] usa Árboles

[15]Se realizan los árboles de clasificación propuestos por Breiman *et al.* (2017).

[16]También definido como Isolation Forest

de aislamiento para alistar datos, los cuales permiten dividir recursivamente mediante cortes paralelos al eje en puntos de partición elegidos aleatoriamente, de esta manera se aislan las posibles anomalías; mientras mayor sera el número de nodos que tengan que utilizarse para separar un dato, menos probable es que este dato sea considerado una anomalía, de manera inversa, los valores extremos se espera que puedan ser aislados de manera mucho más acelerada.

Las diferentes ramas corresponden a diferentes regiones locales del subespacio de los datos, mientras menos ramas utilizadas, menor es la dimensionalidad de los subespacios en los que se han aislado los valores atípicos, en esta línea Aggarwal (2017) indica que la longitud del camino está altamente correlacionada con la dimensionalidad del subespacio utilizado para el aislamiento. Para un conjunto de datos que contiene $2^8 = 256$ puntos [17], la profundidad promedio del árbol será 8, pero un valor atípico a menudo puede aislarse en menos de tres o cuatro divisiones. Es importante indicar que la longitud del camino desde la raíz hasta la hoja se usa como puntaje de atipicidad, permitiendo visualizarse fácilmente en la Figura 2.9, donde la línea roja representa una observación aislada fácilmente, mientras que la línea azul representa un mayor esfuerzo requerido para lograr un aislamiento total de la observación. Este

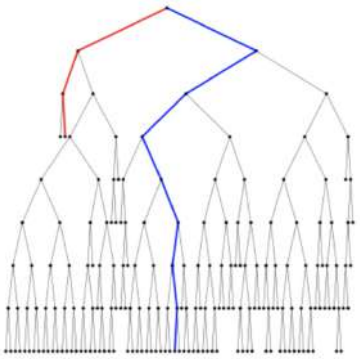

Figura 2.9. Representación de un árbol de aislamiento, tomado de Hariri *et al.* (2019).

proceso de ramificación se realiza de forma recursiva sobre el conjunto de datos hasta que se aísla un solo punto o se alcanza un límite de profundidad predeterminado[18].

[17] Liu *et al.* (2008) establece como el tamaño de submuestra recomendado.

[18] Se define un valor de profundidad determinado ya que en conjuntos de datos altamente concentrados el nivel de profundidad requerida para aislar las observaciones de manera completa requiere un esfuerzo computacional elevado, esta una de las ventajas de esta técnica ya que permite limitar el esfuerzo computacional con poco impacto sobre la calidad de la detección.

Ese proceso es repetido de manera continua con otras submuestras aleatorias hasta completar lo que podría llamarse un bosque de aislamiento.
La profundidad promedio de las ramas que atraviesa este punto de datos se traducirá a un puntaje de anomalía usando la siguiente ecuación.

$$s\left(x,n\right)=2^{\frac{-E\left(h(x)\right)}{c(n)}},$$

donde $E\left(h\left(x\right)\right)$ es el valor promedio de la profundidad de un punto x alcanzado en todos los arboles, mientras que $c\left(n\right)$ se trata de un factor de normalización definido como la profundidad promedio de una búsqueda fallida de un árbol binario.

$$c\left(n\right)=2H\left(n-1\right)-\frac{2\left(n-1\right)}{n}$$

La manera de realizar el aislamiento de los puntos se da mediante una división recurrente con cortes aleatorios, los mismos se generan considerando los valores mínimos y máximos de los datos, como se puede observar en la Figura 2.10. Finalmente,

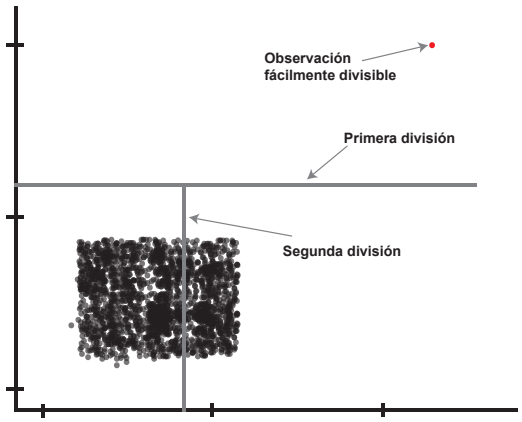

Figura 2.10. Representación gráfica del mapeo realizado por la técnica de árboles de aislamiento, tomado de Aggarwal (2017)

.

Aggarwal (2017) explica como en el caso de que un punto requiera un alto trabajo para aislar la observación (profundizar mucho en la estructura del árbol), es posible detener el proceso de ramificación y se asigna a la observación el valor de profundidad máxima. Para Liu *et al.* (2008) los puntajes pueden clasificarse de manera segura en tres categorías teóricas como se definen en el Cuadro 2.1.

Cuadro 2.1. Establecimiento de umbrales para las anomalías

Valor estimado	Resultado de anomalía
Valor cercano a 1	Existencia de anomalías
Valores menores a 0.5	Observaciones normales
Todos los valores cercanos a 0.5	No existen distinciones claras de anormalidades

2.5.2 Bosques de aislamiento extendidos

Para Aggarwal (2017) uno de los problemas que presenta el método de Bosques de Aislamiento, es que en ocasiones la generación de líneas paralelas al eje podrían generar espacios fantasmas que podrían ser identificados como posibles áreas de concentración para puntos normales, es por lo tanto más intuitivo al ver la representación visual presentada por Hariri *et al.* (2019) y observable en la Figura 2.11, en dicha figura se puede identificar en la sección (b) como parecen crearse sombras en las secciones superior derecha e inferior izquierda, lo cual podría interpretarse como una evidencia de posibles observaciones.

Con base en lo anterior se realiza la aplicación de la técnica Extendida de Bosques

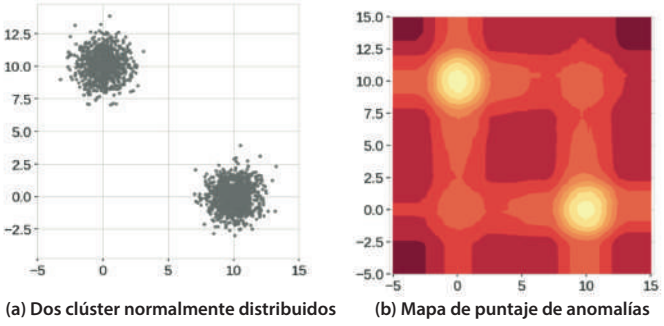

Figura 2.11. Distribución de dos clúster y mapa de calor generado según los puntajes obtenidos, extraído de Hariri *et al.* (2019)

.

de Aislamiento[19], el cual se basa en el uso de hiperplanos con pendientes aleatorias. Con este método se permite obtener puntajes de anomalías más confiables y robustos, e incluso en algunos casos una detección más precisa de la estructura de un conjunto de datos dado, esta técnica permite producir un mapa más simétrico en el cual se concentren los datos, evitando generar bloques de información donde en

[19]Extended Isolation Forest en su nombre original.

el cual se de una posible identificación de puntos anómalos pudiendo considerarse como normales (Hariri *et al.*, 2019).

Esta técnica puede ejemplificarse visualmente en el caso de la Figura 2.12 donde se observa como es posible el aislamiento aún con pendientes aleatorias.

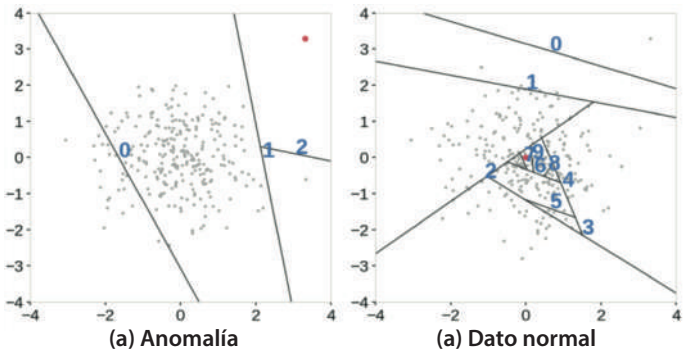

Figura 2.12. Método de aplicación de la versión extendida de los bosques aleatorios, extraído de Hariri *et al.* (2019)

.

2.5.3 Bosques de aislamiento con criterio de selección dividida

Liu *et al.* (2010) detectaron escenarios en los cuales los Bosques de Aislamiento podrían no identificar valores anómalos, por ejemplo, cuando las anomalías forman grupos, se genera un efecto de enmascaramiento debido a su proximidad y densidad, lo cual hace que se vuelvan más difíciles de detectar.

Ante lo anterior, una posible solución propuesta es utilizar una variación de los Bosques de Aislamiento conocida como SCiForest, según los autores este método ofrece un mejor rendimiento de detección a un bajo costo de procesamiento adicional.

Para construir SCiForest se requieren tres elementos principales, el cálculo de los valores del hiperplano, seguido de la clasificación de los valores del hiperplano y finalmente el cálculo del criterio.

Los atributos individuales no siempre son efectivos para separar las anomalías, bajo esta premisa con el método de SCiForest no es necesario tener el hiperplano óptimo en cada nodo desde el inicio del cálculo, ya que mediante suficientes pruebas de hiperplanos generados aleatoriamente se podrá aproximar finalmente el mejor

hiperplano (Liu *et al.*, 2010).

Se determina que en cada división llevada a cabo para la construcción del árbol, un hiperplano de separación f es construido tal como se ve en la Ecuación 18, utilizando el mejor punto de división p y el mejor hiperplano que produce la ganancia de desviación estándar más alta entre los hiperplanos τ que han sido generados aleatoriamente.

$$f(x) = \sum_{j \ \epsilon \ Q} c_j \frac{x_j}{\sigma\left(X'_j\right)} - p, \tag{18}$$

donde Q tiene q índices de atributos seleccionados aleatoriamente sin reemplazo de $\{1, 2, \ldots, d\}$; c_j es un coeficiente aleatoriamente seleccionado entre $[-1, 1]$; X'_j es el j-ésimo valor de atributo de X'.

Luego de que es construido el valor de f, se realiza una evaluación donde es posible obtener dos posibles valores:

$$\begin{aligned} X^l &\leftarrow \left\{x \ \epsilon \ X' \mid f(x) < 0\right\} \\ X^r &\leftarrow \left\{x \ \epsilon \ X' \mid f(x) \geq 0\right\} \end{aligned} \tag{19}$$

De la Ecuación 19, se obtienen dos subconjuntos de datos, donde $X^l \cup X^r = X'$, de acuerdo con f. Este proceso de construcción de árboles continúa recursivamente con los subconjuntos filtrados X^l y X^r hasta que el tamaño de un subconjunto sea menor o igual a dos.

2.5.4 Bosques de aislamiento con criterio de ganancia

Liu *et al.* (2010) proponen que para el empleo de esta técnica se separan los hiperplanos utilizando el criterio de ganancia, el cual parte del supuesto de que los clusteres de las anomalías son propensas a tener su propia distribución, de esta manera el criterio de separación busca aislar dichos puntos según su propia distribución, como se ve en la Figura 2.13.

Adicionalmente, el criterio de separación puede definirme matemáticamente con la Ecuación 20, donde Y es un set real de valores obtenidos al proyectar X sobre un hiperplano f.

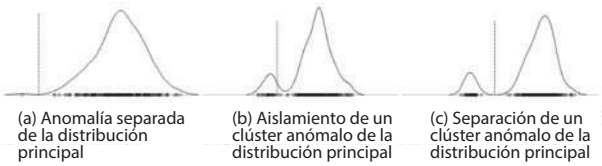

Figura 2.13. Ejemplo de división según distribución, extraído de Liu *et al.* (2010)

.

$$Sd_{ganancia}\left(Y\right) = \frac{\sigma\left(Y\right) - avg\left(\sigma\left(Y^{l}\right), \sigma\left(Y^{T}\right)\right)}{\sigma\left(Y\right)} \tag{20}$$

Es importante notar que $Y^{l} \cup Y^{T} = Y$, asimismo, $\sigma\left(.\right)$ es la función de la desviación estándar, y $avg\left(\left(\sigma\left(Y^{l}\right), \sigma\left(Y^{T}\right)\right)\right)$ simplemente es el promedio de ambos valores. Además, se requiere de un valor de división p, el cual es requerido para separar Y en Y^{l} y Y^{r}, tal que $y^{l} < p \leq y^{r}$, donde $y^{l} \epsilon Y^{l}$ y además $y^{r} \epsilon Y^{r}$.

Con todo lo anterior, la división se realiza con el valor de p que brinde la mayor ganancia de Sd, entre todas las combinación de Y^{l} y Y^{r}. Es importante notar que el valor de $\sigma\left(Y\right)$ es utilizado para normalizar el criterio, y permitir realizar comparaciones entre las diferentes escalas.

El empleo de esta técnica posee un alto potencial para hacer la separación de los conglomerados de anomalías aún cuando este se encuentre muy cerca de los conglomerado normales, sin representar un alto costo a nivel de procesamiento (Liu *et al.*, 2010).

2.6 Aproximación y Proyección de Manifolds Uniformes

La técnica de Aproximación y Proyección de Manifolds Uniformes (UMAP por sus siglas en inglés), es una manera de realizar proyección no lineal para la reestructuración para la Reducción de Dimensionalidad, esto se realiza mediante la localización de distancias en un espacio con muchas variables y su posterior reproducción en un espacio de baja dimensión.

Una ventaja importante de la técnica propuesta por McInnes *et al.* (2018) es que no existen restricciones computacionales en cuanto a la dimensión, lo cual permite aprovechar la potencia disponible en la actualidad con las técnicas de aprendizaje

de máquinas, haciendo uso de esta técnica la cual es escalable para datos masivos.
En términos globales los algoritmos de reducción de dimensionalidad tienden a dividirse en dos categorías, aquellos que buscan preservar la estructura de distancia par a par entre todas las muestras de datos y aquellos que favorecen la preservación de las distancias locales[20] sobre la distancia global, en este caso tal como lo indica Fernandez y Plumbley (2021) UMAP conserva tanto la estructura local como la global.

McInnes *et al.* (2018) indica que las bases teóricas para UMAP se basan en gran medida en la teoría de complejos simples y el análisis topológico de datos, UMAP utiliza aproximaciones locales de variedades y une sus representaciones locales para construir una representación topológica de los datos de alta dimensión.

Los complejos simpliciales son una forma de construir un objeto de k dimensiones, de esta manera un símplex de dimensión k se llama k-simplex y se forma tomando la envolvente convexa de k+1 puntos independientes. De esta forma, un 0-simple es un punto, un 1-simple es un segmento de línea (entre dos ceros simplex), un 2- simple es un triángulo (con tres 1-simples como "caras") y un 3-simple es un tetraedro (con cuatro 2- simples como "caras"), tal como se observa en la Figura 2.14.

Figura 2.14. Ejemplo de simplex de baja dimensión McInnes (2023)

.

Además de lo anterior, McInnes (2023) emplea el Teorema del Nervio[21], esto permite proporcionar una justificación de que este proceso captura la estructura topológica en los datos. Desde el punto de vista computacional, McInnes *et al.* (2018) explica como el algoritmo funciona en términos de conjuntos simpliciales difusos, lo cual permite describirse en términos de construcción y operaciones en grafos ponderados,

[20]En el caso de UMAP se emplea la geometría Riemanniana, la cual es una rama de la geometría diferencial que estudia las variedades Riemannianas, la ventaja que esto conlleva es que permite medir distancias locales, es decir, tener nociones de ángulo, longitud de curvas, superficie y volumen que pueden ser locales.

[21]El complejo simplicial será homotópicamente equivalente a la unión de la cobertura, dos espacios X e Y son homotópicamente equivalentes si pueden transformarse entre sí mediante operaciones de flexión, contracción y expansión.

en este caso se puede describir en dos fases:

- Construcción de un grafo ponderado k-vecino particular
- Posterior diseño de baja dimensión de este grafo el cual permita optimizar su función objetivo

Asimismo, McInnes *et al.* (2018) indican que el comportamiento real de los datos no es tan reelvante, si se asume que la variedad tiene una métrica Riemanniana no heredada del espacio ambiental, lo cual podría permitir encontrar una métrica tal que los datos estén aproximadamente uniformemente distribuidos con respecto a esa métrica.
Si los datos se distribuyeran uniformemente en la variedad, sería fácil seleccionar un radio adecuado, la distancia promedio entre puntos funcionaría bien, asimismo, con una distribución uniforme, se tiene la garantía de cubrir sin "huecos" y sin componentes desconectados innecesariamente. Tal como se observa en la Figura 2.15.

(a). Datos originales **(b).** Grafos con pesos combinados

. .

Figura 2.15. Ejemplos del empleo de la técnica McInnes (2023)

En este caso la construcción de un grafo ponderado de k-vecinos, donde
$X = \{x_1, \dots x_N\}$ es el set de datos de entrada con una medida de similitud tal $d : X \times X \to \mathbb{R}_{\geq 0}$.
Dado un hiperparámetro k, para cada x_i se calculam el conjunto $X = \{x_{i1}, \dots x_{ik}\}$ de los k vecinos más cercanos de x_i bajo la métrica d.
Para cada x_i, se define ρ_i y σ_i donde,

$$\rho_i = min \left\{ d\left(x_i, x_{ij}\right) \mid 1 \leq j \leq k, d\left(x_i, x_{ij}\right) > 0 \right\} \tag{21}$$

además, se define σ_i para ser un valor tal que,

$$\sum_{j=1}^{k} exp\left(\frac{-max\left(0, d\left(x_i, x_{ij}\right) - \rho_i\right)}{\sigma_i}\right) = log_2\left(k\right) \tag{22}$$

En este caso, la selección σ_i deriva de la restricción de conectividad local además x_i conecta al menos con otro punto de datos con una arista de peso 1. Esto es equivalente a que el conjunto simplicial difuso resultante esté localmente conectado en x_i.

Con lo anterior se puede definir un grafo dirigido ponderado $\bar{G} = (V, E, w)$, donde Los vértices V de $\bar{G}$ son simplemente el conjunto X. Luego, se forma el conjunto de aristas dirigidas $E = \left\{\left(x_i, x_{ij}\right) \mid 1 \leq j \leq k, 1 \leq i \leq N\right\}$ y se define la función de peso w estableciendo:

$$w\left(x_i, x_{ij}\right) = exp\left(\frac{-max\left(0, d\left(x_i, x_{ij}\right) - \rho_i\right)}{\sigma_i}\right) \tag{23}$$

Las fuerzas repulsivas se calculan mediante muestreo debido a las limitaciones computacionales. Así, cada vez que se aplica una fuerza atractiva a una arista, uno de los vértices de esa arista es repelido por una muestra de otros vértices.

Tal como se ha indicado UMAP utiliza un algoritmo de diseño de grafo dirigido por fuerzas en un espacio de baja dimensión, donde se utilizan un conjunto de fuerzas atractivas aplicadas a lo largo de las aristas y un conjunto de fuerzas repulsivas aplicadas entre los vértices, el algoritmo procede iterativamente aplicando fuerzas atractivas y repulsivas en cada arista o vértice, esto es equivalente a un problema de optimización no convexo, donde la convergencia a un mínimo local está garantizada mediante la disminución lenta de las fuerzas atractivas y repulsivas.

Autores como Fernandez y Plumbley (2021) emplean la suposición de que si dos regiones parecen separables en la proyección UMAP, también son separables en la representación original, esta separabilidad de la proyección podría facilitar los análisis posteriores, y por lo tanto mejorar la calidad de la detección. Fernandez y Plumbley (2021) indicaron que la alternativa definida como más popular para evaluar los resultados de la proyección fue la técnica de K vecinos más cercanos, con lo cual para el caso de la población no supervisada empleada en la presente investigación

se emplearánn las técnicas de K medias.

Respecto a las debilidades de esta técnica, McInnes *et al.* (2018) indican que UMAP carece de la fuerte interpretabilidad del Análisis de Componentes Principales (APC), ya que las dimensiones del espacio no tienen un significado específico, a diferencia de APC donde las dimensiones son las direcciones de mayor varianza en los datos fuente. Asimismo, UMAP se basa en la distancia entre observaciones en lugar de las características fuente, lo cual no tiene un equivalente de cargas factoriales que técnicas lineales como APC o Análisis de Factores pueden proporcionar.

A pesar de lo anterior, a medida que se muestrea más datos, la cantidad de estructura evidente proveniente del ruido tiende a disminuir y UMAP se vuelve más robusto, sin embargo, se debe tener cuidado con tamaños de muestra pequeños de datos ruidosos o datos con solo estructura de variedad a gran escala.

En el estudio propuesto por Fernandez y Plumbley (2021), se realizó un empleo de poblaciones muy desequilibradas, a pesar de esto la técnica de UMAP demostró mejores resultados que el resto de métodos empleados, tanto a nivel de resultados como de velocidad de procesamiento.

2.7 Distribución en el conjunto de los datos

Para Emmott *et al.* (2015) a medida que aumenta la dimensionalidad de los datos, el volumen que contiene los datos también aumenta, lo cual aumenta el riesgo de que los puntos normales caigan en las colas de la distribución.

En este contexto la limpieza previa de la información, podría generar un aumento razonable de la confianza en la calidad de la evaluación, como parte del procesamiento, autores como Campos *et al.* (2016) indican que si un atributo tiene menos del 10% de valores faltantes, esas instancias se eliminan, de lo contrario, se elimina el individuo en sí. Para el caso del estudio acá presentado no se genera una limpieza de ese tipo ya que el procesamiento previo genera un set de datos completos.

Adicional a lo anterior, es necesario conocer la distribución de la información a analizar, Campos *et al.* (2016) evaluaron el impacto de la normalización en los métodos de detección de valores atípicos, donde se encontró que los datos normalizados dan valores de rendimiento más altos en comparación con el rendimiento en conjuntos de datos no normalizados, otros autores como Zimek *et al.* (2012) indican

que el preprocesamiento de datos afecta considerablemente el resultado, pero dicha mejora en su rendimiento se da cuanto los datos han sido previamente normalizados entre 0 y 1, o bien cuando se esperan atributos estandarizados como la media cero y variancia 1.

En general existe un consenso donde se hace evidente la necesidad de lograr un balance entre una transformación que implique una ligera pérdida de información, en línea con esto para autores como Campos *et al.* (2016) existe un riesgo de normalizar la información, ya que variar la elección de las clases a partir de las cuales se extraen los valores atípicos puede llevar a resultados experimentales sustancialmente diferentes.

Es por lo tanto necesario lograr un balance para lograr una mejora en el poder de predicción de los resultados, principalmente porque algunos métodos de detección se ven más afectados que otros, y por lo tanto su rendimiento relativo no cambia drásticamente.

En el presente estudio se buscará evaluar el efecto que tiene la normalización en el rendimiento de la detección de outliers, tal como fue planteado por autores como Campos *et al.* (2016); Kandanaarachchi *et al.* (2020); Zimek *et al.* (2012). A pesar de esto, el análisis integral de diversas metodologías de normalización, excede los alcances de la presente tesis, sin embargo, para cada conjunto de datos analizado se incluirán 3 variantes en el análisis, el uso de los datos sin normalización, con normalización logarítmica y con aplicación de detección anómala con poblaciones que poseen una distribución gamma.

2.8 Antecedentes de la Contratación Pública

Los principios rectores de la Contratación Pública se pueden rastrear desde la misma concepción de la Constitución Política Costarricense, donde de los análisis de las actas se puede extraer el espíritu de los constituyentes, quienes ordenan tutelar el interés general, para proteger la plena satisfacción de los ciudadanos que son los que se favorecerán de la actuación eficiente de cada sujeto público Monge (2007). Para lograr esta eficiencia se define en el artículo 182 de la Constitución Política la licitación como el mecanismo para la adquisición de las necesidades del Estado.

En línea con lo anterior, Costa Rica ha establecido un proceso de promoción de las contrataciones públicas mediante el uso de una plataforma centralizada, específicamente con la promulgación de la Ley sobre Transparencia de las contrataciones administrativas por medio de la reforma del artículo 40 y de la adición del artículo 40 bis a la Ley número 7494, Ley número 9395); se estableció que toda la actividad de contratación regulada por la Ley General de Contratación Administrativa, Ley número 7494; o bien por un régimen especial, deberá realizarse mediante el sistema digital unificado de compras públicas[22] (DFOE, 2021).

El SICOP fue establecido mediante el Decreto Ejecutivo número 38830-H-MICITT en el año 2015, en el cual se dispone a esta plataforma para la tramitación de procedimientos de contratación administrativa, asimismo, se establece que Radiográfica Costarricense (RACSA) será la empresa proveedora del sistema. Ya para finales del año 2020, un 93% de las instituciones del Sector Público ya eran usuarias de la plataforma SICOP, siendo esta plataforma la más utilizada por las proveedurías del Estado.

En cuanto a los ahorros y beneficios en el uso del SICOP, la CGR (2019) estimó que en el año 2017 se generó un ahorro que representa un 20,8% del total de compras públicas registradas en SIAC y un 0,9% del PIB durante ese año, asimismo, de haber incoporado durante dicho año el 100% de procedimientos pendientes de contratación administrativa tramitados por otros medios, el potencial de ahorro estimado con respecto al PIB hubiera sido de 1,55%.

Dado lo anterior, determinar la eficiencia y eficacia del gasto publico es necesario no solo por los recursos pecuniarios comprometidos, sino también por la credibilidad fiduciaria de la gestion del sector publico.

A nivel del sector público costarricense, y en función de la presente investigación se identifica el proceso de contratación pública en la Figura 2.16, donde se consideran las etapas comprendidas entre el inicio del proceso y la recepción del bien.

En términos usuales los procesos de auditoría pasan por un muestreo estadístico necesario por la eficiencia y eficacia de los costos implicados en los procesos de fiscalización (tanto a nivel de recursos monetarios como del tiempo requerido para un abordaje integral de los procesos). Esta selección generada mediante un muestreo,

[22] A la hora de redacción de esta Tesis, el medio vigente es el SICOP

Figura 2.16. Principales etapas dentro del proceso de contratación pública.

podría llegar a desaprovechar la existencia de bases de datos de mayor dimensión, las cuales dan un potencial de análisis más exhaustivo. Por lo tanto, La búsqueda de herramientas destinadas a un análisis masivo de información, permite una mayor focalización hacia aquellas observaciones que difieran del comportamiento normal generalmente observado (Eilifsen *et al.*, 2014).

Estudios llevados a cabo a nivel latinoamericano han indagado la posibilidad de realizar indicadores asociados a la detección de anomalías, en Colombia Zuleta *et al.* (2019) abordan la creación de indicadores más enfocados hacia una detección directa de comportamientos irregulares[23], y en el caso Mexicano Martínez *et al.* (2019) encontraron relacionaron los comportamientos con anómalos con la falta de competencia, transparencia e integridad en cada procedimiento de contratación, mediante la elaboración de puntuación para cada dependencia.

Finalmente CAF (2021), explica como las autoridades responsables en términos fiscales así como los encargados de las compras públicas tienen interés y responsabilidad en la utilización de la información para mejorar la calidad del gasto y mediante la competencia tomar mejores decisiones, como consecuencia genera confianza en la ciudadanía quienes a su vez ejercen el control social del gasto público.

[23]En este caso, se emplearon indicadores como por ejemplo, monto asignado a empresas fantasmas o sancionadas, o bien porcentaje de las licitaciones nacionales o internacionales con plazos irregularmente cortos.

2.9 Comportamiento anómalo objetivo

Para el desarrollo de la presente Tesis, se busca la identificación de valores que poseen un comportamiento atípico, definido este como aquel con desviaciones del comportamiento esperado.

López-Iturriaga y Sanz (2018) indican que al existir recursos limitados la utilización de alertas tempranas podría categorizar las áreas más sensibles, y enfocarse por lo tanto en aplicar políticas preventivas y correctivas de manera más efectiva.

Se analiza la integridad[24] con una visión de la realidad costarricense, en este caso no se emplea el concepto de corrupción ya que sería necesario realizar estudios específicos para lograr dicha identificación, lo cual queda fuera del alcance de la presente Tesis.

Con lo anterior como punto de partida los valores anómalos que se buscan identificar se relacionan al resultado del desconocimiento de los principios y deberes en el ejercicio de la función pública y tutela del interés general, tal como se define en Costa Rica Integra (2021).

Usualmente, en temas relacionados con integridad, el mecanismo más difundido es el fomento de la denuncia, pero esto deja la responsabilidad en terceras personas, de acá nace la necesidad de generar un modelo que aproveche el marco normativo e institucional. Rabuzin y Modrusan (2019) sugieren que la aplicación de los grandes datos y el uso de métodos de minería de datos ayuda a lograr mejores resultados, además, Purwanto y Emanuel (2020) encontraron que el uso masivo de información puede proporcionar información valiosa para procesos de auditoría en la adquisición gubernamental, aumentando la eficacia de las adquisiciones apoyado en la toma de decisiones con información más actualizada y amplia.

La identificación de los valores anómalos es relevante con el fin de generar una línea de defensa en el nivel superior o de alta gerencia, y con esto aumentar la facilidad de atención de estos casos. Para realizar esto es preciso propiciar el uso de los datos abiertos, así como el fortalecimiento de las medidas de control interno, y propiciar el análisis pro activo de aquellas áreas proclives al surgimiento de posibles conflictos.

En términos de la materialidad de este problema, se ha determinado que el peso de

[24]OCDE (2009) define la integridad como el uso de fondos, recursos, activos y autoridad de acuerdo con los propósitos oficiales previstos, para ser utilizados en línea con el interés público.

la contratación pública es una actividad económica que genera flujos financieros entre un 10% y un 15% del PIB en todo el mundo. Asimismo, según estudios del Foro Económico Mundial, se estima que el soborno por parte de empresas internacionales en países de la OCDE es añade entre un 10% y un 20% a los costos totales de los contratos OCDE (2009).

OCDE (2009) indicó que los esfuerzos para mejorar la gobernanza y la integridad en la contratación pública forman parte de una gestión eficiente y efectiva de los recursos públicos, esto se puede realizar con un alto grado de transparencia en todo el ciclo de contratación, generando condiciones para que todos los posibles proveedores tenga un tratamiento justo y equitativo. Esto deberá de pasar por el establecimiento de mecanismos para prevenir riesgos en la integridad en la contratación pública, refiriendo a posiciones, actividades o proyectos potencialmente vulnerables.

Los pilares clave propuestas por OCDE (2009) para mejorar la integridad en la contratación pública, se basan en la transparencia, buena gestión, prevención de conductas indebidas, cumplimiento y monitoreo, responsabilidad y control, de esta manera algunos elementos pueden considerarse como estructuralmente necesarios, tal como un marco legislativo; infraestructura institucional y administrativa; un régimen de revisión y responsabilidad efectivo; un régimen de sanciones efectivo; y recursos humanos, financieros y tecnológicos adecuados para apoyar todos los elementos del sistema.

Para Ferwerda *et al.* (2017) un elemento importante en los análisis encontrados en la revisión de literatura, es que usualmente los modelos generados para determinar casos de anómalos en Contratación Pública se basan en el análisis de casos ya conocidos de corrupción, usualmente de dichas experiencias se generan banderas rojas que sirven como base para análisis posteriores, pero esto posee un alto riesgo de generar un sesgo de selección[25], lo que limita la capacidad de muchos estudios previamente propuestos.

[25]Consiste en elegir una muestra atípica que posteriormente no permite inferir de nuevo a la población más grande.

2.10 Tratamiento de los datos

En Costa Rica, la plataforma de SICOP utiliza la codificación del catálogo estándar de Productos y Servicios de las Naciones Unidas como base, esto permite hasta cierto nivel rastrear los bienes y servicios mediante una codificación uniforme.
Los códigos empleados en SICOP para las contrataciones constan de 24 dígitos, los primeros 8 dígitos son determinados por las Naciones Unidas, los siguientes 8 son especificados para identificar los productos a contratar (los usuarios los solicitan al requerir un bien o servicio específico, y se asignan aleatoriamente). Finalmente, los últimos 8 dígitos no generan diferencias sustanciales en el tipo de bien que se está analizando, incluso, se podría afirmar que ese nivel de detalle (códigos a 24 dígitos) imposibilita la comparación, ya que una gran mayoría de los mismos aparecen únicamente una vez adjudicados mediante la plataforma como se detalla en el Cuadro 2.2.

Cuadro 2.2. Porcentajes de bienes adjudicados en la plataforma SICOP, según cantidad de ocasiones por cantidad de dígitos del código.

Conteo	8 Dígitos	16 Dígitos	24 Dígitos
1	16.42%	**54.70**%	**81.40**%
2	9.70%	16.85%	10.70%
3	6.86%	8.02%	3.29%
4	5.59%	4.87%	1.51%
5	4.58%	3.16%	0.83%
6 o más	56.85%	13.40%	2.27%

Con base en lo anterior, el análisis de la presente tesis se centrará en la comparabilidad de los primeros 16 dígitos, esto permite un nivel de detalle suficientemente claro como para realizar comparaciones a nivel de toda la plataforma, tal como se ejemplifica en la Figura 2.17.
Los últimos 8 dígitos no contemplados en la Figura 2.17, son asignados de manera secuencial y son creados por los proveedores de los bienes, estos dígitos se conocen como el código del producto y aportan variaciones como por ejemplo marca o tipo de empaque, pero no elementos de fondo[26].
A pesar de la situación planteada, es posible realizar comparaciones en función de

[26]Para más detalle del catálogo de bienes y servicios se puede consultar la página de SICOP.]

Figura 2.17. Ejemplo de identificación de un bien en los primeros 16 dígitos*

* Empleando la clasificación de la raíz indicada por el catálogo estándar de Productos y Servicios de las Naciones Unidas, así como la identificación del producto asignada aleatoriamente.

las ofertas presentadas por los oferentes[27], con este tipo de análisis se aumenta la cantidad de información disponible, logrando que la cantidad de bienes únicos se ubique en 36.53%, asimismo, el porcentaje de códigos que han sido adjudicado y ofertados en una sola ocasión representan el 7.45% del total de las observaciones dentro de la plataforma, siendo estos objetos de contratación muy específicos y por lo tanto son adquiridos con poca frecuencia.

Con base en lo anterior, es posible emplear el código a 16 dígitos para dar trazabilidad con poca pérdida de información. Además, es factible establecer condiciones con las cuales evaluar la normalidad de los procesos, y por lo tanto la identificación de aquellas observaciones consideradas como anomalías en las contrataciones.

Es importante indicar que la información analizada no contempla la totalidad del proceso indicado en la Figura 2.16, ya que la etapa de recepción de los bienes no cuenta con suficiente información a nivel de SICOP, por lo tanto el análisis comprenderá las etapas comprendidas desde el inicio del proceso a la formalización contractual.

[27]Si bien es cierto, es posible presentar ofertas por parte de proveedores que no cumplan con las condiciones técnicas requeridas por parte de los usuarios, es subsanado mediante una criterio de selección con el cual únicamente se toman en cuenta aquellas ofertas que hayan sido analizadas por las unidades usuarias y se determine que las mismas cumplen con los requisitos definidos en el cartel del proceso, mediante una calificación otorgada de oferta elegible.

2.10.1 Creación de indicadores

Con el objetivo de plantear un modelo de analítica con las técnicas utilizadas en esta investigación, se consideró un procedimiento lógico (ver Figura 2.18). Dicho procedimiento tiene como objetivo ilustrar la creación de indicadores como parte de un proceso con el cual se establezcan las condiciones para generar comparabilidad, y por lo tanto establecer una vez aplicados los modelos de detección, aquellos valores que se comporten como observaciones normales, o anomalías fuertes o débiles tal como lo planteó Aggarwal (2017).

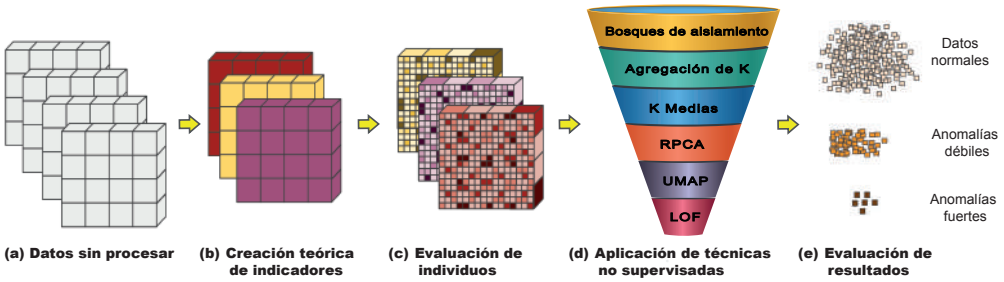

Figura 2.18. Modelo conceptual implementado

Conceptualmente, la presente tesis tomó como base el estudio de principios de Integridad en procedimientos públicos planteado por OCDE (2009), principalmente en la definición de tres grandes temas a evaluar, siendo estos la buena gestión del proceso de compra, la prevención de conductas indebidas y la transparencia[28].

- **Buena gestión del proceso de compra:** Fundamentalmente, la planificación de los procedimientos son clave para reflejar una visión estratégica de las necesidades, esto permite a su vez fomentar la transparencia y la rendición de cuentas, así como mejorar la relación costo-beneficio, esta planificación requiere indispensablemente que los funcionarios estén adecuadamente capacitados en planificación, programación y estimación de costos. OCDE (2009). Un elemento puntual de la planificación indicada, puede resumirse en la realización de una estimación realista del presupuesto, lo cual permita asegurar una aproximación al precio que se encontrará en el mercado, este tipo de indicador puede encontrarse en estudios planteados por el observatorio de Corrupción de Indonesia (ICW) tal como lo citan Purwanto y Emanuel (2020).

[28]La selección de los indicadores se detalla en el Anexo A4

Otro factor considerado en la presente tesis como parte de una buena gestión del proceso de compra, es la cantidad de objeciones que sean presentados al proceso de contratación[29]. En esta línea, OCDE (2009) recomienda proporcionar formas para impugnar la decisión en el proceso de contratación, este tipo de impugnaciones deben de contar con un tiempo suficiente para que se realicen las solicitudes necesarias, lo cual permita transparentar el proceso de contratación.

- **Prevención de conductas indebidas:** Ferwerda *et al.* (2017); Purwanto y Emanuel (2020) indican que procesos con una alta rapidez en el proceso de contratación podrían reflejar vulnerabilidades, asimismo, OCDE (2009) recomienda emplear la contratación electrónica para asegurar que la información sobre la adjudicación se comunique de manera oportuna a todos los posibles oferentes en un plazo razonable.

  Además de lo anterior, Della Porta y Vannucci (2002) explican que las modificaciones de los contratos y las alteraciones de precios posteriores a la asignación del contrato son posibles indicadores de riesgos, esto no descarta que dichos cambios puede ser socialmente beneficiosa, pero deben de existir mecanismos para evitar el abuso de dicha figura tal como lo indican Laffont y Tirole (1990).

  Finalmente, OCDE (2009) concluye que los riesgos de abuso por parte del contratista en el cumplimiento del contrato, podrían traer consecuencias en relación con su calidad, precio y plazo.

- **Transparencia:** OCDE (2009) recomienda que la información proporcionada por los usuarios sea verificada, comparada y monitoreada de cerca en la gestión del contrato para mantener altos estándares de integridad. Es importante que esta transparencia en la información permite una reducción en la asimetría de la información.

  Un profundo conocimiento de las industrias de interés así como del mercado, brinda claridad del posible número de proveedores involucrados en relación a

[29]Es importante indicar que se podrían presentar múltiples objeciones por parte de proveedores no adjudicados no por temas relacionados a la calidad del cartel, sino más bien a esfuerzos para retrasar o cambiar el resultado dado por la Administración, en este caso se consideró como supuesto que este tipo de casos es el mínimo, y por lo tanto las objeciones presentadas se basan en aclaraciones debido a elementos del cartel.

una contratación en específico, y podría generar alertas preventivas en el caso de aquellos mercados donde no exista una competencia real, o la competencia sea limitada Purwanto y Emanuel (2020). Incluso, autores como Della Porta y Vannucci (2002); Fernandez y Plumbley (2021); Rabuzin y Modrusan (2019), explican que en ocasiones la competencia parece real por el número de empresas inscritas en una proceso de contratación, pero podría ser que la mayoría de estas empresas no sean competidores reales, lo cual hace que el proceso de contratación no sea transparente.

Por último, en la presente tesis se consideró el tema de la transparencia según la composición propia del mercado, donde podrían generarse ineficiencias derivadas de una alta concentración de mercado o bien por medio de prácticas no concordantes con procesos idóneamente transparentes.

CAPÍTULO 3
MARCO METODOLÓGICO

Esta investigación tiene como objetivo analizar la capacidad de detección de anomalías en un enfoque no supervisado mediante simulaciones bajo escenarios de diversas características. Para llevar esto a cabo, primero se seleccionaron las técnicas de interés, además se realizó una comparabilidad bajo diferentes escenarios con el fin de proponer un modelo óptimo. Luego se realizó su aplicación para el caso de la Contratación Pública de bienes mediante la plataforma SICOP.

3.1 Simulaciones

En la presente tesis se llevó a cabo un análisis comparativo del comportamiento de diferentes metodologías para la detección de valores anómalos en un enfoque no supervisado, aportando diversos escenarios y condiciones que permiten robustecer la discusión en torno al efecto de las técnicas planteadas para la detección de anomalías.

Con base en lo anterior, se generan evidencias sobre el impacto de la detección de anomalías dependiendo de los siguientes factores: Simetría de la distribución; Tamaño de la población; Porcentaje de anomalías; Correlación entre las variables.

Para realizar este análisis se crearon datos sintéticos[30], considerando en el caso de la distribución Normal la matriz de correlaciones y además los vectores de varianzas y de medias, asimismo, en el caso de la distribución Gamma se utilizaron copulas y parámetros para simular el nivel de correlación entre la población analizada[31].

A partir de simulaciones y estudios llevados a cabo por Goix (2016); Hennig (2018); Mayer *et al.* (2020); Morris *et al.* (2019), así como de los objetivos de esta tesis, se consideró la siguiente metodología para analizar la capacidad de detección de anomalías en un enfoque no supervisado mediante simulaciones, específicamente para el caso de la contratación de bienes mediante el SICOP.

[30]Según Nikolenko (2019) los datos sintéticos consisten en la creación de datos artificiales utilizando técnicas avanzadas de manipulación de datos para generar suficiente información que sea empleada para el entrenamiento de modelos.

[31]Se amplia con mayor detalle el proceso de creación de datos multivariados en el Anexo A1.

Para esto se generaron en el caso de la distribución normal una simulación con base en la matriz de correlación, medias y desviación estándar.
Con el fin de determinar el poder predictivo de los modelos seleccionados bajo escenarios distintos, se consideran 4 factores en la simulación de las bases de datos:

1. **Simetría de la distribución:** bajo la premisa de conocer el impacto de la distribución de los datos se consideró el análisis en función de las distribuciones Normal y Gamma.

2. **Tamaño de la población:** el objetivo es verificar el potencial de predicción cuando la cantidad de observaciones aumenta, considerando la creación de datos en función de la matriz de covariancias y el valor de la mediana de las variables, se generaron siete poblaciones distintas, con valores poblacionales de 10 000, 25 000, 50 000, 100 000, 150 000, 200 000 y 250 000 observaciones.

3. **Porcentaje de las anomalías:** se consideran los siguiente valores de anomalías[32] tanto para las distribuciones normales como gamma: 0,01%, 0.5% 1%, 2%, 3%, 4% y 5%; en el caso de la distribución normal autores como Liu *et al.* (2008) ubican el porcentaje en valores menores al 0,3%, mientras que otros autores como Campos *et al.* (2016); Nooghabi *et al.* (2010); Singh y Lalitha (2018) ubican dicho porcentaje en un máximo de 5%. En el caso de la poblaciones cuya distribución sea Gamma Nachmani *et al.* (2021) considera porcentajes de anomalías entre el 0.01% y el 2%, incluso se llegan a emplear valores superiores al 5%.

4. **Correlación entre las variables:** con el objetivo de conocer el nivel de sensibilidad entre las diferentes variables y el nivel de afectación positiva o negativa que algunos modelos pueden tener en función de la correlación subyacente entre las variables, se plantean cinco escenarios basados según la correlación entre las mismas, dichas correlaciones van desde el 0% hasta el 80% con cambios de 20% entre cada escenario.

[32] Aggarwal y Yu (2001); Hautamaki *et al.* (2004) definen las anomalías como desviaciones atípicas, en el caso de la distribución normal se emplea el supuesto de que las mismas son aquellas observaciones con tres veces la desviación estándar; en el caso de la distribución gamma se consideran las observaciones más desviados en la cola derecha, asimismo en el caso de la distribución Gamma se emplea la indicación de anomalía en el caso de aquellos individuos que tengan al menos dos valores atípicos.

Se tienen siete escenarios relacionados con el tamaño de la población, siete escenarios relativos al porcentaje de anomalías presentes en los datos, cinco escenarios considerando la correlación entre las variables y dos escenarios enfocados en la simetría de la distribución de los datos analizados, lo anterior fue evaluado con 9 variables que conforman el modelo de interés para la presente tesis. En general se contó con un total de 490 escenarios que son sometidos a la evaluación por parte de las 97 variantes de los modelos previamente descritos, para un total de 47 530 modelos diferentes.

3.2 Modelos

Los modelos que serán analizados en la presente tesis son los siguientes:

- **Modelo de ACP:** En este caso para el ACP se tienen dos elementos de interés: el centrado y escalamiento de la media y la desviación típica de las variables, para la presente investigación se calibrará el modelo con la combinación de de esos 2 factores considerando tanto los valores estandarizados como los no estandarizados, tal como se propone en Kassambara y Mundt (2020).

- **Modelo de ACPR:** se consideran 2 elementos de interés tanto el escalado mediante el cual se establece si las variables deben escalarse, como el cambio de signo, por medio del cual se indica si se debe tratar de resolver la indeterminación de signo de las cargas, ya que el enfoque ad hoc establece que el elemento máximo en un vector singular sea positivo.

- **Modelo de K Medias:** Ye *et al.* (2006) consideran el nivel de lejanía entre los clúster, lo cual permite identificar aquellos espacios poco poblados y por lo tanto con mayor probabilidad de ser valores atípicos. En el caso de la presente tesis se emplearon los siguientes valores de K 3, 4, 5, 6, 10, 15, 20, 25.

- **Modelo agregación de K:** Angiulli y Pizzuti (2002) La identificación de valores anómalos se realizó en regiones específicas definidas por un rango de valores de K. Para este propósito, se consideraron tanto un valor mínimo como

un valor máximo de K. En el caso de los valores mínimos de K, se emplearon 10, 20, 25, 30 y 35, mientras que en el caso de los valores máximos de K se utilizaron 40, 60, 80, 100 y 120, finalmente se realizó la combinatoria completa de dichos valores para analizar finalmente un total de 25 combinaciones.

- **Modelo LOF:** Madsen (2018) emplea diversos valores de K vecinos para comparar la densidad, en el caso de la presente investigación se emplearon 8 variantes, con los siguientes valores de K: 20, 40, 60, 80, 100, 120, 140 y 160.

- **Modelo Bosques de Aislamiento:** se emplean 4 diferentes variables de modelos propuestos por Liu *et al.* (2008), a saber el modelo simple, el modelo extendido, el modelo con criterio de selección divida (SCi) y la variante de Bosques de aislamiento con criterio de ganancia.
  Para los 4 modelos indicados se emplean un número diverso de árboles para determinar en qué nivel de repetición el modelo tiende a mejorar su detección de valores anómalos, específicamente se emplean las siguientes cantidades: 10, 20, 30, 40, 50, 60, 70, 80, 90 y 100..

- **Modelo UMAP:** posterior a la aproximación y proyección de Manifolds Uniformes, se empleó la técnica de K medias, en específico los valores de K empleados en el caso de K Medias aplicados son 3, 4, 5, 6, 10, 15, 20, 25.

3.2.1 Evaluación de los modelos bajo escenarios simulados

La detección de anomalías carece de una metodología estándar para comprender y evaluar algoritmos, por lo tanto para evaluar la calidad de detección con técnicas no supervisadas, se pueden emplear experimentos de control, donde una alteración previa de observaciones permite evaluar los resultados obtenidos con el modelo. De esta manera es posible identificar como se evalúan aquellas observaciones previamente alteradas y consideradas a priori como anómalas (Ghafoori, 2018).
Emmott *et al.* (2015) indican que la forma estándar de comparar el rendimiento de las técnicas de detección de anomalías se basa en área bajo la curva de característica operativa del receptor (en adelante AUC-ROC) y la Precisión Promedio, también conocida como el área bajo la curva de precisión-sensibilidad, en adelante (AUC-PR).

Amer y Abdennadher (2011); Ghafoori (2018); Swersky (2018), explican como el AUC-ROC es trazada al establecer un umbral en las puntuaciones, comenzando con el valor más pequeño hasta el mayor (o viceversa), en cada umbral es calculada la tasa de positivos reales y los falsos positivos reales. En ese escenario, una identificación totalmente aleatoria se evidenciaría como una línea en la diagonal del cuadro, mientras que identificaciones más precisas se alejan más de la diagonal hacia el límite izquierdo y superior del cuadrado, teniendo el AUC-ROC óptimo un valor de 1.
Diversos autores, entre ellos Goutte y Gaussier (2005) explican como existen otras técnicas relacionada a la validación de resultados en el contexto de clasificación entre las que se pueden encontrar las indicadas en el Cuadro 3.3.

Cuadro 3.3. Métricas de calidad de clasificación

Prueba	**Fórmula***
Exactitud	$(VP + VN) \,/\, (VP + FP + FN + VN)$
Tasa de Error	$(FP + FN) \,/\, (VP + FP + FN + VN)$
Sensibilidad (RC)	$(VP) \,/\, (VP + FN)$
Especificidad	$(VN) \,/\, (VN + FP)$
Precisión (PR)	$(VP) \,/\, (VP + FP)$
Tasa de Detección	$(VP) \,/\, (VP + FP + FN + VN)$
Exactitud de Detección	$(FP + VP) \,/\, (VP + FP + FN + VN)$
Balance de Detección	$(Sensibilidad + Especificidad) \,/2$
Valor predicción Negativo (VPN)	$(VN) \,/\, (FN + VN)$
Precisión Global	$(VP + VN) \,/\, (VP + FP + FN + VN)$
Error Global	1 - Precisión Global

* Los valores empleados se derivan de la matriz de confusión, de la siguiente manera: VP= Verdaderos Positivos; VN= Verdaderos Negativos, FP= Falsos Positivos; FN= Falsos Negativos.

Autores como Wardhani *et al.* (2019) indican que la medida de evaluación $F1$ es muy empleada como clasificador para evaluar resultados con máquinas de aprendizaje con bases desbalanceadas, este indicador es estimado como se observa en la Ecuación 24, este factor se considera como una media armónica entre la precisión y la sensibilidad.

$$F_1 = \frac{2 \times Precisin \times Sensibilidad}{Precisin + Sensibilidad} \tag{24}$$

A pesar de lo anterior, autores como Craven y Bockhorst (2004); Davis y Goadrich (2006); Fayzrakhmanov *et al.* (2018) consideran a la AUC-PR como la mejor manera de realizar evaluaciones con poblaciones con altos niveles de asimetría, por lo tanto este elemento de medición será considerado como parte de los indicadores principales

a analizar en la sección de Resultados de la presente tesis.
Finalmente, a nivel de la correlación entre las variables empleadas, interesa la utilización del coeficiente de correlación de Phi, el cual ha sido empleado por múltiples autores como Kuhn (1973) para analizar resultados dicotómicos, donde sus valores pueden compararse directamente para determinar su significancia estadística de una manera más precisa que otros tipos de correlaciones usualmente empleadas como las de Pearson.

3.3 Fuentes de información

La datos de las contrataciones públicas mediante información de SICOP fueron obtenidas mediante acuerdo con Radiográfica Costarricense (RACSA) y la Contraloría General de la República.
La información es pública y está disponible en la página del SICOP (https://www.sicop.go.cr/index.jsp/).

3.4 Programas y técnicas a emplear

Los análisis fueron realizados utilizando el programa R Core Team (2021) en su versión 3.6.3. Dicho programa es de uso global para el desarrollo de los análisis estadísticos requeridos. Los paquetes requeridos en R fueron los siguientes:

- **Lectura y manejo de los datos:** `data.table` (Dowle y Srinivasan, 2021), `dplyr` (Wickham *et al.*, 2023), `readr` (Wickham *et al.*, 2022), `lubridate` (Grolemund y Wickham, 2011).

- **Paralelización y mejoras del proceso:** `purrr` (Henry y Wickham, 2022), `snow` (Tierney *et al.*, 2021), `h2o` (LeDell *et al.*, 2022), `progressr` (Bengtsson, 2023).

- **Creación de datos:** `MASS` (Venables y Ripley, 2002a), `JWileymisc` (Wiley, 2022), `copula` (Hofert *et al.*, 2023).

- **Evaluación de resultados:** `pROC` (Robin *et al.*, 2011), `PRROC` (Grau *et al.*, 2015), `traineR` (Rodriguez R. *et al.*, 2022), `caret` (Kuhn, 2022).

- **Modelos empleados:** `isotree` (Cortes, 2022), `factoextra` (Kassambara y Mundt, 2020), `rrcov` (Todorov y Filzmoser, 2009a), `DDoutlier` (Madsen, 2018), `cluster` (Maechler *et al.*, 2022), `robustbase` (Todorov y Filzmoser, 2009b), `umap` (Konopka, 2023).

Es importante recalcar que se realizó una paralelización con 12 núcleos, esto con el objetivo de reducir los tiempos de procesamiento, en este caso debido al alcance de la presente tesis no se analizará el efecto de reducción de tiempo entre el análisis paralelo y de núcleo único.
Finalmente, el equipo empleado para llevar a cabo las simulaciones fue un Macbook Pro con procesador Intel Core i7 y 16GB de RAM.

CAPÍTULO 4

RESULTADOS

4.1 Evaluación de resultados

Con las simulaciones empleadas, se procede a una evaluación para determinar los factores relacionados en términos de duración de los procesos, así como la aplicación de las metodologías de evaluación previamente descritas.

En el caso de la presente tesis, se determinan algunas consideraciones a la hora de hacer las evaluaciones correspondientes, las medidas principales para la detección de atípicos se centran en los resultados de las AUC-ROC, AUC-PR, Precisión Global y la Precisión, esto con base en la literatura especializada en términos de evaluación descritos en la Sección 3.2.1, asimismo, en el Anexo A3 se pueden observar los mejores modelos considerando las calibraciones correspondientes.

Aunado a lo anterior, en el caso de la detección de valores anómalos para el caso de contratación pública de bienes se realizó la evaluación de los indicadores propuestos en el Cuadro 3.3, pero se otorga una mayor relevancia a la detección de valores positivos, esto porque el principal interés teórico de este tipo de investigaciones, se delimita a la capacidad de detección de aquellas contrataciones cuyo comportamiento sea considerado anormal, por lo tanto es de interés de esta investigación centrarse en la calidad de los resultados proporcionados tanto por la globalidad detectada con la Precisión Global (al igual que su inverso el Error Global), además de la Precisión, la cual da una mayor relevancia a la detección de los valores positivos.

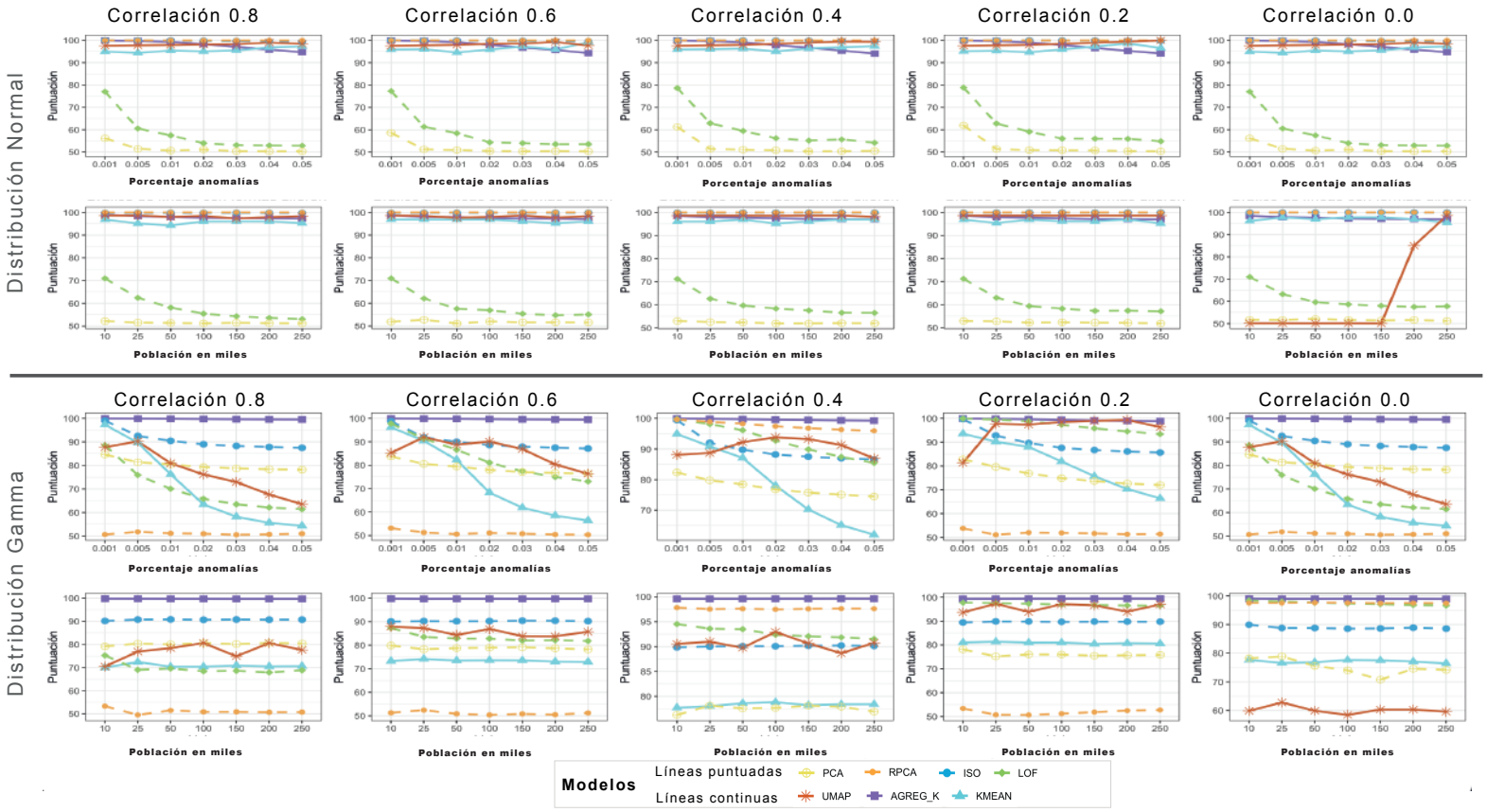

Figura 4.19. Resultados promedios de valores del área bajo la curva ROC*

* En el caso de los modelos cuya distribución es normal parece que los comportamientos son más estables en los diversos niveles de correlación de los datos, mientras que en el caso de los modelos Gamma existen mayores fluctuaciones, más generalizadas ante la ausencia de correlación de las variables.

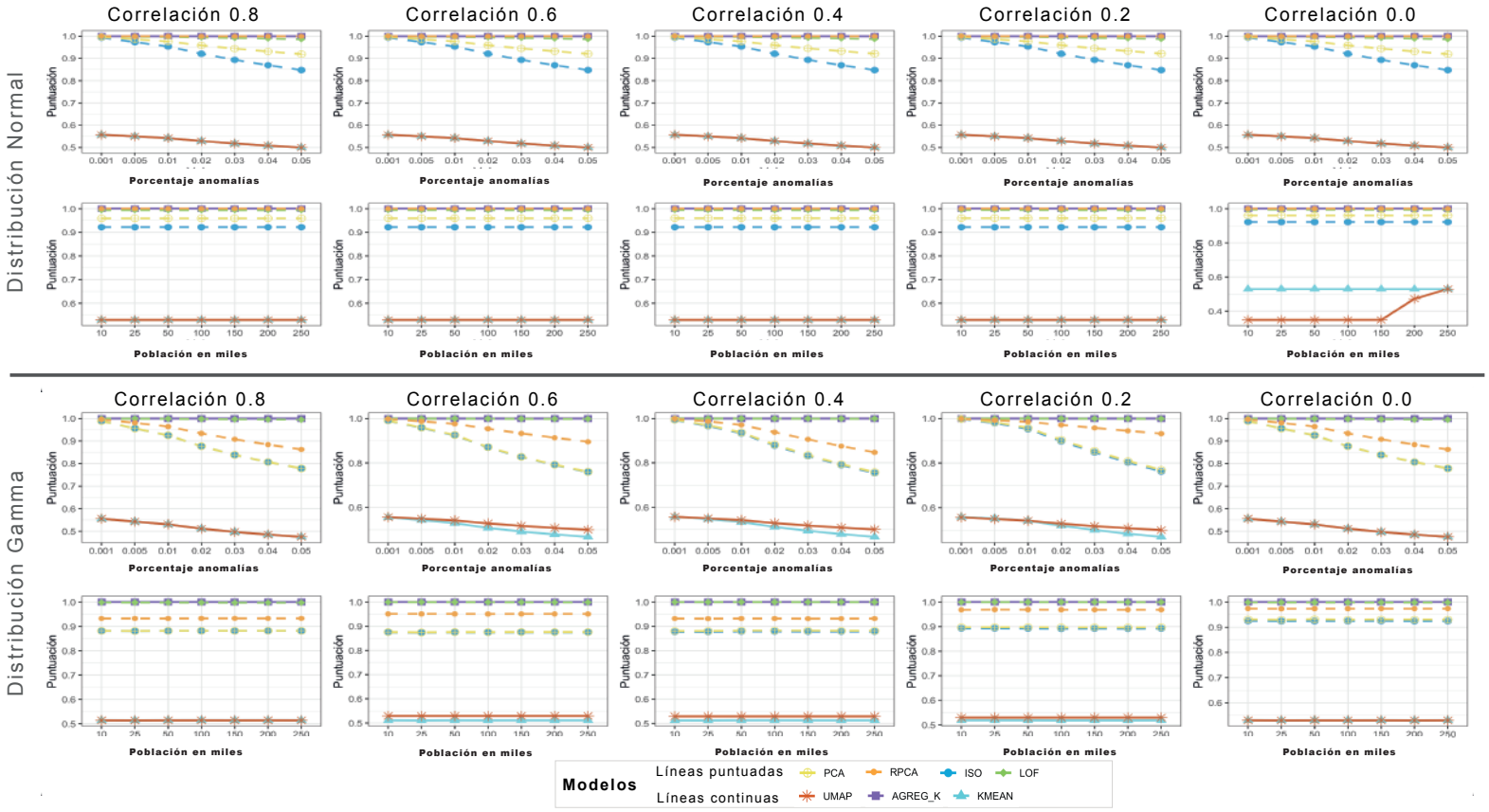

Figura 4.20. Resultados promedios de valores del área bajo la curva PR*

*En el caso de la curva de precisión y sensibilidad, el comportamiento de los modelos parece ser más estable a lo largo de las diferentes técnicas empleadas, destacando particularmente la estabilidad en este indicador de los modelos basados en proximidad y densidad empleados en esta tesis.

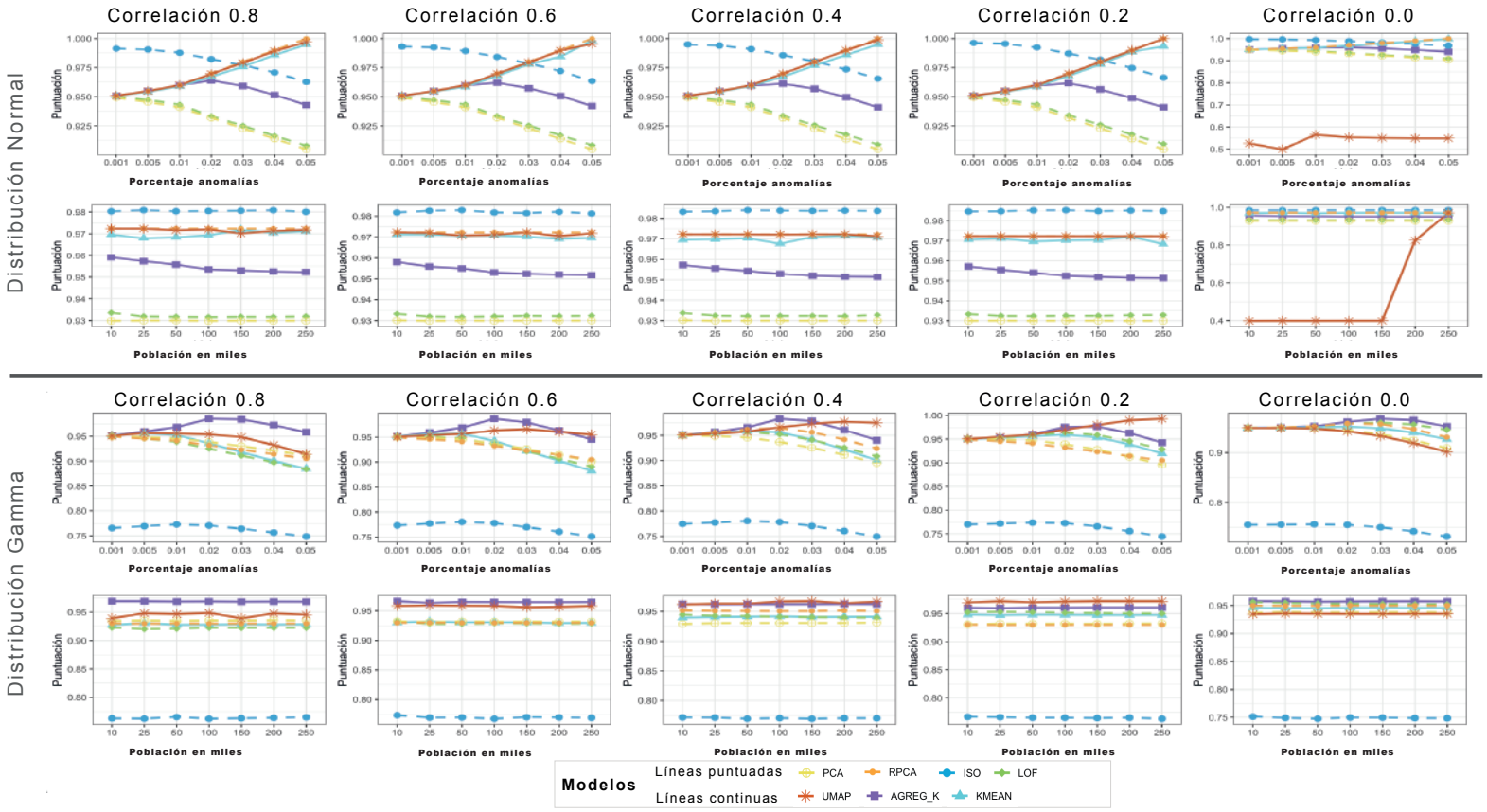

Figura 4.21. Resultados promedios de la Precisión Global*

*Observando los resultados de la precisión global, se observa como afectan algunos factores afectan la calidad de los modelos, siendo notable la variación generada cuando se tiene una ausencia de correlación entre las variables y su impacto en ciertas técnicas.

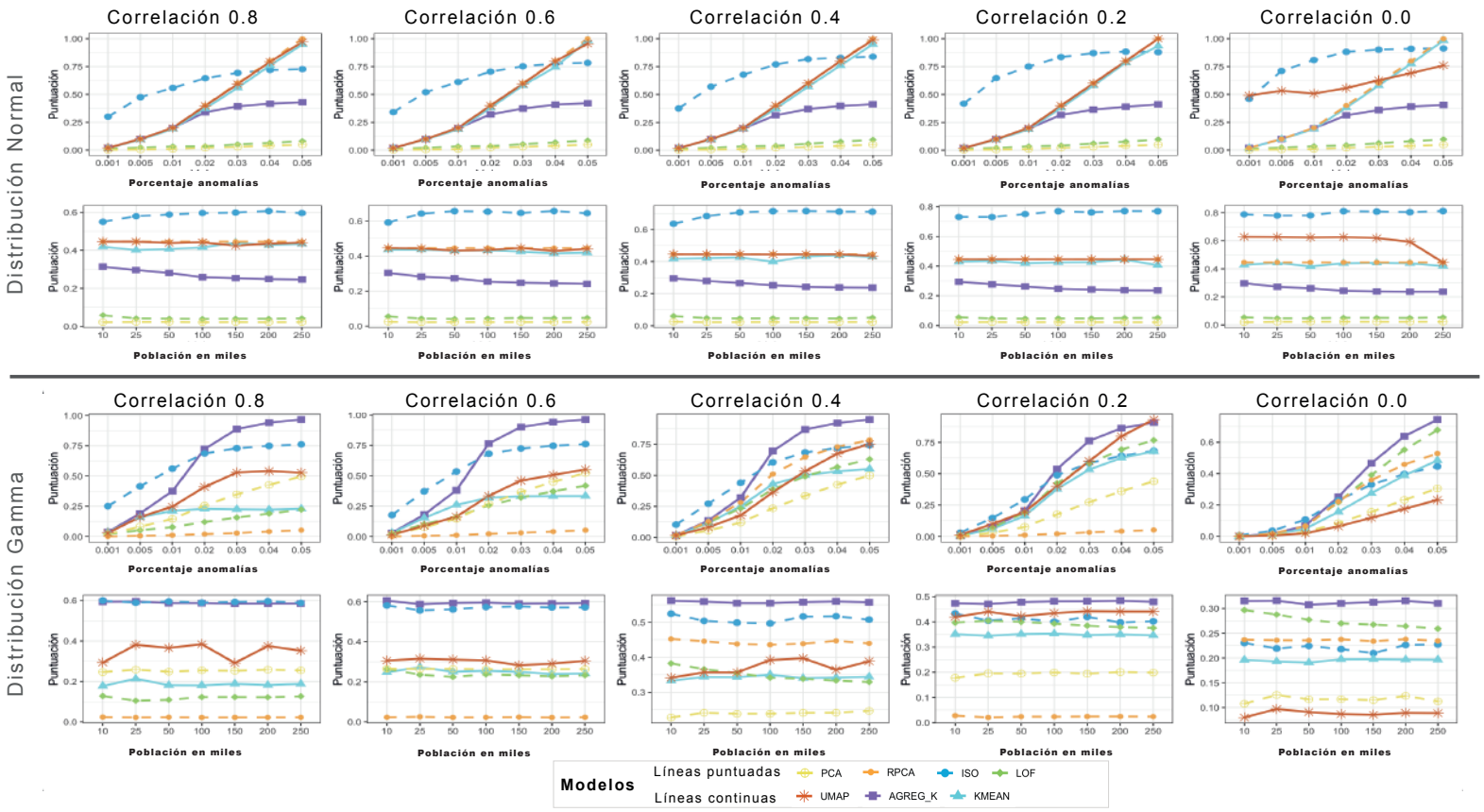

Figura 4.22. Resultados promedios de la Precisión*

*En el caso de la precisión, se denota como mayores tamaños de población y de anoamlías, parece mejorar los resultados de la predicción, lo cual parece concordar con la literatura analizada al respecto.

4.2 Análisis de Componentes Principales

En el caso del modelo ACP (centrado y escalado), la calidad de los modelos, medida por la curva AUC-ROC (ver Figura 4.19), es consistentemente baja en todos los casos analizados, independientemente del tamaño de la población, el porcentaje de anomalías en los datos o el nivel de correlación entre las observaciones. Específicamente, al considerar la distribución normal, los resultados apenas superan el 65% en AUC-ROC. En contraste, para la distribución gamma, los valores son relativamente mejores, oscilando entre el 85% y el 72%. Sin embargo, estos valores decrecen a medida que aumenta el número de valores anómalos en el conjunto de datos. Además, el nivel de correlación entre las variables no parece tener un impacto significativo en estos resultados.

La Figura 4.20 ilustra cómo la mayoría de los modelos se ven afectados por incrementos en el porcentaje de anomalías presentes en la población analizada, particularmente en el caso del modelo ACP. Tanto para la distribución normal como para la distribución gamma, se observa que el aumento en el porcentaje de anomalías no afecta significativamente al modelo ACP con distribución normal, con resultados que superan el 99%. Por otro lado, en el caso de la distribución gamma, los valores oscilan entre el 99% y el 75%, siendo los peores niveles observados cuando el porcentaje de anomalías es más alto.

En lo que respecta a la Precisión Global (ver Figura 4.21), se ha observado una estabilidad y resultados positivos en el caso del modelo ACP, tanto para datos con distribución normal como para aquellos con distribución gamma. No obstante, se han identificado disminuciones en la precisión global a medida que aumenta el porcentaje de anomalías en los datos, así como con el incremento en el tamaño de la población. A pesar de estas afectaciones, el valor de la precisión global se mantiene por encima del 92%, con una variabilidad muy baja.

En cuanto a la Precisión, los resultados observados en los modelos analizados se presentan en la Figura 4.22. Es evidente que, en el caso del modelo ACP, los resultados muestran un incremento en la precisión a medida que aumenta la correlación entre las observaciones, especialmente para la distribución gamma. Sin embargo, este aumento es muy moderado, no superando el 26%. Este hallazgo refuerza la sensibilidad del método de clasificación a las anomalías. Además, para las pobla-

ciones con distribución normal, la precisión observada es menor al 3% en los casos analizados.

Al considerar la cantidad de observaciones anómalas, se observa que, con distribuciones gamma y un 5% de valores anómalos, la precisión aumenta hasta casi el 50% cuando existe una correlación del 80% entre los datos. Esto confirma cómo los valores de precisión aumentan con la correlación entre las variables en este tipo de modelaciones. No obstante, la calidad general de la técnica está entre las peores de las analizadas.

4.3 Análisis de Componentes Principales Robustos

Con el modelo ACPR, se han obtenido resultados notoriamente altos del AUC-ROC al emplear matrices de covarianza robustas en conjuntos de datos con distribución normal, llegando incluso al 100%. Sin embargo, en el caso de datos que siguen una distribución Gamma, la calidad de los resultados está condicionada por la correlación entre variables, el tamaño poblacional y la presencia de observaciones anómalas. Específicamente, para la distribución Gamma, los resultados oscilan entre el 50% y el 71%, y se observa una alta variabilidad en las observaciones.

Es importante indicar que en el caso del modelo ACPR parece ser menos sensible ante los valores extremos y su influencia en la matriz de covarianza (en comparación con el modelo de ACP). A pesar de esto, en el caso de las distribuciones Gamma el modelo ACP parece ser menos afectada por esta influencia ya que las observaciones no anómalas se pueden representar de manera más precisa en la proyección del plano, a su vez, en el caso de la distribución Normal la matriz de covariancia robusta permite resultados menos variables, pero esta variabilidad disminuye a medida que aumenta la cantidad de observaciones anómalas a identificar.

Por otra parte, para el modelo ACPR (ya sea considerando distribuciones simétricas o asimétricas), los resultados son considerablemente altos con el caso del AUC-PR. En el caso de la distribución normal, todos los resultados superan el 99%, mientras que en la distribución Gamma, se sitúan entre el 99% y el 84%. Esto posiciona al modelo ACPR como la elección óptima al evaluar métodos de detección basados en modelos lineales empleados en esta tesis. Además, se ha observado un nivel de

afectación.

El modelo ACPR, cuando se aplica a poblaciones con distribución normal, exhibe un comportamiento opuesto al observado con el ACP. Este modelo se beneficia de niveles más altos de anomalías en la población, alcanzando niveles cercanos al 97% de la Precisión Global. No obstante, al considerar poblaciones con distribución Gamma, se observa una leve disminución en los resultados a medida que se incrementa el porcentaje de anomalías analizadas, aunque la precisión global sigue superando el 90%.

En lo que respecta al modelo ACPR, la calidad de los resultados es notablemente variable. En el caso de la distribución normal, los valores de la Precisión se sitúan en alrededor del 44% a lo largo de diferentes tamaños poblacionales. Sin embargo, se evidencian variaciones significativas al analizar la cantidad de poblaciones anómalas, beneficiando al modelo a medida que aumenta esta cantidad. Este beneficio también se refleja en poblaciones con distribución gamma, donde los resultados aumentan con la cantidad de anomalías presentes en el estudio. En términos de la cantidad de valores anómalos a identificar, las poblaciones normales tienden a beneficiarse con un aumento en la cantidad de anomalías, alcanzando incluso un 100% de precisión al alcanzar el máximo de observaciones anómalas estudiadas en esta tesis.

4.4 K Medias

En el caso del modelo K Medias, se observa que, con una distribución normal, a medida que aumenta el número de observaciones anómalas, también aumenta el nivel de la calidad según el AUC-ROC. Además, para el conjunto de datos analizados, se estabiliza el nivel de detección con tan solo 4 vecinos, lo cual puede ser beneficioso en términos de tiempo de ejecución. En el caso de la distribución Gamma, esta estabilización se logra con un valor de K igual a 10. Sin embargo, a diferencia de la distribución normal, el poder de predicción tiende a disminuir con incrementos en el porcentaje de anomalías presentes en los datos.

En relación al modelo K Medias, se nota que la calidad de los modelos es notablemente baja según el área bajo la curva de Precisión y sensibilid. Tanto para la distribución Normal como para la Gamma, el valor de este indicador se sitúa entre

el 55% y el 50%, lo cual, en comparación con modelos previos, no proporciona una base sólida resultante.

Por otro lado, el modelo K Medias muestra una particularidad que podría resultar ventajosa en el análisis con técnicas de detección de anomalías no supervisadas, ya que se observa una escasa variabilidad entre todas las variantes de K empleadas. En la población con distribución Normal, aumentos en la cantidad de observaciones anómalas mejoran el nivel de la precisión global. Sin embargo, en la distribución Gamma, se generan disminuciones en dicho indicador. Respecto al tamaño poblacional, el porcentaje de detección se mantiene casi constante en los diferentes escenarios empleados, siempre ubicado por encima del 92%.

En cuanto al modelo K Medias, los resultados son notoriamente altos en la Precisión al comparar con otras técnicas. Por ejemplo, ante cambios en la población, para la distribución Normal, los resultados superan el 40%. En el caso de la distribución Gamma, es más sensible a los cambios en el nivel de correlación, alcanzando su punto máximo cuando el nivel de correlación se ubica en el 40%. Asimismo, los resultados más bajos se observan cuando la correlación es del 0%, 60% o 80%. En relación a la cantidad de observaciones anómalas, a mayor cantidad de anomalías, mayor es la precisión, alcanzando su máximo cuando hay un 5% de anomalías. No obstante, los resultados no son tan altos cuando las poblaciones tienen distribuciones gamma.

4.5 Agregación de K Medias

Respecto al modelo de agregación de K, tanto para la distribución Normal como para la distribución Gamma, los resultados se encuentran por encima del 96% del AUC-ROC en los diferentes escenarios analizados. Los valores más bajos se observan cuando el nivel de población alcanza las 250 mil observaciones. En cuanto a la cantidad de observaciones anómalas, se logra identificar más del 93%, pero con una adecuada parametrización, estos valores pueden alcanzar el 99%.

Ahora bien, los valores obtenidos son entre los más altos de todos los modelos analizados cuando se trata de los resultados del área bajo la curva de Precisión y sensibilidad, asimismo es de los modelos con menor afectación ante los escenarios analizados, tanto a nivel de la simetría de la distribución, el tamaño de la población, el por-

centaje de anomalías observadas y el nivel de correlación entre las variables.
Este modelo logra valores de precisión cercanos al 30% en el caso de la distribución Normal, mientras que en el caso de poblaciones Gamma dicha precisión se ubica cerca del 50%, siendo beneficiada con valores de correlación mayor, alcanzando su punto más alto con una correlación del 80%. En el caso de la cantidad de observaciones anómalas, tanto para la distribución normal como gamma, la precisión mejora mientras mayor sea el número de anomalías, es importante indicar que el ámbito en los resultados observados cambia drásticamente según el nivel de la correlación de los datos, por ejemplo, si existe una ausencia de correlación, el ámbito de los resultados se ubica entre 0.1% y un 74% para el caso de la distribución Gamma y entre 2% y un 40% si la distribución es normal, mientras que si la correlación es de 80%, se ubica entre 3% y un 96% para el caso de la distribución Gamma y entre 9% y 42% para el caso de la distribución Normal.

4.6 LOF

Conforme al modelo LOF, se observa que al aumentar el número de valores de K, el AUC-ROC tiende a incrementar tanto para la distribución normal como para la distribución gamma. No obstante, la técnica se ve afectada por el aumento de la población analizada y la cantidad de observaciones anómalas. En general, las poblaciones con distribución gamma presentan mejores resultados. Además, en la distribución gamma, se muestra una mayor sensibilidad al nivel de correlación entre las observaciones. Por ejemplo, con los mismos 160 valores de K como parámetro, un 5% de anomalías y una correlación del 0%, la detección se sitúa en un 95.5%. Sin embargo, con un cambio en la correlación al 80%, el nivel de detección apenas alcanza el 61%.
En lo que respecta a los modelos de Agregación de K y LOF, los valores obtenidos en el AUC-PR son algunos de los más altos entre todos los modelos analizados. Asimismo, estos modelos muestran una menor afectación ante los escenarios examinados, ya sea en términos de simetría de la distribución, tamaño de la población, porcentaje de anomalías observadas o nivel de correlación entre las variables.
El modelo LOF posee una alta afectación en el caso de incrementos en el porcentaje

de anomalías observadas. A pesar de esto, dichos resultados se sitúan por encima del 90% y por debajo del 95% de la Precisión Global. Esta afectación también se da en el caso de la distribución Gamma, siendo este de los modelos con resultados menos positivos. Es posible observar en este modelo un comportamiento similar al observado en el caso del modelo de Análisis de Componentes Principales, siendo modelos con un comportamiento homólogo.

Conforme al modelo LOF, la calidad de los modelos es deficiente conforme a la Precisión. En el caso de la distribución normal, los resultados no superan el 6%. Además, se ven afectados por incrementos en la población y están relacionados con la cantidad de observaciones anómalas. Por otro lado, en la distribución Gamma, la situación es ligeramente diferente ya que a pesar de sufrir los mismos efectos por la cantidad de observaciones anómalas y los incrementos poblacionales, existen rangos donde el resultado de la precisión se ubica por encima del 75%. Sin embargo, en términos generales, la variabilidad de los resultados es muy alta, oscilando entre el 0.1% (con ausencia de correlación y un 0.1% de anomalías) y el 76% (con una correlación del 20% y un 5% de anomalías).

4.7 ISO

Con el modelo de Bosques de Aislamiento, las calificaciones considerando todas las posibles variantes son notablemente altas en el caso del AUC-ROC. Por ejemplo, la variante ALT alcanza valores promedio del 100% para todos los porcentajes de anomalías en la población en el caso de la distribución normal. Sin embargo, en la distribución normal, la calificación más baja se encuentra en la identificación del 5% de valores anómalos para la variante simple y una correlación del 80%, con un valor de 98.92%, que sigue siendo bastante elevado en comparación con otros modelos. En las distribuciones gamma, los resultados de este modelo tienden a disminuir a medida que aumenta el número de anomalías. Aun así, tanto la variante ALT como la variante simple mantienen una calificación por encima del 90%. En comparación con otros modelos, parece ser uno de los mejores modelos analizados en todos los escenarios.

El modelo ISO muestra una mayor afectación del AUC-PR por el porcentaje de ano-

malías presentes en el conjunto de los datos, y esta disminución se da sin importar el modelo o el nivel de correlación. La caída es más pronunciada al identificar un 5% de anomalías, con una calificación del 84.7%. En el caso de la distribución Gamma, esta caída asciende hasta valores cercanos al 75%. Además, parece beneficiarse de observaciones menos correlacionadas.

Conforme a la Precisión Global, el modelo de Bosques de Aislamiento, se denota como aquel con los mejores resultados (al considerar una distribución Normal) y el modelo con los peores resultados (al poseer datos con una distribución Gamma). En ambos escenarios, el modelo experimenta una disminución en su nivel de precisión global al aumentar tanto el porcentaje de anomalías como el tamaño de la población (aunque en este caso es el modelo con menor variabilidad, al menos para la distribución Normal).

En el caso del modelo ISO, la Precisión muestra un comportamiento con caídas y una desaceleración de las mismas a medida que aumenta el porcentaje de anomalías cuando los datos se comportan de manera Normal. Es posible observar que al aumentar el nivel de correlación de los datos, disminuye la calidad de la precisión. Sin embargo, si la distribución es Gamma, el comportamiento observado es inverso, ya que se ve beneficiado por incrementos en el nivel de correlaciones. A pesar de esto, este modelo es el que obtiene mejores resultados al existir pocas observaciones anómalas en el conjunto de datos analizados.

4.8 UMAP

En cuanto a los resultados obtenidos con UMAP se observa un comportamiento homólogo al observado con el K-means, donde ambas técnicas parecen beneficiarse de incrementos en la población y la cantidad de anomalías observadas según el AUC-ROC. Aunque la calidad de los resultados no es tan alta para las poblaciones con distribuciones gamma en comparación con las normales, las variabilidades son relativamente similares. Es importante destacar que existe cierta convergencia que permite obtener valores relativamente altos tanto para la distribución normal como para la distribución gamma al parametrizar adecuadamente la técnica. Asimismo, en el caso de la distribución Normal, la ausencia de correlación parece afectar sig-

nificativamente los datos cuando no existe una población suficientemente alta. Lo mismo se observa en el caso de la distribución gamma, aunque no se logra incrementar el nivel de detección.

En el caso del modelo UMAP, se tiene una calidad relativamente baja respceto a la métrica de AUC-PR. Tanto para la Distribución normal como Gamma, el valor de este indicador se ubica entre el 55% y el 43% (similar al observado en el caso de K Medias). Estos resultados parecen no ser seriamente afectados por los niveles de correlación, pero sí parecen ser más sensibles ante cambios en el porcentaje de anomalías presentes en los datos.

Finalmente, para el modelo UMAP, se observa que en el caso de la distribución normal hay una fuerte afectación de los resultados de la Precisión Global cuando los datos no están correlacionados y además son menores a 200 mil observaciones. Pero posterior a dicho umbral, el valor se establece en un resultado del 97%, lo cual confirma la calidad del modelo al tener incrementos poblacionales. Ahora bien, si se considera el caso de la distribución Gamma, dicho equilibrio se alcanza con la misma condición, permitiendo a su vez tener valores ubicados entre el rango del 93% y el 97%. En el caso de la cantidad de anomalías, la ausencia de correlación afecta en el caso de la distribución Normal, ubicando sus valores cerca del 55%. Mientras que incrementos en la correlación permiten ubicar dicha precisión global entre el 95% y el 100%. Además, en el caso de la distribución Gamma, dichos valores siempre se ubican por encima del 90%, con muy poca variabilidad ante cambios de la correlación y la cantidad de anomalías observadas entre los datos.

Finalmente, en el caso del modelo UMAP los resultados de la Precisión son considerablemente altos en comparación con otras técnicas. Específicamente, cuando se analizan datos con distribuciones normales, el ámbito de la Precisión se ubica entre el 62% y el 44%. Ahora bien, si la distribución es Gamma, dicho ámbito se ubica entre el 7% y el 44%, siendo su punto máximo alcanzado para ambas distribuciones cuando se posee una correlación del 20%.

En el caso de la cantidad de observaciones anómalas, la mayor cantidad de las mismas incrementa el resultado obtenido en la Precisión, mientras que una menor cantidad parece tener un efecto que afecta considerablemente la precisión. Adicionalmente, ante la ausencia de correlación en los datos y al comportarse como una distribución normal, el ámbito resultante en la precisión se ubica entre el 49% y el 76%, mientras

que si la distribución es Gamma el ámbito se ubica entre el 0.03% y un 23%. En ambos casos, el valor de la precisión aumenta con la correlación observada. Además, en el caso de la distribución Gamma, se observa el mismo fenómeno que en el caso de la distribución Normal: la calidad desciende al incrementar la correlación, descendiendo hasta el 92% cuando la correlación se ubica en un 80%. Adicionalmente, la sensibilidad incrementa cuanto mayor es la cantidad de observaciones anómalas, descendiendo en el caso de la correlación más alta analizada, de un 95% cuando la cantidad de anomalías apenas es de un 0.1%, a un 88% cuando la cantidad de anomalías observadas es de un 5%.

4.9 Mejores resultados individuales

Posterior al análisis de los modelos con los criterios de evaluación empleados, se analizó cada una de las variantes empleadas con el objetivo de determinar cuál de las mismas permite una mejor calibración con el conjunto de datos analizados (el detalle puede observarse en el Anexo A3), en este caso los modelos que destacaron con el conjunto de métricas aplicadas son los siguientes:

- **Análisis de Componentes Principales:** muestra los mejores resultados en el caso cuando los datos no son escalados y centrados, para los diversos niveles de correlación de los datos analizados, de todas las técnicas empleadas estos resultados demuestran ser los menos acertados en cada uno de los escenarios analizados para el caso de la distribución Normal, no así en el caso de la distribución Gamma.

- **Análisis de Componentes Principales Robustos:** los resultados óptimos son alcanzados cuando los parámetros son ajustados para considerar los valores de escalado y cambio de signo como verdaderos.

- K **Medias:** el modelo parece encontrar algún grado de estabilidad al acercarse a los 15 valores de K, lo cual es además beneficioso ya que permite disminuir los altos tiempos de procesamiento evidenciados en este tipo de técnicas.

- **Agregación de** K **Medias:** el modelo con mayor calidad de detección es

aquel con valor mínimo de K igual a 10, y un valor máximo de K igual a 60, siendo esta una de las técnicas más potentes analizadas en la presente tesis.

- **Valor Atítpico Local:** la calidad de la predicción aumenta con incrementos en la cantidad de K empleada, este modelo parece comportarse mejor en el caso de poblaciones con distribución Gamma, a pesar de esto en dichas poblaciones la variabilidad de sus resultados aumenta considerablemente mientras mayor sea la cantidad de poblaciones anómalas. Asimismo, en el caso de la distribución Normal la calidad de la predicción cae considerablemente al obtener una mayor cantidad de anomalías. Con lo anterior descrito, el mejor modelo empleado según estos datos es el modelo con 160 valores de K.

- **Bosques de Aislamiento:** parece tener un comportamiento bastante adecuado para la identificación de anomalías sin importar el tipo de distribución de los datos o el nivel de correlación de los mismos, como en la mayoría de modelos, cae su nivel de predicción mientras mayor sea el número de individuos a identificar, a pesar de esto cuando se analiza la desviación estándar de las observaciones se observa la menor variabilidad de los resultados, lo cual permite seleccionar a la variante Simple como la idónea para la aplicación final con los datos reales. Finalmente, en el caso del modelo Simple el modelo que posee una mayor calidad de detección y una menor variabilidad a lo largo de los escenarios es aquel que emplea 100 árboles de aislamiento.

- **Aproximación y Proyección de Manifolds Uniformes:** con un valor de K igual a 20 se logra estabilizar para el caso de la distribución Gamma, dado que ahí parece que los valores son ligeramente mayores de manera transversal entre los escenarios identificados, por otra parte, en el caso de la distribución Normal, existe realmente muy poca variabilidad sin importar la cantidad de valores de K empleados para realizar la identificación final, en este caso no parece existir una señal que permita prever mejores resultados por parte de un modelo o de otro, por lo tanto el valor de K óptimo a emplear será de 20.

CAPÍTULO 5
APLICACIÓN EN COMPRAS PÚBLICAS

Se realizó una aplicación con datos reales de las compras públicas de bienes llevadas a cabo por medio de la plataforma SICOP, considerando un período comprendido entre el año 2020 y el año 2022.

5.1 Análisis exploratorio

Utilizando el período comprendido entre enero del año 2020 y diciembre de 2022, se obtienen más de 298 mil observaciones, correspondiente a un total de 257 instituciones o figuras legales encargadas de las adjudicaciones.

En el presente estudio, se procedió a la imputación de los indicadores mediante el uso de percentiles. Esta técnica de análisis sustituyó los valores del 1% más alto de los indicadores con las calificaciones correspondientes al percentil 99%. Este reemplazo se llevó a cabo debido a ciertas dudas que surgieron acerca de la fiabilidad de la información utilizada[33]. Esta técnica de imputación se implementó con el fin de mitigar el efecto perjudicial de los valores extremos en la calidad de las predicciones de los resultados. Cabe mencionar, además, que en los indicadores empleados no se encontraron valores faltantes.

La Figura 5.23 muestra que la mayoría de los indicadores poseen comportamientos desbalanceados, donde algunas variables poseen distribuciones extremadamente asimétricas y otras poseen distribuciones con colas pesadas, asimismo, los indicadores no están correlacionados debido a que su creación consideró etapas totalmente independientes entre sí.

[33]Un ejemplo revelador de esto se observó en el resultado de un indicador que comparaba el monto adjudicado con el monto estimado por la unidad contratante. El resultado mostraba cifras considerablemente superiores a un millón, a pesar de que sería lógico esperar valores no mayores a una decena. A través de revisiones manuales de dichos casos, se constató que en ocasiones, las Administraciones completaban el espacio del monto estimado con tan solo 1 colón, lo que ocasionaba que cualquier monto adjudicado diera lugar a cifras en miles o millones, lo cual podría enmascarar otros comportamientos cuyas anomalías fueran auténticas.

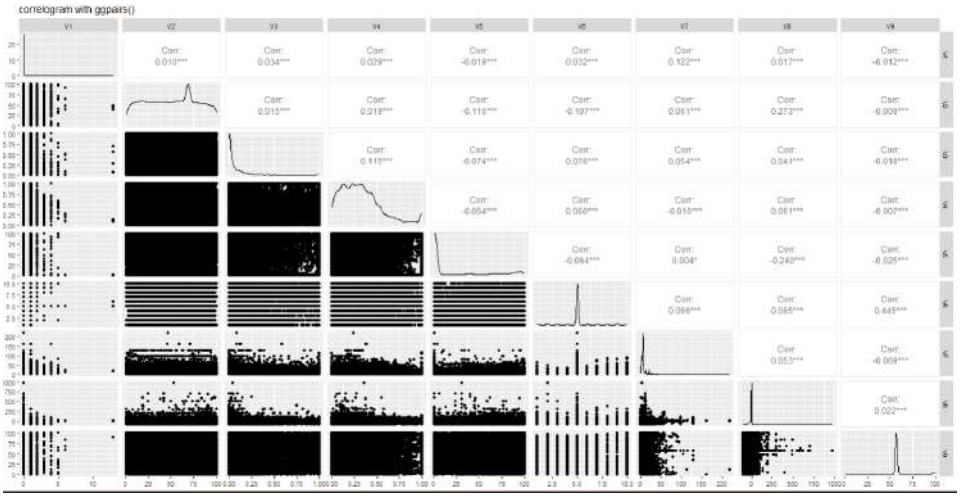

Figura 5.23. Descriptivos de indicadores finales empleados

5.2 Análisis de las compras de bienes (2020 a 2022)

5.2.1 Contextualización General

Los resultados se analizaron desde diferentes aristas, tal como la naturaleza jurídica de las mismas, el tipo de bien contratado, la modalidad o tipo de procedimiento y no por instituciones de manera independiente, esta situación se debe a que se han encontrado algunos factores relacionados a la calidad de los datos, que parecen indicar que el modelo acá presentado en una primera instancia debe de emplearse en la limpieza de los datos para remover los patrones que generan ruido en la información y no como técnicas de predicción, (Amer y Abdennadher, 2011; Ghafoori, 2018).

Es importante considerar que los indicadores no están diseñados para encontrar anomalías que desde el concepto de contratación pública puedan considerarse como acciones inadecuadas realizadas por las Administraciones, en este caso el enfoque permite identificar también aquellas contrataciones que tengan un comportamiento anómalo positivo, por lo tanto cualquier conclusión derivada de la presente tesis debe de ser vista como un paso hacia procesos de fiscalización o valoración posterior, y no tanto como hallazgos puntuales en términos de contratación administrativa.

Los indicadores fueron sometidos a los mejores enfoques de detección de anomalías según los resultados obtenidos en la Sección 4.1, con el objetivo de determinar el comportamiento de cada uno de ellos. Con base en criterio experto se estableció un supuesto de que el 5% de las contrataciones totales podrían considerarse como anómalas, por lo tanto el umbral se generó según el puntaje obtenido para cada

metodología.

Al tratarse de una identificación no supervisada, se emplea un concepto de consenso, tal como los propuestos por Kainulainen *et al.* (2011); López-Iturriaga y Sanz (2018), quienes llegaron a la conclusión que una combinación de resultados en un modelo híbrido, da mayor precisión que cualquiera de las técnicas utilizadas de manera independiente.

Además de la premisa anterior, se generó una nota final con el objetivo de identificar si la anomalía identificada se trató de una anomalía fuerte o una anomalía débil, tal como lo establece Aggarwal (2017). De esta manera por ejemplo, si el umbral definido para definir si una observación es anómala se estableció como 0.5, y el valor máximo observado en dicho fue de 0.9, se asignaron calificaciones ubicadas en el rango de 10 a 100, de esta manera es posible promediar las notas finales obtenidas por todas las anomalías identificadas y de esta forma se determinó cuales anomalías podrían definirse como las más fuertes identificadas en la mayor cantidad de modelos posibles.

5.2.2 Correlación entre los modelos

Un elemento que se comprueba con el análisis general de la aplicación, es el hecho que existe una alta correlación entre ciertos tipos de detección de anomalías, por ejemplo, el modelo de K Medias y UMAP están positivamente correlacionados (Cuadro 5.4), esto tiene lógica ya que como se indicó previamente luego de la proyección bidimensional realizada con UMAP se emplea la técnica de K Medias para identificar las distancias entre los grupos y por lo tanto definir el comportamiento anómalo.

Cuadro 5.4. Correlación* de las anomalías detectadas los modelos analizados

	KMEANS	ACPR	AGR_K	ISO	LOF	UMAP	Nota Final
KMEANS	1.00	-0.01	0.06	0.15	0.01	0.64	0.41
ACPR	-0.01	1.00	-0.00	0.18	-0.01	-0.02	0.20
AGR_K	0.06	-0.00	1.00	0.12	0.16	0.01	0.34
ISO	0.15	0.18	0.12	1.00	0.05	0.14	0.46
LOF	0.01	-0.01	0.16	0.05	1.00	0.00	0.25
UMAP	0.64	-0.02	0.01	0.14	0.00	1.00	0.36
Nota Final	0.41	0.20	0.34	0.46	0.25	0.36	1.00

* Al tratarse de variables dicotómicas, se empleó el coeficiente de correlación phi.

5.2.3 Resultados según variaciones en bien así como la metodología de contratación

Si se analiza la identificación de anomalías según el tipo de procedimiento tal como se detalla en el Cuadro 5.5, es posible observar como por ejemplo, proporcionalmente procedimientos como la Licitación Internacional (LI), o la Licitación Pública Nacional (LN), tienen una gran proporción de anomalías identificadas en modelos como el de Bosques de Aislamiento, o bien el de Componentes Principales Robustos, esto podría ser explicado ya que en realidad ambos tipos de procedimientos representan menos del 3% por ciento del total de las observaciones analizadas, asimismo, debido a la naturaleza de estos procedimientos de contratación, los montos transados deben de haber sido de altas cuantías, con procesos más largos y complejos, lo cual por ejemplo no deberían de compararse de manera simultánea con Contrataciones Directas.

Al igual que en el caso anterior, la detección de anomalías según la modalidad

Cuadro 5.5. Proporción promedio de anomalías según el tipo de procedimiento analizado

Tipo de procedimiento	Modelos analizados					
	ISO	AG_K	ACPR	KMEAN	LOF	UMAP
Contratación Directa	0.03	0.05	0.04	0.05	0.05	0.05
Contratación Especial	0.07	0.15	0.07	0.02	0.06	0.03
Licitación Abreviada	0.10	0.06	0.05	0.04	0.06	0.03
Licitación Internacional	0.37	0.09	0.00	0.00	0.29	0.00
Licitación Publica Nacional	0.25	0.06	0.21	0.11	0.05	0.10
Procedimiento por Principio	0.05	0.05	0.04	0.03	0.07	0.03

Nota: En este caso, la cantidad de observaciones totales para cada tipo de procedimiento es el siguiente: Contratación Directa, 233463; Licitación Abreviada, 45231; Licitación Publica Nacional, 10108; Procedimiento por Principio, 5321; Contratación Especial, 4766; Licitación Internacional, 70.

de procedimiento indicada en el Cuadro 5.6, permite observar un comportamiento similar, donde aquellos procedimientos pocos comunes como los Convenios Marco, la Ejecución por consignación o el arrendamiento de inmuebles sin opción de compra, aparecen identificados en altas proporciones por algunas de las técnicas acá presentadas, esto en si mismo no es un problema, pero al limitar el comportamiento anómalo al 5% de los datos obtenidos, podría estar desplazando a otras contrataciones cuyo comportamiento atípico pasaría como desapercibido por las técnicas presentadas.

Cuadro 5.6. Porcentaje promedio de anomalías según el modalidad de procedimiento analizado

Modalidad de procedimiento	Modelos analizados					
	ISO	AG_K	ACPR	KMEAN	LOF	UMAP
Arrendamiento de inmuebles sin opción de compra	0.00	1.00	0.00	0.00	0.00	0.00
Cantidad definida	0.03	0.04	0.05	0.05	0.05	0.05
Contratación especial	0.07	0.15	0.07	0.02	0.06	0.03
Convenio marco	0.44	0.08	0.40	0.29	0.03	0.29
Ejecución por consignación	0.48	0.07	0.05	0.00	0.03	0.00
Según demanda	0.10	0.06	0.05	0.03	0.05	0.03
Servicios	0.05	0.06	0.04	0.05	0.07	0.06

Nota: En este caso, la cantidad de observaciones totales para cada modalidad de contratación es el siguiente: Cantidad definida, 224916; Según demanda, 64870; Contratación especial, 4766; Servicios, 3145; Convenio marco, 1045; Ejecución por consignación, 216; Arrendamiento de inmuebles sin opción de compra, 1.

Finalmente, respecto a la falta de comparabilidad y posible afectación relacionada a la detección de valores anómalos, se puede observar como empleando el primer dígito de la raíz del producto explicitada en la Figura 2.17, se observa que aquellos bienes menos comunes son detectados en términos porcentuales de una forma más elevada; como ejemplo de esto, el código 6 referido a "Instrumentos musicales, juegos, juguetes, artesanía y equipamiento, material, accesorios y suministros para educación" se detectó tal como se observa en el Cuadro 5.7 como anómalo en más del 50% de las observaciones disponibles en la población analizada. Estas conclusiones son un punto importante a considerar cuando se proceda en estudios futuros con otros análisis más complejos como es el caso de las contrataciones de servicios.

Cuadro 5.7. Porcentaje promedio de anomalías según el código de los bienes analizados

Familia de bienes*	Modelos analizados					
	ISO	AG_K	ACPR	KMEAN	LOF	UMAP
1	0.15	0.05	0.05	0.01	0.05	0.00
2	0.04	0.06	0.04	0.04	0.06	0.00
3	0.01	0.02	0.05	0.00	0.04	0.00
4	0.03	0.04	0.06	0.03	0.05	0.00
5	0.08	0.12	0.04	0.13	0.05	0.25
6	0.52	0.16	0.04	1.00	0.08	1.00

*1: Materia Prima, Químicos, Papel, Combustible; 2: Equipos e Instrumentos Industriales Componentes y Suministros; 3: Equipo y Suministros de Construcción, Transporte e Instalaciones; 4: Equipos y Suministros Médicos, de Laboratorio y de Pruebas y Farmacéuticos; 5: Equipos y Suministros para la Industria Alimentaria, de Limpieza y Servicios; 6: Equipos y Suministros para Negocios, Comunicación y Tecnología.

Nota: En este caso, la cantidad de observaciones totales para cada código de bien es el siguiente: código 4, 131147; código 3, 63360; código 5, 42417; código 2, 35743; código 1, 22155; código 6, 4137.

Con lo analizado queda claro que un elemento importante es que la aplicación de

estas técnicas deberían de realizarse luego de someter a un primer escrutinio de la información, donde se establezcan marcos comparativos similares sin que esto llegue a generar un sesgo de selección de la población muestral analizada.

5.2.4 Análisis según sector analizado

Respecto a los sectores donde se presentan una cantidad de anomalías mayor, no fue posible generar una clasificación exacta de las 257 instituciones o figuras legales encargadas de los procesos de adjudicación, esto ya que las clasificaciones existente por parte del Ministerio de Planificación y Política Económica (MIDEPLAN), no contemplan la totalidad de los Órganos auxiliares, o bien los Fideicomisos que participan en SICOP de manera directa, de esta manera y con el conocimiento que el Sector Público Costarricense es cambiante, se emplea como primera base el listado de las instituciones públicas según naturaleza jurídica actualizado por MIDEPLAN en abril del año 2023, realizando una asignación aproximada de los órganos no clasificados previamente.

Luego de la anterior clasificación, se generó el análisis para determinar los niveles de detección según las diversas técnicas tal como se observa en el Cuadro 5.8, donde llama la atención que el porcentaje de anomalía es aproximadamente uniforme para estos sectores por parte de todos los modelos analizados en el caso de las Instituciones Autómonmas, semiautónomas, así como en el caso de las Municipalidades y Ministerios con sus órganos adscritos. Este tipo de identificación requiere de pasos adicionales con los cuales se pueda profundizar sobre la naturaleza de las anomalías, intentando la construcción de bases de datos en las cuales se tenga certeza de las anomalías positivas (posiblemente relacionado a buenas prácticas institucionales) y negativas (provocado por puntos de mejora en los procesos de adquisición),

Es relevante destacar que la Figura 5.24 ilustra que la calificación mediante el consenso se sitúa, en la mayoría de los casos, entre los porcentajes de anomalías máximas y mínimas identificadas por los diferentes modelos. Este hallazgo sugiere que la aplicación de un consenso en las calificaciones logra una suavización en el porcentaje de detección observado, permitiendo que los modelos que detectan un alto porcentaje de anomalías no eleven la cantidad de observaciones que deban de revisarse mediante

Cuadro 5.8. Porcentaje promedio de anomalías según la distribución por sectores establecidos en MIDEPLAN

Sector institucional	Modelos analizados					
	ISO	AG_K	ACPR	KMEAN	LOF	UMAP
Asociaciones de desarrollo	0.03	0.05	0.06	0.03	0.03	0.03
Concejos municipales de Distrito	0.02	0.02	0.01	0.07	0.03	0.03
Empresas públicas estatales	0.11	0.07	0.08	0.02	0.10	0.03
Empresas públicas no estatales	0.03	0.03	0.02	0.02	0.07	0.02
Entes públicos no estatales	0.04	0.05	0.05	0.04	0.06	0.03
Fideicomisos	0.19	0.04	0.13	0.06	0.02	0.04
Instituciones autónomas	0.07	0.08	0.05	0.06	0.06	0.07
Instituciones semiautónomas	0.03	0.07	0.05	0.06	0.04	0.05
Ministerios y órganos adscritos	0.04	0.05	0.05	0.05	0.05	0.04
Municipalidades	0.03	0.02	0.05	0.04	0.04	0.04
Organismo electoral	0.02	0.04	0.06	0.03	0.05	0.06
Poderes de la República	0.13	0.07	0.06	0.02	0.04	0.14
Órganos adscritos a Instituciones Auton.	0.04	0.08	0.02	0.08	0.10	0.05
Órganos adscritos al sector municipal	0.02	0.05	0.04	0.18	0.04	0.02
Órganos del Poder Legislativo	0.05	0.06	0.04	0.04	0.05	0.05

Nota: En este caso, la cantidad de observaciones totales para cada sector es el siguiente: Instituciones autónomas, 113206; Municipalidades, 112569; Ministerios y órganos adscritos, 43098; Empresas públicas estatales, 6092; Instituciones semiautónomas, 4880; Organismo electoral, 4033; Empresas públicas no estatales, 4032; Órganos adscritos al sector municipal, 3718; Entes públicos no estatales, 3150; Órganos del Poder Legislativo, 1090; Poderes de la República, 771; Órganos adscritos a Instituciones Autónomas, 731; Concejos municipales de Distrito, 612; Fideicomisos, 566; Asociaciones de desarrollo, 411.

análisis manuales.

Este tipo de consenso podría ser de suma importancia al llevar a cabo investigaciones enfocadas en sectores que presentan una mayor incidencia, ya que se aprovecharía la combinación de las mejores técnicas disponibles sin incrementar de manera sustancial los tamaños muestrales para la revisión en campo de los casos anómalos.

5.3 Evaluación con segmentación según el tipo bienes

En función de lo observado previamente, existen diferencias según el valor poblacional relacionado a cada tipo de bienes que se tranza a nivel de la plataforma de SICOP. De esta manera, se procedió a realizar un ejemplo de evaluación empleando una segmentación de la población según los diferentes tipos de bienes adquiridos, con el objetivo de analizar cómo se comporta la detección de anomalías según las subpoblaciones y con ello demostrar las diferencias posibles que pueden encontrarse en la detección a la hora de realizar este análisis segmentado.

Se realizó una comparación entre la detección de anomalías para cada uno de los tipo de bienes, según las técnicas preestablecidas se realizó la comparación entre los

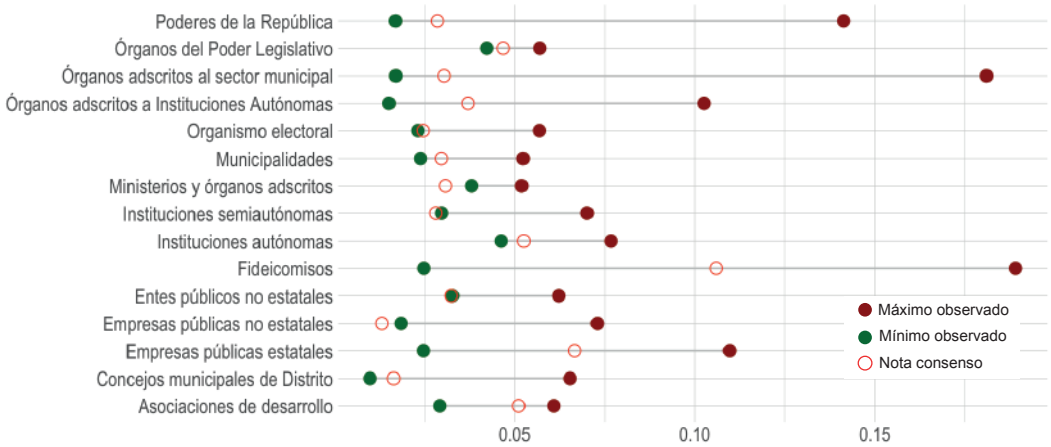

Figura 5.24. Comparación del porcentaje de anomalías detectadas, considerando máximos, mínimos y la nota final de consenso aplicada

resultados obtenidos y los resultados generales de la aplicación, bajo la premisa de que los valores generales son los valores verdaderos que se busca identificar.

Tal como se observa en el Cuadro 5.9, se realiza una matriz de confusión con los resultados obtenidos para cada uno de los códigos de los bienes, donde se dan resultados diferentes en la mayoría de los códigos exceptuando el sexto (el del tamaño muestral más pequeño) donde los resultados únicamente se terminaron categorizando como verdaderos positivos y falsos positivos.

Cuadro 5.9. Resultados de la matriz de confusión entre los valores obtenidos a nivel general y a nivel de las sub bases analizadas

Código del bien	Resultados Matriz Confusión			
	VP	VN	FP	FN
1	2056	14925	1841	3333
2	2193	25022	3910	4618
3	2551	48353	8232	4114
4	8020	93827	15325	13975
5	4363	20610	2618	14826
6	725	0	3412	0

*VP: Verdadero positivo; VN: Verdadero negativo; FP: Falso positivo; FN: Falso negativo.
Nota: En este caso, la cantidad de observaciones totales para cada código de bien es el siguiente: código 4, 131147; código 3, 63360; código 5, 42417; código 2, 35743; código 1, 22155; código 6, 4137.

Es posible observar en la Figura 5.25 en la mayoría de los casos los verdaderos valores negativos son identificados como tales, lo cual tiene lógica debido a la alta proporción de los resultados finales que componen, ahora bien, se observan una alta proporción de falsos negativos en la mayoría de las sub poblaciones, tal como se ve en el caso de los bienes cuyo código inicia con 5.

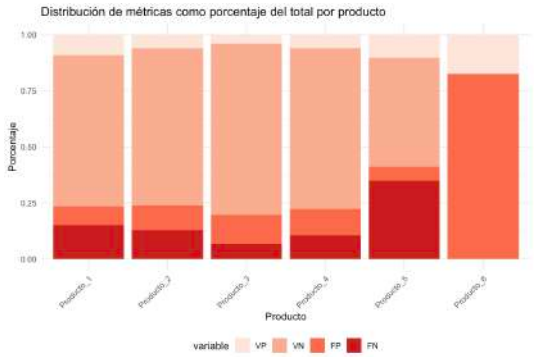

Figura 5.25. Comparación de resultados según tipo de bienes y porcentaje del total

Esta situación se logra evidenciar de mejor manera cuando se analiza el porcentaje de detección para cada bien con cada uno de los modelos empleados tal como se observa en el Cuadro 5.10, en este caso, tomando como ejemplo la detección del modelo de Bosques de Aislamiento, se observa como cambia la detección total según el tipo de de sector analizado en cada una de las submuestras, por ejemplo, al considerar el nivel de detección general las Asociaciones de desarrollo poseían cerca de un 3% del total como anómalo, mientras que al emplear la submuestra en los códigos 1, 4 y 6, no se detectaron anomalías con esta técnica, y por el contrario con el código 2 y 3 el valor estuvo sobre el 13% del total de las observaciones, es importante además el observar como el modelo con mayores observaciones parece que el nivel de detección comienza a converger, tal como se observa con el caso de las Municipalidades y las Instituciones Autónomas, donde la cantidad de observaciones es mayor.

Lo anterior, es posible de observar también con el caso del tipo de procedimiento, en donde la Contratación Directa que es el procedimiento más empleado, tiende a cierta convergencia, mientras que itris como los Procedimientos por Principio muestra una muy alta variabilidad de identificación entre los diversos códigos, identificado esto en el Cuadro 5.11.

Finalmente, en el caso del Cuadro 5.12, se observa el mismo efecto según la modalidad del procedimiento, donde algunos códigos de bienes parecen ser más sensibles de detección al considerar el modelo ISO.

Cuadro 5.10. Resultados de la evaluación por el modelo ISO en cada tipo de subpoblación

	Sector MIDEPLAN	Cod_1	Cod_2	Cod_3	Cod_4	Cod_5	Cod_6
1	Asociaciones de desarrollo	0.00	0.13	0.18	0.00	0.02	0.00
2	Concejos municipales de Distrito	0.01	0.00	0.02	0.03	0.00	0.00
3	Empresas públicas estatales	0.20	0.24	0.33	0.17	0.05	0.25
4	Empresas públicas no estatales	0.06	0.06	0.10	0.03	0.00	0.57
5	Entes públicos no estatales	0.02	0.09	0.24	0.05	0.00	0.00
6	Fideicomisos	0.10	0.10	0.57	0.17	0.29	0.21
7	Instituciones autónomas	0.08	0.06	0.07	0.08	0.07	0.05
8	Instituciones semiautónomas	0.03	0.00	0.02	0.02	0.02	0.06
9	Ministerios y órganos adscritos	0.05	0.07	0.05	0.03	0.03	0.09
10	Municipalidades	0.03	0.03	0.03	0.02	0.03	0.04
11	Organismo electoral	0.04	0.02	0.05	0.02	0.01	0.00
12	Poderes de la República	0.40	0.05	0.08	0.03	0.02	0.00
13	Órganos adscritos a Instituciones Autónomas	0.06	0.06	0.39	0.06	0.02	0.00
14	Órganos adscritos al sector municipal	0.00	0.01	0.01	0.01	0.03	0.08
15	Órganos del Poder Legislativo	0.05	0.05	0.01	0.05	0.04	0.00

Nota: En este caso, la cantidad de observaciones totales para cada sector es el siguiente: Instituciones autónomas, 113206; Municipalidades, 112569; Ministerios y órganos adscritos, 43098; Empresas públicas estatales, 6092; Instituciones semiautónomas, 4880; Organismo electoral, 4033; Empresas públicas no estatales, 4032; Órganos adscritos al sector municipal, 3718; Entes públicos no estatales, 3150; Órganos del Poder Legislativo, 1090; Poderes de la República, 771; Órganos adscritos a Instituciones Autónomas, 731; Concejos municipales de Distrito, 612; Fideicomisos, 566; Asociaciones de desarrollo, 411.

CAPÍTULO 6

CONCLUSIONES

6.1 Limitaciones

La calidad de la información proporcionada por SICOP presenta numerosas áreas de mejora, en virtud de esto es posible que diversos factores contribuyan a estas deficiencias en la calidad. Entre ellos se incluyen la falta de experiencia de los usuarios, una capacidad de revisión insuficiente por parte del Ministerio de Hacienda sobre la información que se ingresa en el sistema, así como una falta de transparencia por parte de las personas involucradas en el proceso de compras.

Un aspecto crítico se relaciona con las limitaciones informáticas del equipo utilizado para llevar a cabo el análisis. Estas limitaciones han restringido la profundidad de los análisis y han impedido un estudio más detallado de modelos con múltiples variables o combinaciones de variables con distintos tipos de distribuciones.

Además, la cantidad de información disponible en la Contraloría General de la República sobre las bases de SICOP es limitada. Esta limitación dificulta la creación

Cuadro 5.11. Resultados de la evaluación por el modelo ISO en cada tipo de subpoblación

	Tipo Procedimiento	Cod_1	Cod_2	Cod_3	Cod_4	Cod_5	Cod_6
1	Contratación Directa	0.03	0.03	0.03	0.02	0.02	0.03
2	Contratación Especial	0.14	0.10	0.05	0.07	0.17	0.17
3	Licitación Abreviada	0.10	0.15	0.13	0.12	0.12	0.19
4	Licitación Internacional		0.87	1.00	0.24		
5	Licitación Publica Nacional	0.26	0.32	0.29	0.23	0.22	0.11
6	Procedimiento por Principio	0.05	0.07	0.10	0.05	0.05	0.56

Nota: En este caso, la cantidad de observaciones totales para cada tipo de procedimiento es el siguiente: Contratación Directa, 233463; Licitación Abreviada, 45231; Licitación Publica Nacional, 10108; Procedimiento por Principio, 5321; Contratación Especial, 4766; Licitación Internacional, 70.

Cuadro 5.12. Resultados de la evaluación por el modelo ISO en cada tipo de subpoblación

Modalidad Procedimiento	Cod_1	Cod_2	Cod_3	Cod_4	Cod_5	Cod_6
Arrendamiento sin opción de compra			0.00			
Cantidad definida	0.03	0.03	0.03	0.02	0.02	0.04
Contratación especial	0.14	0.10	0.05	0.07	0.17	0.17
Convenio marco	0.60	0.06	0.38	0.68	0.48	
Ejecución por consignación	0.00			0.60		
Según demanda	0.14	0.12	0.12	0.12	0.11	0.11
Servicios	0.01	0.04	0.08	0.05	0.09	0.21

Nota: En este caso, la cantidad de observaciones totales para cada modalidad de contratación es el siguiente: Cantidad definida, 224916; Según demanda, 64870; Contratación especial, 4766; Servicios, 3145; Convenio marco, 1045; Ejecución por consignación, 216; Arrendamiento de inmuebles sin opción de compra, 1.

de otros indicadores que podrían facilitar de manera más eficiente la detección de valores anómalos en la información, asimismo, el rendimiento de los modelos se ve afectado debido a lo mezclados que son los comportamientos de las variables seleccionadas, lo cual limita la capacidad final de detección y la falta de estandarización de los modelos más óptimos posibles.

6.2 Discusión

Las técnicas de identificación de valores anómalos bajo enfoques no supervisados pueden considerarse una iniciativa destacada que busca generar debates sobre diversas problemáticas que pueden ser evaluadas desde el punto de vista de la identificación de comportamientos atípicos en la información. Específicamente, en el caso de la detección de anomalías en la contratación pública de bienes, es importante llevar a cabo un profundo análisis acerca de las identificaciones realizadas. No es posible

determinar si dichos comportamientos atípicos son buenos o malos, por lo que es necesario desarrollar indicadores específicos si se pretende implementar análisis de este tipo para la identificación de casos de corrupción u otros delitos relacionados con la contratación pública.

En el ámbito estadístico, el comportamiento observado mediante las técnicas empleadas resalta la necesidad de abordar su aplicación con un profundo conocimiento sobre el comportamiento de las variables analizadas. Una calibración adecuada puede generar resultados completamente opuestos a aquellos obtenidos mediante un enfoque laxo en la selección de parámetros.

A nivel general, se ha observado que algunos modelos requieren un alto poder de procesamiento computacional. Por ejemplo, el caso de K Medias o UMAP demanda largos tiempos de procesamiento incluso con una adecuada paralelización. Sin embargo, estos esfuerzos no siempre se traducen en mejores resultados. Por otro lado, modelos como ACPR pueden ofrecer resultados más rápidos y una alta capacidad de identificación de valores anómalos.

Como se ha analizado a lo largo de la tesis, es importante enfocar los esfuerzos de aplicación de la técnica. Aunque se trata de una metodología pensada para el Aprendizaje Automático, relacionada implícitamente con grandes volúmenes de información, la evidencia ha demostrado que no todos los modelos responden de la misma manera ante el incremento del tamaño de la muestra. Además, mezclar distintos tipos de poblaciones puede llevar a la identificación de peculiaridades relacionadas directamente con la naturaleza intrínseca de la información y no tanto con los patrones o comportamientos que podrían aportar un mayor valor agregado a la aplicación de las técnicas.

En cuanto a las metodologías de evaluación de las simulaciones, tras realizar un análisis integral, se llegó a las mismas conclusiones que los múltiples autores citados en esta tesis. El uso de las áreas bajo la curva ROC y PR representa la forma más completa de medir la calidad de las simulaciones. Aunque en esta investigación se centra principalmente en la detección de valores positivos, no se debe descartar que, en el caso de las metodologías no supervisadas, la identificación de valores normales fortalece los hallazgos y brinda una mayor confianza al posible usuario de los resultados obtenidos.

Como parte de la experiencia en el manejo de información de SICOP, se puede

concluir que la mayoría de los errores y anomalías identificados con los modelos de detección se deben a errores humanos en el ingreso de la información. Estos factores podrían solucionarse mediante capacitación y la implementación de mecanismos adecuados para la gestión y control de la información, así como controles transaccionales del propio sistema, lo cual facilitaría la labor de fiscalización.

6.3 Trabajos futuros

En el contexto de futuras investigaciones, se presentan diversas áreas que podrían profundizar en los estudios aquí expuestos. Por ejemplo, sería relevante verificar el poder predictivo de los modelos en otros tipos de distribuciones de información y determinar su nivel de impacto en poblaciones con distribuciones mixtas, como sería de esperar en casos de aplicación real.

Además, es necesario ahondar en la aplicación de estas técnicas utilizando conjuntos de datos diversos. Esto permitiría fortalecer la precisión y calidad de los modelos empleados, así como generar indicadores cuyo objetivo final no sea únicamente la detección de anomalías, sino más bien especificar cómo estas anomalías pueden ser negativas, como en el caso de la detección de corrupción.

BIBLIOGRAFÍA

Aggarwal, C. C. (2017). An introduction to outlier analysis. En *Outlier analysis*, pp. 1–34. Springer.

Aggarwal, C. C. y Yu, P. S. (2001). Outlier detection for high dimensional data. En *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pp. 37–46.

Amat Rodrigo, J. (2017). Análisis de componentes principales (principal component analysis, pca) y t-sne.

Amat Rodrigo, J. (2020). Detección de anomalías: Autoencoders y pca.

Amer, M. y Abdennadher, S. (2011). Comparison of unsupervised anomaly detection techniques. *Bachelor's Thesis*.

Angiulli, F. y Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. En *European conference on principles of data mining and knowledge discovery*, pp. 15–27. Springer.

Arnold, B. C. y Arvanitis, M. (2021). On a general class of gamma based copulas. *Dependence Modeling*, 9(1):374–384.

Bengtsson, H. (2023). *progressr: An Inclusive, Unifying API for Progress Updates*. R package version 0.13.0.

Breiman, L., Friedman, J. H., Olshen, R. A., y Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., y Sander, J. (2000). Lof: identifying density-based local outliers. En *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

CAF (2021). Experiencia: Datos e inteligencia artificial en el sector público.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., y Houle, M. E. (2016). On the evaluation of unsupervised outlier

detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30:891–927.

Carvalho Rocha, W. F., Nogueira, R., da Silva, G. E. B., Queiroz, S. M., y Sarmanho, G. F. (2013). A comparison of three procedures for robust pca of experimental results of the homogeneity test of a new sodium diclofenac candidate certified reference material. *Microchemical Journal*, 109:112–116.

CGR (2019). Transformación hacia una mayor eficiencia de las compras públicas electrónicas, beneficios y ahorros de la unificación. *Contraloría General de La República*.

Cortes, D. (2022). *isotree: Isolation-Based Outlier Detection*. R package version 0.5.5.

Costa Rica Integra, T. I. y. N. C. f. S. C. N. (2021). Estrategia nacional de integridad y prevención de la corrupción.

Craven, M. y Bockhorst, J. (2004). Markov networks for detecting overalpping elements in sequence data. *Advances in Neural Information Processing Systems*, 17.

Crous, C., Lamprecht, J., Eilifsen, A., Messier Jr, W., Glover, S., y Prawitt, D. (2012). *EBOOK: Auditing and Assurance Services*. McGraw Hill.

Davis, J. y Goadrich, M. (2006). The relationship between precision-recall and roc curves. En *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Della Porta, D. y Vannucci, A. (2002). Corruption and public contracts: some lessons from the italian case. *Corrupt exchanges: empirical themes in the politics and political economy of corruption. Baden-Baden, Germany: Nomos Verlagsgesellschaft*.

DFOE (2021). Uso del sistema integrado de compras públicas (sicop) en las instituciones públicas. *Contraloría General de La República*.

Dowle, M. y Srinivasan, A. (2021). *data.table: Extension of 'data.frame'*. R package version 1.14.2.

Eilifsen, A., Messier, W. F., Glover, S. M., y Prawitt, D. F. (2014). Auditing and assurance services.

Emmott, A., Das, S., Dietterich, T., Fern, A., y Wong, W.-K. (2015). A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158*.

Fayzrakhmanov, R., Kulikov, A., y Repp, P. (2018). The difference between precision-recall and roc curves for evaluating the performance of credit card fraud detection models. En *Proceedings of International Conference on Applied Innovation in IT*, volumen 6, pp. 17–22. Anhalt University of Applied Sciences.

Fernandez, A. y Plumbley, M. D. (2021). Using umap to inspect audio data for unsupervised anomaly detection under domain-shift conditions. *arXiv preprint arXiv:2107.10880*.

Ferwerda, J. y Deleanu, I. (2013). Identifying and reducing corruption in public procurement in the eu.

Ferwerda, J., Deleanu, I., y Unger, B. (2017). Corruption in public procurement: finding the right indicators. *European Journal on Criminal Policy and Research*, 23:245–267.

Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine*, 53(2):1385–1390.

Ghafoori, Z. (2018). *Robust and efficient unsupervised anomaly detection in complex and dynamic environments*. Tesis doctoral.

Gogoi, P., Borah, B., y Bhattacharyya, D. K. (2010). Anomaly detection analysis of intrusion data using supervised & unsupervised approach. *J. Convergence Inf. Technol.*, 5(1):95–110.

Goix, N. (2016). How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*.

Goutte, C. y Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. En *European conference on information retrieval*, pp. 345–359. Springer.

Grau, J., Grosse, I., y Keilwagen, J. (2015). Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, 31(15):2595–2597.

Grolemund, G. y Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

Gutman, J. (2014). Is there room for discretion? reforming public procurement in a compliance-oriented world. *Global Economy & Development Working Paper*, 74.

Harary, F., Hayes, J. P., y Wu, H.-J. (1988). A survey of the theory of hypercube graphs. *Computers & Mathematics with Applications*, 15(4):277–289.

Hariri, S., Kind, M. C., y Brunner, R. J. (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*.

Hautamaki, V., Karkkainen, I., y Franti, P. (2004). Outlier detection using k-nearest neighbour graph. En *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volumen 3, pp. 430–433. IEEE.

Hawkins, D. M. (1980). *Identification of outliers*, volumen 11. Springer.

Hennig, C. (2018). Some thoughts on simulation studies to compare clustering methods. *Arch Data Sci Ser A*, 5(1):1–21.

Henry, L. y Wickham, H. (2022). *purrr: Functional Programming Tools*. R package version 0.3.5.

Hilbert, D. (1935). Über die stetige abbildung einer linie auf ein flächenstück. En *Dritter Band: Analysis · Grundlagen der Mathematik · Physik Verschiedenes*, pp. 1–2. Springer.

Hofert, M., Kojadinovic, I., Maechler, M., y Yan, J. (2023). *copula: Multivariate Dependence with Copulas*. R package version 1.1-2.

Kainulainen, L., Miche, Y., Eirola, E., Yu, Q., Frénay, B., Séverin, E., y Lendasse, A. (2011). Ensembles of local linear models for bankruptcy analysis and prediction. *Case Studies In Business, Industry And Government Statistics*, 4(2):116–133.

Kandanaarachchi, S., Muñoz, M. A., Hyndman, R. J., y Smith-Miles, K. (2020). On normalization and algorithm selection for unsupervised outlier detection. *Data Mining and Knowledge Discovery*, 34(2):309–354.

Kassambara, A. y Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Konopka, T. (2023). *umap: Uniform Manifold Approximation and Projection*. R package version 0.2.10.0.

Kuhn, G. M. (1973). The phi coefficient as an index of ear differences in dichotic listening. *Cortex*, 9(4):450–457.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-93.

Laffont, J.-J. y Tirole, J. (1990). Adverse selection and renegotiation in procurement. *The Review of Economic Studies*, 57(4):597–625.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., y Malohlava, M. (2022). *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*. R package version 3.36.0.2.

Liu, F. T., Ting, K. M., y Zhou, Z.-H. (2008). Isolation forest. En *2008 eighth ieee international conference on data mining*, pp. 413–422. IEEE.

Liu, F. T., Ting, K. M., y Zhou, Z.-H. (2010). On detecting clustered anomalies using sciforest. En *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–290. Springer.

López-Iturriaga, F. J. y Sanz, I. P. (2018). Predicting public corruption with neural networks: An analysis of spanish provinces. *Social Indicators Research*, 140:975–998.

Madsen, J. H. (2018). *DDoutlier: Distance Density-Based Outlier Detection*. R package version 0.1.0.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., y Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).

Martínez, Ana and Torres, Luis M and others (2019). Compras públicas y big data: El caso mexicano.

Mayer, R., Hittmeir, M., y Ekelhart, A. (2020). Privacy-preserving anomaly detection using synthetic data. En *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 195–207. Springer.

McInnes, L. (2023). umap documentation release 0.5.

McInnes, L., Healy, J., y Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Monge, C. E. C. (2007). Contratación pública y corrupción: un análisis particular de los principios rectores de la contratación administrativa). *revista de Ciencias Jurídicas*, (112).

Moon, B., Jagadish, H. V., Faloutsos, C., y Saltz, J. H. (2001). Analysis of the clustering properties of the hilbert space-filling curve. *IEEE Transactions on knowledge and data engineering*, 13(1):124–141.

Morris, T. P., White, I. R., y Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Nachmani, E., Roman, R. S., y Wolf, L. (2021). Non gaussian denoising diffusion models. *arXiv preprint arXiv:2106.07582*.

Nelsen, R. (1999). *An Introduction to Copulas*. Lecture Notes in Statistics 139. Springer New York.

Nikolenko, S. I. (2019). Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512*.

Nooghabi, M. J., Nooghabi, H. J., y Nasiri, P. (2010). Detecting outliers in gamma distribution. *Communications in StatisticsTheory and Methods*, 39(4):698–706.

OCDE (2009). *OECD principles for integrity in public procurement*. OECD Publishing.

OCDE, O. (2019). Enhancing the use of competitive tendering in costa rica's public procurement system. *OCDE*.

Oyeyemi, G. e Ipinyomi, R. (2010). A robust method of estimating covariance matrix in multivariate data analysis. *African Journal of Mathematics and Computer Science Research*, 3(1):001–018.

Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, 36(1):157–160.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Pfeifer, D., Mändle, A., Ragulina, O., y Girschig, C. (2019). New copulas based on general partitions-of-unity (part iii)the continuous case. *Dependence Modeling*, 7(1):181–201.

Purwanto, A. y Emanuel, A. W. R. (2020). Data analysis for corruption indications on procurement of goods and services. En *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 56–60. IEEE.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabuzin, K. y Modrusan, N. (2019). Prediction of public procurement corruption indices using machine learning methods. En *KMIS*, pp. 333–340.

Ramaswamy, S., Rastogi, R., y Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. En *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 427–438.

Ripley, B. D. (1987). *Stochastic simulation*. John Wiley & Sons.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., y Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Rodriguez R., O., Navarro D., A., y Arroyo S., A. (2022). *traineR: Predictive (Classification and Regression) Models Homologator*. R package version 2.0.4.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., y Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING.

Singh, A. K. y Lalitha, S. (2018). Detection of upper outliers in gamma sample. *J. Stat. Appl. Probab. Lett*, 5:53–62.

Song, X., Wu, M., Jermaine, C., y Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on knowledge and Data Engineering*, 19(5):631–645.

Stage, F. K., Carter, H. C., y Nora, A. (2004). Path analysis: An introduction and analysis of a decade of research. *The journal of educational research*, 98(1):5–13.

Su, M.-Y. (2011). Real-time anomaly detection systems for denial-of-service attacks by weighted k-nearest-neighbor classifiers. *Expert Systems with Applications*, 38(4):3492–3498.

Suzuki, J. (2021). *Statistical Learning with Math and Python: 100 Exercises for Building Logic*. Springer Nature.

Swersky, L. (2018). A study of unsupervised outlier detection for one-class classification.

Tierney, L., Rossini, A. J., Li, N., y Sevcikova, H. (2021). *snow: Simple Network of Workstations*. R package version 0.4-4.

Todorov, V. y Filzmoser, P. (2009a). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.

Todorov, V. y Filzmoser, P. (2009b). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.

Venables, W. N. y Ripley, B. D. (2002a). *Modern Applied Statistics with S*. Springer, New York, fourth edición. ISBN 0-387-95457-0.

Venables, W. N. y Ripley, B. D. (2002b). *Modern applied statistics with S*. Springer Science & Business Media.

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., y Lestantyo, P. (2019). Cross-validation metrics for evaluating classification performance on imbalanced data. En *2019 international conference on computer, control, informatics and its applications (ic3ina)*, pp. 14–18. IEEE.

Wickham, H., François, R., Henry, L., Müller, K., y Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.2.

Wickham, H., Hester, J., y Bryan, J. (2022). *readr: Read Rectangular Text Data*. R package version 2.1.2.

Wiley, J. F. (2022). *JWileymisc: Miscellaneous Utilities and Functions*. R package version 1.3.0.

Wold, S., Esbensen, K., y Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Yan, J. (2012). Multivariate modeling with copulas and engineering applications. *Springer handbook of engineering statistics*, pp. 931–945.

Ye, Y., Huang, J. Z., Chen, X., Zhou, S., Williams, G., y Xu, X. (2006). Neighborhood density method for selecting initial cluster centers in k-means clustering. En *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 189–198. Springer.

Zimek, A., Schubert, E., y Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387.

Zuleta, María M and Ospina, Santiago and Caro, Carlos A and others (2019). Índice de riesgo de corrupción en el sistema de compra pública colombiano a partir de una metodología desarrollada por el instituto mexicano para la competitividad.

ANEXOS

A1 Simulación de datos

A1.1 Simulación de distribución normal multivariada

La creación de datos multivariados en el caso de la simetría normal se realizó mediante la metodología empleada por Stage *et al.* (2004), esta distribución multivariada es de las más comunes en p dimensiones, con media y dispersión como lo indica Ripley (1987). Asimismo, Venables y Ripley (2002b) explican como se producen muestras mediante un vector medio empírico especificado y una matriz de covarianza, lo cual permita la replicación de un conjunto de datos, considerando elementos como las medias muestrales y las correlaciones entre las variables analizadas.

A1.2 Simulación de distribución gamma multivariada

Para el caso de la simulación con poblaciones cuya distribución es Gamma, se realizó una creación de datos sintéticos empleando copulas para la creación de los mismos, tal como lo han empleado autores como Arnold y Arvanitis (2021); Pfeifer *et al.* (2019); Yan (2012).

Las cópulas propuestas por Nelsen (1999) son ampliamente utilizadas para modelar funciones de distribución multivariables, dado que asumen que todas las marginales son uniformes en el intervalo unitario. Esto se debe a que cualquier variable aleatoria continua puede ser transformada en una variable aleatoria uniforme en dicho intervalo mediante la transformación integral de probabilidad. Esta característica permite que las cópulas sean empleadas para combinar distintas marginales y construir nuevas distribuciones multivariables.

Yan (2012) indican que este método separa una distribución multivariable en dos componentes, las marginales y una cópula, lo que proporciona un marco muy flexible en la modelización multivariable.

Para el caso de la presente tesis, se consideraron las distribuciones marginales como gamma, además se especificó una cópula gaussiana con diversos niveles lo cual otorgó el nivel de correlación entre los datos analizados consistente con los escenarios planteados para las distribuciones Pfeifer *et al.* (2019)

A2 Tiempos promedio de ejecución de los modelos

Puede observarse en el Cuadro A2 como los modelos que se ejecuta más rápido en todos los escenarios son tanto el modelo PCA como el PCA robusto, siendo consecuente con los indicado por Aggarwal y Yu (2001), asimismo, la técnica de Bosques de Aislamiento evidencia un comportamiento poco demandante para la identificación de valores anómalos en los escenarios analizados, esta poca demanda de recursos es consecuencia de la ventaja subyacente relacionada con el límite de clasificación de los árboles, lo cual limita el análisis y no permite que se profundice más de lo necesario. En el caso de la técnica de LOF, es posible observar como se da un incremento en la duración del procesamiento al aumentar la población, a pesar que el proceso es más expedito que el observado con el caso de aquellos modelos enfocados en la localización de K medias. En este caso los modelos que más tiempo tardan en ejecutarse son justamente el K Medias, y el UMAP (el cual como se detalló en apartados previos, emplea luego de la proyección en el hiper plano de la técnica de K Medias para lograr la identificación de anomalías.

El modelo de agregación de K demuestra tiempos de ejecución no tan extensos como su contraparte de K Medias, pero un poco mayor que las técnicas más eficientes, dicha duración podría evidenciar una ventaja a la hora de realizar los cálculos una vez que sea conocido el comportamiento de las observaciones, ya que al delimitar el posible comportamiento de la cantidad de vecinos buscados, logra aprovechar las ventajas implícitas de estas técnicas.

Aunado a lo anterior, se nota un impacto importante en la duración de los procesos según el tipo de correlación existente entre las variables analizadas pero únicamente en el caso de que los datos tengan una distribución Gamma, ya que por ejemplo, si la distribución es Normal[34] prácticamente no existe variabilidad en los plazos totales de ejecución de la totalidad de los modelos, pero en el caso de las observaciones con distribución Gamma mientras mayor sea el valor de la correlación el tiempo total de ejecución tiende a aumentar, por ejemplo, se aumenta un 9% el tiempo entre las observaciones con 0% de correlación y un 80% de correlación. Adicionalmente, a nivel de los modelos empleados tanto el modelo de LOF como el de Agregación de K,

[34]En el Cuadro A1 pueden observarse los tiempos de ejecución en el caso de la distribución Normal, mientras que en el caso de la Distribución Gamma se pueden observar en el Cuadro A2

fueron los únicos modelos que requirieron menos tiempo de procesamiento cuando los datos poseían una distribución Gamma.

Finalmente, el aumento de los plazos al utilizar los datos con distribución Gamma provoca incrementos sostenidos en la duración total de ejecución de los modelos, pasando de un total de 172.2 horas cuando se emplean poblaciones con distribución Normal, a 214.3 horas al tener datos con distribución Gamma.

Cuadro A1. Tiempo promedio de ejecución en segundos para cada modelo analizado y el caso de la distribución normal

Corr.	Modelo	Tamaño de muestra						
		10mil	25mil	50mil	100mil	150mil	200mil	250mil
0%	AGR_K	1.14	3.18	6.90	15.5	24.2	31.8	40.6
	ISO	2.10	2.78	3.94	6.15	8.25	8.83	10.5
	KMEAN	4.17	11.4	23.7	43.5	64.1	87.0	108
	LOF	0.65	1.55	3.81	9.19	15.5	21.5	27.8
	PCA	0.09	0.29	1.11	2.12	2.65	3.57	3.97
	RPCA	0.20	0.59	1.43	2.83	4.18	5.16	5.83
	UMAP	5.81	15.1	30.9	56.8	84.5	117	146
20%	AGR_K	1.07	2.97	6.41	14.3	22.8	33.7	41.4
	ISO	2.05	2.56	3.51	5.34	7.85	9.94	11.2
	KMEAN	4.25	10.8	23.5	42.6	62.9	86.4	104
	LOF	0.58	1.67	4.05	9.79	14.5	19.3	24.7
	PCA	0.09	0.31	1.21	2.16	2.62	3.48	4.07
	RPCA	0.21	0.62	1.44	2.63	3.78	4.73	5.31
	UMAP	6.07	15.0	29.2	57.7	87.0	120	150
40%	AGR_K	1.11	16.9	11.6	15.8	24.5	31.8	40.5
	ISO	2.12	2.62	3.48	5.40	7.98	10.1	10.7
	KMEAN	4.01	10.3	21.5	42.9	65.6	86.3	106
	LOF	0.56	1.52	3.95	9.21	14.7	20.9	27.3
	PCA	0.09	0.33	1.12	2.24	2.53	3.59	4.20
	RPCA	0.24	0.52	1.43	2.74	4.07	5.21	5.67
	UMAP	5.71	14.5	30.1	60.2	90.1	115	143
60%	AGR_K	1.07	2.92	6.35	14.2	22.6	31.5	40.0
	ISO	2.06	2.55	3.47	5.29	7.00	9.74	11.7
	KMEAN	4.01	10.3	20.9	43.6	64.6	89.8	106
	LOF	0.51	1.48	3.49	8.59	13.6	18.91	24.4
	PCA	0.09	0.31	1.21	2.17	2.62	3.54	4.07
	RPCA	0.27	0.64	1.49	2.89	4.17	5.10	5.42
	UMAP	6.01	14.9	30.1	56.9	86.2	116	147
80%	AGR_K	1.06	2.92	6.29	14.2	22.4	33.6	43.4
	ISO	2.16	2.74	3.85	5.85	7.94	9.04	10.8
	KMEAN	4.05	10.9	23.0	42.1	62.6	87.2	110
	LOF	0.57	1.60	3.75	8.90	14.4	18.7	24.7
	PCA	0.08	0.33	1.10	2.16	2.63	3.72	4.20
	RPCA	0.25	0.56	1.52	2.68	4.05	5.16	5.70
	UMAP	5.59	13.9	28.3	57.4	91.0	117	146

*: El tiempo total de ejecución de los modelos fue de 172.2 horas.

Cuadro A2. Tiempo promedio de ejecución en segundos para cada modelo analizado y el caso de la distribución Gamma

Corr.	Modelo	Tamaño de muestra						
		10mil	25mil	50mil	100mil	150mil	200mil	250mil
0%	AGR_K	1.17	3.96	10.2	27.2	42.7	62.1	83.6
	ISO	2.12	2.74	3.80	5.70	7.77	8.75	11.3
	KMEAN	4.09	10.6	21.9	40.6	61.8	82.5	104
	LOF	0.63	2.22	7.75	21.5	37.1	55.4	72.9
	PCA	0.10	0.33	1.11	2.13	2.62	3.61	3.97
	RPCA	0.27	0.57	1.44	2.56	11.4	4.91	6.00
	UMAP	5.81	14.9	30.8	56.9	87.8	123	147
20%	AGR_K	1.20	3.75	9.48	25.9	45.8	69.3	88.6
	ISO	2.16	2.76	3.78	5.30	6.88	8.41	10.0
	KMEAN	4.16	10.2	20.7	43.8	66.3	91.2	119
	LOF	0.61	2.22	6.74	19.6	35.2	52.0	70.3
	PCA	0.09	0.35	1.21	2.17	2.60	3.65	3.98
	RPCA	0.24	0.57	1.54	2.78	4.10	5.46	5.90
	UMAP	5.97	14.7	30.5	57.2	86.3	121	150
40%	AGR_K	1.16	4.05	10.6	29.5	46.9	69.2	94.2
	ISO	2.15	2.72	3.54	4.92	6.96	9.40	11.1
	KMEAN	4.15	10.6	22.5	41.9	63.1	84.6	109
	LOF	0.63	2.28	7.74	22.7	38.5	60.6	73.7
	PCA	0.10	0.34	1.20	2.15	2.72	3.86	4.15
	RPCA	0.24	0.52	1.42	2.69	3.85	5.12	5.83
	UMAP	6.03	14.8	29.9	62.6	91.9	125	153
60%	AGR_K	1.15	3.65	9.67	27.2	52.1	71.2	96.9
	ISO	2.03	2.55	3.39	5.07	6.79	8.23	10.0
	KMEAN	4.01	10.3	21.1	43.4	65.6	87.1	114
	LOF	0.61	2.26	7.10	21.1	38.1	57.3	84.4
	PCA	0.09	0.35	1.26	2.18	2.58	3.63	4.02
	RPCA	0.28	0.58	1.59	2.83	4.08	5.27	6.04
	UMAP	5.64	14.2	28.6	58.4	91.5	116	152
80%	AGR_K	1.14	3.82	10.9	30.7	49.3	74.2	101
	ISO	2.13	2.75	3.84	5.84	7.90	8.96	11.4
	KMEAN	4.25	10.9	22.1	42.7	63.7	84.6	101
	LOF	0.59	2.34	7.04	21.3	39.2	113	93.7
	PCA	0.09	0.41	1.20	2.17	2.57	3.74	4.17
	RPCA	0.28	0.56	1.52	2.66	3.89	5.21	5.86
	UMAP	5.67	14.7	29.7	54.6	83.6	112	177

*: El tiempo total de ejecución de los modelos fue de 214.3 horas.

A3 Calificaciones promedio de los modelos

A3.1 Modelo ACP

Cuadro A3. Resultados promedios del AUC-ROC para la técnica de ACP, considerando todas las variantes de correlación y población

Factor	Dist.	Escenario*	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	Ausencia	55.9%	51.3%	50.8%	50.8%	50.5%	50.4%	50.5%
		Presencia	61.9%	51.5%	50.9%	50.7%	50.3%	50.4%	50.3%
	Gamma	Ausencia	83.5%	80.6%	78.6%	76.5%	75.3%	74.6%	74.0%
		Presencia	84.1%	80.2%	78.4%	76.3%	75.2%	74.4%	73.8%
Desv. Est.	Normal	Ausencia	3.95%	1.99%	1.03%	1.36%	0.61%	0.45%	0.61%
		Presencia	4.89%	1.58%	1.21%	0.76%	0.51%	0.65%	0.50%
	Gamma	Ausencia	5.90%	3.46%	2.73%	2.64%	2.86%	3.12%	3.41%
		Presencia	6.46%	2.58%	2.00%	2.59%	2.93%	3.30%	3.56%

* El valor de Ausencia indica que no se realizó el Centrado y el Escalado, mientras que la presencia indica que si se realizó tal como se indicó en la Sección 3.2.

A3.2 Modelo ACPR

Cuadro A4. Resultados promedios del AUC-ROC para la técnica de RPCA, considerando todas las variantes de correlación y población

Factor	Dist.	Escenario*	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	Ausencia	100%	100%	100%	100%	100%	100%	100%
		Presencia	100%	100%	100%	100%	100%	100%	100%
	Gamma	Ausencia	71.5%	70.5%	70.2%	69.8%	69.2%	68.9%	68.6%
		Presencia	71.3%	70.6%	70.2%	69.8%	69.3%	68.8%	68.6%
Desv. Est.	Normal	Ausencia	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		Presencia	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Gamma	Ausencia	23.5%	23.8%	23.6%	23.0%	22.7%	22.4%	22.0%
		Presencia	23.6%	23.7%	23.5%	22.%	22.7%	22.4%	22.0%

* El valor de Ausencia indica que no se realizó el Escalado y el Cambio de Signo, mientras que la presencia indica que si se realizó la variante tal como se indicó en la Sección 3.2.

A3.3 Modelo K Medias

Cuadro A5. Resultados promedios del AUC-ROC para la técnica de KMEANS, considerando todas las variantes de correlación y población

Factor	Dist.	Valor de K	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	3	81.7%	81.7%	76.5%	75.3%	81.0%	81.4%	83.3%
		4	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	100%
		5	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	99.9%
		6	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	100%
		10	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	100%
		15	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	100%
		20	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	100%
		25	97.5%	97.7%	97.9%	98.5%	98.9%	99.5%	100%
	Gamma	3	85.7%	79.3%	74.6%	67.5%	62.9%	60.1%	58.2%
		4	87.7%	77.3%	70.3%	62.6%	59.0%	56.8%	55.4%
		5	91.4%	81.6%	74.5%	65.9%	62.0%	59.8%	58.0%
		6	96.6%	89.9%	82.5%	73.8%	68.6%	65.2%	62.6%
		10	97.5%	93.8%	88.2%	77.5%	70.3%	65.4%	61.9%
		15	97.5%	96.3%	91.3%	80.8%	72.7%	67.3%	63.7%
		20	97.5%	96.8%	92.9%	81.2%	72.8%	67.1%	63.8%
		25	97.5%	97.0%	92.9%	81.7%	73.3%	67.1%	63.7%
Desv. Est.	Normal	3	22.2%	22.4%	23.7%	24.1%	23.6%	23.8%	23.5%
		4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		15	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		20	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		25	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Gamma	3	13.1%	11.6%	7.78%	3.18%	1.46%	0.93%	1.08%
		4	11.3%	7.39%	3.23%	3.24%	3.89%	3.53%	3.78%
		5	8.17%	5.85%	4.42%	6.60%	6.63%	6.48%	6.16%
		6	1.65%	4.97%	6.09%	8.61%	9.05%	8.44%	7.73%
		10	0.13%	6.85%	11.9%	13.2%	11.0%	8.83%	7.11%
		15	0.05%	1.92%	7.17%	9.74%	8.86%	7.03%	5.92%
		20	0.03%	1.69%	6.94%	9.94%	8.77%	7.35%	5.55%
		25	0.03%	1.12%	6.94%	9.34%	8.28%	6.63%	5.28%

A3.4 Agregación de K

Cuadro A6. Resultados promedios del AUC-ROC para la técnica de Agregación de K, considerando todas las variantes de correlación y población

Factor	Dist.	Max K*	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	40	99.9%	99.6%	99.0%	97.6%	96.3%	95.0%	93.7%
		60	99.9%	99.6%	99.1%	97.8%	96.6%	95.3%	94.0%
		80	99.9%	99.7%	99.2%	98.0%	96.7%	95.5%	94.2%
		100	99.9%	99.7%	99.2%	98.1%	96.9%	95.7%	94.4%
		120	99.9%	99.7%	99.3%	98.2%	97.0%	95.8%	94.6%
	Gamma	40	99.9%	99.8%	99.7%	99.5%	99.2%	99.1%	98.9%
		60	99.9%	99.8%	99.7%	99.5%	99.2%	99.1%	98.9%
		80	99.9%	99.8%	99.7%	99.5%	99.3%	99.1%	98.9%
		100	99.9%	99.8%	99.7%	99.5%	99.3%	99.1%	98.9%
		120	99.9%	99.8%	99.7%	99.5%	99.3%	99.1%	98.9%
Desv. Est.	Normal	40	0.02%	0.21%	0.37%	0.54%	0.68%	0.81%	0.87%
		60	0.02%	0.20%	0.37%	0.56%	0.73%	0.88%	0.94%
		80	0.01%	0.18%	0.37%	0.58%	0.77%	0.93%	0.99%
		100	0.01%	0.17%	0.37%	0.60%	0.80%	0.97%	1.05%
		120	0.01%	0.16%	0.37%	0.61%	0.83%	1.00%	1.09%
	Gamma	40	0.01%	0.04%	0.09%	0.26%	0.42%	0.56%	0.71%
		60	0.01%	0.04%	0.09%	0.26%	0.42%	0.55%	0.69%
		80	0.01%	0.04%	0.09%	0.26%	0.41%	0.54%	0.68%
		100	0.01%	0.04%	0.09%	0.26%	0.41%	0.54%	0.67%
		120	0.01%	0.04%	0.09%	0.26%	0.41%	0.54%	0.67%

* El valor máximo de K indicado considera el promedio de todos los valores mínimos de K indicados en la Sección 3.2.

A3.5 LOF

Cuadro A7. Resultados promedios del AUC-ROC para la técnica de LOF, considerando todas las variantes de correlación y población

Factor	Dist.	Valor de K	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	20	64.3%	56.2%	56.5%	56.5%	55.9%	55.8%	55.7%
		40	68.9%	57.6%	55.1%	55.5%	55.2%	55.0%	54.8%
		60	73.7%	61.4%	56.0%	55.2%	54.9%	54.7%	54.2%
		80	76.0%	61.7%	57.2%	55.0%	54.7%	54.5%	53.8%
		100	81.9%	62.8%	61.2%	54.9%	54.6%	54.5%	53.5%
		120	84.0%	63.4%	61.3%	55.1%	54.6%	54.4%	53.5%
		140	85.6%	66.9%	61.8%	55.7%	54.6%	54.3%	53.4%
		160	90.6%	67.4%	61.8%	56.2%	54.7%	54.3%	53.4%
	Gamma	20	95.7%	90.9%	88.1%	84.7%	82.1%	80.3%	78.8%
		40	96.5%	92.0%	89.3%	86.1%	83.7%	81.9%	80.6%
		60	96.9%	92.6%	89.9%	86.7%	84.4%	82.7%	81.3%
		80	97.2%	93.0%	90.2%	87.1%	84.8%	83.1%	81.3%
		100	97.5%	93.3%	90.6%	87.4%	85.1%	83.4%	82.0%
		120	97.7%	93.6%	90.8%	87.6%	85.3%	83.6%	82.2%
		140	97.9%	93.8%	91.0%	87.8%	85.5%	83.8%	82.4%
		160	98.1%	94.0%	91.2%	87.9%	85.7%	83.9%	82.6%
Desv. Est.	Normal	20	16.2%	3.77%	3.65%	3.04%	2.80%	2.49%	1.62%
		40	21.3%	7.20%	3.82%	2.76%	2.67%	2.62%	1.92%
		60	23.5%	16.2%	4.04%	2.68%	2.31%	2.47%	2.11%
		80	22.0%	16.5%	6.50%	3.24%	2.33%	2.28%	2.13%
		100	21.5%	16.8%	16.2%	3.76%	2.56%	2.27%	2.25%
		120	19.2%	17.4%	16.3%	4.52%	2.91%	2.47%	1.98%
		140	17.5%	21.5%	16.5%	5.63%	3.34%	2.66%	2.15%
		160	15.4%	21.4%	16.5%	6.72%	3.81%	3.00%	2.26%
	Gamma	20	7.11%	10.6%	11.8%	11.9%	11.8%	11.5%	11.2%
		40	6.17%	10.2%	11.7%	12.4%	12.5%	12.5%	12.3%
		60	5.59%	9.89%	11.6%	12.5%	12.8%	12.9%	12.7%
		80	5.10%	9.59%	11.4%	12.6%	12.9%	13.1%	13.0%
		100	4.68%	9.32%	11.3%	12.5%	13.0%	13.2%	13.2%
		120	4.33%	9.08%	11.1%	12.5%	13.0%	13.3%	13.3%
		140	4.03%	8.87%	11.0%	12.4%	13.0%	13.3%	13.3%
		160	3.78%	8.70%	10.8%	12.4%	13.0%	13.3%	13.4%

A3.6 ISO

Cuadro A8. Resultados promedios del AUC-ROC para la técnica de árboles de aislamiento, considerando todas las variantes de correlación y población

Factor	Dist.	Escenario	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	Alt	100%	100%	100%	100%	99.9%	99.9%	99.9%
		Ext	100%	99.9%	99.9%	99.9%	99.7%	99.6%	99.4%
		Sci	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%
		Sim	99.9%	99.9%	99.9%	99.8%	99.7%	99.5%	99.3%
	Gamma	Alt	99.0%	98.6%	98.3%	97.8%	97.5%	97.2%	97.0%
		Ext	99.9%	99.7%	99.4%	98.8%	98.2%	97.7%	97.1%
		Sci	97.7%	72.9%	62.6%	56.3%	54.3%	53.3%	52.6%
		Sim	99.9%	99.7%	99.4%	98.8%	100%	97.7%	97.2%
Desv. Est.	Normal	Alt	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		Ext	0.00%	0.00%	0.03%	0.09%	0.19%	0.22%	0.29%
		Sci	0.00%	0.01%	0.02%	0.02%	0.03%	0.03%	0.03%
		Sim	0.00%	0.01%	0.04%	0.12%	0.19%	0.26%	0.33%
	Gamma	Alt	1.81%	1.80%	1.92%	2.34%	2.60%	2.76%	2.90%
		Ext	0.03%	0.10%	0.21%	0.55%	0.84%	1.11%	1.40%
		Sci	2.24%	6.25%	2.97%	1.06%	0.94%	0.87%	0.80%
		Sim	0.03%	0.11%	0.25%	0.55%	0.83%	1.10%	1.34%

* Los escenarios indicados se detallaron en la Sección 3.2, y hace referencia a la versión con critetrio de ganancia (Alt), con versión extendida (Ext), con Selección dividida (SCi) y finalmente la variante simple (Sim).

A3.7 UMAP

Cuadro A9. Resultados promedios del AUC-ROC para la técnica de UMAP, considerando todas las variantes de correlación y población

Factor	Dist.	Valor de K	Porcentaje de anomalías						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
Promedio	Normal	3	89.4%	89.5%	91.0%	91.5%	90.4%	91.6%	86.5%
		4	89.4%	89.5%	91.1%	91.0%	92.0%	92.1%	91.9%
		5	89.4%	89.6%	91.1%	91.4%	91.9%	92.4%	92.9%
		6	89.4%	89.5%	91.1%	91.5%	91.7%	92.4%	92.8%
		10	89.4%	89.5%	91.2%	91.5%	91.9%	92.4%	92.9%
		15	89.4%	89.6%	91.1%	91.5%	91.9%	92.4%	92.9%
		20	89.4%	89.6%	91.1%	91.5%	92.0%	92.4%	92.8%
		25	89.4%	89.6%	91.2%	91.5%	91.9%	92.4%	92.8%
	Gamma	3	81.1%	76.4%	79.5%	83.1%	78.8%	77.2%	68.0%
		4	82.4%	83.1%	86.7%	79.6%	81.9%	81.3%	73.9%
		5	74.8%	87.8%	83.8%	83.7%	82.8%	78.2%	78.0%
		6	79.9%	87.9%	78.9%	81.8%	79.3%	79.1%	76.1%
		10	78.9%	89.0%	83.5%	82.4%	81.7%	81.6%	78.0%
		15	83.6%	88.0%	89.6%	86.1%	83.4%	77.0%	77.9%
		20	84.8%	89.6%	86.8%	85.7%	85.0%	79.1%	77.8%
		25	84.1%	90.1%	85.9%	86.9%	83.3%	79.4%	77.0%
Desv. Est.	Normal	3	18.1%	18.3%	17.0%	17.3%	17.8%	17.4%	20.7%
		4	18.1%	18.1%	17.0%	17.2%	17.2%	17.6%	17.6%
		5	18.2%	18.1%	16.9%	17.3%	17.4%	17.6%	17.6%
		6	18.1%	18.1%	16.9%	17.2%	17.4%	17.6%	17.7%
		10	18.1%	18.2%	16.8%	17.2%	17.3%	17.5%	17.6%
		15	18.2%	18.0%	16.9%	17.1%	17.3%	17.6%	17.6%
		20	18.0%	18.1%	16.9%	17.2%	17.3%	17.4%	17.6%
		25	18.0%	18.1%	16.8%	17.1%	17.4%	17.5%	17.7%
	Gamma	3	22.3%	21.5%	20.5%	20.1%	21.4%	21.5%	19.5%
		4	21.1%	19.7%	17.3%	21.9%	17.9%	19.6%	19.7%
		5	23.4%	16.7%	18.2%	18.4%	18.5%	20.8%	17.8%
		6	22.5%	16.0%	20.7%	19.0%	21.0%	20.1%	19.2%
		10	22.3%	12.1%	17.2%	18.8%	17.7%	18.7%	18.1%
		15	20.4%	16.3%	14.2%	16.4%	18.5%	18.7%	18.3%
		20	19.4%	14.3%	16.9%	15.8%	17.0%	18.2%	18.1%
		25	20.3%	14.2%	17.7%	15.3%	18.0%	18.1%	18.0%

A4 Indicadores utilizados para el análisis de los datos

Para la presente investigación se analizó un conjunto de 9 indicadores, mediante los cuales se aproximó el detalle de explorado en la Sección 2.10.1.

- **Conteo de objeciones:** Cantidad total de objeciones presentadas por cada línea analizada.
- **Diferencia entre la estimación y la adjudicación:** Diferencia entre el precio estimado y el adjudicado por parte de la institución o figura legal encargada.
- **Poder de Mercado:** Porcentaje de concentración de mercado según los 8 dígitos del bien contratado.
- **Porcentaje de victoria:** Determina la probabilidad que posee cada proveedor al participar en cada proceso de contratación, esto en función de la cantidad de veces que ha participado en procesos similares y quedado adjudicado.
- **Similitud de direcciones:** Porcentaje de igualdad entre las direcciones físicas registradas en SICOP, según la distancia de Jaro-Winkler.
- **Índice de Alcance:** Composición de indicadores relacionado con la variación entre las cantidades inicialmente solicitadas y adjudicadas, así como las cantidades adjudicadas y contratadas.
- **Ámbito de recepción de ofertas:** Duración para la recepción de las ofertas en el proceso de contratación.
- **Diferencia entre monto adjudicado y contratado:** Variación de los montos en el proceso, durante el proceso de contratación.
- **Diferencias entre monto adjudicado y ofertas recibidas:** Diferencia entre el monto adjudicado y las ofertas recibidas.