EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS R_{NX}

DIEGO FERLEY URREA BURGOS

UNIVERSIDAD CESMAG

FACULTAD DE INGENIERÍA

PROGRAMA DE INGENIERÍA DE SISTEMAS

SAN JUAN DE PASTO

2024

EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS R_{NX}

DIEGO FERLEY URREA BURGOS

Proyecto de grado para optar al título de ingeniero de sistemas

Asesor

Carlos Fernando Gonzáles Guzmán

Ingeniero de Sistemas

UNIVERSIDAD CESMAG

FACULTAD DE INGENIERÍA

PROGRAMA DE INGENIERÍA DE SISTEMAS

SAN JUAN DE PASTO

2024

Nota de Aceptación

_____

_____

_____

_____

Dalila Pachajoa

_____

_____

_____

_____

Héctor Mora

San Juan de Pasto, 25 de noviembre del 2024

Dedicatoria

Dedico este proyecto a mi madre Franca Burgos y a mi hermana Gabriela Urrea, quienes siempre me han acompañado y apoyado en las diversas etapas de mis estudios y vida personal.

Diego Ferley Urrea Burgos

Agradecimientos

Expresamos los más sinceros agradecimiento a todo el grupo de trabajo que colaboró en este proyecto, siendo estos Carlos David Correa Lozano, Juan Andrés lozano Thomé y Diego Ferley Urrea Burgos, quienes se han comprometido con el desarrollo y finalización del trabajo de grado y al profesor Juan Carlos Alvarado Pérez, quién en su rol de asesor, ha acompañado a los estudiantes en todo su proceso investigativo y que con su amplio conocimiento en el área investigada en el presente proyecto, no ha dudado en compartir dicho conocimiento con cada uno de los estudiantes, nuevamente les estamos inmensamente agradecidos.

Página de exclusión de responsabilidad intelectual

“El pensamiento que se expresa en esta obra es exclusiva responsabilidad de sus autores y no compromete la ideología de la Institución Universitaria CESMAG”

RESUMEN ANALÍTICO DE ESTUDIO R.A.E

Facultad	Ingeniería
Programa	Ingeniería de Sistemas
Fecha de elaboración	28 de noviembre de 2021
Autores de la investigación	Diego Ferley Urrea Burgos
Director de la investigación	Carlos Fernando Gonzáles Guzmán Ingeniero en Sistemas
Título de la investigación	EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS RNX

PALABRAS CLAVE

Inteligencia Artificial, Reducción, Dimensión, Machine Learning, Topología, Métodos, Métricas.

DESCRIPCIÓN

En la actualidad, la masiva cantidad de datos generada de diversas fuentes convierte la tarea del análisis de datos en algo complejo y tardío, puesto que los datos pueden contener información que no es relevante para un algoritmo de Machine Learning, generando resultados poco confiables, por lo cual se vuelve necesario procesar los datos antes de aplicarlos al algoritmo, dentro de este proceso de transformación se encuentra la reducción de dimensión (RD), la cual permite obtener un espacio de menor dimensión a partir de un conjunto de datos de mayor dimensión, con el fin de que atributos irrelevantes, no muy relevantes o redundantes sean eliminados. Los métodos RD son de gran ayuda puesto que transforman los datos ingresados en una representación de los mismos mucho más manejable, una noción visual del desempeño de las técnicas RD, es la preservación topológica o la medida en que el espacio de incrustamiento conserva la estructura de los datos de mayor dimensión, por tal motivo se hace necesario medir cuantitativamente dicha preservación de la topología. Por lo anterior, el presente proyecto implementó las curvas de calidad RNX propuestas por John Lee y Michael Verleysen, que permiten evaluar el desempeño de los algoritmos RD, generando una representación gráfica de la preservación topológica.

CONTENIDO

La investigación está conformada de la siguiente manera:

Capítulo 1: En este capítulo se describe la problemática de las grandes cantidades de datos generados actualmente, además de las pocas herramientas de análisis de datos que integren métricas de evaluación como RNX para medir el desempeño de los métodos RD, se realiza la definición del objetivo general y los cuatro objetivos específicos, se justifica la necesidad de

realizar proyectos de este naturaleza ya que como se menciona en el artículo “Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva” realizado por Ana Cristina Umaquinga, Diego Peluffo y Paul Rosero , el crecimiento de los datos generados actualmente alcanza el orden de los petabytes y exabytes, por lo cual se deben realizar avances en esta área. Finalmente se delimita el proyecto en un periodo de 24 meses a partir del periodo A del año 2020 hasta periodo B del año 2021.

Capítulo 2: En este capítulo se presentan antecedentes internacionales, nacionales y regionales tomando estudios que abordaran temáticas de análisis de métodos RD como en el artículo “Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades” realizado por Juliana Valencia Aguirre, Genaro Daza Santacoloma, Carlos D Acosta y Germán Castellanos Domínguez, los cuales realizan un estudio comparativo de diversos tipos de métodos RD como LLE, Isomap, etc, además de la descripción de los supuestos teóricos y la definición de las variables de investigación.

Capítulo 3: En este capítulo se define el tipo de metodología de investigación como tipo positivista ya que tiene un enfoque metodológico predominante cuantitativo, en donde gracias al intervalo de calidad de [0, 100], es posible determinar la precisión con que los métodos RD preservan la topología de los datos en un espacio de baja dimensión, además se definen los conjuntos de datos tanto artificiales como el rollo suizo, las esfera y el toroide y datos reales como MNIST, que se utilizaron para la fase de experimentación de la herramienta con el fin de validar el funcionamiento de los métodos RD.

Capítulo 4: En este capítulo se presentan los resultados de la investigación realizada en donde se exhiben los diversos métodos de RD que se integraron en la herramienta, además del flujo de ejecución de las medidas de calidad RNX y el levantamiento de requisitos funcionales como no funcionales de la herramienta, teniendo en cuenta los pasos del proceso KDD (Knowledge Discovery in Databases) que se consideraron para el desarrollo de los diversos módulos de la herramienta.

Capítulo 5: En este capítulo se encuentra el análisis y discusión de los resultados en donde se evidencia diversos experimentos realizados con los métodos RD, analizando los resultados obtenidos y la evaluación generada por la métrica RNX con las diferentes bases de datos propuestas en el capítulo anterior.

METODOLOGÍA

La metodología de investigación empleada en este proyecto fue de tipo positivista porque tiene un enfoque metodológico predominante cuantitativo, en donde gracias al intervalo de calidad [0, 100] es posible determinar con que precisión, es preservada la topología de los datos en baja dimensión, para poder identificar que métodos RD deben ser usados en situaciones muy específicas.

LÍNEA DE INVESTIGACIÓN Inteligencia Artificial

CONCLUSIONES

La herramienta desarrollada cumple de manera satisfactoria los objetivos planteados ya que la interfaz Drag and Drop desarrollada, permite una interacción eficiente con los diversos módulos implementados, tales como carga de datos, métodos RD, particionador, visualización Scatter Plot, visualización Line Chart, visualización en tabla y métrica de evaluación RNX, logrando así disponer de diversas herramientas para la creación de flujos para la evaluación de la preservación topológica.

Finalmente, el desarrollo de la herramienta permitió obtener conocimientos relacionados a la implementación de interfaces gráficas dinámicas con Python y el análisis de scripts desarrollados en Matlab, además del uso de librerías de análisis de datos disponibles en Python.

RECOMENDACIONES

La herramienta desarrollada es de fácil escalabilidad, para la implementación de nuevos módulos y funcionalidades que implementen otros pasos del proceso KDD, además de necesitar nuevas implementaciones de visualización que permitan generar gráficos con datos de más de tres dimensiones.

Por otro lado, dada la naturaleza de la investigación, en donde se hondo en la fundamentación heurística y matemática tanto de los métodos RD como de las métricas RNX, se cree necesario fomentar entre los estudiantes la importancia de entender las bases de cualquier tópico de investigación.

FUENTES

UMAQUINGA, Cristina; PELUFFO, Diego y ROSERO, Paul. Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva, 2016. Disponible en https://www.researchgate.net/publication/316921329_Propuesta_de_analisis_visual_de_datos_en_Big_Data_usando_reduccion_de_dimension_interactiva

AGUIRRE, Juliana; SANTACOLOMA, Genaro; ACOSTA, Carlos; DOMÍNGUEZ, Germán. Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades, 2010 Disponible en https://dialnet.unirioja.es/servlet/articulo?codigo=5062986

CONTENIDO

LISTA DE TABLAS

pág.

LISTADO DE FIGURAS

Pág.

LISTA DE ANEXOS

pág.

INTRODUCCIÓN

La revolución digital ha hecho posible que la información sea fácil de capturar, procesar, almacenar, distribuir y transmitir, debido al importante progreso tecnológico y computacional en diferentes aspectos cotidianos. De esta manera se han generado grandes volúmenes de información en todas las áreas del conocimiento, para afrontar los retos de esta era digital es necesario contar con herramientas eficientes para el tratamiento de la información. Dicho lo anterior surge la necesidad de procesar gran cantidad de datos con el fin de obtener conocimiento y una forma de procesar eficientemente los datos es aplicando técnicas de MD (Minería de Datos).

Según Riquelme, Ruíz y Gilbert [1] La minería de datos permite extraer información, descubrir conocimiento y determinar patrones ocultos dentro de los datos, está implícita dentro del proceso KDD (Knowledge Discovery in Data Base) o Descubrimiento de Conocimiento en Bases de Datos, sin embargo, es solo una fase esencial del proceso KDD, pues Riquelme afirma que en este proceso también se lleva a cabo la preparación y selección de los datos hasta los resultados que ofrece la MD.

Dentro de la comunidad científica suele decirse que entre mayor sea la cantidad de datos, más eficiente es la MD u otras área como ML (Machine learning) en donde la cantidad de los datos puede permitir el entrenamiento de algoritmos más precisos, sin embargo, puede que el dataset con el que se esté trabajando contenga una cantidad de datos que realmente no son necesarios para el algoritmo que se esté desarrollando si de ML se trata, o en MD sería la obtención de conocimiento a partir de los datos. Según Han, Kamber y Pei, en su libro “Data Mining Concepts and Techniques”, Han; Kamber y Pei.[2] Existe un proceso llamado “Data Preprocessing” en donde se explica el proceso ETL (Extract, Transform and Load) el cual permite a las organizaciones trasladar datos desde múltiples fuentes, reformatearlos, limpiarlos, y cargarlos en otra base de datos, dentro del proceso de transformación de datos se encuentra la reducción de dimensión (RD), la cual permite obtener un espacio de menor dimensión (entiéndase dimensión como variable o atributo) a partir de un conjunto de datos de mayor dimensión, con el fin de que atributos irrelevantes, no muy relevantes o redundantes sean eliminados y removidos.

Como se ha mencionado anteriormente los métodos de RD son un gran aporte para distintas áreas como MD y ML así mismo como para el reconocimiento de

[1] RIQUELME, José; RUÍZ, Roberto y GILBERT, Karina. Minería de datos conceptos y tendencias. Valencia. 2006.
[2] HAN, Jiawey; KAMBER, Michelin y PEI, Jian. Minería de datos: Concepts and Techniques. 2011

patrones y la visualización de información, según Salazar, Peña y otros[3]. La RD es de gran ayuda en el campo del reconocimiento de patrones el cual es considerado como la extracción de características que transforma los datos ingresados en una representación de los mismos mucho más manejable, por otro lado, en la visualización de datos RD es un proceso crucial que le permite al analista tener una visualización de los datos más entendible, de tal forma que éste identifique patrones de una forma más fácil, esta estrategia de minería de datos basada en la visualización de los datos es llamada "visual data mining and Info Vis".

De lo anterior se puede decir que la integración entre la RD y la Info Vis permite realizar un análisis eficiente de los datos en donde es posible inspeccionar de manera gráfica el resultado de aplicar las técnicas RD. Una noción visual del desempeño de la técnica RD aplicada es la preservación topológica o la medida en que el espacio de incrustamiento conserva la estructura de los datos de mayor dimensión. Por lo tanto, se hace necesario medir cuantitativamente dicha preservación topológica, para este propósito y como tema central de este proyecto se implementó las curvas de calidad R_{NX} propuestas por John Lee y Michel Verleysen que según los autores, Lee y Verleysen[4]. Evalúan el desempeño de los algoritmos RD, generando una representación gráfica de la preservación de la topología local y global de un colector mediante un indicador cuantitativo asociado al área bajo las curvas R_{NX}. En el presente proyecto se realizó la evaluación de métodos RD mediante las curvas R_{NX} implementadas por medio del desarrollo de software libre en Python.

[3] SALAZAR, José; PEÑA, Diego, *et al.* Dimensionality reduction for interactive data visualization via Geo-Desic approach. 2016.
[4] LEE, John y VERLEYSEN, Michel. Quality assessment of dimensionality reduction: Rank-based criteria. 2009

1. PROBLEMA DE INVESTIGACIÓN

1.1 Tema objeto de estudio

Curvas R_{NX} para la evaluación de la preservación topológica de métodos Reducción de Dimensión.

1.2 Línea de investigación

Inteligencia Artificial: Que según el documento formato líneas de investigación A-2020, "Es la simulación de inteligencia humana por parte de las máquinas. Dicho de otro modo, es la disciplina que trata de crear sistemas capaces de aprender y razonar como un ser humano, aprendan de la experiencia, averigüen cómo resolver problemas ante unas condiciones dadas, contrasten información y lleven a cabo tareas lógicas".[5]

1.3 Planteamiento del problema

La evolución de la era informática ha ocasionado un incremento significativo en el volumen de los datos, esto hace que los analistas se enfrenten a un gran reto, pues es a partir de los datos que se obtiene información necesaria para cualquier empresa o entidad. Según Pang y otros. [6] Áreas como reconocimiento de patrones, Reducción de la dimensión (RD), minería de datos, machine learning y la visualización de datos, permiten que este reto pueda ser afrontado de forma eficiente.

Aunque según Peña, Salazar, Peluffo y otros afirman que.[7] Actualmente existen modernas herramientas de visualización de datos y métodos de RD convencionales. Sin embargo, siguen siendo pocas las herramientas que integren métodos RD, y las que lo hacen, no tienen implementadas métricas que permitan evaluar si estos métodos se están desempeñando adecuadamente, esto debido a que no existe una implementación modular de las curvas R_{NX}, que son las métricas más comúnmente utilizadas en la literatura científica para medir el desempeño de métodos RD mediante el uso de entornos dinámicos como los que brinda la POO (Programación Orientada a Objetos).

La falta de herramientas e implementaciones modulares, es causada porque para aplicar métodos RD es necesario ajustar ciertos parámetros, lo cual es

---

[5] PROGRAMA DE INGENIERÍA DE SISTEMAS. Líneas de Investigación, ingeniería de sistemas. San Juan de Pasto: Universidad Cesmag, 2020.

[6] PANG, Bo; *et al.* Thump up? Sentiment classification using Machine learning techniques. Philadelphia. 2002.

[7] PEÑA, Diego, SALAZAR, Diego, PELUFFO, Diego, *et al.* Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction. 2016.

hecho por un experto en el área con el fin de que la reducción de la dimensión de los datos sea lo más efectiva posible, así mismo, el experto tiene el conocimiento necesario para entender la literatura científica que menciona a nivel epistemológico y matemático, la forma de evaluar la preservación topológica después de aplicar un método RD, sin embargo, para que el experto pueda hacer la evaluación necesita realizar ciertos pasos, y al no contar con herramientas que le ayuden a hacer el proceso más ágil, conlleva a que la investigación tome más tiempo, por otro lado, aquellas personas que no sean expertas y quieran hacer uso de los métodos RD, al no tener herramientas que le permitan al usuario evaluar el rendimiento de los algoritmos, resultaría en no obtener resultados confiables, afectando significativamente el estudio que se esté llevando a cabo.

1.4 Formulación del problema

¿De qué manera se puede evaluar la preservación topológica de métodos RD, conservando la estructura de los datos de alta dimensión respecto al incrustamiento generado?

1.5 Objetivos

1.5.1 Objetivo General

Evaluar métodos de reducción de dimensión mediante la implementación de métricas R_{NX}, para la preservación topológica de los datos.

1.5.2 Objetivos Específicos.

Analizar diversos enfoques y heurísticas de reducción de dimensión desarrolladas en la literatura científica.

Identificar algunos de los métodos de reducción de dimensión analizados, procedentes de distintos enfoques matemáticos para su integración.

Desarrollar una librería funcional que aborde las diversas rutinas de programación para la evaluación de la preservación topológica de los métodos integrados.

Establecer los resultados del rendimiento de los métodos evaluados por medio de la librería, en una interfaz gráfica intuitiva para el usuario.

1.6 Justificación

Según Umaquinga, Perlufo y otros “En la actualidad se puede evidenciar un crecimiento exponencial del volumen de datos, dando lugar al área emergente denominada Big Data. Paralelamente a este crecimiento, ha aumentado la demanda de herramientas, técnicas y dispositivos para almacenar, transmitir y procesar datos de alta dimensión.”[8]. Por lo que, para la comunidad académica, se hace necesario implementar un proyecto más de esta naturaleza, en donde se tenga en cuenta aspectos relativos a la evaluación y representación de los datos; ya que, en la actualidad, los procesos sistemáticos e informáticos, permiten una exploración minuciosa de los datos a través de herramientas que posibilitan obtener de forma rápida y oportuna, la información suficiente para optimizar los procesos de evaluación.

El propósito de este proyecto fue evaluar el rendimiento y calidad de los resultados efectuados por algoritmos de RD, usando métricas como las curvas de calidad R_{NX} que permiten analizar la preservación topológica de los datos, esto debido, a que cada uno de ellos posee un comportamiento diferente y la importancia de evaluarlos, fue detectar cual es el grado de confiabilidad determinando qué tan eficiente fue el método con respecto a la preservación topológica.

Las curvas de calidad R_{NX} son integradas generalmente mediante fragmentos de código o scripts que les permiten a los expertos hacer sus estudios, sin embargo, no estaban implementadas en lenguajes de programación orientados a la analítica de datos, de tal forma de que pudieran ser utilizadas de una forma ágil mediante bibliotecas o librerías, la implementación de estas métricas en una librería funcional, permitió que éstas puedan ser integradas en sistemas KDD de forma ágil, optimizando el tiempo en el proceso de investigación, por otro lado, se tomó ventaja de paradigmas de programación como la POO (Programación Orientada a Objetos), la cual permitió que la escalabilidad y adaptabilidad de dicha librería fuera posible.

Por lo anterior se utilizó el lenguaje de programación Python para desarrollar una librería que permite la evaluación de métodos de reducción de dimensión mediante curvas de calidad R_{NX}. Además de esto, fue sumamente importante el desarrollo de una interfaz de visualización interactiva e intuitiva para el usuario, para esto, se indagó en las diferentes técnicas de visualización de datos ya que según los mismos autores Umaquinga y otros “Existen decenas de herramientas de software que usan un sinnúmero de técnicas de visualización, entre ellas las técnicas de dimensión lineal o no lineal, que se basan en métodos matemáticos, geométricos, estadísticos, y topológicos”[9]. Por

[8] UMAQUINGA, A.; PERLUFO, Diego, *et al*. Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva. 2016.
[9] ibid

lo tanto, es correcto afirmar que este proyecto sirvió como una herramienta para conocer los procesos que se llevan a cabo relacionados con la RD de los datos y su comportamiento, contribuyendo al mejoramiento, crecimiento y el desarrollo del análisis de las grandes cantidades de datos promoviendo las buenas prácticas en cuanto a la RD y la visualización de los datos.

1.7 Delimitación

La investigación se desarrolló con los métodos de reducción de dimensión clásicos y espectrales basados en kernel como: laplacian eigenmaps, locally linear embedding, classical multidimensional scaling y principal component analysis. Los cuáles fueron evaluados mediante la implementación de las métricas R_{NX} en una librería, para la preservación topológica de los datos, la investigación presenta grandes beneficios para expertos en el uso de estos métodos, reduciendo el tiempo de ejecución de la evaluación y a su vez, ayudando a aquellos que no son expertos, en la correcta evaluación de métodos RD. El proyecto se desarrolló en un tiempo estimado de 24 meses, comenzando en el periodo A del año 2020 y finalizando en el periodo B de 2021.

2. MARCO TEÓRICO

2.1 Antecedentes

Antecedentes internacionales.

Según Bodt, Mulders y otros. En su artículo titulado “Nonlinear dimensionality reduction with missing data using parametric multiple inputations”, publicado en la revista IEEE en el año 2018[10]. Métodos como la incrustación de vecinos estocásticos (SNE) y sus variantes como t-SNE el cual define probabilidades a partir de la distribución t-Student, estos métodos que definen probabilidades basadas en el incrustamiento de vecinos, el problema radica en que éstos no pueden ser usados en data sets incompletos, los cuales se están volviendo cada vez más frecuentes en el aprendizaje de maquina (ML), descartar los datos faltantes es inaceptable puesto que se puede perder información, así mismo el tamaño del data set se vería reducido conforme la cantidad de datos faltantes aumenta, a pesar de los avances en RD no se han generado métodos que puedan abordar satisfactoriamente dichos data sets.

Se propuso una metodología llamada imputación múltiple usando una mezcla multivariada Gaussiana (MIMG), esta metodología permite procesar los datos faltantes bajo el paradigma de imputación múltiple, esto es posible ya que las características de los data sets no son independientes, por lo cual, se puede obtener la distribución condicional de las características faltantes teniendo en cuenta las características observables, las cuales deben ser tomadas en cuenta para obtener información importante sobre la relación de los datos, es decir, los datos faltantes pueden ser reemplazados por la media condicional dada por los valores observados y la mezcla Gaussiana es particularmente atractiva para este propósito, la cual es ajustada primero dentro del data set incompleto y luego versiones imputadas del mismo son creadas muestreando la distribución condicional de los datos faltantes a partir de los datos dados.

La metodología se aplicó al método RD de escala múltiple SNE, sin embargo, ésta puede ser aplicada a cualquier método RD en especial a los que se basan en el instrustramiento de vecinos, para el experimento se compararon diferentes métodos como IMG (condicional mean Imputation using a mixture of Multivariant Gaussian), MIG (Multiple Inputation using a single multivariant Gaussian), mediante la implementación de las curvas $R_{NX,}$ para observar y evaluar su rendimiento, además, se usaron diferentes proporciones P_{mis} de datos faltantes que van desde el 1% hasta el 30%, como resultado, MIMG es el mejor método en promedio para todas las proporciones $P_{mis,}$ exceptuando una cantidad muy limitada, como por ejemplo para los P_{mis} altos MIG tuvo un mejor rendimiento, sin embargo, esto se debía al aumento de componentes

[10] BODT, Cyril; MULDERS, Doumia, et al. Nonlinear dimensionality reduction with missing data using parametric multiple inputations. 2018.

Gaussianos, lo cual puede ocasionar un sobreajuste de los datos observados y por tanto MIMG sigue siendo el mejor método estadísticamente hablando ya que además puede ser usada en cualquier método.

El aporte de este trabajo para la investigación, radicó en que, puede presentarse la situación donde las bases de datos incompletas se hagan presentes, por lo tanto, este trabajo provee una metodología capaz de lidiar con dicha base de datos, así mismo, las curvas R_{NX} tuvieron una participación fundamental al momento de decidir cuál método tuvo un mejor rendimiento.

Anteriormente se mencionó un método llamado escala múltiple, el cual fue propuesto por Lee, Peluffo y Verleysen, en su artículo "Multi-scale similarities in stochastic neigbour embedding: reducing dimensionality while preserving both local and global structure" realizado en el año 2015[11]. La intuición general de la RD es que los datos similares, deben ser representados cercanamente unos de otros, por otro lado, los datos no similares o diferentes deben ser representados lejanamente unos de otros, el término de similaridad es bastante reciente en la RD teniendo métodos genuinos de preservación de similaridad como SNE, t-SNE, NeRV, JSE, los cuales tienen un mejor rendimiento con respecto a los métodos más viejos, sin embargo, este rendimiento es bueno solo cuando se trabaja con vecindarios pequeños alrededor de cada dato, esto debido a que dichos métodos definen vecindarios Gaussianos fáciles de trabajar, por lo tanto involucran anchos de banda (cantidad de información) que no fueron definidos por el usuarios, por lo que puede que se filtre información en escalas mayores o menores de la especificada, mientras que la RD debería entregar óptimos resultados en todas las escalas

Basándose en los anteriores métodos se propuso una definición refinada de similaridad que permita una aplicación a múltiples escalas para la RD, por lo que se introdujeron unas similaridades generalizadas, las cuales son promedios de la probabilidad de los vecindarios Gaussianos fáciles, con crecimiento de ancho de banda cubriendo así, todos los tamaños de vecindarios, desde pequeños hasta grandes, asumiendo también el costo del aumento en la complejidad computacional. Esta similaridad a múltiples escalas permitió la propuesta de nuevos métodos RD como Ms. SNE, Ms. NeRV, Ms. JSE donde Ms. Es Múltiples escalas.

Para el experimento se utilizaron bases de dato artificiales como la esfera, la Bola y la Elipse. Como resultado se obtuvo que los métodos basados en escalas únicas, capturaron de una forma más pobre la información que los de múltiples escalas, los cuales conservaron en mayor proporción la topología

[11] LEE, Jonh; PELUFFO, Diego y VERLEYSEN, Michel. Multi-scale similarities in stochastic neigbout embedding: reducing dimensionality while preserving both local and global structure. 2015.

local y global de los datos, es importante mencionar que Ms. JSE fue el método que mejores resultados arrojo de todos.

Este trabajo fue importante para la investigación, puesto que se mencionan nuevos métodos de RD, cuya revisión es uno de los objetivos de esta investigación, además de aspectos importantes sobre la preservación topológica de los datos, es decir la RD debería conservar ambas topologías y la mejor forma de evaluar dicha conservación es mediante las curvas R_{NX}, las cuales permiten cuantificar la evaluación de los métodos, permitiendo determinar que método tuvo un mejor rendimiento respecto a los demás.

Por otro lado según France y Akkucuk, en su artículo de investigación titulado "A review, Framework, and toolkit for Exploring, Evaluating, and Comparing Visualization Methods" publicado en 2020[12] El cerebro humano solo puede comprender 3 dimensiones, es por eso, que la reducción de la dimensión va de la mano con la infoVis (Visualización de la información), ya que la RD permite obtener una representación en baja dimensión de los datos de alta dimensión, sin embargo, los resultados de la RD no siempre son perfectos y puede que un método tenga un mejor rendimiento que otro, lo cual afecta la visualización y por consiguiente la toma de decisiones a partir de los datos

Se realizó una investigación con el fin de conocer diferentes métodos de RD, así como técnicas de visualización de datos, con el fin de crear un framework llamado QVisVis que incluya diferentes técnicas para evaluar la calidad de las visualizaciones usando RD, esto es posible ya que hay diferentes formas de evaluar los métodos de forma visual al graficar el incrustamiento en baja dimensión, y sombrear por el nivel de la recuperación de los vecindarios de cada uno de los ítems graficados, ya que en la alta dimensión un dato puede encontrarse en un vecindario con una cierta cantidad de vecinos más cercanos, y la RD busca conservar dichos vecinos en la baja dimensión, por lo que este tipo de visualización puede ayudar a identificar la región donde la recuperación de vecindarios es pobre.

Se implementó el método SNE, el cual se evaluó con PCA, MDS, LLE, LE, isomap, KPCA (Versión kernel de PCA para datos no lineales), de donde el framework QVisVis mediante diferentes técnicas de visualización pudo identificar qué métodos conservaban mejor la topología global o local con respecto a los demás, por ejemplo PCA al tratarse de un método global conserva mejor la topología global que LLE el cual conserva mejor la topología local, al ser de forma visual, le permite al usuario tener un mejor entendimiento de la evaluación sin tener que recurrir a los números para su evaluación.

[12] FRANCE, Stephen y AKKUCUK, Ulas. A review, Framework, and R toolkit for Exploring, Evaluating, and Comparing Visualization Methods. 2020.

Esta investigación fue importante para el proyecto debido a la comparación que hace de métodos RD, teniendo en cuenta que los resultados de los mismo no siempre arrojan resultados perfectos, por lo que se recurre a su evaluación, así mismo se hace un recorrido sobre diferentes técnicas de visualización de datos, lo cual es importante ya que esta investigación tiene como uno de sus objetivos mostrar de forma visual la evaluación de diferentes métodos RD.

Antecedentes nacionales

Según Rosero, Díaz, Salazar, y otros en su artículo titulado "Interactive data visualization using dimensionality reduction and similarity-based representations", publicado en febrero de 2017[13]. El área de RD mejora campos como el reconocimiento de patrones y la presentación inteligible de los datos, en la actualidad hay muchas formas de representar los datos, por lo que el objetivo de este trabajo es traer una forma de visualización de datos interactiva usando RD y similaridad basado en representaciones, esto se logra al implementar diversos métodos de RD, ya que al obtener los resultados de su aplicación se obtendrá una representación de los datos en una baja dimensión, ya que la mejor forma de visualizar los datos es mediante representaciones 2D o 3D, este nuevo enfoque de visualización llamado DataVisSim, la cual tiene 3 etapas: combinación de métodos RD, interacción y visualización; al finalizar, se evaluó el nuevo enfoque mediante la utilización de las curvas de calidad R_{NX}, para los resultados, se implementó una interfaz intuitiva para el usuario, en la cual sin importar si este es o no experto pueda interactuar con los métodos RD sin dificultades, como conclusión, la visualización interactiva de los datos permite un mayor entendimiento de los mismos, haciendo que diferentes áreas del conocimiento relacionadas al análisis de datos se vean beneficiadas, para esto, es necesario preservar la topología de los datos de tal forma de que estos puedan ser vistos gráficamente después de su procesamiento. Esta investigación es relevante para el proyecto, puesto que muestra la necesidad de evaluar los métodos RD mediante métricas que permitan conocer qué tan eficaces son estos cuantitativamente, además de la importancia que tiene la visualización de datos en distintas áreas del conocimiento.

El siguiente trabajo a pesar de no ser de los más actuales es de suma importancia ya que según Valencia, Daza y otros, en su artículo titulado "Comparación de métodos de reducción de dimensión basados en análisis por localidades", publicado en el año 2010[14]. Los métodos RD tienen diferentes aplicaciones dependiendo de las características del data set que se quiere trabajar, en este caso, se utilizan métodos RD enfocados a variables locales tales como Locally linear embeddin, Isometric feature mapping y máximum variance unfolding, esto debido a que existen variables globales también y para

---

[13] ROSERO, Paul; DÍAZ, P.; SALAZAR, Jose, *et al.* Interactive data visualization using dimensionality reduction and similarity-based representations. 2017.

[14] VALENCIA, Juliana; DAZA, Genaro, et al. Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades. 2010.

estas puede que otros métodos RD sean más eficientes, y su objetivo es saber cuál de los métodos obtiene mejores resultados, es decir, cual conserva mejor las propiedades locales de los datos, para su evaluación se utilizaron dos criterios, Error de conservación de vecindarios y promedio de vecinos conservados, como resultado la técnica maximun variance unfolding, presentó los mejores resultados al conservar mejor las distancias entre los puntos cercanos, por último se puede notar que hay diferentes factores que pueden influir significativamente en el desempeño de los diferentes métodos. Por lo tanto, para este proyecto es de suma importancia conocer otras medidas de calidad que permiten la evaluación de métodos RD con el fin de tener un marco teórico amplio sobre el tema en cuestión, que, en este caso, son las curvas de calidad R_{NX} y los métodos de reducción de dimensión.

Por otro lado, para dar un mayor aporte al presente estudio en cuanto a reducción de dimensión de datos, los métodos a emplear y la visualización clara de los mismos se encuentra la investigación realizada por Salazar, Unigarro, Peluffo y otros llamada Reducción de dimensionalidad, para datos interactivos visualización a través de un enfoque geodésico[15], aquí se incluye métodos que tratan con datos de manera interactiva para ser visualizados, el proceso conlleva a la aplicación de métodos de RD en datos de alta dimensión y métodos de matrices kernel las cuales se combinan linealmente aplicando el modelo Geo-Desic, la aplicación matemática de estos conceptos resulta en una mezcla a nivel kernel que se podrá posteriormente introducir en un entorno PCA (Principal Component Analysis) y de esta manera poder obtener datos en una baja dimensión. En este sentido, los autores proponen este método para que la información en LD (low Dimension), se pueda visualizar de manera fácil e intuitiva.

El modelo Geo-Desic anteriormente señalado, está basado en la geometría de Los domos geodésicos, también llamados cúpulas geodésicas son un tipo de estructura que consiste en la combinación de elementos poligonales para formar una estructura de forma semiesférica. De esta forma, se consigue un poliedro cuyos vértices se encuentran sobre una esfera, parábola, o un elipsoide. Esta última condición es necesaria para que el domo sea llamado geodésico. El enfoque del artículo está basado en el aspecto geográfico de la tierra que permite seleccionar puntos de forma interactiva mediante combinación de coordenadas geográficas y mediante modelos matemáticos se pueden trasformar para derivar en la reducción eficaz de los datos aplicando mediciones de curvas R_{NX} para la obtención de resultados.

Es por esta razón que estos conceptos representan los pilares más importantes para dar continuidad a la investigación de nuevos métodos para obtener de la manera más óptima datos de alta dimensión y convertirlos en datos de baja dimensión y que sean deducibles y fácilmente analizables para personas con

[15] SALAZAR, José; UNIGARRO, Diego, *et al.* Op. Cit.

mucha o poca experiencia en el análisis de datos, además que sean visibles de manera comprensible para el razonamiento humano.

Antecedentes regionales

Según Anaya, Ordoñez, Alvarado y otros, en su artículo titulado "Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: LDA versus PCA"[16]. Presenta un estudio comparativo entre dos métodos de reducción de dimensión lineal tales como: Análisis de Componentes Principales y Análisis Discriminante Lineal, se nos presenta una definición sobre el término de reducción de dimensión (RD), el cual dicen es el proceso de llevar una muestra de datos de alta dimensión a un espacio de baja dimensionalidad, el cual promete reservar la mayor información intrínseca de las muestras originales.

El problema tratado dentro del estudio es la comparación cuantitativa de los métodos espectrales para la reducción de dimensión, entrando en explicaciones teóricas de cada método y llevándolos a prueba con dos bases de datos, evaluándolos bajo ciertos criterios objetivos con lo cual se pretende determinar cuál de estos obtiene los mejores resultados, con esto concluyen diversas fortalezas y debilidades de cada algoritmo, haciendo necesario también evaluar aspectos relevantes como la presentación de complejidad en tiempo de ejecución. Este proyecto es de gran ayuda dado que ejecuta una evaluación entre dos métodos de reducción de dimensión aportando ideas y posibles nuevas evaluaciones a tener en cuenta al momento de evaluar estos métodos.

Así mismo Salazar y otros en su artículo titulado "Generalized Low-Computational Cost Laplacian Eigenmaps" Publicado en 2018[17]. En el cual se habla sobre el alto costo computacional al realizar los métodos de reducción de dimensión, por lo cual se hace necesario proponer una estrategia para reducir el costo computacional requerido en el cálculo de vectores propios y valores propios necesarios para la ejecución adecuada de estos métodos y además tener un método más interactivo de los datos de cara al usuario, con esto proponen un modelo que permita al usuario visualizar dinámicamente los datos a través de una mezcla ponderada, en los casos en que desarrollaron esta técnica concluyeron que permite generar representaciones de baja dimensión y aunque esta metodología es menos costosa para el cálculo que KPCA, aún requiere de un tiempo de procesamiento considerable para los

[16] ANAYA, Andrés, *et al*. Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: LDA versus PCA . 2016.

[17] SALAZAR, Jose; BASANTE, Cielo; PEÑA, Diego y CRUZ, Lilian. Generalized Low-Computational Cost Laplacian Eigenmaps . 2018.

vectores y valores propios. El aporte que brinda este estudio es un método con el cual ya se ha experimentado que puede ser usado para dar un rendimiento favorable de cara al usuario.

Por otro lado, Peña y otros, en su artículo "Interactive visualization methodology of high-dimensional data with color based model for dimensionality reduction" publicado en 2016[18]. La trasformación de datos de alta dimensión en una versión de menor dimensión es un área de investigación ampliamente estudiado, dada su capacidad de reducir costo computacional, pero a pesar de la existencia de herramientas que alcanzan indicadores de eficiencia, exploración y representación de datos de alta dimensión, carecen de propiedades como interactividad y controlabilidad, lo que hace necesario la intervención de expertos que proporcionen conocimiento previo del sistema. En consecuencia, existe una brecha entre el conocimiento de los usuarios y la base de datos a analizar.

Se presenta un modelo cromático como una alternativa adecuada para reducir la brecha entre los usuarios y la base de datos por que los métodos de DR se pueden seleccionar / mezclar a través de un marco basado en colores.

Este fue un aporte significativo para el proyecto dado que plantea de manera muy interesante la forma en que diversos métodos de reducción de dimensión pueden mezclarse y producir resultados.

2.2 Supuestos teóricos de la investigación

2.2.1 Descubrimiento de conocimiento en bases de datos

Los métodos RD pueden hacer parte de diferentes etapas dentro del proceso de descubrimiento en bases de datos (KDD por sus siglas en inglés) que según Fayyad, Shapiro y Smyth" El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos)" [19]

Hu, y Saskatchewan afirman que[20]. El descubrimiento de conocimiento es el proceso de obtener información que en un principio no parecía ser importante, KDD es un área de investigación activa con la que se puede obtener cantidad de beneficio para cualquier negocio, sin embargo, muchas áreas como la

---

[18] PEÑA; *et al.* Op. Cit.

[19] FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. From Data Mining to Knowledge Discovery in Databases, Citado por NIGRO, Hector; XODO, Daniel, et al. KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario.

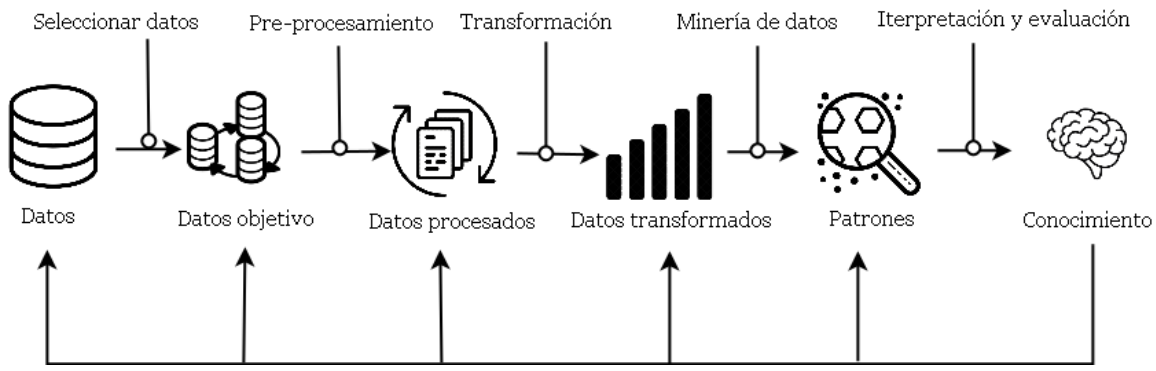
[20] HU, Xiaohua; SASKATCHEWAN, Regina. KNOWLEDGE DISCOVERY IN DATABASES AN ATTRIBUTEORIENTED ROUGH SET APPROACH, 1995.

ciencia, los gobiernos, etc. Se vieron asfixiadas por la cantidad de datos que eran almacenados en las bases de datos, con el fin de obtener patrones significativos a partir de las herramientas de aquel entonces, sin embargo, los avances tecnológicos han permitido que no solo aquellas personas con conocimiento en estadística o análisis de datos, puedan hacer uso de las nuevas herramientas, sino que se han desarrollado diferentes herramientas KDD de fácil aprendizaje que permiten hacer del proceso KDD más intuitivo.

Muchos algoritmos de ML se han implementado en KDD, sin embargo, las bases de datos que contienen la materia prima para ML, suelen tener cantidad de datos muy largos, incompletos, ruidosos y redundantes, que terminan afectando el rendimiento de los algoritmos y, por lo tanto, se terminan obteniendo malas predicciones que afectarían al negocio.

Por lo anteriormente mencionado, KDD contiene una serie de etapas donde es posible colectar los datos y hacerles un tratamiento de tal forma que se obtenga un data set con el cual pueda trabajarse adecuadamente, las etapas se pueden ver en la Figura 1

Figura 1: Pasos que componen el proceso KDD



Fuente: Elaboración propia

Según Fayyad, Shapiro y Smyth.[21] El proceso KDD es interactivo e iterativo en donde se incluyen diferentes pasos guiados por el usuario Fayyad describe dichos pasos:

- **Etapa 1.** Primeramente, se debe entender todo lo relacionado al área de dominio o negocio, siendo esta área en donde se va a aplicar el proceso KDD para la obtención del conocimiento

- **Etapa 2.** Esta etapa es prácticamente la de recolección de los datos, es decir, obtener los datos objetivos con el fin de centrarse en un sub set de donde se obtendrán las muestras y las variables de interés con el fin de realizar el descubrimiento.

[21] FAYYAD, U.; SHAPIRO, G. y SMYTH, P. From Data Mining to Knowledge Discovery in Databases. 1996.

- **Etapa 3.** Una vez se hayan recolectado los datos, es necesario que pasen por un proceso de pre-procesamiento, en donde se busca limpiar los datos, es decir, deshacerse de datos vacíos, ruidosos, redundantes o innecesarios, con el fin de más adelante obtener el mejor rendimiento en el modelado, aquí se aplican estrategias para manejar los datos anteriormente mencionados.

- **Etapa 4.** Se reduce los datos y se proyectan para ver características útiles teniendo en cuenta el objetivo de la terea, haciendo uso de métodos RD es posible reducir el número de variables que no requieran ser consideradas.

- **Etapa 5.** Se tiene en cuenta el objetivo del proceso KDD con algún método de minería de datos, esto con el fin de obtener el mejor rendimiento, es decir, se puede usar clasificación, regresión, etc. Dependiendo del objetivo.

- **Etapa 6.** Se hace un análisis exploratorio, modelado y selección de hipótesis, una vez escogido el o los métodos a utilizar, es posible buscar patrones relevantes, por lo tanto, se decide que modelos y parámetros son los más apropiados, además de relacionar el criterio general del proceso KDD con los métodos de minería de datos utilizado, de tal forma que se pueda presentar de la mejor forma al cliente los resultados.

- **Etapa 7.** Una vez analizado diferentes métodos y rutas que elegir, se realiza la minería de datos ya sea con reglas de clasificación, cluster o regresión para obtener patrones.

- **Etapa 8.** Una vez obtenidos los patrones, se interpretan mediante visualizaciones intuitivas para el usuario, esta etapa puede retornar a cualquiera de las etapas 1 a 7 de forma iterativa.

- **Etapa 9.** Una vez realizadas las anteriores etapas, debe documentarse el proceso y el conocimiento obtenido puede incorporarse a diferentes sistemas con el mismo fin, así como a diferentes grupos interesados en dicho conocimiento, una de las cosas más importantes es contrastar el conocimiento obtenido con el que se tenía anteriormente.

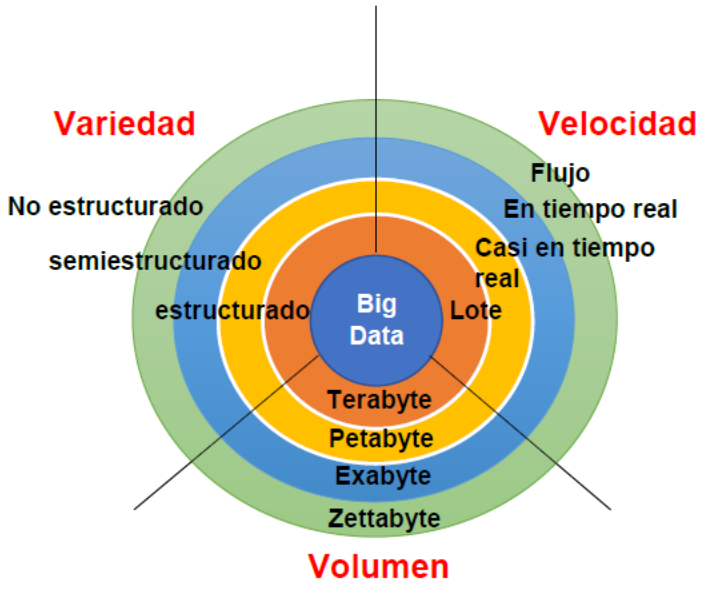
El proceso KDD puede tener iteraciones importantes, así como ciclos de ejecución con el fin de tener un proceso más pulido, cada proceso dentro del proceso es importante, suele creerse que el modelado es uno de los procesos más importantes, sin embargo, hay otros igual o más importantes como la limpieza de los datos.

2.2.2 Big Data

Big data es uno de los conceptos más utilizados en la actualidad en diferentes entornos, según Seref Sagiroglu y Duygu sinanc[22]. Big data es el término para referirse a los sets de datos masivos, estos datos se obtienen de diferentes fuentes y en diferentes formatos, pueden venir de transacciones online, correos electrónicos, videos, audio, imagen, redes sociales, del sector salud, sensores, etc. Su crecimiento masivo hace difícil, capturar, almacenar, manejar, compartir, analizar y visualizar los datos con las herramientas tradicionales.

Philip Russom afirma que[23] a principio del año 2000 Big data se volvió en un gran problema debido a que la velocidad con la que crecían los datos, las tecnologías de almacenamiento y las CPU se vieron abrumadas por los terabytes de datos, de tal forma que las tecnologías de la información se enfrentaron a una crisis que con el tiempo fue solucionada, esta crisis permitió que no solo se desarrollaran CPU con más almacenamiento, velocidad e inteligencia, sino que tenían la capacidad de proveer un mejor manejo de dichos datos, lo cual no pasó desapercibido para las empresas puesto que éstas en la actualidad, exploran Big Data para conseguir información valiosa para sí mismas, con el surgimiento de avanzadas técnicas de análisis de datos, es posible estudiar el Big data para comprender el estado actual de cualquier negocio y predecir su comportamiento a futuro. Big data tiene 3 componentes principales: Variedad, Velocidad y Volumen como se muestra en la Figura 2

Figura 2: Las 3 V de Big Data



Fuente: Elaboración propia

[22] SAGIROGLU, Seref ;SINANC, Duygu. Big Data: A review, IEEE, 2013
[23] RUSSOM, Philip. Big Data analytics. 2011.

2.2.2.1 Variedad

La variedad es la principal razón de que Big Data sea lo que es, esto debido a que los datos vienen de diferentes fuentes, pueden ser texto, video, audio, etc. Así mismo los datos pueden ser de 3 tipos diferentes: estructurados, semi estructurados y no estructurados. Los datos estructurados ya vienen etiquetados y organizados, en cambio, los datos no estructurados pueden ser aleatorios lo cual hace difícil su análisis y por último los semi estructurados, no vienen en campos separados, pero si vienen con etiquetas para la separación de elementos.

2.2.2.2 Volumen

Según Xindoung Wu y Xingquan Zhu[24]. Cada día se generan 2.5 quintillones de bytes de datos y el 90% de los datos en el mundo los últimos 2 años, el tamaño de datos que se genera supera los terabytes y petabytes.

2.2.2.3 Velocidad

Como afirma nuevamente Seref Sagiroglu y Duygu sinanc[25] Big data al tener cantidades de datos tan extremadamente grandes, la velocidad es un atributo importante, no solo aquí sino en muchos otros procesos con el fin de tener una mejor respuesta en todas las actividades que se realicen, como la extracción, almacenamiento etc.

2.2.3 Reducción de dimensión

Como ya se ha mencionado anteriormente, en la actualidad se generan gran cantidad de datos que vienen en diferentes tamaños, formatos y dimensiones, según Kuang, Zhang y otros.[26] Big data provee un mejor entendimiento del mundo real así como servicios de alta calidad a partir del análisis de datos, sin embargo, los datos que allí se alojan están en constante crecimiento y tienen mucha complejidad, el exceso de la dimensión de los datos hace que éstos se vuelvan redundantes y confusos, con el fin de realizar análisis de datos de calidad, extraer las características que mejor definen los datos se ha vuelto en una necesidad innegable, la RD permite la disminución de las variables, lo cual permite abordar ese aumento de dimensión en los datos.

Aunque la RD se encuentra en varias etapas del proceso KDD, es principalmente conocida por su gran labor en el pre-procesamiento de los datos con el fin de mejorar los procesos en las siguientes etapas, según Thippa, Kumar y otros.[27] ML es una de las tecnologías que más crecimiento ha tenido en los últimos años, esto debido, a que es utilizada en diferentes campos como

---

[24]WU, Xindoung; ZHU, Xingquan. Data mining with Big Data, IEEE, 2014

[25] SAGIROGLU, Seref ;SINANC, Duygu. Op. Cit. pp 43.

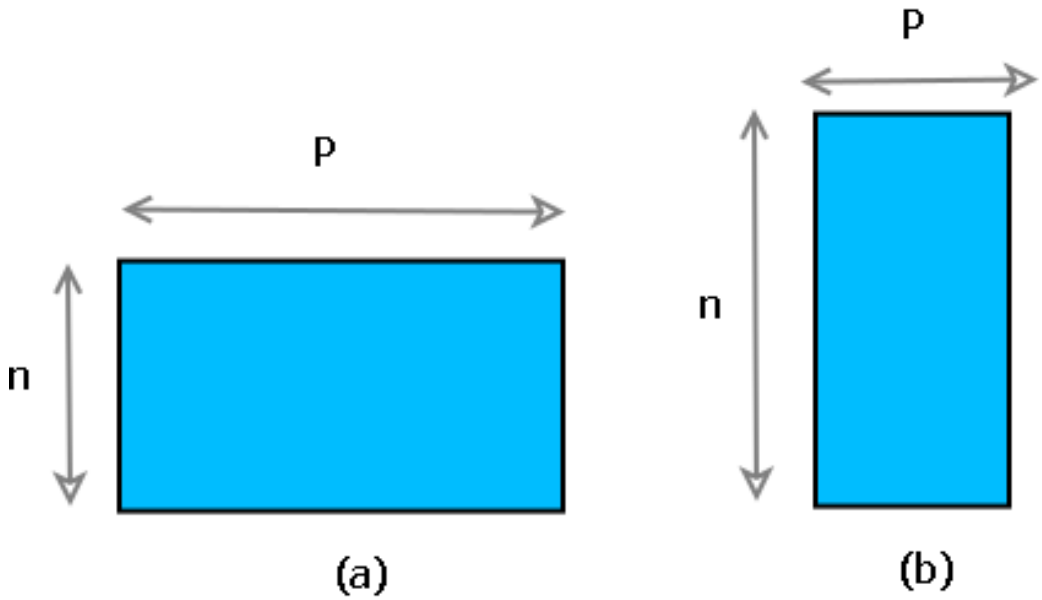
[26] KUANG, Liwei; ZHANG, Laurence, et al. A Holistic Approach to Distributed Dimensionality Reduction of Big Data. IEEE, 2015

[27] THIPPA, G.; KUMAR, M., et al. Analysis of dimensionality reduction techniques on Big Data. IEEE, 2020.

la salud, visión por computadora, negocios, bioinformática, etc. ML usa los datos para permitirle al computador aprender a predecir patrones, es principalmente útil en la medicina ya que permite dar diagnósticos, así como prevenir enfermedades y muertes con una buena precisión y rapidez, todo eso es posible porque la RD juega un papel fundamental al eliminar datos que no sirven para realizar una buena etapa de modelado.

Como se ha mencionado anteriormente, los negocios también se ven beneficiados y más si estos cuentan con bases de datos muy grandes, Bertini, Tau y otros[28], esto debido a que si los datos se encuentran en alta dimensión, no podrán ser visualizados fácilmente y por lo tanto extraer la información más valiosa permitiría una mejor representación de los mismos, en ese orden de ideas Sacha, Zhang y otros afirman que[29] la reducción de dimensión (RD) es una de las mejores técnicas para la abstracción de datos en la infovis, lo que se quiere con la RD Según Fodor[30] es que a partir de una variable aleatoria con dimensión p es decir, $\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_p})^{\boldsymbol{T}}$ se obtenga una menor dimensión de la misma $\boldsymbol{s} = (\boldsymbol{s_1}, \dots, \boldsymbol{s_k})^{\boldsymbol{T}}$ en donde $k < p$, en la Figura 3 se tiene una representación visual.

Figura 3: Objetivo de la reducción de dimensión



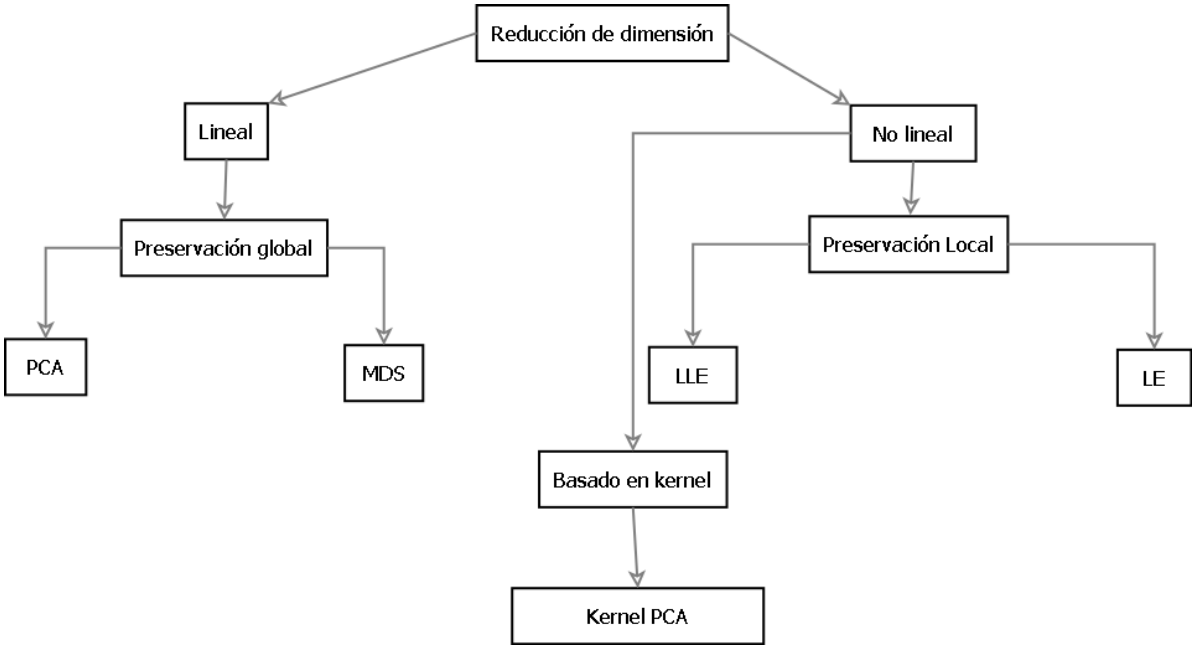
Fuente: Elaboración propia

En donde n es el número de muestras y p el de variables, entendiendo que el número de variables es la dimensión, lo que se quiere con RD es transformar

[28] BERTINI, Enrico; TAU, Andrada; KEIM, Daniel. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, IEEE, 2011.
[29] SACHA, Dominik ; ZHANG, Leishi, et al. Visual interaction with dimensionality reduction: A structured literature analysis, 2016.
[30] FODOR, Imola. A survey of dimensionality reduction techniques. 2002.

(a) en (b). Según Ayesha, Kashif y Talib[31] debido a que los datos pueden tener comportamientos lineales como no lineales, existen métodos RD que se centran en cada uno de ellos, los métodos lineales utilizan funciones lineales simples, por otro lado los métodos no lineales utilizan estructuras no lineales complejas, una de las formas para presentar relaciones no lineales de los datos, es usar métodos basados en kernel la Figura 4 muestra una vista general de los métodos RD.

Figura 4: Taxonomía de las técnicas RD



Fuente: Elaboración propia

2.2.3.1 Métodos lineales

2.2.3.1.1 Análisis de componentes principales (PCA)

PCA es un método global y no parametrizado, es decir, no recibe ningún tipo de parámetro en comparación con otros métodos que reciben el parámetro K, que hace referencia a la escala de vecindarios que se quiere analizar, además es uno de los más usados para el análisis exploratorio de datos que se basa en la matriz de covarianza de los datos, según Shlens[32] el objetivo de PCA es encontrar las características más importantes en los datos, las cuales generalmente se encuentran escondidas con el fin de re-expresar el data set con dichas características. Fodor afirma que[33] PCA reduce la dimensión al encontrar combinaciones lineales ortogonales a partir de los datos originales, estas combinaciones lineales son llamadas componentes principales (PC por sus siglas en ingles), el primer PC denotada s_1 es una combinación lineal con

[31] AYESHA, Shaeela; KASHIF, Muhammad y TALIB, Ramzan. Overview and comparative study of dimensionality reduction techniques for high dimensional data, 2020.
[32] SHLENS, Jonathon. A tutorial on principal component analysis, 2014.
[33] FODOR, Imola. OP. Cit.

la varianza más grande, entonces tenemos $s_1 = X^T w_1$ donde $w_1 = (\omega_{1,1}, \dots \omega_{1,p})^T$ es un vector de coeficientes con dimensión p, solucionando así:

$$w_1 = \arg max_{\|w=1\|} Var\{X^T w\}$$

El segundo PC es la combinación lineal con la segunda varianza más grande, y además este es ortogonal al primer PC y así sucesivamente con los demás PC, la cantidad de PCs que pueden haber es el de la cantidad de variables originales, sin embargo los primero son los que explican mejor la varianza, por lo que los otros pueden ser descartados sin temer a perder información, ya que la varianza depende de la escala de las variables, es decir, puede tener muchos comportamientos, lo esencial es estandarizar cada variable de tal forma que se tenga media 0 y desviación estándar 1, de esa forma las variables originales que podían tener diferentes unidades de medidas, ahora todas tienen unidades de medidas comparables, las variables estandarizadas con la matriz de covarianza se denota como:

$$\sum_{pxp} = \frac{1}{n} XX^T$$

En donde se usa el teorema de la descomposición espectral para reescribir la anterior ecuación de la siguiente forma:

$$\sum = U \Lambda U^T$$

Donde Λ es una matriz diagonal de los valores propios ordenados $\lambda_1 \leq \cdots \leq \lambda_p$ y U es una matriz ortogonal que contiene los vectores propios, una de las propiedades de la descomposición de los valores propios es que la variación total es igual a la suma de todos los valores propios de la matriz de covarianza:

$$\sum_{i=1}^{p} Var(PC_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} traza(\sum)$$

Y que

$$\sum_{i=1}^{k} \lambda_i \, / traza(\sum)$$

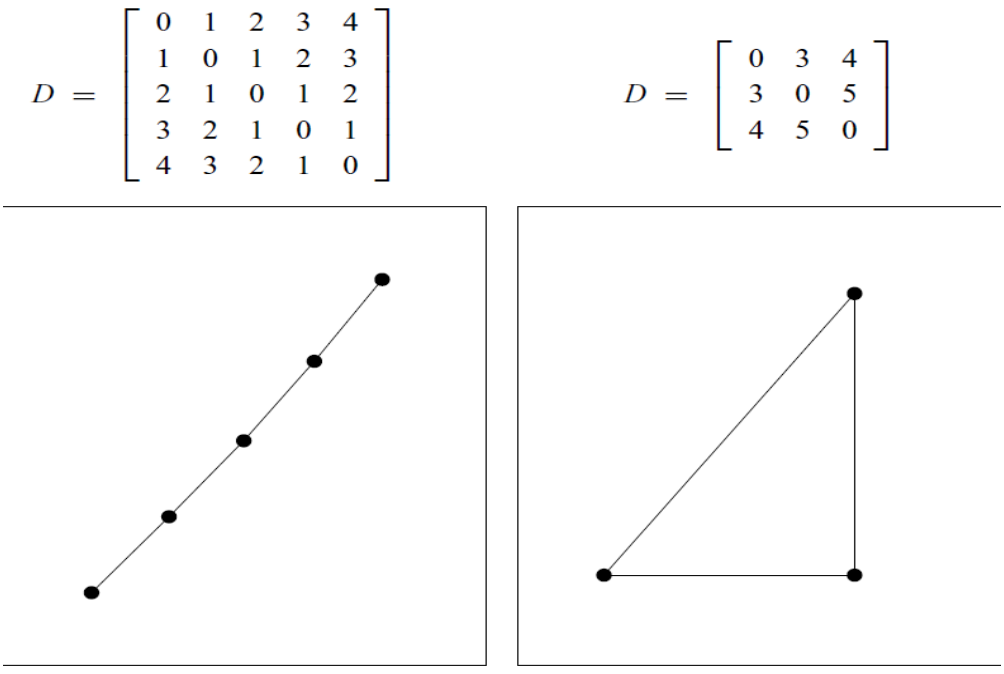
Es la proporción acumulada de la varianza explicada de los primeros PCs, si se representa esa proporción acumulada de forma visual se puede seleccionar los PC apropiados para poder explicar la variación de forma general.

2.2.3.1.2 Escalonamiento multidimensional (MDS)

Arce y De Francisco.[34] El escalamiento multidimensional, es una técnica que pretende representar un conjunto de objetos en un espacio de baja dimensión, la palabra objeto puede referirse a cualquier entidad que deseemos escalar, MDS se ha utilizado para estudiar dimensiones subyacentes a las percepciones del hablar humana, también ha sido utilizado como método de visualización de datos, y generalmente las aplicaciones MDS están diseñadas para descubrir patrones a lo largo de dos o más dimensiones por lo que su uso en psicología es uno de los más conocidos ya que permite determinar lo similar o no similar que son los objetos en el espacio de baja dimensión.

Según Buja y Swayne[35] El escalamiento multidimensional utiliza conceptos como proximidad y función de costo, MDS utiliza la proximidad para referirse a la disimilitud de los datos, en otras palabras la diferencia de los datos, para un objeto etiquetado como $i = 1, \dots, N$, la proximidad de los datos está dada por D_{ij} el objetivo de MDS entonces es mapear el objeto $i = 1, \dots, N$ a puntos incrustados $x_1, \dots, x_N \in \mathbb{R}^k$ de tal forma que D_{ij} sea lo más aproximada posible a la distancia $\|x_i - x_j\|$, en cuestiones practicas $k = 1,2,3$ ya que esas son las dimensiones que el cerebro humano logra comprender.

Figura 5: Contraste entre la matriz de disimilitud y su escalonamiento óptimo.



Fuente: BUJA, Andreas; SWAYNE, Deborah, el al. Ibid.

[34] ARCE, Constantino; DE FRANCISCO, Cristina. Escalonamiento multidimensional: Concepto y aplicaciones, 2010.

[35] BUJA, Andreas; SWAYNE, Deborah, el al. Data visualization with multidimensional Scaling. 2008

MDS busca minimizar la función de costo también llamada “Stress”, que es la medida de la falta de ajuste de D_{ij} y la distancia ajustada $\|x_i - x_j\|$, Stress es la suma de los cuadrados:

$$Stress_D(x_1, \dots, x_N) = \left(\sum_{i \neq j = 1..N} (D_{ij} - \|x_i - x_j\|)^2 \right)^{\frac{1}{2}}$$

En donde asumimos que las disimilitudes de los datos son simétricas.

2.2.3.2 Métodos no lineales

2.2.3.2.1 Incrustamiento Local Lineal (LLE)

Según Saul, Roweis[36] LLE es un método RD no lineal y de aprendizaje no supervisado, el cual, que a diferencia de PCA que conserva una estructura global, busca conservar la geometría local de los datos, tal como PCA y MDS, LLE es fácil de implementar y genera encrustamientos no lineales con buena calidad.

Así mismo Vanderplas y Connoly afirman que[37] Considerando un conjunto de puntos dados por $\boldsymbol{X} = [\boldsymbol{x_i}, \dots \boldsymbol{x_N}], \boldsymbol{x_i} \in \mathbb{R}^{\boldsymbol{D}}$, el cual es mapeado en un espacio menor $\boldsymbol{Y} = [\boldsymbol{y_i}, \dots \boldsymbol{y_N}], \boldsymbol{y_i} \in \mathbb{R}^{\boldsymbol{d}}$ con $d < D$ para cada punto $\boldsymbol{x_i}$ se calculan los $\boldsymbol{n}^{(i)} = [n_j^{(i)}, \dots, n_K^{(i)}]^T$, los cuales son los vecinos más cercanos, en donde además se calcula $\boldsymbol{w}^{(i)} = [w_j^{(i)}, \dots, w_K^{(i)}]^T$ que representa la reconstrucción de los pesos para los k-vecinos de $\boldsymbol{x_i}$ y a partir de esto se calcula la función de costo de reconstrucción:

$$\mathcal{E}_1^{(i)}(\boldsymbol{w}^{(i)}) = \left| \boldsymbol{x_i} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{w_j^{(i)}} \boldsymbol{x}_{\boldsymbol{n_j^{(i)}}} \right|^2$$

En donde se minimiza $\mathcal{E}_1^{(i)}(\boldsymbol{w}^{(i)})$ sujeto a la siguiente condición

$$\sum_j \boldsymbol{w_j^{(i)}} = 1$$

Y una vez determinados los pesos para cada punto, dichos pesos son utilizados para determinar el vector de proyección $\boldsymbol{y_i}$, el cual minimiza la función de costo global:

$$\mathcal{E}_2(\boldsymbol{Y}) = \sum_{i=1}^{N} \left| \boldsymbol{y_i} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{w_j^{(i)}} \boldsymbol{y}_{\boldsymbol{n_j^{(i)}}} \right|$$

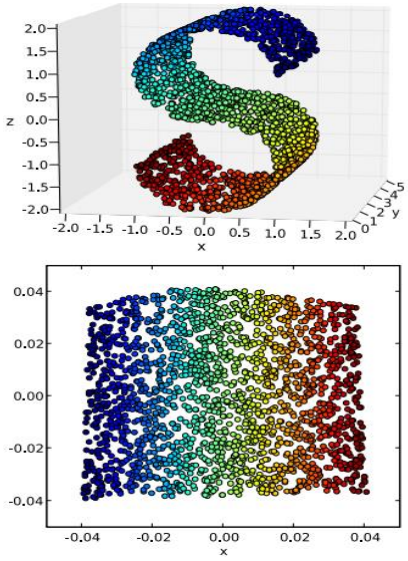
---

[36] SAUL, Lawrence y ROWEYS, Sam. An introduction to localy linear embedding. 2001

[37] VANDERPLAS, Jake y CONNOLLY, Andrew. REDUCING THE DIMENSIONALITY OF DATA: LOCALLY LINEAR EMBEDDING OF SLOAN GALAXY SPECTRA, 2009.

De tal forma que los vecindarios locales de $\boldsymbol{X}$ y $\boldsymbol{Y}$ tienen propiedades similares. Computacionalmente, estos pasos pueden ser implementados de una forma eficiente si se usan métodos del algebra lineal optimizados.

Figura 6: Incrustación de LLE con K = 15.



Fuente: VANDERPLAS, Jake y CONNOLLY, Andrew. Ibid.

En la anterior figura se observa que el conjunto de datos está agrupado por colores, y la incrustación realizada con un parámetro K = 15 logra conservar ese agrupamiento en un especio de 2 dimensiones.

2.2.3.2.2 Kernel Análisis de componentes principales (KPCA)

Según Wu, Su y Carpuat[38] PCA solo tiene un buen rendimiento cuando los datos son lineales, KPCA es una extensión de PCA para datos no lineales, de tal forma que procesa los datos cuya estructura no es lineal usando el método Kernel, KPCA tiene una gran aplicación para los problemas de lenguaje natural.
Según [39] asumiendo que se tiene una transformación no lineal $\phi(x)$ del espacio original D-dimensional a un espacio M-dimensional con $M \gg D$, entonces el punto x_i es proyectado a un punto $\phi(x_i)$, de tal forma se podría realizar PCA, sin embargo, esto es extremadamente costoso e ineficiente computacionalmente hablando por lo que se usan los métodos kernel para computarlo.

Primeramente, se asume que el punto proyecto tiene media 0 es decir:

[38] WU, Dekai; SU, Weifeng y CARPUAT, Marine. A. Kernel PCA Method for Superior Word Sense Disambiguation, 2004.
[39] WANG, Quang. Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models. 2014

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i) = 0$$

La matriz de covarianza de la característica proyectada es MxM y se calcula por:

$$C = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\,\phi(x_i)^T$$

A su vez, se calculan sus valores y vectores propios:

$$Cv_k = \lambda_k v_k$$

Con $k = 1{,}2 \ldots, M$, a partir de las anteriores dos ecuaciones se tiene:

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\,\{\phi(x_i)^T v_k\} = \lambda_k v_k$$

Si se sobre escribe quedaría así:

$$v_k = \sum_{i=1}^{N} a_{ki}\phi(x_i)$$

Ahora se sustituye v_k en la anterior ecuación se obteniene lo siguiente:

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\,\phi(x_i)^T\sum_{i=1}^{N}a_{ki}\phi(x_i) = \lambda_k\sum_{i=1}^{N}a_{ki}\phi(x_i)$$

Definimos una función kernel:

$$\mathcal{K}(x_i, x_j) = \phi(x_i)^T\phi(x_i)$$

Si se multiplica $\phi(x_l)^T$ a ambos lados de la anterior ecuación se tiene:

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{K}(x_l, x_i)\sum_{i=1}^{N}a_{ki}\mathcal{K}(x_i, x_j) = \lambda_k\sum_{i=1}^{N}a_{ki}\mathcal{K}(x_l, x_i)$$

Se puede usar la siguiente notación para la siguiente matriz:

$$K^2 a_k = \lambda_k NK a_k$$

En donde

$$K_{i,j} = \mathcal{K}(x_i, x_j)$$

Y a_k es el vector columna N-dimensional de a_{ki}:

$$a_k = [a_{k1}, \dots, a_{kN}]^T$$

De tal forma que el componente principal basado en kernel puede calcularse usando:

$$y_k(x) = \phi(x_i)^T v_k = \sum_{i=1}^{N} a_{ki} \mathcal{K}(x, x_i)$$

De esta forma el costo computacional al calcular $\phi(x_i)$ explícitamente se resuelve calculando la matriz kernel a partir del dataset de entrenamiento.

2.2.3.2.3 Laplacian Eigenmaps (LE)

Teniendo en cuenta a Belkin y Niyogi. [40] LE como LLE busca preservar la estructura local de los datos en baja dimensión, así mismo, el algoritmo de ambos es muy similar, dado un conjunto de datos con k puntos pertenecientes a la alta dimensión, se construye un grafo ponderado con k nodos, es decir, se crea un nodo por punto y se tienen en cuenta las aristas entre cada punto.

El primer paso entonces es construir el grafo, si los puntos x_i y x_j son cercanos, entonces se enlazan con una arista sus respectivos nodos i y j, hay dos variaciones para hacer esto:

- Se puede saber que vecinos tiene i a partir de los $\mathcal{E}$-vecindarios en donde el nodo i y j están conectados por una arista si $\|x_i - x_j\|^2 < \mathcal{E}$, las ventajas que trae esta forma es que la relación es simétrica por naturaleza, sin embargo, cuando hay grafos con muchas conexiones el valor de $\mathcal{E}$ es difícil de escoger

- La otra forma es obtener los n vecinos más cercanos, en este caso, los nodos i y j están conectados si i se encuentra entre los n vecinos más cercanos y viceversa, su ventaja es que es fácil de escoger, ya que dirige los grafos conectados, sin embargo, es geométricamente menos intuitivo.

Seguidamente se eligen las ponderaciones, aquí nuevamente hay dos formas de hacerlo

- Usando kernel si los nodos i y j están conectados entonces

[40] BELKIN, Mikhail y NIYOGI, Partha. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, 2001.

$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

Donde t es un número real.

- Está forma no recibe ningún parámetro y por lo tanto $W_{ij} = 1$ si y solo si los nodos i y j están conectados por una arista

Por último, se crea el mapa propio o eigenmap, se computan los valores propios y vectores propios para el problema:

$$Ly = \lambda Dy$$

En donde D es la matriz de peso diagonal y sus entradas son las columnas sumadas de W, $D_{ii} = \sum_j W_{ji}\ y\ L = D - W$ es la matriz laplaciana, la cual es simétrica, positiva y semi definida. Si tenemos a $y_i, \dots, y_{k-1}$ la cual es solución de la anterior ecuación y ordenada de forma ascendiente acorde con sus valores propios y con i empezando desde 0, entonces la imagen de x_i proyectada en el espacio de baja dimensión $y_1(i), \dots, y_m(i)$. Siendo de menor dimensión que el espacio de entrada.

2.2.4 Matriz co-ranking

La matriz de co-ranking propuesta por Lee y Verleysen[41] es utilizada por las métricas de evaluación de métodos RD y es un histograma conjunto de los rangos y es la suma de N matrices permutadas con tamaño $N - 1$, para poder entender la matriz de co-ranking es necesario entender ciertos conceptos, generalmente los métodos de reducción de dimensión usan distancias de una forma u otra, en ese orden de ideas, δ_{ij}es la distancia entre dos puntos en alta dimensión mientras que d_{ij}es la distancia de los dos puntos en baja dimensión, a partir de estas distancias se pueden calcular los rangos.

El rango de dos puntos en alta dimensión se denota como $p_{ij} = \left|\{k: \delta_{ik} < \delta_{ij}\ o\ (\delta_{ik} = \delta_{ij}\ y\ 1 \le k < j \le N)\}\right|$, de esta misma forma el rango de los puntos en baja dimensión es $r_{ij} = \left|\{k: d_{ik} < d_{ij}\ o\ (d_{ik} = d_{ij}\ y\ 1 \le k < j \le N)\}\right|$ aquí el rango reflexivo es decir cuando $p_{ii} = r_{ii} = 0$, cuando los rangos en alta y baja dimensión son diferentes, se hace una distinción entre los k-vecinos de un dato en alta dimensión y sus k-vecinos en baja dimensión $v_i^K = \{j: 1 \le p_{ij} \le K\}$ y $n_i^K = \{j: 1 \le r_{ij} \le K\}$ respectivamente, en donde K es el tamaño de los vecindarios. A partir de los anteriores conceptos se crea la matriz de co-ranking Q

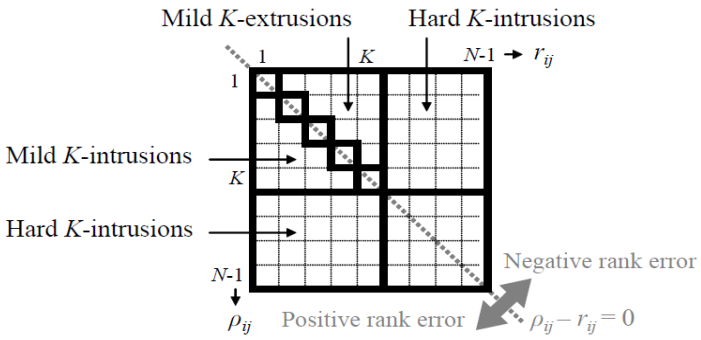
$$Q = [q_{kl}]_{1 \le k,l \le N-1}\ con\ q_{kl} = \left|\{(i,j): p_{ij} = k\ y\ r_{ij} = l\}\right|$$

En resumidas palabras k es el rango en alta dimensión y l en baja dimensión,

[41] LEE, John Y VERLEYSEN, Michel. Quality assessment of nonlinear dimensionality reduction based on K-ary neigborhoods, 2008.

a partir de aquí se define el rango de error que es la diferencia $p_{ij} - r_{ij}$ el cual si es positivo significaría una intrusión en el k-ésimo vecindario representado por n_i^K con respecto al original v_i^K, si la diferencia de los rangos es negativa se obtiene una extrusión, así mismo existe el concepto de vecindarios duros y leves que en otras palabras son vecindarios con gran tamaño y con tamaño moderado respectivamente, de tal forma que también existen las intrusiones y extrusiones duras y leves La matriz de co-ranking se muestra en la siguiente figura.

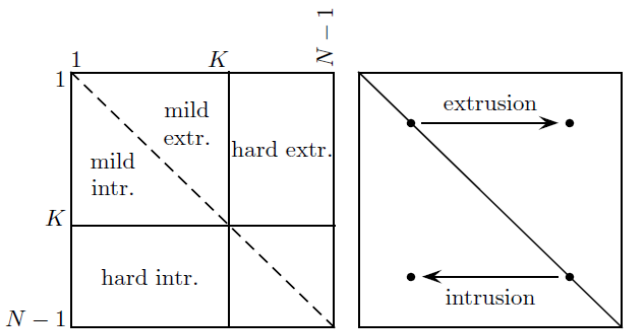
Figura 7: División de los bloques de la matriz de co-ranking.



Fuente: LEE, John Y VERLEYSEN, Michel. Quality assessment of nonlinear dimensionality reduction based on K-ary neigborhoods, 2008.

Los bloques de la matriz están definidos de la siguiente forma $\mathbb{LL}_K$, $\mathbb{LR}_K$, $\mathbb{UL}_K$ y $\mathbb{UR}_K$ y son los bloques inferior izquierdo, inferior derecho, superior izquierdo y superior derecho respectivamente, la diagonal está dada por $\mathbb{D}_K = \{(i,j): 1 \leq i \leq K\}$, así mismo se definen los triángulos inferior y superior ocasionados por la diagonal $\mathbb{LT}_K = \{(i,j): 1 < i \leq K \; y \; j < i\}$ y $\mathbb{UT}_K = \{(i,j): 1 \leq i < K \; y \; j > i\}$, por ultimo $\mathbb{UR}_K$ y $\mathbb{LL}_K$, son las K- extrusiones y K-intrusiones duras respectivamente y $\mathbb{UL}_K$ y $\mathbb{LR}_K$ son las K-extrusiones y K-intrusiones leves. En la Figura 8 se puede observar otra representación de la matriz.

Figura 8: Otra forma de representar la matriz co-ranking.



Fuente: Biehl, Michael; Hammer, Barbara. How to evaluate dimensionality reduction?- Improving the co-ranking matrix. 2011.

2.2.5 Topología

En los conjuntos de datos las variables generalmente dependen unas de otras, al momento de relacionarlas se puede obtener una estructura geométrica, la cual puede ser vista como algún tipo de objeto en dicho espacio, según Marta Stadler[42]. La topología es una de las ramas más recientes de las matemáticas, la cual estudia las propiedades de los objetos o figuras que son invariantes, es decir, sin importar que tipo de transformación tenga la figura, no aparecerán puntos nuevos en su estructura y por lo tanto, en la figura original y en la transformada los puntos próximos a cada punto no cambian, esta propiedad es conocida como continuidad y es necesaria, ya que lo que se quiere es que la transformación y su inversa sean continuas, de tal forma que se trabajaría con homeomorfismos.

Por otro lado, la universidad de Waterloo afirma que[43] la topología también es llamada "geometría de la hoja de caucho" esto debido a que los objetos pueden ser estirados y contraídos como cauchos, pero no pueden romperse, como ejemplo tenemos a un cuadrado que puede ser deformado en un círculo, es decir el cuadrado es topológicamente igual a un círculo y a su vez este es topológicamente equivalente a una elipse, esfera o a un triángulo.

2.2.6.1 Topología General

La topología general considera las propiedades locales de los espacios, esta cercanamente relacionada al análisis, en donde el concepto de continuidad es pertinente para definir espacios topológicos, y en los cuales los límites de las secuencias pueden ser consideradas. A veces las distancias pueden ser definidas en estos espacios, en tales casos estás son llamadas métricas de distancia.

2.2.6.2 Topología combinatoria

A diferencia de la anterior topología, esta considera las propiedades globales de los espacios, está construida por una red de vértices, bordes y caras, es la rama más vieja de la topología, esta rama ha mostrado que los espacios topológicamente equivalentes tienen el mismo invariante numérico a lo que se le llama la característica de Euler. El número (V-E+F) en donde V, E, F son el vértice, borde y cara de un objeto, por ejemplo, un tetraedro y un cubo puede ser topológicamente equivalente a una esfera.

2.2.7 Medidas de calidad

Según Valencia, Daza y otros.[44] La RD está muy relacionada a la tarea de visualización de datos, en donde es importante observar los datos en una, dos

[42] STADLER, Marta. ¿Qué es la topología?. 2002
[43] University of Waterloo. What is topology?.
[44] VALENCIA, Juliana; DAZA, Genaro, et al. Op. Cit. p. 140

o tres dimensiones, para esto la RD debe realizar un incrustamiento en baja dimensión de tal forma que se conserven la estructura intrínseca de los datos, es por ello que se han propuesto métricas que permitan conocer la calidad de la transformación y evaluar si los rendimientos de los métodos son efectivos, una de las formas más simples de evaluación es la visual, sin embargo esta se vuelve subjetiva, por lo tanto, poner la confianza en este tipo de métodos no es muy fiable. Los métodos RD se pueden evaluar mediante la comparación del espacio en alta dimensión con el mismo pero en baja dimensión, sin embargo la medida de dicha variedad no está dada, es decir, no está definida y por lo tanto no puede ser establecida de forma precisa, es por ello, que la calidad de las inmersiones realizadas no se pueden obtener con implementaciones generales, sino que se debe obtener mediante medidas alternativas con el fin de obtener una buena evaluación de las inmersiones de una forma objetiva y confiable.

2.2.7.1 Definición de vecindarios

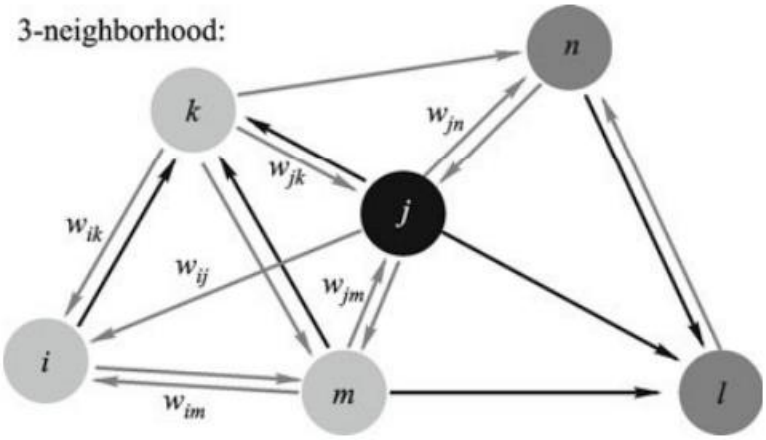
Los métodos de reducción de dimensión, aunque tienen diferentes aplicaciones y enfoques, buscan preservar la topología de los datos, es decir su estructura a partir del concepto de los vecindarios, de tal forma que las métricas planteadas por diferentes autores buscan determinar con que precisión se conservaron dichos vecindarios, según WANG, Jianzhong [45] $\mathfrak{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ es un data set y $\mathcal{E} > 0$ es una tolerancia. El $\mathcal{E}$-vecindario del punto $x_i \in \mathfrak{X}$ es un subconjunto de $\mathfrak{X}$ definido de la siguiente manera:

$$O_i = O_i(\mathcal{E}) = \mathfrak{X} \cap B_{\mathcal{E}}(i) \backslash \{x_i\}$$

Donde $B_{\mathcal{E}}(i) \in \mathbb{R}^D$ es una esfera con centro en x_i de radio $\mathcal{E}$ uno podría afirmar que si $x_j \in O_i$, entonces $x_i \in O_j$, sin embargo, esto no es siempre cierto debido a que la tolerancia $\mathcal{E}$ depende en gran medida de la dimensión del dataset y de la escala con la que son digitados los datos, la siguiente figura muestra como $x_j \in O_i$ no implica que $x_i \in O_j$.

[45] WANG, Jianzhong. Geometric Structure of High Dimensional Data and Dimensionality Reduction. 2012. p. 52

Figura 9: Demostración de la no conmutatividad de vecindarios.



Fuente: WANG, Jianzhong. Geometric Structure of High Dimensional Data and Dimensionality Reduction. 2012.

En la Figura 9 se puede apreciar que los vecinos de i son $k, m\ y\ j$, mientras que los vecinos de j son $k, m\ y\ n$.

2.2.7.2 Error de conservación de vecindarios (ECV)

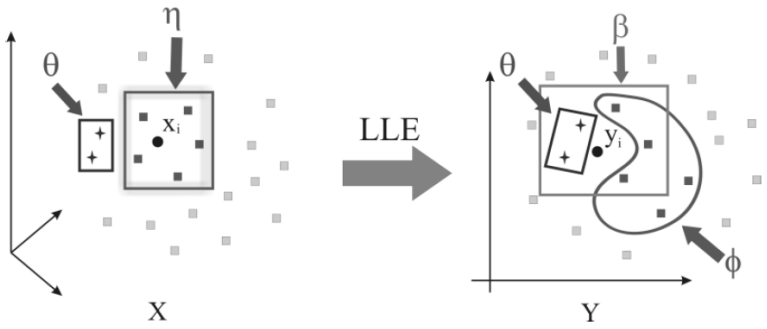
Una vez sabiendo que es un vecindario se puede dar inicio a las métricas existentes tal como lo afirma Valencia y otros [46] ECV se basa en la conservación de la geometría local además de la co-ubicación de los vecindarios, de tal forma que se pueda identificar si hay superposición en el espacio de menor dimensión.

$$ECV(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2n}\sum_{i=1}^{n}\{\frac{1}{k}\sum_{j=1}^{k}\left(D_{(\boldsymbol{X_I},\boldsymbol{\eta_J})} - D_{(\boldsymbol{Y_I},\boldsymbol{\varphi_J})}\right)^2 + \frac{1}{k_n}\sum_{j=1}^{k_n}\left(D_{(\boldsymbol{X_I},\boldsymbol{\theta_j})} - D_{(\boldsymbol{Y_I},\boldsymbol{\gamma_J})}\right)^2\}.$$

Esta métrica hace uso de la distancia Euclidiana D estandarizada con el fin de establecer un valor máximo igual a 1 y $\boldsymbol{\eta}$ son los vecinos más cercanos a cada dato en la alta dimensión, una vez realizada la inmersión o el incrustamiento, se calcula nuevamente un conjunto β de los vecinos más cercanos de cada punto en ese espacio y otro conjunto φ correspondiente a las proyecciones de $\boldsymbol{\eta}$, los vecinos de β que no son de $\boldsymbol{\eta}$ conforman un nuevo conjunto γ el cuál se define como $\boldsymbol{\gamma} = \boldsymbol{\beta} - (\boldsymbol{\beta} \cap \boldsymbol{\varphi})$, es decir, en esta métrica se tiene en cuenta los vecinos que no lograron conservarse en el espacio de baja dimensión.

[46] VALENCIA, Juliana; DAZA, Genaro, et al. Op. Cit. p. 141

Figura 10: Conjuntos espacio de entrada y espacio de salida.



Fuente: VALENCIA, Juliana; DAZA, Genaro, et al. Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades. 2010

2.2.7.3 Promedio de vecindarios conservados (PVC)

Esta métrica utiliza los conjuntos de la previamente mencionada, con la diferencia de que aquí se busca calcular el promedio de los vecindarios conservados tras la aplicación del método RD.

$$PVC = \frac{1}{n}\sum_{i=1}^{n}\frac{|\boldsymbol{\varphi_i} \cap \boldsymbol{\beta_i}|}{k_i}$$

De esta forma se obtienen aquellos vecindarios que se conservaron en ambas dimensiones y se pondera dividiéndolo entre la cantidad de datos analizados.

2.2.7.4 Integridad y Continuidad (T&C)

Según Venna y Kaski.[47] T&C es una medida de calidad que se utiliza para evaluar el rendimiento de métodos RD teniendo en cuenta la integridad de los vecindarios resultantes, esto debido a que los datos cercanos mapeados, generalmente también son cercanos en el espacio de entrada, se dice que los vecindarios han sido conservados, si el set de los k-vecinos más cercanos no cambia en los datos proyectados en menor dimensión, por el contrario, se dice que hay dos tipos de errores, el primero es que la cantidad de vecinos de un dato en menor dimensión, es mayor a los vecinos del mismo dato en mayor dimensión y el segundo es que los vecinos cercanos a dicho dato en alta dimensión, se observan lejanos en baja dimensión, el primer error es la integridad y el segundo es llamado discontinuidad.

La integridad se calcula de la siguiente forma:

$$M_1(k) = 1 - \frac{2}{Nk(2N-3k-1)}\sum_{i=1}^{N}\sum_{\boldsymbol{x_j}\epsilon U_k(\boldsymbol{x_i})}\big(r(\boldsymbol{x_i},\boldsymbol{x_j}) - k\big),$$

Por otro lado, la continuidad está dada por:

[47] Venna, Jarkko; Kaski, Samuel. Neigborhood preservation in nonlinear projection methods: An experimental study. 2001.

$$M_2(k) = 1 - \frac{2}{Nk(2N-3k-1)} \sum_{i=1}^{N} \sum_{\boldsymbol{x_j} \epsilon V_k(\boldsymbol{x_i})} \left(\hat{r}(\boldsymbol{x_i}, \boldsymbol{x_j}) - k\right),$$

Lo anterior permite calcular ambos errores, evaluando la calidad a partir de los mismos.

2.2.7.5 Meta criterio de continuidad local (LCMC)

Mientras que T&C calcula errores para medir la integridad de los vecindarios de un dato, el meta criterio de continuidad local en su lugar según Chen y Buja[48]. Mide que tan parecidos son los vecindarios de alta y baja dimensión, en otras palabras, LCMC define el promedio de los tamaños de la superposición de los vecindarios de alta dimensión y baja dimensión y por lo tanto, puede detectar problemas locales en la reducción de dimensión, dicha superposición de ambas dimensiones se calcula:

$$N_{K'}(i) = \left|\mathcal{N}_{K'}^{D}(i) \cap \mathcal{N}_{K'}^{X}(i)\right|, \;\; N_{K'} = \frac{1}{N} \sum_{i=1}^{N} N_{K'}(i).$$

La idea es obtener resultados pertenecientes al intervalo [0,1] para diferentes K', por lo tanto al normalizarse se obtiene:

$$M_{K'} = \frac{1}{K'} N_{K'}$$

Finalmente se hace un ajuste para superposiciones aleatorias de esta forma:

$$M_{K'}^{adj} = M_{K'} - \frac{K'}{N-1}.$$

2.2.7.6 Rangos de error de la media relativa (MRREs)

John Lee y Michel Verleysen establecen una nueva medida de calidad llamada MRREs[49]. La cual se basa principalmente en T&C, sin embargo, T&C solo se centra en el error de vecindarios grandes, por otro lado, MRREs también se centra en los leves o pequeños, en ese orden de ideas MRREs se calcula siguiendo esa idea.

$$MRRE_{\boldsymbol{Y}\rightarrow\boldsymbol{X}}(K) \triangleq \frac{1}{C} \sum_{i=1}^{N} \sum_{j \in N_K(y(i))} \left| \frac{rank(\boldsymbol{X}, i, j) - rank(\boldsymbol{Y}, i, j)}{rank(\boldsymbol{Y}, i, j)} \right|$$

[48] CHEN, Lisha y BUJA, Andreas. Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis, 2006.

[49] LEE, John; VERLEYSEN Michel. Nonlinear dimensionality reduction, Springer Link, 2007.

$$MRRE_{\boldsymbol{X}\to\boldsymbol{Y}}(K) \triangleq \frac{1}{C}\sum_{i=1}^{N}\sum_{j\in N_K(x(i))}\left|\frac{rank(\boldsymbol{X},i,j)-rank(\boldsymbol{Y},i,j)}{rank(\boldsymbol{X},i,j)}\right|$$

$N_K(x(i))$ Son los k-ésimos vecinos de $x(i)$ y el factor de normalización es

$$C = N\sum_{K=1}^{K}\frac{|2k-N-1|}{k}$$

2.2.7.7 Framework unificado

En secciones anteriores se ha hablado sobre la matriz de co-ranking así como de algunas medidas de calidad existentes, Lee y Verlysen[50]. Proponen un framework unificado a partir de dichas medidas, las medidas anteriormente mencionadas en secciones anteriores pueden ser definidas en términos de la matriz de co-ranking, de esa forma T&C puede definirse como:

$$M_T(K) = 1-\frac{2}{G_K}\sum_{i=1}^{N}\sum_{j\in n_i^k\setminus v_i^k}(p_{ij}-K) = 1-\frac{2}{G_K}\sum_{(k,l)\in\mathbb{LL}_K}(k-K)q_{kl},$$

$$M_C(K) = 1-\frac{2}{G_K}\sum_{i=1}^{N}\sum_{j\in n_i^k\setminus v_i^k}(r_{ij}-K) = 1-\frac{2}{G_K}\sum_{(k,l)\in\mathbb{UR}_K}(l-K)q_{kl},$$

Donde el vector de normalización es

$$G_K = \begin{cases} NK(2N-3K-1) & si\ K < N/2 \\ N(N-K)(N-K-1) & si\ K \geq N/2 \end{cases}$$

De esta forma se calculan ambos errores de T&C teniendo en cuenta la matriz de co-ranking, en donde el error de los intrusos disminuye la integridad y el error de las extrusiones disminuye la continuidad, de la misma forma MRREs se define como:

$$W_n(K) = \frac{1}{H_K}\sum_{i=1}^{N}\sum_{j\in n_i^k}\frac{|p_{ij}-K|}{p_{ij}} = \frac{1}{H_K}\sum_{(k,l)\in\mathbb{UL}_K\cup\mathbb{LL}_K}\frac{|k-l|}{l}q_{kl},$$

$$W_v(K) = \frac{1}{H_K}\sum_{i=1}^{N}\sum_{j\in v_i^k}\frac{|p_{ij}-r_{ij}|}{r_{ij}} = \frac{1}{H_K}\sum_{(k,l)\in\mathbb{UL}_K\cup\mathbb{UR}_K}\frac{|k-l|}{k}q_{kl},$$

Donde el vector de normalización es

[50] LEE, John; VERLEYSEN, Michel. Op. Cit.

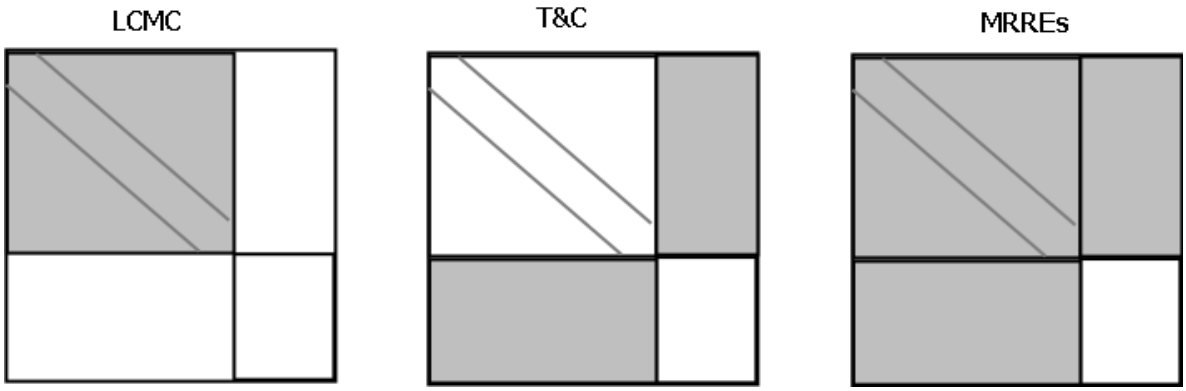
$$H_K = N \sum_{K=1}^{K} \frac{|N - 2k + 1|}{k}$$

De esta forma se calcula el error de todas las k-intrusiones y k-extrusiones sean duras o leves, y por ultimo LCMC se define como

$$U_{LC}(K) = \frac{1}{NK} \sum_{i=1}^{N} \left(\left| n_i^k \cap v_i^k \right| - \frac{K^2}{N-1} \right) = \frac{K}{1-N} + \frac{1}{NK} \sum_{(k,l) \in \mathbb{UL}_K} q_{kl},$$

De una forma simple T&C y MRREs, detectan lo que está mal en un incrustamiento dado, por otro lado, LCMC detecta lo que está bien en dicho incrustamiento, sin embargo, LCMC no puede definir el rendimiento por completo de un método RD.

Figura 11: Criterios de calidad teniendo en cuenta el esquema de la matriz de co-ranking.



Fuente: Elaboración propia

Estas métricas pueden definirse con el concepto de precisión y recuperación (P&R), la precisión es la proporción de elementos relevantes entre los que son recuperados, por otro lado, la recuperación es la proporción de elementos recuperados entre los relevantes, los elementos relevantes son aquellos que provienen del conjunto de vecinos de un punto en alta dimensión, mientras que los recuperados provienen del conjunto de vecinos de un dato en baja dimensión, P&R está relacionado al concepto de falso positivo y falso negativo, en donde los falsos positivos disminuyen la precisión y los falsos negativos la recuperación.

T&C se centra en los falsos positivos y falsos negativos, MRREs los positivos (falsos y verdaderos) y los negativos (falsos y verdaderos) mientras que LCMC se enfoca en los verdaderos positivos como se muestra en la Figura 11. n_i^k Y v_i^k tienen el mismo tamaño ya que cuando un elemento de v_i^k no se encuentra en n_i^k (falso negativo), dicho elemento es reemplazado con un vecino incorrecto (falso positivo), la cantidad de vecindarios es directamente proporcional al número de registros N, ya que para cada registro se crea un

vecindario, de esta forma se tiene

$$\sum_{(k,l)\in \mathbb{UL}_K \cup \mathbb{LL}_K} q_{kl} = \sum_{(k,l)\in \mathbb{UL}_K \cup \mathbb{UR}_K} q_{kl} = KN$$

Y

$$\sum_{(k,l)\in \mathbb{LL}_K} q_{kl} = \sum_{(k,l)\in \mathbb{UR}_K} q_{kl}$$

De tal forma que el número de las k-intrusiones y k-extrusiones duras son equivalentes, y por lo tanto $M_T(K) = M_C(K)$ y $W_n(K) = W_v(K)$, por otro lado, el peso ponderado que toma todas las K-intrusiones y K-extrusiones se escribe como

$$W_N^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in \mathbb{LT}_K \cup \mathbb{LL}_K} \frac{(k-l)^v}{k^w} \; q_{kl}$$

$$W_X^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in \mathbb{UT}_K \cup \mathbb{UR}_K} \frac{(l-k)^v}{l^w} \; q_{kl}$$

Dónde:

$$C_K = N \sum_{K=1}^{K} \frac{\max\{0, N-2k\}^w}{k^v}$$

Los exponentes w y v son enteros y se definen de manera tal que $v \geq w \geq 0$ y se pueden ajustar para definir diferentes longitudes de rango, $si\ v = 1$ y $w = 1$ da el mismo peso de MRREs, por otro lado si $v = 1$ y $w = 0$ se obtiene un peso similar a T&C, los criterios anteriores, se encuentran en una posición entre T&C y MRREs y permiten abarcar más elementos que los anteriores pero menos que los más recientes, de tal forma que para obtener un criterio de calidad mucho más completo se puede definir:

$$Q_{wNX}^{v,w} = 1 - \frac{W_N^{v,w}(K) + W_X^{v,w}(K)}{2}$$

Aquí se mide la calidad, la cual incrementa, si el número de intrusiones y extrusiones disminuye, además otra medida es añadida para ver la calidad general.

$$B_{wNX}^{v,w} = W_N^{v,w}(K) - W_X^{v,w}(K)$$

La cual define el comportamiento del método RD, es decir si $B_{wNX}^{v,w} < 0$ hay un comportamiento de extrusión, por el contrario, si $B_{wNX}^{v,w} > 0$, hay un comportamiento intrusivo, sin embargo, estas medidas pueden redefinirse si nos basamos en la idea de LCMC, de tal forma que se escribiría como:

$$U_N(K) = \frac{1}{KN} \sum_{(k,l)\epsilon \mathbb{UT}_k} qkl\,, \qquad U_X(K) = \frac{1}{KN} \sum_{(k,l)\epsilon \mathbb{LT}_k} qkl$$

Y

$$U_P(K) = \frac{1}{KN} \sum_{(k,l)\epsilon \mathbb{D}_k} qkl$$

Las dos primeras cantidades son las fracciones de los K-intrusos y K-extrusiones leves y la última es la fracción de vectores que se mantuvieron iguales en n_i^k y v_i^k, si estas tres cantidades se suman se obtiene una fracción muy relacionada a LCMC que se escribe:

$$Q_{NX}(K) = U_P(K) + U_N(K) + U_X(K) = U_{LC}(K) + \frac{K}{N-1}$$

La anterior curva finalmente puede ser escrita como el promedio de la intersección de los k-vecinos en alta dimensión con los k-vecinos en baja dimensión para cada registro:

$$Q_{NX}(K) = \sum_{i=1}^{N} \frac{\left|v_i^k \cap n_i^k\right|}{KN}$$

Y el comportamiento queda definido tal que:

$$B_{NX}(K) = U_N(K) - U_X(K)$$

2.2.7.8 Métricas R_{NX}

A partir de los conceptos anteriormente mencionados y explicados, se formuló R_{NX} que es la métrica más comúnmente utilizada por la comunidad científica y académica para evaluar métodos RD, según Lee, Renard y otros.[51] Ésta métrica combina $Q_{NX}(K)$ y $B_{NX}(K)$, $R_{NX}(K)$ es un ajuste de $Q_{NX}(K)$ con el fin de obtener una representación más intuitiva de la calidad de los métodos RD ya que el área bajo la curva se convierte en un buen indicador, la curva está dada por:

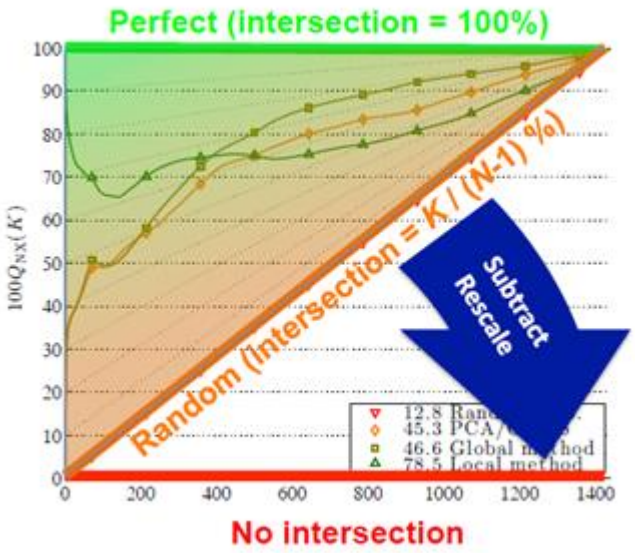
$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N - 1 - K}$$

Así mismo se añade el promedio de $100B_{NX}(K)$ para obtener que error predomina más en la incrustación, la

Figura 12 muestra la re-escala hecha a $Q_{NX}(K)$

[51] LEE, John, RENARD, Emilie, et al. Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. 2013

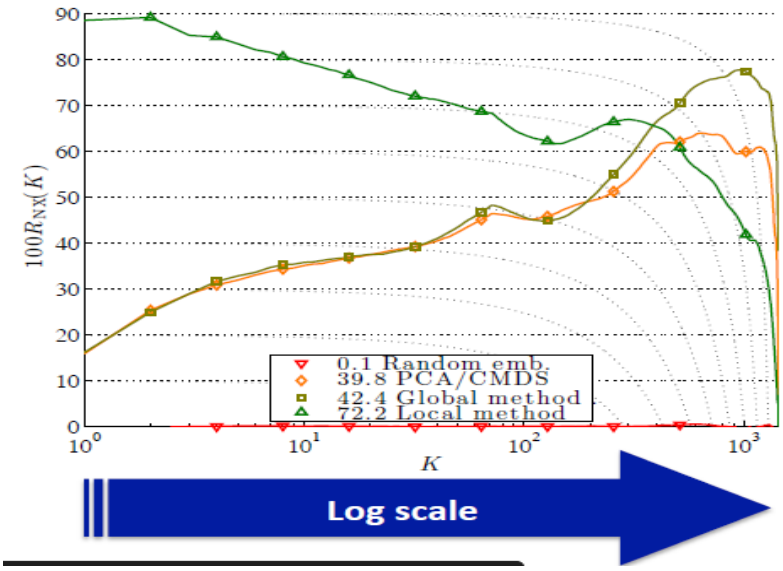
Figura 12: Calidad relativa entre una incrustación perfecta y aleatoria.



Fuente: LEE, John. Nonlinear Dimensionality Reduction: Towards better scalability, 2016.

Como se puede observar $Q_{NX}(K)$ no contempla toda el área bajo la curva y el ajuste de $R_{NX}(K)$ si lo hace y por lo tanto se obtiene una mejor interpretación de la preservación topológica, así mismo $R_{NX}(K)$ se convierte a una escala logarítmica con el fin de obtener escalas más grandes de K como se muestra en la siguiente figura.

Figura 13: Curvas RNX.



Fuente: LEE, John. Ibid.

2.2.8 Software Python

La herramienta será construida con el lenguaje de programación Python que tal como lo afirman Pérez, Yanet y Roberto.[52] Python es un lenguaje de programación interpretado a diferencia de otros lenguajes compilados como Java, C ++ y otros, fue creado por Guido Van Rossum y fue lanzado por primera vez en 1991. Python se basa en una filosofía en donde su escritura es fácilmente leíble, su estructura y orientación a objetos les permite a los programadores escribir código más limpio, el cual se adapta tanto a proyectos pequeños como grandes.

Tiene diferentes paradigmas de programación, siendo los más importantes el paradigma de programación funcional y el de orientado a objetos, Python ha sido definido como "batería incluidas" debido a su amplia librería estándar, en la que se encuentran un gran número de librerías que abarcan una gran variedad de áreas.

Así mismo Oliphant dice que[53] uno de los aspectos importantes de Python es su impacto en la computación científica, muchos ingenieros y científicos han expuesto su preferencia por lenguajes interpretados, ya que les permiten escribir programas computacionales sin profundizar mucho en la sintaxis y retrasos en la compilación de los programas, Python cumple exactamente esta función, siendo un lenguaje que ha contribuido mucho en áreas como la ciencia y la ingeniería, además gracias a su licencia abierta, tiene una gran comunidad desarrolladora que trabaja día a día para aumentar sus capacidades, por ejemplo, la comunidad ha desarrollado librerías como Numpy, Pandas, Matplotlib, Sklearn, etc. Las cuales se usan a diario en esta área.

2.2.9 Herramientas KDD

Las diversas herramientas KDD disponibles en el mercado permiten a sus usuarios realizar procesos de análisis de datos complejos, teniendo como base metodológica los procesos KDD, a continuación, se mencionan algunas de las herramientas más reconocidas en el ámbito del análisis de datos o minería de datos:

- **KNIME, (Konstanz information Miner)**

KNIME[54] se establece como una de las herramientas de código abierto más completas del mercado ya que sus posibilidades en cuanto a manejo de datos son ilimitadas dentro de las capacidades de la herramienta. KNIME está escrito en Java y basado en eclipse, publicado bajo licenciamiento

[52] Pérez, Ivet; Ricardo, Yanet y Becerra, Roberto. El lenguaje de programación python, Cuba, 2014.
[53] Oliphant, Travis. Python for scientific computing, IEEE, 2007.
[54] KNIME. Historia de código abierto [Sitio web] Universidad de Konstanz, Alemania; [Consultado: 19 de octubre de 2021]. Disponible en: https://www.knime.com/knime-open-source-story

GPLv3 de código abierto la cual cuenta con excepciones para poder usar la API de nodo para agregar extensiones propias de cada usuario. Integra ampliamente conceptos de aprendizaje automático y minería de datos, además que su interfaz gráfica de usuario permite añadir rápida y fácilmente los diferentes nodos que la componen, y gracias a su filosofía de código abierto, hace que cada vez se incluya nuevos componentes y nodos a la medida.

- **WEKA (*Waikato Environment for Knowledge Analysis)***

WEKA [55] es una herramienta de software libre desarrollada en el departamento de ciencias de la computación de la universidad de Waikato en Hamilton, Nueva Zelanda, creada para la realización de minería de datos. Weka abarca una gran parte del proceso KDD y así como muchas herramientas que incluyen este proceso, contiene ciertas características como clustering, asociación y visualización de resultados. Al incluir gran variedad de algoritmos para diferentes aspectos de la minería de datos como reglas de inducción, conjuntos difusos, arboles de decisión entre otros, requiere que sus funcionalidades sean aplicadas y visualizadas por usuarios en general, pero con un cierto grado de conocimiento sobre el tema.

- **ORANGE**

ORANGE[56] es una herramienta desarrollada en la universidad de Ljubljana, en Eslovenia. Es un software libre de aprendizaje automático y MD. Una de sus características principales es la visualización de resultados, donde esta permite generar flujos de trabajo interactivos acorde a las necesidades de cada empresa, además su visualización es diversa, es decir que se puede recurrir a la observación de diagramas de dispersión, graficas de barras, arboles de decisión, redes o mapa de calor esto hace de Orange una de las herramientas con licencias publica general (GPL) mejor posicionadas en el mercado de la MD que además cuenta con un proceso interactivo de visualización de datos.

2.2.10 Metodología en cascada

Según Cervantes Ojeda [57] Las actividades fundamentales del proceso de desarrollo de software se llevan a cabo como fases separadas y consecutivas.

---

[55] WEKA. Weka 3: Machine Learning Software in Java [Sitio web]. Nueva Zelanda; [Consultado: 19 de octubre de 2021]. Disponible en: https://www.cs.waikato.ac.nz/ml/weka/

[56] Orange. License [Sitio web]. Eslovania; [Consultado: 20 de octubre de 2021]. Disponible en: https://orangedatamining.com/

[57] Cervantes Ojeda, J. Taxonomía de los modelos y metodologías de desarrollo de software más utilizados, 2012, pp. 37-47

Estas actividades son: especificación (análisis y definición de requerimientos), implantación (diseño, codificación, validación) y mantenimiento. La metodología en cascada, se caracteriza por seguir un procedimiento lineal, es decir que los procesos de desarrollo se realizaran mediante fases, contrario a los modelos iterativos, cada fase se ejecuta una sola vez y de los resultados que se obtengan, se genera una hipótesis de partida para la siguiente fase. El también llamado WATERFALL MODEL se utiliza especialmente en el desarrollo de software.

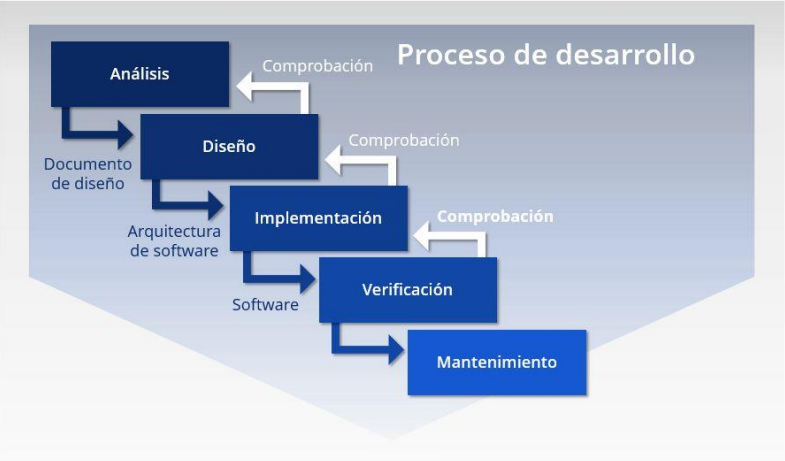
La versión original fue propuesta por Winston W. Royce en 1970 y posteriormente revisada por Barry Boehm en 1980 y Ian Sommerville en 1985[58], donde Royce propone las siguientes fases en el modelo.

- **Requisitos de sistema**
- **Requisitos de software**
- **Análisis**
- **Diseño**
- **Implementación**
- **Prueba**
- **Servicio**

Aunque en la práctica hay diversas versiones del modelo, ya que habitualmente las 7 fases propuestas por Royce, se simplifican a cinco, ya que las fases 1, 2 y 3 se incluyen en la fase de análisis.

En el modelo en cascada se siguen cinco fases lineales tal como se muestra en la Figura 14.

Figura 14: Ciclo de desarrollo. Modelo en cascada.



Fuente: Rojas y Boucchechter (2005)

[58] Cataldi, Z., Lage, F., Pessacq, R. y García Martínez, R. Ingeniería de software educativo. 2013

Cada uno de esos procesos tiene un objetivo específico, siguiendo lo descrito por Rojas y Boucchechter (2005)[59] estos son:

- **Análisis**

  Se realiza una planificación inicial, especificación de requisitos o requerimientos del sistema y se realiza el análisis de la información recolectada, es en esta etapa donde se incluyen estudios de viabilidad, donde se evalúan costes del proyecto, factibilidad y rentabilidad. Al generar una descripción completa de los requerimientos del sistema, también se incluyen las condiciones para su respectiva aceptación.

- **Diseño**

  Se realiza la especificación de los requerimientos del sistema y se traza el diseño del mismo, aquí se formula una solución especifica dependiente del análisis y evaluación de los requerimientos incluidos y aceptados en la fase anterior, posteriormente se incluye la arquitectura de software, generando un plan detallado del diseño del producto, como por ejemplo interfaces, bibliotecas o entornos de trabajo. Con todos estos componentes se puede obtener un borrador y plan de prueba inicial.

- **Implementación**

  Se contempla todo lo referente al desarrollo y programación y se realizan pruebas unitarias, es aquí donde toda la arquitectura de software realizada en la fase anterior, se dispone a traducirse en el respectivo lenguaje de programación de software, se realizan las pruebas unitarias por cada momento de desarrollo, donde los componentes de podrán desarrollar por separado y estos se integran parcialmente al producto general, cabe resaltar que es indispensable tener un producto funcional en esta fase para proseguir a la siguiente.

- **Verificación**

  Se realiza integración del sistema con sus respectivas pruebas, por lo general, se envía una versión de prueba o beta al usuario final, y es aquí donde las validaciones de los requerimientos iniciales incluidos en la fase de análisis se evalúan para determinar si el producto cumple con las exigencias contempladas al principio, si todo se cumple con normalidad se podría dar un lanzamiento más formal.

- **Mantenimiento**

  Entrega final, y recomendaciones para su mantenimiento y mejora.

[59] Rojas, R y Boucchechter, I. Ingeniería del Software. Ciclo de vida, Mallorca, España. (2005).

Ventajas del modelo en cascada

- Sus fases del proyecto están claramente definidas y contienen una estructura sencilla.
- La estimación de costes y carga en cada actividad se pueden generar al inicio del proyecto.
- Se puede presentar de forma cronológica de manera sencilla.

Desventajas del modelo en cascada

- En proyectos de mayor complejidad no permite evidenciar claramente las fases
- El margen para realizar ajustes se reduce si surge un cambio de requerimientos.
- Las fallas se detectan al finalizar el proceso de desarrollo y no durante.

2.3 Variables de investigación

La investigación realizada en este proyecto, hizo una investigación básica sobre los métodos de RD y no se realiza una investigación aplicada sobre un conjunto de datos en particular, dado que la naturaleza de los métodos RD permiten procesar datos que pueden provenir de diversas fuentes, tanto sintéticas como lo son el rollo suizo, la esfera y el toroide, o datos dispersos reales como MNIST. Por tanto, las variables que se han considerado son intrínsecas para cada conjunto de datos, en el caso de los datos artificiales, las variables son las coordenadas posiciónales en X, Y, Z que indican la cercanía entre un punto y otro, además de la etiqueta de color que da el sentido de cercanía entre puntos, para MNIST las variables presentes son la composición de los pixeles representativos de una imagen, además de la etiqueta de correspondencia de un número. Todas estas variables son procesadas dentro de los métodos RD, generando un nuevo conjunto de datos en un espacio de baja dimensión con respecto a los datos originales de entrada.

Con lo anterior podemos extraer dos grupos de variables independientes, unas que se encuentran en el estado de alta dimensión y las otras en el estado de bajo dimensión, el hecho de comparar estos dos grupos de variables con las métricas RNX, produce como resultado un escalar comprendido en un rango de [0, 100] que permite evaluar la preservación topológica resultante siendo este la variable dependiente.

Variable independiente

- Estado de alta dimensión
    - Coordenadas en X
    - Coordenadas en Y
    - Coordenadas en Z
    - Pixeles representativos de la imagen

- Estado de baja dimensión
    - Coordenadas en X reducidas
    - Coordenadas en Y reducidas
    - Pixeles representativos de la imagen

Variable dependiente

- Calidad de la incrustación obtenida con RNX

2.4 Definición nominal de las variables

Variables de alta y baja dimensión

En un conjunto de datos de 3 dimensiones como la esfera, las coordenadas en X, Y, Z, son variables cuantitativas de intervalo, lo que quiere decir que son valores numéricos y pueden ser tanto negativos como positivos, por otro lado, el valor de la etiqueta es una variable cualitativa ordinal puesto que es el color que va a tener el punto y esto determina el lugar en donde este se encuentra.

En conjuntos de datos como lo es MNIST, los pixeles son variables cuantitativas de razón debido a que un pixel no puede tener valores menores que cero, siendo este absoluto, de igual forma, el valor de la etiqueta es una variable cualitativa nominal la cual representa el número (0, 1, 2, 3, 4, 5, 6, 7, 8 o 9).

Calidad de la incrustación obtenida con RNX

RNX es la combinación de varias métricas que permiten evaluar los métodos RD según Lee, Renard y otros [60] Éstas permiten calcular la proporción de vecindarios conservados, además de los errores de intrusiones y extrusiones, para al final obtener la calidad general del método RD evaluado, el valor de la calidad obtenida, es cuantitativo y de razón debido a que el cero es absoluto.

2.5 Definición operativa de las variables

Variables de alta y baja dimensión

Las variables de alta y baja dimensión pueden ser operadas dependiendo del método RD y dichas operaciones pueden ser globales o locales, en este caso, los métodos son los que transforman las variables, por ejemplo, PCA según Fodor [61], realiza una operación global mediante combinaciones lineales con la varianza más grande y son llamadas Componentes Principales, los cuales son ortogonales buscando representar la mayor cantidad de información del

[60] LEE, John, RENARD, Emilie, et al. Op. Cit
[61] FODOR, Imola. Op. Cit.

conjunto de datos en unos pocos Componentes Principales. Contrario a las operaciones globales, se encuentran las locales, en donde métodos como LLE según Vanderplas y Connody [62], mapea las variables a un espacio de menor dimensión, con el fin de que el mapeo conserve la mayor información posible, se calculan los vecinos más cercanos en los datos de alta dimensión además de los pesos para cada vecino de los datos, con el fin de determinar el vector de proyección en baja dimensión.

Teniendo en cuenta lo expuesto anteriormente, la operatividad de las variables depende de los métodos RD, siendo estos las funciones por las que pasan los datos para su proceso de transformación.

Calidad de la incrustación obtenida con RNX

La calidad general de la incrustación generada por los métodos RD, se obtiene mediante el cálculo realizado por las métricas RNX al comparar las variables en el estado de alta dimensión, con las de baja dimensión, para obtener el resultado de la calidad de la incrustación es necesario obtener QNX, el cual según Lee y Verlysen.[63] Es la sumatoria de la intersección de los vecinos conservados en alta y baja dimensión, entre el producto de todos los vecindarios del sistema, es importante mencionar que Q_{NX} ya calcula los errores de intrusiones y extrusiones internamente.

$$Q_{NX} = \sum_{i=1}^{N} \frac{\left| v_i^k \cap n_i^k \right|}{KN}$$

El cual varía de 0 a 1 en donde v_i^k son los k-vecinos en alta dimensión y n_i^k los k-vecinos en baja dimensión, 0 significa que no hay ningún tipo de preservación topológica y 1 que la preservación ha sido perfecta por parte del método, para obtener R_{NX} segun Lee, Renard y otros[64] Q_{NX} es normalizada con el fin de poder comparar diferentes métodos en diferentes escalas de la siguiente forma:

$$R_{NX} = \frac{(N-1)Q_{NX} - K}{N - 1 - K}$$

Además R_{NX} se multiplica por $100/(N-1)$ que es el promedio de $B_{NX}(K)$ para obtener el comportamiento, obteniendo como resultado $100R_{NX}$, para obtener un escalar de 0 a 100, siendo esta la calidad de la incrustación que se quiere conocer.

2.6 Formulación de hipótesis

2.6.1 Hipótesis de investigación

[62] VANDERPLAS, Jake y CONNOLLY, Andrew. Op. Cit.
[63] LEE, John y VERLEYSEN, Michel. Op. Cit
[64] LEE, John, RENARD, Emilie, et al. Op. Cit

Hi: La implementación de las curvas R_{NX} propuesta, evalúa los métodos RD satisfactoriamente, permitiendo conocer cuál es la calidad y el comportamiento de los métodos en situaciones específicas, así como la conservación topológica de los diversos conjuntos de datos.

2.6.2 Hipótesis nula

Ho: La implementación de las curvas R_{NX} propuesta es inadecuada, obteniendo evaluaciones de los métodos imprecisas y que no estaban previstas.

2.6.3 Hipótesis alterna

Ha: La implementación de las curvas R_{NX} solo evalúa satisfactoriamente los métodos RD con un comportamiento no lineal.

3. METODOLOGÍA

3.1 Paradigma

Este proyecto es de tipo positivista porque tiene un enfoque metodológico predominante cuantitativo, en donde gracias al intervalo de calidad [0, 100] es posible determinar con que precisión, es preservada la topología de los datos en baja dimensión, para poder identificar que métodos RD deben ser usados en situaciones muy específicas.[65]

3.2 Enfoque

El enfoque de la investigación es cuantitativo puesto que se pretende realizar la medición de calidad de los métodos RD la cual es cuantificable y puede variar dependiendo de los datos y los ajustes de los parámetros.[66]

3.3 Método

El método de la investigación es científico, puesto que se basa en la búsqueda de la verdad mediante la evaluación de métodos RD, con el fin de determinar en qué datos estos tienen un mejor rendimiento y en el caso de los métodos parametrizados, que ajuste de los parámetros es el mejor para cada uno.[67]

3.4 Tipo de investigación

La investigación es de tipo descriptiva puesto que se buscó medir con la mayor precisión posible dos propiedades importantes de los métodos RD, como lo son la calidad de su incrustación y su comportamiento, es decir si conserva mejor la topología global o local de los datos por medio de las curvas R_{NX}.[68]

3.5 Diseño de investigación

El diseño de investigación es de experimentos verdaderos debido a que se realizaron distintos tipos de pruebas, como el análisis del rendimiento de la herramienta y la confianza de los rendimientos obtenidos por los algoritmos RD.

Finalmente se realizó la documentación pertinente para el funcionamiento de la herramienta y la librería que se desarrollaron, con el fin de que la comunidad científica y académica tenga un entendimiento claro y conciso del uso de la herramienta.

3.6 Población

La población son un grupo de personas con características comunes, las

---

[65] QUIJANO, Armando. Guía de Investigación Cuantitativa. Institución Universitaria CESMAG. Parte 1. Pag 76.
[66] Ibid. p. 76
[67] Ibid. p. 76
[68] Ibid. p. 77

cuales se toman como el foco principal de un proyecto, esta investigación se centró en la parte epistemológica y matemática del objeto de estudio, que en este caso son los métodos RD y por lo tanto no se requirió población ni muestra, esto se debe a que la naturaleza del proyecto exige un tipo de investigación de **Data Driven Approach,** en el cual los datos son uno de los pilares del proyecto, según Tinoco y otros.[69] Los investigadores suelen usar diferentes datasets con el fin de probar sus algoritmos o métodos, sean estos de machine learning, deep learning, etc. Para ello recurren a Big data o repositorios de datos creados por investigadores para diferentes propósitos y esta investigación no fue diferente, puesto que los métodos RD evaluados, fueron probados en diferentes conjuntos de datos mencionados en la siguiente sección, es decir, en este proyecto no se utilizaron datos focales, sino que se hizo uso de diferentes bases de datos ya pre-establecidas.

3.7 Técnicas de recolección de la información

Áreas de la inteligencia computacional inmersas en el Data Driven Aproach tales como Big-Data y Machine Learning parten de datos previamente capturados. Este proyecto al estar dentro del marco de Data Driven Aproach no requirió de una técnica específica para la recolección de la información, puesto que utilizó conjuntos de datos previamente capturados en anteriores investigaciones, dichos conjuntos han sido utilizados y avalados previamente por la comunidad académica y científica, y, particularmente en reducción de dimensión, son utilizados para probar múltiples algoritmos fundamentados en diversas heurísticas, ya que al tener una estructura geométrica como la esfera (Figura 17), rollo suizo (Figura 18), y toroide (Figura 19), las cuales han sido generadas mediante una formulación matemática de manera sintética para generar colectores de estructuras de fácil compresión, es posible realizar inferencias directas sobre el resultado del incrustamiento, ya que dichas estructuras son cercanas al campo de percepción humana y permiten comprender el comportamiento del método aplicado, es decir, si logra la preservación de la topología global o local. En el proceso de experimentación, se ha hecho evidente la importancia del uso de datos pre establecidos, ya que se permite una estandarización en los resultados y por tanto comparar los métodos de manera justa. Algunos de los autores que avalan estos conjuntos de datos son Lee y otros, que en su trabajo[70] realizan experimentos con los conjuntos de datos de esfera, hélice, Fray Faces entre otros, dichos experimentos permiten probar, validar y evaluar el funcionamiento de los diferentes métodos RD, los cuales son implementados en diversos trabajos, gracias a su eficiencia en el proceso de reducción de dimensión, así mismo, Peluffo Diego y otros[71] mencionan los diferentes métodos matemáticos con los que son construidos los métodos RD, por ejemplo, mientras que Kernel PCA maximiza la varianza de los datos en alta dimensión representados por un kernel, LE es el pseudo-inverso del grafo laplaciano, obteniendo así,

[69] TINOCO, et al. A data-driven approach to develop physically soundpredictors: Application to depth-averaged velocities on flowsthrough submerged arrays of rigid cylinders, 2015.
[70] LEE, jhon; et al. Op. Cit., p.27
[71] PELUFFO, Diego; et al. Multiple Kernel Learning for Spectral Dimensionality Reduction, Colombia, 2015

incrustaciones diferentes por los métodos RD en los distintos conjuntos de datos, y por lo tanto, su utilización se hace indispensable en este tipo de investigaciones.

Con lo anteriormente mencionado, en la presente investigación se experimentó con los conjuntos de datos de la esfera, el rollo suizo y el toroide, siendo estos conjuntos de datos artificiales y con MNIST, Fray Faces e Iris, siendo estos conjuntos de datos reales, aunque en la investigación no se realizó la aplicación de una técnica de recolección de información, los conjuntos de datos anteriormente mencionados si fueron construidos con diferentes técnicas por parte de los autores, las cuales serán mencionados en cada uno de ellos.

Conjunto de datos MNIST

MNIST es un banco de imágenes que ha ayudado a muchos investigadores en el área de Deep Learning o aprendizaje profundo al momento de probar sus algoritmos, fue realizada por Lecun y otros.[72] Esta base de datos está conformada por 6000 imágenes en escala de grises de los 10 dígitos (0, 1, 2, 3, 4, 5, 6, 7, 8, 9), este conjunto de datos puede encontrarse en diversos repositorios de datos como Kaggle o incluso en la página del autor, para crear el conjunto de datos Lecun y otros [73] le solicitaron a diversos usuarios escribir los números de 0 a 9 a mano alzada y para que el banco de imágenes quede con dimensiones estándar, los autores normalizaron el tamaño de cada imagen tomada y las ajustaron a dicho tamaño, como resultado se obtuvo que las imágenes fueron normalizadas en una caja de 20x20 pixeles y fueron centradas en una imagen de 28x28 pixeles. El conjunto de datos MNIST ha sido utilizado por diferentes autores para entrenar algoritmos de DL y ML como los K-vecinos más cercanos, SVM (Super Vector Machine), PCA, Redes Neuronales Convolucionales, entre otros.

[72] LECUN, Yann; et al. Gradient-Based learning applied to document recognition, 1998.
[73] LECUN, Yann; et al. THE MNIST DATABASE [en linea]. Of Handwritten digits. Nueva York; [Consultado: 20 de octubre de 2021]. Disponible en: http://yann.lecun.com/exdb/mnist/

Figura 15: Banco de imágenes MNIST.

Fuente: LECUN, Yann; et al. Gradient-Based learning applied to document recognition, 1998.

Conjunto de datos Frey Face

El conjunto de datos Frey Face puede encontrarse en el siguiente recurso,[74] la base de datos está compuesta de 2000 imágenes del rostro de frey en cuadros consecutivos de tamaño 20x28, para esto Fray grabó un corto video y tomó cada cuadro del video para convertirlo en una imagen.

Figura 16: Banco de imágenes Frey Face.

Fuente: Frey. Data for MATLAB hackers

[74] FREY. Data for matlab hackers.

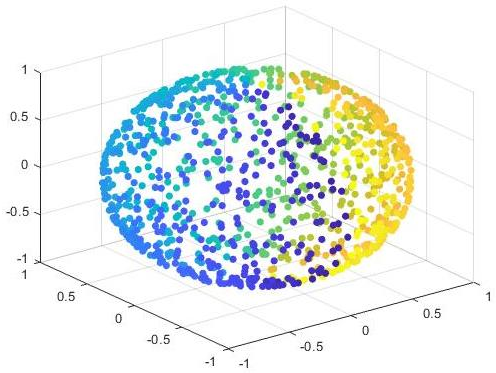
Conjunto de datos Iris

El conjunto de datos Iris puede encontrarse en el repositorio UCI y fue creado por Fisher[75] consta de 150 observaciones las cuales están coloreadas por especies (Iris setosa, versicolor e Iris virginica) de cada clase hay 50 imágenes con 3 características, para construir el data set fue necesario hacer la medición de la altura y ancho del sépalo y pétalo por cada especie de Iris.

Esfera

La esfera es un colector formado con 1000 puntos distribuidos en X, Y, Z

Figura 17: Conjunto de datos artificial: Esfera

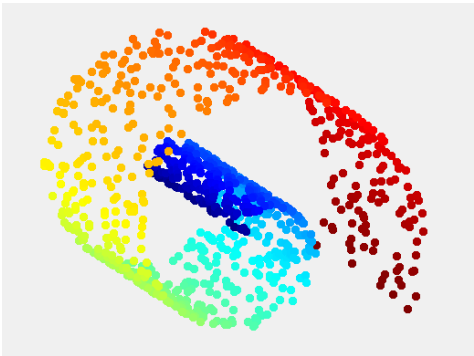


Fuente: Elaboración propia

Rollo suizo

El Rollo suizo está conformado por 1000 puntos dispuestos en X, Y y Z

Figura 18: Conjunto de datos artificial: Rollo suizo.



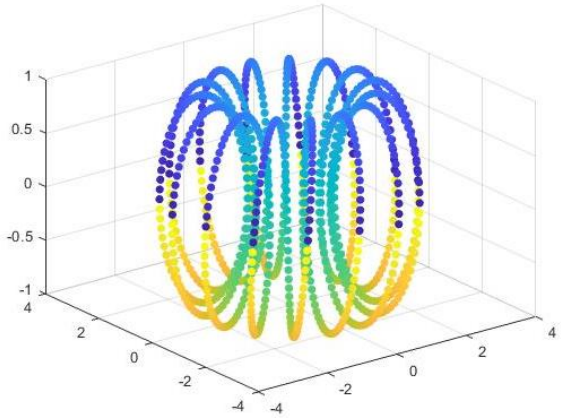
Fuente: Elaboración propia

[75] FISHER, R. Iris Data set. 1936.

Toroide

El Toroide está compuesto por 1000 puntos dispuestos en X, Y y Z

Figura 19: Conjunto de datos artificial: Toroide.



Fuente: Elaboración propia

4. RESULTADOS DE LA INVESTIGACION.

Cómo se ha mencionado en secciones anteriores el proceso KDD comprende una serie de pasos que le permite a los investigadores, obtener conocimiento a partir de uno o varios conjuntos de datos por medio de la aplicación de métodos RD, técnicas de limpieza de datos, algoritmos de MD entre otros, algunos trabajos mencionados en la sección **2.1** muestran la forma en la que los métodos RD pueden ser usados en distintas áreas del conocimiento como la educación, la salud y la psicología, por lo tanto, se desarrolló una herramienta que integra diferentes pasos del proceso KDD tal como las herramientas mencionadas en la sección **2.2.9**, siendo éstas la inspiración para el desarrollo de la herramienta con un modelo Drag and Drop. Gracias a la investigación realizada tanto de métodos RD como de métricas para su evaluación, fue posible identificar los requerimientos tanto funcionales como no funcionales de la herramienta, los cuales son introducidos en la sección **4.1.5**.

4.1.1 Análisis De Diversos Enfoques Y Heurísticas De Reducción De Dimensión

Teniendo en cuenta que el objetivo fue analizar diversos métodos de reducción de dimensión, fue necesario realizar una búsqueda en múltiples fuentes de la literatura científica que abordaron temas como la implementación de métodos RD, sus estudios comparativos y análisis, con lo cual, se obtuvo información pertinente de diversos enfoques de reducción de dimensión, esto con el fin de realizar una revisión de los métodos comúnmente utilizados por la comunidad científica. Por otro lado, el análisis realizado ayudó al grupo investigador a conocer los diversos enfoques y heurísticas documentados, comprendiendo los casos de uso de los métodos y la forma en como estos trabajan.

Con lo anterior mencionado, se encontró que a lo largo de los años diversos autores han propuesto métodos de reducción de dimensión basados en métodos estadísticos, como la maximización de la varianza realizada por PCA en la sección **2.2.3.1.1** y algunas distribuciones como la T-Student, realizada por T-SNE mencionado en los antecedentes de la investigación sección **2.1**, otros se basan en grafos, buscando preservar la topología local a partir del cálculo de vecindarios y pesos para cada dato como LLE, mencionado en **2.2.3.2.1**, algunos otros se basan en el plano Laplaciano como LE **2.2.3.2.3**, o en matrices kernel como Kernel PCA **2.2.3.2.2**, entre otros.

4.1.2 Identificación E Integración De Los Métodos De Reducción De Dimensión Analizados

Una vez analizada la literatura, se logró identificar que, si bien hay una gran variedad de algoritmos de reducción de dimensión e incluso existen combinaciones entre los mismos, con el fin de aprovechar sus características combinadas, se tomaron métodos RD espectrales y basados en Kernel como PCA, MDS, LLE, LE, ISOMAP y Kernel PCA, de los cuales se pueden observar sus características de funcionalidad en la **Tabla 1**, definiendo así, los 6 métodos de reducción de dimensión que fueron integrados en la herramienta, teniendo en cuenta los aspectos de cada uno de ellos como su naturaleza, calidad de conservación topológica frente a otros modelos, además de la taxonomía principal dividida entre los métodos lineales que realizan una preservación global, los métodos no lineales que tienen un mejor rendimiento en la preservación local, los paramétricos que requieren el parámetro de vecindarios (*K)*, o no paramétricos los cuales no requieren parámetros especiales para su utilización. Los métodos RD seleccionados, son los más utilizados por los investigadores y, por lo tanto, fueron usados en el presente proyecto para generar en el producto final un resultado fiable y fácil de interpretar.

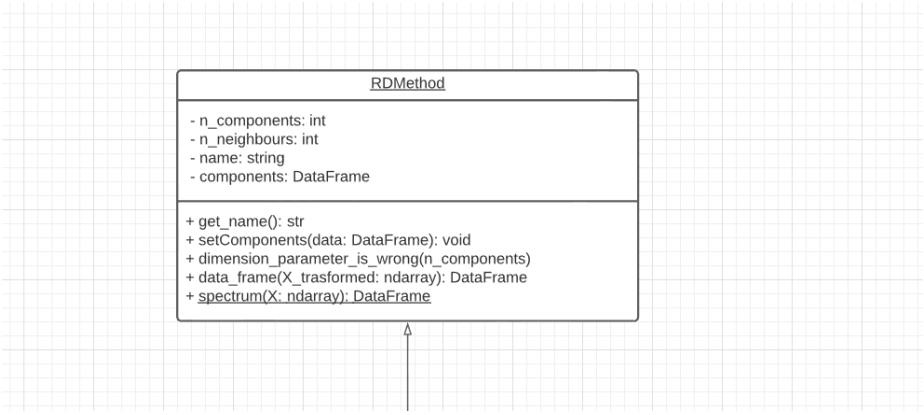
Tabla 1: Características de los métodos RD

No	Método	Local	Global	Lineal	No lineal	Paramétrico	No paramétrico
1	MDS		X	X			X
2	LE	X			X	X	
3	LLE	X			X	X	
4	PCA		X	X			X
5	KPCA		X		X	X	X
6	ISOMAP		X		X	X	

Fuente: Elaboración propia

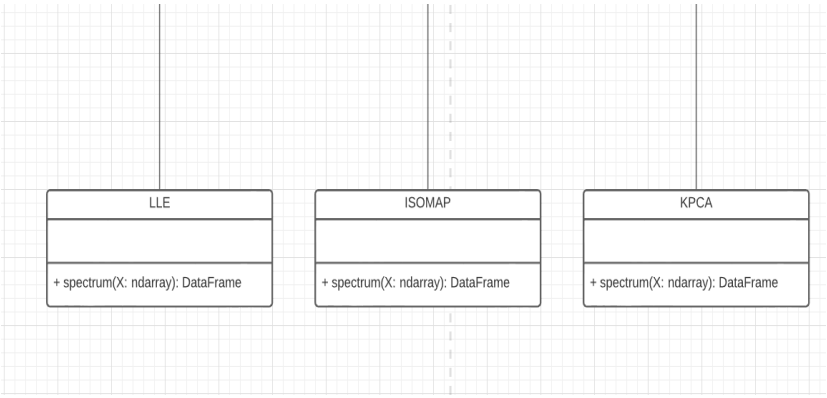
Cabe resaltar que la integración de los métodos RD en la herramienta, fue posible gracias a las implementaciones realizadas de los métodos en la librería de scikit-learn, por lo que en el presente proyecto se desarrollaron módulos que se adaptaron a los requerimientos de los métodos RD desarrollados en dicha librería. El desarrollo se realizó con Python, el cual es un lenguaje de programación con muchos paradigmas, uno de ellos es la POO, de esta forma se creó un esquema como se muestra desde la Figura 20 hasta la Figura 22, en donde el nodo padre abstracto **RDMethod** les hereda a todos los métodos RD las funcionalidades propias de cualquier método de reducción de dimensión, haciendo mucho más fácil el desarrollo y permitiendo escalabilidad en caso de que se quieran incluir más métodos RD a la herramienta.

Figura 20: Diagramas de clases métodos RD clase padre



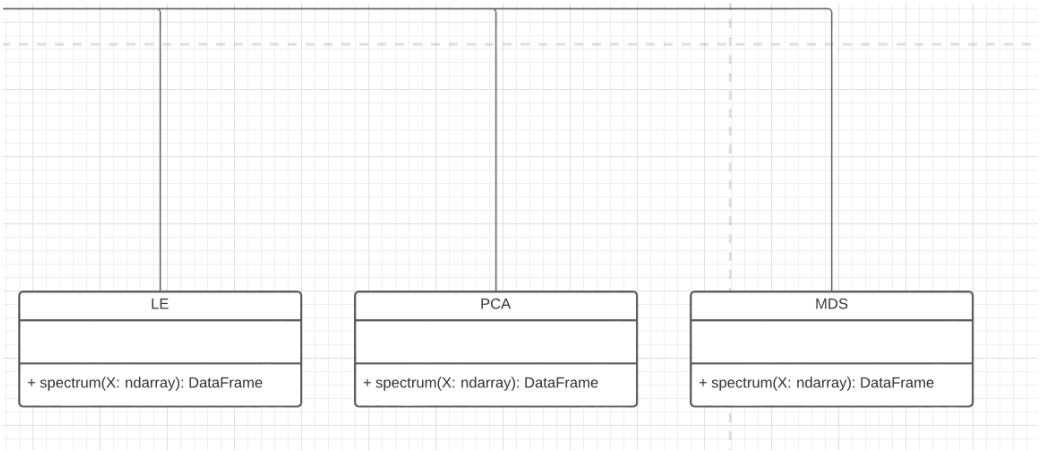
Fuente: Elaboración propia

Figura 21: Diagrama de clases de métodos RD clases hijas LLE, ISOMAP y KPCA



Fuente: Elaboración propia

Figura 22: Diagrama de clases métodos RD clases hijas LE, PCA y MDS



Fuente: Elaboración propia

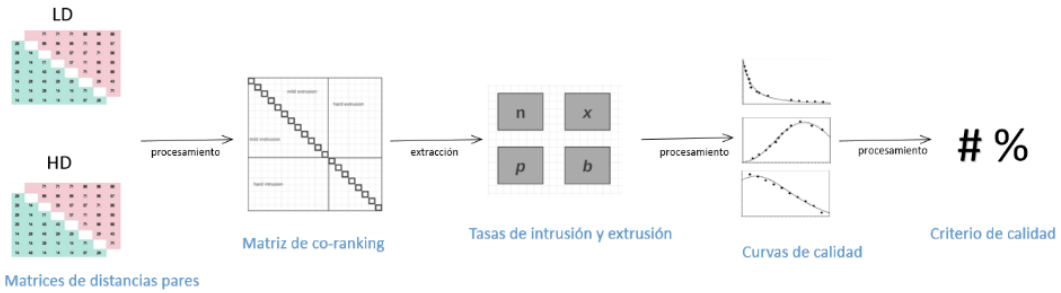
Los módulos desarrollados están pensados para soportar cualquier método RD, sean estos paramétricos o no paramétricos, en donde siempre se toman como inputs las dimensiones a las cuales se quiere reducir el conjunto de datos y los vecindarios (solo para los métodos paramétricos), y como output el conjunto de datos transformado con un tipado específico de DataFrame, el tipado es importante puesto que la idea desde un principio, fue que la herramienta tratara los inputs y outputs en un solo tipo de dato, definiendo pautas de desarrollo específicas que permitan la escalabilidad y legibilidad necesaria para el desarrollo. Finalmente, se realizó un polimorfismo en las clases hijas (métodos RD), siendo este, el método "**spectrum**" que se encuentra en todas las clases, pero dependiendo del objeto tiene diferente funcionalidad, aquí se hace presente la variedad entre los métodos paramétricos y no paramétricos. El código fuente de la integración de los métodos RD puede verse en el siguiente repositorio.
https://github.com/CarlosDCorrea/rnx_tool/tree/master/rnx_node_editor/dimensionality_reduction_methods

4.1.3 Librería Para La Evaluación De La Preservación Topológica De Métodos De Reducción De Dimensión Desarrollada

Las curvas de calidad RNX introducidas en la sección **2.2.7.8**, se implementaron de una forma modular (librería) para evaluar el desempeño de los métodos mediante una representación gráfica, donde se evidencia la preservación topológica. En el presente trabajo, se desarrolló tanto la librería de las métricas RNX como la herramienta para la preservación topológica que integra la librería desarrollada, con el fin de ofrecer a los usuarios un entorno para su fácil uso. Inicialmente, las métricas pueden ser difíciles de entender para personas con poco conocimiento matemático, por lo que en la Figura 27

se ofrece el flujo que sigue RNX desde el ingreso de los datos de entrada hasta su salida.

Figura 23: Representación gráfica del flujo de procesos de la herramienta.

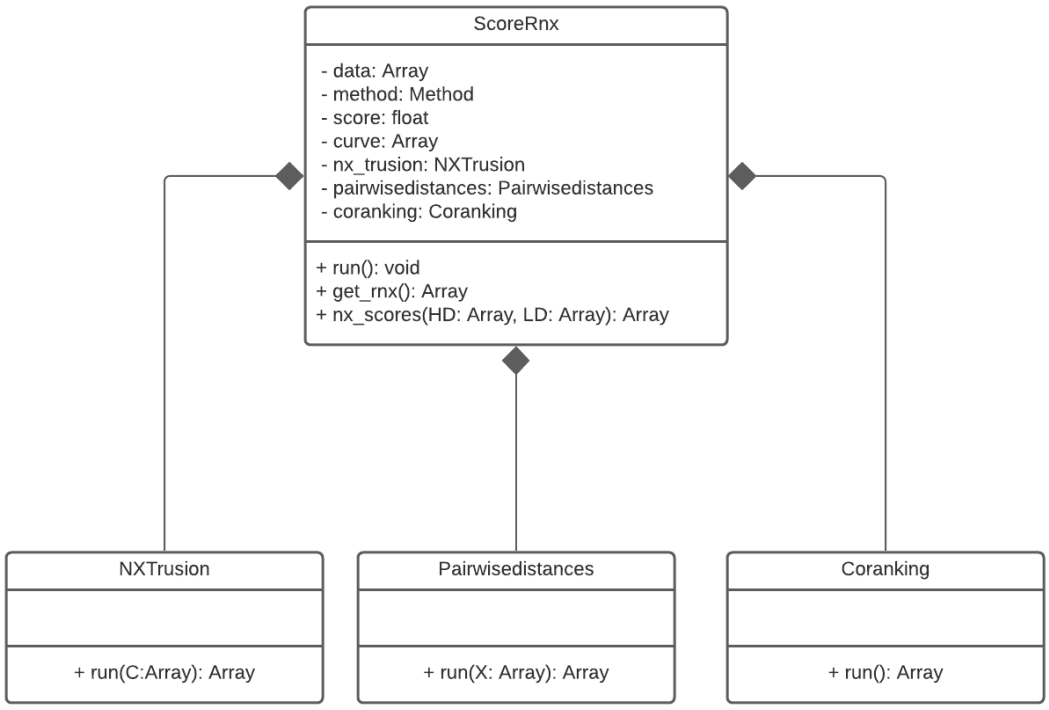


Fuente: Elaboración propia

El flujo consiste en que RNX toma como input los datos en baja y alta dimensión, a partir de ellos construye la matriz de co-ranking que a su vez retorna las tasas de intrusiones y extrusiones con las que se obtienen las curvas y un resultado medible y cuantificable, que representa la evaluación topológica de los datos.

Como se ha mencionado, la librería y la herramienta fueron desarrollos independientes, para poder desarrollar la librería, fue necesario implementar diferentes rutinas de programación utilizando métodos matemáticos implementados por librerías como NumPy, el cual ofrece módulos para operar matrices con un menor tiempo de complejidad (tiempo que un algoritmo tarda en ejecutarse) y que también es de fácil uso, en la Figura 24 se observan los módulos desarrollados de RNX

Figura 24: Diagrama de clases UML de la librería RNX



Fuente: Elaboración propia

El código fuente de la librería se encuentra en GitHub en el siguiente link https://github.com/DiegoUrrea195/score_rnx y además puesto que la librería se encuentra registrada en PyPI, puede ser instalada con el comando *pip install score-rnx* para su uso.

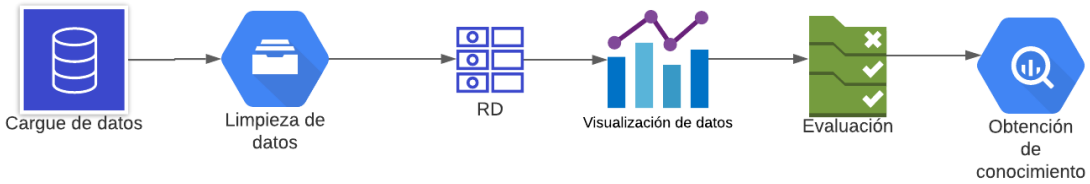
4.1.4 Herramienta Interactiva Para La Evaluación De La Preservación Topológica De Métodos De Reducción De Dimensión

Debido a que pueden haber usuarios que no tienen conocimiento en programación y que por lo tanto, no pueden usar la librería para realizar la evaluación de métodos RD, se desarrolló una herramienta que ofrece un entorno interactivo y fácil de usar para realizar la evaluación, la herramienta desarrollada se inspiró de la metodología y uso de algunos productos software mencionados en **2.2.9**, los cuales mediante una interfaz gráfica Drag and Drop disponen de diversos módulos de análisis y depuración de datos con los cuales, los usuarios pueden interactuar en un espacio de trabajo para desarrollar flujos secuenciales con los datos.

La herramienta se basa en el proceso KDD, el cual es un proceso secuencial y automático donde se mescla el descubrimiento y análisis, usado generalmente en campos científicos, académicos, comerciales, de salud, o

cualquier instancia que genere grandes cantidades de información, la cual adopta una serie de pasos a fin de lograr el conocimiento referente a un conjunto de datos inicial. Estos pasos se dividen en 9, sin embargo, la herramienta no hace uso de todos los pasos de KDD, en la Figura 25 se observan los pasos de KDD usados en la herramienta.

Figura 25 Pasos de KDD implementados por la herramienta



Fuente: Elaboración propia

Con el fin de que esta sección se entienda de la mejor forma posible, se utilizó la siguiente dinámica, debido a que la herramienta sigue el proceso KDD como se mencionó anteriormente, en donde se realiza un cargue, limpieza, procesamiento y evaluación de los datos, se decidió que por cada paso de KDD implementado en la herramienta se explicaría el proceso de la metodología en cascada realizado en dicho paso, por ende, para el paso de KDD de selección de datos se realizó un proceso de análisis, diseño y desarrollo, siendo este proceso el mismo para los otros pasos.

4.1.4.1 Abstracción del escenario

Generalmente, los proyectos orientados al análisis de datos tienen un enfoque dependiendo del área en el que se esté realizando, sea esta área educativa, de salud, industrial, entre otras, por lo que es necesario definir todos los elementos que harán parte de la investigación, así como la delimitación y objetivos iniciales referente al escenario que serán objeto de estudio, la presente investigación no tiene un escenario definido, puesto que el objeto de estudio no es una comunidad o población específica sino que son los propios métodos RD.

4.1.4.2 Selección de datos

La herramienta implementa nodos para el cargue de los datos, los cuales pueden ser reales o artificiales y por lo tanto se requieren dos nodos distintos para cada caso, estos nodos les permiten a los usuarios cargar datos alojados en la herramienta, o ingresar a su gestor de archivos para cargar los datos que ellos deseen. Las extensiones de archivos soportadas son .csv, .xlsx, mat.

- **Análisis: Requerimientos funcionales para el cargue de datos.**

Tabla 2. Requerimiento funcional RF01.

código del requerimiento	RF01
Nombre	Nodo datos artificiales
Propósito	Cargar los data set artificiales
Descripción	El sistema dispondrá de un nodo que incluya los data sets artificiales seleccionados previamente.
Entrada	Data set de pruebas incluidos en la herramienta
Salida	Diferentes nodos según la actividad a realizar
Prioridad	Alta

Fuente: Elaboración propia

Tabla 3. Requerimiento funcional RF02.

código del requerimiento	RF02
Nombre	Selección
Propósito	Cargar los data set seleccionados por el usuario
descripción	El sistema deberá soportar una lista desplegable de selección de los diferentes data set precargados en la herramienta (Toroide, Swisroll, Sphere)
Entrada	Seleccionar el data set de la lista
Salida	En espera para ejecución
Prioridad	Alta

Fuente: Elaboración propia

Tabla 4. Requerimiento funcional RF03.

código del requerimiento	RF03
Nombre	Nodo datos reales
Propósito	Importar data set reales seleccionados por el usuario
Descripción	El sistema permitirá importar archivos con extensión (.xlsx .mat .csv) de forma externa
Entrada	Importación de data set
Salida	Data set almacenado temporalmente
Prioridad	Alta

Fuente: Elaboración propia

Tabla 5. Requerimiento funcional RF04.

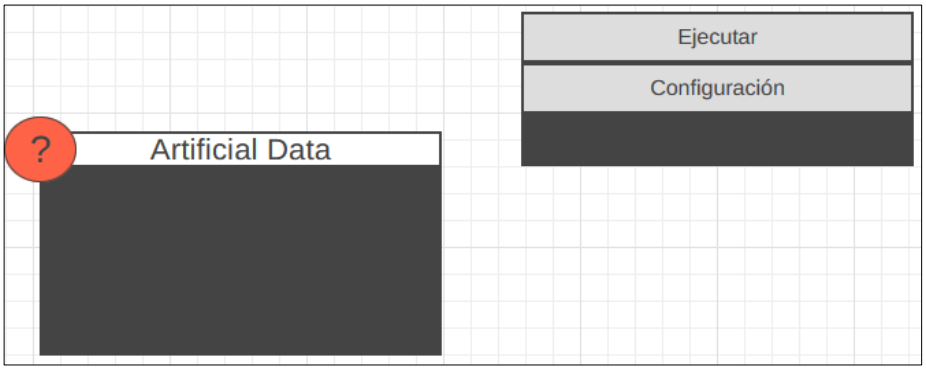
código del requerimiento	RF04
Nombre	Separador
Propósito	Elegir el separador o tabulador para ordenar el data set por columnas
Descripción	Separador para generar la tabulación de data set cuando este no este dividido por columnas y así pueda ser reconocido por los métodos de reducción de dimensionalidad
Entrada	Data set importado
salida	Data set configurado y listo para ejecución
prioridad	Alta

Fuente: Elaboración propia

- **Diseño**

El mockup de la siguiente figura contiene la forma en la que se va a desarrollar el nodo y su menú contextual, el cual define si se quiere configura o ejecutar el nodo.

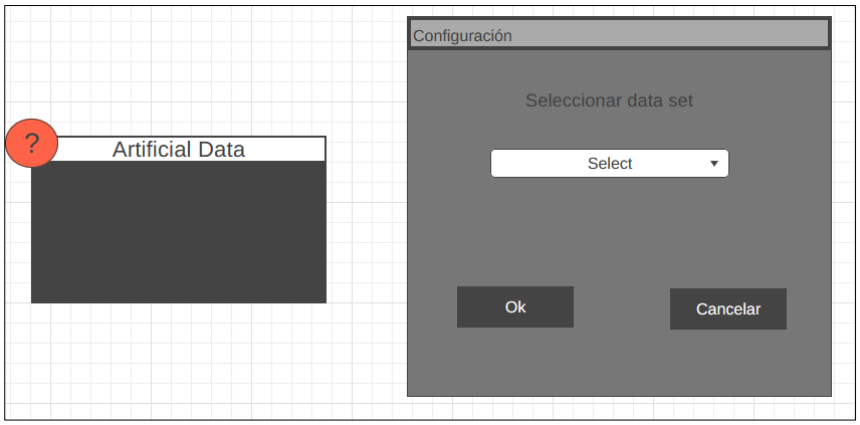
Figura 26: Mockup Nodo de datos artificiales



Fuente: Elaboración propia

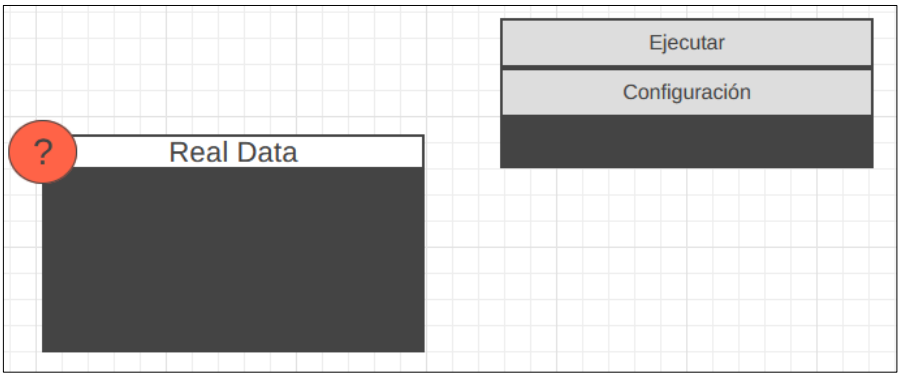
El apartado de configuración, le ofrece al usuario un modal para seleccionar el conjunto de datos artificiales que quiere cargar, cabe mencionar que algunos nodos requieren esta configuración y otros no.

Figura 27: Mockup modal, selección de datos artificiales



Fuente: Elaboración propia

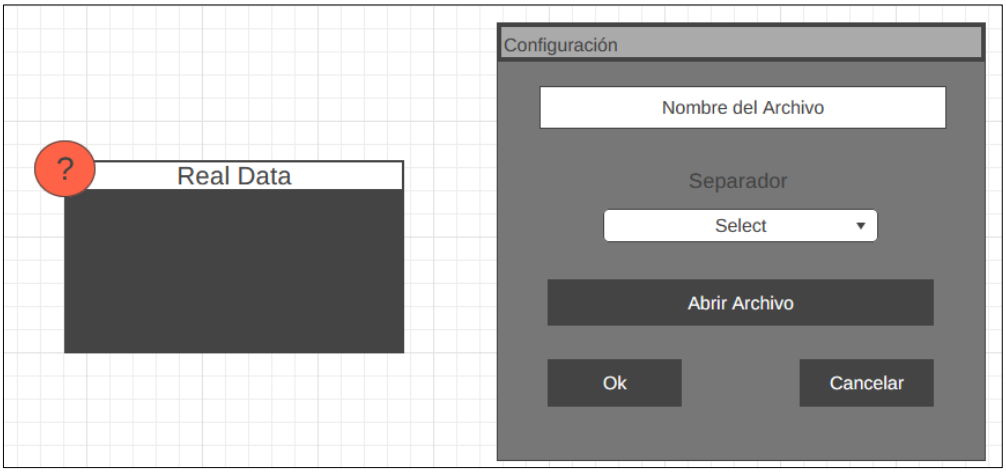
Figura 28: Mockup nodo de datos reales



Fuente: Elaboración propia

El modal de configuración del nodo de datos reales es bastante diferente, pues el usuario debe seleccionar el conjunto de datos del gestor de sus archivos locales, una vez cargados el sistema le mostrará el conjunto de datos que cargará al ejecutar el nodo.

Figura 29: Modal, cargue de datos reales



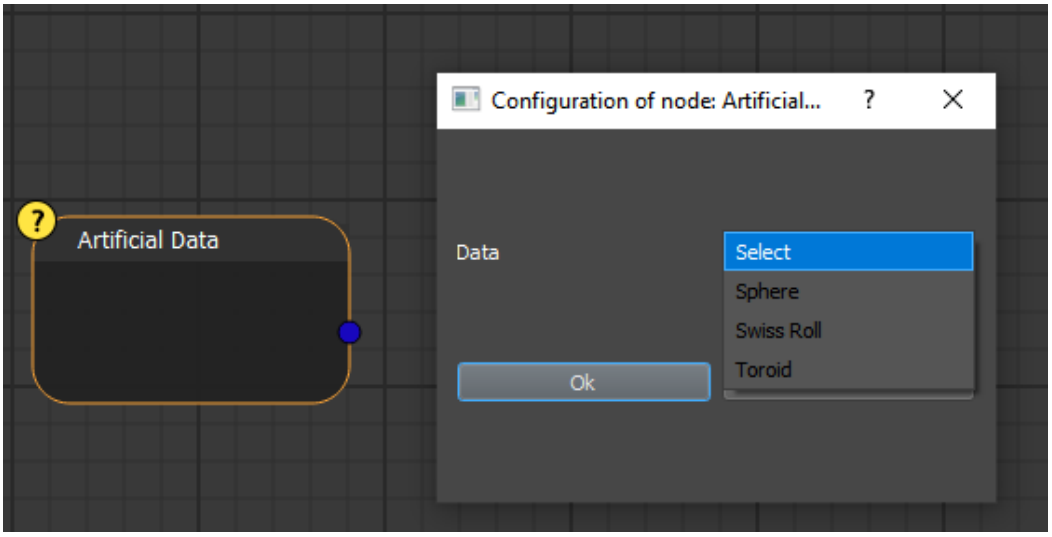
.

Fuente: Elaboración propia.

- **Desarrollo**

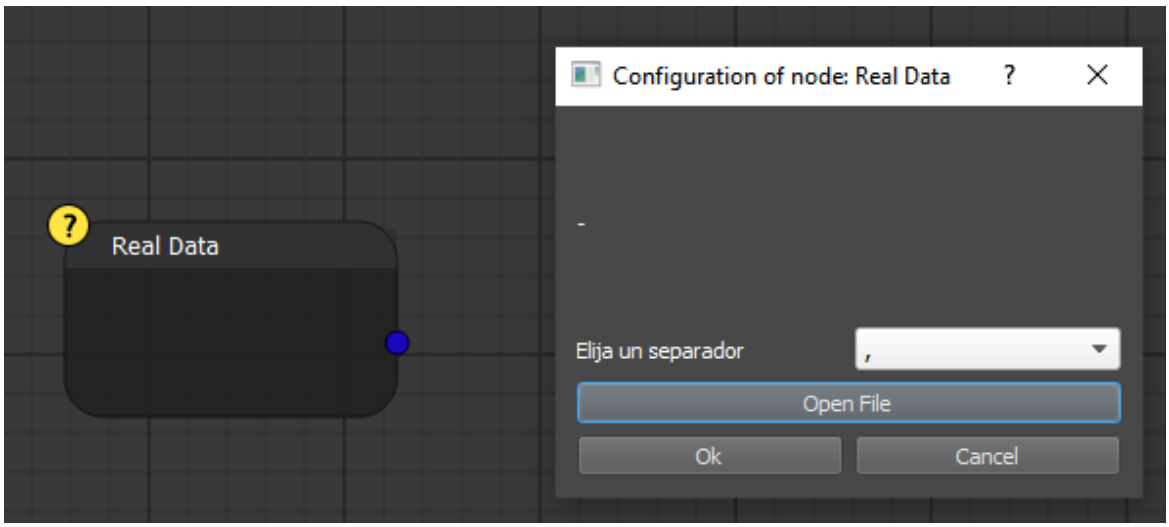
El nodo de datos artificiales solo requiere un output, siendo este el que contendrá los datos cargados, para el nodo de datos reales es lo mismo. En este caso, el nodo permite cargar el Rollo Suizo, la Esfera y el Toroide, pero se pueden ingresar más conjuntos de datos, debido a la escalabilidad de la herramienta.

Figura 30: Nodo de datos artificiales desarrollado



Fuente: Elaboración propia.

Figura 31: Nodo de datos reales desarrollado



Fuente: Elaboración propia.

En ocasiones los conjuntos de datos suelen venir con algún tipo de separador, uno de los más comunes es el separado por coma o CSV, sin embargo, el conjunto de datos puede venir con un separador como la barra inclinada (/) o el punto y coma (;), por ello, la herramienta le permite al usuario seleccionar el tipo de separador que su conjunto de datos usa.

4.1.4.3 Limpieza de datos

Aunque la herramienta no tiene un módulo grande de limpieza de datos, si cuenta con un nodo para eliminar columnas, ya que este proceso es muy usado cuando se usan algoritmos no supervisados en donde los datos respuesta no son necesarios, en este caso gran parte de los métodos RD son no supervisados, aun así, el proceso de limpieza de datos es importante debido a que es necesario determinar el grado de confianza de la información obtenida, para ello se realiza un tratamiento para remoción de valores ambiguos o atípicos, donde se eliminarán variables o atributos que no sean útiles para la implementación de los métodos RD, como los datos categóricos de figuras geométricas como “cuadrado” o “triangulo”.

- **Análisis: Requerimientos funcionales**

Tabla 6. Requerimiento funcional RF05.

código del requerimiento	RF05
nombre	Nodos particionador
propósito	Eliminar columnas de data set que no son identificadas o necesarias por los métodos
descripción	identifica cada columna en el data set y se visualiza por medio de una lista desplegable (chekbox)
entrada	Data set importado
salida	Data set configurado y listo para ejecución
prioridad	Alta

Fuente: Elaboración propia

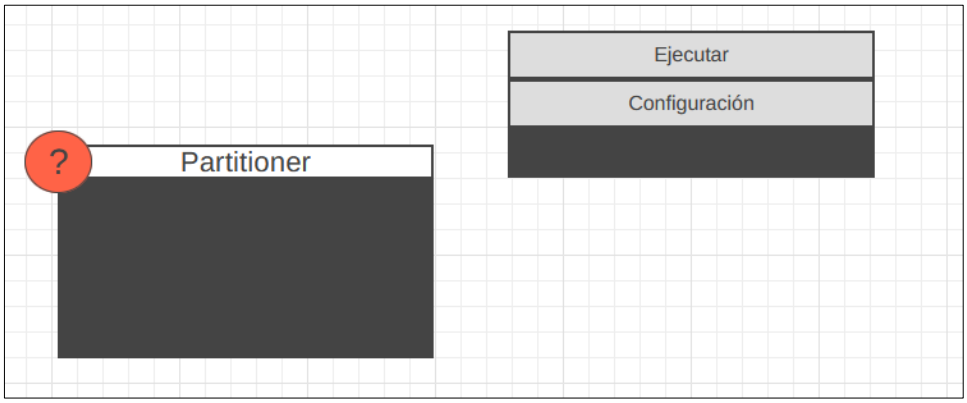
Tabla 7. Requerimiento funcional RF06.

código del requerimiento	RF06
nombre	Configurar particionador
propósito	Permite seleccionar la columna a eliminar
descripción	Identificar en la lista desplegable la columna que se va a eliminar
entrada	Data set real
salida	Data set depurado
prioridad	Media

Fuente: Elaboración propia

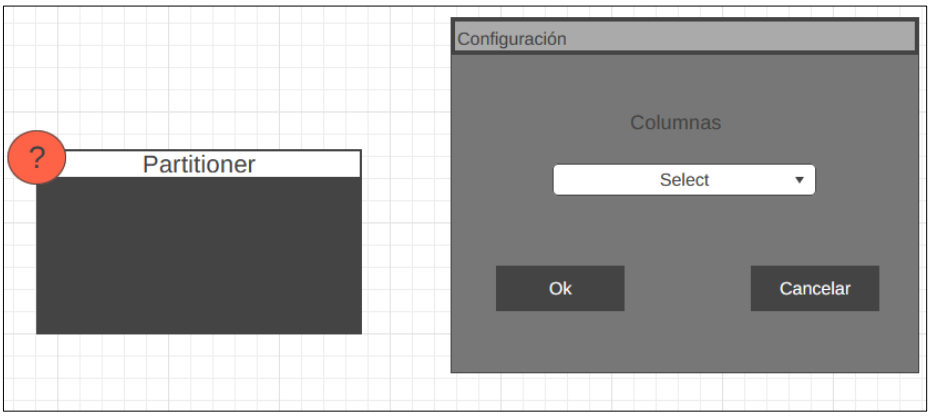
- **Diseño**

Figura 32: Mockup nodo particionador



Fuente: Elaboración propia.

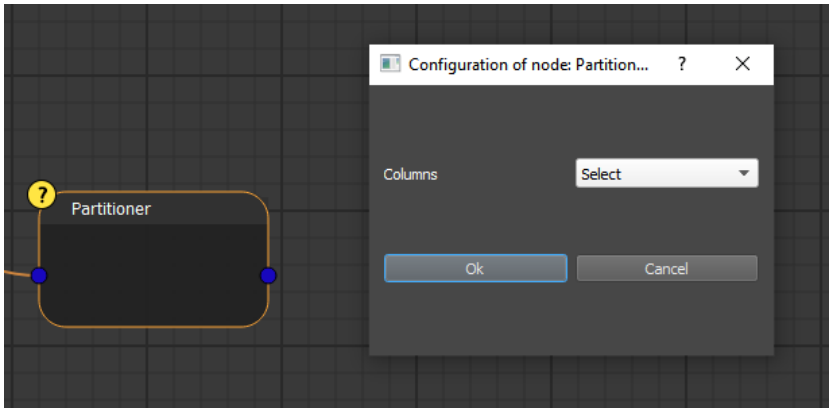
Figura 33: Mockup modal selección de columna



Fuente: Elaboración propia.

- **Desarrollo**

Figura 34: Nodo particionador desarrollado



Fuente: Elaboración propia.

El nodo particionador siempre estará conectado a un proveedor de datos, los proveedores de datos pueden ser los nodos de cargue de datos, métodos RD e incluso otro particionador.

4.1.4.4 Transformación de los datos

Una vez los datos están cargados, limpios y procesados, se pueden aplicar los métodos RD o cualquier otro algoritmo, en este caso, la herramienta cuenta con 6 nodos que representan los 6 métodos RD descritos en la sección 1.7.

- **Análisis: Requerimientos funcionales**

Tabla 8. Requerimiento funcional RF07.

código del requerimiento	RF07
nombre	Nodos métodos RD
propósito	Reducir dimensionalidad de los datos de entrada
descripción	Nodos de los métodos de reducción de dimensión PCA, KPCA, LE, LLE, ISOPAM, MCD.
entradas	Data set cargado
salida	Nuevo data set a partir de los datos de entrada
prioridad	Alta

Fuente: Elaboración propia

Tabla 9. Requerimiento funcional RF08.

código del requerimiento	RF08
nombre	Nodo Paramétrico
propósito	Configurar el número de dimensiones y vecindarios que tendrá el data set
descripción	El sistema deberá permitir como entrada el numero de dimensiones y numero de vecindario.
entradas	Numero de dimensiones.
Salida	Nodo configurado
prioridad	alta

Fuente: Elaboración propia

Tabla 10. Requerimiento funcional RF09.

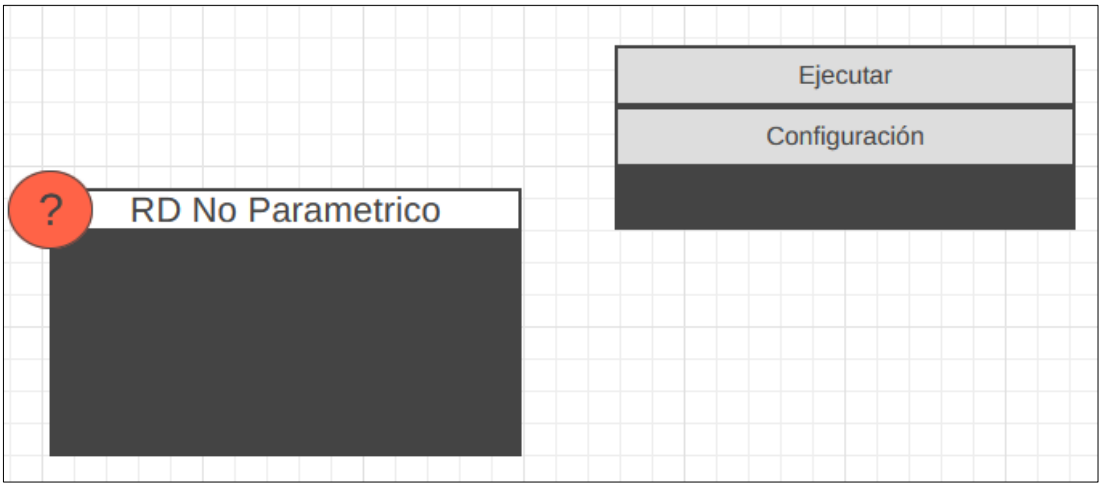
código del requerimiento	RF09
nombre	Nodo No paramétrico
propósito	Configurar el número de dimensiones a aplicar en el data set
descripción	El sistema deberá permitir la entrada número de dimensiones
entradas	Numero de dimensiones.
Salida	Nodo configurado
prioridad	alta

Fuente: Elaboración propia

- **Diseño**

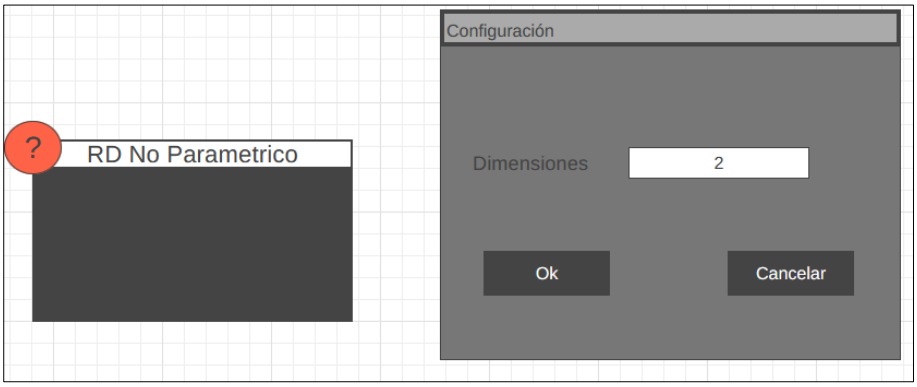
En este caso, la vista de todos los nodos de métodos RD es la misma, el cambio radica entre los métodos paramétricos y los no paramétricos como se observa en la Tabla 1, esto debido a que los métodos paramétricos reciben en su configuración un tamaño de vecindarios y los no paramétricos solo reciben las dimensiones a las que se quieren reducir el conjunto de datos.

Figura 35: Nodos no paramétricos



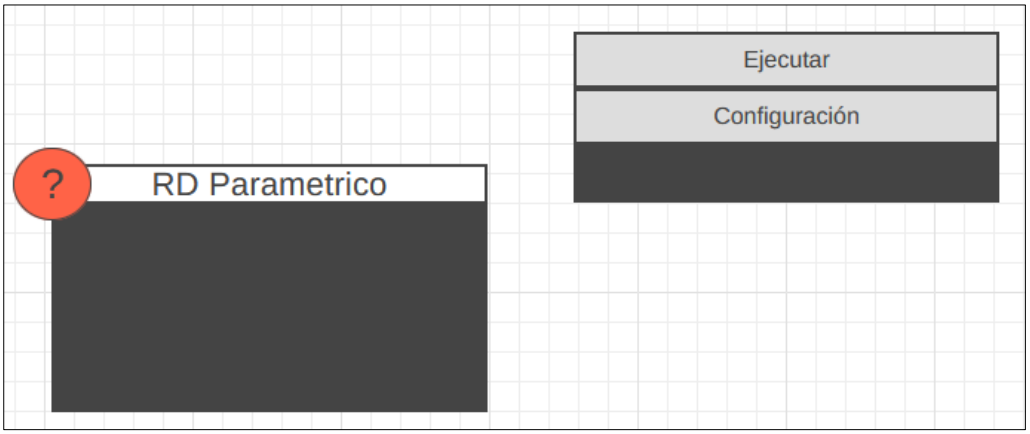
Fuente: Elaboración propia.

Figura 36: Mockup modal nodos no paramétricos



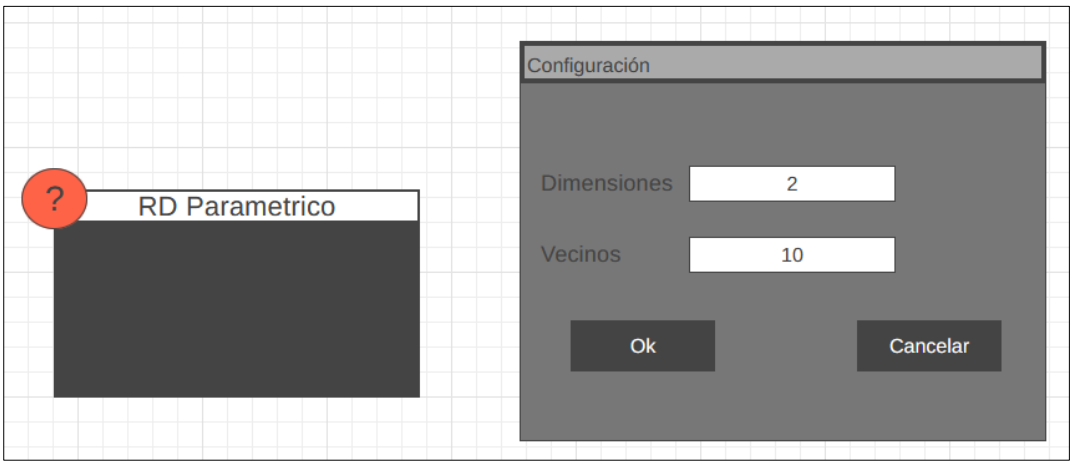
Fuente: Elaboración propia.

Figura 37: Mockup nodos paramétricos



Fuente: Elaboración propia.

Figura 38: Mockup modal nodos paramétricos

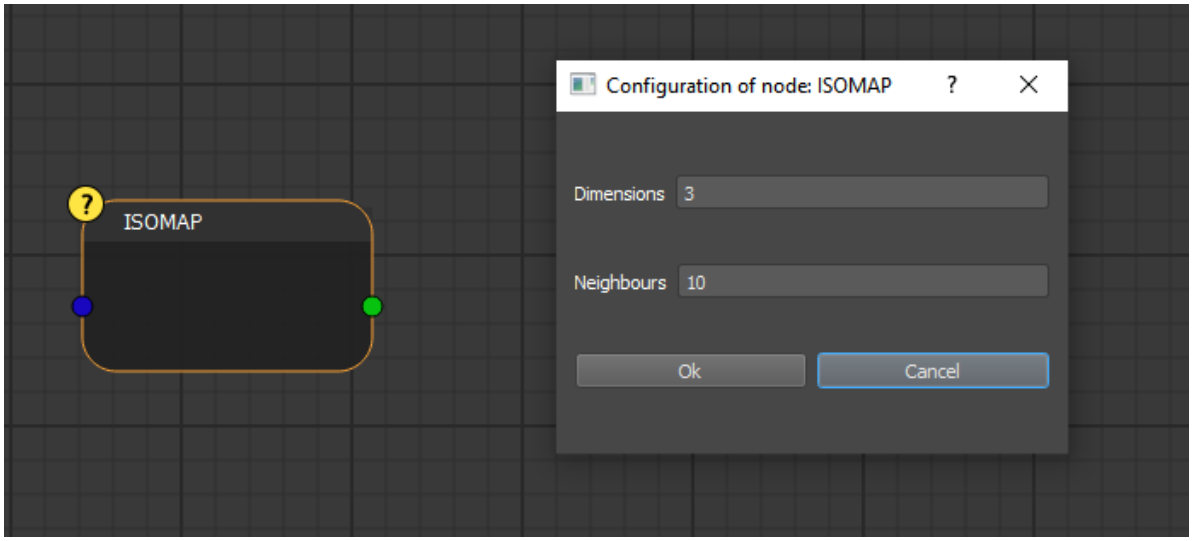


Fuente: Elaboración propia.

- **Desarrollo**

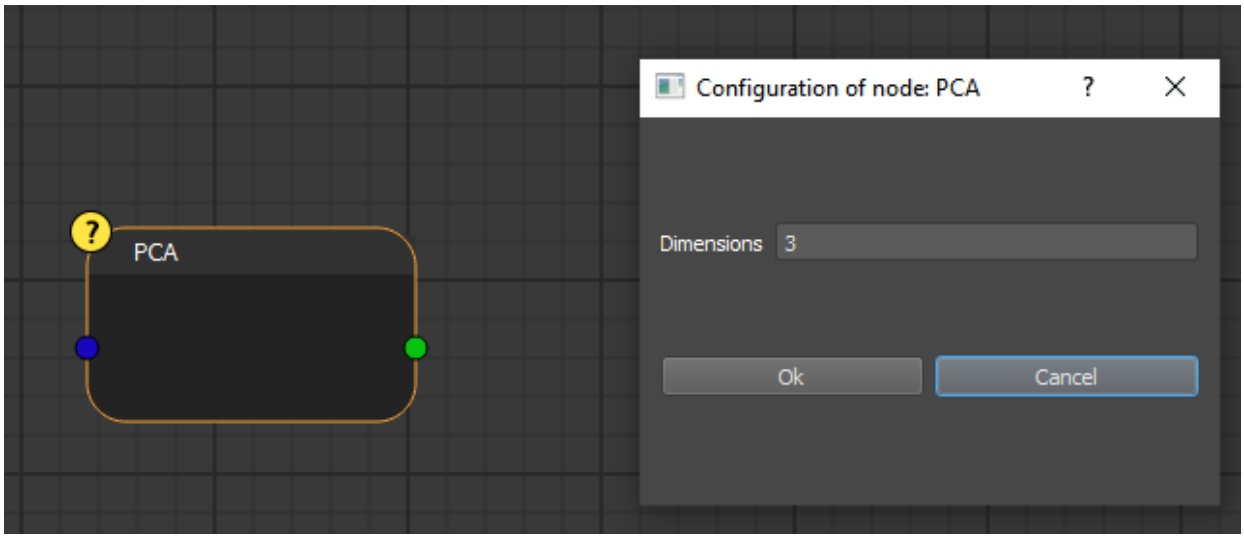
Los nodos de métodos RD, pueden recibir como input, conjuntos de datos, estos pueden venir del output de los nodos de cargue de datos, particionador o incluso otros métodos RD, y retornan un conjunto de datos.

Figura 39: Nodo ISOMAP (paramétrico) desarrollado con su modal de configuración



Fuente: Elaboración propia.

Figura 40: Nodo PCA (no paramétrico) desarrollado con su modal de configuración



Fuente: Elaboración propia.

4.1.4.5 Elección de tareas de minería de datos

En esta etapa se escoge el paradigma apropiado para minería de datos, aquí se puede incluir la clasificación, regresión o agrupación, según se adapte a los objetivos de la investigación. La herramienta no incluye este paso debido a que no se están analizando y evaluando algoritmos de minería de datos, aunque cabe resaltar que debido a la versatilidad de los métodos RD, estos también se pueden encontrar en esta etapa.

4.1.4.6 Elección del algoritmo

Depende de los objetivos iniciales y la naturaleza de la investigación, se pueden establecer elecciones de algoritmos ya sea para predicción, encontrando modelos utilizados para posteriores casos o descripción para solo realizar una observación del comportamiento y resultados. Hay diversas posibilidades para seleccionar estos algoritmos y según esta decisión, se pueden obtener resultados óptimos, para el cual es necesario conocer previamente las propiedades de cada candidato y observar cual se ajusta mejor a las necesidades. este proceso de selección está más relacionado a la selección de algoritmos de MD, y, por lo tanto, no se tiene en cuenta en la presente investigación.

4.1.4.7 Aplicación del algoritmo

Esta sección va orientada a los algoritmos de Minería de Datos como los algoritmos de regresión o clasificación, los cuales son entrenados con los datos previamente depurados para la obtención de patrones que permitan realizar predicciones con alta precisión, la presente investigación no implementa algoritmos de minería de datos, que no sean los métodos RD ya mencionados.

4.1.4.8 Evaluación e interpretación

En esta etapa se verifican los patrones obtenidos por los algoritmos de MD, en este caso, se verifican y comparan los resultados obtenidos por los diferentes métodos RD mediante la implementación de las métricas RNX y se define cuales métodos han tenido un mejor rendimiento en los diferentes conjuntos de datos ya mencionados.

- **Análisis: Requerimientos funcionales.**

Para poder obtener las métricas RNX, fueron necesarias rutinas de programación para matrices de distancias pares, matriz de co-ranking entre otros, además las métricas deben recibir múltiples conexiones con el fin de que se puedan evaluar varios métodos RD a la vez.

Tabla 11. Requerimiento funcional RF10.

código del requerimiento	RF10
Nombre	Nodo evaluador RNX
Propósito	Evaluar mediante métricas RNX los diversos nodos de reducción conectados
Descripción	El sistema deberá permitir crear Mediante el nodo de implementación de las métricas RNX para evaluar los nodos de reducción conectados
Entrada	Nodos de reducción de dimensión
Salida	Evaluation
prioridad	Alta

Fuente: Elaboración propia

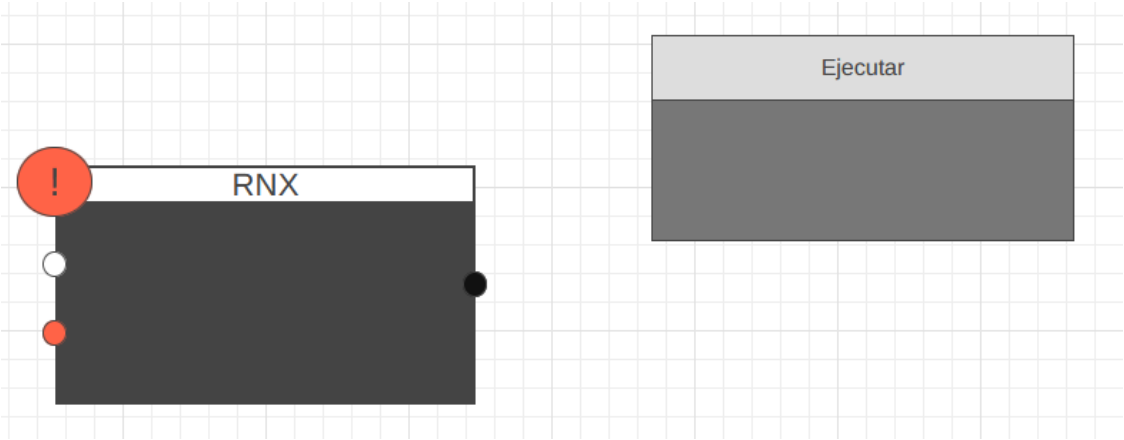
Tabla 12. Requerimiento funcional RF11.

código del requerimiento	RF11
nombre	Múltiples conexiones en nodo RNX
propósito	Permitir agrupar varias evaluaciones
descripción	El sistema deberá permitir que el nodo evaluador soporte como entradas los datos generados por uno o más métodos de reducción y desde el dataset original
entrada	Nodos de reducción de dimensión, dataset
Salida	Evaluaciones de cada nodo conectado
Prioridad	Alta

Fuente: Elaboración propia

- **Diseño**

Figura 41: Mockup nodo RNX



Fuente: Elaboración propia.

- **Desarrollo**

El nodo de RNX recibe dos inputs, los datos originales y los datos después de aplicarle algún método RD, después de realizar el proceso de evaluación, RNX retorna una matriz que contiene los valores en el eje y de la gráfica de evaluación, siendo los valores de la matriz mayores o iguales a 0 y menores o iguales a 100, además se obtiene el promedio de los resultados de la matriz obteniendo así la evaluación de los diferentes métodos.

Figura 42: Nodo RNX desarrollado

Fuente: Elaboración propia.

4.1.4.9 Entendimiento de conocimiento

Al culminar, se realiza una iteración superficial de cada paso aplicado, para establecer si los resultados son los esperados y resuelven la problemática inicial, en caso de no satisfacer al objetivo, se hace necesario ir por cada etapa realizando los ajustes pertinentes. En este caso la obtención de conocimiento hace referencia a conocer los comportamientos de los métodos RD en los

diferentes conjuntos de datos, siendo esta información de gran importancia para futuras investigaciones, o mejoras a los diferentes métodos RD.

4.1.4.10 Herramienta de visualización

La herramienta de visualización se implementa para generar en el usuario el dinamismo que se requiere para realizar diferentes combinaciones entre los nodos según sea necesario, así mismo, mediante la inclusión de la metodología DRAG & DROP, permite que el usuario solamente realice acciones de configuración y unión de nodos, ya que esta forma de generar los grafos hace que el aplicativo sea intuitivo, ilustrativo y educativo al momento de usarlo. La etapa de visualización de datos se puede presenciar antes o después de otras etapas, por ejemplo, la visualización de datos se puede dar antes de realizar limpieza de datos mediante un HeatMap, o después con el mismo método de visualización, en la herramienta desarrollada sucede lo mismo, los métodos de visualización de información pueden usarse antes o después de aplicar un método RD, aunque existe una excepción, y es que el uso del gráfico lineal debe ser únicamente para visualizar las curvas RNX, a continuación se presenta el proceso de desarrollo de estos métodos de visualización y finalmente se especifican aspectos generales del funcionamiento de la herramienta.

- **Análisis: Requerimientos funcionales**

Tabla 13. Requerimiento funcional RF12.

código del requerimiento	RF12
nombre	Nodo scatter plot
propósito	Crear gráficas de dispersion
descripción	El sistema deberá permitir crear nodo para generar graficas de dispersión en 2D y 3D
entrada	Data set
salida	Gráfica
prioridad	Alta

Fuente: Elaboración propia

Tabla 14. Requerimiento funcional RF13.

código del requerimiento	RF13
nombre	Nodo visualización de evaluación RNX
propósito	Graficar métrica RNX
descripción	El sistema deberá permitir crear Nodo para visualizar las curvas de evaluación RNX
entrada	Nodo RNX
salida	Gráfica de curvas RNX
prioridad	Alta

Fuente: Elaboración propia

Tabla 15. Requerimiento funcional RF14.

código del requerimiento	RF14
nombre	Interfaz
propósito	Permitir una interacción sencilla con el usuario
descripción	El sistema deberá tener un listado de nodos y una zona de trabajo
entrada	Interacción del usuario
salida	Proyección en pantalla
prioridad	alta

Fuente: Elaboración propia

Tabla 16. Requerimiento funcional RF15.

código del requerimiento	RF15
nombre	Nodo Data Table
propósito	Visualización de datos
descripción	El sistema deberá permitir crear tabla de datos conectados al nodo
entrada	Data set
salida	Tabla de datos
prioridad	media

Fuente: Elaboración propia

Tabla 17. Requerimiento funcional RF16.

código del requerimiento	RF16
nombre	Workflow
propósito	Espacio de trabajo
descripción	El sistema deberá permitir crear flujos de trabajo que el usuario pueda usar con facilidad
entrada	Interacción del usuario
salida	Workflow
prioridad	Alto

Fuente: Elaboración propia

Tabla 18. Requerimiento funcional RF17.

Código del requerimiento	RF17
nombre	Interconexión de nodos
propósito	Comunicar nodos
descripción	El sistema deberá permitir la interconexión entre los nodos en el WorkFlow
entrada	Interacción del usuario
salida	Proyección en pantalla
prioridad	alto

Fuente: Elaboración propia

Tabla 19. Requerimiento funcional RF18.

código del requerimiento	RF18
nombre	Guardar workflow
propósito	Almacenar WorkFlow en archivos JSON
descripción	El sistema deberá permitir guardar flujos de trabajo en formato JSON
entrada	Workflow
salida	Archivo JSON
prioridad	medio

Fuente: Elaboración propia

Tabla 20. Requerimiento funcional RF19.

código del requerimiento	RF19
nombre	Abrir Workflow
propósito	Cargar Workflow seleccionados por el usuario
descripción	El sistema deberá permitir cargar flujos de trabajo que el usuario allá guardado con anterioridad
entrada	Sistema de archivos
salida	Workflow
prioridad	medio

Fuente: Elaboración propia

- **Análisis: Requerimientos no funcionales**

Debido a que algunos algoritmos tienen tiempos de complejidad bastante altos, siendo uno de estos MDS y las mismas métricas RNX, fue necesario usar programación multihilo, de esta forma mientras algún algoritmo se está ejecutando, su ejecución se hace en un proceso a parte de la ejecución de la interfaz gráfica, impidiendo que esta se congele dando una mala experiencia al usuario.

Tabla 21. Requerimiento no funcional RNF1

código del requerimiento	RNF1
nombre	Usabilidad
propósito	Soporte multiplataforma
descripción	El sistema deberá permitir ser ejecutado en distintos sistemas operativos (Windows, Linux, OSX)
prioridad	alto

Fuente: Elaboración propia

Tabla 22. Requerimiento no funcional RNF2

código del requerimiento	RNF2
nombre	Eficiencia
propósito	Ejecución rápida
descripción	El sistema deberá tener tiempos de ejecución de cada nodo no mayores a 5 segundos en promedio
prioridad	alto

Fuente: Elaboración propia

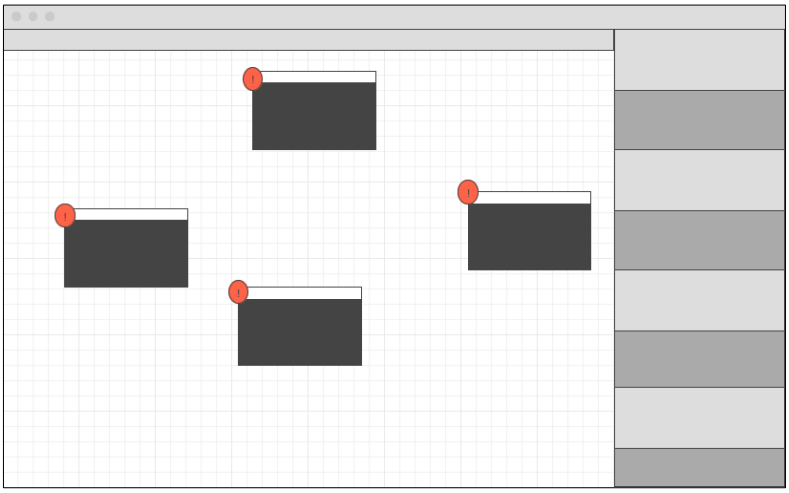
Tabla 23. Requerimiento no funcional RNF3

código del requerimiento	RNF3
nombre	Workflow
propósito	Espacio de trabajo
descripción	El sistema deberá permitir crear flujos de trabajo que el usuario pueda usar con facilidad
prioridad	alto

Fuente: Elaboración propia

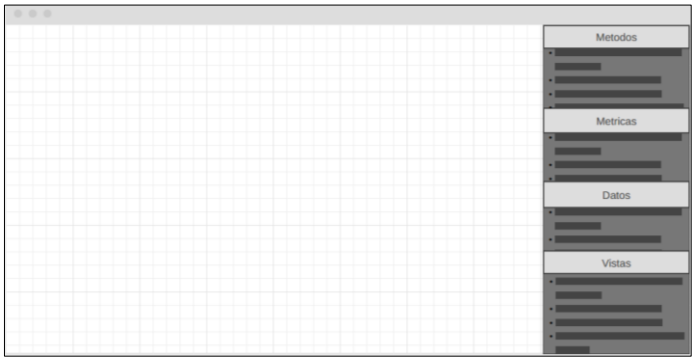
- **Diseño**

Figura 43: Mockup canvas de la herramienta Drag and Drop



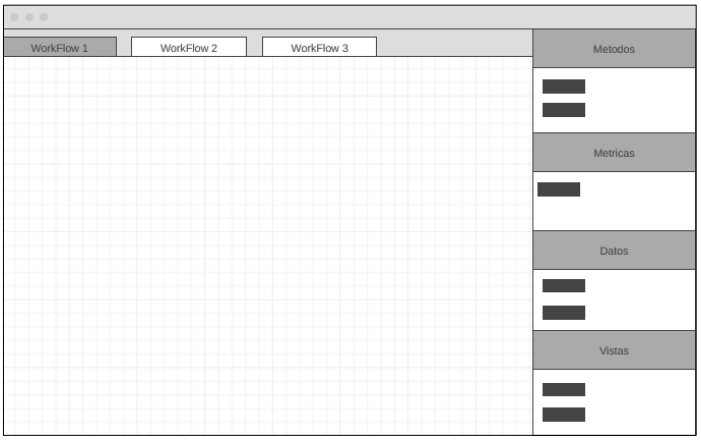
Fuente: Elaboración propia.

Figura 44: Mockup menús de la herramienta Drag and Drop



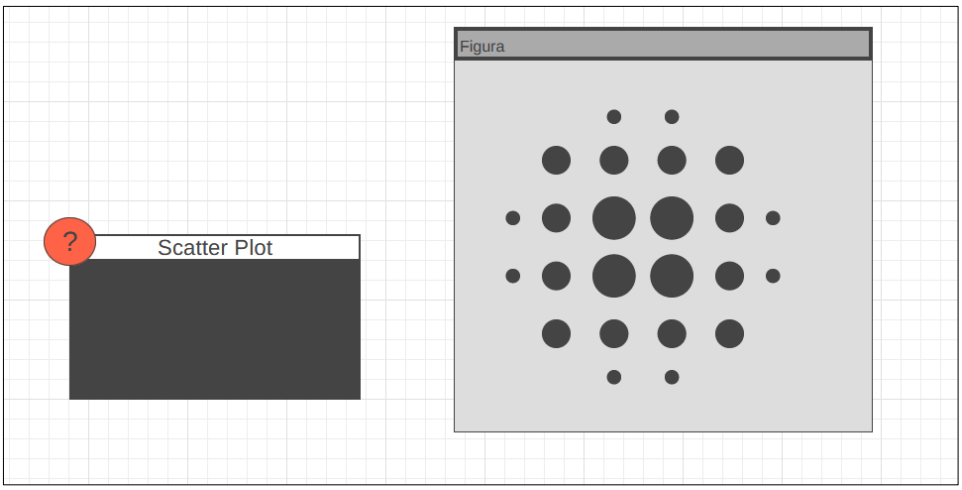
Fuente: Elaboración propia.

Figura 45: Flujos de trabajo



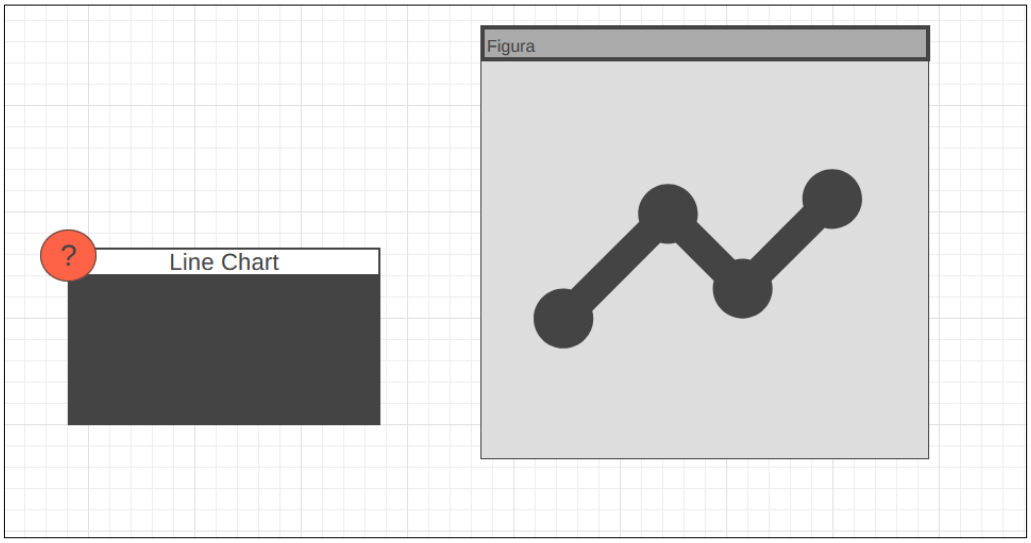
Fuente: Elaboración propia.

Figura 46: Mockup nodo SCATTER PLOT



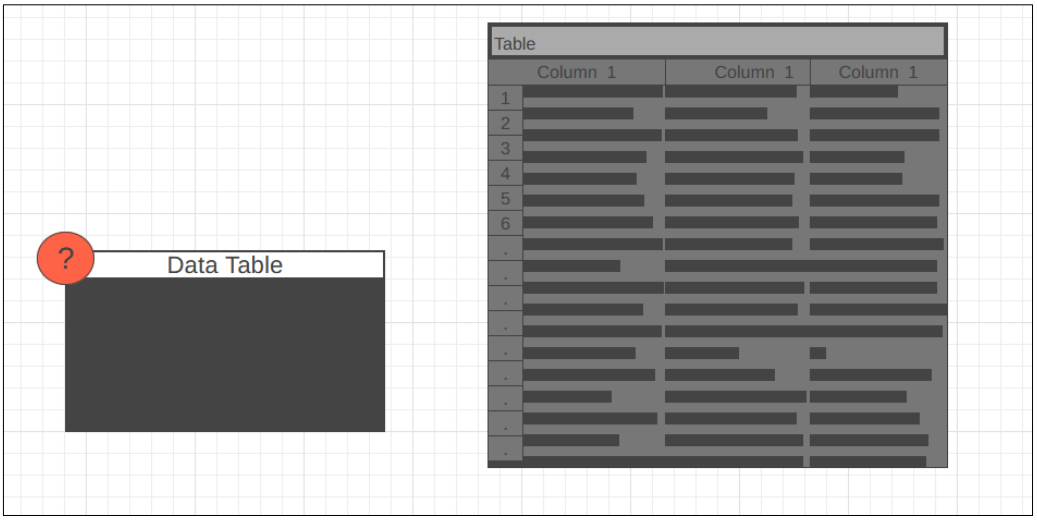
Fuente: Elaboración propia.

Figura 47: Mockup nodo LINE CHART



Fuente: Elaboración propia.

Figura 48: Mockup nodo DATA TABLE

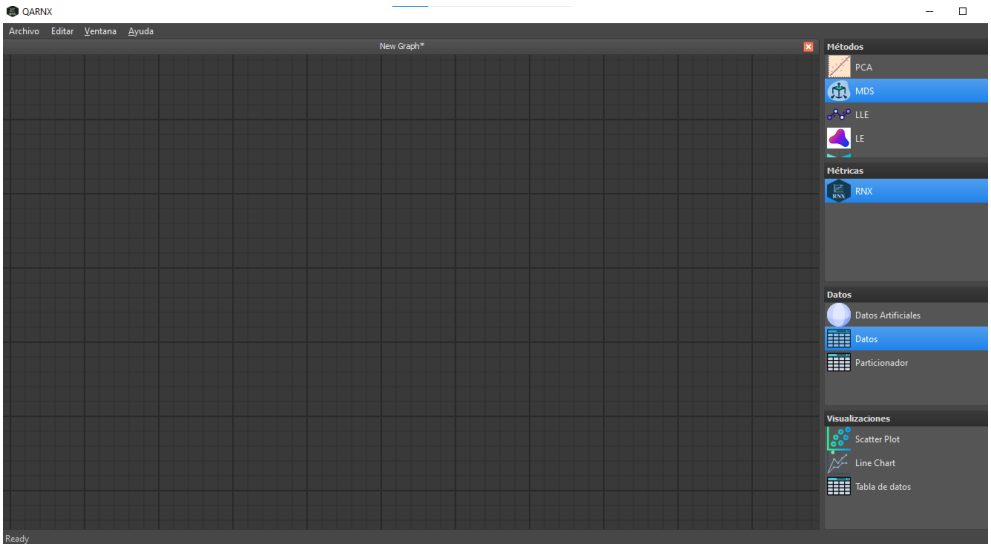


Fuente: Elaboración propia.

- **Desarrollo**

La herramienta cuenta con 4 menús, los cuales contienen los nodos que posteriormente serán arrastrados al flujo de trabajo, los menús están debidamente clasificados y ordenados.

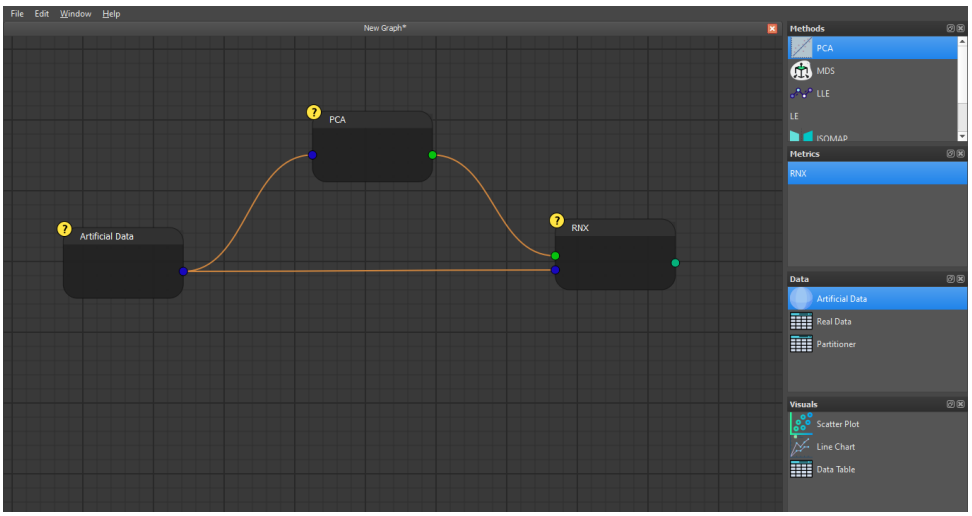
Figura 49: Entorno gráfico de la herramienta desarrollado



Fuente: Elaboración propia

Una vez arrastrados los nodos al flujo de trabajo, estos deben poder conectarse por medio de líneas, las cuales pueden ser curvas o rectas.

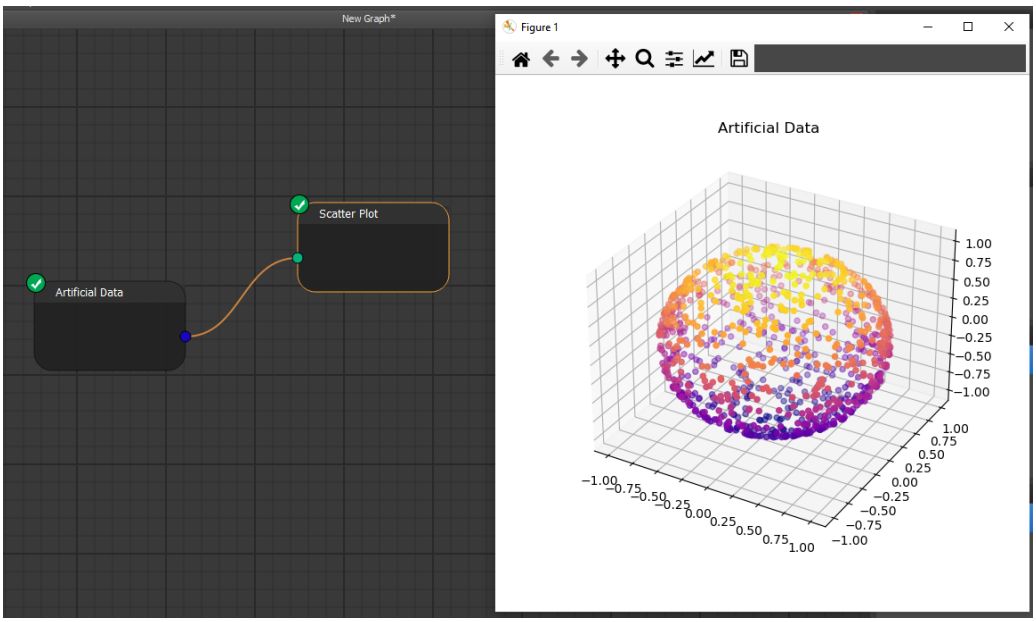
Figura 50: Nodos y conexiones en el canvas



Fuente: Elaboración propia

Con el nodo SCATTER PLOT o diagrama de dispersión, se pueden visualizar los datos en alta y baja dimensión, aunque como se ha mencionado anteriormente, este tipo de diagramas solo pueden visualizar hasta 3 dimensiones, para visualizar más dimensiones es necesario otro tipo de métodos de visualización.

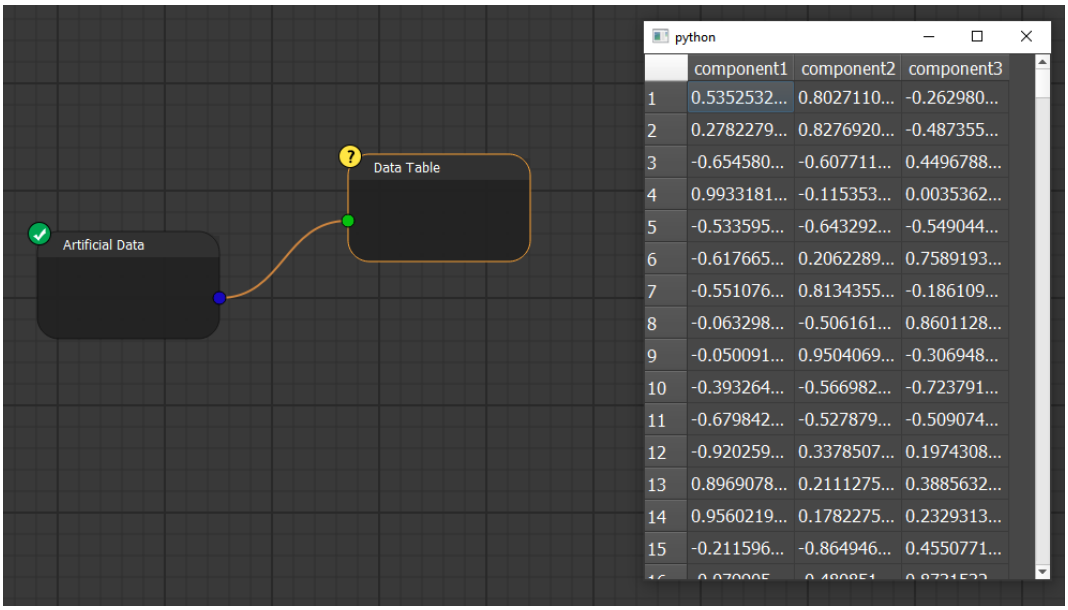
Figura 51: Nodo scatter plot desarrollado



Fuente: Elaboración propia

La visualización tabular también es de suma importancia, esta permite la observación de los datos de tal forma que se pueda visualizar su tipo, sean estos categóricos numéricos entre otros, además de si hay datos nulos dentro del conjunto de datos que, de ser así, no se podría usar en los métodos RD.

Figura 52: Nodo Data Table

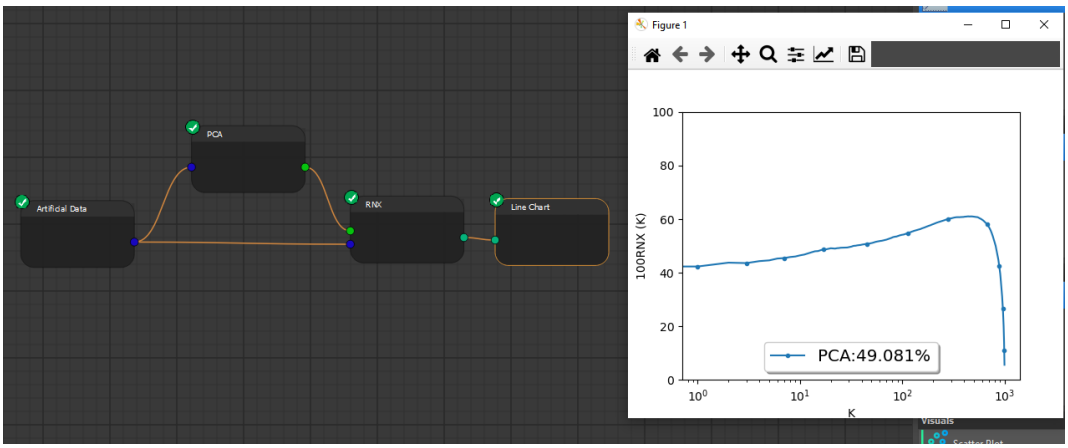


	component1	component2	component3
1	0.5352532...	0.8027110...	-0.262980...
2	0.2782279...	0.8276920...	-0.487355...
3	-0.654580...	-0.607711...	0.4496788...
4	0.9933181...	-0.115353...	0.0035362...
5	-0.533595...	-0.643292...	-0.549044...
6	-0.617665...	0.2062289...	0.7589193...
7	-0.551076...	0.8134355...	-0.186109...
8	-0.063298...	-0.506161...	0.8601128...
9	-0.050091...	0.9504069...	-0.306948...
10	-0.393264...	-0.566982...	-0.723791...
11	-0.679842...	-0.527879...	-0.509074...
12	-0.920259...	0.3378507...	0.1974308...
13	0.8969078...	0.2111275...	0.3885632...
14	0.9560219...	0.1782275...	0.2329313...
15	-0.211596...	-0.864946...	0.4550771...

Fuente: Elaboración propia.

Como se observa en el flujo de la siguiente figura, el gráfico lineal solo se utiliza con el resultado de la ejecución del nodo de las métricas RNX.

Figura 53: Nodo Line Chart



Fuente: Elaboración propia.

Adicionalmente, en el **Anexo B** se encuentran los diagramas UML referentes a el manejo lógico y visual de la herramienta, allí se explica la conexión entre las diferentes clases y componentes gráficos, además, en el **Anexo A** manual de usuario, se observa el funcionamiento de la creación, guardado y cargado de los workflows. Para aquellos interesados en el código fuente de la herramienta, pueden visualizarlo en el siguiente repositorio de GitHub https://github.com/CarlosDCorrea/rnx_tool/tree/master

5 ANÁLISIS Y DISCUSIÓN DE RESULTADOS

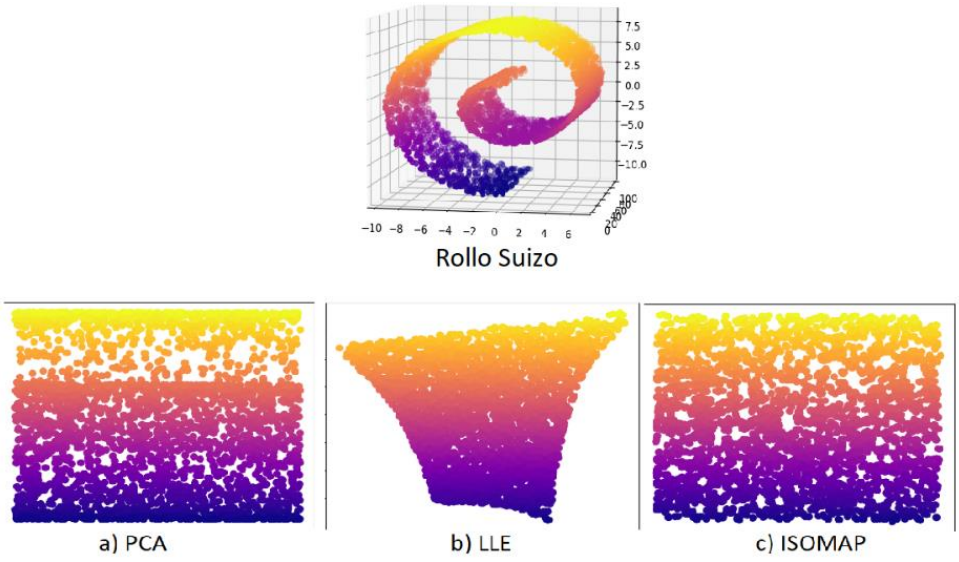
Experimentos con los métodos RD integrados en la herramienta

En esta sección se llevaron a cabo diferentes experimentos con los métodos RD mencionados en la sección **1.7**, los experimentos realizados fueron una parte fundamental de la investigación, puesto que se sabe que no hay un método superior a los demás, sino que cada uno tiene sus virtudes dependiendo del conjunto de datos al que se esté aplicando, es importante recordar que se tienen métodos RD globales (MDS y PCA) y locales (LLE, LE, ISOMAP) además de uno kernelizado (Kernel PCA). Por otro lado, estos métodos fueron usados con conjuntos de datos como MNIST, Fray Faces, la Esfera, el Toroide y el Rollo Suizo mencionados en la sección **3.7**. Finalmente, la evaluación fue realizada con las métricas RNX introducidas en la sección **2.2.7.8**.

Experimentos con el conjunto de datos del Rollo Suizo

Para este experimento se utilizó un número de vecindarios de 10 para los métodos paramétricos, el objetivo entonces, es la visualización de la incrustación realizada y su calidad.

Figura 54: Rollo suizo con su incrustamiento realizado por PCA, LLE e ISOMAP

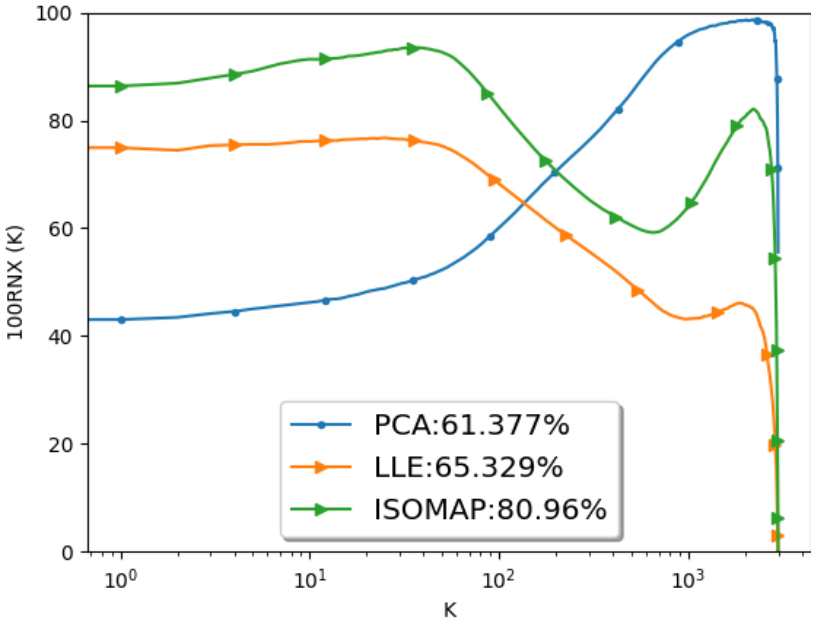


Fuente: Investigación propia

En la anterior figura se puede observar las diferentes incrustaciones realizadas por los métodos RD, debido a que los métodos son globales o locales, además

de lineales o no lineales, su incrustación siempre variará, para determinar cuál fue más efectivo en ese conjunto de datos, en la siguiente figura se muestra la evaluación realizada por RNX

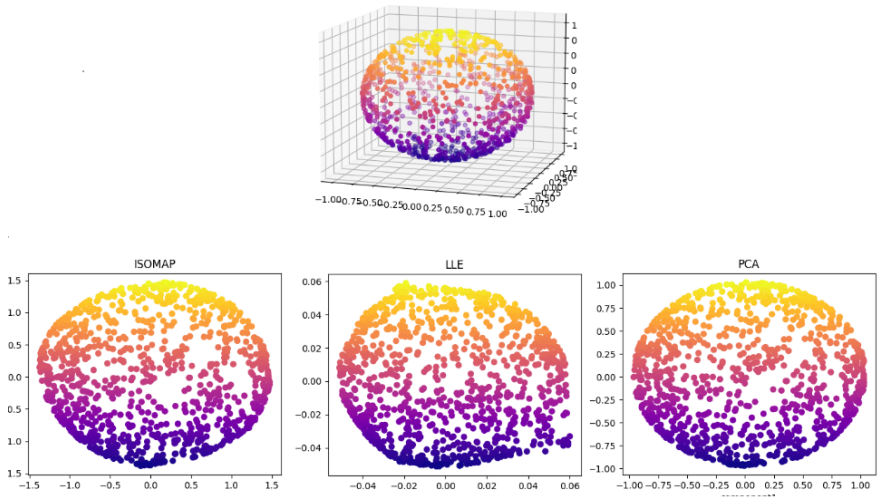
Figura 55: Evaluación de los métodos PCA, LLE, e ISOMAP con RNX conjunto de datos Rollo Suizo.



Fuente: Investigación propia

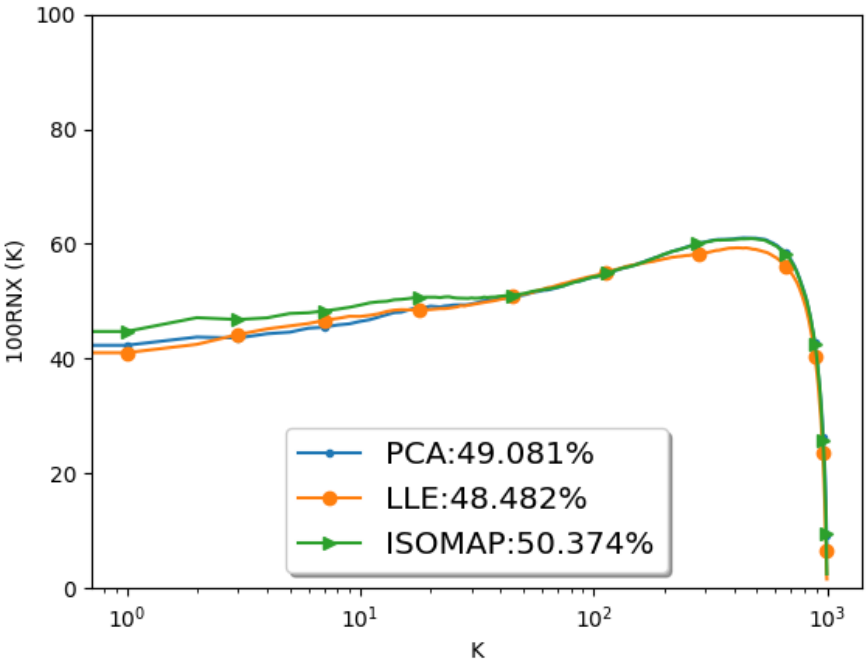
Experimentos con el conjunto de datos Esfera

Figura 56: Esfera con su incrustamiento realizado con PCA, LLE e ISOMAP.



Fuente: Elaboración propia

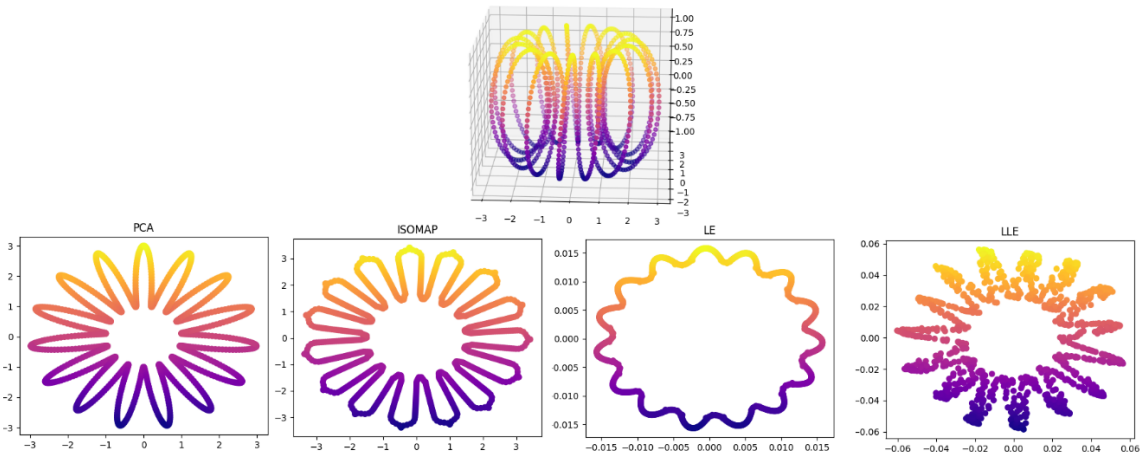
Figura 57: Evaluación de los métodos PCA, LLE, e ISOMAP con RNX conjunto de datos Esfera



Fuente: Elaboración propia

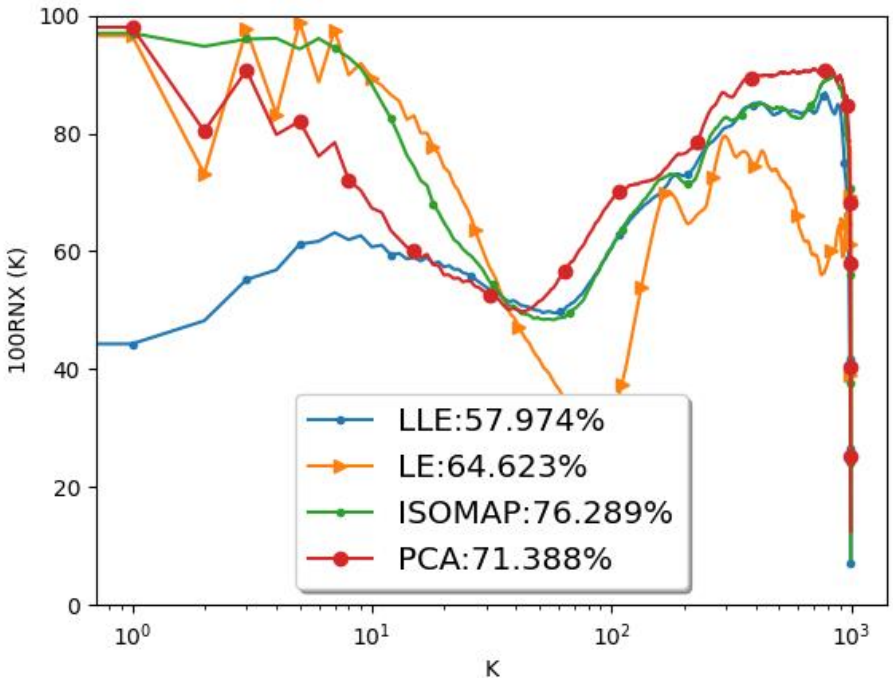
Experimentos con el conjunto de datos del Toroide

Figura 58: Toroide con su incrustamiento realizado por PCA, LLE, LE e ISOMAP.



Fuente: Investigación propia

Figura 59: Evaluación de los métodos PCA, LE, LLE e ISOMAP con RNX conjunto de datos Toroide.



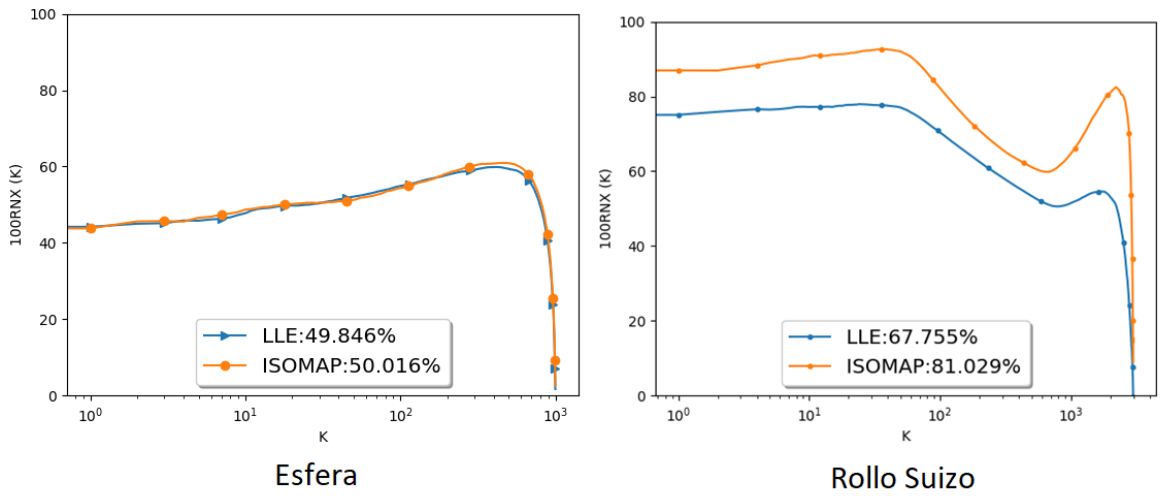
Fuente: Elaboración propia

La Figura 54 representa el incrustamiento realizado por ISOMAP, LLE y PCA en el conjunto de datos Rollo Suizo, y la Figura 55 muestra su evaluación

topológica, en esta se observa que ISOMAP supera a PCA y LLE en la preservación topológica del conjunto de datos, aunque si se observa detenidamente, PCA es mucho mejor que ISOMAP en grandes vecindarios, llegando casi al 100% de preservación topológica en ellos, siendo este comportamiento acorde en los mencionado por los autores en la sección **2.2.3.1.1**, quienes afirman que PCA debido a sus combinaciones lineales, las cuales tienen en cuenta todo el conjunto de datos, tiene un mejor comportamiento en la topología global, por otro lado, la Figura 57 muestra los resultados de la incrustación realizada en la Figura 56, aquí se puede observar que los tres métodos tienen preservaciones generales similares, además otra característica importante, es que la preservación tanto local como global de los métodos difiere muy levemente, aunque el método que mejor conserva la topología de los datos originales sigue siendo ISOMAP, la diferencia realmente no es tan significativa como en otros casos como por ejemplo, en la Figura 59, ISOMAP conserva considerablemente mejor la topología de los datos del conjunto de datos toroide con respecto a los otros métodos, aunque nuevamente, ISOMAP no es mejor que los otros métodos en la evaluación para todos los vecindarios, sino que, como se observa en dicha figura, PCA lo supera en la topología global y LE en algunas parte de la topología local concordando así, con lo mencionado en la sección **2.2.3.2.1**, en donde se afirma que LLE busca conservar mejor la geometría local de los datos, mediante el uso de incrustamientos no lineales basados en grafos.

La incrustación generada en la Figura 58, permite percibir de una mejor forma que las anteriores, las diferentes formas en la que un objeto (representado por datos) puede ser deformado o transformado, en la Figura 60, se presenta una configuración diferente para los métodos LLE e ISOMAP, siendo ésta, el cambio de vecindarios a 12.

Figura 60: Evaluación LLE (12 vecindarios), ISOMAP (12 vecindarios) con 2 dimensiones cada uno.

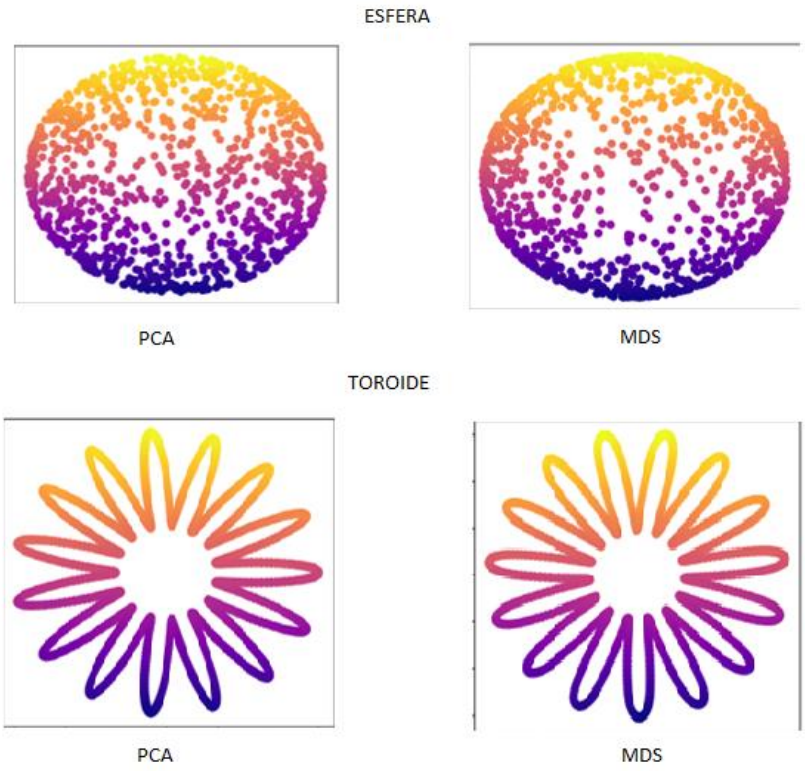


Fuente: Investigación propia

Para vecindarios de 10 y 12 en la esfera, la variación es mínima, aunque se pueden realizar otras configuraciones, es posible que la preservación no varíe mucho para este conjunto de datos, algo similar pasa en el Rollo Suizo, que, aunque, ISOMAP tenga un mejor rendimiento con 12 vecindarios que con 10, dicho cambio es solo de 1%.

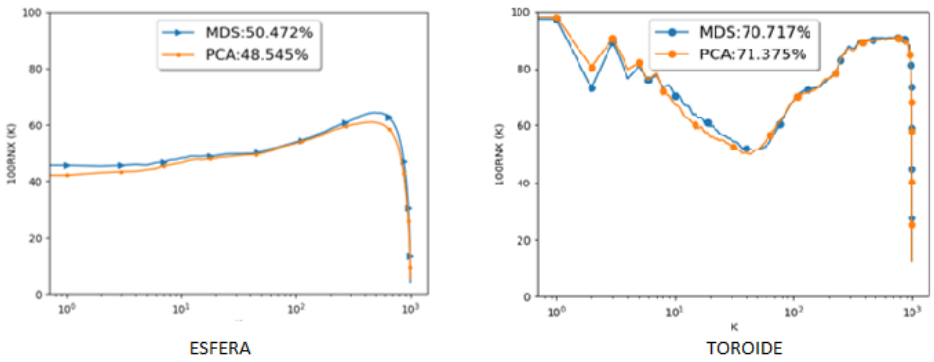
Con el fin de evaluar el comportamiento de los métodos globales, siendo estos PCA y MDS, se hicieron experimentos de estos dos métodos en los conjuntos de datos Esfera y Toroide, MDS fue introducido en la sección **2.2.3.1.2**, en donde se puede apreciar que su tiempo de complejidad es muy grande cuando la cantidad de registros también lo es, por lo que para estos experimentos se decidió utilizar estos conjuntos de datos con 1000 puntos solamente.

Figura 61: incrustamiento de la esfera generado por PCA y MDS



Fuente: Elaboración propia

Figura 62: Evaluación PCA y MDS en conjuntos de datos Esfera y Toroide

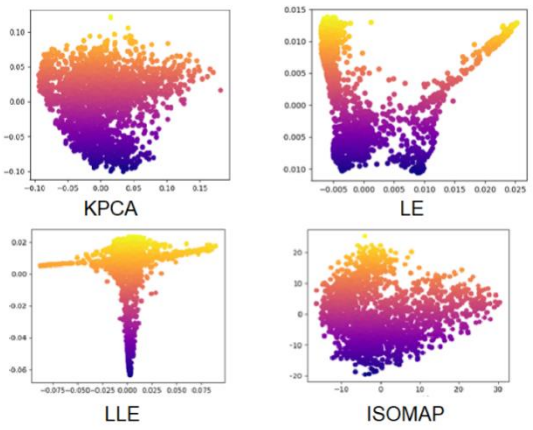


Fuente: Elaboración propia

Tal como se puede observar en la Figura 62, PCA y MDS tienen preservaciones topológicas muy cercanas en ambos conjuntos de datos, en el caso del conjunto de datos de la Esfera, MDS tiene una mejor preservación que PCA y en la del Toroide sucede al contrario, cabe mencionar que la preservación de ambos métodos en Toroide es más alta que en la de la Esfera, pudiendo deducir que probablemente sea mejor usar otros métodos con mejor rendimiento en el conjunto de datos de la Esfera.

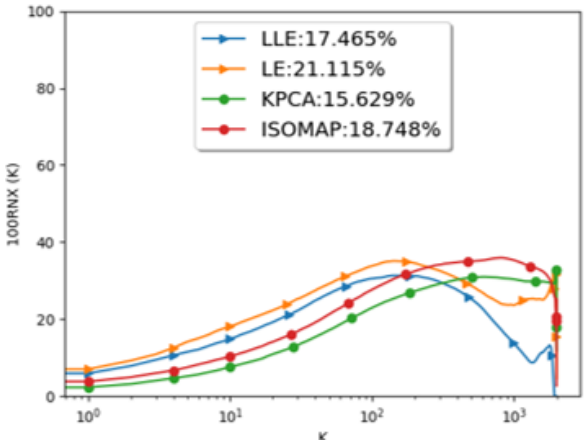
Finalmente, se evaluaron los métodos RD con algunos conjuntos de datos reales, algunos de estos son imágenes tomadas a objetos en diferentes ángulos, algunos otros son fotos del rostro del autor del conjunto de datos, inicialmente, estas imágenes no pueden ser utilizadas por los métodos RD, teniendo éstas que ser pre procesadas para obtener una matriz representativa de las imágenes, por lo que la cantidad de dimensiones suele ser muy alta y por lo tanto, los métodos RD son muy útiles para realizar esa reducción de dimensión, primeramente, en la Figura 64 se realizó la evaluación con el conjunto de datos MNIST, debido a su alta dimensionalidad, este conjunto de datos no puede ser visualizado con métodos de visualización convencionales, hasta que con los métodos RD se reduzca su dimensión a 3 o 2 dimensiones, tal como se muestra en la Figura 63.

Figura 63: Encrustamiento de KPCA, LE, LLE e ISOMAP conjunto de datos MNIST.



Fuente: Elaboración propia

Figura 64: Evaluación de KPCA, LE, LLE e ISOMAP en MNIST.

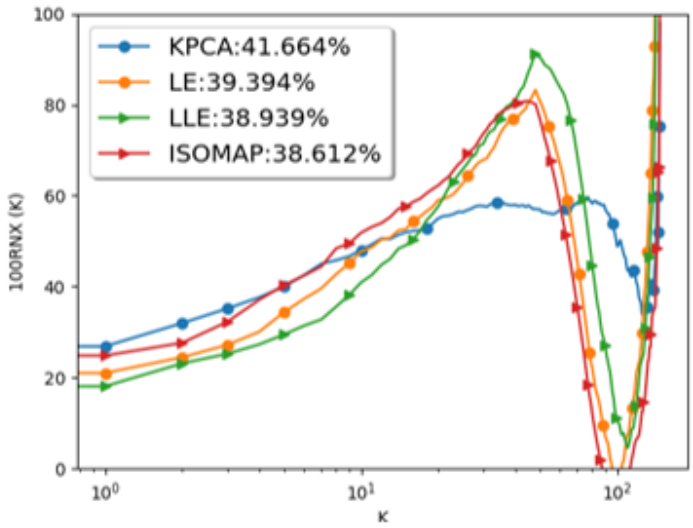


Fuente: Elaboración propia

Como se ha mencionado en secciones anteriores, una de las ventajas de utilizar conjuntos de datos artificiales, es la capacidad que se tiene para observar la transformación del objeto, esto no puede ser percibido con la misma claridad en conjuntos de datos reales, así mismo, las métricas RNX, evalúan preservación topológica y por lo tanto podrían haber características de los métodos que no se observan con las métricas y por lo tanto, los resultados de las evaluaciones no son tan altas como en los otros conjuntos de datos, aun así, RNX, siguen siendo una de las métricas más utilizadas para evaluar los métodos RD y en este caso se obtuvo que LE tuvo una mejor preservación topológica de los datos, siendo esto debido a su proveniencia matemática que tal como se mencionó en **2.2.3.2.3**, LE utiliza grafos realizando un proceso de preservación de localidades pequeñas muy minucioso, teniendo en cuenta las aristas entre cada punto especificado, definiendo así también los vecindarios.

Para finalizar, se hicieron experimentos con el conjunto de datos de Fray Faces introducido en la sección 3.7 Figura 16 y con el de Iris introducido en la misma sección. Algo que tener en cuenta con el conjunto de datos Iris, es que este contiene en su última columna los datos respuesta, para este experimento, se quitó esta columna con el particionador dejando las primeras 3 columnas con valores numéricos.

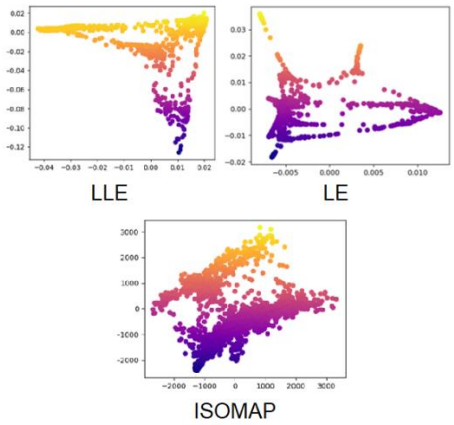
Figura 65: Evaluación de KPCA, LE, LLE e ISOMAP en conjunto de datos Iris



Fuente: Elaboración propia

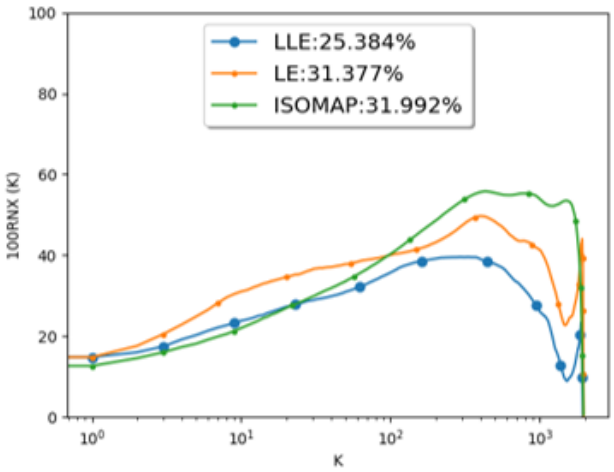
Como se puede observar en la anterior figura, los métodos tienen preservaciones similares para este conjunto de datos, sin embargo, es KPCA el que logra preservar mejor los datos originales, algo interesante es que, aunque LE tiene menor preservación en la topología local que ISOMAP, el promedio obtenido en la global le permite obtener una mayor preservación general que los otros métodos que no son KPCA, contrastando con la literatura científica y académica, en donde se menciona que LE por naturaleza preserva mejor la topología local.

Figura 66: Incrustamiento de LLE, LE e ISOMAP en conjunto de datos Fray Faces



Fuente: Elaboración propia

Figura 67: Evaluación de LLE, LE e ISOMAP en Fray Faces.



Fuente: Elaboración propia

En el anterior experimento, ISOMAP vuelve a obtener un mejor rendimiento que los otros métodos, sin embargo, LE tiene un rendimiento muy similar y además, tuvo una mejor preservación en la topología local, a lo largo de este trabajo, se ha mencionado como métodos RD locales deberían tener una mejor preservación en la localidad, sin embargo, con los experimentos se ha conocido que no siempre es así, todo depende del conjunto de datos y de los vecindarios elegidos para los métodos RD, RNX permite observar los comportamientos esperados e inesperados, y teniendo en cuenta que el objetivo del estudio es evaluar los métodos RD teniendo en cuenta todas las alternativas y resultados obtenidos, RNX sin duda alguna se vuelve en una métrica fundamental en este proceso evaluativo.

6 CONCLUSIONES

- Los métodos RD que utilizan conceptos matemáticos no lineales como los grafos o los kernel, tienden a conservar mejor la geometría local de los datos, siendo la situación contraria, los métodos que utilizan funciones lineales como PCA y MDS que conservan mejor la topología global, por lo que dependiendo de la heurística del método RD, este tendrá un rendimiento y comportamiento diferente.

- Aunque los métodos globales investigados conservan la topología global de los datos en la mayoría de los casos, así mismo como los locales conservan la geometría local, existen situaciones en donde dicho criterio no se cumple, como por ejemplo cuando LE preservó mejor la topología global que la local de los datos tal como se observa en la Figura 65, esto puede deberse a los distintos tipos de configuración con los que los métodos RD pueden ser probados, aunque en la presente investigación se ofrecieron algunas configuraciones para los nodos como el parámetro de vecindarios, existen otros parámetros que pueden ajustar mejor el rendimiento.

- La integración del lenguaje de programación Python, los diferentes paradigmas de programación que este ofrece y el framework PyQt5 que permite la creación de interfaces gráficas, han sido claves en el desarrollo de la herramienta para la evaluación de la preservación topológica de los datos y así mismo, el concepto Drag and Drop adoptado por la herramienta, provee un producto fácil de usar e intuitivo para el usuario.

- Las métricas RNX implementadas en la herramienta y provenientes de la literatura científica, son capaces de evaluar tanto las intrusiones como las extrusiones generadas por los métodos RD a la vez, algo que otras métricas como la de error de conservación de vecindarios introducidas en **2.2.7.2** y promedio de conservación de vecindarios introducida en **2.2.7.3,** no son capaces de hacer, así mismo, la curva generada ofrece un buen indicador para determinar tanto el rendimiento como el comportamiento de los métodos RD de forma precisa, eliminando la base aleatoria generada por QNX.

7 RECOMENDACIONES

- Debido a la naturaleza de la investigación, en donde fue necesario ahondar más profundamente en la fundamentación heurística y matemática tanto de los métodos RD como de las métricas implementadas, se cree necesario fomentar entre los estudiantes la importancia de entender las bases de cualquier tópico de investigación.

- La herramienta desarrollada es de fácil escalabilidad, por lo que está abierta al desarrollo de más módulos y funciones que implementen otros pasos de KDD no implementados en esta investigación, de tal forma que la continuidad del desarrollo e investigación de la herramienta puede quedar en manos de otros estudiantes.

- Las visualizaciones de la herramienta para conjuntos de datos con alta dimensionalidad son limitadas, por lo que se recomienda investigar e implementar más métodos de visualización para estos casos.

- Hay situaciones en donde los conjuntos de datos no son aptos para uso de los métodos RD, por lo que se usan procesos de limpieza de datos para su preparación, la actual herramienta no implementa estos procesos ampliamente, por lo que implementarlos es recomendable.

BIBLIOGRAFIA

ANAYA, Andrés; PERLUFO, Diego; ALVARADO, Juan; RIOS, Jorge; CASTRO, Juan; ROSERO, Paul; PEÑA, Diego; SALAZAR, José y UMAQUINGA, Ana. Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: LDA versus PCA. 2016. Disponible en: https://pdfs.semanticscholar.org/57ca/0d35fd5610f3d5d3c3a8486143495caf152f.pdf

PRIETO, Carlos. Adaptación de las metodologías tradicionales cascada y espiral para la inclusión de evaluación inicial de usabilidad en el desarrollo de productos software en México, 2015. Tomado de: http://repositorio.utm.mx/bitstream/123456789/72/1/2015-MMI-CGPA.pdf

ARCE, Constantino; DE FRANCISCO, Cristina. Escalonamiento multidimensional: Concepto y aplicaciones, 2010. Tomado de: https://www.researchgate.net/publication/41734044_Escalamiento_multidimensional_Concepto_y_aplicaciones

ASPGEMS. I Metodología de desarrollo de software (III)-Modelo en Espiral. 2019 tomado de: https://aspgems.com/metodologia-de-desarrollo-de-software-iii-modelo-en-espiral/

AYESHA, Shaeela; KASHIF, Muhammad y TALIB, Ramzan. Overview and comparative study of dimensionality reduction techniques for high dimensional data, 2020. Disponible en: https://www.sciencedirect.com/science/article/pii/S156625351930377X

BERTINI, Enrico; TATU, Andrada; KEIM, Daniel. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, IEEE, 2011. Disponible en: https://bib.dbvis.de/uploadedFiles/350.pdf

BELKIN, Mikhail y NIYOGI, Partha. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, 2001. Tomado de: http://papers.nips.cc/paper/1961-laplacian-eigenmaps-and-spectral-techniques-for-embedding-and-clustering.pdf

BIEHL, Michael y HAMMER, Barbara. How to evaluate dimensionality reduction? - Improving the co-ranking matrix tomado de: https://www.researchgate.net/publication/51946294_How_to_Evaluate_Dimensionality_Reduction_-_Improving_the_Co-rankingMatrix

BODT, Cyril; MULDERS, Doumia; VERLEYSEN, Michel y LEE, Jonh. Nonlinear dimensionality reduction with missing data using parametric multiple inputations. En: IEEE, 2018, p.1166-1179. Disponible en: https://ieeexplore.ieee.org/document/8447227

BUJA, Andreas y SWAYNE, Deborah, el al. Data visualization with multidimensional Scaling. tomado de: https://www.tandfonline.com/doi/abs/10.1198/106186008X318440

CHEN, Lisha; BUJA, Andreas. Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis, 2006. Tomado de: https://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0111

FAYYAD, U.; SHAPIRO, G. y SMYTH, P. From Data Mining to Knowledge Discovery in Databases. 1996. Disponible en https://doi.org/10.1609/aimag.v17i3.1230

FISHER, R. Iris Data set. 1936 tomado de: http://archive.ics.uci.edu/ml/datasets/Iris

FODOR, Imola. A survey of dimensionality reduction techniques. 2002. Tomado de: https://www.cc.gatech.edu/~isbell/tutorials/dimred-survey.pdf

FRANCE, Stephen y AKKUCUK, Ulas. A review, Framework, and toolkit for Exploring, Evaluating, and Comparing Visualization Methods. En: ResearchGate, 2020. Disponible en: https://www.researchgate.net/publication/331318654_A_Review_Framework_and_R_toolkit_for_Exploring_Evaluating_and_Comparing_Visualization_Methods

FREY. Data for matlab hackers. Tomado de: https://cs.nyu.edu/~roweis/data.html

HAN, Jiawey; KAMBER, Michelin y PEI, Jian. Minería de datos: Concepts and Techniques. 3 ed. Elsevier, 2011. Disponible en: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

HU, Xiaohua; SASKATCHEWAN, Regina. KNOWLEDGE DISCOVERY IN DATABASES AN ATTRIBUTEORIENTED ROUGH SET APPROACH, 1995. Tomado de: https://wiki.eecs.yorku.ca/course_archive/2013-14/F/4403/_media/huphd.pdf

KUANG, Liwei; ZHANG, Laurence, et al. A Holistic Approach to Distributed Dimensionality Reduction of Big Data. IEEE, 2015 tomado de: https://ieeexplore.ieee.org/abstract/document/7134729

KNIME. Historia de código abierto [Sitio web] Universidad de Konstanz, Alemania; [Consultado: 19 de octubre de 2021]. Disponible en: https://www.knime.com/knime-open-source-story

LEE, John y VERLEYSEN, Michel. Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing, 2009, Vol.72, p. 1431-1443. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0925231209000101

LEE, Jonh; PELUFFO, Diego y VERLEYSEN, Michel. Multi-scale similarities in stochastic neigbout embedding: reducing dimensionality while preserving both local and global structure. Neurocomputing, 2015, Vol.169, p. 246-261. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0925231215003641

LEE, John; VERLEYSEN, Michel. Nonlinear dimensionality reduction, Springer Link, 2007 Tomado de: https://link.springer.com/book/10.1007/978-0-387-39351-3

LEE, John, RENARD, Emilie, et al. Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. 2013. Tomado de: https://perso.uclouvain.be/michel.verleysen/papers/neurocomputing13jl.pdf

LEE, John. Nonlinear Dimensionality Reduction: Towards better scalability, 2016. Tomado de: http://pagesperso.litislab.fr/rherault/resources/Rouen_NLDR_2016_JL.pdf

LECUN, Yann; et al. Gradient-Based learning applied to document recognition, 1998. Tomado de: https://www.researchgate.net/publication/2985446_Gradient-Based_Learning_Applied_to_Document_Recognition

LECUN, Yann; et al. THE MNIST DATABASE [en linea]. Of Handwritten digits. Nueva York; [Consultado: 20 de octubre de 2021]. Disponible en: http://yann.lecun.com/exdb/mnist/

NIGRO, Héctor; XODO, Daniel; CORTI, Gabriel; TERREN, Damián. KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario. Disponible en http://sedici.unlp.edu.ar/bitstream/handle/10915/21220/Documento_completo.pdf?sequence=1

OLIPHANT, Travis. Python for scientific computing, IEEE, 2007. Tomado de: https://ieeexplore.ieee.org/abstract/document/4160250

ORANGE. License [Sitio web]. Eslovania; [Consultado: 20 de octubre de 2021]. Disponible en: https://orangedatamining.com/

PANG, Bo; LEE, Lillian y VAITHYANATHAN, Shivakumar. Thump up? Sentiment classification using Machine learning techniques. Asociation for computational linguistics: Philadelphia, Julio de 2002, pp. 79-86. Disponible en: https://www.aclweb.org/anthology/W02-1011.pdf

PEÑA, Diego; SALAZAR José; PELUFFO, Diego; ROSERO, Paul; OÑA, Omar; ANAYA, Andrés; ALVARADO, Juan y THERON, Roberto. Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction. En: ResearchGate, 2016. Disponible en: https://www.researchgate.net/profile/Diego_Peluffo/publication/310464991_Interactive_visualization_methodology_of_high-dimensional_data_with_a_color-based_model_for_dimensionality_reduction/links/5c281bb1a6fdccfc70713111/Interactive-visualization-methodology-of-high-dimensional-data-with-a-color-based-model-for-dimensionality-reduction.pdf

PELUFFO, Diego; et al. Multiple Kernel Learning for Spectral Dimensionality Reduction, Colombia, 2015. Tomado de: https://link.springer.com/content/pdf/10.1007/978-3-319-25751-8_75.pdf

CHALLENGER-PÉREZ, Ivet; DÌAZ-RICARDO, Yanet y BECERRA-GARCÌA, Roberto. El lenguaje de programación python, Cuba, 2014. Tomado de: https://www.redalyc.org/pdf/1815/181531232001.pdf

PROGRAMA DE INGENIERÍA DE SISTEMAS. Líneas de Investigación, ingeniería de sistemas. San Juan de Pasto: Universidad Cesmag, 2020. Disponible en: https://drive.google.com/file/d/1e4lX8IMl3BRPBm5CI8qsZJMbaCuuYPjC/view?ts=5edaae13&gt

RIQUELME, José; RUÍZ, Roberto y GILBERT, Karina. Minería de datos conceptos y tendencias. En: Revista iberoamericana de inteligencia artificial. Valencia: Asociación española para la inteligencia artificial, 2006, vol. 10, núm. 29, pp 11-18, ISSN: 1137-3601. Disponible en: https://www.redalyc.org/pdf/925/92502902.pdf

ROSERO, Paul; DÍAZ, P.; SALAZAR, Jose, *et al.* Interactive data visualization using dimensionality reduction and similarity-based representations. En: ResearchGate, 2017. Disponible en: https://www.researchgate.net/publication/313787026_Interactive_Data_Visualization_Using_Dimensionality_Reduction_and_Similarity-Based_Representations

ROJAS, R y BOUCCHECHTER, I. Ingeniería del Software. Ciclo de vida, Mallorca, España. (2005).

RUSSOM, Philip. Big Data analytics. 2011. Tomado de: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

SALAZAR, José; PEÑA, Diego; PELUFFO, Diego; ROSERO, Paul; DOMINGUEZ, Mauricio; ALVARADO, Juan y THERON, Roberto. Dimensionality reduction for interactive data visualization via Geo-Desic approach. En: ResearchGate 2016. Disponible en: https://www.researchgate.net/profile/Jose_Salazar_Castro/publication/315853808_Dimensionality_reduction_for_interactive_data_visualization_via_a_Geo-Desic_approach/links/5a297977aca2728e05dad422/Dimensionality-reduction-for-interactive-data-visualization-via-a-Geo-Desic-approach.pdf

SALAZAR, Jose; BASANTE, Cielo; PEÑA, Diego y CRUZ, Lilian. Generalized Low-Computational Cost Laplacian Eigenmaps. En: ResearchGate, pp. 661-669, 2018. Disponible en: https://www.researchgate.net/publication/328821143_Generalized_Low-Computational_Cost_Laplacian_Eigenmaps_19th_International_Conference_Madrid_Spain_November_21-23_2018_Proceedings_Part_I

SACHA, Dominik; ZHANG, Leishi, et al. Visual interaction with dimensionality reduction: A structured literature analysis, 2016. Disponible en: https://ieeexplore.ieee.org/document/7536217

SAUL, Lawrence y ROWEIS, Sam. An introduction to localy linear embedding. 2001. tomado de: https://cs.nyu.edu/~roweis/lle/papers/lleintro.pdf

SAGIROGLU, Seref; SINANC, Duygu. Big Data: A review, IEEE, 2013. Tomado de: https://ieeexplore.ieee.org/abstract/document/6567202

DE SILVA, Vin y TENENBAUM, Joshua. Global vs Local methods in Nonlinear Dimensionality Reduction. 2003. Disponible en: http://papers.nips.cc/paper/2141-global-versus-local-methods-in-nonlinear-dimensionality-reduction.pdf

SHLENS, Jonathon. A tutorial on principal component analysis, 2014. Tomado de: https://arxiv.org/pdf/1404.1100v1.pdf

STADLER, Marta. ¿Qué es la topología?. 2002. Tomado de: http://www.ehu.eus/~mtwmastm/sigma20.pdf

TINOCO, et al. A data-driven approach to develop physically soundpredictors: Application to depth-averaged velocities on flowsthrough submerged arrays of rigid cylinders, 2015. Tomado de: https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1002/2014WR016380

THIPPA, G.; KUMAR, M., et al. Analysis of dimensionality reduction techniques on Big Data. IEEE, 2020. Tomado de: https://ieeexplore.ieee.org/abstract/document/9036908

University of Waterloo. What is topology? Tomado de: https://uwaterloo.ca/pure-mathematics/about-pure-math/what-is-pure-math/what-is-topology

UMAQUINGA, Ana; PERLUFO, Diego; ROSERO, Paul; CABRERA, M.; ALVARADO, Juan y ANAYA, Andrés. Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva. En: ResearchGate, 2016. Disponible en: https://www.researchgate.net/publication/316921329_Propuesta_de_analisis_visual_de_datos_en_Big_Data_usando_reduccion_de_dimension_interactiva

VALENCIA, Juliana; DAZA, Genaro, et al. Comparación de Métodos de Reducción de Dimensión. Basados en Análisis por Localidades. En: Dialnet, 2010. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=5062986

VANDERPLAS, Jake y CONNOLLY, Andrew. REDUCING THE DIMENSIONALITY OF DATA: LOCALLY LINEAR EMBEDDING OF SLOAN GALAXY SPECTRA, 2009. Tomado de: https://iopscience.iop.org/article/10.1088/0004-6256/138/5/1365

VENNA, Jarkko y KASKI, Samuel. Neigborhood preservation in nonlinear projection methods: An experimental study. 2001. Tomado de: https://aaltodoc.aalto.fi/bitstream/handle/123456789/2875/article2.pdf?sequence=3&isAllowed=y

WANG, Jianzhong. Geometric Structure of High Dimensional Data and Dimensionality Reduction. 2012. p. 52 Disponible en: https://link.springer.com/book/10.1007/978-3-642-27497-8

WANG, Quang. Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models. 2014. Tomado de: https://arxiv.org/pdf/1207.3538.pdf

WU, Xindoung; ZHU, Xingquan. Data mining with Big Data, IEEE, 2014. Tomado de: https://ieeexplore.ieee.org/abstract/document/6547630/

WEKA. Weka 3: Machine Learning Software in Java [Sitio web]. Nueva Zelanda; [Consultado: 19 de octubre de 2021]. Disponible en: https://www.cs.waikato.ac.nz/ml/weka/

WU, Dekai; SU, Weifeng y CARPUAT, Marine. A. Kernel PCA Method for Superior Word Sense Disambiguation, 2004. Tomado de: https://www.aclweb.org/anthology/P04-1081.pdf

Anexo A: Manual de usuario

EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS R_{NX}

Manual de Usuario

Autores:

CARLOS DAVID CORREA LOZANO
JUAN ANDRÉS LOZANO THOMÉ
DIEGO FERLEY URREA BURGOS

Versión: 0100

Contenidos

Descripción del software

La herramienta QARNX desarrollada en el proyecto denominado EVALUACIÓN DE METODOS DE REDUCCION DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS RNX, se presenta como una herramienta para evaluar el rendimiento y calidad de los resultados obtenidos de los algoritmos de reducción de dimensionalidad con un data set proporcionado, usando métricas de evaluación como las curvas RNX, las cuales permiten analizar la preservación topológica de los datos y cómo estos pueden comportarse.

La herramienta proporciona diversos módulos que permiten al usuario realizar procesos que se encuentran en la metodología KDD para el descubrimiento de conocimiento en bases de datos, mediante la interacción de una interfaz drag and drop, para crear flujos que analicen diversos métodos de reducción con las curvas RNX.

Interfaces del software

Interfaz principal

A continuación, se describe la interfaz principal de usuario y las herramientas que la componen.

Captura de pantalla

Descripción de la funcionalidad de la interfaz

Interfaz principal para el control de los flujos de trabajo que se necesiten crear.

Descripción de los elementos de la interfaz

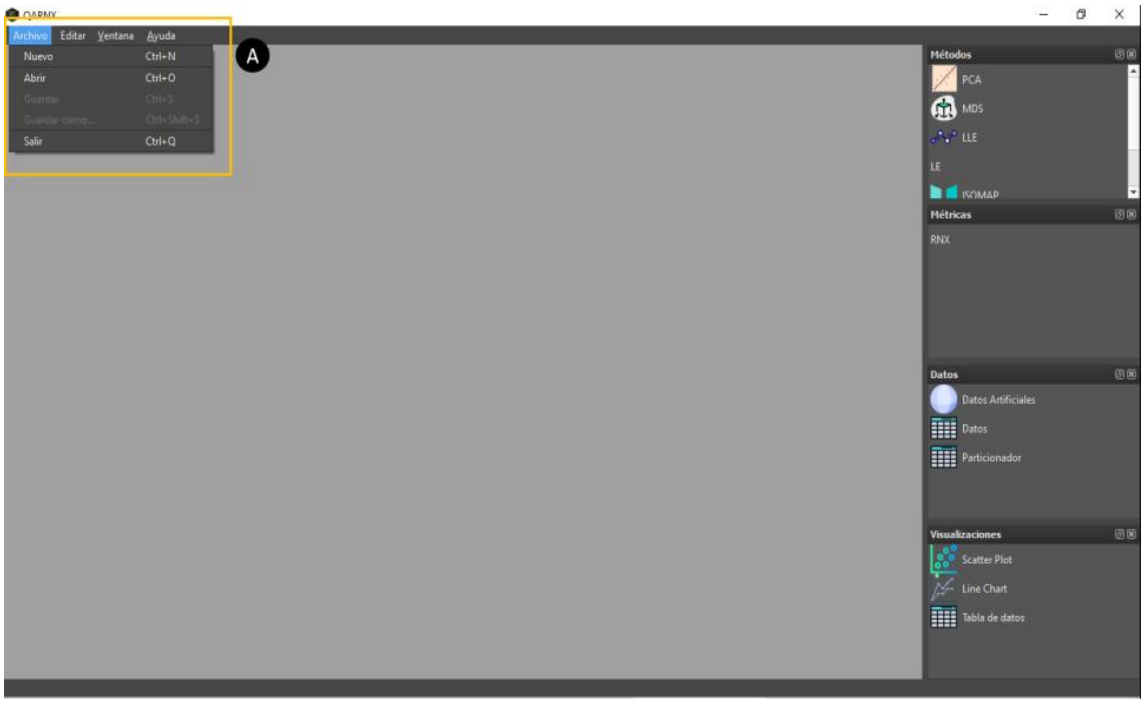
A. Menú de opciones para control de la herramienta.

B. Barra de herramientas para utilizar en las zonas de trabajo creadas.

Interfaz menú control de espacios de trabajo

A continuación, se describe el menú de control de espacios de trabajo.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El menú de archivo permite crear nuevos espacios de trabajo, guardar espacios de trabajo realizados y abrir espacios de trabajo previamente almacenados.

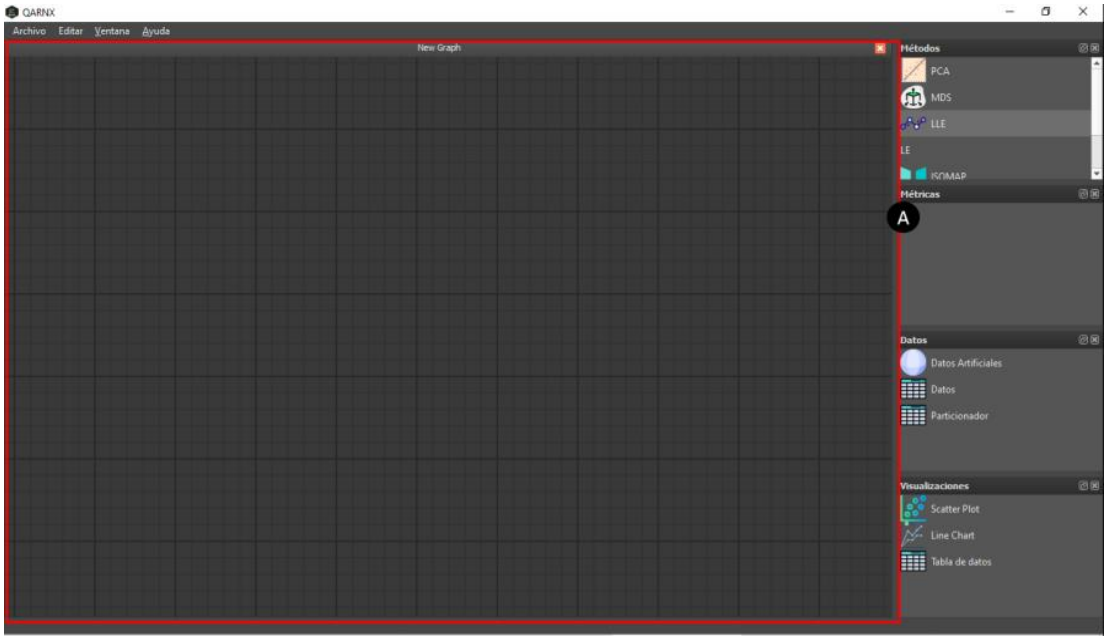
Descripción de los elementos de la interfaz

A. Menú de opciones para abrir, guardar o crear flujos de trabajo.

Interfaz espacio de trabajo

A continuación, se describe la interfaz de espacio de trabajo.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

Los espacios de trabajo permiten añadir elementos desde la barra de herramientas.

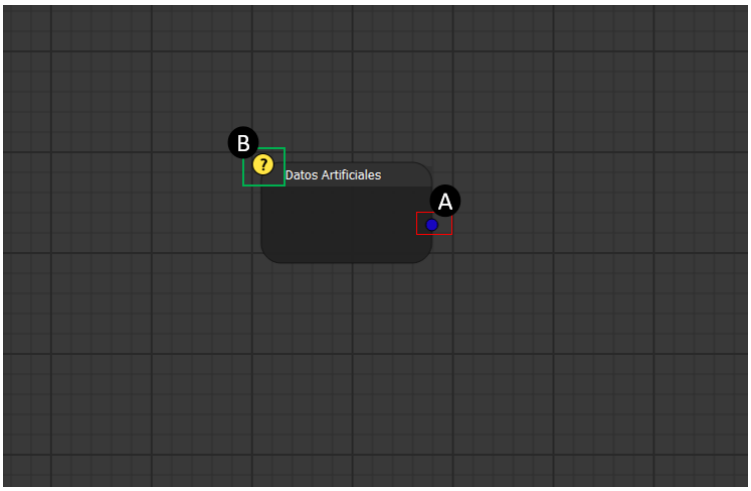
Descripción de los elementos de la interfaz

A. Espacio de trabajo para la creación de los workflows.

Interfaz nodo datos artificiales

A continuación, se describe la interfaz de nodo de datos artificiales.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El nodo de datos artificiales permite cargar datos de prueba disponibles en la herramienta como los son la esfera, rollo suizo y toroide.

Descripción de los elementos de la interfaz

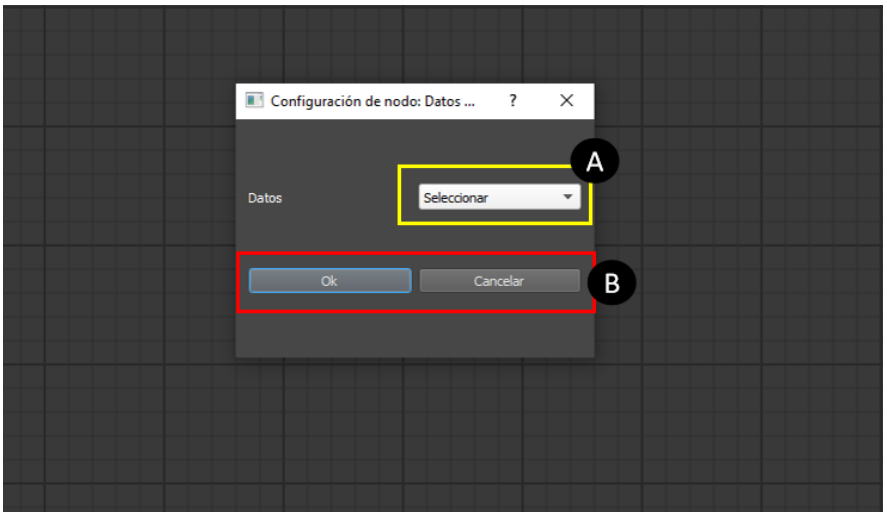
A. Salida de los datos cargados.

B. Icono modal para visualizar el estado del nodo.

Interfaz configuración nodo datos artificiales

A continuación, se describe la ventana de configuración de nodo de datos artificiales.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

La ventana de configuración permite escoger entre los tres diferentes datasets disponibles en la herramienta las opciones a escoger son Esfera, Toroide y Rollo Suizo.

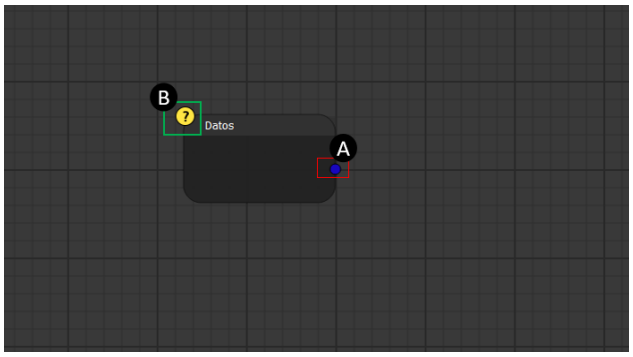
Descripción de los elementos de la interfaz

A. Selector de datasets de pruebas.

B. Botones para cancelar o guardar la configuración establecida.

Interfaz nodo datos reales

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El nodo de datos reales permite al usuario cargar datasets que tenga guardados en su dispositivo, los archivos soportados son: mat, .csv, .xlsx.

Descripción de los elementos de la interfaz

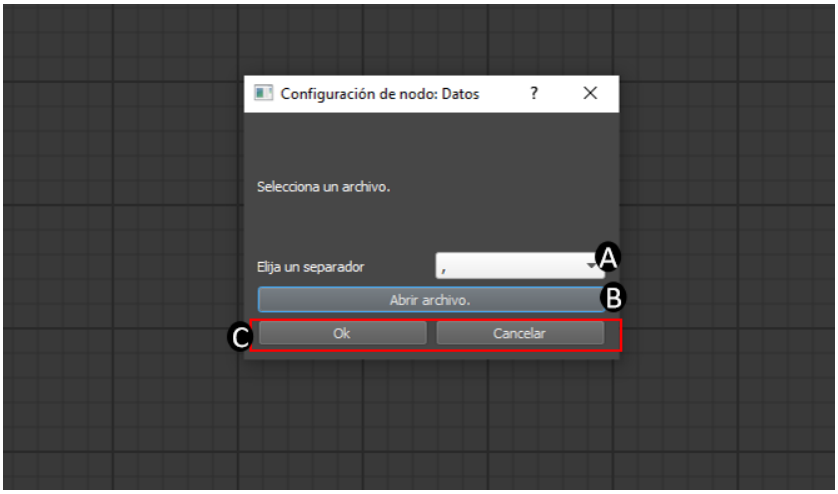
A. Salida de los datos cargados.

B. Icono modal para visualizar el estado del nodo.

Interfaz configuración nodo datos reales

A continuación, se describe la ventana de configuración del nodo de datos reales

Captura de pantalla



Descripción de la funcionalidad de la interfaz

La ventana de configuración permite escoger un archivo del equipo que contengan datos que puedan ser evaluados.

Descripción de los elementos de la interfaz

A. Selector de separador con el que se procesa la lectura del archivo seleccionado.

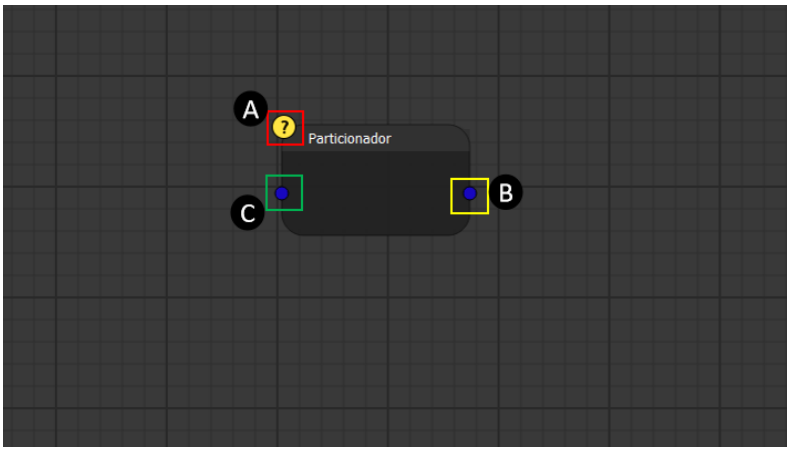
B. Botón abrir el gestor de archivos del sistema para escoger el dataset.

C. Botones para cancelar o guardar la configuración establecida.

Interfaz nodo particionador

A continuación, se describe la interfaz del nodo particionador.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El nodo particionador permite eliminar una columna de un dataset conectado.

Descripción de los elementos de la interfaz

A. Icono modal para visualizar el estado del nodo.

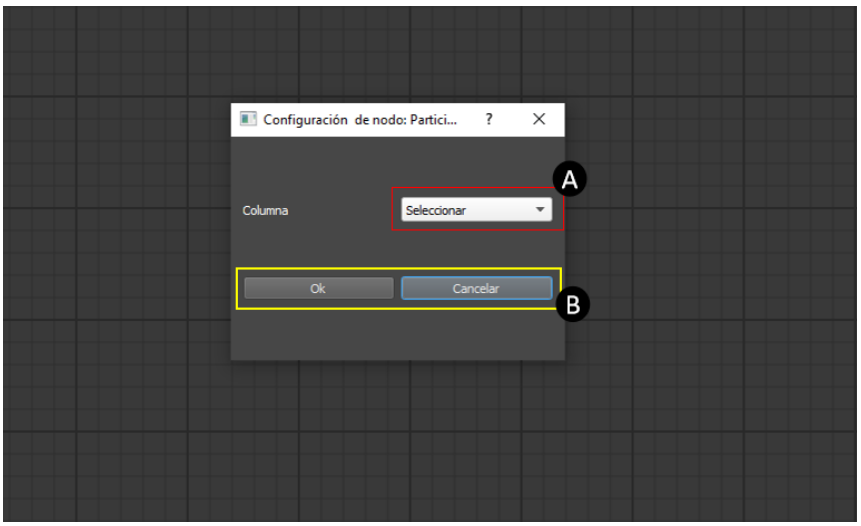
B. Salida del nuevo dataset sin la columna especificada.

C. Entrada del dataset que se quiere limpiar, los nodos que pueden realizar una conexión a esta entrada son artificial data y datos reales.

Interfaz configuración nodo particionador

A continuación, se describe la ventana de configuración del nodo particionador.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

La ventana de configuración permite escoger una columna del dataset cargado al nodo particionador.

Descripción de los elementos de la interfaz

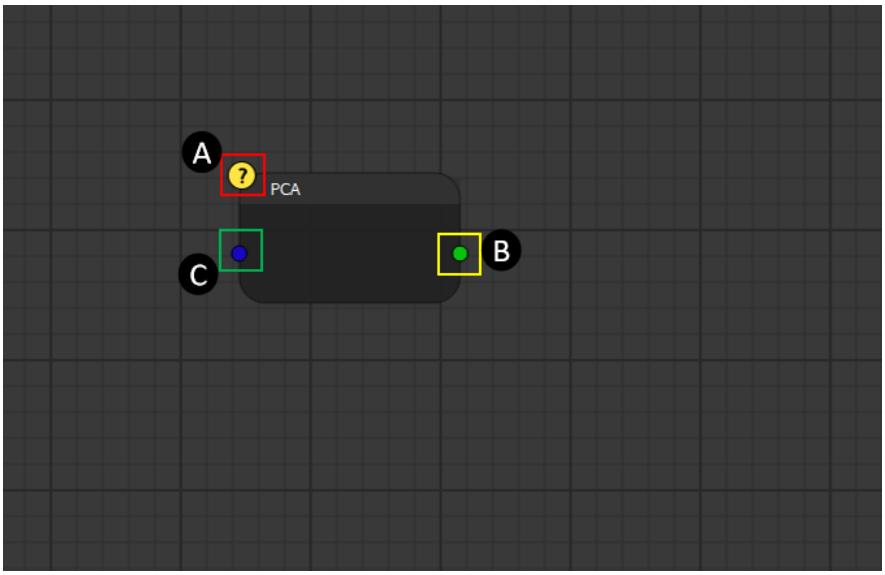
A. Selector de la columna que se eliminara del dataset.

B. Botones para guardar o cancelar la configuración realizada.

Interfaz nodo para métodos de reducción de dimensión no paramétrico

A continuación, se describe nodo de RD no paramétrico.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

Los nodos de RD no paramétricos permiten reducir las dimensiones de un dataset que se haya conectado al nodo.

Descripción de los elementos de la interfaz

A. Icono modal para visualizar el estado del nodo.

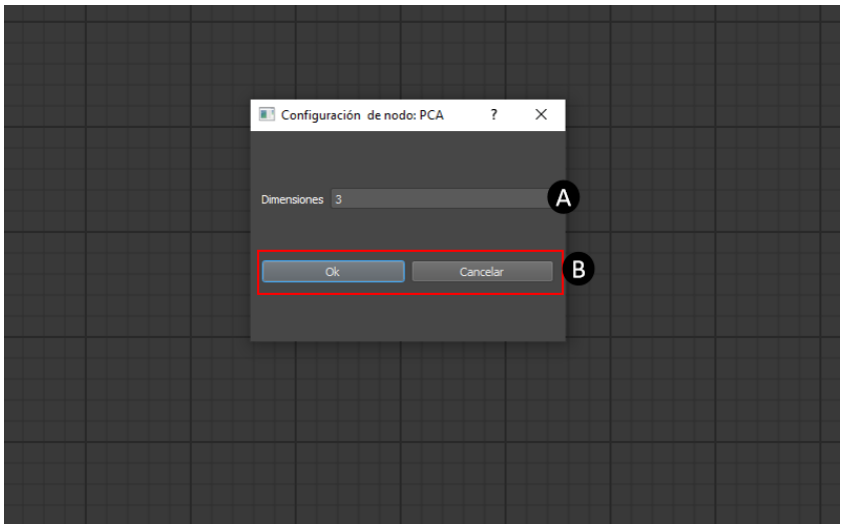
B. Salida de datos en baja dimensión a partir del dataset de entrada.

C. Entrada del dataset en alta dimensión, los nodos que pueden realizar una conexión a esta entrada son artificial data y datos reales.

Interfaz de configuración para nodo de métodos de reducción de dimensión no paramétricos

A continuación, se describe la ventana de configuración de los nodos RD no paramétricos

Captura de pantalla



Descripción de la funcionalidad de la interfaz

La ventana de configuración permite definir la dimensión a la que se desea reducir los datos conectados al nodo.

Descripción de los elementos de la interfaz

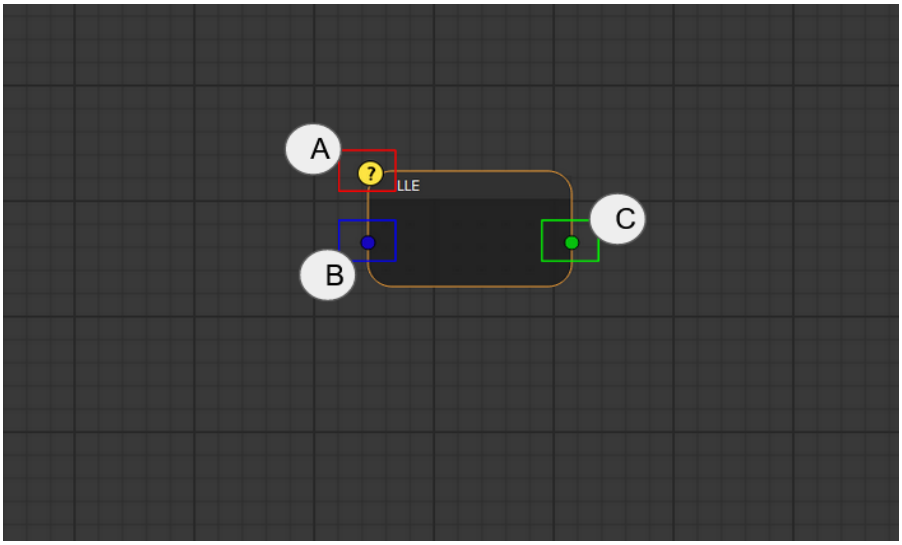
A. Entrada para definir la dimensión a la cual se va a reducir el dataset original.

B. Botones para guardar o cancelar la configuración realizada.

Interfaz nodo para métodos de reducción paramétricos

A continuación, se describe la interfaz nodo RD paramétrico.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

Los nodos de RD no paramétricos permiten reducir las dimensiones de un dataset que se haya conectado al nodo.

Descripción de los elementos de la interfaz

A. Icono modal para visualizar el estado del nodo.

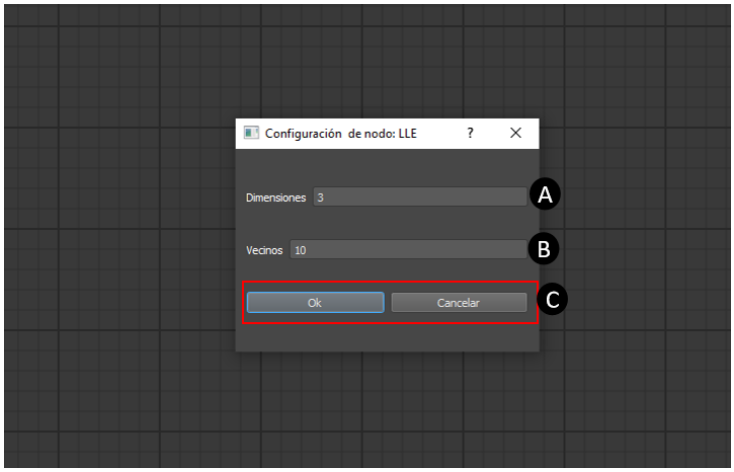
B. Salida de datos en baja dimensión a partir del dataset de entrada.

C. Entrada del dataset en alta dimensión, los nodos que pueden realizar una conexión a esta entrada son artificial data y datos reales.

Interfaz de configuración para nodo de métodos de reducción paramétricos

A continuación, se describe la ventana de configuración de nodo RD paramétrico.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

La ventana de configuración permite definir la dimensión a la que se reducirá el dataset y el valor de los vecindarios que se formaran en el método.

Descripción de los elementos de la interfaz

A. Entrada para definir la dimensión a la que se reducirá el dataset original.

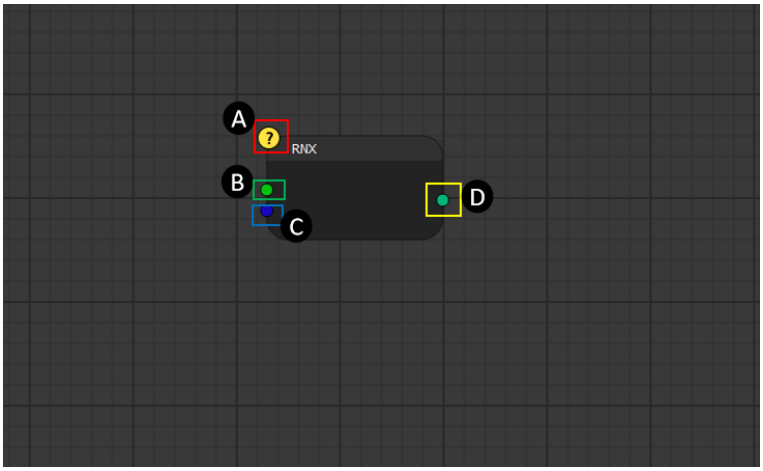
B. Entrada para definir el valor del vecindario que se utilizara en el método.

C. Botones para guardar o cancelar la configuración realizada.

Interfaz nodo métrica RNX

A continuación de describe la interfaz nodo métrica RNX

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El nodo RNX permite evaluar diversos métodos de reducción de dimensión teniendo en cuenta un dataset específico.

Descripción de los elementos de la interfaz

A. Icono modal para visualizar el estado del nodo.

B. Entrada de datos en baja dimensión, los nodos que pueden establecer una conexión a esta entrada son los métodos RD paramétricos y métodos RD no paramétricos.

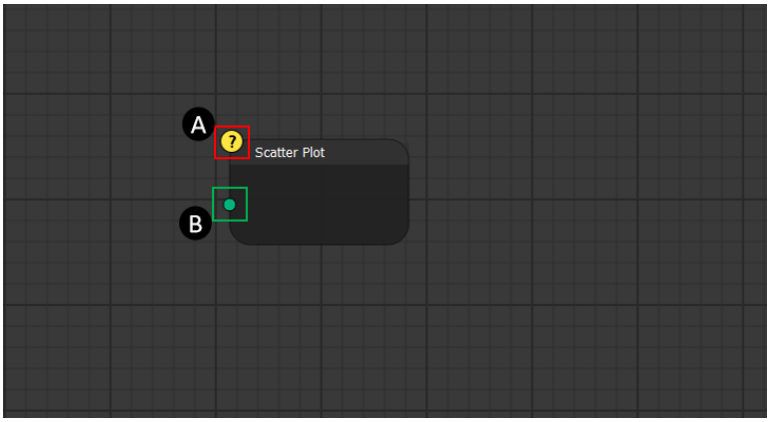
C. Entrada de datos de alta dimensión, los nodos que pueden establecer conexión a esta entrada son nodo datos reales y nodo datos artificiales.

D. Salida de métrica de calidad y curva RNX realizada a los nodos conectados.

Interfaz nodo SCATTER PLOT

A continuación, se describe la interfaz nodo SCATTER PLOT.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El nodo SCATTER PLOT permite visualizar conjuntos de datos en 2 y 3 dimensiones.

Descripción de los elementos de la interfaz

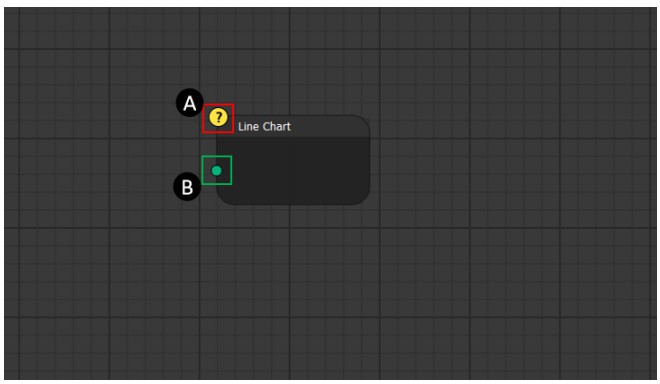
A. Icono modal para visualizar el estado del nodo.

B. Entrada de datos, los nodos que pueden establecer conexión a esta entrada son datos reales, datos artificiales, métodos RD paramétricos, métodos RD no paramétricos.

Interfaz nodo Line chart

A continuación, se describe la interfaz nodo Line Chart.

Captura de pantalla



Descripción de la funcionalidad de la interfaz

El nodo Line Chart permite visualizar las curvas de evaluación generadas por el nodo RNX

Descripción de los elementos de la interfaz

A. Icono modal para visualizar el estado del nodo.

B. Entrada de evaluación RNX, los nodos que pueden establecer una conexión a esta entrada son nodo RNX.

Interfaz nodo tabla de datos

A continuación, se describe la interfaz nodo tabla de datos

Captura de pantalla

Descripción de la funcionalidad de la interfaz

El nodo tabla de datos permite visualizar los datos de entrada en una tabla.

Descripción de los elementos de la interfaz

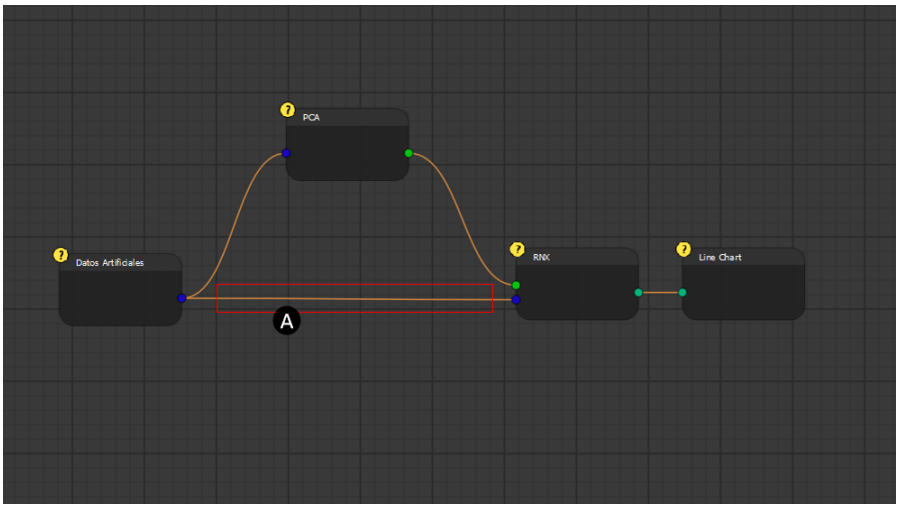
A. Icono modal para visualizar el estado del nodo.

B. Entrada de datos, los nodos que pueden establecer conexión a esta entrada son datos reales, datos artificiales, métodos RD paramétricos, métodos RD no paramétricos.

Interfaz conexión entre nodos

A continuación, se describe la conexión entre nodos en el workflow

Captura de pantalla



Descripción de la funcionalidad de la interfaz

La conexión entre nodos se realiza tomando los puntos de entrada o salida presentes en cado nodo

Descripción de los elementos de la interfaz

A. Línea de conexión entre nodos.

Anexo B: Manual de Sistema

QARNX

Manual de Sistema

Universidad CESMAG

Autores:

Carlos David Correa Lozano

Juan Andrés Lozano Thomé

Diego Ferley Urrea Burgos

1. Datos Generales

Herramienta para evaluación de métodos RD mediante métricas R_{NX}.

1.1. Nombre del Proyecto General

Evaluación de Métodos de Reducción de Dimensión para la Preservación topológica de los Datos Mediante Métricas R_{NX}.

1.2. Título del Software

QARNX

1.3. Tipo de Producción Software

Producción Tecnológica

1.4. Categoría del Software

Minería de datos.

1.5. Tecnología de Despliegue

Para el software QARNX, es necesario elementos de despliegue orientados a aplicaciones de escritorio los cuales son:

1.5.1. Hardware

CPU: Intel i3

Memoria RAM: Mínimo 1GB

1.5.2 Software

Sistema Operativo: Windows, Linux, OS

1.7. Tecnología usada en el Desarrollo

1.7.1. Sistema de Desarrollo

Entorno de desarrollo: PyCharm Community Edition

Sistema de gestión de paquetes: Anaconda

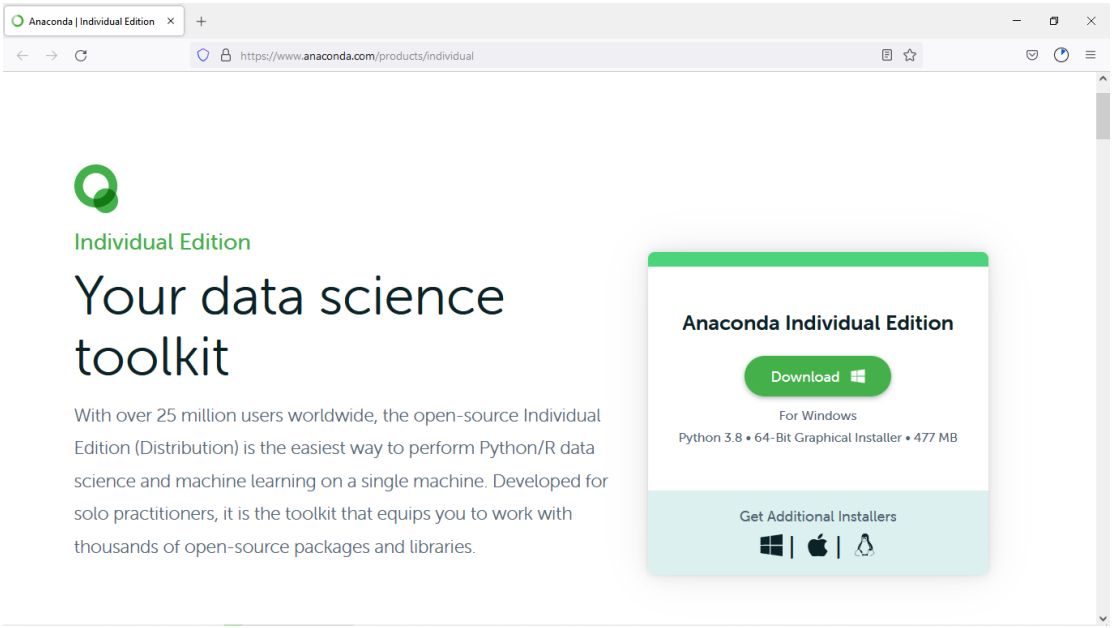
1.7.2. Lenguaje de Programación

- Python

1.8 Configuración del Entorno de Despliegue

1.8.1 Instalación de anaconda

- Descargar instalador de anaconda en: https://www.anaconda.com/products/individual

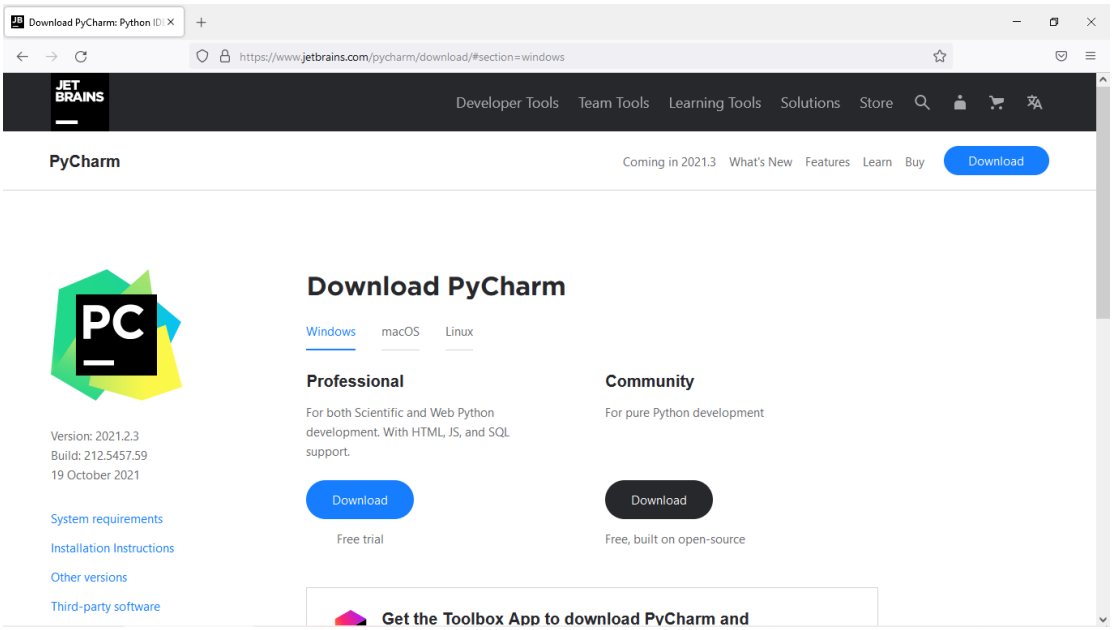


- Seguir instrucciones de instalación del ejecutable descargado

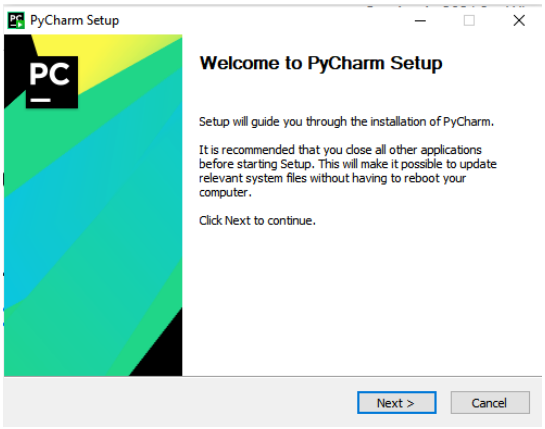
- Probar correcto funcionamiento anaconda en nuestro dispositivo utilizando el comando “pip -V” en la terminal de comandos

1.8.2 Instalación de IDE PyCharm

- Descargar instalador de PyCharm en: https://www.jetbrains.com/pycharm/download/#section=windows

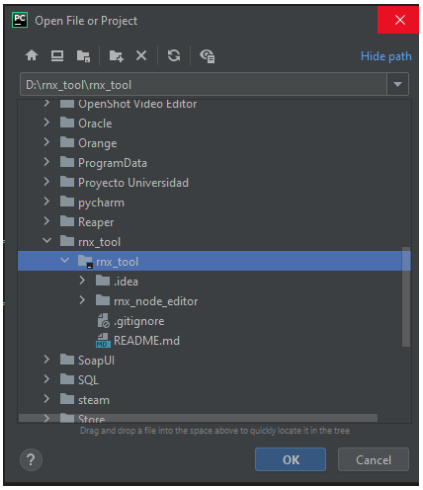


- Seguir instrucciones de instalación del ejecutable descargado

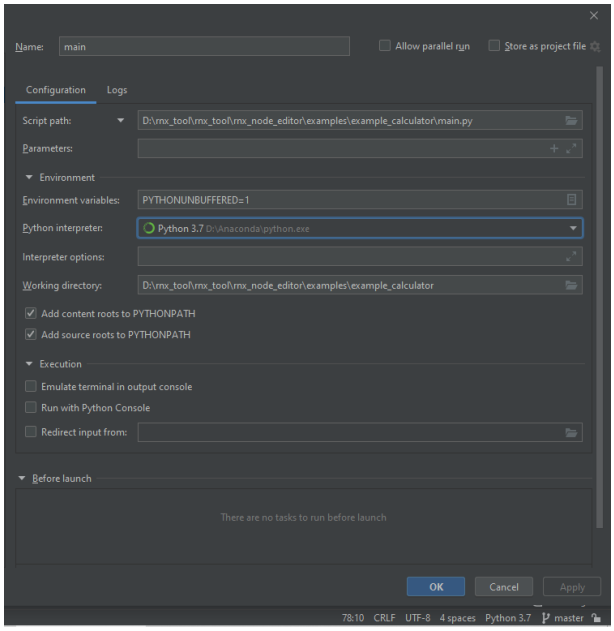


1.8.3 Configuración de entorno de programación

- Abrir desde el editor el proyecto



- Escoger el intérprete para la ejecución del proyecto



2. Participantes

A continuación, se describen a los participantes e interesados en el desarrollo QARNX:

2.1 Organizaciones Participantes

ORG001	
Organización	Universidad CESMAG
Dirección	Cra 20ª #14-54, Pasto, Nariño
Teléfono	7216535
Comentario	Ninguno

2.2 Personas Participantes

STK001	
Nombre	Carlos David Correa Lozano
Organización	Universidad CESMAG
Rol	Investigador
Es desarrollador	Si
Es cliente	Si
Es usuario	No
Comentario	Ninguno

STK001	
Nombre	Juan Andrés Lozano Thomé
Organización	Universidad CESMAG
Rol	Investigador
Es desarrollador	Si
Es cliente	Si
Es usuario	No
Comentario	Ninguno

STK001	
Nombre	Diego Ferley Urrea Burgos
Organización	Universidad CESMAG
Rol	Investigador

Es desarrollador	Si
Es cliente	Si
Es usuario	No
Comentario	Ninguno

STK001	
Nombre	Juan Carlos Alvarado Pérez
Organización	Universidad CESMAG
Rol	Asesor
Es desarrollador	No
Es cliente	Si
Es usuario	No
Comentario	Ninguno

3. Objetivos del Sistema

OBJ001	**Disposición de datos para trabajar**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Descripción	Permitir cargar datos de diversos tipos para su procesamiento en la herramienta
Sub-objetivo	Ninguno
Importancia	Vital
Urgencia	Inmediata

OBJ002	Implementar los métodos de reducción desarrollados en la literatura científica
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Descripción	Permitir procesar datos de alta dimensión a un nuevo data set en baja dimensión
Sub-objetivo	Ninguno
Importancia	Vital
Urgencia	Inmediata
Estado	Implementado
Estabilidad	Alta
Comentario	Ninguno

OBJ003	Evaluar la preservación topológica de los métodos implementados
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Descripción	Evaluar los diversos métodos de RD mediante las métricas RNX
Sub-objetivo	Ninguno
Importancia	Vital
Urgencia	Inmediata
Estado	Implementado
Estabilidad	Alta
Comentario	Ninguno

OBJ004	Interacción rápida y sencilla con la herramienta
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Descripción	Permitir una interacción de fácil manejo para el usuario mediante una metodología drag and drop
Sub-objetivo	Ninguno
Importancia	Vital
Urgencia	Inmediata
Estado	Implementado
Estabilidad	Alta
Comentario	Ninguno

OBJ005	Visualizar los diversos resultados generados en la herramienta
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Descripción	Disponer de visualizaciones que permitan interpretar rápida y eficientemente los resultados obtenidos
Sub-objetivo	Ninguno
Importancia	Vital
Urgencia	Inmediata
Estado	Implementado
Estabilidad	Alta
Comentario	Ninguno

4. Catálogo de Requisitos

4.1. Requisitos Funcionales

FRQ001	Nodo datos artificiales
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ001
Descripción	El sistema dispondrá de un nodo que incluya los data sets artificiales seleccionados previamente.
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ002	**Selección**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ001

Descripción	El sistema deberá soportar una lista desplegable de selección de los diferentes data set pre-cargados en la herramienta (Toroide, Swisroll, Sphere)
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ003	**Nodo datos reales**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ001
Descripción	El sistema permitirá importar archivos con extensión (.xlsx .mat . csv) de forma externa
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ004	Separador
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ001
Descripción	El sistema deberá integrar un separador para generar la tabulación de data set ya sea por “.” “**,**” “**;**”
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ005	Nodo particionador
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ001
Descripción	El sistema dispondrá de un nodo para eliminar columnas de un data set que no son necesarias para el proceso de los métodos

Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ006	**Configurar particionador**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ001
Descripción	El sistema dispondrá de un selector para identificar en una lista desplegable la columna que se va a eliminar
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ007	Nodos métodos RD
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ002
Descripción	El sistema dispondrá de nodos de métodos de reducción de dimensión PCA, KPCA, LE, LLE, ISOMAP, MDS
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ008	Nodo RD Paramétrico
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ002
Descripción	El sistema deberá permitir configurar los parámetros de numero de dimensiones y vecindarios con los que trabajará el método

Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ009	**Nodo RD no Paramétrico**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ002
Descripción	El sistema deberá permitir configurar el parámetro del número de dimensiones con el que trabajará el método
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ010	Nodo evaluador RNX
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ003
Descripción	El sistema dispondrá de un nodo para evaluar los nodos de reducción mediante las métricas Rnx
Importancia	Vital
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ011	Múltiple conexión en nodo Rnx
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ003
Descripción	El sistema deberá permitir que el nodo evaluador soporte como entradas los datos generados por uno o más métodos de reducción y desde el dataset original
Importancia	Importante
Urgencia	Hay presión

Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ012	**Nodo SCATTER PLOT**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ005
Descripción	El sistema deberá permitir generar gráficas de dispersión en 2D y 3D
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ013	**Nodo visualización de evaluación RNX**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ005
Descripción	El sistema deberá permitir generar la visualización de las curvas de evaluación Rnx

Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ014	**Interfaz**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ004
Descripción	El sistema deberá disponer de un listado de nodos disponibles y una zona de trabajo
Importancia	Vital
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ015	Nodo Data Table
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ005
Descripción	El sistema deberá permitir crear una tabla de datos que esté conectada al nodo
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ016	Workflow
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ004
Descripción	El sistema deberá permitir crear flujos de trabajo que el usuario pueda usar con facilidad
Importancia	Importante

Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ017	**Interconexión de nodos**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ004
Descripción	El sistema deberá permitir la interconexión entre los nodos en el workflow
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ018	Guardar workflow
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ004
Descripción	El sistema deberá permitir guardar flujos de trabajo en formato JSON
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

FRQ019	Abrir workflow
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	OBJ004
Descripción	El sistema deberá permitir cargar flujos de trabajo que el usuario allá guardado con anterioridad
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado

Comentario	Ninguno

4.2. Requisitos no Funcionales

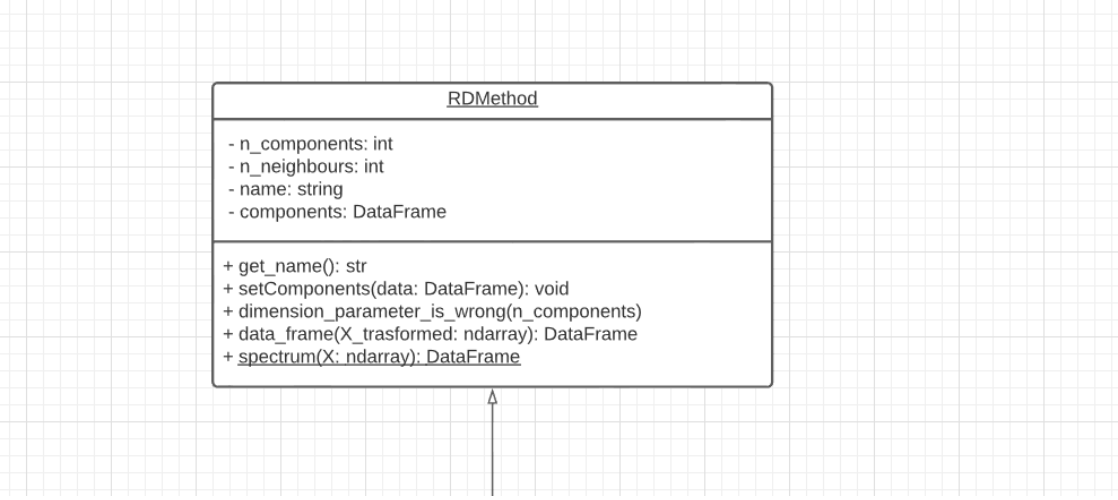
NFR001	**Usabilidad**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	Ninguna
Descripción	El sistema deberá permitir ser ejecutado en distintos sistemas operativos (Windows, Linux, OSX)
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

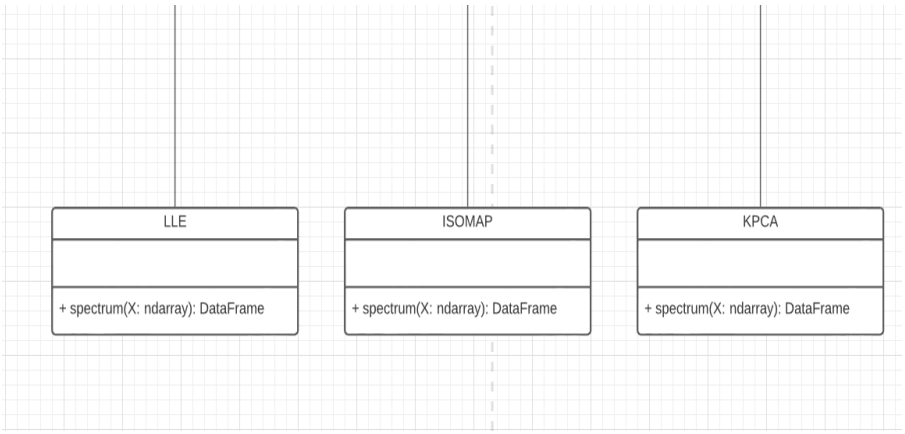
NFR002	**Eficiencia**
Versión	1.0 (27/09/2021)
Autores	Carlos David Correa Lozano Juan Andrés Lozano Thomé Diego Ferley Urrea Burgos
Dependencias	Ninguna
Descripción	El sistema deberá tener tiempos de ejecución de cada nodo no mayores a 5 segundos en promedio

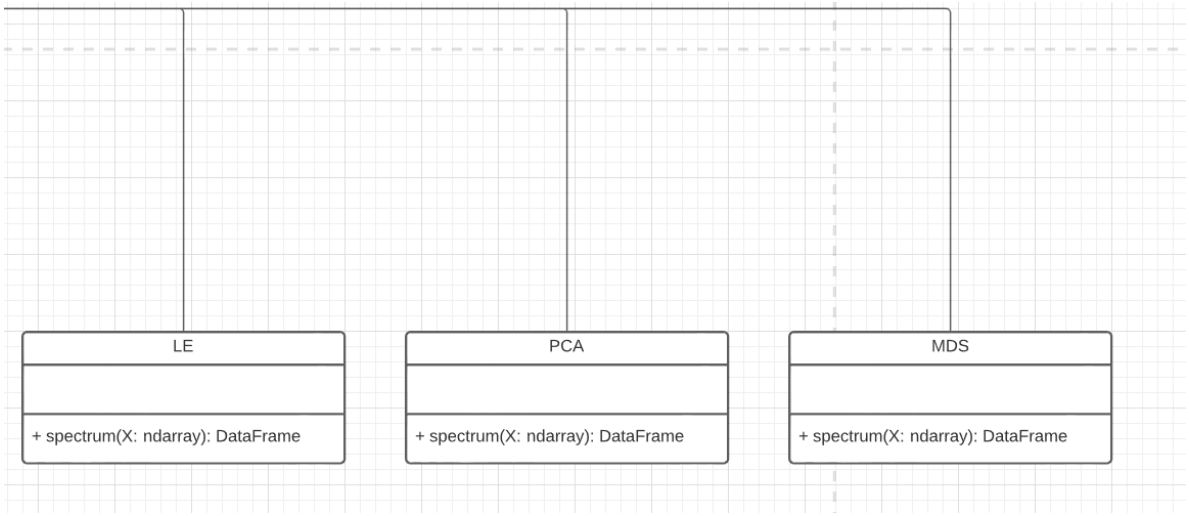
Importancia	Importante
Urgencia	Hay presión
Estabilidad	Alta
Estado	Implementado
Comentario	Ninguno

5. Diagramas UML

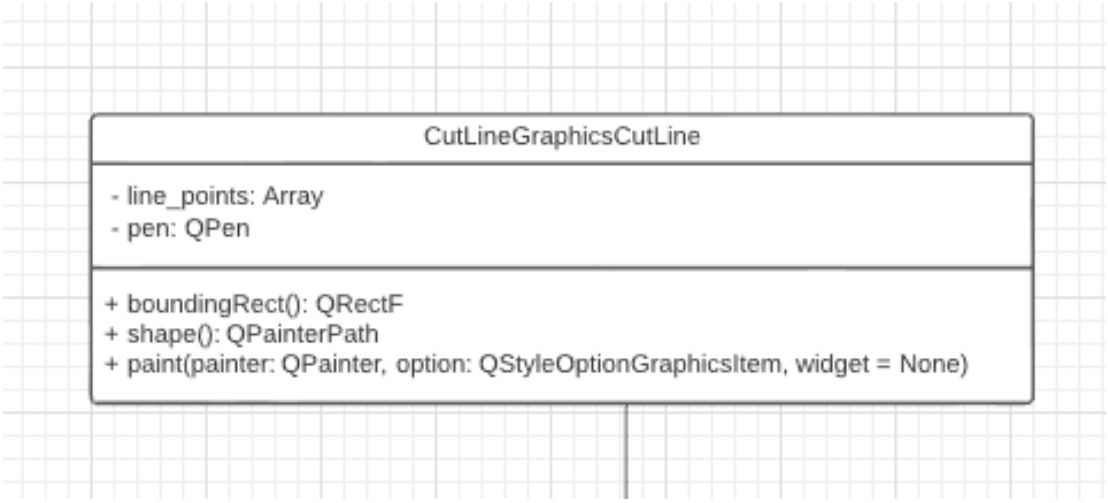
5.1 UML Clases Métodos RD

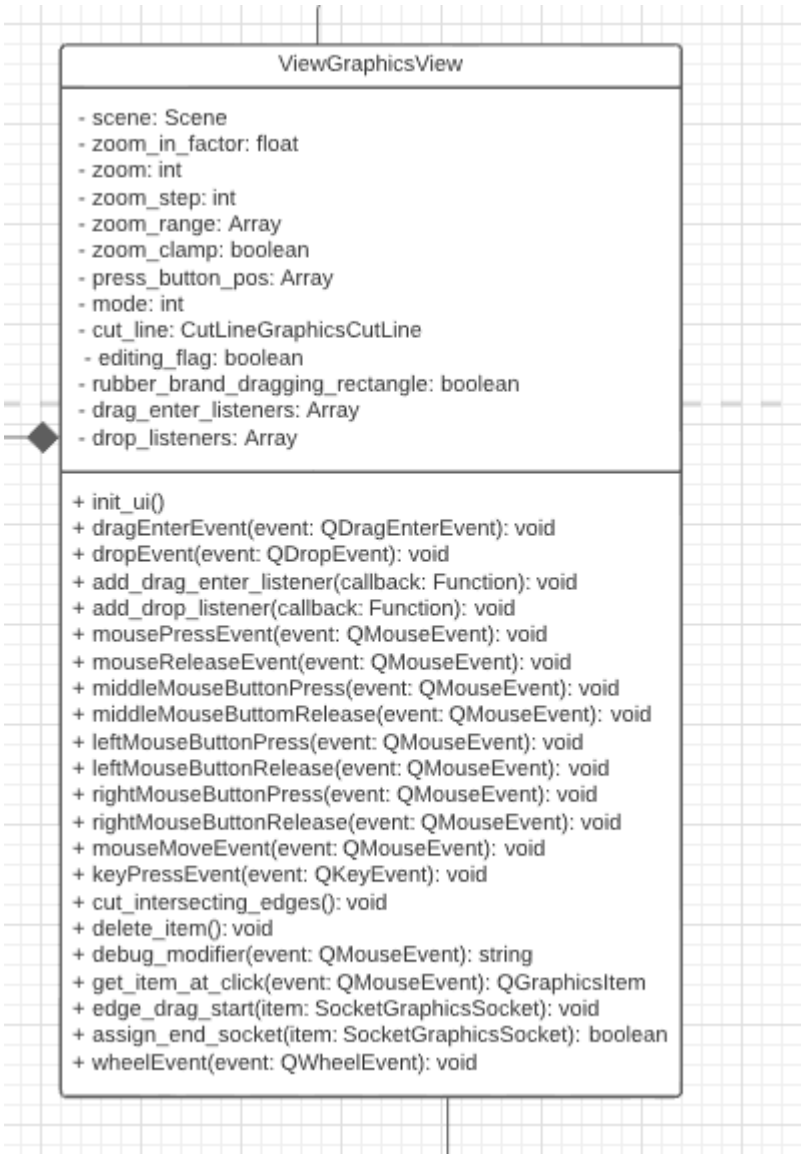


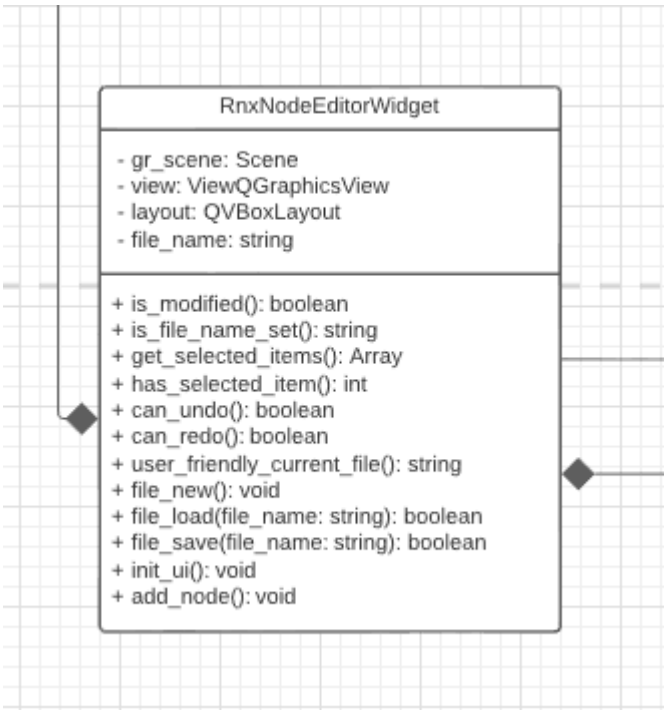


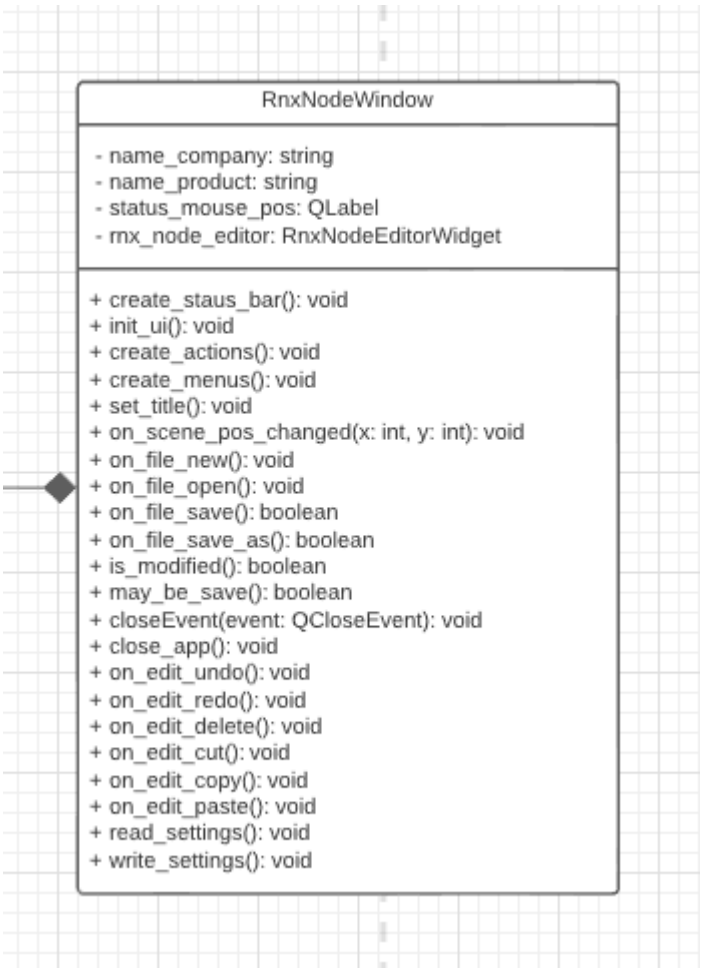


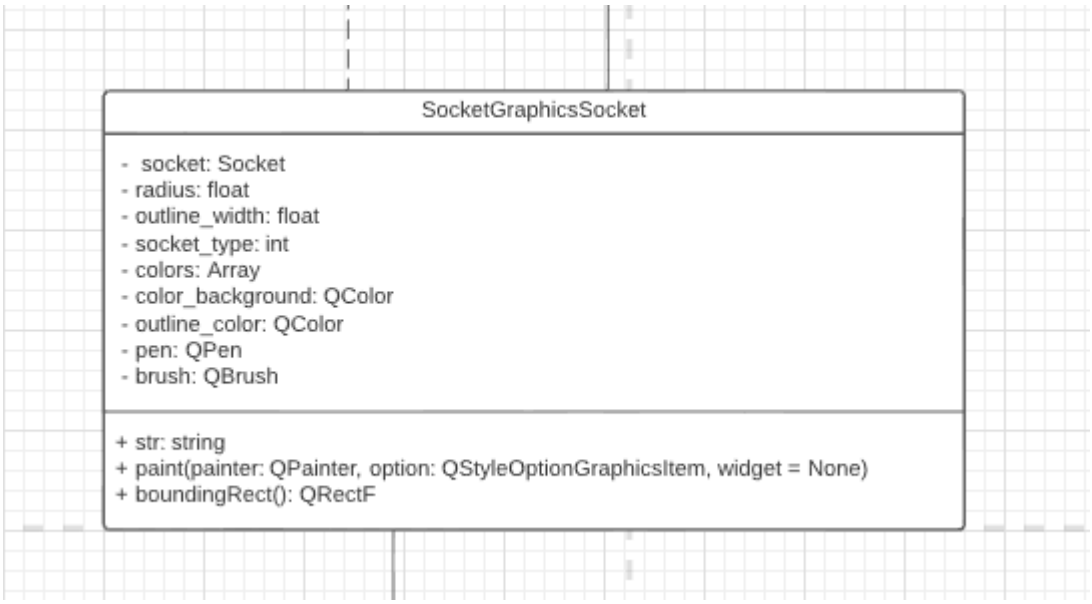
5.2 UML Clases Herramienta

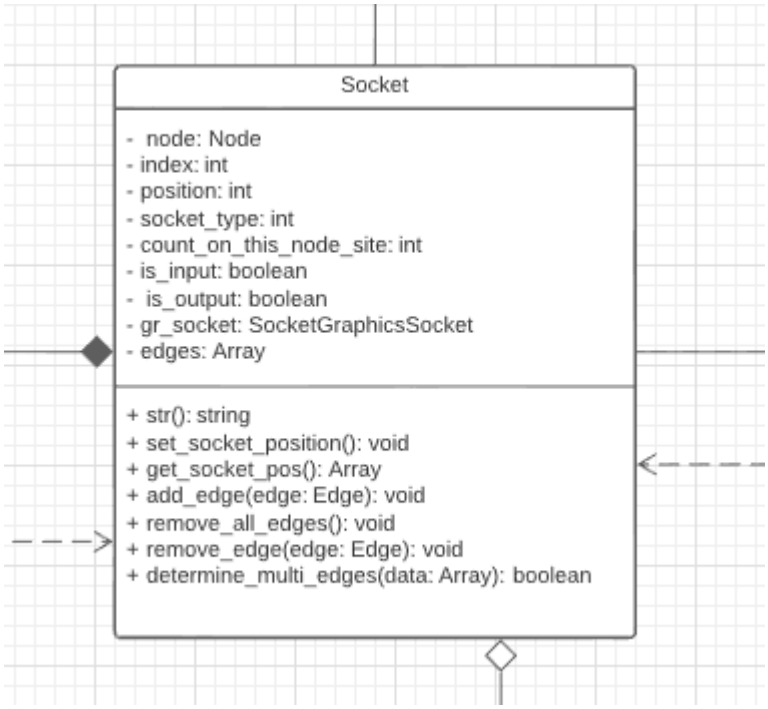


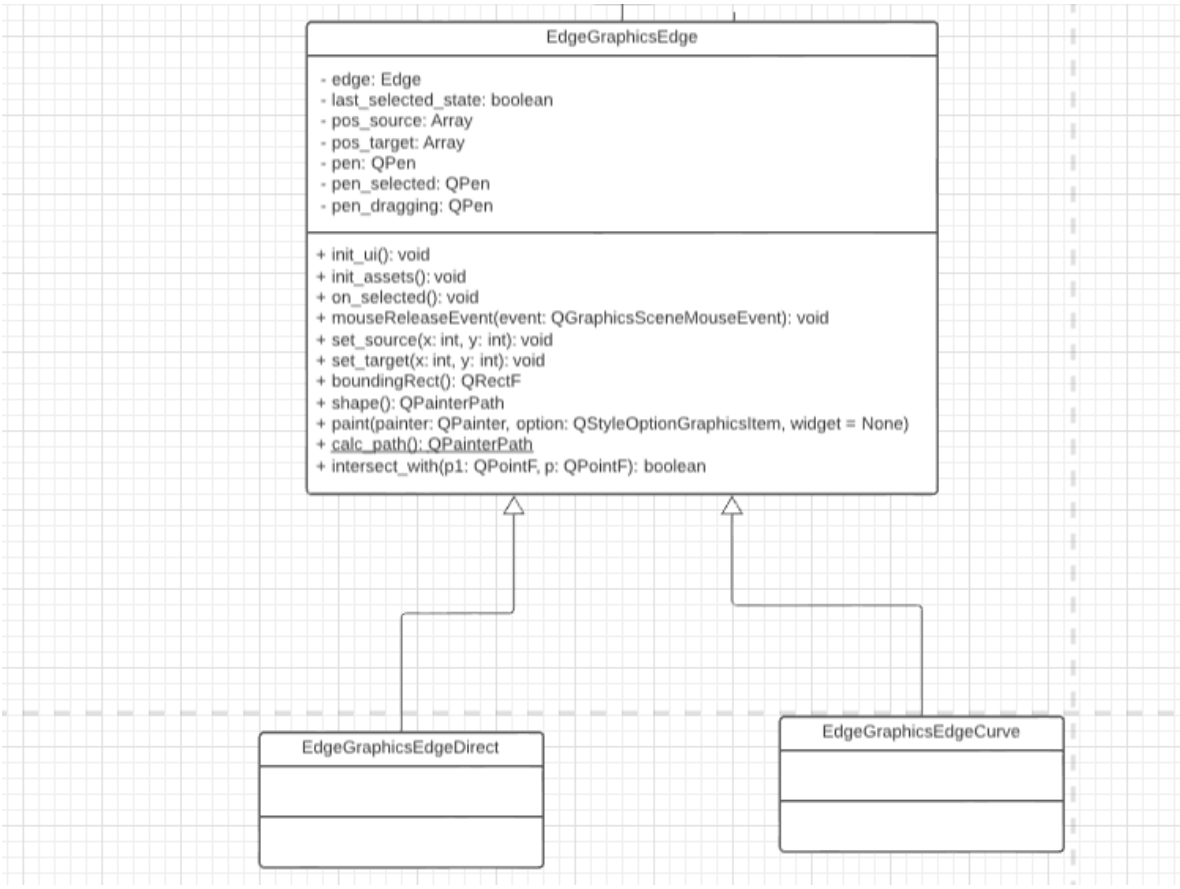

ViewGraphicsView
- scene: Scene
- zoom_in_factor: float
- zoom: int
- zoom_step: int
- zoom_range: Array
- zoom_clamp: boolean
- press_button_pos: Array
- mode: int
- cut_line: CutLineGraphicsCutLine
- editing_flag: boolean
- rubber_brand_dragging_rectangle: boolean
- drag_enter_listeners: Array
- drop_listeners: Array
+ init_ui()
+ dragEnterEvent(event: QDragEnterEvent): void
+ dropEvent(event: QDropEvent): void
+ add_drag_enter_listener(callback: Function): void
+ add_drop_listener(callback: Function): void
+ mousePressEvent(event: QMouseEvent): void
+ mouseReleaseEvent(event: QMouseEvent): void
+ middleMouseButtonPress(event: QMouseEvent): void
+ middleMouseButtomRelease(event: QMouseEvent): void
+ leftMouseButtonPress(event: QMouseEvent): void
+ leftMouseButtonRelease(event: QMouseEvent): void
+ rightMouseButtonPress(event: QMouseEvent): void
+ rightMouseButtonRelease(event: QMouseEvent): void
+ mouseMoveEvent(event: QMouseEvent): void
+ keyPressEvent(event: QKeyEvent): void
+ cut_intersecting_edges(): void
+ delete_item(): void
+ debug_modifier(event: QMouseEvent): string
+ get_item_at_click(event: QMouseEvent): QGraphicsItem
+ edge_drag_start(item: SocketGraphicsSocket): void
+ assign_end_socket(item: SocketGraphicsSocket): boolean
+ wheelEvent(event: QWheelEvent): void

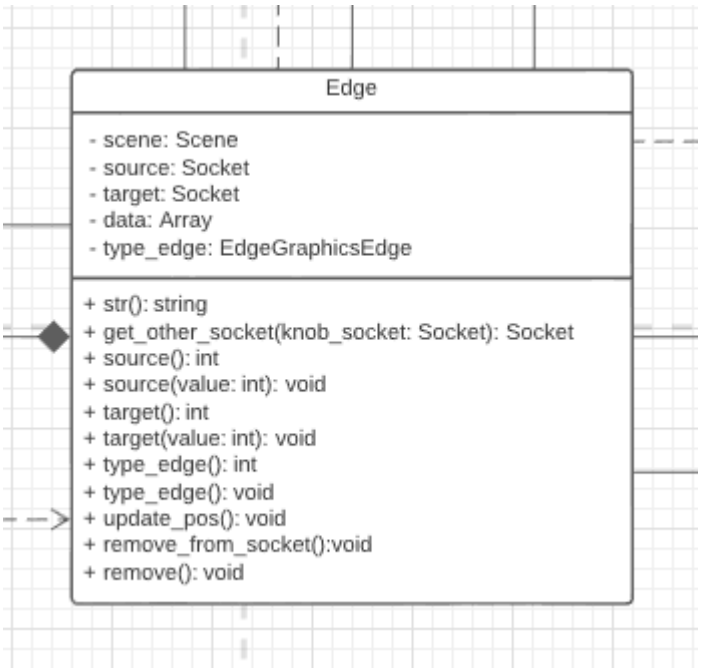

RnxNodeEditorWidget
- gr_scene: Scene
- view: ViewQGraphicsView
- layout: QVBoxLayout
- file_name: string
+ is_modified(): boolean
+ is_file_name_set(): string
+ get_selected_items(): Array
+ has_selected_item(): int
+ can_undo(): boolean
+ can_redo(): boolean
+ user_friendly_current_file(): string
+ file_new(): void
+ file_load(file_name: string): boolean
+ file_save(file_name: string): boolean
+ init_ui(): void
+ add_node(): void

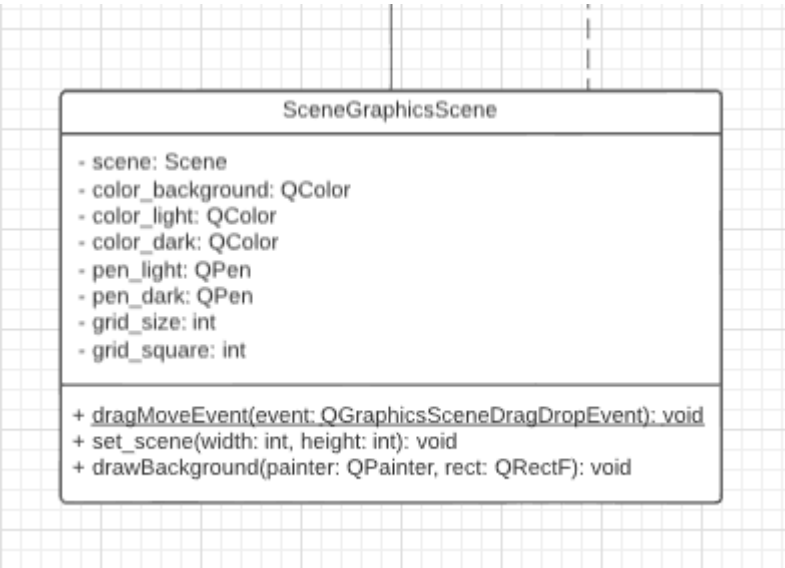

RnxNodeWindow
- name_company: string
- name_product: string
- status_mouse_pos: QLabel
- rnx_node_editor: RnxNodeEditorWidget
+ create_staus_bar(): void
+ init_ui(): void
+ create_actions(): void
+ create_menus(): void
+ set_title(): void
+ on_scene_pos_changed(x: int, y: int): void
+ on_file_new(): void
+ on_file_open(): void
+ on_file_save(): boolean
+ on_file_save_as(): boolean
+ is_modified(): boolean
+ may_be_save(): boolean
+ closeEvent(event: QCloseEvent): void
+ close_app(): void
+ on_edit_undo(): void
+ on_edit_redo(): void
+ on_edit_delete(): void
+ on_edit_cut(): void
+ on_edit_copy(): void
+ on_edit_paste(): void
+ read_settings(): void
+ write_settings(): void

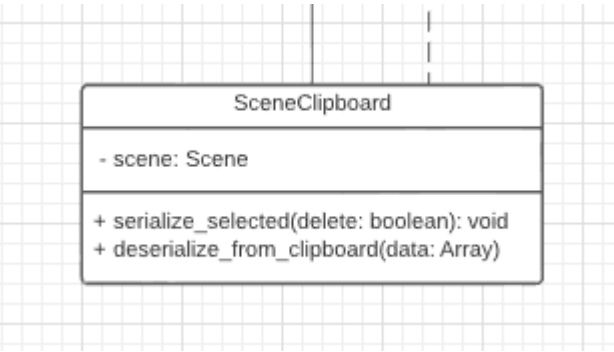

SocketGraphicsSocket
- socket: Socket
- radius: float
- outline_width: float
- socket_type: int
- colors: Array
- color_background: QColor
- outline_color: QColor
- pen: QPen
- brush: QBrush
+ str: string
+ paint(painter: QPainter, option: QStyleOptionGraphicsItem, widget = None)
+ boundingRect(): QRectF

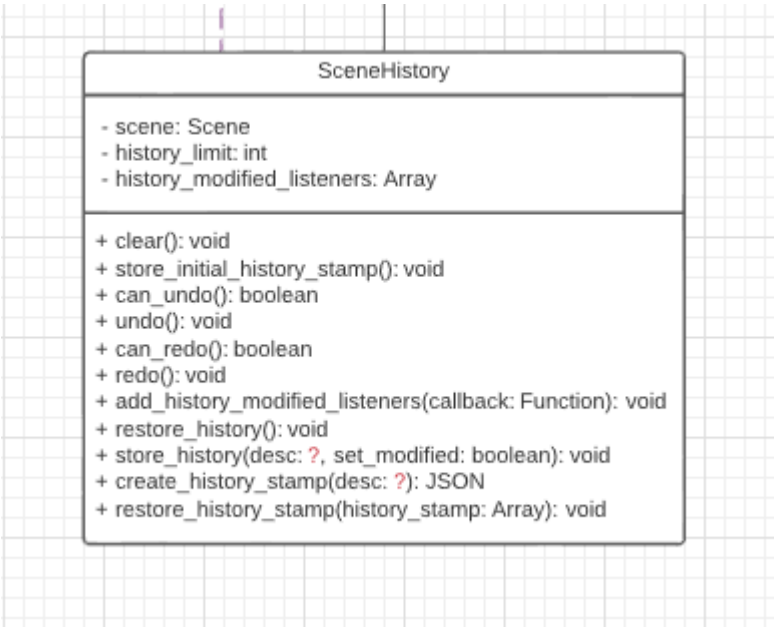

Socket
- node: Node
- index: int
- position: int
- socket_type: int
- count_on_this_node_site: int
- is_input: boolean
- is_output: boolean
- gr_socket: SocketGraphicsSocket
- edges: Array
+ str(): string
+ set_socket_position(): void
+ get_socket_pos(): Array
+ add_edge(edge: Edge): void
+ remove_all_edges(): void
+ remove_edge(edge: Edge): void
+ determine_multi_edges(data: Array): boolean

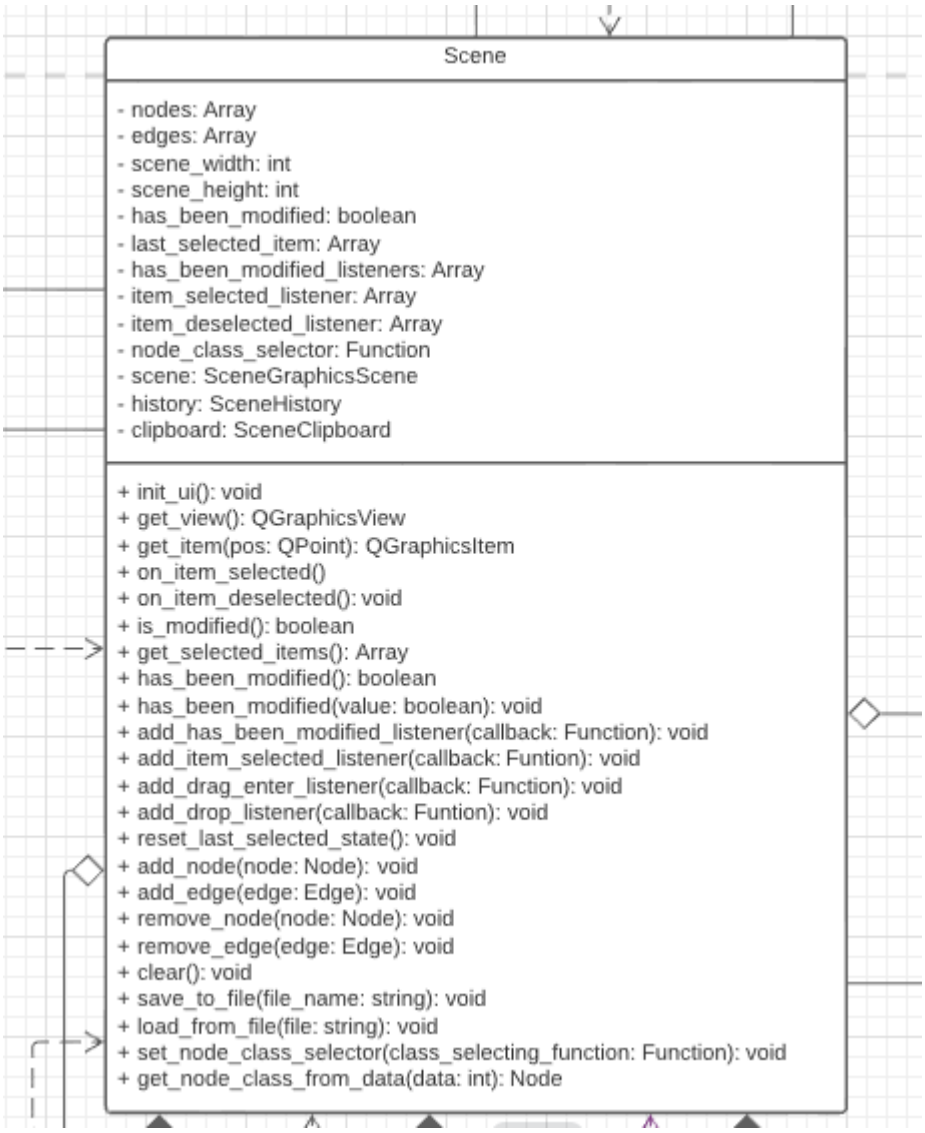

EdgeGraphicsEdge
- edge: Edge
- last_selected_state: boolean
- pos_source: Array
- pos_target: Array
- pen: QPen
- pen_selected: QPen
- pen_dragging: QPen
+ init_ui(): void
+ init_assets(): void
+ on_selected(): void
+ mouseReleaseEvent(event: QGraphicsSceneMouseEvent): void
+ set_source(x: int, y: int): void
+ set_target(x: int, y: int): void
+ boundingRect(): QRectF
+ shape(): QPainterPath
+ paint(painter: QPainter, option: QStyleOptionGraphicsItem, widget = None)
+ calc_path(): QPainterPath
+ intersect_with(p1: QPointF, p: QPointF): boolean
EdgeGraphicsEdgeDirect
EdgeGraphicsEdgeCurve

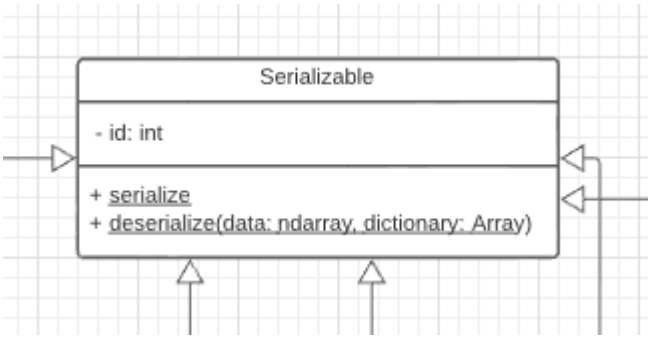

Edge
- scene: Scene
- source: Socket
- target: Socket
- data: Array
- type_edge: EdgeGraphicsEdge
+ str(): string
+ get_other_socket(knob_socket: Socket): Socket
+ source(): int
+ source(value: int): void
+ target(): int
+ target(value: int): void
+ type_edge(): int
+ type_edge(): void
+ update_pos(): void
+ remove_from_socket():void
+ remove(): void

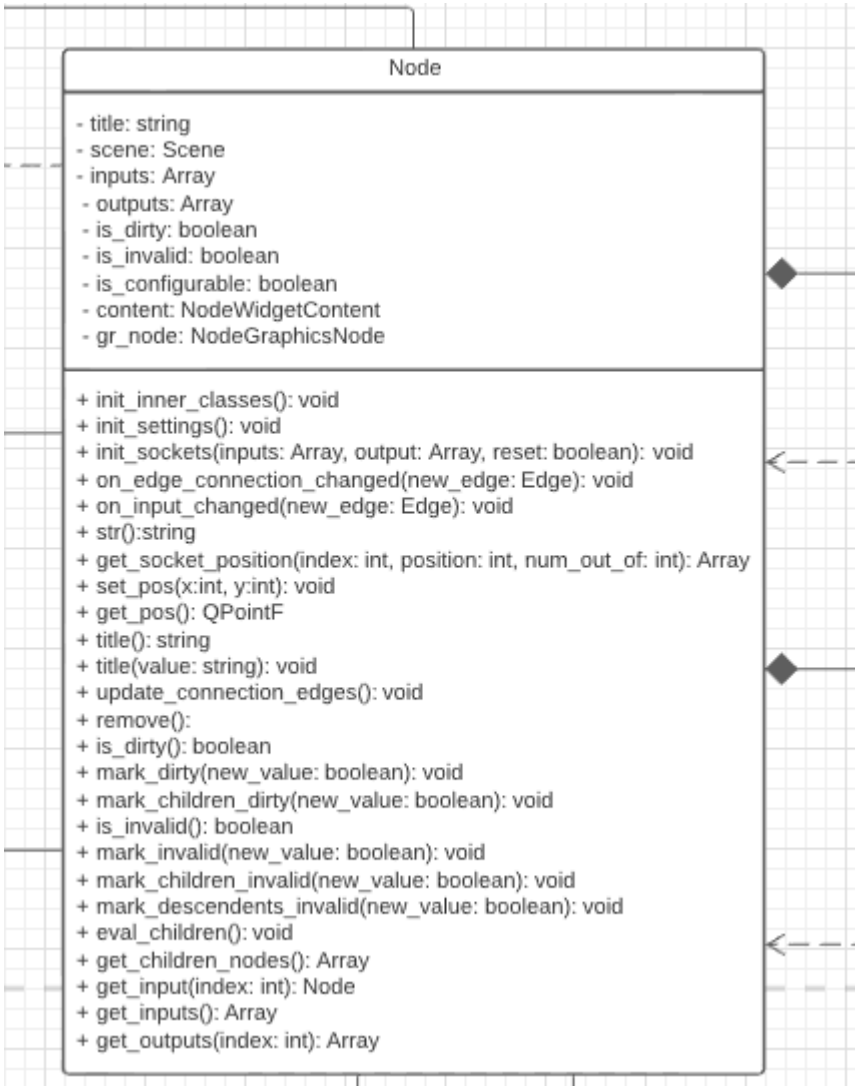

SceneGraphicsScene
- scene: Scene
- color_background: QColor
- color_light: QColor
- color_dark: QColor
- pen_light: QPen
- pen_dark: QPen
- grid_size: int
- grid_square: int
+ dragMoveEvent(event: QGraphicsSceneDragDropEvent): void
+ set_scene(width: int, height: int): void
+ drawBackground(painter: QPainter, rect: QRectF): void

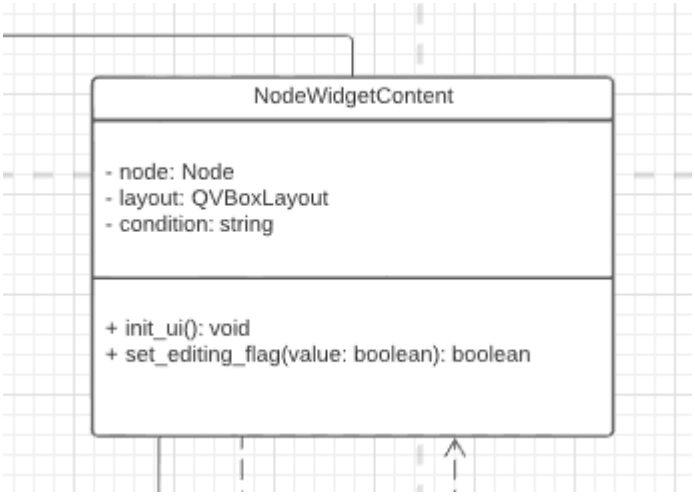

SceneClipboard
- scene: Scene
+ serialize_selected(delete: boolean): void
+ deserialize_from_clipboard(data: Array)

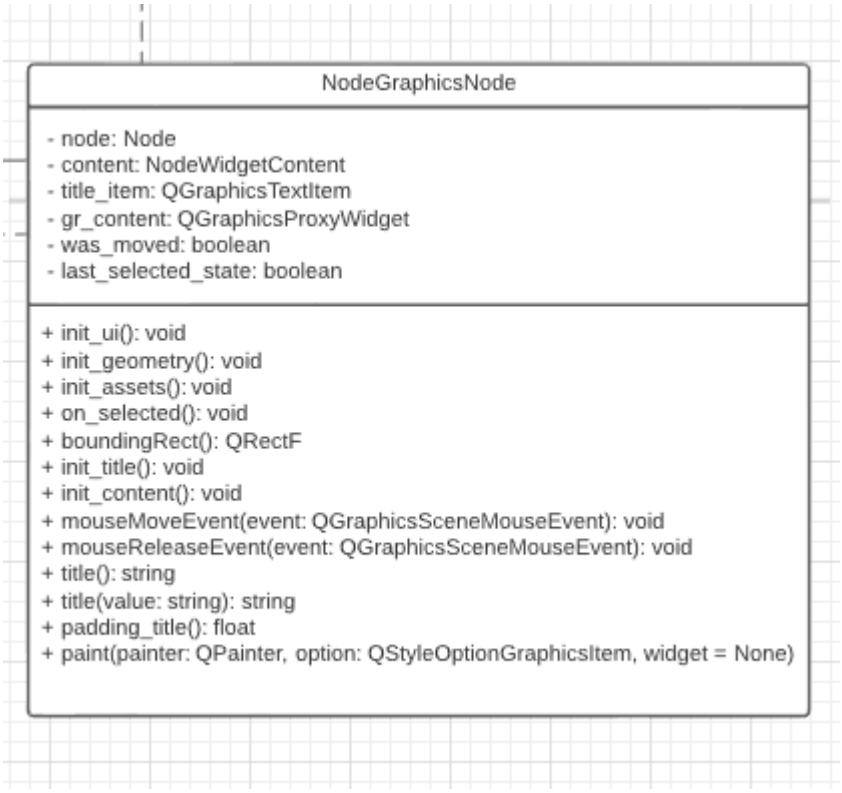

SceneHistory
- scene: Scene
- history_limit: int
- history_modified_listeners: Array
+ clear(): void
+ store_initial_history_stamp(): void
+ can_undo(): boolean
+ undo(): void
+ can_redo(): boolean
+ redo(): void
+ add_history_modified_listeners(callback: Function): void
+ restore_history(): void
+ store_history(desc: ?, set_modified: boolean): void
+ create_history_stamp(desc: ?): JSON
+ restore_history_stamp(history_stamp: Array): void


Scene
- nodes: Array
- edges: Array
- scene_width: int
- scene_height: int
- has_been_modified: boolean
- last_selected_item: Array
- has_been_modified_listeners: Array
- item_selected_listener: Array
- item_deselected_listener: Array
- node_class_selector: Function
- scene: SceneGraphicsScene
- history: SceneHistory
- clipboard: SceneClipboard
+ init_ui(): void
+ get_view(): QGraphicsView
+ get_item(pos: QPoint): QGraphicsItem
+ on_item_selected()
+ on_item_deselected(): void
+ is_modified(): boolean
+ get_selected_items(): Array
+ has_been_modified(): boolean
+ has_been_modified(value: boolean): void
+ add_has_been_modified_listener(callback: Function): void
+ add_item_selected_listener(callback: Funtion): void
+ add_drag_enter_listener(callback: Function): void
+ add_drop_listener(callback: Funtion): void
+ reset_last_selected_state(): void
+ add_node(node: Node): void
+ add_edge(edge: Edge): void
+ remove_node(node: Node): void
+ remove_edge(edge: Edge): void
+ clear(): void
+ save_to_file(file_name: string): void
+ load_from_file(file: string): void
+ set_node_class_selector(class_selecting_function: Function): void
+ get_node_class_from_data(data: int): Node


Serializable
- id: int
+ serialize
+ deserialize(data: ndarray, dictionary: Array)


Node
- title: string
- scene: Scene
- inputs: Array
- outputs: Array
- is_dirty: boolean
- is_invalid: boolean
- is_configurable: boolean
- content: NodeWidgetContent
- gr_node: NodeGraphicsNode
+ init_inner_classes(): void
+ init_settings(): void
+ init_sockets(inputs: Array, output: Array, reset: boolean): void
+ on_edge_connection_changed(new_edge: Edge): void
+ on_input_changed(new_edge: Edge): void
+ str():string
+ get_socket_position(index: int, position: int, num_out_of: int): Array
+ set_pos(x:int, y:int): void
+ get_pos(): QPointF
+ title(): string
+ title(value: string): void
+ update_connection_edges(): void
+ remove():
+ is_dirty(): boolean
+ mark_dirty(new_value: boolean): void
+ mark_children_dirty(new_value: boolean): void
+ is_invalid(): boolean
+ mark_invalid(new_value: boolean): void
+ mark_children_invalid(new_value: boolean): void
+ mark_descendents_invalid(new_value: boolean): void
+ eval_children(): void
+ get_children_nodes(): Array
+ get_input(index: int): Node
+ get_inputs(): Array
+ get_outputs(index: int): Array


NodeWidgetContent
- node: Node
- layout: QVBoxLayout
- condition: string
+ init_ui(): void
+ set_editing_flag(value: boolean): boolean


NodeGraphicsNode
- node: Node
- content: NodeWidgetContent
- title_item: QGraphicsTextItem
- gr_content: QGraphicsProxyWidget
- was_moved: boolean
- last_selected_state: boolean
+ init_ui(): void
+ init_geometry(): void
+ init_assets(): void
+ on_selected(): void
+ boundingRect(): QRectF
+ init_title(): void
+ init_content(): void
+ mouseMoveEvent(event: QGraphicsSceneMouseEvent): void
+ mouseReleaseEvent(event: QGraphicsSceneMouseEvent): void
+ title(): string
+ title(value: string): string
+ padding_title(): float
+ paint(painter: QPainter, option: QStyleOptionGraphicsItem, widget = None)

Anexo C: Artículo sobre la herramienta presentado en CACIED

El proyecto descrito en este documento fue presentado en modalidad de ponencia en el Quinto Congreso Andino en Computación, Informática y Educación – CACIED y obtuvo el primer lugar en la sesión de ponencias en la sala 2 día 2. El articulo presentado a continuación, fue el consolidado de la investigación realizada en el proyecto.

QARNX: Una herramienta para la evaluación de la preservación topológica de métodos de reducción de dimensión

QARNX: A tool for assessment of topological preservation of dimensionality reduction methods

Carlos David Correa Lozano

Carlosdcorrea3@gmail.com

Cdcorrea.4457@unicesmag.edu.co

https://orcid.org/0000-0002-1299-9507

Diego Ferley Urrea Burgos

urreadiego767@gmail.com

dfurrea.3928@unicesmag.edu.co

https://orcid.org/0000-0002-3020-3074

Juan Andrés Lozano Thomé

jandresl092@gmail.com

https://orcid.org/0000-0001-9230-1991

Juan Carlos Alvarado Pérez

jcalvarado@unicesmag.edu.co

https://orcid.org/0000-0001-9497-5893

Resumen

Los avances tecnológicos y computacionales, han hecho más fácil los procesos de capturar, almacenar, procesar y transportar grandes volúmenes de información, de igual forma, la comunidad científica y académica, ha propuesto diversas heurísticas de RD (reducción de dimensión), como métodos espectrales, estadísticos, basados en Kernel, entre otros. La RD permite reducir la cantidad de dimensiones que se encuentran en un conjunto de datos, entiéndase dimensión como atributo o característica, uno de los métodos de RD más conocidos a nivel mundial es PCA (Principal Component Analysis), que, debido a su efectividad, se ha convertido en la base e inspiración de muchos otros métodos RD, así mismo, existen métodos RD como LLE, ISOMAP y LE que buscan preservar la topología local de los datos. La reducción de dimensión tiene como consecuencia la perdida de información, debido a esto, hay diversas propuestas para la evaluación de estos métodos, algunas se basan en la preservación de la distancia y otras en la topológica. Este trabajo introduce QARNX, un software de uso libre cuyo objetivo es la evaluación de la preservación topológica de los datos, mediante la implementación de las métricas RNX, QARNX está inspirada en el proceso KDD, por lo que la herramienta contiene, 1) módulos para cargar datos reales con extensión .xlxs, .csv y .mat, y para cargar datos artificiales, como lo son la esfera, el rollo suizo y el toroide, 2) módulo de particionamiento de los datos, 3) módulos de visualización de datos, 4) módulos de reducción de dimensión y finalmente 5) módulo para la evaluación de la preservación topológica de los métodos utilizados. Los módulos anteriormente mencionados, especialmente el de evaluación RNX, el cual hasta ahora no tenía una implementación en Python, han sido integrados en QARNX, siendo ésta una herramienta Drag and Drop, permitiéndole al usuario una interacción rápida y sencilla con la misma, para realizar los flujos de evaluación, finalmente, se realizaron experimentos con algunos de los métodos RD ya mencionados con diferentes conjuntos de datos artificiales y reales, obteniendo así los resultados de la evaluación, ofreciéndole al usuario una forma intuitiva para determinar qué método RD tuvo el mejor rendimiento.

Palabras Clave: métricas RNX, Reducción de dimensión, Evaluación de calidad, Proceso KDD

Keywords: RNX metrics, dimension reduction, Quality assessment, KDD Process

Introducción

En la actualidad, se generan diariamente grandes volúmenes de información provenientes de todas las áreas del conocimiento, ya que los humanos no tienen la capacidad de procesamiento que tienen las máquinas, en los últimos años, han habido grandes avances en herramientas que integran procesos como el cargue, preprocesamiento, procesamiento y visualización de datos, uno de los conceptos más conocidos en el mundo de la Ciencia de Datos, es el proceso KDD (Knowledge Discovery in Databases), el cual, como se afirma en [1], incluye diferentes métodos como la minería de datos, la reducción de dimensión y la visualización de la información con el fin de brindar una mejor interpretación de los datos, siendo ésta más intuitiva para el ser humano, la RD aunque se encuentra en varias etapas del proceso KDD, su papel en la etapa del preprocesamiento de los datos se hace más notorio [2], es aquí, donde la RD con el fin de eliminar datos irrelevantes o redundantes, transforma los datos reduciendo su dimensión y al mismo tiempo buscando preservar la mayor información de los mismos, beneficiando a áreas como el ML (Machine Learning), MD (Minería de datos) y reconocimiento de patrones tal como algunos autores lo afirman [3, 4]. Uno de los mayores beneficios que trae la RD, es su capacidad de representar los datos multidimensionales en planos tridimensionales y bidimensionales, esto teniendo en cuenta que el ser humano no puede entender más allá de la tercera dimensión, esto es posible por el incrustramiento que generan los métodos RD, siendo el incrustramiento, la proyección de los datos de alta dimensión en espacios de baja dimensionalidad tal como algunos trabajos lo mencionan [5, 6], sin embargo, la reducción de dimensión realizada por los métodos RD, tiene como costo, la perdida de información del conjunto de datos que se esté trabajando, debido a ello, diversos autores han propuestos métricas para evaluar los métodos RD, algunas de ellas se encuentran en [5, 7, 8, 9], que por lo general se basan en comparar los datos en alta dimensión HD (por sus siglas en inglés), con los mismos datos pero en baja dimensión LD (por sus siglas en inglés). Este trabajo introduce QARNX (Quality Assesment RNX) una herramienta Drag and Drop, que evalúa los métodos RD, mediante las implementaciones de la librería Scikit Learn de diferentes métodos RD, métodos de visualización y las métricas RNX propuestas por Jhon Lee y Michael Verleysen [10], la herramienta ofrece una interfaz interactiva e intuitiva para el usuario, permitiéndole al mismo, evaluar los métodos y visualizar los resultados de la evaluación.

Este artículo está organizado de la siguiente forma. En la siguiente sección se introducen conceptos y procesos preliminares para entender el funcionamiento de QARNX, siendo estos de conjuntos de datos, RD, visualización y métricas de calidad y finalmente se introduce QARNX teniendo en cuenta la metodología en la que está basada, siendo esta, el proceso KDD, finalmente se dan algunas conclusiones y recomendaciones teniendo en cuenta las experiencias obtenidas en el proceso de investigación.

Metodología

QARNX es una herramienta que evalúa los métodos RD, para la preservación topológica de los datos, la herramienta implementa las curvas RNX propuestas por Jhon Lee y Michel Verleysen [10], siendo ésta, la primera implementación modular de estas métricas en Python, el objetivo general de la herramienta, es ofrecer un entorno gráfico fácil de usar e intuitivo para usuario, que integre diversos pasos del proceso KDD, teniendo así, un flujo de trabajo definido, en la Figura se muestran los pasos del proceso KDD que QARNX implementa.

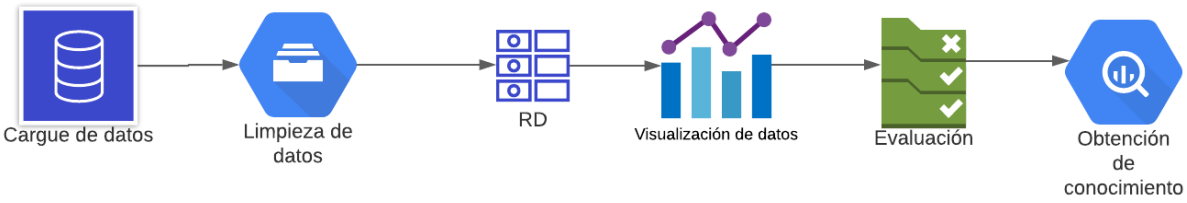


Figura 1: Proceso KDD en QARNX

Módulo para cargar datos

Este módulo está conformado por dos nodos, los cuales representan el cargue de datos artificiales y reales, los datos artificiales, son conjuntos de datos cuya visualización tiene alguna representación geométrica, y generalmente son generados por fórmulas matemáticas, en la Figura , se observan algunos de ellos

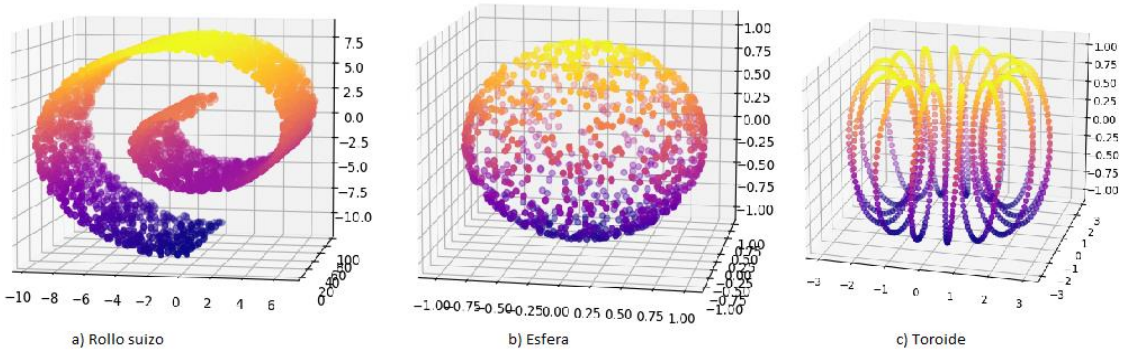
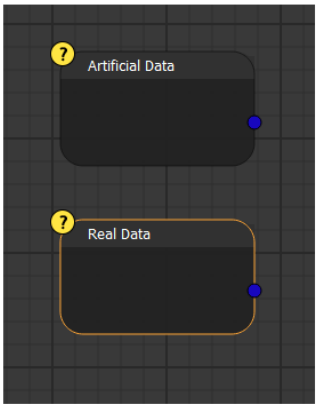
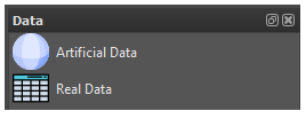


Figura 2: Conjuntos de datos artificiales

Cuando se trata de evaluar la preservación topológica de los métodos RD, lo más común es utilizar estos conjuntos de datos debido a su estructura, por otro lado, se QARNX, permite cargar datos reales, tales como de diabetes, sobre el estado de salud cardiovascular, entre otros. Debido a que estos últimos vienen en diferentes formatos, la herramienta permite cargar archivos en formato CSV (Comma Separated Values), xlxs (de Excel) y .mat (de MatLab), en la Figura se puede observar la representación visual de los nodos y su respectivo menú



a) Nodos en escena



b) Menú de nodos para cargar datos

Figura 3: Representación visual de los nodos para cargar datos y su menú

Finalmente, para que los nodos puedan ser ejecutados satisfactoriamente, primero deben ser configurados, aunque existen excepciones, ya que algunos de estos no requieren previa configuración, en la siguiente Figura se pueden ver las ventanas de configuración de ambos nodos.

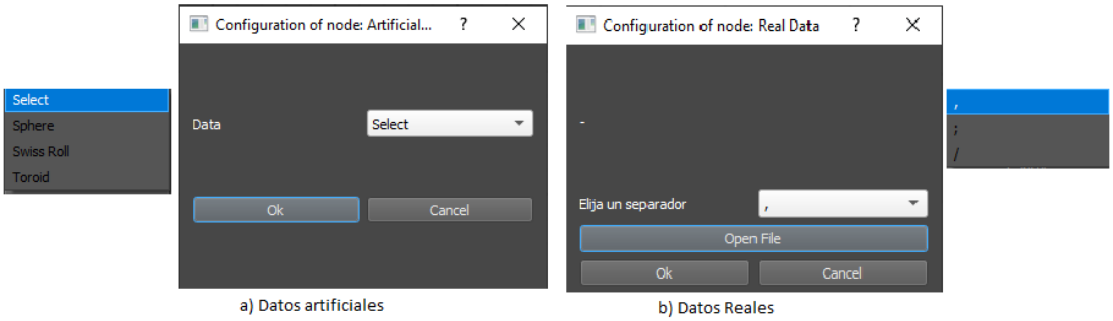


Figura 4: Ventanas de configuración de nodos para cargue de datos

Módulo de Reducción de dimensionalidad

Como ya se ha mencionado anteriormente, la reducción de dimensión permite representar los datos provenientes de una alta dimensión (HD) en espacios de baja dimensión (LD), tal como se expresa en [11], la idea general es que a partir de una variable aleatoria con una dimensión p por ejemplo, que $\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_p})^{\boldsymbol{T}}$, se obtenga un conjunto de datos de la forma $\boldsymbol{s} = (\boldsymbol{s_1}, \dots, \boldsymbol{s_k})^{\boldsymbol{T}}$, de tal forma que $k < p$. La Figura muestra el objetivo de la RD.

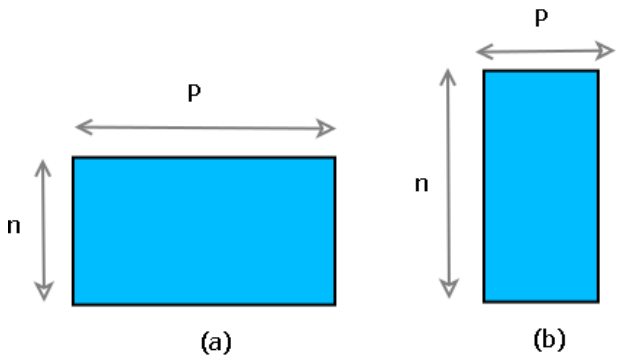


Figura 5: Objetivo de la reducción de dimensión.

Existen métodos RD cuyo rendimiento es mejor cuando los datos tienen un comportamiento lineal, los cuales, utilizando funciones lineales simples, logran cumplir con su objetivo en este tipo de datos, sin embargo, los datos no siempre se comportan de forma lineal, por lo que, a lo largo de los años, se han propuesto implementaciones RD no lineales [12], los cuales utilizan estructuras no lineales

más complejas, como lo son las funciones Kernel. De tal forma que la variedad de los métodos RD propuestos por la comunidad científica y académica se debe al comportamiento de los datos.

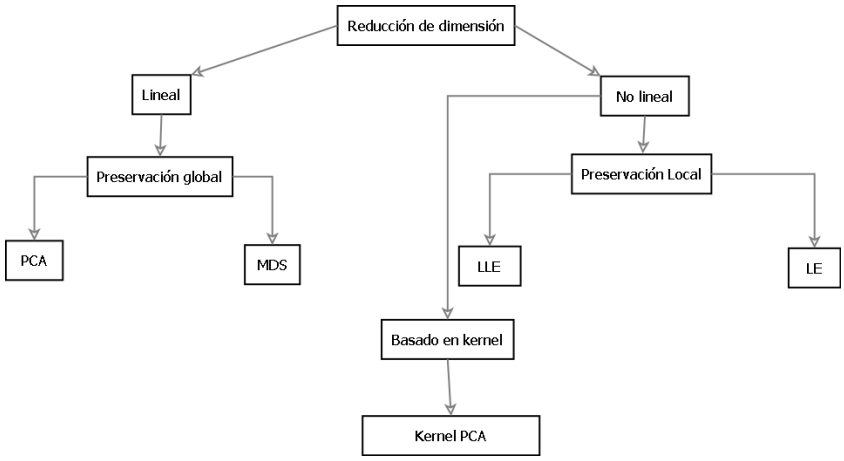


Figura 6: Taxonomía de los métodos RD

Los métodos que se pueden observar en la Figura, son los que han sido implementados en QARNX gracias a las implementaciones realizadas en Scikit Learn, aunque gracias a la escalabilidad de la herramienta, pueden incluirse muchos más. La RD se encuentra en la etapa de preprocesamiento de datos y aunque no se limita a esta, es aquí en donde más logra apreciarse su potencial, pues gracias a la transformación de los datos que realiza la RD, es que en las últimas etapas de KDD pueden realizarse clasificaciones de calidad, tal como ya se mencionó en [3]. La Figura muestra la representación visual de los nodos de los métodos RD y su menú.

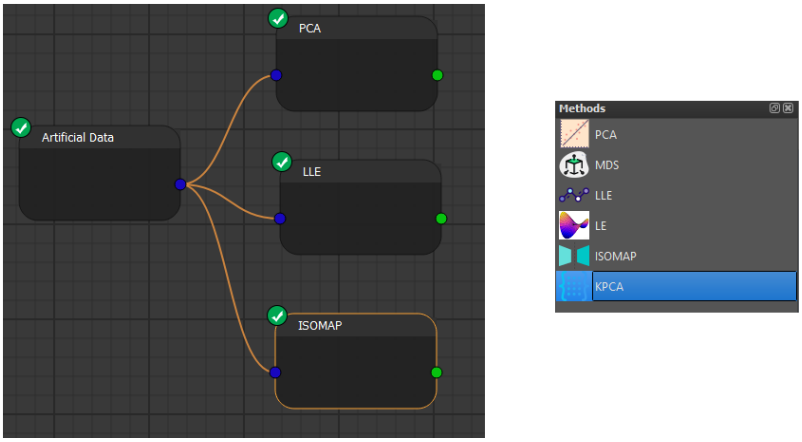


Figura 7: Representación visual de los nodos de reducción de dimensión, su menú y conectividad

Debido a que estos métodos RD vienen de diferentes heurísticas, no todos manejan la misma lógica y por lo tanto su configuración varía dependiendo del método, en [12] se menciona que hay métodos lineales y no lineales, agregándole a esto que los métodos como PCA y MDS (Multidimensional Scalling) son métodos globales, es decir, buscan conservar la topología global de los datos, por otro lado, LLE, LE, e ISOMAP, son métodos locales, y por lo tanto buscan preservar mejor la topología local de los datos tal como se menciona en [13, 14, 15], por lo que se desarrollaron ventanas de configuración para los casos, las cuales se muestran en la Figura.

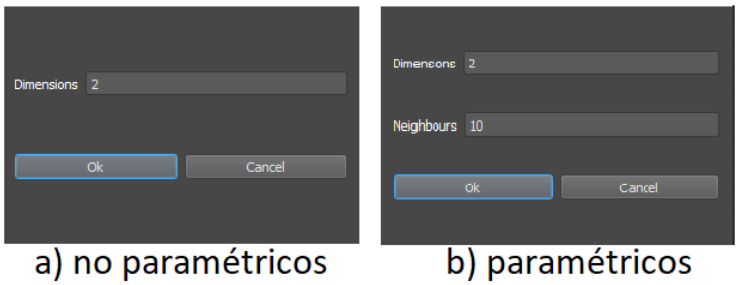


Figura 8: Ventana de configuración para métodos RD

Módulo de visualización de datos

El módulo de visualización de datos contiene 2 nodos para visualizar datos de forma gráfica, siendo estas visualizaciones un diagrama de dispersión de puntos y el otro de visualización línea, y un nodo para visualizar los datos de forma tabular, esto

sirve para que el usuario pueda ver las filas y columnas del conjunto de datos, el diagrama de dispersión de puntos, puede ser utilizado tanto para los datos en HD como para los de LD, tal como se muestra en la Figura.

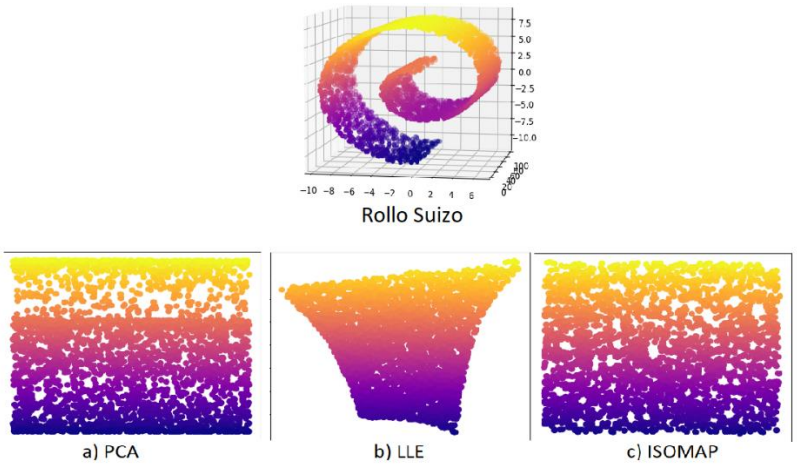


Figura 9: Diagrama de dispersión con el Rollo Suizo y el incrustramiento generado por PCA, LLE (10 vecindarios) e ISOMAP (10 vecindarios)

La figura 9 muestra la transformación de los datos del rollo suizo de 3 a 2 dimensiones generada por cada uno de los métodos RD, cada uno tiene una preservación diferente, pero dicha preservación no puede ser deducida solo por esta representación visual, es aquí donde las métricas RNX son importantes para representar esa preservación de forma numérica y exacta. Por otro lado, la visualización tabular de los datos se vuelve una parte fundamental del proceso, pudiendo identificar cualquier anomalía en el conjunto de datos, en la Figura se tiene un conjunto de datos real, que contiene datos vacíos.

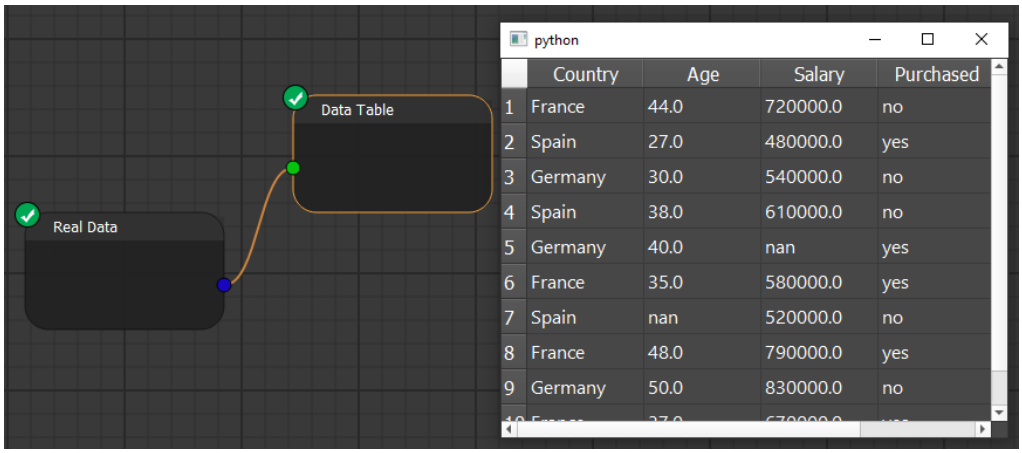


Figura 10: visualización tabular del conjunto de datos.

Debido a que el gráfico lineal sirve para observar el resultado de la evaluación de RNX, será mostrada después de la evaluación.

Módulo de evaluación

El módulo de evaluación solo tiene un nodo el cual es RNX, como se menciona en [16], RNX es la combinación de QNX y BNX mencionada en [10] re escaladas, con el fin de que el área bajo la curva se convierta en un buen indicador del rendimiento y comportamiento de los métodos RD, RNX está dada por:

$$R_{NX} = \frac{(N-1)Q_{NX}(K)-K}{N-1-K} \qquad (1)$$

Por último, se agrega el promedio de $100B_{NX}(K)$ para poder identificar mejor qué error predomina más en la incrustación, en la figura Figura 11: Flujo para evaluar los métodos de reducción de dimensión se encuentra el flujo necesario para realizar la evaluación.

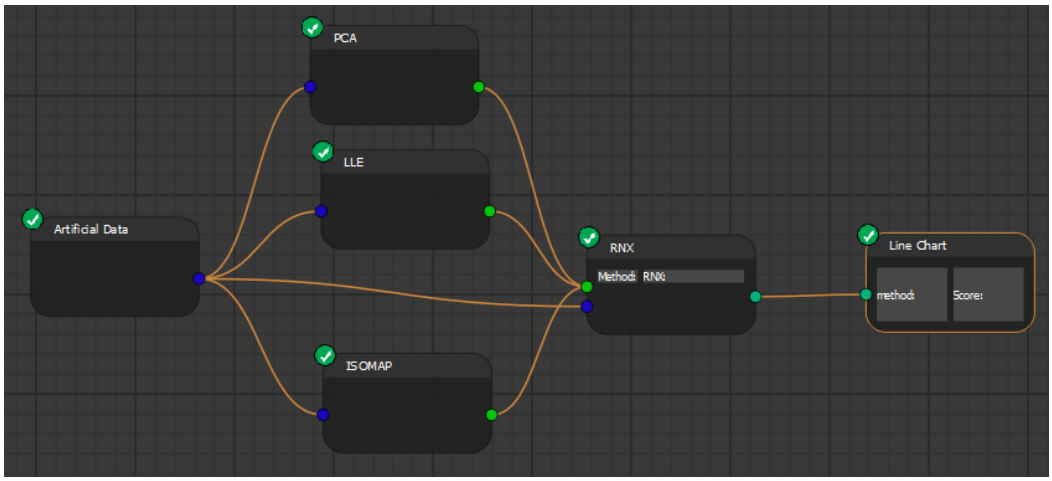


Figura 11: Flujo para evaluar los métodos de reducción de dimensión

Como se mencionó anteriormente, el diagrama lineal, se utiliza para mostrar el resultado de la evaluación de la preservación topológica de los métodos, en la Figura se encuentra dicha visualización, mostrando las curvas RNX.

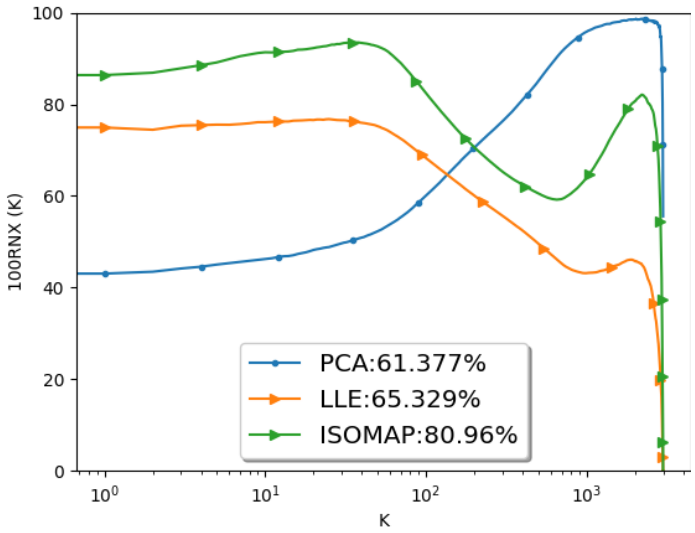


Figura 12: Curvas RNX con los resultados de la evaluación de PCA, LLE (10 vecindarios) e ISOMAP (10 vecindarios) con el conjunto de datos del Rollo Suizo

Experimentación y Resultados

A continuación, se harán diferentes experimentos, en donde se utilizarán los métodos RD con los conjuntos de datos artificiales, obteniendo así los resultados de sus rendimientos, primeramente, se evaluarán los métodos RD PCA, LLE e ISOMAP con el rollo suizo y LLE e ISOMAP con 10, 12 y 15 vecindarios, en la Figura se mostró el incrutamiento dado por estos métodos RD.

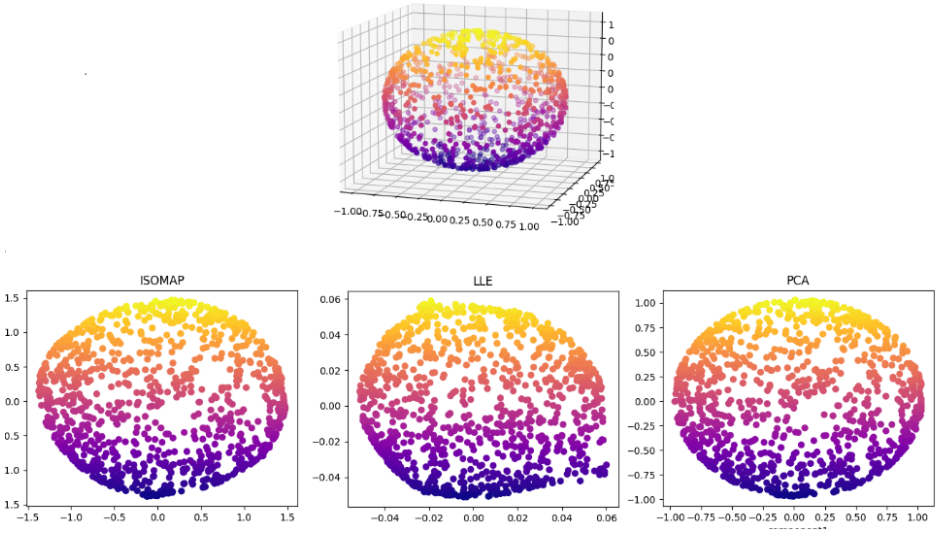


Figura 13: Incrustamiento generado de PCA, LLE e ISOMAP en el conjunto de datos esfera

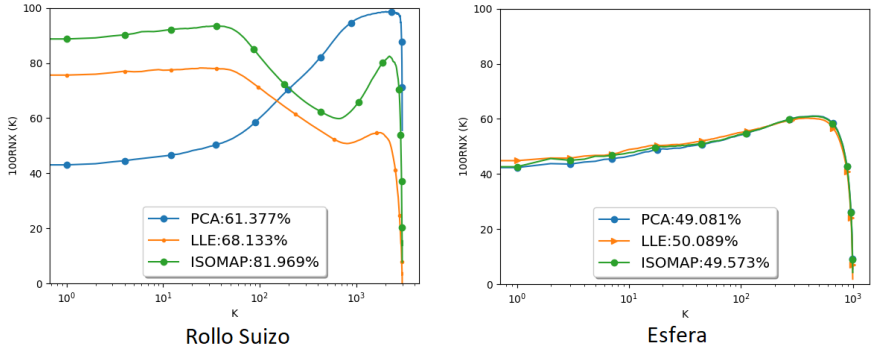


Figura 14: Evaluación PCA, LLE (15 vecindarios) e ISOMAP (15 vecindarios) con 2 dimensiones cada uno.

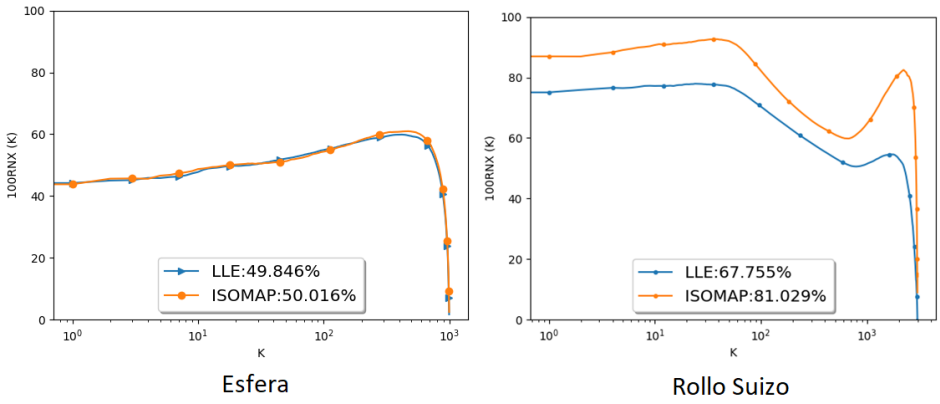


Figura 15: Evaluación LLE (12 vecindarios), ISOMAP (12 vecindarios) con 2 dimensiones cada uno.

.

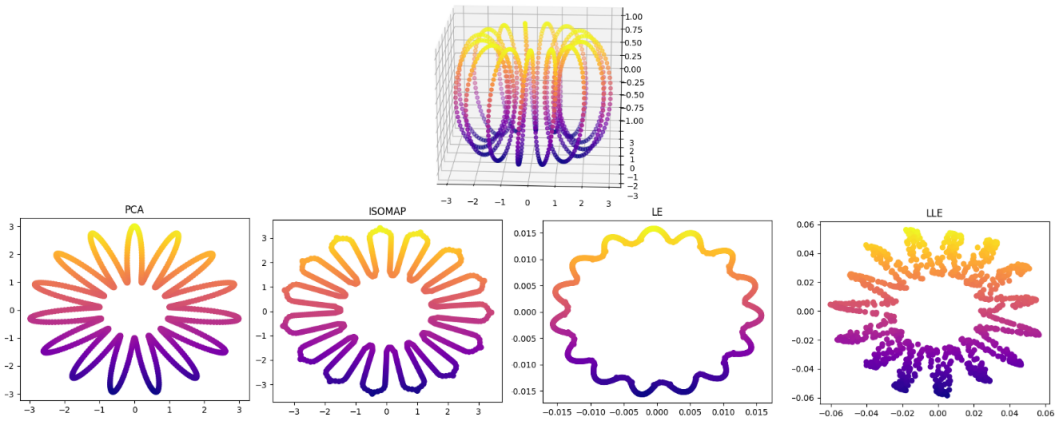


Figura 16: incrustamiento generado por PCA, ISOMAP, LE y LLE (10 vecindarios) en el conjunto de datos toroide

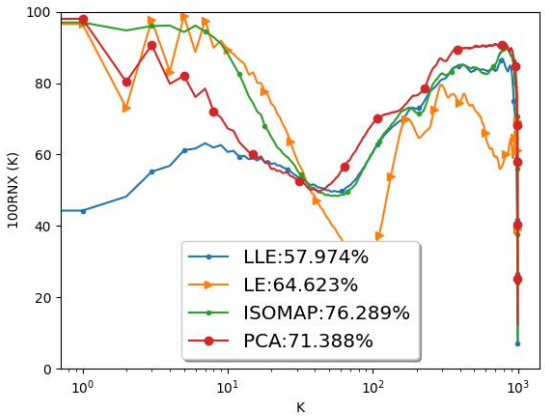


Figura 17: Evaluación PCA, LE, ISOMAP, LLE (10 vecindarios) con el conjunto de datos toroide

Mientras que métodos como PCA tienen un mejor rendimiento en la topología global, LLE e ISOMAP conservan más la local como se muestra en las figuras Figura y Figura, en el conjunto de datos Rollo Suizo, por otro lado, aquellos métodos RD que requieren una configuración previa (número de vecindarios), aunque leve, existe un cambio en la preservación lograda, por ejemplo ISOMAP con 12 vecindarios, tiene una mejor preservación topológica que con 15 vecindarios tal como lo muestra las figuras Figura y Figura en el conjunto de datos Esfera. Finalmente, en la Figura el método que mejor preservó la topología del toroide fue ISOMAP con un 76.289%.

Discusión y conclusiones

QARNX Evalúa satisfactoriamente los métodos RD, ofreciendo una interfaz agradable e intuitiva para el usuario, los flujos que la herramienta permite realizar, son flexibles y las visualizaciones tanto de las incrustaciones como de las evaluaciones son claras y fáciles de interpretar y los resultados obtenidos son acordes a lo que muchas autores afirman en sus trabajos de investigación, finalmente, QARNX se convierte en una herramienta que cualquier persona puede utilizar, tenga o no conocimiento sobre métodos RD y las métricas que los evalúan.

Agradecimientos

Agradecemos a Carlos David Correa Lozano, Diego Ferley Urrea Burgos y Juan Andrés Lozano Thomé, creadores de QARNX y del presente documento, y al docente Juan Carlos Alvarado Pérez, quién en su rol de Asesor de proyecto de grado, ha estado en todo el proceso investigativo de los estudiantes.

Referencias

[1] J. Riquelme, R. Ruíz and K. Gilbert., "Minería de Datos: Conceptos y Tendencias," *Revista Iberoamericana de Inteligencia Artificial,* vol. 10, pp. 11-18, 2006.

[2] J. HAN, M. KAMBER and J. PEI, Data Mining: Concepts and Techniques., ELSEVIER, 2011.

[3] J. Salazar, P. Diego and otros, "Dimensionality reduction for interactive data visualization via Geo-Desic approach," *IEEE Latin American Conference on Computational Intelligence,* pp. 1-6, 2016.

[4] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Association for Computational Linguistics,* pp. 79--86, 2002.

[5] J. Valencia, G. Daza, C. Acosta and G. Castellanos, "Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades," *Tecno Lógicas,* pp. 131-150, 2010.

[6] D. Michael, S. Laurent, Q. Christine, F. Stanley and T. Cole, "Dimensionality Reduction By UMAP To Visualize Physical And Genetic Interactions," *nature COMUNICATIONS,* 2020.

[7] V. Jarkko and K. Samuel, "Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study," *Proceedings of the International Conference on Artificial Neural Networks,* 2001.

[8] C. Lisha and B. Andreas, "Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis," *Journal of the American Statistical Association,* pp. 209-219, 2009.

[9] L. John and V. Michel, Nonlinear Dimensionality Reduction, Nueva York: Springer, New York, NY, 2007.

[10] J. Lee and V. Michel, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing,* vol. Vol.72, pp. 1431-1443, 2009.

[11] I. FODOR, "A survey of dimensionality reduction techniques," *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory,* 2002.

[12] S. AYESHA, M. KASHIF and R. TALIB, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion,* vol. 59, pp. 44-58, 2020.

[13] L. SAUL and S. ROWEYS, "An introduction to localy linear embedding.," *Journal of Machine Learning Research,* 2001.

[14] S. Jonathon, "A Tutorial on Principal Component Analysis," *International Journal of Remote Sensing,* 2015.

[15] M. BELKIN and P. NIYOGI, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems,* 2001.

Anexo D: Carta del asesor para jurados

San Juan de Pasto, 4 de octubre del 2024

Estimado
Comité de Investigación
UNIVERSIDAD CESMAG
Ciudad: Pasto-Nariño

Cordial saludo.

Ref. Certificación Asesoría y entrega Proyecto de grado

Por medio de la presente remitimos el informe final del proyecto de investigación denominado: "**EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS R_{NX}**", realizada por los estudiantes ***Carlos David Correa Lozano, Diego Ferley Urrea Burgos y Juan Andrés Lozano Thomé***, de décimo semestre del programa de Ingeniería de Sistemas, quienes han asistido a las asesorías para el desarrollo y consolidación del proyecto de investigación, por lo tanto, se somete el mismo a la evaluación de los jurados lectores quienes emitirán un concepto sobre el mismo dentro de los plazos establecidos.

Atentamente,

Ing. Carlos Fernando Gonzáles Guzmán
Docente asesor
Programa Ingeniería de Sistemas
Universidad Cesmag

1

UNIVERSIDAD CESMAG NIT: 800.109.387-7 VIGILADA MINEDUCACION	CARTA DE ENTREGA TRABAJO DE GRADO O TRABAJO DE APLICACIÓN – ASESOR(A)	**CÓDIGO:** AAC-BL-FR-032
		VERSIÓN: 1
		FECHA: 26/NOV/2024

San Juan de Pasto, 26 de noviembre de 2024

Biblioteca
REMIGIO FIORE FORTEZZA OFM. CAP.
Universidad CESMAG
Pasto

Saludo de paz y bien.

Por medio de la presente se hace entrega del Trabajo de Grado / Trabajo de Aplicación denominado "**EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS Rnx**", presentado por el autor Diego Ferley Urrea Burgos del Programa Académico Ingeniería de Sistemas al correo electrónico biblioteca.trabajosdegrado@unicesmag.edu.co. Manifiesto como asesor(a), que su contenido, resumen, anexos y formato PDF cumple con las especificaciones de calidad, guía de presentación de Trabajos de Grado o de Aplicación, establecidos por la Universidad CESMAG, por lo tanto, se solicita el paz y salvo respectivo.

Atentamente,

(Firma del Asesor)

Carlos Fernando Gonzáles Guzmán
C.C 98379634 Pasto
Ing. Sistemas
Teléfono de contacto: 3122279731
Correo electrónico: cfgonzalez@unicesmag.edu.co

INFORMACIÓN DEL (LOS) AUTOR(ES)	
Nombres y apellidos del autor: Diego Ferley Urrea Burgos	**Documento de identidad:** 1233193928
Correo electrónico: urreadiego767@gmail.com	**Número de contacto:** 3155152941
Nombres y apellidos del asesor: Carlos Fernando Gonzáles Guzmán	**Documento de identidad:** 98379634
Correo electrónico: **cfgonzalez@unicesmag.edu.co**	**Número de contacto:** **3122279731**
Título del trabajo de grado: EVALUACIÓN DE MÉTODOS DE REDUCCIÓN DE DIMENSIÓN PARA LA PRESERVACIÓN TOPOLÓGICA DE LOS DATOS MEDIANTE MÉTRICAS Rnx	
Facultad y Programa Académico: Ingeniería de Sistemas	

En mi (nuestra) calidad de autor(es) y/o titular (es) del derecho de autor del Trabajo de Grado o de Aplicación señalado en el encabezado, confiero (conferimos) a la Universidad CESMAG una licencia no exclusiva, limitada y gratuita, para la inclusión del trabajo de grado en el repositorio institucional. Por consiguiente, el alcance de la licencia que se otorga a través del presente documento, abarca las siguientes características:

a) La autorización se otorga desde la fecha de suscripción del presente documento y durante todo el término en el que el (los) firmante(s) del presente documento conserve (mos) la titularidad de los derechos patrimoniales de autor. En el evento en el que deje (mos) de tener la titularidad de los derechos patrimoniales sobre el Trabajo de Grado o de Aplicación, me (nos) comprometo (comprometemos) a informar de manera inmediata sobre dicha situación a la Universidad CESMAG. Por consiguiente, hasta que no exista comunicación escrita de mi(nuestra) parte informando sobre dicha situación, la Universidad CESMAG se encontrará debidamente habilitada para continuar con la publicación del Trabajo de Grado o de Aplicación dentro del repositorio institucional. Conozco(conocemos) que esta autorización podrá revocarse en cualquier momento, siempre y cuando se eleve la solicitud por escrito para dicho fin ante la Universidad CESMAG. En estos eventos, la Universidad CESMAG cuenta con el plazo de un mes después de recibida la petición, para desmarcar la visualización del Trabajo de Grado o de Aplicación del repositorio institucional.
b) Se autoriza a la Universidad CESMAG para publicar el Trabajo de Grado o de Aplicación en formato digital y teniendo en cuenta que uno de los medios de publicación del repositorio institucional es el internet, acepto(amos) que el Trabajo de Grado o de Aplicación circulará con un alcance mundial.
c) Acepto (aceptamos) que la autorización que se otorga a través del presente documento se realiza a título gratuito, por lo tanto, renuncio(amos) a recibir emolumento alguno por la publicación, distribución, comunicación pública y/o cualquier otro uso que se haga en los términos de la presente autorización y de la licencia o programa a través del cual sea publicado el Trabajo de grado o de Aplicación.
d) Manifiesto (manifestamos) que el Trabajo de Grado o de Aplicación es original realizado sin violar o usurpar derechos de autor de terceros y que ostento(amos) los derechos patrimoniales de autor sobre la misma. Por consiguiente, asumo(asumimos) toda la responsabilidad sobre su contenido ante la Universidad CESMAG y frente a terceros, manteniéndose indemne de cualquier reclamación que surja en virtud de la misma. En todo caso, la Universidad CESMAG se

compromete a indicar siempre la autoría del escrito incluyendo nombre de(los) autor(es) y la fecha de publicación.

e) Autorizo(autorizamos) a la Universidad CESMAG para incluir el Trabajo de Grado o de Aplicación en los índices y buscadores que se estimen necesarios para promover su difusión. Así mismo autorizo (autorizamos) a la Universidad CESMAG para que pueda convertir el documento a cualquier medio o formato para propósitos de preservación digital.

NOTA: En los eventos en los que el trabajo de grado o de aplicación haya sido trabajado con el apoyo o patrocinio de una agencia, organización o cualquier otra entidad diferente a la Universidad CESMAG. Como autor(es) garantizo(amos) que he(hemos) cumplido con los derechos y obligaciones asumidos con dicha entidad y como consecuencia de ello dejo(dejamos) constancia que la autorización que se concede a través del presente escrito no interfiere ni transgrede derechos de terceros.

Como consecuencia de lo anterior, autorizo(autorizamos) la publicación, difusión, consulta y uso del Trabajo de Grado o de Aplicación por parte de la Universidad CESMAG y sus usuarios así:

- Permito(permitimos) que mi(nuestro) Trabajo de Grado o de Aplicación haga parte del catálogo de colección del repositorio digital de la Universidad CESMAG por lo tanto, su contenido será de acceso abierto donde podrá ser consultado, descargado y compartido con otras personas, siempre que se reconozca su autoría o reconocimiento con fines no comerciales.

En señal de conformidad, se suscribe este documento en San Juan de Pasto a los 26 días del mes de noviembre del año 2024

Firma del autor	Firma del autor
Nombre del autor: Diego Ferley Urrea Burgos	Nombre del autor:
Firma del autor	Firma del autor
Nombre del autor:	Nombre del autor:
Firma del asesor Nombre del asesor: Carlos Fernando Gonzáles Guzmán	