|  |
| --- |
| **UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS**  **Facultad 3**  **Centro de Estudio en Gestión de Proyectos y Toma de Decisiones**    **ALGORITMOS PARA LA SUMARIZACIÓN LINGÜÍSTICA DE DATOS PARA LA AYUDA A LA TOMA DE DECISIONES**  **Tesis presentada en opción al Titulo**  **Máster en Gestión de Proyectos**  **Ing YYYYYY ZZZZ**  **La Habana**  **20xx** |

|  |
| --- |
| **UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS**  **Facultad 3**  **Centro de Estudio en Gestión de Proyectos y Toma de Decisiones**    **ALGORITMOS PARA LA SUMARIZACIÓN LINGÜÍSTICA DE DATOS PARA LA AYUDA A LA TOMA DE DECISIONES**  **Tesis presentada en opción al Título**  **Máster en Gestión de Proyectos**  **Autor: Ing. YYY XXXX**  **Tutores: Prof. Titular, Dr.C ZZZZZ**  **Prof. Titular, Dr. C QQQQQ**  **La Habana**  **20xx** |

AGRADECIMIENTOS

*, yyyyy.*

DEDICATORIA

*xxxxxx.*

SÍNTESIS

El desarrollo de la tecnología de la información causa incremento en los registros de datos en los disímiles escenarios de toma de decisiones, cuyo procesamiento y necesaria interpretación conlleva al uso de herramientas o técnicas que permitan el descubrimiento de patrones de comportamiento y su interpretabilidad. Este problema se incrementa en escenarios donde con frecuencia se subutiliza la información que se dispone, donde no se cuenta con suficientes herramientas que permitan el descubrimiento de las dependencias no triviales ocultas en los datos ni con herramientas que permitan representar estas dependencias en una forma legible y entendible por el ser humano; afectando en cierta medida, el proceso de toma de decisiones por parte de los decisores.

En este contexto surge la sumarización lingüística de datos como una de las técnicas de descubrimiento de conocimiento descriptivo con un enfoque prometedor para producir resúmenes a partir de una base de datos utilizando lenguaje natural, donde autores como Yager y Zadeh fueron pioneros y marcaron pautas en el desarrollo de estas técnicas. No obstante, es un área de conocimiento con muchas oportunidades de mejora y líneas abiertas a la investigación.

En esta tesis se analizan las tendencias en la sumarización lingüística de datos y las insuficiencias de los métodos existentes tanto para la generación como para la evaluación de resúmenes lingüísticos de datos. Se proponen nuevos algoritmos para la generación de resúmenes lingüísticos a partir de datos que combinan diferentes técnicas de computación emergente, entre las que se destacan la teoría neutrosófica, los conjuntos aproximados, el aprendizaje de reglas de asociación y el aprendizaje de grafos probabilísticos. También se discuten diferentes enfoques que permiten la aplicación de la sumarización lingüística de datos en el descubrimiento y representación de situaciones con presencia de datos anómalos. Para facilitar el enfoque multilingüe de los resúmenes lingüísticos generados, los algoritmos propuestos hacen uso de los lenguajes naturales controlados (LNC), donde se proponen LNC asociados a los idiomas español, inglés, japonés y árabe. Los algoritmos propuestos se comparan con otras técnicas reportadas en la bibliografía a partir de su desempeño en 12 bases de datos, demostrando la superioridad de la propuesta respecto a la eficiencia, el tratamiento de la incertidumbre, la interpretabilidad de los resúmenes lingüísticos y su generación en múltiples idiomas.

Otro aporte de la investigación radica en que se proponen nuevas extensiones a los indicadores para la evaluación de los resúmenes lingüísticos que complementan a los indicadores existentes e incorporan elementos de la teoría neutrosófica mejorando el tratamiento de la indeterminación. Finalmente se presentan las conclusiones y recomendaciones de la investigación.

ÍNDICE

[Introducción 1](#_Toc84598920)

[1 Capítulo 1: Fundamentación teórica 10](#_Toc84598921)

[1.1 Conceptos generales sobre sumarización lingüística de datos 10](#_Toc84598922)

[1.2 Métodos o técnicas para la generación de los resúmenes lingüísticos de datos 14](#_Toc84598923)

[1.3 Principales técnicas y métodos de validación empleados en las investigaciones 18](#_Toc84598924)

[1.4 Áreas de aplicación de los resúmenes lingüísticos 20](#_Toc84598925)

[1.5 Conclusiones del capítulo 24](#_Toc84598926)

[2 Capítulo 2: Algoritmos para la sumarización lingÜística de datos 26](#_Toc84598927)

[2.1 Notaciones y definiciones fundamentales empleadas en el capítulo. 26](#_Toc84598928)

[2.2 Estructura de los resúmenes que se construyen siguiendo enfoque multilingüe. 27](#_Toc84598929)

[2.3 Algoritmo *PCA\_LDS* basado en análisis de componentes principales 32](#_Toc84598931)

[2.4 Algoritmo *LPA\_LDS* basado en gráficos probabilísticos 36](#_Toc84598932)

[2.5 Algoritmo RST\_LDS basado en conjuntos aproximados 38](#_Toc84598933)

[2.6 Diferentes enfoques que vinculan la generación de resúmenes lingüísticos y la minería de datos anómalos 42](#_Toc84598934)

[2.7 Valoración general de los algoritmos propuestos 45](#_Toc84598935)

[2.8 Conclusiones del capítulo 46](#_Toc84598936)

[3 Capítulo 3: Nuevos métodos de evaluación de los resúmenes lingüísticos 48](#_Toc84598937)

[3.1 Indicadores tradicionales para la evaluación de los resúmenes lingüísticos 48](#_Toc84598938)

[3.2 Propuestas de extensiones de *T* para evaluar resúmenes lingüísticos 53](#_Toc84598939)

[3.3 Conjunto de pruebas 1: Validación y comparación de los indicadores tradicionales y los extendidos 59](#_Toc84598940)

[3.4 Conclusiones del capítulo 66](#_Toc84598941)

[4 Capítulo 4 Validación de la eficacia de los algoritmos propuestos 68](#_Toc84598942)

[4.1 Parámetros de algoritmos y bases de datos empleadas en la experimentación 68](#_Toc84598943)

[4.2 Conjunto de pruebas 2: validación de variable independiente respecto a la dimensión indicadores de calidad de los resúmenes lingüísticos. 70](#_Toc84598944)

[4.3 Conjunto de pruebas 3: validación de la variable independiente en la dimensión “Eficiencia de los algoritmos”. 82](#_Toc84598945)

[4.4 Conjunto de pruebas 4: validación de la variable independiente en la dimensión “Desempeño global de los algoritmos” 83](#_Toc84598946)

[4.5 Conjunto de pruebas 5: validación de variable dependiente “Eficacia de los algoritmos para la ayuda a la toma de decisiones” 85](#_Toc84598947)

[4.6 Conjunto de pruebas 6: validación de la variable dependiente en la dimensión “Enfoque multilingüe de los algoritmos”. 95](#_Toc84598948)

[4.7 Conclusiones del capítulo 97](#_Toc84598949)

[5 Conclusiones generales 98](#_Toc84598950)

[6 Recomendaciones 100](#_Toc84598951)

[7 Glosario de términos 101](#_Toc84598952)

[8 Bibliografía 104](#_Toc84598953)

[9 Anexos 118](#_Toc84598954)

[9.1 Anexo 1. Producción científica del autor 118](#_Toc84598955)

[9.2 Anexo 2. Protocolo de revisión sistemática aplicado 119](#_Toc84598956)

[9.3 Anexo 3. Caracterización de la bibliografía consultada en la revisión sistemática 121](#_Toc84598957)

[9.4 Anexo 4. Taxonomía de técnicas para la detección de datos anómalos 124](#_Toc84598958)

[9.5 Anexo 5. Taxonomía del concepto de incertidumbre 124](#_Toc84598959)

[9.6 Anexo 6. Evolución de la teoría neutrosófica 124](#_Toc84598960)

[9.7 Anexo 7. Escenarios de pruebas para la comparación de los indicadores de calidad 125](#_Toc84598961)

[9.8 Anexo 8. Resumen de bases de datos empleadas en la experimentación 126](#_Toc84598962)

[9.9 Anexo 9. Lenguajes Naturales Controlados 127](#_Toc84598963)

[9.10 Anexo 10. Ejemplos de diccionarios 133](#_Toc84598964)

[9.11 Anexo 11. Especialistas que participaron en los casos de estudio. 135](#_Toc84598965)

[9.12 Anexo 12. Encuestas aplicadas a los especialistas de los casos de estudio 135](#_Toc84598966)

[9.13 Anexo 13. Resultado de los algoritmos. 137](#_Toc84598967)

[*9.14* Anexo 14. Resultados de la aplicación del *Page´L Trent Test* 142](#_Toc84598968)

[9.15 Anexo 15. Expertos que participaron en la validación del enfoque multilingüe 143](#_Toc84598969)

[9.16 Anexo 16. Resúmenes en múltiples idiomas generados en el caso de estudio 144](#_Toc84598970)

**Í****NDICE DE TABLAS**

[Tabla 1: Operacionalización de la variable dependiente. 5](#_Toc84598971)

[Tabla 2: Operacionalización de las variables independientes. 5](#_Toc84598972)

[Tabla 3: Clasificación de resúmenes lingüísticos (Kacprzyk y Zadrożny 2009). 12](#_Toc84598973)

[Tabla 4: Funciones de pertenencia que representan a los cuantificadores. 27](#_Toc84598974)

[Tabla 5: Comparación teórica de los algoritmos. 45](#_Toc84598975)

[Tabla 6: Ejemplo de descripción de un objeto de la clase de prueba ER1. 60](#_Toc84598976)

[Tabla 7: Comparación de indicadores respecto a tratamiento de incertidumbre. 66](#_Toc84598977)

[Tabla 8: Comparación de los algoritmos respecto a la media aritmética del *T1*. 71](#_Toc84598978)

[Tabla 9: Comparación de los algoritmos respecto a la media aritmética del *T2*. 72](#_Toc84598979)

[Tabla 10: Comparación de los algoritmos respecto a la media aritmética del *T3*. 73](#_Toc84598980)

[Tabla 11: Comparación de los algoritmos respecto a la media aritmética del *T4*. 73](#_Toc84598981)

[Tabla 12: Comparación de los algoritmos respecto a la media aritmética del *T5*. 75](#_Toc84598982)

[Tabla 13: Comparación de los algoritmos respecto a la media aritmética del *Te1a*. 75](#_Toc84598983)

[Tabla 14: Comparación de los algoritmos respecto a la media aritmética del *Te1b*. 76](#_Toc84598984)

[Tabla 15: Comparación de los algoritmos respecto a la media aritmética del *Te3*. 77](#_Toc84598985)

[Tabla 16: Comparación de los algoritmos respecto a la media aritmética del *Te4*. 78](#_Toc84598986)

[Tabla 17: Comparación de los algoritmos respecto a la media aritmética del *Te5*. 79](#_Toc84598987)

[Tabla 18: Comparación de los algoritmos respecto a la dispersión. 80](#_Toc84598988)

[Tabla 19: Comparación de los algoritmos respecto a la fortaleza de las dependencias descubiertas. 81](#_Toc84598989)

[Tabla 20: Comparación de los algoritmos respecto al tiempo de ejecución. 82](#_Toc84598990)

[Tabla 21: Complejidad de los algoritmos. 83](#_Toc84598991)

[Tabla 22: Clases de prueba para los escenarios de prueba. 125](#_Toc84598992)

[Tabla 23: Ejemplo de objetos de la clase de prueba ER1. 126](#_Toc84598993)

[Tabla 24: Descripción de las bases de datos. 126](#_Toc84598994)

[Tabla 25: Ejemplo de aplicación del lenguaje controlado en resúmenes lingüísticos. 128](#_Toc84598995)

[Tabla 26: Especialistas que participaron en el caso de estudio sobre embarazadas cardiópatas. 135](#_Toc84598996)

[Tabla 27: Expertos que participaron en el caso de estudio sobre gestión de proyectos. 135](#_Toc84598997)

[Tabla 28 Encuesta aplicada a los especialistas en el tratamiento de embarazadas cardiópatas. 135](#_Toc84598998)

[Tabla 29: Encuesta aplicada a los expertos en gestión de proyectos. 136](#_Toc84598999)

[Tabla 30: Resultado de los algoritmos respecto a la media y desviación estándar de los indicadores en las bases de datos sobre gestión de proyectos. 137](#_Toc84599000)

[Tabla 31: Resultado de los algoritmos respecto a la media y desviación estándar de los indicadores en las bases de datos sobre auditorías. 139](#_Toc84599001)

[Tabla 32: Resultado de los algoritmos respecto a la media y desviación estándar de los indicadores en las bases de datos médicas. 140](#_Toc84599002)

[Tabla 33: Resultado de los algoritmos respecto al tiempo de ejecución. 141](#_Toc84599003)

[Tabla 34: Resultado de los algoritmos respecto al Cubrimiento de diferentes situaciones en las bases de datos. 141](#_Toc84599004)

[Tabla 35: Resultado de los algoritmos respecto a la Fortaleza de las dependencias en los resúmenes. 142](#_Toc84599005)

[Tabla 36: Resumen de posiciones de los algoritmos respecto a cada indicador de comparación. 142](#_Toc84599006)

[Tabla 37: Expertos que participaron en la revisión de los resúmenes lingüísticos en diferentes idiomas. 143](#_Toc84599007)

**ÍNDICE DE FIGURAS**

[Figura 1: Variable lingüística de los cuantificadores (*Q*) basadas en funciones campana Gauss. 28](#_Toc84599008)

[Figura 2: Variable lingüística de los cuantificadores (*Q*) basada en funciones triangulares. 29](#_Toc84599009)

[Figura 3: Comportamiento del](#_Toc84599010) *T*5. 52

[Figura 4: Conjunto borroso incorporando los valores de veracidad, neutralidad y falsedad](#_Toc84599011) 54

[Figura 5: Comportamiento del *T*1 y *Te1a* en resúmenes con filtro. 61](#_Toc84599012)

[Figura 6: Comportamiento del *T1* y *Te1a* en resúmenes sin filtro. 62](#_Toc84599013)

[Figura 7: Comportamiento del T3 y Te3. 64](#_Toc84599014)

[Figura 8: Comportamiento del T4 y Te4. 65](#_Toc84599015)

[Figura 9: Curvas del *T*5 tradicional y *Te*5 extendido. 66](#_Toc84599016)

[Figura 10: Media aritmética del indicador](#_Toc84599017) *T1*. 70

[Figura 11: Media aritmética del indicador](#_Toc84599018) *T2*. 71

[Figura 12: Media aritmética del indicador](#_Toc84599019) *T3*. 72

[Figura 13: Media aritmética del indicador](#_Toc84599020) *T4*. 74

[Figura 14: Comportamiento del indicador](#_Toc84599021) *T5* 74

[Figura 15: Comportamiento del indicador](#_Toc84599022) *Te1a* 75

[Figura 16: Media aritmética del indicador](#_Toc84599023) *Te1b*. 76

[Figura 17: Media aritmética del indicador](#_Toc84599024) *Te3*. 77

[Figura 18: Media aritmética del indicador](#_Toc84599025) *Te4*. 78

[Figura 19: Media aritmética del indicador](#_Toc84599026) *Te5*. 79

[Figura 20: Complementariedad de los algoritmos en el caso de estudio sobre embarazadas cardiópatas.](#_Toc84599027) 86

[Figura 21: Resultados de la agregación de las preferencias de los especialistas médicos sobre embarazadas cardiópatas.](#_Toc84599028) 88

[Figura 22: Dispersión de opiniones de especialistas por criterio, en el caso de estudio sobre embarazadas cardiópatas.](#_Toc84599029) 88

[Figura 23: Complementariedad de los algoritmos en el caso de estudio sobre proyectos.](#_Toc84599030) 89

[Figura 24: Resultados de la agregación de las preferencias de los expertos en gestión de proyectos.](#_Toc84599031) 95

[Figura 25: Dispersión de opiniones de expertos por criterio, en el caso de estudio sobre gestión de proyectos.](#_Toc84599032) 95

[Figura 26: Resultado de la evaluación asignada por los expertos, a los resúmenes lingüísticos en los idiomas inglés, japonés y árabe.](#_Toc84599033) 96

[Figura 27: Publicaciones por países de la bibliografía consultada. 121](#_Toc84599034)

[Figura 28: Bibliografía por tipo de publicación e indexado. 121](#_Toc84599035)

[Figura 29: Fuentes y editoriales principales de la bibliografía analizada en profundidad 121](#_Toc84599036)

[Figura 30: Tendencias de los métodos o técnicas para generar resúmenes lingüísticos de datos. 122](#_Toc84599037)

[Figura 31: Estrategias y técnicas de validación empleadas en la bibliografía estudiada. 122](#_Toc84599038)

[Figura 32: Alcance de la validación empleada en la bibliografía estudiada. 123](#_Toc84599039)

[Figura 33: Análisis de publicaciones respecto a los entornos de aplicación. 123](#_Toc84599040)

[Figura 34: Taxonomía de técnicas para el descubrimiento de datos anómalos. 124](#_Toc84599041)

[Figura 35: Taxonomía que representa el concepto de incertidumbre. 124](#_Toc84599042)

[Figura 36: Evolución de la teoría neutrosófica. 124](#_Toc84599043)

[Figura 37: Variable lingüística de los cuantificadores usados en las pruebas. 126](#_Toc84599044)

# Introducción

Varios autores (Dijkman y Wilbik 2017) (Amghar y Chikh 2018) (Kaczmarek-Majer et al. 2019) (Peláez-Aguilera et al. 2019) aseveran que uno de los problemas latentes en los procesos de toma de decisiones no estructuradas, está asociado al descubrimiento de patrones de comportamiento en datos provenientes de sistemas de información. Por lo general estos datos no son legibles ni entendibles a simple vista por los decisores, lo cual unido al volumen de datos que se genera, dificulta los procesos de toma de decisiones (Wilbik, Barreto y Backus 2020). En este sentido, constituye objeto de interés, el desarrollo de técnicas y herramientas que permitan el descubrimiento de las dependencias no triviales ocultas en los datos que faciliten su interpretabilidad (Wu y Mendel 2010) (Igde et al. 2017) (Amghar y Chikh 2018) (Wilbik, Barreto y Backus 2020). Pues los escenarios donde se disponga de estas técnicas, pueden obtener mejores resultados respecto a aquellos que no lo hacen.

En este contexto, dentro de la inteligencia artificial[[1]](#footnote-2) se han creado técnicas como la Sumarización Lingüística de Dato[[2]](#footnote-3) (SLD) que permiten identificar las relaciones subyacentes en los datos, tales como: correlaciones, tendencias, agrupaciones o anomalías (Kacprzyk y Zadrożny 2000). Estas técnicas se caracterizan por generar resúmenes lingüísticos a partir de bases de datos estructuradas utilizando lenguaje natural (Yager 1991) facilitando por su naturaleza la toma de decisiones.

Como parte de esta investigación, se desarrolló una revisión sistemática (Pérez Pupo et al. 2021) que abarcó el análisis de más de 71 trabajos, para caracterizar las diferentes técnicas de SLD. En este estudio se identifica que autores como Yager (Yager 1982), Zadeh (Zadeh 1983) y Kacprzyk (Kacprzyk, J. 1999) han sido precursores en el desarrollo de estas técnicas, cuyas publicaciones marcaron pautas en su conceptualización y desarrollo. Según Yager (Yager 1982), las técnicas de sumarización lingüística de datos constituyen un importante método para entender el comportamiento de grandes volúmenes de datos, diseñadas para ayudar a los humanos a entender el entorno de aplicación de una manera sencilla, generando conocimiento descriptivo útil para su empleo futuro. En general, estas técnicas son reconocidas como parte de las técnicas de *soft computing* y son empleadas para la ayuda a la toma de decisiones (Wilbik et al. 2018).

En la revisión se puede identificar que las técnicas de SLD han sido empleadas en diferentes escenarios, entre los que se destacan la gestión de ventas (Kacprzyk y Strykowski 1999), análisis financiero (Kacprzyk y Zadrożny 2016a), la salud (Wilbik et al. 2018) y el análisis de logs (Kacprzyk y Zadrożny 2016b), en este sentido se identifican los siguientes elementos:

* Los entornos de toma de decisiones donde se han aplicado estas técnicas, se caracterizan por ser entornos bajo incertidumbre[[3]](#footnote-4) con presencia de fenómenos como: ambigüedad, indeterminación o imprecisión (García et al. 2017). Sin embargo, en las técnicas reportadas en la bibliografía se aprecia que solo analizan el grado de certeza en la información obviando el tratamiento de la indeterminación o la falsedad.
* La mayoría de los algoritmos reportados en la bibliografía se concentran en la construcción de resúmenes que reflejen comportamientos frecuentes en los datos, pero no los de baja frecuencia. Esto afecta su aplicabilidad en entornos donde se necesite detectar datos anómalos como detección de fraudes en aseguramiento de ingresos, detección de errores en la planificación de proyectos (Castro Aguilar 2017), entre otros.
* A pesar de la alta aplicabilidad de la gestión de proyectos en disímiles sectores de la sociedad, hay pocas investigaciones de SLD en este contexto; lo cual es una de las motivaciones de esta investigación, considerando que se dispone de un repositorio de bases de datos (Piñero et al. 2019) que facilitan la validación de los resultados.

También, se realizó un análisis respecto a la eficacia de los algoritmos para la generación de resúmenes lingüísticos. Entendiendo la eficacia como la capacidad de los algoritmos para generar resúmenes con calidad, interpretables y que faciliten la toma de decisiones bajo un enfoque multilingüe.

Como resultado de este análisis se identificó que la mayoría de los trabajos, se basan en el tratamiento simple de cadenas en idioma inglés (Dijkman y Wilbik 2017) (Marín y Sánchez 2016) (Ramos-Soto y Martin-Rodillab 2019). Esta situación afecta la eficacia de los algoritmos para la generación de los resúmenes respecto a su interpretabilidad y presentación en múltiples idiomas. Se identifican en la bibliografía tres enfoques fundamentales para la generación de resúmenes lingüísticos:

* Extensiones a lenguajes de consultas de bases de datos (Kacprzyk, J. y Zadrożny 2005).
* Generación de resúmenes a partir del aprendizaje de reglas (Wu y Mendel 2010).
* Generación de resúmenes usando metaheurísticas (Donis-Díaz, Bello, et al. 2014).

Entre las limitaciones principales de los trabajos reportados en la bibliografía bajo estos enfoques se puede señalar que:

* Los algoritmos reportados en la bibliografía tales como los basados en reglas (Wu y Mendel 2010) y en metaheurísticas (Donis-Díaz, Muro, et al. 2014), no explotan la información estadística asociada a la correlación entre las variables durante el proceso de búsqueda. Elemento que afecta la eficacia de los algoritmos porque pueden generar resúmenes candidatos que no tienen una representación real en los datos. En estos trabajos generalmente se incluyen etapas intermedias de limpieza de resúmenes candidatos.
* En las propuestas realizadas por (Donis-Díaz, Bello y Kacprzyk 2015) se favorece la generación de resúmenes con cuantificadores “Muchos” y “La mayoría”, limitando la generación de resúmenes lingüísticos con otros cuantificadores. Este enfoque no facilita la identificación de situaciones con poca ocurrencia en la base de datos, afectando la eficacia de los algoritmos en la solución de problemas de detección de datos anómalos, tales como la detección de epidemias, de fallas, de aseguramiento de ingresos, entre otros.
* Las metaheurísticas reportadas en la bibliografía (Donis-Díaz, Bello y Kacprzyk 2015) tienen alto costo computacional, elemento que influye negativamente en la aplicabilidad de estos algoritmos en entornos reales de toma de decisiones.

Respecto a los indicadores de calidad para la evaluación de los resúmenes se identifica que los más empleados son los descritos en (Kacprzyk y Zadrożny 2010):

* El grado de verdad (*T*1), el cual mide el porcentaje de veracidad de un resumen respecto a los datos.
* El grado de imprecisión (*T*2), criterio de validez que complementa el grado de verdad, pues un resumen lingüístico impreciso, aunque tenga un alto grado de verdad, no es muy útil.
* El grado de cobertura (*T*3), que indica cuántos objetos en la base de datos están cubiertos o representados por el resumen evaluado.
* El grado de adecuación (*T*4), el cual mide la relevancia de un resumen. Es decir, la capacidad de identificar relaciones de alta correlación entre las variables, aun cuando haya poca ocurrencia de eventos que vinculen a estas variables.
* La longitud de un resumen (*T*5), que es el grado de calidad del resumen respecto al número de variables que involucra.

En el cálculo de estos indicadores de calidad, se utiliza el grado de certidumbre considerando el cero como umbral, sin considerar la indeterminación o la falsedad (Pérez Pupo et al. 2021). Esta limitante de los indicadores reportados en la bibliografía constituye una de las motivaciones y línea de trabajo en esta investigación.

Por otra parte, se analizaron las diferentes técnicas empleadas para la validación en los trabajos reportados en la bibliografía. Se señala en este aspecto que la mayoría de los trabajos solo aplican análisis descriptivo en casos de estudio. Pocos trabajos aplican métodos cuantitativos, técnicas de triangulación de datos y triangulación de métodos, elemento que denota baja fortaleza en la validación de las investigaciones (Pérez Pupo et al. 2021).

A partir de estos análisis se identifica el siguiente problema de investigación:

**Problema de investigación**

¿Cómo mejorar la eficacia de los algoritmos de sumarización lingüística de datos para la ayuda a la toma de decisiones?

El **objeto de investigación** es la: sumarización lingüística de datos.

**Objetivo general**

Desarrollar nuevos algoritmos para mejorar la eficacia en la generación y evaluación de resúmenes lingüísticos de datos respecto al tratamiento de la indeterminación, la interpretabilidad y su aplicabilidad en la toma de decisiones bajo un enfoque multilingüe.

**Objetivos específicos**

* Construir un marco teórico referencial que permita identificar las tendencias en los métodos de generación, representación y evaluación de los resúmenes lingüísticos de datos.
* Desarrollar nuevos algoritmos para la generación de resúmenes lingüísticos a partir de la hibridación de técnicas de *soft computing* mejorando: la evaluación de los resúmenes respecto a los indicadores de calidad, el tratamiento de la indeterminación, la interpretabilidad bajo un enfoque multilingüe y el descubrimiento de situaciones anómalas.
* Desarrollar extensiones que mejoren los indicadores de evaluación de resúmenes lingüísticos respecto al tratamiento de la indeterminación y la capacidad de discriminar situaciones diferentes.
* Validar los métodos propuestos a partir de la comparación de éstos con otras técnicas reportadas en la bibliografía aplicando técnicas de triangulación de métodos y triangulación de datos.

El **campo de investigación**: algoritmos para la generación y evaluación de resúmenes lingüísticos de datos.

**Hipótesis**

El desarrollo de nuevos algoritmos para la generación de resúmenes lingüísticos y nuevos indicadores de evaluación de resúmenes bajo un enfoque multilingüe con centro en el tratamiento de la incertidumbre, mejorará la eficacia en la generación de los resúmenes lingüísticos de datos y su aplicabilidad en la toma de decisiones.

**Métodos de investigación**

En la investigación se aplican los siguientes métodos teóricos:

* Histórico-lógico: en la primera parte de la investigación se construye un marco teórico referencial a partir de realizar una revisión sistemática sobre métodos para la generación y evaluación de resúmenes lingüísticos de datos.
* Hipotético deductivo: en el transcurso de la investigación, la hipótesis es resuelta siguiendo métodos fundamentados científicamente y luego se realizan pruebas estadísticas para demostrar la validez de los resultados.

Además, se aplican los siguientes métodos empíricos:

* Experimental: se diseñan experimentos para la comparación de los algoritmos propuestos con los reportados en la bibliografía, en las comparaciones se aplican técnicas de triangulación de métodos y triangulación de datos. Se comparan los algoritmos respecto a su desempeño en 12 bases de datos de diferentes dominios de aplicación.
* Los algoritmos propuestos se aplican en casos de estudio sobre toma de decisiones en proyectos y sobre estudios clínicos en embarazadas cardiópatas.

**Operacionalización de variables y diseño de experimentos**

Tabla 1: Operacionalización de la variable dependiente.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable dependiente** | **Dimensión** | **Medida** | **Unidad de medida** |
| Eficacia de los algoritmos para la ayuda a la toma de decisiones | Aplicabilidad en la toma de decisiones | Aplicabilidad en diferentes escenarios | Sí o No |
| Facilidad para la toma de decisiones | Cualitativo |
| Interpretabilidad de resúmenes |
| Enfoque multilingüe de los algoritmos | Generación de resúmenes lingüísticos en diferentes idiomas | Sí o No |

Tabla 2: Operacionalización de las variables independientes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable independiente** | **Dimensión** | **Medida** | **Unidad de medida** |
| Indicadores de evaluación de resúmenes | Efectividad indicadores de evaluación | Tratamiento de la incertidumbre[[4]](#footnote-5) | Sí o No |
| Comportamiento en escenarios de pruebas de los indicadores | Evaluación cualitativa |
| Algoritmos para la generación de resúmenes lingüísticos de datos | Calidad de los resúmenes | Grado de verdad | [0, 1] |
| Grado de imprecisión |
| Grado de cobertura |
| Grado de adecuación |
| Longitud de un resumen |
| Fortaleza de las relaciones |
| Grado de cubrimiento de situaciones con alta y baja ocurrencia en BD |
| Eficiencia de algoritmos | Complejidad computacional | Complejidad |
| Tiempo de ejecución | Minutos |
| Desempeño integral de los algoritmos | Evaluación integral de los algoritmos considerando eficacia y eficiencia | Orden de los algoritmos |

Para evaluar la variable dependiente y las independientes se diseñan conjuntos de pruebas por cada dimensión como se muestra a continuación.

En el capítulo 3 se evalúa la variable independiente “Indicadores de evaluación de resúmenes”, para ello se aplica el conjunto de pruebas 1:

1. Se compara el comportamiento de los indicadores de evaluación tradicionales con las nuevas extensiones propuestas considerando las medidas: “Capacidad para el tratamiento de la incertidumbre” y “Comportamiento en los escenarios de prueba”. Para las comparaciones se simulan 29 posibles escenarios sobre los que se realizan los siguientes análisis:

* Comportamiento del indicador grado de verdad y sus extensiones.
* Comportamiento del indicador de soporte y sus extensiones.
* Comportamiento del indicador grado de adecuación y sus extensiones.
* Comportamiento de los indicadores asociados al tamaño del resumen.

En el capítulo 4 se evalúa la variable independiente “Algoritmos para la generación y evaluación de resúmenes lingüísticos de datos”, combinando técnicas de triangulación de métodos y triangulación de datos, y se realizan los siguientes conjuntos de pruebas:

1. Se evalúa la variable dependiente respecto a la dimensión “Calidad de los resúmenes”. En este caso se comparan los algoritmos propuestos con diferentes algoritmos reportados en la bibliografía tales como: la metaheurística propuesta en (Donis-Díaz, Bello y Kacprzyk 2015) y el algoritmo basado en reglas de asociación propuestos en (Dijkman y Wilbik 2017) de la siguiente forma:

* Se comparan los resultados obtenidos por los algoritmos, respecto a los indicadores tradicionales *T1, T2, T3, T4* y *T5* y respecto a los indicadores extendidos *Te1a, Te1b, Te3, Te4* y *Te5*.
* Se aplican técnicas de triangulación de datos en las siguientes 12 bases de datos:
  + 7 bases de datos públicas en el Repositorio de Investigaciones en Gestión de Proyectos (Piñero et al. 2019), asociadas a evaluación y toma de decisiones en gestión de proyectos.
  + 2 bases de datos relacionadas con auditorías y controles internos realizados a entidades estatales y no estatales del país.
  + 3 bases de datos asociadas a información clínica, 2 de ellas sobre predicción de complicaciones en pacientes embarazadas cardiópatas y la 3ra sobre enfermos de covid-19 en Cuba.
* Comparación de los algoritmos respecto al indicador “Cubrimiento de diferentes situaciones en la base de datos”. Este indicador es importante porque apoya el análisis de los algoritmos respecto a su aplicabilidad en entornos asociados a la minería de datos anómalos.
* Comparación de los algoritmos respecto a la “Fortaleza de las dependencias descubiertas”. Este último indicador fue empleado en (Donis-Díaz, Bello y Kacprzyk 2015) para la comparación de algoritmos, mide la aparición de cuantificadores “la mayoría”, “casi la totalidad” y “muchos” entre los resúmenes.

1. Análisis de la dimensión “Eficiencia de los algoritmos”, donde se comparan los algoritmos propuestos con otros reportados en la bibliografía respecto a la complejidad computacional y al tiempo de ejecución.
2. Análisis de la dimensión “Desempeño integral de los algoritmos” a partir de aplicar el test de ranking *Page´L Trent Test* (Salkind 2007) que integra los resultados de los conjuntos de pruebas 2 y 3.

En las pruebas de comparación de los algoritmos, se emplean pruebas estadísticas bien fundamentadas y la herramienta *IBM SPSS Statistics* versión 25 (McCormick y Salcedo 2017), cumpliendo los siguientes pasos (Grau, Correa y Rojas 1999) para una significación de 0.05 e intervalos de confianza del 99%:

1. Se aplica prueba de normalidad (empleando los *test* de *Shapiro-Wilk y Kolmogorov-Smirnov*) y análisis descriptivo.
2. En caso de que las muestras cumplan con la distribución normal, se aplican pruebas paramétricas empleando *T-Student* para dos muestras relacionadas.
3. En caso de que las muestras no cumplan con la distribución normal, se aplican pruebas no paramétricas empleando *Friedman* para *n* muestras relacionadas.
4. Si *Friedman* indica que no hay diferencias significativas en las *n* muestras relacionadas, entonces se concluye que no hay diferencias significativas.
5. Si *Friedman* indica que hay diferencias significativas en las *n* muestras relacionadas, entonces se aplican pruebas *post-hoc* *Test Wilcoxon* para dos muestras relacionadas.

En este análisis de resultados, los algoritmos se organizan según el orden establecido en el análisis descriptivo.

También en el capítulo 4 se evalúa la variable dependiente “Eficacia de los algoritmos para la ayuda a la toma de decisiones” a partir de los siguientes conjuntos de pruebas:

1. Evaluación de la variable dependiente en la dimensión “Aplicabilidad de los algoritmos propuestos en la toma de decisiones”. Para esto se combinan los algoritmos propuestos en un meta-algoritmo *Hybrid\_LDS* y se aplican en los casos de estudio: “toma de decisiones en la gestión de proyectos” y “toma de decisiones en embarazadas cardiópatas”. Además, en cada caso de estudio se analizan los indicadores “Interpretabilidad de los resúmenes” y “Facilidad para la toma de decisiones” a partir de aplicar una técnica de análisis multicriterio.
2. Análisis de la variable dependiente respecto a la dimensión “Enfoque multilingüe” donde se analiza la facilidad de los algoritmos para la generación de resúmenes en los idiomas: español, inglés, japonés y árabe. Se seleccionan estos idiomas porque se dispone de especialistas que puedan validar la calidad de los resúmenes lingüísticos generados.

**Novedad**

* Se proponen nuevas extensiones a los indicadores de evaluación de resúmenes lingüísticos complementando a los reportados en la bibliografía. Estos nuevos indicadores incorporan elementos de la teoría de conjuntos aproximados y de la teoría neutrosófica, mejorando el tratamiento de la indeterminación y el comportamiento de los indicadores en diferentes situaciones.
* Se proponen nuevos algoritmos para la generación de resúmenes lingüísticos a partir de datos estructurados que combinan diferentes técnicas de computación emergente, entre las que se destacan la teoría neutrosófica, los conjuntos aproximados, el aprendizaje de reglas de asociación y el aprendizaje de grafos probabilísticos, mejorando la eficacia de los algoritmos existentes. Se discuten diferentes enfoques de aplicación de la sumarización lingüística de datos en el tratamiento y la representación de situaciones anómalas. Además, se propone el uso de los lenguajes naturales controlados combinados con los algoritmos propuestos para facilitar la representación de los resúmenes lingüísticos en múltiples lenguajes.

**Aporte práctico de la investigación.**

* Se incorporan los algoritmos propuestos a una biblioteca de análisis de datos para la ayuda a la toma de decisiones, integrada a un ecosistema que incluye facilidades de cuadro de mando para la gestión de proyectos y entornos productivos (Pérez Pupo, Villavicencio, Piñero Pérez, et al. 2020) (Pérez et al. 2018) (Pérez Pupo, Santos, et al. 2018) (Piñero Pérez, Pérez Pupo y Piñero Cruz 2019).
* Se desarrolla y se publica un repositorio de bases de datos para el desarrollo de investigaciones en gestión de proyectos (Piñero et al. 2019).

La producción científica del autor asociada a esta investigación se describe en el Anexo 1.

**Estructura de la tesis**

La tesis está organizada en cuatro capítulos. En el primer capítulo se presenta el marco teórico de la investigación y se realiza un análisis de las tendencias en la sumarización lingüística de datos. En el capítulo 2 se presentan nuevos algoritmos para la generación de resúmenes lingüísticos que combinan diferentes técnicas de *soft computing*. Luego en el capítulo 3 se proponen nuevos indicadores que permiten la evaluación de resúmenes lingüísticos y complementan los indicadores existentes. En el capítulo 4 se comparan los algoritmos propuestos con otras investigaciones reportadas en la bibliografía. Finalmente se presentan las conclusiones y recomendaciones de la investigación.

# Capítulo 1: Fundamentación teórica

En este capítulo se construye el marco teórico referencial que sustenta la investigación, donde se realiza un análisis de la evolución y las tendencias de la sumarización lingüística de datos. El capítulo se encuentra organizado de la siguiente forma: un primer epígrafe donde se exponen los conceptos generales y principales formas de representación de los resúmenes lingüísticos. Luego, se analizan las principales técnicas reportadas en la bibliografía asociadas a la generación de resúmenes lingüísticos de datos. En el tercer epígrafe se realiza un análisis de los métodos de validación empleados en las investigaciones de la bibliografía estudiada. En un cuarto epígrafe se revisan las principales áreas o escenarios de aplicación de los resúmenes lingüísticos para la toma de decisiones. Finalmente se presentan las conclusiones del capítulo.

Para la elaboración del marco teórico referencial se aplican buenas prácticas de investigación (Kelley y Kelley 2019) (Böger et al. 2021) (Grau, Correa y Rojas 1999) siguiendo los pasos:

1. Diseño de un protocolo de revisión sistemática que permite construir un marco teórico referencial asociado a la temática, ver el Anexo 2.
2. Análisis a profundidad de la bibliografía consultada por cada una de las categorías establecidas y caracterización respecto a los siguientes elementos:

* Estructura de los resúmenes lingüísticos y técnicas para su generación.
* Técnicas y métodos de validación empleados en las investigaciones.
* Principales áreas de aplicación.

1. Aplicación de técnicas estadísticas para el análisis de los resultados de la revisión.
2. Identificar líneas abiertas de investigación.

En la revisión sistemática se analiza la evolución histórica de la sumarización lingüística en los siguientes períodos: publicaciones antes del 2000, publicaciones entre el 2000-2015 y publicaciones a partir del 2015.

## Conceptos generales sobre sumarización lingüística de datos

Zadrożny identificó que la esencia de esta técnica es que un conjunto de datos se puede resumir lingüísticamente con respecto a un atributo o atributos seleccionados, mediante proposiciones cuantificadas lingüísticamente (Kacprzyk, J. y Zadrożny 2005).

Boran, Akay y Yager en (Boran, Akay y Yager 2016) plantean que la SLD es una poderosa técnica de descubrimiento de conocimiento descriptivo sobre conjuntos de datos, capaz de extraer conocimiento potencial, útil y abstracto de datos tanto numéricos como categóricos, cuyo objetivo es descubrir las relaciones existentes entre atributos en una base de datos.

Por otro lado, Anna Wilbik y otros autores plantean que el objetivo de los resúmenes lingüísticos es proporcionar de forma ágil la comprensión de grandes cantidades de datos, al describir en forma lingüística las principales propiedades de los datos (Wilbik y Dijkman 2016) (Eciolaza, Pereira-Fariña y Trivino 2013).

A partir del estudio realizado se identifica que autores europeos (Zadeh 1983) (Kacprzyk y Strykowski 1999) y norteamericanos (Yager 1982), fueron pioneros en el desarrollo de la teoría en la década del 80 y del 90; mientras que en la actualidad, se observa una mayor dispersión de los trabajos respecto a áreas geográficas, ver la Figura 27 en el Anexo 3. Los siguientes elementos caracterizan la bibliografía consultada:

* La mayoría de las investigaciones consultadas son artículos en revistas, existiendo pocas tesis de doctorado o libros sobre la temática, ver Anexo 3 Figura 28, demostrando que aún está en proceso de madurez.
* La mayoría de las publicaciones consultadas están indexadas en *Scopus* y *Web Of Science* (*WOS*), ver Anexo 3 Figura 28.
* Las fuentes de información fundamentales para las investigaciones consultadas fueron *IEEE*, *Elsevier* y *Springer*, ver Anexo 3 Figura 29.

En la bibliografía consultada se identifica un conjunto de elementos que caracterizan o forman parte de la notación establecida en esta área temática (Yager 1982) (Donis-Díaz, Bello y Kacprzyk 2015) (Wu y Mendel 2010):

* *D*: Base de datos, por ejemplo sobre “trabajadores” o “proyectos”.
* *Y* = *{y1, ..., yn},* conjunto de objetos (registros) en la base de datos.
* *A* *= {A1, ..., Am},* conjunto de atributos que describen a los objetos *Y*, ejemplo “salario”.
* *Aj(yi)*: denota el valor del atributo *Aj* para el objeto *yi*, por ejemplo *“*joven” para el atributo “edad”.
* *Q*: cuantificador, es un conjunto borroso con el universo de discurso en el intervalo [0, 1] expresando una cantidad, por ejemplo, “la mayoría”, “el 60%” o “más de la mitad”.
* *R*: calificador o filtro, es otro atributo que determina un subconjunto borroso del objeto *yi*, por ejemplo “joven” para el atributo “edad”.
* *S*: sumarizador, es un atributo con un valor lingüístico (predicado borroso) definido en el dominio del atributo *Aj*, por ejemplo “bajo salario” para el atributo “salario”.
* *T*: grado de verdad (validez) del resumen, es un número del intervalo [0, 1] que evalúa el grado de verdad del resumen; los resúmenes con un alto valor *T* son de interés.

Las sintaxis más empleadas para la construcción de los resúmenes son las siguientes (Donis-Díaz, Bello y Kacprzyk 2015) (Kacprzyk y Zadrożny 2016a) (Wilbik, Kaymak y Dijkman 2017):

* *Qy’s are S*: resúmenes lingüísticos sin filtro, en los que la declaración tiene la forma “*Q* objetos son *S”*. Por ejemplo: “La mayoría de los trabajadores son puntuales”.
* *QRy’s are S*: resúmenes lingüísticos con filtro, en los que la declaración tiene la forma “*Q* objetos *R* son *S”*, donde *R* puede estar formado por uno o varios atributos que califican al objeto. Ejemplo: “La mayoría de los trabajadores jóvenes son impuntuales”.

Las estructuras de los resúmenes lingüísticos se han conceptualizado en protoformas, la siguiente subsección resume las tendencias en este sentido.

**Evolución y las tendencias en las protoformas para la construcción de resúmenes**

Las protoformas que más se han empleado en la bibliografía para la construcción de los resúmenes lingüísticos siguen la estructura propuesta por Zadeh en (Zadeh 2002) y extendidas posteriormente por J. Kacprzyk y S. Zadrożny en (Kacprzyk, Janusz y Zadrożny 2005). Estas protoformas las denominamos protoformas clásicas, ver Tabla 3.

Tabla 3: Clasificación de resúmenes lingüísticos (Kacprzyk y Zadrożny 2009).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tipo** | **Protoforma** | **Conocido** | **Duda** | **Comentarios** |
| 0 | *QRy’s* are *S* | Todo | *T* | Resúmenes condicionales a través de consultas *ad-hoc* |
| 1 | *Qy’s* are *S* | *S* | *Q* | Resúmenes simples a través de consultas *ad-hoc* |
| 2 | *QRy’s* are *S* | *S R* | *Q* | Resúmenes condicionales a través de consultas *ad-hoc* |
| 3 | *Qy’s* are *S* | *Q* *SEstructura* | *SValor* | Resúmenes sencillos orientados a valores |
| 4 | *QRy’s* are *S* | *Q SEstructura*, *R* | *SValor* | Resúmenes orientados a valores condicionales |
| 5 | *QRy’s* are *S* | Nada | *Q R S* | Reglas difusas generales |

En estas protoformas se establecen jerarquías donde se transita por los niveles de mayor a menor abstracción y se caracterizan por los siguientes elementos (Zadeh 1983) (Kacprzyk, Janusz y Zadrożny 2005):

* El término *SEstructura* significa que en el resumen se conocen las variables que lo conforman; mientras que *SValor* denota que se desconoce el valor del sumarizador.
* Protoforma 0: responde a la estructura *QRy’s are S*. Es la de menor nivel de abstracción, pues se asume que todos los elementos que conforman el resumen son conocidos. Lo que se desea conocer es su grado de verdad *T*.
* Protoforma 1: estructura *Qy’s are S*, se conoce el sumarizador (*S*) y se quiere descubrir el cuantificador (*Q*). Las protoformas 1 y 2 pueden generar resúmenes a partir de sentencias SQL o de sus disímiles extensiones (Kacprzyk y Zadrożny 2000).
* Protoforma 2: estructura *QRy’s are S*, se conocen el sumarizador (*S*) y el filtro(*R*),se deseadescubrir el cuantificador (*Q*)*.* Al igual queen las protoformas 0 y 1, estos resúmenes pueden ser obtenidos a través de consultas a la base de datos.
* Protoforma 3: estructura *Qy’s are S,* se conocen el cuantificador (*Q*) y la estructura del resumen*,* se quiere descubrir el sumarizador (*S*).
* Protoforma 4: estructura *QRy’s are S*, se conocen el cuantificador (*Q*) y la estructura del resumen*,* se quieren descubrir el sumarizador (*S*) y el filtro (*R*).
* Protoforma 5: estructura *QRy’s are S*. Es el de mayor nivel de abstracción, pues no se conoce ningún elemento y por tanto se trata de descubrir todo. Estos son los resúmenes más complejos y son el objetivo principal de este trabajo.

Las protoformas clásicas y las basadas en series de tiempo son las más empleadas en la construcción de los resúmenes lingüísticos (Hudec, Bednárová y Holzinger 2018). En esta investigación los resúmenes son generados a partir de las protoformas clásicas.

Las protoformas clásicas sumarizan los atributos del conjunto de datos o las relaciones entre ellos (Yager 1982) (Kacprzyk, Janusz y Zadrożny 2005). Estas protoformas han sido usadas en disímiles escenarios de aplicación como se muestra en la sección 1.4.

En la bibliografía consultada se encuentran investigaciones que particularizan las protoformas para problemas específicos (Hudec, Bednárová y Holzinger 2018). Por ejemplo, en algunas investigaciones relacionadas con series de tiempo se introducen protoformas con la estructura: *Q Bs are A QT time*, donde *QT* es un cuantificador aplicado al atributo tiempo. Un ejemplo de resumen basado en estas protoformas es: “cerca de la mitad de los negocios pequeños tienen tiempo de respuesta pequeño la mayor parte del tiempo”.

Existen algunas extensiones asociadas a secuencia de eventos o actividades, como el caso de (Wilbik y Dijkman 2016) donde podemos encontrar resúmenes como “Cuando el tiempo de procesamiento era largo, el caso contenía una secuencia como `*abcdefgh*`”. Otro ejemplo de los mismos autores es el análisis de logs asociados a registros de procesos de apelación de la municipalidad de Dutch, donde se obtienen resúmenes tales como “Muchos casos tienen tiempo de operación corto” (Wilbik, Kaymak y Dijkman 2017).

Otras extensiones a las protoformas, se presentan en escenarios de aplicación donde se quiera incorporar épocas o fenómenos temporales, donde la bibliografía consultada emplea protoformas que no usan cuantificadores lingüísticos. Este tipo de resumen tiene la estructura *P, datos son A*, donde *P* es un término temporal y *A* es un término lingüístico borroso. Por ejemplo, *“*regularmente en otoño, las precipitaciones son altas”.

También podemos encontrar modelos híbridos como por ejemplo en (Kacprzyk y Zadrożny 2016b), donde se introducen protoformas temporales para el análisis de logs en servidores web. En este caso se generan resúmenes tales como “Recientemente, entre todos los segmentos, la mayoría están aumentando lentamente”; y protoformas temporales extendidas generando resúmenes como “Inicialmente, entre todos los segmentos cortos, la mayoría están aumentando lentamente”.

Existen otras variantes de protoformas relacionadas con datos sobre el clima (Ramos-Soto y Martin-Rodillab 2019), donde los autores combinan protoformas con los términos, “pero” y “especialmente” para establecer sentencias de contraste y de énfasis respectivamente. Este enfoque permite obtener resúmenes lingüísticos con mayor expresividad y constituye una línea abierta a la investigación. A continuación, se muestran diferentes ejemplos:

Ejemplos de protoforma de contraste:

“Muchos valores en Ancares son normales”, **pero**

“La mayoría de los valores de la tercera semana en Ancares son calientes”

“La mayoría de los valores de la cuarta semana en Ancares son fríos”

Ejemplo de protoforma de énfasis:

“Muchos valores en Pontevedra-Campolongo son calientes”, **especialmente**

“Muchos valores de la segunda semana en Pontevedra-Campolongo son muy calientes”

En las investigaciones consultadas, generalmente los autores construyen los resúmenes para lenguajes específicos. Se identifica en este sentido como una línea abierta a la investigación la posibilidad de representar un mismo resumen en diferentes idiomas, facilitando la generación multilingüe de los resúmenes.

## Métodos o técnicas para la generación de los resúmenes lingüísticos de datos

En la bibliografía consultada se identifican diferentes métodos o técnicas para la construcción de resúmenes lingüísticos, entre las que se destacan (Pérez Pupo et al. 2021) (Piñero et al. 2020):

* Construcción de resúmenes a partir de técnicas de estadística descriptiva (Boran, Akay y Yager 2016), (Khedidja, Allel y Mohand 2020).
* Generación de resúmenes lingüísticos a partir de lenguaje de consultas, inicialmente desarrollados por J. Kacprzyk y colaboradores (Kacprzyk, Janusz 1999).
* Generación de resúmenes lingüísticos a partir de reglas de asociación (Kacprzyk y Zadrożny 2003) y de reglas de producción (Dubois y Prade 1992).
* Generación de resúmenes lingüísticos a partir de metaheurísticas, entre los primeros trabajos está el de George y colaboradores (George et al. 1996).
* Generación de resúmenes a partir de técnicas de agrupamiento (Wilbik y Dijkman 2016) y combinaciones con problemas de generación de resúmenes a partir de datos anómalos (*outliers*) (Pérez Pupo, Piñero Pérez, Vacacela, et al. 2020).
* Generación de resúmenes a partir conjuntos aproximados (Pérez Pupo, Piñero Pérez, Bello, et al. 2020).

En el Anexo 3, la Figura 30 muestra un resumen de cuál ha sido la tendencia de los métodos de generación de resúmenes lingüísticos de datos.

### Enfoque estadístico y enfoque basado en consultas

Las técnicas de estadística descriptiva generalmente son empleadas para la construcción de resúmenes basados en las protoformas 1 y 2 (Boran, Akay y Yager 2016). Estas técnicas se caracterizan generalmente por generar resúmenes cuantitativos y con una limitada capacidad de descubrimiento de nuevo conocimiento (Khedidja, Allel y Mohand 2020).

Otro ejemplo del enfoque basado en estadística descriptiva es (Kaczmarek-Majer et al. 2019), donde se presentan resultados preliminares sobre resúmenes lingüísticos aplicados a la duración media diaria de las llamadas salientes en teléfonos inteligentes. En este trabajo los algoritmos fueron aplicados a pacientes con trastorno bipolar. El estudio tuvo como objetivo el desarrollo de una herramienta de diagnóstico basada en el uso de teléfonos inteligentes para la detección y predicción de los primeros síntomas de un cambio de fase. Los resúmenes lingüísticos obtenidos en esta investigación son personalizados, basados en la protoforma “Entre todos los objetos que pertenecen al paciente *A, Q* son *P”*, donde *Q* es el cuantificador y *P* es el sumarizador. Un elemento positivo de esta investigación fue el uso de resúmenes lingüísticos en lugar del procesamiento directo con los registros históricos del paciente en función de la ética médica.

En la generación de resúmenes lingüísticos a partir de lenguaje de consultas se destaca el paquete FQUERY desarrollado por Kacprzyk y Zadrożny en (Kacprzyk y Zadrożny 2000) y (Kacprzyk, J. 1999). Estos autores utilizan consultas difusas que extienden al SQL para la construcción de resúmenes con las protoformas descritas en la Tabla 3, es un paquete integrado al gestor Microsoft Access (Kacprzyk, Zadrożny y Dziedzic 2014) (Kacprzyk y Zadrożny 2016b). Otros ejemplos son propuestos en (Kacprzyk y Zadrożny 1995) y (Rasmussen y Yager 1999) donde se generan resúmenes lingüísticos a partir de un lenguaje de consulta llamado *SummarySQL*. Este enfoque tiene como ventaja la flexibilidad y versatilidad de los lenguajes de consultas como los basados en SQL.

Tanto el enfoque estadístico como el basado en consultas se basan simplemente en una observación de los objetos, identificando relaciones explícitas, pero no implícitas, por lo que pueden presentar dificultades para identificar situaciones anómalas.

### Enfoque basado en reglas

El enfoque de generación de resúmenes lingüísticos de datos a partir de reglas tiene su base en técnicas de construcción de reglas de producción, como los árboles de decisión y otros. En este sentido se destacan autores como Dubois y Prade (Dubois y Prade 1992), Anna Wilbik y colaboradores (Wilbik, Kaymak y Dijkman 2017), Kacprzyk y Zadrożny (Kacprzyk y Zadrożny 2003), Wu, Mendel y Joo (Wu, Mendel y Joo 2010).

En (Dubois y Prade 1992) se generan reglas de inferencia gradual de la forma "Cuanto más *X* es *F*, más *Y* es *G*", que expresan un cambio progresivo del grado en que la entidad *Y* satisface la propiedad gradual *G* cuando se modifica el grado en que la entidad *X* satisface la propiedad gradual *F*. Luego las reglas son transformadas empleando conjuntos borrosos hasta convertirlas en resúmenes lingüísticos. Esta variante resulta interesante porque facilita la descripción de procesos dinámicos, cambiantes en el tiempo.

Muchos autores proponen la generación de resúmenes lingüísticos a partir de reglas de asociación y reglas de producción (Kacprzyk y Zadrożny 2003), (Wu, Mendel y Joo 2010), (Wilbik, Kaymak y Dijkman 2017). En estos trabajos se logra encontrar resúmenes en las protoformas más complejas, pero tienen las siguientes limitantes:

* Con frecuencia se basan en la construcción de una base de reglas candidatas que luego son transformadas en resúmenes lingüísticos, generando un número elevado de resúmenes que son eliminados en una etapa posterior.
* Generalmente no explotan suficientemente la información asociada a la correlación de los datos en el problema en cuestión.

Ejemplos de trabajos donde se evidencian estas dificultades se relacionan a continuación:

* En (Wilbik, Kaymak y Dijkman 2017), los autores generan resúmenes lingüísticos inspirados en el algoritmo *Apriori* *(**T**a**hyudin, Haviluddin y Nambo 2019)*. Un elemento interesante en esta estrategia es cómo los autores especificanlos posibles cuantificadores y etiquetas lingüísticas en formas de frases simples, lo que facilita la interpretabilidad de los resúmenes obtenidos. En la presente investigación se extiende la propuesta de estos autores y se propone un nuevo algoritmo (Pérez Pupo, Santos, et al. 2018) que combina algoritmos de aprendizaje de reglas de asociación con otros enfoques que consideran el comportamiento estadístico de los datos.
* En (Smits et al. 2018) los autores aplican el algoritmo *Apriori* para generar reglas calculando su cardinalidad, y posteriormente identifican el cuantificador que mejor las describe para conformar los resúmenes. Para ello calculan la cardinalidad relativa para los atributos numérico y la moda para los valores categóricos.

### Enfoque basado en metaheurísticas

Otro enfoque en la construcción de resúmenes lingüísticos es el basado en metaheurísticas, donde se destacan autores como George y colaboradores (George et al. 1996), seguidos por J. Kacprzyk y P. Strykowski (Kacprzyk y Strykowski 1999) y posteriormente Donis-Díaz, Bello, et al. (Donis-Díaz, Bello y Kacprzyk 2015) (Donis-Díaz, Bello, et al. 2014). Bajo este enfoque, los algoritmos genéticos han sido los más utilizados.

En (Donis-Díaz, Muro, et al. 2014) se propone un modelo híbrido de Algoritmo Genético (AG) con un escalador de colinas para la identificación de resúmenes lingüísticos. Estos mismos autores experimentan además con diferentes metaheurísticas de inteligencia colectiva como las colonias de hormigas (ACO por sus siglas en inglés) (Donis-Díaz, Bello y Kacprzyk 2015). En la representación de los individuos o soluciones ambos algoritmos conciben que cada individuo está formado por resúmenes lingüísticos. En particular, durante la identificación de los cuantificadores en (Donis-Díaz, Bello y Kacprzyk 2015), se propone una búsqueda exhaustiva hasta encontrar el cuantificador que logre mayor representatividad de una combinación específica *(filtros-sumarizador)* en el conjunto de datos. Además, ambas propuestas incluyen métodos adicionales basados en técnicas de búsqueda local para la mejora de las soluciones que incorporan mayor complejidad en el proceso de búsqueda.

Como debilidad de este enfoque, se le señala que estas técnicas no aprovechan la información asociada a las correlaciones entre los datos. Además, en ambos trabajos los autores establecen un algoritmo de generación de resúmenes con un alto costo computacional que se aprecia a partir de la cantidad de evaluaciones de resúmenes lingüísticos que debe realizar este algoritmo.

### Enfoque basado en agrupamientos y descubrimiento de datos anómalos

En el enfoque basado en el agrupamiento de los datos se destacan trabajos como (Wilbik y Dijkman 2016), donde construyen matrices de partición y emplean los algoritmos *Single Linkage (SL)* y *Non-Euclidean Relational Fuzzy CMeans (NERFCM).* Otros trabajos como (Dijkman y Wilbik 2017), se basan en la construcción de un árbol a partir de considerar todas las combinaciones de filtros y sumarizadores, luego podan las ramas que representa una combinación con bajo soporte en la base de datos. La naturaleza de estos trabajos, implica un alto costo computacional durante el proceso de búsqueda.

Los métodos basados en este enfoque generalmente construyen los *clusters* y luego por cada grupo generan un resumen lingüístico que explica el comportamiento del mismo. Entre las principales limitantes de este enfoque se identifica que generalmente dependen de parámetros que acotan la cantidad de grupos a construir. En otros casos las estrategias basadas en distancias tienden a formar híper-esferas.

Por otra parte, hay pocos trabajos que vinculan la generación de resúmenes con la minería de datos anómalos (Pérez Pupo, Piñero Pérez, Vacacela, et al. 2020), (Castro Aguilar 2017). Ejemplo de este enfoque es presentado por Duraj y colaboradores en (Duraj, Szczepaniak y Chomatek 2020) y (Duraj, Szczepaniak y Ochelska-Mierzejewska 2016), donde construyen de forma exhaustiva todas las posibles combinaciones de cuantificadores y sumarizadores predefinidos. Como limitante de esta propuesta, se identifica que no usan heurísticas u otros métodos que simplifiquen la complejidad en la búsqueda.

### Síntesis del análisis de los métodos para la generación de resúmenes lingüísticos

Como resumen de este análisis se identifica que antes del 2015 los métodos más utilizados son los basados en consultas difusas, existiendo numerosas publicaciones que hacen referencia al paquete FQUERY (Kacprzyk y Zadrożny 1995); una de las ventajas de este enfoque son las facilidades que aporta el uso de lenguajes de consultas.

Esta situación se comporta diferente en la actualidad, pues en el período 2015-2020 hay mayor diversidad en el uso de diferentes técnicas, sin que se identifiquen diferencias significativas en la aplicación de éstas. Este comportamiento, muestra la inclusión de nuevos métodos basados en otras teorías, aportando resultados prometedores con potencialidades para el uso de los resúmenes en otras problemáticas asociadas a problemas de predicción, entre otros. Un resumen de los principales métodos de generación de resúmenes lingüísticos se puede ver en el Anexo 3 Figura 30.

A partir del análisis de las diferentes técnicas se identifican las siguientes oportunidades de investigación:

* Continuidad en el trabajo con las metaheurísticas, a partir de un mejor uso de la información asociada al problema en cuestión y técnicas de evaluación parcial que disminuyan la complejidad de la búsqueda.
* El uso de conjuntos aproximados y otras técnicas de *soft computing*, que impliquen un mejor aprovechamiento de información asociada a la dependencia entre variables.
* El uso de los resúmenes lingüísticos en el descubrimiento y análisis de datos anómalos, con disímiles aplicaciones. En esta área se reportan pocas publicaciones.
* El análisis de nuevas protoformas y mejoras en los métodos de visualización de los resúmenes, potenciando la combinación con otras técnicas de visualización científica.

## Principales técnicas y métodos de validación empleados en las investigaciones

En la bibliografía estudiada, también se analizaron cuáles fueron las principales estrategias de validación empleadas en las investigaciones sobre sumarización lingüística de datos (Pérez Pupo et al. 2021).

A partir de este análisis se identifica que antes del 2015, la mayoría de las investigaciones solo se validaban a partir de la observación y el análisis cualitativo simple en casos de estudio, ver Anexo 3 Figura 31. Mientras que, en los últimos cinco años, hay investigaciones que usan técnicas básicas de estadística descriptiva, mediante el uso de los indicadores de calidad de los resúmenes en la mayoría de los casos (Kaczmarek-Majer et al. 2019) (Peláez-Aguilera et al. 2019) (Amghar y Chikh 2018). En esta temática, muy pocos estudios aplican técnicas estadísticas de rigor basadas en test paramétricos y no paramétricos, ejemplo de excepciones son (Donis-Díaz, Muro, et al. 2014) y (Donis-Díaz, Bello y Kacprzyk 2015).

Respecto a la validez externa, se realiza un análisis observando cómo cada una de las investigaciones consultadas, hacen uso de alguna de las siguientes estrategias de validación: triangulación de datos, triangulación de métodos, triangulación de expertos y estudio de casos, ver Anexo 3 Figura 32.

En particular, respecto a la triangulación de datos, solo en los últimos años se encuentran investigaciones que lo aplican, se muestran a continuación algunas de estas investigaciones:

* En (Ramos-Soto y Martin-Rodillab 2019) utilizan datos de 15 estaciones meteorológicas de Galicia.
* En (Amghar y Chikh 2018) se aplica la investigación a las bases de datos *Pima Indian Diabetes Dataset* y Wisconsin Breast Cancer Dataset (WBCD), ambas del UCI Repository of Machine Learning (Dua y Graff 2021).
* Kacprzyk y Zadrożny (Kacprzyk y Zadrożny 2016a) realizan análisis de cotizaciones de fondos de inversión en bolsas de valores de Varsovia, Zagreb y Moscú.
* En (Khedidja, Allel y Mohand 2020) se aplica a datos sobre vuelos reales y datos sobre la ciudad inteligente del proyecto *neOCampus*, de la universidad *Toulouse* en Francia (*Université de Toulouse*).

Pocos autores han empleado técnicas de triangulación de métodos en sus investigaciones, comparando sus propuestas con otros métodos reportados en la bibliografía; tal es el caso de los trabajos publicados en (Donis-Díaz, Muro, et al. 2014) y (Khedidja, Allel y Mohand 2020).

También se identifican pocas publicaciones donde se aplique triangulación de expertos. En este caso se podría citar (Rojas Valenzuela 2018), donde el autor aplica encuestas para la validación de sus resultados.

En resumen, en la bibliografía consultada, se identifica que las técnicas de validación que han predominado en las investigaciones son las basadas en el estudio de casos. Este elemento permite identificar la necesidad de aumentar el uso de métodos cuantitativos en el diseño de experimentos, que permitan elevar la validez interna y externa de las investigaciones.

## Áreas de aplicación de los resúmenes lingüísticos

La sumarización lingüística de datos se ha aplicado en numerosos dominios y en diferentes tipos de aplicación (Anexo 3 Figura 33), ya sea aplicación descriptiva, como en problemas de predicción o clasificación.

A partir del análisis realizado se identifica que el área de mayor cantidad de aplicaciones en el período 2000 al 2015, fue la asociada a los sectores financieros y económicos. Sin embargo, la tendencia en los últimos cinco años se concentra en la medicina, dispositivos inteligentes y redes de sensores.

Respecto a los tipos de aplicación, en la mayoría de la bibliografía consultada, el uso de los resúmenes lingüísticos se concentra en la descripción de fenómenos y la toma de decisiones en áreas como:

* Economía y ventas: en (Kacprzyk y Zadrożny 2016a) Kacprzyk y Zadrożny resumieron lingüísticamente, series de tiempo de registros y servidores web sobre cotizaciones de fondos de inversión en bolsas de valores de Varsovia, Zagreb y Moscú, con el fin de identificar desajustes económicos en Europa en el período 2002-2010, considerando tres aspectos de las tendencias: duración, dinámica y variabilidad de los fondos de inversión. Otro ejemplo es (Piñero et al. 2020) donde se aplica en el análisis económico-financiero de organizaciones orientadas a proyectos.
* Registros de logs: en (Kacprzyk y Zadrożny 2016a) Kacprzyk y Zadrożny los utilizan para la descripción de registros de logs de servidores web. En (Wilbik y Dijkman 2016) Wilbik y Dijkman lo aplican a eventos de logs que contienen información sobre cómo algunas actividades son realizadas por casos particulares de usuarios. Un año más tarde utilizan logs sobre procesos de apelación de la municipalidad de Dutch en el caso de estudio de su validación (Dijkman y Wilbik 2017).
* Redes de dispositivos y sensores: gestión de datos de sensores en (Khedidja, Allel y Mohand 2020); en (Díaz-Hermida y Vidal 2018) se aplica a restricciones temporales en una red; datos coleccionados de relojes inteligentes que miden índices del corazón en sesiones de rehabilitación (Peláez-Aguilera et al. 2019). Otras mediciones de relojes inteligentes sobre manía-euthymia y manía-depresión (Kaczmarek-Majer et al. 2019). En la red de comercio internacional (Genç et al. 2020). En (Jain, Keller y Bezdek 2016) se recopilan datos de sensores de hogares inteligentes de personas mayores sobre sus actividades de la vida diaria.
* En el sector de la salud y la medicina se han realizado varios estudios: con datos sobre gestión de glucosa en la Unidad de Cuidados Intensivos (UCI) del Centro Médico Universitario de Maastricht, Países Bajos (Wilbik et al. 2018); sobre afecciones de corazón isquémico en pacientes del programa de rehabilitación del corazón en Andalucía (Peláez-Aguilera et al. 2019); sobre actividad física (Sanchez-Valdes, Alvarez-Alvarez y Trivino 2016); sobre diabetes y cáncer (Amghar y Chikh 2018). En (Wilbik et al. 2018) proponen utilizar los resúmenes lingüísticos para analizar el cumplimiento de los protocolos clínicos, en este estudio los autores demuestran que los resúmenes lingüísticos pueden proporcionar información comprensible para el personal médico. Otra investigación realizada en este sentido es (Peláez-Aguilera et al. 2019), donde los resúmenes son aplicados a datos proporcionados por pacientes de un programa de rehabilitación cardíaca; en este caso presentan una metodología para evaluar los flujos de frecuencia cardíaca de pacientes con cardiopatía isquémica utilizando un enfoque lingüístico, donde las variables utilizadas para la conformación de los resúmenes constituyen métricas de interés identificados por el equipo de rehabilitación cardíaca. En (Kaczmarek-Majer et al. 2019) se aplica al análisis de llamadas telefónicas de pacientes para identificar cambios en el comportamiento de pacientes con trastorno bipolar.
* En la meteorología: en (Khedidja, Allel y Mohand 2020) es aplicado a colecciones de temperaturas de una ciudad inteligente; en (Ramos-Soto y Martin-Rodillab 2019) es aplicado a datos sobre observaciones del tiempo (temperatura) de 15 estaciones meteorológicas de Galicia; en (Heble-Lahera et al. 2020) son datos de observaciones del tiempo (estado del cielo, viento y temperatura) de la Agencia Meteorológica de Galicia MeteoGalicia.
* En la agricultura como (Rojas Valenzuela 2018), donde se plantea desarrollar un algoritmo generador de resúmenes lingüísticos para desplegar información relevante de cultivos de paltos, proporcionada por los sensores de humedad instalados en la zona cultivada.
* En la industria: en (Donis-Díaz, Muro, et al. 2014) Donis-Díaz, Bello y Morales lo aplican a datos de tensión de rotura por fluencia. Esta es una de las propiedades mecánicas más importantes consideradas en el diseño de nuevos aceros utilizados en las industrias aeronáutica, energética y petroquímica. Mide el nivel de estrés en el que falla la estructura de acero cuando se expone a condiciones bastante agresivas (como altas temperaturas del vapor) durante períodos de tiempo de hasta 30 años. Por otra parte, en (Hudec, Bednárová y Holzinger 2018) se generan resúmenes a datos sobre contaminación por 30 días en dos distritos.
* En gestión de patentes: en (Igde et al. 2017) los autores generan resúmenes lingüísticos a datos relacionados con registros de patentes europeas, incluidos los códigos de Clasificación Internacional de Patentes (*IPC: International Patent Classification*) del sector de Tecnología de la Información y las Comunicaciones (*TIC: Information and Communications Technology*), publicados por la OCDE[[5]](#footnote-6) y estructurados a partir de la base de datos en línea PATSTAT[[6]](#footnote-7).
* Gestión del consenso o análisis de grupos: se usan para brindar a los moderadores informaciones relevantes acerca de los participantes y de los factores que influyen en el acuerdo o desacuerdo del grupo (Kacprzyk y Zadrożny 2018), (Kacprzyk y Zadrożny 2016c). Otro trabajo también dedicado a intensiones de personas se presenta en (Kacprzyk, Yager y Merigo 2019).
* Temas legales: Wilbik y colaboradores generan resúmenes relacionados con procesos de apelación de la municipalidad de Dutch (Wilbik, Kaymak y Dijkman 2017), (Dijkman y Wilbik 2017).
* En gestión de tráfico y movilidad urbana: donde en (Degtiarev y Remnev 2016) y (Gilsing et al. 2020) se proponen un sistema de reportes de accidentes de carros y de movilidad urbana.

Existen ejemplos de uso de los resúmenes donde se fusionan varios entornos de aplicación, como en (Jain et al. 2019), donde se realizan análisis de datos provenientes de sensores instalados en los apartamentos de ancianos para ayudar a médicos a entender los datos provenientes de diferentes equipos, y tomar decisiones que beneficien a sus pacientes. En este caso los resúmenes lingüísticos conducen a alertas de salud derivadas de los datos del sensor en tiempo real, y ayudan a establecer una conexión entre los cambios en los datos del sensor y la salud de los pacientes ancianos.

En menor medida se han empleado resúmenes en problemas de predicción o clasificación. Un ejemplo de aplicación de los resúmenes lingüísticos para la predicción es (Chiang, Chow y Wang 2000), donde a partir de datos sobre uso de recursos (uso del CPU de una sede de IBM), se generan resúmenes para predecir los rangos de utilización de diferentes recursos en un momento específico. En (Wilbik, Barreto y Backus 2020) se identifica la necesidad de aplicar los resúmenes lingüísticos para la predicción.

Se identifica como una línea abierta a la investigación el uso, en mayor medida, de las técnicas de sumarización lingüística de datos en la solución de problemas de predicción como parte de procesos de ayuda a la toma de decisiones.

Por otra parte, el uso de los resúmenes en la solución de problemas de clasificación resulta ser un enfoque interesante, debido al aprovechamiento de la alta capacidad de interpretabilidad de los mismos. Ejemplo, en (Amghar y Chikh 2018) se aplica esta técnica para resolver problemas de diagnóstico médico clasificando al paciente como enfermo o no enfermo, y se comparan los resúmenes con los métodos KNN y SVM. En este caso las bases de datos médicas están relacionadas a diabetes (*Pima Indian Diabetes Dataset*) y cáncer (*Wisconsin Breast Cancer Dataset WBCD*). En (Chiang, Chow y Wang 2000), además de la predicción también se realiza clasificación porque según los resultados, los usuarios pueden clasificar los objetos en diferentes grupos.

En la Figura 33 del Anexo 3 se evidencia que la sumarización lingüística de datos es ampliamente utilizada en el sector financiero y económico, sobre todo en el primer quindenio[[7]](#footnote-8) del presente siglo; sin embargo, tienen aplicación creciente en las redes de sensores, en el análisis de logs y en el sector de la salud.

Es importante resaltar que, en las investigaciones reportadas en la bibliografía, no se utilizan datos provenientes de una colección estándar que facilite la comparación de métodos y la validación. Generalmente se emplean datos provenientes de cada escenario en cuestión. Además, con frecuencia se diseñan protoformas para cada escenario de aplicación basadas en idioma inglés.

A partir de la metodología aplicada como parte del diseño de experimentos y el análisis detallado de las diferentes fuentes bibliográficas, se establece en esta investigación que la sumarización lingüística de datos es: un área de conocimiento interdisciplinar, que forma parte del *soft computing*, e involucra un conjunto de áreas de conocimiento entre las que se destacan la lógica borrosa, el aprendizaje automático, la estadística, las bases de datos y otras; tiene como objetivo la ayuda a la toma de decisiones a partir de la construcción automática de resúmenes en lenguaje natural, que reflejen el conocimiento implícito y las relaciones intrínsecas que existen entre las variables en bases de datos heterogéneas.

## Conclusiones del capítulo

En el capítulo se construye un marco teórico referencial enfocado en las siguientes perspectivas: la notación y la estructura de los resúmenes, las técnicas de generación, las estrategias de validación de las investigaciones y las áreas de aplicación. Se arriba a las siguientes conclusiones:

* En la bibliografía consultada se identifican diferentes protoformas, generalmente la variabilidad en estas se debe a particularidades de los escenarios de aplicación. La mayoría de las protoformas están diseñadas para el trabajo en el idioma inglés y las más generalizadas son las propuestas por J. Kacprzyk y S. Zadrożny. Se identifica en este sentido la oportunidad de construir nuevos algoritmos para la generación de resúmenes lingüísticos bajo un enfoque multilingüe.
* Las técnicas de generación de resúmenes son diversas, se destacan antes del 2015 el uso de métodos basados en consultas. En los últimos años hay mayor variedad en los métodos de generación de resúmenes lingüísticos, se ha incrementado el uso de técnicas asociadas al aprendizaje automático y la minería de datos. Estas nuevas tendencias incorporan a los algoritmos mayor robustez; pero se identifica que quedan líneas abiertas a la investigación a partir de la posibilidad de hibridación de diferentes técnicas, y una mejor explotación de la información asociada a los datos, tales como la correlación entre las variables o aspectos específicos de los dominios de aplicación.
* Se identifican oportunidades de mejoras en los algoritmos reportados en la bibliografía respeto a la eficacia, pues algunos como los basados en reglas y en metaheurísticas, no explotan la información estadística asociada a la correlación entre las variables durante el proceso de búsqueda, y en ocasiones generan resúmenes candidatos que no tienen una representación real en los datos.
* Las investigaciones reportadas en la bibliografía que emplean metaheurísticas, tienen alto costo computacional, elemento que influye negativamente en la aplicabilidad de estos algoritmos en entornos reales de toma de decisiones.
* Respecto a la validación de las investigaciones publicadas, se identifica que hay pocos trabajos que aplican las técnicas de triangulación de datos, triangulación de métodos u otras que ayuden a verificar la capacidad de generalización de las investigaciones. Esto constituye una oportunidad de mejora para el desarrollo de investigaciones que comparen con profundidad, los diferentes métodos de generación de resúmenes, y permitan identificar las potencialidades de los mismos para diferentes escenarios.
* Los resúmenes lingüísticos de datos han sido empleados generalmente de forma descriptiva, siendo las áreas de aplicación más explotadas la toma de decisiones en la gestión económica y las ventas desde etapas tempranas del desarrollo de la teoría. No obstante, a partir del 2016, hay un incremento significativo de aplicaciones en la medicina, en el análisis de datos de dispositivos inteligentes y en el análisis de redes de sensores como parte de la internet de las cosas. Se identifica como una línea abierta a la investigación el uso, en mayor medida, de las técnicas de sumarización lingüística de datos en combinación con otras técnicas en escenarios de toma de decisiones.

# Capítulo 2: Algoritmos para la sumarización lingÜística de datos

En este capítulo se proponen nuevos algoritmos para la generación de resúmenes lingüísticos a partir de datos que combinan diferentes técnicas de computación emergente. En las secciones 2.3, 2.4 y 2.5 se proponen los algoritmos *PCA\_LDS, LPA\_LDS y RST\_LDS* respectivamente,publicados en*(**P**é**rez Pupo, Piñero Pérez, Bello Pérez, et al. 2020)* (Pérez Pupo et al. 2019). En la sección 2.6 se discuten diferentes enfoques que combinan la sumarización lingüística de datos con la minería de datos anómalos, donde se propone el algoritmo *LDS\_outliers* *(**P**é**rez Pupo, Piñero Pérez, Vacacela, et al. 2020)*. Los algoritmos propuestos generan resúmenes que pueden ser expresados en múltiples idiomas haciendo uso de los lenguajes naturales controlados.

## Notaciones y definiciones fundamentales empleadas en el capítulo.

Notación común de los algoritmos propuestos:

*U:* conjunto de datos (*dataset*) formado por *n* objetos, cada objeto representa una fila del *dataset* *U* tal que *n* = *|U|.*

*H:* conjunto de atributos que describen a los objetos en *U* tal que *p = |H|.*

*ALV*: conjunto de variables lingüísticas (ver Definición 1) que describen a los atributos *H* tal que existe una variable lingüística por cada atributo en *H*.

*m =* cantidad máxima de conjuntos borrosos de una variable lingüística.

*LVQ:* variable lingüística para los cuantificadores de los resúmenes.

*LNC*: lenguaje natural controlado (ver Definición 2) empleado, que incluye la gramática *LNCGrammar y* eldiccionario *LNCDictionary* con frases simples que describen a las variables y los atributos del problema en cuestión.

*T =*  tal que son los indicadores tradicionales de evaluación de calidad de los resúmenes lingüísticos (Zadeh 1983), y *Te* son los nuevos indicadores que extienden a los tradicionales, ambos conjuntos presentados en el capítulo 3.

* :* umbral para el cálculo de indicadores extendidos.

*tnorm y snorm*: operadores de agregación t-norma y co-norma respectivamente.

**Definición** 1**:** una variable lingüística, (Zadeh 1976), es definida por un quíntuplo *(x, T(x), X, G, M)* en el cual *x* es el nombre de la variable, *T(X)* es el conjunto de términos lingüísticos, *X* es el universo de discurso, *M* es una regla semántica que asocia a cada valor lingüístico *Z* su significado *M(Z)*, donde *M(Z)* denota un conjunto borroso en *X,* y *G* es el conjunto de reglas sintácticas de generación de términos compuestos, a partir de los términos atómicos que configuran las sentencias que dan lugar a cada valor lingüístico.

**Definición** 2**(****K****u****hn 2014)** (Kittredge 2003): Se define que un lenguaje natural controlado (LNC) es un lenguaje construido, basado en cierto lenguaje natural, más restrictivo en lo concerniente al léxico, sintaxis y/o semántica mientras se preserva la mayoría de sus propiedades naturales, de manera que los hablantes de la lengua base pueden de manera intuitiva y correcta, entender textos en el lenguaje natural controlado, al menos en un grado sustancial. Visto de forma práctica, un LNC está definido por una gramática que establece la sintaxis y un diccionario que describe el léxico y la semántica.

## Estructura de los resúmenes que se construyen siguiendo enfoque multilingüe.

En esta sección se define la estructura de las protoformas construidas por los algoritmos propuestos en la investigación, basadas en las protoformas “*Qy’s are S”* y “*QRy’s are S”*  empleadas en (Kacprzyk, Janusz y Zadrożny 2005), ver Tabla 3 en el Capítulo 1.

A continuación, se describen las variables lingüísticas *LVQ* que representan a los cuantificadores *Q* empleados en la investigación, a partir de la Definición 1:

* *T(x)* = *{“Muy pocos (menos del 5%)”, “Pocos (alrededor del 15%)”, “Algunos (alrededor del 33%)”, “Aproximadamente la mitad”, “Muchos (cerca del 65%)”, “La mayoría (alrededor del 83%)”, “Casi la totalidad”}*
* *X* es el universo de discurso definido en el intervalo [0, 1].
* *G* regla sintáctica que describe la relación entre los conjuntos borrosos, que puede ser representada por funciones campana de Gauss o funciones triangulares, ver Tabla 4.

Tabla 4: Funciones de pertenencia que representan a los cuantificadores.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Término lingüístico | Funciones de pertenencia campana Gauss (a, b, c) | | | Funciones de pertenencias triangulares (a,b,c) | | Interpretación |
| Muy pocos | | (1, 0, 0.05) | (0.0,0.0,0.17) | | menos del 5% | |
| Pocos | | (1, 0.17, 0.05) | (0.0,0.17,0.33) | | alrededor del 15% | |
| Algunos | | (1, 0.33, 0.05) | (0.17,0.33,0.5) | | alrededor del 33% | |
| Aproximadamente la mitad | | (1, 0.5, 0.05) | (0.33,0.5,0.67) | | alrededor del 50% | |
| Muchos | | (1, 0.67, 0.05) | (0.5,0.67,0.83) | | cerca del 65% | |
| La mayoría | | (1, 0.83, 0.05) | (0.67,0.83,1.0) | | alrededor del 83% | |
| Casi la totalidad | | (1, 1, 0.05) | (0.83,1.0,1.0) | | Casi todos | |

* *M(z)* es una regla semántica que asocia su significado a cada valor lingüístico *Z*.

En el caso deque se empleen funciones de pertenencia basadas en campana Gauss para representar los cuantificadores, estos se basarían en la Ecuación y quedarían representados como muestra la Figura 1.

|  |  |
| --- | --- |
|  | Ecuación 1 |

Figura 1: Variable lingüística de los cuantificadores (*Q*) basadas en funciones campana Gauss.

En el caso deque los cuantificadores estén basados en funciones de pertenencia triangulares, estarían basados en la Ecuación y la Figura 2.

|  |  |
| --- | --- |
| siendo *Q* función del extremo izquierdo  siendo *Q* función del extremo derecho  resto de las funciones internas de *Q* | Ecuación 2 |

Figura 2: Variable lingüística de los cuantificadores (*Q*) basada en funciones triangulares.

Los posibles valores de *S* y *R* también estarán representados por variables lingüísticas, donde las diferentes alternativas de representación dependen del problema en cuestión y constituyen entradas a los algoritmos. Un elemento novedoso que se propone en esta investigación es la construcción de resúmenes lingüísticos bajo un enfoque multilingüe empleando Lenguajes Naturales Controlados (*LNC*), ver Definición 2.

Para caracterizar a los lenguajes podemos emplear el esquema de clasificación de PENS (Kuhn 2014) , que se basa en las siguientes dimensiones:

* Precisión, que mide el grado en que el significado de un texto puede ser extraído de éste, tal que cada lenguaje puede ser clasificado en: impreciso (P1), menos impreciso (P2), confiables en la interpretación (P3), interpretado de forma determinista (P4) y con una semántica determinada (P5).
* Expresividad, que describe el rango de proposiciones que un lenguaje es capaz de expresar. Considerando este criterio, los lenguajes pueden ser clasificados en: inexpresivos (E1), baja expresividad (E2), expresividad media (E3), con alta expresividad (E4) y con máxima expresividad (E5).
* Naturalidad, que mide cuán cercano está el lenguaje a la lengua natural, en términos de legibilidad y comprensibilidad de los hablantes de la lengua natural de base. Bajo este criterio, los lenguajes son clasificados en: no naturales (N1), con predominio de elementos no naturales (N2), con predominio de elementos naturales (N3), con sentencias naturales (N4) y con textos naturales (N5).
* Simplicidad, relacionada con el esfuerzo necesario para implementar completamente la sintaxis y la semántica del lenguaje a través de un modelo formal. Considerando este criterio, los lenguajes pueden ser clasificados en: muy complejos (S1), sin descripciones exhaustivas (S2), con grandes descripciones (S3), con descripciones cortas (S4) y con descripciones muy cortas (S5).

Respecto a cada dimensión, los extremos conceptuales de los atributos representan a los lenguajes naturales. Por ejemplo, el idioma inglés es clasificado como P1E5N5S1 mientras que el lenguaje de la lógica proposicional es clasificado como P5E1N1S5. Los LNC definidos en esta investigación, no se sitúan en ninguno de los dos extremos, sino en caracterizaciones intermedias de cada dimensión; siendo la clasificación P3E3N4S4 la más apropiada para caracterizarlos.

El uso de los lenguajes naturales controlados tiene como objetivo lograr mayor expresividad en los resúmenes generados. Ejemplos de resúmenes generados en el LNC para el idioma español se presentan a continuación, la gramática utilizada para su construcción está en la sección 9.9.1 del Anexo 9, y un fragmento del diccionario empleado está en la sección 9.10.1:

1. *“Muchos registros en la base de datos, reportan que en el 100.0% de las veces, las embarazadas cardiópatas que tienen ausencia de soplo holosistólico tienen ausencia de cianosis.”*
2. *“La mayoría de los registros en la base de datos, reportan que en el 90.4% de las veces, los proyectos que tienen muy alta calidad del dato tienen muy baja cantidad de competencias con baja evaluación.”*

Los algoritmos de generación de resúmenes lingüísticos que se proponen en este capítulo, tienen como elemento esencial, que generan un conjunto de resúmenes candidatos *CandidateSummaries.* Cada resumencandidatoconstituyeun objeto *P(Q, R, S, T)* formado por el cuantificador *Q,* los filtros *R*, los sumarizadores *S* y un conjunto de indicadores de calidad *T*. La concepción de los resúmenes lingüísticos a partir de esta estructura de datos facilita la representación de los resúmenes bajo un enfoque multilingüe empleando diferentes *LNC.*

En esta investigación se construyen lenguajes naturales controlados para los idiomas: español, inglés, japonés y árabe en las secciones de la 9.9.1 a la 9.9.4 del Anexo 9. También se definieron los diccionarios para construir los resúmenes en español, inglés, japonés y árabe, un fragmento de cada uno de ellos se muestra en el Anexo 10. En el Algoritmo 1 se describe cómo se emplean estos *LNC* para la construcción de resúmenes lingüísticos a partir de los resúmenes candidatos.

En este algoritmo se recorre el conjunto de resúmenes candidatos, donde cada resumen tiene la estructura *P*(Q*, R, S*, *T*). Entre los pasos 4 y 6 se traducen los diferentes elementos que conforman el resumen empleando el diccionario *LNCDictionary* definido para el *LNC* en el dominio específico de aplicación.

Una vez que todos los elementos que conforman el resumen han sido traducidos al *LNC* especificado, en el paso 7 se procede a unir o concatenar las partes del resumen, para que quede finalmente construida la oración según la gramática especificada en *LNCGrammar*.

*Algoritmo* 1 *humanize\_summaries (CandidateSummaries, ALV, LVQ, LNC)*

*Entradas*

*CandidateSummaries: listado de resúmenes candidatos*

*ALV, LVQ, LNC: parámetros descritos en la sección 2.1 sobre notaciones.*

*Inicio:*

1. *summaries = []*
2. *Para cada P*(Q*, R, S*, *T*) *en CandidateSummaries*
3. *translations = []*
4. *Para cada element de P*(Q*, R, S*, *T*)
5. *element\_translation=search\_dictionary(element,* *ALV, LVQ , LNCDictionary)*
6. *translations* * element\_translation*

*Fin del paso 4*

1. *sentence = generate\_sentence(translations,* *LNCGrammar)*
2. *summaries  sentence*

*Fin del paso 2*

1. *Return summaries*

La complejidad de este algoritmo es *C(n,p,m)* = *n·(p+1)·m·(p+1) = n·m·(p2+2p+1)* dado porque, el ciclo del paso 2 se ejecuta como máximo *n* veces, el ciclo del paso 4 se ejecuta como máximo *p+1* veces, y la complejidad de la búsqueda en el diccionario del paso 5 está acotada por *O(m·(p+1)).* Por tanto, elAlgoritmo 1 tiene una complejidad de *O(n·m·p2)*.



## Algoritmo *PCA\_LDS* basado en análisis de componentes principales

Este algoritmo, genera resúmenes lingüísticos a partir de la combinación del algoritmo *PCA* (Dunteman 1989) (Lasisi y Attoh-Okine 2018) para la detección de componentes principales[[8]](#footnote-9), con algoritmos para la construcción de reglas de asociación. Con el algoritmo *PCA\_LDS* se pretende generar resúmenes lingüísticos teniendo en cuenta las relaciones o dependencias identificadas entre las variables para reducir el espacio de búsqueda (Pérez Pupo, Piñero Pérez, Bello Pérez, et al. 2020) (Pérez Pupo et al. 2019).

*Algoritmo* 2 *PCA\_LDS*

*Entradas:*

*U, ALV,LVQ, T, LNC,  : parámetros definidos en la sección 2.1 sobre notaciones.*

*percentil\_component: valor de percentil que permite seleccionar las componentes relevantes.*

*percentil\_elbow: valor de percentil empleado para seleccionar las variables relevantes en el contexto de una componente.*

*support, confidence: representan las medidas de calidad soporte y confianza respectivamente, para evaluar las reglas de asociación* *(**M**e**dina et al. 2007).*

*Inicio*

1. *Rules* = 
2. *components = do\_pca(U)*
3. *UB = transform\_fuzzy(U, ALV)*
4. ** = *calculate\_percentil(components, percentil\_component)*
5. *choosed\_components = choose\_components(components, )*
6. *Por cada i-ésima componente principal en choosed\_components hacer*:
7. *Velbow* = *calculate\_percentil(componentsi , percentil\_elbow)*
8. *ACi = get\_attributes(componentsi, Velbow)*
9. *UCi = get\_data(U, ACi)*
10. *rulesCi* = *do\_rules (UCi, support, confidence)*
11. *Rules*  *rulesCi*

*Fin del ciclo iniciado en el paso 6*

1. *Candidates=generate\_summaries\_from\_rules(U, Rules, ALV, LVQ, T)*
2. *calculate\_quantifiers(Candidates, LVQ)*
3. *calculate\_T(Candidates, )*
4. *Summaries = humanize\_summaries(Candidates, ALV, LVQ, LNC)*
5. *ranking\_summaries = sort\_summaries(Summaries)*
6. *Return ranking\_summaries*

En el paso 2 se aplica el algoritmo *PCA* *(**N**a**ik 2018)* y se generan los componentes principales, identificándose los atributos del conjunto *H* que se relacionan a partir del análisis del conjunto de datos *U*. Internamente el *PCA* calcula una matriz de covarianza de complejidad *O(p2n)* y la descompone en autovalores, con una complejidad *O(p3).* Siendo así, la complejidad de este paso es *O(p2n + p3).*

En el paso 3 se recibe un conjunto de datos (*U*) como entrada y el conjunto de variables lingüísticas *ALV* que describen cada uno de los atributos del problema en cuestión. Cada uno de los datos del conjunto de entrada *U* son transformados en etiquetas lingüísticas a partir de aplicar el principio de máxima membresía, considerando la pertenencia de cada dato a los conjuntos borrosos de la variable lingüística que representa al atributo en cuestión. De este proceso se obtiene el conjunto de datos transformado (*UB*). Este paso tiene una complejidad *O(p·n·m)*.

En el paso 4, se calcula el umbral ** a partir de los componentes calculados en *components* (variable generada en el paso 2) y el parámetro *percentil\_component* establecido en la entrada. La definición del parámetro *percentil\_component* dependerá del escenario en cuestión. Un enfoque común en diferentes contextos es la identificación del percentil 80, que representa la selección del 20% de los factores más importantes bajo el supuesto que, estos factores pueden impactar en el 80% de los problemas.

Luego en el paso 5, se emplea ** para seleccionar las componentes principales cuyo valor de fortaleza sea mayor que este umbral. La selección de las componentes relevantes permite la reducción del espacio de búsqueda. Continuando el ejemplo anterior, este umbral podría ser el valor correspondiente al percentil 80 de las componentes, lo que representaría la selección del 20% de las componentes más relevantes. En este proceso de selección, se necesita realizar un ordenamiento cuya complejidad es *O(p* log *p)*.

Entre los pasos 6 y 12, a partir de cada una de las componentes seleccionadas, se generan resúmenes lingüísticos como se explica a continuación:

En el paso 6.1, para cada *i*-ésima componente se calcula un umbral que llamaremos *“codo” (Velbow)*, el cual en el paso 6.2 se usa para seleccionar los atributos que representan las variables más relevantes del componente en cuestión. Luego en el paso 6.3 se recupera para la *i*-ésima componente, el conjunto de datos , que representa la proyección de los objetos del conjunto *U* respecto a los atributos . De esta forma se logra la reducción del espacio de búsqueda, porque solo se trabajará con las variables representativas según la componente que se esté analizando. La complejidad de estos pasos está acotada superiormente por *O(n·p)*.

Luego en el paso 6.4 se aplica una técnica para la generación de reglas de asociación, pero solo sobre los datos del conjunto , lo que permite reducir el espacio de búsqueda. En este paso del algoritmo, no se preestablece ninguna técnica de generación de reglas de asociación en particular, pero se recomienda que en la implementación práctica del algoritmo *PCA\_LDS,* se consideren las ventajas y limitaciones de diferentes técnicas para la generación de las reglas de asociación (Medina et al. 2007).

En la experimentación de esta investigación se decide emplear *Apriori* *(**W**a**silewska 2007)* para la generación de reglas de asociación. Esta decisión estuvo motivada porque en la bibliografía consultada (Dijkman y Wilbik 2017) usan *Apriori* para la generación de resúmenes lingüísticos y es interés de esta investigación, evitar posibles sesgos en la comparación por el uso de una u otra técnica de aprendizaje de reglas de asociación. La construcción de reglas de asociación tiene complejidad *O(2p+1),* tomando en consideración la complejidad del algoritmo *Apriori* *(**T**a**hyudin, Haviluddin y Nambo 2019)*.

Considerando las complejidades computacionales de los pasos independientes que conforman el paso 6, se deduce que la complejidad de este paso es *O(p· [(p* log *p)* + *2p+1])*, acotada superiormente por *O(p·2p+1)*.

En el paso 7 se procede a generar resúmenes candidatos a partir del conjunto de reglas generado en los pasos anteriores. En este proceso, los antecedentes de las reglas son convertidos en filtros *R* y los consecuentes en sumarizadores *S*; por tanto, en este momento cada resumen candidato tiene la forma *(R, S)*. Es de interés señalar que este algoritmo permite tanto la generación de resúmenes sin filtros cubriendo las protoformas 1 y 3, como la generación de resúmenes con filtros, cubriendo el resto de las protoformas. La cantidad de reglas generadas está acotada superiormente por la cantidad total de casos *n,* y la cantidad de elementos en filtros y sumarizadores está acotada superiormente por la cantidad total de atributos *p*; por tanto, la complejidad de este paso está acotada por *O(n·p).*

En el paso 8, para cada uno de los resúmenes candidatos se procede a determinar el cuantificador, para ello se aplica el Algoritmo 3 con una complejidad *O(n2·p)*.

En el paso 9 del Algoritmo 2*,* se procede al cálculo de los indicadores *T*, que permiten evaluar la calidad de los resúmenes lingüísticos obtenidos. La complejidad computacional de este paso está acotada superiormente por *O(n·p),* y está dada por la complejidad de los métodos de cálculo de los indicadores.

Al concluir los pasos 8 y 9 del algoritmo *PCA\_LDS,* se tiene el conjunto de resúmenes candidatos *Candidates,* dondecada candidato cumple con la estructura *P(Q, R, S, T)*.

*Algoritmo* 3 *calculate\_quantifiers(CandidatesSummaries, LVQ)*

*Entradas*

*CandidateSummaries: listado de resúmenes candidatos, cada uno con la forma (R, S)*

*LVQ: parámetros definidos en la sección 2.1 sobre notaciones.*

*Inicio:*

1. *Para cada Candidate CandidatesSummaries hacer //*
2. *R = Candidate.filters*
3. *S = Candidate.summaries*
4. *Si R = *
5. *F = calculate\_F(frecuencia de aparición del sumarizador S en UB)*
6. *relative\_frequency = F / | UB |*
7. *sino*
8. *F = calculate\_F (frecuencia de aparición del par R, S en el conjunto UB)*
9. *relative\_frequency = F / | UB |*
10. *Fin del paso condicional 1.3*
11. *Determinar el cuantificador Q a partir de aplicar el criterio de máxima membresía de la frecuencia relativa en la variable lingüística, LVQ*
12. *Candidate(R, S)  Q*
13. *Fin del ciclo iniciado en paso 1*
14. *Return CandidateSummaries*

Luego en el paso 10 se generan los resúmenes finales *Summaries* a partir de humanizar los resúmenes candidatos *humanize\_summaries(Candidates, ALV, LVQ, LNC)* usando el Algoritmo 1 con una complejidad *O(n·m·p2)*.

En el paso 11 los resúmenes son ordenados en función de los valores de *T* según se especifique y retornados en un ranking. La complejidad computacional de este paso está acotada superiormente por *O(n* log *n).*

Finalmente, la complejidad computacional del *PCA\_LDS*es *O(max(p·2p+1, n2·p, n·m·p2)).*

Este algoritmo debe ser aplicado en escenarios con datos continuos, porque está basado en variantes tradicionales del algoritmo PCA. Constituye una línea abierta a la investigación identificar otras técnicas que puedan ser empleadas con formas similares al PCA, pero que permita el trabajo en dominios con datos heterogéneos.

## Algoritmo *LPA\_LDS* basado en gráficos probabilísticos

En este algoritmo, el conjunto de datos de entrada es transformado en valores borrosos aplicando el principio de máxima membresía. Luego se aplican técnicas de aprendizaje de grafos probabilísticos (Pérez Pupo et al. 2019) (Chow y Liu 1968) que posibilitan el descubrimiento de relaciones entre variables. A partir del aprendizaje de los grafos, se genera un conjunto de árboles que son empleados para la construcción de resúmenes candidatos. Luego, los resúmenes lingüísticos candidatos son humanizados permitiendo la generación de los resúmenes lingüísticos con un enfoque multilingüe, ver Algoritmo 4.

*Algoritmo* 4 *LPA\_LDS*

*Entradas*

*U, ALV,LVQ, T, LNC,  : parámetros definidos en la sección 2.1 sobre notaciones.*

*Inicio:*

1. *Candidates* = 
2. *UB = transform\_fuzzy(U, ALV)*
3. *prob\_graph = do\_probabilistic\_graph(UB)*
4. *Por cada uno de los prob\_graph i árboles en prob\_graph hacer*
5. *Si prob\_graph i tiene más de un vértice*
6. *candidate\_summariesi = do\_candidate\_from\_branchs (prob\_graph i)*
7. *Candidates  candidate\_summariesi*

*Fin de la condicional iniciada en paso 4.1*

*Fin del ciclo iniciado en el paso 4*

1. *CandidateSummaries =* *generate\_summaries(U, Candidates, ALV, LVQ, T*)
2. *calculate\_quantifiers(CandidateSummaries, LVQ)*
3. *calculate\_T(CandidateSummaries, )*
4. *Summaries = humanize\_summaries(CandidateSummaries, ALV, LVQ, LNC)*
5. *ranking\_summaries* = *sort\_summaries(Summaries)*
6. *Return* *ranking\_summaries*

En esencia, el algoritmo en el paso 2 transforma los datos de entrada empleando variables lingüísticas *ALV,* de forma similar a como se hizo en el paso 3 del Algoritmo 2, cuya complejidad es *O(p·n·q)*.

Luego en el paso 3, al conjunto de datos transformados se le aplica un algoritmo para el aprendizaje del modelo probabilístico (Soto 2003) que mejor se aproxime al comportamiento de los datos. Ejemplos de algoritmos que pueden ser empleados en este paso se relacionan a continuación:

* Chow Liu, algoritmo desarrollado en 1968, inicialmente propuesto para la construcción de árboles a partir de datos (Chow y Liu 1968).
* Algoritmo Rebane-Pearl (Rebane y Pearl 2013), extiende el algoritmo Chow Liu y posibilita el aprendizaje de poli-árboles, permitiendo describir interacciones de mayor orden (Chow y Liu 1968).
* Algoritmo *Polytree Aproximation Algorithm* (PA) 1998, que basa el aprendizaje en el cálculo de la información mutua marginal y la información mutua condicional.
* Algoritmo *Learning Polytree Algorithm (*LPA), es el empleado en la experimentación y el que se recomienda en esta investigación. Fue propuesto en (Soto y Ochoa 2000) y su principal aporte es la mejora de la eficiencia respecto a los algoritmos que le preceden (Soto 2003). Este algoritmo calcula las dependencias marginales y condicionales empleando métodos de cálculo basados en el concepto de entropía. La complejidad de este algoritmo es *O(n3)* *(**S**o**to 2003)*.

Los algoritmos del paso 3 que aprenden los modelos probabilísticos generan un poli-árbol donde cada árbol puede generar un resumen candidato. Luego en el paso 4, por cada uno de los árboles se genera un objeto candidato, de forma tal que el nodo en la raíz del árbol es identificado como el atributo sumarizador y los nodos en las ramas son identificados como filtros. La cantidad de árboles que se pueden generar está acotada superiormente por la cantidad total de atributos *p,* la cantidad de elementos en filtros y sumarizadores máximo es *p-1*. Por tanto, la complejidad de este paso está acotada por *p(p - 1),* siendo *O(p2)*.

En el paso 5, se generan resúmenes candidatos a partir de la lista de objetos *Candidates* transformando cada objeto en un objeto resumen cuyos atributos son filtros y sumarizadores.

Luego se aplican los pasos del 6 al 10 que coinciden con los pasos del 8 al 12, propuestos en el Algoritmo 2. Al concluir los pasos 6 y 7 del algoritmo *LPA\_LDS,* se tiene un conjunto de resúmenes candidatos *CandidateSummaries,* dondecada resumen candidato tiene la estructura *P(Q, R, S, T).* Luego en el paso 8 se aplica el Algoritmo 1 para generar los resúmenes lingüísticos con *humanize\_summaries(CandidateSummaries, ALV, LVQ, LNC)*, la complejidad computacional de este paso es *O(n·m·p2)*.

A modo general, la complejidad del algoritmo *LPA\_LDS* es *O(n3).*

Este algoritmo supone obtener buenos resultados donde las variables del conjunto de datos a analizar estén relacionadas. El *LPA\_LDS* hereda las limitaciones de los algoritmos de aprendizaje de grafos probabilísticos que se use en el paso 3, que se concentran en el descubrimiento de relaciones entre variables de primer, segundo y tercer orden. Si se tienen en cuenta relaciones de orden superior, los requerimientos de espacio y tiempo aumentan exponencialmente con respecto al orden de las relaciones*(**S**o**to 2003)*.

## Algoritmo RST\_LDS basado en conjuntos aproximados

La teoría de los conjuntos aproximados también se presenta como una alternativa válida para el descubrimiento de relaciones entre atributos en la generación de resúmenes lingüísticos.

Esta teoría fue propuesta en 1986 por Z. Pawlak para su aplicación en inconsistencia de datos. Por lo general, los conjuntos aproximados se utilizan en dos alternativas: para datos discretos (Pawlak y Skowron 2007) basados ​​en relaciones de equivalencia, o para relaciones de indiscernibilidad extendidas (Bello, Falcon y Verdegay 2019). Se exponen diferentes extensiones de aplicaciones de conjuntos aproximados en (Pawlak y Skowron 2007), (Guoyin Wang 2002) y (Zhai, Zhang y Zhang 2016).

Sea un sistema de información *SI = (U, AD)* donde *U* es unconjunto de datos descrito formado por los conjuntos de atributos *A* y *D,* usualmente denominados en la teoría de conjuntos aproximados como atributos predictores y atributos de decisión respectivamente, se denomina *H* al conjunto total de atributos del problema en cuestión, tal que *H* = *AD, XU* es un conjunto de objetos y *B A* un conjunto seleccionado de atributos. De la información contenida en *B,* la variable *X* puede ser aproximada de la forma siguiente:

* La aproximación inferior de *X* con respecto a *B* es:

*B*∗*(X)= {x* ∈ *U: B(x)* ⊆ *X}* Ecuación 3

* La aproximación superior de *X* con respecto a *B* es:

*B\*(X) = {x ∈ U: B(x)* ∩ *X  Φ}* Ecuación 4

* La región frontera puede ser definida como:

*BNB(X)= B\*(X) - B∗(X)* Ecuación 5

* La región negativa de la decisión *d* con respecto a *B* es:

*NEGB(X)= U – B∗(X)* Ecuación 6

* Relación de indiscernibilidad: define una relación de equivalencia *INDB* (Pawlak 1982), (Bello y Verdegay 2012) y esta relación es denotada por:

*INDB*(*B*) = {(*x*, *y*) ∈*U*×*U*: *a*(*x*) = *a*(*y*) para cada *a*∈*B*} Ecuación 7

* La región positiva de la decisión *d* con respecto a *B* es:

*POSB* (*d*) *=*  {*B∗*(*X*)*: X* ∈ *U*/*IND*(*d*)*, d* ∈ *D*}Ecuación 8

Otro concepto útil es la dependencia en grado *k*, que se explica en la Definición 3.

**Definición** 3**:** Intuitivamente, un conjunto de atributos de decisión *D*, depende totalmente de un conjunto de atributos *B*, denotados por *B  D*, si todos los valores de los atributos *D* están determinados unívocamente por los valores del atributo en *B*. En otras palabras, *D* depende totalmente de *B*, si existe una dependencia funcional entre los valores de *D* y *B* (Pawlak 1982). *D* depende de *B* en un grado *k* donde *k*  [0,1] y se denota por *B k D*, ver Ecuación . Si *k* = 1 entonces *D* depende totalmente de *B*, mientras que si *k<1* entonces *D* depende parcialmente de *B*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Ecuación 9 |  | Donde: | Ecuación 10 |

El Algoritmo 5 está basado en la teoría de los conjuntos aproximados para la generación de resúmenes lingüísticos (Pérez Pupo, Piñero Pérez, Bello, et al. 2020). Inicialmente, las variables del conjunto de datos de entrada son transformadas a variables lingüísticas aplicando el principio de máxima membresía. Se constituyen dos conjuntos de atributos que representarán a los filtros y sumarizadores de los resúmenes lingüísticos a construir. Estos conjuntos se inicializan con todas las variables del conjunto de datos como subconjuntos independientes y durante el proceso de búsqueda se van transformando. Posteriormente, se calcula la dependencia en grado *k* entre grupos de atributos que representan filtros y sumarizadores; aquellos que superen el umbral establecido conformarán resúmenes candidatos. Luego, los resúmenes lingüísticos candidatos son filtrados y humanizados mediante un lenguaje natural controlado, permitiendo la generación de los resúmenes lingüísticos finales, ver Algoritmo 5.

*Algoritmo* 5 *RST\_LDS*

*Entradas:*

*U, ALV, LVQ, T, LNC,  : parámetros definidos en la sección 2.1 sobre notaciones.*

*Sistema de información SI = (U, A  D)*

*A, D: atributos del sistema de información.*

*k1: alfa corte k1, establece el límite superior de la dependencia en grado k en la generación de resúmenes.*

*k2: alfa corte k2, establece el límite inferior de la dependencia en grado k en la generación de resúmenes.*

*Inicio*

1. *Transformar SI = (U, A  D) en un dataset lingüístico considerando los conjuntos borrosos y los principios de máxima membresía.*
2. *A\_ItemSet = attributes\_to\_sets(AD)*
3. *D\_ItemSet = A\_ItemSet*
4. *CS = {}*
5. *Mientras Stackset no esté vacío*
6. *A\_ItemSet = Stackset.pop*
7. *Por cada B  A\_ItemSet*
8. *Por cada X D\_ItemSet: B  X = *
9. , , *B k D, k(B, X)  k2*
10. *Por cada Ot(B,X) POSB(D)*
11. *Si (k(B, X)  k1) entonces*
12. *CS = CS  Ot(B,X)*
13. *Sino*
14. *Si (k(B, X)  k2 ) entonces*
15. *Stackset.push({B  X})*

*Fin del condicional iniciado en el paso 6.9*

*Fin del condicional iniciado en el paso 6.6*

*Fin del ciclo iniciado en el paso 6.5*

*Fin del ciclo iniciado en el paso 6.3*

*Fin del ciclo iniciado en el paso 6.2*

*Fin del ciclo iniciado en el paso 6*

1. *CandidateSummaries = generate\_summaries\_from\_candidates(**U, CS, ALV, LVQ, T)*
2. *calculate\_quantifiers(CandidateSummaries,* *LVQ)*
3. *calculate\_T(CandidateSummaries,* *)*
4. *Summaries = humanize\_summaries(CandidateSummaries,* *ALV, LVQ , LNC)*
5. *ranking\_summaries = sort\_summaries(Summaries)*
6. *Return ranking\_summaries*

El paso 1 se transforma el sistema de información *S* en un conjunto de datos lingüísticos considerando los conjuntos borrosos y los principios de máxima membresía, tiene complejidad *O(p·n·q)*.

En el paso 2, cada elemento *a**A D* es transformado en un conjunto {*a*} e insertado en un conjunto *A\_ItemSet*, Ejemplo: {*a1, a2, a3*} *=>* {{*a1*}*,* {*a2*}, {*a3*}}. Luego en los pasos 3 y 4 se inicializan la variable *D\_ItemSet* y la pila *Stackset*. La complejidad de estos pasos es *O(p).*

El paso 6 constituye el corazón de este algoritmo, en los pasos 6.2 y 6.3 se crean pares de conjuntos *B* y *X* que representan a posibles filtros y sumarizadores respectivamente. Luego en el paso 6.4 se calcula la región positiva *POSB(D)* aplicando la Definición 3, así como la Ecuación y la Ecuación .La condición *k(B, X)  k2*, ayuda a encontrar resúmenes entre atributos de dependencia parcial, pero considerando los conjuntos de atributos que tienen un nivel mínimo de relación. Luego, en el paso 6.6, *Ot*  *U*, *Ot*(*B*,*X*) representa los valores de los atributos *a**B**X* en el objeto *Ot;* si hay dependencia mayor que el umbral *k1* en el contexto *Ot*, entonces los valores de los atributos *Ot*(*B*,*X*) se utilizarán como resúmenes candidatos. Si la dependencia encontrada cumple que *k(B,X)*[*k2, k1*], se genera el conjunto *Z*=*B**X* y este nuevo conjunto se añade a la pila para continuar el proceso de búsqueda, donde este nuevo conjunto pasará a formar parte de filtros en nuevos resúmenes. La complejidad del ciclo iniciado en el paso 6 está acotada superiormente por *O*(2*p*)*.*

En el paso 7 se recorren los pares de conjuntos que están en el conjunto *CS* de resúmenes candidatos, cada conjunto representa un posible resumen lingüístico. Por tanto, la complejidad de este paso está acotada por un *O(n).* Luego se aplican los pasos del 8 al 12 similar a los pasos del 8 al 12 propuestos en el Algoritmo 2.

Al finalizar el paso 9, se tiene un conjunto de resúmenes candidatos *CandidateSummaries* y luego en el paso 10 se aplica el Algoritmo 1 para generar los resúmenes lingüísticos, con una complejidad computacional *O(n·m·p2).*

Finalmente, la complejidad del algoritmo *RST\_LDS* es *O(max (n2·p, n·m·p2)).* Este algoritmo, por su naturaleza, permite el descubrimiento de resúmenes lingüísticos que representan situaciones con muy baja frecuencia de aparición en la base de datos y con una alta dependencia entre los atributos. Estos resúmenes permiten identificar datos anómalos, lo cual amplía su aplicabilidad en diferentes contextos, por ejemplo, en escenarios asociados al aseguramiento de ingresos (Castro Aguilar 2017) o la seguridad informática.

## **Diferentes enfoques que vinculan la generación de resúmenes lingüísticos y la minería de datos** anómalos

Otro escenario interesante para la aplicación de los resúmenes lingüísticos es su uso para el descubrimiento de datos anómalos y la interpretabilidad de éstos. Existen diferentes definiciones de datos anómalos, en este trabajo nos basaremos en la Definición 4 propuesta por (Aggarwal y Sathe 2017).

**Definición** 4**:** Un dato anómalo es una observación que se desvía mucho del resto de las observaciones, apareciendo como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de los datos (Aggarwal y Sathe 2017).

Ejemplos de escenarios donde es útil el descubrimiento de datos anómalos son: el descubrimiento de errores en la planificación de proyectos, procesos de aseguramiento de ingresos (Castro Aguilar 2017), los problemas de detección de fallas entre otros. En esta investigación se proponen los siguientes enfoques para la generación de resúmenes lingüísticos orientados a la interpretación y el análisis de datos anómalos:

* Primer enfoque, centrado en la generación de resúmenes lingüísticos que permitan identificar situaciones poco frecuentes en la base de datos, que puedan ser situaciones anómalas según expertos de un dominio de aplicación.
* Segundo enfoque, centrado en la aplicación de algoritmos para el descubrimiento de datos anómalos y luego la construcción de resúmenes lingüísticos que faciliten la interpretabilidad de las situaciones anómalas.

El primer enfoque puede ser logrado a partir de algunos de los algoritmos propuestos en este trabajo como por ejemplo, los algoritmos *RST\_LDS* y *LPA\_LDS*. Estos algoritmos permiten identificar relaciones fuertemente dependientes entre variables, aun cuando estas relaciones se presenten en una muy baja frecuencia de los objetos en la base de datos.

El segundo enfoque puede ser desarrollado combinando diferentes técnicas para la generación de datos anómalos con algoritmos para la generación de resúmenes lingüísticos. La Figura 34 del Anexo 4 muestra una taxonomía de métodos empleados para el descubrimiento de datos anómalos, que clasifica las técnicas en métodos supervisados, métodos semi-supervisados y métodos no supervisados (Castro Aguilar 2017) (Ranga Suri, Murty M y Athithan 2019).

Los métodos no supervisados son los más empleados, entre ellos se destacan: los métodos basados en agrupamientos (Zhao et al. 2020) (Alves et al. 2017) (Cursi y Yang 2019), métodos basados en distancias como (Nguyen et al. 2020) (Pokhrel, Pokharel y Kumar Timalsina 2019) (Ning, Chen y Chen 2018) (Wang et al. 2020), modelos basados en la densidad (Kamalov y Leung 2020) (Li et al. 2020) (Verma, Taneja y Arora 2017), modelos estadísticos basados en la distribución (Zhang et al. 2014).

Un ejemplo de un algoritmo que se basa en el segundo enfoque lo constituye el Algoritmo 6, el cual combina un algoritmo de detección de datos anómalos basado en la proximidad, que usa la distancia de *Mahalanobis,* y luego transforma los datos anómalos en resúmenes lingüísticos.

**Algoritmo** 6**: Algoritmo** ***LDS\_outliers***

*Entradas:*

*U, ALV,LVQ, T, LNC, : parámetros definidos en la sección 2.1 sobre notaciones.*

*C: conjunto de semillas centroides.*

*Distancia(d,P): función de distancia desde d al conjunto de puntos P.*

*percentil: percentil usado para la determinación de datos anómalos.*

*k1: alfa corte k1, establece límite superior de la dependencia en grado k en la generación de resúmenes.*

*k2: alfa corte k2, establece límite inferior de la dependencia en grado k en la generación de resúmenes.*

*Inicio*

1. *Out = U // se inicializa con todos los datos*
2. *clusters = Cluster(U, centers=C)*
3. *centers = clusters.centers*
4. *Para cada clusteri en clusters, hacer*
5. *B0 = Calculate\_threshold(clustersi)*
6. *Oi = out\_centers(Bo, clusteri)*
7. *Out = Out  Oi*

*Fin del ciclo paso 4*

1. *Out = Ranking\_outlier(Out, percentil)*
2. *CandidateSummaries＝Aplicar RST\_LDS(Out, ALV, k1,k2,LVQ, T, )*
3. *CandidateSummaries = empower\_summaries(CandidateSummaries, , Outf)*
4. *Summaries = filter\_with\_experts(CandidateSummaries)*
5. *Return Summaries*

En el primer paso se inicializa el conjunto de datos anómalos *Out* con todos los datos, se pretende eliminar de este conjunto todos los datos que no tengan comportamiento anómalo. En el paso 2 se construyen *C* clústeres a partir de los datos, este algoritmo podría aplicarse con diferentes métodos de agrupación. La selección del algoritmo apropiado para el agrupamiento, corresponde a los expertos del dominio en cuestión y depende de la naturaleza de los datos, conociendo las potencialidades y desventajas de cada algoritmo (Pasuoathi et al. 2021). La complejidad de este paso dependerá del algoritmo de agrupamiento que se aplique. Por ejemplo, para datos numéricos se podría utilizar el algoritmo *kmeans,* conociendo que los métodos basados en distancias tienden a formar híper-esferas, tendría una complejidad *O(n2)* *(**N**e**rurkar et al. 2018)* *(**D**a**latu 2016)* *(**Z**h**ang, Xiaoliang et al. 2020).*

En el paso 4, para cada uno de los clústeres obtenidos, se calcula el umbral *B0,*basado en el concepto “conjunto *b0* compacto” definido en (Ruiz-Shulcloper 2009). Luego en el paso 4.2 se detecta, para cada clúster, el conjunto de datos que tienen comportamiento anómalo y se intercepta este conjunto con los elementos del conjunto *Out*. Al finalizar este paso, se tiene en *Out* solo los datos que pueden ser anómalos.

En el paso 5, los objetos identificados como posibles valores anómalos son ordenados en un ranking, donde los elementos más alejados de los centros de los clústeres quedan ubicados en los primeros lugares del ranking. Este paso está acotado superiormente por O(*m·log(m)*), siendo *m* la cantidad de datos anómalos *m<<n*.

En el paso 6 se aplica el algoritmo *RST\_LDS* para la generación de resúmenes a partir del conjunto de datos anómalos *Out*. Luego en el paso 7, los resúmenes lingüísticos generados se enriquecen y se les asocian los datos anómalos que están representados con mayor fuerza en el resumen. Finalmente, en el paso 8 se produce un proceso de aprendizaje activo[[9]](#footnote-10), donde los expertos identifican aquellos resúmenes lingüísticos que representen realmente datos anómalos y que sean interesantes para los procesos de toma de decisiones. En general la complejidad computacional del Algoritmo 6 está acotada superiormente por *O(n3).*

## Valoración general de los algoritmos propuestos

Los algoritmos *LPA\_LDS*, *PCA\_LDS* y *RST\_LDS* fueron diseñados sobre la base de mejorar la eficacia de los algoritmos de generación de resúmenes lingüísticos reportados en la bibliografía consultada. Un breve resumen comparativo de estos algoritmos se muestra en la Tabla 5.

Tabla 5: Comparación teórica de los algoritmos.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *PCA\_LDS* | *LPA\_LDS* | *RST\_LDS* |
| Complejidad | *O(max(p·2p+1, n2·p, n·m·p2))* | *O(n3)* | *O(max (n2·p, n·m·p2))* |
| Tipos de datos | Continuos | Heterogéneos | Heterogéneos |
| Capacidad de descubrir datos anómalos | No | Sí | Sí |

En particular, el *PCA\_LDS* aprovecha las potencialidades del algoritmo *PCA* para identificar las relaciones entre variables como primer paso del proceso de construcción de resúmenes lingüísticos, de esta forma se reduce el espacio de búsqueda, una de las limitaciones presentes en los algoritmos que descubren resúmenes lingüísticos basados en reglas de asociación (Kacprzyk y Zadrożny 2003) (Wilbik, Kaymak y Dijkman 2017). Este algoritmo parte de datos continuos, por lo que se identifica como oportunidad de mejora, el uso de otras técnicas que permitan la reducción del espacio de búsqueda partiendo de datos heterogéneos.

El algoritmo *LPA\_LDS* aprovecha las potencialidades de los algoritmos de aprendizaje de grafos probabilísticos. En este sentido existen varios algoritmos que pueden ser empleados para el descubrimiento de los grafos probabilísticos, siendo el *LPA* la variante que se emplea en la experimentación de esta investigación por las siguientes razones: mejora la eficiencia respecto a los algoritmos que le preceden y resuelve la dirección de las dependencias (Soto 2003)*,* *(**S**o**to y Ochoa 2000)*. Este algoritmo, por su naturaleza, es capaz de descubrir relaciones fuertes entre variables, aunque sean poco frecuentes en la base de datos; esta característica permite su uso también para el descubrimiento de datos anómalos.

El algoritmo *RST\_LDS* aprovecha los principios de dependencia en grado *k* de la teoría de los conjuntos aproximados, para el descubrimiento de relaciones entre conceptos que luego son transformadas en resúmenes lingüísticos. Al igual que el *LPA\_LDS*, es capaz de identificar relaciones fuertes entre variables, tanto en situaciones con alta frecuencia de aparición en la base de datos, como en situaciones con baja frecuencia de aparición. Una de las ventajas de los algoritmos *RST\_LDS* y *LPA\_LDS* respecto al *PCA\_LDS* es que permiten el trabajo con datos heterogéneos.

En la Tabla 5 se puede observar que el algoritmo de mayor complejidad entre los propuestos es el *LPA\_LDS,* su complejidad depende fundamentalmente de la cantidad de objetos de la base de datos (*n)*;mientras que la complejidad de los algoritmos *PCA\_LDS* y *RST\_LDS* dependen de la cantidad de objetos de la base de datos (*n*) y de la cantidad de variables (*p*), siendo *RST\_LDS* el de menor complejidad.

Los algoritmos propuestos para la generación de resúmenes lingüísticos, presentan una complejidad inferior a otros algoritmos reportados en la bibliografía basados en el uso de metaheurísticas (Donis-Díaz, Bello y Kacprzyk 2015) (George et al. 1996) (Kacprzyk y Strykowski 1999) (Donis-Díaz, Muro, et al. 2014) (Donis-Díaz, Bello, et al. 2014). Además, los algoritmos propuestos están concebidos para ser integrados con lenguajes naturales controlados para facilitar la construcción de resúmenes lingüísticos en diferentes idiomas, garantizando su carácter multilingüe.

## Conclusiones del capítulo

En este capítulo se proponen varios algoritmos para la generación de resúmenes lingüísticos que combinan diferentes técnicas de *soft computing*, se arriba a las siguientes conclusiones:

* El algoritmo *PCA\_LDS* combina la generación de reglas de asociación con el análisis de componentes principales, de esta forma reduce las dimensiones del espacio de búsqueda durante la generación de resúmenes lingüísticos. Una limitante de este algoritmo es que requiere que los datos de entrada sean numéricos.
* El algoritmo *LPA\_LDS* descubre las relaciones más fuertes entre las variables empleando el aprendizaje de grafos probabilísticos. Luego genera resúmenes lingüísticos a partir de los grafos aprendidos, reduciendo el espacio de búsqueda. Este algoritmo es capaz de trabajar con datos heterogéneos. Tiene como limitante, que solo descubre relaciones de primer, segundo y tercer orden, porque el consumo de espacio y tiempo aumenta exponencialmente con respecto al orden de las relaciones.
* El algoritmo *RST\_LDS*, aplica la dependencia en grado k y otros conceptos de la teoría de conjuntos aproximados para la generación de resúmenes lingüísticos. Este algoritmo permite el trabajo con datos heterogéneos y por su naturaleza, permite generar resúmenes lingüísticos que representen tanto situaciones con alta frecuencia en la base de datos como resúmenes que representen situaciones con baja ocurrencia.
* En el capítulo se proponen dos enfoques que combinan la generación de resúmenes lingüísticos con el descubrimiento de datos anómalos. El primer enfoque es centrado en el uso de algoritmos que sean capaces de identificar situaciones pocos frecuentes en la base de datos, que puedan representar datos anómalos. El *RST\_LDS* y *LPA\_LDS* son ejemplos de algoritmos que pueden detectar datos anómalos basados en este primer enfoque. Por otro lado, el segundo enfoque combina algoritmos para la generación de datos anómalos y luego la construcción de resúmenes lingüísticos que faciliten la interpretabilidad de los datos anómalos, como el algoritmo *LDS\_outliers*.
* Los algoritmos propuestos incluyen una etapa de humanización que emplea lenguajes naturales controlados en los idiomas español, inglés, japonés y árabe para mejorar la generación e interpretabilidad de los resúmenes lingüísticos bajo un enfoque multilingüe.

# Capítulo 3: Nuevos métodos de evaluación de los resúmenes lingüísticos

En este capítulo se analizan diferentes enfoques para la evaluación de los resúmenes lingüísticos y se organiza en tres secciones. En la primera sección 3.1 se analiza y describe un conjunto de indicadores reportados en la bibliografía. En la segunda sección 3.2 se propone un conjunto de nuevos indicadores que complementan o extienden a los tradicionales, siendo este uno de los elementos que se aporta como novedad en la presente investigación. Luego en la sección 3.3 se realiza la comparación de las extensiones propuestas con los indicadores tradicionales, a partir de evaluar el comportamiento de ambos grupos, en un conjunto de clases de pruebas que cubren diferentes escenarios de resúmenes. Finalmente se presentan las conclusiones del capítulo.

## Indicadores tradicionales para la evaluación de los resúmenes lingüísticos

Existen artículos donde se propone el cálculo del valor de verdad de un resumen, como los trabajos publicados por Yager (Yager 1982), Zadeh (Zadeh 1983) y Kacprzyk y Zadrożny (Kacprzyk y Zadrożny 2000). A éstos, le sucedieron otros artículos (Kacprzyk, Yager y Zadrożny 2000) (Kacprzyk y Yager 2001) (Kacprzyk y Wilbik 2010)donde se exponen otros indicadores para evaluar un resumen como el grado de imprecisión, de cobertura, de adecuación y la longitud de un resumen.

Antes del análisis de estos indicadores, revisemos las notaciones necesarias para su comprensión.

Sea *Y = {y1, ..., yn}* una base de datos de objetos(ej. “persona”)descritos por atributos *V* (ej. “edad”)el cual puede tomar valores *X = {x1, x2, ..., xm,},* ejemplo *{1, 2, ..., 100}*. Luego, *di=V(yi)* denota el valor del atributo *V* para el objeto *yi*. Para un conjunto de datos de *m* atributos *V={V1, V2, ..., Vm}*, de ahí que *Vj(yi)* denota el valor del atributo *Vj* para el objeto *yi*, y el atributo *Vj* toma valores en el dominio del conjunto *Xj*.

A partir de *Y* se puede formar el dataset *D* tal que, ver Ecuación :

|  |  |
| --- | --- |
| *D = {[V1(y1), V2(y1), ..., Vm(y1)],..., [V1(yn), V2(yn), ..., Vm(yn)]}* | Ecuación 11 |

Sea un sumarizador *S = {S1, S2, ..., Sm}* descrito por varios atributos *Si*, entonces, , *i* = 1, 2, …,*n,* representa el grado de certeza del sumarizador y puede ser definido como muestra la Ecuación :

|  |  |
| --- | --- |
| *minj∈{1, 2, ..., m}[ μSj(Vj(yi)) ]* | Ecuación 12 |

En caso de que el resumen contenga filtros, se define según la Ecuación :

|  |  |
| --- | --- |
| *minj∈{1, 2, ..., m}[ μSj(Vj(yi)) ∧ μRg(Vg(yg)) ]* | Ecuación 13 |

Donde “∧” es una t-norma y *μRg(Vg(yg))* representa el grado de certeza de los filtros del resumen.

### T1: grado de verdad

El “grado de verdad *T*1” determina la validez de un resumen, criterio introducido por Yager en (Yager 1982). En la Ecuación se muestra cómo calcular este indicador para los resúmenes sin filtro y la Ecuación para los que lo tienen.

|  |  |
| --- | --- |
|  | Ecuación 14 |
|  | Ecuación 15 |
|  | Ecuación 16 |
|  | Ecuación 17 |

Donde: es el grado de pertenencia del filtro de un resumen.

es el grado de pertenencia del sumarizador de un resumen.

∩ representa una t-norma, que pudiera ser la operación “mínimo” o una productoria.

*n* es la cantidad de objetos de la base de datos.

*µQ(r)* es la función de pertenencia que representa el cuantificador lingüístico *Q* *(**K**a**cprzyk y Zadrożny 2000)*.

El grado de verdad (*T*1) mide el porcentaje de veracidad de un resumen respecto a los datos; si este valor estuviera cercano a 1, el resumen podría ser importante para la toma de alguna decisión. Pero también pudiera estar cercano a 0, pues este indicador considera todos los casos cuyo grado de verdad es superior a cero, por lo que valora también a los resúmenes de muy baja certeza. Sin embargo, en muchos escenarios reales donde el grado de indeterminación o falsedad del resumen podría influir también en una decisión a tomar, caracterizados por la influencia de numerosos especialistas, los cuales presentan a menudo, criterios neutros u opuestos a las variables y/o eventos de un escenario en cuestión, esta variante no sería viable, pues no considera elementos de indeterminación tratados en la teoría neutrosófica (Smarandache y Pramanik 2018).

### T2: grado de imprecisión

El “grado de imprecisión (*T*2)” es un criterio de validez obvio e importante (Kacprzyk y Zadrożny 2010), pues un resumen lingüístico muy impreciso con un alto grado de verdad no es muy útil (Kacprzyk y Zadrożny 2010). Ejemplo, el resumen “en casi todos los días de invierno, la temperatura es bastante fría” tiene un valor de verdad muy alto (el *T*1 tendería a 1); sin embargo, es muy impreciso, pues no es útil o no genera información valiosa. Para calcular este indicador se emplean la Ecuación y la Ecuación . Como se puede observar en la Ecuación , este indicador depende del sumarizador del resumen y no de la base de datos; es decir, para calcularlo no se recorren los registros, solo se considera la cardinalidad de cada conjunto lingüístico del sumarizador, así como su dominio.

|  |  |
| --- | --- |
|  | Ecuación 18 |
|  | Ecuación 19 |

Donde: *m* es la cantidad de conjuntos borrosos del sumarizador *S = {S1, S2, ... Sj, ..., Sm}.*

denota la cardinalidad del conjunto no borroso correspondiente.

es el dominio, que representa todos los valores finitos del conjunto borroso del sumarizador *Sj.*

Este indicador mide el grado de imprecisión de un resumen, pero no incluye elementos de indeterminación, lo cual podría ser complementado con elementos de la teoría neutrosófica (Smarandache y Pramanik 2018).

### T3: grado de cobertura

El “grado de cobertura (*T*3)”, indica cuántos objetos en la base de datos que cumplen con el filtro *R* están cubiertos por el resumen cuyo sumarizador es *S.* Su interpretación es simple, por ejemplo, si el valor fuera 0.15, entonces significa que el 15% de los objetos son consistentes con el resumen en cuestión. El valor de este indicador depende claramente del contenido de la base de datos (Kacprzyk y Zadrożny 2010) y se calcula como muestran las siguientes ecuaciones: Ecuación , Ecuación y Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 20 |
|  | Ecuación 21 |
|  | Ecuación 22 |

Este indicador tiene su base en la probabilidad condicional (van Rooij y Schulz 2019) de aparición de los sumarizadores dado que tengan lugar los filtros, y en el concepto de conjunto soporte tratado en la lógica borrosa (Cpałka 2017). En esta investigación se considera que puede ser relevante conocer cuánto cubre un resumen a la base de datos. En este sentido se realiza una propuesta de extensión de este indicador en la sección 3.2.4.

### T4: grado de adecuación o idoneidad

El “grado de adecuación (*T*4)” es el indicador más relevante según (Kacprzyk y Zadrożny 2010), ver las ecuaciones: Ecuación , Ecuación , Ecuación y Ecuación . Supongamos que el resumen contiene una descripción (conjunto borroso) *S = (S1, S2, ..., Sm),* el cual es particionado en *m* sumarizadores compuestos por atributos *V1, V2, ..., Vm*, tal que cada sumarizador se corresponde a un conjunto borroso, entonces se denota:

|  |  |
| --- | --- |
|  | Ecuación 23 |

Luego se puede calcular el *T*4 según las siguientes ecuaciones:

|  |  |
| --- | --- |
|  | Ecuación 24 |
| para *j* = 1,…, *n* | Ecuación 25 |
|  | Ecuación 26 |

Ejemplificándolo con una base de datos sobre trabajadores, si el 50% de ellos son menores de 25 años y el 50% son altamente calificados, entonces se puede esperar que los empleados que sean menores de 25 años y altamente calificados estén cerca del 25%. Sin embargo, si el grado de adecuación es de, por ejemplo, 0.39 (es decir, el 39% tiene menos de 25 años y está altamente calificado), entonces el resumen encontrado refleja una relación interesante, no esperada en nuestros datos.

El *T*4 es una combinación lineal del *T*3 y otros elementos.

### T5: longitud de un resumen

La “longitud de un resumen (*T*5)” es un indicador importante porque mientras más largo sea el resumen, más difícil será su comprensión, ver Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 27 |

Para visualizar cómo se comporta este indicador, se construye una secuencia de datos aleatoriamente, que simulen el tamaño de resúmenes lingüísticos tomando valores entre 1 y 8. Estos valores se evalúan en la Ecuación . Posteriormente, utilizando el tabulador Excel, se construye un gráfico de líneas, donde en el eje de las ´*x´* se representa el tamaño de los resúmenes, y en el de las ´*y´* se representan los valores del *T*5*.* Seobtiene que este indicador tiene un comportamiento como el que muestra la Figura 4.

Este es un indicador que responde a una función exponencial, como se representa en la Figura 4. Para resúmenes de tamaño 1, el *T*5 alcanza su máximo valor, siendo éste un tamaño de resumen no muy adecuado, sobre todo si de él depende tomar una decisión. Sin embargo, para los resúmenes que involucran de 2 a 4 variables, el valor del *T*5 cae abruptamente, siendo éste el rango de valores más esperado de un resumen. Un resumen de tamaño 2 es bueno porque involucra a más de una variable de la base de datos; a su vez, un resumen con muchos componentes ya deja de ser legible o comprensible.

En (Kacprzyk y Wilbik 2010)se incluyen otros indicadores como:

* El “grado total de validez (*Tt*)”, el cual se presenta como una agregación lineal de los restantes indicadores, ver Ecuación :

|  |  |
| --- | --- |
|  | Ecuación 28 |

* El resumen óptimo *S\**∈ {*S*}, que sería el máximo argumento de , ver Ecuación , tal que:

|  |  |
| --- | --- |
|  | Ecuación 29 |

Donde *k* es el número total de resúmenes lingüísticos generados, y *arg* es una función que devuelve el resumen lingüístico obtenido como resultado de la operación.

## Propuestas de extensiones de *T* para evaluar resúmenes lingüísticos

En esta sección, se propone un grupo de extensiones a los indicadores tradicionales de evaluación de resúmenes lingüísticos, combinando elementos de la teoría de los conjuntos aproximados (Dubois y Prade 1990) (Bello y Verdegay 2012) y de la teoría neutrosófica[[10]](#footnote-11) (Smarandache y Pramanik 2018).

Para mayor comprensión de las propuestas, se introduce en la subsección 3.2.1 un conjunto de definiciones y notaciones que se emplean para la formulación de las extensiones propuestas. En el resto de las subsecciones se proponen extensiones para los indicadores *T1, T2, T3, T4* y *T5* respectivamente.

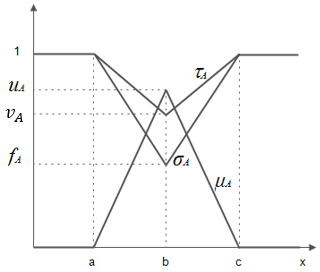
### Definiciones y notaciones empleadas en las extensiones propuestas

A continuación, se presenta un conjunto de definiciones y notaciones que forman parte de la propuesta realizada en la presente investigación y publicada por la autora en (Pérez Pupo, Piñero Pérez, et al. 2018).

**Definición** 5**. Conjunto neutrosófico:** sea *X* un universo de discurso, un conjunto neutrosófico *A* sobre *X* es un objeto de la forma: *A = { μA(x), A(x), σA(x) : x  X},* ver Figura 3 y se cumple que:

* *μA(x) [0,1]* es una función de membresía que representa el grado de certeza de la pertenencia del valor *x* al conjunto *A*, verEcuación *.*
* *A(x)*  *[0,1]* función de membresía que representa el grado de indeterminación del valor *x* al conjunto *A*, verEcuación .
* *σA(x)  [0,1]* función de membresía que representa el grado de no pertenencia (o falsedad) del valor *x* al conjunto *A*, ver Ecuación .
* *0  μA(x) + A(x) + σA(x)  3  x  X*

|  |  |
| --- | --- |
|  | Ecuación 30 |
|  | Ecuación 31 |
|  | Ecuación 32 |

Siendo *yi Y* tal que, *Y* representa a la base de datos de *n* objetos, *|Y| = n* y cadaobjeto *yi* está formado por *k* atributos. Para cada atributo *z* de un objeto *yi* existe una variable lingüística *Vz = {A1, A2, …, Ag}* formada por “*g”* conjuntos borrosos.

**Definición** 6**:** Siendo *Vz* una variable lingüística asociada al *z-ésimo* atributo de los objetos de la base de datos *Y;* la pertenencia de un objeto *yiY,*al conjunto borroso *AVz,* se denota *[0,1]* y evalúa el *z-ésimo* atributo del objeto *yi* en el conjunto borroso *A.*

**Definición** 7**:** El grado de pertenencia del objeto *yiY* a un resumen lingüístico *SUMMARY,* se denota , constituye una medida en que el objeto es cubierto o representado por un resumen lingüístico. Siendo *SUMMARY* un resumen lingüístico en su forma generalizada “*QRy´s are S”* y formado por la familia de conjuntos borrosos de *Q*, *R={ Rp}* y *S={S1, S2, ...,Sm}*, donde *Q* constituye un conjunto borroso de la variable lingüística del cuantificador *Q*, *R* constituye la familia de conjuntos borrosos que conforman los filtros y *S* la familia de conjuntos borrosos que representan los sumarizadores. El grado de pertenencia , se calcula a partir de la aplicación de una *t*-norma al conjunto de los valores de pertenencia de los atributos del objeto *yi* a la familia de los conjuntos borrosos que forman el *SUMMARY*, ver Ecuación , Ecuación y la Definición 6.

|  |  |
| --- | --- |
| *SUMMARY* tiene la estructura *“Qy´s are S”, R = * | Ecuación 33 |
| *SUMMARY* tiene la estructura“*QRy´s are S”* | Ecuación 34 |
| Donde representa la evaluación del objeto *yi* en la función de pertenencia de los conjuntos borrosos de *R* y , representa la evaluación del objeto *yi* en la función de pertenencia de los conjuntos borrosos de *Sj*. | |

**Definición** 8**:** Sea *SUMMARY* un resumen lingüístico en su forma generalizada “*QRy´s are S”*, se define como aproximación inferior en grado ** de *SUMMARY* y se denota al conjunto de objetos *yi Y,* tal que *Y* representa a la base de datos, cuyo grado de pertenencia al *SUMMARY* es mayor que un *-corte α[0,1],* Error: Reference source not found.

**Definición** 9**:** Sea *SUMMARY* un resumen lingüístico en su forma generalizada “*QRy´s are S”*, se define como aproximación superior de *SUMMARY* y se denota al conjunto de objetos *yi  Y,* tal que *Y* representa a labase de datos, cuyo grado de pertenencia al *SUMMARY* es mayor que 0. Ver Ecuación .

Figura 4: Comportamiento del *T*5.

Figura 3: Conjunto borroso incorporando los valores de veracidad, neutralidad y falsedad

|  |  |
| --- | --- |
|  | Ecuación 35 |

|  |  |
| --- | --- |
|  | Ecuación 36 |

A partir de estas aproximaciones se define la región límite como muestra la Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 37 |

Estas tres últimas ecuaciones están inspiradas en los conceptos de aproximación inferior y aproximación superior de la Teoría de Conjuntos Aproximados (Bello y Verdegay 2012). Los objetos que pertenecen a la aproximación inferior representan los objetos *yi  Y*, donde no existe duda que pertenecen al resumen *SUMMARY* considerando el *-corte;* mientras que aquellos que pertenecen a la aproximación superior son los objetos de *yi  Y* que pertenecen al resumen lingüístico en algún grado mayor que 0, que incluye tanto a los objetos de la aproximación inferior como a los de la región límite . Los objetos que pertenecen a la región límite representan objetos sobre los que se tiene cierta duda del nivel de representatividad de un determinado resumen sobre él.

**Definición** 10**:** Sea *yi  Y* un objeto y *SUMMARY* un resumen lingüístico, se define como el grado de indeterminación asociado a la pertenencia del objeto *yi* al *SUMMARY*. Se calcula para los resúmenes lingüísticos con estructura “*Qy´s are S”* como muestra la Ecuación , y para los resúmenes lingüísticos con estructura“*QRy´s are S”* como muestra la Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 38 |
|  | Ecuación 39 |

**Definición** 11**:** Sea *yi  Y* un objeto y *SUMMARY* un resumen lingüístico, se define como el grado de no pertenencia del objeto *yi* al *SUMMARY.* Se calcula para los resúmenes lingüísticos con estructura “*Qy´s are S”* como muestra la Ecuación , y para los resúmenes lingüísticos con estructura “*QRy´s are S”* como muestra la Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 40 |
|  | Ecuación 41 |

### Extensiones para el cálculo del grado de verdad Te1a

El indicador *Te1a*es también una medida de la certeza del resumen, complementa al indicador *T1,*no lo reemplaza. Tanto el *T1* como el *Te1a* consideran el cuantificador durante el cálculo. La Ecuación muestra cómo el indicador *Te1a* debe ser calculado.

La principal diferencia entre el *T1* y el *Te1a* radica en que, el *T1* considera en el numerador a todos los objetos con una pertenencia al resumen mayor que cero; mientras que el *Te1a* solo considera en el cálculo del numerado, a aquellos objetos con una pertenencia al resumen mayor que el *-corte* establecido. Además, para el caso de los resúmenes con filtros, el *T1* ubica en el denominador a la cardinalidad de la base de datos, mientras que el *Te1a* contabiliza en el denominador solo aquellos objetos con una pertenencia mayor que cero. De esta forma se persigue proponer un indicador más centrado en la evaluación de la certeza del resumen, que sea capaz de discriminar mejor los resúmenes y objetos con mayor grado de certeza de aquellos que no lo tengan.

|  |  |
| --- | --- |
|  | Ecuación 42 |

es el conjunto borroso que representa al cuantificador del resumen. Forma parte de la variable lingüística de los cuantificadores definida en el capítulo anterior, ver Ecuación .

Por otra parte, se propone la extensión *Te1b* complementando los valores de *T*1 y *Te1a*. Se calcula a partir de la media de los valores de certeza de objetos cuya pertenencia al resumen supere el *-corte*, ver Ecuación . Este indicador tiene como principal diferencia con el *T*1 y *Te1a*, que solo incluye la evaluación de pertenencia de filtros y sumarizadores sin considerar al cuantificador. De esta forma se trata de separar mejor, a resúmenes con alta frecuencia de aparición conjunta de filtros y sumarizadores, de aquellos que representan a muchos objetos pero con baja evaluación de pertenencia a sus respectivos filtros y sumarizadores.

|  |  |
| --- | --- |
|  | Ecuación 43 |

### Extensiones al grado de imprecisión Te2

Se proponen las extensiones *Te2a*, *Te2b, Te2c* y *Te2d* que complementan al *T*2,no lo reemplazan. El principal aporte de estos indicadores es la incorporación de elementos de la teoría neutrosófica. A diferencia del *T*2*,* en el cálculo de estas extensiones se emplean los objetos de la base de datos.

Se proponen estas extensiones con el objetivo de calcular con mayor precisión las medidas de certidumbre, indeterminación y no pertenencia, considerando tanto un enfoque integrado de estas medidas como su influencia de forma independiente. De esta forma se facilita el cálculo de la calidad de los resúmenes, en escenarios con presencia de fenómenos de neutralidad e indeterminación en la toma de decisiones.

En el caso del *Te2a* se calcula la incertidumbre considerando los grados de pertenencia, no pertenencia e indeterminación de cada objeto cuya pertenencia al resumen es mayor que el *-corte* establecido, como se muestra en la Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 44 |

El indicador *Te2b*, calcula la incertidumbre, pero solo considerando los grados de no pertenencia e indeterminación de los objetos cuya pertenencia al resumen es mayor que el *-corte* establecido, Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 45 |

El indicador *Te2c*, constituye la media de la indeterminación de los objetos cuya pertenencia al resumen es mayor que el *-corte* establecido, ver Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 46 |

El indicador *Te2d*, constituye la media de la no pertenencia al resumen de los objetos que superan el *-corte* (Ecuación ).

|  |  |
| --- | --- |
|  | Ecuación 47 |

### Extensión al cálculo del grado de cobertura Te3.

Se propone el indicador *Te3* como extensión, puede reemplazar al indicador tradicional *T*3. Se construye con el objetivo de determinar de una forma más estricta, el grado de cobertura de un resumen lingüístico. El *T3* tradicional considera en el cálculo de la cobertura a todos los objetos con pertenencia mayor que cero, mientras que esta extensión da una evaluación mayor a aquellos resúmenes que garanticen la cobertura de mayor cantidad de objetos, pero estimulando a aquellos que logran además, altos valores de pertenencia. La Ecuación resume la forma de cálculo de este indicador.

|  |  |
| --- | --- |
|  | Ecuación 48 |

### Extensión al cálculo del grado de adecuación Te4

La extensión *Te4*se propone como complemento al *T4* tradicional, no como reemplazo. Se propone esta extensión para medir la adecuación en escenarios de toma de decisiones con presencia de fenómenos como la indeterminación o la neutralidad. En este sentido, este indicador incorpora elementos de la teoría neutrosófica y calcula la adecuación a partir de combinar los grados de certeza, indeterminación y no pertenencia de los objetos al resumen.

Es una extensión que promueve la identificación de relaciones peculiares entre variables, relaciones que no están presentes en la mayoría de los objetos de la base de datos pero que en los objetos donde aparecen, son fuertes. Establece una relación inversamente proporcional entre el grado de pertenencia de los objetos al resumen y la cantidad de objetos que pertenecen a la aproximación superior. El *Te4*, asigna mayor valor a aquellos resúmenes que representen objetos con pertenencia mayor que el *-corte* y no a resúmenes que representen a muchos objetos en el intervalo [0, *-corte*]; es decir, no a resúmenes con muchos objetos en la aproximación superior. Otro elemento que diferencia esta extensión del *T*4 tradicional, es que en el grado de pertenencia se incluyen también el grado de no pertenencia y la indeterminación, ver Ecuación .

|  |  |
| --- | --- |
|  | Ecuación 49 |

### Extensión a la evaluación de la longitud de los resúmenes Te5

El *Te5* se propone como una extensión que reemplaza al indicador *T*5 tradicional. La principal diferencia entre estos indicadores es la función matemática sobre la que soportan su cálculo. Se propone este indicador con el objetivo de lograr una mayor suavidad en los valores de evaluación ante el aumento del tamaño de los resúmenes, mejorando los saltos bruscos que se experimentan en el *T*5 tradicional. El *Te5* está basado en la curva de Gauss, Ecuación , y asigna mayor valor a los resúmenes que tienen longitud entre dos y tres, por eso el parámetro *b = 2.5,* mientras queel valor *k = 0.125.* Estos dos valoresse establecieron a partir de prueba y error hasta lograr la suavidad deseada.

|  |  |
| --- | --- |
|  | Ecuación 50 |

## Conjunto de pruebas 1: Validación y comparación de los indicadores tradicionales y los extendidos

### Escenario de pruebas

Para la comparación de los indicadores tradicionales y los extendidos, se establece un entorno de pruebas donde se simulan 29 posibles escenarios, ver Anexo 7 Tabla 22. Cada uno de los escenarios es representado por una clase de prueba, donde cada clase de prueba representa una combinación de posibles valores de evaluación de filtros y sumarizadores, cubriendo todos los posibles escenarios en que se puedan expresar los resúmenes lingüísticos.

Los escenarios de pruebas están caracterizados por los siguientes elementos:

* Se emplean dos resúmenes prototipo. El primer prototipo tiene la estructura *“Qy’s R1 & R2 are S1* & *S2”* formado por dos filtros *R1 y R2*, y dos sumarizadores *S1 y S2.* El segundo prototipo con la estructura *“Qy’s are S1* & *S2”*, es decir, sin filtros y con dos sumarizadores *S1 y S2.*
* Cada clase de prueba representará una combinación de valores posibles de cuantificador-filtros-sumarizadores al evaluar un objeto en un resumen lingüístico. Para representar los rangos de valores de cada clase de prueba se emplean los términos ***bajo, medio, alto***querepresentan a los intervalos [0, 0.33), [0.33, 0.66) y [0.66, 1] respectivamente.

Por ejemplo, la clase de prueba ER1, ver Tabla 22 del Anexo 7, representa el escenario de resúmenes que tienen cuantificador ***poco*** y al conjunto de objetos que al ser evaluados en el resumen queden evaluados como ***bajo*** respecto a los filtros y como ***bajo*** respecto a los sumarizadores. La Tabla 6 muestra un ejemplo de objeto que pertenece a la clase de prueba ER1. En esta tabla se representa un resumen formado por dos filtros (R1, R2), dos sumarizadores (S1, S2) y el objeto 1, que al ser evaluado en este resumen, obtiene baja evaluación de pertenencia tanto en filtros como en sumarizadores, porque los valores de evaluación están en el intervalo de [0, 0.33).

Tabla 6: Ejemplo de descripción de un objeto de la clase de prueba ER1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | μM(x) Pertenencia | | | | **M (x) Indeterminación | | | | σM(x) No pertenencia | | | |
| Objeto | R1 | R2 | S1 | S2 | R1 | R2 | S1 | S2 | R1 | R2 | R1 | S2 |
| 1 | 0.15 | 0.23 | 0.23 | 0.17 | 0.80 | 0.72 | 0.72 | 0.78 | 0.72 | 0.65 | 0.65 | 0.71 |

* Cada clase de prueba está conformada por un conjunto de 10 instancias, tal que las primeras 7 representan a objetos que se corresponden con las características de la clase de prueba en cuestión y las otras 3 se generan con valores aleatorios. Un ejemplo de los objetos evaluados en un resumen para la clase de prueba ER1 se muestra en la Tabla 23 del Anexo 7.
* Para representar el cuantificador en los escenarios de prueba, se emplea una variable lingüística formada por tres conjuntos borrosos representados en el Anexo 7 Figura 37. Se definen solo 3 conjuntos lingüísticos para no introducir una innecesaria complejidad en la experimentación con un aumento de clases de prueba.
* Las clases de prueba ER28 y ER29 representan clases porosas, es decir, clases cuyos objetos, al evaluarse en el resumen, tienen muchos valores iguales a cero en los filtros y los sumarizadores.

### Análisis del comportamiento del indicador Grado de Verdad y sus extensiones

En esta subsección se caracteriza el comportamiento de los indicadores *T1* tradicional y la extensión *Te1a.* La Figura 5 grafica estos dos indicadores ante resúmenes del tipo *QRy´s are S* mientras que la Figura 6 muestra el comportamiento ante resúmenes con estructura *Qy´s are S*, es decir, resúmenes con filtro y sin filtro respectivamente.

Para el análisis del comportamiento de los indicadores, nos concentramos en el análisis lógico de la respuesta esperada de los mismos según su semántica y la respuesta real ante los objetos de la clase de prueba.

Figura 5: Comportamiento del *T*1 y *Te1a* en resúmenes con filtro.

A partir del análisis de la Figura 5 se identifica que, los resúmenes lingüísticos representados por las clases de prueba desde ER1 hasta ER8, tienen como semántica “pocos objetos que cumplen con los filtros *R1* y *R2* tienen los sumarizadores *S1* y *S2*”, por lo que se espera que los indicadores *T1* y *Te1a* alcancen valores medios o altos. Esto se infiere porque es cierto que son pocos los objetos que cumplen con esas condiciones, como lo indica el cuantificador; pues en estas clases de prueba, los objetos toman valores bajos o medios de certeza en los conjuntos borrosos de las variables representadas tanto por los filtros como por los sumarizadores. Por tanto, con seguridad se podría decir que es cierto que son pocos los objetos que cumplen con la expresión. En este contexto, nótese cómo el *Te1a* reporta mejor comportamiento que el *T1*.

Por otro lado, la semántica de los resúmenes lingüísticos representados por las clases de prueba desde ER19 hasta ER26, se interpreta como “la mayoría de los objetos que cumplen con los filtros *R1* y *R2* tienen los sumarizadores *S1* y *S2*”. En este caso la respuesta esperada de los indicadores es baja, porque no se corresponde el cuantificador “la mayoría” con la cantidad de objetos que cumplen con las condiciones del resumen; pues en este rango de clases de prueba, también los objetos toman valores bajos o medios de certeza en los conjuntos borrosos de las variables de los filtros y los sumarizadores. En este sentido el *Te1a* también reporta mejor comportamiento en estas clases de prueba.

Siguiendo los razonamientos anteriores, en las clases de prueba ER9, ER11, ER12, ER14, ER15, ER18, ER28 y ER29, el indicador tradicional y la extensión tienen un comportamiento similar y en correspondencia con el resultado esperado; mientras que en la clase ER13 el *Te1a* reportan un mejor comportamiento. Por otra parte, en las clases ER10, ER16, ER17 y ER27 para el juego de datos empleado, el *T1* tradicional reporta un mejor comportamiento respecto a la extensión *Te1a* propuesta.

Figura 6: Comportamiento del *T1* y *Te1a* en resúmenes sin filtro.

Para el caso de los resúmenes sin filtro representados en la Figura 6, se identifica que: ante las clases de prueba ER1, ER4 y ER8, la respuesta de los indicadores *T1* y *Te1a* deben alcanzar valores medios o altos. Esto se infiere porque la semántica del resumen en estas clases de prueba se interpreta como “Pocos objetos cumplen con tener valores de sumarizadores S1 y S2”. Esta expresión es cierta porque en estas clases de prueba los sumarizadores de los objetos toman valores bajos o medios de certeza. Por tanto, con seguridad se podría decir que es cierto que son pocos los objetos que cumplen con la expresión. En este contexto, nótese que el *Te1a* tiene un mejor comportamiento respecto al *T1* en ER4 y ER8, mientras que en ER1 el *T1* tuvo un mejor comportamiento.

En las clases de prueba ER11, ER14 y ER17, la respuesta de los indicadores *T1* y *Te1a* deben alcanzar valores medios o altos. Esto se infiere porque la semántica del resumen en estas clases de prueba se interpreta como “Aproximadamente el 50% de los objetos cumplen con tener valores de sumarizadores S1 y S2”. Esta expresión es cierta porque en estas clases de prueba los sumarizadores de los objetos toman valores medios de certeza. Por tanto, con seguridad se podría decir que es cierto que son aproximadamente el 50% de los objetos los que cumplen con la expresión. En este sentido, nótese que en las clases ER11 y ER14 el *Te1a* muestra mejor comportamiento, pues representa valores del conjunto lingüístico “aproximadamente el 50%” y el *T1* representa valores altos; mientras que en la clase ER17 es el *T1* el que muestra mejor comportamiento.

En la clase de prueba ER24 los indicadores *T1* y *Te1a* deben alcanzar valores altos. Esto se infiere porque la semántica del resumen para este caso se interpreta como “La mayoría de los objetos cumplen con los sumarizadores S1 y S2”. Esta expresión es cierta porque en esta clase de prueba, los sumarizadores toman valores altos de certeza. Por tanto, con seguridad se podría decir que es cierto que la mayoría de los objetos cumplen con la expresión. En este contexto, nótese que el *Te1a* tiene un mejor comportamiento que el *T1*.

En las clases de prueba ER28 y ER29 los indicadores *T1* y *Te1a* deben alcanzar valores bajos. Esto se infiere porque la semántica del resumen en estas clases de prueba se interpreta como “La mayoría de los objetos cumplen con los sumarizadores S1 y S2”. Como estas clases son porosas, al ser evaluados sus objetos en el resumen los valores de los sumarizadores son iguales a cero. Por tanto, se podría decir que no es cierto que la mayoría de los objetos cumplen con la expresión. En este contexto, nótese como el *T1* tiene un mejor comportamiento que el *Te1a*.

En las clases de prueba ER2, ER3, ER5, ER6, ER7, ER9, ER10, ER12, ER13, ER15, ER16, ER18, ER19, ER20, ER21, ER22, ER23, ER25, ER26 y ER27 el indicador tradicional y la extensión tienen un comportamiento similar y en correspondencia con el resultado esperado.

Se identifica que los resúmenes con la estructura *QRy´s are S* presentan mayores distancias entre las respuestas del *T1* y *Te1a* (media de las distancias: 0.42) que en los resúmenes con la estructura *Qy´s are S* (media de las distancias: 0.29).

### Análisis del comportamiento del indicador Grado de Soporte y su extensión

En esta subsección se caracteriza el comportamiento de los indicadores *T3* tradicional y la extensión *Te3.* A partir del análisis de la Figura 7, se puede decir que el indicador *T3* prácticamente no refleja variaciones, lo cual no es un buen comportamiento.

Figura 7: Comportamiento del T3 y Te3.

Como se puede apreciar, el *Te3* reporta un mejor comportamiento en la mayoría de las clases de prueba. En las clases donde los datos generados están por debajo del umbral, es esperado que los resultados de este indicador sean bajos. Además, este indicador debe alcanzar valores altos en las clases de prueba ER9, ER18 y ER27 porque es donde hay valores altos tanto en filtros como en sumarizadores, lo cual garantiza que exista un alto soporte para los resúmenes representados en estas clases. El comportamiento de estas clases de prueba es claramente representado por el indicador extendido *Te3*.

En general, se puede decir que el *Te3* extendido es más estricto que el tradicional, y logra diferenciar mejor los escenarios representados por las clases de prueba.

### Análisis del comportamiento del indicador Grado de Adecuación y su extensión

En la Figura 8 se puede observar que en la mayoría de las clases de prueba el *Te4* tiene un mayor poder discriminatorio de las diferentes situaciones representadas en las clases de prueba respecto al *T4.* En las clases ER9, ER18 y ER27 el indicador *Te4* presenta picos altos porque los filtros y los sumarizadores presentan altos valores de certeza. En las clases ER28 y ER29 también el *Te4* presenta picos altos porque son clases porosas, donde hay una alta certeza en pocos casos y un bajo soporte.

Figura 8: Comportamiento del T4 y Te4.

En particular el *T4* en las clases de prueba ER1 y ER3, tiene un comportamiento no esperado, porque en estos escenarios de prueba se valida con objetos que tienen una baja certeza respecto a la evaluación de los filtros, lo cual implica que exista una baja certeza en la pertenencia del objeto al resumen. Sin embargo, en estos casos se obtiene un valor de *T4* mayor que el valor del *Te4,* cuando el comportamiento esperado es un valor muy bajo del *T4.* En estas clases de prueba el indicador *Te4* tiene un mejor comportamiento que el *T4.*

### Análisis del comportamiento del indicador Longitud de un Resumen y su extensión

El análisis del *T*5 se aprecia mejor en las curvas representadas en la Figura 9. Nótese que la curva correspondiente al *T*5 tradicional, la azul, cae abruptamente. Además, asigna muy alta puntuación (el máximo valor) para resúmenes de tamaño 1; mientras que para resúmenes de tamaño 2 y 3, los más comunes, asigna una baja puntuación. Por otra parte, el *Te*5 extendido asigna valores altos a los resúmenes que involucran hasta 4 variables; y la caída de la curva es más suave, con un comportamiento similar a la curva campana. Por ello se concluye que el *Te*5 extendido tiene un mejor comportamiento que el *T*5 tradicional.

|  |  |  |
| --- | --- | --- |
| Tamaño del resumen | *T*5 Tradicional | *Te*5 Extendido |
| 1 | 1 | 0.755 |
| 2 | 0.5 | 0.969 |
| 3 | 0.25 | 0.969 |
| 4 | 0.125 | 0.755 |
| 5 | 0.063 | 0.458 |
| 6 | 0.031 | 0.216 |
| 7 | 0.016 | 0.08 |
| 8 | 0.008 | 0.023 |

Figura 9: Curvas del *T*5 tradicional y *Te*5 extendido.

### Resumen de comparación de indicadores respecto a tratamiento de la incertidumbre.

La Tabla 7 muestra una comparación de los indicadores tradicionales y extendidos respecto al tratamiento de la incertidumbre.

Tabla 7: Comparación de indicadores respecto a tratamiento de incertidumbre.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tratamiento** | **T1** | **T2** | | **T3** | | | **T4** | | | **Te1a** | | **Te1b** | | **Te2a** | | **Te2b** | | **Te2c** | | **Te2d** | **Te3** | | **Te4** | |
| Certidumbre | | |  | |  |  | |  |  | |  | |  | |  | |  | |  | | |  | |  |
| Indeterminación | | |  | |  |  | |  |  | |  | |  | |  | |  | |  | | |  | |  |
| Falsedad | | |  | |  |  | |  |  | |  | |  | |  | |  | |  | | |  | |  |
| Umbral parametrizable | | |  | |  |  | |  |  | |  | |  | |  | |  | |  | | |  | |  |

La Tabla 7 muestra que los indicadores tradicionales solo tratan la certeza y lo hacen empleando un umbral fijo (cero); mientras que los indicadores extendidos emplean un umbral parametrizable, lo cual flexibiliza su uso en diferentes dominios de aplicación. Por otra parte, los indicadores extendidos logran medir con mayor precisión la incertidumbre combinando los grados de certeza, indeterminación y falsedad, a diferencia de los indicadores tradicionales. Además, se incorporan indicadores como Te2c y Te2d que permiten medir de forma particular la indeterminación y la falsedad respectivamente.

## Conclusiones del capítulo

En este capítulo se proponen nuevos indicadores para la evaluación de los resúmenes lingüísticos, de los cuales se concluye que:

* Los indicadores que se proponen en este capítulo extienden a los indicadores tradicionales propuestos en la bibliografía, combinando elementos de la teoría neutrosófica que posibilitan el tratamiento de la certidumbre, la indeterminación y la falsedad en la evaluación de los resúmenes.
* La introducción de conceptos como: el parámetro *-corte*, la aproximación superior y aproximación inferior, permite manejar con mayor flexibilidad el nivel de rigor en la evaluación de los resúmenes lingüísticos.
* En la medición del grado de verdadde los resúmenes lingüísticos, la extensión *Te1a* reportó un comportamiento significativamente mejor que la variante tradicional *T1* en las clases de prueba analizadas.
* El indicador *T2* es complementado con otros cuatro indicadores que ayudan a medir la imprecisión del resumen, pues se mide con mayor precisión los grados de indeterminación y falsedad a través de los indicadores *Te2c* y *Te2d* respectivamente. Por otra parte, los indicadores *Te2a* y *Te2b* brindan dos formas de medir la incertidumbre al combinar los grados de certeza, indeterminación y falsedad en una única medida.
* Respecto al grado de cobertura, se puede decir que el *Te3* extendido es más estricto que el *T3* tradicional, permite diferenciar mejor los escenarios representados por las clases de prueba.
* Así mismo, en el grado de adecuación, el *Te4* tiene una mayor diferenciación de los escenarios representados por las clases de prueba respecto al *T4*, porque promueve la identificación de relaciones peculiares entre variables, como las relaciones que no están presentes en la mayoría de los objetos de la base de datos pero que son fuertes en aquellos objetos donde aparecen. El nuevo indicador complementa al original al brindar una medida de adecuación que combina el grado de certidumbre con la indeterminación y la falsedad.
* Respecto a los indicadores para medir la longitud de un resumen, el *T5* tradicional tiene un comportamiento abrupto, de modo que cae abruptamente desde resúmenes de longitud 1 a resúmenes de longitud 3. Mientras que su propuesta de reemplazo, el *Te*5 extendido, muestra una curva más suave que favorece a los resúmenes de longitud 2 y 3 con un comportamiento más deseado.

# Capítulo 4 Validación de la eficacia de los algoritmos propuestos

En este capítulo, se aplica una estrategia de validación para la comparación de los algoritmos propuestos con algunos de los reportados en la bibliografía, aplicando métodos cuantitativos del diseño de experimentos. Para este análisis de comparación, se tienen en cuenta los indicadores para evaluar la calidad de los resúmenes lingüísticos referenciados en la bibliografía, los extendidos propuestos en el capítulo 3, y otros indicadores como el “cubrimiento” de las diferentes situaciones en la base de datos, la “fortaleza de las relaciones que descubren”[[11]](#footnote-12) y el tiempo de ejecución.

Posteriormente se analiza el desempeño global de los algoritmos, considerando el comportamiento de todos los indicadores empleados en las comparaciones. Se demuestra a través de un caso de estudio la aplicabilidad de los algoritmos propuestos. Finalmente se presentan las conclusiones del capítulo.

## Parámetros de algoritmos y bases de datos empleadas en la experimentación

En la definición de los parámetros para las corridas experimentales, se establece como principio que los parámetros que son comunes a varios algoritmos se fijan con un valor para mitigar sesgo en la experimentación. Se emplean los siguientes parámetros:

* *LVQ*: se emplean las funciones triangulares definidas en la Tabla 4 de la sección 2.2 sobre las notaciones, esta decisión se tomó para lograr mayor simplicidad en las pruebas.
* *ALV:* se construye un conjunto de variables lingüísticas por cada base de datos que es empleado por todos los algoritmos.
* *:* se fija el valor 0.33 y es aplicado por igual a todos los algoritmos.
* *percentil\_component y percentil\_elbow*: se fijan con valores 0.2 y 0.1 respectivamente. Son usados por el algoritmo *PCA\_LDS.* Se toman valores sobre lo bajo para no afectar significativamente la extensión de la búsqueda en el algoritmo durante la experimentación.
* *support* y *confidence*: son parámetros empleados en los algoritmos *PCA\_LDS* y *Apriori\_LDS*, estos valores se fijan en un valor medio.
* *k1* y *k2:* se fijan en los valores 0.9 y 0.8 respectivamente. Son empleados por el algoritmo *RST\_LDS.* Se consideran valores medianamente altos, para no restringir demasiado la extensión del proceso de búsqueda.
* *percentil*: parámetro usado en el Algoritmo 6 *Outlier\_LDS*. Se fija en 0.92 considerando los resultados reportados en (Castro Aguilar 2017).

Para la corrida de los algoritmos, se utilizaron bases de datos de tres entornos que se describen a continuación y que se muestran de forma detallada en Tabla 24 del Anexo 8.

Bases de datos asociadas a la evaluación y toma de decisiones en gestión de proyectos, tomadas del repositorio publicado en (Pérez Pupo 2021) (Piñero et al. 2019):

* *BD\_gp\_eval\_proy\_hard*: base de datos con 229 registros y 8 indicadores que representan casos reales de evaluación de proyectos en diferentes cortes.
* *BD\_gp\_eval\_proy\_integral*:base de datos con 8 430 registros y 19 atributos que representan a la evaluación integral de proyectos.
* *BD\_gp\_indicadores*: base de datos formada por 174 registros y 14 atributos asociados a la evaluación de programas de proyectos.
* *BD\_gp\_competencias*: base de datos formada por 20 574 instancias y 12 atributos que representan la evaluación de las competencias de los participantes en proyectos.
* *BD\_gp\_rrhh*: base de datos de 3 931 instancias y 9 atributos asociados a la evaluación de recursos humanos en los cortes de un proyecto.
* *BD\_gp\_requerimientos*:base de datos asociada a la gestión de requerimientos de un proyecto formada por 1 310 instancias y 7 atributos.
* *BD\_gp\_toma\_decisiones*:base de datos asociada a la toma de decisiones en gestión de proyectos y formada por 1 011 instancias y 17 atributos.

Bases de datos asociadas a procesos de auditorías suministradas por la Contraloría General de la República a partir de las comprobaciones nacionales de los años 2017 y 2018:

* + *BD\_audit\_sinOrganismos*: formada por 1 778 instancias y 16 atributos que registran daños económicos detectados en instituciones, así como el impacto del daño económico en las monedas CUP, CUC y USD y otras monedas.
  + *BD\_audit\_sinProvMinistInstit*: formada por 1 778 instancias y 16 atributos que registra daños económicos detectados en diferentes instituciones, organismos y ministerios, así como el impacto del daño económico en CUP, CUC, USD y otras monedas.

Bases de datos asociadas a toma de decisiones médicas. Las dos primeras asociadas a registros clínicos de embarazadas cardiópatas, ofrecidos por los expertos médicos (Tabla 26 del Anexo 11) que laboran en el Hospital Docente Ginecobstétrico “Ramón González Coro” y en el Instituto de Cardiología y Cirugía Cardiovascular en La Habana. La tercera base de datos es sobre casos de covid-19 en Cuba, publicado por el Ministerios de Salud Pública.

* *BD\_med\_congenito*:base de datos creada a partir de registros clínicos sobre síntomas, diagnósticos, tratamientos y pronósticos de embarazadas cardiópatas con enfermedades congénitas. Está formada por 464 instancias y 20 atributos.
* *BD\_med\_valvular*: base de datos creada a partir de registros clínicos sobre sobre síntomas, diagnósticos, tratamientos y pronósticos de embarazadas cardiópatas con enfermedades valvulares. Formada por 304 instancias y 17 atributos.
* *BD\_med\_covidCuba*: base de datos publicada por el Ministerio de Salud Pública de Cuba asociada a casos de covid-19 en Cuba que registra el sexo, la edad, el país de procedencia de los casos, municipio, provincia y tipo de contagio de algunos de los enfermos por covid-19. Incluye 18 151 instancias y 6 atributos.

## Conjunto de pruebas 2: validación de variable independiente respecto a la dimensión indicadores de calidad de los resúmenes lingüísticos.

En Tabla 30, Tabla 31 y Tabla 32 del Anexo 13 se muestran los datos utilizados en las corridas de los experimentos realizados en la validación.

### Comparación de los algoritmos respecto al indicador T1

Respecto a la media aritmética del indicador *T1*, los algoritmos quedan ubicados en el siguiente orden: GA\_LDS, PCA\_LDS, Apriori\_LDS, RST\_LDS, ACO\_LDS y LPA\_LDS.

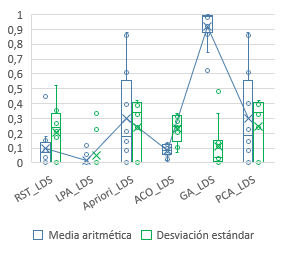


Figura 10: Media aritmética del indicador *T1*.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 8. Se comparan los algoritmos empleando pruebas no paramétricas. Para la comparación se aplicó el test no paramétrico *Wilcoxon* y se demuestra que el algoritmo GA\_LDSreporta resultados significativamente mejores que el resto de los algoritmos, está ubicadoen el“grupo A”. Luego se ubican los algoritmos PCA\_LDS y Apriori\_LDS en los grupos B y C respectivamente. No se encontraron diferencias significativas en los algoritmos RST\_LDS y ACO\_LDS, que quedaron ubicados en el “grupo D”. El algoritmo de peor resultado fue el LPA\_LDS ubicado en el “grupo E”.

La Figura 10 evidencia los resultados de las pruebas estadísticas, presentando el GA\_LDScomo el algoritmo con mejores resultados.

1. Tabla 8: Comparación de los algoritmos respecto a la media aritmética del *T1*.

|  |  |  |  |
| --- | --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |  |
| RST\_LDS | .017 | .002 |  |
| LPA\_LDS | .000 | .000 |  |
| PCA\_LDS | .200\* | .026 |  |
| ACO\_LDS | .200\* | .180 |  |
| GA\_LDS | .016 | .001 |  |
| Apriori\_LDS | .200\* | .029 |  |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test*  GA\_LDS (grupo A) –  PCA\_LDS (grupo B) | 0.004 | Hay diferencias significativas |
| *Wilcoxon test*  PCA\_LDS (grupo B) –  Apriori\_LDS (grupo C) | 0.046 | Hay diferencias significativas |
| *Wilcoxon test*  Apriori\_LDS (grupo C)-  RST\_LDS (grupo D) | 0.050 | Hay diferencias  significativas |
| *Wilcoxon test* (grupo D)  RST\_LDS - ACO\_LDS | 0.753 | No hay diferencias  Significativas |
| *Wilcoxon test*  RST\_LDS (grupo D)-  LPA\_LDS (grupo E) | 0.007 | Hay diferencias  significativas |

### Comparación de los algoritmos respecto al indicador T2

Respecto a la media aritmética del indicador *T2*, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, ACO\_LDS, GA\_LDS, PCA\_LDS, Apriori\_LDS y LPA\_LDS.

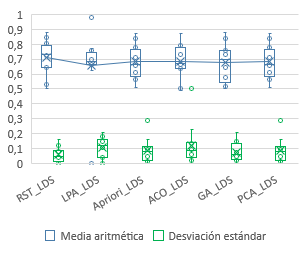


Figura 11: Media aritmética del indicador *T2*.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 9. Se comparan los algoritmos empleando pruebas no paramétricas. Para la comparación se aplicó el test no paramétrico *Friedman* y se demuestra que todos los algoritmos están ubicados en el mismo “grupo A” sin diferencias significativas entre ellos, Tabla 9. No obstante, en la Figura 11 y en los resultados de los análisis estadísticos se muestra que el RST\_LDSobtiene resultados mejores que el resto con una baja dispersión en el conjunto de las bases de datos.

1. Tabla 9: Comparación de los algoritmos respecto a la media aritmética del *T2*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | .137 | .024 |
| LPA\_LDS | .000 | .003 |
| Apriori\_LDS | .079 | .016 |
| ACO\_LDS | .108 | .213 |
| GA\_LDS | .144 | .054 |
| PCA\_LDS | .088 | .019 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Friedman test* (grupo A)  RST\_LDS, ACO\_LDS,  GA\_LDS, PCA\_LDS, Apriori\_LDS y LPA\_LDS. | 0.853 | No hay diferencias significativas |

### Comparación de los algoritmos respecto al indicador T3

Respecto a la media aritmética del indicador *T3*, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, GA\_LDS, PCA\_LDS, Apriori\_LDS, LPA\_LDS y ACO\_LDS.

Se identifica que los resultados no cumplen con una distribución normal. Para la comparación se aplicó el test no paramétrico *Wilcoxon* y se demuestra que los algoritmos RST\_LDS y GA\_LDSreportan resultados significativamente mejores que el resto de los algoritmos, están ubicadosen el“grupo A”.

1. La Figura 12 evidencia los resultados de las pruebas estadísticas, donde el algoritmo RST\_LDS obtiene los mejores resultados con la menor dispersión. Mientras que la

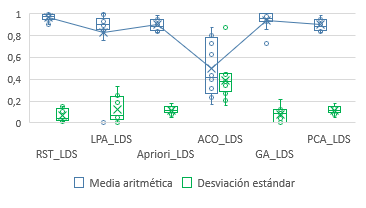


Figura 12: Media aritmética del indicador *T3*.

muestra el resultado de las pruebas no paramétricas.

Luego, sin diferencias significativas entre ellos, se ubican los algoritmos PCA\_LDS, Apriori\_LDS y LPA\_LDS en el “grupo B”. ACO\_LDS fue el algoritmo de peor resultado, ubicado en el “grupo C”.

2. Tabla 10: Comparación de los algoritmos respecto a la media aritmética del *T3*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | .200\* | .024 |
| LPA\_LDS | .000 | .000 |
| Apriori\_LDS | .119 | .080 |
| ACO\_LDS | .200\* | .100 |
| GA\_LDS | .013 | .002 |
| PCA\_LDS | .119 | .080 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test* (grupo A)  RST\_LDS - GA\_LDS | 0.423 | No hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo A) – PCA\_LDS (grupo B) | 0.004 | Hay diferencias significativas |
| *Wilcoxon test* (grupo B)  PCA\_LDS - Apriori\_LDS  PCA\_LDS - LPA\_LDS | 1.000  1.000 | No hay diferencias significativas |
| *Wilcoxon test*  PCA\_LDS (grupo B) – ACO\_LDS (grupo C) | 0.002 | Hay diferencias significativas |

### Comparación de los algoritmos respecto al indicador T4

Respecto a la media aritmética del indicador *T4*, los algoritmos quedan ubicados en el siguiente orden: GA\_LDS, LPA\_LDS, RST\_LDS, ACO\_LDS, Apriori\_LDS y PCA\_LDS. En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 11. La Figura 13 evidencia los resultados de las pruebas estadísticas.

1. Tabla 11: Comparación de los algoritmos respecto a la media aritmética del *T4*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | | Shapiro-Wilk Sig. | |
| RST\_LDS | | .200\* | | .609 |
| LPA\_LDS | | .133 | | .007 |
| Apriori\_LDS | | .010 | | .000 |
| ACO\_LDS | | .097 | | .029 |
| GA\_LDS | | .100 | | .033 |
| PCA\_LDS | | .011 | | .000 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test* (grupo A)  GA\_LDS - LPA\_LDS | 0.130 | No hay diferencias significativas |
| *Wilcoxon test*  GA\_LDS (grupo A) - RST\_LDS (grupo B) | 0.005 | Hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo B) - ACO\_LDS (grupo C) | 0.021 | Hay diferencias significativas |
| *Wilcoxon test*  ACO\_LDS (grupo C) - Apriori\_LDS (grupo D) | 0.013 | Hay diferencias significativas |
| *Wilcoxon test* (grupo D)  Apriori\_LDS - PCA\_LDS | 0.317 | No hay diferencias significativas |

Para la comparación se aplicó el test no paramétrico *Wilcoxon* y se demuestra que los algoritmos GA\_LDS y LPA\_LDSreportan resultados significativamente mejores que el resto de los algoritmos, están ubicadosen el“grupo A”. Luego se ubican los algoritmos RST\_LDS y ACO\_LDS en los grupos B y C respectivamente. No se encontraron diferencias significativas en los algoritmos Apriori\_LDS y PCA\_LDS, quedaron ubicados en el “grupo D” reportando peores resultados.

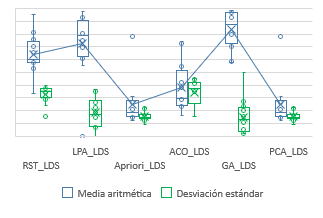
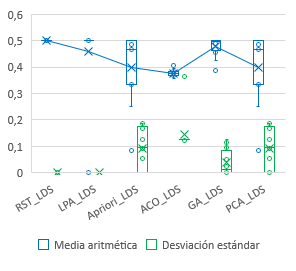


Figura 13: Media aritmética del indicador *T4*.

### Comparación de los algoritmos respecto al indicador T5

Respecto a la media aritmética del indicador *T5*, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, GA\_LDS, LPA\_LDS, PCA\_LDS, Apriori\_LDS y ACO\_LDS.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 12. Para la comparación se aplicó el test no paramétrico *Wilcoxon* y se demuestra que los algoritmos RST\_LDS, GA\_LDS y LPA\_LDS reportan resultados significativamente mejores, están bicadosen el“grupo A”. Posteriormente, no se encontraron diferencias significativas en los algoritmos PCA\_LDS, Apriori\_LDS y ACO\_LDS, los cuales quedaron en el “grupo B” reportando peores resultados.

Figura 14: Comportamiento del indicador *T5*

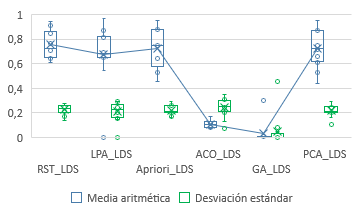
La Figura 14 evidencia los resultados de las pruebas estadísticas, presentando el RST\_LDScomo el algoritmo con mejores resultados y muy baja dispersión.

1. Tabla 12: Comparación de los algoritmos respecto a la media aritmética del *T5*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | . | . |
| LPA\_LDS | .000 | .000 |
| Apriori\_LDS | .006 | .007 |
| ACO\_LDS | .037 | .170 |
| GA\_LDS | .000 | .000 |
| PCA\_LDS | .006 | .007 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test* (grupo A)  RST\_LDS - GA\_LDS  RST\_LDS - LPA\_LDS | 0.068  0.317 | No hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo A) -  PCA\_LDS (grupo B) | 0.012 | Hay diferencias significativas |
| *Wilcoxon test* (grupo B)  PCA\_LDS - Apriori\_LDS  PCA\_LDS - ACO\_LDS | 1.000  0.345 | No hay diferencias significativas |

### Comparación de los algoritmos respecto al indicador Te1a

Respecto a la media aritmética del indicador *Te1a*, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, Apriori\_LDS, PCA\_LDS, LPA\_LDS, ACO\_LDS y GA\_LDS.

La Figura 15 evidencia los resultados de las pruebas a través de un diagrama de cajas y bigotes mientras que la

Figura 15: Comportamiento del indicador *Te1a*

muestra los resultados de las comparaciones de los algoritmos empleando pruebas no paramétricas.

Tabla 13: Comparación de los algoritmos respecto a la media aritmética del *Te1a*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | .200\* | .318 |
| LPA\_LDS | .005 | .006 |
| Apriori\_LDS | .200\* | .758 |
| ACO\_LDS | .200\* | .444 |
| GA\_LDS | .000 | .000 |
| PCA\_LDS | .200\* | .908 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Friedman test* (grupo A)  RST\_LDS, Apriori\_LDS, PCA\_LDS | 0.433 | No hay diferencias significativas |
| *Wilcoxon test* (grupo A)  RST\_LDS - LPA\_LDS | 0.109 | No hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo A) –  ACO\_LDS (grupo B) | 0.002 | Hay diferencias significativas |
| *Wilcoxon test*  ACO\_LDS (grupo B) –  GA\_LDS (grupo C) | 0.034 | Hay diferencias significativas |

En el análisis se identifica que los resultados no cumplen con una distribución normal. En la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que no existen diferencias significativas entre los algoritmos RST\_LDS, Apriori\_LDS, PCA\_LDS y LPA\_LDS, ubicados en el “grupo A” y reportando resultados significativamente mejores al resto. Luego, en el “grupo B” está ubicado el algoritmo ACO\_LDS y en el “grupo C” el GA\_LDS reportando peores resultados.

### Comparación de los algoritmos respecto al indicador Te1b

Respecto a la media aritmética del indicador *Te1b*, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, Apriori\_LDS, GA\_LDS, PCA\_LDS, LPA\_LDS y ACO\_LDS.

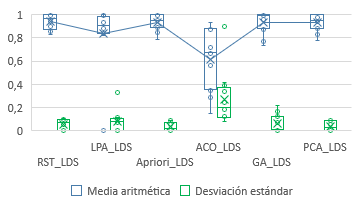


Figura 16: Media aritmética del indicador *Te1b*.

En el análisis se identifica que los resultados no cumplen con una distribución normal.

Tabla 14: Comparación de los algoritmos respecto a la media aritmética del *Te1b*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | .031 | .020 |
| LPA\_LDS | .000 | .000 |
| Apriori\_LDS | .040 | .051 |
| ACO\_LDS | .200\* | .206 |
| GA\_LDS | .003 | .002 |
| PCA\_LDS | .143 | .057 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Friedman test* (grupo A)  RST\_LDS, Apriori\_LDS  GA \_LDS, PCA\_LDS, LPA\_LDS | 0.720 | No hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo A) - ACO\_LDS (grupo B) | 0.002 | Hay diferencias significativas |

En la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que no existen diferencias significativas entre los algoritmos RST\_LDS, Apriori\_LDS, GA\_LDS, PCA\_LDS y LPA\_LDS ubicados en el “grupo A” y reportando resultados significativamente mejores al ACO\_LDS, el cual está ubicado en el “grupo B” reportando peores resultados.

La Figura 16 evidencia los resultados de las pruebas estadísticas, presentando al RST\_LDScomo el algoritmo de mejores resultados con el mayor valor de la media, aunque el Apriori\_LDS obtuvo la menor dispersión. Mientras que la Tabla 14 muestra los resultados de comparación de los algoritmos empleando pruebas paramétricas.

### Comparación de los algoritmos respecto al indicador Te3

Respecto a la media aritmética del indicador *Te3*, los algoritmos quedan ubicados en el siguiente orden: PCA\_LDS, Apriori\_LDS, RST\_LDS, ACO\_LDS, LPA\_LDS y GA\_LDS.

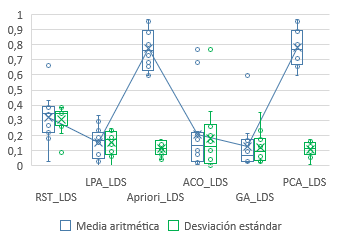
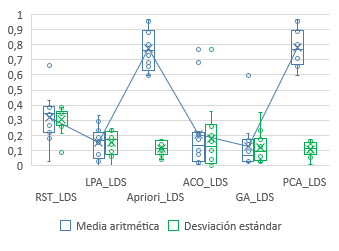


Figura 17: Media aritmética del indicador *Te3*.

En el análisis se identifica que los resultados no cumplen con una distribución normal.

Se comparan los algoritmos empleando pruebas no paramétricas. Ver

y 

.

Tabla 15: Comparación de los algoritmos respecto a la media aritmética del *Te3*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | .200\* | .688 |
| LPA\_LDS | .200\* | .664 |
| Apriori\_LDS | .200\* | .120 |
| ACO\_LDS | .002 | .002 |
| GA\_LDS | .037 | .001 |
| PCA\_LDS | .200\* | .243 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test*  PCA\_LDS (grupo A) - Apriori\_LDS (grupo B) | 0.026 | Hay diferencias significativas |
| *Wilcoxon test*  Apriori\_LDS (grupo B) –  RST\_LDS (grupo C) | 0.002 | Hay diferencias significativas |
| *Wilcoxon test* (grupo C)  RST\_LDS – ACO\_LDS | 0.182 | No hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo C) –  LPA\_LDS (grupo D) | 0.002 | Hay diferencias significativas |
| *Wilcoxon test* (grupo D)  LPA\_LDS – GA\_LDS | 0.169 | No hay diferencias significativas |

Para la comparación se aplicó el test no paramétrico *Wilcoxon* y se demuestra que el algoritmo PCA\_LDSreporta resultados significativamente mejores que el resto de los algoritmos, ubicadoen el“grupo A” y seguido por el algoritmo Apriori\_LDS ubicado en un grupo B. Luego se ubican los algoritmos RST\_LDS y ACO\_LDS en un grupo C sin diferencias significativas entre ellos, así como LPA\_LDS y GA\_LDS en el grupo D reportando peores resultados.

### Comparación de los algoritmos respecto al indicador Te4

Respecto a la media aritmética del indicador *Te4*, los algoritmos quedan ubicados en el siguiente orden: PCA\_LDS, Apriori\_LDS, GA\_LDS, RST\_LDS, LPA\_LDS y ACO\_LDS.

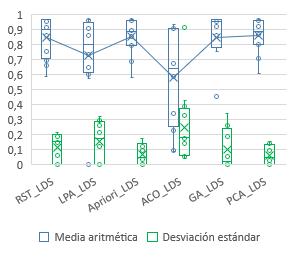
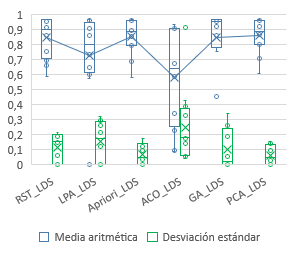


Figura 18: Media aritmética del indicador *Te4*.

La 

evidencia los resultados de las pruebas estadísticas, presentando el PCA\_LDS y Apriori\_LDS como los algoritmos con mejores resultados y más baja dispersión.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 16. En la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que no existen diferencias significativas entre los algoritmos PCA\_LDS, Apriori\_LDS, GA\_LDS, RST\_LDS y LPA\_LDS ubicados en el “grupo A” y reportando resultados significativamente mejores al ACO\_LDS, el cual está ubicado en el “grupo B” reportando peores resultados.

1. Tabla 16: Comparación de los algoritmos respecto a la media aritmética del *Te4*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | .186 | .028 |
| LPA\_LDS | .200\* | .009 |
| Apriori\_LDS | .200\* | .052 |
| ACO\_LDS | .200\* | .079 |
| GA\_LDS | .002 | .000 |
| PCA\_LDS | .200\* | .064 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Friedman test* (grupo A)  PCA\_LDS, Apriori\_LDS, GA\_LDS, RST\_LDS, LPA\_LDS | 0.767 | No hay diferencias significativas |
| *Wilcoxon test*  PCA\_LDS (grupo A) –  ACO\_LDS (grupo B) | 0.002 | Hay diferencias significativas |

### Comparación de los algoritmos respecto al indicador Te5

Respecto a la media aritmética del indicador *Te5*, los algoritmos quedan ubicados en el siguiente orden: GA\_LDS, ACO\_LDS, RST\_LDS, LPA\_LDS, PCA\_LDS y Apriori\_LDS.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 17. En la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que no existen diferencias significativas entre los algoritmos GA\_LDS, ACO\_LDS, RST\_LDS y LPA\_LDS ubicados en el “grupo A” y reportando resultados significativamente mejores. Mientras los algoritmos PCA\_LDS y Apriori\_LDS están ubicados en el “grupo B” sin diferencias significativas entre ellos, reportando peores resultados.

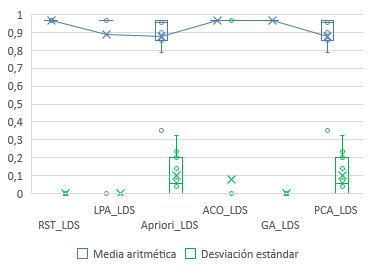


Figura 19: Media aritmética del indicador *Te5*.

La Figura 19 evidencia los resultados de las pruebas estadísticas.

1. Tabla 17: Comparación de los algoritmos respecto a la media aritmética del *Te5*.

|  |  |  |
| --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | Shapiro-Wilk Sig. |
| RST\_LDS | . | . |
| LPA\_LDS | .000 | .000 |
| Apriori\_LDS | .005 | .000 |
| ACO\_LDS | . | . |
| GA\_LDS | . | . |
| PCA\_LDS | .005 | .000 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Friedman test* (grupo A)  GA\_LDS, ACO\_LDS, RST\_LDS | 1.000 | No hay diferencias significativas |
| *Wilcoxon test* (grupo A)  GA\_LDS – LPA\_LDS | 0.317 | No hay diferencias significativas |
| *Wilcoxon test*  GA\_LDS (grupo A) –  PCA\_LDS (grupo B) | 0.018 | Hay diferencias significativas |
| *Wilcoxon test* (grupo B)  PCA\_LDS – Apriori\_LDS | 1.000 | No hay diferencias significativas |

### Comparación de los algoritmos respecto al indicador “Cubrimiento de diferentes situaciones en la base de datos”.

En diferentes problemas reales es importante poder identificar la frecuencia de aparición de determinados fenómenos o situaciones en las bases de datos. En esta investigación, se introduce el indicador “Cubrimiento de diferentes situaciones en la base de datos” que mide la capacidad de los algoritmos para generar resúmenes lingüísticos que cubran diferentes fenómenos en los datos. Este indicador constituye una medida de versatilidad de los algoritmos para su aplicación en diferentes contextos.

Para el cálculo de este indicador se analiza la frecuencia relativa de aparición de las siguientes situaciones:

* Muy pocos registros en la base de datos.
* Pocos registros en la base de datos.
* Algunos registros en la base de datos.
* Aproximadamente la mitad de los registros en la base de datos.
* Muchos de los registros en la base de datos.
* La mayoría de los registros en la base de datos.

Luego, por cada base de datos se calcula la dispersión de las frecuencias relativas calculadas empleando la desviación estándar. Lo deseado en este caso es que haya un balance en la cantidad de resúmenes generados, cubriendo cada una de las diferentes situaciones descritas. Los algoritmos con mejores resultados serán los que tengan menor dispersión.

1. Tabla 18: Comparación de los algoritmos respecto a la dispersión.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | | Shapiro-Wilk Sig. | |
| RST\_LDS | | .012 | | .014 |
| LPA\_LDS | | .049 | | .347 |
| Apriori\_LDS | | .200\* | | .977 |
| ACO\_LDS | | .000 | | .000 |
| GA\_LDS | | .001 | | .001 |
| PCA\_LDS | | .200\* | | .522 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test*  ACO\_LDS (grupo A) –  RST\_LDS (grupo B) | 0.028 | Hay diferencias significativas |
| *Wilcoxon test*  RST\_LDS (grupo B) – LPA\_LDS (grupo C) | 0.003 | Hay diferencias significativas |
| *Wilcoxon test* (grupo C)  LPA\_LDS - Apriori\_LDS  LPA\_LDS - PCA\_LDS | 0.307  0.239 | No hay diferencias significativas |
| *Wilcoxon test*  LPA\_LDS (grupo C) – GA\_LDS (grupo D) | 0.003 | Hay diferencias significativas |

En las pruebas estadísticas, respecto a este indicador, los algoritmos quedan ubicados en el siguiente orden: ACO\_LDS, RST\_LDS, LPA\_LDS, Apriori\_LDS, PCA\_LDS y GA\_LDS. Los datos de estas corridas se muestran en la Tabla 34 del Anexo 13.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 18. En la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que existen diferencias significativas entre ellos, donde el algoritmo ACO\_LDS está ubicado en el “grupo A” reportando mejores resultados, seguido por el RST\_LDS en el “grupo B”. En el “grupo C” y sin diferencias significativas entre ellos están LPA\_LDS, Apriori\_LDS y PCA\_LDS. El algoritmo GA\_LDS quedó ubicado en el “grupo D” reportando los peores resultados.

### Comparación de los algoritmos respecto a la “Fortaleza de las dependencias descubiertas”.

En la publicación de Donis-Días y colaboradores (Donis-Díaz, Bello y Kacprzyk 2015), se emplea el indicador “número de proposiciones con cuantificador `mayoría´ y `muchos´”. En esta sección se utiliza este indicador para complementar la comparación de los algoritmos propuestos con otros trabajos reportados en la bibliografía. Para mayor comprensión denominamos a este indicador “Fortaleza de las dependencias descubiertas”.

El indicador “Fortaleza de las dependencias descubiertas” mide la capacidad de los algoritmos para la detección de resúmenes con alta fortaleza en la relación filtro-sumarizador. Se mide este indicador a partir del análisis de las frecuencias de aparición de los cuantificadores que indiquen fuertes relaciones de dependencia, tales como los cuantificadores “la mayoría”, “casi la totalidad” y “muchos”. Para el cálculo de este indicador se aplicaron los siguientes pasos:

1. Por cada base de datos se calcula la frecuencia relativa de aparición de cada cuantificador en los resúmenes lingüísticos que genera cada algoritmo.
2. Por cada base de datos se calcula la suma pesada de la frecuencia relativa utilizando los siguientes pesos: 0 a los cuantificadores “muy pocos”, “pocos” y “algunos”; mientras que a los cuantificadores “aproximadamente la mitad”, “muchos”, “la mayoría” y “casi la totalidad” le fueron asignados los pesos 0.08, 0.1, 0.4 y 0.42 respectivamente.

Los datos de estas corridas se muestran en la Tabla 35 del Anexo 13, mientras que los resultados de las pruebas estadísticas se muestran en la Tabla 19.

1. Tabla 19: Comparación de los algoritmos respecto a la fortaleza de las dependencias descubiertas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | | Shapiro-Wilk Sig. | |
| RST\_LDS | | . | | . |
| LPA\_LDS | | .001 | | .000 |
| Apriori\_LDS | | .200\* | | .977 |
| ACO\_LDS | | .200\* | | .603 |
| GA\_LDS | | .012 | | .000 |
| PCA\_LDS | | .200\* | | .977 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test*  RST\_LDS (grupo A) –  LPA\_LDS (grupo B) | 0.014 | Hay diferencias significativas |
| *Wilcoxon test*  LPA\_LDS (grupo B) – GA\_LDS (grupo C) | 0.024 | Hay diferencias significativas |
| *Wilcoxon test*  GA\_LDS (grupo C) – PCA\_LDS (grupo D) | 0.005 | Hay diferencias significativas |
| *Wilcoxon test* (grupo D)  PCA\_LDS - Apriori\_LDS | 1.000 | No hay diferencias significativas |
| *Wilcoxon test*  PCA\_LDS (grupo D) – ACO\_LDS (grupo E) | 0.008 | Hay diferencias significativas |

Respecto a este indicador, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, LPA\_LDS, GA\_LDS, PCA\_LDS, Apriori\_LDS y ACO\_LDS. En el análisis se identifica que los resultados no cumplen con una distribución normal. Por tanto, en la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que existen diferencias significativas entre ellos, donde el algoritmo RST\_LDS está ubicado en el “grupo A” reportando mejores resultados, seguido por el LPA\_LDS en el “grupo B” y GA\_LDS en el “grupo C”. En el “grupo D” y sin diferencias significativas entre ellos están PCA\_LDS y Apriori\_LDS. El algoritmo ACO\_LDS está ubicado en el “grupo E” reportando los peores resultados.

## Conjunto de pruebas 3: validación de la variable independiente en la dimensión “Eficiencia de los algoritmos”.

En las corridas de experimentación, todos los algoritmos se ejecutaron en la misma computadora, con las siguientes características: I5, 4 núcleos y 8 GB RAM; de esta forma se evitan posibles sesgos en la medición por el uso de diferentes tecnologías.

Respecto al indicador tiempo de ejecución, los algoritmos quedan ubicados en el siguiente orden: RST\_LDS, PCA\_LDS, Apriori\_LDS, LPA\_LDS, GA\_LDS y ACO\_LDS. Los datos de estas corridas se muestran en la Tabla 33 del Anexo 13.

En el análisis se identifica que los resultados no cumplen con una distribución normal, ver Tabla 20.

1. Tabla 20: Comparación de los algoritmos respecto al tiempo de ejecución.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmos | Kolmogorov-Smirnova Sig. | | Shapiro-Wilk Sig. | |
| RST\_LDS | | .017 | | .000 |
| LPA\_LDS | | .000 | | .000 |
| Apriori\_LDS | | .044 | | .002 |
| ACO\_LDS | | .006 | | .003 |
| GA\_LDS | | .021 | | .002 |
| PCA\_LDS | | .024 | | .010 |

|  |  |  |
| --- | --- | --- |
| Algoritmos | p-value | Resultados |
| *Wilcoxon test*  RST\_LDS (grupo A) –  PCA\_LDS (grupo B) | 0.002 | Hay diferencias significativas |
| *Wilcoxon test* (grupo B)  PCA\_LDS - Apriori\_LDS  PCA\_LDS - LPA\_LDS | 0.158  0.814 | No hay diferencias significativas |
| *Wilcoxon test*  PCA\_LDS (grupo B) –  GA\_LDS (grupo C) | 0.004 | Hay diferencias significativas |
| *Wilcoxon test*  GA\_LDS (grupo C) –  ACO\_LDS (grupo D) | 0.002 | Hay diferencias significativas |

En la comparación de los algoritmos, se aplicaron pruebas no paramétricas y se demuestra que existen diferencias significativas entre ellos, donde el algoritmo RST\_LDS está ubicado en el “grupo A” reportando mejores tiempos de ejecución. Los algoritmos PCA\_LDS, Apriori\_LDS y LPA\_LDS sin diferencias significativas entre ellos en el “grupo B”. Posteriormente el GA\_LDS en el “grupo C” y el ACO\_LDS en el “grupo D” reportando los peores resultados.

En la Tabla 21 se muestra la complejidad computacional de los algoritmos propuestos en la investigación y de los reportados en la bibliografía utilizados en la validación. En ella se observa que los algoritmos propuestos mejoran la complejidad computacional de los reportados en la bibliografía. Existe una correspondencia entre la complejidad computacional de los algoritmos con el tiempo de ejecución.

Tabla 21: Complejidad de los algoritmos.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algoritmos | | Complejidad | | Media T. (Ɐ BD) | | Donde:  *n:* tamaño de la base de datos*.*  *p:* cantidad de variables*.*  *k:* tamaño medio de los resúmenes, *k<p.*  *g:* generaciones del algoritmo*.*  *d:* número de individuos.  *z:* cantidad de resúmenes por individuos*.* |
| RST\_LDS (A) | *O(n2·p)* | | 2,11 | |  | |
| PCA\_LDS (B) | *O(max(p·2p+1, n2·p))* | | 43,99 | |  | |
| Apriori\_LDS (B) | *O(2p+1)* | | 45,94 | |  | |
| LPA\_LDS (B) | *O(n3)* | | 52,76 | |  | |
| GA\_LDS (C) | *O(kgnd)* | | 1212,62 | |  | |
| ACO\_LDS (D) | *O(zgnd2)* | | 19692,03 | |  | |

Con el análisis anterior, se demuestra que los algoritmos propuestos mejoran la eficiencia respecto a los reportados en la bibliografía empleados en la experimentación. El algoritmo que mostró una mejor eficiencia fue RST\_LDS, mientras que los algoritmos basados en metaheurísticas son los menos eficientes.

## Conjunto de pruebas 4: validación de la variable independiente en la dimensión “Desempeño global de los algoritmos”

En esta sección se analizan todos los algoritmos a partir de su comportamiento global, considerando todos los indicadores empleados en las comparaciones de las secciones anteriores. Como parte del análisis del desempeño global de los algoritmos, se utilizó el método *Page’s L Test* *(**B**a**rtz-Beielstein et al. 2020),* el cual demuestra, por varias vías, la posición o ranking de estos algoritmos. A continuación, se explica este procesamiento.

**Procedimiento para la aplicación del *Page’s L Test*:**

1. Selección del criterio para el establecimiento del ranking, en nuestro caso son los indicadores *T1*, *T2*, *T3*, *T4*, *T5*, *Te1a*, *Te1b*, *Te3*, *Te4*, *Te5*, el “Cubrimiento de diferentes situaciones en la base de datos”, la “Fortaleza de las dependencias descubiertas” y el tiempo de ejecución.
2. Asignación de los valores del ranking, de forma tal que se asigna la posición a cada algoritmo respecto a su resultado en cada una de las pruebas de comparación realizadas anteriormente considerando los elementos siguientes:
   1. Por cada indicador, a cada algoritmo se le asigna un valor de ranking que se corresponde con la posición donde quedó ubicado según cada indicador. El mejor algoritmo queda ubicado en el primer lugar y luego se ubica el resto según los resultados.
   2. Si algoritmos, respecto a un mismo indicador, quedaron ubicados en el mismo lugar, se les asignan a todos el mismo valor correspondiente al punto medio de las posiciones que ocuparían si tuvieran diferencias significativas. Por ejemplo, si dos algoritmos están en el primer lugar respecto a un determinado indicador, se le asigna a cada uno el valor de 1.5.
3. Ordenamiento de las muestras a partir de las sumas de los rankings.
4. Aplicación del test no paramétrico *Page’s L Test.* En todos los casos, el nivel de significación estadística es de 0.05, lo que significa que la hipótesis nula se rechazará siempre que el resultado sea , en este paso se emplea laherramienta *Statex* *(**M**a**ndel 2020)*.

**Aplicación del *Page’s L Test*:**

Para la aplicación del *Page’s L Test*, se utilizaron como datos de entrada el resumen global de la posición de cada algoritmo respecto a cada indicador analizado en las comparaciones que se presenta en el Anexo 14, Tabla 36.

Hipótesis nula

. Hipótesis alternativa

Los resultados mostraron que sí hay diferencias significativas: ,, , por tanto, se rechaza la hipótesis nula y se demuestra que los datos cumplen con el orden .

Como parte de la confirmación de estos resultados, la propia herramienta informática *Statex* aplica las técnicas de Chi-Cuadrado y la simulación de Monte Carlo, generando los siguientes resultados.

Resultados del Chi-Cuadrado[[12]](#footnote-13):

*,.* Se rechaza la hipótesis nula y se demuestra que los datos cumplen con el orden .

Se demuestra que, respecto a los indicadores analizados,los algoritmos mantienen el siguiente ranking:

|  |  |  |
| --- | --- | --- |
| * 1. RST\_LDS   2. GA\_LDS | * 1. PCA\_LDS   2. Apriori\_LDS | * 1. LPA\_LDS   2. ACO\_LDS |

En el análisis global se concluye que el algoritmo propuesto RST\_LDS obtiene los mejores resultados mientras que el algoritmo ACO\_LDS es el que obtiene la evaluación más baja. Respecto a la comparación entre los algoritmos PCA\_LDS y Apriori\_LDS se demuestra que la extensión propuesta PCA\_LDS obtiene mejores resultados. Los peores resultados globales son los reportados por LPA\_LDS y ACO\_LDS.

## Conjunto de pruebas 5: validación de variable dependiente “Eficacia de los algoritmos para la ayuda a la toma de decisiones”

En esta sección se demuestra la variable dependiente en la dimensión “Aplicabilidad en la toma de decisiones”. Para validar el indicador “Aplicabilidad en diferentes escenarios”, de esta dimensión, se aplican los algoritmos propuesto en los casos de estudio: `toma de decisiones en gestión de proyectos´ y `toma de decisiones en el tratamiento médico de embarazadas cardiópatas´.

Para esta prueba, los algoritmos fueron integrados al subsistema “Análisis de datos (*Prodanalysis*)” del ecosistema de software Cuadro de Mando para la Toma de Decisiones (Piñero Pérez, Pérez Pupo y Piñero Cruz 2019), donde también se implementó el meta-algoritmo *Hybrid\_LDS* que aplica los algoritmos propuestos (*RST\_LDS*, *LPA\_LDS* y *PCA\_LDS*) para su uso de forma combinada. Con este meta-algoritmo se aprovechan las potencialidades de cada algoritmo generando un único conjunto de resúmenes lingüísticos sin repetición, y en cada caso de estudio se muestran elementos que reflejan su complementariedad. Finalmente, en ambos casos de estudio se aplica el siguiente procedimiento.

**Procedimiento** 1

1. Se aplica el *Hybrid\_LDS* en las bases de datos de cada caso de estudio.
2. Los resúmenes lingüísticos obtenidos, son sometidos a la evaluación de expertos en el área de conocimiento según el caso de estudio, aplicando la técnica de análisis multicriterio (Zhang, S. et al. 2020). En ese sentido se evalúa la variable dependiente en la dimensión “Aplicabilidad en la toma de decisiones” respecto a los indicadores: “Interpretabilidad de los resúmenes” y “Facilidad para la toma de decisiones”.
   1. En cada caso de estudio se aplica una encuesta que recoge las preferencias de los expertos respecto a un conjunto de criterios.
   2. Para la agregación de las preferencias de los expertos se aplica la técnica de computación con palabras 2-tuplas (Herrera y Martínez 2000) (Zhang, S. et al. 2020) con el siguiente conjunto de términos lingüísticos *LBTL = {Muy bajo, Bajo, Medio, Alto, Muy alto}.* Se decide aplicar esta técnica porque no tiene pérdida de información, a diferencia de otras técnicas de computación con palabras como escala de ordinales y el principio de extensión (Tong y Bonissone 1980) (Bonissone y Decker 1986).
   3. Se analizan los resultados considerando las preferencias agregadas de los expertos y su concordancia en la evaluación.

### Caso de estudio en entorno clínico: embarazadas cardiópatas.

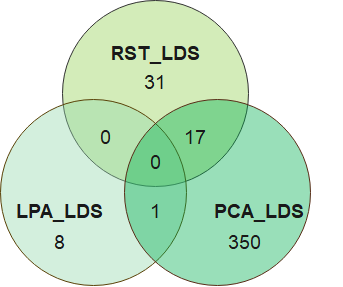
Se muestra a continuación la aplicación de los algoritmos propuestos en el diagnóstico y tratamiento de embarazadas cardiópatas. Se escoge este caso de estudio porque se dispone de datos del programa nacional de tratamientos para embarazadas cardiópatas, provistos por el Hospital Docente Ginecobstétrico “Ramón González Coro” y el Instituto de Cardiología y Cirugía Cardiovascular en La Habana, y porque también se cuenta con la voluntad de los especialistas de estos centros para colaborar con esta investigación. Se trabajó con las bases de datos *BD\_med\_congenito* y *BD\_med\_valvular*, asociadas al diagnóstico y tratamiento de las embarazadas cardiópatas con enfermedades congénitas y enfermedades valvulares respectivamente.

Figura 20: Complementariedad de los algoritmos en el caso de estudio sobre embarazadas cardiópatas.

Al aplicar el paso 1 del Procedimiento , se evidencia la complementariedad de los algoritmos en el caso de estudio respecto a la coincidencia y exclusividad de los resúmenes lingüísticos obtenidos por los algoritmos combinados en el *Hybrid\_LDS* en la base de datos *BD\_med\_congenito*, ver Figura 20. A continuación, se muestra un subconjunto de resúmenes obtenidos luego de aplicar el *Hybrid\_LDS* en estas bases de datos.

Subconjunto de resúmenes sobre embarazadas cardiópatas con enfermedades congénitas:

* Muy pocos registros en la base de datos, reportan que en el 100.0% de las veces, las embarazadas cardiópatas que tienen como tratamiento aplicar cierre percutáneo tienen como tratamiento no aplicar cuartoplastia.
* La mayoría de los registros en la base de datos, reportan que en el 96.0% de las veces, las embarazadas cardiópatas que tienen ausencia de fracción de eyección del ventrículo izquierdo (FEVI) tienen ausencia de soplo sistólico tricúspide.
* Casi la totalidad de los registros en la base de datos, reportan que en el 96.0% de las veces, las embarazadas cardiópatas que tienen ausencia de cianosis y ausencia de soplo sistólico aórtico tienen ausencia de soplo sistólico tricúspide.

Subconjunto de resúmenes sobre embarazadas cardiópatas con enfermedades valvulares:

* Algunos registros en la base de datos, reportan que en el 100.0% de las veces, las embarazadas cardiópatas que tienen presencia de primer ruido cardíaco fuerte tienen como tratamiento no aplicar heparina.
* Casi la totalidad de los registros en la base de datos, reportan que en el 100.0% de las veces, las embarazadas cardiópatas que tienen presencia de soplo sistólico aórtico y ausencia de soplo diastólico aórtico tienen como pronóstico parto vaginal.
* Muy pocos registros en la base de datos, reportan que en el 92.0% de las veces, las embarazadas cardiópatas que tienen prolapso valvular mitral como diagnóstico tienen como pronóstico parto vaginal.
* Muy pocos registros en la base de datos, reportan que en el 100.0% de las veces, las embarazadas cardiópatas que tienen doble lesión aórtica como diagnóstico tienen como tratamiento no aplicar heparina.

En la aplicación del método de análisis multicriterio (ver Procedimiento ), participaron 4 especialistas en el tratamiento de embarazadas cardiópatas, ver Anexo 11 Tabla 26. Estos especialistas completaron el formulario de evaluación recogido en la Tabla 28 del Anexo 12 donde evalúan los resúmenes respecto a 10 criterios. Los criterios C3 y C6 están orientados a medir el indicador “interpretabilidad de resúmenes” mientras que el resto de los criterios están orientados a evaluar la “facilidad para la toma de decisiones”.

A partir del análisis se obtienen los resultados de la Figura 21, donde se muestran las preferencias agregadas de los expertos y los resultados de la Figura 22, que muestra la dispersión en las preferencias. En general se obtienen los siguientes resultados:

Figura 21: Resultados de la agregación de las preferencias de los especialistas médicos sobre embarazadas cardiópatas.

* Respecto a la “interpretabilidad de los resúmenes” se obtuvo una alta evaluación con una baja dispersión en las opiniones.
* Respecto a la “facilidad para la toma de decisiones”, la mayoría de los criterios fueron evaluados con muy altos valores y a su vez presentan baja dispersión en las opiniones; excepto los criterios C4 y C7, que aunque fueron evaluados con altas calificaciones, presentan mayor dispersión respecto al resto, como se muestra en la Figura 22.

Figura 22: Dispersión de opiniones de especialistas por criterio, en el caso de estudio sobre embarazadas cardiópatas.

Se demuestra que los resúmenes generados por los algoritmos propuestos son fácilmente interpretables por los expertos y útiles para la toma de decisiones.

También se aplicó el coeficiente de correlación *RWG* (Benavente Reche 2009) para el análisis de la concordancia entre los expertos, cuyo valor fue *RWG* = 0.97, demostrando que existe concordancia entre ellos.

### Caso de estudio en entorno orientado a proyectos.

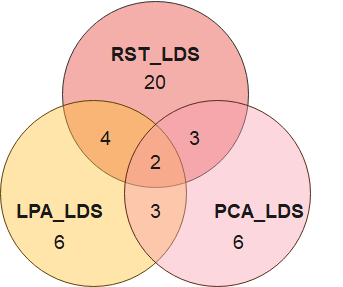
Se muestra a continuación la aplicación de los algoritmos propuestos, en el

Figura 23: Complementariedad de los algoritmos en el caso de estudio sobre proyectos.

* Centro de Consultoría y Desarrollo de Arquitecturas Empresariales de la UCI. Se escoge este caso de estudio porque en él se dispone de suficientes datos recogidos en el sistema de información para la gestión de proyectos GESPRO (López 2017) (Pérez Pupo, García Vacacela, Piñero Pérez, et al. 2020) entre los años 2010 – 2020, y porque también se cuenta con la colaboración de los expertos en esta área de conocimiento. Se trabajó con datos asociados a la evaluación, seguimiento y control de proyectos.

Al aplicar el paso 1 del Procedimiento , se evidencia la complementariedad de los algoritmos en el caso de estudio respecto a la coincidencia y exclusividad de los resúmenes lingüísticos obtenidos por los algoritmos combinados en el *Hybrid\_LDS* con datos asociados a la evaluación de proyectos, ver Figura 23.

En este escenario se aplicaron los algoritmos para analizar tres elementos fundamentales: la relación que existe entre los rasgos de personalidad de los recursos humanos y su desempeño en los resultados de los proyectos, la toma de decisiones durante los cortes de proyectos, y la detección de errores durante los procesos de planificación (Pérez Pupo, Piñero Pérez, Bello Pérez, et al. 2020).

#### Rasgos de personalidad y desempeño de los recursos humanos

Para evaluar los rasgos de personalidad de los recursos humanos y su relación con el desempeño en diferentes roles de proyectos, se incorporaron nuevos datos a los ya recogidos en el sistema de información GESPRO. Con este fin se aplicaron los instrumentos: CDE (López 2017), 16PF (Bahner y Clark 2020), BigFive (Caprara et al. 1993), estos son instrumentos validados y ampliamente empleados en numerosas investigaciones. Se muestra a continuación algunos de los resúmenes lingüísticos obtenidos a partir de unir los datos del sistema de información con los datos provenientes de cada tipo de cuestionario (Pérez Pupo, García Vacacela, Piñero Pérez, et al. 2020) (Pérez et al. 2018).

Ejemplos de resúmenes lingüísticos obtenidos combinando los datos del sistema de información con los datos del “Cuestionario sobre estilos de dirección (CED)”:

* La mayoría de los especialistas que tienen rol de programador y alto rendimiento y trabajan en condiciones de tensión se cumple que son pasivos.
* Casi todos los especialistas que tienen rol de analista en segundo nivel y rendimiento medio y trabajan en condiciones normales se cumple que son orientados hacia las personas y pueden realizar tareas en el rol de programador.

Ejemplos de resúmenes obtenidos a partir de los datos integrados con el instrumento “Inventario de personalidad 16 PF en Forma C”:

* En casi la totalidad de los especialistas que tienen alto rendimiento y rol de programador se cumple que coincidieron como grupo en valores normales en la protensión y fuerza en el ego.
* En muchos especialistas que tienen alto rendimiento y rol de calidad se cumple que coincidieron como grupo en valores normales en la socialización.
* En la mayoría de los especialistas que tienen rol de analistas en un segundo nivel y rendimiento medio se cumple coincidieron como grupo en valores normales en la animación.
* En muchos especialistas que tienen alto rendimiento y rol de arquitecto en un segundo nivel se cumplen que coincidieron como grupo en valores normales en el perfeccionismo.

Ejemplos de resúmenes obtenidos a partir de los datos integrados con el instrumento “BFQ, Cuestionario Big Five”:

* En la mayoría de los especialistas que tienen rol de programador y alto rendimiento se cumple que son poco comprensivos y son poco tolerantes.
* En casi la totalidad de los especialistas que tienen rol de calidad y alto rendimiento se cumple que son moderadamente creativos y son muy afectivos.
* En la mayoría de los especialistas que tienen rol de arquitecto y alto rendimiento se cumple que son moderadamente responsables y son muy poco activos.
* En muchos especialistas que tienen rol de analista como segundo nivel y alto rendimiento se cumple que son moderadamente creativos y poco comprensivos.

Ejemplos de resúmenes obtenidos a partir de los datos integrados con el instrumento “Cuestionario sobre estilos de dirección utilizando variables lingüísticas”:

* En muchos especialistas que tienen rol de programador y alto rendimiento en condiciones normales y presentan mezcla técnica se cumple que son personas exactas y precisas y calmadas y lógicas y completan trabajos importantes siguiendo métodos probados.
* En casi la totalidad de los especialistas que tienen alto rendimiento en condiciones de tensión y rol de arquitectos en un segundo nivel se cumple que son pasivos y pueden realizar tareas en el rol de programador.
* En muchos especialistas que tienen alto rendimiento en tareas de calidad se cumple que se orientan hacia las personas en condiciones normales.
* En muchos especialistas que tienen alto rendimiento en tareas de calidad se cumple que se orientan hacia las tareas en condiciones de tensión.

A partir del análisis de los resúmenes lingüísticos generados de la aplicación de los instrumentos de personalidad, los decisores concluyeron que los resúmenes:

* Permitieron descubrir relaciones entre los rasgos de la personalidad y la mayoría de los roles principales de un proyecto de software.
* Permiten identificar relaciones entre los rasgos de la personalidad y el rendimiento en los roles asignados.
* Brindan información relevante, actualizada y de forma rápida para apoyar la toma de decisiones en el proceso de adquisición de los recursos humanos en proyectos de software.

#### Toma de decisiones durante los cortes de proyectos

También se aplicaron resúmenes lingüísticos en la toma de decisiones durante los cortes de proyectos. En este sentido se trabajaron con los datos recopilados en el sistema de información GESPRO, en particular en las tablas y registros que recogen información indicadores de evaluación, control y seguimiento de proyectos (Piñero et al. 2020). Se muestran a continuación ejemplos de los resúmenes lingüísticos obtenidos en este caso y la interpretación que hicieron los decisores ante los mismos:

* Muchos proyectos que tienen perfecto el tiempo real del trabajo real o alta cantidad de RH evaluados de mal o aproximadamente el 50% de los RH con competencia baja o mala correspondencia entre el tiempo planificado y el tiempo realizado tienen mal el índice de ejecución.

Como resultado del análisis de este resumen, los decisores identificaron que, durante la ejecución de los proyectos con baja evaluación, hay una elevada cantidad de recursos con bajas competencias. Además, que muchos de los proyectos evaluados de mal, tienen como causa un mal desempeño de los recursos humanos. A partir de este resumen, para lograr el éxito de los proyectos se identificaron como posibles medidas asociadas al desempeño:

* Elevar las competencias técnicas de los recursos humanos, así mejorar su desempeño.
* Aplicar mejores métodos para la motivación del personal, como recompensas y penalizaciones.
* Mover recursos humanos más competentes de los proyectos menos priorizados para los más priorizados. Detener si es necesario los proyectos de menor prioridad.
* Aproximadamente el 50% de los proyectos que tienen mal índice de ejecución o baja cantidad de RH evaluados de bien o regular la evaluación tienen mala ejecución.

Los decisores identificaron en el 50 % de los proyectos que están mal en el momento del análisis o del corte, persisten dificultades en su producción en proceso, lo cual acarreará mayores conflictos con los clientes. Se recomendó que se tomaran medidas de forma inmediata que les permitiera recuperar los atrasos presentes.

* Pocos proyectos que tienen aproximadamente el 50% de los RH con baja competencia tienen perfecto el tiempo trabajado.

A partir de este resumen y otros similares, los decisores notaron que existe alta dependencia de las competencias laborales con la eficiencia en el trabajo, ratificándose que son pocos los proyectos que tienen el 50% de los recursos humanos con bajas competencias y que logran una excelente gestión de tiempo. En este sentido, identificaron como posibles medidas:

* Elevar las competencias técnicas de los recursos humanos, así mejorar su desempeño.
* Reordenar los equipos de la organización en función de sus competencias.
* Aproximadamente el 50% de los proyectos que tienen alta cantidad de RH evaluados de mal tienen perfecto el tiempo planificado.
* Muchos de los proyectos que tienen perfecto el tiempo planificado tienen alta cantidad de RH evaluados de mal.

Este otro grupo de resúmenes lingüísticos a criterio de los decisores evidenciaron que existían proyectos donde los recursos humanos están alterando (falseando) los datos de la gestión del tiempo, porque no es posible que sus recursos humanos estén evaluados de mal y la gestión del tiempo esté excelente. Como medida, se propuso revisar los proyectos afectados y corregir la dificultad en la generación de la información.

* Aproximadamente el 50% de los proyectos que tienen regular el aprovechamiento de los RH o baja cantidad de RH evaluados de bien o regular desempeño de los RH o moderado fondo de tiempo tienen mala eficacia.
* Aproximadamente el 50% de los proyectos que tienen media la cantidad de RH evaluados de mal tienen mala eficacia.

Los resúmenes sexto y séptimo evidencian que existe una marcada influencia de los recursos humanos en la calidad del proyecto. Cuando los recursos humanos están dedicando poco tiempo al desarrollo de las actividades, uno de los criterios que se afecta inmediatamente es la calidad. En este caso se propuso a los gestores de proyectos:

* Revisar la calidad y el tiempo que dedican los recursos humanos a las tareas asignadas.
* Planificar horas extras al equipo.
* Concentrar a los recursos humanos principales liberándolos de otras tareas que afecten su desempeño.

#### Detección de errores en la planificación de proyectos

En el entorno de gestión de proyectos también se evaluó la capacidad de los algoritmos para la identificación de errores en la planificación de proyectos, detectando errores en los cronogramas con impacto en el tiempo y en los costos fundamentalmente. Para esta prueba se construyó una base de datos experimental de tareas y de recursos en proyectos. A esta base de datos se le introdujeron de forma controlada, algunos cambios que afectaron a menos del 5% de los registros asociados a modificaciones en los tiempos de duración de las tareas, o cambios en la cantidad de recursos humanos o no humanos que requerían.

Se aplicó el meta-algoritmo pero solo empleando los algoritmos orientados a la detección de datos anómalos con el objetivo de validar la capacidad de detección de datos anómalos (Pérez Pupo, Piñero Pérez, Vacacela, et al. 2020). Algunos de los resúmenes lingüísticos obtenidos se muestran a continuación:

1. Aproximadamente el 50% de los proyectos anómalos tienen muy alta cantidad de tareas planificadas para los recursos humanos.

Este resumen significa que alrededor de la mitad de los “proyectos anómalos” han planificado en exceso la cantidad de tareas asignadas a los recursos humanos, creando sobrecarga de trabajo, aumentando la probabilidad de riesgos que afecten el cumplimiento de los cronogramas de proyecto. En este sentido, para disminuir la sobrecarga de tiempo, se tomaron las siguientes medidas:

* Priorizar las tareas y renegociar los tiempos en caso necesario.
* Revisar qué tareas se pueden paralelizar y proceder con otros recursos humanos en las tareas menos priorizadas (subcontratar).

1. Cercano al 33% de los proyectos anómalos tienen muy alta tarifa horaria de recursos humanos.

Este segundo resumen representa que el costo de recursos humanos en esos proyectos está sobre-planificado.

1. Cercano al 33% de los proyectos anómalos tienen muy alta cantidad de recursos materiales planificados.
2. Cercano al 33% de los proyectos anómalos que tienen alto plan de recursos materiales tienen alta tarifa horaria del equipamiento contratado.

Los resúmenes tercero y cuarto representan que en esos proyectos se han planificado en exceso los recursos materiales, y algunos de ellos, tienen una taza de costo de equipo sobre planificada. De ahí, como medidas aplicadas están:

* Identificación y selección de alternativas más económicas sin afectar significativamente la calidad.
* Revisión de los contratos y ajuste de los métodos de gestión logística priorizando los proyectos adecuadamente.

1. Cercano al 33% de los proyectos anómalos que tienen alta cantidad de tareas planificadas para los recursos humanos tienen muy alta cantidad de tiempo real dedicado.

De este quinto resumen lingüístico se interpreta que, en algunos casos, el número de recursos humanos se planificó por debajo del número real de recursos humanos utilizados.

A partir del análisis de estos resúmenes lingüísticos sobre elementos anómalos durante la planificación de proyectos, se identifica que en su mayoría, se refieren a la sobreestimación de los recursos humanos en las tareas del proyecto, por lo tanto, incurren en costos más altos por usar más recursos de lo planeado. La detección de estos tipos de resúmenes lingüísticos ayuda a los gerentes de proyectos a corregir errores en la programación del proyecto y detectar el costo excesivo del mismo.

#### Resumen del análisis multicriterio en el entorno de gestión de proyectos

En la aplicación del método de análisis multicriterio (ver Procedimiento 1), participaron 5 expertos en Gestión de Proyectos, ver Tabla 27 del Anexo 11, los cuales completaron el formulario de evaluación recogido en la Tabla 29 del Anexo 12 donde evalúan los resúmenes respecto a 11 criterios. De ellos, los criterios C3 y C7 están orientados a medir el indicador “interpretabilidad de resúmenes” mientras que el resto está orientado a evaluar la “facilidad para la toma de decisiones”.

Figura 24: Resultados de la agregación de las preferencias de los expertos en gestión de proyectos.

A partir del análisis se obtienen los resultados de la Figura 24 donde se muestran las preferencias agregadas de los expertos y los resultados de la Figura 25 que muestra la dispersión en las preferencias. En general se obtienen los siguientes resultados:

* Respecto a la “interpretabilidad de los resúmenes” se obtuvo una alta evaluación con una baja dispersión en las opiniones.
* Respecto a la “facilidad para la toma de decisiones”, la mayoría de los criterios fueron evaluados con altos valores a la vez que presentan baja dispersión en las opiniones; excepto los criterios C4 y C6, que aunque fueron evaluados con altas calificaciones, presentan mayor dispersión respecto al resto, como se muestra en la Figura 25.

Figura 25: Dispersión de opiniones de expertos por criterio, en el caso de estudio sobre gestión de proyectos.

Se demuestra que los resúmenes generados por los algoritmos propuestos son fácilmente interpretables por los expertos y útiles para la toma de decisiones.

También se aplicó el coeficiente de correlación *RWG* (Benavente Reche 2009) para el análisis de la concordancia entre los expertos, cuyo valor fue *RWG* = 0.95, demostrando que existe concordancia entre ellos.

## Conjunto de pruebas 6: validación de la variable dependiente en la dimensión “Enfoque multilingüe de los algoritmos”.

El caso de estudio de gestión de proyectos se consideró para demostrar la capacidad de los algoritmos de generar resúmenes en múltiples idiomas con los lenguajes naturales controlados definidos o referenciados en el capítulo 2. Se consideró este escenario porque:

* La gestión de proyectos es un área temática de fácil comprensión por especialistas de disímiles esferas. Los términos tratados en esta área temática son de uso común en la sociedad.
* Se dispone de especialistas que conocen la gestión de proyectos, su lengua materna es japonés o árabe (ver Anexo 15 Tabla 37), y dominan el español porque han vivido en Cuba por más de 15 años.

A partir del caso de estudio analizado se trabaja con la base de datos *BD\_gp\_eval\_proy\_hard* con un conjunto de 8 430 instancias, de la cual se generaron resúmenes en español, inglés, japonés y árabe. En el Anexo 16 se muestran ejemplos de estos resúmenes, siendo los mismos de un idioma a otro según la enumeración.

Los resúmenes generados fueron evaluados por un conjunto de expertos en los idiomas especificados (ver Anexo 15) respecto a los siguientes criterios: gramática, sencillez, legibilidad y ortografía. Para la agregación de las preferencias de los expertos se aplica la técnica de computación con palabras 2-tuplas (Herrera y Martínez 2000) (Zhang, S. et al. 2020) con el siguiente conjunto de términos lingüísticos *LBTL = {Muy bajo, Bajo, Medio, Alto, Muy alto}.* En la Figura 26, se representa la agregación de las evaluaciones asignadas por los expertos en estos idiomas a los resúmenes lingüísticos generados en el caso de estudio, oscilando entre “alta” y “muy alta”.

Figura 26: Resultado de la evaluación asignada por los expertos, a los resúmenes lingüísticos en los idiomas inglés, japonés y árabe.

Muy alto

Alto

Medio

Muy bajo

Bajo

En esta figura se observa que los criterios mejor evaluados globalmente fueron la sencillez y la ortografía. El árabe y el japonés resultaron mejor evaluados que el inglés, pero en los tres idiomas, la mayoría de los criterios obtuvieron calificaciones de alto o muy alto. En general, los expertos consultados se mostraron satisfechos con los resultados, pero se identifica como una línea abierta de la investigación, continuar trabajando en la legibilidad.

## Conclusiones del capítulo

En este capítulo se combinan técnicas cuantitativas y cualitativas que permiten comparar el comportamiento de los algoritmos propuestos *RST\_LDS, PCA\_LDS* y *LPA\_LDS* con los algoritmos *Apriori\_LDS*, *ACO\_LDS* y *AG\_LDS* reportados en la bibliografía y se concluye:

* El algoritmo GA\_LDS reportó los mejores resultados en los indicadores *T1,T4* y *Te5*; pero obtuvo los peores resultados en los indicadores *Te1a*, *Te3* y “Cubrimiento de diferentes situaciones en la base de datos”*.* Mientras que el *ACO\_LDS* solo logró superar a los demás algoritmos en el indicador “Cubrimiento de diferentes situaciones en la base de datos” y obtuvo los peores resultados en los indicadores *T3*, *T5, Te1b,Te4*, “Fortaleza de las dependencias” y “Tiempo de ejecución”*.* En el análisis del comportamiento global con el *Page’s L Test,* estos algoritmos quedaron ubicados entre los últimos lugares siendo *ACO\_LDS* el de peores resultados.
* Se demuestra que los algoritmos *PCA\_LDS* y *RST\_LDS* superaron a los reportados en la bibliografía en la mayoría de los experimentos realizados.
* El *RST\_LDS* quedó ubicado como el mejor algoritmo en el análisis global y obtuvo mejores resultados que el resto de los algoritmos en los indicadores *T2*, *T3*, *T5,Te1a,Te1b,Te5*, *Fortaleza de las dependencias* y *Tiempo de ejecución*. Este algoritmo nunca quedó ubicado como el peor en los indicadores analizados.
* En la mayoría de los experimentos realizados el algoritmo *PCA\_LDS* fue ligeramente superior al *Apriori\_LDS*, pero sin diferencias significativas entre ellos.
* El algoritmo *LPA\_LDS* es el segundo respecto a los peores resultados en el análisis global y fue el de peores resultados en los indicadores *T1* y *T2*.
* Se demuestra la aplicabilidad de los algoritmos propuestos a través de casos de estudio sobre tratamiento de embarazadas cardiópatas y sobre evaluación de gestión de proyectos. En la experimentación se evidencia la complementariedad entre los algoritmos *RST\_LDS*, *LPA\_LDS* y *PCA\_LDS* yse construye el meta-algoritmo *Hybrid\_LDS* como parte del ecosistema Cuadro de Mando para la Toma de Decisiones.
* Se valida la interpretabilidad de los resúmenes desde un enfoque multilingüe y se someten los resultados obtenidos al análisis de expertos en los idiomas español, inglés, japonés y árabe; considerando los criterios gramática, sencillez, legibilidad y ortografía. En la mayoría de los criterios, la evaluación fue entre alta y muy alta, siendo la ortografía el criterio de mejor evaluación y la legibilidad de los resúmenes en inglés, el criterio que obtuvo la menor evaluación.

# Conclusiones generales

En la bibliografía analizada se identifica que hay consenso entre los diferentes investigadores respecto a los elementos que conforman los resúmenes lingüísticos. Además, se identificó que la mayoría de las protoformas y algoritmos reportados en la bibliografía están diseñados para el idioma inglés, afectando su capacidad de poder representar los resúmenes en múltiples idiomas.

En la bibliografía consultada, las técnicas más empleadas para la generación de resúmenes lingüísticos emplean enfoques basados en consultas y en el uso de metaheurísticas. En el análisis de la bibliografía se identifican líneas abiertas a la investigación asociadas a la hibridación de técnicas que eleven la eficacia en la generación de resúmenes con menos costo computacional.

En esta tesis se proponen los algoritmos *RST\_LDS, PCA\_LDS, LPA\_LDS* y *LDS\_outliers* que combinan diferentes técnicas de computación emergente, entre las que se destacan la teoría neutrosófica, los conjuntos aproximados, el aprendizaje de reglas de asociación y el aprendizaje de grafos probabilísticos. Estos algoritmos facilitan la presentación de los resúmenes en múltiples idiomas empleando lenguajes naturales controlados.

Se demuestra que los algoritmos *PCA\_LDS* y *RST\_LDS* superaron a los algoritmos reportados en la bibliografía en la mayoría de los experimentos realizados, siendo el algoritmo *RST\_LDS* el que reportó los mejores resultados.

En la investigación se proponen dos enfoques que posibilitan aplicar los resúmenes lingüísticos en la minería de datos anómalos. El algoritmo *RST\_LDS* representa el primer enfoque centrado en identificar situaciones poco frecuentes en la base de datos, que puedan representar datos anómalos. Mientras que el algoritmo *LDS\_outliers* representa al segundo enfoque donde se aplican métodos para el descubrimiento de datos anómalos y luego la construcción de resúmenes lingüísticos a partir de éstos.

Los nuevos indicadores propuestos para la evaluación de los resúmenes lingüísticos complementan a los indicadores tradicionales reportados en la bibliografía, mejorando el tratamiento de la indeterminación y la falsedad. Se demuestra que:

* La introducción de conceptos como el parámetro *-corte*, así como la aproximación superior y la aproximación inferior en el cálculo de los indicadores, permiten regular con mayor flexibilidad el nivel de rigor en la evaluación de los resúmenes lingüísticos.
* El indicador *Te1a* reporta un comportamiento mejor que la variante tradicional *T1* en las clases de prueba analizadas.
* Los indicadores propuestos para medir la imprecisión, complementan al *T2*, no lo reemplaza. Las extensiones *Te2c* y *Te2d* permiten medir de forma diferenciada la indeterminación y la falsedad respectivamente. El indicador *Te2a* mide la impresión combinando la certeza, la indeterminación y la falsedad mientras que el *Te2b* calcula la imprecisión solo considerando la indeterminación y la falsedad.
* El indicador *Te3* extendido asociado al grado de cobertura es más estricto que el *T3* tradicional y diferencia mejor los escenarios representados en las clases de prueba.
* El indicador *Te4* relacionado con el grado de adecuación combina el grado de certidumbre con la indeterminación y la falsedad; discrimina mejor los diferentes escenarios representados en las clases de prueba respecto al indicador *T4*.
* El indicador *Te*5 extendido tiene un mejor comportamiento que su variante tradicional *T5*, la curva de valores que representa el *Te*5 es más suave.

Se demuestra la aplicabilidad de los algoritmos propuestos para la toma de decisiones a partir de dos casos de estudio y el empleo de métodos multicriterios. En la experimentación se evidencia la complementariedad entre los algoritmos *RST\_LDS*, *LPA\_LDS* y *PCA\_LDS* yse construye el meta-algoritmo *Hybrid\_LDS* como parte del ecosistema Cuadro de Mando para la Toma de Decisiones.

Finalmente se valida la capacidad de los algoritmos propuestos para la generación de resúmenes lingüísticos en los idiomas español, inglés, japonés y árabe respecto a los criterios gramática, sencillez, legibilidad y ortografía. En la mayoría de los criterios, la evaluación fue entre alta y muy alta, siendo la legibilidad de los resúmenes en inglés el criterio que obtuvo la menor evaluación.

# Recomendaciones

* Continuar aplicando los métodos de sumarización lingüística de datos en otros escenarios para la ayuda a la toma de decisiones, explotando su flexibilidad para ser combinados con otras técnicas.
* Continuar desarrollando nuevas formas de construcción de resúmenes lingüísticos y métodos para su visualización, cubriendo mayor cantidad de idiomas.
* Generar nuevos algoritmos basados en protoformas que potencien la integración de múltiples resúmenes, y faciliten no solo la generación de resúmenes en forma de oraciones, sino también la generación de textos largos que apoyen la toma de decisiones.

# Glosario de términos

**Aprendizaje activo:** caso especial deaprendizaje automático en el que un algoritmo de aprendizaje puede consultar de forma interactiva a un experto o usuario para etiquetar nuevos puntos de datos con resultados deseados (Rubens et al. 2016).

**Computación con Palabras**: es una metodología que permite realizar un proceso de computación y razonamiento utilizando palabras en lugar de números (Mendel et al. 2010). Dicha metodología permite crear y enriquecer modelos de decisión en los cuales la información vaga e imprecisa es representada a través de variables lingüísticas (Zadeh 1996).

**Corte de proyecto**: constituye un punto de control del proyecto donde se evalúa el estado del mismo. Generalmente se hace en forma de reuniones periódicas en las que participen los principales interesados del proyecto y en correspondencia con la etapa en que se encuentre el mismo.

**Decisiones no estructuradas**: responden a situaciones diversas y problemas de decisiones para los cuales se requieren modelos o procesos específicos de solución. El conjunto, de estas decisiones, representa el complemento de las decisiones estructuradas, estas últimas caracterizadas por ser decisiones que responden a problemas que ocurren regularmente, prácticamente cotidianas en el contexto de aplicación y para las cuales existen métodos definidos de solución.

**Entorno de toma de decisiones con incertidumbre**: se entiende como un escenario real de toma de decisiones no estructuradas donde en los datos que se manejan hay presencia de uno o varios de los siguientes fenómenos: vaguedad, incertidumbre, indeterminación, información incompleta, imprecisión o inconsistencia.

**Imprecisión**: está relacionada con la calidad de los datos, o sea, mide en qué grado los valores usados son precisos y puede ser provocada por errores en la medición de variables asociadas a un problema en cuestión.

**Incertidumbre**: se refiere a la veracidad del conocimiento utilizado por el sistema y se produce por la falta de certeza de la información. Se refleja en efectos como la imprecisión, la indeterminación, la inconsistencia y la incompletitud de la información, ver taxonomía en Anexo 5 Figura 35.

**Inconsistencia**: los datos existentes en el dominio de aplicación pueden ser contradictorios; por ejemplo, dos ejemplos prácticamente iguales pertenecen a clases diferentes.

**Indeterminación**: se entiende como el grado que caracteriza a los elementos que no tienen una pertenencia absoluta a un conjunto (certeza) y tampoco se pueda asegurar que no pertenezcan al propio conjunto de forma absoluta (falsedad). O sea, la Indeterminación incluye a los elementos que se encuentran entre los conjuntos opuestos <A> y <antiA>, y se denota por <neutA> (Smarandache 2017), (Prof. Dr. Florentin Smarandache 2021).

**Información incompleta:** considera la posibilidad de que no estén disponibles todos los datos requeridos para un razonamiento.

**Inteligencia artificial**: Conjunto de técnicas dedicadas a la creación de hardware y software que simulan el pensamiento humano y el comportamiento de los sistemas naturales para modelar soluciones a problemas para los cuales no hay algoritmos o los que existen tienen complejidad no polinomial. No pretende sustituir al hombre, sino apoyarlo en procesos complejos de toma de decisiones no estructuradas (Piñero Pérez y Pérez Pupo 2021).

**Neutrosofía:** rama de lafilosofía que estudia el origen, naturaleza y alcance de neutralidades, también sus interacciones con espectros ideacionales diferentes. Establecida por Florentin Smarandache en 1995. Esta teoría considera cada noción o idea <A> junto con su opuesto o negación <antiA>, y el espectro de neutralidades <neutA> (Smarandache 2017). Ver gráfico de evolución de esta teoría en el Anexo 6 Figura 36.

**Teoría neutrosófica:** es un área de conocimiento que trata la neutralidad o la indeterminación en el contexto de la toma de decisiones, constituye una generalización de la teoría de la lógica difusa (Zadeh 1965) y la teoría de la lógica intuicionista (Atanassov 1986). La teoría neutrosófica ofrece una representación matemática de la incertidumbre e indeterminación, donde cada proposición tiene un grado de certeza, indeterminación y falsedad (Smarandache 2019).

**Triangulación de datos**: consiste en considerar diferentes bases de datos para la validación de los métodos propuestos. Esto eleva la capacidad de generalización de los resultados de comparación entre los métodos propuestos (Moon 2019).

**Triangulación de expertos**: supone la participación de varios expertos en el proceso de validación, elemento que puede ayudar a mitigar el sesgo de considerar solamente las opiniones de un único experto. Se utiliza esta estrategia como complemento al análisis experimental combinando los métodos cuantitativos con los cualitativos. Los métodos cualitativos toman en consideración variables que no pueden ser expresadas cuantitativamente.

**Triangulación de métodos**: uso de diferentes métodos durante la validación, o sea, consiste en la comparación de los métodos propuestos con otros métodos reportados en la bibliografía (Moon 2019).

**Vaguedad**: puede estar motivada por el empleo de conceptos cuyos límites no están bien definidos; por ejemplo, el termino joven para denotar la edad de una persona, o bajo para la estatura.

# Bibliografía

AGGARWAL, C.C. y SATHE, S., 2017. *Outlier Ensembles: An Introduction*. S.l.: Springer. ISBN 978-3-319-54765-7.

ALVES, W., MARTINS, D., BEZERRA, U. y KLAUTAU, A., 2017. A Hybrid Approach for Big Data Outlier Detection from Electric Power SCADA System. *IEEE Latin America Transactions*, vol. 15, no. 1, pp. 57-64. ISSN 1548-0992. DOI 10.1109/TLA.2017.7827888.

AMGHAR, D. y CHIKH, Amine.M., 2018. Extracting a Linguistic Summary from a Medical Database. *International Journal of Intelligent Systems and Applications*, vol. 10, no. 12, pp. 16–26. ISSN 2074904X, 20749058. DOI 10.5815/ijisa.2018.12.02.

BAHNER, C.A. y CLARK, C.B., 2020. Sixteen Personality Factor Questionnaire (16PF). En: V. ZEIGLER-HILL y T.K. SHACKELFORD (eds.), *Encyclopedia of Personality and Individual Differences* [en línea]. Cham: Springer International Publishing, pp. 4958-4961. [Consulta: 29 marzo 2021]. ISBN 978-3-319-24612-3. Disponible en: https://doi.org/10.1007/978-3-319-24612-3\_86.

BARTZ-BEIELSTEIN, T., DOERR, C., BERG, D. van den, BOSSEK, J., CHANDRASEKARAN, S., EFTIMOV, T., FISCHBACH, A., KERSCHKE, P., LA CAVA, W., LOPEZ-IBANEZ, M., MALAN, K.M., MOORE, J.H., NAUJOKS, B., ORZECHOWSKI, P., VOLZ, V., WAGNER, M. y WEISE, T., 2020. Benchmarking in Optimization: Best Practice and Open Issues. En: arXiv: 2007.03488, *arXiv:2007.03488 [cs, math, stat]* [en línea], [Consulta: 12 agosto 2021]. Disponible en: http://arxiv.org/abs/2007.03488.

BELLO, R., FALCON, R. y VERDEGAY, J.L., 2019. *Uncertainty Management with Fuzzy and Rough Sets: Recent Advances and Applications*. S.l.: Springer. ISBN 978-3-030-10463-4.

BELLO, R. y VERDEGAY, J.L., 2012. Rough sets in the Soft Computing environment. *Information Sciences, ISSN 0020-0255*, vol. 212, pp. 1–14. ISSN 0020-0255. DOI 10.1016/j.ins.2012.04.041.

BENAVENTE RECHE, A.P., 2009. *Medidas de acuerdo y de sesgo entre jueces*. PhD Thesis. S.l.: Facultad de Psicología, Universidad de Murcia.

BÖGER, B., FACHI, M.M., VILHENA, R.O., COBRE, A.F., TONIN, F.S. y PONTAROLO, R., 2021. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *American Journal of Infection Control*, vol. 49, no. 1, pp. 21-29. ISSN 0196-6553. DOI 10.1016/j.ajic.2020.07.011.

BONISSONE, P.P. y DECKER, K.S., 1986. Selecting uncertainty calculi and granularity: An experiment in trading-off precision and complexity. *Machine intelligence and pattern recognition*. S.l.: Elsevier, pp. 217–247.

BORAN, F.E., AKAY, D. y YAGER, R.R., 2016. An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, vol. 61, pp. 356–377. DOI 10.1016/j.eswa.2016.05.044.

CAPRARA, G.V., BARBARANELLI, C., BORGOGNI, L. y PERUGINI, M., 1993. The “big five questionnaire”: A new questionnaire to assess the five factor model. *Personality and Individual Differences*, vol. 15, no. 3, pp. 281-288. ISSN 0191-8869. DOI 10.1016/0191-8869(93)90218-R.

CASTRO AGUILAR, G.F., 2017. *Modelo para el aseguramiento de ingresos en organizaciones orientadas a proyectos basado en minería de datos anómalos* [en línea]. Doctoral. La Habana, Cuba: Universidad de las Ciencias Informáticas. [Consulta: 2 abril 2021]. Disponible en: https://repositorio.uci.cu/jspui/handle/123456789/7933.

CHIANG, D.-A., CHOW, L.R. y WANG, Y.-F., 2000. Mining time series data by a fuzzy linguistic summary system. *Fuzzy sets and Systems*, vol. 112, no. 3, pp. 419–432. DOI 10.1016/S0165-0114(98)00003-7.

CHOW, C. y LIU, C., 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462-467. ISSN 1557-9654. DOI 10.1109/TIT.1968.1054142.

CPAŁKA, K., 2017. *Design of Interpretable Fuzzy Systems*. S.l.: Springer. SCI. ISBN 978-3-319-52881-6.

CURSI, F. y YANG, G., 2019. A Novel Approach for Outlier Detection and Robust Sensory Data Model Learning. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. S.l.: s.n., pp. 4250-4257. DOI 10.1109/IROS40897.2019.8967653.

DALATU, P.I., 2016. Time Complexity of K-Means and K-Medians Clustering Algorithms in Outliers Detection. *Global Journal of Pure and Applied Mathematics*, vol. 12, no. 5, pp. 4405–4418. ISSN 0973-1768.

DEGTIAREV, K.Y. y REMNEV, N.V., 2016. Linguistic resumes in software engineering: the case of trend summarization in mobile crash reporting systems. *Procedia Computer Science*, vol. 102, pp. 121–128. DOI 210.1016/j.procs.2016.09.378.

DÍAZ-HERMIDA, F. y VIDAL, J.C., 2018. Fuzzy quantification for linguistic data analysis and data mining. En: arXiv: 1807.07389, *II Data Science & Engineering ConsortiumMeeting* [en línea]. Universidade de Santiago de Compostela, Spain: Centro Singular de Investigaci ́on en Tecnolox ́ıas da Informaci ́on(CiTIUS), arXiv:1807.07389, [Consulta: 8 abril 2021]. Disponible en: http://arxiv.org/abs/1807.07389.

DIJKMAN, R. y WILBIK, A., 2017. Linguistic summarization of event logs – A practical approach. *Information Systems*, vol. 67, pp. 114–125. ISSN 0306-4379. DOI 10.1016/j.is.2017.03.009.

DONIS-DÍAZ, C.A., BELLO, R., KACPRZYK, J. y OTHERS, 2014. Linguistic data summarization using an enhanced genetic algorithm. *Czasopismo Techniczne*, vol. 2013, no. Automatyka Zeszyt 2 AC (10) 2013, pp. 3–12. DOI 10.4467/2353737XCT.14.019.2602.

DONIS-DÍAZ, C.A., BELLO, R. y KACPRZYK, Janusz, 2015. Using Ant Colony Optimization and Genetic Algorithms for the Linguistic Summarization of Creep Data. En: P. ANGELOV, K.T. ATANASSOV, L. DOUKOVSKA, M. HADJISKI, V. JOTSOV, J. KACPRZYK, N. KASABOV, S. SOTIROV, E. SZMIDT y S. ZADROŻNY (eds.), *Intelligent Systems’2014*. Cham: Springer International Publishing, pp. 81–92. ISBN 978-3-319-11313-5. DOI 10.1007/978-3-319-11313-5\_8.

DONIS-DÍAZ, C.A., MURO, A.G., BELLO-PÉREZ, R. y MORALES, E.V., 2014. A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data. *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 2035–2042. ISSN 0957-4174. DOI 10.1016/j.eswa.2013.09.002.

DUA, D. y GRAFF, C., 2021. *UCI Machine Learning Repository* [en línea]. S.l.: University of California, Irvine, School of Information and Computer Sciences. Disponible en: http://archive.ics.uci.edu/ml.

DUBOIS, D. y PRADE, H., 1990. Rough fuzzy sets and fuzzy rough sets. *International Journal of General System*, vol. 17, no. 2-3, pp. 191–209. DOI 10.1080/03081079008935107.

DUBOIS, D. y PRADE, H., 1992. Gradual inference rules in approximate reasoning. *Information Sciences*, vol. 61, no. 1, pp. 103-122. ISSN 0020-0255. DOI 10.1016/0020-0255(92)90035-7.

DUNTEMAN, G.H., 1989. *Principal Components Analysis*. S.l.: SAGE, ISBN: 978-0-8039-3104-6. ISBN 978-0-8039-3104-6.

DURAJ, A., SZCZEPANIAK, P.S. y CHOMATEK, L., 2020. Intelligent Detection of Information Outliers Using Linguistic Summaries with Non-monotonic Quantifiers. *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. S.l.: Springer, pp. 787–799. ISBN 978-3-030-50152-5. DOI 10.1007/978-3-030-50153-2\_58.

DURAJ, A., SZCZEPANIAK, P.S. y OCHELSKA-MIERZEJEWSKA, J., 2016. Detection of Outlier Information Using Linguistic Summarization. En: T. ANDREASEN, H. CHRISTIANSEN, J. KACPRZYK, H. LARSEN, G. PASI, O. PIVERT, G. DE TRÉ, M.A. VILA, A. YAZICI y S. ZADROŻNY (eds.), *Flexible Query Answering Systems 2015*. Cham: Springer International Publishing, pp. 101–113. ISBN 978-3-319-26154-6. DOI 10.1007/978-3-319-26154-6\_8.

ECIOLAZA, L., PEREIRA-FARIÑA, M. y TRIVINO, G., 2013. Automatic linguistic reporting in driving simulation environments. *Applied Soft Computing*, vol. 13, no. 9, pp. 3956–3967. DOI 10.1016/j.asoc.2012.09.007.

GARCÍA, J.A.L., PEÑA, A.B., PIÑERO PÉREZ, P.Y. y PÉREZ, R.B., 2017. Project Control and Computational Intelligence: Trends and Challenges. *International Journal of Computational Intelligence Systems*, vol. 10, no. 1, pp. 320–335. ISSN 1875-6883. DOI 10.2991/ijcis.2017.10.1.22.

GENÇ, S., AKAY, D., BORAN, F.E. y YAGER, R.R., 2020. Linguistic summarization of fuzzy social and economic networks: an application on the international trade network. *Soft Computing*, vol. 24, no. 2, pp. 1511–1527. DOI 10.1007/s00500-019-03982-9.

GEORGE, R., SRIKANT, R., HERRERA, F. y VERDEGAY, J.L., 1996. Data summarization using genetic algorithms and fuzzy logic. *Genetic Algorithms and Soft Computing*, pp. 599–611.

GILSING, R., WILBIK, A., GREFEN, P., TURETKEN, O. y OZKAN, B., 2020. A Formal Basis for Business Model Evaluation with Linguistic Summaries. *Enterprise, Business-Process and Information Systems Modeling*. S.l.: Springer, pp. 428–442. ISBN 978-3-030-49417-9.

GRAU, R., CORREA, C. y ROJAS, M., 1999. *Metodología de la investigación*. Corporación universitaria de Ibague. S.l.: s.n. ISBN 958-8028-10-8.

GUOYIN WANG, 2002. Extension of rough set under incomplete information systems. *2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE’02. Proceedings (Cat. No.02CH37291)*. S.l.: s.n., pp. 1098-1103 vol.2. DOI 10.1109/FUZZ.2002.1006657.

HEBLE-LAHERA, C., CASCALLAR-FUENTES, A., RAMOS-SOTO, A. y DIZ, A.B., 2020. Empirical study of fuzzy quantification models for linguistic descriptions of meteorological data. *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. S.l.: IEEE, pp. 1–7. ISBN 978-1-72816-933-0. DOI 10.1109/FUZZ48607.2020.9177716.

HERRERA, F. y MARTÍNEZ, L., 2000. A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on fuzzy systems*, vol. 8, no. 6, pp. 746–752. ISSN 1063-6706. DOI 10.1109/91.890332.

HUDEC, M., BEDNÁROVÁ, E. y HOLZINGER, A., 2018. Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language. *Journal of Official Statistics*, vol. 34, no. 4, pp. 981–1010. DOI 10.2478/jos-2018-0048.

IGDE, E.Y., AYDOĞAN, S., BORAN, F.E. y AKAY, D., 2017. Linguistic Summarization of Structured Patent Data. *International Scholarly and Scientific Research & Innovation*. S.l.: s.n., pp. 1062-1065.

JAIN, A., KELLER, J.M. y BEZDEK, J.C., 2016. Quantitative and qualitative comparison of periodic sensor data. *2016 IEEE-embs international conference on biomedical and health informatics (bhi)*. S.l.: IEEE, pp. 37–40. ISBN 978-1-5090-2455-1. DOI 10.1109/BHI.2016.7455829.

JAIN, A., POPESCU, M., KELLER, J., RANTZ, M. y MARKWAY, B., 2019. Linguistic summarization of in-home sensor data. *Journal of Biomedical Informatics*, vol. 96, pp. 103240. ISSN 1532-0464. DOI 10.1016/j.jbi.2019.103240.

KACPRZYK, Janusz, 1999. An interactive fuzzy logic approach to linguistic data summaries. *18th International Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 99TH8397)*. S.l.: IEEE, pp. 595–599. ISBN 0-7803-5211-4. DOI 10.1109/NAFIPS.1999.781763.

KACPRZYK, J., 1999. Fuzzy logic for linguistic summarization of databases. *FUZZ-IEEE’99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315)*. S.l.: s.n., pp. 813–818 vol.2. DOI 10.1109/FUZZY.1999.793053.

KACPRZYK, J. y STRYKOWSKI, P., 1999. Linguistic summaries of sales data at a computer retailer via fuzzy logic and a genetic algorithm. *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, vol. 2, pp. 937–943. DOI 10.1109/CEC.1999.782523.

KACPRZYK, J. y WILBIK, A., 2010. Linguistic summaries of time series: on some extended aggregation techniques. *Studia i Materiały Polskiego Stowarzyszenia Zarzdzania Wiedza*, vol. 2010, no. 31, pp. 326-337.

KACPRZYK, J. y YAGER, R.R., 2001. Linguistic summaries of data using fuzzy logic. *International Journal of General System*, vol. 30, no. 2, pp. 133–154. DOI 10.1080/03081070108960702.

KACPRZYK, J., YAGER, R.R. y MERIGO, J.M., 2019. Towards Human-Centric Aggregation via Ordered Weighted Aggregation Operators and Linguistic Data Summaries: A New Perspective on Zadeh’s Inspirations. *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 16-30. ISSN 1556-603X. DOI 10.1109/MCI.2018.2881641.

KACPRZYK, J., YAGER, R.R. y ZADROŻNY, S., 2000. A Fuzzy Logic Based Approach to Linguistic Summaries of Databases. *International Journal of Applied Mathematics and Computer Science*, no. Vol. 10, no 4, pp. 813–834. ISSN 1641-876X.

KACPRZYK, J. y ZADROŻNY, S., 1995. Fquery for Access: Fuzzy Querying for a Windows-Based DBMS. *Fuzziness in database management systems* [en línea]. S.l.: Springer, pp. 415–433. Disponible en: https://link.springer.com/chapter/10.1007/978-3-7908-1897-0\_18.

KACPRZYK, J. y ZADROŻNY, S., 2000. On a fuzzy querying and data mining interface. *Kybernetika*, vol. 36, no. 6, pp. 657–670.

KACPRZYK, J. y ZADROŻNY, S., 2003. Linguistic summarization of data sets using association rules. *Fuzzy Systems, 2003. FUZZ’03. The 12th IEEE International Conference on*. S.l.: IEEE, pp. 702–707.

KACPRZYK, J. y ZADROŻNY, S., 2005. Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. En: B. GABRYS, K. LEIVISKÄ y J. STRACKELJAN (eds.), *Do Smart Adaptive Systems Exist? Best Practice for Selection and Combination of Intelligent Methods* [en línea]. Berlin, Heidelberg: Springer, Studies in Fuzziness and Soft Computing, pp. 321–340. [Consulta: 10 enero 2020]. ISBN 978-3-540-32374-7. Disponible en: https://doi.org/10.1007/3-540-32374-0\_16.

KACPRZYK, Janusz y ZADROŻNY, S., 2005. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Sciences*, vol. 173, no. 4, pp. 281–304. DOI 10.1016/j.ins.2005.03.002.

KACPRZYK, J. y ZADROŻNY, S., 2009. Linguistic database summaries using fuzzy logic, towards a human-consistent data mining tool. , no. 20, pp. 10.

KACPRZYK, J. y ZADROŻNY, S., 2010. Linguistic data summarization: a high scalability through the use of natural language? *Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design*. S.l.: IGI Global, pp. 214–237.

KACPRZYK, J. y ZADROŻNY, S., 2016a. Fuzzy logic-based linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems under imprecision. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 1, pp. 37–46. DOI 10.1002/widm.1175.

KACPRZYK, J. y ZADROŻNY, S., 2016b. Linguistic summarization of the contents of Web server logs via the Ordered Weighted Averaging (OWA) operators. *Fuzzy Sets and Systems*, vol. 285, pp. 182–198. DOI 10.1016/j.fss.2015.07.020.

KACPRZYK, J. y ZADROŻNY, S., 2016c. On a fairness type approach to consensus reaching support under fuzziness via linguistic summaries. *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. S.l.: s.n., pp. 1999–2006. DOI 10.1109/FUZZ-IEEE.2016.7737937.

KACPRZYK, J. y ZADROŻNY, S., 2018. Reaching Consensus in a Group of Agents: Supporting a Moderator Run Process via Linguistic Summaries. *Soft Computing Applications for Group Decision-making and Consensus Modeling*. S.l.: Springer, pp. 465–485.

KACPRZYK, J., ZADROŻNY, S. y DZIEDZIC, M., 2014. A Novel View of Bipolarity in Linguistic Data Summaries. En: L.T. KÓCZY, C.R. POZNA y J. KACPRZYK (eds.), *Issues and Challenges of Intelligent Systems and Computational Intelligence* [en línea]. Cham: Springer International Publishing, Studies in Computational Intelligence, pp. 215–229. [Consulta: 16 diciembre 2019]. ISBN 978-3-319-03206-1. Disponible en: https://doi.org/10.1007/978-3-319-03206-1\_16.

KACZMAREK-MAJER, K., HRYNIEWICZ, O., DOMINIAK, M. y ŚWIĘCICKI, Ł., 2019. Personalized linguistic summaries in smartphone-based monitoring of bipolar disorder patients. [en línea]. S.l.: Atlantis Press, [Consulta: 11 diciembre 2019]. ISBN 978-94-6252-770-6. DOI 10.2991/eusflat-19.2019.56. Disponible en: https://www.atlantis-press.com/proceedings/eusflat-19/125914826.

KAMALOV, F. y LEUNG, H.H., 2020. Outlier Detection in High Dimensional Data. *Journal of Information & Knowledge Management*, vol. 19, no. 01, pp. 2040013. ISSN 0219-6492. DOI 10.1142/S0219649220400134.

KELLEY, G.A. y KELLEY, K.S., 2019. Systematic reviews and meta-analysis in nutrition research. *British Journal of Nutrition*, vol. 122, no. 11, pp. 1279-1294. ISSN 0007-1145, 1475-2662. DOI 10.1017/S0007114519002241.

KHEDIDJA, B., ALLEL, H. y MOHAND, L., 2020. Data Summarization for Sensor Data Management: Towards Computational-Intelligence-Based Approaches. *International Journal of Computing and Digital Systems*, vol. 9, no. 5, pp. 825–833. ISSN 2210-142X. DOI 10.12785/ijcds/090505.

KITTREDGE, R.I., 2003. Sublanguages and Controlled Languages. *The Oxford Handbook of Computational Linguistics 2nd edition, 2016, ISBN: 9780199573691*, pp. 430-447. DOI 10.1093/oxfordhb/9780199573691.013.015.

KUHN, T., 2014. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics, ISSN: 0891-2017*, vol. 40, no. 1, pp. 121–170. ISSN 0891-2017. DOI 10.1162/COLI\_a\_00168.

LASISI, A. y ATTOH-OKINE, N., 2018. Principal components analysis and track quality index: A machine learning approach. *Transportation Research Part C: Emerging Technologies*, vol. 91, pp. 230-248. ISSN 0968-090X. DOI 10.1016/j.trc.2018.04.001.

LI, Y., ZHA, D., VENUGOPAL, P., ZOU, N. y HU, X., 2020. PyODDS: An End-to-end Outlier Detection System with Automated Machine Learning. *Companion Proceedings of the Web Conference 2020* [en línea]. New York, NY, USA: Association for Computing Machinery, pp. 153–157. [Consulta: 2 abril 2021]. ISBN 978-1-4503-7024-0. DOI 10.1145/3366424.3383530. Disponible en: https://doi.org/10.1145/3366424.3383530.

LÓPEZ, P., 2017. *Procedimiento para la aplicación de test de personalidad como apoyo a la gestión de recursos humanos en proyectos informáticos*. Máster. La Habana, Cuba: Universidad de las Ciencias Informáticas.

MANDEL, I., 2020. Expert system for Statistics STATEX – 30 years after. [en línea]. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. [Consulta: 29 marzo 2021]. ID 3664990. Disponible en: https://papers.ssrn.com/abstract=3664990.

MARÍN, N. y SÁNCHEZ, D., 2016. On generating linguistic descriptions of time series. *Fuzzy Sets and Systems*, vol. 285, pp. 6–30. DOI 10.1016/j.fss.2015.04.014.

MCCORMICK, K. y SALCEDO, J., 2017. *SPSS Statistics for Data Analysis and Visualization*. S.l.: John Wiley & Sons, ISBN: 978-1-119-00355-7. ISBN 978-1-119-00355-7.

MEDINA, J.E., HERNÁNDEZ, J., HERNÁNDEZ, R., PÉREZ, A., HECHAVARRÍA., A. y GONZÁLEZ, R., 2007. Generación de conjuntos de ítems y reglas de asociación. [en línea]. Reporte técnico sobre minería de datos. Cenatav: Serie Gris, 2143. Disponible en: www.cenatav.co.cu.

MENDEL, J.M., ZADEH, L.A., TRILLAS, E., YAGER, R., LAWRY, J., HAGRAS, H. y GUADARRAMA, S., 2010. What Computing with Words Means to Me [Discussion Forum]. En: Conference Name: IEEE Computational Intelligence Magazine, *IEEE Computational Intelligence Magazine*, vol. 5, no. 1, pp. 20-26. ISSN 1556-6048. DOI 10.1109/MCI.2009.934561.

MOON, M.D., 2019. Triangulation: A Method to Increase Validity, Reliability, and Legitimation in Clinical Research. *Journal of Emergency Nursing*, vol. 45, no. 1, pp. 103-105. ISSN 0099-1767, 1527-2966. DOI 10.1016/j.jen.2018.11.004.

NAIK, G., 2018. *Advances in Principal Component Analysis. Research and Development*. Springer. S.l.: s.n. 1. ISBN 978-981-10-6704-4.

NERURKAR, P., SHIRKE, A., CHANDANE, M. y BHIRUD, S., 2018. Empirical Analysis of Data Clustering Algorithms. *Procedia Computer Science*, vol. 125, pp. 770-779. ISSN 1877-0509. DOI 10.1016/j.procs.2017.12.099.

NGUYEN, D., NGUYEN, P., DO, K., RANA, S., GUPTA, S. y TRAN, T., 2020. Unsupervised Anomaly Detection on Temporal Multiway Data. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. S.l.: s.n., pp. 1059-1066. DOI 10.1109/SSCI47803.2020.9308219.

NING, J., CHEN, L. y CHEN, J., 2018. Relative Density-Based Outlier Detection Algorithm. *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence* [en línea]. New York, NY, USA: Association for Computing Machinery, pp. 227–231. [Consulta: 2 abril 2021]. ISBN 978-1-4503-6606-9. DOI 10.1145/3297156.3297236. Disponible en: https://doi.org/10.1145/3297156.3297236.

PASUOATHI, S., SHANMUGANATHAN, V., MADASAMY, K., YESUDHAS, H. y KIM, M., 2021. Trend analysis using agglomerative hierarchical clustering approach for time series big data. *The Journal of Supercomputing*, no. 77, pp. 6505-6524. DOI 10.1007/S11227-020-03580-9.

PAWLAK, Z., 1982. Rough sets. *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341-356. ISSN 1573-7640. DOI 10.1007/BF01001956.

PAWLAK, Z. y SKOWRON, A., 2007. Rough sets: Some extensions. *Information Sciences*, vol. 177, no. 1, pp. 28-40. ISSN 0020-0255. DOI 10.1016/j.ins.2006.06.006.

PELÁEZ-AGUILERA, M.D., ESPINILLA, M., FERNÁNDEZ, M.R. y MEDINA, J., 2019. Fuzzy Linguistic Protoforms to Summarize Heart Rate Streams of Patients with Ischemic Heart Disease. *Hindawi*, vol. 2019, pp. 11. DOI 0.1155/2019/2694126.

PÉREZ, I., LÓPEZ, P., VARONA, E., PIÑERO, P. y GARCÍA, R., 2018. Construcción de resúmenes lingüísticos a partir rasgos de la personalidad y el desempeño en el desarrollo de software. *Revista Cubana de Ciencias Informáticas*, vol. 12, pp. 135–150. ISSN 1994-1536.

PÉREZ PUPO, I., 2021. Repositorio de investigaciones en gestión de proyectos. [en línea]. Disponible en: https://gespro.uci.cu/projects/repositorio-de-investigaciones-en-gestion-de-proyectos/dmsf.

PÉREZ PUPO, I., GARCÍA VACACELA, R., PIÑERO PÉREZ, P., SADEQ, G. y PEÑA ABREU, M., 2020. Experiencias en el uso de técnicas de softcomputing en la evaluación de proyectos de software. *Revista Investigación Operacional*, vol. 41, no. 1, pp. 106–117. ISSN 2224-2405.

PÉREZ PUPO, I., PIÑERO PÉREZ, P.Y., BELLO PÉREZ, R.E., GARCÍA VACACELA, R.C., PIÑERO RAMÍREZ, P.E. y PIÑERO RAMÍREZ, C.M., 2020. Aplicaciones de la sumarización lingüística de datos en la toma de decisiones en gestión de proyectos. *V Conferencia Internacional en Ciencias Computacionales e Informáticas (CICCI’ 2020)*. S.l.: Informática CICCI´2020,

PÉREZ PUPO, I., PIÑERO PÉREZ, P.Y., BELLO, R., ACUÑA, L.A. y VACACELA, R.G., 2020. Linguistic Summaries Generation with Hybridization Method Based on Rough and Fuzzy Sets. *International Joint Conference on Rough Sets. Lecture Notes in Computer Science ISSN 0302-9743, Springer, 12179 LNAI, ISBN 978-3-030-52704-4*. S.l.: Lecture Notes in Computer Science, Springer, pp. 385–397. ISBN 978-3-030-52704-4. DOI 10.1007/978-3-030-52705-1\_29.

PÉREZ PUPO, I., PIÑERO PÉREZ, P.Y., GARCÍA VACACELA, R., BELLO, R., SANTOS ACOSTA, O. y LEYVA VÁZQUEZ, M.Y., 2018. Extensions to Linguistic Summaries Indicators based on Neutrosophic Theory: Applications in Project Management Decisions. *Neutrosophic Sets and Systems, University of New Mexico*, vol. 22, pp. 87-100.

PÉREZ PUPO, I., PIÑERO PÉREZ, P.Y., MARTÍN, N. y BELLO PÉREZ, R.E., 2021. Tendencias en la sumarización lingüística de datos. *Revista cubana de transformación digital*, vol. 2, no. 1, pp. 79-101. ISSN 2708-3411.

PÉREZ PUPO, I., PIÑERO PÉREZ, P.Y., VACACELA, R.G., BELLO, R. y ACUÑA, L.A., 2020. Discovering Fails in Software Projects Planning Based on Linguistic Summaries. *International Joint Conference on Rough Sets. Lecture Notes in Computer Science ISSN 0302-9743, Springer, 12179 LNAI, ISBN 978-3-030-52704-4* [en línea]. S.l.: Lecture Notes in Computer Science, Springer, pp. 365–375. DOI 10.1007/978-3-030-52705-1\_27. Disponible en: https://link.springer.com/content/pdf/10.1007%2F978-3-030-52705-1\_27.pdf.

PÉREZ PUPO, I., SANTOS ACOSTA, O., BELLO PÉREZ, R.E. y PIÑERO PÉREZ, P., 2019. Algorithms for linguistic data summarization, help in decision-making in project-oriented organizations. *XXII Ibero-American Conference on Software Engineering, CIbSE 2019*. S.l.: Springer, pp. 633–640.

PÉREZ PUPO, I., SANTOS, O., GARCÍA, R., PIÑERO PÉREZ, P.Y. y RAMÍREZ, E.C., 2018. Descubrimiento de resúmenes lingüísticos para ayuda a la toma decisiones en gestión de proyecto. *Revista Cubana de Ciencias Informáticas*, vol. 12, pp. 163–175. ISSN 1994-1536.

PÉREZ PUPO, I., VILLAVICENCIO, N., PIÑERO PÉREZ, P., GARCÍA VACACELA, R. y GARCÍA SANCHEZ, R., 2020. Proerp, ecosistema de software para la toma de decisiones en gestión de proyectos. *Experiencias Iberoamericanas de Ingeneria de Proyectos*. Guayaquil, Ecuador: Universidad Católica de Santiago de Guayaquil, pp. 832-845. ISBN 978-9942-82-511-7.

PIÑERO, P., PÉREZ PUPO, I., GARCÍA VACACELA, R. y TOSCANINI, P., 2020. *Caracterización de los estándares de gestión de proyectos y su impacto en la gestión económico financiera de las organizaciones orientadas a proyectos.* Guayaquil, Ecuador: Universidad Católica de Santiago de Guayaquil, ISBN: 978-9942-82-514-8. ISBN 978-9942-82-514-8.

PIÑERO, P., PUPO, I., RIVERO HECHAVARRÍA, C.C., ROJAS, C., SOSA, R. y TORRES, S., 2019. Repositorio de datos para investigaciones en gestión de proyectos. *Revista Cubana de Ciencias Informáticas*, vol. 13, no. 1, pp. 176–191. ISSN 2227-1899.

PIÑERO PÉREZ, P., PÉREZ PUPO, I. y PIÑERO CRUZ, P., 2019. Suite BusinessRedmine Sistema para la toma de decisiones empresariales y la gestión de proyectos 19.05. 4076-12-2019. Centro Nacional de Derecho de Autor de Cuba (CENDA). 4076-12-2019.

PIÑERO PÉREZ, P.Y. y PÉREZ PUPO, I., 2021. Curso Inteligencia Artificial para la toma de decisiones. Maestría en Gestión de Proyectos, Universidad de las Ciencias informáticas, Cuba. *MGPI* [en línea]. Disponible en: https://gespro.uci.cu/.

POKHREL, R., POKHAREL, P. y KUMAR TIMALSINA, A., 2019. Anomaly-Based – Intrusion Detection System using User Profile Generated from System Logs. *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 2, pp. p8631. ISSN 2250-3153. DOI 10.29322/IJSRP.9.02.2019.p8631.

PROF. DR. FLORENTIN SMARANDACHE, 2021. *Definition of indeterminacy* [en línea]. mail. 5 junio 2021. S.l.: s.n. Disponible en: http://fs.unm.edu/FlorentinSmarandache.htm.

RAMOS-SOTO, A. y MARTIN-RODILLAB, P., 2019. Enriching linguistic descriptions of data: A framework for composite protoforms. *Fuzzy Sets and Systems*, pp. 26. DOI 10.1016/j.fss.2019.11.013.

RANGA SURI, N.N.R., MURTY M, N. y ATHITHAN, G., 2019. *Outlier Detection: Techniques and Applications. A Data Mining Perspective*. S.l.: s.n. ISRL. ISBN 978-3-030-05127-3.

RASMUSSEN, D. y YAGER, R.R., 1999. Finding fuzzy and gradual functional dependencies with SummarySQL. *Fuzzy Sets and Systems*, vol. 106, no. 2, pp. 131–142. ISSN 0165-0114. DOI 10.1016/S0165-0114(97)00268-6.

REBANE, G. y PEARL, J., 2013. The Recovery of Causal Poly-Trees from Statistical Data. *Third Conference on Uncertainty in Artificial Intelligence* [en línea]. S.l.: AUAI Press, arXiv:1304.2736, ISBN: 978-0-444-87417-7, [Consulta: 24 marzo 2021]. Disponible en: http://arxiv.org/abs/1304.2736.

ROJAS VALENZUELA, Á.R., 2018. *Resúmenes lingüísticos para riego de cultivos* [en línea]. Tesis. Santiago, Chile: Universidad Técnica Federico Santa María, Departamento de Informática. Disponible en: https://repositorio.usm.cl.

RUBENS, N., ELAHI, M., SUGIYAMA, M. y KAPLAN, D., 2016. Active learning in recommender systems. *Recommender Systems Handbook*. S.l.: Springer. DOI 10.1007/978-1-4899-7637-6, 2da, ISBN 978-1-4899-7637-6.

RUIZ-SHULCLOPER, J., 2009. *Reconocimiento lógico combinatorio de patrones: teoría y aplicaciones*. PhD Thesis. Universidad Central «Marta Abreu» de Las Villas, Santa Clara: Centro de Investigaciones de Tecnologías de Avanzadas CENATAV.

SALKIND, N.J., 2007. *Encyclopedia of Measurement and Statistics*. S.l.: SAGE, ISBN: 978-1-4129-1611-0. ISBN 978-1-4129-1611-0.

SANCHEZ-VALDES, D., ALVAREZ-ALVAREZ, A. y TRIVINO, G., 2016. Dynamic linguistic descriptions of time series applied to self-track the physical activity. *Fuzzy Sets and Systems*, vol. 285, pp. 162–181. DOI 10.1016/j.fss.2015.06.018.

SHESKIN, D.J., 2003. *Handbook of Parametric and Nonparametric Statistical Procedures*. S.l.: CRC Press, ISBN: 1-58488-440-1. ISBN 1-58488-440-1.

SMARANDACHE, F., 2017. *Nidus idearum. Scilogs, III: Viva la Neutrosophia!* [en línea]. Editions Pons. Brussels, Belgium: Pons Publishing. ISBN 978-1-59973-508-5. Disponible en: https://digitalrepository.unm.edu/math\_fsp/306.

SMARANDACHE, F. y PRAMANIK, S., 2018. *New Trends in Neutrosophic Theory and Applications, Volume II*. Pons. Brussels, Belgium, EU: Infinite Study. ISBN 978-1-59973-559-7.

SMITS, G., NERZIC, P., PIVERT, O. y LESOT, M.-J., 2018. Efficient Generation of Reliable Estimated Linguistic Summaries. *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. S.l.: s.n., pp. 1–8. DOI 10.1109/FUZZ-IEEE.2018.8491604.

SOTO, M., 2003. *A single connected factorized distribution algorithm and its cost of evaluation*. Doctorado. Universidad de La Habana, Cuba: Instituto de Cibernética, Matemática y Física. Departamento Matemática Interdisciplinaria.

SOTO, M. y OCHOA, A., 2000. A factorized distribution algorithm based on polytrees. *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512)*. S.l.: s.n., pp. 232-237 vol.1. DOI 10.1109/CEC.2000.870300.

TAHYUDIN, I., HAVILUDDIN, H. y NAMBO, H., 2019. Time Complexity Of A Priori And Evolutionary Algorithm For Numerical Association Rule Mining Optimization. [en línea], [Consulta: 29 marzo 2021]. ISSN 2277-8616. Disponible en: http://repository.unmul.ac.id/handle/123456789/3898.

TONG, R.M. y BONISSONE, P.P., 1980. A linguistic approach to decisionmaking with fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 10, no. 11, pp. 716–723.

VAN ROOIJ, R. y SCHULZ, K., 2019. Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, vol. 28, no. 1, pp. 55-71. ISSN 1572-9583. DOI 10.1007/s10849-018-9275-5.

VERMA, A., TANEJA, A. y ARORA, A., 2017. Fraud detection and frequent pattern matching in insurance claims using data mining techniques. *2017 Tenth International Conference on Contemporary Computing (IC3)*. S.l.: s.n., pp. 1-7. DOI 10.1109/IC3.2017.8284299.

WANG, R., ZHU, Q., LUO, J. y ZHU, F., 2020. Local dynamic neighborhood based outlier detection approach and its framework for large-scale datasets. *Egyptian Informatics Journal* [en línea], [Consulta: 2 abril 2021]. ISSN 1110-8665. DOI 10.1016/j.eij.2020.06.001. Disponible en: https://www.sciencedirect.com/science/article/pii/S1110866520301328.

WASILEWSKA, A., 2007. Apriori algorithm. *Lecture Notes, http://www. cs. sunysb. edu/∼ cse634/lecture\_notes/07apriori. pdf, accessed* [en línea], vol. 10. Disponible en: https://www3.cs.stonybrook.edu/ cse634/lecture\_notes/07apriori.pdf.

WILBIK, A., BARRETO, D. y BACKUS, G., 2020. On Relevance of Linguistic Summaries–A Case Study from the Agro-Food Domain. *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. S.l.: Springer, pp. 289–300. ISBN 978-3-030-50146-4. DOI 10.1007/978-3-030-50146-4\_22.

WILBIK, A. y DIJKMAN, R.M., 2016. On the generation of useful linguistic summaries of sequences. *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. S.l.: s.n., pp. 555–562. DOI 10.1109/FUZZ-IEEE.2016.7737736.

WILBIK, A., KAYMAK, U. y DIJKMAN, R.M., 2017. A method for improving the generation of linguistic summaries. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. S.l.: s.n., pp. 1–6. DOI 10.1109/FUZZ-IEEE.2017.8015752.

WILBIK, A., VANDERFEESTEN, I., BERGMANS, D., HEINES, S. y MOOK, W. van, 2018. Linguistic Summaries for Compliance Analysis of a Glucose Management Clinical Protocol. *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. S.l.: s.n., pp. 1–7. DOI 10.1109/FUZZ-IEEE.2018.8491449.

WILCOXON, F., 1992. Individual Comparisons by Ranking Methods. En: S. KOTZ y N.L. JOHNSON (eds.), *Breakthroughs in Statistics: Methodology and Distribution* [en línea]. New York, NY: Springer, ISBN: 978-1-4612-4380-9, Springer Series in Statistics, pp. 196-202. [Consulta: 8 abril 2021]. ISBN 978-1-4612-4380-9. Disponible en: https://doi.org/10.1007/978-1-4612-4380-9\_16.

WU, D. y MENDEL, J.M., 2010. Linguistic summarization using IF–THEN rules and interval type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 1, pp. 136–151. ISSN 1063-6706. DOI 10.1109/TFUZZ.2010.2088128.

WU, D., MENDEL, J.M. y JOO, J., 2010. Linguistic summarization using if-then rules. *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*. S.l.: IEEE, pp. 1–8.

YAGER, R.R., 1982. A new approach to the summarization of data. *Information Sciences*, vol. 28, no. 1, pp. 69–86. DOI 10.1016/0020-0255(82)90033-0.

YAGER, R.R., 1991. On Linguistic Summaries of Data. *Knowledge Discovery in Databases*, pp. 378–389.

ZADEH, L.A., 1976. The concept of a linguistic variable and its applications to approximate reasoning. , pp. 43-80.

ZADEH, L.A., 1983. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with applications*, vol. 9, no. 1, pp. 149–184.

ZADEH, L.A., 1996. Fuzzy logic = computing with words. En: Conference Name: IEEE Transactions on Fuzzy Systems, *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 103-111. ISSN 1941-0034. DOI 10.1109/91.493904.

ZADEH, L.A., 2002. A prototype-centered approach to adding deduction capability to search engines–the concept of protoform. *Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium*. S.l.: IEEE, pp. 2–3. DOI 10.1109/IS.2002.1044219.

ZHAI, J., ZHANG, S. y ZHANG, Y., 2016. An extension of rough fuzzy set. *Journal of Intelligent & Fuzzy Systems*, vol. 30, no. 6, pp. 3311-3320. ISSN 1064-1246. DOI 10.3233/IFS-152079.

ZHANG, J., ZHANG, S., CHANG, K.H. y QIN, X., 2014. An outlier mining algorithm based on constrained concept lattice. *International Journal of Systems Science*, vol. 45, no. 5, pp. 1170-1179. ISSN 0020-7721. DOI 10.1080/00207721.2012.745029.

ZHANG, S., WEI, G., ALSAADI, F.E., HAYAT, T., WEI, C. y ZHANG, Z., 2020. MABAC method for multiple attribute group decision making under picture 2-tuple linguistic environment. *Soft Computing*, vol. 24, no. 8, pp. 5819–5829. DOI 10.1007/s00500-019-04364-x.

ZHANG, Xiaoliang, HE, Y., JIN, Y., QIN, H., AZHAR, M. y HUANG, J.Z., 2020. A Robust k-Means Clustering Algorithm Based on Observation Point Mechanism. *Complexity*, vol. 2020, pp. e3650926. ISSN 1076-2787. DOI 10.1155/2020/3650926.

ZHAO, Y., TARUS, S.K., YANG, L.T., SUN, J., GE, Y. y WANG, J., 2020. Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives. *Information Sciences*, vol. 515, pp. 132-155. ISSN 0020-0255. DOI 10.1016/j.ins.2019.10.019.

# Anexos

## Anexo 1. Producción científica del autor

Se presentan un conjunto de publicaciones del autor relacionadas a la investigación:

1. **I. Pérez Pupo**, Piñero, Vacacela, Bello, & Acuña. “Discovering Fails in Software Projects Planning Based on Linguistic Summaries”. Lecture Notes in Computer Science ISSN 0302-9743, Springer, ISBN 978-3-030-52704-4, pp. 365–375, doi: 10.1007/978-3-030-52705-1\_27, 2020. (SCOPUS)
2. **I. Pérez Pupo**, Piñero, Bello, Acuña, & Vacacela. “Linguistic Summaries Generation with Hybridization Method Based on Rough and Fuzzy Sets”. Lecture Notes in Computer Science ISSN 0302-9743, Springer, ISBN 978-3-030-52704-4, 2020, pp. 385–397, doi: 10.1007/978-3-030-52705-1\_29, 2020. (SCOPUS)
3. **I. Pérez Pupo**, Vacacela, Piñero, Sadeq, & Peña Abreu, M. “Experiencias en el uso de técnicas de softcomputing en la evaluación de proyectos de software”. *Revista Investigación Operacional*, *41*(1), 106–117, 2020. (SCOPUS)
4. Piñero, **I. Pérez Pupo**, Hechavarría, Lusardo, Sosa & Torres. “Repositorio de datos para investigaciones en gestión de proyectos”. *Revista Cubana de Ciencias Informáticas*, vol. 13, no. 1, pp. 176–191, ISSN 2227-1899, 2019. (SCIELO)
5. **I. Pérez Pupo**, Gómez, Varona, Piñero & Vacacela. “Construcción de resúmenes lingüísticos a partir de rasgos de la personalidad y el desempeño en el desarrollo de software”. *Revista Cubana de Ciencias Informáticas*, ISSN: 2227-1899, vol. 12, pp. 135–150, 2018. (SCIELO)
6. **I. Pérez Pupo**, Acosta, Vacacela, Piñero & Ramírez. “Descubrimiento de resúmenes lingüísticos para ayuda a la toma decisiones en gestión de proyectos”. *Revista Cubana de Ciencias Informáticas*, ISSN: 2227-1899, vol. 12, pp. 163–175, 2018. (SCIELO)
7. **I. Pérez Pupo**, Piñero, Vacacela, Bello, Santos & Leyva. “Extensions to Linguistic Summaries Indicators based on Neutrosophic Theory, Applications in Project Management Decisions”. *Neutrosophic Sets and Systems*, ISSN: 2331-6055, vol. 22, pp. 87–100, ISBN: 978-1-59973-587-0, 2018. (COPERNICUS) (SCOPUS)

Otras publicaciones en revistas o libros relevantes para la investigación

1. Piñero, **I. Pérez Pupo**, Vacacela & Toscanini. “Caracterización de los estándares de gestión de proyectos y su impacto en la gestión económico financiera de las organizaciones orientadas a proyectos”. Guayaquil, Ecuador: Universidad Católica de Santiago de Guayaquil, (Book) ISBN: 978-9942-82-514-8, 2020.
2. **I. Pérez Pupo,** Piñero, Martín & Bello. “Tendencias en la sumarización lingüística de datos”. *Revista Cubana de Transformación Digital*, 2(1), 79–101, 2021.

Publicaciones en eventos

1. **I. Pérez Pupo,** Villavicencio, Piñero, Vacacela & García. “Proerp, ecosistema de software para la toma de decisiones en gestión de proyectos”. (Book Section) in *Experiencias Iberoamericanas de Ingeniería de Proyectos*, Ecuador: Universidad Católica de Santiago de Guayaquil, pp. 832–845, ISBN: 978-9942-82-511-7, 2020.
2. **I. Pérez Pupo,** Santos, Bello & Piñero. “Algorithms for linguistic data summarization, help in decision-making in project-oriented organizations”. In *XXII Ibero-American Conference on Software Engineering*, CIbSE 2019, pp. 633–640, Springer, 2019. (SCOPUS)
3. **I. Pérez Pupo**, Piñero, Bello, Vacacela, et al. “Aplicaciones de la sumarización lingüística de datos en la toma de decisiones en gestión de proyectos”. In *V Conferencia Internacional en Ciencias Computacionales e Informáticas* (CICCI’ 2020). Informática CICCI´2020.
4. **I. Pérez Pupo,** Piñero. “Ecosistema de software inteligente para la toma de decisiones en gestión de proyectos”, en *IX Congreso Iberoamericano de Ingeniería de Proyecto, CIIP´2019,* Guayaquil, Ecuador, 27-29 noviembre, 2019.
5. Hechavarría, **I. Pérez Pupo**, Piñero & Huergo. “Proceso de limpieza de datos en la construcción del repositorio para investigaciones en gestión de proyectos”. In *IV Conferencia Internacional en Ciencias Computacionales e Informática* (CICCI’ 2018), La Habana, Cuba, pp. 24–26, 2018.

## Anexo 2. Protocolo de revisión sistemática aplicado

Protocolo de revisión sistemática aplicado en la construcción del marco teórico referencial:

1. Definición del objeto de estudio, campo de acción y objetivo de la investigación.
   1. Objeto de investigación: sumarización lingüística de datos
   2. Objetivo general: Desarrollar nuevos algoritmos para mejorar la eficacia en la generación y evaluación de resúmenes lingüísticos de datos respecto al tratamiento de la indeterminación, la interpretabilidad y su aplicabilidad en la toma de decisiones bajo un enfoque multilingüe.
   3. Campo de investigación: algoritmos para la generación y evaluación de resúmenes lingüísticos de datos.
2. Definición de un gestor bibliográfico: zotero.
3. Definición de fuentes de información académica para el desarrollo de la revisión: “*Semantic Scholar*”, “*Google Scholar*”, “*Scopus*” y otros meta-buscadores académicos basados en ciencia abierta.
4. Definición de frases claves para la realización de las búsquedas: “linguistic data summarization”, “linguistic data summaries”, “linguistic summarization”, “linguistic summaries”; también la combinación de éstas con “review”, “applied”.
5. Definición de las metas del análisis bibliométrico en forma de preguntas de investigación y criterios de inclusión – exclusión:
   1. ¿Cómo ha sido la tendencia de las publicaciones por año?
   2. ¿Cuáles son los principales autores?
   3. ¿Cuáles son las afiliaciones y países de los principales autores?
   4. ¿Cómo se distribuyen las publicaciones considerando los tipos de documentos en: artículos, libros, tesis y memorias de conferencias o congresos?
   5. Exclusión de trabajos publicados en espacios con poco nivel de arbitraje.
   6. Exclusión de trabajos de minería de textos no asociados al uso de técnicas de sumarización lingüística de datos.
6. Clasificar y filtrar las publicaciones en el siguiente conjunto de categorías:
   * Clásicas: se refiere a publicaciones pioneras en sumarización lingüística de datos, donde se exponen los principios fundamentales que marcan pautas en la teoría.
   * Extensiones a las teorías: se refiere a publicaciones que extienden la teoría planteada en las publicaciones entendidas como clásicas, no marcan pautas que cambien significativamente los métodos propuestos con anterioridad, aunque si desarrollan aportes al conocimiento.
   * Resultados de aplicación: publicaciones que se concentran en el empleo de la teoría existente en escenarios prácticos concretos.
   * Revisiones terciarias: se refiere a artículos de revisión de las tendencias y la evolución en la temática en cuestión.
7. Sintetizar las principales tendencias.
8. Realizar análisis detallado de cada uno de los trabajos y caracterizarlos respecto a:
   * Estructura de los resúmenes que se generan (protoformas).
   * Métodos o técnicas para la generación de los resúmenes lingüísticos de datos.
   * Principales técnicas y métodos de validación empleados en las investigaciones.
   * Áreas de aplicación de la propuesta.

## Anexo 3. Caracterización de la bibliografía consultada en la revisión sistemática

Figura 27: Publicaciones por países de la bibliografía consultada.

Figura 28: Bibliografía por tipo de publicación e indexado.

Figura 29: Fuentes y editoriales principales de la bibliografía analizada en profundidad

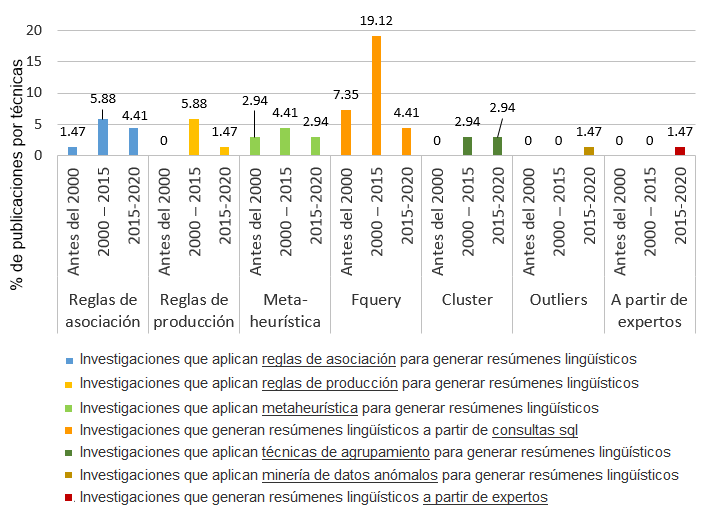


Figura 30: Tendencias de los métodos o técnicas para generar resúmenes lingüísticos de datos.

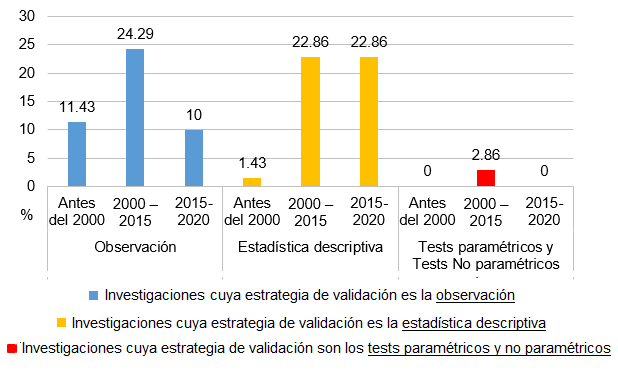


Figura 31: Estrategias y técnicas de validación empleadas en la bibliografía estudiada.

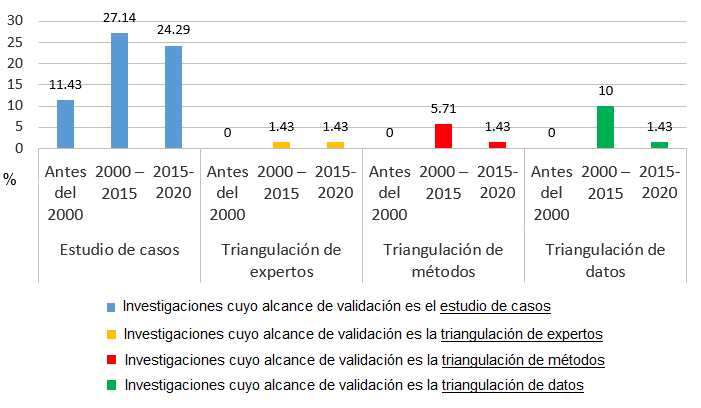


Figura 32: Alcance de la validación empleada en la bibliografía estudiada.

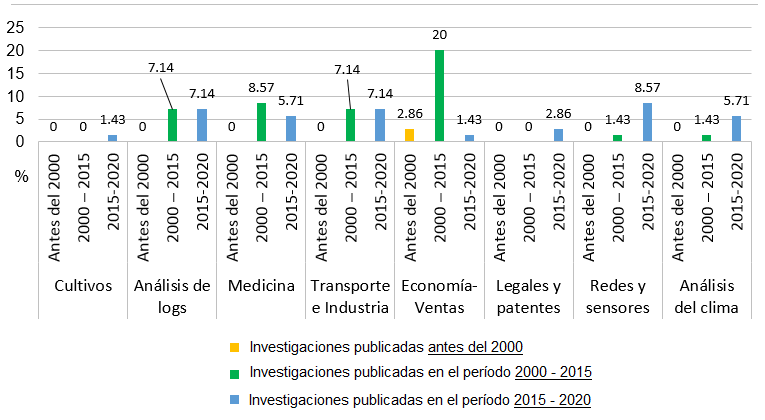


Figura 33: Análisis de publicaciones respecto a los entornos de aplicación.

## Anexo 4. Taxonomía de técnicas para la detección de datos anómalos

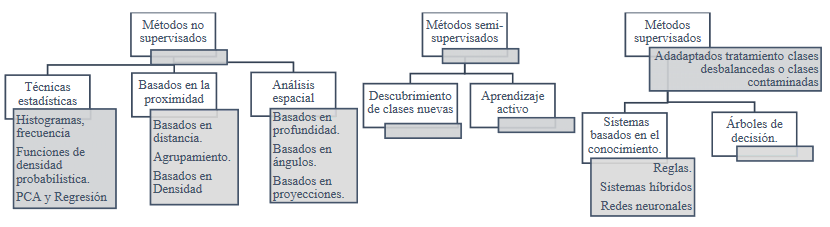
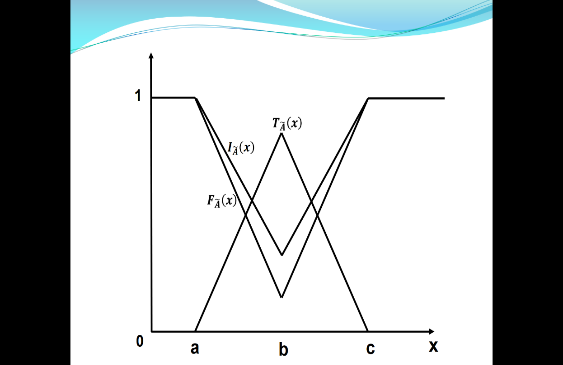
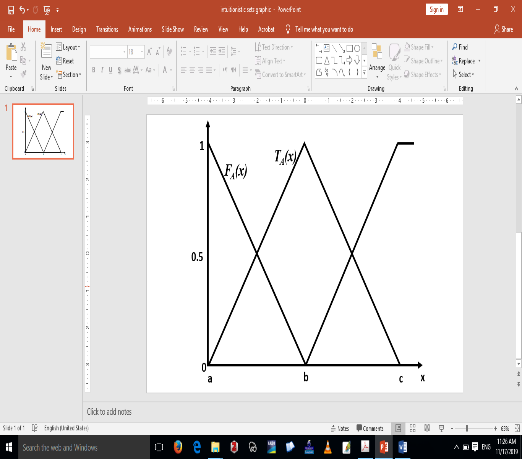
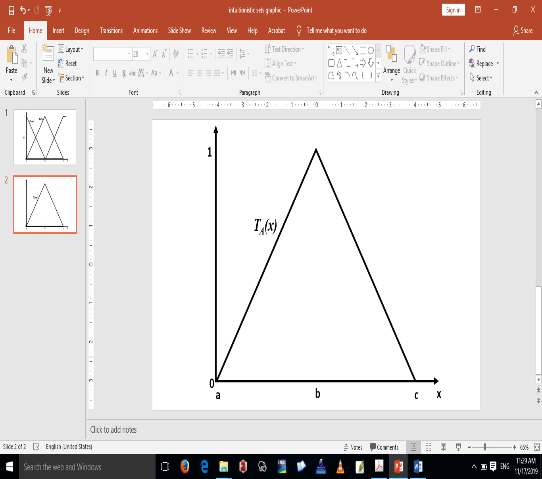


Figura 34: Taxonomía de técnicas para el descubrimiento de datos anómalos.

## Anexo 5. Taxonomía del concepto de incertidumbre

Figura 35: Taxonomía que representa el concepto de incertidumbre.

## Anexo 6. Evolución de la teoría neutrosófica



verdad (T), indeterminación (I) falsedad (F)

Figura 36: Evolución de la teoría neutrosófica.

## Anexo 7. Escenarios de pruebas para la comparación de los indicadores de calidad

Conjunto de escenarios de prueba para la comparación de los indicadores de calidad de los resúmenes lingüísticos. Se conforman 29 escenarios de prueba que cubren las diferentes situaciones durante el proceso de evaluación de resúmenes, ver Tabla 22.

Tabla 22: Clases de prueba para los escenarios de prueba.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cuantificador | Filtro | Sumarizador |
| Clase ER1 | Pocos | Bajo | Bajo |
| Clase ER2 | Pocos | Bajo | Medio |
| Clase ER3 | Pocos | Bajo | Alto |
| Clase ER4 | Pocos | Medio | Bajo |
| Clase ER5 | Pocos | Medio | Medio |
| Clase ER6 | Pocos | Medio | Alto |
| Clase ER7 | Pocos | Alto | Bajo |
| Clase ER8 | Pocos | Alto | Medio |
| Clase ER9 | Pocos | Alto | Alto |
| Clase ER10 | Aproximadamente el 50% | Bajo | Bajo |
| Clase ER11 | Aproximadamente el 50% | Bajo | Medio |
| Clase ER12 | Aproximadamente el 50% | Bajo | Alto |
| Clase ER13 | Aproximadamente el 50% | Medio | Bajo |
| Clase ER14 | Aproximadamente el 50% | Medio | Medio |
| Clase ER15 | Aproximadamente el 50% | Medio | Alto |
| Clase ER16 | Aproximadamente el 50% | Alto | Bajo |
| Clase ER17 | Aproximadamente el 50% | Alto | Medio |
| Clase ER18 | Aproximadamente el 50% | Alto | Alto |
| Clase ER19 | La mayoría | Bajo | Bajo |
| Clase ER20 | La mayoría | Bajo | Medio |
| Clase ER21 | La mayoría | Bajo | Alto |
| Clase ER22 | La mayoría | Medio | Bajo |
| Clase ER23 | La mayoría | Medio | Medio |
| Clase ER24 | La mayoría | Medio | Alto |
| Clase ER25 | La mayoría | Alto | Bajo |
| Clase ER26 | La mayoría | Alto | Medio |
| Clase ER27 | La mayoría | Alto | Alto |
| Clase ER28 | La mayoría | Porosa | Porosa |
| Clase ER29 | La mayoría | Porosa | Porosa |

En la Figura 37 se representa el dominio de la variable lingüística de los cuantificadores empleados en las pruebas.

Figura 37: Variable lingüística de los cuantificadores usados en las pruebas.

Tabla 23: Ejemplo de objetos de la clase de prueba ER1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Clase de prueba: ER1** | | | | | | | | | | | | |
| Objetos | μM(x) Pertenencia | | | | **M(x) Indeterminación | | | | σM(x) No pertenencia | | | |
| R1 | R2 | S1 | S2 | R1 | R2 | S1 | S2 | R1 | R2 | R1 | S2 |
| 1 | 0.15 | 0.23 | 0.23 | 0.17 | 0.80 | 0.72 | 0.72 | 0.78 | 0.72 | 0.65 | 0.65 | 0.71 |
| 2 | 0.05 | 0.13 | 0.2 | 0.17 | 0.89 | 0.82 | 0.75 | 0.78 | 0.81 | 0.74 | 0.68 | 0.71 |
| 3 | 0.09 | 0.1 | 0.16 | 0.12 | 0.86 | 0.85 | 0.79 | 0.83 | 0.77 | 0.77 | 0.71 | 0.75 |
| 4 | 0.02 | 0.06 | 0.11 | 0.19 | 0.92 | 0.88 | 0.84 | 0.76 | 0.83 | 0.80 | 0.76 | 0.69 |
| 5 | 0.18 | 0.19 | 0.22 | 0.03 | 0.77 | 0.76 | 0.73 | 0.91 | 0.70 | 0.69 | 0.66 | 0.82 |
| 6 | 0.15 | 0.33 | 0.03 | 0 | 0.80 | 0.63 | 0.91 | 0.94 | 0.72 | 0.57 | 0.82 | 0.85 |
| 7 | 0.1 | 0.05 | 0.25 | 0 | 0.85 | 0.89 | 0.71 | 0.94 | 0.77 | 0.81 | 0.64 | 0.85 |
| 8 | 0.65 | 0.7 | 0.88 | 0.65 | 0.33 | 0.28 | 0.11 | 0.33 | 0.30 | 0.26 | 0.10 | 0.30 |
| 9 | 0.64 | 0.69 | 0.76 | 1 | 0.34 | 0.29 | 0.23 | 0.00 | 0.31 | 0.26 | 0.20 | 0.00 |
| 10 | 0.76 | 0.71 | 0.99 | 0.98 | 0.23 | 0.27 | 0.01 | 0.02 | 0.20 | 0.25 | 0.01 | 0.02 |

## Anexo 8. Resumen de bases de datos empleadas en la experimentación

En la Tabla 24 se muestra un resumen de estas bases de datos.

Tabla 24: Descripción de las bases de datos.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descripción de las bases de datos** | | Cantidad de | | Cantidad de atributos de tipo | |
| instancias | atributos | Numérico | Nominal |
| **Bases de datos sobre proyectos** | | | | | |
| BD\_gp\_eval\_proy\_hard |  | 229 | 8 | 8 | 0 |
| BD\_gp\_eval\_proy\_integral |  | 8 430 | 19 | 7 | 12 |
| BD\_gp\_indicadores |  | 174 | 14 | 2 | 12 |
| BD\_gp\_competencias |  | 20 574 | 12 | 2 | 11 |
| BD\_gp\_rrhh |  | 3931 | 9 | 0 | 9 |
| BD\_gp\_requerimientos |  | 1 310 | 7 | 7 | 0 |
| BD\_gp\_toma\_decisiones |  | 1 011 | 17 | 9 | 8 |
| **Bases de datos sobre auditorías** | | | | | |
| BD\_audit\_sinOrganismos |  | 1 778 | 16 | 4 | 14 |
| BD\_audit\_sinProvMinistInstit |  | 1 778 | 16 | 4 | 14 |
| **Bases de datos médicas** | | | | | |
| BD\_med\_congenito |  | 464 | 20 | 20 | 0 |
| BD\_med\_valvular |  | 304 | 17 | 17 | 0 |
| BD\_med\_covidCuba |  | 18 151 | 6 | 6 | 0 |

## Anexo 9. Lenguajes Naturales Controlados

### LNC para resúmenes en español

<Resumen lingüístico>::= <D> <conector descriptor de datos > <Q> <conector cuantificador> <y’> <conector filtros> <R> <conector sumarizador> <S> |

<Q> <conector de género> <y’> <conector filtros> <R> <conector sumarizador> <S> |

<Q> <conector de género> <y’> <conector sumarizador> <S>

<D>::= “casi la totalidad” | “la mayoría” | “muchos”| “aproximadamente la mitad” | “algunos” | “pocos” | “muy pocos”

<conector de género>::= “de los” | “de las”

<conector descriptor de datos>::= “registros en la base de datos, reportan que”

<conector cuantificador>::= “de las veces, “

<conector filtros>::= “que cumplen” | “que tienen” | “que son”

<conector sumarizador>::= “tienen” | “se cumple que”

<y’>::= <sujeto>

<sujeto>::= <frase simple>

<Q>::= <cuantificador lingüístico> | <cuantificador numérico>| <cuantificador mixto>| <cuantificador por ciento>

<cuantificador lingüístico>::= “casi la totalidad” | “la mayoría” | “muchos” | “aproximadamente la mitad” | “algunos” | “pocos” | “muy pocos”

<cuantificador numérico>::= “más del 95%” | “alrededor del 85%” | “alrededor del 75%” | “aproximadamente el 50%” | “cercano al 33%” | “menos del 17%” | “menos del 5%”

<cuantificador mixto>::= “muy pocos (menos del 5%)” | “pocos (alrededor del 15%)” | “algunos (alrededor del 33%)” | “aproximadamente la mitad” |

“muchos (cerca del 65%)” | “la mayoría (alrededor del 83%)” | “casi la totalidad”

<cuantificador porciento>::= <conector porciento> <valor numérico en porciento>

< conector porciento>::= “en el”

<S>::= <frase>

<R>::= <frase>

<frase>::= <frase> <operador lógico> < frase simple> | <frase simple>

<operador lógico>::= <conjunción> | <disyunción>

Ejemplo de resúmenes y su representación en la gramática (Tabla 25):

1. “Muchos registros en la base de datos, reportan que en el 100.0% de las veces, las embarazadas cardiópatas que tienen ausencia de soplo holosistólico tienen ausencia de cianosis.”
2. “La mayoría de los registros en la base de datos, reportan que en el 90.404% de las veces, los proyectos que tienen muy alta calidad del dato tienen muy baja cantidad de competencias con baja evaluación.”

Tabla 25: Ejemplo de aplicación del lenguaje controlado en resúmenes lingüísticos.

|  |  |  |
| --- | --- | --- |
| Partes del LNC | Resumen 1: sobre  embarazadas cardiópatas | Resumen 2: sobre  evaluación de proyectos |
| <*D*> | muchos | la mayoría |
| <conector de género> | de los | de los |
| <conector descriptor de datos> | registros en la base de datos, reportan que en el | |
| <*Q*> | 100.0% | 90.404% |
| <conector cuantificador> | de las veces, | de las veces, |
| <*y’*> | las embarazadas cardiópatas | los proyectos |
| <conector filtros> | que tienen | que tienen |
| <*R*> | ausencia de soplo holosistólico | muy alta calidad del dato |
| <conector sumarizador> | tienen | tienen |
| <*S*> | ausencia de cianosis | muy baja cantidad de competencias con baja evaluación |

### LNC para resúmenes en inglés

<Resumen lingüístico>::= <D> <conector descriptor de datos > <Q> <conector cuantificador> <y’> <conector filtros> <R> <conector sumarizador> <S> |

<Q> <y’> <conector filtros> <R> <conector sumarizador> <S> |

<Q> <y’> <conector sumarizador> <S>

<D>::= “almost all” | “most” | “many” | “around a half of” | “some” | “few” | “very few”

<conector descriptor de datos>::= “records show, that”

<conector cuantificador>::= “of times,”

<conector filtros>::= “with”

<conector sumarizador>::= “have”

<y’>::= <sujeto>

<sujeto>::= <frase simple>

<Q>::= <cuantificador lingüístico> | <cuantificador numérico> | <cuantificador mixto> | <cuantificador por ciento>

<cuantificador lingüístico>::= “almost all” | “most” | “many”| “around a half of” | “some” |

“few” | “very few”

<cuantificador numérico>::= “more than 95% of” | “around of 85% of” | “around of 75% of” | “approximately el 50%” | “close to 33%” | “less than 17%” | “less than 5%”

<cuantificador mixto>::= “very few (less than 5%)” | “few (around of 15%)” | “some of (close to 33%)” | “around a half of” | “many (close to 65%)” | “most of (around of 83%)” | “almost all of”

<cuantificador porciento>::= <conector porciento> <valor numérico en porciento>

< conector porciento>::= “in”

<S>::= <frase>

<R>::= <frase>

<frase>::= <frase> <operador lógico> <frase simple> | <frase simple>

<operador lógico>::= <conjunción> | <disyunción>

Ejemplo: “*Very few records show, that 94.0% of times projects with regular data quality, have bad productivity”*.

Donde:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *<D>:* | Very few | *<conector descriptor de datos >:* | | records show, that |
| *<Q>:* | 94.0% | *<conector cuantificador>:* | | of times, |
| *<y’>:* | projects | | *<conector filtros>:* | with |
| *<R>:* | regular data quality | | *<conector sumarizador>:* | have |
| *<S>:* | bad productivity | | | |

### LNC para resúmenes en japonés

<Resumen lingüístico>::= <conector descriptor de datos> <D> <conector cuantificador D> <Q> <conector cuantificador Q> <R> の <y’> は <S> と報告しています。|

<S> の <y’> の <Q> は、<R>です。|

<Q> の <y’> は <S> です。

<D>[[13]](#footnote-14)::= “ほとんど” | “たくさん” | “約半分” | “いくつか” | “少数” |

“非常に少数”

<Q>::= <cuantificador lingüístico> | <cuantificador por ciento>

<cuantificador lingüístico>[[14]](#footnote-15)::= “ほとんど” | “たくさん” | “約半分” |

“いくつか” | “少数” | “非常に少数”

<cuantificador porciento>::= <Valor numérico en porciento>

<conector descriptor de datos>::= “データベース内の” [[15]](#footnote-16)

<conector cuantificador D>::= “記録が、” [[16]](#footnote-17)

<conector cuantificador Q>::= “% の確率で、” [[17]](#footnote-18)

<y’>::= <sujeto>

<sujeto>::= <frase simple>

<S>::= <frase>

<R>::= <frase>

<frase>::= <frase> | <partículas> | <frase simple>

<partículas>::= “は” | “が” | “と” | “の”

**Ejemplo 1:** データベース内の非常に数少ない記録が、94.0％の確率で、データ品質の悪いのプロジェクトは悪い評価と報告しています。

LNC:<conector descriptor de datos> <D> <conector cuantificador D> <Q> <conector cuantificador Q> <R> の<y’> は <S> と報告しています。

Traducción literal: “Muy pocos registros en la base de datos, reportan que en el 94.0% de las veces, los proyectos que tienen mala calidad del dato tienen mala evaluación”.

Donde:

|  |  |
| --- | --- |
| *<R>:* データ品質の悪い | → mala calidad de los datos |
| *<y’>:* プロジェクト | → proyectos |
| *<S>:* 悪い評価 | → mala evaluación |
| *<D>:*非常に数少ない*<Q>: 94.0* | → muy pocos |

**Ejemplo 2:** データ品質の悪いプロジェクトのほとんどは、悪い評価です。

LNC :<S> の <y’> の <Q>は、<R>です。 (Estructura: *QRy’s are S*)

Traducción literal: “*La mayoría de los proyectos con datos de mala calidad tienen una mala calificación”.*

Donde:

|  |  |
| --- | --- |
| *<R>:* データ品質の悪い | mala calidad de los datos |
| *<y’>:* プロジェクト | proyectos |
| *<S>:* 悪い評価 | mala evaluación |
| *<Q>:* ほとんど | mayoría |

**Ejemplo 3:** ほとんどのプロジェクトは悪い評価です。

LNC: <Q> の <y’> は <S> です。 (Estructura: *Qy’s are S*)

Traducción literal: “*La mayoría de los proyectos tienen una mala calificación.”*

Donde:

*<Q>:* ほとんど *<y’>:* プロジェクト *<S>:* 悪い評価

### LNC para resúmenes en árabe

<Resumen lingüístico>::= <S> <conector sumarizador> <R> <conector filtros> <y’> <conector cuantificador> <Q> <conector descriptor de datos> <D> |

<S> <conector sumarizador> <R> <conector filtros> <y’> <Q> | <S> <conector sumarizador> <y’> <Q>

<D>::= “*أغلبية*” | “*الكثير*” | “*ما يقارب النصف من*” | “*بعض*” | “*القليل*” | “قليل جدا”

<conector descriptor de datos>::= “*السجلات تظهر أن*”

<conector cuantificador>::= “*من المرات،*”

<conector filtros>::= “*التي لديها*”

<conector sumarizador>::= “*لديه*”

<y’>::= <sujeto>

<sujeto>::= <frase simple>

<Q>::= <cuantificador lingüístico> | <cuantificador numérico>| <cuantificador mixto>| <cuantificador por ciento>

<cuantificador lingüístico>::= “*تقريباً كل*” | “*أغلبية*” | “*الكثير من*”| “*تقريباً نصف*” | “*بعض*” | “*القليل من*” | “*قليل جداً من*”

<cuantificador numérico>::= “*أكثر من 95%*” | “*حوالي 85%*” | “*حوالي 75%*” |”*حوالي 50%*” | “ *ما يقارب 33%* “ | “*أقل من 17%*” | “*أقل من 5%*”

<cuantificador mixto>::= “*قليل جداً من* (*أقل من 5%*)” | “*القليل* (*حوالي 15%*)” | “*بعض* (*حوالي 33%*)” | “*تقريباً النصف*” | “*الكثير* (*ما يقارب 65%*)” | “*أغلبية* (*حوالي 83%*)” | “*تقريباً كل*”

<cuantificador porciento>::= <conector porciento> <valor numérico en porciento>

< conector porciento>::= “*في*”

*<*S>::= <frase>

<R>::= <frase>

<frase>::= <frase> <operador lógico> <frase simple> | <frase simple>

<operador lógico>::= <conjunción> | <disyunción>

**Ejemplo:** قليل جداً من السجلات في قاعدة البيانات تشير إلى أن هناك نسبة 94.0% أن المشاريع التي لديها جودة بيانات سيئة لديها تقييم سيء

Traducción: “Muy pocos registros en la base de datos, reportan que en el 94.0% de las veces, los proyectos que tienen mala calidad del dato tienen mala evaluación.”

Donde:

|  |  |
| --- | --- |
| *<R>:* جودة بيانات سيئة | → mala calidad del dato |
| *<y’>:* مشاريع | → proyectos |
| *<S>:* تقييم سيء | → mala evaluación |
| *<D>:* قليل جدا  *<Q>:* %94.0 | → muy pocos  → 94.0% |

## Anexo 10. Ejemplos de diccionarios

Estos son fragmentos de diccionarios construidos en los idiomas utilizados para la validación de los resúmenes lingüísticos con enfoque multilingüe, en el entorno de gestión de proyectos.

### Fragmento de diccionario de LNC en español para el dominio de gestión de proyectos

fuzzy\_001:

name: icd

type: :symbolic

n\_sets: 0

description: el Índice de Calidad del Dato

label:

icd\_1: 1

icd\_2: 2

icd\_3: 3

translation:

icd\_1: buena calidad del dato

icd\_2: mala calidad del dato

icd\_3: calidad del dato regular

short\_translation:

icd\_1: '1'

icd\_2: '2'

icd\_3: '3'

### Fragmento de diccionario de LNC en ingés para el dominio de gestión de proyectos

fuzzy\_001:

name: icd

type: :symbolic

n\_sets: 0

description: data quality index

label:

icd\_1: 1

icd\_2: 2

icd\_3: 3

translation:

icd\_1: good data quality

icd\_2: poor data quality

icd\_3: regular data quality

short\_translation:

icd\_1: '1'

icd\_2: '2'

icd\_3: '3'

### Fragmento de diccionario de LNC en japonés para el dominio de gestión de proyectos

fuzzy\_001:

name: icd

type: :symbolic

n\_sets: 0

description: データ品質指数

label:

icd\_1: 1

icd\_2: 2

icd\_3: 3

translation:

icd\_1: データ品質の良い

icd\_2: データ品質の悪い

icd\_3: データ品質が中級程度

short\_translation:

icd\_1: '1'

icd\_2: '2'

icd\_3: '3'

### Fragmento de diccionario de LNC en árabe para el dominio de gestión de proyectos

fuzzy\_001:

name: icd

type: :symbolic

n\_sets: 0

description: مؤشر جودة البيانات

label:

icd\_1: 1

icd\_2: 2

icd\_3: 3

translation:

icd\_1: جودة بيانات جيدة

icd\_2: جودة بيانات رديئة

icd\_3: جودة بيانات متوسطة

short\_translation:

icd\_1: '1'

icd\_2: '2'

icd\_3: '3'

## Anexo 11. Especialistas que participaron en los casos de estudio.

Tabla 26: Especialistas que participaron en el caso de estudio sobre embarazadas cardiópatas.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nombre del experto | Años de experiencia | | Especialidad | Cargo | Centro de trabajo |
| Dr. Pedro Antonio Román Rubio | 32 | Cardiología | | Jefe de Servicio de Cardiopatía y Embarazos | Servicio de Cardiopatía y Embarazos, Instituto de Cardiología y Cirugía Cardiovascular |
| Dra. Susana Hernández García | 20 | Rehabilitación Cardiovascular | | Especialista Principal de Rehabilitación Cardiovascular | Instituto de Cardiología y Cirugía Cardiovascular |
| Dr. Román Vasallo Peraza | 5 | Cardiología | | Cardiólogo del Servicio de Cardiopatía y Embarazos | Instituto de Cardiología y Cirugía Cardiovascular |
| Dra. Yilena Gil Jiménez | 5 | Obstetricia | | Obstetra del Servicio de Cardiopatía y Embarazos | Hospital Docente Ginecobstétrico “Ramón González Coro” |

Tabla 27: Expertos que participaron en el caso de estudio sobre gestión de proyectos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Expertos | Años experiencia | Especialidad | Cargo | Afiliación |
| MSc. Daisy Oropesa Méndez | 21 | Informática | Docentes  e  investigadores | UCI |
| Dr.C Roberto Delgado Victore | 44 | Civil | UCI |
| Dr.C Surayne Torres López | 13 | Ciencias Informáticas | UCI |
| Dr.C Armando Pérez Fuentes | 35 | Deporte | UCI |
| Dr.C Pascual Verdecia | 35 | Minería | UCI |

## Anexo 12. Encuestas aplicadas a los especialistas de los casos de estudio

Tabla 28 Encuesta aplicada a los especialistas en el tratamiento de embarazadas cardiópatas.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criterios | | Especialistas | | | |
| E1 | E2 | E3 | E4 |
| C1 | ¿En qué grado los resúmenes lingüísticos que se obtienen muestran información correcta, cierta? \*\* | 3 | 4 | 3 | 3 |
| C2 | ¿En qué grado los resúmenes lingüísticos que se obtienen muestran información relevante, novedosa? \*\* | 2 | 3 | 3 | 3 |
| C3 | ¿En qué grado consideran los decisores que los resúmenes lingüísticos son interpretables? \* | 4 | 3 | 4 | 3 |
| C4 | ¿En qué medida el uso de los resúmenes lingüísticos permite a los especialistas médicos obtener información sobre el comportamiento de las afecciones cardiovasculares en los pacientes rápidamente? \*\* | 4 | 2 | 3 | 4 |
| C5 | ¿En qué medida el uso de resúmenes lingüísticos puede contribuir a ofrecer una mejor explicación sobre el diagnóstico y decisión del tratamiento a los pacientes? \*\* | 4 | 4 | 3 | 3 |
| C6 | ¿En qué grado consideran los especialistas médicos que los resúmenes lingüísticos son legibles y fáciles de leer? \* | 4 | 4 | 4 | 4 |
| C7 | ¿En qué grado los resúmenes lingüísticos le permiten realizar cambios en el tratamiento a los pacientes? \*\* | 3 | 3 | 2 | 3 |
| C8 | ¿En qué grado recomienda el uso de los resúmenes lingüísticos a otros especialistas médicos? | 4 | 4 | 4 | 4 |
| C9 | ¿En qué grado los resúmenes lingüísticos que se obtienen facilitan la toma de decisiones médicas? \*\* | 4 | 4 | 4 | 4 |
| C10 | ¿En qué grado se siente confiable usando los resúmenes lingüísticos durante el proceso de toma de decisiones médicas? \*\* | 3 | 4 | 3 | 3 |

\* Estos criterios permiten conocer las preferencias respecto al indicador “Interpretabilidad de resúmenes”.

\*\* Estos criterios permiten conocer las preferencias respecto al indicador “Facilidad para la toma de decisiones”.

Tabla 29: Encuesta aplicada a los expertos en gestión de proyectos.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Criterios | | Expertos | | | | |
| E1 | E2 | E3 | E4 | E5 |
| C1 | ¿En qué grado los resúmenes lingüísticos que se obtienen muestran información correcta, cierta? \*\* | 4 | 3 | 4 | 4 | 3 |
| C2 | ¿En qué grado los resúmenes lingüísticos que se obtienen muestran información relevante, novedosa? \*\* | 3 | 3 | 3 | 2 | 3 |
| C3 | ¿En qué grado consideran los decisores que los resúmenes lingüísticos son interpretables? \* | 4 | 4 | 4 | 4 | 3 |
| C4 | ¿En qué medida el uso de los resúmenes lingüísticos permite a los decisores obtener información sobre los proyectos rápidamente? \*\* | 3 | 4 | 4 | 2 | 4 |
| C5 | ¿En qué medida el uso de resúmenes lingüísticos puede contribuir a ofrecer una mejor explicación de las decisiones tomadas? \*\* | 4 | 4 | 3 | 3 | 4 |
| C6 | ¿En qué medida el uso de los resúmenes lingüísticos mejora la eficiencia de seguimiento y control de los proyectos? \*\* | 3 | 4 | 3 | 2 | 3 |
| C7 | ¿En qué grado consideran los decisores que los resúmenes lingüísticos son legibles o entendibles, y fáciles de leer? \* | 3 | 4 | 3 | 3 | 4 |
| C8 | ¿En qué grado los resúmenes lingüísticos le permiten realizar ajustes rápidos en la planificación de los proyectos? \*\* | 3 | 3 | 3 | 2 | 3 |
| C9 | ¿En qué grado recomienda el uso de los resúmenes lingüísticos a otros decisores? | 4 | 4 | 4 | 3 | 4 |
| C10 | ¿En qué grado los resúmenes lingüísticos que se obtienen facilitan la toma de decisiones en los proyectos? \*\* | 4 | 4 | 4 | 4 | 4 |
| C11 | ¿En qué grado se siente confiable usando los resúmenes lingüísticos durante el proceso de toma de decisiones? \*\* | 3 | 3 | 4 | 2 | 3 |

\* Estos criterios permiten conocer las preferencias respecto al indicador “Interpretabilidad de resúmenes”.

\*\* Estos criterios permiten conocer las preferencias respecto al indicador “Facilidad para la toma de decisiones”.

## Anexo 13. Resultado de los algoritmos.

Tabla 30: Resultado de los algoritmos respecto a la media y desviación estándar de los indicadores en las bases de datos sobre gestión de proyectos.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Media y StdDev** | | *T1* | *T3* | *T4* | *T5* | *Te1a* | *Te1b* | *Te2a* | *Te3* | *Te4* | *Te5* |
| **BD\_gp\_eval\_proy\_hard** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.00 | 0.90 | 0.60 | 0.50 | 0.61 | 1.00 | 0.02 | 0.35 | 0.97 | 0.97 |
| StdDev | 0.00 | 0.14 | 0.34 | 0.00 | 0.27 | 0.00 | 0.00 | 0.26 | 0.01 | 0.00 |
| LPA\_LDS | Media | 0.01 | 0.91 | 0.70 | 0.50 | 0.65 | 1.00 | 0.02 | 0.33 | 0.97 | 0.97 |
| StdDev | 0.00 | 0.07 | 0.19 | 0.00 | 0.30 | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 |
| Apriori\_LDS | Media | 0.00 | 0.84 | 0.31 | 0.50 | 0.53 | 1.00 | 0.02 | 0.60 | 0.97 | 0.97 |
| StdDev | 0.00 | 0.13 | 0.22 | 0.00 | 0.29 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 |
| ACO\_LDS | Media | 0.07 | 0.80 | 0.75 | 0.37 | 0.11 | 0.88 | 0.06 | 0.20 | 0.91 | 0.97 |
| StdDev | 0.26 | 0.34 | 0.35 | 0.13 | 0.25 | 0.12 | 0.04 | 0.20 | 0.06 | 0.00 |
| GA\_LDS | Media | 1.00 | 0.86 | 0.70 | 0.50 | 0.00 | 1.00 | 0.02 | 0.21 | 0.96 | 0.97 |
| StdDev | 0.04 | 0.13 | 0.17 | 0.04 | 0.00 | 0.00 | 0.00 | 0.18 | 0.00 | 0.00 |
| PCA\_LDS | Media | 0.01 | 0.84 | 0.31 | 0.50 | 0.53 | 1.00 | 0.02 | 0.60 | 0.97 | 0.97 |
| StdDev | 0.00 | 0.13 | 0.22 | 0.00 | 0.29 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 |
| **BD\_gp\_eval\_proy\_integral** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.18 | 0.95 | 0.59 | 0.50 | 0.76 | 0.89 | 0.12 | 0.37 | 0.75 | 0.97 |
| StdDev | 0.38 | 0.12 | 0.34 | 0.00 | 0.21 | 0.09 | 0.09 | 0.35 | 0.21 | 0.00 |
| LPA\_LDS | Media | 0.00 | 0.88 | 0.76 | 0.50 | 0.82 | 0.88 | 0.13 | 0.17 | 0.65 | 0.97 |
| StdDev | 0.00 | 0.19 | 0.23 | 0.00 | 0.21 | 0.08 | 0.08 | 0.16 | 0.26 | 0.00 |
| Apriori\_LDS | Media | 0.21 | 0.87 | 0.21 | 0.47 | 0.75 | 0.95 | 0.07 | 0.73 | 0.88 | 0.96 |
| StdDev | 0.39 | 0.16 | 0.23 | 0.09 | 0.23 | 0.03 | 0.03 | 0.17 | 0.05 | 0.04 |
| ACO\_LDS | Media | 0.03 | 0.62 | 0.55 | 0.40 | 0.13 | 0.69 | 0.26 | 0.17 | 0.61 | 0.97 |
| StdDev | 0.07 | 0.44 | 0.45 | 0.12 | 0.21 | 0.08 | 0.04 | 0.16 | 0.07 | 0.00 |
| GA\_LDS | Media | 0.94 | 0.96 | 0.80 | 0.50 | 0.01 | 0.96 | 0.06 | 0.17 | 0.86 | 0.97 |
| StdDev | 0.11 | 0.08 | 0.20 | 0.00 | 0.00 | 0.07 | 0.07 | 0.18 | 0.19 | 0.00 |
| PCA\_LDS | Media | 0.22 | 0.87 | 0.21 | 0.47 | 0.73 | 0.95 | 0.07 | 0.74 | 0.89 | 0.96 |
| StdDev | 0.39 | 0.16 | 0.23 | 0.09 | 0.23 | 0.03 | 0.03 | 0.17 | 0.05 | 0.04 |
| **BD\_gp\_indicadores** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.09 | 0.99 | 0.60 | 0.50 | 0.71 | 0.86 | 0.16 | 0.34 | 0.66 | 0.97 |
| StdDev | 0.28 | 0.02 | 0.29 | 0.00 | 0.23 | 0.10 | 0.10 | 0.28 | 0.21 | 0.00 |
| LPA\_LDS | Media | 0.05 | 0.89 | 0.75 | 0.50 | 0.66 | 0.82 | 0.19 | 0.16 | 0.57 | 0.97 |
| StdDev | 0.22 | 0.18 | 0.26 | 0.00 | 0.15 | 0.11 | 0.11 | 0.21 | 0.30 | 0.00 |
| Apriori\_LDS | Media | 0.61 | 0.95 | 0.21 | 0.34 | 0.75 | 0.95 | 0.07 | 0.90 | 0.89 | 0.90 |
| StdDev | 0.42 | 0.09 | 0.14 | 0.17 | 0.27 | 0.05 | 0.05 | 0.14 | 0.10 | 0.14 |
| ACO\_LDS | Media | 0.06 | 0.27 | 0.24 | 0.37 | 0.10 | 0.29 | 0.25 | 0.10 | 0.23 | 0.97 |
| StdDev | 0.21 | 0.39 | 0.35 | 0.13 | 0.24 | 0.25 | 0.21 | 0.10 | 0.20 | 0.00 |
| GA\_LDS | Media | 0.92 | 0.94 | 0.81 | 0.49 | 0.02 | 0.87 | 0.12 | 0.14 | 0.75 | 0.97 |
| StdDev | 0.15 | 0.13 | 0.27 | 0.05 | 0.09 | 0.17 | 0.12 | 0.23 | 0.26 | 0.00 |
| PCA\_LDS | Media | 0.61 | 0.95 | 0.21 | 0.34 | 0.75 | 0.95 | 0.07 | 0.90 | 0.89 | 0.90 |
| StdDev | 0.42 | 0.09 | 0.14 | 0.17 | 0.27 | 0.05 | 0.05 | 0.14 | 0.10 | 0.14 |
| **BD\_gp\_competencias** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.09 | 0.99 | 0.53 | 0.50 | 0.65 | 0.86 | 0.15 | 0.43 | 0.70 | 0.97 |
| StdDev | 0.27 | 0.01 | 0.37 | 0.00 | 0.27 | 0.10 | 0.10 | 0.36 | 0.22 | 0.00 |
| LPA\_LDS | Media | 0.00 | 0.84 | 0.55 | 0.50 | 0.54 | 0.84 | 0.18 | 0.29 | 0.65 | 0.97 |
| StdDev | 0.00 | 0.26 | 0.35 | 0.00 | 0.26 | 0.11 | 0.11 | 0.25 | 0.22 | 0.00 |
| Apriori\_LDS | Media | 0.08 | 0.87 | 0.12 | 0.49 | 0.64 | 0.79 | 0.23 | 0.65 | 0.58 | 0.97 |
| StdDev | 0.24 | 0.11 | 0.09 | 0.05 | 0.26 | 0.08 | 0.08 | 0.16 | 0.18 | 0.00 |
| ACO\_LDS | Media | 0.02 | 0.82 | 0.24 | 0.38 | 0.08 | 0.36 | 0.61 | 0.68 | 0.09 | 0.97 |
| StdDev | 0.11 | 0.27 | 0.23 | 0.13 | 0.25 | 0.11 | 0.10 | 0.28 | 0.14 | 0.00 |
| GA\_LDS | Media | 0.87 | 0.93 | 0.82 | 0.46 | 0.00 | 0.73 | 0.22 | 0.09 | 0.47 | 0.97 |
| StdDev | 0.14 | 0.09 | 0.15 | 0.10 | 0.00 | 0.22 | 0.12 | 0.13 | 0.34 | 0.00 |
| PCA\_LDS | Media | 0.08 | 0.87 | 0.12 | 0.49 | 0.64 | 0.78 | 0.24 | 0.71 | 0.61 | 0.97 |
| StdDev | 0.24 | 0.11 | 0.09 | 0.05 | 0.25 | 0.09 | 0.09 | 0.13 | 0.15 | 0.00 |
| **BD\_gp\_requerimientos** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.45 | 0.99 | 0.33 | 0.50 | 0.87 | 0.99 | 0.03 | 0.66 | 0.96 | 0.97 |
| StdDev | 0.53 | 0.03 | 0.38 | 0.00 | 0.24 | 0.00 | 0.00 | 0.38 | 0.00 | 0.00 |
| LPA\_LDS | Media | 0.00 | 0.95 | 0.81 | 0.50 | 0.87 | 1.00 | 0.03 | 0.19 | 0.96 | 0.97 |
| StdDev | 0.00 | 0.05 | 0.15 | 0.00 | 0.20 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 |
| Apriori\_LDS | Media | 0.86 | 0.98 | 0.26 | 0.25 | 0.96 | 0.99 | 0.03 | 0.96 | 0.96 | 0.79 |
| StdDev | 0.35 | 0.08 | 0.12 | 0.18 | 0.16 | 0.00 | 0.00 | 0.11 | 0.00 | 0.23 |
| ACO\_LDS | Media | 0.13 | 0.44 | 0.38 | 0.37 | 0.10 | 0.66 | 0.05 | 0.17 | 0.69 | 0.97 |
| StdDev | 0.34 | 0.46 | 0.45 | 0.13 | 0.25 | 0.39 | 0.04 | 0.26 | 0.39 | 0.00 |
| GA\_LDS | Media | 0.62 | 0.95 | 0.59 | 0.39 | 0.30 | 0.90 | 0.05 | 0.60 | 0.92 | 0.97 |
| StdDev | 0.48 | 0.12 | 0.50 | 0.13 | 0.46 | 0.12 | 0.04 | 0.35 | 0.06 | 0.00 |
| PCA\_LDS | Media | 0.86 | 0.98 | 0.26 | 0.25 | 0.96 | 0.99 | 0.03 | 0.96 | 0.96 | 0.79 |
| StdDev | 0.35 | 0.08 | 0.12 | 0.18 | 0.16 | 0.00 | 0.00 | 0.11 | 0.00 | 0.23 |
| **BD\_gp\_toma\_decisiones** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.01 | 0.90 | 0.70 | 0.50 | 0.73 | 0.97 | 0.05 | 0.22 | 0.89 | 0.97 |
| StdDev | 0.00 | 0.15 | 0.37 | 0.00 | 0.27 | 0.06 | 0.06 | 0.28 | 0.17 | 0.00 |
| LPA\_LDS | Media | 0.00 | 0.93 | 0.83 | 0.50 | 0.70 | 0.93 | 0.09 | 0.15 | 0.74 | 0.97 |
| StdDev | 0.00 | 0.06 | 0.15 | 0.00 | 0.24 | 0.09 | 0.09 | 0.19 | 0.29 | 0.00 |
| Apriori\_LDS | Media | 0.00 | 0.83 | 0.28 | 0.50 | 0.46 | 0.98 | 0.04 | 0.60 | 0.92 | 0.97 |
| StdDev | 0.00 | 0.12 | 0.13 | 0.00 | 0.21 | 0.01 | 0.01 | 0.08 | 0.05 | 0.00 |
| ACO\_LDS | Media | 0.12 | 0.24 | 0.23 | 0.38 | 0.17 | 0.88 | 0.06 | 0.02 | 0.91 | 0.97 |
| StdDev | 0.33 | 0.16 | 0.16 | 0.13 | 0.35 | 0.13 | 0.04 | 0.00 | 0.06 | 0.00 |
| GA\_LDS | Media | 0.98 | 0.98 | 0.96 | 0.50 | 0.01 | 0.99 | 0.03 | 0.03 | 0.95 | 0.97 |
| StdDev | 0.03 | 0.04 | 0.05 | 0.00 | 0.00 | 0.02 | 0.02 | 0.03 | 0.02 | 0.00 |
| PCA\_LDS | Media | 0.00 | 0.83 | 0.28 | 0.50 | 0.44 | 0.98 | 0.04 | 0.60 | 0.92 | 0.97 |
| StdDev | 0.00 | 0.12 | 0.13 | 0.00 | 0.22 | 0.01 | 0.01 | 0.08 | 0.05 | 0.00 |
| **BD\_gp\_rrhh** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.01 | 0.95 | 0.65 | 0.50 | 0.72 | 0.83 | 0.19 | 0.27 | 0.59 | 0.97 |
| StdDev | 0.00 | 0.15 | 0.24 | 0.00 | 0.23 | 0.08 | 0.08 | 0.21 | 0.19 | 0.00 |
| LPA\_LDS | Media | 0.00 | 0.75 | 0.61 | 0.50 | 0.68 | 0.84 | 0.17 | 0.23 | 0.60 | 0.97 |
| StdDev | 0.00 | 0.25 | 0.29 | 0.00 | 0.21 | 0.07 | 0.07 | 0.18 | 0.23 | 0.00 |
| Apriori\_LDS | Media | 0.01 | 0.84 | 0.17 | 0.50 | 0.56 | 0.84 | 0.17 | 0.62 | 0.69 | 0.97 |
| StdDev | 0.00 | 0.18 | 0.18 | 0.00 | 0.27 | 0.02 | 0.02 | 0.04 | 0.03 | 0.00 |
| ACO\_LDS | Media | 0.07 | 0.31 | 0.31 | 0.38 | 0.08 | 0.35 | 0.12 | 0.04 | 0.35 | 0.97 |
| StdDev | 0.24 | 0.38 | 0.38 | 0.13 | 0.22 | 0.34 | 0.12 | 0.04 | 0.33 | 0.00 |
| GA\_LDS | Media | 0.74 | 0.73 | 0.58 | 0.43 | 0.01 | 0.77 | 0.21 | 0.17 | 0.46 | 0.97 |
| StdDev | 0.33 | 0.21 | 0.23 | 0.12 | 0.05 | 0.12 | 0.09 | 0.14 | 0.28 | 0.00 |
| PCA\_LDS | Media | 0.01 | 0.84 | 0.16 | 0.50 | 0.61 | 0.83 | 0.18 | 0.65 | 0.71 | 0.97 |
| StdDev | 0.00 | 0.18 | 0.18 | 0.00 | 0.25 | 0.02 | 0.02 | 0.05 | 0.03 | 0.00 |

Tabla 31: Resultado de los algoritmos respecto a la media y desviación estándar de los indicadores en las bases de datos sobre auditorías.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Media y StdDev** | | *T1* | *T3* | *T4* | *T5* | *Te1a* | *Te1b* | *Te2a* | *Te3* | *Te4* | *Te5* |
| **BD\_audit\_sinOrganismos** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.08 | 1.00 | 0.81 | 0.50 | 0.91 | 0.97 | 0.05 | 0.18 | 0.91 | 0.97 |
| StdDev | 0.27 | 0.01 | 0.35 | 0.00 | 0.17 | 0.08 | 0.08 | 0.34 | 0.15 | 0.00 |
| LPA\_LDS | Media | 0.01 | 0.96 | 0.93 | 0.50 | 0.82 | 0.94 | 0.08 | 0.04 | 0.88 | 0.97 |
| StdDev | 0.00 | 0.07 | 0.08 | 0.00 | 0.25 | 0.10 | 0.09 | 0.06 | 0.14 | 0.00 |
| Apriori\_LDS | Media | 0.39 | 0.95 | 0.17 | 0.33 | 0.87 | 0.89 | 0.13 | 0.88 | 0.80 | 0.85 |
| StdDev | 0.41 | 0.11 | 0.14 | 0.18 | 0.20 | 0.09 | 0.08 | 0.17 | 0.15 | 0.20 |
| ACO\_LDS | Media | 0.11 | 0.39 | 0.29 | 0.36 | 0.11 | 0.56 | 0.05 | 0.22 | 0.59 | 0.97 |
| StdDev | 0.31 | 0.46 | 0.42 | 0.12 | 0.27 | 0.41 | 0.07 | 0.36 | 0.43 | 0.00 |
| GA\_LDS | Media | 0.99 | 1.00 | 0.97 | 0.50 | 0.00 | 0.99 | 0.03 | 0.03 | 0.96 | 0.97 |
| StdDev | 0.01 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 |
| PCA\_LDS | Media | 0.40 | 0.95 | 0.17 | 0.33 | 0.87 | 0.88 | 0.13 | 0.89 | 0.80 | 0.85 |
| StdDev | 0.41 | 0.11 | 0.14 | 0.18 | 0.20 | 0.09 | 0.09 | 0.16 | 0.14 | 0.20 |
| **BD\_audit\_sinProvMinistInstit** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.04 | 0.99 | 0.76 | 0.50 | 0.84 | 0.94 | 0.08 | 0.22 | 0.86 | 0.97 |
| StdDev | 0.17 | 0.02 | 0.36 | 0.00 | 0.20 | 0.10 | 0.10 | 0.35 | 0.19 | 0.00 |
| LPA\_LDS | Media | 0.00 | 1.00 | 0.95 | 0.50 | 0.69 | 0.96 | 0.06 | 0.07 | 0.91 | 0.97 |
| StdDev | 0.00 | 0.00 | 0.07 | 0.00 | 0.27 | 0.08 | 0.08 | 0.07 | 0.12 | 0.00 |
| Apriori\_LDS | Media | 0.39 | 0.95 | 0.17 | 0.33 | 0.87 | 0.89 | 0.13 | 0.88 | 0.80 | 0.85 |
| StdDev | 0.41 | 0.11 | 0.14 | 0.18 | 0.20 | 0.09 | 0.08 | 0.17 | 0.15 | 0.20 |
| ACO\_LDS | Media | 0.10 | 0.17 | 0.17 | 0.38 | 0.14 | 0.15 | 0.19 | 0.01 | 0.10 | 0.97 |
| StdDev | 0.28 | 0.37 | 0.37 | 0.13 | 0.34 | 0.21 | 0.25 | 0.01 | 0.18 | 0.00 |
| GA\_LDS | Media | 0.99 | 1.00 | 0.97 | 0.50 | 0.00 | 0.99 | 0.03 | 0.03 | 0.96 | 0.97 |
| StdDev | 0.02 | 0.01 | 0.04 | 0.00 | 0.00 | 0.02 | 0.01 | 0.04 | 0.02 | 0.00 |
| PCA\_LDS | Media | 0.40 | 0.95 | 0.17 | 0.33 | 0.87 | 0.88 | 0.13 | 0.89 | 0.80 | 0.85 |
| StdDev | 0.41 | 0.11 | 0.14 | 0.18 | 0.20 | 0.09 | 0.09 | 0.16 | 0.14 | 0.20 |

Tabla 32: Resultado de los algoritmos respecto a la media y desviación estándar de los indicadores en las bases de datos médicas.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Media y StdDev** | | *T1* | *T3* | *T4* | *T5* | *Te1a* | *Te1b* | *Te2a* | *Te3* | *Te4* | *Te5* |
| **BD\_med\_congenito** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.15 | 0.97 | 0.58 | 0.50 | 0.66 | 1.00 | 0.02 | 0.40 | 0.97 | 0.97 |
| StdDev | 0.35 | 0.04 | 0.40 | 0.00 | 0.25 | 0.00 | 0.00 | 0.38 | 0.00 | 0.00 |
| LPA\_LDS | Media | 0.12 | 0.89 | 0.81 | 0.50 | 0.67 | 0.89 | 0.02 | 0.17 | 0.86 | 0.97 |
| StdDev | 0.33 | 0.33 | 0.37 | 0.00 | 0.29 | 0.33 | 0.01 | 0.25 | 0.32 | 0.00 |
| Apriori\_LDS | Media | 0.14 | 0.90 | 0.15 | 0.47 | 0.66 | 1.00 | 0.02 | 0.80 | 0.96 | 0.95 |
| StdDev | 0.35 | 0.11 | 0.16 | 0.10 | 0.21 | 0.00 | 0.00 | 0.12 | 0.00 | 0.07 |
| ACO\_LDS | Media | 0.12 | 0.87 | 0.42 | 0.37 | 0.07 | 0.90 | 0.05 | 0.77 | 0.92 | 0.97 |
| StdDev | 0.13 | 0.87 | 0.43 | 0.37 | 0.07 | 0.90 | 0.06 | 0.77 | 0.92 | 0.97 |
| GA\_LDS | Media | 1.00 | 1.00 | 0.99 | 0.50 | 0.00 | 1.00 | 0.02 | 0.02 | 0.97 | 0.97 |
| StdDev | 0.00 | 0.00 | 0.02 | 0.03 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 |
| PCA\_LDS | Media | 0.14 | 0.90 | 0.15 | 0.47 | 0.66 | 1.00 | 0.02 | 0.80 | 0.96 | 0.95 |
| StdDev | 0.35 | 0.11 | 0.16 | 0.10 | 0.21 | 0.00 | 0.00 | 0.12 | 0.00 | 0.07 |
| **BD\_med\_valvular** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.05 | 0.95 | 0.57 | 0.50 | 0.64 | 1.00 | 0.02 | 0.38 | 0.97 | 0.97 |
| StdDev | 0.21 | 0.03 | 0.37 | 0.00 | 0.25 | 0.00 | 0.00 | 0.35 | 0.00 | 0.00 |
| LPA\_LDS | Media | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| StdDev | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Apriori\_LDS | Media | 0.88 | 0.98 | 0.78 | 0.08 | 0.90 | 1.00 | 0.02 | 0.97 | 0.97 | 0.35 |
| StdDev | 0.33 | 0.05 | 0.16 | 0.13 | 0.20 | 0.00 | 0.00 | 0.08 | 0.00 | 0.32 |
| ACO\_LDS | Media | 0.12 | 0.73 | 0.73 | 0.41 | 0.08 | 0.93 | 0.04 | 0.07 | 0.93 | 0.97 |
| StdDev | 0.33 | 0.38 | 0.38 | 0.12 | 0.13 | 0.10 | 0.03 | 0.00 | 0.05 | 0.00 |
| GA\_LDS | Media | 1.00 | 1.00 | 0.97 | 0.50 | 0.01 | 1.00 | 0.02 | 0.05 | 0.97 | 0.97 |
| StdDev | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 |
| PCA\_LDS | Media | 0.88 | 0.98 | 0.78 | 0.08 | 0.90 | 1.00 | 0.02 | 0.97 | 0.97 | 0.35 |
| StdDev | 0.33 | 0.05 | 0.16 | 0.13 | 0.20 | 0.00 | 0.00 | 0.08 | 0.00 | 0.32 |
| **BD\_med\_covidCuba** | |  |  |  |  |  |  |  |  |  |  |
| RST\_LDS | Media | 0.00 | 0.98 | 0.95 | 0.50 | 0.94 | 0.99 | 0.03 | 0.03 | 0.95 | 0.97 |
| StdDev | 0.00 | 0.10 | 0.15 | 0.00 | 0.14 | 0.03 | 0.03 | 0.09 | 0.06 | 0.00 |
| LPA\_LDS | Media | 0.01 | 0.99 | 0.98 | 0.50 | 0.97 | 1.00 | 0.03 | 0.02 | 0.96 | 0.97 |
| StdDev | 0.00 | 0.03 | 0.07 | 0.00 | 0.15 | 0.01 | 0.01 | 0.09 | 0.01 | 0.00 |
| Apriori\_LDS | Media | 0.00 | 0.86 | 0.15 | 0.50 | 0.77 | 0.95 | 0.07 | 0.68 | 0.86 | 0.97 |
| StdDev | 0.00 | 0.15 | 0.15 | 0.00 | 0.17 | 0.06 | 0.06 | 0.04 | 0.12 | 0.00 |
| ACO\_LDS | Media | 0.06 | 0.27 | 0.26 | 0.39 | 0.13 | 0.72 | 0.21 | 0.02 | 0.67 | 0.97 |
| StdDev | 0.21 | 0.21 | 0.20 | 0.13 | 0.31 | 0.19 | 0.11 | 0.02 | 0.19 | 0.00 |
| GA\_LDS | Media | 1.00 | 0.95 | 0.93 | 0.50 | 0.00 | 1.00 | 0.02 | 0.02 | 0.96 | 0.97 |
| StdDev | 0.03 | 0.09 | 0.10 | 0.00 | 0.00 | 0.01 | 0.01 | 0.06 | 0.02 | 0.00 |
| PCA\_LDS | Media | 0.00 | 0.86 | 0.15 | 0.50 | 0.71 | 0.94 | 0.08 | 0.71 | 0.87 | 0.97 |
| StdDev | 0.00 | 0.15 | 0.15 | 0.00 | 0.10 | 0.07 | 0.07 | 0.01 | 0.11 | 0.00 |

Tabla 33: Resultado de los algoritmos respecto al tiempo de ejecución.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **T. de ejecución** | RST\_LDS | | LPA\_LDS | Apriori\_LDS | ACO\_LDS | GA\_LDS | PCA\_LDS |
| **BD sobre proyectos** | | |  | | | | |
| BD\_gp\_competencias |  | 10.97 | 219.22 | 118.33 | 79262.00 | 5378.45 | 125.40 |
| BD\_gp\_eval\_proy\_hard |  | 3.64 | 21.53 | 6.67 | 43429.29 | 1665.43 | 8.20 |
| BD\_gp\_eval\_proy\_integral |  | 0.29 | 7.76 | 2.51 | 2015.21 | 1705.83 | 2.99 |
| BD\_gp\_indicadores |  | 0.13 | 2.80 | 1.19 | 1469.50 | 52.45 | 1.40 |
| BD\_gp\_requerimientos |  | 0.17 | 1.27 | 19.42 | 6024.71 | 336.69 | 20.23 |
| BD\_gp\_toma\_decisiones |  | 1.97 | 34.73 | 3.57 | 6921.28 | 342.19 | 4.55 |
| BD\_gp\_rrhh |  | 0.92 | 19.87 | 3.57 | 19218.98 | 826.71 | 4.93 |
| **BD sobre auditorías** | | |  | | | | |
| BD\_audit\_sinOrganismos |  | 1.7323 | 40.2193 | 65.5803 | 11939.9741 | 534.1207 | 65.9898 |
| BD\_audit\_sinProvMinistInstit |  | 1.2865 | 24.3813 | 64.1094 | 11786.9016 | 542.1494 | 62.8049 |
| **BD sobre medicina** | | |  | | | | |
| BD\_med\_congenito |  | 0.5562 | 5.0624 | 16.0917 | 4217.5541 | 154.7409 | 15.6919 |
| BD\_med\_valvular |  | 0.2897 | 4.4956 | 211.9272 | 2227.936 | 97.8265 | 155.8819 |
| BD\_med\_covidCuba |  | 3.34 | 251.717 | 38.3428 | 47791 | 2914.8029 | 59.8399 |

Tabla 34: Resultado de los algoritmos respecto al Cubrimiento de diferentes situaciones en las bases de datos.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **T. de ejecución** | RST\_LDS | | | LPA\_LDS | | Apriori\_LDS | | ACO\_LDS | | GA\_LDS | | PCA\_LDS |
| **BD sobre proyectos** | | | |  | | | | | | | | |
| BD\_gp\_competencias |  | 0.08 | 0.15 | | 0.16 | | 0.04 | | 0.23 | | 0.19 | |
| BD\_gp\_eval\_proy\_hard |  | 0.08 | 0.09 | | 0.19 | | 0.03 | | 0.38 | | 0.19 | |
| BD\_gp\_eval\_proy\_integral |  | 0.10 | 0.18 | | 0.14 | | 0.04 | | 0.36 | | 0.14 | |
| BD\_gp\_indicadores |  | 0.08 | 0.20 | | 0.26 | | 0.04 | | 0.25 | | 0.26 | |
| BD\_gp\_requerimientos |  | 0.14 | 0.18 | | 0.32 | | 0.04 | | 0.32 | | 0.32 | |
| BD\_gp\_toma\_decisiones |  | 0.19 | 0.21 | | 0.25 | | 0.05 | | 0.37 | | 0.25 | |
| BD\_gp\_rrhh |  | 0.11 | 0.18 | | 0.28 | | 0.02 | | 0.22 | | 0.38 | |
| **BD sobre auditorías** | | | |  | | | | | | | | |
| BD\_audit\_sinOrganismos |  | 0.28 | 0.31 | | 0.23 | | 0.02 | | 0.38 | | 0.24 | |
| BD\_audit\_sinProvMinistInstit |  | 0.23 | 0.27 | | 0.23 | | 0.05 | | 0.38 | | 0.24 | |
| **BD sobre medicina** | | | |  | | | | | | | | |
| BD\_med\_congenito |  | 0.11 | 0.21 | | 0.22 | | 0.38 | | 0.38 | | 0.22 | |
| BD\_med\_valvular |  | 0.11 | 0.38 | | 0.33 | | 0.04 | | 0.38 | | 0.33 | |
| BD\_med\_covidCuba |  | 0.36 | 0.36 | | 0.38 | | 0.03 | | 0.37 | | 0.38 | |

Tabla 35: Resultado de los algoritmos respecto a la Fortaleza de las dependencias en los resúmenes.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **T. de ejecución** | RST\_LDS | | | LPA\_LDS | | Apriori\_LDS | | ACO\_LDS | | GA\_LDS | | PCA\_LDS |
| **BD sobre proyectos** | | | |  | | | | | | | | |
| BD\_gp\_competencias |  | 0.42 | 0.41 | | 0.28 | | 0.31 | | 0.40 | | 0.28 | |
| BD\_gp\_eval\_proy\_hard |  | 0.42 | 0.41 | | 0.35 | | 0.34 | | 0.34 | | 0.35 | |
| BD\_gp\_eval\_proy\_integral |  | 0.42 | 0.41 | | 0.32 | | 0.24 | | 0.40 | | 0.32 | |
| BD\_gp\_indicadores |  | 0.42 | 0.42 | | 0.35 | | 0.09 | | 0.38 | | 0.35 | |
| BD\_gp\_requerimientos |  | 0.42 | 0.41 | | 0.33 | | 0.16 | | 0.39 | | 0.33 | |
| BD\_gp\_toma\_decisiones |  | 0.42 | 0.41 | | 0.31 | | 0.02 | | 0.42 | | 0.31 | |
| BD\_gp\_rrhh |  | 0.42 | 0.41 | | 0.25 | | 0.13 | | 0.24 | | 0.25 | |
| **BD sobre auditorías** | | | |  | | | | | | | | |
| BD\_audit\_sinOrganismos |  | 0.42 | 0.42 | | 0.30 | | 0.15 | | 0.42 | | 0.30 | |
| BD\_audit\_sinProvMinistInstit |  | 0.42 | 0.42 | | 0.30 | | 0.07 | | 0.42 | | 0.30 | |
| **BD sobre medicina** | | | |  | | | | | | | | |
| BD\_med\_congenito |  | 0.42 | 0.42 | | 0.39 | | 0.42 | | 0.42 | | 0.39 | |
| BD\_med\_valvular |  | 0.42 | 0.42 | | 0.33 | | 0.28 | | 0.42 | | 0.33 | |
| BD\_med\_covidCuba |  | 0.42 | 0.42 | | 0.26 | | 0.03 | | 0.40 | | 0.26 | |

## Anexo 14. Resultados de la aplicación del *Page´L Trent Test*

Tabla 36: Resumen de posiciones de los algoritmos respecto a cada indicador de comparación.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | 1 | 2 | 3 | 4 | 5 | 6 |
| T1 | GA\_LDS | PCA\_LDS | Apriori\_LDS | RST\_LDS | ACO\_LDS | LPA\_LDS |
| T2 | RST\_LDS | ACO\_LDS | GA\_LDS | PCA\_LDS | Apriori\_LDS | LPA\_LDS |
| T3 | RST\_LDS | GA\_LDS | PCA\_LDS | Apriori\_LDS | LPA\_LDS | ACO\_LDS |
| T4 | GA\_LDS | LPA\_LDS | RST\_LDS | ACO\_LDS | Apriori\_LDS | PCA\_LDS |
| T5 | RST\_LDS | GA\_LDS | LPA\_LDS | PCA\_LDS | Apriori\_LDS | ACO\_LDS |
| Te1a | RST\_LDS | Apriori\_LDS | PCA\_LDS | LPA\_LDS | ACO\_LDS | GA\_LDS |
| Te1b | RST\_LDS | Apriori\_LDS | GA\_LDS | PCA\_LDS | LPA\_LDS | ACO\_LDS |
| Te3 | PCA\_LDS | Apriori\_LDS | RST\_LDS | ACO\_LDS | LPA\_LDS | GA\_LDS |
| Te4 | PCA\_LDS | Apriori\_LDS | GA\_LDS | RST\_LDS | LPA\_LDS | ACO\_LDS |
| Te5 | GA\_LDS | ACO\_LDS | RST\_LDS | LPA\_LDS | PCA\_LDS | Apriori\_LDS |
| Cubrimiento de diferentes situaciones | ACO\_LDS | RST\_LDS | LPA\_LDS | Apriori\_LDS | PCA\_LDS | GA\_LDS |
| Fortaleza de las dependencias | RST\_LDS | LPA\_LDS | GA\_LDS | PCA\_LDS | Apriori\_LDS | ACO\_LDS |
| Tiempo de ejecución | RST\_LDS | PCA\_LDS | Apriori\_LDS | LPA\_LDS | GA\_LDS | ACO\_LDS |

## Anexo 15. Expertos que participaron en la validación del enfoque multilingüe

Tabla 37: Expertos que participaron en la revisión de los resúmenes lingüísticos en diferentes idiomas.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Expertos que revisaron los resúmenes en idioma árabe: | | | | | | | | |
| Nombre | 1ra lengua | 2da lengua | | Profesión | | | | Tiempo |
| Gaafar Sadeq Saeed Mahdi | Árabe | Español | Traductor en la embajada de Arabia Saudita en Cuba | | | 18 años de dominio del español | | |
| Salah Hasan Saleh Al- subhi | Árabe | Español | Traductor en la embajada de Qatar en Cuba | | | 20 años de dominio del español | | |
| Expertos que revisaron los resúmenes en idioma japonés: | | | | | | | | |
| Nombre | 1ra lengua | 2da lengua | | Profesión | | | Tiempo | |
| Kazuhiro Negawa | Japonés | Español | | Periodista | | | Más de 15 años de dominio del español | |
| Yuki Asaka | Japonés | Español | | Traductora en la embajada de Japón en Cuba | | | | 13 años de dominio del español |
| Expertos que revisaron los resúmenes en idioma inglés: | | | | | | | | |
| Nombre | 1ra lengua | 2da lengua | Profesión | | Tiempo | | | |
| Roberto A. Espí | Inglés | Español | Ingeniero Informático | | Domina ambos idiomas desde su infancia | | | |
| Beatriz Muñoz | Español | Inglés | Traductora del ESTI | | Más de 30 años de dominio del inglés | | | |

## Anexo 16. Resúmenes en múltiples idiomas generados en el caso de estudio

### Resúmenes en español:

1. Muy pocos registros en la base de datos, reportan que en el 94.0% de las veces, los proyectos que tienen calidad del dato regular tienen mala productividad.
2. Pocos registros en la base de datos, reportan que en el 99.0% de las veces, los proyectos que tienen en tiempo la ejecución tienen buena calidad del dato.
3. Aproximadamente la mitad de los registros en la base de datos, reportan que en el 99.0% de las veces, los proyectos que tienen mal rendimiento de la ejecución tienen mala evaluación del proyecto.
4. Muy pocos registros en la base de datos, reportan que en el 94.0% de las veces, los proyectos que tienen buena eficacia en la gestión del alcance tienen buen rendimiento de la planificación.
5. Pocos registros en la base de datos, reportan que en el 96.0% de las veces, los proyectos que tienen mal rendimiento de la planificación tienen muy atrasada la ejecución.
6. Aproximadamente la mitad de los registros en la base de datos, reportan que en el 99.0% de las veces, los proyectos que tienen rendimiento regular de los recursos humanos tienen buena calidad del dato.
7. Algunos registros en la base de datos, reportan que en el 95.0% de las veces, los proyectos que tienen buena productividad tienen buena evaluación del proyecto.
8. Muy pocos registros en la base de datos, reportan que en el 100.0% de las veces, los proyectos que tienen productividad regular tienen buena calidad del dato.
9. Aproximadamente la mitad de los registros en la base de datos, reportan que en el 99.0% de las veces, los proyectos que tienen mala evaluación del proyecto tienen mala productividad.
10. Algunos registros en la base de datos, reportan que en el 96.0% de las veces, los proyectos que tienen buena evaluación del proyecto tienen buen rendimiento de la planificación.
11. Casi la totalidad de los proyectos que tienen mala evaluación del proyecto tienen mala productividad.
12. Casi la totalidad de los proyectos que tienen buena evaluación del proyecto tienen buen rendimiento de la planificación.
13. Casi la totalidad de los proyectos que tienen evaluación regular del proyecto tienen buena calidad del dato.
14. Casi la totalidad de los proyectos que tienen calidad del dato regular y muy atrasada la ejecución tienen mala productividad.
15. Casi la totalidad de los proyectos que tienen buen rendimiento de la ejecución y buen rendimiento de la planificación tienen buena calidad del dato.
16. Casi la totalidad de los proyectos que tienen mala productividad y buena calidad del dato tienen muy atrasada la ejecución.
17. Casi la totalidad de los proyectos que tienen mal rendimiento de la planificación y mala productividad tienen mala evaluación del proyecto.
18. Casi la totalidad de los proyectos que tienen mal rendimiento de la planificación y mala evaluación del proyecto tienen mala productividad.
19. Casi la totalidad de los proyectos que tienen mal rendimiento de la planificación y mala eficacia en la gestión del alcance y mal rendimiento de los recursos humanos tienen mala evaluación del proyecto.
20. Casi la totalidad de los proyectos que tienen mal rendimiento de la planificación y mala eficacia en la gestión del alcance y mala productividad tienen mala evaluación del proyecto.

### Resúmenes en inglés:

1. Very few records show, that 94.0% of times projects with regular data quality, have bad productivity.
2. Few records show, that 99.0% of times projects with in time execution, have good data quality.
3. Around a half of records show, that 99.0% of times projects with bad execution performance, have bad project evaluation.
4. Very few records show, that 94.0% of times projects with good effectiveness in scope management, have good planning performance.
5. Few records show, that 96.0% of times projects with bad planning performance, have execution very late.
6. Around a half of records show, that 99.0% of times projects with regular performance of human resources, have good data quality.
7. Some records show, that 95.0% of times projects with good productivity, have good project evaluation.
8. Very few records show, that 100.0% of times projects with regular productivity, have good data quality.
9. Around a half of records show, that 99.0% of times projects with bad project evaluation, have bad productivity.
10. Some records show, that 96.0% of times projects with good project evaluation, have good planning performance.
11. Almost all projects with bad project evaluation have bad productivity.
12. Almost all projects with good project evaluation have good planning performance.
13. Almost all projects with regular project evaluation have good data quality.
14. Almost all projects with regular data quality and execution very late have bad productivity.
15. Almost all projects with good execution performance and good planning performance have good data quality.
16. Almost all projects with bad productivity and good data quality have execution very late.
17. Almost all projects with bad planning performance and bad productivity have bad project evaluation.
18. Almost all projects with bad planning performance and bad project evaluation have bad productivity.
19. Almost all projects with bad planning performance and bad effectiveness in scope management and bad performance of human resources have bad project evaluation.
20. Almost all projects with bad planning performance and bad effectiveness in scope management and bad productivity have bad project evaluation.

### Resúmenes en japonés:

1. データベース内の非常に数少ない記録が、94.0％の確率で、データ品質が中級程度のプロジェクトは生産性が悪いと報告しています。
2. データベース内の少数の記録が、99.0％の確率で、実行時間内に終了するプロジェクトはデータ品質が良いと報告しています。
3. データーベース内の約半分の記録が、99.0%の確率で実行パフォーマンスの悪いプロジェクトはプロジェクト評価が低いと報告しています。
4. データーベース内の非常に少数の記録が、94.0％の確率で、目的達成の効率がよいプロジェクトは、計画性が高いと報告しています。
5. データベース内の数少ない記録が、96.0％の確率で、計画性の悪いプロジェクトは実行がとても遅いと報告しています。
6. データーベースの内約半数の記録が、99.0％の確率で、人事パフォーマンスが中級程度のプロジェクトはデーター品質が良いと報告しています。
7. データーベース内のいくつかの記録が、95.0％の確率で、生産性の高いプロジェクトは評価が高いと報告しています。
8. データーベース内の非常に少数の記録が、100.0％の確率で、生産性が中級程度のプロジェクトは、データー品質が良いと報告しています。
9. データーベース内の約半分の記録が、99.0％の確率で、評価の低いプロジェクトは生産性が低いと報告しています。
10. データベース内のいくつか記録が、96.0％の確率で、評価の高いプロジェクトは計画性が高いと報告しています。
11. 評価の低いプロジェクトのほとんどは、生産性が低いです。
12. 評価の高いプロジェクトのほとんどは、計画性が高いです。
13. 評価が中級程度のプロジェクトのほとんどは、データ品質が高いです。
14. データー品質が中級程度で実行がとても遅れているプロジェクトのほとんどは、生産性がとても低いです。
15. 実行パフォーマンスが良く、計画性の高いプロジェクトのほとんどは、データ品質が良好です。
16. 生産性は低いがデータ品質の高いプロジェクトのほとんどは、実行が遅れています。
17. 計画性が低く、生産性も低いプロジェクトのほとんどは、プロジェクト評価が低いです。
18. 計画性が低く、プロジェクト評価の低いプロジェクトのほとんどは、生産性が低いです。
19. 計画性が低く、目的達成の効率が悪く、人事パフォーマンスの低いプロジェクトのほとんどは、プロジェクト評価が低いです。
20. 計画性が低く、目的達成の効率が悪く、生産性の低いプロジェクトのほとんどは、プロジェクト評価が低いです。

### Resúmenes en árabe:

1. قليل جدا من\_السجلات في قاعدة البيانات، تظهر أنه في\_94.0%\_من المرات,\_المشاريع\_التي لديها\_جودة بيانات متوسطة\_لديها\_إنتاجية سيئة\_
2. القليل من\_السجلات في قاعدة البيانات، تظهر أنه في\_99.0%\_من المرات,\_المشاريع\_التي لديها\_تنفيذ في الوقت المحدد\_لديها\_جودة بيانات جيدة\_
3. حوالي نصف\_السجلات في قاعدة البيانات، تظهر أنه في\_99.0%\_من المرات,\_المشاريع\_التي لديها\_أداء تنفيذ ضعيف\_لديها\_تقييم سيء\_
4. قليل جدا من\_السجلات في قاعدة البيانات، تظهر أنه في\_94.0%\_من المرات,\_المشاريع\_التي لديها\_كفاءة جيدة في إدارة الأهداف\_لديها\_أداء تخطيط جيد\_
5. القليل من\_السجلات في قاعدة البيانات، تظهر أنه في\_96.0%\_من المرات,\_المشاريع\_التي لديها\_أداء تخطيط سيء\_لديها\_تنفيذ متأخر جداً\_
6. حوالي نصف\_السجلات في قاعدة البيانات، تظهر أنه في\_99.0%\_من المرات,\_المشاريع\_التي لديها\_أداء مقبول للمواردالبشرية\_لديها\_جودة بيانات جيدة\_
7. بعض\_السجلات في قاعدة البيانات، تظهر أنه في\_95.0%\_من المرات,\_المشاريع\_التي لديها\_إنتاجية جيدة\_لديها\_تقييم جيد\_
8. قليل جدا من\_السجلات في قاعدة البيانات، تظهر أنه في\_100.0%\_من المرات,\_المشاريع\_التي لديها\_إنتاجية مقبولة\_لديها\_جودة بيانات جيدة\_
9. حوالي نصف\_السجلات في قاعدة البيانات، تظهر أنه في\_99.0%\_من المرات,\_المشاريع\_التي لديها\_تقييم سيء\_لديها\_إنتاجية سيئة\_
10. بعض\_السجلات في قاعدة البيانات، تظهر أنه في\_96.0%\_من المرات,\_المشاريع\_التي لديها\_تقييم جيد\_لديها\_أداء تخطيط جيد\_
11. تقريبا كل\_المشاريع\_التي لديها\_تقييم سيء\_لديها\_إنتاجية سيئة\_
12. تقريبا كل\_المشاريع\_التي لديها\_تقييم جيد\_لديها\_أداء تخطيط جيد\_
13. تقريبا كل\_المشاريع\_التي لديها\_تقييم مقبول\_لديها\_جودة بيانات جيدة\_
14. تقريبا كل\_المشاريع\_التي لديها\_جودة بيانات متوسطة وتنفيذ متأخر جداً\_لديها\_إنتاجية سيئة\_
15. تقريبا كل\_المشاريع\_التي لديها\_أداء تنفيذ جيد وأداء تخطيط جيد\_لديها\_جودة بيانات جيدة\_
16. تقريبا كل\_المشاريع\_التي لديها\_إنتاجية سيئة وجودة بيانات جيدة\_لديها\_تنفيذ متأخر جداً\_
17. تقريبا كل\_المشاريع\_التي لديها\_أداء تخطيط سيء وإنتاجية سيئة\_لديها\_تقييم سيء\_
18. تقريبا كل\_المشاريع\_التي لديها\_أداء تخطيط سيء وتقييم سيء\_لديها\_إنتاجية سيئة\_
19. تقريبا كل\_المشاريع\_التي لديها\_أداء تخطيط سيء وكفاءة سيئة في إدارة الأهداف وأداء ضعيف للموارد البشرية\_لديها\_تقييم سيء\_
20. تقريبا كل\_المشاريع\_التي لديها\_أداء تخطيط سيء وكفاءة سيئة في إدارة الأهداف وإنتاجية سيئة\_لديها\_تقييم سيء\_

1. Ver concepto deinteligencia artificial en el glosario de términos*.* [↑](#footnote-ref-2)
2. En inglés: *Linguistic Data Summarization (LDS).* [↑](#footnote-ref-3)
3. Incertidumbre se refiere a la veracidad del conocimiento y se produce por la falta de certeza de la información, ver concepto en glosario de términos. [↑](#footnote-ref-4)
4. “Incertidumbre” se refiere a la veracidad del conocimiento y se produce por la falta de certeza de la información. En la operacionalización de la variable se refiere a la capacidad de medir la certidumbre, la indeterminación y la falsedad, ver concepto completo en glosario de términos. [↑](#footnote-ref-5)
5. Organización para la Cooperación Económica y Desarrollo (OCDE: Organization for Economic Co-operation and Development). [↑](#footnote-ref-6)
6. Base de datos estadísticos de patentes de la Oficina Europea de Patentes (PATSTAT: Patent Statistical Database of European Patent Office) [↑](#footnote-ref-7)
7. Quindenio: período de 15 años. [↑](#footnote-ref-8)
8. El término análisis de componentes principales (ACP) es más conocido por sus siglas en inglés: *PCA* (*Principal Component Analisys*)(Naik 2018). [↑](#footnote-ref-9)
9. Ver definición de aprendizaje activo en el glosario de términos. [↑](#footnote-ref-10)
10. Ver definición de teoría neutrosófica en el glosario de términos y la evolución de esta teoría en el Anexo 6 Figura 36. [↑](#footnote-ref-11)
11. Se refiere a la fortaleza de las relaciones que descubren los algoritmos en la base de datos. [↑](#footnote-ref-12)
12. Prueba no paramétrica que mide la discrepancia entre la distribución observada y una teórica (bondad de ajuste). Cuanto mayor sea la Chi-Cuadrado, menos verosímil es la hipótesis nula que supone la igualdad de las distribuciones; mientras más se acerca a cero este valor, más ajustadas están las distribuciones. [↑](#footnote-ref-13)
13. La traducción de los términos es equivalente a la versión en español en la sección 9.9.1 de este mismo anexo. [↑](#footnote-ref-14)
14. La traducción de los términos es equivalente a la versión en español en la sección 9.9.1 de este mismo anexo. [↑](#footnote-ref-15)
15. En español: “en la base de datos” [↑](#footnote-ref-16)
16. En español: “registros” [↑](#footnote-ref-17)
17. En español: “% de probabilidad” [↑](#footnote-ref-18)