UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Facultad 4

Departamento de Investigaciones en Gestión de Proyectos

Modelo para el aseguramiento de ingresos en organizaciones orientadas a proyectos basado en minería de datos anómalos

Tesis presentada en opción al grado científico de

Doctor en Ciencias Técnicas

GILBERTO FERNANDO CASTRO AGUILAR

La Habana

2017

UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Facultad 4

Departamento de Investigaciones en Gestión de Proyectos

Modelo para el aseguramiento de ingresos en organizaciones orientadas a proyectos basado en minería de datos anómalos

Tesis presentada en opción al grado científico de

Doctor en Ciencias Técnicas

Autor: MSc. GILBERTO FERNANDO CASTRO AGUILAR

Tutores: Prof. Titular, Dr. C Pedro Yobanis Piñero Pérez

Prof. Titular, Dr. C Antonio de Jesús Romillo Tarke

Prof. Titular, Dr. C Natalia Martínez Sánchez

La Habana

2017

AGRADECIMIENTOS

Agradezco a Dios por darme la vida, la salud, la familia, los amigos y la oportunidad de haber emprendido este camino de la ciencia lleno de momentos gratos y de nuevos amigos.

Quiero dar un agradecimiento muy especial a mis tutores Natalia, Antonio y Pedro, que han sido un apoyo importante para la consecución de este objetivo y sobre todo me han sabido guiar en el sendero de la ciencia y del conocimiento.

Un agradecimiento y un cariño especial para el personal de la Universidad de las Ciencias Informáticas (UCI), por el trato y las atenciones recibidas durante las estadías en el campus universitario.

DEDICATORIA

El presente trabajo va dedicado de manera especial a mi bella esposa, a mi hija, mis padres y mis suegros, que han sabido entenderme durante todos estos años dedicados a la consecución de este proyecto, han sido mi inspiración y va por ellos.

A mis distinguidos tutores, que, con sabiduría y entusiasmo, vertieron todo su postulado en mi persona para alcanzar la culminación de esta meta importante en mi vida.

SÍNTESIS

El incremento de la competitividad en los mercados globales ha provocado la necesidad de mejoras en las organizaciones orientadas a proyectos, dirigidas a mejorar la salud financiera y los ingresos de las mismas.En este contexto, surge el aseguramiento de ingresos, como un campo interdisciplinar que combina técnicas de estadística, bases de datos, *soft computing* y minería de datos anómalos, orientado a la reducción de los costos y la maximización de los ingresos en las organizaciones que los apliquen. Se propone en este trabajo un modelo para el aseguramiento de ingresos para organizaciones orientadas a proyectos, que permita la detección de errores de planificación y la maximización de los ingresos durante el desarrollo de proyectos.Como parte de la novedad del trabajo el modelo propuesto combina técnicas de gestión de riesgos, alcance, tiempo, minería de datos anómalos y técnicas de *soft computing*. En el trabajo se realizan pruebas de validación cruzada comparando diferentes técnicas, para la detección de situaciones anómalas, en la planificación de los proyectos. En la comparación se emplean bases de datos de gestión de proyectos de desarrollo de soluciones informáticas. El modelo propuesto se introduce como un módulo en la plataforma para la dirección integrada de proyectos GESPRO y se presentan los resultados de su aplicación.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En la actualidad, una de las formas de organización que ha ganado fuerza es la dirección integrada de proyectos,por su aplicabilidad en diferentes escenarios[1, 2].Esto ha motivado la proliferación de organizaciones orientadas a proyectos[1] en disímiles áreas de la sociedad y de estándares como la guía del PMBOK[3](*Project Management Body of Knowledge*) (Compendio del Saber de la Gestión de Proyectos) y CMMI [4] *(Capability Maturity Model Integration)* (Integración de modelos de madurez de capacidades) que recogen buenas prácticas para la gestión de dichas organizaciones.

Pero, a pesar de los esfuerzos por mejorar la eficaciaen la gestión de dichas organizaciones, persisten numerosas dificultades que generan pérdidas de ingresos. Este fenómeno es particularmente relevante en las organizaciones orientadas al desarrollo de las TICs (Tecnologías de la Información y la Comunicación). Un estudio realizado en el 2015 por *The Standish Group International Incorporated*,muestra que históricamente las cifras de proyectos exitosos, fallidos y renegociados, se ha movido ligeramente alrededor del29%, 19% y 52% respectivamente [5-7], ver Anexo 1.Detrás de los proyectos no exitosos existen pérdidas significativas de ingresos para organizaciones con un elevado impacto económico y social. En este contexto, se

[1]Organizaciones orientadas a proyectos son aquellas que desarrollan nuevos productos o servicios, organizando los recursos humanos y no humanos en forma de proyectos con objetivos, fechas de inicio y fin bien determinados.

identifican entre las causas fundamentales del fracaso de los proyectos las insuficiencias en los procesos de planificación, de control y seguimiento, así como insuficientes mecanismos para la gestión de riesgos y la gestión de los recursos humanos [8-10].Algunas de estas causas, pueden ser mitigadas si se analizan los datos anómalos contenidos en los sistemas de información de las propias organizaciones[11-13].

Por otra parte, desde finales de la década de los 70' surge la disciplina de “Aseguramiento de ingresos”, orientadaa la protección y recuperación de los recursos financieros en diferentes organizaciones [14-17]. Asociada a esta disciplina, surgen la Asociación Global de Profesionales de Aseguramiento de Ingresos (GRAPA) [18] y el fórum de discusión de expertos en esta disciplinaTMForum [14, 15].Pero, las técnicas propuestas por esta disciplina aún son insuficientes tanto para las empresas de telecomunicaciones donde surgen, como para las organizaciones orientadas a proyectos,donde tampoco han sido suficientemente aplicadas. Entre las principales deficiencias de esta disciplina se señalan [14-18]:

- Dependen de recursos humanos para su aplicación,que éstos a su vez, también están sujetos a posibles errores de operación y no se maneja la certidumbre de las decisiones o de la detección de situaciones anómalas.
- En las organizaciones orientadas a proyectos, se presentan fenómenos como la heterogeneidad en los datos, la imprecisiónyla incertidumbre; que las técnicas tradicionales de aseguramiento de ingresos y análisis de datos

anómalos no gestionan adecuadamente, afectando la eficacia en la detección de las situaciones anómalas generadoras de pérdidas de ingresos:

  - La imprecisión en este escenario se muestra en la incompletitud y el ruido en los datos. Generalmente los datos recogidos dependen de operaciones manuales, sujeta a posibles errores de edición y de operación.
  - La incertidumbre se presenta porque los datos registrados dependen en gran medida de la percepción y experticia de los usuarios de los sistemas.

- Con frecuencia las soluciones planteadas se basan solamente en enfoques reactivos y no usan adecuadamente estrategias activas o proactivas del aseguramiento de ingresos; afectando de esta forma la eficiencia de los procesos de detección de situaciones anómalas.
- Muchas de las soluciones para el aseguramiento de ingresos se basan en los sistemas basados en reglas de producción [17]. En este contexto, este tipo de sistema basado en el conocimiento presenta las siguientes dificultades:
  - Dificultades para tratar con el elevado dinamismo y la diversidad de las organizaciones orientadas a proyectos.
  - Dificultades con el cubrimiento del dominio y visión parcial del espacio de búsqueda. En el dominio del aseguramiento de ingresos las entradas varían mucho y requieren por lo general de numerosas reglas para considerar todas las posibles situaciones anómalas.Este factor unido al

lento reaprendizaje afecta la eficacia del proceso de detección de situaciones anómalas.

    - Poco nivel de reutilización: en la definición de las reglas en el contexto del aseguramiento de ingresos, fenómeno provocado porque con frecuencia se especifica el nombre exacto de las tablas y los atributos donde deben realizar la búsqueda. Esto afecta la eficacia para la aplicación de las técnicas de detección de situaciones anómalas en nuevos escenarios.

- Con frecuencia las soluciones implantadas constituyen cajas negras soportadas por herramientas privativas que afectan la soberanía tecnológica de las organizaciones. No se sabe a ciencia cierta todo el impacto que tiene para la organización la gestión de la información con estas herramientas externas.

Se ha identificado que muchos de estos problemas afectan la eficiencia y la eficacia de los procesos de aseguramiento de ingresos desde la perspectiva de la capacidad de detección de los datos anómalos. Para entender mejor esta situación,se introducen a continuación los conceptos eficiencia y eficacia empleados en este trabajo.

- Eficiencia: en el contexto de esta investigación, la eficiencia evalúa el tiempo empleado por los algoritmos, para la detección de datos anómalos, en los procesos de aseguramiento de ingresos.

- Eficacia: en el contexto de esta investigación, la eficacia refleja la capacidad para la detección de situaciones anómalas en los datos, generalmente provocadas por acciones de fraude o fallos operacionales, o la detección y estimación de riesgos que afectan los ingresos.

Problema de investigación

A partir de estos análisis se identifica el siguiente **problema de investigación**: las insuficiencias en las técnicas existentes para el aseguramiento de ingresos,está afectando la eficacia yla eficiencia en la detección de situaciones anómalas,generadoras de pérdidas de ingresos en las organizaciones orientadas a proyectos.

El **objeto de investigación** es: el aseguramiento de ingresos en las organizaciones orientadas a proyectos.

Objetivo general

Desarrollar un modelo[2]para el aseguramiento de ingresos que combine técnicas de gestión de riesgos y minería de datos anómalos,mejorando la eficiencia y la eficaciaen los procesos de detección y prevención de situaciones que afectan a las organizaciones orientadas a proyectos.

[2]Modelo es una representación de un hecho o fenómeno que permite mostrar las características generales de dicho fenómeno a partir de modelar sus componentes principales y las relaciones entre ellas. Los modelos tienen un criterio de uso, aplicabilidad, una representación gráfica y una forma de instrumentación.

Objetivos específicos

- Identificar las tendencias principales en el aseguramiento de ingresos en las organizaciones orientadas a proyectos, sus potencialidades y deficiencias.
- Desarrollar un modelo para el aseguramiento de ingresos que permita la introducción de estrategias proactivo, activo y reactiva; basado en la combinación de técnicas de gestión de riesgos, *soft computing* y minería de datos anómalos.
- Validar el modelo propuesto a partir de comparar sus diferentes componentes con otras técnicas, demostrando la mejora en la eficacia y la eficiencia en ladetección de datos anómalos que generan pérdidas de ingresos.

El **campo de investigación** es: técnicas para la predicción y detección de datos anómalos en organizaciones orientadas a proyectos.

Tipo de investigación

El tipo de investigación es explicativa, el trabajo pretende identificar las mejores técnicas para la detección de datos anómalos y la predicción de situaciones que afectan los procesos de aseguramiento de ingresos en las organizaciones orientadas a proyectos. Diferentes técnicas y algoritmos son aplicados sobre bases de datos de gestión de proyectos, descubriendo relaciones de causalidad y combinaciones eficaces y eficientes para la detección de situaciones anómalas generadoras de pérdidas de ingresos.

Hipótesis

El desarrollo de un modelo que combine técnicas de gestión de riesgos y minería de datos anómalos permitirá mejorar la eficiencia y la eficacia en la detección de situaciones anómalas generadoras de pérdidas de ingresos en las organizaciones orientadas a proyectos.

La **variable independiente** es: modelo para el aseguramiento de ingresos.

Las**variables dependientes** son: eficacia y la eficiencia en la detección de situaciones que provocan pérdidas de ingresos.

Métodos de investigación

Métodos teóricos empleados:

- Histórico-lógico: en la primera parte de la investigación se desarrolla un estudio del estado del arte de las técnicas de aseguramiento de ingresos y detección de datos anómalos,así como su aplicación en las organizaciones orientadas a proyectos. A partir del problema concreto se plantean los objetivos específicos e hipótesis.
- Hipotético deductivo: en el transcurso de la investigación la hipótesis es resuelta siguiendo métodos bien fundamentados científicamente y luego se realizan pruebas estadísticas para demostrar la validez de los resultados.
- Sistémico: todos los componentes de la propuesta y los entes beneficiados con la misma se analizan como un sistema integrado con relaciones entre los

mismos. La solución propuesta propone un enfoque holístico con impacto de la propuesta en la sociedad tanto desde el punto de vista económico como social.

Métodos empíricos empleados:

- Experimental: se diseñan experimentos a partir de los cuales se recopilan datos de los registros digitales y se realiza un diagnóstico de los mismos orientado al análisis de factores que generan pérdidas de ingresos.
- Encuesta: se realizan encuestas a especialistas de departamentos de aseguramientos de empresas de telecomunicaciones y a especialistas en gestión de proyectos para evaluar la aplicabilidad de los algoritmos.

Muestreo

- Población: se toman datos de bases de registros de proyectos y planificación de recursos de bases de datos de organizaciones orientadas a proyectos. Se aplican técnicas de triangulación de datos, para ello se emplean cinco bases de datos y organizaciones reales orientadas a proyectos.
- Muestra: la muestra para los experimentos se toma de forma probabilística a partir de bases de datos de organizaciones orientadas a proyectos. Se emplean técnicas de validación cruzada para de esta forma mitigar la presencia de variables extranjeras y sesgos en la experimentación. Se aplica la investigación en la empresa QuitusServices y entidades desarrolladoras de proyectos de software.

Diseño de experimentos

Para el diseño de experimentos se emplean técnicas de triangulación metodológica combinando técnicas para la triangulación de datos, la triangulación de expertos y la triangulación metodológica de métodos.Para la comparación de las muestras se realizan pruebas de normalidad usando el test shapiro-wills y se aplican técnicas paramétricas o no paramétricas en dependencia del análisis de normalidad de los datos. Para la realización de todas las pruebas se empleael lenguaje R y sus bibliotecas de algoritmos.

Novedad

- Un nuevo modelo para el aseguramiento de ingresos con centro en la detección de situaciones anómalas y que combina técnicas de gestión de proyectos, *soft computing* y minería de datos anómalos, aplicable en organizaciones orientadas a proyectos y propuestas de soluciones de detección.

Aporte práctico de la investigación

- Implementación del modelo propuesto y los algoritmos propuestos como un módulo para la plataforma GESPRO basada en tecnologías libres y biblioteca de algoritmos basados en R para la detección de datos anómalos.

Estructura de la tesis

La tesis está organizada en trescapítulos. En el primer capítulo se realiza un análisis acerca de las técnicas de aseguramiento de ingresos en las organizaciones orientadas a proyectos, así como de los métodos para la predicción y detección de

datos anómalas. En el segundo capítulo se presenta el modelo propuesto y se discute acerca de su alcance. En el tercer capítulo se discuten los resultados de la aplicación del modelo en escenarios concretos y de los algoritmos analizados. Finalmente se presentan las conclusiones, las recomendaciones y los anexos del trabajo.

1. CAPÍTULO: ASEGURAMIENTO DE INGRESOS Y LA MINERÍA DE DATOS ANÓMALOS

El aseguramiento de ingresos como área de conocimiento surge desde finales de la década de los 70', en el sector de las telecomunicaciones, como disciplina orientada a la protección y recuperación de los recursos financieros de las organizaciones[14-16]. Se extiende por su aplicabilidad a disímiles áreas del desarrollo social, entre los que destacan la salud, la agricultura, la gobernabilidad[19- 22].

Según GRAPA [18], en el sector de las telecomunicaciones, entre los factores que tienen mayor influencia en el aseguramiento de ingresos se encuentran los asociados ala capacitación del personal y su certificación, ver Figura 1.

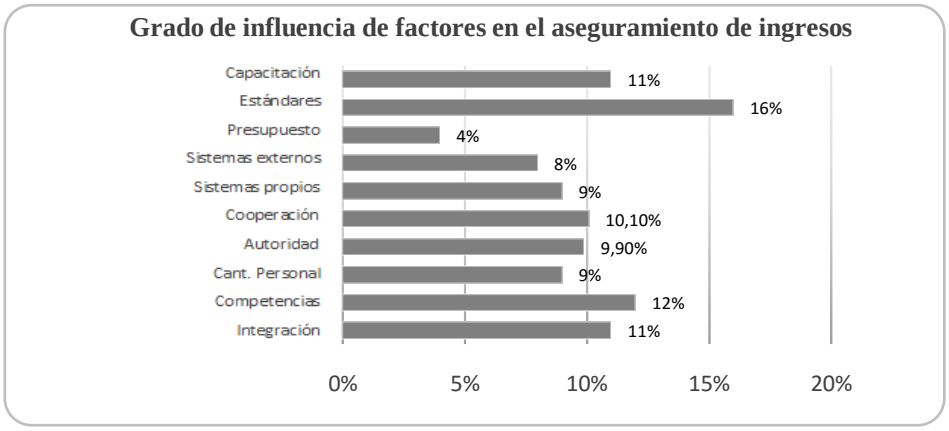


Figura 1. Factores que influyen en el aseguramiento de ingresos en lastelecomunicaciones. Tomado de [18].

Se debe notar que GRAPA no identifica los problemas financieros como factores que influyen, motivado por la solvencia de las empresas de telecomunicaciones. Sin embargo, los factores asociados al presupuesto, las tecnologías empleadas, la

cantidad de personal y sus competencias sí constituyen factores relevantes en las organizaciones orientadas a proyectos [23, 24].

Por otra parte, Acosta [16] plantea que la actividad de aseguramiento de ingresos consiste en buscar, identificar y eliminar las causas técnicas y estructurales que dan origen a las fugas de ingresos. Plantea que este proceso se divide en dos etapas: primera etapa asociada a la maximización de los ingresos, a través del control sistemático, y una segunda etapa dirigida a minimizar los costos e identificar nuevas oportunidades de ingresos. En las empresas de telecomunicaciones estas etapas ocurren de forma secuencial mientras que en las organizaciones orientadas a proyectos los procesos de planificación, control y seguimiento ocurren de forma continua y por iteraciones.

Otro espacio para el análisis de estándares de aseguramiento de ingresos es TMForum [14, 15]. Este espacio promueve un modelo de evaluación de la madurez de las organizaciones en la implantación de los procesos de aseguramiento de ingresos.

Respecto a las técnicas computacionales empleadas para el aseguramiento de ingresos, tanto TMForum como GRAPA reconocen el uso de técnicas de minería de datos anómalos [14],en la Tabla 1 se presenta un resumen de las más usadas.

Tabla 1. Técnicas más empleadas en los procesos de aseguramiento de ingresos.

Procesos	**Técnicas de minería de datos anómalos empleadas**
Proceso de análisis de riesgos	Muestreo, Análisis de grupo, Análisis de distribución, Análisis de tendencia central, Regresión

Proceso de detección temprana	Muestreo, Análisis de grupo, *Chaid/Cart*, RNA
Proceso de diseño de controles	Muestreo,*Chaid/Cart*, Análisis de distribución, Análisis de tendencia central
Proceso de análisis de causa raíz	Análisis de grupo, Análisis de distribución, Análisis de tendenciacentral
Proceso de pronóstico	Muestreo, Análisis de grupo, *Chaid/Cart*,RNA y Regresión

Pero en la mayoría de los casos las grandes empresas [17, 16, 25], contratan a empresas consultoras, las cuales generalmente implementan procedimientos de aseguramiento de ingresos según métodos propios. Entre las técnicas más empleadas se encuentran: aplicación de juicio de expertos y la aplicación de reglas de producción propuestas por expertos, generalmente muy específicas para cada escenario.

Así mismo tanto GRAPA como TMForum coinciden en la existencia de líneas abiertas a la investigación entre las que se encuentran:

- La proactividad en el aseguramiento de ingresos y la prevención por sobre las técnicas clásicas para la recuperación de ingresos.
- Las investigaciones en *big data* y su aplicación en el aseguramiento de ingresos.

Las diferencias relacionadas, entre las organizaciones orientadas a proyectos y las empresas de telecomunicaciones, provocan que no se puedan usar de forma literal los estándares de aseguramiento de ingresos de las telecomunicaciones en las organizaciones orientadas a proyectos.

Aseguramiento de ingresos en organizaciones orientadas a proyectos

La gestión de proyectos en la etapa moderna se inicia entre los años 50 y 60, fuertemente motivada por grandes proyectos militares que requerían la coordinación de equipos y áreas diferentes en la construcción de complejos sistemas. Pero el mayor avance en el desarrollo de la disciplina de la Gestión de Proyectos ha sido la creación de escuelas o instituciones dedicadas a la formalización y estandarización para desarrollar nuevos métodos de organización y trabajo. Entre estas instituciones se encuentrael *Project Management Institute* (PMI) con su estándarPMBOK [3], el *Software Engineering Institute* (SEI) con el estándar de CMMI [4],el *International Project Management Association* (IPMA) [26],la escuela de la Universidad Politécnica de Madrid representada por Heredia[27] y la *International Organization for Standardization* (ISO) con sus normas 10006 y 21500 [28].

El PMI en el PMBOK[3] define proyecto como “un esfuerzo temporal que se lleva a cabo para crear un producto, servicio o resultado único. La naturaleza temporal de los proyectos indica un principio y un final definidos”. En este concepto no aparecen elementos asociados a la necesaria optimización de los proyectos respecto a los recursos y costos asociados que afectan los ingresos de las organizaciones.La guía del PMBOK establece 47 procesos y por cada proceso propone técnicas a emplear que se analizan en la Tabla 2,respecto a su impacto en el aseguramiento de ingresos.

La ISO 21500[28]por su parte fue creada por la Organización de Estándares Internacionales y propone 39 procesos agrupados en 5 grupos de procesos y 10 áreas de conocimiento con un alto nivel de similaridad con el PMBOK [29, 30].

En la siguiente Tabla 2 se hace un análisis tanto del PMBOK como de la ISO respecto a los procesos y técnicas que pueden influir con en el aseguramiento de ingresos.

Tabla 2. Procesos del PMBOK, ISO 21500 asociados al aseguramiento de ingresos.

Área	**PMBOK 5ta edición**	**ISO 21500**	**Técnicas que proponen que pueden influir en el aseguramiento de ingresos**
Integración	4.4 Monitorear y controlar el trabajo del proyecto	4.3.5 Controlar el trabajo del proyecto	-Técnicas analíticas para pronosticar resultados potenciales sobre la base de variaciones en las variables del proyecto: análisis de regresión, métodos de clasificación, análisis causal, series temporales, simulación,análisis de modos de fallo y efectos,análisis de reservas,tendencias y análisis de variación.
	4.5 Control de cambio	4.3.6 Controlar cambios	-Juicio de expertos.
Alcance	5.5 Validar el alcance		-Inspección: incluye actividades de medir, examinaryvalidar. Determinando si eltrabajoylosentregablescumplenconlos criterios de aceptación del producto.
	5.6 Controlar el alcance	4.3.14 Controlar el alcance	-Análisis de variación:técnica para determinar lacausayelgradode la diferencia entre lalíneabaseyeldesempeño.
Tiempo	6.4 Estimar los recursos	4.3.16 Estimar recursos	-Juicio de expertos y análisis de alternativas a partir de datos de estimaciones publicados.
	6.5 Estimar la duración de las actividades	4.3.22 Estimar la duración de las actividades	-Juicio de expertos. -Estimación análoga: basada en datos históricos de las actividades (presupuesto, tamaño, carga y complejidad).

			-Estimación paramétrica: basada en algoritmo de análisis estadístico para calcular el costo o la duración. -Técnicas grupales de toma de decisiones. -Estimación por tres valores y análisis de reservas.
	6.6 Desarrollar el cronograma	4.3.23 Desarrollar el cronograma	-Método de la ruta crítica y cadena crítica. -Técnicas de optimización de recursos. -Técnicas grupales de toma de decisiones. -Análisis de reservas, análisis de escenarios y simulación. -Adelantos y retrasos, compresión del cronograma.
	6.7 Controlar el cronograma	4.3.24 Controlar el cronograma	-Revisiones del desempeño: análisis de tendencias, análisis de la ruta crítica, análisis de la cadena crítica, análisis del valor ganado.
Costo	7.2 Estimar costos	4.3.25 Estimar costos	-Juicio de expertos, estimación análoga y estimación paramétrica. -Estimación por tres valores y análisis de reservas.
	7.4 Controlar costos	4.3.27 Controlar costos	-Gestión del valor ganado e indicadores de pronóstico. -Análisis de desempeño y análisis de reservas.
Calidad	8.2 Realizar el aseguramiento de la calidad	4.3.33 Realizar el aseguramiento de la calidad	-Auditorías de calidad. -Análisis de procesos: incluye el análisis de la causa raíz.
	8.3 Realizar el control de la calidad	4.3.34 Realizar el control de la calidad	-Herramientas de control de calidad: diagramas de afinidad, gráficas de programación de decisiones de proceso, dígrafos de interrelaciones, diagramas de árbol, matrices de priorización, diagramas de red, diagramas matriciales.
Recursos humanos	9.4 Gestionar el equipo.	4.3.20 Gestionar el equipo.	-Evaluaciones de desempeño del proyecto.
Riesgos	11.2 Identificar riesgos	4.3.28 Identificar riesgos	-Juicio de expertos, análisis FODA, lista de verificación. -Técnicas de diagramación: diagramas de causa y efecto, diagramas de flujo de procesos, diagramas de influencias. -Técnicas de recopilación de información.

	11.3 Realizar el análisis cualitativo 11.4 Realizar el análisis cuantitativo	4.3.29 Evaluar riesgos	-Juicio de expertos. -Evaluación de probabilidad e impacto de los riesgos. -Evaluación de la calidad de los datos sobre riesgos. -Categorización de riesgos y priorización. -Técnicas de análisis cuantitativo: análisis de sensibilidad, árboles de decisión, modelado y simulación.
	11.5 Planear la respuesta a los riesgos	4.3.30 Tratar los riesgos	-Juicio de expertos. -Estrategias para riesgos negativos o amenazas. -Estrategias para riesgos positivos u oportunidades. -Estrategias de respuesta a contingencias.
	11.6 Controlar riesgos	4.3.31 Controlar los riesgos	-Revaluación y auditorías a los riesgos. -Análisis de variación, tendencias y análisis de reserva. -Medición del desempeño técnico.

Se puede apreciar que, aunque PMBOK e ISO 21500 incluyen actividades y técnicas que pueden influir en el aseguramiento de ingresos,estas están basadas en análisis manuales con una fuerte influencia de expertos. Además, no se incluyen técnicas computacionales, como la minería de datos anómalos, que permitan identificar situaciones anómalas en las planificaciones o en las estimaciones del proyectoque puedan afectar los ingresos de los proyectos. El mayor aporte de estas metodologías al aseguramiento de ingresos se puede encontrar en el área de gestión de riesgos. No obstante, se señala que las propias técnicas que proponen para el análisis cualitativo son rígidas y no permiten un adecuado tratamiento de la incertidumbre y la ambigüedad existente en los procesos de gestión de riesgos.

El SEI, por su parte en CMMI [4] presenta el concepto de proyecto como “un conjunto de recursos interrelacionados y organizados para la entrega de uno o más productos a un cliente o usuario, en un plazo definido y bajo un plan de desarrollo”. Este estándar está diseñado para medir la capacidad y madurez de organizaciones orientadas a proyectos de software, pero al igual que el PMBOK en su definición no incorpora suficientes elementos asociados a técnicas de minería de datos u optimización que puedan contribuir de esa forma en la detección de situaciones anómalas generadoras de pérdidas de ingresos en las organizaciones. En la siguiente tabla se analizan las prácticas genéricas y específicas de CMMI que mayor influencia tienen en el aseguramiento de ingresos.

Tabla 3. Prácticas genéricas y específicas que influyen en el aseguramiento de ingresos.

Área de proceso	**Práctica genérica**	**Práctica específica**
REQM (Requisitos)	SG1 (Requisitos)	SP1.3-1
PP (Plan de proyecto)	SG1 (Estimación)	SP1.4-1
	SG2(Desarrollo del Plan)	SP2.1-1, SP2.4-1
	SG3 (Revisiones del plan)	SP3.1-1
PMC (Monitoreo y control)	SG1 (Monitoreo del plan)	SP1.1-1, SP1.3-1
	SG2 (Acciones correctivas)	SP2.3-1
SAM (Proveedores)	SG2 (Acuerdos con proveedores)	SP2.1-1
MA (Medición y análisis)	SG2 (Resultados de mediciones)	SP2.2-1
CM (Gestión de la configuración)	SG2 (Control de cambios)	SP2.2-1
RD (Desarrollo de requisitos)	SG3 (Validación requerimientos)	SP3.5-2

PI (Integración de producto)	SG3 (Diseño del producto)	SP3.3-1
VER (Verificación)	SG3 (Verificar productos)	SP3.3-1
VAL (Validación)	SG2 (Validar productos)	SP2.2-1
OPF (Organización centrada en procesos)	SG1 (Determinar procesos a mejorar)	SP1.3-1
RSKM (Riesgos)	SG1 (Planificar gestión riesgos)	SP1.1-1, SP1.2-1, SP1.3-1
	SG2 (Identificar, analizar riesgos)	SP2.1-1, SP2.2-1
	SG3 (Gestión de riesgos)	SP3.1-1, SP3.2-1
OID (Innovación en organización)	SG1 (Seleccionar mejoras)	SP1.1-1, SP1.2-1, SP1.3-1, SP1.4-1
	SG2 (Implementar mejoras)	SP2.1-1, SP2.2-1, SP2.3-1, SP2.4-1
CAR (Análisis causal)	SG1 (Detectar, causas defectos)	SP1.1-1, SP1.2-1
	SG2 (Gestionar, causas defectos)	SP2.1-1, SP2.2-1, SP2.3-1

CMMI relaciona las prácticas genéricas y específicas aplicables a la gestión de riesgos; sin embargo, no propone algoritmos concretos para lograrlo. Se centraen el trabajo manual y la documentación exhaustiva de los procesos, más que en la determinación de fallas y errores a partir del análisis de los datos.

Atendiendo a los conceptos antes explicados se entiende la gestión de proyectos como la define el Departamento de Investigaciones de Gestión de Proyectosde la Universidad de las Ciencias Informáticas [29], que presenta la gestión de proyectos como: “un área interdisciplinaria donde convergen elementos de psicología, técnicas de dirección, gestión económica, gestión logística, conocimientos técnicos del área concreta donde se aplique, las ciencias matemáticas y las tecnologías de la información y las comunicaciones para alcanzar un objetivo bien determinado, con un

conjunto de recursos limitados, en un tiempo determinado, con una calidad deseada y a través de un conjunto de acciones organizadas de forma óptima o cuasi óptima manteniendo un balance entre costo, tiempo y calidad".

Esta misma escuelaexplica que los proyectos se conciben para lograr el cumplimiento de objetivos estratégicos de los entes que los generan y pueden ser agrupados en forma de programas para lograr un objetivo común. Añaden además, que la organización por programas permite hacer un uso más eficiente de los recursos asignados a los proyectos, considerando las prioridades que se establezcan para los mismos y que la gestión de proyectos debe estar soportada por el uso de sistemas de información que ayuden a la toma de decisiones [31-36].

Eneste trabajo se asume este concepto porque es más completo. Cuando el autor se refiere a acciones organizadas de forma óptima o cuasi óptima está incluyendo elementos que optimicen los recursos humanos y no humanos y por tanto permiten la reducción de los costos de las organizaciones [37-42].

Esta escuela al igual que otros autores plantea que existen un conjunto de errores en la gestión de proyectos que tienen un alto impacto en el aseguramiento de ingresos [43-49]. Entre los errores más comunes señalan:

1. Errores en la definición del alcance del proyecto, que generalmente provoca malas estimaciones en los costos afectando los ingresos y las utilidades.

2. Errores en la planificación, el control y el seguimiento de los proyectos, respecto al cubrimiento solo parcial de los requisitos definidos en el alcance o porque el plan no es comprendido por los miembros de la organización.
3. Poca atención a los riesgos del proyecto, que con frecuencia provoca un sobregiro en los costos del proyectoafectando los ingresos y las utilidades.
4. Deficiente gestión de cambios que provoca aumentos o modificaciones significativas en el alcance del proyecto, afectando los ingresos y las utilidades.
5. Deficiencias en la gestión de los proveedores que generalmente provocan entregas tardías de recursos, materiales de baja calidad, afectaciones en las entregas al cliente, reclamaciones y pérdidas de ingresos.
6. Empleo de la tecnología equivocada, en el caso de los proyectos de tecnologías de la información, se refleja en malas decisiones arquitectónicas que generan atrasos, generan el re-trabajo y afectan las utilidades de las organizaciones.
7. Poca interacción con los clientes, que afecta la comunicación y provoca una disminución gradual del interés por parte de los clientes, se desarrollan planes que no se cumplen por incumplimiento de ambas partes hasta provocar la cancelación del proyecto con pérdidas de ingresos y prestigio de las organizaciones.

Como estrategia para resolver estos problemas en organizaciones orientadas a proyectos, este departamento, propone en su “Banco de problemas” el desarrollo de investigaciones asociadas a la aplicación de técnicas de minería de datos [50-52].

Análisis bibliométrico asociado a la minería de datos anómalos en la gestión de proyectos

Se realizó una búsqueda en las bases de datos de la *Web of Science Book* (*Thomson Reuters*) en 2016 y bajo diferentes criterios de búsqueda se encuentran los resultados de las tablas: Tabla 4, Tabla 5 y Tabla 6. Hay que señalar que cuando se realizó la búsqueda de los términos “*Revenue Assurance*” + “*Project Management*” no se encontraron coincidencias.

Tabla 4. Criterio de búsqueda: publicaciones en los últimos cinco años.

Búsqueda	***Revenue Assurance, 5 years***		***Project management and outliers, 5 years***	
Años	**Cantidad**	**% de 31**	**Cantidad**	**% de 29**
2016	3	9,68	2	6,90
2015	3	9,68	9	31,03
2014	11	35,48	4	13,79
2012	9	29,03	9	31,03
2011	5	16,13	5	17,24

Tabla 5.Criterio de búsqueda:*(Document types).*

Criterio de búsqueda	***Revenue Assurance, 10 years***		***Project management and Outliers mining, 5 years***	
Documents types	**Cantidad**	**% de84**	**Cantidad**	**% de 29**
ARTICLE	56	66,67	22	78,57

PROCEEDINGS PAPER	24	28,57	6	21,43
REVIEW	2	2,38		
BOOK CHAPTER	2	2,38		

Tabla 6. Criterios de búsqueda: (*Web of Science Categories, Project management and outliers, últimos cinco años*).

Criterio de búsqueda	***Revenue assurance, 5 years***		***Project management and outliers, last 5 years***	
Web of Sciences categories	**Cantidad**	**% de 31**	**Cantidad**	**% de 29**
FORESTRY	2	6,45	3	10,34
COMPUTER SCIENCE SOFTWARE ENGINEERING AND INFORMATION SYSTEMS, INFORMATION SCIENCE	4	12,90	12	41,38
HEALTH CARE SCIENCES SERVICES, PHARMACOLOGY PHARMACY	9	29,03	2	6,90
ENGINEERING MULTIDISCIPLINARY	5	16,13	12	41,38
ENERGY FUELS	4	12,90		0,00
TELECOMMUNICATIONS	2	6,45		0,00
OPERATIONS RESEARCH MANAGEMENT SCIENCE AND ECONOMICS	5	16,13		0,00

Como se puede apreciar, no son frecuentes las publicaciones en este sentido, elemento que a opinión del autor de este trabajo está marcado porque existe un alto grado de privatización de las tecnologías de aseguramiento de ingresos que provoca una lenta evolución y bajo nivel de publicación de las técnicas. En el caso particular de las organizaciones orientadas a proyectos prevalecen métodos tradicionales de análisis, fuertemente motivados por la formación básica de los especialistas principales que dirigen estas áreas y que no estimulan la aplicación de técnicas

avanzadas de análisis de datos. En general se considera que esta es un área abierta a la investigación con un alto impacto económico y social [51- 52].

Minería de datos anómalos, aplicabilidad en el aseguramiento de ingresos

Se discuten en esta sección algunas de las técnicas de minería de datos anómalos (*outliers mining*) y se analiza su aplicabilidad en el problema de aseguramiento de ingresos.En lo adelante se emplea el término en español.

Ben-Gal en [53] presenta diferentes definiciones de *datos anómalos,* propuestas por varios autoresentre las que se destacan:

- Hawkins define dato anómalocomo 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.
- Barnet and Lewis definen a una observación *dato anómalo*, como una observación que se desvía marcadamente de otros miembros de la muestra en la cual se encuentra.
- Johnson define que un *dato anómalo*es una observación en los datos que aparece como inconsistente con respecto al resto de los datos.

Las técnicas para la detección de datos anómalos se muestran en disímiles escenarios de aplicación [54-60] entre los que destacan: detección de fraudes en tarjetas de créditos y las telecomunicaciones, errores en las planificaciones, detección de precios de productos manipulados, en el procesamiento de imágenes,

en fraudes en ensayos clínicos, análisis de irregularidades en procesos de votación, en la detección de intrusos en redes, en cambios severos del clima, en análisis criminalista en otras áreas.

Existen varios enfoques para la caracterización de los escenarios de aplicación de las técnicas de minería de datos anómalos[13, 61-65] entre los que se destacan:

- Respecto a la naturaleza de los datos de entrada: que se refiere a la heterogeneidad de los datos. Los datos pueden ser: objetos, puntos, vectores, eventos, observaciones, entidades entre otros. Además, cada instancia puede ser escrita como un conjunto de atributos con dominio: binario, categórico, continuo, series de tiempo, datos espaciales, entre otros.
- La dimensionalidad de los datos y el volumen de información a procesar.
- La incertidumbre y la imprecisión de los datos: que se refiere a la calidad de los datos y la certeza de los valores que representan.
- Tipos de *datos anómalos:* los datos anómalos pueden ser de tres tipos diferentes: puntuales, contextuales o colectivos.
- Respecto a la forma de presentación de los *datos anómalos*: estrategias basadas en puntuación o *ranking*o métodos basados en etiquetado.

Respecto a los tipos de *datos anómalos* se describen las siguientes subcategorías:

- *Datos anómalos*puntuales: cada registro puede ser considerado de forma independiente como un registro anómalo o no con respecto al resto de los

datos. Éste es el que siguen la mayoría de los algoritmos tradicionales [66-71]. En este enfoque existen las siguientes subclasificaciones:

- Respecto a la distribución probabilística los *datos anómalos*son clasificados en *datos anómalos*univariados y *datos anómalos*multivariados.
- Respecto al alcance del*dato anómalo*: se subdividen en *datos anómalos*locales (*datos anómalos* respecto a sus vecinos) o globales (datos inconsistentes con respecto al resto de todos los demás registros).

- *Datos anómalos* colectivos: representan una subcolección de datos que son anómalos respecto al conjunto completo de datos, pero si esa subcolección se analiza de forma independiente ninguno de sus registros destaca por sí sólo un *dato anómalo*.
- Los *datos anómalos* contextuales se identifican porque existen en el problema en cuestión atributos de contexto que pueden definir si determinado registro es o no un dato anómalo. Por ejemplo,respectoa la temperatura histórica en diferentes momentos del año del hemisferio norte, una temperatura baja entre los meses de junio y agosto es un *dato anómalo*.

Respecto a la organización de las estrategias empleadas para la identificación de *datos anómalos*se plantean los siguientes enfoques*:*

- Enfoque que agrupa los métodos en: paramétricos que incluyen métodos estadísticos para el análisis de datos univariados y multivariados, y no-paramétricos que incluyen técnicas de agrupamiento [53].
- Enfoque que divide los métodos en dos grupos: métodos clásicos de *datos anómalos* (basados en distancia, basados en densidad y estadísticos) y métodos basados en la distribución espacial de los objetos según Karanjit Singh [54].
- Enfoque que organiza los métodos en: métodos supervisados, métodos semi-supervisados y métodos no supervisadossegún Manish Gupta [61]. Este es el enfoque que se sigue en este trabajo y que se explica en la siguiente subsección.
- Enfoque que diferencia los métodos estadísticos, los métodos basados en la proximidad, los métodos supervisados y los meta-modelos [13, 72-74].

Métodos no supervisados para la detección de datos anómalos

Métodos basados en técnicas estadísticas

En este enfoque se destaca el uso de métodos de estadística descriptiva, el análisis de histogramas y el análisis basado en funciones de densidad probabilística. Se explican a continuación algunos de estos métodos y sus limitaciones.

Método basado en el análisis de las funciones de distribución probabilística

Los métodos paramétricos univariados y multivariados se basan en el análisis de la distribución probabilística de los datos: en este fin, estos métodos, determinan a priori la distribución de los datos y se identifican como *datos anómalos* aquellos objetos cuyas variables no cumplan con la distribución predeterminada. Los métodos

univariados consideran que las variables son independientes mientras que los multivariados consideran las interacciones entre las variables [75-97].La base de estos métodos está en la desigualdad de Markov [80] y en la desigualdad de *Chebyshev* [81].Ejemplos de aplicaciones de este enfoque se presentan en el trabajo de Deneshkumar [67] en la identificación de factores que influyen en la diabetes y en trabajo de Zhiguo Li, Robert y otros [98].

Como limitaciones fundamentales de estos modelos se encuentran:

- El trabajo con datos heterogéneos, lasrestricciones temporales y el momento de detección del*dato anómalo*, si es en tiempo real o no.
- Presuponen el conocimiento de la función de densidad probabilística de los datos, elemento que puede provocar detección excesiva de datos sospechosos.
- Además, estos enfoques en su esencia, no contemplan la incertidumbre de la información basando su análisis en criterios duros, para la detección de si un dato es anómalo o no.

Métodos basado en el análisis de las desviaciones de los datos

Estos métodos se basan en el análisis de la variación de las medidas de tendencia central de los datos a partir de remover determinados puntos del conjunto [99]. Estos métodos suponen que si se quitan los *datos anómalos* del conjunto de datos como estos representan valores extremos entonces se logran variaciones significativas de

la media y la varianza de los datos restantes. Tienen como inconveniente su complejidad computacional 2^N, siendo N el tamaño del conjunto de datos.

Métodos basados en el análisis de histogramas de frecuencia

Los histogramas son simples de construir y particularmente útiles en escenarios univariados basados en densidad. Para el uso de este método es recomendable la discretización del conjunto de datos y aquellos conjuntos con baja densidad son reportados como datos anómalos. Entre las dificultades para la aplicación de este método se encuentran:

- La determinación de la amplitud óptima de los grupos de frecuencia. Histogramas muy amplios o muy estrechos pueden no modelar con la granularidad requerida el espacio de búsqueda dificultando la detección de los *datos anómalos*.
- Estas son técnicas basadas en el análisis local y con frecuencia no consideran las características globales del espacio de búsqueda, no trabajan bien en espacios de alta dimensionalidad.

Métodosbasados en regresión lineal

Los métodos basados en la regresión lineal tienen algunas dificultades para su aplicación en la detección de datos anómalos entre las que se encuentran [82,100-101]:

- Depende de que exista una alta correlación entre las variables que describen a los datos y que funcionan mejor en espacios de baja dimensionalidad.

- Aportan poca interpretabilidad a los resultados finales donde se espera poder explicar a los interesados, porque determinado dato es un dato anómalo.
- Además, se debe tener cuidado porque también se puede producir un sobre-ajuste de los datos provocando disfrazar los verdaderos *datos anómalos*.

Métodos basados en la proximidad de los datos

Las tres categorías de métodos más reconocidos en este enfoque son:métodos basados en distancias, métodos basados en agrupamientos y métodos basados en densidad.

Métodos basados en distancia

Estos métodos se basan en detectar la proximidad entre los puntos, consideran como *datos anómalos* a aquellos puntos con mayor distancia al resto de sus vecinos. Pioneros en estos métodos fueron Knorr y Ng [84, 85]además han sido ampliamente usados como se muestra a continuación.

Amol Ghoting y otros en [86] proponen algoritmos basados en distancia para la detección de *datos anómalos* en espacios de alta dimensionalidad. Bajo este enfoque se encuentran trabajos [53, 66, 87, 102- 111] que comparan a la distancia Euclidiana y la distancia de Mahalanobis y que reportan mejores resultados para esta última [107].

Prasanta Gogoi en [88] propone otro método de detección de *datos anómalos* basado en la detección de simetrías de las relaciones entre vecinos más cercanos y

la simetría de las distancias hacia los mismos tomando como base la distancia Euclidiana entre los objetos.

Otro método basado en distancias es el método basado en la construcción de celdas y el análisis de la proximidad entre los datos contenidos en las mismas [13]. Como principal deficiencia a este método se señala que es exponencial respecto a la dimensionalidad, aunque su complejidad es lineal respecto al conjunto de datos.

Otros autores proponen el uso de heurísticas e índices que ayuden a disminuir la complejidad de los algoritmos y la cantidad de comparaciones requeridas en los métodos tradicionales basados en distancias [89].

Estos métodos reportan sus mejores resultados en escenarios con relativamente pocos datos a diferencia de los otros métodos basados en proximidad.La mayor diferencia entre estos métodos y los basados en agrupamientos está en la granularidad empleada durante el proceso de análisis, elemento que permite que, en ocasiones, los algoritmos basados en distancia sean más robustos [68]. Entre las ventajas de los métodos basados en distancias se encuentran:

- No necesitan conocer a priori la distribución de los datos y pueden ser aplicados en espacios de búsqueda sobre los que se pueda definir una medida de distancia.
- Estos métodos generalmente permiten mayor nivel de granularidad que otros métodos facilitando la diferenciación entre datos ruidosos y *datos anómalos*.

Como limitación fundamental de estos métodos se puede señalar su alta complejidad computacional $O(n^2)$ provocada por las numerosas comparaciones que necesitan hacer.

Métodos basados en densidad

Los métodos basados en densidad se centran en identificar las regiones del espacio basadas en la densidad de los datos en el espacio. Los métodos basados en densidad pueden ser muy útiles respecto a su capacidad de interpretabilidad cuando los espacios son mostrados como combinación de atributos [90-94].

Uno de los métodos basados en densidad más reconocidos es el método “Factor local de datos anómalos” (LOF) [90]. Este método se basa en la cuantificación del grado de anomalía de los datos, se plantea además que este grado puede ser ajustado ante diferentes escenarios de variaciones de densidad.Otro método similar esel método local de correlación integral (LOCI)[92], donde se define la densidad $M(X, \varepsilon)$de un dato Xen función del número dedatos dentro de un radio εpre-definido.

Diferentes autores como Changyong Lee&Hakyeon Lee en [95-96] y Ania Cravero en [99], presentan los métodos basados en densidad combinados con otras técnicas. Por ejemplo en [95-96] se combinan estadígrafos básicos con el algoritmo LOF y se detectan tanto datos anómalos univariados como multivariados.

Métodos basados en agrupamientos

Estas técnicas se subdividen en métodos jerárquicos, métodos basados en partición, métodos basados en cuadrículas y métodos basados en restricciones [53, 67, 112-115]. A continuación, se describen algunos de ellos:

1. Agrupamiento jerárquico, que produce unadescomposición jerárquica del conjunto de datos,creando un gráfico conocido como dendograma que representa la forma de agrupación. Estos métodos generalmente generan un grupo con los elementos que están demasiado dispersos y requieren que sea cuidadoso al identificar a este grupo en particular.
2. Métodos basados en particiones, realizan divisiones sucesivas del conjunto de datos. Los objetos se organizan en *k* grupos, de modo que la desviación de cada objeto debe reducirse al mínimo en relación con el centro de la agrupación.

Singh Vijendra y Pathak Shivani en [75] proponen como enfoque para la detección de *datos anómalos*un algoritmo en dos pasos, primero aplican técnicas estadísticas y luego agrupamientos. Como algoritmo para la formación de los grupos proponen el uso de *K-means*clásico dependiendo de la selección a priori de la cantidad de grupos y de los centroides.

Shivani P. Patel y Vinita Shah en [116] proponen un enfoque híbrido. Emplean una variación de *K-means* y seleccionan los centroides de los grupos antes de ejecutar el

algoritmo. Además, proponen el cálculo dinámico del umbral de distancia permitiendo la generación de más grupos en tiempo de ejecución.

Como principal ventaja de los métodos basados en agrupamientos se señala que permiten un análisis global de los datos, detectando pequeños grupos de datos aislados. Mientras que como principal limitación se les señala que en ocasiones estos métodos por su naturaleza no logran discernir con claridad si se está en presencia de datos realmente anómalos o si son datos ruidosos o débiles.

Métodos basados en el análisis espacial de los datos

Estos métodos son bastante cercanos a los métodos de agrupamientos, se basan en el principio de que un *dato anómalo* en el espacio es un objeto que al representar sus atributos en el espacio estos son significativamente diferentes de sus objetos vecinos a él. Estos métodos son clasificados en dos subcategorías: métodos cuantitativos y métodos gráficos. Los métodos cuantitativos se basan en pruebas que distinguen a los datos anómalos espaciales del resto de los datos. Los métodos gráficos están basados en la visualización espacial de los datos, su implementación es más sencilla en espacios de baja dimensionalidad [53, 66]. Se relacionan a continuación algunos métodos de este enfoque.

Método basado en la profundidad

Este método se basa en construir poliedros convexos a partir de los puntos externos al conjunto de datos y obtener estos conjuntos por iteraciones sucesivas [13]. Luego se define un umbral de decisión que determina cuáles son los puntos considerados

datos anómalos.Este método adolece de la dificultad de encontrar los *datos anómalos* interiores en el conjunto de datos, solo determina con cierta facilidad los que se encuentran geométricamente por fuera del conjunto de datos.

Método basado en los ángulos

Este método supone que los *datos anómalos* son aquellos datos que, al formar diferentes ángulos, con cualquiera del resto de los puntos del conjunto de datos, se encuentra que no varían significativamente la amplitud de dichos ángulos. Este método se basa en la relación que existe entre la distancia entre los puntos y el coseno del ángulo inscrito entre los segmentos que forman dichos puntos [117,118].

Método basado en proyecciones de subespacios dimensionales

Zhana Bao en [77, 119] propone un interesante método para la detección de datos anómalos basado en métodos de minería de subespacios como una solución práctica para el cubrimiento de la alta dimensionalidad. Este método resuelve algunos de los problemas de los métodos basados en distancia tradicionales. Pero tiene un elemento que debe cuidarse en su aplicación que es el diseño de las proyecciones a construir. Esto significa que se debe considerar las relaciones entre las variables que conforman a los objetos en el proceso de diseño de las proyecciones, las variables altamente correlacionadas deben formar parte de la misma proyección.

Resumen de dificultades de los métodos basados en análisis espaciales

En [53] se referencia a un artículo publicado por Lu donde se hace una comparación entre tres métodos espaciales de detección de datos anómalos, uno de los algoritmos basado en la mediana como una medida más robusta que la media y muestran buenos resultados de estos enfoques en la reducción de riesgos de detectar *datos anómalos*falsos negativos. En otro trabajo Karanjit Singh y Dr. Shuchita Upadhyaya en [54] identifican diferentes factores que dificultan la detección de los *datos anómalos*con este enfoque y señalan:

- Imprecisión en los límites de las regiones y el efecto de intercambio de *datos anómalos.*
- El dinamismo de los escenarios que provoca que en algunos dominios los registros representativos de comportamiento normal en un momento del tiempo no se mantengan en el tiempo.
- Ambigüedad en los conceptos de *dato anómalo*para diferentes dominios de solución.
- Ruido presente en los datos que con frecuencia hace parecer iguales a los registros *datos anómalos*y los registros normales dificultando su detección.

Métodos semi-supervisados para la detección de datos anómalos

Estos métodos son aplicables en escenarios semi-supervisados, donde se conocen algunos *datos anómalos* o al menos donde se conocen *datos anómalos* de algunas

clases, pero no de otras.Algunas de las técnicas de este enfoque se presentan a continuación.

Método basado en detección de nuevas clases

Una estrategia empleada es descubrir datos que tengan un comportamiento que los diferencie de clases ya conocidas. Se procede a entrenar los algoritmos basados en técnicas de aprendizaje supervisado con bases de ejemplos con datos normales garantizando que no contengan *datos anómalos*. La idea es encontrar aquellas clases que se diferencien tanto como sea posible de las clases representadas en los datos de entrenamiento. Un ejemplo de aplicación de este método se presenta en [120,121] donde se ha adaptado el método SVM para este fin.

Combinación de detección de nuevas clases con métodos de detección de clases raras

En algunos escenarios existen casos correspondientes a clases raras o anómalas en el conjunto de entrenamiento, pero aún existen nuevas clases no detectadas y que necesitan descubrirse. Un ejemplo de estos escenarios es la detección de intrusos. Los casos correspondientes a clases raras ya etiquetadas proveen información relevante que puede ser usada por los métodos supervisados, pero se requiere combinar estos con los no supervisados para detectar las nuevas clases. Ejemplos de estos métodos se presentan en [122-124].

Método basado en aprendizaje activo

En [125-126] se propone un método basado en una técnica que los autores han llamado aprendizaje activo. En este método los datos son clasificados por

iteraciones. En cada iteración solo unos conjuntos de datos son identificados y clasificados con la intervención de expertos humanos que clasifican o ratifican la clasificación realizada por los algoritmos de los datos analizados. En la primera iteración los autores proponen el uso de métodos no supervisados, los datos ya clasificados pueden ser empleados utilizando técnicas supervisadas en nuevas iteraciones.

Algunas dificultades asociadas a estos métodos son la incertidumbre y la ambigüedad durante las primeras iteraciones. Técnicas de consenso de expertos que pueden ser empleadas en estos casos para disminuir la posibilidad de incertidumbre en las clasificaciones y garantizar un adecuado aprendizaje de los patrones de comportamiento de los datos anómalos identificados.

Métodos supervisados para la detección de datos anómalos

Las técnicas tradicionales de clasificación también pueden ser empleadas en la detección de datos anómalos. En [127-131] se presentan aplicaciones de estos métodos en la detección de defectos en sistemas, detección de fraude financiero y en detección de robots actuando sobre la web. En general se recomienda su uso siempre que sea posible y considerando las situaciones específicas que se presentan ante los problemas de detección de datos anómalos.

Una de estas situaciones se conoce como escenario con clases contaminadas: en muchos escenarios reales se conocen con seguridad algunos datos normales mientras que hay conjuntos de datos que se señalan como normales pero que

realmente no lo son o al menos que hay dudas si son o no anómalos. Un ejemplo de aplicación con esta situación es presentada en [131-135].

Otra de estas situaciones es la presencia de clases desbalanceadas que se despliegan en escenarios cuando no está balanceada la cantidad de casos de diferentes clases [121, 128]. En este escenario generalmente son aplicadas dos estrategias diferentes: la primera basada en la penalización de las clases con numerosos datos [132, 136-143] y la segunda basada en el re-muestreo adaptativo [133]. Respecto al re-muestreo adaptativo se puede señalar que como deficiencia tiene la pérdida de casos de entrenamiento en detrimento de la calidad del clasificador.

Métodos basados en la modificación de métodos basados en el conocimiento

La mayoría de los algoritmos de los sistemas basados en el conocimiento pueden ser modificados de forma sencilla considerando los costos. Una forma de lograrlo es penalizar con un costo ante cada error en la clasificación de cada instancia del conjunto de entrenamiento. Ejemplos de la aplicación de este enfoque se presenta en [111].

Métodos supervisados basados en árboles de decisión

En los árboles de decisión la idea fundamental es particionar el conjunto de datos considerando los valores de sus atributos, siguiendo como criterio de selección de atributos durante el proceso de partición medidas como la entropía u otras. Combinaciones de estrategias basadas en el uso de métodos basados en la

sensibilidad de los costos de clasificación con los árboles de decisión están centradas en diferenciar los costos de las regiones con mayor cantidad de datos correspondientes a clases raras.Ejemplos de estas aplicaciones de este algoritmo se encuentran en [133-134].

Métodos basados en el aprendizaje de reglas de asociación

A.M.Rajeswari, M.Sridevi y C.Deisy en [136] proponen la aplicación de un método basado en el descubrimiento de reglas de asociación para la detección de *datos anómalos* en bases de datos educacionales. Emplean el soporte y la confianza como medidas para identificar las reglas descubiertas que pueden representar *datos anómalos*. Definen en su propuesta cuatro pasos: el primer paso es la fuzzificación de los datos de entrada, el segundo y tercer paso es la identificación de items raros a partir del *ranking* de los estudiantes, finalmente el cuarto paso es la generación de reglas de asociación. Las reglas de asociación raras pueden representar *datos anómalos.*

Métodos basados en redes neuronales

En [53] Ben-Gal hace una comparación de diferentes enfoques y señala que los métodos basados en redes neuronales han demostrado ser efectivos y son particularmente buenos en grandes volúmenes de datos.

Graham Williams, Rohan Baxter y otros [137] proponen el empleo de una red neuronal artificial basada en perceptrón multicapas para la detección de *datos anómalos*. En la topología de la red proponen el empleo de tres capas ocultas y un

mismo número de neuronas tanto para la entrada como para la salida. Las capas de entrada y salida tienen *n* neuronas correspondiendo las mismas a las características del espacio de búsqueda, una neurona por cada una de los atributos que describen a los registros u objetos de análisis.

Métodos basados en conjuntos aproximados

En [138, 139] se presenta el algoritmo *NREOD* para la detección de *datos anómalos* que combina la teoría de los conjuntos aproximados con el cálculo de la entropía. En el primer paso aplican la teoría de conjuntos aproximados para determinar los conjuntos de objetos que pertenecen a las aproximaciones inferior y superior de cada clase de equivalencia. En un segundo paso calculan la entropía de estos conjuntos.

Como elemento a criticar de este trabajo es que: solo es capaz de detectar *datos anómalos* siguiendo un enfoque puntual del análisis de los datos obviando el enfoque contextual de los *datos anómalos* o el enfoque colectivo en la detección de los mismos. Como otro elemento se puede identificar su complejidad computacional que es $O\left(n^2 m^2\right)$.

Métodos basados en meta-heurísticas para la detección de datos anómalos

La búsqueda exhaustiva en espacios grandes de datos anómalos tiene un alto costo computacional. Además, existen escenarios con mucho ruido donde hay un alto nivel de datos esparcidos, problema que es frecuente en el caso de existir alta dimensionalidad. Una estrategia posible es detectar los datos anómalos en dos pasos, primero disminuir la dimensionalidad del espacio a partir de identificar

proyecciones de los datos y luego en las regiones proyectadas detectar los anómalos. El objetivo de la identificación de las proyecciones está en lograr disminuir el nivel de esparción de los datos, logrando espacios más compactos con menor nivel de ruido y que faciliten la detección de los verdaderos datos anómalos. Para lograr esto, algunos autores han propuesto el uso de metaheurísticas y de otros métodos basados en la combinación de técnicas[140-143]. En [142] los autores proponen una variante rápida para un método de búsqueda local basado en la heurística del glotón. En el propio artículo se refieren a las dificultades del método clásico de búsqueda local basado en el glotón por su alta complejidad temporal en grandes volúmenes de datos. En el mismo artículo se relacionan otros trabajos que modelan el proceso de detección de fraudes sobre datos simbólicos como un problema de optimización y usan algoritmos de búsqueda local para resolverlo.

Conclusiones del capítulo

Se presentan a continuación las conclusiones parciales del capítulo:

- El origen del aseguramiento de ingresos estuvo asociado a las empresas de telecomunicaciones, pero su aplicación se ha extendido a numerosas áreas del conocimiento humano, y se identifica una línea abierta a la investigación: la aplicación de estas técnicas en las organizaciones orientadas a proyectos y que combinen los enfoques proactivos, activos y reactivos.
- Para la aplicación de las técnicas de aseguramiento de ingresos se deben considerar las especificidades de cada escenario y la naturaleza de sus datos.

- En las organizaciones orientadas a proyectos ocurren errores frecuentes que afectan los planes y la ejecución de los proyectos, provocando gran cantidad de proyectos cancelados o renegociados con alto impacto económico y social. Una estrategia para resolver estos problemas se centra en el desarrollo de técnicas para el aseguramiento de ingresos que combinen técnicas de minería de *datos anómalos*, las buenas prácticas en gestión de proyectos y las técnicas de *soft computing*.

2. CAPÍTULO: MODELO PARA EL ASEGURAMIENTO DE INGRESOS EN ORGANIZACIONES ORIENTADAS A PROYECTOS

En este capítulo se presenta un modelo para la implantación de técnicas de aseguramiento de ingresos en organizaciones orientadas a proyectos. Está dividido en las siguientes secciones: conceptualización del modelo propuesto, pasos para la instrumentación del modelo y conclusiones.

Conceptualización del modelo propuesto

A partir del análisis de los conceptos anteriormente planteados, se considera en este trabajo al "Aseguramiento de ingresos" como: el conjunto de técnicas, políticas y modelos, aplicados con el objetivo de aumentar los ingresos y disminuir los costos de las organizaciones que las apliquen siguiendo enfoques reactivos, activos y proactivos. Se muestra como un área interdisciplinar donde convergen las tecnologías de bases de datos, la estadística, la minería de *datos anómalos*, las técnicas de *softcomputing*, técnicas de la computación emergente y del área específica de aplicación.

Además, es importante conocer que no existen soluciones deterministas o únicas, para el aseguramiento de ingresos, aplicables a todas las organizaciones. Esto ocurre porque en cada escenario existen variables, factores internos y externos específicos, con elevado impacto en los ingresos y en las decisiones para su gestión.

Por estas razones se propone un modelo compuesto por tres componentes Figura 2, que puede ser adaptado según las especificidades de cada escenario real de aplicación.

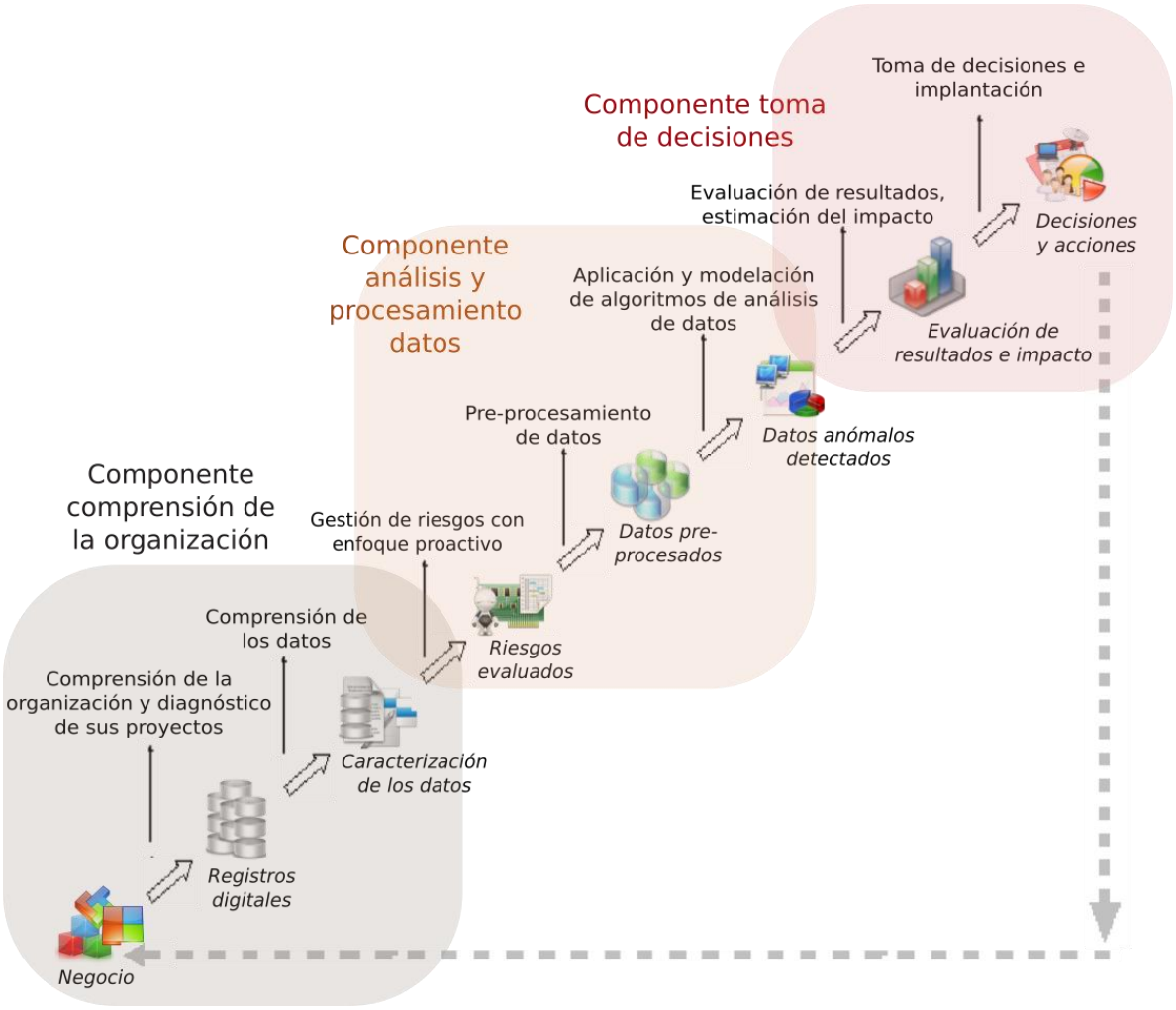


Figura 2. Representación gráfica del modelo para el aseguramiento de ingresos.

Las premisas del modelo propuesto son:

- Requiere la presencia de expertos para la instrumentación del mismo.

- Requiere conocimientos de análisis de datos y aseguramiento de ingresos para su aplicación, porque combina, técnicas de gestión de proyectos con técnicas de minería de *datos anómalos* para la detección de situaciones que afectan los procesos de aseguramiento de ingresos en las organizaciones.

Los componentes del modelo interactúan bajo un enfoque sistémico con una estrecha relación entre ellosFigura 2:

- Componente comprensión de las organizaciones: incluye técnicas y diagramas que permiten guiar el proceso de comprensión del negocio de las organizaciones objeto de análisis. Este componente genera como salida, un documento de diagnóstico, una taxonomía con los principales problemas que pueden generar pérdidas de ingresos y el modelo de datos que caracteriza los datos recogidos en los sistemas de información de la organización.
- Componente análisis y procesamiento de datos: este componente incluye grupos de algoritmos y técnicas para el análisis de los datos que reflejan la actividad de las organizaciones analizadas. Recibe como entrada las salidas del primer componente y genera como salida el informe de análisis de resultados, el listado de los riesgos priorizados y el listado de los *datos anómalos*.
- Componente toma de decisiones: componente asociado a la evaluación de resultados obtenidos en el segundo componente y la implantación de las soluciones. Recibe como entrada las salidas del componente de análisis y

genera como salidas las decisiones tomadas y las lecciones aprendidas. Todas las salidas de este componente constituyen entradas para el primer componente y permitiendo ajustar y mejorar el comportamiento del modelo en el escenario real de aplicación.

Para la instrumentación del modelo se siguen los procesosde la Figura 3 descritos enIDEF0 [138] (*Integration Definition for Function Modeling*).

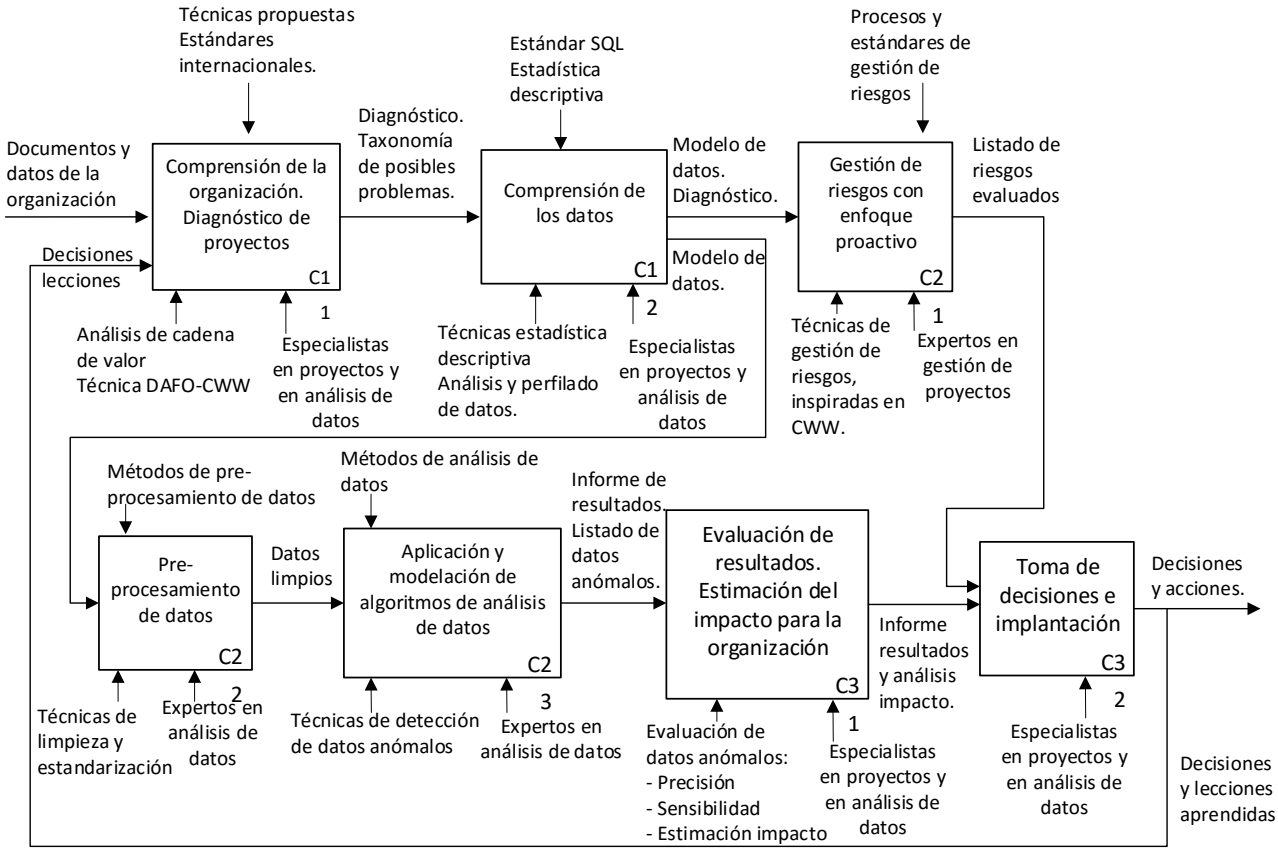


Figura 3.Instrumentación del modelo para el aseguramiento de ingresos en IDEF0.

Descripción de los procesos de instrumentación del modelo

Proceso 1. Comprensión de la organización: proceso que tiene como objetivo el diagnóstico de la organización y definir una taxonomía que ayude a identificar las situaciones anómalas que afecten los ingresos.

Proceso 2. Comprensión de los datos: tiene como objetivo la caracterización y la construcción del modelo de datos recogidos en los sistemas de información de la organización. Analiza la naturaleza de los datos, la existencia de valores ausentes y los errores más frecuentes en los mismos.

Proceso 3. Gestión de riesgos: se basa en la aplicación de técnicas de gestión de riesgos combinadas con *soft computing*para la planificación y la evaluación cualitativa de los riesgos generando listado de riesgos priorizados.

Proceso 4. Pre-procesamiento de datos: en este proceso se aplican las técnicas de limpieza de datos para la eliminación de posibles ruidos que afecten la detección de los verdaderos datos anómalos. Se ejecutan actividades de limpieza, estandarización y la selección de los atributos.

Proceso 5. Aplicación y modelación de algoritmos de análisis de los datos: tiene como objetivo el diseño y la aplicación de algoritmos para la detección de *datos anómalos* que reflejen pérdidas de ingresos.

Proceso 6. Evaluación de los resultados, estimación de impacto: en este proceso se procede a estimar el impacto de los mismos.

Proceso 7. Toma de decisiones: se toman decisiones y los resultados son recogidos en forma de lecciones aprendidas, lo que permitela sostenibilidad en la aplicación de las técnicas de aseguramiento de ingresos en la organización.

Proceso 1. Comprensión dela organización y diagnóstico de sus proyectos

En este proceso se realizan actividades para el diagnóstico de la organización a partir de aplicar las técnicas: análisis de cadena de valor[3][144], DAFO-CWW [145]. Se aplica el siguiente algoritmo.

Algoritmo para el diagnóstico

Se basa en la combinación de las técnicas de la cadena de valor, latécnica DAFO para el análisis de las fortalezas y debilidades y el modelo 2-tuplas de computación con palabras.En este caso esta técnica tiene como objetivo identificar los procesos de la organización donde ocurren la mayor cantidad de fugas de ingresos y se combina esta técnica con la matriz DAFO-CWW, verAlgoritmo 1.

Algoritmo 1:Diagnóstico y organización respecto a elementos que influyen en los ingresos

[3]Técnica de análisis empresarial mediante la cual se descompone una empresa en sus partes constitutivas, buscando identificar fuentes de ventaja competitiva en las actividades generadoras de valor [144]. Su objetivo es lograr que la empresa desarrolle las actividades de su cadena de valor de forma menos costosa y diferenciada que sus rivales. En este caso, se aplica esta técnica para identificar los procesos donde ocurren la mayor cantidad de fugas de ingresos.

Paso 1. Seleccionar al conjunto de m expertos con conocimientos suficientes del proceso de generación de valores de la organización. Este conjunto estará denotado por E, identificándose el i- ésimo experto como e_i.

Paso 2. El conjunto de expertos E en mesa de trabajo construyen la secuencia L, de actividades primarias[4] y de soporte[5], que forman la cadena de valor de la organización. Para cada actividad L_j calcular el margen[6] m_j, tal que: m_j = Ingreso total$_j$–Costo$_j$.

Paso 3. A partir del análisis de la cadena de valor, construir una matriz DAFO identificando elementos que influyen en cada actividad primaria tanto en la reducción de los costos como en la mejora de los ingresos y que afectan el margen m_j de cada actividad.

Paso 4. Al concluir este paso se tienen seis grupos de elementos que influyen en los ingresos de la organización como se muestra en la Tabla 7:

Tabla 7.MatrizDAFO empleada por los expertospara evaluar las actividades de la cadena de valor.

Análisis externo	***Análisis Interno***	
	Fortalezas	***Debilidades y errores***
	- *Ventajas naturales, competencias.* - *Fortalezas en control que evitan las pérdidas de ingresos.*	- *Errores de organización.* - *Fuentes de posibles fraudes.* - *Costos elevados afectan los ingresos.*
Oportunidades - *Mejoras tecnológicas* - *Posicionamiento estratégico*	*Estrategia (FO)(Max-Max)* *Grupo1:Identificar nuevos servicios o productos, generadores de nuevos ingresos basados en las oportunidades y fortalezas.*	*Estrategia (DO)(Min-Max)* *Grupo2: Identificar fuentes de errores que afectan el aprovechamiento de las oportunidades o que generan pérdidas de ingresos.*
Amenazas - *Altos riesgos* - *Cambios en el entorno*	*Estrategia (FA)(Max-Min)* *Grupo5: Identificación,a partir del análisis de las amenazas, de riesgos externos y fuentes que pueden provocar pérdidas de ingresos.* *Grupo6: Identificación de actividades que*	*Estrategia (DA)(Min-Min)* *Grupo3: Identificar errores o posibles acciones anómalasque potencien las amenazas afectando a los ingresos.* *Grupo4: Identificar actividades para*

[4]Actividades primarias: son las relacionadas con la producción, la logística, comercialización y los servicios de post-venta.

[5]Actividades de soporte: son aquellas que apoyan el desarrollo de las actividades primarias, tales como las de administración, las de desarrollo tecnológico y las de compras de bienes y servicios.

[6]Margen diferencia entre el valor total y los costos totales incurridos para realizar la actividad generadora de valor.

	basadas en las fortalezas ayuden a evitar o mitigar las pérdidas de ingresos por amenazas externas.	*mitigar o evitar las amenazas.*

Paso 5. Evaluar cada uno de los elementos contenidos en los 6 grupos identificados, empleando técnicas de computación con palabras. Se define un conjunto básico de términos lingüísticos LBTL = {Ninguno, Muy bajo, Bajo, Medio, Alto, Muy alto, Perfecto}, para la evaluación de los elementos, basado en el grado de impacto en los ingresos ya sea positiva o negativamente, ver Figura 4.

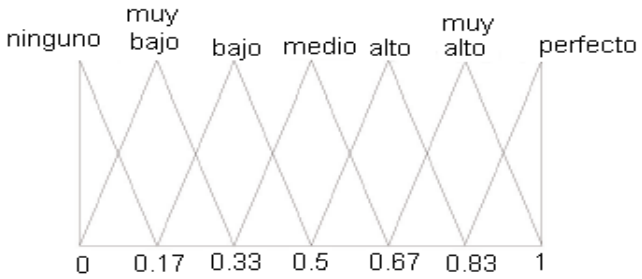


Figura 4. Variable lingüística "*Evaluación de Impacto",*para la evaluación de las fortalezas y debilidades.

Los expertos evalúan cada elemento como muestra la Tabla 8.

Tabla 8. Estructura de la evaluación de los expertos.

Grupo	**Elementos**	**Expertos**		
		e_1	…	e_m
Grupo1	*Elemento₁*	x_1^{11}	…	x_1^{1m}
	⋮	⋮	⋮	⋮
Grupo6	*Elemento$_p$*	x_1^{p1}	…	x_1^{pm}

Paso 6. Siguiendo el modelo 2-tuplas de computación con palabras se agregan las evaluaciones de los expertos consolidando las mismas por cada elemento a evaluar. Al finalizar se tiene un listado de elementos que influyen en los ingresos, agrupados por su naturaleza en los seis grupos.

Paso 7. Para mitigar, evitar o potenciar cada elemento en función de su impacto en los ingresos, se propone un conjunto de acciones a ejecutar, generalmente acciones de grupos diferentes pueden ejecutarse en paralelo.Los elementos de los grupos 2, 3 y 5 concentran los factores internos o externos, las posibles fuentes de errores y otras situaciones anómalas que afectan los ingresos. Los elementos de los grupos 4

y 6 constituyen acciones recomendadas y que serán empleadas siguiendo un enfoque proactivo en el proceso 3 de este modelo. Los elementos del grupo 1 están orientados a explotar las oportunidades y serán considerados siguiendo un enfoque proactivo en el proceso 3 de este modelo.

Proceso 2. Comprensión de los datos

En este proceso se emplean técnicas de estadística descriptiva combinadas con las facilidades de los sistemas de información y el lenguaje SQL para la recuperación de datos, como se muestra a continuación:

- Técnicas de recopilación de información: análisis de documentos, análisis de las bases de datos, entrevistas personales o grupales con los especialistas con más experiencia en el trabajo con los datos de la entidad para la construcción del modelo de datos.
- Técnica de perfilado de datos: que permiten entender la estructura de los datos sus características y la calidad de los mismos [146].
- Identificación de situaciones generadoras de datos anómalos con reflejo en las bases de datos de la organización, se muestran a continuación algunas de estas situaciones:

Situación 1. Se consideran datos anómalos puntuales, los registros de tareas que no tengan asignados recursos con las competencias requeridas o que por su volumen requieran más recursos humanos de los asignados.

Situación 2. Se consideran datos anómalos puntuales, registros de tareas cuya estimación de tiempo o costo esté por encima o muy por debajo de los valores previstos.

Situación 3. Son datos anómalos puntuales, los registros de tareas que no respetan en el cronograma la precedencia lógica o que tengan una holgura de espera excesivamente alta respecto a otras tareas.

Situación 4. Se consideran datos anómalos puntuales, los registros de requisitos en el EDT[7] del proyecto para los cuales no hay tareas registradas en el cronograma del proyecto, dedicadas al desarrollo de los mismos.

Situación 5. Se consideran datos anómalos colectivos, los registros de proyectos que, aunque son similares a otros por su alcance, pueden tener costos estimados muy por encima de la media.

Situación 6. Se consideran datos anómalos de contexto, las planificaciones registradas que muestran una sobrecarga de recursos humanos o no humanos en escenarios de desarrollo de múltiples proyectos simultáneamente.

Proceso 3. Gestión de riesgos con un enfoque proactivo

En este proceso se aplican las técnicas y procesos del PMBOK [3]. Pero, se introduce una técnica nueva en los procesos de análisis cualitativo de los riesgos. En

[7]Estructura de desglose de trabajo, técnica del PMBOK [3]

particular se propone una técnica basada en el modelo de computación con palabras 2-tuplas [147] en lugar de la técnica propuesta por el PMBOK, ver el Algoritmo 2.

Algoritmo 2: Evaluación de los riesgos usando 2-tuplas

Paso 1. Seleccionar el conjunto de m expertos. Este conjunto estará denotado por E, identificándose el i- ésimo como e_i.

Paso 2. Identificación de los riesgos con impacto en el aseguramiento de ingresos.

Paso 3. Se establecen tres criterios de evaluación de riesgos: probabilidad de ocurrencia, el impacto y la facilidad de detección (se considera en este sentido lo establecido por la resolución 60 [148]).

Paso 4. Evaluar cada uno de los riesgos identificados empleando técnicas de computación con palabras. Se define un conjunto básico de términos lingüísticos LBTL = {Ninguno, Muy bajo, Bajo, Medio, Alto, Muy alto, Perfecto}, para la evaluación de los elementos, basado en el grado de impacto en los ingresos. Ver Figura 4. Los expertos evalúan cada elemento como muestra laTabla 9.

Tabla 9. Tabla de evaluación de riesgos, aplicando la computación con palabras.

Riesgos	Expertos						
	e_1			…	e_m		
	P	I	D		P	D	I
$Riesgo_1$	X_{11P}	X_{11I}	X_{11D}	…	X_{1mP}	X_{1mI}	X_{1mD}
⋮	⋮	⋮	⋮	⋮			⋮
$Riesgo_n$	X_{n1P}	X_{n1I}	X_{n1D}	…	X_{nmP}	X_{nmI}	X_{nmD}

Paso 5. Siguiendo el modelo 2-tuplas de computación con palabras se agregan las evaluaciones de los expertos consolidando las mismas por cada elemento a evaluar.

Paso 6. Al finalizar se tiene que devolver los riesgos identificados según su influencia.

Proceso 4. Pre-procesamiento de datos registrados en el sistema de información

El proceso de pre-procesamiento se basa en la identificación de errores en la codificación de los datos que puedan generar dificultades en la detección de

situaciones anómalas. Por ejemplo: errores de edición, datos ausentes, entre otros. Se procede en este paso a ejecutar el siguiente Algoritmo 3.

Algoritmo 3. Algoritmo para la limpieza de datos

Paso 1. Identificación de los errores y definición de la taxonomía de errores para clasificar a los diferentes tipos de errores encontrados.

Paso 2. Se aplica la taxonomía para la clasificación de errores propuesta por [149-152]. Se aplica la siguiente lista de chequeo para la detección de los posibles errores en el caso de la gestión de proyectos.

Tabla 10. Lista de chequeo: muestra taxonomía y posibles errores en los datos.

Categoría	Posibles errores
Datos incompletos	Tareas que no tienen asignada una fecha de inicio. Tareas que no tienen asignada una fecha de fin. Valores de tiempo estimado ausentes. Valores de tiempo dedicado ausentes. Tareas que no están asignadas a ningún usuario.
Datos incorrectos	Tareas con fecha final menor que la fecha inicial. Tareas con valor de tiempo estimado igual a cero. Tareas cerradas con tiempo dedicado en cero. Inclusión de caracteres extraños en los nombres de las tareas para separar distintas partes: comillas, signos de comparación, paréntesis, corchetes, entre otros. Utilización de siglas: no conformidades (NC y NCF), casos de uso (CU), diseño de casos de prueba (DCP), interfaz de usuario (IU). Problemas de edición o tipografía. Errores ortográficos de acentuación.
Datos inconsistentes	Nombres de tareas muy generales que no permiten exactamente saber qué se va a realizar. Incorrecta clasificación de las tareas de Desarrollo-Producción, gestión o formación.

	Utilización de términos no estandarizados en la descripción de las tareas. Utilización para las tareas de implementación indistintamente: implementar, arreglar método, realizar cambios, validar interfaz. Utilización para las tareas de diseño y administración de base de datos indistintamente: migrar, migración, analizar e implementar, hacer script. Utilización para las tareas de diseño de interfaz de usuario: realizar diseño, diseñar.

Paso 3. *Medición del volumen de errores en los datos, basado en la métrica para la evaluación de la calidad de los datos [153].*

Paso 4. *Aplicación de técnicas de limpieza de datos tomando como base tres estrategias: la eliminación de los errores, la sustitución de cadenas para los problemas de estandarización y la imputación de datos para casos específicos de valores ausentes. Aplican las técnicas propuestas en [154].*

Paso 5. *Análisis de los resultados obtenidos: terminada la limpieza se analizan los resultados obtenidos, cuántos errores fueron erradicados y qué nivel de calidad del dato presentan las bases de datos una vez concluida la limpieza.*

Paso 6. *Propuestas de mejoras a las herramientas informáticas, en aras de mantener el nivel de calidad obtenido con la limpieza.*

Proceso 5. Aplicación y modelación de algoritmos de análisis de los datos

En este proceso se introduce un método combinado para la detección de datos anómalos apoyando los enfoques activosy reactivos. Estos enfoques se garantizan por la agilidad de la respuesta de los algoritmos propuestos que es además una de las variables analizadas en esta investigación.

Además, se aplican algoritmos que permiten la detección de datos anómalos basado en una estrategia de tratamiento independiente de los datos con múltiples algoritmos y luego la unión de los resultados. Este algoritmo efectúa varias iteraciones y en cada una de ellas aplica un algoritmo especializado en la detección de un tipo de

situación. Una vez ejecutados todos los algoritmos se combinan los resultados encontrados y se pasa al siguiente paso. Ver Algoritmo 4.

Algoritmo 4.Meta algoritmo basado en combinación de diferentes técnicas

1. *AlgoritmoBasadoCombinacionMetodos (D, A)*

   Entradas:

   D: representa el conjunto de datos a analizar.

   A: representa el conjunto de algoritmos, siendo A_i un algoritmo y se denota $A_i \in A$.

   A_{activo} : representa un método de aprendizaje activo con la intervención de expertos.

2. *Inicio*

3. *i =1*

4. *$D_1 = D$*

5. *Mientras queden algoritmos sin aplicar hacer, en caso contrario ir al paso 11.*

6. *Seleccionar el algoritmo A_i*

7. *Seleccionar el conjunto de datos a partir del conjunto original $D_i=D$*

8. *$P_i = A_i(D_i)$ // Detección de posibles datos anómalos*

9. *i ++*

10. *Regresar al paso 5*

11. *$O_i = A_{activo}(P_i) \forall i$ // Aplicación del aprendizaje activo en la verificación de los datos anómalos*

12. *$O = \cup O_i$ //combinación de los datos anómalos detectados en cada iteración*

13. *Devolver los datos anómalos contenidos en O y marcarlos para aprendizaje*

14. *Fin*

La complejidad de este algoritmo es *$max(O(A_i)) \forall i$, siendo A_i el i-ésimo algoritmo que incluye.* Se presentan a continuación algunas variantes de algoritmos a emplear en el meta-algoritmo propuesto, con el objetivo de encontrar aquellos que reporten los mejores resultados para cada problema específico [155-157].

Algoritmo de aprendizaje activo

Se ejecuta al finalizar la aplicación del resto de los algoritmos para la detección de datos anómalos. Recibe como entrada el listado de datos sospechosos de ser anómalos. En este momento los expertos en aseguramiento de ingresos de la organización, validan si los datos sospechosos son o no realmente anómalos y se marcan los datos analizados. Además, se aprenden lecciones a emplear en futuros procesos de análisis. La técnica fundamental a aplicar en este caso es el juicio de expertos.

Algoritmo basado en distancia

En este sentido se propone el Algoritmo 5 empleando la distancia de Mahalanobis y partir de considerar los elementos analizados en el capítulo 1 asociados a las comparaciones entre la distancia Euclideana y la distancia de Mahalanobis.

Algoritmo 5. Algoritmo empleado basado en distancia

1. *AlgoritmoDistancia (D)*

   Entradas:

   k: vecinos más cercanos

   D: conjunto de datos

   MaxDistancia(d,S) función de distancia máxima

   Vecindad(d, S, k): k elementos más cercanos a d; Vecinos(d): conjunto de los vecinos de d

   PrimerOutlier(S): retorna elementos ordenados descendentemente según la distancia a sus k vecinos más cercanos

   MaxUmbral(O, c) permite le trabajo con umbrales dinámicos refinando la búsqueda, por defecto es la función identidad (no modifica el umbral c)

2. *Inicio*

3. *c = 0 (umbral de corte)*

4. *O = {}*

5. *Para cada d en D*

6. *Vecinos(d) = {}*

7. *Para cada b en D tal que b ≠ d*

8. *Si | Vecinos(d) | < k ó Distancia(b,d)<MaxDistancia(d, Vecinos(d))*

9. *Entonces Vecinos(d) = Vecindad(d, Vecinos(d) U b, k)*

10. *Si | Vecinos(d) | > k y c > Distancia(b,d) Entonces Volver a línea 7*

11. *Fin del ciclo iniciado en línea 7*

12. *PrimerOutlier(OU b)*

13. *c= MaxUmbral(O, c)*

14. *Fin del ciclo iniciado en línea 5*

15. *Devolver los datos anómalos contenidos en O*

16. *Fin*

La complejidad de este algoritmo es $O(n^2)$, siendo $n = |D|$, la cantidad de registros.

Algoritmos híbridos que combinan agrupamientos con métodos basados en distancias

Se propone en este caso el uso de algoritmos híbridos aprovechando las ventajas de las estrategias basadas en distancia con las estrategias basadas en agrupamientos. Ver Algoritmo 6.

Algoritmo 6. Algoritmo empleado basado en agrupamientos y combinado con distancia

1. *AlgoritmoClustersDistancia (D)*

Entradas:

D: conjunto de datos; C cantidad de centros esperados; O conjunto de datos anómalos

Distancia(d,S): función de distancia de d al conjunto de puntos S

u: umbral de corte; n: cantidad de datos anómalos a retornar

PrimerOutlier(S): retorna los elementos de S ordenados descendentemente según la distancia

2. *Inicio*
3. *O = {}*
4. *C = {}*
5. *clusters =ClusterMethod(D, centers=C)*
6. *centros=clusters$centers //devuelve los centros de los agrupamientos encontrados*
7. *Para cada b en D*

   SiDistancia(b,centros)> uEntonces O = O U b
8. *Fin del ciclo iniciado en línea 7*
9. *O = PrimerOutlier(O)*
10. *Devolver los datos anómalos contenidos en O*
11. *Fin*

La complejidad de este algoritmo es $O(n^2)$, siendo $n = |D|$, la cantidad de registros.

En este caso se propone la experimentación con los siguientes algoritmos:

- *Algorithm kmeans_euclidean:* un algoritmo híbrido, agrupamiento Kmeans aplicado en el paso 2 combinado con distancia Euclidiana en los pasos 4 y 6.
- *Algorithm kmeans_norm_euclidean:* un algoritmo híbrido, agrupamiento Kmeans aplicado en el paso 2 combinado con distancia Euclidiana en los pasos 4 y 6. Pero considerando los datos normalizados.

Algoritmo híbrido kmeans_stats

Este es un algoritmo híbrido que combina técnicas de agrupamientos, con métodos basados en distancia y técnicas de reconocimiento de patrones. Ver Algoritmo 7.

Algoritmo 7. Algoritmo híbrido combina agrupamiento, distancia y heurísticas

1. *AlgoritmoClustersDistancia (D)*

   Entradas:

   D: conjunto de datos; C conjuntos de centros sembrado; O conjunto de datos anómalos

   Distancia(d,S): función de distancia de d al conjunto de puntos S

   *B_0 :*umbral *basado en las ecuaciones(1) o (2)*

   PrimerOutlier(S): retorna los elementos de S ordenados descendentemente según la distancia

2. *Inicio*

   3. *O = {}*
   4. *C = centros a sembrar //se toman en consideración información del problema en cuestión*
   5. *clusters =ClusterMethod(D, centers=C)*
   6. *centros=clusters$centers //devuelve los centros de los agrupamientos encontrados*
   7. *O=clusters$out_centers_$B_o$ //devuelve datos fuera del umbral B_0, Distancia(b,centros)> $B_o$respecto a //cada centro*
   8. *O = PrimerOutlier(O)*
   9. *Devolver los datos anómalos contenidos en O*

10. *Fin*

La complejidad de este algoritmo es $O(n^2)$, siendo $n = |D|$, la cantidad de registros. En particular para mejorar la eficiencia de este algoritmo,se usa como estrategia sembrar los centros inicialmente según los tipos de tareas. Se emplea además el

cálculo del umbral B_0 para disminuir la cantidad de comparaciones en el algoritmo, siguiendo las ecuaciones (1) ó (2) tomadas de [159].

$$B_0 = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} Distancia(d_i, d_j) \quad (1)$$

siendo d_i, d_j, datos del conjunto *D*, *m*= $|D|$

$$B_0 = \frac{1}{m} \sum_{i=1}^{m} \max_{j=1}^{m} (Distancia(d_i, d_j) \quad (2)$$

Algoritmo híbrido Combine_outlier

Algoritmo híbrido que combina técnicas de agrupamientos basadas en heurísticas.

Algoritmo 8. Algoritmo híbrido combine_outlier

1. AlgoritmoClustersDistancia (D)

Entradas:

D: conjunto de datos; Ggrupos previstos según escenario de aplicación, constituyen subespacios

Lista_atributos_G_i: lista de atributos ordenados por grado de dispersión para cada grupo G_i

ClusterJerarquicoMethod(D, List): se recomienda aplicar algoritmo agnes [160]

PrimerOutlier(O): retorna los elementos de O(anómalos) ordenados descendentemente

2. *Inicio*

3. *Para cada grupo G*

4. *Cálculo de estadígrafos básicos para cada atributo por grupo. // Incluye aplicación de técnicas de estadística descriptiva*

5. *Lista_atributos_G_i = Listado de atributos ordenado descendentemente según la dispersión*

6. *Fin del ciclo iniciado en la línea 3*

7. *Lista_atributos_G_i= A_{activo}(Lista_atributos_G_i) $\forall$ i // Aplicación del aprendizaje activo en la validación // y actualización del orden de atributos, según relevancia // en la introducción de anomalías, base dispersión*

8. *clusters = ClusterJerarquicoMethod(D, Lista_atributos_G_i) $\forall$ i, // Este clúster particiona el conjunto de // datospartiendode la organización por*

// subespacios y la información del paso 7

9. *O= A_{activo}(clusters) // Aplicación del aprendizaje activo para determinación de clústers con datos anómalos*
10. *O = PrimerOutlier(O)*
11. *Devolver los datos anómalos contenidos en O*
12. *Fin*

La complejidad de este algoritmo es $O(n^2)$, siendo $n = |D|$, la cantidad de registros. Para mejorar la eficiencia de este algoritmo se emplean conocimientos del escenario de aplicación para conocer de antemano la cantidad de grupos esperados, conocimiento quees empleado para la búsqueda por subespacios, en el caso específico de las organizaciones orientadas a proyectos se debe formar al menos un grupo por cada tipo de tarea. Este algoritmo va formando particiones sucesivas de forma independiente por cada uno de los subespacios, por medio de reordenamientos del conjunto de datos hasta concluir con un conjunto de particiones donde queden agrupados convenientemente los casos $b \in D$ con propiedades similares y grupos que contienen a los datos anómalos.

Otros algoritmos híbridos reportados con los que se recomienda experimentar

Algorithm Angle: basado en el enfoque de análisis espacial de los datos, en particular método basado en ángulos. Realiza la detección de valores anómalos basados en ángulos en un marco de datos especificado. Este algoritmo es recomendado para escenarios de alta dimensionalidad [117, 118].

Algorithm crossclustering: algoritmobasado en agrupamiento parcialcon la estimación automática del número de clústeres y la identificación de valores atípicos, combinado con algoritmos evolutivos, provee estimación automática de grupos y estimación automática de elementos anómalos. Calcula un algoritmo de agrupamiento parcial que combina los algoritmos de varianza mínima y de acoplamiento completo de Ward, proporcionando una estimación automática de un número adecuado de conglomerados y la identificación de elementos atípicos[113].

Algorithm kmodr: Algoritmo basado en el uso de métodos de agrupación simultánea.Es una implementación del algoritmo 'k-means-' propuesto por Chawla y Giovanni en 2013 con enfoque unificado para la agrupación y detección de valores atípicos. Útil para crear grupos potencialmente más apretados que los k-means estándar y encontrar simultáneamente datos anómalos a bajo costo en un espacio multidimensional[115].

Proceso 6. Evaluación de los resultados, estimación de impacto para la organización, análisis detallado

Este proceso se subdivide en los siguientes dos subprocesos: la evaluación de los algoritmos propuestos para la detección de los datos anómalos y la estimación del impacto en los ingresos de los datos anómalos y los riesgos detectados.

Evaluación de los algoritmos propuestos para la detección de los datos anómalos

En este proceso se evalúa la calidad de los algoritmos para el escenario específico identificando aquellos que tuvieron un mejor comportamiento en la detección de

anómalos y generalizarlos. Para la evaluación se propone el uso y adaptación de las métricas de precisión y sensibilidad [161, 162], considerando que los datos anómalos por definición son raros y excepcionales, que las tradicionales empleadas para evaluar la calidad de los métodos de clasificación, no tienen un buen comportamiento en estos casos.

Además, es importante conocer que en la propuesta realizada se emplea un algoritmo de aprendizaje activo y por esta razón los falsos positivos son menos relevantes porque generalmente son eliminados de la respuesta final por los expertos.A continuación se explicanlas métricas propuestas usando la siguiente notación:sea *A* un algoritmo de detección de datos anómalos que devuelve una lista ordenada de posibles datos anómalos y ρelumbral de recuperación de datos anómalos sobre esa lista. Se denota al conjunto de datos anómalos recuperados como $S(\rho)$, mientras que *T* representa el conjunto de datos anómalos verdaderos.

La precisión es definida como el porcentaje de datos anómalos reportados, ecuación (3).Mientras que la sensibilidad (*recall)* es definida como el porcentaje de los verdaderos datos anómalos, los cuales fueron reportados como datos anómalos en el umbral ρ, ver ecuación(4).

$$\mathrm{Pr}esición(\rho) = 100\frac{|\mathrm{S}(\rho) \cap \mathrm{T}|}{\mathrm{S}(\rho)} \qquad (3) \qquad Sensibilidad(\rho) = 100\frac{|\mathrm{S}(\rho) \cap \mathrm{T}|}{T} \qquad (4)$$

Como parte de este paso se propone además la combinación de las medidas de Sensibilidad y Precisión.Se denomina Eficacia_Detección a esta medida y se calcula usando un operador OWA [162] como un caso particular de dos valores tal que $a_i \in$ *{sensibilidad, precisión}.* Este método unifica los criterios clásicos de decisión con incertidumbre en un solo modelo. Es decir, esta unificación abarca los criterios optimista, el pesimista, el de Laplace y el de Hurwicz en una sola expresión (163, 167). Este operador puede ser definido de la forma siguiente:

Definición 1: un operador OWA es una función $F : \Re^n \to \Re$ con un vector asociado *W* de dimensión*n*tal que $w_i \in [0,1], \sum_{i=1}^{n} w_i = 1$ y que cumple la ecuación (5).

$F(a_1, a_2, \ldots, a_n) = \sum_{j=1}^{n} w_j b_j \quad F(a_1, a_2, \ldots a_n) = \sum_{i=1}^{n} w_i b_j$ (5)Donde bj es el j-ésimo más grande de los $a_i$que se deseanagregar

Estimación del impacto en los ingresos de la organización de los datos anómalos detectados

Al llegar a este paso se cuenta con tres grupos de eventos orientados a la recuperación de ingresos:

- Riesgos que son evitados o mitigados, para disminuir la fuga de ingresos.
- Medidas tomadas para la mitigación o eliminación de riesgos con un costo de implementación que debe considerarse también en el aseguramiento de ingresos.

- Listado de situaciones anómalas detectadas que reflejan situaciones de fugas de ingresos ya sea por fraude, errores de operación u otras causas.

El objetivo de este paso es estimar el impacto económico de estas situaciones para evaluar la efectividad de los procesos de aseguramiento de ingresos. Se propone a continuación un conjunto de técnicas y luego en la Tabla 11 se sugiere en qué situación deben ser empleadas.

Técnica de estimación por tres valores

Se basa en lograr estimar tres valores: monto más probable (M) a ser recuperado basado en una evaluación realista del experto, monto a recuperar basado en un enfoque optimista (O) tomando como base el mejor escenario posible y el monto a recuperar pesimista (P) basado en el análisis del peor escenario para la recuperación de los ingresos. Luego se procede a calcular el valor estimado usando alguna de las ecuacionessiguientes [3].

$$ce_1 = \frac{O + 4M + P}{6} \quad (6) \qquad ce_2 = \frac{O + M + P}{3} \quad (7)$$

Técnica de estimación ascendente

En este caso se realiza una estimación de lo recuperado por cada una de las situaciones anómalas detectadas y se procede a consolidar este resultado sumando los montos en unidades monetarias asociados a cada actividad. En caso de que sea difícil la estimación del impacto de una actividad se procede a descomponer la

misma en componentes de nivel inferior para un análisis más detalladoy luego se consolidan las estimaciones respetando la jerarquía construida [3].

Técnica basada en el análisis de redes

Esta técnica se basa en la construcción de una red de forma tal que cada nodo representa un estado posible de la organización ante diferentes situaciones. Las aristas representan las posibles decisiones y cada arista está etiquetada por un vector con las siguientes características [162-165].

- La primera componente representa la probabilidad de tomar la decisión o de que ocurra un riesgo.
- La segunda componente representa alguno de los siguientes elementos: costo de tomar la decisión, impacto económico positivo en caso de ocurrir una oportunidad, impacto económico negativo en caso de ocurrir una amenaza.

Con esta estructura se puede aplicar un conjunto de técnicas clásicas asociadas al trabajo con redes que permitiría estimar el impacto económico de diferentes escenarios del aseguramiento de ingresos entre las que se encuentran:

- Aplicando análisis de redes bayesianas [166-168] se puede estimar la probabilidad de que ocurran diferentes situaciones y estimar el impacto económico de cada camino posible.

- Aplicando algoritmo de flujo máximo[169] se puede conocer el impacto económico de las decisiones caminos a seguir por la organización.

- Aplicando algoritmo de Dijkstra [169] se puede conocer el camino con costo mínimo de la organización que podría ser útil para representar el camino del aseguramiento de ingresos que garantice los ingresos pero que tenga el menor costo de implementación posible.

- Aplicando el algoritmo de Floyd [170] se pueden determinar todos los caminos con costo mínimo de la organización. Éste es similar a la aplicación de Dijkstra, el cual podría ser útil para asegurar ingresos con el menor costo de implementación posible.

Se pueden usar otros algoritmos para el trabajo sobre redes y se considera que esta línea de investigación, se debe continuar en futuros trabajos.

Recomendación asociada a las técnicas a emplear

Se presenta a continuación una tabla de decisión que sugiere qué técnicas se pueden emplear para la evaluación de cada una de las situaciones con datos anómalos identificados en los procesos anteriores, ver Tabla 11.

Tabla 11.Tabla de decisión para la selección de las técnicas de estimación del impacto.

Situación	Técnica a emplear
Conjunto de riesgos que de ocurrir tienen un impacto en los ingresos de la organización.	Estimación por tres valores. Técnica basada en el análisis de redes.

Conjunto de actividades orientadas a la gestión de riesgos y al tratamiento proactivo del aseguramiento de ingresos.	Estimación por tres valores. Estimación ascendente. Técnica basada en el análisis de redes.
Situaciones anómalas detectadas en los datos, que muestran fugas de ingresos.	Estimación ascendente. Estimación por tres valores. Técnica basada en el análisis de redes.

Proceso 7. Toma de decisiones e implantación

En este proceso se propone el uso de sistemas de información combinada con el juicio de expertos para la toma de decisiones. Para la toma de decisiones se deben seguir las siguientes recomendaciones:

- Uso de sistemas de información que permitan la gestión por cortes y el uso de indicadores objetivos que cubran las áreas de conocimiento.
- Involucrar a los miembros en la búsqueda de las soluciones.
- Priorización en la toma de decisiones con centro en las actividades de la cadena de valor con mayor impacto en los ingresos y las utilidades.

Como sistema de información se recomienda el uso de la plataforma GESPRO por la versatilidad de la misma y la gran cantidad de funcionalidades para el aseguramiento de ingresos [153, 171-174] entre las que se encuentran:

- Módulo para el análisis de datos y el aseguramiento de ingresos, que integra bibliotecas en R para la detección de datos anómalos.

- Módulo para la gestión de riesgos, aplicable para el análisis proactivo.
- Cuadro de mando con indicadores y alertas tempranas, orientado a la detección de insuficiencias en la planificación y la ejecución de proyectos.
- Gestión del alcance y de la calidad respecto al cubrimiento de los requisitos en el cronograma y el control de la calidad.
- Gestión de los costos de los proyectos y predicción de costos de proyectos en función del comportamiento de los datos.

En este paso también pueden ser empleados sistemas de recomendaciones, se recomienda que se trabaje esta línea en investigaciones futuras.

Se propone el siguiente algoritmo para la toma de decisiones.

Algoritmo 9. Algoritmo para la toma de decisiones por cortes

1. *AlgoritmoControlTomaDecisiones()*

   Entradas:

   O conjunto de datos anómalos, O_i situaciones anómalas detectadas en el i-ésimo corte

   D conjunto de decisiones, D_i decisiones del i-ésimo corte

2. *Inicio*
3. *En cada corte i de evaluación*
4. *Análisis de resultados de las situaciones anómalas O_{i-1}identificadas en el corte anterior y su evolución*
5. *Análisis de resultados de las nuevas situaciones anómalas $O_i$identificadas en el corte actual*
6. *Si no hay situaciones anómalas o si las que existen están siendo tratadas satisfactoriamente, finalizar chequeo e ir al paso 10*
7. *Bajar en la cascada, analizar las causas de las dificultades por cada proyecto o área del conocimiento*

8. *Identificadas las causas priorización para el tratamiento de las situaciones anómalas según su impacto económico actual en el corte i*
9. *Proceder a la toma de decisiones considerando la prioridad,su evolución desde el corte i-1 y el pronóstico del corte i+1 ver Figura 5*
10. *Documentar los acuerdos y lecciones aprendidas* $D_i = D_{i-1} \cup D_i$
11. *Fin*

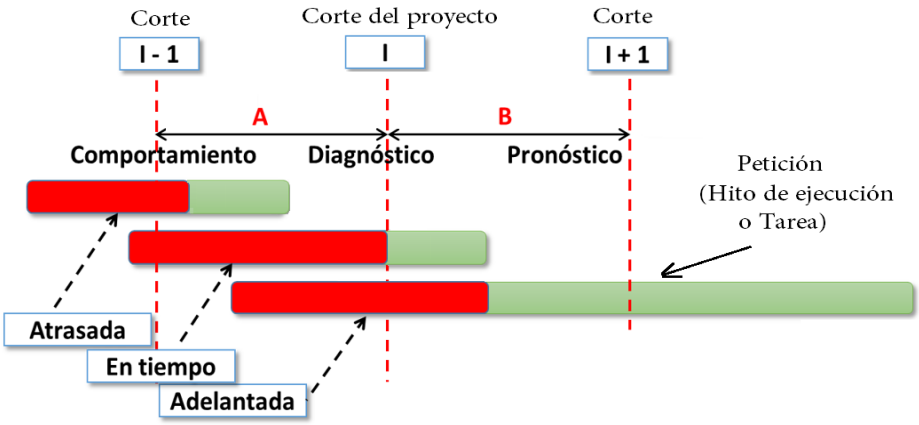


Figura 5.Vista de análisis de un proyecto y sus tareas por cortes, análisis del impacto en los ingresos.

Conclusiones del capítulo

- El modelo propuesto combina técnicas de gestión de proyectos, técnicas de minería de datos anómalos y computación con palabras siguiendo los enfoques: proactivo, reactivo y activo en el aseguramiento de ingresos y requiere de la presencia de múltiples expertos para su aplicación.
- El modelo propuesto, se encuentra integrado a la plataforma GESPRO y se propone el uso de esta plataforma para su aplicación en la toma de decisiones por la versatilidad y las funcionalidades para el aseguramiento de ingresos.

- En el proceso 7 para la toma de decisiones pueden ser empleados sistemas de recomendaciones entre otras técnicas de la computación emergente. Se recomienda que se trabaje esta línea de investigación en investigaciones futuras.

3. CAPÍTULO: EXPERIMENTACIÓN Y VALIDACIÓN DE LOS RESULTADOS

Para el diseño de experimentos se emplean técnicas de triangulación metodológica combinando técnicas para la triangulación de datos, la triangulación de expertos y la triangulación de métodos. Se siguieron los siguientes pasos:

1. Descripción de las bases de datos para la experimentación permitiendo la aplicación de técnicas de triangulación de datos y validación cruzada.
2. Validación de la variable eficacia. A partir de la aplicación de técnicas de triangulación metodológica de métodos para ladeterminación de la mejor configuración de cada uno de los algoritmos propuestos en el capítulo 2 para la detección de situaciones anómalas. A partir de experimentación sobre las bases de datos seleccionadas.
3. Validación de las variables dependientes. A partir de la aplicación de técnicas de triangulación metodológica de métodos para la determinación de los algoritmos con mejores resultados en la detección de situaciones anómalas en las bases de datos seleccionadas.
4. Validación de la variable dependiente eficacia. A partir de aplicación de técnicas de triangulación metodológica de métodos y de expertos, en la comparación del enfoque proactivo para la gestión de riesgos del modelo propuesto con la técnica propuesta por el PMBOK.

5. Validación de la variable independiente. A partir de aplicación de técnicas de triangulación metodológica de expertos, para la validación del modelo por expertos y su aplicación en un caso de estudio.
6. El objetivo de estos experimentos consiste en determinar cuál es la mejor configuración de cada algoritmo para poder compararlos luego entre ellos y finalmente determinar los de mejores resultados.

Para la comparación de los resultados de los algoritmos con las 5 bases de datos y sus particiones se comparan las poblaciones formadas por los resultados de los algoritmos usando test no paramétrico de Wilcoxon para dos muestras relacionadas con 95% de intervalo de confianza. Se emplea Wilcoxon porque el test de Shapiro Wills de normalidad, dio que las muestran no cumplen con la distribución normal con un p-value= 0.00032.

En el análisis de los algoritmos, se identifican los grupos de forma tal que los algoritmos en un mismo grupo no tienen diferencias significativas entre ellos. Además, se organizan los grupos, respecto a la calidad de los resultados obtenidos de forma tal que *"grupo a"*>*"grupo b"* > *"grupo c"* > *"grupo d".* O sea, los que están en el *"grupo a"* reportan los mejores resultados respecto a la variable analizada.

Descripción de las bases de datos

Se comparan los algoritmos a partir de analizar su desempeño con las bases de datos: *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix"*del repositorio de bases de datos para investigaciones, del Laboratorio de Investigaciones en Gestión de

Proyectos [152, 178].Cada una de estas cinco bases de datos está formada por 20 particiones aplicando técnicas de validación cruzada.

Tabla 12. Descripción de las bases de datos empleadas en la experimentación.

Base de datos	**Cantidad de registros**	**Cantidad de atributos**	**Cantidad de particiones**	**Atributos modificados**	**Porciento de datos anómalos**
“alone_rate”	9470	23	20	"rate_rrhh"	5% de los registros 473 modificados en cada partición
“mul_plan”	9470	23	20	serv_plan_quantity, "rrhh_plan_quantity" "eqp_plan_quantity" "inf_plan_quantity" "mat_plan_quantity"	5% de los registros 473 modificados en cada partición
“mul_rate”	9470	23	20	rate_equipment", "rate_rrhh", "rate_service" "rate_material"	5% de los registros 473 modificados en cada partición
“mul_mix”	9470	23	20	"rate_rrhh", "rrhh_plan_quantity" "rate_material" "mat_plan_quantity" "rrhh_plan_quantity", "rrhh_real_quantity",	5% de los registros 473 modificados en cada partición
“col_mix”	9470 88 proyectos	23	20	"rate_rrhh" "rrhh_plan_quantity" : "rate_material",	95% de los registros en cada proyecto seleccionado.

	Can	Par	Can
1	128	11	45
2	35	12	45
3	389	13	44

					4	58	14	3
					5	25	15	10
					6	52	16	801
					7	801	17	2
					8	364	18	37
					9	102	19	50
					10	12	20	185

Par: significa partición, Can: cantidad de registros modificados, en total se generaron 100 bases de datos diferentes.

Validación de variable eficacia, determinación de la mejor configuración de los algoritmos

*Resultados cuasi-experimento 1 post prueba, configuración del algoritmo***Angle**

Se aplica el cuasi-experimento 1 con post prueba en las bases de datos *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix"*. Comparando diferentes configuraciones de un algoritmo basado en ángulos [117, 118] "*Algorithm Angle"*, respecto a la variable eficacia, en la detección de datos anómalos en bases de datos orientadas a proyectos.

Se aplica el algoritmo *Angle* con diferentes valores de $k \in \{3, 5, 7, 9\}$ y valor de percentil para determinar umbral de las distancias entre los ángulos $\rho \in \{0.92, 0.95\}$. En elAnexo 2Tabla 21 se muestran los resultados una vez aplicado el test de Wilcoxon. Se encuentran diferencias significativas en la mayoría de las bases de datos, pero la configuración con resultados más estables fue An*gle_5_0.95.*

Resultados cuasi-experimento 2 post prueba, configuración del algoritmocrossclustering

Se aplica el cuasi-experimento 2 post prueba en las bases de datos *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix"*. Comparando diferentes configuraciones del algoritmo "*Algorithm crossclustering"* [113], respecto a la variable eficacia.

Se aplica el algoritmo *Crossclustering* con diferentes valores de *k_min* $\in$ *{3, 4, 5} y k_max* $\in$ *{5, 7, 9}.* En elAnexo 2Tabla 27 se muestran los resultados una vez aplicado el test de Wilcoxon. Se encuentran diferencias significativas en las diferentes configuraciones de los algoritmos. La versión más estable fue *Crossclustering_5_3.*

Resultados cuasi-experimento 3 post prueba, configuración del algoritmo basado en Distance_mahalanobis

Se aplica el cuasi-experimento 3 post prueba en las bases de datos *"alone_rate, col_mix, mul_rate, mul_mix"*. Comparando diferentes configuraciones de un algoritmo basado en distancias "*Algorithm Distance_mahalanobis"*, respecto a la variable eficacia, en la detección de datos anómalos en bases de datos orientadas a proyectos.

Se aplica el algoritmo *Distance_mahalanobis* con diferentes valores de *k* $\in$ *{3, 5, 7, 9}* y valor de percentil para determinar umbral de las distancias $\rho \in$ *{0.92, 0.95}.* En elAnexo 2Tabla 33 se muestran los resultados una vez aplicado el test de Wilcoxon. Se encuentran diferencias significativas en algunas configuraciones de los algoritmos.La versión más estable fue *Distance_mahalanobis_3_0.*

Resultados cuasi-experimento 4 post prueba, configuración del algoritmobasado en kmeans_euclidean

Se aplica el cuasi-experimento 4 post prueba en las bases de datos *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix"*. Comparando diferentes configuraciones de un algoritmo híbrido "*Algorithm kmeans_euclidean"*, respecto a la variable eficacia, en la detección de datos anómalos en bases de datos orientadas a proyectos.

Se aplica el algoritmo *kmeans_euclidean* con diferentes valores de $k \in \{3, 5, 7, 9\}$ y valor de percentil para determinar umbral de las distancias $\rho \in \{0.92, 0.95\}$. En elAnexo 2Tabla 38 se muestran los resultados una vez aplicado el test de Wilcoxon. Se encuentran diferencias significativas en algunas configuraciones de los algoritmos. La versión más estable fue *Kmeans_Euclidean_9.*

Resultados cuasi-experimento 5 post prueba, configuración del algoritmobasado en kmeans_norm_euclidean

Se aplica el cuasi-experimento 5 post Prueba en las bases de datos *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix"*. Comparando diferentes configuraciones de un algoritmo híbrido "*Algorithm kmeans_norm_euclidean"*, respecto a la variable eficacia, en la detección de datos anómalos en bases de datos orientadas a proyectos.

Se aplica el algoritmo *kmeans_norm_euclidean* con diferentes valores de $k \in \{3, 5, 7, 9\}$ y valor de percentil para las distancias $\rho \in \{0.92, 0.95\}$. En elAnexo 2Tabla 44 se muestran los resultados una vez aplicado el test de Wilcoxon. Se encuentran diferencias significativas en algunas configuraciones de los algoritmos. La versión más estable fue *kmeans_norm_euclidean_9_0.92*.

Resultados cuasi-experimento 6 post prueba, configuración del algoritmobasado en kmeans_stats

Se aplica el cuasi-experimento 6 post prueba en las bases de datos *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix"*. Comparando diferentes configuraciones del

algoritmo “*Algorithm kmeans_stats”*, respecto a la variable eficacia, en la detección de datos anómalos en bases de datos orientadas a proyectos.

Se prueba con diferentes valores de $k \in \{3, 5, 7, 9\}$. En elAnexo 2Tabla 50 se muestran los resultados una vez aplicado el test de Wilcoxon.No se encuentran diferencias significativas, aunque la configuración con mejores resultados fue kmeans_stats_3 que será empleada en la comparación con otros algoritmos.

Resultados cuasi-experimento 7 post prueba, configuración del algoritmobasado en agrupamientos kmodr

Se aplica el cuasi-experimento 7 post prueba en las bases de datos *“alone_rate, col_mix, mul_plan, mul_rate, mul_mix”*. Comparando diferentes configuraciones del algoritmo *kmod* [115] “*Algorithm kmodr”*, respecto a la variable eficacia.

Se aplica el algoritmo *kmodr* con diferentes valores de $k \in \{3, 5, 7, 9\}$. En elAnexo 2Tabla 51 se muestran los resultados una vez aplicado el test de Wilcoxon. No se encuentran diferencias significativas, aunque la configuración con mejores resultados fue *kmodr_3* que será empleada en la comparación con otros algoritmos.

Resultados cuasi-experimento 8 post prueba, configuración del algoritmobasado en agrupamientos Combine_outlier

Se aplica el cuasi-experimento 8 post prueba en las bases de datos *“alone_rate, col_mix, mul_plan, mul_rate, mul_mix”*. Comparando diferentes configuraciones del algoritmo “*Algorithm Combine_outlier*”, respecto a la variable eficacia.

Se aplica el algoritmo *Combine_outlier* con diferentes valores de percentil para determinar umbral de las distancias $\rho \in$ *{3, 5, 7, 9}.* En elAnexo 2Tabla 52 se muestran los resultados una vez aplicado el test de Wilcoxon. No se encuentran diferencias significativas, aunque la configuración más estable fue *Combine_outlier_0.92*.

Validación de las variables dependientes comparación de los algoritmos

Se comparan las mejores combinaciones de los algoritmos que fueron detectadas en los cuasi-experimentos del 1 al 8. Se aplica el cuasi-experimento 9 con post prueba en las bases de datos *"alone_rate, col_mix, mul_plan, mul_rate, mul_mix".* Comparando los diferentes algoritmos, respecto a las variables eficacia y eficiencia, en la detección de datos anómalos puntuales en un atributo aislado.

Resultados de la eficacia: variable porciento de datos anómalos detectados correctamente

En la Tabla 13 se muestran los resultados una vez aplicado el test de Wilcoxon. En esencia se encuentran diferencias significativas entre los algoritmos.

Tabla 13. Comparación de múltiples algoritmos respecto a la eficacia, aplicando Wilcoxon.

Grupo	col_mix	alone_	mult_mix_	mult_plan_	mult_rate_
a	combine_outlier_0.92 kmeans_stats_3_0	combine_outlier_0.92 kmeans_stats_3	combine_outlier_0.92 kmeans_stats_3_0	combine_outlier_0.92 kmeans_stats_3	combine_outlier_0.92 kmeans_stats_3
b	distance_mahalanobis_3_0.92	distance_mahalanobis_3_0.92	distance_mahalanobis_3_0.92 angle_5_0.95	kmodr_3 angle_5_0.95	kmodr_3_0

c	Angle_5_0.95 kmodr_3_0	Kmodr_3_0 Angle_5_0.95	kmodr_3 kmeans_norm_euclidean_9_0.92 crossclustering_5_3	crossclustering_5_3_0	angle_5_0.95 distance_mahalanobis_3_0.92
d	Crossclustering_5_3 kmeans_norm_euclidean_9_0.92 kmeans_euclidean_9_0.92	Crossclustering_5_3 Kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92	crossclustering_5_3
e		kmeans_euclidean_9_0.92		kmeans_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92
f					kmeans_euclidean_9_0.92

En lasFigura 6, Figura 7 y la Figura 8 se aprecia que los algoritmos *kmeans_stats* y *combine_outlier* tienen resultados muy similares en todas las bases de datos excepto en la de datos anómalos colectivos (*col_mix*), donde el algoritmo *combine_outlier* es ligeramente superior.El de peor resultado fue *kmeans_euclidean_9_0.92.* Los datos detallados de las pruebas se muestran en elAnexo 3.

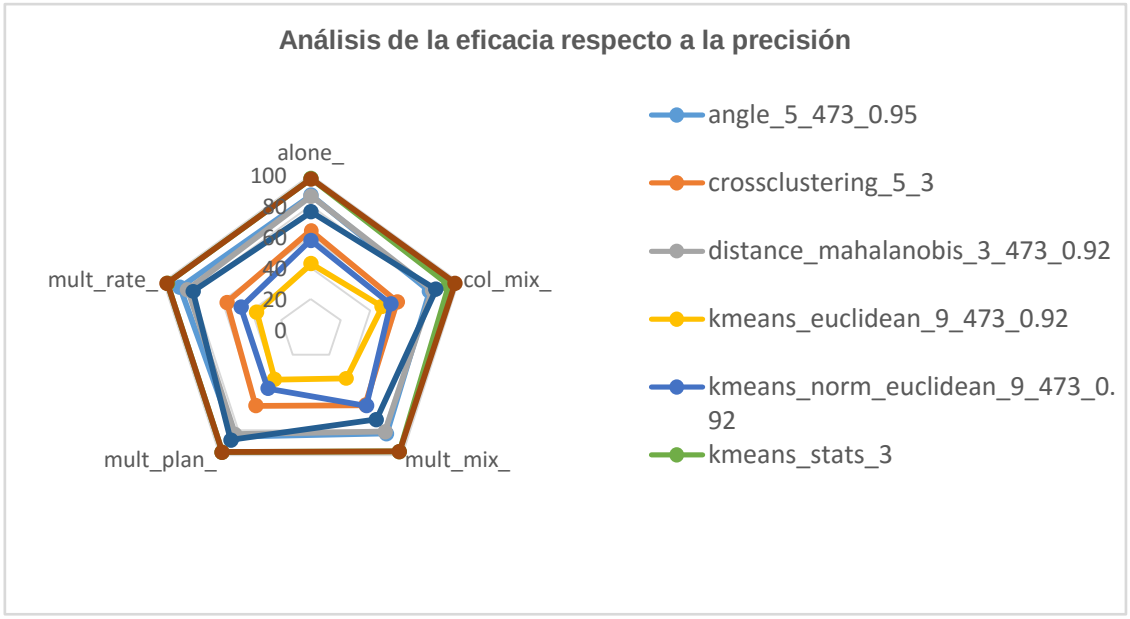


Figura 6.Eficacia respecto a la precisión.A mayor área, mejor es el resultado.

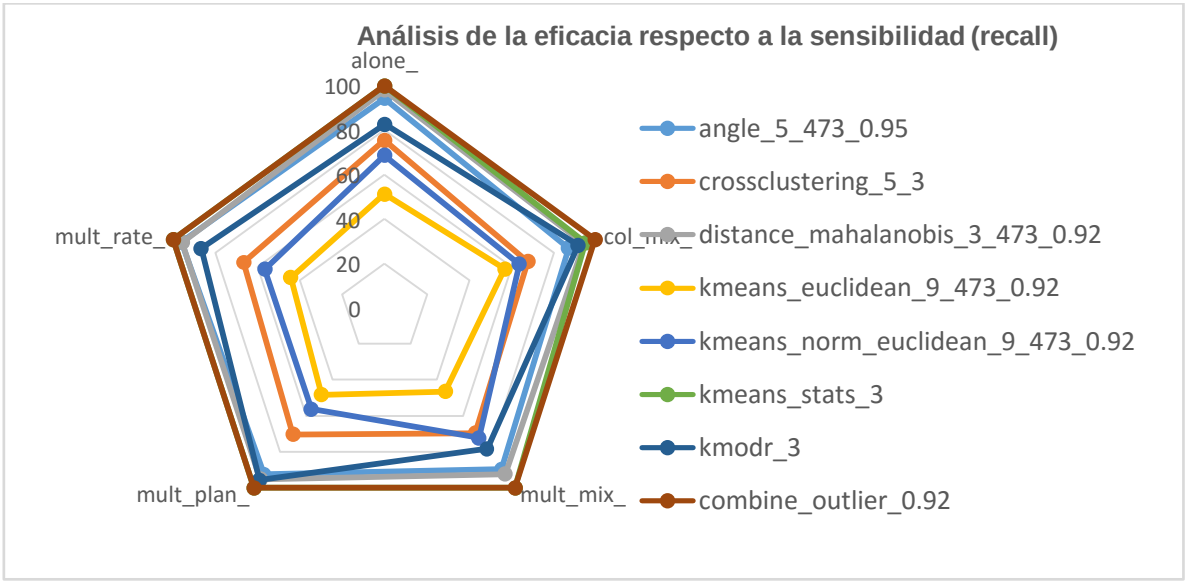


Figura 7.Eficacia respecto a la sensibilidad.A mayor área, mejor es el resultado.

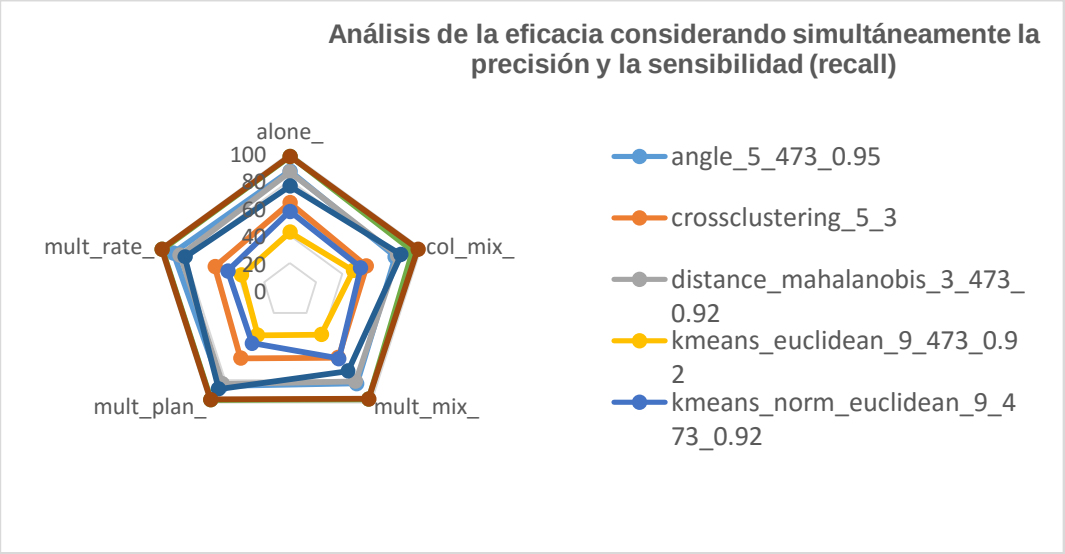


Figura 8.Eficacia considerando la precisión y la sensibilidad simultáneamente.

A mayor área, mejor es el resultado.

Resultados de la eficiencia: variable tiempo de los algoritmos

En la Tabla 14 se muestran los resultados una vez aplicado el test de Wilcoxon. En esencia se encuentran diferencias significativas entre los algoritmos. La Figura 9 muestra la eficiencia de los diferentes algoritmos respecto a las bases de datos empleadas. Los algoritmos con mejores resultados fueron *distance_mahalanobis_3_473_0.92y kmeans_stats_3.* Mientras, que los peores resultados los presentan*angle_5_473_0.95 y crossclustering_5_3.*Los datos detallados de las pruebas se muestran en elAnexo 4.

Tabla 14. Comparación de los algoritmos respecto a la eficiencia.

Grupo	col_mix	alone_	mult_mix_	mult_plan_	mult_rate_
a	distance_mahalanobis_3_0.9	kmeans_stats_3 distance_mahalanobis_3_0.92	kmeans_stats_3	kmeans_stats_3	distance_mahalanobis_3_0.92

b	kmeans_stats_3 kmeans_euclidean_9_0.92	combine_outlier_0.92	distance_mahalanobis_3_0.92 kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.92	kmeans_stats_3
c	combine_outlier_0.92	kmeans_euclidean_9_0.92	combine_outlier_0_0.92	combine_outlier_0.92	kmeans_euclidean_9_0.92
d	kmodr_3_0	kmodr_3_0	kmodr_3_0	kmodr_3_0	combine_outlier_0.92
e	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92	kmodr_3
f	crossclustering_5_3	crossclustering_5_3	crossclustering_5_3	crossclustering_5_3	kmeans_norm_euclidean_9_0.92
g	angle_5_0.95	angle_5_0.95	angle_5_0.95	angle_5_0.95	crossclustering_5_3
					angle_5_0.95

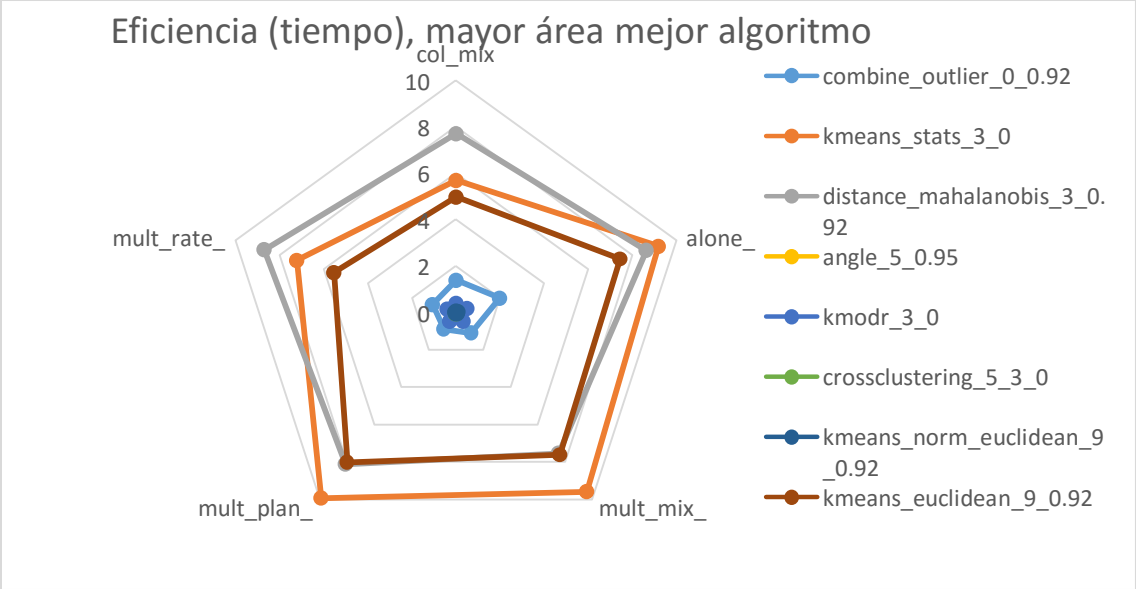


Figura 9. Estabilidad en la eficiencia de los algoritmos. En este caso a mayor área mejores resultados.

Validación de variable dependiente, comparación del modelo con la técnica del PMBOK

Se aplica el cuasi-experimento 10Post prueba:en la aplicación de técnicas de análisis proactivo para la evaluación cualitativa de riesgos causantes de pérdidas de

ingresos. Se compara la técnica tradicional propuesta en el PMBOK con la técnica propuesta basada en el modelo 2-tuplas de computación con palabras. En este caso para la validación se toman 14 proyectos de software que se describen brevemente en el Anexo 5. En la evaluación de los riesgos participaron 6 expertos, cuyos datos se reflejan en la Tabla 15.

Tabla 15. Caracterización de los expertos encuestados para la evaluación de los riesgos.

Total de expertos	Cantidad Doctores	Cantidad de Másters	Promedio de años dedicados	Desviación Estándar	Mínimo cantidad de años dedicados	Máxima cantidad de años dedicados
6	4	2	18,8	9,8	14	37

Se tuvieron en cuenta 18 de los riesgos más comunes en este escenario, que cubren todas las áreas de conocimiento de la gestión de proyectos [8, 9], ver Tabla 16.

Tabla 16. Relación de áreas de conocimiento y riesgos identificados para la validación.

Riesgos	Áreas de conocimiento
Pérdida de recursos humanos	Recursos humanos
Bajo de nivel de formación de recursos humanos	Recursos humanos
Pocos incentivos al equipo, baja producción	Recursos humanos
Mala conformación de equipos	Recursos humanos
Cliente desinteresado, que no participa en los encuentros	Interesados y Comunicaciones
Tardía entrega de información por cliente	Interesados y Alcance
Atrasos en entrega de los proveedores	Interesados y Adquisiciones
Aumento de los precios de los recursos	Adquisiciones y Costos
Dificultades energéticas, afectan la producción	Adquisiciones
Dificultades con el transporte, afecta plan	Adquisiciones
Fenómenos atmosféricos afectan la productividad	Riesgos

Rotura de equipos y lento mantenimiento	Riesgos
Trámites engorrosos para la comercialización	Interesados
Dificultades con la elicitación de requisitos	Alcance y Calidad
Dificultades con la definición de la arquitectura	Alcance y Calidad
Falta de liderazgo en los jefes de proyectos	Integración
Elevada presión externa, provoca errores de planificación y ejecución	Tiempo, Integración, Calidad
Bajos niveles de reutilización	Tiempo y Calidad

Todos los proyectos están terminados y se conoce cómo se comportaron los riesgos [176-178], esta información se considera como la respuesta deseada que deben aportar los métodos de evaluación cualitativa que se comparan en el trabajo.

Para la experimentación se tomaron un grupo de expertos en gestión de proyectos que no participaron en los proyectos seleccionados y se les dio suficiente información que caracteriza a cada proyecto. Luego se procedió a aplicar los siguientes pasos:

Paso 1. Los expertos sentados en una mesa de trabajo, evaluaron por consenso cada riesgo en cada proyecto empleando el método propuesto por el PMBOK [3], (método Riesgos-PMBOK).

Paso 2. Los expertos de forma independiente, evaluaron según su criterio cada riesgo en cada proyecto empleando el método propuesto en este trabajo. Se emplearon técnicas de computación con palabras, usando la misma variable representada en la Figura 4, (método Riesgos-CWW).

Paso 3. Se calculó el error cuadrático medio de las dos variantes de evaluación basado en la ecuación siguiente:

$$E_p = \frac{1}{18}\sum_{i=1}^{18}(D_i - Y_i)^2 \quad (8)$$

Siendo D_i lo que realmente ocurrió en los riesgos, la salida deseada por los sistemas.

Siendo Y_i la salida real de los diferentes métodos de evaluación ante cada riesgo.

E_p error cuadrático medio cometido en el proyecto $p \in [1..14]$.

Finalmente se compararon ambos errores cuadráticos medios de cada uno de los proyectos, ver laFigura 10y la tabla con los resultados en la Tabla 17.

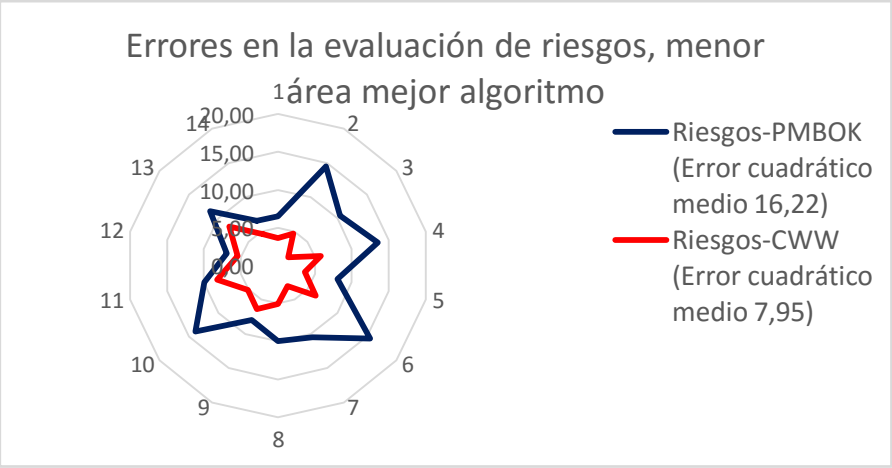


Figura 10. Gráfico radial que representa el error cuadrático medio en la evaluación de riesgos.

Resultados de evaluación de error cuadrático medio en la comparación de los métodos de evaluación cualitativa de riesgos.

Tabla 17. Valor del error cuadrático medio en la evaluación de los proyectos.

Proyectos	**Riesgos-PMBOK (Error cuadrático medio 16,22)**	**Riesgos-CWW (Error cuadrático medio 7,95)**

1	6,50	3,63
2	14,50	4,65
3	10,50	1,69
4	13,50	5,80
5	8,00	3,63
6	15,50	6,31
7	10,50	2,94
8	10,00	5,02
9	8,00	6,34
10	14,00	5,10
11	10,00	8,21
12	7,00	5,46
13	11,50	8,21
14	6,50	4,56

En el análisis de los resultados se identifica que en el algoritmo Riesgos-PMBOK el error cuadrático medio fue 16.22 mientras que en el algoritmo Riesgos-CWW fue de 7.95. Como muestra la Figura 10, los mejores resultados se obtienen con el método de evaluación Riesgos-CWW. Además, dicho método basado en computación con palabras también resultó más intuitivo y permite una mejor interpretación de los resultados al usar términos lingüísticos en lugar de valores numéricos.

Validación de variable independiente y aplicación en un caso de estudio

Se aplica el cuasi-experimento 11 post prueba: observación de resultados de la aplicación del modelo en casos de estudio. El modelo propuesto se implementó como parte del módulo de análisis de datos AnalysisPRO de la Suite de Gestión de

Proyectos Xedro-GESPRO [165-166, 176]. Este módulo integrado a la plataforma GESPRO es utilizado en la actualidad por la “Red de Centros Productivos” de la UCI, por la empresa XETID y por la empresa Copextel TecnoStart. Además, este módulo se emplea en el programa de posgradoMaestría en Gestión de ProyectosInformáticos que se imparte en la Universidad de las Ciencias Informáticas en los cursos de: “Gestión de organizaciones orientadas a proyectos” y “Herramientas para la toma de decisiones”.

Para el caso específico de la UCI, con la aplicación del modelo en cuestión se han beneficiado un total de 14 centros de desarrollo de tecnologías de la información, en los cuales se han gestionado más de 200 proyectos y donde convergen más de 500 usuarios de la herramienta GESPRO.

Además, se implementó el modelo en la compañía ecuatoriana QuitusServices [179], dedicada a la prestación de servicios de tecnologías de la información y las comunicaciones. Esta empresa presta servicios de consultorías para la informatización de entidades, vende productos de software, brinda servicios de mantenimiento y además gestiona ventas de productos de hardware. La Tabla 18 muestra el resumen de los montos recuperados por la compañía. Se presenta un análisis de la recuperación de ingresos en un mes, a partir del análisis de comportamiento de ingresos, costos y fugas de ingresos de la compañía durante los seis meses de aplicación del modelo.

Se emplearon para el análisis las mediciones de los costos e ingresos mínimos, promedios y máximos durante los meses de aplicación del modelo. Se calculó la media de números borrosos triangulares y se estimaron las fugas de ingresos por cada una de las actividades principales de la compañía. Luego se agregaron los números borrosos triangulares finales usando la técnica de estimación por tres valores.

Tabla 18.Resumen de análisis de ingresos recuperados a partir de aplicar el modelo.

	Costo USD **(mínimo; promedio; máximo)**	**Ingreso USD** **(mínimo; promedio; máximo)**	**Recuperación USD**
Ingresos recuperados por cliente según actividades en 1 mes.			$267,67
Servicios Desarrollo de software personalizado.			$37,46
Identificación de requisitos, actividad clave.	($28;$112;$280)	($42;$168;$420)	$5,2
Diseño de planos, actividad clave.	($112;$224;$560)	($168;$336;$840)	$7,31
Programación de los requisitos, actividad clave.	($112;$224;$560)	($168;$336;$840)	$13,32
Validación de los requisitos, actividad clave.	($56;$168;$280)	($84;$252;$420)	$7,1
Implementación de mejoras, actividad clave.	($14;$56;$168)	($21;$84;$252)	$2,2
Actividad de montaje.	($8;$32;$128)	($12;$48;$192)	$2,32
Servicios de mantenimiento.			$105,59
Gestión de los contratos de servicio de mantenimiento.	($8;$32;$128)	($12;$48;$192)	$2
Identificación de requisitos de mantenimiento, actividad clave.	($32;$64;$320)	($48;$64;$396)	$1,04
Diseño de planos de mantenimiento,	($128;$256;$392)	($192;$384;$588)	$6,91

actividad clave.			
Programación de los requisitos de mantenimiento, actividad clave.	($56;$320;$1280)	($84;$480;$1920)	$21,59
Validación de los requisitos de mantenimiento, actividad clave.	($8;$128;$320)	($12;$192;$480)	$5,22
Implementación del mantenimiento, actividad clave.	($8;$128;$320)	($12;$192;$480)	$5,84
Gestión de los cobros del servicio y mediciones por horas de servicios.	($8;$32;$128)	($12;$48;$192)	$1,76
Centro de llamada, actividad clave.	($8;$32;$64)	($12;$48;$96)	$1,36
Subcontrato para desarrollo de mantenimiento.	($8;$32;$128)	($12;$48;$192)	$2,1
Subcontrato del servicio *hosting*, actividad clave.	($8;$32;$128)	($8;$48;$192)	$1,2
Viáticos y servicios de transportación para mantenimiento.	($64;$192;$1280)	($96;$288;$1920)	$16,95
Gestión de capacitación de los servicios de mantenimiento.	($64;$192;$1280)	($96;$288;$1920)	$10,03
Gestión de las relaciones públicas.	($64;$192;$1280)	($96;$288;$1920)	$10,56
Asesoría legal para resolución de conflictos.	($64;$192;$1280)	($96;$288;$1920)	$16,7
Actualizaciones de productos, actividad clave.	($8;$64;$192)	($12;$96;$288)	$2,33
Servicio de informatización de empresas.			$23,28
Análisis de la empresa a capacitar, actividad clave.	($64;$128;$192)	($96;$192;$288)	$6,23
Diseño de cronograma para consultoría.	($32;$64;$128)	($48;$96;$192)	$2,49
Consultoría a empresas sobre informatización de empresas, actividad clave.	($64;$320;$980)	($96;$480;$1920)	$14,56
En venta de hardware.			$32,4

Gestión de precios.	($64;$128;$192)	($96;$192;$288)	$5,81
Publicación y presentación del producto.	($32;$64;$128)	($48;$96;$192)	$1,81
Venta de productos.	($32;$64;$128)	($48;$96;$192)	$3,77
Entrega de productos.	($8;$64;$192)	($12;$96;$288)	$3,86
Gestión de stock.	($64;$128;$320)	($96;$192;$480)	$5,63
Gestión de los proveedores.	($64;$128;$192)	($96;$198;$288)	$5,8
Licitación de proveedores.	($8;$64;$128)	($12;$96;$192)	$2,58
Resolución de conflictos con proveedores.	($8;$32;$128)	($12;$48;$192)	$1,93
Evaluación de calidad de los proveedores.	($8;$32;$128)	($12;$48;$192)	$1,18
Marketing.			$68,94
Gestión de la actividad de marketing.	($320;$640;$2560)	($480;$960;$3840)	$45,94
Análisis y lanzamiento nuevos productos.	($64;$192;$320)	(96;288;480)	$10,24
Análisis y marketing de nuevos servicios.	($64;$192;$640)	($96;$288;$960)	$12,77

Para la validación de la aplicabilidad del modelo se emplean técnicas de triangulación metodológica y se combinan las técnicas de evaluación estadística aplicadas con anterioridad con técnicas de juicio de expertos. Se emplean 25 expertos en gestión de proyectos de diferentes instituciones donde se ha aplicado el modelo y otros expertos internacionales en los temas de aseguramiento de ingresos miembros de TMForum.En general se encuestan expertos de los siguientes países: Israel, Ecuador, Estados Unidos, Cuba y Mozambique. Se muestra en la Tabla 19 un resumen que caracteriza a los expertos encuestados y en la Figura 11 se muestra el histograma de frecuencias representado los años de experiencia de los expertos.

Tabla 19. Caracterización de los expertos encuestados para valoración del modelo.

Total de expertos	Cantidad de Doctores	Cantidad de Máster	Promedio de años dedicados	Desviación estándar	Mínima cantidad de años dedicados	Máxima cantidad de años dedicados	Países representados
25	7	18	11,36	6,26	6	37	5

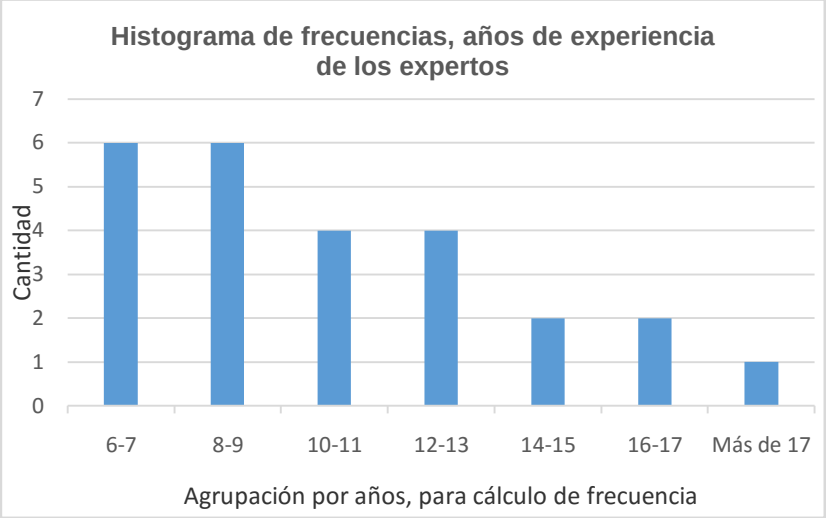


Figura 11. Histograma de frecuencias por años de experiencia de los expertos.

Se encuesta a los expertos acerca del modelo propuesto y su aplicabilidad en los entornos de gestión de proyectos respecto a los siguientes criterios:

C1. Nivel de integración con buenas prácticas de gestión de proyectos basadas en estándares.

C2. Nivel de cubrimiento de enfoques proactivos, para el aseguramiento de ingresos.

C3. Nivel de cubrimiento del enfoque reactivo basado en minería de *datos anómalos*.

C4. Nivel de cubrimiento del enfoque activo basado en minería de *datos anómalos*.

C5. Nivel de tratamiento de la imprecisión y la incertidumbre durante la toma de decisiones.

C6. Nivel de aplicación de técnicas de limpieza de datos que garanticen un adecuado proceso de toma de decisiones.

C7. Nivel de reutilización basado en la obtención de bibliotecas de algoritmos.

C8. Nivel de uso de tecnologías de código abierto, potenciando la reutilización y la soberanía tecnológica de las organizaciones que usen el modelo.

C9. Nivel de implementación del modelo propuesto para su uso en escenarios de gestión de proyectos.

C10. Facilidad para la generalización del modelo en diferentes entornos de gestión de proyectos (construcción, informática, investigación, formación).

C11. Eficacia del algoritmo respecto a la capacidad de detección de datos anómalos.

C12. Eficiencia del sistema respecto a los tiempos de respuesta experimentales.

Para la evaluación se pidió a los expertos que evaluaran cada criterio empleando el siguiente conjunto de términos lingüísticos *LBTL = {Ninguno, Muy bajo, Bajo, Medio, Alto, Muy alto, Altísimo}.* Para unificar la evaluación de los expertos respecto a cada

criterio se empleó la técnica de computación con palabras modelo 2-tuplas. Se obtuvieron los resultados mostrados en la Tabla 20.

Tabla 20. Resultados de la evaluación de expertos del modelo propuesto respecto a los criterios definidos.

Criterios	**Resultado de aplicar 2-tuplas**	**Varianza en la respuesta de los expertos**	**Resultados de aplicar el Cochran test análisis de concordancia a partir de varianzas [175]**
C1	(Alto; 0.36)	0,24	Cochran test for outlying variance: data: x ~ criterios C = 0.11574, df = 25, k = 12, p-value = 1 alternative hypothesis: Group 10 has outlying variance. Cochran test for inlying variance data: x ~ criterios C = 0.065972, df = 25, k = 12, p-value < 2.2e-16 alternative hypothesis: Group 3 has inlying variance Group 6 has inlying variance.
C2	(Alto; 0.48)	0,26	
C3	(Muy Alto; -0.24)	0,19	
C4	(Medio; -0.04)	0,21	
C5	(Alto; -0.48)	0,26	
C6	(Alto; -0.24)	0,19	
C7	(Alto; 0.32)	0,23	
C8	(Muy Alto; -0.28)	0,21	
C9	(Alto; 0.12)	0,28	
C10	(Medio; -0.32)	0,39	
C11	(Alto; 0.32)	0,23	
C12	(Alto; 0.48)	0,26	

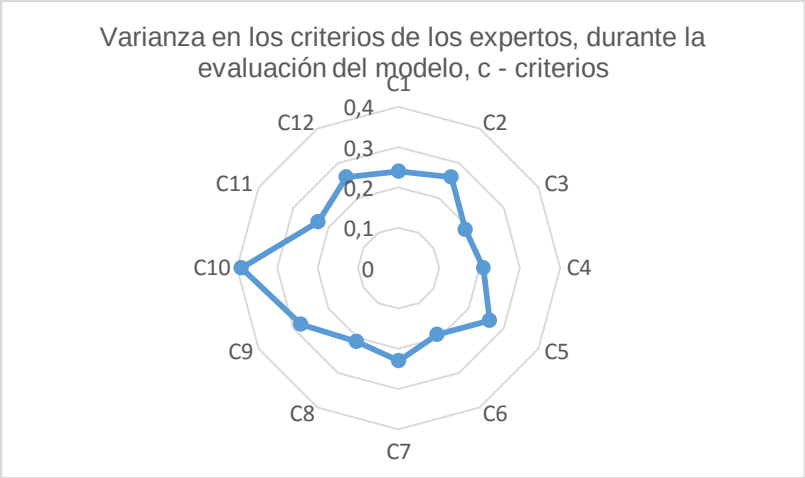


Figura 12.Varianza de la concordancia de los expertos respecto a cada criterio.

Como se aprecia tanto en la Tabla 20, ver resultado test “Cochran” en la Figura 12 no hay una variación significativa respecto a los criterios de los expertos. No obstante, el criterio con mayor variación respecto a la opinión de los expertos fue “Facilidad para la generalización del modelo en diferentes entornos de gestión de proyectos”, elemento que se entiende asociado a la diversidad en la naturaleza de los datos de los diferentes entornos de gestión de proyectos. Se aprecia la mayor concordancia de expertos en la aplicación de las técnicas de limpieza de datos y en el cubrimiento por parte del modelo propuesto del enfoque reactivo.

Se aprecia, además, la valoración positiva del modelo por parte de los expertos, representada por el valor “Alto” en la calificación conjunta de la mayoría de los criterios. Se identifica que los criterios con una evaluación más baja son el C4 y el C10. El C4 referido a la posibilidad de la aplicación del modelo en tiempo real, elemento que no se trabaja suficientemente en la investigación. El C10 asociado a la aplicación del modelo en diversos escenarios, y como se explicó antes influye la diversidad de escenarios y de la naturaleza de los datos en los mismos.

Conclusiones del capítulo

En el proceso de experimentación y análisis de resultados se arriba a las siguientes conclusiones:

- Respecto a los métodos de evaluación cualitativa de riesgos se concluye que el método basado en computación con palabras reporta mejores resultados que la técnica tradicional propuesta por el PMBOK.
- Se demuestra queen las bases de datos empleadas para la experimentación los algoritmos con mejores resultados, respecto a la variable eficacia, fueron el *Combine_outlier_0.92 y kmeans_stats_3.* Mientras, que los peores resultados los presenta *kmeans_euclidean_9_0.92.*
- Se demuestra que en las bases de datos empleadas para la experimentación, los algoritmos con mejores resultados, respecto a la variable eficiencia, fueron de *distance_mahalanobis_3_473_0.92y kmeans_stats_3.* Mientras, que los peores resultados los presentan *angle_5_473_0.95 y crossclustering_5_3.*
- Se desarrolló una biblioteca de algoritmos basados en R y se integran los mismos en el módulo PRODAnalysis, integrado en la plataforma GESPRO.
- La evaluación integral del modelo arrojó que el mismo cumple con todos los indicadores previstos a analizar en la variable independiente y en particular se demuestra que: integra buenas prácticas de gestión de proyectos, complementando las mismas con técnicas de *soft computing* y de minería de datos anómalos.

CONCLUSIONES GENERALES

- El origen del aseguramiento de ingresos estuvo asociado a las empresas de telecomunicaciones, pero su aplicación se ha extendido a numerosas áreas del conocimiento humano, y se identificó en el trabajo una línea abierta a la investigación, la aplicación de estas técnicas en las organizaciones orientadas a proyectos y que combinen los enfoques proactivos, activos y reactivos.
- Para la aplicación de las técnicas de aseguramiento de ingresos se deben considerar las especificidades de cada escenario y la naturaleza de sus datos.
- Se introduce un modelo que permite la detección de situaciones anómalas generadoras de pérdidas de ingreso en organizaciones orientadas a proyectos, basado en la combinación de técnicas de gestión de proyectos, de minería de datos anómalos y computación con palabras.
- Se demuestra que en las bases de datos empleadas para la experimentación los algoritmos con mejores resultados, respecto a la variable eficacia, fueron el *Combine_outlier_0.92* y *kmeans_stats_3*. Mientras, que el de peor resultado fue el *kmeans_euclidean_9_0.92*.
- Se demuestra que en las bases de datos empleadas para la experimentación los algoritmos con mejores resultados, respecto a la variable eficiencia, fueron de *distance_mahalanobis_3_0.92* y *kmeans_stats_3*. Mientras, que los peores resultados los presentan *angle_5_0.95 y crossclustering_5_3*.

- Respecto a los métodos de evaluación cualitativa de riesgos, se concluye que el método basado en computación con palabras reporta mejores resultados que la técnica tradicional propuesta por el PMBOK.
- Se demuestra la aplicabilidad del modelo a partir de su aplicación en escenarios reales y la validación por un conjunto de expertos, se implementó el mismo en la plataforma GESPRO, mediante el módulo PRODAnalysis que explota las funcionalidades de R.

RECOMENDACIONES

- En el proceso 7 de la instrumentación del modelo propuesto, sobre la toma de decisiones, pueden ser empleados sistemas de recomendaciones, entre otras técnicas de la computación emergente. Se recomienda que se trabaje esta temática en investigaciones futuras.
- Se debe continuar investigando acerca de la aplicación del modelo propuesto en escenarios en tiempo real usando estrategias para el cómputo de altas prestaciones y potenciando la aplicación del modelo en un enfoque activo del aseguramiento de ingresos.

PRODUCCIÓN CIENTÍFICA DEL AUTOR

Publicaciones en revistas indexadas

1. Rodríguez C. R., Peña M., **Castro G. F.**, Piñero P. Y. (2017). Sistema clasificador borroso basado en algoritmos genéticos para evaluar el estado de ejecución de proyectos. Revista Cubana de Ciencias Informáticas Vol. 11, No. 3, Editorial “Ediciones Futuro”, Pág. 174- 188, ISSN: 2227-1899 RNPS: 2301 http://rcci.uci.cu.
2. **Castro Gilberto F.,** Pérez, I., et al. (2016). PRODanalysis, un Sistema para el Aseguramiento de Ingresos Basado en Minería de Outliers. INNOVA Research Journal, Vol 1, No. 7, 18-36. ISSN 2477-9024. Disponible en: http://www.journaluidegye.com/magazine/index.php/innova/article/view/34.
3. Torres, S., **Castro Gilberto** F. et al, (2016) Rough Sets for Human Resource Competence Evaluation and Experiences. Applied Mathematics, 2016, 7, 1317-1325. in SciRes. http://www.scirp.org/journal/am, http://dx.doi.org/10.4236/am.2016.712116
4. **Castro Gilberto F**., Pérez, I., et al. (2016). Aplicación de la minería de datos anómalos en organizaciones orientadas a proyectos. Revista Cubana de Ciencias Informáticas Vol. 10, Editorial “Ediciones Futuro”, ISSN: 2227-1899 | RNPS: 2301 http://rcci.uci.cu Pág. 195-209 Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1456.

5. **Castro Gilberto F**, García, Roberto, et al. (2016). Método para el aseguramiento de ingresos de desarrollo de software. Revista Cubana de Ciencias Informáticas Vol. 10, Editorial “Ediciones Futuro”, Pág. 43-57, ISSN: 2227-1899 RNPS: 2301 http://rcci.uci.cu.

6. **Castro Gilberto F.,** Pérez, I., et al (2016). Platform for Project Evaluation Based on Soft-Computing Techniques. Springer International Publishing AG 2016. CCIS 658, pp. 1–15. DOI: 10.1007/978-3-319-48024-4_18. Volume 658 of the series Communications in Computer and Information Science pp 226-240. Disponible en: http://link.springer.com/chapter/10.1007/978-3-319-48024-4_18?no-access=true.

Publicaciones en eventos internacionales

7. **Castro, G. F.**; Pérez, I., Piñero, P. R., Piñero, P. Y., et. at. (2016). Plataforma para aseguramiento de ingresos, aplicación en gestión de proyectos y telcos. IV Taller Internacional Las TIC en la Gestión de las Organizaciones, Informática 2016.

8. **Castro Gilberto F.,** Pérez, I., Piñero P (2016). Platform for Project Evaluation Based on Soft-Computing Techniques. In 2dn International Conference on technologies and innovation CITI 2016, Held in Universidad Agraria de Ecuador, Guayaquil 23-25 November 2016.

REFERENCIAS BIBLIOGRÁFICAS

1 Project Management Institute. IBM: Keys to Building a Successful Enterprise Project Management Office. New York: Project Management Institute, Inc. 2015Disponible en: http://www.pmi.org/-/media/pmi/documents/public/pdf/white-papers/ibm-coe-whitepaper.pdf

2 Project Management Institute. The View From Above: The Power of Portfolio Management. Washinton DC: Project Management Institute, Inc. 2013 Disponible en: http://www.pmi.org/-/media/pmi/documents/public/pdf/white-papers/portofolio-management.pdf

3Project Management Institute. A Guide to the Project Management Body of Knowledge (PMBOK® Guide) (Vol. 5 Edition). Newtown Square, Pennsylvania 19073-3299 EE.UU. Project Management Institute.ISBN 978-1-935589-67-9.2013. Disponible en: http://www.pmi.org/

4 Software Engineering Institute. *CMMI para Desarrollo, Versión 1.3. Mejora de los procesos para el desarrollo de mejores productos y servicios.*Technical Report, Software Engineering Institute, EE.UU. 2010 Disponible en: http://www.sei.cmu.edu/library/assets/whitepapers/Spanish%20Technical%20Report%20CMMI%20V%201%203.pdf

5The Standish Group International. *The CHAOS Manifesto.* The Standish Group International, Incorporated.2014 Disponible en: https://www.projectsmart.co.uk/white-papers/chaos-report.pdf

6 The Standish Group International. *Standish Group 2015 Chaos Report*. 2015. Disponible en: https://www.infoq.com/articles/standish-chaos-2015.

7The Standish Group International.*Big Bang Boom, Chaos Report.*New York: The Standish Group International, Inc. 2014.Disponible en: https://www.standishgroup.com/sample_research_files/BigBangBoom.pdf

8 Villavicencio, N., Peña, M., Burneo, S., Pérez, I.. Experiencias en la integración de procesos en las organizaciones orientadas a proyectos de software. Revista Cubana de Ciencias Informáticas Vol. 10, No. Especial UCIENCIA, ISSN: 2227-1899 | RNPS: 2301 http://rcci.uci.cu Pág. 171-185. 2016. http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1461

9 Paselli, L.The Project Management Advisor: 18 Mayor Project Screw-Ups, and How to Cut Them Off at the Pass. Ed. Financial Times Prentice Hall, 2004, 167 pages, ISBN: 0131490478.

10 Mossalam, A., & Arafa, M. The role of project manager in benefits realization management as a project constraint/driver. Housing and Building National Research Center, HBRC Journal, 56-67. 2014.

11 Castro, G.F., Pérez, I., Piñero, P. Y., García, R.Método para el aseguramiento de ingresos en entornos de desarrollo de software. Revista Cubana de Ciencias Informáticas Vol. 10, No. Especial UCIENCIA, ISSN: 2227-1899 | RNPS: 2301. 2016.http://rcci.uci.cu Pág. 43-57. Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1463

12 Mattison, R. The Telco Revenue Assurance Handbook. XiT Press, Oakwood Hills, Illinois, USA. ISBN: 1-4116-2801-2. 2005. Disponible en: http://www.grapatel.com/A-GRAPA/07-Library/RABook.asp#top

13 Aggarwal, CH. C. Datos anómalo Analysis. IBM T.J. Watson Research Center Yorktown Heights New York USA, ISBN 978-1-4614-6396-2 (eBook), DOI 10.1007/978-4614-6396-2 Springer science + Business Media, New York, Heidelberg Dordrecht London. 2013.

14TM Forum.Revenue Assurance a survey pre-result blog: Lack of cross-functional mandate holds back change, say Revenue Assurance professionals.2015. Disponible en: https://inform.tmforum.org/features-and-analysis/2014/12/revenue-assurance-survey-2014-maturity-rise/

15TM Forum Revenue Assurance practitioner blog: Do we need a new approach to revenue assurance in the digital world? & Seeing is believing: Setting revenue assurance KPIs. 2014. Disponible en: https://inform.tmforum.org/

16 Acosta, K. Aseguramiento de ingresos: una actividad fundamental en las empresas de telecomunicaciones, Ingeniería Industrial, vol. XXIX, núm. 2, 2008, pp. 1-6, ISSN: 0258-5960. Disponible en: http://www.redalyc.org/pdf/3604/360433566002.pdf

17 CNT Portal colaborativo. Departamento de Aseguramiento de Ingresos. 2015. Disponible en: http://corporativo.cnt.gob.ec/cnt-ep-contribuye-con-el-estado/

18GRAPA. The Global Revenue Assurance Professional Association (GRAPA) Professionalizing the Information, Communications and Technology Industry. 2016. Disponible en: http://www.grapatel.com/

19 Khan, N. Internship Report on Revenue Assurance and Fraud Management. ID: 10104009. BRAC Business School. 2014. Disponible en: http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/3180/10104009.pdf?sequence=1

20Massyn, R. H. A provisional taxonomy of revenue assurance: a grounded theory approach. M.Sc. in Philosophy, University of Johannesburg, TH 621.3820683. 2010. Disponible en: https://ujdigispace.uj.ac.za

21 Almache,C. J.Desarrollo e Implementación de un Sistema de Gestión para Aseguramiento de Ingresos en Telefonía Fija. Caso ETAPA. Trabajo de graduación previo a la obtención del título de Máster en Administración de Empresas, Universidad del Azuay, Ecuador. 2009.

22 Castro, G. F.; Pérez, I., et. at. Plataforma para aseguramiento de ingresos, aplicación en gestión de proyectos y telcos. IV Taller Internacional Las TIC en la Gestión de las Organizaciones, Informática 2016.2016.

23 Villavicencio, N. Modelo integrado para la mejora de la productividad en organizaciones orientadas a proyectos de tecnologías de la información.Tesis para optar al grado de: Máster en Diseño, Gestión y Dirección de Proyectos, Fundación Universitaria Iberoamericana FURNIBER, Área de ingeniería, proyectos prevención y calidad. 2016.

24Villavicencio, N., Peña,M., et al. Experiencias en la integración de procesos en las organizaciones orientadas a proyectos de software. Revista Cubana de Ciencias Informáticas (RCCI), Vol. 10, No. Especial UCIENCIA, ISSN: 2227-1899 | RNPS: 2301, Pág. 171-185. 2016. Disponible en: http://rcci.uci.cu

25Mattison, R. The Revenue Assurance Standards – Release 2009, GRAPA. XiT Press, Oakwood Hills, Illinois, USA. ISBN-13: 978-0557254750. 2009. Disponible en: http://www.grapatel.com/members/viewfile.asp?file=stdbook

26 IPMA. International Project Management Association.2015. Disponible en: http://www.ipma.world/

27 Heredia, R. Dirección Integrada de Proyecto - DIP -. Segunda edición, Universidad Politécnica de Madrid, ISBN: 84-7484-108-9, 605 páginas. 1995.

28 ISO. ISO 21500:2012 Guidance on Project Management. International Organization for Standardization.2012. Disponible en: http://www.iso.org/iso/catalogue_detail?csnumber=50003

29 Piñero,P. Y., Pérez, I., et al.Sistema de Información para la Gestión de Organizaciones Orientadas a Proyectos. DOI: 10.13140/2.1.3491.1522, V Congreso Iberoamericano de Ingeniería de Proyectos, Loja Ecuador. 2014.Disponible en: http://congreso.riipro.org/index.php/CIIP/V-CIIP/paper/viewFile/105/31

30 STS Sauter Training and Simulation. Comparing PMBOK Guide 4th, PMBOK Guide 5th and ISO 21500, STS Sauter Training and Simulation. 2016. Disponible en: http://www.pmi-netherlands-chapter.org/images/stories/PMI-data/chapter-news/pmbokiso.pdf

31 Chang, C.-Y. Risk-bearing capacity as a new dimension to the analysis of project governance. International Journal of Project Management, 33(1), 1195 – 1205. DOI: 10.1016/j.ijproman.2015.02.003. 2015.

32Kerzner, H. Project management: a systems approach to planning, scheduling, and controlling. John Wiley & Sons, ISBN: 0884222245414. 2013. Disponible en: http://www.academia.edu/download/21360035/enma604-syllabus.pdf

33 Martinez Elarre, M.Consolidating the presence of EOS Project Management in Europe. Faculty of Economics and Business, Universidad Pública de Navarra. 2016. Disponible en: https://academica-e.unavarra.es/bitstream/handle/2454/21151/Martinez%20Elarre%20Miren.pdf?sequence=1&isAllowed=y

34 Paquin, J. P., Gauthier, C., & Morin, P. P. The downside risk of project portfolios: The impact of capital investment projects and the value of project efficiency and project risk management programmes. International Journal of Project Management, 34(8), 1460-1470, ISSN: 0263-7863, DOI: 10.1016/j.ijproman.2016.07.009. 2016.

35Turner, R. Gower handbook of project management. Ashgate 4th edition, 912 pages, ISBN: 978-0566088063. 2016.

36Alotaibi, A. B., & Mafimisebi, O. P. Project Management Practice: Redefining Theoretical Challenges in the 21st Century. Journal of Economics and Sustainable Development, 7(1), ISSN 2222-1700. 2016. Disponible en: https://www.researchgate.net/publication/299590063

37Klakegg, O. J. Project Risk Management: Challenge Established Practice. Administrative Sciences, 6(4), 21; doi:10.3390/admsci6040021. 2016. Disponible en: http://www.mdpi.com/2076-3387/6/4/21/htm.

38Klakegg, O.J., Williams, T., Shiferaw, A.T. Taming the ‘trolls’: Major public projects in the making. Int. J. Proj. Manag. 34(2), 282–296, doi:10.1016/j.ijproman.2015.03.008, ISSN 0263-7863. 2016. Disponible en: http://dx.doi.org/10.1016/j.ijproman.2015.03.008.

39 Johansen, A.; Eik-Andresen, P.; Dypvik Landmark, A.; Ekambaram, A.; Rolstadås, A. Value of Uncertainty: The Lost Opportunities in Large Projects. Adm. Sci. 2016, 6(3), 11, doi:10.3390/admsci6030011.

40Klakegg, O. J., Torp, O., Austeng, K. (). Good and simple – a dilemma in analytical processes? International Journal of Managing Projects in Business, ISSN: 1753-8378, DOI: http://dx.doi.org/10.1108/17538371011056057.

41Stasiak-Betlejewska, R., & Potkány, M.Construction Costs Analysis and its Importance to the Economy. Procedia Economics and Finance, Business Economics and Management 2015 Conference, BEM2015, 34, 35-42, doi: http://dx.doi.org/10.1016/S2212-5671(15)01598-1. 2015. Available online at www.sciencedirect.com.

42Johnson, M. Demystifying Communications Risk: A Guide to Revenue Risk Management in the Communications Sector. Routledge New edition edition, 270 pages, ISBN: 978-1409429418. 2016.

43Burke, R. Project Management: Planning and Control Techniques. Wiley 5 edition, 428 pages, New Jersey, USA, ISBN: 978-1118561256. 2013.

44Schwalbe, K. Information technology project management. Cengage Learning 7 edition, 656 pages, ISBN: 978-1285847092. 2015.

45Phillips, J. PMP, Project Management Professional (Certification Study Guides). Sybex 7 edition, 696 pages, McGraw-Hill Osborne Media, ISBN: 978-1118531822. 2013.

46 Leach, L. P. Critical chain project management. The North River Press 1st edition, 246 pages, Artech House, ISBN: 978-0884271536. 2014.

47Verzuh, E. The fast forward MBA in project management. Wiley 5 edition, 528 pages. ISBN: 978-1119086574. 2015.

48Fischer, H., Dreisiebner, S., et al.Revenue vs. Costs of MOOC Platforms. Discussion of Business Models for xMOOC Providers Based on Empirical Findings and Experiences During Implementation of the Project iMOOX. In 7th International Conference of Education, Research and Innovation (ICERI2014). IATED (pp. 2991-3000). 2014.

49 Wojnar, K. Comparison between ISO 21500 and PMBOK® Guide 5th Edition. Theoretical background and practical usage of ISO 21500 in IT projects. 2013.

50 Acanda, J. Modelo para la evaluación para programas de proyectos basados en técnicas de soft comuting. Trabajo final presentado en opción al título de Máster en Gestión de Proyectos Informáticos. Facultad 5, Departamento de Investigaciones en Gestión de Proyectos.2015.

51 Castro, G.F., Pérez**,** I., et al. Aplicación de la minería de datos anómalos en organizaciones orientadas a proyectos. Revista Cubana de Ciencias Informáticas Vol. 10, No. Especial UCIENCIA, ISSN: 2227-1899 | RNPS: 2301 http://rcci.uci.cu. Pág. 195-209. 2016. Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1456

52 Piñero, P. Y., Bermúdez,A. Banco de problemas de investigaciones en Gestion de Proyectos. Conferencia Científica Uciencia 2016, II Taller Internacional de Gestión de Proyectos, Panel Aplicaciones de la Inteligencia Artificial a la Gestión de Proyectos, ISBN: 978-959-286-054-4.2016.

53Ben-Gal, I. Outlier detection. Data Mining andKnowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, ISBN 0-387-24435-2. Department of Industrial Engineering, Tel-Aviv University. 2005.

54Karanjit,S. and Upadhyaya, S.Outlier Detection: Applications And Techniques.*IJCSI International Journal of Computer Science Issues,* Vol. 9, Issue 1, No 3, ISSN (Online): 1694-0814, 2012. www.IJCSI.org

55 Deneshkumar, V., Senthamaraikannan, V., et al.Identification of Datos anómalos in Medical DiagnosticSystem Using Data Mining Techniques.*International Journal of Statistics and Applications,* 4(6): 241-248, DOI: 10.5923/j.statistics.20140406.01. 2014.Disponible en:http://www.researchgate.net/publication/274721695

56Chen, X.Optimizing MPBSM Resource Allocation Based on Revenue Management: A China Mobile Mobile Information Systems Volume 2015, Hindawi Publishing Corporation, Article ID 892705, 10 pages. 2015. Disponible en: http://dx.doi.org/10.1155/2015/892705

57 Guerriero, F., Miglionico, G., et al.Strategicandoperationaldecisionsinrestaurantrevenuemanagement.European

Journal of Operational Research, vol. 237, no. 3, pp. 1119–1132. 2014. Disponible en: http://dx.doi.org/10.1016/j.ejor.2014.02.048

58Schwartz, Z., Stewart, W., & Backlund, E. A. Visitation at capacity-constrained tourism destinations: Exploring revenue management at a national park. Tourism Management, 33(3), 500-508. DOI: 10.1016/j.tourman.2011.05.008. 2012.

59Wrubel, E., & Gross, J. Contracting for Agile Software Development in the Department of Defense: An Introduction (Vols. CMU/SEI-2015-TN-006, http://www.sei.cmu.edu). Pittsburgh, PA 15213-3890, EEUU: Carnegie Mellon University.2015.

60 Ferrara,E., De Meo,P., et al.Detecting criminal organizations in mobile phone networks. Expert Systems and Applications, 41(13), 5733-5757. 2014. Disponible en:http://dx.doi.org/10.1016/j.eswa.2014.03.024, http://www.elsevier.com/locate/eswa.

61 Manish,G., Jing,G., et al.Outlier Detection for Temporal Data.ISBN: 9781627053754 (paperback), ISBN: 9781627053761 (ebook), DOI 10.2200/S00573ED1V01Y201403DMK008. 2014. Disponible en:www.morganclaypool.com.

62 Souza, A. M., & Amazonas, J. R.An outlier detect algorithm using big data processing and internet of things architecture. Procedia Computer Science, 52, 1010-1015, ISSN: 1877-0509, DOI: 10.1016/j.procs.2015.05.095. 2015.

63 Whyte, J., Stasis, A., & Lindkvist, C. Managing change in the delivery of complex projects: Configuration management, asset information and ‘big data’.

International Journal of Project Management, 34(2), 339 – 351. 2016. Disponible en: http://dx.doi.org/10.1016/j.ijproman.2015.02.006.

64 Kriegel, H. P., Kröger, P., & Zimek, A. Outlier detection techniques. Tutorial at KDD, 10. 2010. Disponible en: http://www.imada.sdu.dk/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf

65Ro, K., Zou, C., Wang, Z., & Yin, G. Outlier detection for high-dimensional data. Biometrika, 102(3), 589-599. 2015. Disponible en: http://web.stat.nankai.edu.cn/zjwang/publications/2015/rzwy2015-bio.pdf

66Barmade, A., & Nashipudinath, M. M. An Efficient Strategy to Detect Outlier Transactions. International Journal of Soft Computing and Engineering (IJSCE), 3(6), 174-178, ISSN: 2231-2307. 2014.Disponible en: https://pdfs.semanticscholar.org/6574/a7dc14ee3d46ffd60d018c985e6fa5681e82.pdf

67 Deneshkumar, V., Senthamaraikannan, K., & Manikandan, M.Identification of Outliers in Medical Diagnostic System Using Data Mining Techniques. International Journal of Statistics and Applications, 4(6), 241-248.DOI: 10.5923/j.statistics.20140406.01. 2014. Disponible en: http://www.researchgate.net/publication/274721695

68Zimmermann, A.A feature construction framework based on dato anómalo detection and discriminative pattern mining. 2014. Disponible en: http://www.researchgate.net/publication/264049231.

69 Aggarwal, C. C. (Ed.).Managing and mining sensor data. Springer Science & Business Media. ISBN 978-1-4614-6308-5. 2013.

70Aggarwal, C. C., & Zhai, C. (Eds.). Mining text data. Springer Science & Business Media. ISBN 978-1-4614-3222-7. 2012.

71Aggarwal, C. C. Outlier analysis. In Data mining (pp. 237-263). Springer International Publishing.2015.

72 Aggarwal, C. C., & Reddy, C. K. (Eds.). Data clustering: algorithms and applications. Chapman and Hall/CRC.2013. Disponible en: http://www.crcnetbase.com/doi/pdf/10.1201/b15410-1

73 Aggarwal, C. C.High-Dimensional Outlier Detection: The Subspace Method. In Outlier Analysis (pp. 135-167). Springer New York. 2013. Disponible en: http://www.charuaggarwal.net/High_Dimensional_Outlier_Detection_Survey.pdf

74 Singh, J. & Aggarwal, S. Survey on Outlier Detection in Data Mining. International Journal of Computer Application, 67(19),0975 – 8887.2013.

75 Vijendra, S., & Shivani, P.Robust Outlier Detection Technique in Data Mining: A Univariate Approach. arXiv preprint arXiv:1406.5074. Faculty of Engineering and Technology, Mody Institute of Technology and Science, Lakshmangarh, Sikar, Rajasthan, India. 2014.Disponible en: https://arxiv.org/ftp/arxiv/papers/1406/1406.5074.pdf

76 Ren, G.Detection of Outliers in a time series of available parking spaces.*Mathematical Problems in Engineering*, vol. 2013, Article ID 416267, 12 pages. 2013.

77Hu, W. and Bao,J. The interval outliersdetectionalgorithms on astronautical time series data.*Mathematical Problems in Engineering*, vol. 2013, Article ID 979035, 6 pages, 2013. Disponible en: http://dx.doi.org/10.1155/2013/979035

78Shpigelman. A Unified Framework for Outlier Detection in Trace Data Analysis. IEEE Transactions on semiconductor manufacturing, vol. 27, no. 1, Impact Factor: 0.98 · DOI: 10.1109/TSM.2013.2267937.2014.

79 Hubert, M., Rousseeuw, P. J., & Segaert, P.Multivariate functional outlier detection. Statistical Methods & Applications, 24(2), 177-202. 2015. Disponible en: https://lirias.kuleuven.be/bitstream/123456789/481784/1/MFOD_revision.pdf

80 Markov, A.A. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons. 1971.

81Hardy, G. H., Littlewood, J. E., Pólya, G.Inequalities. Cambridge Mathematical Library. Cambridge: Cambridge University Press. ISBN 0-521-35880-9. MR 0944909.1988.

82Srivastava, M. S., & Von Rosen, D. Outliers in multivariate regression models. *Journal of Multivariate Analysis*, *65*(2), 195-208. 1988.

83Bro, R., & Smilde, A. K.Principal component analysis. Analytical Methods, 6(9), 2812-2831, DOI: 10.1039/C3AY41907J. 2014.

84Knorr, E. M., & Ng, R. T. Finding intensional knowledge of distance-based outliers. In VLDB (Vol. 99, pp. 211-222), ISBN:1-55860-615-7. 1999.

85Knorr, E. M., Ng, R. T., & Tucakov, V. Distance-based outliers: algorithms and applications. The VLDB Journal—The International Journal on Very Large Data Bases, 8(3), 237-253, DOI: 10.1007/s007780050006, ISSN1066-8888.2000.

86Ghoting, A., Parthasarathy, S., & Otey, M. E. Fast mining of distance-based outliers in high-dimensional datasets. In Proceedings of the 2006 SIAM

International Conference on Data Mining (pp. 609-613). Society for Industrial and Applied Mathematics.DOI: http://dx.doi.org/10.1137/1.9781611972764.70. 2006. Disponible en: http://epubs.siam.org/doi/pdf/10.1137/1.9781611972764.70.

87Hautamaki, V., Karkkainen, I., & Franti, P.Outlier detection using k-nearest neighbour graph. In Pattern Recognition, ICPR 2004. Proceedings of the 17th International Conference, vol. 03, 430-433, doi:10.1109/ICPR.2004.1334558, ISSN: 1051-4651, ISBN: 0-7695-2128-2. 2004. Disponible en: http://doi.ieeecomputersociety.org/10.1109/ICPR.2004.1334558

88Gogoi, P., Borah, B., et al.Outlier Identification Using Symmetric Neighborhoods. 2nd International Conference on Communication Computing & Security [ICCCS 2012], Procedia Technology 6, 239 – 246, Elsevier Ltd. Selection and/or peer-review under responsibility of the Department of Computer Science & Engineering, National Institute of Technology Rourkela, doi: 10.1016/j.protcy.2012.10.029. 2012.Disponible en: http://dx.doi.org/10.1016/j.protcy.2012.10.029

89Ramaswamy, S., Rastogi, R., & Shim, K.Efficient algorithms for mining outliers from large data sets. In ACM Sigmod Record (Vol. 29, No. 2, pp. 427-438). ACM. 2000. Disponible en: ftp://ftp10.us.freebsd.org/users/azhang/disc/disc01/cd1/out/papers/sigmod/efficientalgorisrrak.pdf.

90Breunig, M. M., Kriegel, H. P., et al.LOF: identifying density-based local outliers. In ACM sigmod record, 29(2), 93-104, I SBN:1-58113-217-4, doi: 10.1145/342009.335437.2000. Disponible en:

http://people.cs.vt.edu/badityap/classes/cs6604-Fall13/readings/breunig-2000.pdf.

91 Kriegel, H. P., Kröger, P., & Zimek, A. Outlier detection techniques. Tutorial at KDD, 10. 2010. Disponible en: http://www.imada.sdu.dk/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf

92Papadimitriou, S., Kitagawa, H., et al. Loci: Fast outlier detection using the local correlation integral. In Data Engineering, 2003. Proceedings. 19th International Conference on (pp. 315-326). IEEE. 2003.Disponible en: http://www.dtic.mil/dtic/tr/fulltext/u2/a461085.pdf

93Howe, D. Clustering and anomaly detection in tropical cyclones.School of Information Technologies, Faculty of Engineering & Information Technologies, the University of Sidney. 2013. Disponible en: https://sydney.edu.au/engineering/it/research/conversazione-2013/HOWE-David.pdf

94Zolhavarieh, S., Aghabozorgi, S., et al.A Review of Subsequence Time Series Clustering.*The Scientific World Journal*, Volume 2014, Article ID 312521, 19 pages, http://dx.doi.org/10.1155/2014/312521. 2014. Disponible en: http://www.researchgate.net/publication/263470292

95Lee, Ch.& Lee, H. Novelty-focussed document mapping to identify new service opportunitie. The Service Industries Journal, Vol. 35, No. 6, 345–361, Impact Factor: 2.58 · DOI: 10.1080/02642069.2015.1003368.2015. Disponible en: http://dx.doi.org/10.1080/02642069.2015.1003368

96Lee, Ch., Kang, B.,et al.Novelty-focused patent mapping for technology opportunity analysis. Technological Forecasting & Social Change, 355–365, Elsevier, Impact Factor: 1.71 · DOI: 10.1016/j.techfore.2014.05.010.2014. Disponible en: http://dx.doi.org/10.1016/j.techfore.2014.05.010

97Milton, P.,Georgina, S., et al.Cluster Ensembles for Big Data Mining Problems.2015. Disponible en: http://www.researchgate.net/publication/281461828

98Zhiguo, L.,Robert, B., et al.A Unified Framework for Outliers Detection in Trace Data Analysis. IEEE Transactions on semiconductor manufacturing, vol. 27, no. 1, Impact Factor: 0.98, DOI: 10.1109/TSM.2013.2267937.2014.

99 Cravero, A., Sepúlveda, S. Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio. Workshop internacional EIG2009. 2009. Disponible https://pdfs.semanticscholar.org/1731/e33c6f8c99d3f00f3b372d4f359bdc4d4df1.pdf

100 Vankeerberghen, P., Smeyers-Verbeke, J., et al. Robust regression and outlier detection for non-linear models using genetic algorithms. Chemometrics and Intelligent Laboratory Systems, Volume 28, Issue 1, Pages 73-87. 1995. Disponible en: http://dx.doi.org/10.1016/0169-7439(95)80041-7

101 Rousseeuw, P. J., & Leroy, A. M. Robust regression and outlier detection (Vol. 589). John wiley & sons, ISBN: 9780471852339, DOI: 10.1002/0471725382. 2005.

102 Radovanović, M., Nanopoulos, A., & Ivanović, M.Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE transactions on knowledge and data engineering, 27(5), 1369-1382. 2015. Disponible en: http://perun.dmi.rs/radovanovic/publications/2015-tkde-outliers.pdf

103Navile, D., & Ravikumar, G. K.Outlier Detection in High Dimension Data Based On Multimodality And Neighbourhood Size Using KNN Method. 2016. Disponible en: http://ijecs.in/issue/v5-i5/35%20ijecs.pdf

104 Rakhe, S. S., & Vaidya, A. S.Enhanced Outlier Detection for High Dimensional Data Using Different Neighbor Metrics. International Journal of Engineering Science, 1568. 2016. Disponible en: http://ijesc.org/upload/952f9f9d64768bec295e4934b7bcd2d2.Enhanced%20Outlier%20Detection%20for%20High%20Dimensional%20Data%20Using%20Different%20Neighbor%20Metrics.pdf

105 Kumar, M. K. S., Ramakrishna, M. P., & Naik, M. G. M.Unsupervised Outlier Detection Using Reverse Neighbors Counts. International Journal of Computer Science Engineering (IJCSE), 5(3), 193-201, ISSN: 2319-7323. 2016. Disponible en: http://www.ijcse.net/docs/IJCSE16-05-03-129.pdf

106 Yang, Y., Ong, S. H., & Foong, K. W. C. A robust global and local mixture distance based non-rigid point set registration. Pattern Recognition, 48(1), 156-173, Elsevier Science Inc. New York, NY, USA, doi:10.1016/j.patcog.2014.06.017, ISSN: 0031-3203. 2015.

107 Hazewinkel, M. ed. Mahalanobis distance. Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4. 2001.

108 Duda, R. O., Hart, P. E., & Stork, D. G. Pattern classification. John Wiley & Sons. Second Edition. ISBN: 0-471-05669-3. 2012.

109Lane, T. and Brodley, C. Temporal Sequence Learning and Data Reduction for Anomaly Detection, ACM Transactions on Informa- tion and Security, 2(3), pp. 295–331. 1999.

110Liu,X., Zhang, P., et al. Sequence Matching for Suspicious activity Detection in Anti-money Laundering. Lecture Notes in Computer Science, Vol. 5075, pp. 50–61. DOI: 10.1007/978-3-540-69304-8_6, ISSN: 0302-9743. 2008.

111 Mani, I., & Zhang, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. Proceedings of the ICML Workshop on Learning from Imbalanced Datasets. 2003. Disponible en: https://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf

112 Kumar, V., Kumar, S., Kumar, A.Outlier Detection: A Clustering-Based Approach. International Journal of Science and Modern Engineering, 1(7), 16-19, ISSN: 2319-6386. 2013.

113 Tellaroli, P., Bazzi,M., et at. CrossClustering: apartial clustering algorithm with automatic estimation of the number of clusters. Package 'CrossClustering' Version 3.0. PLOS One (In Press). 2016. Disponible en: https://cran.r-project.org/web/packages/CrossClustering/CrossClustering.pdf

114 Chawla, S. and Gionis, A.k-means--: A unified approach to clustering and outlier detection. SIAM International Conference on Data Mining (SDM13).DOI: http://dx.doi.org/10.1137/1.9781611972832.21 2013. Disponible en: http://www.pmg.it.usyd.edu.au/outliers.pdf

115 Howe, D. Ch. K-Means with Simultaneous Outlier Detection. Package ‘kmodR’ Version 0.1.0. 2016. Disponible en: https://cran.r-project.org/web/packages/kmodR/kmodR.pdf

116Patel, S. P., Shah, V., & Vala, J. Outlier Detection in Dataset using Hybrid Approach. International Journal of Computer Applications (0975 – 8887), 122(8). 2015.

117 Jimenez, J. Angle-Based Outlier Detection. Package ‘abodOutlier’ Version 0.1.MIT License. 2016. Disponible en: https://cran.r-project.org/web/packages/abodOutlier/abodOutlier.pdf

118 Kriegel, H-P., Schubert, M., et al. Angle-Based Outlier Detection in High-dimensional Data. KDD’08, Las Vegas, Nevada, USA. ACM 978-1-60558-193-4/08/08. 2008. Disponible en: http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD2008.pdf

119Bao, Z.Two Phases Outlier Detection in Different Subspaces. DOI: 10.1145/2663714.2668046. 2014. Disponible en: http://www.researchgate.net/publication/268040848

120Schölkopf, B., Williamson, R. C., et al. Support vector method for novelty detection. In NIPS (Vol. 12, pp. 582-588). 1999. Disponible en: https://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf

121Tang,Y., Zhang,Y.-Q., et al. SVMs Modeling for Highly Imbalanced Classification, IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics, 39(1), pp. 281– 288. 2009.

122Al-Khateeb, T., Masud, M. M., et al.Recurring and Novel Class Detection using Class-Based Ensemble for Evolving Data Stream. IEEE Transactions on Knowledge and Data Engineering, 28(10), 2752-2764.2016. Disponible en: http://ieeexplore.ieee.org/abstract/document/7350165/?reload=true

123Masud, M. M., Al-Khateeb, T. M., et al.Detecting recurring and novel classes in concept-drifting data streams. In Data Mining (ICDM), 2011 IEEE 11th International Conference on (pp. 1176-1181). IEEE. 2011. Disponible en: https://pdfs.semanticscholar.org/398d/ac12db5417221f14cb2b0338ba6b2a1e1b6e.pdf

124Masud, M. M., Chen, Q., et al.Classification and adaptive novel class detection of feature-evolving data streams. IEEE Transactions on Knowledge and Data Engineering, 25(7), 1484-1497. 2013.

125Pelleg, D., & Moore, A. W. Active Learning for Anomaly and Rare-Category Detection. In NIPS Conference (pp. 1073-1080). 2004. Disponible en: https://papers.nips.cc/paper/2554-active-learning-for-anomaly-and-rare-category-detection.pdf

126Melville, P., & Mooney, R. J. Diverse ensembles for active learning. In Proceedings of the twenty-first international conference on Machine learning (p. 74). ACM. 2004. Disponible en: http://machinelearning.wustl.edu/mlpapers/paper_files/icml2004_MelvilleM04.pdf

127Wu, G., & Chang, E. Y. Class-boundary Alignment for Imbalanced Dataset Learning. Proceedings of the ICML Workshop on Learning from Imbalanced Data

Sets *II, Washington, DC* (pp. 49-56).2003. Disponible en: www.site.uottawa.ca/~nat/Workshop2003/Wu-final.pdf.

128Drummond, C., & Holte, R. C.C4.5, Class Imbalance, and Cost Sensitivity: Why Undersampling beats Oversampling. ICML Workshop on Learning from Imbalanced Data Sets II (Vol. 11).2003. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.6858&rep=rep1&type=pdf

129Liu, B., Lee, W. S., et al.Partially supervised classification of text documents. In ICML (Vol. 2, pp. 387-394). 2002.

130 Li, X. L. Partially Supervised Text Categorization. In Handbook of Research on Text and Web Mining Technologies (pp. 75-95). IGI Global. 2009.

131Li, X. L., Liu, B., & Ng, S. K.Negative training data can be harmful to text classification. In Proceedings of the 2010 conference on empirical methods in natural language processing (pp. 218-228). Association for Computational Linguistics.2010. Disponible en: https://pdfs.semanticscholar.org/b7a3/cd23c49b0973dcec2b7145003c605c06e199.pdf

132Zadrozny, B., & Elkan, C.Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 204-213). ACM. 2001. Disponible en:http://sci2s.ugr.es/keel/pdf/specific/congreso/zadrozny01learning.pdf

133Ting, K. M. An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering, 14(3), 659-665. 2002.

134Weiss, G. M., & Provost, F. Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research, 19, 315-354.2003. Disponible en: http://www.jair.org/media/1199/live-1199-2209-jair.pdf

135Zhang, J., Gao, Q., & Wang, H. SPOT: A system for detecting projected outliers from high-dimensional data streams. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on (pp. 1628-1631). IEEE. 2008. Disponible en: https://pdfs.semanticscholar.org/faaf/c37801c4132c2724a2d067468a4e6a80da60.pdf

136Rajeswari, A.M., Sridevi, M., et al.Outliers Detection on Educational Data using Fuzzy Association Rule Mining",Int. Conf. on Adv. in Comp., Comm., and Inf. Sci. (ACCIS-14) (1–9). 2014. Disponible en: http://www.researchgate.net/publication/263814468

137Williams, G., Baxter, R., et al. Hongxing He, Simon Hawkins and Lifang Gu, "A Comparative Study of RNN for Outlier Detection in Data Mining", *CSIRO Mathematical and Information Sciences,* CSIRO Technical Report CMIS-02/102. 2002. Disponible en: http://datamining.csiro.au.

138 Menzel, C., & Mayer, R. J.The IDEF family of languages. In Handbook on architectures of information systems (pp. 209-241). Springer Berlin Heidelberg, DOI 10.1007/978-3-662-03526-9_10, ISBN 978-3-662-03528-3. 1998.

139Kuna, H.D.,et al.Outlier detection in audit logs for application systems", *Information Systems 44,* 22–33, Elsevier, Impact Factor: 1.24 · DOI: 10.1016/j.is.2014.03.1. 2014. Disponible en: http://www.researchgate.net/publication/262915159

140 Cucina, D., Di Salvatore, A., & Protopapas, M.'Meta-heuristic Methods for Outliers Detection in Multivariate Time Series (No. 003). 2008. Disponible en: http://www.dss.uniroma1.it/en/system/files/pubblicazioni/54_RT_4_2013_Meta-heuristic%20Methods%20for%20Outliers%20Detection%20in%20Multivariate%20Time%20Series.pdf

141 Krishna, G. J., & Ravi, V. Outlier Detection using Evolutionary Computing. In Proceedings of the International Conference on Informatics and Analytics (p. 17). ACM. 2016. Disponible en: https://www.researchgate.net/profile/Gutha_Krishna/publication/309664931_Outlier_Detection_using_Evolutionary_Computing/links/583a537308aed5c61489e900.pdf

142He, Z., Xu, X., et al.A fast greedy algotrihm for dato anómalomining. 2015. Disponible en: http://www.researchgate.net/publication/231521185

143Tang, G., Pei, J., et al.Mining multidimensional contextual Outliers from categorical relational data. DOI: 10.1145/2484838.2484883.2013. Disponible en: http://www.researchgate.net/publication/266654131.

144 Porter, M., & Kramer, M. Estrategia y sociedad. Harvard business review, 84(12), 42-56, ISSN 0717-9952. 2006.

145 Piñero, P. Y.Matriz DAFO cuantificada y computación con palabras. Tema 1: Dirección Estratégica. Curso de gestión de organizaciones orientadas a proyectos. Maestría en Gestión de Proyectos Informáticos, COPED 2009, Universidad de las Ciencias Informáticas.2016.

146 Dorr, B., & Herbert, P. Data profiling: Designing the blueprint for improved data quality. DataFlux Corporation, Cary, NC, SUGI 30: Data Warehousing, Management and Quality, Philadelphia, Pennsylvania, EEUU, 10 p. 2005.

147 Herrera, F. & Martınez, L. A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems, 8(6):746–752, ISSN: 1941-0034. 2000.

148 Contraloría General.Normas del Sistema de Control Interno, Resolución No. 60/11, Contraloría General de la República de Cuba. 2011.

149 López, B. Limpieza de Datos: Reemplazo de valores ausentes y Estandarización. Tesis Doctoral, Facultad de Matemática y Computación. Santa Clara, Universidad Central “Marta Abreu” de Las Villas, p. 45. 2011.

150Porrero, B. L., & Vázquez, R. P. Taxonomía de errores en las bases de datos cubanas. Revista Cubana de Ciencias Informáticas, 2(1-2), ISSN: 1994-1536, pp. 63-69. 2008. Disponible en: http://www.redalyc.org/pdf/3783/378343635008.pdf

151Torres, S. Modelo de evaluación de competencias a partir de evidenciasdurante la gestión de proyectos. Tesis en opción del grado científico de Doctor en Ciencias Técnicas, Laboratorio de investigaciones en Gestión de Proyectos, Universidad de las Ciencias Informáticas, La Habana, Cuba. 2015.

152 García, R., Pérez, I., et al. Experiencias usando algoritmos genéticos en la planificación de proyectos. Revista Cubana de Ciencias Informáticas Vol. 10, No. Especial UCIENCIA, ISSN: 2227-1899 | RNPS: 2301 http://rcci.uci.cu Pág. 71-86. 2016. Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1462

153 González, R. S., Pupo, I. P., et al. Ecosistema de Software GESPRO-16.05 para la Gestión de Proyectos. Revista Cubana de Ciencias Informáticas, 10, 239-251, ISSN: 2227-1899 | RNPS: 2301. 2016.

154 Escobar, M. E. Procedimiento de limpieza de datos en el GESPRO. Máster en Gestión de Proyectos Informáticos, Universidad de las Ciencias Informáticas, Departamento de Investigaciones en Gestión de Proyectos, MSc. Surayne Torres López, La Habana, Cuba. 2016.

155 Wolpert, D. H. The supervised learning no-free-lunch theorems. In Soft computing and industry (pp. 25-42). Springer London, DOI: 10.1007/978-1-4471-0123-9_3, ISBN: 978-1-4471-1101-6. 2002.

156 Oltean, M. Searching for a practical evidence of the no free lunch theorems. In International Workshop on Biologically Inspired Approaches to Advanced Information Technology (pp. 472-483), DOI: 10.1007/978-3-540-27835-1_34, ISBN: 978-3-540-23339-8. Springer Berlin Heidelberg. 2004.

157 Wolpert, D. H., & Macready, W. G. No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82, ISSN: 1089-778X, DOI: 10.1109/4235.585893. 2002.

158 Bouguessa, M. A probabilistic combination approach to improve outlier detection. In Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on (Vol. 1, pp. 666-673), IEEE, DOI: 10.1109/ICTAI.2012.95, ISSN: 1082-3409, CD-ROM ISBN: 978-0-7695-4915-6.2012.

159Ruiz-Shulcloper, J. Reconocimiento lógico combinatorio de patrones: teoría y aplicaciones. Tesis en opción al grado científico de Doctor en Ciencias. Santa Clara, Centro de Investigaciones de Tecnologías de Avanzadas CENATAV.2009.

160Kaufman, L. and Rousseeuw, P.J. (=: “K&R(1990)”) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.1990.

161 Emad, A., Christopher, T., et al. Breast tumor classification using a new OWA operator, Expert Systems With Applications 61 page 302–313. 2016. Disponible en: www.elsevier.com/locate/eswa

162 Yager, R. R.On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on systems, Man, and Cybernetics, 18(1), 183-190, ISSN: 0018-9472, DOI: 10.1109/21.87068. 1988.

163 Merigó, J. New extensions to the OWA operators and its application in decision making. Department of Business Administration, University of Barcelona. Barcelona, University of Barcelona. PhD. 2008.

164 Johnsonbaugh, R. Matemática Discreta, 4ta edición, vol. 1, ISBN: 970-17-0253-0.Prentice Hall, Mexico. 1999.

165 Piñero, P, S. Torres, et al.Paquete de Herramientas para la Gestión de Proyectos, En Registro Centro Nacional de Registro de Derecho de Autor de Cuba, No Registro CENDA: 1540-2010, La Habana, Cuba. 2010.

166 Piñero, P, S. Lugo J. A, Menéndez J. et al. Solución de software XEDRO GESPRO v13.05, En Registro Centro Nacional de Registro de Derecho de Autor de Cuba, No Registro CENDA: 2336 -06-2015, La Habana, Cuba. DCN-002/2016. 2015.

167Merigó, J. M., & Yager, R. R.Norm aggregations and OWA operators. In Aggregation functions in theory and in practice, Volume 228 of the series Advances in Intelligent Systems and Computing, pp. 141-151, DOI: 10.1007/978-3-642-39165-1_17, ISBN: 978-3-642-39164-4. Springer Berlin Heidelberg. 2013.

168 Nielsen, T. D., & Jensen, F. V. Bayesian networks and decision graphs. Springer Science & Business Media, 2da edición, ISBN: 978-0-387-68282-2, páginas: XVI, 448. 2009.

169 Aho, A., Ullman, J. Data Structures and Algorithms: Pearson; 1st edition, ISBN: 978-0201000238, 427 pages. 1983.

170 Vattai, Z. A. FLOYD-warshall in Scheduling Open Networks. Procedia Engineering, 164, 106-114. 2016. Disponible en: http://dx.doi.org/10.1016/j.proeng.2016.11.598

171 Torres, S., Castro, G.F., et al.Rough Sets for Human Resource Competence Evaluation and Experiences. Applied Mathematics, 7, 1317-1325.2016. Disponible en: http://dx.doi.org/10.4236/am.2016.712116

172Castro, G.F., Pérez, I., et al.Platform for Project Evaluation Based on Soft-Computing Techniques. Springer International Publishing AG 2016. CITI 2016, CCIS 658, pp. 1–15. DOI: 10.1007/978-3-319-48024-4_18. Volume 658 of the series Communications in Computer and Information Science pp 226-240. 2016.

Disponible en: http://link.springer.com/chapter/10.1007/978-3-319-48024-4_18?no-access=true

173 Castro, G.F., Pérez, I., et al. Aplicación de la minería de datos anómalos en organizaciones orientadas a proyectos. Revista Cubana de Ciencias Informáticas Vol. 10, No. Especial UCIENCIA, ISSN: 2227-1899 | RNPS: 2301 http://rcci.uci.cu. Pág. 195-209. 2016. Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1456

174 Biblioteca R. Paquetes “Outliers” y “outlierD” para la minería de Outliers. http://cran.r-project.org/web/packages/Outliers/Outliers.pdf. 2015.

175Snedecor, G.W., Cochran, W.G. Statistical Methods (seventh edition). Iowa State University Press, Ames, Iowa. 1980.

176Castro, G. F., Pérez, I., et. at.PRODanalysis, un Sistema para el Aseguramiento de Ingresos Basado en Minería de Outliers. INNOVA Research Journal, Vol 1, No. 7, 18-36. ISSN 2477-9024. 2016. Disponible en: http://www.journaluidegye.com/magazine/index.php/innova/article/view/34

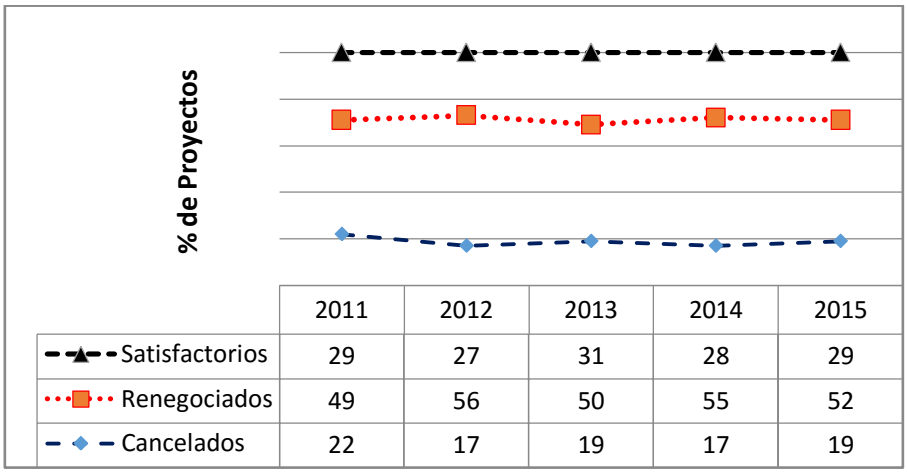
177Pérez,I., Piñero, P. Y. Repositorio de bases de datos para investigaciones en gestión de proyectos. Conferencia Científica Uciencia 2016, II Taller Internacional de Gestión de Proyectos, Panel Aplicaciones de la Inteligencia Artificial a la Gestión de Proyectos, ISBN: 978-959-286-054-4. 2016.

178Pérez,I. Propuesta de metodología para el diseño e implantación de repositorios de activos de software reutilizables. Maestría en Gestión de Proyectos Informáticos, Laboratorio de investigaciones en Gestión de Proyectos, Universidad de las Ciencias Informáticas. 2012.

179 QuitusServices, Portal corporativo compañíade servicios informáticos. Guayaquil-Ecuador, Last update [Enero 2017], Online: [Julio 2017], https://businessredmine.herokuapp.com/portal/quitusservices, 2017.

ANEXOS

Anexo 1.Evolución histórica de los proyectos, según *Standish group*



	2011	2012	2013	2014	2015
Satisfactorios	29	27	31	28	29
Renegociados	49	56	50	55	52
Cancelados	22	17	19	17	19

Figura 13. Evolución histórica de los proyectos, según *Standish group*.

Datos tomados de [11]

Anexo 2.Análisis de la mejor configuración de los algoritmosanalizados

Resultados del algoritmoAngle

Tabla 21.Algoritmo Angle.Resultados de comparación.

Grupo	*col_mix*	*alone_*	*mult_mix_*	*mult_plan_*	*mult_rate_*
a	angle_3_0.92 angle_3_0.95 angle_5_0.92 angle_5_0.95 angle_7_0.92 angle_7_0.95 angle_9_0.92 angle_9_0.95	angle_3_0.95 angle_5_0.95 angle_7_0.95 angle_9_0.95 angle_3_0.92	angle_5_0.95	angle_9_0.95 angle_7_0.95	angle_3_0.95 angle_5_0.95 angle_7_0.95
b		angle_5_0.92 angle_7_0.92 angle_9_0.92	angle_5_0.92 angle_7_0.95 angle_9_0.95 angle_7_0.92	angle_7_0.92 angle_9_0.92 angle_5_0.95	angle_3_0.92 angle_5_0.92 angle_9_0.95 angle_7_0.92 angle_9_0.92

			angle_3_0.95 angle_9_0.92 angle_3_0.92		
c				angle_5_0.92 angle_3_0.95	
d				angle_3_0.92	

Base de datos 'col_mix_'. Análisis de la variable: eficacia

Tabla 22. Algoritmo Angle, resultados de comparación sobre la base de datos *'col_mix'.*

	angle_3_0.95	angle_5_0.95	angle_5_0.92	angle_9_0.92	angle_3_0.92	angle_7_0.95	angle_7_0.92
angle_9_0.95	p-value = 0.9679	p-value = 0.5016	p-value = 0.6009	p-value = 0.3271	p-value = 0.5197	p-value = 0.9359	p-value = 0.4209
angle_3_0.95		p-value = 0.9405	p-value = 0.4939	p-value = 0.3547	p-value = 0.7225	p-value = 1	p-value = 0.4445
angle_5_0.95			p-value = 0.9039	p-value = 0.546	p-value = 0.6012	p-value = 0.7938	p-value = 1
angle_5_0.92				p-value = 0.8228	p-value = 0.8092	p-value = 0.8788	p-value = 0.6874
angle_9_0.92					p-value = 0.6292	p-value = 0.8519	p-value = 0.7089
angle_3_0.92						p-value = 0.9702	p-value = 0.4688
angle_7_0.95							p-value = 0.6791

Base de datos 'alone_'. Análisis de la variable: eficacia

Tabla 23. Algoritmo Angle, resultados de comparación sobre la base de datos *'alone'.*

	angle_5_0.95	angle_7_0.95	angle_9_0.95	angle_3_0.92	angle_5_0.92	angle_7_0.92	angle_9_0.92
angle_3_0.95	p-value = 0.1841	p-value = 0.06191	p-value = 0.09291	p-value = 0.126	p-value = 0.0004808	p-value = 0.001239	p-value = 0.00299
angle_5_0.		p-value =	p-value =	p-value =	p-value =	p-value =	p-value =

95		0.6291	0.2179	0.184	0.06193	0.0005144	0.01374
angle_7_0.95			p-value = 0.5461	p-value = 0.654	p-value = 0.02491	p-value = 0.03824	p-value = 0.005726
angle_9_0.95				p-value = 0.8813	p-value = 0.1453	p-value = 0.03822	p-value = 0.0196
angle_3_0.92					p-value = 0.09645	p-value = 0.05222	p-value = 0.08592
angle_5_0.92						p-value = 0.433	p-value = 0.2349
angle_7_0.92							p-value = 0.9702

Base de datos 'mult_mix'. Análisis de la variable: eficacia

Tabla 24. Algoritmo Angle, resultados de comparaciónsobre la base de datos *'mult_mix'.*

	angle_5_0.92	angle_7_0.95	angle_9_0.95	angle_7_0.92	angle_3_0.95	angle_9_0.92	angle_3_0.92
angle_5_0.95	p-value = 0.0002873	p-value = 0.4859	p-value = 0.658	p-value = 0.05257	p-value = 0.02977	p-value = 0.1474	p-value = 0.0438
angle_5_0.92		p-value = 0.9039	p-value = 0.9679	p-value = 0.2761	p-value = 0.06415	p-value = 0.2273	p-value = 0.07017
angle_7_0.95			p-value = 0.7652	p-value = 0.3271	p-value = 0.2049	p-value = 0.2959	p-value = 0.1004
angle_9_0.95				p-value = 0.8721	p-value = 0.1614	p-value = 0.2707	p-value = 0.1559
angle_7_0.92					p-value = 0.1989	p-value = 0.2772	p-value = 0.1354
angle_3_0.95						p-value = 0.3547	p-value = 0.6873
angle_9_0.92							p-value = 0.4688

Base de datos 'mult_plan'. Análisis de la variable: eficacia

Tabla 25. Algoritmo Angle, resultados de comparaciónsobre la Base de datos *'mult_plan'.*

	angle_7_0.95	angle_7_0.92	angle_9_0.92	angle_5_0.95	angle_5_0.92	angle_3_0.95	angle_3_0.92

angle_9_0.95	p-value = 0.6784	p-value = 0.1688	p-value = 0.007632	p-value = 0.2203	p-value = 0.1117	p-value = 0.1488	p-value = 0.1024
angle_7_0.95		p-value = 0.007686	p-value = 0.7987	p-value = 0.7331	p-value = 0.3635	p-value = 0.1626	p-value = 0.1119
angle_7_0.92			p-value = 0.4234	p-value = 0.6907	p-value = 1	p-value = 0.3318	p-value = 0.2668
angle_9_0.92				p-value = 0.6906	p-value = 0.3942	p-value = 0.1626	p-value = 0.1221
angle_5_0.95					p-value = 0.0006517	p-value = 0.5228	p-value = 0.4459
angle_5_0.92						p-value = 0.9058	p-value = 0.9811
angle_3_0.95							p-value = 0.0006467

Base de datos 'mult_rate'. Análisis de la variable: eficacia

Tabla 26. Algoritmo Angle, resultados de comparaciónsobre la base de datos *'mult_rate'.*

	angle_5_0.95	angle_7_0.95	angle_3_0.92	angle_5_0.92	angle_9_0.95	angle_7_0.92	angle_9_0.92
angle_3_0.95	p-value = 0.6148	p-value = 0.5062	p-value = 0.000188	p-value = 0.02387	p-value = 0.0979	p-value = 0.01235	p-value = 0.04187
angle_5_0.95		p-value = 0.08676	p-value = 0.04996	p-value = 0.0001312	p-value = 0.07759	p-value = 0.001687	p-value = 0.06195
angle_7_0.95			p-value = 0.1913	p-value = 0.04635	p-value = 0.3239	p-value = 0.009994	p-value = 0.09093
angle_3_0.92				p-value = 0.8562	p-value = 0.7652	p-value = 0.6274	p-value = 0.3411
angle_5_0.92					p-value = 0.7439	p-value = 0.3044	p-value = 0.2574
angle_9_0.95						p-value = 0.7936	p-value = 0.007159
angle_7_0.92							p-value = 0.5015

Resultados del algoritmo *Crossclustering*

Tabla 27.Algoritmo *Crossclustering.*Resultados de comparación.

Grupo	*col_mix*	*alone_*	*mult_mix_*	*mult_plan_*	*mult_rate_*
a	crossclustering_9_5 crossclustering_5_3 crossclustering_5_4 crossclustering_7_3 crossclustering_9_3 crossclustering_7_4 crossclustering_9_4	crossclustering_5_3 crossclustering_7_3 crossclustering_9_3 crossclustering_5_4	crossclustering_5_3 crossclustering_7_3 crossclustering_9_3	crossclustering_5_3 crossclustering_7_3 crossclustering_9_3	crossclustering_5_3 crossclustering_7_3 crossclustering_9_3 crossclustering_5_4 crossclustering_7_4 crossclustering_9_4
b		crossclustering_7_4 crossclustering_9_4 crossclustering_9_5	crossclustering_5_4 crossclustering_7_4 crossclustering_9_4	crossclustering_5_4 crossclustering_7_4 crossclustering_9_4 crossclustering_9_5	crossclustering_9_5
c			crossclustering_9_5		

Base de datos 'alone_'. Análisis de la variable: eficacia

Tabla 28. Algoritmo *Crossclustering*,resultados de comparación sobre la base de datos *'alone'.*

	crossclustering_5_4_0	crossclustering_7_3_0	crossclustering_7_4_0	crossclustering_9_3_0	crossclustering_9_4_0	crossclustering_9_5_0
crossclustering_5_3_0	p-value = 0.06735	p-value = 0.9687	p-value = 0.02194	p-value = 1	p-value = 0.01981	p-value = 0.004848
crossclustering_5_4_0		p-value = 0.03858	p-value = 0.3759	p-value = 0.05691	p-value = 0.4703	p-value = 0.2043

crossclusteri ng_7_3_0			p-value = 0.01132	p-value = 0.9249	p-value = 0.006491	p-value = 0.005491
crossclusteri ng_7_4_0				p-value = 0.01591	p-value = 0.9479	p-value = 0.6358
crossclusteri ng_9_3_0					p-value = 0.008361	p-value = 0.003286
crossclusteri ng_9_4_0						p-value = 0.5712

Base de datos 'col_mix_'. Análisis de la variable: eficacia

Tabla 29. Algoritmo *Crossclustering*,resultados de comparaciónsobre la base de datos *'col_mix'.*

	crossclusteri ng_5_4_0	crossclusteri ng_7_3_0	crossclusteri ng_7_4_0	crossclusteri ng_9_3_0	crossclusteri ng_9_4_0	crossclusteri ng_9_5_0
crossclusteri ng_5_3_0	p-value = 0.7958	p-value = 1	p-value = 0.6997	p-value = 0.9165	p-value = 0.8016	p-value = 0.8971
crossclusteri ng_5_4_0		p-value = 0.932	p-value = 0.8971	p-value = 0.9433	p-value = 0.8258	p-value = 0.5403
crossclusteri ng_7_3_0			p-value = 0.6231	p-value = 0.9811	p-value = 0.7007	p-value = 0.6832
crossclusteri ng_7_4_0				p-value = 0.7945	p-value = 1	p-value = 0.5345
crossclusteri ng_9_3_0					p-value = 0.8128	p-value = 0.6472
crossclusteri ng_9_4_0						p-value = 0.6413

Base de datos 'mult_mix'. Análisis de la variable: eficacia

Tabla 30. Algoritmo *Crossclustering*, resultados de comparaciónsobre la base de datos *'mult_mix'.*

	crossclusteri ng_5_4_0	crossclusteri ng_7_3_0	crossclusteri ng_7_4_0	crossclusteri ng_9_3_0	crossclusteri ng_9_4_0	crossclusteri ng_9_5_0
crossclusteri ng_5_3_0	p-value = 0.01796	p-value = NA	p-value = 0.01796	p-value = NA	p-value = 0.01796	p-value = 0.0009815

crossclusteri ng_5_4_0		p-value = 0.01796	p-value = NA	p-value = 0.01796	p-value = NA	p-value = 0.01796
crossclusteri ng_7_3_0			p-value = 0.01796	p-value = NA	p-value = 0.01796	p-value = 0.0009815
crossclusteri ng_7_4_0				p-value = 0.01796	p-value = NA	p-value = 0.01796
crossclusteri ng_9_3_0					p-value = 0.01796	p-value = 0.0009815
crossclusteri ng_9_4_0						p-value = 0.01796

Base de datos 'mult_plan'. Análisis de la variable: eficacia

Tabla 31. Algoritmo *Crossclustering*, resultados de comparaciónsobre la base de datos *'mult_plan'.*

	crossclusteri ng_5_4_0	crossclusteri ng_7_3_0	crossclusteri ng_7_4_0	crossclusteri ng_9_3_0	crossclusteri ng_9_4_0	crossclusteri ng_9_5_0
crossclusteri ng_5_3_0	p-value = 0.002218	p-value = NA	p-value = 0.002218	p-value = 0.8092	p-value = 0.0003385	p-value = 0.0001204
crossclusteri ng_5_4_0		p-value = 0.002218	p-value = NA	p-value = 0.0005167	p-value = 0.8813	p-value = 0.07314
crossclusteri ng_7_3_0			p-value = 0.002218	p-value = 0.8092	p-value = 0.0003385	p-value = 0.0001204
crossclusteri ng_7_4_0				p-value = 0.0005167	p-value = 0.8813	p-value = 0.07314
crossclusteri ng_9_3_0					p-value = 0.0002536	p-value = 0.0001032
crossclusteri ng_9_4_0						p-value = 0.02771

Base de datos 'mult_rate'. Análisis de la variable: eficacia

Tabla 32. Algoritmo *Crossclustering*, resultados de comparaciónsobre la base de datos *'mult_rate'.*

	crossclusteri ng_5_4_0	crossclusteri ng_7_3_0	crossclusteri ng_7_4_0	crossclusteri ng_9_3_0	crossclusteri ng_9_4_0	crossclusteri ng_9_5_0
crossclusteri ng_5_3_0	p-value = 0.1797	p-value = NA	p-value = 0.1621	p-value = 0.6274	p-value = 0.1621	p-value = 0.01173
crossclusteri		p-value =	p-value =	p-value =	p-value =	p-value =

ng_5_4_0		0.1797	0.6749	0.5016	0.6749	0.08936
crossclustering_7_3_0			p-value = 0.1621	p-value = 0.6274	p-value = 0.1621	p-value = 0.01173
crossclustering_7_4_0				p-value = 0.2321	p-value = NA	p-value = 0.2707
crossclustering_9_3_0					p-value = 0.2321	p-value = 0.0006241
crossclustering_9_4_0						p-value = 0.2707

Resultados del algoritmo *Distance_mahalanobis*

Tabla 33.Algoritmo *Distance_mahalanobis*. Resultados de comparación.

Grupo	***col_mix***	***alone_***	***mult_mix_***	***mult_rate_***
a	distance_mahalanobis_3_0.92 distance_mahalanobis_5_0.92 distance_mahalanobis_7_0.92 distance_mahalanobis_9_0.92 distance_mahalanobis_3_0.95 distance_mahalanobis_5_0.95 distance_mahalanobis_7_0.95 distance_mahalanobis_9_0.95	distance_mahalanobis_3_0.92 distance_mahalanobis_7_0.92 distance_mahalanobis_9_0.92 distance_mahalanobis_5_0.92	distance_mahalanobis_3_0.92 distance_mahalanobis_5_0.92 distance_mahalanobis_7_0.92 distance_mahalanobis_9_0.92	distance_mahalanobis_3_0.92 distance_mahalanobis_5_0.92 distance_mahalanobis_7_0.92 distance_mahalanobis_9_0.92
b		distance_mahalanobis_3_0.95 distance_mahalanobis_5_0.95 distance_mahalanobis_7_0.95 distance_mahalanobis	distance_mahalanobis_3_0.95 distance_mahalanobis_5_0.95 distance_mahalanobis_7_0.95 distance_mahalanobis	distance_mahalanobis_3_0.95 distance_mahalanobis_5_0.95 distance_mahalanobis_7_0.95 distance_mahalanobis

		_9_0.95	_9_0.95	_9_0.95

Base de datos 'alone_'. Análisis de la variable: eficacia

Tabla 34. Algoritmo *Mahalanobis*,resultados de comparaciónsobre labase de datos *'alone'.*

	distance_mahalanobis_3_0.95	distance_mahalanobis_5_0.92	distance_mahalanobis_5_0.95	distance_mahalanobis_7_0.92	distance_mahalanobis_7_0.95	distance_mahalanobis_9_0.92	distance_mahalanobis_9_0.95
distance_mahalanobis_3_0.92	p-value = 8.782e-05	p-value = 0.875	p-value = 8.782e-05	p-value = 0.9055	p-value = 8.807e-05	p-value = 0.831	p-value = 8.782e-05
distance_mahalanobis_3_0.95		p-value = 8.845e-05	p-value = 0.6009	p-value = 8.77e-05	p-value = 0.9198	p-value = 8.695e-05	p-value = 0.9652
distance_mahalanobis_5_0.92			p-value = 8.782e-05	p-value = 0.8197	p-value = 8.845e-05	p-value = 0.8969	p-value = 8.795e-05
distance_mahalanobis_5_0.95				p-value = 8.733e-05	p-value = 0.887	p-value = 8.695e-05	p-value = 0.4795
distance_mahalanobis_7_0.92					p-value = 8.832e-05	p-value = 0.9839	p-value = 8.782e-05
distance_mahalanobis_7_0.95						p-value = 8.77e-05	p-value = 0.9721
distance_mahalanobis_9_0.92							p-value = 8.795e-05

Base de datos 'col_mix_'. Análisis de la variable: eficacia

Tabla 35. Algoritmo *Mahalanobis*,resultados de comparaciónsobre la base de datos *'col_mix'.*

	distance_mahalanobis_3_0.95	distance_mahalanobis_5_0.92	distance_mahalanobis_5_0.95	distance_mahalanobis_7_0.92	distance_mahalanobis_7_0.95	distance_mahalanobis_9_0.92	distance_mahalanobis_9_0.95
distance_mahalanobis_3_0.92	p-value = 0.4962	p-value = 1	p-value = 0.4962	p-value = 1	p-value = 0.4962	p-value = 1	p-value = 0.4962
distance_m		p-value =	p-value = 1	p-value =	p-value = 1	p-value =	p-value = 1

ahalanobis_3_0.95		0.593		0.4962		0.4962	
distance_mahalanobis_5_0.92			p-value = 0.4962	p-value = 1	p-value = 0.4962	p-value = 1	p-value = 0.4962
distance_mahalanobis_5_0.95				p-value = 0.4962	p-value = 1	p-value = 0.1025	p-value = 1
distance_mahalanobis_7_0.92					p-value = 0.4962	p-value = 1	p-value = 0.4652
distance_mahalanobis_7_0.95						p-value = 0.4652	p-value = 1
distance_mahalanobis_9_0.92							p-value = 0.4962

Base de datos 'mult_mix'. Análisis de la variable: eficacia

Tabla 36. Algoritmo *Mahalanobis*, resultados de comparaciónsobre la base de datos *'mult_mix'.*

	distance_mahalanobis_3_0.95	distance_mahalanobis_5_0.92	distance_mahalanobis_5_0.95	distance_mahalanobis_7_0.92	distance_mahalanobis_7_0.95	distance_mahalanobis_9_0.92	distance_mahalanobis_9_0.95
distance_mahalanobis_3_0.92	p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05
distance_mahalanobis_3_0.95		p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05	p-value = NA
distance_mahalanobis_5_0.92			p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05b	p-value = NA	p-value = 8.82e-05
distance_mahalanobis_5_0.95				p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05	p-value = NA
distance_mahalanobis_7_0.92					p-value = 8.82e-05	p-value = NA	p-value = 8.82e-05

distance_mahalanobis_7_0.95						p-value = 8.82e-05	p-value = NA
distance_mahalanobis_9_0.92							p-value = 8.82e-05

Base de datos 'mult_rate'. Análisis de la variable: eficacia

Tabla 37. Algoritmo *Mahalanobis*, resultados de comparaciónsobre la base de datos *'mult_rate'.*

	distance_mahalanobis_3_0.95	distance_mahalanobis_5_0.92	distance_mahalanobis_5_0.95	distance_mahalanobis_7_0.92	distance_mahalanobis_7_0.95	distance_mahalanobis_9_0.92	distance_mahalanobis_9_0.95
distance_mahalanobis_3_0.92	p-value = 8.832e-05	p-value = NA	p-value = 8.832e-05	p-value = 0.9721	p-value = 8.658e-05	p-value = 0.7505	p-value = 8.658e-05
distance_mahalanobis_3_0.95		p-value = 8.832e-05	p-value = 0.9381	p-value = 8.845e-05	p-value = 0.9198	p-value = 8.658e-05	p-value = 0.9198
distance_mahalanobis_5_0.92			p-value = 8.832e-05	p-value = 0.9721	p-value = 8.658e-05	p-value = 0.7505	p-value = 8.658e-05
distance_mahalanobis_5_0.95				p-value = 8.82e-05	p-value = 0.9652	p-value = 8.795e-05	p-value = 0.9652
distance_mahalanobis_7_0.92					p-value = 8.795e-05	p-value = 0.9108	p-value = 8.795e-05
distance_mahalanobis_7_0.95						p-value = 8.795e-05	p-value = NA
distance_mahalanobis_9_0.92							p-value = 8.795e-05

Resultados del algoritmo *kmeans_euclidean*

Tabla 38.Algoritmo *kmeans_euclidean*. Resultados de comparación.

Grupo	***col_mix***	***alone_***	***mult_mix_***	***mult_plan_***	***mult_rate_***

a	kmeans_euclidean_3_0.95 kmeans_euclidean_9_0.92 kmeans_euclidean_9_0.95 kmeans_euclidean_3_0.92 kmeans_euclidean_7_0.92 kmeans_euclidean_5_0.92	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.92
b	kmeans_euclidean_7_0.95 kmeans_euclidean_5_0.95	kmeans_euclidean_7_0.92 kmeans_euclidean_9_0.95	kmeans_euclidean_7_0.92 kmeans_euclidean_9_0.95	kmeans_euclidean_9_0.95 kmeans_euclidean_7_0.92	kmeans_euclidean_7_0.92 kmeans_euclidean_9_0.95
c		kmeans_euclidean_7_0.95	kmeans_euclidean_5_0.92	kmeans_euclidean_5_0.92 kmeans_euclidean_7_0.95	kmeans_euclidean_7_0.95
d		kmeans_euclidean_5_0.9	kmeans_euclidean_5_0.95 kmeans_euclidean_3_0.92	kmeans_euclidean_5_0.95	kmeans_euclidean_5_0.92
e		kmeans_euclidean_5_0.95	kmeans_euclidean_3_0.95	kmeans_euclidean_3_0.92	kmeans_euclidean_5_0.95
f		kmeans_euclidean_3_0.92		kmeans_euclidean_3_0.95	kmeans_euclidean_3_0.92
g		kmeans_euclidean_3_0.95			kmeans_euclidean_3_0.95

Base de datos 'alone_'. Análisis de la variable: eficacia

Tabla 39. Algoritmo *kmeans_euclidean*,resultados de comparaciónsobre la base de datos *'alone'.*

	kmeans_euclidean_3_0.95	kmeans_euclidean_5_0.92	kmeans_euclidean_5_0.95	kmeans_euclidean_7_0.92	kmeans_euclidean_7_0.95	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.95
kmeans_eu	p-value =	p-value =	p-value =	p-value =	p-value =	p-value =	p-value =

clidean_3_0.92	8.82e-05	0.0004483	8.845e-05	0.0001201	8.845e-05	8.845e-05	0.0003376
kmeans_euclidean_3_0.95		p-value = 0.0001029	p-value = 8.845e-05	p-value = 0.0001031	p-value = 8.845e-05	p-value = 8.857e-05	p-value = 8.845e-05
kmeans_euclidean_5_0.92			p-value = 0.001502	p-value = 0.01112	p-value = 0.05681	p-value = 0.0001032	p-value = 0.1258
kmeans_euclidean_5_0.95				p-value = 0.0001396	p-value = 8.77e-05	p-value = 8.845e-05	p-value = 0.0003898
kmeans_euclidean_7_0.92					p-value = 0.01111	p-value = 0.001157	p-value = 0.6317
kmeans_euclidean_7_0.95						p-value = 0.0001026	p-value = 0.1671
kmeans_euclidean_9_0.92							p-value = 0.0008909

Base de datos 'col_mix_'. Análisis de la variable: eficacia

Tabla 40. Algoritmo *kmeans_euclidean*,resultados de comparaciónsobre la base de datos *'col_mix'*.

	kmeans_euclidean_3_0.95	kmeans_euclidean_5_0.92	kmeans_euclidean_5_0.95	kmeans_euclidean_7_0.92	kmeans_euclidean_7_0.95	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.95
kmeans_euclidean_3_0.92	p-value = 0.6496	p-value = 0.9039	p-value = 0.5564	p-value = 0.8446	p-value = 0.8092	p-value = 0.7583	p-value = 0.8446
kmeans_euclidean_3_0.95		p-value = 0.5277	p-value = 0.2146	p-value = 0.4925	p-value = 0.3957	p-value = 0.5227	p-value = 0.5277
kmeans_euclidean_5_0.92			p-value = 0.1989	p-value = 0.2461	p-value = 0.9547	p-value = 0.1841	p-value = 0.1841
kmeans_euclidean_5_0.95				p-value = 0.06415	p-value = 0.2145	p-value = 0.05341	p-value = 0.04862

kmeans_euclidean_7_0.92					p-value = 0.1978	p-value = 0.2145	p-value = 0.5566
kmeans_euclidean_7_0.95						p-value = 0.04862	p-value = 0.1978
kmeans_euclidean_9_0.92							p-value = 0.286

Base de datos 'mult_mix'. Análisis de la variable: eficacia

Tabla 41. Algoritmo *kmeans_euclidean*, resultados de comparaciónsobre la base de datos *'mult_mix'.*

	kmeans_euclidean_3_0.95	kmeans_euclidean_5_0.92	kmeans_euclidean_5_0.95	kmeans_euclidean_7_0.92	kmeans_euclidean_7_0.95	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.95
kmeans_euclidean_3_0.92	p-value = 8.832e-05	p-value = 8.795e-05	p-value = 0.07301	p-value = 8.832e-05	p-value = 8.82e-05	p-value = 8.832e-05	p-value = 0.0002188
kmeans_euclidean_3_0.95		p-value = 8.857e-05	p-value = 0.00642	p-value = 8.845e-05	p-value = 8.832e-05	p-value = 8.845e-05	p-value = 0.0001398
kmeans_euclidean_5_0.92			p-value = 8.845e-05	p-value = 8.795e-05	p-value = 0.002932	p-value = 0.0001033	p-value = 0.0477
kmeans_euclidean_5_0.95				p-value = 8.832e-05	p-value = 8.845e-05	p-value = 8.845e-05	p-value = 0.001019
kmeans_euclidean_7_0.92					p-value = 8.807e-05	p-value = 0.001017	p-value = 0.7275
kmeans_euclidean_7_0.95						p-value = 0.0001032	p-value = 0.07314
kmeans_euclidean_9_0.92							p-value = 0.0004493

Base de datos 'mult_plan'. Análisis de la variable: eficacia

Tabla 42. Algoritmo *kmeans_euclidean*, resultados de comparaciónsobre la base de datos *'mult_plan'.*

	kmeans_euclidean_3_0.95	kmeans_euclidean_5_0.92	kmeans_euclidean_5_0.95	kmeans_euclidean_7_0.92	kmeans_euclidean_7_0.95	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.95
kmeans_euclidean_3_0.92	p-value = 8.832e-05	p-value = 8.832e-05	p-value = 0.001506	p-value = 8.845e-05	p-value = 0.0001032	p-value = 8.845e-05	p-value = 8.832e-05
kmeans_euclidean_3_0.95		p-value = 8.857e-05	p-value = 8.832e-05	p-value = 8.857e-05	p-value = 0.0001032	p-value = 8.845e-05	p-value = 8.857e-05
kmeans_euclidean_5_0.92			p-value = 8.857e-05	p-value = 8.807e-05	p-value = 0.9748	p-value = 0.0001201	p-value = 8.857e-05
kmeans_euclidean_5_0.95				p-value = 8.845e-05	p-value = 0.0007779	p-value = 8.82e-05	p-value = 8.857e-05
kmeans_euclidean_7_0.92					p-value = 8.807e-05	p-value = 0.0003372	p-value = 0.08285
kmeans_euclidean_7_0.95						p-value = 0.0001026	p-value = 8.82e-05
kmeans_euclidean_9_0.92							p-value = 0.0005167

Base de datos 'mult_rate'. Análisis de la variable: eficacia

Tabla 43. Algoritmo *kmeans_euclidean*, resultados de comparaciónsobre la base de datos *'mult_rate'.*

	kmeans_euclidean_3_0.95	kmeans_euclidean_5_0.92	kmeans_euclidean_5_0.95	kmeans_euclidean_7_0.92	kmeans_euclidean_7_0.95	kmeans_euclidean_9_0.92	kmeans_euclidean_9_0.95
kmeans_euclidean_3_0.92	p-value = 0.0001203	p-value = 0.0001398	p-value = 0.0001399	p-value = 0.0002137	p-value = 8.832e-05	p-value = 0.0001399	p-value = 0.0001888
kmeans_euclidean_3_0.95		p-value = 0.0001033	p-value = 8.832e-05	p-value = 8.845e-05	p-value = 8.832e-05	p-value = 0.0001033	p-value = 0.0001399
kmeans_euclidean_5_			p-value = 0.002202	p-value = 0.004271	p-value = 0.004045	p-value = 0.001323	p-value = 0.02387

0.92							
kmeans_euclidean_5_0.95				p-value = 0.0002191	p-value = 0.0001031	p-value = 0.0002185	p-value = 0.0006799
kmeans_euclidean_7_0.92					p-value = 0.05206	p-value = 0.008962	p-value = 0.1116
kmeans_euclidean_7_0.95						p-value = 0.001941	p-value = 0.05214
kmeans_euclidean_9_0.92							p-value = 0.03334

Resultados del algoritmo *kmeans_norm_euclidean*

Tabla 44.Algoritmo *kmeans_norm_euclidean*.Resultados de comparación.

Grupo	*col_mix*	*alone_*	*mult_mix_*	*mult_plan_*	*mult_rate_*
a	kmeans_norm_euclidean_3_0.92 kmeans_norm_euclidean_3_0.95 kmeans_norm_euclidean_9_0.92 kmeans_norm_euclidean_7_0.92 kmeans_norm_euclidean_5_0.92 kmeans_norm_euclidean_9_0.95 kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.92
b	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_7_0.92 kmeans_norm_euclidean_9_0.95	kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.92kmeans_norm_euclidean_9_0.95
c		kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_5_0.92

		kmeans_norm_euclidean_3_0.92	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_9_0.95	kmeans_norm_euclidean_7_0.95
d		kmeans_norm_euclidean_9_0.95 kmeans_norm_euclidean_7_0.95 kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_5_0.95
e		kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_3_0.92kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_3_0.92 kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_3_0.92
f				kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_3_0.95

Base de datos 'alone_'. Análisis de la variable: eficacia

Tabla 45. Algoritmo *kmeans_norm_euclidean*,comparaciónsobre la base de datos *'alone'.*

	kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.95
kmeans_norm_euclidean_3_0.92	p-value = 8.845e-05	p-value = 8.832e-05	p-value = 8.857e-05	p-value = 0.1671	p-value = 8.832e-05	p-value = 0.01688	p-value = 0.0001031
kmeans_norm_euclidean_3_0.95		p-value = 8.857e-05	p-value = 0.0275	p-value = 0.0001031	p-value = 0.02508	p-value = 0.000189	p-value = 0.0006719
kmeans_norm_euclidean_5_0.92			p-value = 8.857e-05	p-value = 0.6012	p-value = 8.857e-05	p-value = 0.5257	p-value = 8.857e-05
kmeans_norm_euclidean_5_0.95				p-value = 0.0001401	p-value = 0.3506	p-value = 0.0002185	p-value = 0.001503
kmeans_norm_euclidean_7_0.92					p-value = 0.000189	p-value = 0.6149	p-value = 0.0001629
kmeans_norm_euclidean_7_0.95						p-value = 0.0003385	p-value = 0.02688

kmeans_norm_euclidean_9_0.92							p-value = 0.0007788

Base de datos 'col_mix_'. Análisis de la variable: eficacia

Tabla 46. Algoritmo *kmeans_norm_euclidean*,comparaciónsobre la base de datos *'col_mix'.*

	kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.95
kmeans_norm_euclidean_3_0.92	p-value = 0.4769	p-value = 0.193	p-value = 0.1117	p-value = 0.07873	p-value = 0.02168	p-value = 0.2432	p-value = 0.06735
kmeans_norm_euclidean_3_0.95		p-value = 0.1401	p-value = 0.3258	p-value = 0.2461	p-value = 0.02772	p-value = 0.6292	p-value = 0.04947
kmeans_norm_euclidean_5_0.92			p-value = 0.8092	p-value = 0.7771	p-value = 0.5349	p-value = 0.4939	p-value = 0.8228
kmeans_norm_euclidean_5_0.95				p-value = 0.446	p-value = 0.9039	p-value = 0.3271	p-value = 0.6791
kmeans_norm_euclidean_7_0.92					p-value = 0.157	p-value = 0.6292	p-value = 0.896
kmeans_norm_euclidean_7_0.95						p-value = 0.06415	p-value = 0.218
kmeans_norm_euclidean_9_0.92							p-value = 0.5503

Base de datos 'mult_mix'. Análisis de la variable: eficacia

Tabla 47. Algoritmo *kmeans_norm_euclidean*, comparaciónsobre la base de datos *'mult_mix'.*

	kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.95
kmeans_norm_euclidean_3_0.92	p-value = 0.06195	p-value = 8.857e-05	p-value = 0.0002702	p-value = 0.0001033	p-value = 0.0002191	p-value = 0.0001033	p-value = 0.0001204

kmeans_norm_euclidean_3_0.95		p-value = 8.857e-05	p-value = 8.857e-05	p-value = 8.857e-05	p-value = 0.0001032	p-value = 8.857e-05	p-value = 0.0001399
kmeans_norm_euclidean_5_0.92			p-value = 0.01409	p-value = 0.001604	p-value = 0.1353	p-value = 0.0001629	p-value = 0.002821
kmeans_norm_euclidean_5_0.95				p-value = 0.0003381	p-value = 0.007184	p-value = 0.0001033	p-value = 0.0007796
kmeans_norm_euclidean_7_0.92					p-value = 0.008956	p-value = 0.0004493	p-value = 0.8405
kmeans_norm_euclidean_7_0.95						p-value = 0.001018	p-value = 0.02063
kmeans_norm_euclidean_9_0.92							p-value = 0.003592

Base de datos 'mult_plan'. Análisis de la variable: eficacia

Tabla 48. Algoritmo *kmeans_norm_euclidean*, comparaciónsobre la base de datos *'mult_plan'.*

	kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.95
kmeans_norm_euclidean_3_0.92	p-value = 8.845e-05	p-value = 8.845e-05	p-value = 0.9108	p-value = 0.0001033	p-value = 0.0001401	p-value = 8.832e-05	p-value = 0.008962
kmeans_norm_euclidean_3_0.95		p-value = 8.857e-05	p-value = 0.0001629	p-value = 8.845e-05	p-value = 8.857e-05	p-value = 8.845e-05	p-value = 0.0002354
kmeans_norm_euclidean_5_0.92			p-value = 8.82e-05	p-value = 0.001604	p-value = 0.0004483	p-value = 0.0001032	p-value = 0.05592
kmeans_norm_euclidean_5_0.95				p-value = 0.0001032	p-value = 0.0001033	p-value = 8.857e-05	p-value = 0.00338
kmeans_norm_euclide					p-value = 0.00013	p-value = 0.00034	p-value = 0.00020

an_7_0.92							
kmeans_norm_euclidean_7_0.95						p-value = 8.82e-05	p-value = 0.030
kmeans_norm_euclidean_9_0.92							p-value = 0.0001203

Base de datos 'mult_rate'. Análisis de la variable: eficacia

Tabla 49. Algoritmo *kmeans_norm_euclidean*, comparaciónsobre la base de datos *'mult_rate'.*

	kmeans_norm_euclidean_3_0.95	kmeans_norm_euclidean_5_0.92	kmeans_norm_euclidean_5_0.95	kmeans_norm_euclidean_7_0.92	kmeans_norm_euclidean_7_0.95	kmeans_norm_euclidean_9_0.92	kmeans_norm_euclidean_9_0.95
kmeans_norm_euclidean_3_0.92	p-value = 0.0126	p-value = 8.857e-05	p-value = 0.01236	p-value = 0.00014	p-value = 0.000219	p-value = 8.857e-05	8.857e-05
kmeans_norm_euclidean_3_0.95		p-value = 0.0001033	p-value = 0.0003902	p-value = 8.845e-05	p-value = 0.0001032	p-value = 8.857e-05	p-value = 8.832e-05
kmeans_norm_euclidean_5_0.92			p-value = 0.000293	p-value = 0.02276	p-value = 0.2043	p-value = 0.000188	p-value = 0.000103
kmeans_norm_euclidean_5_0.95				p-value = 0.0002536	p-value = 0.0007788	p-value = 8.832e-05	p-value = 8.857e-05
kmeans_norm_euclidean_7_0.92					p-value = 0.009996	p-value = 8.845e-05	p-value = 0.3905
kmeans_norm_euclidean_7_0.95						p-value = 0.0001029	p-value = 8.832e-05
kmeans_norm_euclidean_9_0.92							p-value = 0.0005167

Resultados del algoritmo *kmeans_stats*

Tabla 50.Algoritmo *kmeans_stats.* Resultados de la comparación.

Grupo	***col_mix***	***alone_***	***mult_mix_***	***mult_plan_***	***mult_rate_***

a	kmeans_stats_5	kmeans_stats_3	kmeans_stats_3	kmeans_stats_3	kmeans_stats_3
	kmeans_stats_7	kmeans_stats_5	kmeans_stats_5	kmeans_stats_9	kmeans_stats_5
	kmeans_stats_9	kmeans_stats_9	kmeans_stats_7	kmeans_stats_5	kmeans_stats_7
	kmeans_stats_3	kmeans_stats_7	kmeans_stats_9	kmeans_stats_7	kmeans_stats_9

* Usando el test de Wilcoxon no se encontraron diferencias significativas.

Resultados del algoritmo *kmodr*

Tabla 51.Algoritmo *kmodr. R*esultados de la comparación.

Grupo	***col_mix***	***alone_***	***mult_mix_***	***mult_plan_***	***mult_rate_***
a	kmodr_9	kmodr_9	kmodr_3	kmodr_7	kmodr_3
	kmod_5	kmod_3	kmod_5	kmod_3	kmod_5
	kmodr_7	kmodr_7	kmodr_9	kmodr_5	$kmodr_7$
	kmodr_3	kmodr_5	kmodr_7	kmodr_9	kmodr_9

* Usando el test de Wilcoxon no se encontraron diferencias significativas

Resultados del algoritmo *Combine_outlier*

Tabla 52.Algoritmo *Combine_outlier*. Resultados de la comparación.

Grupo	***col_mix***	***alone_***	***mult_mix_***	***mult_plan_***	***mult_rate_***
a	combine_outlier_ 0.92 combine_outlier_ 0.95	combine_outlier_ 0.92 combine_outlier_ 0.95	combine_outlier_ _0.92 combine_outlier_ 0.95	combine_outlier_ _0.92 combine_outlier_ 0., 95	combine_outlier_ _0.92 combine_outlier_ 0.95

* Usando el test de Wilcoxon, no se encontraron diferencias significativas

Anexo 3.Análisis comparación de los algoritmos respecto a la eficacia

Base de datos alone. Análisis de la variable: eficacia

Tabla 53.Comparación de múltiples algoritmos, sobre la base de datos *alone.*

	kmeans_stats_3_0	distance_ mahalanobis_3_0.92	kmodr_3_0	angle_5_0.95	crosscluste ring_5_3_0	kmeans_n orm_euclid ean_9_0.9 2	kmeans_e uclidean_9 _0.92
combine_o utlier_0_0. 92	p-value = NA	p-value = 3.117e-08	p-value = 6.1e-09	p-value = 3.514e-08	p-value = 8.208e-06	p-value = 3.524e-08	p-value = 3.484e-08

kmeans_stats_3_0		p-value = 8.306e-05	p-value = 6.1e-09	p-value = 8.82e-05	p-value = 0.001474	p-value = 8.832e-05	p-value = 8.782e-05
distance_mahalanobis_3_0.92			p-value = 0.4592	p-value = 0.0002535	p-value = 0.004042	p-value = 8.845e-05	p-value = 8.857e-05
kmodr_3_0				p-value = 0.4596	p-value = 0.7564	p-value = 0.2368	p-value = 2.423e-06
angle_5_0.95					p-value = 0.004045	p-value = 8.845e-05	p-value = 8.845e-05
crosscluster ing_5_3_0						p-value = 0.06195	p-value = 8.857e-05
kmeans_norm_euclidean_9_0.92							p-value = 0.001412

Base de datos col_mix. Análisis de la variable: eficacia

Tabla 54.Comparación de múltiples algoritmos, sobre la base de datos *col_mix.*

	kmeans_stats_3_0	distance_mahalanobis_3_0.92	kmodr_3_0	angle_5_0.95	crosscluster ing_5_3_0	kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92
combine_outlier_0_0.92	p-value = 0.8316	p-value = 0.6729	p-value = 0.1389	p-value = 3.114e-05	p-value = 9.78e-05	p-value = 1.54e-06	p-value = 7.391e-07
kmeans_stats_3_0		p-value = 1	p-value = 0.207	p-value = 0.004847	p-value = 0.0146	p-value = 0.002225	p-value = 0.004847
distance_mahalanobis_3_0.92			p-value = 0.03977	p-value = 0.02768	p-value = 0.007394	p-value = 0.000327	p-value = 0.0007748
kmodr_3_0				p-value = 0.09264	p-value = 0.01711	p-value = 0.002876	p-value = 0.0007838
angle_5_0.95					p-value = 0.1074	p-value = 0.001325	p-value = 0.005491
crosscluster ing_5_3_0						p-value = 0.5461	p-value = 0.2598

kmeans_norm_euclidean_9_0.92							p-value = 0.2432

Base de datos mult_mix. Análisis de la variable: eficacia

Tabla 55.Comparación de múltiples algoritmos, sobre la base de datos *mult_mix.*

	kmeans_stats_3_0	distance_mahalanobis_3_0.92	angle_5_0.95	kmodr_3_0	kmeans_norm_euclidean_9_0.92	crossclustering_5_3_0	kmeans_euclidean_9_0.92
combine_outlier_0_0.92	p-value = NA	p-value = 3.377e-08	p-value = 1.656e-07	p-value = 3.968e-09	p-value = 3.514e-08	p-value = 3.534e-08	p-value = 3.524e-08
kmeans_stats_3_0		p-value = 8.646e-05	p-value = 0.0001958	p-value = 3.968e-09	p-value = 8.82e-05	p-value = 8.845e-05	p-value = 8.832e-05
distance_mahalanobis_3_0.92			p-value = 0.7652	p-value = 0.1124	p-value = 8.845e-05	p-value = 8.845e-05	p-value = 8.857e-05
angle_5_0.95				p-value = 0.1321	p-value = 0.0003385	p-value = 0.0003385	p-value = 8.857e-05
kmodr_3_0					p-value = 0.06547	p-value = 0.08777	p-value = 4.946e-07
kmeans_norm_euclidean_9_0.92						p-value = 0.2471	p-value = 0.0001031
crossclustering_5_3_0							p-value = 8.857e-05

Base de datos mult_plan. Análisis de la variable: eficacia

Tabla 56.Comparación de múltiples algoritmos, sobre labase de datos *mult_plan.*

	kmeans_stats_3_0	kmodr_3_0	angle_5_0.95	crossclustering_5_3_0	kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92
combine_outlier_0_0.92	p-value = NA	p-value = 2.376e-09	p-value = 1.684e-06	p-value = 3.534e-08	p-value = 3.524e-08	p-value = 3.484e-08

kmeans_stats_3_0		p-value = 2.376e-09	p-value = 0.0006484	p-value = 8.845e-05	p-value = 8.832e-05	p-value = 8.782e-05
kmodr_3_0			p-value = 0.1309	p-value = 0.007466	p-value = 4.619e-05	p-value = 1.689e-07
angle_5_0.95				p-value = 0.0001033	p-value = 8.845e-05	p-value = 8.857e-05
crossclustering_5_3_0					p-value = 0.000189	p-value = 8.857e-05
kmeans_norm_euclidean_9_0.92						p-value = 0.0001509

Base de datos mult_rate. Análisis de la variable: eficacia

Tabla 57.Comparación de múltiples algoritmos, sobre la base de datos *mult_rate.*

	kmeans_stats_3_0	kmodr_3_0	angle_5_0.95	distance_mahalanobis_3_0.92	crosscluster ing_5_3_0	kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92
combine_outlier_0_0.92	p-value = NA	p-value = 2.376e-09	p-value = 7.421e-08	p-value = 3.455e-08	p-value = 1.65e-07	p-value = 3.504e-08	p-value = 3.524e-08
kmeans_stats_3_0		p-value = 2.376e-09	p-value = 0.0001296	p-value = 8.745e-05	p-value = 0.0001954	p-value = 8.807e-05	p-value = 8.832e-05
kmodr_3_0			p-value = 0.02499	p-value = 0.01282	p-value = 0.004945	p-value = 0.0001493	p-value = 1.693e-07
angle_5_0.95				p-value = 0.3045	p-value = 0.0001401	p-value = 8.857e-05	p-value = 8.857e-05
distance_mahalanobis_3_0.92					p-value = 0.0001401	p-value = 8.857e-05	p-value = 8.845e-05
crosscluster ing_5_3_0						p-value = 0.04014	p-value = 8.857e-05
kmeans_norm_euclidean_9_0.92							p-value = 0.001014

Anexo 4.Análisis comparación de los algoritmos respecto a la eficiencia

Base de datos alone. Análisis de la variable: eficiencia (tiempo)

Tabla 58.Comparación de múltiples algoritmos, sobre la base de datos *alone.*

	kmodr_3_0	kmeans_stats_3_0	kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92	distance_mahalanobis_3_0.92	crossclustering_5_3_0	angle_5_0.95
combine_outlier_0_0.92	p-value = 3.545e-08	p-value = 3.498e-08	p-value = 3.557e-08	p-value = 3.522e-08	p-value = 3.54e-08	p-value = 3.564e-08	p-value = 3.561e-08
kmodr_3_0		p-value = 3.524e-08	p-value = 3.566e-08	p-value = 3.519e-08	p-value = 3.535e-08	p-value = 3.566e-08	p-value = 3.567e-08
kmeans_stats_3_0			p-value = 8.782e-05	p-value = 0.001437	p-value = 0.6277	p-value = 8.82e-05	p-value = 8.845e-05
kmeans_norm_euclidean_9_0.92				p-value = 8.782e-05	p-value = 8.857e-05	p-value = 8.845e-05	p-value = 8.857e-05
kmeans_euclidean_9_0.92					p-value = 0.01218	p-value = 8.845e-05	p-value = 8.857e-05
distance_mahalanobis_3_0.92						p-value = 8.832e-05	p-value = 8.857e-05
crossclustering_5_3_0							p-value = 8.857e-05

Base de datos col_mix. Análisis de la variable: eficiencia (tiempo)

Tabla 59.Comparación de múltiples algoritmos, sobre la base de datos *col_mix.*

	kmodr_3_0	kmeans_stats_3_0	kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92	distance_mahalanobis_3_0.92	crossclustering_5_3_0	angle_5_0.95
combine_outlier_0_0.92	p-value = 3.567e-08	p-value = 3.558e-08	p-value = 3.563e-08	p-value = 3.559e-08	p-value = 3.551e-08	p-value = 3.568e-08	p-value = 3.568e-08
kmodr_3_0		p-value =	p-value =	p-value =	p-value =	p-value =	p-value =

		3.566e-08	3.566e-08	3.568e-08	3.566e-08	3.568e-08	3.568e-08
kmeans_stats_3_0			p-value = 8.857e-05	p-value = 0.1254	p-value = 0.002074	p-value = 8.857e-05	p-value = 8.857e-05
kmeans_norm_euclidean_9_0.92				p-value = 8.82e-05	p-value = 8.807e-05	p-value = 8.857e-05	p-value = 8.857e-05
kmeans_euclidean_9_0.92					p-value = 0.0001102	p-value = 8.857e-05	p-value = 8.857e-05
distance_mahalanobis_3_0.92						p-value = 8.857e-05	p-value = 8.857e-05
crossclustering_5_3_0							p-value = 8.857e-05

Base de datos mult_mix.Análisis de la variable: eficiencia (tiempo)

Tabla 60.Comparación de múltiples algoritmos, sobre la base de datos *mult_mix.*

	kmodr_3_0	kmeans_stats_3_0	kmeans_norm_euclidean_9_0.92	kmeans_euclidean_9_0.92	distance_mahalanobis_3_0.92	crossclustering_5_3_0	angle_5_0.95
combine_outlier_0_0.92	p-value = 3.549e-08	p-value = 3.466e-08	p-value = 3.561e-08	p-value = 3.509e-08	p-value = 3.52e-08	p-value = 3.557e-08	p-value = 3.553e-08
kmodr_3_0		p-value = 3.487e-08	p-value = 3.551e-08	p-value = 3.498e-08	p-value = 3.543e-08	p-value = 3.547e-08	p-value = 3.563e-08
kmeans_stats_3_0			p-value = 8.832e-05	p-value = 8.128e-05	p-value = 0.0009871	p-value = 8.832e-05	p-value = 8.82e-05
kmeans_norm_euclidean_9_0.92				p-value = 8.82e-05	p-value = 8.845e-05	p-value = 8.857e-05	p-value = 8.845e-05
kmeans_euclidean_9_0.92					p-value = 0.7097	p-value = 8.807e-05	p-value = 8.845e-05
distance_mahalanobis						p-value = 8.845e-05	p-value = 8.845e-05

_3_0.92							
crosscluste ring_5_3_0							p-value = 8.857e-05

Base de datos mult_plan. Análisis de la variable: eficiencia (tiempo)

Tabla 61.Comparación de múltiples algoritmos, sobre la base de datos *mult_plan.*

	kmodr_3_0	kmeans_stat s_3_0	kmeans_nor m_euclidean _9_0.92	kmeans_eucl idean_9_0.9 2	crossclusteri ng_5_3_0	angle_5_0.9 5
combine_out lier_0_0.92	p-value = 3.534e-08	p-value = 3.212e-08	p-value = 3.553e-08	p-value = 3.348e-08	p-value = 3.547e-08	p-value = 3.563e-08
kmodr_3_0		p-value = 3.432e-08	p-value = 3.563e-08	p-value = 3.523e-08	p-value = 3.559e-08	p-value = 3.562e-08
kmeans_stat s_3_0			p-value = 8.832e-05	p-value = 0.00012	p-value = 8.857e-05	p-value = 8.832e-05
kmeans_nor m_euclidean _9_0.92				p-value = 8.857e-05	p-value = 8.845e-05	p-value = 8.857e-05
kmeans_eucl idean_9_0.9 2					p-value = 8.845e-05	p-value = 8.845e-05
crossclusteri ng_5_3_0						p-value = 8.845e-05

Base de datos mult_rate. Análisis de la variable: eficiencia (tiempo)

Tabla 62.Comparación de múltiples algoritmos, sobre la base de datos *mult_rate.*

	kmodr_3_0	kmeans_st ats_3_0	kmeans_n orm_euclid ean_9_0.9 2	kmeans_e uclidean_9 _0.92	distance_m ahalanobis _3_0.92	crosscluste ring_5_3_0	angle_5_0. 95
combine_o utlier_0_0. 92	p-value = 3.561e-08	p-value = 3.554e-08	p-value = 3.566e-08	p-value = 3.558e-08	p-value = 3.53e-08	p-value = 3.566e-08	p-value = 3.568e-08
kmodr_3_0		p-value = 3.563e-08	p-value = 3.564e-08	p-value = 3.564e-08	p-value = 3.562e-08	p-value = 3.568e-08	p-value = 3.568e-08

kmeans_stats_3_0			p-value = 8.845e-05	p-value = 0.0003573	p-value = 0.003557	p-value = 8.845e-05	p-value = 8.857e-05
kmeans_norm_euclidean_9_0.92				p-value = 8.845e-05	p-value = 8.845e-05	p-value = 8.857e-05	p-value = 8.857e-05
kmeans_euclidean_9_0.92					p-value = 0.0001145	p-value = 8.857e-05	p-value = 8.857e-05
distance_mahalanobis_3_0.92						p-value = 8.857e-05	p-value = 8.857e-05
crosscluster ing_5_3_0							p-value = 0.0001204

Anexo 5. Descripción de los proyectos usados en laevaluación de riesgos

Proyecto 1: Proyecto asociado al montaje de una plataforma para la gestión de proyectos basada en tecnología libre en una empresa con más de 600 trabajadores. Cliente ubicado cerca de la entidad desarrolladora y el equipo competente.

Proyecto 2: Proyecto asociado al montaje de una plataforma para la gestión de proyectos en una entidad dedicada a la formación y al desarrollo de productos informáticos, con media presión externa; equipo de desarrollo con competencias adecuadas. Cliente ubicado geográficamente cerca de la entidad desarrolladora.

Proyecto 3 Proyecto asociado al montaje de un programa de formación en una variante a distancia que depende de la existencia de un alto nivel profesional. Presión media y ubicación de los despliegues cercana al equipo de desarrollo.

Proyecto 4 Proyecto asociado al montaje de un programa de formación en una variante presencial depende de la existencia de un alto nivel profesional. Este proyecto se encuentra en la misma área geográfica donde están ubicados los clientes.

Proyecto 5 Proyecto dedicado al desarrollo de una plataforma basada en servicios, pensada para la integración de soluciones desarrolladas en múltiples plataformas, con media presión externa, donde se requiere alto nivel de especialización del personal; a pesar de ello ni los líderes ni los miembros del equipo habían desarrollado un proyecto similar. El cliente está ubicado geográficamente cerca de la entidad desarrolladora.

Proyecto 6 Proyecto dedicado al desarrollo de una plataforma para la recuperación de datos de sistemas de información y la generación de reportes dinámicos,con media presión externa, equipo de desarrollo capacitado.

Proyecto 7 Proyecto dedicado al desarrollo de una plataforma para la generación automática de libros electrónicos, poca presión externa, equipo competente. El cliente está ubicado geográficamente en el mismo lugar de la entidad desarrolladora.

Proyecto 8 Proyecto dedicado al desarrollo de una plataforma para el diagnóstico y el tratamiento de enfermedades de presión arterial, con tecnología propietaria, con media presión externa, equipo de desarrollo medianamente capacitado. Cliente está ubicado geográficamente en la misma ciudad donde se encuentra la entidad desarrolladora.

Proyecto 9 Proyecto dedicado al desarrollo de una plataforma para informatizar el sistema de gestión de datos estadísticos, con media presión externa, equipo de desarrollo medianamente capacitado. Cliente está ubicado geográficamente a distancia media donde se encuentra la entidad desarrolladora.

Proyecto 10 Proyecto dedicado al desarrollo de una plataforma para informatizar el portal web de una oficina estadística,con media presión externa, equipo de desarrollo medianamente capacitado. El cliente está ubicado geográficamente a distancia media donde se encuentra la entidad desarrolladora.

Proyecto 11 Proyecto dedicado al desarrollo de una plataforma para la gestión de la información estadística, con media presión externa, y equipo con competencias medias. Cliente está ubicado a distancia media del equipo de proyecto.

Proyecto 12 Proyecto dedicado al desarrollo de una plataforma para la captura, adquisición y procesamiento de datos de plantas procesadoras de gas, una elevada presión externa, equipo de desarrollo medianamente capacitado. Cliente está ubicado geográficamente muy distante al lugar donde se encuentra la entidad desarrolladora.

Proyecto 13 Proyecto dedicado al desarrollo de una plataforma para la captura, adquisición y procesamiento de datos de industrias de petróleo, con tecnología libre, una elevada presión externa, equipo de desarrollo medianamente capacitado, conformado por aproximadamente 50 personas. El cliente está ubicado geográficamente muy distante al lugar donde se encuentra la entidad desarrolladora.

Proyecto 14 Proyecto dedicado al desarrollo de una plataforma para la informatización de los registros civiles y notarías, con tecnología propietaria, con muy alta presión externa, equipo de desarrollo medianamente capacitado y líderes de proyectos con baja experiencia, conformado por aproximadamente 50 personas. Cliente ubicado geográficamente muy distante al lugar donde se encuentra la entidad desarrolladora.

Anexo 6. Vista del aseguramiento de ingresos en Xedro-GESPRO

Figura 14. Vista de gestión de riesgos en la plataforma GESPRO.

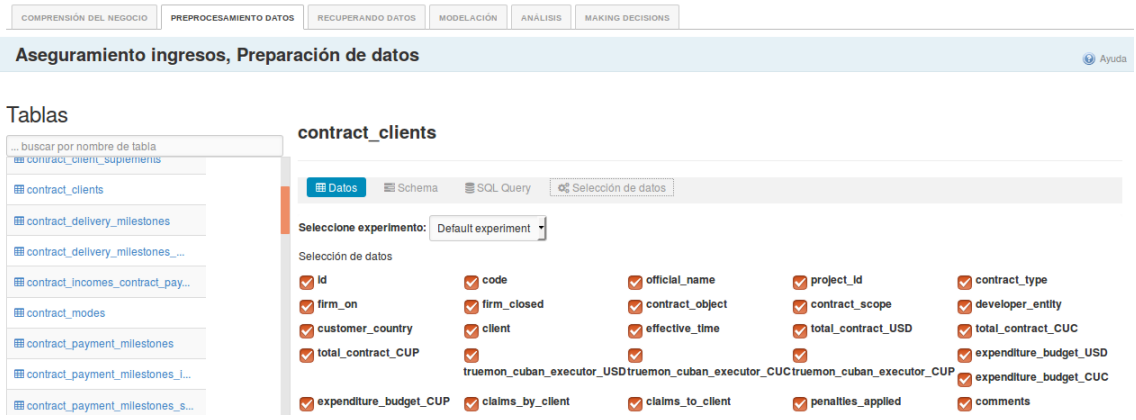


Figura 15. Vista del módulo PRODAnalysisen el GESPRO.