UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

ESCUELA DE POSGRADO

FACULTAD DE CIENCIAS MATEMÁTICAS

UNIDAD DE POSGRADO

"APLICACIÓN DE LA MINERIA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS"

CASO: PODER JUDICIAL

PARA OBTENER EL GRADO ACADÉMICO DE MAGISTER EN COMPUTACIÓN E INFORMÁTICA

AUTOR

Zoraida Emperatriz Mamani Rodríguez

Lima – Perú

2015

FICHA CATALOGRÁFICA

MAMANI RODRIGUEZ, Zoraida Emperatriz

APLICACION DE LA MINERIA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS

PROGRAMA: C.0.3. TECNOLOGÍA DE LA INFORMACION Y COMUNICACIÓN
LINEA DE INVESTIGACION:
C.0.3.21 Técnicas de Programación
C.0.3.22 Ingeniería de Software
C.0.3.23 Administración de Base de Datos
C.0.3.28 Comunicaciones / redes y sistemas distribuidos

Tesis: FACULTAD DE CIENCIAS MATEMÁTICAS UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Formato 28 x 20 cm. Paginas X, 104
(Lima, Perú 2015)

APLICACION DE LA MINERIA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS CASO: PODER JUDICIAL

Autor: Zoraida Emperatriz Mamani Rodríguez

Tesis presentada a consideración del jurado examinador nombrado por la Unidad de Posgrado de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos como parte de los requisitos para obtener el grado académico de **Magister en Computación e Informática**.

Aprobado por:

Dr. Pedro Celso Contreras Chamorro **Presidente**	Dra. Lumila Pró Concepción **Miembro**
Dr. Erwin Kraenau Espinal **Miembro**	Mg. Augusto Cortez Vásquez **Miembro**

Mg. Luz Corina del Pino Rodríguez
Miembro Asesor

DEDICATORIA:

Dedico esta tesina a toda mi familia.

A mis padres Francisco y Alicia por darme la vida y el soporte intelectual, a mis hermanas por su apoyo constante y a Dios por darme salud y fortaleza para continuar por la senda profesional que me he trazado.

AGRADECIMIENTOS

A todos los docentes de la Universidad y personas en general que de alguna u otra manera me brindaron su apoyo incondicional en las diversas etapas de mi vida universitaria y profesional.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMATICAS

UNIDAD DE POST GRADO

MAESTRIA EN COMPUTACION E INFORMATICA

APLICACION DE LA MINERIA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS

Autor: MAMANI RODRIGUEZ, Zoraida Emperatriz

Asesor: DEL PINO RODRIGUEZ, Luz Corina

Grado: Tesis para optar el Grado de Magister en Computación

Fecha: Setiembre 2015

RESUMEN

La minería de datos distribuida está contemplada en el campo de la investigación que implica la aplicación del proceso de extracción de conocimiento sobre grandes volúmenes de información almacenados en bases de datos distribuidas. Las organizaciones modernas requieren de herramientas que realicen tareas de predicción, pronósticos, clasificación entre otros y en línea, sobre sus bases de datos que se ubican en diferentes nodos interconectados a través de internet, de manera que les permita mejorar la calidad de sus servicios.

En ese contexto, el presente trabajo realiza una revisión bibliográfica de las técnicas clustering k-means, elabora una propuesta concreta, desarrolla un prototipo de aplicación y concluye fundamentando los beneficios que obtendrían las organizaciones con su implementación.

Palabras Claves: Minería de Datos Distribuida, Algoritmo Clustering, K-means, Repositorio de Información, Detección de Patrones, Proceso de Negocio, Organizaciones Modernas

NATIONAL UNIVERSITY OF SAN MARCOS

FACULTY OF MATHEMATICS
GRADE UNIT POST
MASTER OF COMPUTING AND INFORMATICS

APPLICATION OF DISTRIBUTED DATA MINING USING CLUSTERING ALGORITHM K-MEANS TO IMPROVE THE QUALITY OF SERVICES OF MODERN ORGANIZATIONS

Author:	MAMANI RODRIGUEZ, Zoraida Emperatriz
Advisory:	DEL PINO RODRIGUEZ, Luz Corina
Academic degree:	Thesis for the degree of MA in computing
Date:	September 2015

ABSTRACT

Distributed data mining is referred to in the research field involves the application of knowledge extraction process on large amounts of information stored in distributed databases. Modern organizations require tools that perform tasks of prediction, forecasting, classification among others, online, on their databases that are located at different sites interconnected over the Internet, so that allows them to improve the quality of their services.

In this context, this paper makes a literature review of clustering techniques k-means, made a concrete proposal, develop a prototype implementation and concludes basing the benefits to be gained with its implementation organizations.

Keywords: Distributed Data Mining, Clustering Algorithm, K-means, Repository of Information, Detecting Patterns, Business Process, Modern Organizations

INDICE

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCION

La minería de datos distribuida es una disciplina de alto interés de los investigadores debido a las limitaciones que ofrecen las centralizadas a las realidades organizacionales actuales.

Las organizaciones competitivas deben mantenerse dispuestas al cambio, a la mejora continua de los servicios que brindan, respetando los estándares de la industria. Por ello están interesadas en utilizar herramientas informáticas que apoyen sus objetivos.

Es en este contexto que se plantea la presente investigación la cual desarrolla un prototipo basado en minería de datos distribuida asimismo propone una adaptación de un algoritmo de agrupamiento distribuido basado en la técnica k-medias. La propuesta se centra básicamente en la detección de patrones de comportamiento basado en un proceso de negocio particular correspondiente a una organización del sector judicial para lo cual aplica la minería de datos distribuida debido a su naturaleza física.

El estudio considera la evaluación basada en data nominal dispersa en repositorios en las sedes de la organización; para lo cual define un modelo dimensional para un proceso de negocio específico, este modelo es sembrado en los repositorios de información de la organización previamente. El proceso de detección de patrones implica la lectura de la data modelada de las a la cual se le denomina instancias a continuación se aplica la técnica clustering sobre estas instancias y se devuelve los resultados hacia una interfaz web para la interpretación respectiva.

A continuación se detalla el contenido del presente trabajo; En el capítulo I se describe el planteamiento del problema el cual comprende la fundamentación y formulación del problema, objetivos generales y específicos, justificación de la investigación, hipótesis y variables. El capítulo II desarrolla el marco teórico; temas como minería de datos, redes P2P, técnicas de modelado, técnicas de clustering distribuido y organizaciones modernas son tratados en este capítulo. En el capítulo III se aborda el estado del arte. En el capítulo IV se desarrolla la propuesta finalmente el capítulo V contiene las conclusiones y recomendaciones.

CAPITULO I. PLANTEAMIENTO DEL PROBLEMA

1.1. Fundamentación y formulación del problema

En las últimas décadas las compañías y organizaciones vienen presentando necesidades de mayor escala y mayor complejidad; esto se presenta debido a las innovaciones y/o mejoras constantes que deben aplicar a sus procesos de negocios en los diversos niveles de su estructura organizacional a efectos de liderar el mercado de su rubro, mantenerse competitivas y/o brindar un servicio de calidad y eficiente a sus clientes y/o usuarios. Estas organizaciones convencionales así como las modernas compañías offshore de hoy en día presentan una estructura funcional global; distribuidas físicamente en amplios espacios geográficos que pueden abarcar varios continentes inclusive.

Efectivamente las nuevas formas de hacer negocios a nivel global va asociado al enorme y constante crecimiento e innovación de tecnologías emergentes; permitiendo que las organizaciones puedan aplicar modelos colaborativos y orientadas a constituirse en organizaciones inteligentes.

Es conveniente destacar que nada o muy poco se conseguiría sin las tecnologías de base de datos, las cuales brindan soporte a los sistemas de información (SI) mediante la administración de los datos de operación. El conglomerado de datos del negocio son recopilados, almacenados, procesados en archivos de datos físicos bajo cierto formato definido por un motor de base de datos; estos datos pueden tener capacidades que fluctúan en Yobibytes ($280 \approx 1.209 \times 1024$bytes).

En la década de los 90s surgen las tecnologías datawarehouse y metodologías que desarrollaban un conjunto de procesos con el objetivo de descubrir conocimiento en bases de datos aplicando técnicas de minería de datos. Este

campo de investigación ha sido ampliamente abordado por investigadores obteniéndose como resultado muchas propuestas interesantes las cuales se dejan ver a través de las innovadoras funcionalidades que nos ofrecen los sistemas de administración de bases de datos propietarios asimismo vienen presentando similar tendencia los de libre distribución.

1.2. Problema general:

¿Cómo influye la ausencia de aplicación de la minería de datos distribuida usando algoritmo de clustering k-means en la mejora de la calidad de servicios que ofrecen las organizaciones modernas?

1.3. Problemas específicos:

¿Cuáles son las limitaciones que presentan la aplicación de la minería de datos no distribuida en las organizaciones modernas?

¿De qué manera se ven limitadas las organizaciones modernas en aspectos relacionadas a la toma de decisiones así como en el soporte a nivel transaccional de sus actividades de negocios debido a la ausencia de soluciones de minería de datos distribuida que proporcionen algoritmos de clustering bajo el enfoque k-means?

1.4. Objetivos

1.4.1. Objetivo General

- Desarrollar un prototipo que aplique minería de datos distribuida mediante el uso de un algoritmo de clustering basado en la técnica k-means.

1.4.2. **Objetivos Específicos**

- Elaborar una propuesta algorítmica de clustering distribuido basado en la técnica k-means
- Desarrollar un prototipo que aplique minería de datos distribuida

1.5. Justificación de la investigación

1.5.1. Justificación Teórica

A nivel nacional no se tienen registros sobre investigaciones sobre el tema en cuestión, sin embargo a nivel internacional si lo hay y estos serán utilizados como base referencial para el cumplimiento de nuestra propuesta.

Consideramos fundamental disponer de un modelo que permita efectuar minería de datos distribuidas sobre infraestructuras de redes que se presenta en internet.

Bajo este contexto se propone el diseño de un algoritmo de clustering bajo el enfoque de la técnica k-means.

1.5.2. Justificación Práctica

De acuerdo a los objetivos de la investigación planteados, se desarrollara un prototipo que aplique minería de datos distribuida y como consecuencia permita fundamentar los beneficios que obtendrían las organizaciones modernas mediante su implementación.

1.6. Fundamentación y formulación de la Hipótesis

Las necesidades de las organizaciones modernas, quienes presentan una estructura organizacional dispersa en nodos o sitios distribuidos geográficamente e interconectados mediantes redes de comunicaciones como

la internet, conlleva a orientar los modelos computacionales existentes a apoyar dichas necesidades tomando en cuenta dicho escenario se plantea la siguiente hipótesis:

"La aplicación de la minería de datos distribuida usando algoritmo de clustering basado en la técnica k-means influye en la mejora de la calidad de servicios de las organizaciones modernas."

1.7. Identificación de variables

1.7.1. Variable Independiente

- La aplicación de la minería de datos distribuida usando algoritmo de clustering basado en la técnica k-means.

1.7.2. Variable Dependiente

- Mejora de la calidad de servicios de las organizaciones modernas

CAPITULO II. MARCO TEÓRICO

Las bases teóricas de la presente investigación se encuentra enmarcada en: minería de datos, minería de datos distribuida, técnicas de clustering distribuido k-means, Redes P2P, Organizaciones modernas.

2.1. Minería de datos

"Minería de datos se define como el proceso de extraer conocimiento útil y comprensible, desde grandes cantidades de datos almacenados por medios automáticos o semi-automáticos. Este proceso incluye no sólo el análisis inteligente de los datos con técnicas de minería de datos, sino también los pasos previos, como el filtrado y preprocesado de los datos, y los posteriores, como la interpretación y validación del conocimiento extraído". [Hernandez at el, 2004, citado por [1]].

Minería de datos es el proceso de descubrimiento automático de conocimiento contenido en la información almacenada en grandes bases de datos.

Este proceso comprende la identificación de relaciones, patrones, asociaciones, segmentos, clasificaciones y tendencias en grandes repositorios de información.

En [2] se presenta la figura 2.1 la cual esquematiza el ciclo de la minería de datos; este inicia con la comprensión de un problema practico de inteligencia de negocios que surge de la actividad cotidiana en las organizaciones y que habitualmente estará asociado a determinadas técnicas y modelos teóricos del análisis de datos. La fase de comprensión de datos cuyo objetivo está en determinar los datos óptimos de acuerdo a los posibles modelos aplicables para la resolución de los problemas identificados en la fase anterior. La fase de preparación de los datos consiste en adecuar los datos para que sean utilizables de modo adecuado en la aplicación de los modelos teóricos identificados para la

resolución del problema. La Fase Modelado consiste en poner en práctica los modelos identificados como óptimos para la resolución del problema. Entre las técnicas a aplicar se tiene las técnicas de minería de datos predictivas y descriptivas. La Fase de Evaluación implica evaluar los modelos definidos en la fase anterior mediante diagnosis formal aplicando técnicas estadísticas e inferenciales adecuadas. La fase de implantación consiste en poner a disposición del usuario solicitante del requerimiento el uso de la herramienta siempre que se haya superado de forma rigurosa las fases anteriores.

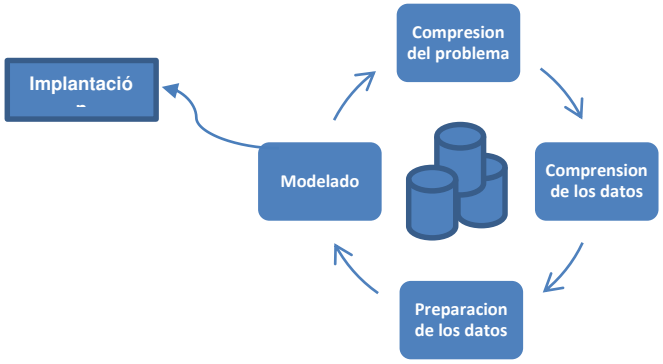


Figura 2.1: Ciclo de la Minería de Datos[1]

2.1.1. El Proceso de Minería de Datos

En [Fayyad at el, 1996 citado por [1]] se define KDD como *"Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos"*.

El proceso de KDD es el proceso de usar métodos algoritmos de Minería de datos para extraer (identificar) lo que es considerado conocimiento de acuerdo a las especificaciones de medidas y umbrales, usando la base de datos junto con

[1] [2], "Técnicas de Minería de Datos e Inteligencia de Negocios"

algún pre-procesamiento requerido, sub-muestreo y transformaciones de esa base de datos.

El proceso de KDD es interactivo e iterativo, envuelve numerosos pasos con muchas decisiones que son hechas por el usuario. En la figura 2.2 se muestra el proceso de KDD enfatizando la naturaleza interactiva del proceso.

El proceso KDD puede ser abordado en forma iterativa e inclusive pares de etapas pueden ser tratadas como ciclo repetitivos.

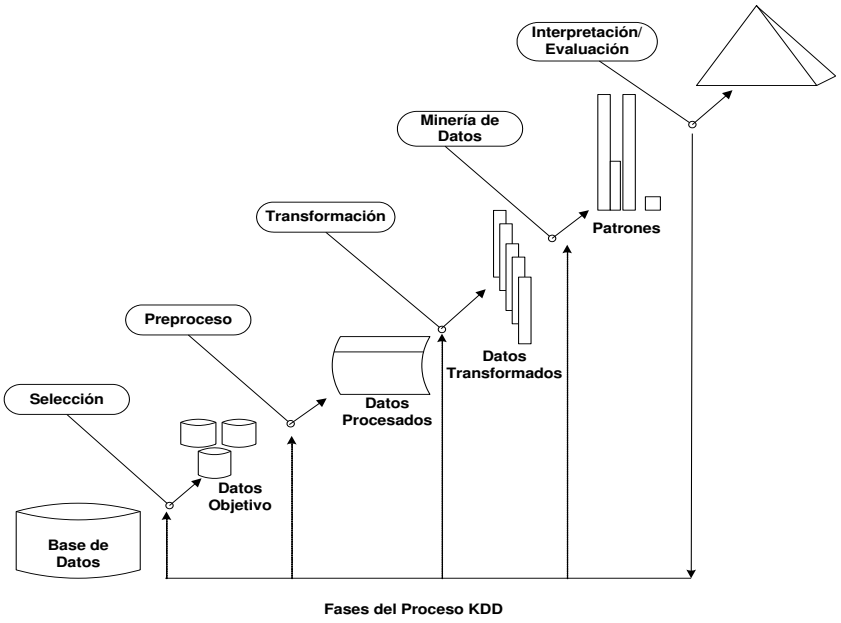


Figura 2.2: Diagrama de las fases del proceso de KDD[2]

A continuación se describe cada una de las fases del proceso de KDD:

1. **Recopilación**: Diseñar el Datawarehouse o DataMart. Esta fase incluye la identificación de las fuentes de información, la definición de la arquitectura tecnológica, el diseño del modelo de datos y la carga del datawarehouse.

[2] [Fayyad at el, 1996, citado por [1]],"Minería de datos: Segmentación de clientes usando el algoritmo de clustering K-Means"

2. **Pre procesamiento**: de los datos: operaciones básicas tales como la eliminación del ruido, estrategias para manejas valores ausentes, normalización de los datos.
3. **Transformación y reducción de los datos**: Incluye la búsqueda de características útiles de los datos según sea el objetivo final, la reducción del número de variables y la proyección de los datos sobre espacios de búsqueda en los que sea más fácil encontrar una solución. Este es un paso crítico dentro del proceso global, que requiere un buen conocimiento del problema y una buena intuición, y que, con frecuencia, marca la diferencia entre el éxito o fracaso de la minería de datos.
4. **Elección de la tarea de Minería de datos:** decidir si el objetivo de KDD es clasificación, regresión, clustering, etc.
5. **Elección de el/los algoritmo(s) de Minería de datos:** seleccionar el/los método(s) que serán utilizados para la búsqueda de patrones en los datos. Esto incluye decidir cuáles modelos y parámetros pueden ser más apropiados y aparear un método de Minería de Datos adecuado con el criterio general del proceso de KDD.
6. **Minería de datos:** buscar patrones de interés en una forma de representación particular o un conjunto de tales representaciones: reglas de clasificación o árboles, regresión, clustering y así sucesivamente.
7. **Interpretación:** Explicación de los patrones minados, y posible retorno a alguno de los pasos.
8. **Consolidación del conocimiento descubierto:** Incorporar este conocimiento dentro del funcionamiento del sistema, o simplemente documentarlo y reportarlo a las partes interesadas. Esta etapa también

incluye revisar y resolver conflictos potenciales con conocimiento previamente conocido.

2.2. Minería de datos distribuida

La Minería de datos distribuida es el proceso de descubrimiento de conocimiento en arquitecturas de datos que son totalmente diferentes al enfoque centralizado. Esto comprende las fuentes de datos distribuidas, el cómputo distribuido y las comunicaciones.

Al igual que en el enfoque centralizado se hacen uso de técnicas de minería para la identificación de relaciones, patrones, asociaciones, segmentos, clasificaciones y tendencias pero para entornos distribuidos.

Las bases de datos distribuidas (BDD) se pueden clasificar en homogéneas y heterogéneas. Las BDD homogéneas son aquellas en las que el mismo esquema está repetido en cada servidor y son, por tanto, los registros los que se encuentran repartidos en los diferentes nodos. Mientras que, las BDD heterogéneas son aquellas en las que cada parte o nodo almacena un subconjunto de las tablas o incluso atributos diferentes de una misma tabla. La minería de datos sobre bases de datos distribuidas (ya sean homogéneas o heterogéneas) se conoce como Minería de Datos Distribuida (MDD)[3].

2.2.1. Arquitectura de MDD

[3] define tres variantes de arquitectura; la figura 2.3 explica la primera variante la cual consiste en que cada procesador o nodo distribuido dispone de un componente de minería encargado de minar los datos en la base de datos local,

[3] [22] Estado del Arte de la Minería de Datos Distribuida

obteniéndose de esta forma, un modelo de minería de datos parcial en cada uno de los nodos. Posteriormente, estos modelos parciales se combinan para obtener el modelo de minería de datos global.

Las figuras 2.4 y 2.5 presentan las dos variantes restantes; estas son similares pues ambas consisten en implementar un modelo global de minería de datos en la parte superior del sistema distribuido que actúa sobre una vista integrada de las distintas bases de datos locales. La diferencia entre estas variantes radica en la forma en que se genera la vista integrada sobre la que actúa la capa de minería de datos.

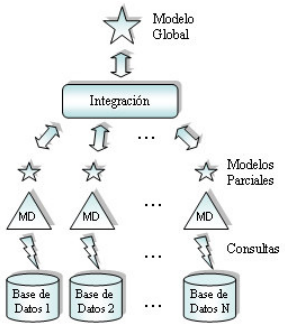


Figura 2.3: Arquitectura MDD con componentes locales (Fuente: [3])

La variante de la figura 2.4 realiza consultas en cada base de datos distribuida de manera independiente, según el conjunto de datos a analizar, posteriormente dichas consultas integradas conforman la vista de datos sobre la que operan los algoritmos de minería de datos.

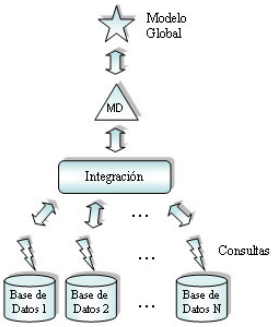


Figura 2.4: Arquitectura MDD con componentes globales e integración de resultados (Fuente: [3])

La figura 2.5 presenta la tercera variante en la cual se integran todas las bases de datos distribuidas y las consultas se realizan sobre esta vista integrada de datos y no en cada base de datos distribuida de manera independiente.

En las variantes de las figuras 2.4 y 2.5, no se obtienen modelos de minería de datos parciales sino únicamente el modelo de minería de datos global.

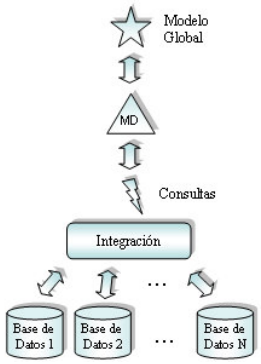


Figura 2.5: Arquitectura MDD con componentes globales e integración de locales. (Fuente: [3])

2.3. Redes P2P

En [3] se da una definición de peer-to -peer (P2P) como:

"P2P es una clase de aplicaciones que aprovechan los recursos de almacenamiento, los ciclos, los contenidos, la presencia humana disponible en la Internet. Debido a que el acceso a los recursos descentralizados significa operar en un entorno de conectividad inestable y direcciones IP impredecibles, los nodos peer-to -peer deben operar fuera del DNS y deben contar con una autonomía significativa al total de los servidores centrales; recursos compartidos, descentralización, dinamismo, autonomía y auto-organización son los principios que caracterizan a las aplicaciones P2P".

En un sistema P2P los recursos se comparten entre todos los participantes que forman una red superpuesta. No hay necesidad de una coordinación centralizada y por lo tanto evitan un punto central de fallo. Esta característica hace que los sistemas P2P sean más escalables y robustas. Como no hay ningún coordinador

centralizado responsable de la organización cada participante actúa de manera autónoma. Cada uno de los nodos se trata como una entidad completamente independiente que cuenta con la autonomía necesaria para interactuar con sus vecinos con una gran flexibilidad. La volatilidad de las conexiones de red es otra característica clave de los sistemas P2P, Los nodos que operan fuera del DNS, se caracteriza principalmente por su carácter estático, donde raramente cambian su topología. La dinamicidad es una característica hace que las aplicaciones P2P encajen perfectamente con el modelo de Internet. Los nodos pueden unirse y salirse de la red P2P en cualquier momento de manera flexible sin afectar o dañar todo el sistema en su conjunto.[4]

2.4. Técnicas de Modelado en Minería De Datos

En [2] se precisa que las técnicas de modelado se basan en el uso de algoritmos. Se distingue tres clases principales: Clasificación, Asociación y Segmentación.

Loa modelos de clasificación usan el valor de una o más variables de entrada para predecir el valor de una o más variables de destino. Algunos ejemplos de estas técnicas son los Arboles de Decisiones, regresión y redes neuronales.

Los modelos de asociación encuentran patrones en los datos en los que una o más entidades (eventos o atributos) se asociación con una o más entidades. Los modelos constituyen conjuntos de reglas que definen estas relaciones. Aquí las variables de los datos pueden funcionar como entrada y salida. Se podrían encontrar estas asociaciones manualmente, pero los algoritmos de reglas de asociación lo hacen mucho más rápido y pueden explorar patrones más complejos.

[4] [3] Survey on Distributed Data Mining in P2P Networks

Los modelos Apriori y Carma son ejemplos del uso de estos algoritmos.

Otro tipo de modelo de asociación es el modelo de detección de secuencias, que encuentra patrones secuenciales en datos estructurados temporalmente.

Los modelos de segmentación también conocidos como conglomerados dividen los datos en segmentos o conglomerados de registros que tienen patrones similares de campos de entrada. Como solo se interesan por los campos de entrada, los modelos de segmentación no contemplan el concepto de campos de salida o destino. Ejemplos de modelos de segmentación se tienen las redes de Kohonen, K-medias, los conglomerados en dos pasos y la detección de anomalías.

En la figura 2.6 se presenta una segunda forma de clasificar las técnicas de minería de datos: predictivas, descriptivas y auxiliares.

Las técnicas predictivas clasifican las variables en variables dependientes y variables independientes, similar al análisis de la dependencia o métodos explicativos del análisis multivariante. Estas técnicas especifican el modelo para los datos en base a un conocimiento teórico previo. Este modelo debe ser contrastado después del proceso de análisis de datos antes de aceptarlo como válido. En este grupo se encuentran las técnicas de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, modelos logísticos, arboles de decisión, redes neuronales, etc.

Las técnicas descriptivas inicialmente todas las variables tienen el mismo estatus similar a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante.

Estas técnicas no están sujetas a ningún modelo formal. Inicialmente no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de los datos.

En este grupo se incluyen las técnicas de Clustering y segmentación las cuales también son consideradas en el grupo de técnicas de clasificación, las técnicas de asociación, las técnicas de análisis exploratorio de datos, las técnicas de la reducción de la dimensión como: factorial, componentes principales, correspondencias, etc. y las técnicas de escalamiento multidimensional.

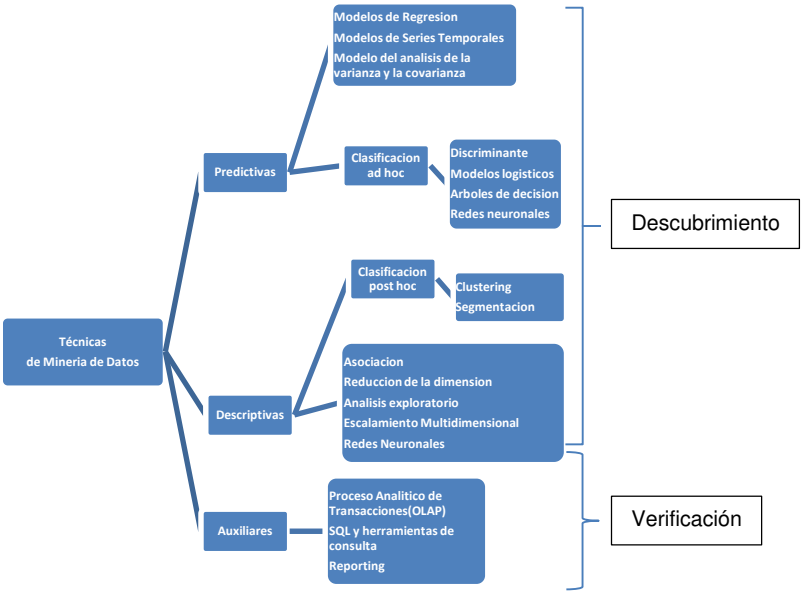


Figura 2.6: Clasificación de las Técnicas de Minería de Datos[5]

El tercer grupo corresponde a las técnicas auxiliares; estas son herramientas de apoyo más superficiales y limitadas; consisten de nuevos métodos basados en

[5] [2], "Técnicas de minería de datos e inteligencia de negocios"

técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la verificación.

Asimismo se observa en la figura 2.6 que las técnicas de clasificación pueden pertenecer tanto al grupo de técnicas predictivas (discriminante, arboles de decisión y redes neuronales) como a las descriptivas (clustering y segmentación). Mientras que a las técnicas de clasificación predictivas se les denomina técnicas de clasificación ad hoc, a las descriptivas como técnicas de clasificación post hoc por que realizan clasificación sin especificación propia de los grupos.

[Fayyad at el, 1996, citado por [1] identifica técnicas descriptivas y predictivas. Las descriptivas construyen modelos de comportamiento sobre patrones de hechos ocurridos en el pasado para su presentación a un usuario en una forma humanamente comprensible. Las predictivas construyen uno o más modelos sobre datos ocurridos en el pasado para predecir el comportamiento de nuevos conjuntos de datos, los cuales pueden corresponderse o no con hechos futuros.

A continuación se ofrece una clasificación más detallada de las técnicas de minería de datos:

2.4.1. Clasificación

[Medina, 2005, citado por [1]] describe la clasificación como el proceso que *"...permite caracterizar las instancias del conjunto de datos en clases sobre la base de propiedades conocidas de ellas. Su propósito es descubrir si una instancia del conjunto de datos pertenece a una de varias clases previamente definidas"*.

Un ejemplo de clasificación según [Medina, 2005, citado por [1]] a partir del conjunto de datos anterior y considerando las clases predefinidas por la variable "Salario", es el siguiente:

Si (Ojos = Castaños) y (Edad ≤ 50) Entonces (Salario = Alto)

Si (Ojos = Castaños) y (Edad > 50) Entonces (Salario = Medio)

Si (Ojos = Verdes) Entonces (Salario = Medio)

Si (Ojos = Azules) y (Edad ≤ 50) Entonces (Salario = Alto)

Si (Ojos = Azules) y (Edad > 50) Entonces (Salario = Bajo)

Lo cual se puede comprobar al agrupar las instancias convenientemente.

Id.	Edad	Salario	Ojos	Sexo
3	47	Alto	Castaños	F
6	27	Alto	Castaños	M
1	62	Medio	Castaños	F
2	53	Medio	Verdes	M
4	32	Medio	Verdes	M
5	21	Alto	Azules	F
8	46	Alto	Azules	M

Tabla 2.1: Clasificación de salarios[6]

Las tareas de clasificación pueden ser tratadas tanto de forma descriptiva como predictivas. En el primer caso, se desea conocer las variables y valores más significativos que describen a las instancias de cada tipo de clase. Mientras tanto, en las tareas predictivas se desea examinar las características de una nueva

[6] [Medina, 2005, citado por [1]],"Minería de datos: Segmentación de clientes usando el algoritmo de clustering K-Means"

instancia y asignarle una de las clases predefinidas. Un caso especial de tarea predictiva es el completar los datos nulos de un campo en una BD.

2.4.2. Análisis de Dependencias

En el análisis de dependencias se agrupan un conjunto de datos que permiten detectar dependencias entre dos o más variables o campos en el conjunto de datos del modelo[6].

Uno de los métodos más simples de análisis de dependencias es el de correlación. La correlación es una medida de la relación entre dos o más variables. Los coeficientes de correlación suelen medirse en el intervalo [-1.00,+1.00]. El valor de +1.00 representa una perfecta correlación positiva; o sea, los valores de las variables correlacionadas aumentan o disminuyen de forma conjunta. Mientras tanto, un coeficiente de correlación de -1.00 representa una perfecta correlación negativa; o sea, las variables correlacionadas varían en direcciones opuestas.

Una de las formas más utilizadas en el Análisis de Dependencias entre variables categóricas es mediante la búsqueda de Reglas de Asociación, tales como:

De todas las instancias con valores A y B, el 83 % también presentan el valor C.

De esta forma, por ejemplo, pueden descubrirse que determinados síntomas suelen ocurrir junto a ciertos tipos de enfermedades, o que algunos productos de un supermercado presentan comportamientos de ventas correlacionadas, sugiriendo el estudio posterior de las razones de tales situaciones.

2.4.3. Clustering

Clustering o Agrupamiento divide la información en grupos diferentes. El objetivo del agrupamiento es encontrar grupos (clusters) que son muy diferentes unos de otros, y cuyos miembros son muy similares unos de otros. A diferencia de la clasificación, al empezar, usted no sabe que grupos van a aparecer, o por qué tipo de atributos se aglomeraran los datos. Como consecuencia, los grupos deben ser interpretados por una persona que tenga conocimiento del negocio. Los grupos o clusters determinados pueden ser usados para clasificar datos nuevos [Edelstein, 2005, citado por [1]].

"...Clustering es una de las principales tareas en Data Mining y consiste en particionar un conjunto de datos en colecciones de objetos o instancias de manera tal que dentro de cada partición los objetos sean "similares" entre sí, y a su vez se "diferencien" de los objetos contenidos en las otras particiones. Similitud y disimilitud son expresadas a través de funciones de distancia / similitud y a las particiones resultantes se las denomina clusters"[7].

Es importante no confundir Agrupamiento con segmentación. La segmentación hace referencia al problema general de identificar grupos que tienen características comunes. Clustering es una forma de segmentar los datos en grupos que no se habían definido con anterioridad, mientras que la clasificación es una manera de segmentar datos, asignándolos en grupos que ya se habían definido previamente.

Los algoritmos de clustering tienen las siguientes características:

[7] [Navas, 2004, citado por [1]], "Minería de datos: Segmentación de clientes usando el algoritmo de clustering K-Means"

- Escalabilidad: normalmente corren con pocos datos.
- Capacidad de manejar diferentes tipos de atributos: numéricos (lo más común), binarios, nominales, ordinales.
- Requerimientos mínimos para especificar parámetros, como el número de clusters.
- Manejo de ruido: muchos son sensibles a datos erróneos.
- Independiente del orden de los datos.
- Puede funcionar eficientemente con alta dimensionalidad.

En la figura 2.7 se muestran dos grupos donde los objetos son similares entre sí.

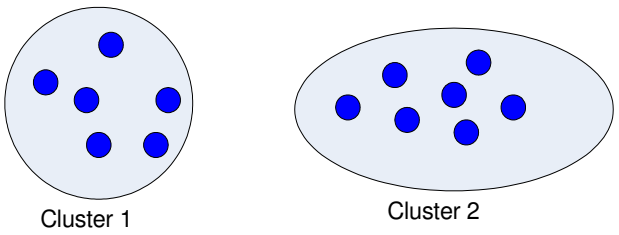


Figura 2.7: Ejemplo de clustering[8]

2.4.3.1. Algoritmos de Clustering

Para elegir un algoritmo de clustering para una determinada aplicación deben tomarse en cuenta los siguientes factores[9]:

- **Objetivo de la aplicación**: El objetivo de la aplicación a menudo afecta la elección de un algoritmo de clustering. Por ejemplo, si se desea encontrar la ubicación óptima de las sucursales de una cadena de supermercados, el objetivo será encontrar la mínima distancia entre la ubicación de los clientes y cada sucursal. Para el reconocimiento de imágenes, es deseable hallar los clusters que deben tener cierta uniformidad de color, densidad, etc.

[8] [1],"Minería de datos: Segmentación de clientes usando el algoritmo de clustering K-Means"
[9] [Navas, 2004,citado por [1]], "Minería de datos: Segmentación de clientes usando el algoritmo de clustering K-Means"

Los algoritmos de particionamiento como k-means, son útiles para el primer caso, y tienden a descubrir clusters con forma esférica y tamaño similar. En cambio para el segundo caso responden mejor los algoritmos basados en densidad.

- **Elección entre calidad y velocidad***:* Existe siempre un problema al elegir entre velocidad de procesamiento y calidad de los clusters obtenidos. Un algoritmo adecuado debe cumplir con estas dos características, pero a veces la cantidad de instancias a procesar juega un papel importante en el tiempo de ejecución del algoritmo de clustering. Un algoritmo que produce clusters de calidad, por lo general es incapaz de manejar grandes cantidades de información. A su vez, cuando se trabaja con grandes volúmenes de datos se realiza una especie de compresión sobre la información inicial, perdiendo así calidad.
- **Características de los datos**: Las características de los datos a los cuales se quiere aplicar el clustering, también son un factor importante para la elección del método adecuado.
- **Tipos de datos de los atributos**: La similitud entre dos objetos es medida según la diferencia entre los valores de sus atributos. Cuando todos los atributos son numéricos las medidas de distancia como Euclídea o Manhattan pueden ser fácilmente calculadas. En cambio, cuando los atributos son binarios, categóricos u ordinales, el cálculo se complica. Por lo tanto, la mayoría de los algoritmos se aplican sólo a datos numéricos.
- **Dimensionalidad**: La dimensionalidad se refiere al número de atributos de cada objeto. Muchos algoritmos funcionan bien con pocos atributos, pero al aumentar las dimensiones los resultados se degeneran. La degeneración puede darse por la disminución de la velocidad o el empobrecimiento de la calidad de los clusters.

- **Cantidad de ruido en los datos**: Algunos algoritmos son muy sensibles a ruido o elementos aislados, no pudiendo funcionar correctamente ante la presencia de los mismos.

  Se explican a continuación los principales algoritmos de clustering:

2.4.3.2. Algoritmos de Particionamiento

Los algoritmos de particionamiento toman un conjunto D de n objetos en un espacio de d dimensiones, y construyen *k* particiones de los objetos, donde cada partición representa un grupo o cluster. Cada grupo tiene al menos un elemento y cada elemento pertenece a un solo grupo. Estos métodos, crean una partición inicial e iteran hasta un criterio de paro. Ejemplo: *k*-means.

2.4.3.3. Algoritmos Jerárquicos

Este tipo de algoritmo de clustering crea descomposiciones jerárquicas formando un árbol que divide la base de datos recursivamente en conjuntos cada vez más pequeños. El árbol puede formarse de dos formas: Bottom-Up o Top-Down.

El método aglomerativo o *bottom-up*, empieza con un grupo por cada objeto y une los grupos más parecidos hasta llegar a un solo grupo u otro criterio de paro.

El método divisorio o *top-down*, empieza con un solo grupo y lo divide en grupos más pequeños hasta llegar a grupos de un solo elemento u otro criterio de paro.

2.4.3.4. Algoritmos basados en densidades

Algoritmos de clustering han sido desarrollados en base a la noción de "densidad". Estos generalmente estiman clusters como regiones con gran densidad de objetos, separados en regiones de baja densidad de objetos (estos

elementos aislados representan ruido). Este tipo de métodos es muy útil para filtrar ruido y encontrar clusters de diversas formas. Ejemplos: DBSCAN, DENCLUE (DENsity-based CLUstEring)

2.4.3.5. Algoritmos basados en Grid

Los métodos basados en densidad suelen tener grandes problemas cuando se trabaja con bases de datos muy grandes. Para mejorar la efectividad del clustering, un método basado en grillas usa una estructura de grilla de datos. El método divide el espacio en un número finito de celdas, formando una grilla, en donde se realizan todas las operaciones del clustering. Ejemplos: STING, CLIQUE.

2.4.3.6. Algoritmo K-Means

El algoritmo de clustering K-Means (Figura 2.8) toma como parámetro *k*, este representa el número de clusters que formará. El algoritmo inicia seleccionando *k* elementos aleatoriamente, los cuales representan el centro o media de cada cluster.

Algoritmo de k-means
Selecciona k objetos aleatoriamente **Repetir** Re(asigna) cada objeto al cluster más similar con el valor medio Actualiza el valor de las medias de los clusters **Hasta no hay cambio**

Figura 2.8: Algoritmo de clustering k-means
(Fuente: [Morales, 2003, citado por [1]])

A cada objeto restante se le asigna el cluster con el que más se parece, basándose en la distancia entre el objeto y la media del cluster. Después calcula

la nueva media del cluster e itera hasta no cambiar de media [Morales, 2003, citado por [1]].

El algoritmo k-means utiliza un medida de similaridad basada en el error cuadrático (E). El objetivo del algoritmo k-means es minimizar el valor de E que representa la menor distancia entre los elementos y el centro del cluster. Así por ejemplo para "K" clusters, C_1, C_2, ... , C_k , son centroides de los clusters entonces el error cuadrático medio está dado por:

$$E = \frac{\sum_{i=1}^{k} \sum_{p \varepsilon C_i} |p - m_i|^2}{k}$$

donde:
p representa al objeto
m_i a la media del cluster C_i
k es el número de cluster

Medida de similaridad

La medida de similaridad mide la distancia entre el objeto y la media de un cluster. K-Means utiliza la fórmula de distancia euclideana que se describe a continuación.

Distancia Euclidiana

$$d(x,y) = (|x_1 - y_1|^2 + |x_2 - y_2|^2 + \ldots + |x_n - y_n|^2)^{1/2}$$

dónde: d(x, y): Representa la distancia entre dos elementos x e y

En la figura 2.9 se muestra un ejemplo de clustering k-means. En el cuadro superior izquierdo se muestra el conjunto de objetos que se van a particionar en k =3 clusters. En el cuadro abajo a la derecha se muestran los clusters ya generados. Estos cuadros visualizan los objetos dentro de cada cluster, estos son similares entre si y a su vez son disimiles con los objetos de los otros clusters.

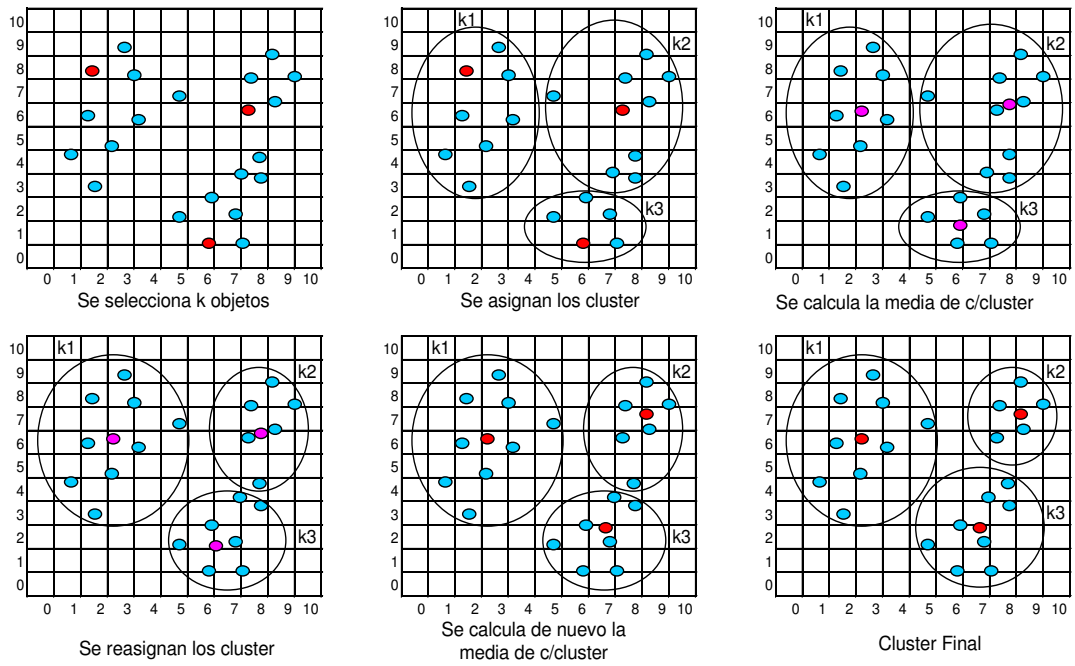


Figura 2.9: Ejemplo de clustering k-means
(Fuente: [1])

2.5. Técnicas de clustering distribuido

El clustering distribuido o agrupamiento distribuido; es el proceso mediante el cual los registros de una base de datos son particionadas en grupos; los elementos de cada grupo comparten características comunes y son estas características las que las distinguen de otros grupos [3].

El objetivo de un proceso de agrupamiento consiste en minimizar la similitud de los elementos pertenecientes a un grupo y maximizarla entre grupos.

Un algoritmo de agrupamiento considera tres pasos:

El primer paso consiste en procesar modelos en nodos locales para lo cual hace uso de algoritmos de agrupamiento local.

El segundo paso agrega los modelos locales resultantes en un nodo central.

El tercer paso consiste en que, en el nodo central se procesa el modelo global. A continuación este modelo global es enviado a todos los nodos locales participantes con la finalidad de generar grupos optimizados.

2.6. Organizaciones Modernas

"Una organización es una estructura social formal y estable, que toma los recursos del entorno y los procesa para producir salidas. Esta definición técnica se enfoca en tres elementos de una organización. El capital y la mano de obra son los factores primarios de producción proporcionados por el entorno. La organización (empresa) transforma estas entradas en productos y servicios en una función de producción."[10]

Los entornos consumen los productos y servicios a cambio del suministro de entradas.

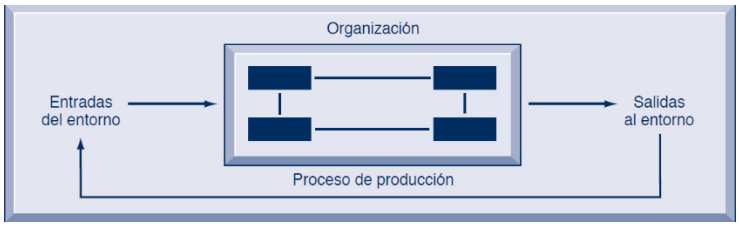


Figura 2.10: La definición microeconómica técnica de la organización (Fuente: [5])

En la figura 2.11 grafica la definición microeconómica de las organizaciones, la empresa transforma el capital y la mano de obra (los factores primarios de producción proporcionados por el entorno) por medio del proceso de producción en productos y servicios (salidas al entorno). El entorno consume los productos y servicios, además de proporcionar el capital y la mano de obra adicionales como entradas en el lazo de retroalimentación.

[10] [5], "Sistema de Información Gerencial"

2.6.1. Características

En [4] describe las características de las organizaciones las cuales se detallan a continuación:

Son Sistemas Sociales

Dentro de cada organización las personas asumen distintos roles según la actividad que deban desempeñar: empleados administrativos, vendedores, operarios industriales, peones, personal de servicios, etc.

Para que esas actividades sean desempeñadas correctamente, deben existir;

- personas que determinen previamente cuáles van a ser las tareas;

- los que ejecutan las directivas que reciben;

- quienes controlan los trabajos efectuados.

El desempeño de todas estas funciones debe hacerse respetando las jerarquías y ocupaciones asignadas, para que el trato entre las personas no encuentre motivos de roces y mal entendidos que podrían provocar situaciones enojosas o irritantes.

Las organizaciones deben ser sistemas sociales reglados por principios de convivencia humana. Las relaciones personales entonces, se efectuarán en un clima armónico y agradable.

Son universales

Ya no se concibe que las actividades colectivas se realicen sin de esbozo organizativo, Las ventajas indudables del orden hacen que ante toda realización

que necesite la colaboración de varias personas, estas busquen coordinar esfuerzos para efectuarla en la mejor forma o con la mayor eficiencia.

No sólo son proclives a ser organizadas las actividades económicas sino también todo otro tipo de labor, ya sea cultural, recreativa, deportiva, etc.

Ejercen influencia sobre sus componentes

En toda organización, existe un modelo funcional donde se establecen los cargos y roles que cada persona debe desempeñar en ella.

Ese modelo funcional se pone en práctica a través de las órdenes que emanan de las jerarquías superiores, las tareas de quienes las ejecutan y el control de los responsables que verifican la corrección de los trabajos realizados.

Quienes se desempeñan dentro de estas estructuras, no sólo reciben conocimientos a través de su experiencia en la misma, sino que en el diario convivir, pueden llegar a tomar como propias las costumbres y hábitos de trabajo, haciendo suyas las motivaciones de la organización.

Trascendentes en el tiempo

Las organizaciones tienen posibilidades de ser más longevas que sus creadores. El período de vida de la persona es limitado con respecto al de las instituciones que pudieran concebir.

El hombre puede llegar a vivir cien años pero mucho antes tiene derecho a obtener los beneficios de la jubilación. Si se descuenta el período de crecimiento y el de capacitación, es muy difícil que llegue a estar cincuenta años dirigiendo una organización.

Las organizaciones se constituyen para perdurar en el tiempo. En aquellas que persiguen fines de lucro, son los hijos, nietos, bisnietos, autoridades se van sucediendo generalmente mediante el voto de sus componentes.

La Iglesia Católica es un ejemplo de una organización que perdura en el tiempo con sus casi dos mil años de vida.

Se personifican

Las organizaciones son consideradas legalmente como personas jurídicas capaces de adquirir derechos y contraer obligaciones.

A los efectos de contratar con terceros, lo hace a través de sus representantes, convienen, pactan, acuerdan, estipulan, etc. las diferentes cláusulas de los tratados respectivos.

Además recibe un nombre, el que se utiliza para identificarla. ¿Cuántas empresas comerciales son famosas por su nombre?

Son representadas

Al ser personas de existencia ideal, las organizaciones contratan con terceros por medio de sus representantes.

Los representantes, generalmente los directivos de la organización no concuerdan a nombre personal sino que son meros administradores.

Los administradores representan a la organización mientras dure su mandato, el que normalmente puede ser renovado o revocado según los estatutos o reglamentos pertinentes.

De estructura dinámica

Las organizaciones evolucionan constantemente para adaptarse a los requerimientos del medio (cambios tecnológicos, sociales, culturales, etc.) y a la demanda de aquellos a quienes está destinada su producción.

Por otra parte los adelantos tecnológicos proveen permanentemente de la materia prima necesaria a fin de renovar estructuras y modernizar procesos.

Aplican la división del trabajo

Cuando el hombre se dio cuenta de que especializándose podía efectuar una mayor y mejor producción, apareció la división del trabajo; parcelamiento de la producción en procesos que efectúan distintas personas, especialistas en sus tareas, con lo que se intenta lograr una mayor eficiencia y un mejor producto terminado.

La aplicación de esta división del trabajo ha traído como consecuencia, la necesidad de organizarlo para que el resultado no sea un caos sino un producto eficiente, consecuencia de la eficaz colaboración de todos y cada uno de los que efectúan la labor.

Son complejas

En toda organización existen distintas funciones o departamentos que serán más o menos numerosos según sus objetivos.

En la faz económica, por ejemplo, muchas organizaciones tratan de lograr una integración vertical (producir desde la materia prima, y en muchos casos vender el producto terminado directamente al público). En este caso, la organización

será más compleja que la de un mono productor. Pero en ambos tendrán, además del departamento de producción, otros que lo complementarán.

Sinergéticas

Vimos que la complejidad es una características de las organizaciones, tengan estás mayor o menor cantidad de departamentos. Esos departamentos deben actuar en forma coordinada para lograr una mejor eficiencia, no sólo en el aspecto económico (menores costos) sino también en la prestación de servicios a sus componentes o a los destinatarios en general. Esta coordinación hace que el resultado no sea un caos, sino un producto eficiente, consecuencia de la eficaz colaboración de todos y cada uno de los que efectúan la labor.

Aplican el efecto multiplicador

Un atleta que corre mil metros tardará más tiempo en llegar a la meta que diez personas que corran cien metros, una a continuación de la otra. En ambos casos la distancia recorrida será la misma, pero como consecuencia de la “división del trabajo”, se habrá logrado un mejor rendimiento.

Dicho ejemplo podemos aplicarlo a cualquier tipo de organización; al dividirse la tarea entre varios, el resultado no será igual a la suma de esfuerzos individuales, sino mucho mayor, lo que se conoce con el nombre de efecto multiplicador.

Eficientes

Todos los caracteres que hemos analizado de las organizaciones buscan llegar a este último carácter básico y fundamental: la eficiencia.

Una organización incorrecta trae como consecuencia mayores costos operativos y un entorpecimiento de las labores, lo que evidentemente atenta contra su eficiencia.

En un mundo competitivo como el actual, las organizaciones deben buscar la mayor eficiencia en sus operaciones.

Las Organizaciones del sector público cada vez están más interesados en mejorar su eficiencia operativa, compartir información entre entidades del sector a través de la colaboración, e integración de sus procesos reduciendo las barreras operativas y jurisdiccionales manteniendo el control y la reducción de costos.

Este tipo de organizaciones en el mundo viene demandando de modelos de implementación modernos.

2.6.2. Elementos

Los elementos clave de una organización son: su gente, su estructura, sus procesos de negocios, sus políticas y su cultura.

Estructura

En la figura 2.12 se presenta la Pirámide Organizacional la cual se encuentra compuesta por distintos niveles y áreas.

Sus estructuras revelan una clara división de labores. La autoridad y responsabilidad en una empresa de negocios se organizan como una jerarquía, o estructura de pirámide.

Los niveles superiores de esta jerarquía consisten en empleados gerenciales, profesionales y técnicos, mientras que los niveles base de la pirámide consisten en personal operacional.

La gerencia de nivel superior toma decisiones estratégicas de largo alcance sobre productos y servicios, además de asegurar el desempeño financiero de la empresa. La gerencia de nivel medio lleva a cabo los programas y planes de la gerencia de nivel superior y la gerencia operacional es responsable de supervisar las actividades diarias de la empresa. Los trabajadores del conocimiento, como los ingenieros, científicos o arquitectos, diseñan productos o servicios y crean nuevo conocimiento para la empresa, en tanto que los trabajadores de datos (secretarias o asistentes administrativos) ayudan con la calendarización y las comunicaciones en todos los niveles de la empresa. Los trabajadores de producción o de servicio son los que elaboran el producto y ofrecen el servicio.

Figura 2.11: Estructura de la Organización[11]

[11] [5], “Sistema de Información Gerencial”

Los expertos se emplean y capacitan para distintas funciones de negocios. Las principales funciones de negocios, o tareas especializadas que realizan las organizaciones comerciales, consisten en ventas y marketing, manufactura y producción, finanzas y contabilidad, y recursos humanos.

Una organización coordina el trabajo mediante su jerarquía y sus procesos de negocios que son tareas y comportamientos relacionados en forma lógica para realizar el trabajo. Desarrollar un nuevo producto, cumplir con un pedido y contratar un empleado son ejemplos de procesos de negocios.

Los procesos de negocios de la mayoría de las organizaciones incluyen reglas formales para realizar tareas, que se han desarrollado a través de un largo periodo. Estas reglas guían a los empleados en una variedad de procedimientos, desde escribir una factura hasta responder a las quejas de los clientes. Algunos de estos procesos de negocios están por escrito, pero otros son prácticas de trabajo informales (como el requerimiento de regresar las llamadas telefónicas de los compañeros de trabajo o clientes) que no se han documentado. Los sistemas de información automatizan muchos procesos de negocios.

FUNCION	PROPOSITO
Ventas y marketing	Vender los productos y servicios de la organización
Manufactura y producción	Producir y ofrecer productos y servicios
Finanzas y contabilidad	Administrar los activos financieros de la organización y mantener sus registros financieros
Recursos humanos	Atraer, desarrollar y mantener la fuerza laboral de la organización; mantener los registros de los empleados

Tabla 2.2: Principales Funciones de Negocios[12]

[12] [5],"Sistema de Información Gerencial"

Por ejemplo, la forma en que un cliente recibe crédito o una factura se determina con frecuencia mediante un sistema de información que incorpora un conjunto de procesos de negocios formales.

Administración

El trabajo de la gerencia es dar sentido a las distintas situaciones a las que se enfrentan las organizaciones, tomar decisiones y formular planes de acción para resolver los problemas organizacionales. Los gerentes perciben los desafíos de negocios en el entorno; establecen la estrategia organizacional para responder a esos retos y asignan los recursos tanto financieros como humanos para coordinar el trabajo y tener éxito. En el transcurso de este proceso, deben ejercer un liderazgo responsable.

Un gerente debe hacer algo más que administrar lo que ya existe, debe crear nuevos productos y servicios, e incluso volver a crear la organización de vez en cuando.

Una buena parte de la responsabilidad de la gerencia es el trabajo creativo impulsado por el nuevo conocimiento e información. La tecnología de la información puede desempeñar un poderoso papel para ayudar a los gerentes a diseñar y ofrecer nuevos productos y servicios, y para redirigir y rediseñar sus organizaciones.

2.6.3. Activos complementarios

Capital organizacional y el Modelo de negocios correcto

En la figura 2.13 se presenta la variación en rendimiento con respecto a la inversión en tecnologías de información Los estudios de rendimientos de las

inversiones en tecnología de información Las empresas no obtienen los mismo rendimientos; algunas de ellas invierten y reciben mucho (cuadrante 2); otras invierten una cantidad igual y reciben pocos rendimientos (cuadrante 4). Asimismo, otras empresas invierten poco y reciben mucho (cuadrante 1), mientras que otras invierten y reciben poco (cuadrante 3).

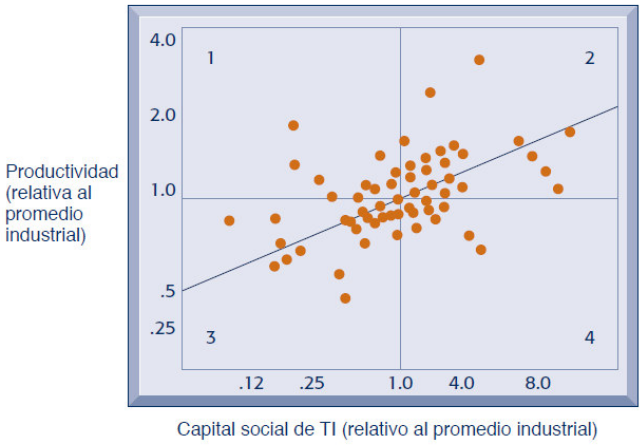


Figura 2.12: Variación en rendimiento sobre la inversión en TI[13]

Esto sugiere que el hecho de invertir en tecnología de la información no garantiza por sí solo buenos rendimientos. ¿Qué explica esta variación entre las empresas?

La respuesta radica en el concepto de los activos complementarios. Las inversiones en tecnología de la información por sí solas no pueden aumentar la efectividad de las organizaciones y los gerentes, a menos que se apoyen con valores, estructuras y patrones de comportamiento en la organización, además de otros activos complementarios.

[13] [5], “Sistema de Información Gerencial”

Las empresas comerciales necesitan cambiar la forma en que hacen sus negocios para que realmente puedan cosechar las ventajas de las nuevas tecnologías de la información.

Algunas empresas no adoptan el modelo de negocios correcto que se adapte a la nueva tecnología, o buscan preservar un modelo de negocios antiguo condenado al fracaso por la nueva tecnología. Por ejemplo, las compañías discográficas se rehusaron a cambiar su antiguo modelo de negocios basado en las tiendas de música tradicionales para la distribución, en vez de adoptar un nuevo modelo de distribución en línea. Como resultado, las ventas de música legales en línea están controladas por una compañía de tecnología llamada Apple Computer, en lugar de por las compañías discográficas.

Los activos complementarios son aquellos requeridos para derivar valor a partir de una inversión primaria (Teece, 1988; citado por [5]). Por ejemplo, para aprovechar el valor de los automóviles se requieren inversiones complementarias considerables en carreteras, caminos, estaciones de gasolina, instalaciones de reparación y una estructura regulatoria legal para establecer estándares y controlar a los conductores.

La investigación sobre la compra de tecnología de información de negocios indica que las empresas que apoyan este tipo de gasto con el de activos complementarios, como nuevos modelos y procesos de negocios, el comportamiento gerencial, la cultura organizacional o la capacitación, reciben mayores rendimientos, en tanto que las empresas que no realizan estas inversiones complementarias reciben menos rendimientos o ninguno sobre sus adquisiciones de tecnología de la información (Brynjolfsson, 2003;Brynjolfsson y

Hitt, 2000; Davern y Kauffman, 2000; Laudon, 1974; citado por [5]). Estas inversiones en organización y administración también se conocen como capital organizacional y gerencial.

La tabla 2.3 muestra una lista de las principales inversiones complementarias que necesitan hacer las empresas para aprovechar el valor de sus gastos en tecnología de la información. Parte de éstos implica los activos tangibles, como edificios, maquinaria y herramientas.

Activos organizacionales	- Cultura organizacional de apoyo, que aprecia la eficiencia, la eficacia y la efectividad - Modelo de negocios apropiado - Procesos de negocios eficientes - Autoridad descentralizada - Derechos de toma de decisiones distribuidas - Sólido equipo de desarrollo de SI
Activos gerenciales	- Sólido apoyo de la gerencia de nivel superior en cuanto a la inversión en tecnología y el cambio. - Incentivos para la innovación gerencial - Entornos de trabajo en equipo y colaborativo - Programas de capacitación para mejorar las habilidades de decisión gerencial - Cultura gerencial que aprecia la flexibilidad y la toma de decisiones basadas en el conocimiento
Activos sociales	- Internet y la infraestructura de telecomunicaciones - Programas educacionales enriquecidos con TI que elevan el alfabetismo computacional de la fuerza laboral - Estándares (tanto de gobierno como del sector privado) - Leyes y regulaciones que creen entornos de mercados justos y estables - Empresas de tecnología y servicios en mercados adyacentes para ayudar en la implementación

Tabla 2.3: Activos Complementarios para optimizar los rendimientos de las inversiones en TI[14]

[14] [5],"Sistema de Información Gerencial"

Sin embargo, el valor de estas compras de tecnología de la información depende en gran parte de las inversiones complementarias en la administración y organización.

Las inversiones complementarias organizacionales clave son una cultura de negocios de apoyo, la cual aprecia la eficiencia, la eficacia y la efectividad, un modelo de negocios apropiado, procesos de negocios eficientes, la descentralización de la autoridad, los derechos de decisión altamente distribuidos y un sólido equipo de desarrollo de sistemas de información.

Los activos complementarios gerenciales importantes son: un sólido apoyo de la gerencia de nivel superior con respecto al cambio, sistemas de incentivos que supervisan y recompensan la innovación individual, un énfasis en el trabajo en equipo y la colaboración, programas de capacitación y una cultura gerencial que aprecie la flexibilidad y el conocimiento.

Las inversiones sociales importantes (no las que hace la empresa, sino la sociedad en general, otras empresas, gobiernos y otros participantes clave del mercado) son Internet y la cultura de apoyo de Internet, los sistemas educativos, los estándares de redes y computación, las regulaciones y leyes, y la presencia de las empresas de tecnología y servicios.

2.6.4. Procesos de Negocios

Los procesos de negocios se refieren a la forma en que se organiza, coordina y enfoca el trabajo para producir un producto o servicio valioso. Los procesos de negocios son el conjunto de actividades requeridas para crear un producto o servicio. Estas actividades se apoyan mediante flujos de material, información y

conocimiento entre los participantes en los procesos de negocios. Los procesos de negocios también se refieren a las formas únicas en que las organizaciones coordinan el trabajo, la información y el conocimiento, y cómo la gerencia elije coordinar el trabajo.

En mayor grado, el desempeño de una empresa depende de qué tan bien están diseñados y coordinados sus procesos de negocios, los cuales pueden ser una fuente de solidez competitiva si le permiten innovar o desempeñarse mejor que sus rivales.

Los procesos de negocios también pueden ser desventajas si se basan en formas obsoletas de trabajar que impidan la capacidad de respuesta a la eficiencia.

Podemos ver a toda empresa como un conjunto de procesos de negocios, algunos de los cuales forman parte de procesos más grandes que abarcan más actividades. Muchos procesos de negocios están enlazados con un área funcional específica. Por ejemplo, la función de ventas y marketing es responsable de identificar a los clientes y la función de recursos humanos de contratar empleados. La tabla 2.4 describe algunos procedimientos comunes de negocios para cada una de las áreas funcionales de una empresa.

Otros procesos de negocios cruzan muchas áreas funcionales distintas y requieren de una coordinación entre los departamentos. Por ejemplo, considere el proceso de negocios aparentemente simple de cumplir el pedido de un cliente (vea la figura 2.14). Al principio, el departamento de ventas recibe un pedido. El cual pasa primero a contabilidad para asegurar que el cliente pueda pagarlo, ya sea mediante una verificación de crédito o una solicitud de pago inmediato antes

del envío. Una vez que se establece el crédito del cliente, el departamento de producción extrae el artículo del inventario o lo elabora. Después el producto se envía (y para esto tal vez haya que trabajar con una empresa de logística, como UPS o FedEx). El departamento de contabilidad genera un recibo o factura y se emite un aviso al cliente para indicarle que la mercancía se ha enviado. El departamento de ventas recibe la notificación del envío y se prepara para dar soporte al cliente, ya sea contestando llamadas o dando seguimiento a las reclamaciones de garantía.

Manufactura y producción	- Ensamblar el producto - Verificar la calidad - Producir listas de materiales
Ventas y marketing	- Identificar a los clientes - Hacer que los clientes estén conscientes del producto - Vender el producto
Finanzas y contabilidad	- Pagar a los acreedores - Crear estados financieros - Administrar cuentas de efectivo
Recursos humanos	- Contratar empleados - Evaluar el desempeño laboral de los empleados - Inscribir a los empleados en planes de beneficios

Tabla 2.4: Ejemplos de Procesos de Negocios Funcionales[15]

Lo que en un principio parece un proceso simple, cumplir un pedido, resulta ser una serie bastante complicada de procesos de negocios que requieren la coordinación estrecha de los principales grupos funcionales en una empresa.

Lo que es más, para desempeñar con eficiencia todos estos pasos en el proceso de cumplimiento del pedido se requiere una gran cantidad de información, la cual debe fluir con rapidez, tanto dentro de la empresa desde un encargado de tomar decisiones a otro; con los socios de negocios, como las empresas de entrega; y

[15] [5],"Sistema de Información Gerencial"

con el cliente. Los sistemas de información basados en computadora hacen esto posible.

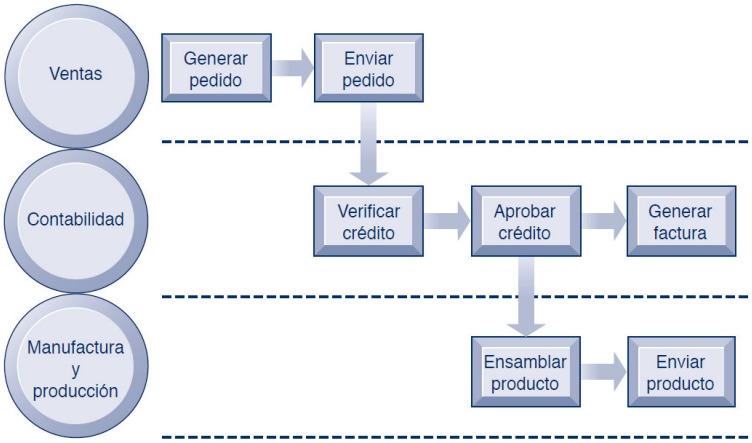


Figura 2.13: El proceso de cumplimiento de pedidos (Fuente: [5])

2.6.5. Negocio Electrónico, Comercio Electrónico y Gobierno Electrónico

Las aplicaciones empresariales como los ERP, SCM, CRM (figura 2.5) crean cambios muy arraigados en cuanto a la forma en que la empresa realiza sus actividades comerciales; ofrecen muchas oportunidades para integrar los datos de negocios importantes en un solo sistema. Con frecuencia son costosas y difíciles de implementar. Asimismo tenemos internet, intranet y extranet que son utilizadas como herramientas alternativas para incrementar la integración y agilizar el flujo de información dentro de la empresa, y con los clientes y proveedores.

El uso de tecnologías emergentes están transformando las relaciones de las empresas con los clientes, empleados, proveedores y socios de logística en relaciones digitales mediante el uso de redes e Internet.

El negocio electrónico, o e-business, se refiere al uso de la tecnología digital e Internet para ejecutar los principales procesos de negocios en la empresa. El e-

business incluye las actividades para la administración interna de la empresa y para la coordinación con los proveedores y otros socios de negocios. También incluye el comercio electrónico, o e-commerce.

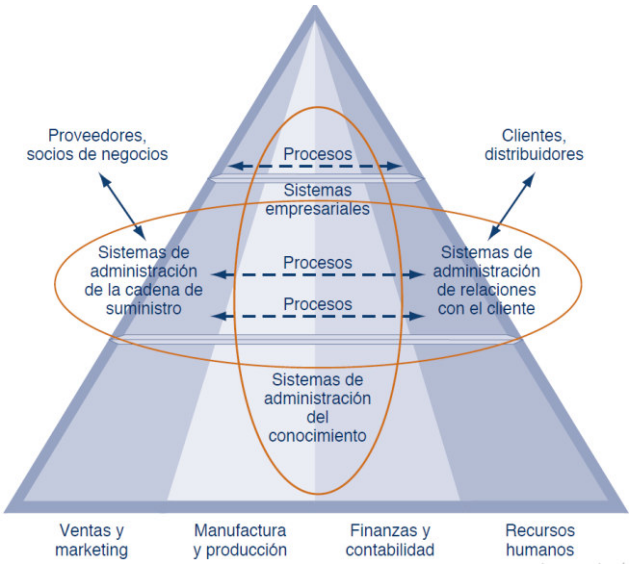


Tabla 2.5: Arquitectura de aplicaciones empresariales (Fuente: [5])

El e-commerce es la parte del e-business que trata sobre la compra y venta de bienes y servicios a través de Internet. También abarca las actividades que dan soporte a esas transacciones en el mercado, como publicidad, marketing, soporte al cliente, seguridad, entrega y pago.

Las tecnologías asociadas con el e-business también han provocado cambios similares en el sector público. Los gobiernos en todos los niveles están usando la tecnología de Internet para ofrecer información y servicios a los ciudadanos, empleados y negocios con los que trabajan. El gobierno electrónico, o e-government, se refiere a la aplicación de las tecnologías de Internet y de redes para habilitar de manera digital las relaciones del gobierno y las agencias del sector público con los ciudadanos, empresas y otras ramas del gobierno.

Además de mejorar el ofrecimiento de los servicios gubernamentales, el e-government aumenta la eficiencia de las operaciones del gobierno y también confiere a los ciudadanos el poder de acceder a la información con facilidad, junto con la habilidad de conectarse en red con otros ciudadanos por medios electrónicos. Por ejemplo, los ciudadanos en ciertos estados pueden renovar sus licencias de manejo o solicitar beneficios por desempleo en línea, e Internet se ha convertido en una poderosa herramienta para movilizar de manera instantánea los grupos de interés para acciones políticas y recaudación de fondos.

COLABORACION

Este término ampliamente utilizado en el sector empresarial y gubernamental consiste en trabajar con otros para lograr objetivos compartidos y explícitos. La colaboración puede ser de corto plazo, en donde dura unos cuantos minutos, o de un plazo más largo, dependiendo de la naturaleza de la tarea y de la relación entre los participantes. La colaboración puede ser de uno a uno o de varios a varios.

Entre los beneficios que se obtiene al aplicar la colaboración en la organización son: el incremento de la productividad, la mejora de la calidad del producto y o servicio que se brinda, la innovación, la mejora del servicio al cliente y mayor rentabilidad como resultado de lo anterior.

CAPITULO III. ESTADO DEL ARTE

3.1. Taxonomía de Enfoques de Minería de Datos Distribuida

En [3] se presenta una taxonomía de los enfoques de minería de datos distribuida. En la figura 3.1 se observa dos grupos: el primero conocido como coordinador centralizado este reúne tres subgrupos como: clustering distribuido, arm distribuido y clasificador de aprendizaje distribuido; el segundo grupo conocido como datamining peer to peer reúne dos subgrupos: datamining complejo y operaciones primitivas.

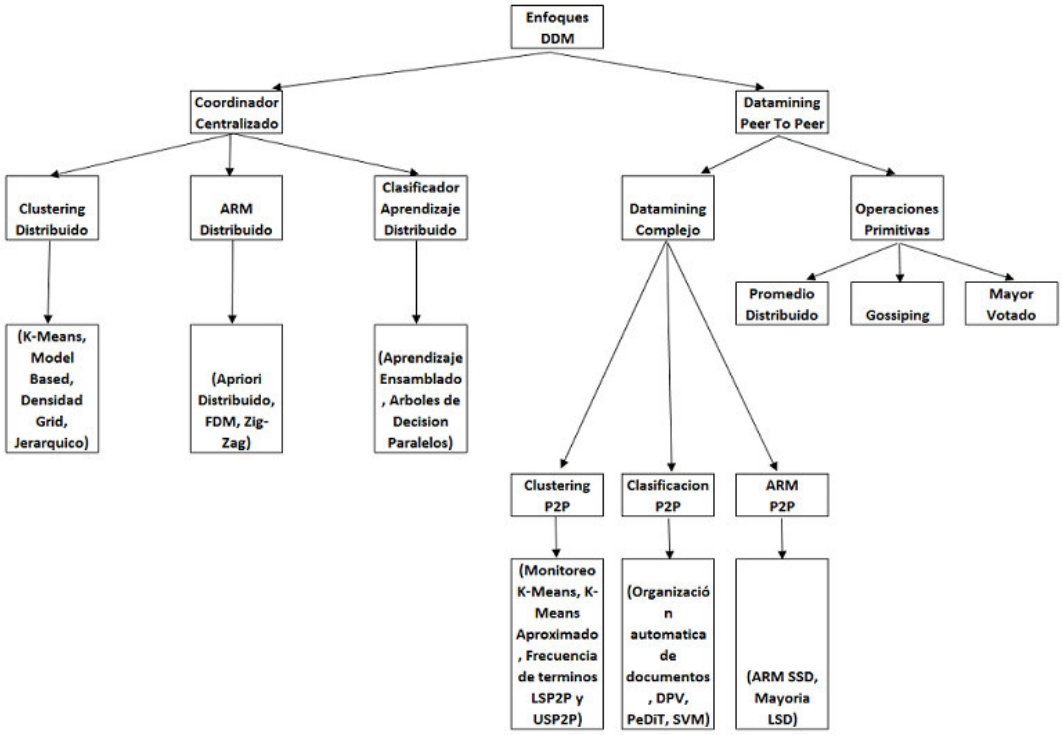


Figura 3.1: Taxonomía de Enfoques de Minería de Datos Distribuida[16]

[16] [3], "Survey on Distributed Data Mining in P2P Networks "

3.2. Algoritmos

A continuación describiremos los algoritmos de agrupamiento distribuido: k-means, modelo base, densidad grid y jerárquico.[2]

3.2.1. Algortimo k-means

1. Se eligen aleatoriamiente k centroides.
2. Los centroides son enviados a todos los nodos participantes; a continuación se realiza el agrupamiento k-means en cada nodo.
3. Cada nodo extrae información estadística de los elementos de sus grupos
4. Las estadísticas son transferidas hacia un controlador central quien se encargara de consolidar los modelos provenientes de los nodos locales.

El hecho de transferir solo información estadística hacia un nodo central y no la data completa permite mantener la confidencialidad y seguridad de la información. Uno de las desventajas de este modelo se basa en el hecho de tener que enviar en forma continua la información estadística de los nodos locales hasta lograr convergencia en los resultados lo cual puede generar tráfico en la red y hacer lento el proceso.

3.2.2. Modelo Base

Este algoritmo utiliza agrupamiento EM (expectation maximization) a nivel local, que es similar a K-means, excepto en que la decisión sobre el agrupamiento final se basa en el uso de funciones adicionales como la función de gaussiana.

Inicialmente, el sistema local procesa sus elementos individuales, mediante el algoritmo de agrupamiento EM local a continuación cada grupo es modelado

como una suma de funciones gaussianas. Las funciones resultantes son transferidas a un coordinador central, quien reúne las funciones para generar la función global sobre la densidad de la probabilidad de la imagen global.

Esta información se envía a cada nodo local con la finalidad que cada uno de ellos pueda utilizarla y reevaluar sus resultados de ser necesario.

El algoritmo emplea buenas medidas de confidencialidad y precisión. Sin embargo tiene un problema básico que consiste en que dos grandes grupos conectados mediante un componente con densidad mínima puede resultar constituyéndose en un mismo grupo sin que lo sea.

3.2.3. Densidad Grid

Este algoritmo hace uso del algoritmo CLIQUE con ciertas mejoras enfocadas en el agrupamiento distribuido.

El enfoque basado en densidad para el agrupamiento distribuido consiste en:

Inicialmente cada atributo definido en la consulta del usuario es explorado y en lugar de definir ciertos valores globales para el tamaño de un grid estos son determinados dinámicamente basándose en información estadística.

Los grupos son representados como grids rellenados y debido al proceso dinámico de cuadricular el área se tiene que en zonas de intensa densidad la granularidad es muy fina y en zonas de baja población la densidad es gruesa.

El algoritmo genera grupos sólidos, sin embargo estos asumen que la data está centralizada en un repositorio desde el cual se distribuye a todos los nodos.

3.2.4. Jerárquicos

El algoritmo jerárquico es muy similar al enfoque basado en “densidad grid”. La idea principal que persigue este algoritmo es empezar con un conjunto de puntos distintos, cada uno formando su propio grupo. A continuación empieza recursivamente a unir dos grupos cercanos hasta asegurar que todos los puntos lleguen a pertenecer a un mismo grupo. De este modo en algoritmos jerárquicos paralelos se utilizan dendogramas para crear grupos y sus distancias mínimas y máximas entre ellos. La unión de grupos se basa en distancias mínimas las cuales son transmitidas junto un TID (objeto identificador). La propiedad reducción es utilizada para crear el modelo global.

3.2.5. P2P Clustering

En [3] se hace referencia a otros autores en relación a investigaciones de algoritmos de agrupamiento P2P logrando considerar las siguientes propuestas:

Un algoritmo exacto para monitoreo de agrupamiento k-means

Este algoritmo consiste en monitorear la distribución de los centroides de los nodos locales dispersos y realizar el proceso k-means cuando se actualizan los grupos.

El algoritmo considera dos fases:

La primera fase consiste en monitorear la distribución de los datos mediante un algoritmo exacto.

La segunda fase consiste en calcular los centroides mediante un enfoque centralizado.

Algoritmo k-means basado en probar y hacer

Esta propuesta consiste en transmitir los centroides a todos los nodos en la red utilizando el mecanismo probar y hacer. Se requiere una sincronización de todos los nodos en cada iteración lo cual genera congestión en la red.

Algoritmo LSP2P basado en P2P K-means

Algoritmo de sincronización local. Esta propuesta distribuye los centroides a los nodos utilizando "informantes". Los centros de cada nodo son actualizados utilizando la información recibida de sus vecinos inmediatos. Este algoritmo produce resultados de agrupamiento de alta precisión pero esta no presenta garantía analítica.

Algoritmo USP2P basado en P2P K-means

Algoritmo de muestreo uniforme ofrece garantía probabilística mediante muestreo. Este algoritmo asume que la red es estática y busca lograr alta precisión.

Tanto LSP2P y USP2P asumen que la distribución de los datos en los nodos es uniforme. Sin embargo estas propuestas podrían no funcionar muy bien en redes grandes; asimismo si se orienta a colecciones de textos estos suelen no presentar distribuciones uniformes por tanto no funcionarían para estos casos.

Algoritmo de agrupamiento de texto para redes P2P basado en frecuencia de términos.

Primero se busca en todos los documentos el conjunto de términos de mayor frecuencia.

Para cada conjunto de términos se calcula los centroides locales.

Cada nodo envía un centroide global aproximado junto con las direcciones de sus vecinos.

Finalmente los documentos locales son agrupados basándose en el centro cluster inicial.

La sobrecarga de comunicación es muy baja y la calidad del grupo no se decrementaría si la cantidad de nodos en la red se incrementa.

Algoritmo distribuido eficiente basado en HDP2P

Este algoritmo ha sido pensado para realizar agrupamiento en grandes bases de datos, la propuesta se basa en una comunicación de los vecinos próximos y una sincronización local, con lo cual se logra obtener resultados de agrupamiento efectivo a bajos costos de comunicación.

3.2.6. Algoritmo K-means P2P

En [6] se propone un algoritmo iterativo basado en el intercambio de mensajes entre nodos conectados directamente para resolver el problema de agrupamiento k-means en redes P2P.

Se eligen aleatoriamente un conjunto de centroides y se distribuyen sobre todos los nodos.

Para cada iteración, cada nodo P_i ejecuta un proceso basado en dos pasos:

Paso 1:

Idéntico a la primera iteración del algoritmo k-means estándar; en el cual cada nodo P_i asigna cada uno de sus puntos a su centroide más cercano.

Sea $\{w^{(i)}_{j,k}\}$, $\forall$ j ε [1,k] el conjunto de centroides locales

Sea $|w^{(i)}_{j,k}|$, número de puntos en un nodo P_i en una iteración k

Cada nodo P_i almacena los centroides locales y el número de puntos en una iteración k con el objetivo de responder las consultas de sus vecinos.

Paso 2:

Un nodo P_i envía un mensaje a los nodos vecinos conteniendo su identificador y el número de la iteración actual en la cual se encuentra.

Id	K

Mensaje_consulta Id: Identificador del nodo, K: N° Iteración

Cada mensaje respuesta de un nodo vecino contiene el conjunto de centroides locales y el número de puntos de la definen para una iteración k

$\{w^{(i)}_{j,k}\}$	$\lvert w^{(i)}_{j,k}\rvert$
1	2

Mensaje_respuesta 1: Conjunto de centroides, 2: Número de Puntos en una iteración k

Una vez que todos los nodos vecinos de P_i responden o dejan de ser vecinos P_i actualiza sus centroides; realizando el cálculo del peso promedio de todos los centroides de los nodos vecinos más el suyo; a continuación pasa a la siguiente iteración.

Se repite Paso 1, Paso 2 hasta que los centroides de las iteraciones k y k+1 no presenten cambios significativos con lo cual el algoritmo habrá concluido.

Variantes:

Supongamos que el nodo P_i recibe un mensaje de consulta del nodo P_h durante la iteración k en P_i y k' en P_h

Si k'<=k

<u>Entonces</u>

El nodo Pi envía mensaje de la forma: mensaje_respuesta={$\{w^{(i)}_{j,k'}\}$,$|w^{(i)}_{j,k'}|$}

o en su defecto es asignado a una cola de espera de la forma:

cola_de_espera={{id,k'},….}

Fin_Si

Para cada iteración el nodo P_i verificara la cola de espera y responderá cualquier mensaje de poder hacerlo.

Cualquier nodo P_i puede pasar a un estado finalizado luego de concluir una iteración k si sus centroides de una iteración k y k+1 no presentan cambios significativos.

Los nodos P_i que llegaron al estado finalizado; podrán continuar atendiendo peticiones de consultas provenientes de otros nodos vecinos.

Ningún nodo P_i dispone de una condición explicita sobre la cual toda la actividad del algoritmo se pueda detener.

Una vez que todos los nodos entraron a un estado terminado, toda comunicación cesa; con lo cual el algoritmo habrá terminado.

El algoritmo puede ser ajustado para ciertas variantes como agregación de nuevos nodos a la red; modificaciones dinámicas en los datos.

En el caso de que se incorporasen nuevos nodos a la red, este nuevo nodo P_n puede unirse al procesamiento del algoritmo de agrupamiento en curso mediante la sincronización a la mínima iteración en curso de sus nodo vecinos.

Si algún nodo Pi presenta modificaciones dinámicas en sus datos, entonces se procede a recalcular los centroides y pasar a la siguiente iteración.

Se cuenta con varios experimentos de implementación del algoritmo lo cual han permitido determinar que la precisión del proceso de agrupamiento en comparación con el enfoque centralizado es mayor al 90%. Por otro lado la

escalabilidad del algoritmo es muy buena logrando mantener la misma precisión del agrupamiento agregando más nodos a la red.

3.3. Aplicativos

En [7] y [8], se presenta WEKA como una colección del estado del arte de algoritmos y herramientas de pre procesamiento de datos de aprendizaje maquina; WEKA son siglas de "Waikato Environment for Knowledge Analysis"; desarrollado en la Universidad de Waikato en Nueva Zelanda en el año 1993. En 1997 fue reimplementado en Java y distribuido bajo licencia GNU, En el año 2005 recibe el premio SIGKDD por su contribución al desarrollo de la Minería de Datos y al Descubrimiento de conocimiento. En el año 2006, la corporación Pentaho adquirió una licencia exclusiva para incorporar WEKA en su suite de inteligencia de negocios. A la fecha WEKA cuenta con un aproximado de 2487213 descargas lo que demuestra el interés por este framework.

Asimismo [7] menciona a grandes rasgos otros frameworks de código libre implementados en java e integrables con WEKA como son: YALE (2001) más conocido como RapidMiner, KNIME(2004) y ELKI(2008).

En [9] [10] y [11] se detallan frameworks como BIOWEKA, MEKA y MULAN; el primero está enfocado a incorporar funcionalidades y métodos bioinformáticos manteniendo el alineamiento con WEKA. MEKA presenta una variante frente a WEKA y está centrada en la implementación de métodos de clasificación multi-etiqueta así como MULAN; ambos frameworks tanto MEKA como MULAN están interesados en el estudio de datos multi-etiqueta; debido a que un ítem de datos puede ser miembro de múltiples categorías o definido por muchas etiquetas y/o clases; naturaleza actual de diversos problemas del mundo real como la

anotación semántica de imágenes y videos; categorización de sitios web marketing directo, genómica funcional y categorización de la música en géneros y emociones.

En [12] se desarrolla ampliamente las herramientas de minería de datos disponibles en el mercado; desde herramientas propietarias hasta las de código libre.

Las herramientas comerciales son amplias de manera que se ha consignado mediante dos listas; en las tres figuras se precisa el nombre de la herramienta, el tipo de la herramienta y la url del sitio web respectivo.

Tool	Type	Link
ADAPA (Zementis)	DMS	www.zementis.com
Alice (d'Isoft)	DMS	www.alice-soft.com
Bayesia Lab	SPEC	www.bayesia.com
C5.0	SPEC	www.rulequest.com
CART	SPEC	www.salford-systems.com
Data Applied	DMS	data-applied.com
DataDetective	DMS	www.sentient.nl/?dden
DataEngine	DMS	www.dataengine.de
Datascope	DMS	www.cygron.hu
DB2 Data Warehouse	BI	www.ibm.com/software/data/infosphere/warehouse
DeltaMaster	BI	www.bissantz.com/deltamaster
Forecaster XL	EXT	www.alyuda.com
GhostMiner	DMS	www.fqs.pl/business_intelligence/products/ghostminer
IBM Cognos 8 BI	BI	www.ibm.com/software/data/cognos/data-mining-tools.html
IBM SPSS Modeler	DMS	www.spss.com/software/modeling/modeler
IBM SPSS Statistics	MAT	www.spss.com/software/statistics
iModel	DMS	www.biocompsystems.com/products/imodel
InfoSphere Warehouse	BI	www.ibm.com/software/data/infosphere/warehouse
JMP	DMS	www.jmpdiscovery.com
KnowledgeMiner	SPEC	www.knowledgeminer.net
KnowledgeStudio	DMS	www.angoss.com
KXEN	DMS	www.kxen.com
Magnum Opus	SPEC	www.giwebb.com
MATLAB	MAT	www.mathworks.com
MATLAB Neural Network Toolbox	EXT	www.mathworks.com
Model Builder	DMS	www.fico.com
ModelMAX	SOL	www.asacorp.com/products/mmxover.jsp

Figura 3.2: Lista de Herramientas comerciales (1)[17]

El autor ha clasificado las herramientas según tipo:

DMS: Suite de minería de datos; este tipo de herramienta está ampliamente enfocado en minería de datos e incluye gran número de métodos.

[17] [12],"Datamining tools"

SPEC: Es un tipo de herramienta de minería de datos pero enfocada en un método en particular como por ejemplo redes neuronales.

MAT: Este tipo de herramienta no está enfocada precisamente en minería de datos pero ofrece un conjunto de métodos y rutinas de visualización de resultados.

BI: No está enfocada estrictamente en minería de datos pero incluye métodos básicos de minería de datos orientadas a métodos estadísticos para aplicaciones de negocios.

SOL: Describe un conjunto de herramientas personalizables a diversos campos de aplicación como text mining, procesamientos de imágenes, descubrimiento de fármacos; análisis de imágenes en microscopios. La ventaja de estas soluciones es el buen soporte de las técnicas de extracción de características específicas, evaluación de resultados, visualización e importación de formatos.

Tool	Type	Link
Molegro Data Modeler	SOL	www.molegro.com
NAG Data Mining Components	LIB	www.nag.co.uk/numeric/DR/DRdescription.asp
NeuralWorks Predict	SPEC	www.neuralware.com/products.jsp
Neurofusion	LIB	www.alyuda.com
Neuroshell	SPEC	www.neuroshell.com
Oracle Data Mining (ODM)	DMS	www.oracle.com/technology/products/bi/odm/index.html
Partek Discovery Suite	DMS	www.partek.com/software
Partek Genomics Suite	SOL	www.partek.com/software
PolyAnalyst	DMS	www.megaputer.com/polyanalyst.php
PolyVista	BI	www.polyvista.com
Random Forests	SPEC	www.salford-systems.com
RapAnalyst	SPEC	www.raptorinternational.com/rapanalyst.html
R-PLUS	MAT	www.experience-rplus.com
SAP Netweaver Business Warehouse (BW)	BI	www.sap.com/platform/netweaver/components/businesswarehouse
SAS Enterprise Miner	DMS	www.sas.com/products/miner
See5	SPEC	www.rulequest.com
SPAD Data Mining	DMS	eng.spadsoft.com
SQL Server Analysis Services	DMS	www.microsoft.com/sql
STATISTICA	DMS	www.statsoft.com/products/data-mining-solutions/G259
SuperQuery	DMS	www.azmy.com
Teradata Database	BI	www.teradata.com
Think Enterprise Data Miner (EDM)	DMS	www.thinkanalytics.com
TIBCO Spotfire	DMS	spotfire.tibco.com
Unica PredictiveInsight	DMS	www.unica.com
WizRule and WizWhy	SPEC	www.wizsoft.com
XAffinity	SPEC	www.exclusiveore.com

Figura 3.3: Lista de Herramientas Comerciales (2)[18]

[18] [12],"Datamining Tools"

La figura 3.4 presenta una lista de herramientas de minería de datos de código libre clasificados según tipo; en la lista se puede observar la herramienta weka; cuyas librerías han sido utilizadas en el proyecto de la presente investigación.

Tool	Type	Link
ADaM*	LIB	datamining.itsc.uah.edu/adam
CellProfilerAnalyst	SOL	www.cellprofiler.org/index.htm
D2K*	DMS	alg.ncsa.uiuc.edu
Gait-CAD	INT	sourceforge.net/projects/gait-cad
GATE	SOL	gate.ac.uk/download
GIFT	RES	www.gnu.org/software/gift
Gnome Data Mine Tools	DMS	www.togaware.com/datamining/gdatamine
Himalaya	RES	himalaya-tools.sourceforge.net
ImageJ	SOL	rsbweb.nih.gov/ij
ITK	SOL	www.itk.org
JAVA Data Mining Package	LIB	sourceforge.net/projects/jdmp
JavaNNS	SPEC	www.ra.cs.uni-tuebingen.de/software/JavaNNS/welcome_e.html
KEEL	INT	www.keel.es
Kepler	MAT	kepler-project.org
KNIME	INT	www.knime.org
LibSVM	LIB	www.csie.ntu.edu.tw/ cjlin/libsvm
MEGA	SOL	www.megasoftware.net/m_distance.html
MLC++	LIB	www.sgi.com/tech/mlc
Orange	LIB	www.ailab.si/orange
Pegasus	RES	www.cs.cmu.edu/ pegasus
Pentaho	BI	sourceforge.net/projects/pentaho
Proximity	SPEC	kdl.cs.umass.edu/proximity/index.html
PRTools	EXT	www.prtools.org
R	MAT	www.r-project.org
RapidMiner	DMS	www.rapidminer.com
Rattle	INT	rattle.togaware.com
ROOT	LIB	root.cern.ch/root
ROSETTA	SPEC	www.lcb.uu.se/tools/rosetta/index.php
Rseslibs	RES	logic.mimuw.edu.pl/ rses
Rule Discovery System*	SPEC	www.compumine.com
RWEKA	INT	cran.r-project.org/web/packages/RWeka/index.html
TANAGRA	INT	eric.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html
Waffles	LIB	waffles.sourceforge.net
WEKA	DMS, LIB	sourceforge.net/projects/weka
XELOPES Library*	LIB	www.prudsys.de/en/technology/xelopes

Figura 3.4: Lista de Herramientas de Código Libre[19]

[19] [12],"Datamining Tools"

3.4. Casos de estudio

En [5] se describe el caso de la compañía Hallmark Cards que utiliza la inteligencia de negocios esta utiliza el software de SAS Analytics para mejorar la comprensión de los patrones de compras de sus clientes que podrían conducir a un aumento de las ventas en las más de 3 000 tiendas Hallmark Gold Crown en Estados Unidos. Hallmark quería fortalecer su relación con sus compradores frecuentes. Mediante el uso de la minería de datos y el modelado predictivo, la compañía determinó cómo comercializar con varios segmentos de consumidores durante los días festivos y ocasiones especiales, además de que aprendió a ajustar las promociones de manera improvisada. Hallmark puede determinar qué segmentos de clientes se dejan influir más por el correo directo, cuándo es mejor usar el correo electrónico y qué mensajes específicos debe enviar a cada grupo.

Asimismo la compañía Capital One realiza más de 30 000 experimentos cada año en donde utiliza distintas tasas, incentivos, paquetes de correo directo y otras variables para identificar a los mejores clientes potenciales a quienes debe dirigir sus ofrecimientos de tarjetas de crédito.

Es más probable que estas personas contraten tarjetas de crédito y paguen a Capital One los saldos que acumulen en sus cuentas. El análisis predictivo también ha funcionado muy bien en la industria de las tarjetas de crédito para identificar a los clientes con riesgo de cancelar sus cuentas.

Dealer Services, que ofrece financiamiento de inventario para los concesionarios de autos usados, trata de usar el análisis predictivo para investigar a sus clientes potenciales. Miles de concesionarios de autos usados, que antes tenían franquicias de General Motors y Chrysler, buscan financiamiento de compañías

como Dealer Services para poder hacer sus propios negocios. Mediante el uso del software WebFOCUS de Information Builders, la compañía está creando un modelo que predecirá los mejores prospectos de préstamos y eliminará entre 10 y 15 de las horas que se requieren para revisar una aplicación financiera. El modelo revisa los datos como el tamaño y tipo de concesionario, el número de oficinas, los patrones de pago, el historial de cheques devueltos sin fondos y las prácticas de inventario, todo lo cual se revalida y actualiza a medida que cambian las condiciones.

FedEx utiliza el software Enterprise Miner de SAS Institute junto con las herramientas de análisis predictivo para desarrollar modelos que pronostiquen cómo responderán los clientes a los cambios en los precios y a los nuevos servicios, cuáles clientes presentan un mayor riesgo de cambiar a la competencia y cuántos ingresos se generarán debido a las nuevas ubicaciones de las sucursales o buzones.

A nivel nacional se cuenta con investigaciones relacionadas al presente tema los cuales se describen a continuación:

En la tesis de los ingenieros en sistemas Arturo Eguila Canales y Alex Parco Iquiapaza (2007) titulada " Implementación de una herramienta de inteligencia de negocios para la administración de justicia sobre una metodología ad-hoc ", se ha llegado a las siguientes conclusiones:

La investigación se enfoca en el análisis y diseño para la implementación de una herramienta de inteligencia de negocios para la toma de decisiones en el área de defensoría de oficio del ministerio de justicia, con el objetivo de obtener un

mejor control y gestión de la información del sistema de defensores de oficio de forma que ayude a mejorar la calidad del servicio que presta la entidad.

El trabajo de investigación describe las siguientes conclusiones:

Se presentó el diseño de un datamart (prototipo) para la gestión del área de la defensoría del ministerio de justicia.

La implementación del datamart ayudara a los directores en las consultas para satisfacer sus requerimientos de forma sencilla a través de una herramienta de fácil aprendizaje.

En la tesis de los ingenieros en sistemas Ángeles Bocanegra, Oscar y Melgarejo Quispe, César (2012) titulada " Algoritmo de Clustering utilizando K-Means e índice de validación Rose Turi para la segmentación de clientes de la Caja Rural PRYMERA ", se ha llegado a las siguientes conclusiones:

Se implementó el algoritmo K-Means como técnica para obtener grupos de clientes, y el índice de Rose Turi como indicador de la cantidad óptima de grupos, obteniéndose la mejor distribución de clientes en segmentos que comparten características similares internamente y a su vez son heterogéneos entre sí.

Mediante la implementación del índice de Rose Turi se mejoró la eficacia del algoritmo K-Means, dado que ya no es necesario que se tenga un conocimiento previo del número de clusters que se quiere obtener, el cual muchas veces es brindado de forma empírica por los usuarios del algoritmo. Mediante la implementación del índice se tiene una certeza del 100% de que se obtuvo la óptima cantidad de clusters.

Se realizó un análisis de la eficiencia del índice de Rose Turi versus el índice de Davies-Bouldin, tomando como criterio de comparación el tiempo de procesamiento de un Dataset con 5000 registros, obteniéndose después de varias corridas que el índice de Rose Turi brindó resultados en un 17% más rápido que el índice de Davies-Bouldin.

De acuerdo a la naturaleza matemática del algoritmo K-Means y el índice de Rose Turi, se puede aplicar la solución planteada a diversos tipos de problema, sólo se debe tener en cuenta que los datos del Dataset deben ser numéricos.

A la fecha tanto a nivel nacional como internacional se ha desarrollado una diversidad de trabajos en relación a minería de datos; sin embargo la minería de datos distribuida aún sigue siendo investigada y se cuenta con determinados pilotos dada su complejidad.

CAPITULO IV. DESARROLLO DE LA PROPUESTA

El presente trabajo está enfocado en aplicar la minería de datos distribuida en una organización moderna con la finalidad de mejorar la calidad de los servicios que esta brinda. Para el desarrollo de la misma se ha enfocado en un tipo de organización particular: la organización pública y muy puntualmente en la organización judicial y su proceso de negocio petitorio.

La metodología de desarrollo que se aplica consiste en:

A.- Analizar el caso de estudio

Este primer paso nos permitirá conocer la organización y el proceso de negocio seleccionado para su análisis.

B.- Elaborar el Modelo dimensional

Este segundo paso consolidara la comprensión del proceso de negocio; como resultado se tiene el diseño del modelo dimensional sobre el cual se circunscribe la aplicación y los resultados.

C.- Aplicación DDM

Este tercer paso plantea tres propuestas algorítmicas basadas en el clustering e implementa una de ellas mediante un prototipo.

D.- Determinar resultados

El último paso revelara los resultados alcanzados luego de aplicar los tres pasos anteriores; asimismo permitirá fundamentar los objetivos del presente trabajo.

4.1. Análisis del Caso de Estudio

El caso de Estudio del presente trabajo está enfocado en la organización judicial del Estado Peruano.

"El Estado Peruano es la organización jurídico-política, de la sociedad concebida como Nación. Incluye su gobierno, sus instituciones públicas, sus leyes y las reglas de juego válidas para la vida social en general." [13]

"El Estado es uno e indivisible. Su gobierno es unitario, representativo y descentralizado, y se organiza según el principio de la separación de poderes." [14]

En la figura 4.1 se presenta la estructura del estado peruano. La organización del estado responde al principio de la división de poderes: Poder Ejecutivo, Poder legislativo y el Poder Judicial.

El Poder Ejecutivo está representado por el presidente de la república, este es el jefe del estado y personifica a la nación.

La dirección y la gestión de los servicios públicos están confiadas al Consejo de Ministros; y a cada ministro en los asuntos que competen a la cartera a su cargo.

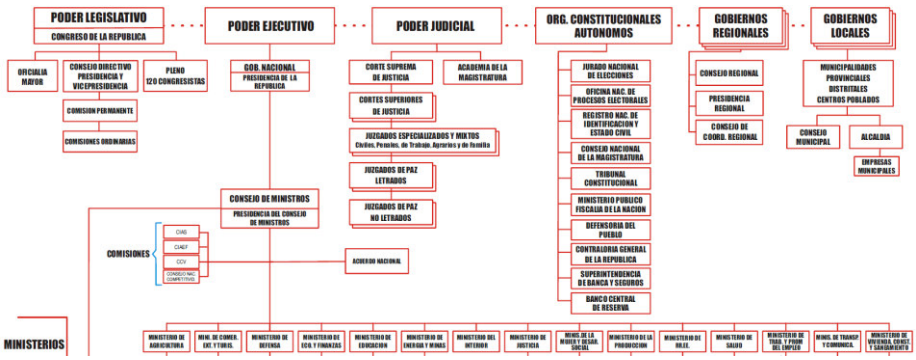


Figura 4.1: Parte del organigrama del Estado Peruano
[Fuente: (tomado de [15])]

El Plan Nacional de Modernización de la Gestión Pública (PNMGP)

"..Es un instrumento orientador de la modernización de la gestión pública en el Perú, este establece la visión, los principios y los lineamientos para una actuación coherente y eficaz del sector público, al servicio de los ciudadanos y el desarrollo del país. El PNMGP está dirigida a todas las entidades del Poder Ejecutivo nacional, Organismos autónomos así como Gobiernos Regionales y Locales, sin afectar las autonomías que les confiere la ley, entidades todas que están llamadas a formular planes y emprender acciones de modernización de su gestión a fin de mejorar su desempeño al servicio de los ciudadanos." [15]

El PNMGP analiza la situación general de los últimos diez años del estado; peruano, determina las principales deficiencias de la gestión pública en el Perú; uno de los ítems resaltantes y de nuestro interés es *"Carencia de sistemas y métodos de gestión de la información y el conocimiento"*; describe la implicancia de la gestión del conocimiento en la organización publica y que en el Estado no existe de manera institucionalizada un sistema de gestión de la información y del conocimiento, no existe un sistema de recopilación y trasferencia de buenas prácticas; por ello se repiten los mismos errores y se redunda en la búsqueda de soluciones a problemas que ya han sido resueltos. Otro de los ítems de nuestro interés radica en *"Débil articulación intergubernamental e intersectorial"*; Ello se refiere a la colaboración de los procesos de las diversas entidades públicas las cuales debieran complementarse a través de un mecanismo de integración digital.

Este instrumento PNMGP define como visión a un Estado moderno al servicio de las personas: orientado al ciudadano, Eficiente, Unitario y descentralizado, Inclusivo y Abierto de manera que garantice un acceso a bienes y servicios públicos de calidad.

La figura 4.2 presenta cinco pilares y tres ejes transversales los cuales permitirá lograr los objetivos del PNMGP. Los cinco pilares comprende: i) La políticas públicas nacionales y el planeamiento, ii) El presupuesto por resultados, iii) La gestión por procesos y la organización institucional iv) El servicio civil meritocrático y v) El seguimiento, monitoreo, evaluación y la gestión del conocimiento.

Figura 4.2: Pilares centrales de la Política de Modernización de la gestión pública (Fuente: [15])

Los ejes transversales son: i) El gobierno Abierto, ii) El Gobierno electrónico y iii) Gobierno institucional; un gobierno colaborativo multinivel orientado hacia una gestión de cambio.

El Gobierno Electrónico es uno de los tres ejes transversales considerado en los pilares centrales de la política de modernización de la gestión pública por considerarse como un instrumento de valor el cual mediante el uso de las Tecnologías de Información y las comunicaciones(TICs) en los órganos de la

administración pública, permite mejorar los servicios ofrecidos a los ciudadanos, orientar la eficacia y eficiencia de la gestión pública, incrementar sustantivamente la transparencia del sector público, la participación de los ciudadanos y el impulso de un gobierno abierto al servicio del estado.

El Poder Legislativo reside en el Congreso de la República, tiene como atribución velar por el respeto de la constitución y de las leyes.

El Poder Judicial ejerce la administración de justicia a través de sus órganos jerárquicos con arreglo a la constitución y a las leyes.

Principios de la Administración de Justicia

En [14] se detalla los principios y derechos de la función jurisdiccional las que se listan a continuación:

1. La unidad y exclusividad de la función jurisdiccional.

No existe ni puede establecerse jurisdicción alguna independiente, con excepción de la militar y la arbitral.

No hay proceso judicial por comisión o delegación.

2. La independencia en el ejercicio de la función jurisdiccional.

Ninguna autoridad puede avocarse a causas pendientes ante el órgano jurisdiccional ni interferir en el ejercicio de sus funciones. Tampoco puede dejar sin efecto resoluciones que han pasado en autoridad de cosa juzgada, ni cortar procedimientos en trámite, ni modificar sentencias ni retardar su ejecución. Estas disposiciones no afectan el derecho de gracia ni la facultad de investigación del

Congreso, cuyo ejercicio no debe, sin embargo, interferir en el procedimiento jurisdiccional ni surte efecto jurisdiccional alguno.

3. La observancia del debido proceso y la tutela jurisdiccional.

Ninguna persona puede ser desviada de la jurisdicción predeterminada por la ley, ni sometida a procedimiento distinto de los previamente establecidos, ni juzgada por órganos jurisdiccionales de excepción ni por comisiones especiales creadas al efecto, cualquiera sea su denominación.

4. La publicidad en los procesos, salvo disposición contraria de la ley.

Los procesos judiciales por responsabilidad de funcionarios públicos, y por los delitos cometidos por medio de la prensa y los que se refieren a derechos fundamentales garantizados por la Constitución, son siempre públicos.

5. La motivación escrita de las resoluciones judiciales en todas las instancias, excepto los decretos de mero trámite, con mención expresa de la ley aplicable y de los fundamentos de hecho en que se sustentan.

6. La pluralidad de la instancia.

7. La indemnización, en la forma que determine la ley, por los errores judiciales en los procesos penales y por las detenciones arbitrarias, sin perjuicio de la responsabilidad a que hubiere lugar.

8. El principio de no dejar de administrar justicia por vacío o deficiencia de la ley.

En tal caso, deben aplicarse los principios generales del derecho y el derecho consuetudinario.

9. El principio de inaplicabilidad por analogía de la ley penal y de las normas que restrinjan derechos.

10. El principio de no ser penado sin proceso judicial.

11. La aplicación de la ley más favorable al procesado en caso de duda o de conflicto entre leyes penales.

12. El principio de no ser condenado en ausencia.

13. La prohibición de revivir procesos fenecidos con resolución ejecutoriada. La amnistía, el indulto, el sobreseimiento definitivo y la prescripción producen los efectos de cosa juzgada.

14. El principio de no ser privado del derecho de defensa en ningún estado del proceso.

Toda persona será informada inmediatamente y por escrito de la causa o las razones de su detención. Tiene derecho a comunicarse personalmente con un defensor de su elección y a ser asesorada por éste desde que es citada o detenida por cualquier autoridad.

15. El principio de que toda persona debe ser informada, inmediatamente y por escrito, de las causas o razones de su detención.

16. El principio de la gratuidad de la administración de justicia y de la defensa gratuita para las personas de escasos recursos; y, para todos, en los casos que la ley señala.

17. La participación popular en el nombramiento y en la revocación de magistrados, conforme a ley.

18. La obligación del Poder Ejecutivo de prestar la colaboración que en los procesos le sea requerida.

19. La prohibición de ejercer función judicial por quien no ha sido nombrado en la forma prevista por la Constitución o la ley. Los órganos jurisdiccionales no pueden darle posesión del cargo, bajo responsabilidad.

20. El principio del derecho de toda persona de formular análisis y críticas de las resoluciones y sentencias judiciales, con las limitaciones de ley.

21. El derecho de los reclusos y sentenciados de ocupar establecimientos adecuados.

22. El principio de que el régimen penitenciario tiene por objeto la reeducación, rehabilitación y reincorporación del penado a la sociedad.

Organigrama Estructural del Poder Judicial

En la figura 4.3 se presenta como se encuentra organizado el sector judicial en el Perú. La presidencia del Poder Judicial recae en el Presidente de la Corte Suprema y es La Sala Plena de la Corte Suprema el órgano máximo de deliberación del Poder Judicial.

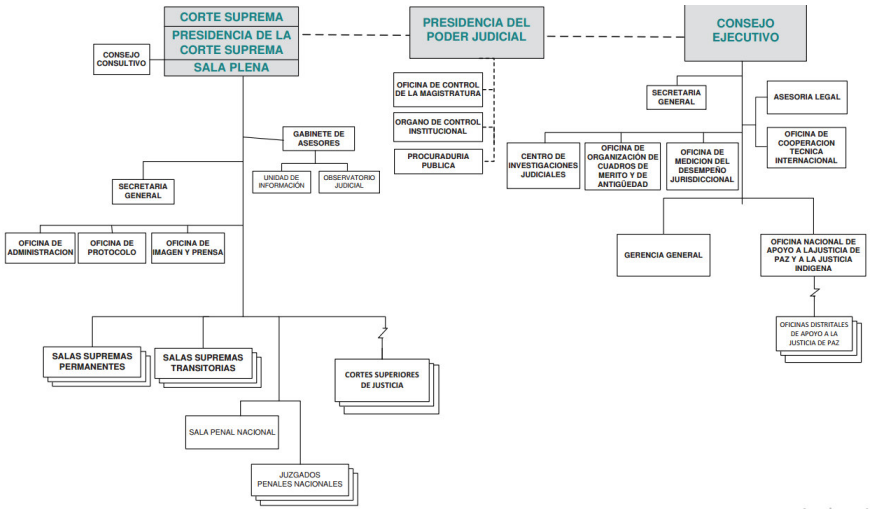


Figura 4.3: Organigrama del Poder Judicial
(Fuente: [16])

A continuación se describirá algunos términos de importancia de la presente investigación:

Órganos jurisdiccionales:[20]

El Poder Judicial está integrado por órganos jurisdiccionales que administran justicia en nombre de la Nación, y por órganos que ejercen su gobierno y administración.

Los órganos jurisdiccionales son: la Corte Suprema de Justicia y las demás cortes y juzgados que determine su ley orgánica.

En la figura 4.4 presenta los distritos judiciales los cuales agrupan órganos jurisdiccionales a nivel nacional quienes desarrollan la actividad jurisdiccional de su competencia.

[20] [14], “Constitución Política del Perú; Art 143”

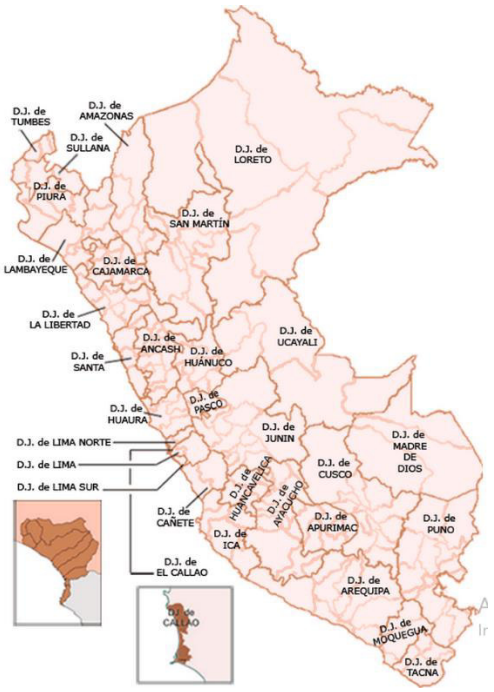


Figura 4.4: Distritos judiciales del Poder Judicial
(Fuente: [17])

Casación:

Es un término asignado a la última instancia, cuando la acción se inicia en una Corte Superior o ante la propia Corte Suprema conforme a ley. Asimismo, conoce en casación las resoluciones del Fuero Militar, con las limitaciones que establece el artículo 173 de la constitución política del Perú.

Petitorio:

"El petitum es lo que se pide sea reconocido o declarado en la sentencia a favor del demandante."[21]

"La petición es la declaración de voluntad, integra el contenido sustancial de la pretensión, determinando los límites cuantitativos (acumulación de pretensiones)

[21] [18]," El petitorio implícito en los procesos de familia: A propósito del tercer pleno casatorio, Pag 4"

y cualitativos (naturaleza de la pretensión: declarativa, constitutiva o de condena) del deber de congruencia del fallo, la parte dispositiva de la sentencia (Sendra, 2007: 209; citado por [18]*)".*

"El petitum es el elemento fundamental de la pretensión del actor en relación con la congruencia de la sentencia ya que ni su objeto inmediato ni mediato puede modificarse a lo largo del proceso ni en la resolución judicial. En pocas palabras, la sentencia debe inexcusablemente ser congruente con la petición." (Ezquiaga, 2000: 53; Sendra, 2007: 209-210; citado por [18]*).*

Agenda Estratégica 2013-2014

El organismo judicial peruano ha sido considerado en el presente estudio por tratarse de un modelo de organización moderna cuyos objetivos se encuentran establecidos en la agenda estratégica 2013-2014; documento que a su vez se encuentra alineado al Plan de desarrollo Institucional al 2018, al Plan Bicentenario y al Plan Nacional para la Reforma Integral de la Administración de Justicia. Esta agenda está enfocada en tres ejes: El eje ciudadano, El eje interno y el eje externo.

A continuación definimos extractos de la agenda; básicamente aquellos centrados en el uso de tecnologías de información:

El eje ciudadano tiene como uno de los objetivos principales la predictibilidad de las decisiones judiciales para lo cual se define un conjunto de acciones estratégicas.

El eje interno precisa como objetivos primordiales: Gestión de la calidad jurisdiccional, El uso de las tecnologías de información y justicia electrónica con monitoreo permanente, El uso y análisis de la información estadística y de los indicadores de producción y de calidad.

Como se puede apreciar el organismo judicial está muy interesado en aprovechar las tecnologías de la información y las comunicaciones para la mejora de los servicios que brinda a los ciudadanos. De esta manera lograr el acceso e inclusión social en la impartición de justicia. [19].

La reducción de los tiempos procesales permitirá la reducción de la carga procesal del sector judicial es otro de los objetivos de dicha institución para lo cual se viene realizando estudios de los proceso judiciales que presentan mayor

carga. En la figura 4.5 se presenta cuatro de los diez procesos identificados con mayor carga procesal de todas las especialidades[22] en seis sedes judiciales.

Figura 4.5: Procesos con mayor carga procesal[23]

[22] [27], “plan de mejora de los servicios judiciales, pag. 34-35”

4.2. Modelo Dimensional

La elaboración del diseño del modelo dimensional seguirá el procedimiento basado en cuatro pasos de Kimball[20]:

1. Seleccionar el proceso del negocio a modelar
2. Definir el nivel de granularidad del proceso del negocio
3. Escoger las dimensiones que aplican en cada fila de la tabla de hechos
4. Identificar los hechos numéricos que poblaran la tabla de hechos

Seleccionar el proceso del negocio a modelar

Uno de los objetivos de la organización judicial es la reducción de la carga procesal, lograr la predictibilidad en las decisiones judiciales mediante la jurisprudencia uniforme.

El proceso de negocio está centrado en el petitorio que realiza cada parte interesada en el caso judicial; aquella que inicia el proceso, la parte demandante.

Definir el nivel de granularidad del proceso del negocio

La granularidad está determinada por cada transacción; esto quiere decir que cada unidad mínima de la carga procesal generara mínimamente una entrada en la tabla de hechos del datamart.

Escoger las dimensiones que aplican en cada fila de la tabla de hechos

Se ha considerado ocho dimensiones primarias: dim_cuadernillo, dim_recurso, dim_especialidad, dim_materia, dim_organo_jurisdiccional, dim_tiempo, dim_tema, dim_norma.

En la tabla 4.1 se describe las dimensiones a considerar en el modelo propuesto:

Dimensión		Descripción
D01 dim_cuadernillo:	-	Contiene información concerniente al expediente judicial
D02 dim_recurso:	-	Contiene información referente al recurso formulado en el petitorio: Casación, Apelación, Queja, etc.
D03 dim_especialidad:	-	La especialidad a la cual corresponde el proceso: Civil, Constitucional, Penal, Laboral, Familia, entre otros.
D04 dim_materia:	-	Clase a la que pertenece el proceso
D05 dim_organo_jurisdiccional:	-	Corresponde a las Salas y juzgados que integran el organismo judicial según la ley orgánica del mismo, cada caso judicial abierto está vinculado a un órgano jurisdiccional.
D06 dim_tiempo:	-	Contiene información referente al calendario; esta permite analizar la información basándose en periodos de evaluación.
D07 dim_tema	-	Contiene información referente al pedido que realiza la parte procesal; no puede cambiar durante el tiempo de vida del proceso judicial.
D08 dim_norma	-	Corresponde a la parte normativa, fundamentos de ley, sobre la cual las partes procesales fundamentan su pedido.
F01 fact_petitorio	-	Contiene los hechos concerniente al proceso de negocio “Petitorio”

Tabla 4.1: Descripción de Dimensiones y Hechos de la propuesta
(Fuente: Elaboración propia)

Todas las dimensiones definen una llave artificial; no se hará uso de la llave natural por temas de estandarización de los datos y desempeño.

El modelo está pensado para ambos enfoques de minería de datos; tanto la centralizada así como la distribuida igualmente debería funcionar sin tener que realizar mayores cambios al modelo.

Identificar los hechos que poblaran la tabla de hechos

Adicionalmente a las dimensiones una peculiaridad está dada en el uso de una dimensión degenerada: cuadernillo_id; esto producto de la granularidad fina

definida para el modelo; asimismo se ha incorporado una clave artificial, numérica, correlativa: petitorio_id con el objetivo de mantener un desempeño de consulta aceptable así como lograr una independencia con respecto a las llaves naturales de los ambientes transaccionales ,los cuales serán las fuentes que permitirán poblar el modelo dimensional.

La arquitectura según la normalización de sus dimensiones del modelo dimensional resultante es un esquema copo de nieve. La figura 4.6 presenta el diseño propuesto.

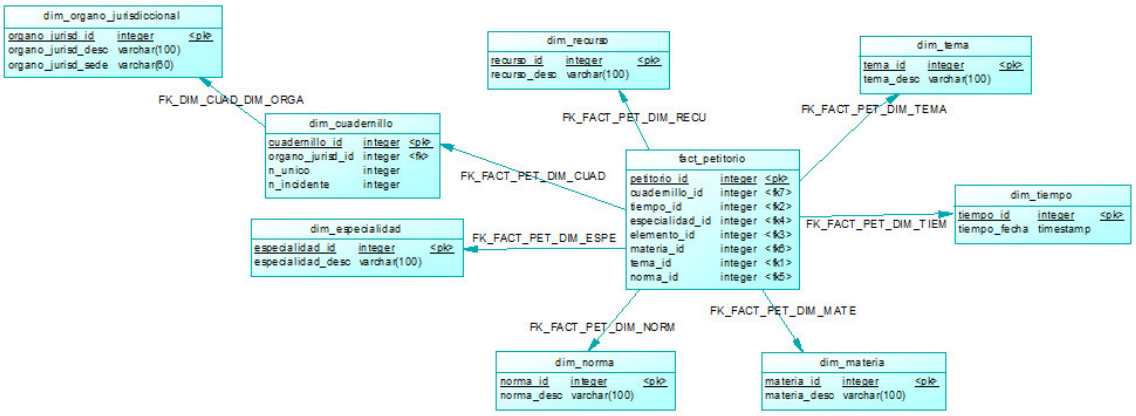


Figura 4.6: Modelo Dimensional Copo de Nieve
(Fuente: Elaboracion Propia)

4.3. Aplicación MDD

Considerando el esquema distribuido que presenta el negocio en cuestión; una base de datos en cada sede judicial, a ello se suma la sensibilidad y confidencialidad de la información que este tipo de organización maneja, infraestructura de tecnologías de información legadas; todo ello nos orienta al uso de minería de datos distribuida con lo cual se mantendría la confidencialidad de la información al evitar su centralización, se reutilizaría la infraestructura

tecnológica y arquitectura de datos actual que presenta el negocio, se requiere mínimo presupuesto y se podrían ver resultados a corto plazo.

La técnica de minería de datos elegida es K-medias; esta técnica pertenece a la clase de técnicas de tipo clustering o agrupamiento la cual permitirá agrupar los hechos del datamart propuesto según patrones de afinidad. El número de grupos o clusters resultantes a obtener es configurable en nuestro caso definiremos cinco clusters inicialmente.

La técnica K-medias presenta una serie de variantes de implementación; por tanto se formulan tres propuestas aplicables al contexto del presente estudio.

Algoritmo propuesta A (APA)

La primera propuesta consiste en el siguiente flujo:

1.- Sea $\Phi = \cup_{i=1}^{n} \beta_i \quad \forall\, i\, \varepsilon\, [1, n]$ el conjunto global de la información distribuida en n sedes, donde cada β_i representa la base de datos de la sede i.

2.- Sea μ el modelo dimensional propuesto el cual se considera desplegado en las n sedes.

3.- Se elige las sedes a incluir en la evaluación y se indicara el número de clusters a generar.

4.- Sea $\delta = \cup_{k=1}^{m} d_k \quad \forall\, k\, \varepsilon\, [1, m]$ el dataset resultante de la ejecución de μ en las m sedes seleccionadas; considérese m<=n respectivamente asimismo las instancias de δ son datos cualitativos y nominales.

5.- De δ elegir c centroides aleatoriamente

6.- Determinar c conglomerados considerando la distancia euclidiana de la instancia i al centroide s ($I_i \rightarrow C_s$)

$\forall\, I_i\, \varepsilon\, \delta$; determinar la distancia de I_i a cada centroide considerándose que la instancia i pertenece al conglomerado s si y solo si la distancia de la instancia i al centroide s es la mínima distancia determinada del conjunto de distancias calculadas para i con respecto al conjunto de c centroides.

$$I_i\, \varepsilon\, C_s \;\therefore distancia(I_i, C_s) = \min\{d_{i1}, d_{i2}, d_{i3}, d_{i4}, \dots, d_{ic}, \dots d_{i(c-1)}, d_{ic}\}$$

7.- Repetir el paso 5 y paso 6 hasta que los nuevos centroides determinados en la iteración p sean los mismos de la iteración p-1 con lo cual se podría concluir el proceso iterativo.

8.- Presentar la distribución final:

$K_1 = \cup_{i=1}^{a_1} I_i \quad \forall\, i\, \varepsilon\, [1, a_1]$, $K_2 = \cup_{i=1}^{a_2} I_i \quad \forall\, i\, \varepsilon\, [1, a_2]$,…,$K_c = \cup_{i=1}^{a_m} I_i \quad \forall\, i\, \varepsilon\, [1, a_m]$

Donde: $a_1 + a_2 + a_3 \dots + a_m \leq n$, $s \leq c$ y K_c son conjuntos de instancias disjuntos.

Algoritmo propuesta B (APB)

1.- Sea μ el modelo dimensional propuesto el cual se considera desplegado en las n sedes.

2.- Desde el nodo de interés se elegirá las sedes a incluir en la evaluación se precisara el número de clusters a generar.

3.- Invocar el proceso de Evaluación el cual seguirá el siguiente flujo:

3.1.- Cada nodo procesa el requerimiento de forma local siguiendo los pasos del 5 al 8 de la propuesta APA y remite los conglomerados resultantes al nodo central.

3.2.- El nodo central reúne los centroides finales de cada nodo participante en la evaluación y repite el paso del 5 al 7 de la propuesta APA considerando cada centroide como una instancia.

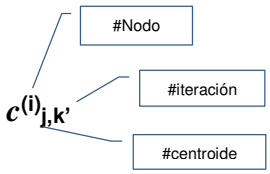
3.3.- Se aplica el paso 8 de la propuesta APA considerando los resultados obtenidos del paso 3.2.

Algoritmo propuesta C (APC)

Este algoritmo es una variante de la propuesta de [6] adaptada a la necesidad de negocio planteada.

1.- El nodo principal envía un mensaje inicio del algoritmo adjuntando el identificador del nodo y el número de clúster a generar msg_inicio{id,k} a todos los nodos participantes, considere k=2 clústers mínimos y un máximo de k_{max} clúster. Donde k_{max} representa el numero máximo de clusters sobre el cual se podría realizar la optimización de clusters utilizando un índice de validación.

2.- Los nodos participantes deberán enviar sus k centroides iniciales determinados aleatoriamente considerando el identificador del nodo, el total de centroides, el número de iteración k', el conjunto de centroides y el total de instancias determinadas por cada uno; msg_iteracion{id,k,k',$\{c^{(i)}_{j,k'}\}$,$|c^{(i)}_{j,k'}|$} hacia el nodo principal.



3.- En cada iteración de ejecución local los nodos participantes enviaran información de la iteración bajo la misma estructura definida en el paso 2. La finalización de la ejecución del algoritmo en los nodos locales es el mismo que rige la finalización en un algoritmo k-means estándar en adición deberá remitir el índice de validación determinado.

4.- El nodo principal recepciona los mensajes informativos procedentes de cada nodo y lo almacena en una cola de espera.

5.- En cada iteración el nodo principal da lectura a la cola de mensajes, selecciona los centroides de todos los nodos participantes para la iteración k' de

su interés y realiza un recalculo de los centroides incluyendo los suyos para procesar la siguiente iteración.

6.- Se repite el paso 5 mientras los centroides correspondientes a las iteraciones k', k'+1 presentan cambios significativos en el nodo principal. Al finalizar debe calcular el índice de validación.

7.- Mientras k clusters sea menor a k_{max} clusters, el nodo principal incrementa k en una unidad y repite los pasos del 1 al 6.

8.- Al terminar con la evaluación de los k clusters, el nodo principal determinara el número de clusters óptimos en función del índice de validación calculado.

El índice de validación a aplicar puede ser el índice de Silueta, Dunn, Daves Bouldin, o Rose Turi.

4.4. Aplicación de la técnica Clustering APA

A continuación recurriremos al uso de una muestra de datos basado en seis instancias la cual nos permitirá explicar el mecanismo de la técnica de clustering:

Se ha hecho uso de las librerías de distribución gratuita weka[23]; Este código fuente recibió ciertas adecuaciones como: el desarrollo de una nueva clase DatabasesLoader, la agregación de métodos sobrecargados en las clases SimpleKMeans y DatabaseUtils todo ello con la intención de obtener la información requerida bajo el contexto propuesto.

La distribución formatea los datos antes de procesar alguna técnica de minería. Esto consiste en definir una estructura conocida como formato arff y sobre este conjunto de datos (al cual llamaremos dataset en lo sucesivo) formateados procesa la técnica elegida.

En la figura 4.6 se presenta la estructura del dataset cuyos elementos se describen a continuación:

Nombre de la relación o del Dataset (QueryResult**):** En el ejemplo se ha asignado un valor predeterminado.

Sección de Atributos (@Attribute**):** Esta sección contiene la lista unificada de los atributos de la relación y para cada atributo una lista de valores únicos determinados en el dataset.

Sección de los Datos (@data**):** El detalle de las instancias que conforman la información a analizar:

[23] [21],” Machine Learning Group at the University of Waikato”

```
@relation QueryResult

@attribute anio_cuadernillo {2008,2009}
@attribute elemento_general {Casacion}
@attribute especialidad {'Contencioso Administrativo'}
@attribute proceso_judicial {'Contencioso Administrativo',Abreviado,Especial}
@attribute materia {NINGUNO,'Impugnación Judicial de Acto Administrativo','Pago de Acciones Laborales'}
@attribute desc_norma_legal {'D. Leg - 276','Ley - 24041','D. Ley - 19990','Ley - 27584','D. S. - 005-90-PCM','Ley - 20530'}
@attribute tema_legal {'RESTITUCION DE PLAZA',REINCORPORACION,'PENSION DE JUBILACION','NULIDAD DE RESOLUCION ADMINISTRATIVA','PAGO DE SUBSIDIO DE LUTO','INCORPORACION AL REGIMEN PENS

@data
2008,Casacion,'Contencioso Administrativo','Contencioso Administrativo',NINGUNO,'D. Leg - 276','RESTITUCION DE PLAZA'
2008,Casacion,'Contencioso Administrativo',Abreviado,'Impugnación Judicial de Acto Administrativo','Ley - 24041',REINCORPORACION
2008,Casacion,'Contencioso Administrativo','Contencioso Administrativo','Pago de Acciones Laborales','D. Ley - 19990','PENSION DE JUBILACION'
2009,Casacion,'Contencioso Administrativo',Especial,NINGUNO,'Ley - 27584','NULIDAD DE RESOLUCION ADMINISTRATIVA'
2009,Casacion,'Contencioso Administrativo',Especial,NINGUNO,'D. S. - 005-90-PCM','PAGO DE SUBSIDIO DE LUTO'
2009,Casacion,'Contencioso Administrativo',Especial,'Impugnación Judicial de Acto Administrativo','Ley - 20530','INCORPORACION AL REGIMEN PENSIONARIO'
```

Figura 4.7: Formato arff de la muestra de datos
(Fuente: [21], Elaboración propia)

En la tabla 4.2 se presenta una distribución de la data muestral; aquí se recurre al uso de una estructura de datos de tipo HashTable para la representación de los datos nominales:

A los valores de los atributos de la primera instancia le asigna un índice cero(0.0).

Para el resto de instancias; valida cada valor del atributo, si se trata de un nuevo valor y este no se encuentra en la tabla hash entonces lo agrega y le genera un nuevo índice secuencial caso contrario devuelve el índice del valor que ya existe en la tabla hash.

Atributos								
Instancias		0	1	2	3	4	5	6
	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	0.0	0.0	0.0	1.0	1.0	1.0	1.0
	2	0.0	0.0	0.0	0.0	2.0	2.0	2.0
	3	1.0	0.0	0.0	2.0	0.0	3.0	3.0
	4	1.0	0.0	0.0	2.0	0.0	4.0	4.0
	5	1.0	0.0	0.0	2.0	1.0	5.0	5.0

Tabla 4.2: Distribución de la data muestral en una Hashtable
(Fuente: Elaboración propia)

En la figura 4.8 (líneas 121-123) se puede observar el mecanismo de asignación de un índice secuencial a cada valor único de cada atributo de las instancias que conforma la data a evaluar.

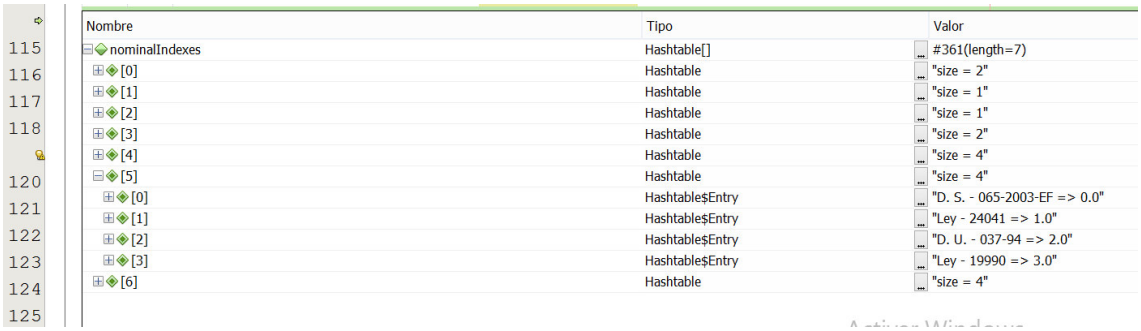


Figura 4.8: Asignación de índice secuencial a los valores de los atributos
(Fuente: Elaboración propia)

El proceso de clustering inicia eligiendo de forma aleatoria n centroides; en el ejemplo asignamos dos centroides.

Considérese la selección aleatoria de los centroides:

m_ClusterCentroids:

C0	3	1.0	0.0	0.0	2.0	0.0	3.0	3.0
C1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabla 4.3: Selección aleatoria de centroides
(Fuente: Elaboración propia)

Lo que sigue es el proceso de determinación de los clusters o conglomerados basado en la distancia euclidiana para lo cual explicaremos la primera iteración:

Sea el orden de distribución del dataset el siguiente:

	Atributos							
Instancias		0	1	2	3	4	5	6
	0	1.0	0.0	0.0	2.0	0.0	4.0	4.0
	1	0.0	0.0	0.0	1.0	1.0	1.0	1.0
	2	0.0	0.0	0.0	0.0	2.0	2.0	2.0
	3	1.0	0.0	0.0	2.0	1.0	5.0	5.0
	4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	5	1.0	0.0	0.0	2.0	0.0	3.0	3.0

Calculo de la distancia entre I_4 y el primer centroide C_0

								$\sum_{diff^2}$	$d_1=\sqrt{\sum_{diff^2}}$
I_0	1.0	0.0	0.0	2.0	0.0	4.0	4.0		
C_0	1.0	0.0	0.0	2.0	0.0	3.0	3.0		
$diff=I_0-C_0$	0.0	0.0	0.0	0.0	0.0	1.0	1.0		
$diff^2=(I_0-C_0)^2$	0.0	0.0	0.0	0.0	0.0	1.0	1.0	2	**1.41421356237309504880168872420 97**

Tabla 4.4: Calculo de la distancia entre I_0 y el primer centroide C_0
(Fuente: Elaboración propia)

Calculo de la distancia entre I_4 y el segundo centroide C_1

									$\sum_{diff^2}$	$d_2=\sqrt{\sum_{diff^2}}$
I_0	4	1.0	0.0	0.0	2.0	0.0	4.0	4.0		
C_1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
$diff=I_0-C_1$		1.0	0.0	0.0	1.0	0.0	1.0	1.0		
$diff^2=(I_0-C_0)^2$		1.0	0.0	0.0	1.0	0.0	1.0	1.0	4	**2**

Tabla 4.5: Calculo de la distancia entre I_0 y el segundo centroide C_1
(Fuente: Elaboración propia)

Por ejemplo considérese que se está evaluando la instancia I_0; las tablas 4.4 y 4.5 presentan el cálculo de la distancia de la instancia I_0 con respecto a cada centroide C_0 y C_1.

La distancia mínima está dada por el mínimo valor determinado entre d1 y d2 donde d1 es la distancia de I_0 a C_0 y d2 es la distancia de I_0 a C_1; lo cual se puede representar como min(d1,d2)=min(1.41,2)=1.41, valor de la distancia que corresponde al centroide C_0; por tanto la instancia I_0 está más cerca del centroide C_0 que del centroide C_1 y se considera que pertenece al conglomerado C_0.

El procedimiento se repite hasta evaluar la última instancia. Al finalizar el ciclo se tiene una lista de distribución de los centroides asignados a las instancias como se muestra en la siguiente figura:

m_ClusterAssigments:

centroides	C_0	C_1	C_1	C_0	C_1	C_0
Instancias	i_0	i_1	i_2	i_3	i_4	i_5

Tabla 4.6: Distribución de centroides a Instancias según la muestra
(Fuente: Elaboración propia)

Asimismo se tiene una estructura que contiene la distribución de las instancias a los centroides como se muestra a continuación:

C_0	{ I_0, I_3, I_5}
C_1	{I_1, I_2, I_4}

Tabla 4.7: Distribución de instancias a centroides según la muestra
(Fuente: Elaboración propia)

		0	1	2	3	4	5	6
C_0	I_0	1.0	0.0	0.0	2.0	0.0	4.0	4.0
	I_3	1.0	0.0	0.0	2.0	1.0	5.0	5.0
	I_5	1.0	0.0	0.0	2.0	0.0	3.0	3.0
		0	1	2	3	4	5	6
C_1	I_1	0.0	0.0	0.0	1.0	1.0	1.0	1.0
	I_2	0.0	0.0	0.0	0.0	2.0	2.0	2.0
	I_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabla 4.8: Distribución detallada de instancias a centroides
(Fuente: Elaboración propia)

El siguiente paso consiste en determinar un nuevo conjunto de centroides.

Este procedimiento consiste en determinar la valoración nominal de cada conglomerado. Para lo cual primero inicializara el m_ClusterCentroids:

m_ClusterCentroids:

C0	null	null	null	null	null	null	null	null
C1	null	null	null	null	null	null	null	null

Tabla 4.9: Estructura de centroides inicializada
(Fuente: Elaboración propia)

La valoración se realiza creando un arreglo cuyo tamaño está determinado por la cantidad distinta de valores nominales para cada atributo. A continuación acumula el peso 1.0 por cada ocurrencia del mismo tipo determinada en el conglomerado y ese acumulado lo asigna al índice que representa el tipo de valor nominal evaluado.

En la tabla 4.10 se presenta la valoración detallada. Por ejemplo, para el dataset muestral el atributo cero (a_0) contiene dos valores distintos. Entonces el tamaño del vector es dos; el índice cero contendrá el acumulado de ocurrencias igual a cero en el conglomerado C_0 (ver Tabla 4.8) para ese atributo (a_0) y el índice uno contendrá el acumulado de ocurrencias igual a uno en el conglomerado C_0 (ver Tabla 4.8). Se repite el procedimiento para todos los atributos.

Atributos	Valoración					
a_0	0.0	3.0				
	0	1				
a_1	3.0					
	0					
a_2	3.0					
	0					
a_3	0.0	0.0	3.0			
	0	1	2			
a_4	2.0	1.0	0.0			
	0	1	2			
a_5	0.0	0.0	0.0	1.0	1.0	1.0
	0	1	2	3	4	5
a_6	0.0	0.0	0.0	1.0	1.0	1.0
	0	1	2	3	4	5

Tabla 4.10: Análisis determinación de valoración del Conglomerado C_0
(Fuente: Elaboración propia)

Para el conglomerado C_0 se obtuvo:

m_Val:

Val	1.0	0.0	0.0	2.0	0.0	3.0	3.0
Atributos	0	1	2	3	4	5	6

Tabla 4.11: Valoración del Conglomerado C_0
(Fuente: Elaboración propia)

Este m_Val es uno de los nuevos centroides y se consigna en la estructura ***m_ClusterCentroids*** igualmente se procede para el segundo centroide.

La tabla 4.12 presenta los nuevos centroides los cuales serán considerados en la siguiente iteración.

m_ClusterCentroids:

C0	1.0	0.0	0.0	2.0	0.0	3.0	3.0
C1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabla 4.12: Conjunto de nuevos centroides
(Fuente: Elaboración propia)

La siguiente iteración dos; consiste en determinar los nuevos conglomerados para ello determina las mínimas distancias de cada instancia del dataset con respecto a cada centroide y se consignan en la estructura *m_ClusterAssigments(iteración 2)*.

m_ClusterAssigments(iteración 1):

centroides	C_0	C_1	C_1	C_0	C_1	C_0
Instancias	i_0	i_1	i_2	i_3	i_4	i_5

m_ClusterAssigments(iteración 2):

centroides	C_0	C_1	C_1	C_0	C_1	C_0
Instancias	i_0	i_1	i_2	i_3	i_4	i_5

Se compara cada uno de los valores de las dos estructuras si estos son equivalente para cada instancia entonces el indicador de control del proceso iterativo se establece en *"fin"* con lo cual termina el ciclo.

4.5. Prototipo

El prototipo para la aplicación de minería de datos es una aplicación web desarrollada en el lenguaje de programación java bajo el modelo MVC.

La figura 4.9 explica la arquitectura del prototipo; cuyo flujo se detalla a continuación:

El usuario inicia el proceso de clustering mediante el uso de un formulario web IU_Petitorio.jsp. Este evento es atendido por un Servlet: CTRL_Petitorio.

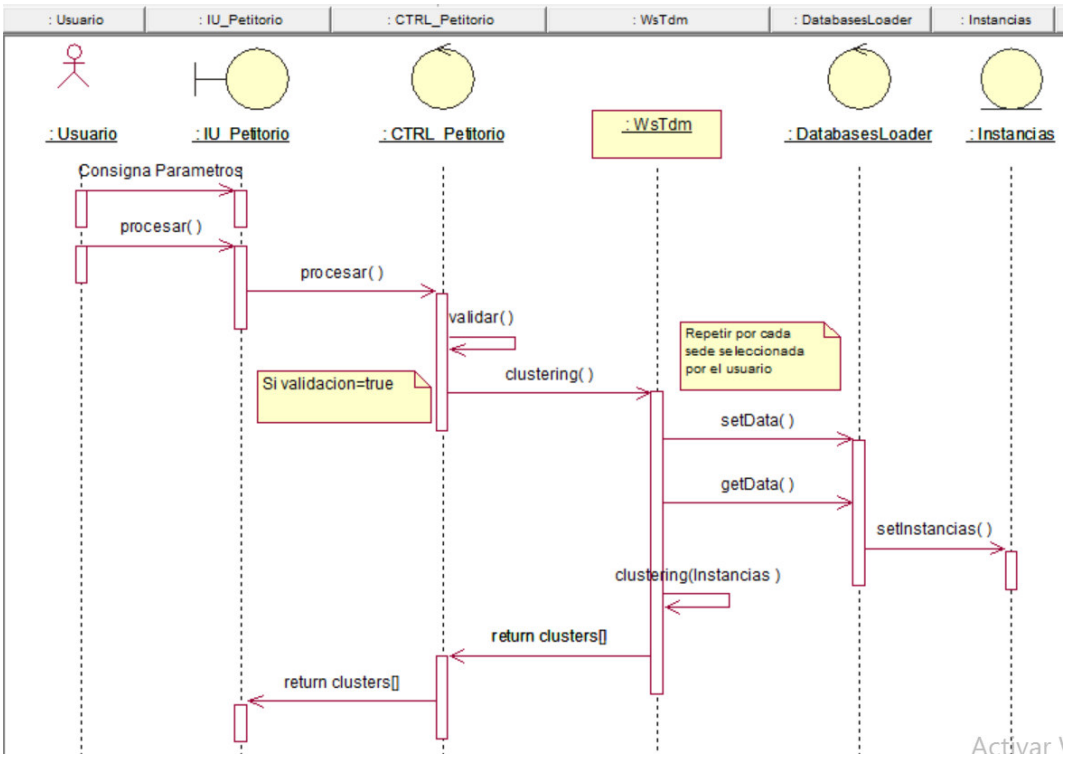


Figura 4.9: Arquitectura del Prototipo
(Fuente: Elaboración propia)

El servlet definido cumplirá el rol de controlador-delegador; este toma la petición procedente de la interfaz web y realiza las siguientes tareas:

1. Valida los datos consignados por el usuario como las sedes a intervenir en la evaluación, el número de clusters de interés.
2. Invoca al servicio web WsTdm la ejecución del método clustering.

3. Recibe el resultado del servicio web WsTdm siendo estos transferidos al usuario a través del formulario web: IU_Petitorio.jsp

El servicio web realizara las siguientes acciones con la finalidad de atender el requerimiento del servlet:

1. Ejecuta el método: setData() con la finalidad de establecer las características de las fuentes de datos implicadas en la evaluación.
2. Ejecuta el método getData() con el objetivo de recuperar la información en un objeto Instancias
3. Ejecuta el método clustering() pasándole como parámetro el objeto Instancias
4. Retorna los clusters resultantes de la evaluación al servlet.

Implementación del Web Service WsTdm

Las figuras 4.10 y 4.11 contienen el código fuente del servicio web definido en el presente prototipo; este define el método clustering01.

```
1.  package com.zemr.tesis;
2.  import java.util.ArrayList;
3.  import javax.jws.WebService;
4.  import javax.jws.WebMethod;
5.  import javax.jws.WebParam;
6.  import javax.xml.ws.Holder;
7.  import weka.clusterers.SimpleKMeans;
8.  import weka.core.Instances;
9.  import weka.core.Utils;
10. import weka.core.converters.DatabasesLoader;
11. import java.util.List;
12. @WebService(serviceName = "WsTdm")
13.
14. public class WsTdm {
15.     @WebMethod(operationName = "clustering01")
16.       public List<miresultado> clustering01(
17.       @WebParam(name = "parameter1") int n_dataset,
18.       @WebParam(name = "parameter2") List<String> s_url,
19.       @WebParam(name = "parameter3") List<String> s_user,
20.       @WebParam(name = "parameter4") List<String> s_pwd,
21.       @WebParam(name = "parameter5") List<String> s_query,
22.       @WebParam(name = "parameter6") int n_clusters) {
23.           List<miresultado> result=new ArrayList(); double lb_error=0.0;
24.        try { DatabasesLoader nodo =new DatabasesLoader(n_dataset);
25.             for(int i=0;i<n_dataset;i++)
26.             nodo.setSource(s_url.get(i),s_user.get(i),s_pwd.get(i),s_query.get(i));
```

```
27.          Instances instances=nodo.getData();SimpleKMeans s=new
     SimpleKMeans();s.setNumClusters(n_clusters);
28.          s.buildClusterer(instances);int i=s.getClusterCentroids().numInstances();int mclusterSizes[]=new
     int[i];
29.          mclusterSizes=s.getClusterSizes();lb_error=s.getSquaredError();int[][] clusterOrden;
30.          clusterOrden= new
     int[mclusterSizes.length][mclusterSizes.length];clusterOrden=ordBurbuja(mclusterSizes);
31.           for(int v=0;v<clusterOrden.length;v++)
32.           {   miresultado unResult =new miresultado();
33.              unResult.setUid(v);unResult.setNombre("Cluster");
34.              unResult.setPorcentaje((clusterOrden[v][1]*100.00)/(Utils.sum(mclusterSizes)*1.0));
35.              unResult.setErrorCuadrado(lb_error);unResult.setInstancias(clusterOrden[v][1]);
     unResult.setDetalle(s.getClusterCentroids().instance(clusterOrden[v][0]).toString());result.add(unResult);
36.           }} catch (Exception e) {System.out.println("\n"+e.getMessage()); }
37.         return result; }
38.
39.      private int[][] ordBurbuja(int[] mcluster)
40.      {   int[][] cluster= new int[mcluster.length][mcluster.length];
41.         int t_valor,t_ind, inferior;
42.         boolean intercambio=true;
43.         for(int i=0;i<mcluster.length;i++)
44.         {   cluster[i][0]=i;
45.            cluster[i][1]=mcluster[i];}
46.         inferior=cluster.length -2;
47.         while(intercambio)
48.         { intercambio=false;
49.         for(int i=0;i<=inferior;i++)
50.         {    if (cluster[i][1]<cluster[i+1][1])
51.              {   t_valor=cluster[i][1];
52.                 t_ind=i;
53.                 cluster[i][0]=cluster[i+1][0];
54.                 cluster[i][1]=cluster[i+1][1];
55.                 cluster[i+1][0]=t_ind;
56.                 cluster[i+1][1]=t_valor;
57.                 intercambio=true;}}
58.         inferior--;}
59.         return cluster;
60.      }}
```

Figura 4.10: Web Service WsTdm
(Fuente: Elaboración propia)

Implementación del Servlet: CTRL_Petitorio

La figura 4.12 contiene el código fuente del servlet definido en el presente prototipo; este define todas las tareas requeridas para solicitar la técnica de minería clustering al servicio web.

Asimismo se encarga de preparar los resultados devueltos por el servicio web para enviarlo al formulario web mediante el uso de un objeto RequestDispacher. El uso de un objeto HttpSession se realiza con la finalidad de consignar mediante el uso de tags el traslado de información hacia la interfaz web IU_Petitorio.jsp.

```
1.  package servlet;
2.  import com.zemr.tesis.WsTdm_Service;
3.  import java.io.*;
4.  import java.util.ArrayList;
5.  import java.util.List;
6.  import javax.servlet.ServletException;
7.  import javax.servlet.http.HttpServlet;
8.  import javax.servlet.http.HttpServletRequest;
9.  import javax.servlet.http.HttpServletResponse;
10. import javax.servlet.RequestDispatcher;
11. import javax.xml.ws.WebServiceRef;
12. import com.zemr.tesis.Miresultado;
13. import org.json.simple.*;
14. public class CTRL_Petitorio extends HttpServlet {
15.     @WebServiceRef(wsdlLocation = "WEB-
    INF/wsdl/localhost_8080/AppWsTdm/WsTdm.wsdl")
16.     private WsTdm_Service service;
17.     @Override
18.     protected void doPost(HttpServletRequest request, HttpServletResponse response)
19.             throws ServletException, IOException {
20.         int n_dataset=0,n_clusters=0,n_corte=4;
21.         String ls_csj;
22.         try {
23.             RequestDispatcher dispacher =
    request.getRequestDispatcher(request.getParameter("nombreJSP"));
24.             List<String> v_url=new ArrayList<String>();
25.             List<String> v_user=new ArrayList<String>();
26.             List<String> v_pwd=new ArrayList<String>();
27.             List<String> v_query=new ArrayList<String>();
28.             n_clusters=Integer.valueOf(request.getParameter("ncluster"));
29.             for(int i=1;i<=n_corte;i++)
30.             { ls_csj=String.valueOf(request.getParameter("CSJ0"+i));
31.                 if (ls_csj.equals("null")==false)
32.                 { v_query.add("call sp_fact_petitorio("+0+")");
33.                   v_url.add("jdbc:sybase:Tds:server:2638/bd11_"+"0"+i);
34.                   v_user.add("dba");
35.                   v_pwd.add("sql");
36.                   n_dataset++;
37.                 }
38.             }
39.             com.zemr.tesis.WsTdm port = service.getWsTdmPort();
40.         java.util.List<Miresultado> v_resultado=new ArrayList<Miresultado>();
41.         v_resultado=port.clustering01(n_dataset,v_url,v_user,v_pwd,v_query,n_clusters);
42.
43.         request.setAttribute("n_dataset",String.valueOf(n_dataset));
44.             for(int i=0;i<n_clusters;i++){
45.               request.setAttribute("r"+String.valueOf(i),v_resultado.get(i).getNombre()+
46.                     String.valueOf(v_resultado.get(i).getUid()+1)+" Instancias="+
47.                     String.valueOf(v_resultado.get(i).getInstancias())+" "+
48.                     String.valueOf(v_resultado.get(i).getDetalle())
49.                     );}
50.             resultado m_resultado=new resultado();
51.             m_resultado.setErrorCuadrado(v_resultado.get(0).getErrorCuadrado());
52.             m_resultado.setMiResult(v_resultado);
53.
54.             request.setAttribute("m_resultado",m_resultado);
55.
56.             String json = null;
57.             JSONArray js = new JSONArray();
58.             JSONObject   j;
```

```
59.             for (int x = 0; x < v_resultado.size(); x++) {
60.                 j = new JSONObject();
61.                 j.put( "name",String.valueOf("Cluster "+(v_resultado.get(x).getUid()+1) ));
62.                 j.put( "y",Math.round(v_resultado.get(x).getPorcentaje() * 100.0) / 100.0);
63.                 js.add(j);}
64.             json = js.toJSONString();
65.             request.setAttribute("json",String.valueOf(json));
66.             System.out.println(json);
67.             dispacher.forward(request, response);
68.         } catch (Exception e) {e.printStackTrace(); System.out.println("\n"+e.getMessage());}}
69.     }
```

Figura 4.11: Servlet Ctrl_Petitorio
(Fuente: Elaboración propia)

Implementación de la Interfaz Gráfica de Usuario: IU_Petitorio.jsp

La figura 4.13 presenta el diseño del prototipo el cual se explica a continuación:

El formulario se divide en tres secciones: la primera sección presenta el título del proyecto. La segunda sección contiene la lista parcial de sedes judiciales consideradas en el presente proyecto y los botones Procesar y Limpiar. La tercera sección contiene los resultados de la evaluación del clustering; Se tiene un ítem por cada cluster o conglomerado generado.

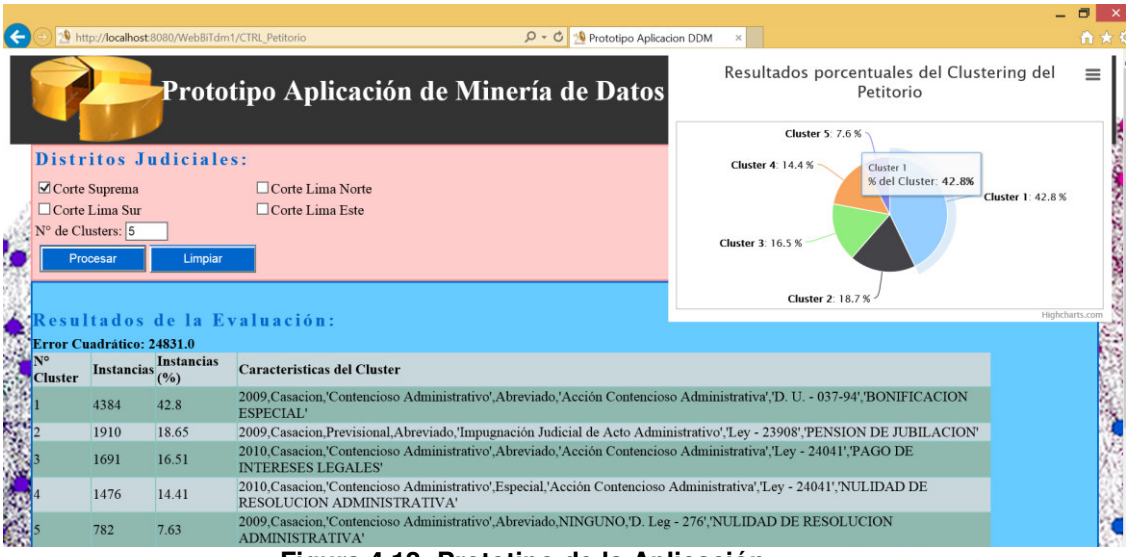


Figura 4.12: Prototipo de la Aplicación
(Fuente: Elaboración propia)

La estructura del ítem consiste de: un número y/o identificador del cluster, el total de instancias que lo conforman, el porcentaje que representa este con respecto

al total de instancias evaluadas y por último el detalle del conglomerado; las características que definen el conglomerado o cluster resultante.

Las siguientes líneas de código corresponden al código fuente del formulario web. Al tratarse de una página jsp esta contiene código html, hojas de estilo y javascript los cuales permiten generar el diseño presentado en la tabla 4.8.

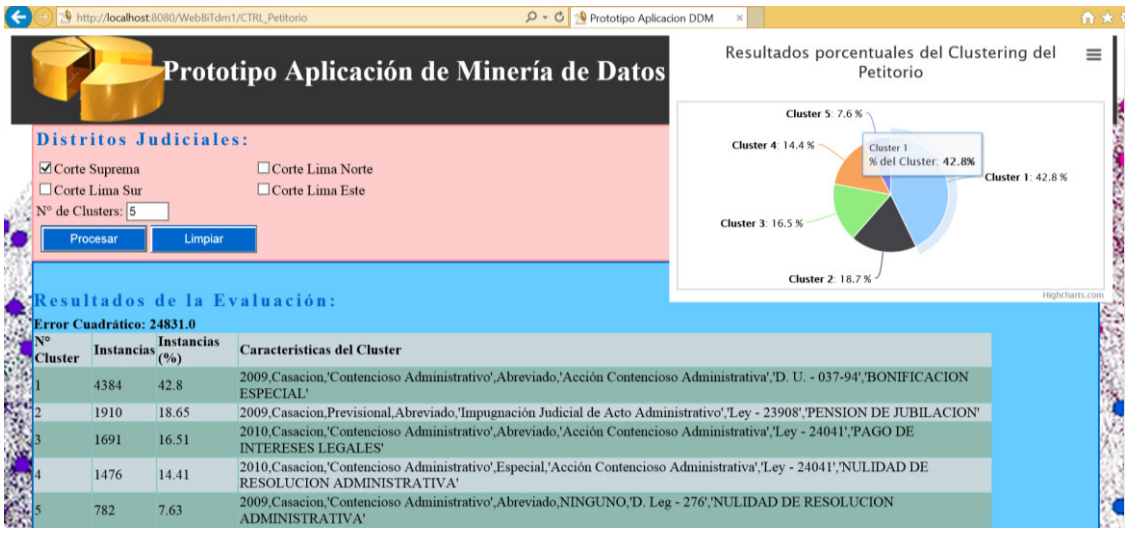
```
1.  <%@page import="java.util.Iterator"%>
2.  <%@page import="servlet.resultado"%>
3.  <%@page contentType="text/html" pageEncoding="UTF-8"%>
4.  <!DOCTYPE html>
5.  <html>
6.      <%
7.          resultado m_resultado =  (resultado)request.getAttribute("m_resultado");
8.          String n_dataset      =  (String)request.getAttribute("n_dataset");
9.          String result="";
10.     %>
11.     <head>
12.         <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
13.         <title>Prototipo Aplicacion DDM</title>
14.
15.         <link rel="shortcut icon" href="itinnovacionLogo.jpg">
16.         <link href="style03.css" rel="stylesheet">
17.         <script src="jquery-1.11.1.min.js"></script>
18.         <link rel="stylesheet" href="nivo-slider/themes/default/default.css"/>
19.         <link rel="stylesheet" href="nivo-slider/nivo-slider.css"/>
20.         <script type="text/javascript" src="nivo-slider/jquery.nivo.slider.js"></script>
21.         <script src="js/highcharts.js"></script>
22.         <script src="js/modules/exporting.js"></script>
23.         <style type="text/css">
24.         ${demo.css}
25.         </style>
26.  <script type="text/javascript">
27. $(function () {
28.     var datos = ${json};
29.     $('#estadistica1').highcharts({
30.         chart: {
31.             plotBackgroundColor: null,
32.             plotBorderWidth: 1,
33.             plotShadow: false
34.         },
35.         title: {
36.             text: 'Resultados porcentuales del Clustering del Petitorio'
37.         },
38.         tooltip: {
39.             pointFormat: '{series.name}: <b>{point.percentage:.1f}%</b>'
40.         },
41.         plotOptions: {
42.             pie: {
43.                 allowPointSelect: true,
44.                 cursor: 'pointer',
45.                 dataLabels: {
46.                     enabled: true,
47.                     format: '<b>{point.name}</b>: {point.percentage:.1f} %',
48.                     style: {
49.                         color: (Highcharts.theme && Highcharts.theme.contrastTextColor) || 'black'
50.                     }
51.                 }
52.             }
53.         },
54.         series: [{
55.             type: 'pie',
56.             name: '% del Cluster',
57.             data: datos
58.         }]
59.     });
60. });
```

```
61.  </script>
62.      </head>
63.      <body>
64.          <div id="topContent">
65.              <div id="estadistica1">
66.              </div>
67.              <h1 id="logo"><a href="#">Prototipo Aplicación de Minería de Datos</a></h1>
68.          </div><!-- topContent -->
69.          <div id="container">
70.              <div id="cabecera">
71.               <form action="CTRL_Petitorio" method="POST">
72.                  <input type="hidden" name="nombreJSP" value="/IU_Petitorio.jsp" />
73.                   <table class="sedes">
74.                      <tr>
75.                        <th><h4>Distritos Judiciales:</h4></th>
76.                      </tr>
77.                      <tr>
78.                         <td>
79.                            <input type="checkbox" name="CSJ01" value="1" checked>Corte Suprema<br>
80.                         </td>
81.                         <td>
82.                            <input type="checkbox" name="CSJ02" value="2">Corte Lima Norte<br>
83.                         </td>
84.                      </tr>
85.                      <tr>
86.                         <td>
87.                            <input type="checkbox" name="CSJ03" value="3">Corte Lima Sur<br>
88.                         </td>
89.                         <td>
90.                            <input type="checkbox" name="CSJ04" value="4">Corte Lima Este<br>
91.                         </td>
92.                      </tr>
93.                      <tr>
94.                         <td>N° de Clusters: <input type="number" name="ncluster" value="5" min="2" max="150" step="1"
     size="2"><br></td>
95.                      </tr>
96.                   </table>
97.                   <div id="tarea">
98.                    <td> </td>
99.                    <input type="submit" class="button_1" name="procesar" value="Procesar">
100.                   <input type="reset" class="button_1" name="reset" value="Limpiar">
101.                  </div>
102.           </form>
103.          </div><!-- cabecera -->
104.          <div id="detalle">
105.              <h4></h4>
106.              <%
107.                  if (m_resultado != null && m_resultado.getMiResult().size()>0)   {
108.                     out.println("<h4><b>Resultados de la Evaluación:</b></h4>");
109.                     out.println("<td><b>Error Cuadrático: "+String.valueOf(m_resultado.getErrorCuadrado())+"</b></td>");}%>
110.              <table class="tProductos">
111.                  <%if (m_resultado != null && m_resultado.getMiResult().size()>0)   {
112.                     out.print("<td><b>N°Cluster</b></td>");
113.                     out.print("<td><b>Instancias</b></td>");
114.                     out.print("<td><b>Instancias(%)</b></td>");
115.                     out.print("<td><b>Caracteristicas del Cluster</b></td>");
116.                   for(int i=0;i<m_resultado.getMiResult().size();i++){
117.                     out.println("<tr>");
118.                        result=String.valueOf(m_resultado.getMiResult().get(i).getUid() + 1);
119.                        out.println("<td>" + result + "</td>");
120.                        result=String.valueOf(m_resultado.getMiResult().get(i).getInstancias());
121.                        out.println("<td>" + result + "</td>");
122.                        result=String.valueOf(Math.round(m_resultado.getMiResult().get(i).getPorcentaje()*100.0)/100.0);
123.                        out.println("<td>" + result + "</td>");
124.                        result=String.valueOf(m_resultado.getMiResult().get(i).getDetalle());
125.                        out.println("<td>" + result + "</td>");
126.                     out.println("</tr>");
127.                   }}%>
128.              </table>
129.          </div>
130.      </div><!-- container -->
131.  </body>
132. </html>
```

Figura 4.13: Código fuente del formulario web IU_Petitorio.jsp
(Fuente: Elaboración propia)

4.6. Resultados

En la figura 4.14 se presenta los resultados obtenidos de la evaluación de la información concerniente al petitorio correspondiente a procesos judiciales de recursos casatorios de la especialidad laboral.



N° Cluster	Instancias	Instancias (%)	Características del Cluster
1	4384	42.8	2009,Casacion,'Contencioso Administrativo',Abreviado,'Acción Contencioso Administrativa','D. U. - 037-94','BONIFICACION ESPECIAL'
2	1910	18.65	2009,Casacion,Previsional,Abreviado,'Impugnación Judicial de Acto Administrativo','Ley - 23908','PENSION DE JUBILACION'
3	1691	16.51	2010,Casacion,'Contencioso Administrativo',Abreviado,'Acción Contencioso Administrativa','Ley - 24041','PAGO DE INTERESES LEGALES'
4	1476	14.41	2010,Casacion,'Contencioso Administrativo',Especial,'Acción Contencioso Administrativa','Ley - 24041','NULIDAD DE RESOLUCION ADMINISTRATIVA'
5	782	7.63	2009,Casacion,'Contencioso Administrativo',Abreviado,NINGUNO,'D. Leg - 276','NULIDAD DE RESOLUCION ADMINISTRATIVA'

Figura 4.14: Resultados de la Evaluación
(Fuente: Elaboración propia)

La figura 4.15 grafica la distribución porcentual de los resultados los cuales se analizan a continuación:

La interpretación de los conglomerados resultantes nos llevan a determinar que el 42.8% de la carga procesal corresponden a procesos casatorios iniciados en el año 2009 correspondiente a la subespecialidad contencioso administrativo en la cual la parte demandante solicita una bonificación especial fundamentando su pedido en el D.U 037-94.

El segundo conglomerado explica que se tiene un 18.65% de casos iniciados en el año 2009 solicitan Pensión de Jubilación y se fundamentan en la ley 23908 del código procesal civil; este pedido lo realizan mediante una impugnación judicial de acto administrativo.

El tercer conglomerado que equivale al 16.51% de la carga evaluada se trata de un pedido de Pago de Intereses Legales sosteniendo su pedido en la ley 24041, esta formulación se realiza mediante un proceso Abreviado con acción contencioso administrativo.

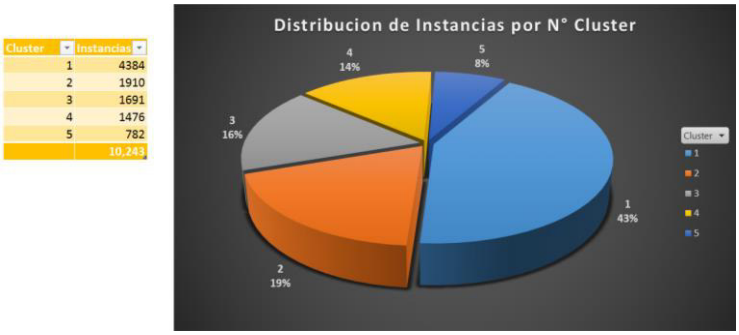


Figura 4.15: Distribución porcentual de la Evaluación
(Fuente: Elaboración propia)

El cuarto conglomerado recae en un 14.41% quienes solicitan Nulidad de Resolución Administrativa respaldando su pedido mediante ley 24041, este petitorio lo formulan mediante un proceso especial de acto administrativo.

El quinto conglomerado presenta el porcentaje más bajo, el 7.63% de la carga procesal tiene como requerimiento una Nulidad de Resolución Administrativa y lo realizan mediante un proceso abreviado de tipo acción contencioso administrativa, se fundamenta el petitorio en el Decreto Ley 276 del Código Civil.

Es preciso señalar que estos resultados apoyarían con el cumplimiento de los objetivos definidos en la agenda estratégica institucional en relación a los siguientes puntos:

Fortalecer los mecanismos de disminución del volumen procesal, reducción de plazos procesales y nivel de litigiosidad.

Los conglomerados resultantes del análisis efectuado; permite detectar patrones que presentan la carga procesal; y con ello se podría tomar acciones como establecer órganos jurisdiccionales especializados; aperturar órganos transitorios; incorporar juristas especializados según las características que presentan los conglomerados.

Predictibilidad de las decisiones judiciales

Es posible realizar un análisis enfocado en los procesos resueltos y centrados en el fallo; los patrones característicos de los conglomerados resultantes pueden ser reutilizados en la emisión de fallos futuros para casos judiciales que presenten el mismo comportamiento.

Acceso e inclusión social en la impartición de justicia

El bajo porcentaje de aprobación de parte de la ciudadanía que recibe este organismo del estado[24] se debe a varios factores y estos tienen relación en parte con los dos objetivos anteriormente definidos. Su cumplimiento permitirá cubrir este tercer objetivo de manera que la población pueda percibir que si existe la impartición de justicia y esta es eficiente.

[24] [28],” Aprobación de Poderes del Estado cayó críticamente desde 2011”

CAPITULO V. CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

1. El presente trabajo desarrolla un prototipo que aplica minería de datos distribuida sobre datos nominales para determinar patrones de comportamiento basado en el petitorio que presenta la carga procesal de los periodos 2008 al 2010 correspondiente a órganos jurisdiccionales casatorios.
2. Se ha propuesto un algoritmo de clustering distribuido adaptable a la entidad judicial por su naturaleza organizacional: esquemas de negocios dispersos físicamente, confidencialidad de la información, reúso de infraestructura tecnológica, complejidad organizacional, bajo presupuesto.
3. La implementación de la presente propuesta en la organización del presente caso de estudio apoyaría en el cumplimiento de los objetivos señalados en la agenda estratégica institucional como el fortalecimiento de mecanismos para la reducción del volumen procesal, plazos procesales y nivel de litigiosidad; con lo cual se lograría una mejora en la calidad de los servicios que esta brinda a los ciudadanos.
4. El presente trabajo está orientado a aportar en la factibilidad de aplicar minería de datos distribuida en los organismos públicos permitiendo cubrir los objetivos centrados en el gobierno electrónico contemplado en el PNMSP.

5.2. RECOMENDACIONES

1. La aplicación de la presente propuesta se ha enfocado a un organismo particular y por la naturaleza del mismo se ha enfocado en un proceso de negocio puntual. Esta idea puede servir de referencia para aplicarlos a la diversidad de procesos que presenta este tipo de organización tanto en sus procesos primarios enfocado al ámbito jurisdiccional así como en los procesos de apoyo como son los proceso de gestión administrativa entre otros.
2. El presente trabajo ha logrado implementar la propuesta de la técnica APA sin embargo futuros trabajos deberían enfocarse al desarrollo de las propuestas APB y APC pues se consideran que permitirían consolidar la aplicación de la minería de datos distribuida en las organizaciones.

BIBLIOGRAFÍA

[1] F. Hurtado Leguia, «Minería de datos: Segmentación de clientes usando el algoritmo de clustering K-Mean,» 2005.

[2] L. C. Perez, Tecnicas de mineria de datos e inteligencia de negocios IBM SPSS Modeler, Madrid: IBERGARCETA PUBLICACIONES, S.L, 2014.

[3] Rekha Sunny T y Sabu M. Thampi , «Survey on Distributed Data Mining in P2P Networks,» 2010.

[4] oni.escuelas, «oni,» 15 10 2014. [En línea]. Available: http://www.oni.escuelas.edu.ar/.

[5] Laudon, Kenneth C. Laudon, Jane P., Sistema de Informacion Gerencial, Mexico: 12, 2012.

[6] Souptik Datta y Otros, «Distributed Data Mining in Peer-to-Peer Networks,» 2005.

[7] «Definicion Weka,» 2012. [En línea]. Available: http://en.wikipedia.org/wiki/Weka_(machine_learning). [Último acceso: 6 Enero 2012].

[8] Ian H. Witten, Eibe Frank, Mark A.Hall, DataMining Practical Machine Learning Tools and Techniques, 3 ed., Elsevier Inc., 2011, p. 665.

[9] «Bioweka,» 2012. [En línea]. Available: http://www.mybiosoftware.com/bioinformatics-platform/2193. [Último acceso: 6 Enero 2012].

[10] «Meka,» 2012. [En línea]. Available: http://meka.sourceforge.net/. [Último acceso: 6 Enero 2012].

[11] «Mulan,» 2012. [En línea]. Available: http://mulan.sourceforge.net/. [Último acceso: 6 Enero 2012].

[12] Mikut,Ralf y Reischl,Markus, «Data mining tools,» 26 10 2014. [En línea]. Available: http://zakki.dosen.narotama.ac.id/files/2012/02/Data-Mining-Tool-Reviews-March-2011.pdf.

[13] «unsa,» 17 10 2014. [En línea]. Available: http://admision.unsa.edu.pe/descargas/constitucion.pdf.

[14] «tc,» 17 10 2014. [En línea]. Available: http://www.tc.gob.pe/constitucion.pdf.

[15] «pcm,» 17 10 2014. [En línea]. Available: http://www.pcm.gob.pe/.

[16] «peru,» 17 10 2014. [En línea]. Available: http://www.peru.gob.pe/docs/PLANES/10051/PLAN_10051_Organigrama_del_Poder_Judicial_2013.pdf.

[17] «poder judicial,» 17 10 2014. [En línea]. Available: www.pj.gob.pe.

[18] E. Veramendi Flores, «El petitorio implícito en los procesos de familia: A propósito del tercer pleno casatorio».

[19] Agenda estrategica pj, «pj,» 28 10 2014. [En línea]. Available: http://www.pj.gob.pe/wps/wcm/connect/CorteSuprema/s_cortes_suprema_home/as_poder_judicial/as_corte_suprema/as_presidencia/as_plan_de_gestion/.

[20] kimball,ralph ross,margy, The data warehouse toolkit, United States of America: John Wiley and Sons, Inc., 2008.

[21] «weka,» 26 10 2014. [En línea]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

[22] A. R. S. y. D. R. F. Ingrid Wilford-Rivera, «Estado del Arte de la Mineria de Datos Distribuida,» 2008.

[23] Kargupta H., Hamzaoglu I. y Stafford B., «Scalable, distributed data mining using an agent based architecture.,» Menlo Park, USA, 1997.

[24] J. Roberts, La empresa moderna, 2006.

[25] «KDnuggets,» 26 10 2014. [En línea]. Available: http://www.kdnuggets.com/polls/index.html.

[26] «spss,» 26 10 2014. [En línea]. Available: http://www.spssfree.com/spss/analisis4.html.

[27] memoria institucional 2013, «pmsj,» 27 10 2014. [En línea]. Available: http://www.pmsj.org.pe/nweb/publicas/memo2013/index.html.

[28] C. Rosales Ferreyros, «el comercio,» 28 10 2014. [En línea]. Available: http://elcomercio.pe/politica/actualidad/aprobacion-poderes-estado-cayo-criticamente-desde-2011-noticia-1737853.

ANEXOS

ANEXO N°1: MATRIZ DE CONSISTENCIA

TÍTULO: "APLICACION DE LA MINERIA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS**"**

PROBLEMA	OBJETIVOS	HIPÒTESIS	VARIABLES	METODOLOGÌA
Problema General:	**Objetivo General:**	**Hipótesis General:**		
¿Cómo influye la ausencia de aplicación de la minería de datos distribuida usando algoritmo de clustering k-means en la mejora de la calidad de servicios que ofrecen las organizaciones modernas?	Desarrollar un prototipo que aplique minería de datos distribuida mediante el uso de un algoritmo de clustering basado en la técnica k-means.	*"La aplicación de la minería de datos distribuida usando algoritmo de clustering basado en la técnica k-means influye en la mejora de la calidad de servicios de las organizaciones modernas."*	**Variable Independiente:** La aplicación de la minería de datos distribuida usando algoritmo de clustering basado en la técnica k-means. **Variable Dependiente:** Mejora de la calidad de servicios de las organizaciones modernas.	**Tipo de Investigación:** La investigación será básica no experimental **Diseño de Investigación:** Descriptivo, y analítico no experimental.
Problemas Específicos:	**Objetivos Específicos:**	**Hipótesis Especificas:**		
1. ¿Cuáles son las limitaciones que presentan la aplicación de la minería de datos no distribuida en las organizaciones modernas?	1. Elaborar una propuesta algorítmica de clustering basado en la técnica k-means.	1. La aplicación de la minería de datos distribuida influye en la mejora de la calidad de servicios de las organizaciones modernas.		
2. ¿De qué manera se ven limitadas las organizaciones modernas en aspectos relacionadas a la toma de decisiones así como en el soporte a nivel transaccional de sus actividades de negocios debido a la ausencia de soluciones de minería de datos distribuida que proporcionen algoritmos de clustering bajo el enfoque k-means?	2. Desarrollar un prototipo que aplique la minería de datos distribuida	**2.** El algoritmo de clustering basado en la técnica k-means para minería de datos distribuida influye en la mejora de la calidad de servicios de las organizaciones modernas.		

APENDICE A: GLOSARIO

DATO NOMINAL: Son variables numéricas cuyos valores representan una categoría o identifican un grupo de pertenencia. Este tipo de variables sólo nos permite establecer relaciones de igualdad/desigualdad entre los elementos de la variable. La asignación de los valores se realiza en forma aleatoria por lo que no cuentan con un orden lógico.

SERVLET: El servlet es una clase en el lenguaje de programación Java cuyas tareas se ejecutan en un servidor de aplicaciones y no en el navegador web.

SERVICIO WEB: Es una tecnología que utiliza un conjunto de protocolos y estándares que sirven para intercambiar datos entre aplicaciones.

ETL: Proceso de extracción, transformación y carga; es una actividad del ciclo de vida dimensional de datawarehouse.

WEKA: Colección de algoritmos de aprendizaje máquina para tareas de minería de datos.

YALE: Colección de algoritmos de aprendizaje máquina para tareas de minería de datos.