

Universidad de las Ciencias Informáticas

Facultad 6



Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

**Aplicación de minería de datos descriptiva, sobre el área de
Deporte del Sistema de Información de Gobierno (SIGOB).**

Autor(es):

Anabel Suárez Savón

Tutor(es):

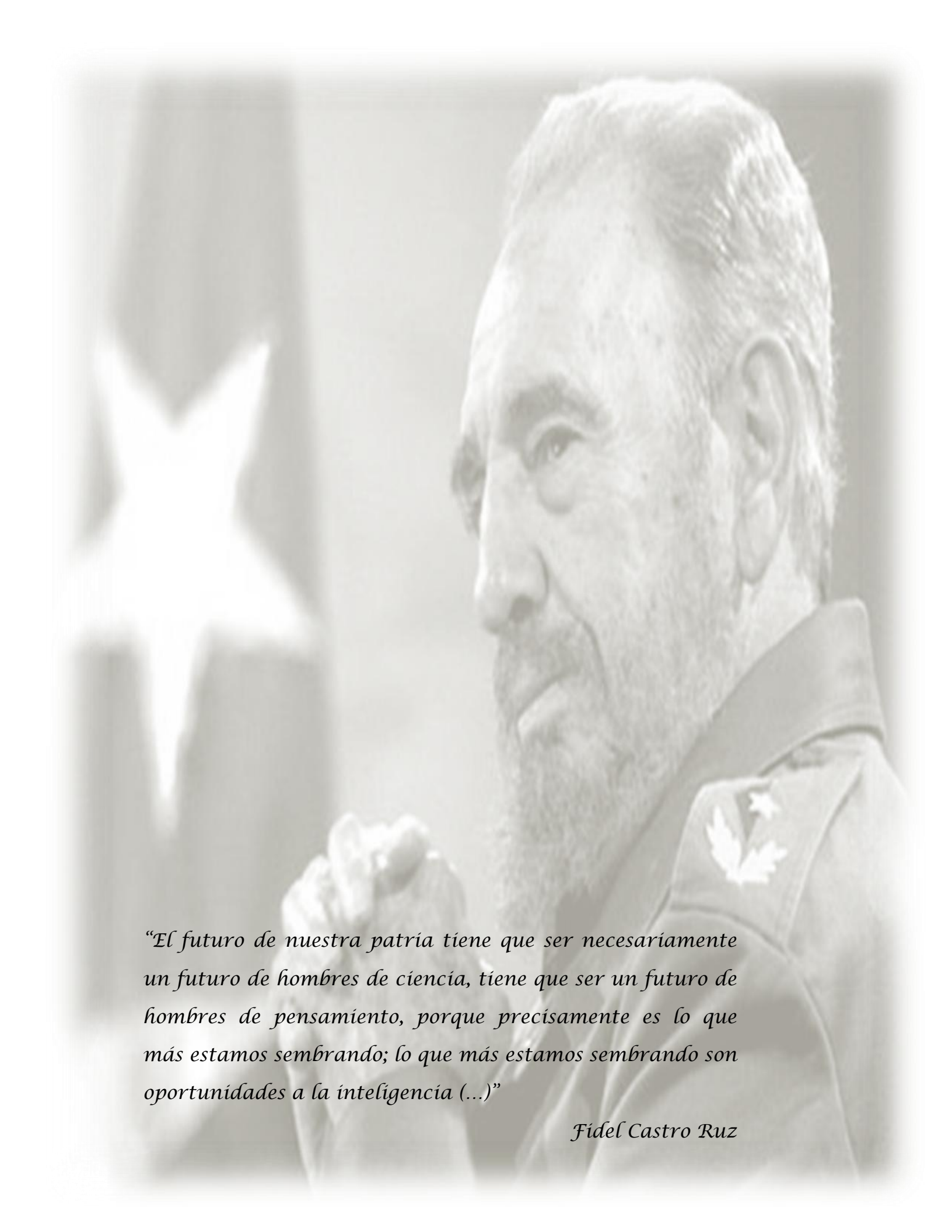
Msc. Maikel Yelandy Leyva Vázquez

Lic. Lisdan Rodríguez Pérez

Ing. Yurima Ibañez Alfonso

La Habana, Junio del 2013

“Año 55 de la Revolución”



“El futuro de nuestra patria tiene que ser necesariamente un futuro de hombres de ciencia, tiene que ser un futuro de hombres de pensamiento, porque precisamente es lo que más estamos sembrando; lo que más estamos sembrando son oportunidades a la inteligencia (...)”

Fidel Castro Ruz

DECLARACIÓN DE AUTORÍA

DECLARACIÓN DE AUTORÍA

Declaro ser autora de la presente tesis que tiene por título: Aplicación de minería de datos descriptiva, sobre el área de Deporte del Sistema de Información de Gobierno y admito los derechos patrimoniales del mismo con carácter exclusivo a la Universidad de las Ciencias Informáticas.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Anabel Suárez Savón (autor)

Msc. Maikel Yelandy Leyva Vázquez (tutor)

Ing. Yurima Ibañez Alfonso (tutora)

Lic. Lisdan Rodríguez Pérez (tutor)

DATOS DE CONTACTO

Tutor: Msc. Maikel Yelandy Leyva Vázquez

Formación Académica: Graduado en la Universidad de Holguín.

Centro Laboral: Universidad de las Ciencias Informáticas (UCI).

Correo Electrónico: mleyvaz@uci.cu

Tutor: Lic. Lisdan Rodríguez Pérez.

Formación Académica: Graduado en la Universidad Central de las Villas.

Centro Laboral: Universidad de las Ciencias Informáticas (UCI).

Correo Electrónico: lisdanrp@uci.cu

Tutor: Ing. Yurima Ibañez Alfonso.

Formación Académica: Graduada en la Universidad de las Ciencias Informáticas.

Centro Laboral: Universidad de las Ciencias Informáticas (UCI).

Correo Electrónico: yibanez@uci.cu

AGRADECIMIENTOS

En mi andar por los caminos estudiantiles hay muchas personas que de una forma u otra han marcado mi vida.

Las primeras personas a quienes les quiero agradecer son mi sostén, mi alegría, mi ejemplo y admiración, los que me dieron la vida y me apoyaron en cualquier tipo de decisión: mis padres que todo mi mundo llenaron siempre de ilusión.

A todos mis familiares que por ser tantos no voy a mencionar nombres les agradezco su apoyo y ayuda en todo momento sin importar las condiciones.

A mis hermanas que aunque son un poco malcriadas me llenan de felicidad.

Por apoyarme y estar junto a mí en todo momento, dándome aliento en los más difíciles, le agradezco a mi novio.

A mi mejor amiga que me acompañó estos 5 años de esfuerzo y lucha.

A mis compañeros de la tierra roja siempre los llevo en mi corazón.

A mis compañeros de grupo desde el 6106 hasta el actual 6507 sin olvidar a aquellos que ya no están en la escuela con nosotros.

Al Club del Frijol que aunque costo tiempo lo logramos.

A mis tutores gracias por su esfuerzo y dedicación en el desarrollo de esta tesis.

A Fidel por crear esta universidad que hoy nos forja como hombres y mujeres capacitados para emprender nuestros sueños futuros.

DEDICATORIA

Mi vida colman de flores y de ejemplos sin cesar, cuidando de mis caídas cuando salgo a caminar, me llenan toda de orgullo y me enseñan sin parar, a mis padres esta tesis se la quiero dedicar.

No me alcanzan las palabras con que pueda expresar, la satisfacción que siento cuando conmigo están, son mi luz, todo mi aliento ,mi apoyo incondicional, son los que desde un principio en mi siempre han de confiar, sin temor a lo que pase siguen siendo mis papás.

RESUMEN

La Minería de Datos se ha convertido en una herramienta muy poderosa que es utilizada para la extracción de conocimiento en grandes bases de datos. Su aplicación es múltiple en sectores como la biología, la medicina, el turismo, la cultura, el deporte, entre otros.

La presente investigación presenta un modelo que permita analizar el comportamiento de los deportistas cubanos en los eventos competitivos, haciendo uso de las técnicas y algoritmos de Minería de Datos. Para la extracción del conocimiento se consulta el Mercado de Datos del área de Deporte del Sistema de Información de Gobierno y se guía la investigación mediante la metodología CRISP-DM con el apoyo de la herramienta de libre distribución Weka en su versión 3.6.2. Se describe su aplicación en un entorno real demostrando la aplicabilidad y la utilidad de la propuesta y finalmente se presentan las conclusiones y las recomendaciones de trabajos futuros.

Palabras claves: CRISP-DM, Minería de Datos, Sistema de Información de Gobierno, Weka.

<i>INTRODUCCIÓN</i>	1
<i>CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA</i>	5
1.1 Introducción.....	5
1.2 Mercado de Datos (MD)	5
1.3 El descubrimiento de conocimientos (KDD).....	5
1.3.1 Metas de KDD	6
1.4 Minería de Datos (DM)	8
1.5 Metodologías de desarrollo para proyectos de DM.....	9
1.5.1 SEMMA	9
1.5.2 CRISP-DM.....	10
1.5.3 Selección de la metodología.....	12
1.6 Técnicas de minería de datos.....	12
1.6.1 Técnicas de tipo Supervisadas o predictivas	13
1.6.2 Técnicas de tipos No supervisadas o descriptivas.....	14
1.6.3 Selección de la técnica de DM.....	15
1.7 Herramientas para el desarrollo del modelo	17
1.7.1 SPSS Clementine.....	17
1.7.2 KNIME.....	17
1.7.3 SAS Enterprise Miner	18
1.7.4 RapidMiner	18
1.7.5 Weka	18
1.7.6 Selección de la herramienta de DM.....	19
1.8 Conclusiones.....	20
<i>CAPÍTULO 2: PROPUESTA DE SOLUCIÓN</i>	21
2.1 Introducción.....	21
2.2 Arquitectura propuesta para el MD deporte	21
2.3 Fases propuestas.....	22
2.3.1 Comprensión del negocio	23
2.3.2 Comprensión de datos.....	25
2.3.3 Preparación y selección de datos	26
2.3.4 Modelado	29

2.3.5 Evaluación.....	30
2.3.6 Despliegue	31
2.4 Conclusiones.....	32
<i>CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.....</i>	<i>33</i>
3.1 Introducción.....	33
3.2 Aplicación de la técnica sobre el MD de Deporte.....	33
3.2.1 Selección de la técnica de modelado.....	38
3.2.2 Construcción del modelo	39
3.2.3 Evaluación de los resultados	41
3.3 Conclusiones.....	46
<i>CONCLUSIONES GENERALES.....</i>	<i>47</i>
<i>RECOMENDACIONES.....</i>	<i>48</i>
<i>TRABAJOS CITADOS.....</i>	<i>49</i>
<i>BIBLIOGRAFÍA.....</i>	<i>51</i>
<i>ANEXOS.....</i>	<i>54</i>

Índice de Tablas

Tabla 1 Datos escogidos para la obtención de la vista minable	35
Tabla 2 Opciones de configuración del algoritmo.....	38
Tabla 3 Resultado de K-means para 4 clúster con varias semillas.....	39

Índice de Figura

Figura 1 Fase de un proyecto de DM	9
Figura 2 Arquitectura de la investigación	22
Figura 3 Fases de CRISP-DM.....	22
Figura 4 Actividades de la fase Comprensión del negocio	23
Figura 5 Actividades de la fase Comprensión de datos.....	25
Figura 6 Actividades de la fase Preparación y selección de los datos.....	27
Figura 7 Actividades de la fase Modelado.....	29
Figura 8 Actividades de la fase Evaluación	31
Figura 9 Actividades de la fase Despliegue.....	31
Figura 10 Años en que se realizan las competiciones.....	34
Figura 11 Deportes en competencia	34
Figura 12 Niveles a los que se realizan las competiciones.....	35
Figura 13 Sexo de los participantes	35
Figura 14 Transformación realizada al hecho participante en eventos competitivos	37
Figura 15 Vista minable final	38
Figura 16 Resultados obtenidos.....	40
Figura 17 Ejemplo de gráfica de dispersión, sexo-mayores	43
Figura 18 Ejemplo de gráfica de dispersión, sexo-juveniles	43

Figura 19 Ejemplo de gráfica de dispersión, sexo-menores	44
Figura 20 Niveles que predominan en los eventos competitivos para mayores.....	45
Figura 21 Niveles que predominan en competiciones de participantes de la etapa Juveniles	45
Figura 22 Niveles que predominan en competiciones de participantes de la etapa Escolares	46
Figura 23 Comportamiento de los participantes escolares en cada deporte por año.....	54
Figura 24 Comportamiento de los participantes juveniles en cada deporte por año	54
Figura 25 Comportamiento de los participantes mayores en cada deporte por año	55
Figura 26 Comportamiento de los participantes escolares en cada año por sexo	55
Figura 27 Comportamiento de los participantes juveniles en cada año por sexo	56
Figura 28 Comportamiento de los participantes mayores en cada año por sexo.....	56
Figura 29 Comportamiento de los participantes escolares en cada año por nivel	57
Figura 30 Comportamiento de los participantes juveniles en cada año por nivel.....	57
Figura 31 Comportamiento de los participantes mayores en cada año por nivel.....	58

INTRODUCCIÓN

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y coleccionar datos, debido básicamente al gran poder de procesamiento de las máquinas así como a su bajo costo de almacenamiento. La habilidad para convertir los datos acumulados durante años en información estratégica e integrada es vital para las instituciones, surgiendo sistemas con nuevas tecnologías que permiten a los usuarios, analizar los datos en busca de información que les ayude a tomar decisiones claves.

Desde su surgimiento, la Universidad de las Ciencias Informáticas (UCI) se ha desempeñado no solo como una institución educacional, sino también como un centro productor de software con proyectos de carácter nacional e internacional. El Centro de Tecnologías de Gestión de Datos (DATEC) de la UCI, en conjunto con la Oficina Nacional de Estadísticas e Información (ONEI), decidió crear el proyecto Sistema de Información de Gobierno (SIGOB), para contribuir a la toma de decisiones en diferentes esferas de la sociedad.

En este proyecto se encuentra implícito el sector del Deporte que representa uno de los rasgos más auténticos de la identidad nacional y se considera una de las áreas más significativas para nuestro país. La misma tiene como objetivo fundamental el perfeccionamiento del sistema de participación deportiva así como la profundización en aspectos técnicos y organizativos, centrando su labor en la extensión masiva del deporte a lo largo y ancho de toda la isla para promover el desarrollo y el esparcimiento de la actividad física a todos los niveles de la población. El trabajo de este organismo se extiende además a las instituciones de alto rendimiento, con especial atención a los atletas y todo tipo de profesionales con el fin de lograr una mejor formación.

Para archivar toda esta información cuenta con un sistema de gestión que almacena los datos históricos del comportamiento de las principales actividades que se efectúan en la misma, por lo cual el volumen de sus fuentes es muy grande y no permite el monitoreo de todos los datos contenidos en este sector. Actualmente nuestro país se encuentra inmerso en el proceso de perfeccionamiento que incluye el sector del Deporte como se plantea en los lineamientos de la Política Económica y Social del Partido y la Revolución (1), aprobados en el Sexto Congreso del Partido. Alineado a este documento resulta importante descubrir conocimiento que

contribuya al perfeccionamiento de dicha actividad y permita mantener a nuestro país como potencia en este sector. Además es necesaria la detección de patrones de comportamiento que tengan un impacto significativo y que apoyen el proceso de toma de decisiones para realizar análisis de tipo descriptivo (2), que de valor agregados a los datos contenidos en esta área.

Atendiendo a la problemática antes descrita el presente trabajo está enfocado en el siguiente **problema a resolver**: ¿Cómo contribuir al proceso de extracción del conocimiento en el mercado de datos de Deporte del sistema de información de gobierno?, teniendo como **objeto de estudio** las técnicas de minería de datos, enmarcándose en el **campo de acción** técnicas de minería de datos descriptivas, sobre el MD de Deporte del Sistema de Información de Gobierno.

El **objetivo general** de la investigación es aplicar la técnica de minería de datos descriptiva, sobre el mercado de datos de Deporte del Sistema de Información de Gobierno que contribuya al proceso de extracción del conocimiento.

Perfilando como **objetivos específicos**:

- ✓ Analizar aspectos teóricos que sustentan el proceso de minería de datos, las particularidades de las técnicas seleccionadas así como de su implementación.
- ✓ Definir el modelo para la minería de datos descriptiva sobre el área de Deporte del SIGOB de la ONEI.
- ✓ Validar la solución propuesta después de la aplicación de la técnica de minería de datos descriptiva al mercado de datos de Deporte.

Para darle cumplimiento al objetivo general se trazaron las siguientes **tareas de investigación**:

- ✓ Realización de estudios bibliográficos sobre las técnicas, herramientas y metodologías empleadas en la minería de datos.

- ✓ Análisis del negocio y del problema a resolver.
- ✓ Caracterización de los datos que serán empleados en el estudio.
- ✓ Definición de los resultados esperados con la aplicación de la minería de datos descriptiva.
- ✓ Realización del pre procesado de los datos.
- ✓ Interpretación de los resultados obtenidos con la aplicación de la técnica de minería de datos descriptiva al mercado de datos de Deporte.

Con el correcto cumplimiento de las tareas se espera obtener el siguiente resultado:

- ✓ La descripción de la técnica para la minería de datos descriptiva aplicada a problemas de segmentación, así como la detección de patrones de conocimiento que permitan evaluar la utilidad de la técnica en el mercado de datos de Deporte.

Para adquirir los conocimientos necesarios que permitan el cumplimiento del objetivo trazado y desempeñar las tareas planeadas, se lleva a cabo una investigación en la que se utilizan los siguientes métodos científicos:

Métodos teóricos:

Analítico sintético: permite realizar un análisis sobre las técnicas descriptivas de DM existentes para la extracción de conocimiento, sintetizando los algoritmos aplicables para el desarrollo de funcionalidades que incluyan análisis inteligente en el MD deporte.

Histórico lógico: este método se utiliza para estudiar las tendencias de las técnicas y algoritmos de DM que se utilizan para el descubrimiento de conocimiento, así como los principales conceptos en esta área y el estado actual. Permite además conocer las herramientas y las metodologías para el desarrollo de proyectos de DM.

Método empírico: Se aplica a los datos contenidos en el MD de Deporte para demostrar la aplicabilidad de la propuesta.

Estructura del documento

Capítulo 1: Fundamentación teórica: Este capítulo aborda sobre los conceptos de mercado de datos (MD), minería de datos (DM por sus siglas en inglés) y descubrimiento del conocimiento (KDD por sus siglas en inglés), además de profundizar en las técnicas, metodologías y herramientas que se utilizarán en el desarrollo de la investigación.

Capítulo 2: Propuesta de la solución sobre el área deporte del sistema de información de gobierno (SIGOB): se define la arquitectura del proyecto de DM, además de los procesos y las actividades que conforman la propuesta.

Capítulo 3: Descripción de los resultados. En este capítulo se analizan los resultados obtenidos en el proceso de minado con la aplicación de la técnica al MD de Deporte.

Al finalizar el documento se presentan las Conclusiones y Recomendaciones derivadas de la investigación, las Referencias Bibliográficas, así como los Anexos que apoyan la comprensión y dan información adicional sobre el trabajo realizado.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA.

1.1 Introducción

En el presente capítulo se precisan un conjunto de elementos que conforman la fundamentación teórica de la investigación. Se profundiza sobre el KDD y los conceptos asociados a uno de sus principales procesos, la DM, se realiza un análisis de las técnicas y herramientas que ayudarán a solucionar el problema planteado, guiados por el estudio de las metodologías que se utilizan en el proceso de DM, caracterizando cada una para la selección de la más óptima para dicha investigación.

1.2 Mercado de Datos (MD)

Un MD almacena la información de un área o departamento específico dentro del negocio, se define como “una base de datos departamental creado para un área específica de una entidad, es alimentado por la integración de un conjunto de fuentes de información y para poder analizar satisfactoriamente la información es necesario una manera de estructurar los datos que contiene” (3).

Con el objetivo de facilitar la construcción y utilización de un almacén de datos se crean los MD, que representan un subconjunto del almacén de datos, referentes a los requisitos de un departamento o área de negocio concreto.

Características de los MD

- ✓ Según las necesidades de los usuarios el diseño del MD se realiza siguiendo una estructura consistente.
- ✓ La información histórica que posee es referente a un determinado departamento.
- ✓ Contiene el grado de granularidad necesaria porque presenta mayor nivel de detalle.
- ✓ Da costes adicionales en hardware, software y accesos de red.
- ✓ Debido a que hay grupos de usuarios que solo acceden a un subconjunto preciso de datos, se hace más fácil el acceso a las herramientas de consulta y divide los datos para controlar mejores accesos (3).

1.3 El descubrimiento de conocimientos (KDD)

En los últimos tiempos debido a la capacidad de generar información, ha crecido considerablemente la cantidad de datos almacenados en bases de datos, este aumento ha provocado que existan

organizaciones que no pueden analizar eficientemente sus fuentes, dificultando el trabajo con todos los datos que registran. Para solucionar este tipo de problemas surge el proceso de KDD que abre las puertas al descubrimiento de información primordial para estas instituciones.

Antes de conocer en qué consiste este proceso se precisa saber que el **conocimiento** es una combinación de ideas, reglas, procesos e información que se aplica para guiar acciones y decisiones. El mismo es una interpretación realizada por la mente, que será válida cuando pueda explicar las interacciones de un problema con su contexto (5).

El **proceso de KDD** consiste en usar métodos de DM para identificar lo que se considera como conocimiento, se encarga de la preparación de los datos y la interpretación de los resultados obtenidos. Además es el proceso que encuentra el significado de los patrones, haciendo que el valor real de los datos que se encuentra en la información se pueda extraer de ellos, obteniendo información que ayude a tomar decisiones en los fenómenos que nos rodean (4).

1.3.1 Metas de KDD

El objetivo del KDD es encontrar patrones relevantes e identificarlos como posible conocimiento presentando la información a los usuarios de forma clara y comprensible. Para llevarlo a cabo se traza las siguientes metas:

- ✓ Procesar automáticamente grandes cantidades de datos crudos.
- ✓ Identificar los patrones más significativos y relevantes.
- ✓ Presentar los patrones como conocimiento apropiado para satisfacer las necesidades del usuario.

Para el cumplimiento adecuado de las metas antes expuestas se fundamenta el proceso en las siguientes etapas:

- ✓ Determinar las fuentes de información: Se obtiene lo que puede ser útil y el lugar dónde conseguir las.
- ✓ Diseñar el esquema de un almacén de datos (Data Warehouse): Unificación de la información recogida.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA.

- ✓ Permitir la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar de los que se han estudiado. Esta es la etapa que puede llegar a consumir el mayor tiempo.

- ✓ **Selección, limpieza y transformación** de los datos que se van a analizar:
 - La **selección** es la fusión tanto horizontal (filas) como vertical (atributos).

 - La **limpieza** y pre procesamiento de datos se logra diseñando una estrategia adecuada para manejar ruidos, valores incompletos, secuencias de tiempo y casos extremos. Seleccionar y aplicar el método de minería de datos apropiado. Incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar.

 - La **transformación** de los datos al formato requerido por el algoritmo específico de DM busca patrones que puedan expresarse como conocimiento o simplemente que expresen dependencias de los datos, los resultados obtenidos dependen de su clasificación y forma de representarlo.

- ✓ Evaluación, interpretación, transformación y representación de los patrones extraídos:

- ✓ Al Interpretar los resultados y posiblemente regresar a los pasos anteriores puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias. Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes irrelevantes.

- ✓ Difusión y uso del nuevo conocimiento.

- ✓ Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente

para almacenarlo y reportarlo a las personas interesadas. En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de DM.

1.4 Minería de Datos (DM)

La DM es fase más relevante dentro del proceso del KDD, consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la DM prepara, sondea y explora los datos para sacar la información oculta en ellos; bajo este nombre se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable y está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos (5).

Etapas de la DM

Aunque en DM cada caso concreto puede ser radicalmente distinto al anterior las fases comienzan con los datos en bruto y finaliza con el conocimiento extraído, el cual fue adquirido como resultado de las siguientes etapas (Figura1):

Determinación de los objetivos: Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista.

Pre procesamiento de los datos: Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del 70% del tiempo total de un proyecto.

Determinación del modelo: Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas.

Análisis de los resultados: Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.



Figura 1 Fase de un proyecto de DM

1.5 Metodologías de desarrollo para proyectos de DM

Cuando se va a realizar un proyecto de DM siempre es necesario contar con una metodología que guíe todo el proceso. De esta manera diversas empresas han especificado y propuesto procesos de modelado con el objetivo de guiar al desarrollador a través de una serie de pasos dirigidos a obtener buenos resultados, las mismas nos facilitan:

- ✓ Hacer una buena planificación y dirección del proyecto.
- ✓ Realizar un seguimiento de calidad del proyecto, para lograr la satisfacción del cliente.
- ✓ Desarrollar nuevos proyectos de DM con características similares.

Para guiar la investigación se analizan las siguientes metodologías:

1.5.1 SEMMA

Esta metodología se define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a los cinco pasos básicos del proceso: Sample (Muestra), Explorer (Explorar), Modify (Modificar), Model (Modelo) y Assess (Evaluar) (6).

Pasos básicos.

Muestreo

Extracción de la población sobre la cual se va a aplicar el análisis. En ocasiones se trata de una muestra aleatoria pero puede ser también un subconjunto de datos del almacén de datos que cumplan unas condiciones determinadas. El objeto de trabajar con una muestra de la población en lugar de con toda ella, es simplificar el estudio y la disminución de la carga del proceso (6).

Exploración

Una vez determinada la población que sirve para la obtención del modelo se deberá determinar cuáles son las variables explicativas que van a servir como entradas al modelo. Para ello es importante hacer una exploración de la información disponible de la población que permita eliminar variables que no influyen y agrupar aquellas que presentan efectos similares. El objetivo es simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo (6).

Manipulación

Tratamiento realizado sobre los datos de forma previa al modelado, en base a la exploración realizada, de forma que se definan claramente las entradas del modelo a realizar (selección de variables explicativas, agrupación de variables similares) (6).

Modelado

Permite establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibilitan inferir el valor de las mismas con un nivel de confianza determinado (6).

Valoración del modelo

Después de todo el trabajo realizado, se comparan los modelos a partir de los cuales se obtienen patrones de negocio (6).

1.5.2 CRISP-DM

La metodología CRISP-DM por sus siglas en inglés, Cross Industry Standard Process for Data Mining) fue concebido a fines de 1996 por tres compañías DaimlerChrysler (liderando aplicaciones de minería a

negocios), SPSS (servicios de DM) y NCR (Compañía que se dedica al desarrollo de los almacenes de datos). Es una metodología de libre distribución que está basada en la experiencia práctica, de cómo las personas efectúan proyectos de DM. La misma, estructura un proyecto de DM en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

Características de la Metodología CRISP-DM

La metodología de CRISP-DM está descrita en términos de un proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada e instancia de procesos. En el nivel superior, el proceso de DM es organizado en seis fases; cada fase consta de varias tareas genéricas de segundo nivel, las cuales deben cubrir el proceso entero de minería de datos y todas las aplicaciones posibles; el proceso debe ser válido para acontecimientos normales y desarrollos imprevistos (7).

El nivel de tarea especializada, es el lugar para describir cómo las acciones en las tareas genéricas deben ser realizadas en ciertas situaciones específicas. Además describe como una tarea se diferencia en distintas situaciones, cómo la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo (7).

Fases de la metodología CRISP-DM

Comprensión del negocio: Se enfoca en la comprensión de los objetivos del proyecto desde una perspectiva del negocio y a partir de este conocimiento se define el problema de la DM de la investigación (8).

Comprensión de los datos: Comienza con la recopilación de los datos iniciales, se realiza un análisis para detectar problemas con la calidad y veracidad de los mismos y finalmente definir las posibles variables seleccionadas para formar hipótesis en cuanto a la información oculta y desconocida (8).

Preparación y selección de datos: Se selecciona el conjunto de datos finales para la construcción del modelo a partir de los datos iniciales. Las tareas de la fase muchas veces tienen que ser repetidas y en cualquier orden, estas pueden ser: la selección de tablas y atributos, así como la transformación y la limpieza de datos para las herramientas de modelado (8).

Modelado: Se selecciona la técnica de modelado, se interpreta y se obtienen un grupo de patrones. En ocasiones algunas técnicas tienen limitantes con el tipo de datos de los atributos y a veces es necesario volver a la fase de preparación de los datos (8).

Evaluación: Después de obtener el o los patrones de conocimiento y antes de proceder al despliegue final del proyecto, es importante evaluar y revisar cada uno de los pasos de la DM para demostrar que responden a los objetivos del negocio de la investigación (8).

Despliegue: La creación del modelo no es generalmente el final del proyecto, el conocimiento obtenido debe ser organizado y presentado en el modo en el que el cliente pueda usarlo (8).

1.5.3 Selección de la metodología

Después de realizar el análisis sobre las características de cada una de las metodologías que son utilizadas en la DM, se propone CRISP-DM como metodología para guiarla investigación respaldando su elección con las siguientes ventajas:

- ✓ Es de libre distribución, permitiendo trabajar con cualquier herramienta de modelado.
- ✓ Proporciona facilidades en la planificación, documentación y comunicación en los proyectos de DM.
- ✓ Facilita y organiza el trabajo a partir de especificar un grupo de tareas en cada una de las fases.
- ✓ Presenta una precisa y sólida distribución de tareas de carácter general con sus resultados, así como una guía para su desarrollo.

1.6 Técnicas de minería de datos

Las técnicas de la minería de datos provienen de la Inteligencia Artificial y de la Estadística. Dichas técnicas no son más que algoritmos, más o menos sofisticados, que se aplican sobre un conjunto de datos para obtener resultados. Estas son clasificadas y asociadas a uno de los siguientes grupos:

Supervisadas o predictivas: En este tipo de técnica, a partir de un conjunto de ejemplos, denominados de entrenamiento de un cierto dominio, se pueden construir criterios para determinar el valor del atributo clase en un ejemplo cualquiera del dominio. Esos criterios están basados en los valores de uno o varios de los otros pares (atributo; valor) que interviene en la definición de los ejemplos. Es sencillo transmitir esa idea al caso en el que el atributo que juega el papel de la clase sea uno cualquiera o con más de dos valores (6).

No supervisadas o descriptivas: Este tipo de técnica aborda el aprendizaje sin supervisión, trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la guía de un atributo en especial. Su proceder es el de los sistemas que realizan agrupamiento conceptual, además de permitir sintetizar conocimiento cualitativo o cuantitativo (6).

1.6.1 Técnicas de tipo Supervisadas o predictivas

Redes neuronales: Es una técnica que modela computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas del cerebro. Las redes de neuronas constituyen una nueva forma de analizar la información con una diferencia fundamental respecto a las técnicas tradicionales: son capaces de detectar y aprender complejos patrones y características dentro de los datos. Una red neuronal se compone de unidades llamadas neuronas conectadas entre sí. Cada neurona recibe una serie de entradas y proporciona una salida, que será la entrada de la siguiente neurona a la que está conectada. Las conexiones entre las neuronas tienen un peso que se usa para ponderar el valor que cada neurona transmite. Al recibir las entradas, cada neurona calcula la suma ponderada de sus entradas y, en la mayoría de los casos, aplica una función de activación para transformar el valor obtenido en una salida válida (9).

Clasificación: Es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a variable(s) específica(s) las cuales se están tratando de predecir. Para este tipo de técnica existen algoritmos como los árboles de decisión que destacan por su uso, eficacia y eficiencia (5).

Algoritmos genéticos: Son métodos numéricos de optimización, en los que aquella variable o variables que se pretenden optimizar junto con las variables de estudio constituyen un segmento de información. Aquellas configuraciones de las variables de análisis que obtengan mejores valores para la variable de respuesta, corresponderán a segmentos con mayor capacidad reproductiva. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Se puede además introducir elementos aleatorios para la modificación de las variables (mutaciones). Al cabo de cierto número de iteraciones, la población estará constituida por buenas soluciones al problema de optimización, pues las malas soluciones han ido descartándose, iteración tras iteración (9).

Series de tiempo: Permite el estudio de la evolución de una variable a través del tiempo para poder realizar predicciones, a partir de ese conocimiento y bajo el supuesto de que no van a producirse cambios estructurales. La Secuencia de los métodos de análisis de series de tiempo son usados para relacionar los eventos con el tiempo (5).

Árbol de decisión: Es un modelo de predicción que dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Los árboles de decisión son una técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas. Cada nodo interno denota una prueba sobre un atributo. Cada rama representa el resultado de una prueba. Las hojas denotan las clases. Es una magnífica herramienta para el control de la gestión empresarial, pues ayudan a las empresas a determinar cuáles son sus opciones al mostrarles las distintas decisiones y sus resultados (9).

1.6.2 Técnicas de tipos No supervisadas o descriptivas

Análisis de varianza: Mediante el mismo se evalúa la existencia de diferencias significativas entre las medias de una o más variables en poblaciones distintas (5).

Análisis discriminante: Permite la clasificación de individuos en grupos que previamente se han establecido, permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto una mejor identificación con las variables que definan la pertenencia al grupo (5).

Asociación: Incluye técnicas conocidas como linkage analysis, utilizadas para buscar patrones que tienen una probabilidad alta de repetición, como ocurre al analizar una canasta en la búsqueda de productos afines (5).

Agrupamiento: Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes de tal manera que se maximice la similitud entre los vectores de un mismo grupo y se minimice la similitud entre los grupos, además esta técnica puede ser combinada fácilmente con cualquier otra. El objetivo fundamental del agrupamiento es determinar el comportamiento de un nuevo vector, a partir de las características del mismo, se define a qué grupo pertenecerá y que acción podrá realizar (9).

1.6.3 Selección de la técnica de DM

Clustering o agrupamiento es la técnica seleccionada en la investigación ya que la misma es el punto de partida en cualquier proceso de DM, siendo la primera técnica a aplicar en cualquier problema de esta índole, se ajusta a la solución del objetivo general y su aplicación se pueden obtener patrones que describan el comportamiento del MD de Deporte del SIGOB.

Algoritmos de la técnica seleccionada

COBWEB: Se caracteriza porque utiliza aprendizaje incremental, realizando las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol de clasificación donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso. La actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la reestructuración de todo el árbol. La clave para saber cómo y dónde se debe actualizar el árbol la proporciona una medida denominada utilidad de categoría, que mide la calidad general de una partición de instancias en un segmento. Este método pertenece a la familia de aprendizaje conceptual considerando cada instancia como un modelo. Su principal desventaja radica en que solo trabaja con atributos nominales (10).

EM: Este algoritmo se utiliza para segmentar conjuntos de datos, clasificado como un método particionado que realiza clustering probabilístico. Se trata de buscar la FDP (función de densidad de probabilidad) desconocida a la que pertenecen el conjunto completo de datos, los resultados de esta FDP se aproximan mediante combinaciones lineales. Cada grupo se corresponde con las respectivas muestras de datos que pertenecen a cada una de las densidades mezcladas, finalmente permite obtener un conjunto de clusters que agrupan el conjunto de proyectos original cuyos parámetros tienen una distribución normal. Este algoritmo solo permite el trabajo con datos numéricos siendo esto una de sus principales desventajas (10).

Algoritmo escogido para la investigación

Simple K-means es el algoritmo de agrupamiento más conocido y se encuentra clasificado como un algoritmo particional. Está basado en la minimización de la distancia interna entre los elementos de los K grupos definidos y esto constituye una de sus debilidades porque en ocasiones la cantidad de particiones no es la más óptima. Es importante destacar las características, pasos y ventajas de este algoritmo que fundamentan su selección para la investigación.

Este algoritmo tiene como ventaja que trabaja con atributos (numéricos, binarios, nominales, ordinales), funciona eficientemente con una gran cantidad de datos y los grupos pueden ser interpretados y convertir esta información en conocimiento.

Características del Simple K-means.

- ✓ Escalabilidad: normalmente corre con grandes cantidades de datos.
- ✓ Manejo de ruido: sensible a datos erróneos
- ✓ Grupos de formas arbitrarias: basado en distancias numéricas.

Pasos para aplicar el algoritmo Simple K-means.

1. Especificar la cantidad de grupos (K) que se van a crear y se selecciona por cada uno elementos de manera aleatoria que serán los centros de los grupos.
2. Cada una de las instancias, es asignada al grupo con características similares más cercano.
3. Se calcula el valor del punto medio de todas sus clases y se toma como el centro de sus respectivos grupos.

4. Se repite el paso anterior, hasta que el valor de los centroides no varíe más, después de cada iteración.

La técnica escogida para la investigación a pesar de su sensibilidad para el trabajo con datos erróneos, nos permite trabajar con todo el volumen de datos contenido en las fuentes del MD de Deporte, siendo eficiente a la hora de procesar atributos numéricos y nominales. Además muestra de forma clara la representación gráfica de los atributos que se evalúan para determinar los patrones de conocimientos.

1.7 Herramientas para el desarrollo del modelo

En la actualidad existe una gran cantidad de herramientas tanto libres como comerciales para el desarrollo de proyectos de DM. Estas herramientas son utilizadas para resolver problemas del mundo real en la ingeniería, la ciencia, los negocios y otros entornos. Son utilizadas para resolver situaciones donde el volumen de datos es muy grande por la cantidad de variables que se manipulan, o la extracción de conocimiento se hace compleja. A continuación se exponen algunas de las herramientas más completas y utilizadas en la DM.

1.7.1 SPSS Clementine

Es una herramienta visual comercializada por SPSS constituye uno de los sistemas más populares en el mercado. Esta herramienta fue comprada en julio del 2009 por la compañía IBM¹, por lo que pasó a conocerse como IBM SPSS² Modeller. Es un potente software que combina modernas técnicas de modelación con poderosas herramientas de acceso, manipulación y exploración de datos en una interfaz simple. Posibilita de forma rápida desarrollar y desplegar modelos que apoyen la toma de decisiones. Entre sus características a destacar está el hecho de que a diferencia de otras herramientas que se centran en el modelado, ella apoya el ciclo completo de KDD y está diseñada bajo la metodología CRISP-MD, es un software privativo y con grandes requerimientos de hardware, siendo estas sus principales desventajas (11).

1.7.2 KNIME

Es un entorno totalmente gratuito para el desarrollo y ejecución de técnicas de DM. Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, en la actualidad continúa su desarrollo, además de prestar servicios de formación y consultoría. El mismo

está desarrollado sobre la plataforma Eclipse y programado esencialmente en Java, su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos (12).

1.7.3 SAS Enterprise Miner

Es una herramienta proporcionada por SAS agiliza el análisis de DM creando modelos descriptivos y predictivos de alta precisión basados en el análisis de una gran cantidad de datos. SAS Enterprise Miner 5 posee una arquitectura cliente/servidor, basado en Java, puede ser desarrollado tanto en la plataforma de Windows como en Linux. Este software tiene una interfaz de usuario fácil de usar, soporta el proceso de DM para crear modelos descriptivos y predictivos, sobre la base del análisis de las grandes cantidades de datos que tienen las empresas. Su diseño está inspirado en la metodología SEMMA. Presenta la implementación de algoritmos que proveen modelos predictivos y descriptivos, tales como árboles de decisión, redes neuronales, asociación, agrupamiento, entre otros, además de tener incluido un potente visualizador gráfico para representar los resultados mediante gráficos en dos o tres dimensiones; así como un generador automático de reportes que resume los resultados en un informe (13).

1.7.4 RapidMiner

Contiene nuevos formatos de entrada de datos con operadores para Microsoft Excel y SPSS, es una herramienta flexible para aprender y explorar la DM. La interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área. Desde la perspectiva de la visualización ofrece representaciones de datos en dispersión en 2D y 3D (13).

1.7.5 Weka

Es un software desarrollado bajo la licencia GPL, es de código abierto e incluye una interfaz gráfica compuesta por diversos entornos, desarrollada por un grupo de investigadores de la Universidad de Waikato de Nueva Zelanda. Se destaca por la cantidad de algoritmos que presenta así como por la eficacia de los mismos. Aunque la herramienta está implementada en Java no presenta problemas de portabilidad mientras que el sistema disponga de la máquina virtual adecuada, convirtiéndolo en un sistema multiplataforma. Entre sus principales características se encuentra el poseer una interfaz gráfica de usuario compuesta de cuatro entornos que permiten diferentes funcionalidades y formas de análisis (11).

1.7.6 Selección de la herramienta de DM

Luego de realizar el estudio de las herramientas de DM más utilizadas en el mundo, se puede determinar que SPSS Clementine y SAS Enterprise Miner constituyen aplicaciones líderes en el mercado, basadas en las metodologías CRISP-DM y SEMMA respectivamente, estas presentan la desventaja de que son herramientas comerciales y su adquisición puede ser altamente costosa. Por otro lado se encuentra RapidMiner y Weka que son aplicaciones de código abierto y de libre distribución, las cuales no comprometen su uso con una metodología en particular y son multiplataforma. De acuerdo con las características antes descritas se decidió utilizar para el desarrollo de esta investigación una herramienta de código abierto, de libre distribución e independiente de cualquier metodología.

Por las ventajas que presenta a su favor y la trayectoria satisfactoria en los procesos de DM se opta por la herramienta de análisis de datos Weka en su versión 3.6.2. Fundamentando su elección en las siguientes razones:

- ✓ Es un software específicamente diseñado y utilizado para investigación y fines educativos. Por esta razón, los elementos que brinda de salida, no están orientados exclusivamente hacia la obtención de herramientas para el dominio de aplicación, sino también hacia la obtención de información acerca del proceso de minería y de la calidad de los resultados obtenidos.
- ✓ Es una herramienta bajo el esquema de licenciamiento público, su uso es totalmente gratis, lo cual facilita su aprovechamiento en esta investigación.
- ✓ Adicionalmente a ser una herramienta de uso libre, su código fuente es abierto, lo que significa que no solo se puede hacer uso de los algoritmos implementados, sino también puede analizarse la implementación realizada de cada uno de ellos.

Herramienta para la transformación de los datos

Pentaho Data Integration en su versión 4.2.1 es una herramienta multiplataforma, lo que permite ejecutarla en cualquier sistema operativo. Está basada en dos tipos de objetos; las transformaciones que contienen una colección de pasos en un proceso de Extracción transformación y Carga (ETL) y los trabajos que poseen una colección de transformaciones y/o trabajos. Además brinda soporte para metadatos, así como

funciones que permiten operar con los campos en el flujo de datos, renombrando, calculando campos en función de otros, correlacionando valores y realizando búsquedas auxiliares en bases de datos.

Además esta herramienta es fácil de usar, brinda la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. Añade diseño para mejorar la productividad del desarrollador, ya que se pueden agregar puntos de ruptura condicionales en la ejecución de las transformaciones, dando la posibilidad de pausar y resumir la ejecución de la transformación, así como especificar el número de filas que se van a usar en las ejecuciones de prueba. Además, se pueden añadir registros personalizados. Cuenta con una gran comunidad de usuarios.

1.8 Conclusiones

A partir de los resultados del estudio realizado en este capítulo se llegaron a las siguientes conclusiones:

- ✓ El KDD está compuesto por varias fases donde la Preparación de datos constituye un papel fundamental para las próximas etapas del proceso.
- ✓ La DM es un poderoso campo para la obtención de conocimiento a partir de la aplicación de técnicas y algoritmos. Permite determinar patrones de comportamiento en grandes volúmenes de datos.
- ✓ Las herramientas disponibles permiten automatizar gran parte de la tarea de encontrar los patrones de comportamiento ocultos en los datos y aplicar las transformaciones necesarias para esto. Dentro de estas herramientas se selecciona Weka en su versión 3.6.2 para apoyar el proceso de DM y Pentaho Data Integration en su versión 4.2.1 para realizar las transformaciones necesarias sobre los datos.
- ✓ Se selecciona la metodología CRISP-DM debido a su amplitud y flexibilidad en proyectos de DM, ya que cuenta con las tareas necesarias para organizar cada paso a realizar en dicha investigación.

CAPÍTULO 2: PROPUESTA DE SOLUCIÓN.

2.1 Introducción

En este capítulo se analizan las fases de la metodología antes propuesta para la investigación. Además se propone la arquitectura a utilizar y se especifican cada uno de los pasos que se siguen para aplicar la técnica descriptiva.

2.2 Arquitectura propuesta para el MD deporte

La arquitectura planteada para la investigación está compuesta por tres subsistemas y niveles, donde cada uno de los subsistemas es ubicado dentro de un nivel. A continuación se describe cada una de los subsistemas:

- El **subsistema de integración** se abastece del MD del área deporte y se encarga de integrar, estandarizar y limpiar los datos que serán extraídos, utilizando la herramienta Pentaho Data Integration en su versión 4.2.1, los cuales serán empleados para la generación del modelo de conocimiento. Solo accederán a este subsistema los clientes que administrarán el proceso de integración mediante el protocolo de conexión TCP/IP¹.
- El **subsistema de almacenamiento**: almacena los datos que son tratados por el subsistema de integración en ficheros de texto.
- El **subsistema de visualización**: consulta los datos que están en el subsistema de almacenamiento mediante la herramienta Weka en su versión 3.6.2. A dicho subsistema acceden los clientes para obtener los modelos deseados.

¹Protocolo de Control de Transmisión/Protocolo de Internet



Figura 2 Arquitectura de la investigación

2.3 Fases propuestas

En la investigación se proponen un conjunto de fases para la aplicación de DM descriptiva sobre el área de Deporte del SIGOB, dicha propuesta está basada en la metodología CRISP-DM que cuenta con seis fases (Figura 3).

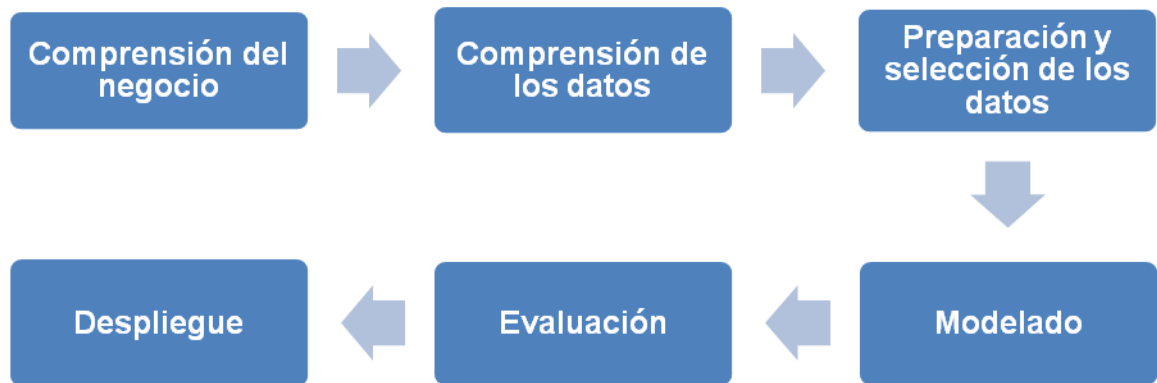


Figura 3 Fases de CRISP-DM

A continuación se describen cada una de estas fases:

2.3.1 Comprensión del negocio

Esta fase cuenta con seis tareas o actividades que nos permiten analizar el negocio de manera sencilla y breve, aportando claridad al desarrollo de la investigación. La descripción de sus actividades (Figura 4) se detalla a continuación.

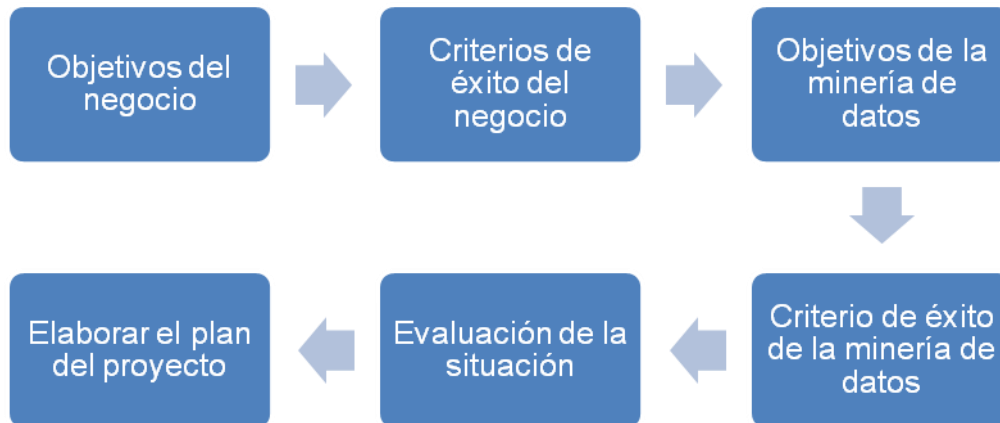


Figura 4 Actividades de la fase Comprensión del negocio

Objetivos del negocio

Esta actividad nos adentra en el negocio antes planteado en busca de lo que se quiere lograr, con la aplicación de la técnica propuesta sobre el área de Deporte del SIGOB, siendo fundamental entender el negocio propuesto. En dicha investigación el principal objetivo es:

Encontrar patrones de comportamiento en el área de Deporte del SIGOB que permitan comprender a los usuarios la eficiencia y fortaleza de la técnica propuesta.

Criterios para lograr el éxito del negocio

EL éxito de la investigación depende de una buena descripción desde el punto de vista del negocio, es útil para obtener el resultado propuesto sobre el MD de Deporte del SIGOB. Los criterios que se tuvieron en cuenta son:

- ✓ Hacer uso de la herramienta Weka para la obtención del modelo.

CAPÍTULO 2: PROPUESTA DE SOLUCIÓN.

- ✓ Aplicar la DM realizando los pasos que se indican en la metodología.
- ✓ Evaluación de la situación existente en el MD de Deporte del SIGOB.

Objetivos de la DM

- ✓ Se proponen aspectos técnicos que nos permitan cumplir con los objetivos del negocio.
- ✓ Obtener conocimiento que tengan un impacto significativo sobre el área de Deporte del SIGOB.

Criterios de éxito de la MD

Se define que se debe tener en cuenta desde la vista técnica para obtener un resultado satisfactorio en el negocio. Para la investigación el criterio de éxito en la DM es:

Evaluar que la técnica sea aplicada correctamente sobre el MD de Deporte del SIGOB, en busca de la solución de mejores resultados.

Evaluación de la situación

Se estudian los recursos disponibles en la investigación, se determinan restricciones y disponibilidad de cada uno de los recursos, así como los riesgos que pueden aparecer en el desarrollo del negocio. En el MD de Deporte se tienen en cuenta una serie de medidas de seguridad para la protección de los datos, ya que cualquier modificación en los mismos puede afectar el resultado de la investigación:

- ✓ Restringir el acceso a la base de datos del personal no autorizado.
- ✓ Proteger los recursos computacionales destinados a la realización de la investigación.
- ✓ Documentar cada fase de la metodología.
- ✓ La investigación será entregada en formato digital.

Elaborar el plan del proyecto

Se describe el plan para alcanzar los objetivos de la DM, se exponen los pasos a seguir en cada momento de la investigación, incluyendo las técnicas y herramientas a utilizar.

Plan para el MD de Deporte:

- ✓ Analizar los datos contenidos en el MD de Deporte en busca de atributos relevantes que describan los participantes en los eventos competitivos que en esta área se realizan.
- ✓ Seleccionar los atributos antes analizados y transformarlos de ser necesario.
- ✓ Obtener la vista minable que contienen los datos necesarios para obtener el modelo.
- ✓ Aplicar la técnica y algoritmo propuesto para obtener el conocimiento.
- ✓ Presentar los resultados obtenido mediante graficas que evidencien su comprensión.

2.3.2 Comprensión de datos

Esta tarea es esencial dentro de la investigación, nos permite determinar cuáles son los datos más relevantes dentro del negocio y cuales aportan la información que se quiere encontrar. A continuación se describen las actividades (Figura 5) contenidas en esta fase.



Figura 5 Actividades de la fase Comprensión de datos

Recolectar y describir los datos iniciales

Se confecciona una tabla con el conjunto de datos escogidos, sus localizaciones, y la manera en que se obtienen. Además se describen los mismos incluyendo cantidad y tipo de dato, evaluando si satisface las necesidades.

Para la selección de los datos que están almacenados en el MD de Deporte se escogen tres hechos, que son el objeto a analizar, con atributos generalmente de tipo cuantitativo. Sus valores se obtienen por la

aplicación de una función estadística que resume un conjunto de valores en un único valor. Los hechos seleccionados son aquellos que mas tuplas contienen dentro del MD.

- ✓ participantes_ eventos competitivos.
- ✓ títulos_ ganados_ Cuba.
- ✓ olimpiadas_ deporte_ cubano.

Explorar y verificar la calidad de los datos

Esta tarea está dirigida a responder interrogantes de la DM usando la visualización, se incluyen gráficos para indicar las características de datos. Se examina la calidad de los datos en relación a si son correctos o contiene errores, que tan reales son estos y si existen valores omitidos.

La exploración absoluta de los datos es lo primero que se realiza antes de cualquier análisis que tiene como fin obtener conocimiento a partir de ellos. En el desarrollo de la investigación no fue necesario realizar esta tarea de la metodología CRISP-DM, ya que no fue preciso almacenar los datos, porque fueron tomados directamente del MD de Deporte del SIGOB. Para la verificación de los datos se realiza un estudio en cuanto a:

- ✓ Representación de la realidad.
- ✓ Campos sobrados.
- ✓ Existencia de campos vacíos.

Determinando que los mismos están listos para realizar las transformaciones pertinente, ya que no cuentan con ninguna de las dificultades antes señaladas.

2.3.3 Preparación y selección de datos

En esta tarea se analizan los datos que son relevantes para la obtención del conocimiento, se le realizan las transformaciones necesarias en cuanto a (cambio de nombre y tipo de dato) y se eliminan aquellos atributos que son nulos o por su irrelevancia no aportan información útil para la investigación.

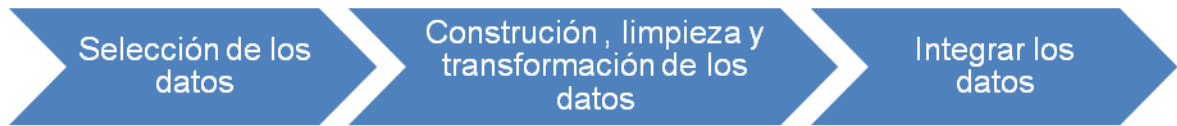


Figura 6 Actividades de la fase Preparación y selección de los datos

Selección de los datos

Se decide los datos que serán excluidos y utilizados para el análisis, de acuerdo a su importancia respecto a los objetivos de la DM, su calidad, y las restricciones técnicas. Antes de conocer que dimensiones y hechos serán utilizados en la investigación detallamos que utilidad tienen las dimensiones:

Las dimensiones se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado. Constituyen las perspectivas de análisis de la información y presentan entre sus características principales la definición de jerarquías entre sus atributos, cuyo objetivo es modelar explícitamente la forma en que se puede consolidar el proceso de análisis de la información. En el MD de Deporte se escogen las dimensiones que permiten describir los hechos escogidos.

Dimensiones

- sexo
- nivel
- deporte
- año
- evento_ deportivo

Hechos

- participantes_ eventos_ competitivos
- títulos_ ganados_ Cuba
- olimpiadas_ deporte_ cubano

Construcción, limpieza y transformación de los datos

Se incluye la construcción de operaciones de preparación de datos y las transformaciones de sus valores. De igual manera, se describe la creación de registros completamente nuevos y que para el modelado

CAPÍTULO 2: PROPUESTA DE SOLUCIÓN.

puedan tener sentido, se describen las decisiones y acciones que fueron tomadas para limpiar o solucionar los problemas de calidad de datos detectados. Los atributos transformados durante el desarrollo de la investigación son aquellos los cuales sus nombres son muy extensos, además al transformar dichos atributos no se afectan el sentido ni el resultado:

- cantidad_participantes_mayores
- cantidad_participantes_juveniles
- cantidad_participantes_menores
- cantidad_medalla_oro
- cantidad_medalla_plata
- cantidad_medalla_bronce
- evento_deportivo
- cantidad_participantes_femeninas
- cantidad_participantes

Integrar los datos

Se resumen la información que es producto de la combinación de varias tablas, para crear la vista minable que será utilizada posteriormente para la obtención del conocimiento. Para la investigación se obtienen tres vistas minable producto de la unión de las siguientes tablas:

Tablas unidas para obtener la vista minable para los participantes_eventos_competitivos

- sexo
- nivel
- deporte
- año
- participantes_eventos_competitivos

Tablas unidas para obtener la vista minable de los títulos_ganados_Cuba

- sexo
- nivel

- deporte
- año
- evento_ deportivo
- títulos_ ganados_ Cuba

Tablas unidas para obtener la vista minable de las olimpiadas_ deporte_ cubano

- deporte
- año
- olimpiadas_ deporte_ cubano

Las fases anteriormente descritas se profundizan en el capítulo 3 con la aplicación de cada una al MD de Deporte, en el caso de las que tres fases restantes, se describen y aplican en su totalidad en el próximo capítulo.

2.3.4 Modelado

Esta fase responde al objetivo propuesto para la investigación, aquí se selecciona la técnica a aplicar y el diseño del modelo que se quiere generar. Se obtiene el modelo y las descripciones que se le realizan al mismo. Las actividades de esta fase (Figura 7) se describen a continuación

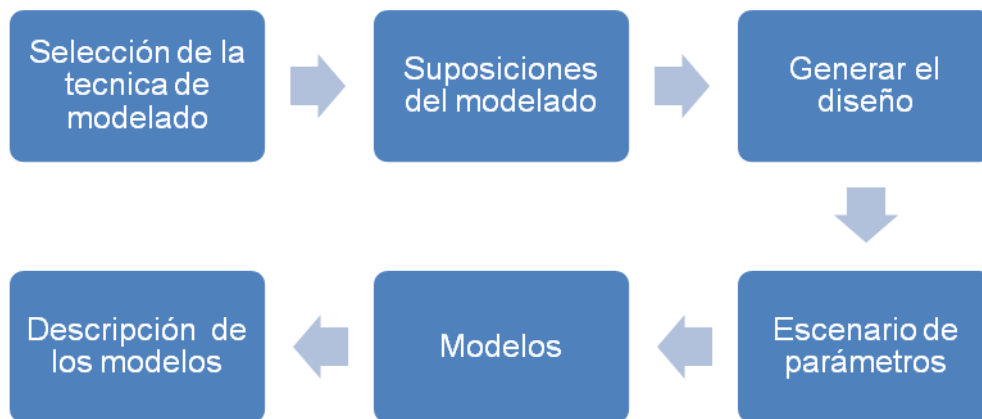


Figura 7 Actividades de la fase Modelado

Selección de la técnica de modelado

Se escoge la técnica que será empleada en la investigación para la obtención de los patrones de conocimientos. La técnica escogida para la investigación es agrupamiento y el algoritmo Simple K-Means.

Suposiciones del modelo

Se analizan todas las condiciones que debe cumplir la vista minable para aplicar la técnica de DM.

Generar el diseño

Se describe el plan intencionado para el entrenamiento, la prueba, y la evaluación de los modelos. Un componente primario del plan determina como dividir un conjunto de datos disponible en, datos de entrenamiento, datos de prueba, y conjunto de datos de validación.

Escenario de parámetros

Se listan los parámetros y sus valores escogidos, así como el razonamiento para elegir los parámetros de ajustes.

Modelos

Se listan los modelos reales producidos por la herramienta de modelado, no un informe. Los modelos se obtienen del resultado de generar instancias, además de la interpretación de las graficas que se generan después de unir diferentes atributos.

Descripción de los modelos

Se describen los modelos obtenidos, informándose su interpretación y documentándose cualquier dificultad encontrada con sus significados.

2.3.5 Evaluación

Esta fase es la que nos permite volver a revisar el proceso antes descrito, además de evaluar si los modelos son los correctos. Sus actividades (Figura 8) se describen a continuación.



Figura 8 Actividades de la fase Evaluación

Evaluación y aprobación del modelo

Se resumen los resultados de la DM en términos de criterios de éxito del negocio. Después de la valoración de los modelos, se toma una decisión con respecto a lo obtenido.

Revisar el proceso

Se resume la revisión de proceso destacándose las actividades que han sido omitidas y/o aquellas que deberían ser repetidas.

2.3.6 Despliegue

Esta fase nos permite obtener la documentación de todo el proceso antes descrito, para cumplir este fin se apoya en las siguientes actividades (Figura 9).



Figura 9 Actividades de la fase Despliegue

Planificación de la explotación

De acuerdo al desarrollo de los resultados de la DM en el negocio, se determina una estrategia para su despliegue, donde se incluyen los pasos necesarios y como realizarlos.

Planificar el monitoreo y el mantenimiento

Se resume la estrategia de supervisión y mantenimiento, incluyendo los pasos necesarios y como realizarlos, a fin de evitar largos periodos innecesarios de uso incorrecto de los resultados de minería de datos.

Producir reportes finales

Se redacta un informe escrito final del compromiso de la minería de datos, lo que incluye todo el desarrollo anterior, y el resumen y la organización de los resultados. A menudo se realizará una reunión en la conclusión en la que los resultados son presentados verbalmente. De igual modo se resumen las experiencias importantes ganadas durante el proyecto.

2.4 Conclusiones

En el este capítulo se define:

- ✓ El análisis de las características del problema permitió definir una arquitectura dividida en tres subsistemas (Integración, Almacenamiento y Visualización).
- ✓ A partir del estudio metodología CRISP-DM y de las características del problema se definieron las fases (Comprensión del negocio, Compresión de los datos, Selección y preparación de los datos, Modelado, Evaluación, Despliegue) para guiar la investigación.
- ✓ Las características particulares del problema permitieron definir actividades dentro de cada fase que admite la ejecución de proceso de minería de datos descriptiva.

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

3.1 Introducción

En este capítulo se describe la aplicación de la técnica y el algoritmo seleccionado para la generación del modelo, el cual es interpretado, obteniendo un grupo de patrones que describen el comportamiento del área de Deporte del SIGOB.

3.2 Aplicación de la técnica sobre el MD de Deporte

Se escoge para realizar la investigación el MD de Deporte del SIGOB, ya que el mismo contiene información suficiente para realizar un análisis descriptivo que permitan demostrar la eficiencia y los resultados arrojados por la técnica propuesta. Dentro de este se seleccionan los hechos participantes_ eventos_ competitivos, títulos_ ganados_ Cuba, olimpiadas_ deporte_ cubano por contener las mayor cantidad de datos dentro del MD de Deporte del SIGOB.

Se realiza la descripción de cada una de las fases de la metodología CRISP-DM, aplicada al hecho (participantes_ eventos_ competitivos) para la comprensión de las mismas. Para seleccionar las variables que serán escogidas para generar la vista minable que se utilizará para extraer el conocimiento, se evalúa el nivel de relevancia que tienen dentro del MD. En este caso se escogen las variables, año (Figura 10), cantidad de deportes en competencia (Figura 11), el nivel (Figura 12) y el sexo (Figura 13), debido a que cada una indistintamente nos aporta información que caracteriza y describe a un participante en específico. Estas variables nos brindan similitudes entre varios atributos de un mismo conjunto, para lograr realizar lo antes descrito se plantean las siguientes preguntas:

- ✓ ¿Qué sexo pueden tener los participantes?
- ✓ ¿A qué niveles se realizan los eventos competitivos?
- ✓ ¿Qué deportes están presentes en los eventos competitivos?
- ✓ ¿En qué año son realizados los eventos competitivos?

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

anno_id	anno_codigo character varying(4)	anno_nombre character varying(4)	anno_numero integer
26	1825	1825	1825
27	1826	1826	1826
28	1827	1827	1827
29	1828	1828	1828
30	1829	1829	1829
31	1830	1830	1830
32	1831	1831	1831
33	1832	1832	1832
34	1833	1833	1833
35	1834	1834	1834
36	1835	1835	1835
37	1836	1836	1836
38	1837	1837	1837
39	1838	1838	1838

Figura 10 Años en que se realizan las competencias

	dim_depor [PK] serial	deporte character varying(50)	codigo_deporte integer	dim_deporte_padre integer	descripcion character varying(150)
1	73	Total	1		Total
2	74	Ajedrez	2	73	Ajedrez
3	75	Ajedrez (Olimpiada)	3	73	Ajedrez (Olimpiada)
4	76	Atletismo	4	73	Atletismo
5	77	Bádminton	5	73	Bádminton
6	78	Baloncesto	6	73	Baloncesto
7	79	Balonmano	7	73	Balonmano
8	80	Beatle	8	73	Beatle
9	81	Beisbol	9	73	Beisbol
10	82	Bolos	10	73	Bolos
11	83	Boliche	11	73	Boliche
12	84	Boxeo	12	73	Boxeo
13	85	Canotaje	13	73	Canotaje
14	86	Cancha	14	73	Cancha

Figura 11 Deportes en competencia

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

	dim_nivel_id [PK] serial	nivel character varying(50)	descripcion character varying(150)
1	6	Internacional	Nivel internacional
2	7	Nacional	Nivel nacional
3	8	Provincial	Nivel provincial
4	9	Municipal	Nivel municipal
5	10	Comunitaria	Nivel comunitario
*			

Figura 12 Niveles a los que se realizan las competiciones

	oid	dim_sexo_i [PK] serial	sexo_codigo character(1)	sexo_nombre character varying(15)	sexo_denominacion character varying(15)
1	67149	1	M	Masculino	Hombre
2	67150	2	F	Femenino	Mujer
3	67151	3	ND	No Definido	No Definido
4	67152	4	T	Ambos	Ambos

Figura 13 Sexo de los participantes

Descripción de los datos (Tabla 1) para la obtención de la vista minable de los participantes que asisten a eventos competitivos.

Tabla 1 Datos escogidos para la obtención de la vista minable

Variables	Tabla	Descripción
Sexo	Sexo	Sexo de cada uno de los participantes.
Nivel	Nivel	Nivel al que pertenece el participante.
Anno	Año	Año en que se celebra la olimpiada del deporte cubano.

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

deporte	Deporte	Deportes en competencia
cantidad_escolares	Participante en eventos competitivos	Participantes de primarias.
cantidad_juveniles	Participante en eventos competitivos	Participantes de secundarias y pre-universitario.
cantidad_participantes_mayores	Participantes en eventos competitivos.	Participantes universitarios

Transformaciones realizadas a los datos

Listado de variables

Sexo

Sexo = 1 entonces Masculino

Sexo = 2 entonces Femenino

Sexo = 3 entonces No definido

Sexo = 4 entonces Ambos

Nivel

Para esta variable se crean 5 nuevas columnas para identificar cada uno de los niveles, en caso de que el participante participe en eventos competitivos a uno de estos niveles, en la columna del nivel aparecerá el 1 en caso contrario siempre aparecerá 0.

Descripción de la transformación realizada al hecho “participantes_ eventos_ competitivos” (Figura 14).

- ✓ Se extraen del MD Deporte las variables sexo, deporte, nivel y cantidad de participantes (mayores, juveniles y escolares) del hecho “participantes_ eventos_ competitivos” y las dimensiones “deporte, sexo, año y nivel”.
- ✓ Se realizan 5 mapeos de valores para crear nuevas columnas que contengan cada uno de los niveles a los cuales se realizan los eventos competitivos.

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

- ✓ Se utiliza el componente Seleccionar/Renombra valores para realizar cambios de nombres en los atributos sexo_nombre, año_número, cant_mayores, cant_juveniles y cant_menores.
- ✓ Se realizan un mapeo de valores para lograr homogeneidad en los valores de los atributos de la columna sexo.
- ✓ Se utiliza el componente Seleccionar/Renombra para cambiar el tipo de dato en la variable sexo de nominal a numérico
- ✓ Se obtienen los datos en un fichero de texto con extensión .csv que nos permite utilizar estos datos en la herramienta Weka.

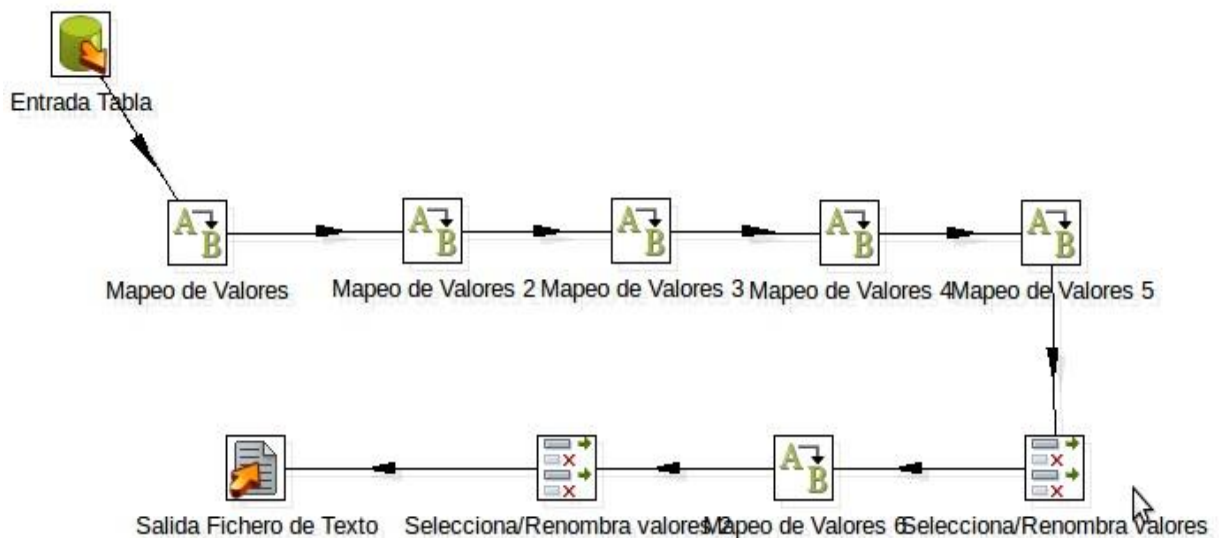


Figura 14 Transformación realizada al hecho participante en eventos competitivos

Después de las transformaciones antes descrita, se obtiene la vista_minable_final (Figura 15), la cual contiene las variables principales para obtener el conocimiento.

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

	A	B	C	D	E	F	G	H	I	J	K
1	mayores	escolares	juveniles	años	deporte	sexo	Internacional	Nacional	Provincial	Municipal	Comunitaria
2		297		1976	Ajedrez	3	0	1	0	0	0
3		1535		1976	Atletismo	3	0	1	0	0	0
4				1976	Bádminton	3	0	1	0	0	0
5		553		1976	Baloncesto	3	0	1	0	0	0
6		360		1976	Balonmano	3	0	1	0	0	0
7				1976	Beate	3	0	1	0	0	0
8		392		1976	Beisbol	3	0	1	0	0	0
9				1976	Boliche	3	0	1	0	0	0
10		568		1976	Boxeo	3	0	1	0	0	0
11		810		1976	Canotaje	3	0	1	0	0	0
12				1976	Caza subma	3	0	1	0	0	0
13		578		1976	Ciclismo	3	0	1	0	0	0
14		33		1976	Clavados	3	0	1	0	0	0
15				1976	Cancha	3	0	1	0	0	0
16				1976	Equitación	3	0	1	0	0	0
17		324		1976	Esgrima	3	0	1	0	0	0

Figura 15 Vista minable final

3.2.1 Selección de la técnica de modelado

La técnica de DM seleccionada es agrupamiento mediante el algoritmo Simple K-means el cual se encuentra implementado en la clase `weka.clusterers.SimpleKMeans.java`. A continuación se muestra en la siguiente tabla las opciones de configuración.

Tabla 2 Opciones de configuración del algoritmo

Opción	Configuración
<code>numClusters(n)</code>	Número de clúster o grupos que se forman
<code>Seed(n)</code>	Semilla a partir de la cual se genera el número aleatorio para inicializar los centros de los clusters

Tipos de datos que admite el algoritmo y las propiedades de la implementación.

- ✓ Admite atributos simbólicos y numéricos.

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

- ✓ Para obtener los centroides iniciales se emplea un número aleatorio obtenido a partir de la semilla empleada. Los k ejemplos correspondientes a los k números enteros siguientes al número aleatorio obtenido serán los que conformen dichos centroides.
- ✓ Se escoge la cantidad de grupos donde el error cuadrático (valor generado automáticamente por el algoritmo) sea menor.

3.2.2 Construcción del modelo

En esta fase se ejecutará la herramienta Weka sobre el conjunto de datos en la fase de modelado para crear el modelo previsto.

Construcción del modelo aplicando el algoritmo Simple K-means

A continuación se presenta el modelo que se generó a partir de la aplicación del algoritmo Simple K-Means sobre los datos almacenados en la vista minable. Para poder ejecutar el algoritmo se le debe especificar el parámetro K que será el número de clúster o grupos que se van a formar y además se debe seleccionar un número n que será denominado semilla, para realizar una distribución aleatoria inicial a partir de la cual el algoritmo comience las iteraciones siguientes. Para la selección de este número se realizaron 10 intentos consecutivos probando distintas semillas y se seleccionó aquella que minimizaba la suma del error cuadrático. En la siguiente tabla se muestran los resultados de los 10 intentos.

Tabla 3 Resultado de K-means para 4 clúster con varias semillas

Semilla	Error cuadrático
1	11870
2	13082
3	11870
4	11870
5	11870
6	12126
7	11410

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

8	11410
9	12123
10	11870

Los resultados obtenidos (Figura 16) por la herramienta Weka tras la ejecución de Simple K-Means con 4 clúster y un valor de la semilla = 7, es el siguiente:

```

=== Run information ===

Scheme:      weka.clusterers.SimpleKMeans -N 4 -A "weka.core.EuclideanDistance -R first-last"
Relation:    part_evento_competitivo
Instances:   9660
Attributes:  11
             mayores
             escolares
             juveniles
             año
             deporte
             sexo
             Internacional
             Nacional
             Provincial
             Municipal
             Comunitaria

Test mode:   evaluate on training data

=== Model and evaluation on training set ===

kMeans
=====
Number of iterations: 7
Within cluster sum of squared errors: 11410.272871489882
Missing values globally replaced with mean/mode

Cluster centroids:
Attribute    Full Data      Cluster#
             (9660)        0          1          2          3
             (3864)        (966)      (966)      (3864)
-----
mayores      1641.1932     1238.3972  5201.3215  2027.6054  1057.3541
escolares    3382.7066     3020.6871  7142.1719  3682.1557  2729.9974
juveniles    1348.734      1285.0598  1891.3399  1360.5486  1273.8029
año          1997.3429     1995.1786  2006       2006       1995.1786
deporte      Pentatlón     Pentatlón  Pentatlón  Pentatlón  Pentatlón
sexo         2             2          2          2          2
Internacional 0             0          0          0          0
Nacional      0.4           0          0          0          1
Provincial    0.4           1          0          0          0
Municipal     0.1           0          0          1          0
Comunitaria   0.1           0          1          0          0

Clustered Instances
0      3864 ( 40%)
1      966 ( 10%)
2      966 ( 10%)
3      3864 ( 40%)

```

Figura 16 Resultados obtenidos

3.2.3 Evaluación de los resultados

La herramienta Weka proporciona 3 modos de prueba para realizar las opciones de evaluación en la técnica de agrupamiento:

- ✓ Use training set.
- ✓ Supplied test set.
- ✓ Percentage Split.

Es utilizado el modo de prueba Use training set, con esta opción Weka entrena el método con todos los datos disponibles y luego lo aplica otra vez sobre los mismos. Pero es válido destacar que los modelos descriptivos en general, son difíciles de evaluar pues inicialmente el modelo va a describir un tipo de comportamiento y además no cuentan con una clase determinada con la que se pueda medir el grado de acierto del mismo. La mejor manera de evaluar este modelo es ver si tiene un comportamiento útil en el área que se vaya a aplicar.

Resultados obtenidos:

- ✓ El sexo con la definición “Femenino” fue el más representativo dentro de los eventos competitivos.
- ✓ Los niveles con mayor cantidad de eventos competitivos realizados son Nacional y Provincial.
- ✓ Las medallas de oro obtenidas por nuestro país en los últimos años en juegos Panamericanos y Centroamericanos y del Caribe las obtienen atletas del sexo femenino.
- ✓ Las participantes femeninas en las olimpiadas del deporte cubano en las últimas ediciones ha crecido, a la vez que disminuye la participación masculina.
- ✓ El atletismo es el deporte que mayor cantidad de medallas aporta a nuestro país.
- ✓ Los participantes escolares cuentan con la mayor asistencia por tradición en eventos competitivos.

Características de los grupos formados:

- ✓ Grupo 0 (40%): El grupo 0 en su mayoría está integrado por participantes del sexo femenino, que asistieron a eventos a nivel Nacional predominando el deporte “pentatlón”. La cantidad de

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

participantes por etapas es: Mayores (1238), Juveniles (1285) y Escolares (3020) que se encuentran mayormente concentrados en el año 1995.

- ✓ Grupo 1 (10%): El grupo 1 en su mayoría está integrado por participantes del sexo femenino, que asistieron a eventos a nivel Comunitario predominando el deporte “pentatlón”. La cantidad de participantes por etapas es: Mayores (5201), Juveniles (1891) y Escolares (7142) que se encuentran mayormente concentrados en el año 2006.
- ✓ Grupo 2 (10%): El grupo en su mayoría está integrado por participantes del sexo femenino, que asistieron a eventos a nivel Municipal predominando el deporte “pentatlón”. La cantidad de participantes por etapas es: Mayores (2027), Juveniles (1360) y Escolares (3682) que se encuentran mayormente concentrados en el año 2006.
- ✓ Grupo 3 (40%): El grupo en su mayoría está integrado por participantes del sexo femenino, que asistieron a eventos a nivel Nacional predominando el deporte “pentatlón”. La cantidad de participantes por etapas es: Mayores (1057), Juveniles (1273) y Escolares (2729) que se encuentran mayormente concentrados en el año 1995.

A partir de las gráficas de dispersión (GD) que se muestra en las Figuras (17, 18 y 19) se conjugaron un conjunto de variables que permitieron la descripción de un grupo de comportamientos en cuanto a la mayor participación por sexo en las etapas (mayores, juveniles y escolares).

CAPÍTULO 3: DESCRIPCIÓN DE LOS RESULTADOS.

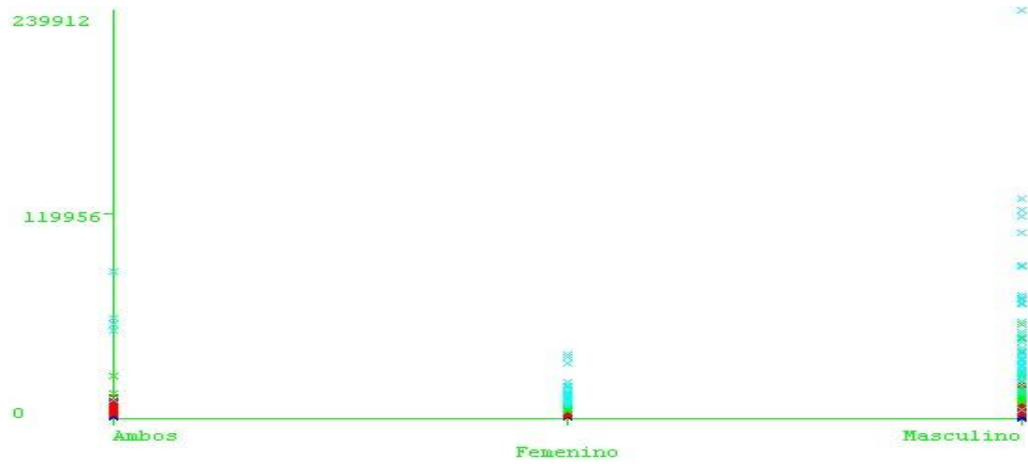


Figura 17 Ejemplo de gráfica de dispersión, sexo-mayores

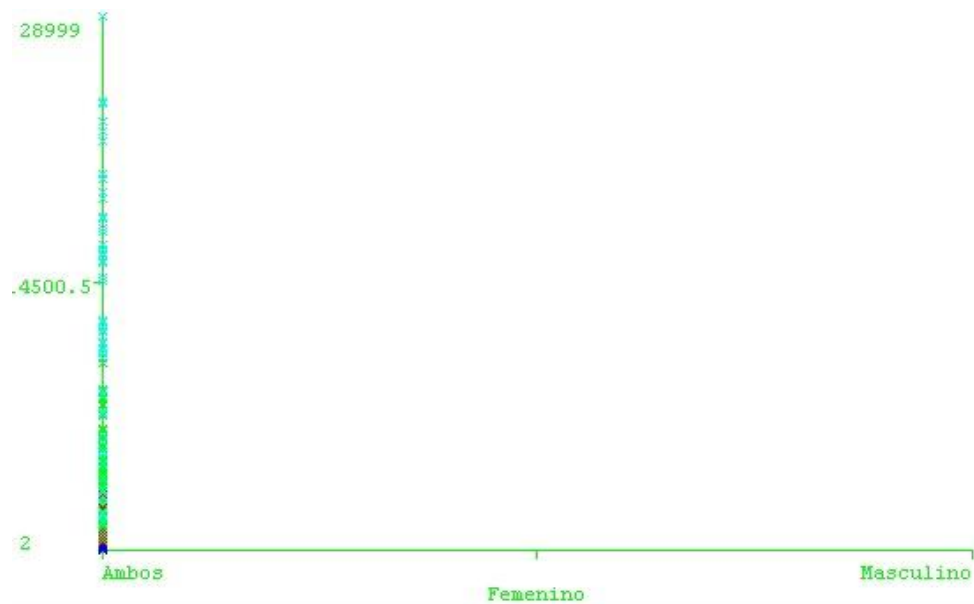


Figura 18 Ejemplo de gráfica de dispersión, sexo-juveniles

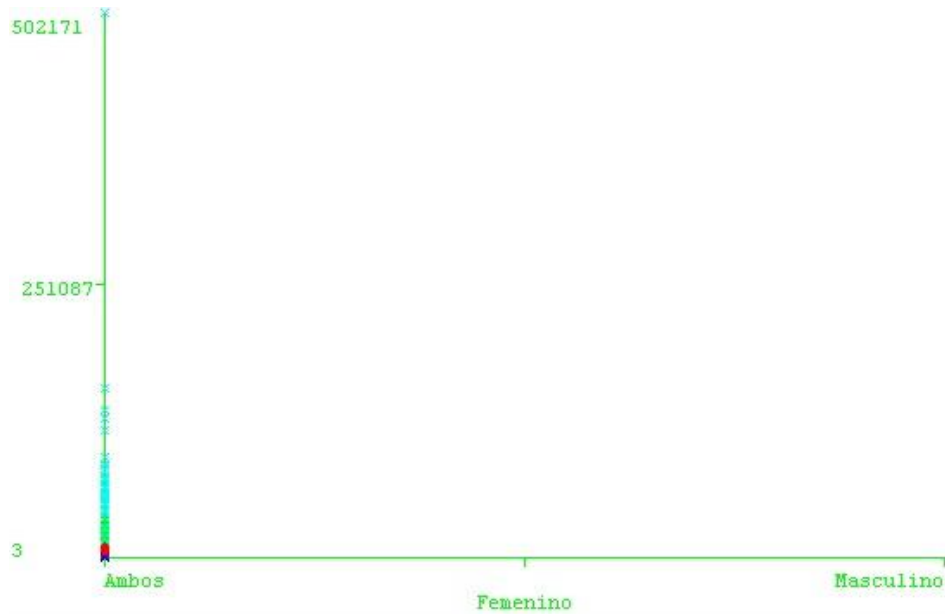


Figura 19 Ejemplo de gráfica de dispersión, sexo-menores

Los pasos de la evaluación anterior trata con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el cual es deficiente.

Para la evaluación de los resultados finales, se realiza una comparación entre los patrones de comportamiento y los datos que se encuentran almacenados en el MD del área Deporte, haciendo uso de gráficos de pastel. A continuación se muestran dos ejemplos en las Figuras (20, 21 y 22), donde se comprueba que el nivel que predomina en las etapas (Mayores, Juveniles y Escolares), son los obtenidos con la interpretación realizada, a partir de las GD. Aclarar quede igual manera se comprobó la veracidad de las demás interpretaciones efectuadas, obteniendo un resultado satisfactorio.

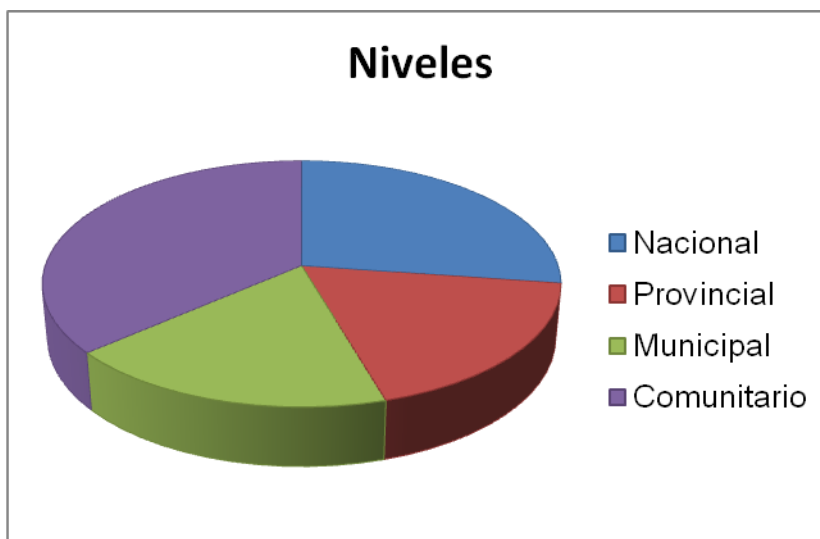


Figura 20 Niveles que predominan en los eventos competitivos para mayores

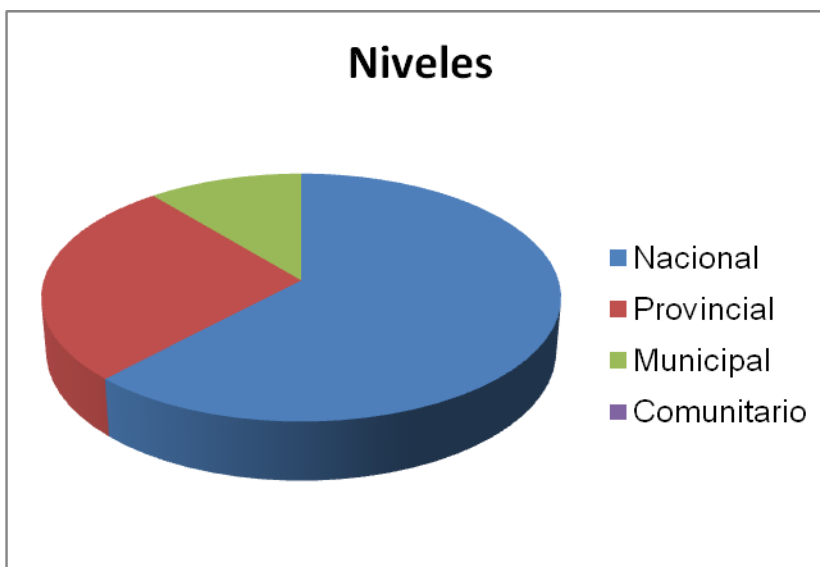


Figura 21 Niveles que predominan en competencias de participantes de la etapa Juveniles

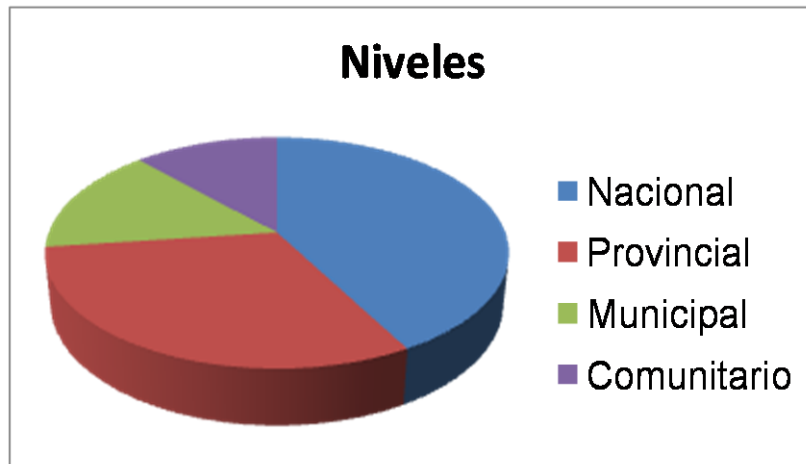


Figura 22 Niveles que predominan en competencias de participantes de la etapa Escolares

3.3 Conclusiones

- ✓ Con la aplicación del algoritmo Simple K-means de la técnica de minería de datos agrupamiento, se obtuvo un conjunto de patrones que a partir de su interpretación se alcanzaron resultados esperados.
- ✓ El algoritmo utilizado permite analizar atributos tanto numéricos como nominales haciendo más fácil la interpretación de los resultados obtenidos en el área de Deporte perteneciente al proyecto SIGOB.

CONCLUSIONES GENERALES

A partir de los resultados obtenidos en la investigación se llegaron a las siguientes conclusiones:

- ✓ A partir de la revisión de los fundamentos teóricos se escoge como técnica a utilizar Clustering apoyada en la metodología CRISP-DM y como herramienta de desarrollo Weka en su versión 3.6.2.
- ✓ El modelo obtenido por la autora con la aplicación de minería de datos descriptiva sobre el mercado de datos de Deporte integra las actividades que intervienen en este proceso y presenta una arquitectura en 3 niveles.
- ✓ Con la aplicación del modelo obtenido a los datos contenidos en el MD de Deporte se demostró la aplicabilidad y validez de la propuesta.

RECOMENDACIONES

Las recomendaciones de la investigación están dirigidas a sugerir acciones para complementar el producto obtenido. A pesar de haber cumplido los objetivos para el buen desempeño y puesta en marcha de la investigación se recomienda:

- ✓ Aplicar técnicas de minería de datos predictivas, sobre el mismo conjunto de datos para la obtención de nuevos patrones de conocimiento.
- ✓ Aplicar la técnica de minería de datos descriptiva, agrupamiento, sobre otros mercados de datos para la obtener patrones de comportamiento.

TRABAJOS CITADOS

1. **PCC.** *Lineamientos la Política Económica y Social del Partido y la Revolución.* La Habana : s.n., 2013.
2. **Vega, D. Morales.** "Integración de modelos de agrupamiento y reglas de asociación obtenidos de múltiples fuentes de datos," *Computación y Sistemas*, vol. 16. 2012.
3. **Kimball, Ralph y Ross, Margy.** *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* New York : John Wiley & Sons, Inc., 2002. ISBN 0-471-20024-7.
4. **Turmero, Iván.** *Minería de Datos (El arte de sacar conocimiento de grandes volúmenes de datos).* Puerto Ordaz : s.n., 2011.
5. **Vallejos, Sofia J.** *Diseño y Administración de Datos.* Argentina : s.n., 2006.
6. **Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth, y otros.** *Guía paso a paso de Minería de Datos.* 2000.
7. **Corría Ramírez, Isidro Manuel y Shelton Nadal, Ronald.** *Estrategia de Trabajo para el Desarrollo del Módulo de Minería de Datos de un Call Center, Aplicando la metodología CRISP-DM.* 2004.
8. **Orallo, José Hernández, José Ramírez, y Cesar Ferri.** *Introducción a la Minería de Datos.* Madrid : s.n., 2004.
9. **Garre M., Cuadrado J.J., Sicilia, M.A., Charro M, Rodríguez D.** *Segmented Parametric Software Estimation Models: Using the EM algorithm with the ISBSG 8 database, Information Technology Interfaces.* Croacia : s.n., 2005.
10. **León Rodríguez, Kirenia Helen y Davila Hernández, Frank.** *Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial.* 2011.
11. **Sarasa.** *Minería de Datos aplicada a la Gestión Docente del Instituto Superior Politécnico José Antonio Hecheverría.* 2008.

12. **Berzal, Juan Carlos Cubero & Fernando.** *Herramientas de Minería de Datos ,Introducción a KNIME.* Granada : s.n., 2005.
13. Data mining with SAS Enterprise Miner. [En línea] 2011. <http://www.sas.com>.
14. SAS Institute Inc. SAS. [En línea] 2011. <http://www.sas.com>.
15. **Fernández, E.** *Asistente para la Gestión de Documentos de Proyectos de Explotación de Datos.* 2006.
16. **López, Cesar Pérez y González, Daniel Santin.** *Minería de Datos, Técnicas y herramientas.* 2008.
17. **González, Erika Vilches y Broitman, Iván A. Escobar.** *Minería de Datos.* 2007.
18. **Gómez, José Ignacio González.** *Generalidades de la Minería de Datos.* 2007.
19. **Cadena, Lisa Leonor Pinzón.** *Aplicando minería de datos al marketing educativo.* 2011.
20. **José Hernández Orallo, Ramírez Quintana, M^a José, Ferri Ramírez, César.** *Introducción a la Minería de Datos.* Madrid : Pearson (Prentice Hall, 2005. ISBN: 8420540919.
21. **Pérez López, César, Santín González, Daniel.** *Minería de datos. Técnicas y Herramientas.* Madrid : Thomson, 2007. ISBN: 8497324922.

BIBLIOGRAFÍA

1. **PCC.** *Lineamientos la Política Económica y Social del Partido y la Revolución.* La Habana : s.n., 2013.
2. **Vega, D. Morales.** "Integración de modelos de agrupamiento y reglas de asociación obtenidos de múltiples fuentes de datos," *Computación y Sistemas*, vol. 16. 2012.
3. **Kimball, Ralph y Ross, Margy.** *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* New York : John Wiley & Sons, Inc., 2002. ISBN 0-471-20024-7.
4. **Turmero, Iván.** *Minería de Datos (El arte de sacar conocimiento de grandes volúmenes de datos).* Puerto Ordaz : s.n., 2011.
5. **Vallejos, Sofia J.** *Diseño y Administración de Datos.* Argentina : s.n., 2006.
6. **Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth, y otros.** *Guía paso a paso de Minería de Datos.* 2000.
7. **Corría Ramírez, Isidro Manuel y Shelton Nadal, Ronald.** *Estrategia de Trabajo para el Desarrollo del Módulo de Minería de Datos de un Call Center, Aplicando la metodología CRISP-DM.* 2004.
8. **Orallo, José Hernández, José Ramírez, y Cesar Ferri.** *Introducción a la Minería de Datos.* Madrid : s.n., 2004.
9. **Garre M., Cuadrado J.J., Sicilia, M.A., Charro M, Rodríguez D.** *Segmented Parametric Software Estimation Models: Using the EM algorithm with the ISBSG 8 database, Information Technology Interfaces.* Croacia : s.n., 2005.
10. **León Rodríguez, Kirenia Helen y Davila Hernández, Frank.** *Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial.* 2011.
11. **Sarasa.** *Minería de Datos aplicada a la Gestión Docente del Instituto Superior Politécnico José Antonio Hecheverría.* 2008.

12. **Berzal, Juan Carlos Cubero & Fernando.** *Herramientas de Minería de Datos ,Introducción a KNIME.* Granada : s.n., 2005.
13. Data mining with SAS Enterprise Miner. [En línea] 2011. <http://www.sas.com>.
14. SAS Institute Inc. SAS. [En línea] 2011. <http://www.sas.com>.
15. **Fernández, E.** *Asistente para la Gestión de Documentos de Proyectos de Explotación de Datos.* 2006.
16. **López, Cesar Pérez y González, Daniel Santin.** *Minería de Datos, Técnicas y herramientas.* 2008.
17. **González, Erika Vilches y Broitman, Iván A. Escobar.** *Minería de Datos.* 2007.
18. **Gómez, José Ignacio González.** *Generalidades de la Minería de Datos.* 2007.
19. **Cadena, Lisa Leonor Pinzón.** *Aplicando minería de datos al marketing educativo.* 2011.
20. **José Hernández Orallo, Ramírez Quintana, M^a José, Ferri Ramírez, César.** *Introducción a la Minería de Datos.* Madrid : Pearson (Prentice Hall, 2005. ISBN: 8420540919.
21. **Pérez López, César, Santín González, Daniel.** *Minería de datos. Técnicas y Herramientas.* Madrid : Thomson, 2007. ISBN: 8497324922.
22. **Varela, Ricardo Herrera.** *Bibliomining: Minería de Datos y descubrimiento de conocimiento en bases de datos aplicadas al ámbito bibliotecario.* Madrid : Universidad Carlos III, 2006.
23. **Moreno, Gimena.** Explotación de Datos del Web Mining. [En línea] 2007. <http://gamoreno.wordpress.com/2007/08/24/explotacion-de-datos-del-web-mining/>.
24. **Machinery, Association for Computing.** [En línea] 2008. <http://www.acm.org/> ..
25. SQL Server Data Mining Team. [En línea] 2009. <http://www.sqlserverdatamining.com/ssdm/>.
26. **Tan, P.-N., Steinbach, M., & Kuman, V.** *Introduction to Data Mining. Boston.* Boston : MA: Pearson Education, Inc, 2006.

27. **Tang, Z., & MacLennan, J.** *Data Mining with SQL Server 2005*. Indianapolis : IN: Wiley Publishing, Inc., 2005.
28. **Witten, I. H., & Frank, E.** *Data Mining: Practical Machine Learning Tools and Techniques (Second ed.)*. . San Francisco : CA: Elsevier, Inc, 2005.
29. **Michalewicz, CitarZ.** *Adaptive Business Intelligence*. New York: Springer. 2007.
30. **Durán, Elena y Costaguta, Rosanna.** *Minería de datos para descubrir estilos de aprendizaje*. 2008.
31. **Morate, Diego García.** *Manual de Weka*. 2007.

ANEXOS

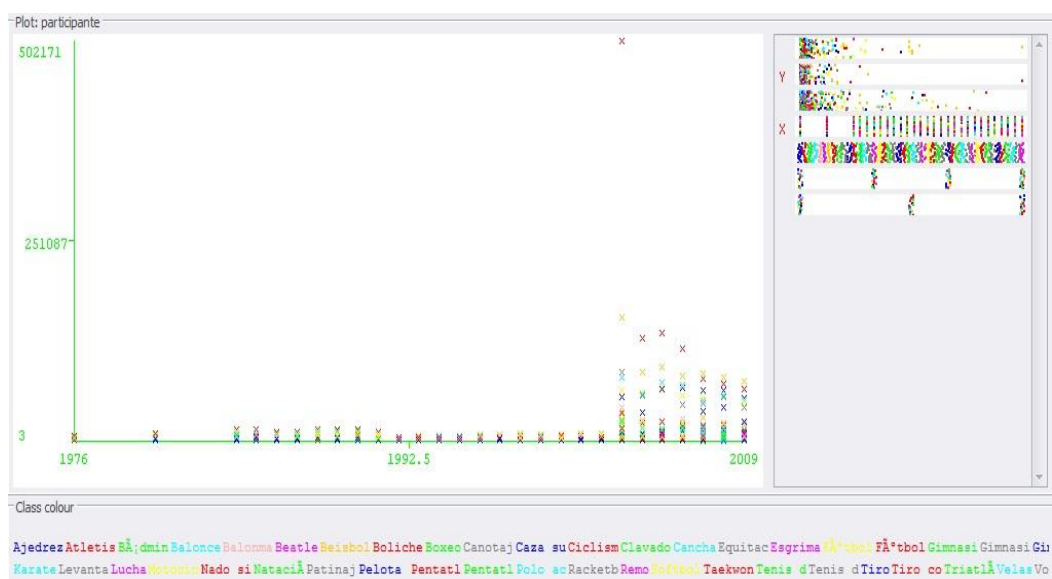


Figura 23 Comportamiento de los participantes escolares en cada deporte por año

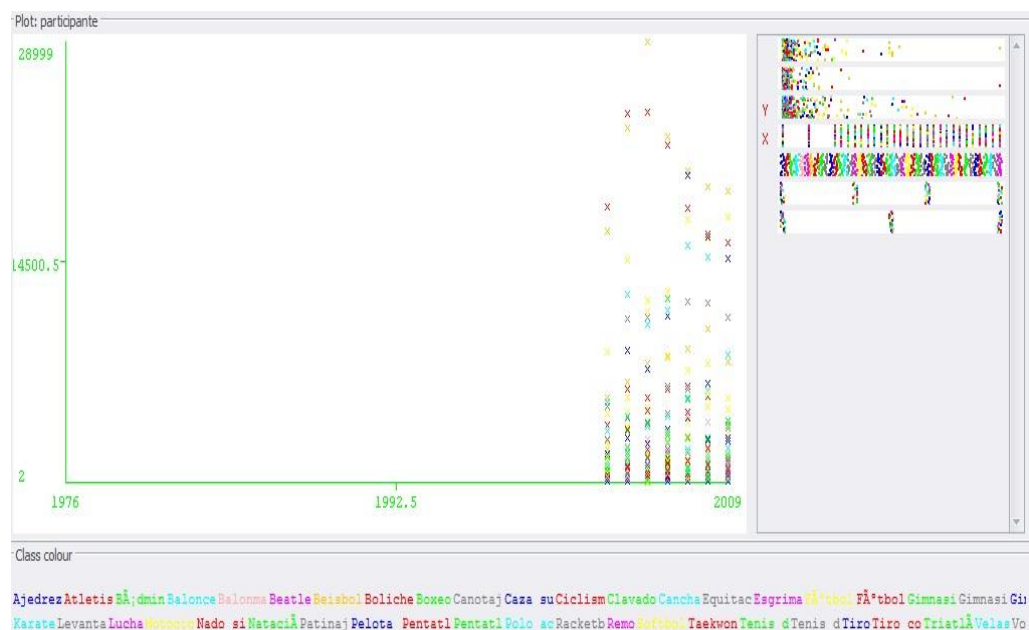


Figura 24 Comportamiento de los participantes juveniles en cada deporte por año

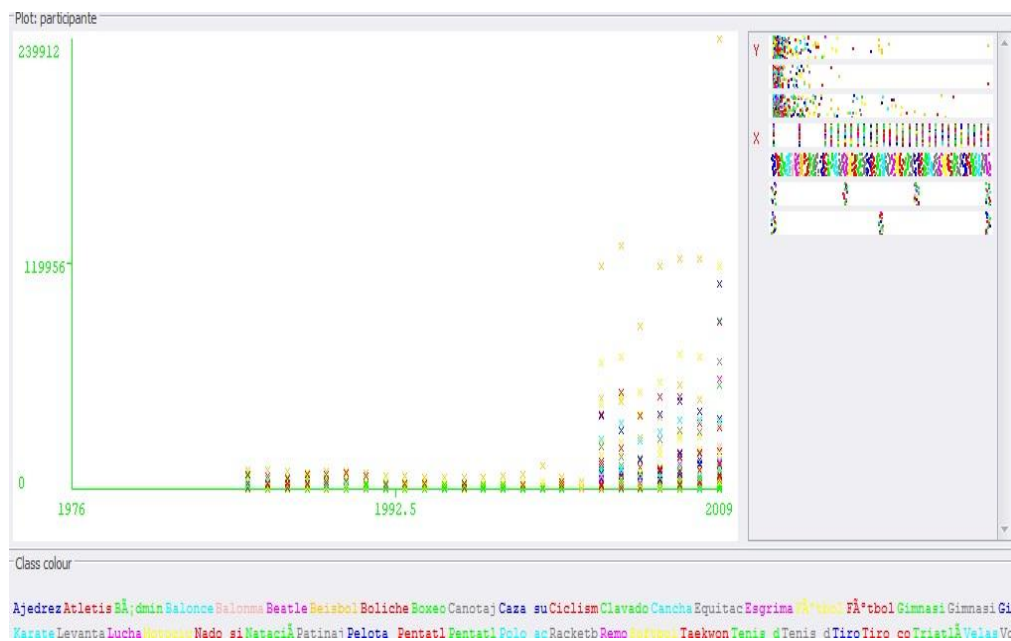


Figura 25 Comportamiento de los participantes mayores en cada deporte por año

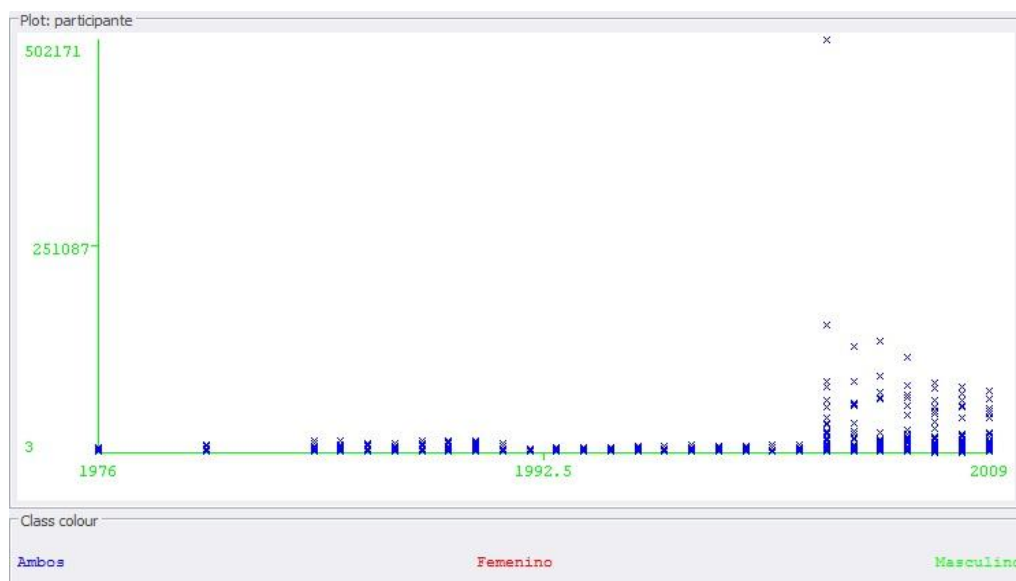


Figura 26 Comportamiento de los participantes escolares en cada año por sexo



Figura 27 Comportamiento de los participantes juveniles en cada año por sexo

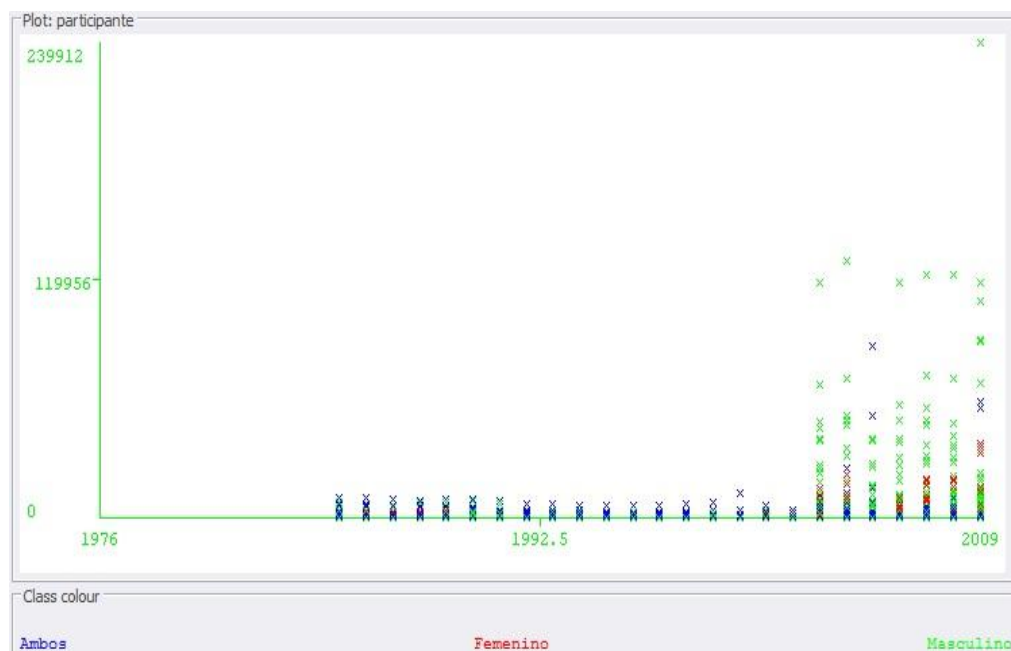


Figura 28 Comportamiento de los participantes mayores en cada año por sexo

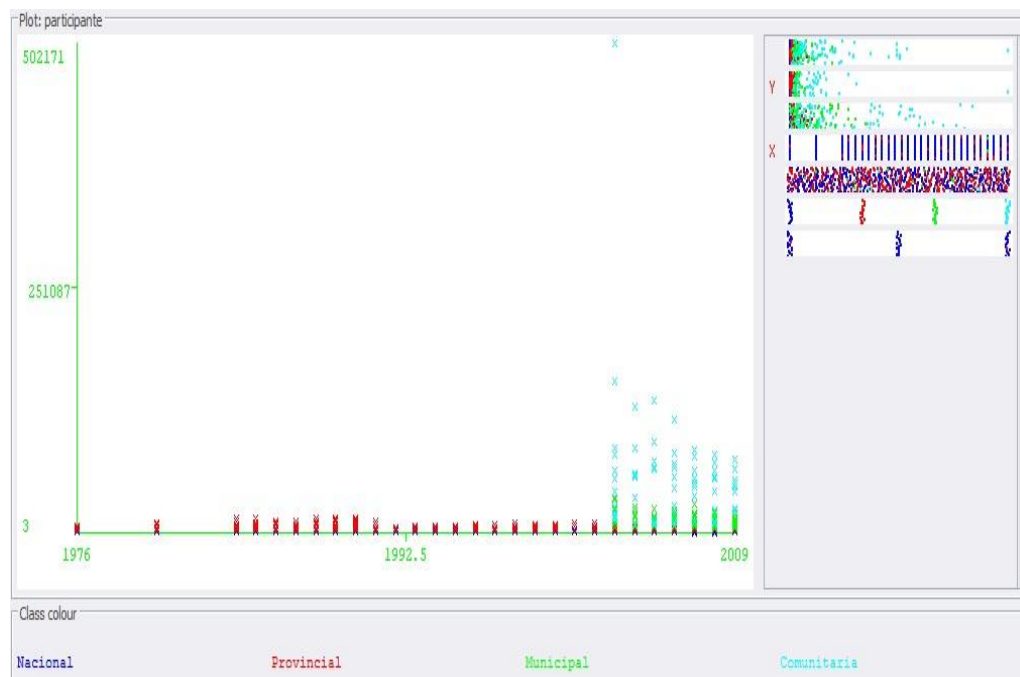


Figura 29 Comportamiento de los participantes escolares en cada año por nivel

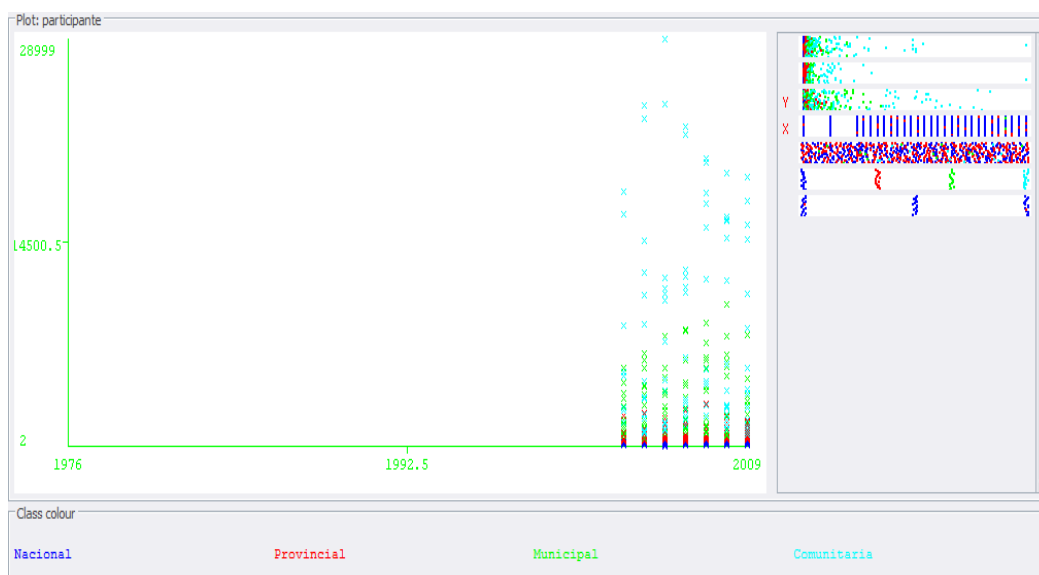


Figura 30 Comportamiento de los participantes juveniles en cada año por nivel

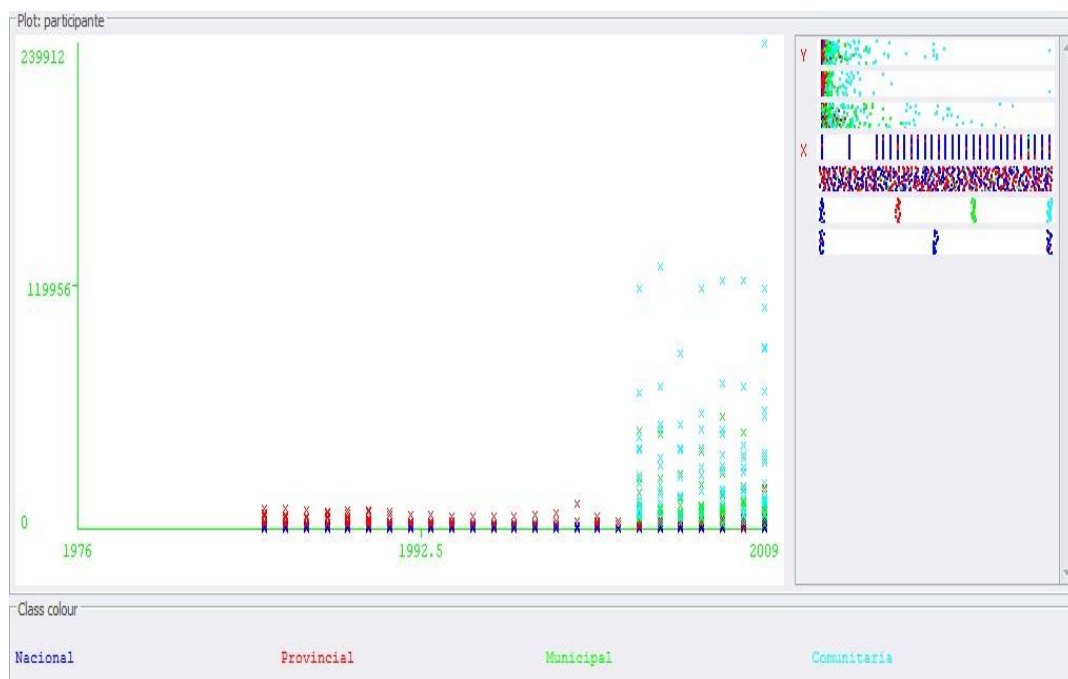


Figura 31 Comportamiento de los participantes mayores en cada año por nivel