Patrones de comportamiento en el voleibol de alto rendimiento

El Levantador, la mente del juego

Trabajo Final de Grado
Ingeniería de Sistemas
Instituto Universitario Aeronáutico

Año: 2013

Silvana Inés Leguizamón

Asesores:

Ing. Daniela E. Díaz
Prof. Carlos D. Cardona
Dr. Manuel Acevedo

Declaración de derechos de autor

Dedicatoria

A mi padre Andrés.

Agradecimientos

A Dios por darme las fuerzas.

A mi familia por la paciencia.

A mis amigos por la comprensión.

A mis compañeros del IUA por el apoyo.

A los profesores por la ayuda.

Resumen

El voleibol como deporte profesional y la minería de datos como herramienta para extraer conocimiento encuentran su punto de fusión en la búsqueda de patrones de comportamiento que permitan definir estrategias de juego. La preparación de un partido de voleibol es una de las tareas fundamentales del entrenador; el "levantador" del equipo es quien ejecuta ese diseño táctico. Este jugador, "la mente del juego", es el encargado de organizar las acciones de ataque de su equipo. Establecer el pensamiento táctico del levantador oponente y anticipar sus elecciones permitiría a los entrenadores definir una estrategia correcta para contrarrestar su juego, aumentando la posibilidad de victoria del propio equipo en un partido y un campeonato.

Por medio de esta investigación se pretende ayudar al entrenador de un equipo de nivel profesional a descubrir patrones de comportamiento observando características o atributos que influyen en las decisiones del levantador del equipo adversario.

La recolección de datos estadísticos durante el transcurso del juego ha llegado hoy en día a un nivel óptimo: es posible leer en símbolos lo que sucede paso a paso, acción tras acción en cada uno de los encuentros disputados. Es posible además corregir esta lectura a través de los videos, agrupar datos, ordenarlos, manipularlos. Aun así lo que no es viable hoy en día es anticipar jugadas adversarias identificando tendencias que relacionen varios atributos a la vez.

El presente estudio, se ocupa de explorar los datos disponibles, construir modelos, validarlos, estudiarlos y exponer los resultados a disposición de quien esté encargado del análisis estratégico del juego, es decir el entrenador del equipo y su staff.

Con el uso del ordenador se ejecutan algoritmos complejos, procesando un elevado volumen de datos que la mente humana por sí sola no podría manipular. Los mismos son confrontados entre sí, evaluados y ponderados según su eficiencia.

Como resultado de este proceso, se logran identificar patrones de comportamiento que anticipan la manera de actuar del levantador desde diferentes perspectivas.

La definición de estrategias de juego es fundamentada con los resultados obtenidos de los modelos generados; éstos dan sustento a las elecciones escogidas.

En este trabajo se demuestra que en el deporte, específicamente en el voleibol profesional, las herramientas informáticas pueden ser protagonistas en la toma de decisiones de un entrenador.

Índice de contenidos

Índice de figuras

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

1.1 ANTECEDENTES

Coordinación, destreza física, disciplina, son términos ligados al deporte. Datos, información, memoria, conocimiento, inteligencia, se relacionan al desarrollo de habilidades intelectuales. Muchas veces estas cualidades se interpretan como enfrentadas o contrapuestas. Sin embargo, en la práctica del voleibol profesional es crucial el desarrollo de ambas. En este deporte, donde no hay contacto físico y existen roles y posiciones definidas, el juego se desarrolla según las decisiones del levantador o armador del equipo. Él es el que piensa y decide quién golpeará la pelota, por dónde será más fácil penetrar en campo contrario y la manera de hacerlo. No se encarga de ejecutar todas las acciones pero sí de definirlas o inducirlas. La habilidad táctica de un equipo consiste en comprender al adversario, identificar las fortalezas y debilidades y saber actuar de manera favorable.

Para poder elaborar y ejecutar una estrategia se necesitan la inteligencia, el conocimiento y la capacidad de relacionar y anticipar. La definición de un planteo táctico inicia en la recolección de datos, el análisis de los mismos y la extracción de información pero encuentra su punto máximo en la emulación del conocimiento. Allí donde no basta el talento físico sino que es necesaria también la destreza intelectual, la informática puede convertirse hoy en día en una herramienta trascendental. La capacidad de almacenamiento, la velocidad de procesamiento, abren la puerta a un abanico de posibilidades que la mente humana por sí sola no podría manejar. La minería de datos definida como proceso de análisis que permite la identificación y extracción de conocimiento a partir de los datos puede servir de apoyo en la definición del comportamiento táctico de un equipo de voleibol.

“El voleibol es considerado como uno de los deportes más populares. Actualmente se practica en todas partes del mundo y su difusión se extiende cada día más. Es un juego de carácter multiforme debido al rápido cambio de situaciones que se presentan en su práctica. Estos cambios exigen una adecuada preparación física que como consecuencia significa salud y asegura además el desarrollo de las cualidades psíquicas y morales”. (1)

“El motivo de la superioridad del juego de voleibol sobre otros juegos análogos debe buscarse, además de en su difusión masiva, en la misma concepción del juego que en Rusia

y Checoslovaquia (actualmente República Checa y Eslovaquia) ha adquirido una fisonomía y un contenido de deporte completo y no de simple pasatiempo o de preparación útil para otras actividades deportivas.

Pero, en su calidad de deporte, entendido en todas sus posibilidades técnicas y estratégicas, revela un sorprendente contenido competitivo y espectacular al conjugarse las dotes individuales y colectivas". (2)

Son las bondades enunciadas, junto a la experiencia totalmente personal de haber dedicado gran parte de mi vida al voleibol las que despertaron mi inquietud por querer fusionar en este trabajo los años de estudio académico con aquellos no menos importantes como deportista profesional.

Desde el punto de vista informático esta unión puede llevarse a cabo utilizando las técnicas de minería de datos, capaces de brindar herramientas que permitan descubrir conocimiento, patrones e identificar tendencias. Estas particularidades serán claves en la definición de un planteo táctico. Desde el punto de vista deportivo se tiene acceso a datos e información de equipos de nivel profesional e internacional. Todo ello ha sido recolectado a través de los distintos torneos internacionales en los que ha participado y participa la selección nacional masculina de voleibol de Puerto Rico, cuyo staff ha contribuido con la materia prima necesaria en la concreción de este proyecto.

1.2 HISTORIA Y CONCEPTOS CLAVES DEL VOLEIBOL

1.2.1 Historia universal

El voleibol nace en 1895, iniciado por William Morgan (director físico de la Asociación Cristiana de Jóvenes en Massachusetts). Originalmente fue llamado mintonette y se trataba de una actividad física recreativa para hombres con sobrepeso.

En 1896 el profesor H.T. Halsted de la Universidad de Springfield, sugiere que se le cambie el nombre a voleibol.

En 1947 en París, con representantes de 14 países, se funda la Federación Internacional de Voleibol. Paul Libaud de Francia se convierte en su primer presidente.

A finales de la década de 1940, tienen lugar en Praga los primeros campeonatos mundiales masculinos y los primeros campeonatos europeos femeninos, donde se adopta el reglamento internacional de voleibol.

En 1957 se decide incluir el voleibol dentro del programa de los Juegos Olímpicos a partir de Tokio 1964.

Después de los juegos olímpicos de Seúl, Corea, en 1988, se determina la anotación con el sistema rolling point. Es decir, cualquier acción ganada genera un punto.

1.2.2 Características

El voleibol, es un deporte donde dos equipos se enfrentan sobre un terreno de juego liso separados por una red central, tratando de pasar el balón por encima de la red hacia el suelo del campo contrario. El balón puede ser tocado o impulsado con golpes limpios, pero no puede ser parado, sujetado, retenido o acompañado. Cada equipo dispone de un número limitado de toques para devolver el balón hacia el campo contrario. Habitualmente el balón se golpea con manos y brazos, pero también con cualquier otra parte del cuerpo. Una de las características más peculiares del voleibol es que los 6 jugadores tienen que ir rotando sus posiciones a medida que van consiguiendo puntos, cada vez que ganan el saque.

1.2.3 Los partidos

Los partidos de voleibol son dirigidos por un árbitro principal, un árbitro asistente y cuatro jueces de línea.

Los partidos se disputan al mejor de cinco parciales (sets). En el momento en que uno de los dos equipos acumula tres sets ganados, gana el partido y se da por concluido el enfrentamiento. Un equipo gana un set cuando alcanza o supera los 25 puntos con una ventaja de dos o más puntos (ejemplo: con 25-23 se gana, pero con 25-24 habría que esperar al 26-24 y así sucesivamente hasta que uno de los dos equipos consiga los dos puntos de ventaja). De ser necesario el quinto set de desempate, se baja la meta a 15 puntos y también se necesita una ventaja de dos o más puntos para ganar. Los campos se sortean antes del partido, así como el saque inicial. Luego de cada set los equipos cambian de campo y se va alternando el primer saque (en el segundo set, saca el equipo que comenzó recibiendo en el primer parcial). Si debe jugarse un quinto set, set decisivo, se procede a un nuevo sorteo y además se realiza un cambio de campo cuando uno de los equipos anota el punto 8.

Estratégicamente cada set puede dividirse en tres momentos diferentes:

- Del inicio al punto 8 (momento A1)
- Del punto 9 al punto 16 (momento A2)

- Del punto 17 al final (momento A3)

En el quinto set, los momentos son:

- Del inicio al punto 8 (momento B1)
- Del punto 9 al final (momento B2)

1.2.4 Los equipos

Cada equipo juega con seis jugadores que pueden ser sustituidos con condiciones.

Tres de los jugadores forman la línea delantera, en tareas de ataque y los otros tres se colocan detrás y actúan de defensores o zagueros.

El equipo completo puede estar formado por un máximo de 14 jugadores (12 más 2 líberos), un entrenador, un entrenador asistente, un masajista, un médico y un estadista.

Cada jugador se identifica por un número distinto, del 1 al 20, número que aparece tanto en la parte delantera como en la trasera de la camiseta. Uno de los jugadores será el capitán del equipo y se identifica por una banda visible debajo de su número. Un líbero es un jugador defensivo que puede entrar y salir continuamente del campo sustituyendo a cualquiera de los otros jugadores cuando por rotación se encuentran en posición defensiva. No puede sacar y tampoco puede realizar alguna acción de ataque o bloqueo. Los líberos no pueden ser capitán y son los únicos que pueden y tienen que vestir una indumentaria distinta, generalmente de distintos colores al resto del equipo.

1.2.5 Las rotaciones

Antes de empezar cada set el entrenador entrega al árbitro asistente el "Orden de saque", que es la formación del equipo en la primera rotación del set.

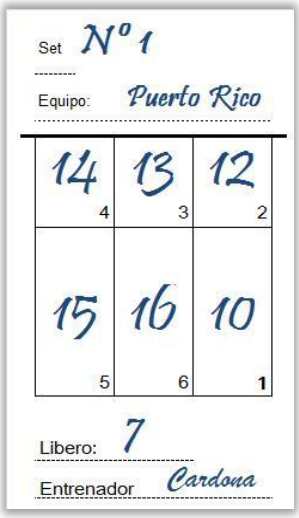


Fig. 1-1: Orden de Saque

En la figura 1-1 se puede ver el "orden de saque" que el equipo de Puerto Rico entregó para el primer set.

Allí se establece la primera rotación con el jugador nº 10 en la posición 1, el jugador nº 12 en la posición 2, el jugador nº 13 en la posición 3 y así sucesivamente.

También se establece al jugador nº 7 como "libero".

La hoja tiene que ser firmada por el entrenador principal del equipo

Muy básicamente el juego se divide en dos estructuras:

- El Complejo 2 o K2, que incluye las acciones de: saque, bloqueo, defensa, levantada, contra ataque y cobertura.
- El Complejo 1 o K1, que incluye las acciones de: recepción, levantada, ataque y cobertura.

El objetivo del K2 es ganar un Break Point (BP) y el objetivo del K1 es ganar un cambio de saque o Side Out (SO).

Cada vez que un equipo anota un punto, será el encargado de poner en juego el balón.

Cuando se logra un punto después del saque se gana un Break Point, vuelve a sacar el mismo jugador y se mantiene la misma fase de juego.

Cuando se arrebata el saque al contrario, se gana un punto (SO) y los seis jugadores tienen que rotar su posición en el campo en el sentido de las agujas del reloj, cambiando la fase de juego. De esta manera todos los jugadores van alternando en las posiciones de delanteros y zagueros.

Las denominadas "fases de juego" son seis y establecen la rotación del equipo según sea la posición del levantador. Ej. La fase 5 es cuando el levantador está en la posición 5, la fase 6 es cuando el levantador está en la posición 6, etc.

A modo de ejemplo vemos en la figura 1-2 la mecánica de una rotación. El equipo que está en recepción, con el levantador (nº10) en Fase 1, gana el cambio de saque (K1), anota un punto "side out" y "rota" enviando a la posición 1 el jugador nº 12 que estaba delantero en la posición 2. Este jugador será el encargado de sacar para comenzar a jugar un K2 en la Fase 6.

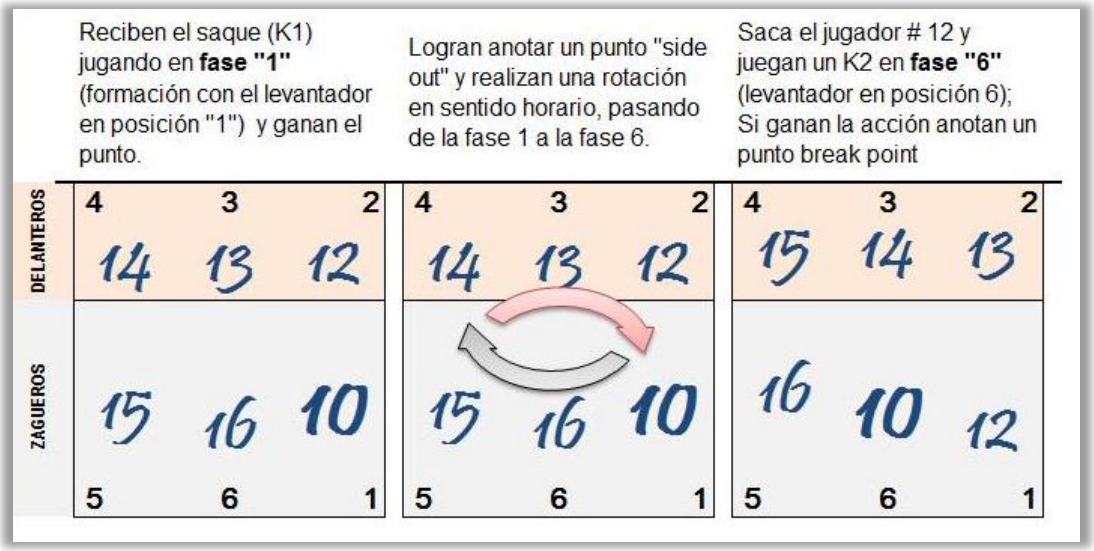


Fig. 1-2: Mecánica de una rotación

Para que la disposición sea correcta y no se cometa una infracción de posición, no es necesaria una determinada geometría, sino simplemente que al iniciar cada punto, con el golpe de saque, cada *delantero* tenga al menos un pie más adelantado respecto a la línea central que su *zaguero* correspondiente y los laterales al menos un pie más cerca de su línea lateral que el jugador en posición central. A partir de ese momento cada jugador puede moverse libremente siguiendo el juego, incluso pueden intercambiar posiciones siempre que se trate de un cambio entre delanteros o zagueros (un jugador zaguero no puede cambiar su posición con un jugador delantero).

Con estas reglas, las posiciones iníciales de los jugadores en cada punto pueden variarse y ajustarse a las diferentes estrategias. El planteo táctico se define según las capacidades técnicas de cada jugador y se implementa para seguir la estrategia establecida.

Como un equipo realiza una rotación cada vez que gana un punto después que el equipo adversario realizó el saque, entonces en cada set se juegan doce mini partidos diferentes que se podrán repetir entre tres y cuatro veces según sea el resultado final del set.

En la figura 1-3, el Equipo “A” comienza el partido con la posesión del saque (K2) en la Fase 1 y el Equipo “B” inicia recibiendo (K1) en la Fase 6.

Cuando el “B” gane un punto (SO), rota una posición (a la Fase 5) y tendrá el saque; mientras que el “A” deberá recibir en la misma formación que perdió su saque (la Fase 1).

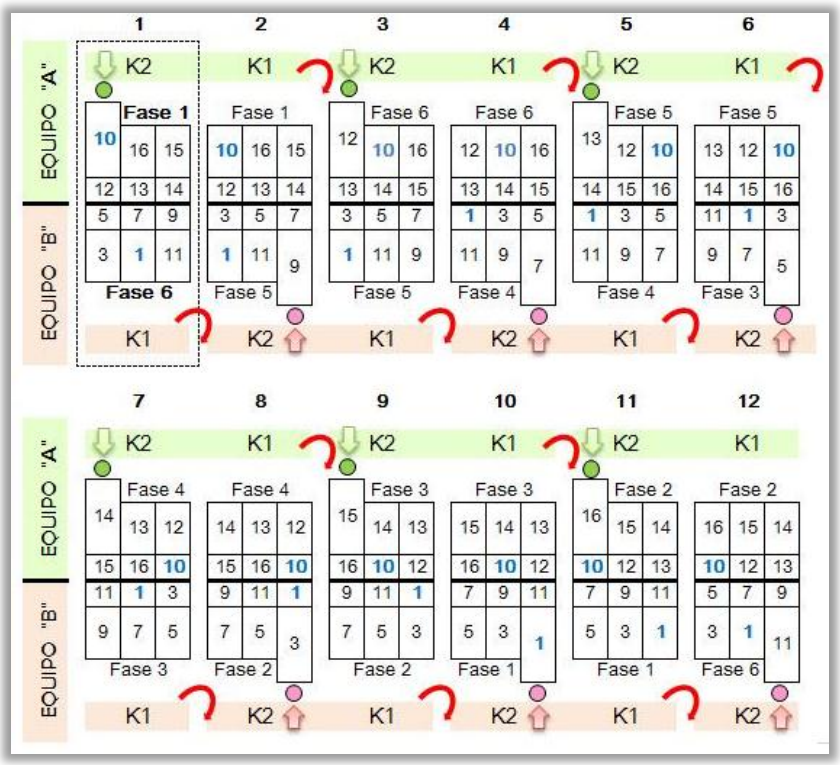


Fig. 1-3 Los 12 mini partidos que se generan en una rotación completa

1.2.6 El juego

Juegan el partido seis jugadores titulares y un líbero.

En el voleibol de alto rendimiento generalmente los equipos utilizan el siguiente módulo:

Módulo 5 - 1 (un levantador y cinco atacantes).

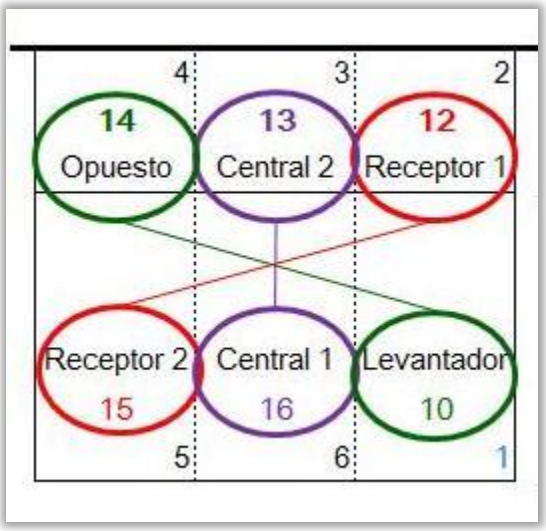


Fig. 1-4: Modulo 5 - 1

La figura 1-4 muestra el Levantador en diagonal con un atacante principal, llamado Opuesto (que ataca también cuando es zaguero desde la zona 1).

Se observa una diagonal con dos atacantes especialistas en recepción del saque (que pueden atacar también cuando son zagueros desde la zona 6).

Completa el esquema una diagonal con dos jugadores muy altos denominados centrales, que son especialistas en bloqueos y ataques rápidos por el centro de la red (zona 3). Estos centrales, luego de realizar su turno de saque, salen para dar lugar al líbero que los reemplazará hasta que las rotaciones del equipo lo lleven nuevamente a la zona de delanteros).

1.2.7 La recepción del saque

Los dos receptores y el líbero son los encargados de recibir el saque y pasar la pelota al levantador, que deberá organizar las acciones de ataque del equipo.

La recepción puede realizarse desde el lado derecho, desde el centro o desde el lado izquierdo del campo y debe llegar a una zona establecida para que allí la encuentre el levantador; si la recepción es perfecta, esta zona es entre las posiciones 2 y 3 a medio metro distante de la red.

1.2.8 El ataque

El receptor delantero y el opuesto atacan por los extremos de la red levantadas altas (denominados ataques de "tercer tiempo"), levantadas "semi rápidas" (ataques de "segundo tiempo") y levantados rápidas llamadas "Super".

Los centrales atacan por el medio de la red diferentes variantes de levantadas rápidas (ataques de "primer tiempo"), estos ataques pueden ser: corta adelante del levantador ("1"), corta atrás del levantador ("A"), tendida rápida entre posiciones 3 y 4 ("3" o "Flecha") y corta adelante del levantador pasada hacia el hombro izquierdo del atacante ("C").

El receptor zaguero ataca una levantada "semi rápida" desde la zona 6, llamada "pippe".

También se realizan algunos esquemas combinando ataques de primer y segundo tiempo en diferentes zonas de la red.

1.3 SITUACIÓN PROBLEMÁTICA

"En términos generales, planificar es prever con suficiente anticipación los hechos, las acciones, es decir establecer de antemano qué es lo que se debe hacer de acuerdo al logro de determinados objetivos. Obviamente la planificación está íntimamente relacionada con la organización del trabajo. Esto se logra teniendo en cuenta el material del cual se dispone

y se debe realizar de tal forma que el proceso mediante el cual se acomete esa tarea se haga de manera sistemática y racional, acorde a:

- Las necesidades del deportista.
- Los recursos disponibles en el momento.
- Los resultados previsibles en el futuro". (3)

Una tarea fundamental (y quizás la más importante) para un entrenador profesional es la preparación o planificación de un partido. La estrategia que seguirá el equipo durante el transcurso del juego es la clave para tener posibilidades de ganar un encuentro.

"Es importantísimo que el entrenador, estudiando los datos del desarrollo físico y las posibilidades de perfeccionamiento del potencial de sus alumnos, esté en condiciones de efectuar un pronóstico de sus resultados deportivos". (4)

El entrenador necesita conocer las características y tendencias de los principales jugadores del equipo propio y adversario lo cual le dará la posibilidad de anticipar situaciones y preparar la defensa para enfrentar al oponente. El análisis del adversario debe hacerse tanto de manera analítica, es decir jugador por jugador, como de manera grupal.

En el voleibol, el levantador o armador es el jugador encargado de organizar las acciones de ataque de su equipo, dirige la ofensiva del mismo. Es el único jugador por el cual pasan todas las jugadas de un partido, en sus manos está la manera en que el juego se desarrolla y en su mente el curso que seguirá el partido. Su rol es fundamental. Poder analizar, interpretar, estudiar, comprender y anticipar lo que este jugador es capaz de hacer puede ser la clave en la victoria o derrota de un partido de alto nivel.

"El papel del levantador como planificador del juego, en el segundo toque, es muy importante, y hace falta diseñar modelos para que a lo largo de las rotaciones ese puesto permanezca bien cubierto". (5)

El entrenador define una estrategia a priori muchas veces modificable durante el transcurso del juego. Para hacer esto utiliza los datos estadísticos recogidos en partidos previos, en entrenamientos propios y en información externa (tiene en cuenta incluso la información sobre las condiciones físicas de los jugadores, posibles lesiones, etc.). El levantador interpreta y ejecuta la estrategia; para ello tiene en cuenta las características suyas propias y del equipo en general.

"El uso de sistemas informáticos se ha extendido al control del juego de voleibol. Se obtienen así datos estadísticos de las distintas acciones de juego. La estadística individual tiene como función poder darle al entrenador una clara visión de cómo se desempeña cada

jugador dentro del campo de juego en cada acción que realiza. Cada entrenador puede realizar el seguimiento de datos que le interese, durante o post partido, del equipo propio o del rival, lo cual puede permitir potenciar el rendimiento de los jugadores". (6)

Las estadísticas existen y están disponibles, son confiables porque pueden ser corregidas a través de los videos con los que cuentan todos los equipos de nivel internacional o profesional. Hace falta saber interpretar las mismas y extraer información relevante.

Lo que se destaca a este punto es la necesidad de enfocar específicamente el estudio de los datos disponibles hacia el comportamiento del levantador, definido como "la mente del juego". Si él será quien precisará el curso del juego, seguramente él es también la clave para interpretar y anticipar el desempeño del adversario. Poder contar con herramientas que estudien específicamente este rol primordial, puede ser el punto de partida para ayudar a un entrenador a predecir estrategias válidas y sacar ventaja al momento de la competencia.

Actualmente se dispone de datos descriptivos sobre porcentajes de error, efectividad, preferencias de movimiento, etc. pero no existe información suficiente relativa a tendencias o evolución a través del tiempo. No se tienen en cuenta las relaciones que se manifiestan entre los datos existentes, la influencia de unos sobre otros. Se cuenta con estadísticas organizadas pero no se dispone de la lectura anticipada de lo que "puede o podría suceder si". Este trabajo intenta ayudar a superar esta dificultad, es decir optimizar el proceso de preparación o planificación de un partido de voleibol. A nivel internacional el entrenamiento físico y técnico es necesario pero no suficiente. Cuanto antes se conozca lo que hará el adversario más rápido se prepararán los mecanismos de defensa y contraataque y más eficiente será nuestra elección.

1.4 PROBLEMA

Queda manifestada la necesidad de estudiar el levantador del equipo e interpretar su manera de entender y guiar un encuentro. Por lo tanto, el problema planteado puede enunciarse sucintamente de la siguiente manera:

- *¿Existe relación entre las acciones de juego que se presentan durante un partido de voleibol y el comportamiento del levantador?*
- *¿Qué factores influyen sobre las elecciones tácticas del levantador?*

1.5 OBJETO DE ESTUDIO Y CAMPO DE ACCIÓN

El objeto de estudio del presente trabajo es el voleibol de alto nivel orientado específicamente hacia el levantador de un equipo profesional o de nivel internacional. Este rol es determinante en las sucesivas secuencias tácticas de juego; por lo tanto, el estudio del mismo podrá aportar y colaborar en la toma de decisiones de un entrenador profesional.
"La minería de datos o Data Mining consiste en la extracción de información que reside de forma implícita en los datos. Se trata de información desconocida pero que puede resultar útil para efectuar algún proceso. Por lo tanto la minería de datos se ocupa de recabar, extraer, sondear, preparar y explorar los datos para sacar toda la información que ocultan". (7)
Dentro de la cantidad de datos recolectados en el transcurso de uno y varios partidos de voleibol, existe seguramente información oculta que podría ser descubierta a través de la utilización de la minería de datos. Identificar variables claves y sus correlaciones, para descubrir patrones de comportamiento llegando a la creación de modelos abstractos será la tarea primordial de este estudio. Aportar nuevo conocimiento útil al entrenador de un equipo de voleibol podrá ayudar al mismo en la definición de estrategias de juego. Esto optimizará el proceso de preparación del partido e incrementará la probabilidad de victoria del propio equipo.
El trabajo se llevará a cabo con la ayuda de diferentes herramientas según la etapa de desarrollo que se esté realizando.

- SPSS: se utilizará como software estadístico principal en el análisis inteligente inicial de los datos provenientes del sistema operacional. Es necesario conocer de qué datos se dispone, verificar la calidad de los mismos, para luego comenzara definir los tipos de modelos que puedan desarrollarse.
- Integration Services: servirá como herramienta principal en la carga, extracción y transformación de los datos. Este software, que forma parte del ambiente de SQL Server 2008 R2 y presenta el entorno de Visual Studio, permitirá generar paquetes que automaticen las tareas de limpieza y transformación de datos. Se deberá lograr consolidar una base de datos lista para ser minada y modelada.
- Analysis Services: será la herramienta primordial del proyecto. Brinda un conjunto de algoritmos de minería de datos. Su entorno, también perteneciente a Visual Studio permite crear, administrar y examinar modelos. Soporta el lenguaje DMX

(Extensiones de Minería de Datos) que sirve además para realizar consultas predictivas y de contenido.

- Reporting Services: será la plataforma de informes, basada en servidor capaz de proporcionar servicios que permitan confeccionar reportes, administrarlos y entregarlos al usuario final. Funciona en el entorno de Visual Studio y está totalmente integrado con las herramientas anteriores y componentes de SQL Server 2008 R2.

1.6 OBJETIVOS

1.6.1 General

Identificar patrones significativos y relevantes del levantador de un equipo de voleibol. Determinar si existe correspondencia entre su comportamiento y las acciones previas a su intervención. Se deberán interpretar los resultados y presentar los mismos como conocimiento para ayudar en la toma de decisiones de un entrenador de un equipo de voleibol de alto rendimiento.

1.6.2 Específicos

- Encontrar relaciones entre los atributos que describen los esquemas de juego en un partido de voleibol.
- Ayudar en la lectura anticipada de la elección del levantador en función de:
    - La evaluación de la recepción.
    - La llamada del atacante central[1]
    - La zona desde donde llega el pase al levantador.
    - Su posición en la cancha.
    - El momento del partido.
- Aportar conocimiento al entrenador de un equipo de voleibol que le facilite el proceso de planificación y preparación de un partido.

[1] Llanada del atacante central es el tipo de juego rápido que anticipa el atacante central

1.7 IDEA A DEFENDER / PROPUESTA A JUSTIFICAR / SOLUCIÓN A COMPROBAR

Idea a defender: Conocer, analizar e interpretar las características y tendencias de juego del levantador de un equipo de voleibol de alto nivel es determinante en la preparación y definición de las estrategias de juego.

Propuesta a Justificar: Existen patrones repetitivos en el rol del levantador de un equipo de voleibol. Estas tendencias pueden ocurrir debido a relaciones entre las acciones de juego, a situaciones del momento del juego y también a rasgos individuales del levantador y del resto de los jugadores del mismo equipo.

Solución a comprobar: La identificación de tendencias de juego del levantador de un equipo de voleibol (ya sea propio u oponente), permitirá definir estrategias tácticas que aumenten la probabilidad de victoria del equipo.

1.8 DELIMITACIÓN DEL PROYECTO

Reconociendo en el levantador de un equipo de voleibol un rol fundamental y determinante, este proyecto se focalizará en el estudio de este jugador. Relacionar el armador o levantador de un equipo con sus jugadores, con sus preferencias, habilidades y particularidades puede definir de manera factible, oportuna y eficaz la forma de preparar un partido de voleibol de alto nivel.

Este estudio no tendrá en cuenta el análisis individual de cada jugador respecto al adversario. No enfrentará ni comparará roles entre contrincantes para tomar ventaja de las deficiencias de un equipo. Tampoco aportará datos on-line durante el transcurso de un juego, se limitará a un estudio a posteriori para poder predecir la manera más apropiada de enfrentar un futuro encuentro.

El producto final de este proyecto serán reportes que muestren consultas sobre el contenido de los modelos evaluados. Los mismos servirán al entrenador para observarlos desde la óptica que estime conveniente y le ayudarán a definir planteos tácticos.

1.9 APORTE TEÓRICO

Desde el punto de vista teórico el presente estudio pretende relacionar la minería de datos con el voleibol de alto rendimiento dando un paso que va más allá de las inferencias estadísticas tradicionales. El estudio puede definirse como una primera etapa en este

nuevo enfoque que podría en el futuro avanzar hacia la implementación de técnicas de la Inteligencia Artificial y Sistemas Expertos.

Se reconoce claramente que en una segunda etapa debería ampliarse el estudio hacia el resto de los roles que se identifican en esta disciplina. Si bien se consideran secundarios en un primer momento son todos interdependientes en el comportamiento del equipo en su totalidad. Para llegar a emular el conocimiento de un experto será necesario tener en cuenta todos los roles que se identifican en un equipo de voleibol. El valor agregado en un deporte de conjunto que pretende destacarse internacionalmente está formado por la fuerza del equipo trabajando de manera colaborativa en el logro de grandes objetivos.

"Si te pones a pensar durante un partido, te pierdes la jugada en la que estás inmerso. Para pensar ya tienes los entrenamientos de la semana. En esos momentos es cuando te tiene que venir la inspiración. Cuando estás en el partido, lo único que puedes hacer es reaccionar a lo que ves" (Albert Lewis).

No es posible automatizar las infinitas situaciones de juego de un partido de voleibol. Si lo pretendiéramos nos estaríamos olvidando de otros componentes del deporte como el azar, los rasgos psicológicos de los individuos que lo practican, los componentes externos (por ejemplo características peculiares del ambiente), los errores humanos (por ejemplo arbitrales) y muchos más. Aun así la informática puede brindar un medio para planificar, programar y ayudar en la sistematización de jugadas ofreciendo de esta manera una ventaja competitiva.

1.10 APORTE PRÁCTICO

Desde un punto de vista práctico, servirá de ayuda a quien se desempeñe como entrenador de un equipo de voleibol de alto nivel. Aportará una herramienta que permita identificar patrones y ayude a la definición de estrategias de juego. Es decir colaborará en la toma de decisiones.

Según las conclusiones halladas, se podrán construir reportes que sirvan al usuario final para realizar consultas y estudiar los patrones detectados. El análisis de los resultados hallados ayudará al staff técnico a realizar planteos tácticos y definir estrategias.

Este trabajo no tendrá una aplicación práctica allí donde el voleibol sea simplemente un deporte recreativo y de distensión.

1.11 MÉTODOS Y MEDIOS DE INVESTIGACIÓN

A partir del planteo del problema se puede determinar claramente el tipo de estudio o investigación que se llevará adelante con el fin de lograr el objetivo y la respuesta a la pregunta propuesta.

El método seleccionado deberá servir para predecir el comportamiento del levantador. Para llevar adelante esta tarea se deberá estudiar si existe relación y qué tan importante es entre las variables que definen el juego o las situaciones particulares del partido y la manera en que el levantador actúa en respuesta a ello.

Por lo anteriormente mencionado se desprende la necesidad de realizar una investigación del tipo correlacional, donde se pretenderá visualizar cómo se relacionan o vinculan las variables definidas entre sí o si por el contrario no existe relación alguna que me permita definir o anticipar el comportamiento del levantador. Se intentará predecir el valor aproximado de la variable objetivo (aquella que represente el comportamiento del levantador) a partir del valor de las variables que la modifican.

Este tipo de estudio no solo encuentra relaciones entre las variables sino que además define qué tan importante es la relación o unión encontrada, es decir cuál es el grado de asociación. Atendiendo entonces a la propuesta planteada resultará sin dudas ser el adecuado en la resolución de este proyecto.

Es oportuno mencionar la necesidad del conocimiento del dominio del problema como arma fundamental para llevar adelante este tipo de investigaciones. En este caso particular habrá un aporte relacionado estrechamente a lo vivencial, la experiencia personal en el ambiente deportivo objeto de estudio de este trabajo es también un condimento o valor agregado que puede facilitar el entendimiento y la búsqueda de resultados positivos.

Se utilizarán también métodos empíricos que ayudarán a la recopilación de información, sobre todo a través de entrevistas con expertos.

1.12 MÉTODOS Y MEDIOS DE INGENIERÍA

Entre las metodologías existentes para llevar a cabo tareas de explotación de la información y la concreción del estudio propuesto, se pueden mencionar:

- SEMMA (Sample, Explore, Modify, Model, Assess)
- CRISP-DM (Cross Industry Standard Process for Data Mining)
- P3TQ (Product Place Price Time Quantity)

La elección de la metodología se realizó observando en primer lugar aquella más comúnmente utilizada pero avalando esto por medio de un somero estudio de las características fundamentales de cada una de ellas.

SEMMA hace énfasis en aspectos técnicos más que en el análisis y comprensión del problema que se está abordando. Fue creada especialmente para trabajar con software de minería de datos de la compañía SAS. Durante su desarrollo se organizan nodos para cada una de las fases del proyecto. En esos nodos hay herramientas que llevan a cabo las tareas necesarias para poder avanzar. Se trata de una metodología en cascada pero a su vez es iterativa. En este caso, descartamos la misma por estar vinculada a un proveedor de software específico y por su orientación técnica fundamentalmente.

P3TQ o metodología Catalyst plantea la formulación de dos modelos, el Modelo de Negocio (MII) y el Modelo de Explotación de la Información (MIII).El primero intenta identificar el problema u oportunidad y definir los requerimientos. El segundo proporciona los pasos para ejecución y construcción de modelos de minería de datos a partir de MII. Esta metodología, no es pertinente en este tipo de trabajo dado que no se está analizando un problema u oportunidad en un negocio. El análisis de la cadena de valor organizacional definido por las relaciones precio/lugar/producto/tiempo/cantidad no es apropiado en este contexto.

CRISP-DM fue creada por el grupo de empresas SPSS, NCR, y DaimlerChrysler es actualmente la más utilizada como guía en el desarrollo de proyectos de minería de datos. El proceso se estructura en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. Se trata de un proceso en cascada, no se definen explícitamente las iteraciones pero suelen realizarse al igual que las interacciones. En general se ataca el problema completo (o a lo sumo se lo divide en sub problemas) y se convierte en un ciclo de vida incremental. De esta forma se van construyendo modelos de minería de datos para cada uno de los sub problemas identificados. Cada fase es descompuesta en tareas y actividades, la metodología las define pero no especifica cómo llevarlas a cabo.

Debido a que se trata de una metodología muy usada, general y fundamentalmente con posibilidades de adaptarse a cualquier tipo de proyecto, es la elegida para realizar este trabajo. No son las características estándares de los proyectos de minería de datos las que definen este estudio, por lo tanto la versatilidad de la metodología la convierten prácticamente en la única admitida.

2 PRIMERA PARTE MARCO CONTEXTUAL

2.1 ENTORNO DEL OBJETO DE ESTUDIO

En la actualidad el deporte profesional se interpreta como consumo y espectáculo de masas. Por lo tanto si se trata de consumo, se habla también de beneficio económico. El deporte de alta competencia es uno de los mayores espectáculos de nuestro tiempo, es uno de los medios más asequibles y económicos de entretenimiento de la sociedad. La crisis económica mundial no impide que el mundo del deporte profesional mantenga su trayectoria. Grandes empresas, en su apartado de publicidad tienen como objetivo principal patrocinar equipos deportivos profesionales. Al haber inversión económica, se exigen también resultados.

El voleibol profesional definido previamente como el objeto de estudio de este trabajo es parte del consumo y espectáculo deportivo.

La Federación Internacional de Voleibol regla este deporte atendiendo a las necesidades sociales y económicas.

La manera más sencilla de contar con apoyo económico y mantener el interés por el espectáculo es manteniendo el interés de los consumidores. En el profesionalismo lo único que vale es ganar. No en vano suele decirse que es mejor ganar jugando mal que perder jugando bien.

En este ambiente totalmente competitivo la informática colabora aportando velocidad de procesamiento y capacidad de almacenamiento brindando de esta manera un ambiente capaz de manipular datos, ordenarlos y presentarlos como información y conocimiento.

2.2 MI RELACIÓN CON EL VOLEIBOL

Mi interés por el presente trabajo nace de mis años de experiencia en el deporte, específicamente en voleibol profesional, junto a mi situación familiar, que involucra el deporte como sostén y modo de vida. La manera más coherente de aplicar los conocimientos adquiridos durante los años de estudio es volcarlos en aquel ambiente que conozco y vivo desde adentro.

2.3 ANÁLISIS DE LOS PROBLEMAS OBSERVADOS

La idea no surge como consecuencia de un análisis de problemas existentes, sino como un desafío de querer encontrar aquello que se siente "latir" en las estadísticas descriptivas de un partido. La cantidad de encuentros disputados y sobre todo la cantidad de aquellos observados, me llevaron a adquirir una capacidad de análisis y comprensión que trasciende al simple observador que disfruta de la emoción de cada punto.

2.4 ANTECEDENTES DE PROYECTOS SIMILARES

No se conoce la existencia de estudios que utilicen minería de datos en el voleibol profesional. Pueden nombrarse a modo de ejemplos aplicaciones de estas técnicas en el mundo del deporte profesional, específicamente en el fútbol y en el basquetbol.

- A.C.Milan (Associazione Calcio Milan)

  Este club, reconocido como uno de los equipos de fútbol más importantes del mundo utiliza redes neuronales para prevenir lesiones y optimizar el acondicionamiento físico de los atletas. Cuenta con un sistema de monitoreo que recoge datos, los procesa y permite predecir la probabilidad de lesión de un jugador. Relaciona además esta situación con el estilo de juego, hábitos y costumbres. Este análisis se convierte en un beneficio deportivo, evitando sorpresas físicas en torneos importantes y un beneficio económico al momento de negociar un contrato.
- NBA(National Basketball Association)

  Es sin dudas la asociación de basquetbol más importante del mundo. Los equipos pertenecientes a esta liga disponen de un software de minería de datos, "Advanced Scout", elaborado por IBM en los años 90. El mismo permite a los entrenadores analizar el juego de los equipos y detectar patrones estadísticos o eventos que no se observan cuando miran un juego en vivo o en un video. La descripción de los encuentros jugada por jugada permite revelar patrones ocultos de interés en la definición de un planteo táctico.

3 SEGUNDA PARTE MARCO TEÓRICO

3.1 MARCO TEÓRICO DEL OBJETO DE ESTUDIO

El concepto de deporte se define como actividad física ejercida como juego o competición, cuya práctica supone entrenamiento y sujeción a normas. Es una demostración de destreza física y mental.

"La sociedad primitiva, continuadamente beligerante, obligaba al individuo a ser un combatiente tenaz y fuerte por lo cual la actividad física era inexcusable. Es en esta actividad donde se encuentran los primeros indicios de manifestaciones deportivas". (8)

Carl Diem (1882-1962), administrador e historiador deportivo alemán, creador de la tradición de la antorcha olímpica, encuentra su origen en lo cultual, en los rituales ejercidos por el hombre primitivo hacia los dioses y coincide con Huizinga (filósofo e historiador holandés 1872-1945) quien en su libro Homo Ludens de 1938 relaciona el deporte con el juego.

Pierre de Coubertin, pedagogo e historiador francés, fundador de los Juegos Olímpicos modernos lo comprende como "progreso, menosprecio del riesgo, perseverancia, un esfuerzo muscular que puede llegar hasta el heroísmo. No como juego espontáneo sino como esfuerzo de superación". (9)

Se puede afirmar que el deporte y la actividad física han formado parte del hombre desde el inicio mismo de la civilización. La actividad física era necesaria para formar un cuerpo robusto competente para combatir, cazar, para mejorar la manera de vivir y convivir con el ambiente. Hoy además, se destaca en el deporte la cualidad de bienestar ligada a la salud humana, continúa siendo una necesidad del hombre pero relacionada también a la salud física y mental.

"Pensemos en la función biológica y social del deporte demostrada en la utilidad que presta a la salud del individuo y al desarrollo de la fuerza vital de la población (mejorando la higiene, ayudando a incrementar la producción) al reducir los porcentajes de ausencia por enfermedad de los trabajadores, hechos que además redundan en un menor costo de la sanidad nacional y en el incremento del producto bruto interno y en que el Comité de Investigaciones del Consejo Nacional del Deporte y Educación Física de la UNESCO comprobó que el deporte retrasa el envejecimiento". (10)

También la psicología y la filosofía expresan la importancia de la práctica deportiva destacando el beneficio que recibe quien lo ejerce. ”Lo que más estimo de los deportes es la confianza en sí mismo que procuran al hombre que los cultiva” (Henri Bergson filósofo francés 1859-1941). (11)

En España, José María Cagigal (1928-1983), recordado por hacer un aporte importante al humanismo deportivo enuncia que el deporte de los años 70 ya no era el mismo que inventaron los ingleses. “Si a partir de la segunda mitad del siglo XIX hasta la década de los 60 se ha podido hablar de un deporte moderno de principal inspiración británica, caracterizado por la organización de clubes, federaciones, por la reglamentación y codificación, por ciertos valores como el contacto social, la entrega, el afán de superación; en el último cuarto de siglo hemos iniciado un nuevo período del deporte, en el que junto a las citadas características y estructuras del llamado deporte moderno aparecen netamente definidas otras funciones, roles, estructuras, valores tan dispares a los anteriores como el gran espectáculo, política, técnica, ciencia, profesión, exigencia internacional, los cuales nos sitúan ante un deporte mucho más variado, gigantesco, multifuncional. Podríamos hablar de un deporte contemporáneo”. (12)

Cagigal sitúa la frontera entre ambos deportes, moderno y contemporáneo, en las Olimpíadas de Roma de 1960.

Ernest Herningway, escritor estadounidense (1899 -1961) definió también esta concepción de deporte contemporáneo agregando: “Cuando un deporte es suficientemente atractivo para inducir a la gente a pagar para verlo se tiene en él el germen del profesionalismo”.

Llegamos así a comprender brevemente la evolución del deporte a través del tiempo, partiendo del hombre primitivo, valorando la importancia en la completitud del hombre pero apuntando particularmente al profesionalismo. Dentro de esta concepción es que se desarrolla el siguiente trabajo, atendiendo a las necesidades del deporte profesional estudiando específicamente una de las tantas disciplinas actuales: el voleibol.

“El voleibol ha tenido un desarrollo vertical desde su inicio a principios del siglo XIX. Esto se refleja en la cantidad de países miembros inscriptos en la Federación Internacional de Voleibol. Este deporte ocupa una posición de vanguardia en su expansión mundial al lado del fútbol y del básquetbol como juego deportivo internacional.

A ese desarrollo exterior compete también su posterior avance, al pasar de recreación a deporte olímpico, con grandes exigencias físicas, técnicas y tácticas y con una composición metodológica-pedagógica.

Junto a esto el voleibol es practicado en todos los campos del deporte popular como un medio de aprovechamiento del tiempo libre, la mantención de la salud y la rehabilitación, ofreciendo una eficacia doble.

El juego recorrió diferentes etapas en este desarrollo, las cuales han estado caracterizadas por determinadas posibilidades de la técnica y de la táctica (forma de juego) y con concepciones correspondientes a su forma de entrenamiento". (13)

En nuestro país le voleibol se juega de manera amateur en clubes, barrios y escuelas. También se practica de manera profesional en la Liga Argentina al igual de lo que sucede en otros países de los 5 continentes.

La Federación Internacional de Voleibol (FIVB), es la encargada de reglamentar y planificar la actividad deportiva del voleibol internacional, aquella de alto rendimiento a la cual se enfocará este análisis. A esta federación se adhieren las respectivas confederaciones continentales quienes reglan a su vez los países adheridos. Las 5 confederaciones continentales son: NORCECA (North, Central American & Caribbean), CEV (European Confederation), AVC (Asian Confederation), CSV (South American Confederation), CAVB (African Confederation). Los torneos que se destacan a nivel internacional son Liga Mundial, campeonatos continentales, Copa del Mundo, Mundial, Olimpíadas.

Entre los países que cuentan con ligas profesionales importantes se pueden mencionar: Italia, Brasil, Rusia, Polonia, Japón, Corea, Turquía, Alemania, Francia, Grecia, España, Puerto Rico, Argentina, Azerbaiyán, Austria, Chipre y Bélgica.

El ranking mundial actual de los primeros 25 países, vigente a enero 2012**,** según los puntos obtenidos durante las competencias internacionales se muestra en la tabla 3.1.

FIVB Senior World Ranking

Men

Rk.	Teams	Points
1	Brazil	252.50
2	Russia	234.50
3	Italy	191.00
4	Poland	182.00
5	Cuba	160.25
6	USA	148.00
7	Serbia	136.25
8	Argentina	121.75
9	Bulgaria	88.75

FIVB Senior World Ranking

Women

Rk.	Teams	Points
1	USA	245.00
2	Brazil	217.50
3	Japan	197.25
4	Italy	190.75
5	China	169.00
6	Serbia	145.00
7	Russia	131.25
8	Germany	108.25
9	Dominican. Rep.	72.75

10	China	57.25	10	Cuba	69.00
11	Egypt	55.00	11	Turkey	64.25
12	Iran	50.25	12	Thailand	62.00
13	Germany	46.50	13	Republic Of Korea	56.25
14	Cameroon	45.25	14	Poland	55.75
15	Japan	41.00	15	Kenya	50.75
16	Venezuela	40.75	16	Algeria	47.25
17	Puerto Rico	38.50	17	Peru	42.50
18	Canada	37.25	18	Argentina	40.00
18	Tunisia	37.25	19	Puerto Rico	33.50
20	Republic of Korea	36.50	20	Netherlands	30.25
21	France	35.75	21	Egypt	28.00
22	Australia	33.50	22	Canada	25.75
23	Mexico	33.00	23	Kazakhstan	23.75
23	Czech Republic	30.00	23	Czech Rep.	23.75
25	Algeria	26.25	25	Senegal	23.50

Tabla 3.1: FIVB Senior World Ranking

El análisis del voleibol de manera estratégica apunta sobre todo a ser una ayuda al voleibol profesional e internacional, a aquel que se juega en torneos mundiales donde la diferencia entre ganar o perder puede depender de sutilezas pero el resultado, es decir ser campeón o subcampeón no es jamás una insignificancia.

Para llevar adelante el análisis mencionado, es oportuno recurrir a la ayuda de la informática, que se valió del deporte para verificar sus resultados durante los primeros años de desarrollo de esta tecnología.

“La expansión de la computadora no solo no es ajena sino que viene de la mano del deporte. Un deporte en especial, el ajedrez, era utilizado para evaluar los niveles de performance de los primeros monstruos mecánicos que funcionaban a válvulas y tarjetas perforadas. Deep Blue, la supercomputadora que venció al número uno del ajedrez mundial Gary Kasparov, es la hija pródiga de aquellas primeras máquinas inteligentes. Ya en 1964, en los Juego Olímpicos de Tokio se utilizó una de aquellas para gestionar los resultados de las pruebas deportivas.

En los juegos Olímpicos de Invierno de Nagano, se realizaron estudios de biomecánica bajo la dirección de Kazuhiko Watanabe de la Universidad de Hiroshima. Se llevaron a cabo estudios con fines científicos e históricos de las pruebas de esquí de fondo, salto con esquí, bobsleigh, patinaje de velocidad y esquí acrobático”. (14)

El voleibol empieza a conectarse con la informática a través de la estadística, en el mundial de 1982 cuando se establece por primera vez un centro de procesamiento de datos en una competencia internacional de esta disciplina. Se realizaban informes estadísticos que eran entregados a los equipos una vez concluidos los partidos. A partir de allí y a medida que se desarrolla la tecnología surgen programas básicos preparados para manipular datos, calcular porcentajes, promedios, midiendo de esa manera rendimientos.

En la actualidad existen numerosos programas capaces de recolectar datos en el momento del juego y a posteriori. Entre ellos, los destacados a nivel internacional son: Volleyball Information System (VIS), Data Volley, Data Video, Elecrtronic Scoresheet, Beach Volleyball Information System (BVIS). Es posible calcular porcentajes, medias, identificar información relevante. La lectura de videos otorga confiabilidad a los datos porque permite la corrección o modificación de los mismos aun cuando el partido ha concluido. Es a partir de estos programas que se intentará llevar adelante la investigación planteada para poder obtener a través de ellos información clave para un entrenador de voleibol.

Se hará uso de nuevas herramientas que ofrece hoy la informática para que los datos relevados durante un partido dejen de ser solamente datos y se transformen en información y conocimiento útil al entrenador de un equipo de voleibol. Una de las funciones destacadas de un entrenador que se desenvuelve en competencias internacionales o profesionales es la preparación táctica antes de un partido. Poder contar con información de apoyo a la toma de decisiones en la definición de estrategias de juego puede ser una ayuda importante y determinante en el resultado del encuentro.

Teniendo presentes las nuevas tendencias y tecnologías disponibles encontramos las técnicas de minería de datos como herramientas efectivas en la búsqueda de la solución al problema propuesto. Éstas permiten extraer conocimiento útil previamente inexplorado desde un importante volumen de datos.

Analizar influenciadores claves, detectar categorías, resaltar excepciones, realizar pronósticos, analizar escenarios y calcular predicciones son actividades que sustentan la minería de datos.

La aplicación de técnicas de minería de datos a través de clustering (agrupamiento de datos en clases), árboles de decisión (técnica de aprendizaje supervisado que admite atributos discretos y continuos) y otros algoritmos de aprendizaje inductivo se utilizarán para detectar si existen patrones ocultos y reglas que caractericen el perfil del levantador para poder anticipar y definir estrategias de juego. El entrenador de un equipo de voleibol

contará así con un poyo concreto en la toma de decisiones para la preparación de un partido.

Como en todos los deportes, existe una tendencia a la especialización, es por esto que se inicia una primera investigación táctica de este deporte por aquel rol (levantador) que es el nexo entre los demás roles.

Debido a la complejidad casi infinita de situaciones que se pueden presentar durante un juego de voleibol y teniendo en cuenta la ineludible y primordial significación del levantador en la determinación de las secuencias de juego, se focalizará la atención en la evaluación de este rol desde distintas perspectivas: grupal, individual, en relación a compañeros, a jugadores adversarios y a la situación o momento del partido, a su posición en la cancha y a la fase del juego que se esté desarrollando.

Cabe puntualizar entonces que atendiendo a los instrumentos disponibles y mencionados se trabajará en conjunto respetando las reglas del deporte (del voleibol) y las de la informática, iniciando por un objetivo puntual (el comportamiento del levantador) pero dejando abierta la posibilidad de continuar hacia la completitud del equipo propio y adversario en un encuentro, un campeonato y una Liga Mundial.

3.2 MARCO TEÓRICO DEL CAMPO DE ACCIÓN

Sin hacer un recorrido exhaustivo de los conceptos de minería de datos, es preciso enunciar dos de las definiciones tradicionales más importantes que la describen.

Desde una óptica general: “La Minería de datos es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos” (Fayyad y otros, 1996).

Desde el punto de vista empresarial se define como: “La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo a la toma de decisión” (Molina y otros, 2001).

En cualquier proyecto de minería de datos es esencial el conocimiento del dominio del problema, es fundamental comprender qué es lo que se pretende hacer, cuál es el objetivo, cuáles son los datos disponibles y dónde es posible encontrarlos.

A partir de allí se podrán preparar los datos, explorarlos, generar modelos, validarlos e implementarlos para posteriormente actualizarlos cada vez que fuera necesario. Es decir se tratará de un proceso cíclico. Si bien existen varias tecnologías disponibles para llevar a

cabo este tipo de proyectos, se eligió Microsoft SQL Server 2008 R2. Su entorno integrado es útil para completar cada paso del proceso planteado. Dentro de ese ambiente está Business Intelligence Development Studio que no es otra cosa más que Microsoft Visual Studio 2008 con el agregado de la capacidad de realizar soluciones de inteligencia de negocio. El mismo permite desarrollar proyectos de Analysis Services, Integration Services, y Reporting Services.

La selección de este entorno no se debe a un análisis completo de las posibilidades o disponibilidades en el mercado, sino fundamentalmente al hecho de que se puede acceder a versiones libres y sobre todo a los conocimientos previos relacionados a Microsoft Visual Studio. Es oportuno mencionar que se han interrogado expertos en el tema con el fin de apoyar esta decisión. Entre ellos el profesor Esteban Alonso de la Universidad Austral de Buenos Aires, Facultad de Ingeniería.

Gráficamente y genéricamente se representan los pasos a seguir en la figura 3.1.

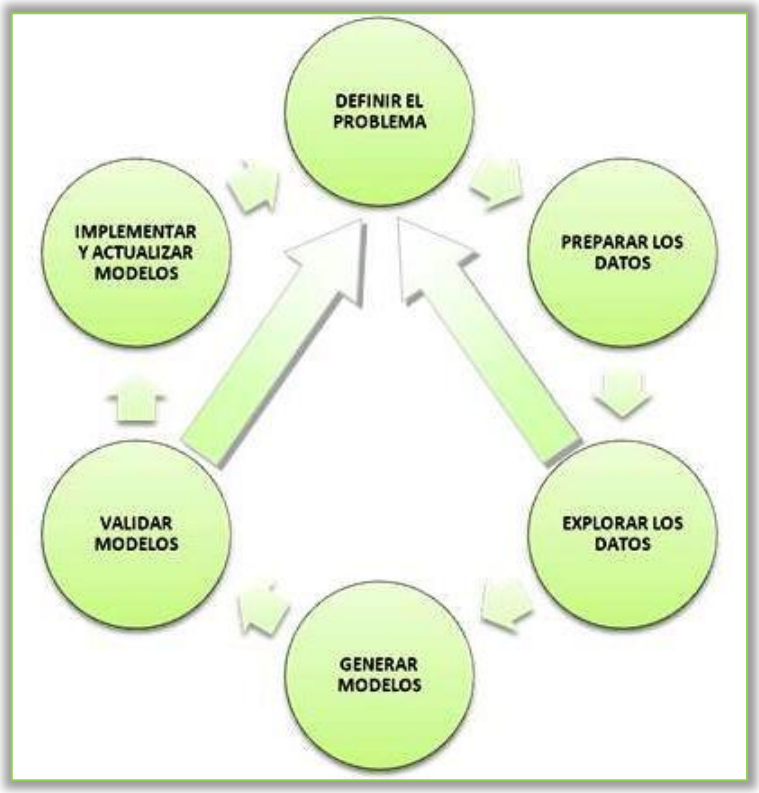


Fig. 3-1: Proceso cíclico Microsoft SQL Server 2008 R2

1) *Definir el problema:* en este paso se intentará definir claramente el problema y las métricas por las que se evaluará el modelo. Se deberá identificar qué es lo que se está buscando, qué resultado o atributo se desea predecir o si se pretende encontrar solamente asociaciones y patrones interesantes.

2) *Preparar los datos:* se realizará la recopilación de datos y verificación de calidad de los mismos. La herramienta de captura de datos durante un partido es el programa Data Volley:
   http://www.dataproject.com/VolleyBall/DataVolley2007.aspx
   Este programa es muy dúctil a la hora de definir y personalizar la manera de captar datos. Se deberá prestar atención a posibles inconsistencias si los datos no hubieran sido relevados todos por la misma persona. Será necesario unificar las acciones para poder sumar datos al modelo.
3) *Explorar los datos:* se deberán limpiar los datos, seleccionarlos, integrarlos para obtener la vista minable o dataset.
4) *Generar Modelos:* aquí se aplicarán las técnicas de minería de datos a los datasets obtenidos anteriormente. Se definirán las columnas de entradas, el atributo que se intenta predecir y los parámetros que indicarán al algoritmo elegido cómo procesar los modelos. El procesamiento del modelo es el entrenamiento del mismo, es decir el algoritmo se aplicará a la estructura de datos para extraer patrones.
5) *Explorar y Validar los modelos:* En este paso se debe verificar si los modelos son útiles a las necesidades. Se podrán separar datos en conjuntos de datos de prueba y entrenamiento para evaluar con precisión el rendimiento de los mismos.
6) *Implementar y actualizar los modelos:* como último paso se deberán presentar los patrones descubiertos de manera útil al entrenador o a quien deba tomar las decisiones del planteo táctico del juego. Si fuera necesario se deberán corregir modelos, actualizarlos y volver a presentarlos.

4 TERCERA PARTE CONCRECIÓN DEL MODELO

La ejecución práctica del proyecto se desarrolló, como se dijo antes, utilizando la metodología CRISP-DM. A continuación se exponen los pasos realizados en cada una de las fases y los resultados obtenidos.

4.1 FASE DE COMPRENSIÓN DEL NEGOCIO

Junto con los expertos en el tema (entrenador y estadísticos del equipo) se realizó una primera reunión informal donde se planteó la idea propuesta.

Luego se sucedieron reuniones específicas donde cada uno aportó su know-how a fin de lograr una comprensión acabada de lo que se buscaría hacer.

Las tareas realizadas y enumeradas a continuación permitieron obtener un conocimiento completo del entorno, objetivos, herramientas y material disponible.

4.1.1 Determinar objetivos del negocio

- *Background:* en este contexto se puede decir que el negocio es el voleibol de alto nivel. Éste es un deporte de conjunto, donde se requieren habilidades físicas y cualidades psíquicas. Existen roles muy bien definidos, uno de ellos es fundamental en el desarrollo de una estrategia de juego: el levantador.
- *Objetivos del Negocio:* el objetivo desde el punto de vista del negocio es ganar la mayor cantidad de partidos internacionales.
- *Criterios de éxito del negocio:* el negocio será exitoso si logra posicionar el voleibol de Puerto Rico entre los 10 primeros equipos de nivel mundial.

4.1.2 Valoración de la situación

- *Inventario de recursos:* se cuenta con datos provenientes del sistema operacional. Se tiene la disponibilidad y ayuda de los operadores y del estadístico principal para identificar variables relevantes, interpretar resultados y guiar la investigación en cuanto a la manera de poder encontrar los datos disponibles necesarios para el desarrollo del problema planteado.
- *Requisitos, supuestos y restricciones:* es necesario contar con partidos de nivel internacional donde el levantador que haya salido al campo de juego haya sido siempre la misma persona. Lo que se analiza es el comportamiento de un jugador, no del equipo en general, se supone que el equipo se comportará según las decisiones del levantador (que es quien lleva adelante la estrategia planteada por el entrenador), por este motivo es ese rol el que estará en estudio. En el caso en que el levantador principal del equipo sea sustituido durante el transcurso de un partido se deberán eliminar los registros correspondientes a las acciones del sustituto porque no serán considerados a la hora de armar el modelo.

- *Riesgos y contingencias:*

RIESGO	PROBABILIDAD OCURRENCIA	CONTINGENCIA
No se cuenta con experiencia en proyectos de Minería de Datos	100%	Se realizó una diplomatura en Business Intelligence para adquirir capacitación
Escasa cantidad de datos disponibles	60%	El análisis de las conclusiones deberá ser realizado junto a quienes conocen plenamente el dominio del problema y no hacer una lectura fría de los datos estadísticos descriptivos.
Escasa práctica en el uso de herramientas de minería de datos	70%	Antes de iniciar el proyecto se siguieron tutoriales y se realizó ejercitación necesaria para adquirir destreza.
Error al estimar el tiempo total del proyecto	50%	Si el tiempo excediera los límites académicos permitidos se buscará ayuda de terceros para reforzar la capacitación y poder así concluir el mismo.

Tabla 4.1: Riesgos y contingencias

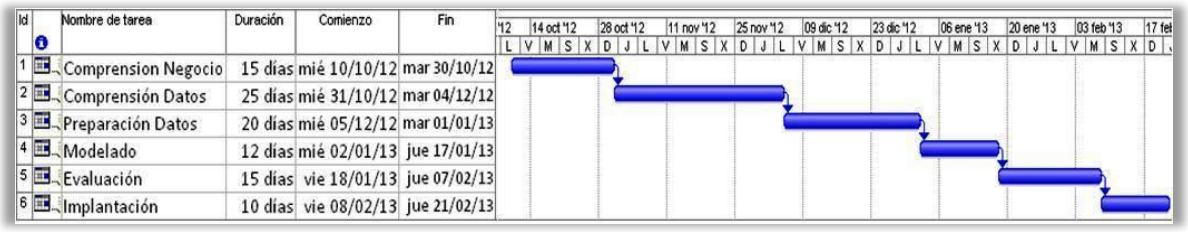
- *Costes y Beneficios:* En este caso al tratarse de un trabajo académico se utilizarán herramientas en versión trial, un posterior reconocimiento de los resultados podría derivar en la evaluación efectiva del costo total.

4.1.3 Determinar los objetivos de la minería de datos

- *Meta:* colaborar con el entrenador en la definición de estrategias de juego.
- *Objetivos:* predecir el comportamiento del levantador en complejo K1 según las siguientes perspectivas: llamada del central, fase, tipo y zona de recepción, momento del juego.
- *Criterios de éxito:* el proyecto será exitoso si se descubren relaciones desconocidas hasta el momento en el comportamiento del levantador que resulten útiles para el entrenador en la preparación táctica de los partidos.

Explicación: Sabiendo que existe un alto equilibrio entre los equipos de nivel profesional cuando juegan el complejo K1, puede decirse entonces que la ventaja la tendrá el quipo que logre ganar más puntos en el complejo K2 (denominados break point). Si un equipo necesita hacer puntos en el complejo K2, indiscutiblemente el adversario estará en complejo K1. Nunca los dos equipos e encuentran en la misa situación. Es necesario estudiar al adversario en complejo K1 para poder hacer con el propio equipo puntos en el complejo k2 y de esa manera poder obtener la victoria.

4.1.4 Realizar el plan del proyecto



Id	Nombre de tarea	Duración	Comienzo	Fin
1	Comprension Negocio	15 días	mié 10/10/12	mar 30/10/12
2	Comprensión Datos	25 días	mié 31/10/12	mar 04/12/12
3	Preparación Datos	20 días	mié 05/12/12	mar 01/01/13
4	Modelado	12 días	mié 02/01/13	jue 17/01/13
5	Evaluación	15 días	vie 18/01/13	jue 07/02/13
6	Implantación	10 días	vie 08/02/13	jue 21/02/13

Fig.4-1: Plan del proyecto

4.2 FASE DE COMPRENSIÓN DE LOS DATOS

Los datos con los que se cuenta están agrupados en planillas Excel, los mismos provienen del sistema operacional. La recolección se realiza en dos etapas. La primera de ellas es la carga on-line de los partidos disputados, la segunda etapa es la visualización a través del programa Data Video. Este sistema permite seleccionar y agrupar imágenes según las especificaciones que se deseen estudiar. En esta instancia, se trabajará con datos del Complejo 1 o K1 (previamente definido en los conceptos importantes). Esta elección se debe a la decisión del experto (el entrenador de equipo) quien sostiene que como primer paso clave en la victoria de un encuentro es necesario anticipar al adversario en la fase de cambio de saque.

4.2.1 Recolección de datos iníciales

Fueron entregadas diversas planillas Excel correspondientes a los partidos disputados durante el año 2012. No se considera válido ir históricamente más atrás de esta fecha por dos motivos fundamentales. El primero de ellos es que el levantador que ha representado el equipo en cuestión durante este año es siempre el mismo, no ocurrió así en el 2011 donde la participación del actual levantador fue salteada. El segundo motivo es que se considera que un año de juego a nivel internacional aporta experiencia, da solidez a los jugadores y estos factores influyen en el rendimiento posterior. La experiencia adquirida y

los entrenamientos constantes ayudan a la modificación permanente y perfeccionamiento. Se busca siempre la corrección de malos hábitos y costumbres intentado superar rendimientos anteriores. De eso se trata el deporte cuando es ejercido a nivel profesional.

Los equipos contra los que se enfrentó Puerto Rico durante el transcurso del año 2012 fueron los siguientes: Canadá (dos derrotas 0-3 y 0-3), México (tres victorias 3-0, 3-0, 3-0), Estado Unidos (dos derrotas 0-3 y 1-3), Irán (victoria 3-2), Serbia (derrota 3-0), Japón (victoria 3-2), Corea (victoria 3-2), Australia (derrota 1-3), China (victoria 3-2), República Dominicana (victoria 3-0), Cuba (derrota 2-3), Venezuela (dos victorias 3-0 y 3-2).

4.2.2 Descripción de los datos

Para poder describir los datos de manera comprensible se copia en la tabla 4.2 una muestra del archivo síntesis en formato Excel que se obtiene al finalizar un partido.

Fase	Set Num	Set Moment	Serer	Rec Play Num	Rec Zone	Rec Eval	Setter Pos	Setter Atk	Setter call	Setter	Setter Eval	Atk Pos	Atk Play Num	Atk Eval
1	1	A1	S	1L	Dx	#	23		1C	Se	#	4	13	#
1	1	A2	S	0L	C	!	1		0T	AL	#	4	13	NE
1	2	A1	S	0L	C	+	2		1C	Se	C	4	15	NE
1	3	A2	S	1L	Dx	#	23		A	A	#	23	5	#
1	3	A3	S	3L	Sx	!	2		0T	AL	#	4	13	=
1	3	A3	S	0L	C	+	23		1C	Se	C	2	4	NE
1	4	A2	S	0L	C	+	2		1C	Se	C	4	15	NE
1	4	A3	S	1L	Dx	#	23		A	A	#	23	5	#
1	5	B1	S	3L	Sx	!	2		0T	AL	#	4	13	=
1	5	B2	S	0L	C	+	23		1C	Se	C	2	4	NE
6	1	A1	S	3L	Sx	#	3		1T	1S	#	3	5	#
6	1	A2	S	3L	Sx	#	23		1C	3S	C	1C	5	NE
6	1	A3	S	3L	Sx	#	3		A	Se	C	4	3	NE
6	2	A1	S	1L	Dx	#	3		1C	S	C	4	3	NE

Tabla 4.2: Muestra de la recolección de datos Puerto Rico vs. Canadá

La definición del tipo de variable, rangos permitidos, etc. se explica usando terminología del programa SPSS que es el que se utilizará para la preparación y exploración del dataset.

En una primera instancia se muestran los datos tal como están almacenados en su origen, posteriormente se analizarán las modificaciones que se consideren pertinentes según el objetivo del trabajo y las necesidades respectivas.

Ingresando el dataset a SPSS se obtuvo el resultado reflejado en la figura 4-2.

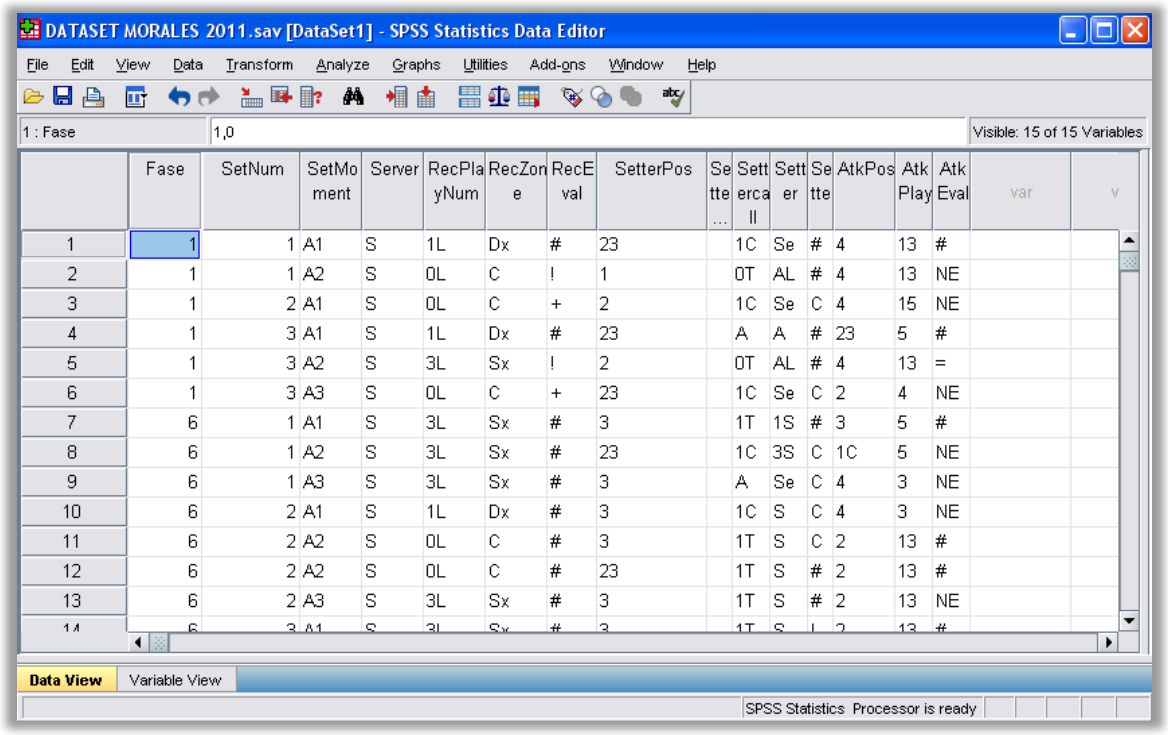


Fig. 4-2: Dataset

La vista de las variables se recoge en la tabla 4.3.

NAME	TYPE	WIDTH	DECIMALS	LABEL	VALUES	MISSING	COLUMNS	ALIGN
Fase	Numeric	11	0	Fase	None	None	11	Right
SetNum	Numeric	11	0	Set Num	None	None	11	Right
SetMoment	Numeric	8	2	Set Moment	None	None	8	Right
Server	String	2	0	Server	None	None	2	Left
RecPlayNum	String	22	0	Rec Play Num	None	None	22	Left
RecZone	String	2	0	Rec Zone	None	None	2	Left
RecEval	String	1	0	Rec Eval	None	None	1	Left

SetterPos	Numeric	11	0	Setter Pos	None	None	11	Right
SetterAtk	String	1	0	Setter Atk	None	None	1	Left
Settercall	String	22	0	Setter call	None	None	22	Left
Setter	String	22	0	Setter	None	None	22	Left
SetterEval	String	1	0	Setter Eval	None	None	1	Left
AtkPos	Numeric	11	0	Atk Pos	None	None	11	Right
AtkPlayNum	Numeric	11	0	Atk Play Num	None	None	11	Right
AtkEval	String	22	0	Atk Eval	None	None	22	Left

Tabla 4.3: Vista de metadatos

Según el momento de este estudio, los datos pueden ser llamados hasta aquí datos brutos.

Según la escala de medida las variables encontradas pueden clasificarse en cuantitativas discretas (Fase, SetNum, SetMoment, SetterPos, AtkPos, AtkPlayNum) y cualitativas nominales o categóricas (Server, RecPlayNum, RecZone, RecEval, SetterAtk, SetterCall, Setter, SetterEval, AtkEval). Se destaca que el presente estudio no utilizará variables cuantitativas continuas, no es pertinente dentro del dominio del problema.

En cuanto a la dimensionalidad de los datos a este punto puede decirse que nos encontramos con datos multivariados, dado que las diferentes propiedades del dataset se miden en un conjunto específico de objetos, en este caso la levantada a partir de las demás propiedades.

Para concluir esta actividad se realizó la unión de los files recibidos consolidando todo en un único dataset. Utilizando SPSS y a través de las funciones MERGE y ADD CASES, se formó el dataset con 1079 registros. La estructura de la tabla resultante es de una fila por caso. Es decir, para cada levantada o variable Setter existe una fila con sus atributos que la describen.

Se observa que en este punto del estudio, debido a la manera de leer los datos de SPSS los nombres de las variables aparecen sin el espacio en el medio. Esta característica probablemente no permanecerá vigente al momento de ingresar los datos a la base de datos dado que los mismos se integrarán a partir de Excel y allí los nombres se encuentran con espacios en el medio, tal como se usan en el sistema operacional.

4.2.3 Reporte de exploración de los datos:

Para realizar esta tarea se verificaron en primer lugar los tipos de datos y se definieron los rangos o valores permitidos. Esta labor es fundamental para poder aplicar las técnicas estadísticas adecuadas en cada caso. La vista de variables quedó redefinida como muestra la figura 4-3.

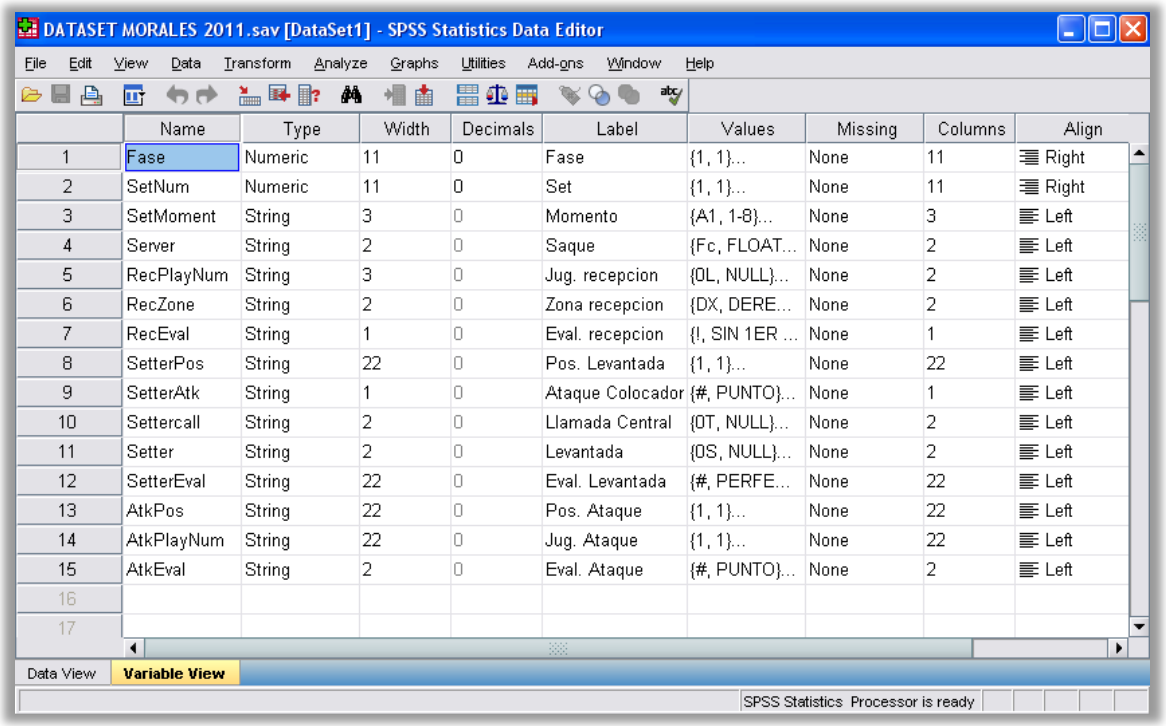


Fig. 4-3: Redefinición dataset

Todas las variables son ahora de tipo nominal, debido a que los atributos a los que hacen referencia son todos del tipo cualitativo. Explícitamente los valores posibles se expresan en la tabla 4.4.

VARIABLE	DESCRIPCIÓN	VALORES POSIBLES
Set Num	Indica el set en que se realiza la acción	1,2,3,4,5
Set Moment	Indica el momento del set en que se realiza la acción	A1, A2, A3, B1, B2
Server	Saque o servicio	S (salto) , SF (salto float), FL(float), FC(float corto)
Rec Play Num	Número del jugador que realiza la acción	[1,20]

Rec Zone	Zona del campo desde donde llega la recepción al levantador	SX (izquierda), DX(derecha), C(centro)
Rec Eval	Evaluación de la recepción	#(perfecta), *(permite solo primer tiempo), +(permite primer tiempo), -(alta), EP(error punto), !(sin primer tiempo), /(over pass)
Setter Pos	Posición desde donde levanta el levantador	1,2,3,4,5,6,23 (entre 2 y 3) ,34 (entre 3 y 4) ,61 (entre 6 y 1) ,65 (entre 6 y 5)
Setter Atk	Evaluación del ataque del levantador	#(perfecta), =(error punto) , 0 (transición)
Setter call	Llamada del central	1T (adelante), 1C(separada), 3T(flecha), A(atrás), 0T(null)
Setter	Tipo de Levantada	Se(semi), AL(alta), A(atrás), 3S(flecha), P(pipe), S(super), 1S(adelante), 1C(alejada)
Setter Eval	Evaluación de la levantada.	#(perfecta), =(error), B(baja), C(corta), L(larga)
Atk Pos	Posición del campo desde donde se ejecuta el ataque	1, 2, 3, 4, 5, 6, 23 (entre 2 y 3) , 34 (entre 3 y 4), 65 (entre 6 y 5) , 61 (entre 6 y 1)
Atk Play Num	Número del jugador que ataca	[1,20]
Atk Eval	Evaluación del ataque.	#(perfecta), =(error punto), 0(transición)

Tabla 4.4: Tabla de Datos

Como primer paso para la exploración de datos se utilizaron técnicas de exploración visual a través del software SPSS. Debido al tipo de variables cualitativas la seleccionada fue diagrama de barras.

Teniendo presente que el objetivo fundamental será predecir el tipo de levantada se inicia la exploración por esa variable primordial.

Análisis de la variable Setter

N	Valid	1079
	Missing	0

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NULL	87	8,1	8,1	8,1
	1C	52	4,8	4,8	12,9
	1S	48	4,4	4,4	17,3
	3S	112	10,4	10,4	27,7
	A	36	3,3	3,3	31,0
	AL	161	14,9	14,9	46,0
	P	74	6,9	6,9	52,8
	S	273	25,3	25,3	78,1
	Se	195	18,1	18,1	96,2
	SE	41	3,8	3,8	100,0
	Total	1079	100,0	100,0	

Tabla 4.5: Distribución de frecuencias Setter

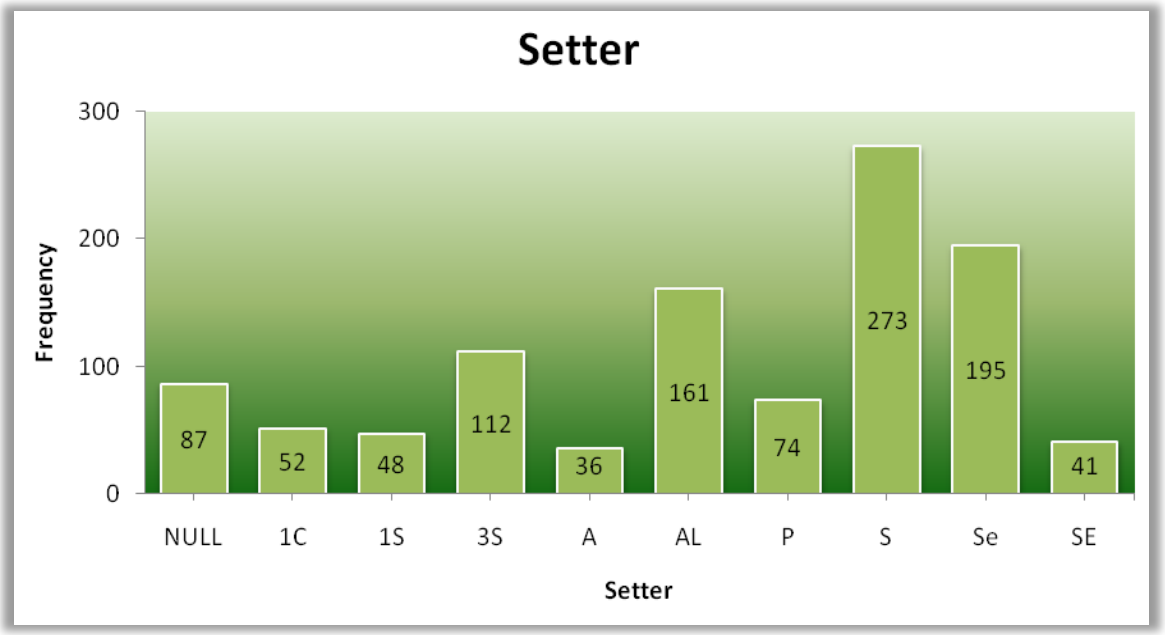


Fig. 4-4: Distribución de frecuencias Setter

Se observa que existe una categoría de valores nulos, será necesario indagar el motivo de ese faltante, si se trata de datos perdidos, de campos que no deben considerarse, etc. Se descubre también que existe probablemente un error en la registración o una falta de criterio homogéneo en la manera de ingresar el dato Se dado que la tabla de frecuencias identifica dos categorías diferentes una Se y otra SE. Probablemente las dos están

indicando el mismo tipo de jugada, es decir "Semi". Se nota también como particularidad que no se relevaron casos perdidos, es decir todos los registros han sido clasificados.

Se continúa el análisis por la variable Setter call que seguramente será uno de los indicadores importantes en el estudio debido a la importancia que tiene en el juego conocer dónde se jugará el primer tiempo, es decir la pelota más rápida de la acción.

Análisis de la variable Setter call

N	Valid	1079
	Missing	0

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NULL	249	23,1	23,1	23,1
	1C	251	23,3	23,3	46,3
	1T	148	13,7	13,7	60,1
	3T	311	28,8	28,8	88,9
	A	120	11,1	11,1	100,0
	Total	1079	100,0	100,0	

Tabla 4.6: Distribución de frecuencias Setter call

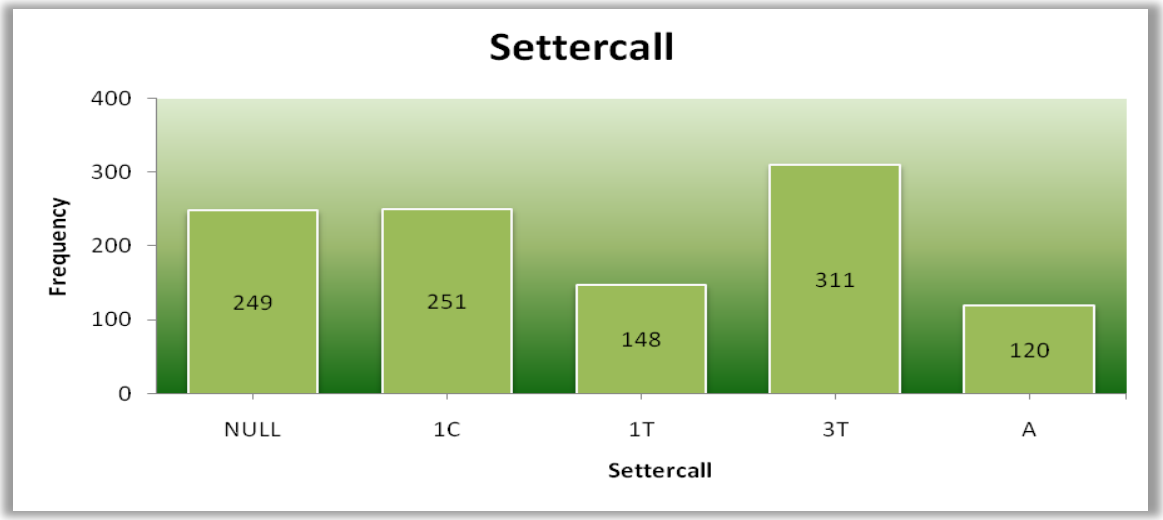


Fig. 4-5: Distribución de frecuencias Settercall

Aquí también se lee que existe una categoría donde hay valores nulos para el conjunto de datos resultante. Se investigará el motivo de este faltante para poder tratar los datos de manera adecuada. Se nota además que no hay registros perdidos que no hayan podido ser clasificados.

Como tercer y cuartoanálisis se visualizan las distribuciones de frecuencias de las variables Rec Eval y Rec Zone.

Análisis de la variable Rec Eval

N	Valid	1079
	Missing	0

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-	35	3,2	3,2	3,2
	!	142	13,2	13,2	16,4
	*	16	1,5	1,5	17,9
	/	29	2,7	2,7	20,6
	#	573	53,1	53,1	73,7
	+	258	23,9	23,9	97,6
	=	26	2,4	2,4	100,0
	Total	1079	100,0	100,0	

Tabla 4.7: Distribución de frecuencias Rec Eval

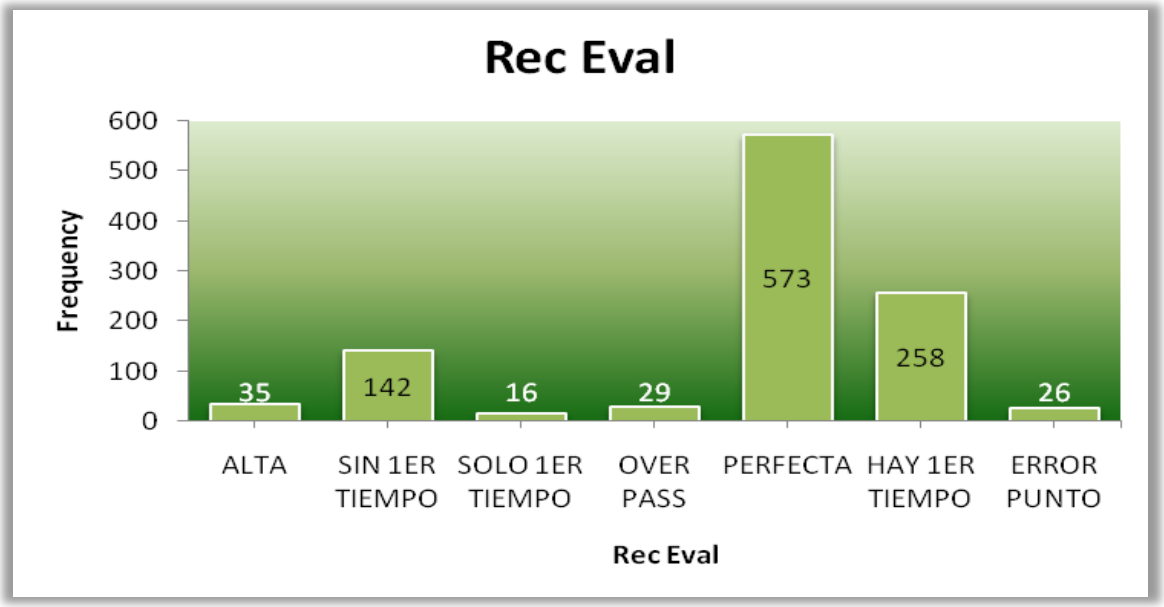


Fig. 4-6: Distribución de frecuencias Rec Eval

Las dos categorías más importantes son la evaluación calificada como "perfecta" y la de "hay primer tiempo" dado que en estos dos casos es el levantador con sus propias características el que decide dónde pasará la pelota. Entre ambas categorías se obtiene el 77% de los datos. Se decidirá a posteriori si se podrán poner en consideración el resto de

las categorías o será oportuno quedarse solamente con estas dos. Se nota nuevamente que no hay registros faltantes, todos han sido clasificados.

Análisis de la variable Rec Zone

N	Valid	1079
	Missing	0

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	C	483	44,8	44,8	44,8
	Dx	144	13,3	13,3	58,1
	DX	109	10,1	10,1	68,2
	Sx	293	27,2	27,2	95,4
	SX	50	4,6	4,6	100,0
	Total	1079	100,0	100,0	

Tabla 4.8: Distribución de frecuencias Rec Zone

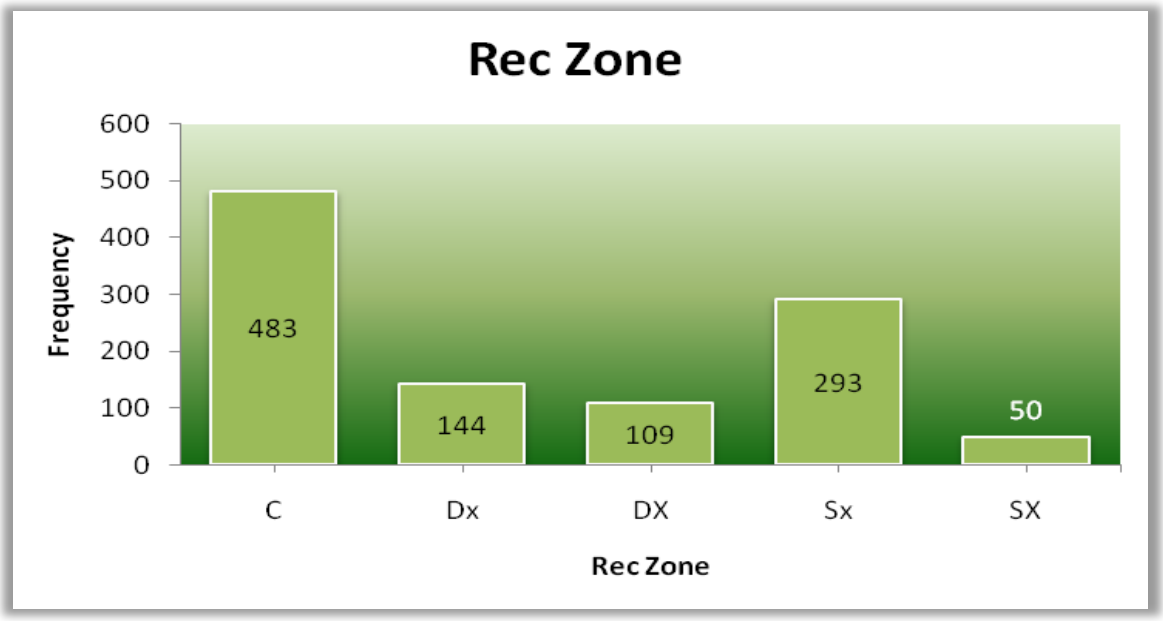


Fig. 4-7: Gráfico "Rec Zone"

Aquí se observa nuevamente que existen valores ambiguos en la manera de llamar la recepción que llega por la derecha y la que llega por la izquierda. Se encuentran valores Dx y DX para indicar derecha y valores Sx y SX para indicar izquierda. Deberá unificarse este

dato para poder ser tratado oportunamente. Una vez más se verifica que todos los registros han sido clasificados.

Para empezar a comprender la relación entre las variables disponibles, su dependencia o independencia y como una primera observación hacia lo que podría definir luego el tipo de algoritmo a utilizar se realizaron pruebas de Chi-Cuadrado Pearson logrando los siguientes resultados.

Comenzando por la relación entre llamada del central (Setter call) y levantada (Setter) se obtienen las tablas 4.9 y 4.10.

		Settercall					Total
		NULL	1C	1T	3T	A	
Setter	NULL	84	3	0	0	0	87
	1C	0	52	0	0	0	52
	1S	0	0	48	0	0	48
	3S	0	11	0	101	0	112
	A	0	0	0	0	36	36
	AL	148	0	0	5	8	161
	P	0	16	16	42	0	74
	S	0	95	78	69	31	273
	SE	17	74	6	94	45	236
	Total	249	251	148	311	120	1079

Tabla 4.9: Tabla de frecuencias cruzadas: Setter / Settercall

	Value	Df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	1950,216a	40	**,000**	**,000**[b]	0	0
Likelihood Ratio	1692,658	40	**,000**	**,000**[b]	0	0
Fisher's Exact Test	1583,128			**,000**[b]	0	0
N of Valid Cases	1079					

a. 7 cells (12,7%) have expected count less than 5. The minimum expected count is 2, 22.

b. Based on 10000 sampled tables with starting seed 1314643744.

Tabla 4.10: Test de Chi-Cuadrado

Chi cuadrado no debe aplicarse si más del 20% de las frecuencias esperadas es inferior a 5 o alguna inferior a uno, en este caso no se viola la enunciada restricción por lo tanto se considera válido el test. En la tabla 4.10 se muestra también el test Fisher Exact y se

mantiene el mismo resultado, las variables son dependientes porque de Asymp. Sig. es menor que 0,05.

Realizando las pruebas para las variables Server y RecEval se obtienen las tablas 4.11 y 4.12.

		Rec Eval							Total
		-	!	*	/	#	+	=	
Server	Salto	27	130	11	0	389	146	21	724
	Salto float	8	12	5	29	184	112	5	355
Total		35	142	16	29	573	258	26	1079

Tabla 4.11: Tabla de frecuencias cruzadas: Server / RecEval

	Value	Df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	114,487[a]	6	**,000**	**,000**[b]	0	0
Likelihood Ratio	129,295	6	**,000**	**,000**[b]	0	0
Fisher's Exact Test	124,155			**,000**[b]	0	0
N of Valid Cases	1079					

a. 2 cells (2, 0%) have expected count less than 5. The minimum expected count is 4,66

b. Based on 10000 sampled tables with starting seed 1502173562...

Tabla 4.12: Test de Chi- Cuadrado

Este test sirve a una verificación que podría decirse lógica, dado que la dificultad que existe al recibir un saque en salto no es la misma que la de recibir un saque float, es correcto que haya relación entre las variables, son dependientes. Aquí también el valor de Asymp. Sig. es menor que 0,05.

Como conclusión se verifica que la variable objetivo está relacionada con las demás. Esta información sirve a priori para descartar técnicas de minería de datos en las que los supuestos indiquen la necesidad de que las variables que predicen el objetivo sean independientes. Por ejemplo no se podrán aplicar técnicas de regresión lineal.

4.2.4 Verificar la calidad de los datos

Desde el punto de vista de la consistencia de los datos individuales en el campo Setter se detectaron valores nulos.

Teniendo en cuenta el dominio del problema se llega a la conclusión de que cuando la evaluación de la recepción es doble negativo (error punto), negativo (permite solo pelota alta) o barra (no permite levantada), no existe posibilidad de que el levantador elija una opción de ataque, por lo tanto se descartarán en el estudio estos campos, sin afectar la calidad de los datos disponibles.

También existen datos faltantes en el mismo campo Setter debido a una recepción perfecta y ataque por parte del levantador, en este momento del estudio se decide mantener estos campos porque hacen a la estrategia de juego y se evaluará posteriormente si es conveniente filtrarlos en la construcción del modelo.

4.3 FASE DE PREPARACIÓN DE LOS DATOS

Esta fase se encuentra estrechamente vinculada a la etapa de modelización, por lo tanto las tareas están relacionadas a las técnicas de minería de datos que se estima serán utilizadas.

4.3.1 Selección de Datos

4.3.2 Calidad de Datos

Se considera irrelevante el campo AtkPlayNum, porque no aporta información en la definición de la estrategia de juego. Tampoco se tendrá en cuenta el campo RecPlayNum. Ambos identifican el número del jugador que realiza la acción. No se vinculan al objetivo propuesto por lo tanto oportunamente se descartarán en el modelo.Calidad de datos

Se mencionaron las causas de los campos con valores nulos con lo cual no existen casos particulares a tener en cuenta. Se considera bueno el dataset disponible.

4.3.3 Estructurar Datos

Analizando la cantidad de registros disponibles, se destaca que la misma es acotada. El motivo es la cantidad de partidos disponibles. El análisis es específico y orientado hacia un rol, un jugador en especial, por lo tanto esto también influye en la cantidad de datos a estudiar.

Debido a estos particulares se analizó la necesidad de reducir la dimensión de la variable target u objetivo, es decir el campo Setter. Teniendo en cuenta la importancia en la

diferenciación entre jugadas de primer tiempo y jugadas de segundo tiempo se decide que será necesario redefinir la variable Setter agrupándola solamente en dos categorías: Primer Tiempo FT (1S, 1C, 3S y A) y Segundo tiempo o 2T (P, S, SE) dejando inalterada la opción de jugada alta, es decir el tipo AL, denominada Tercer Tiempo.

Además de esto se unificará el criterio para la denominación de la variable que indica el lugar desde donde llega la recepción.

Estas modificaciones se realizarán a través de Integration Services creando procesos de ETL (Extract, Transform, Load / Extracción, Transformación y Carga).

4.3.4 Integrar datos

La integración de los datos quedará plasmada luego de la ejecución de los procesos de ETL que deberán generarse.

4.3.5 Formateo de los datos

Las transformaciones de los datos se realizaron creando diferentes paquetes en Integration Services, unificados en un proceso denominado Proceso de Carga. La finalización o salida de este proceso es un data set homogéneo, listo para ser minado.

- Paquete UnirTablas (figura 4-8).

  Tabla Origen: MORALES 2012.

  Tabla destino: UnionJuegosExcel.

  Aquí se tomó la planilla Excel MORALES 2012proveniente del actual sistema operacional, se cargaron las hojas correspondientes a la recopilación de datos de cada uno de los partidos jugados durante el año 2012 por el levantador del equipo Fernando Morales. Se creó una nueva base de datos, Data Volley, a través de la conexión con SQL Server 2008 y se generó una tabla unión conteniendo el total de juegos disputados.

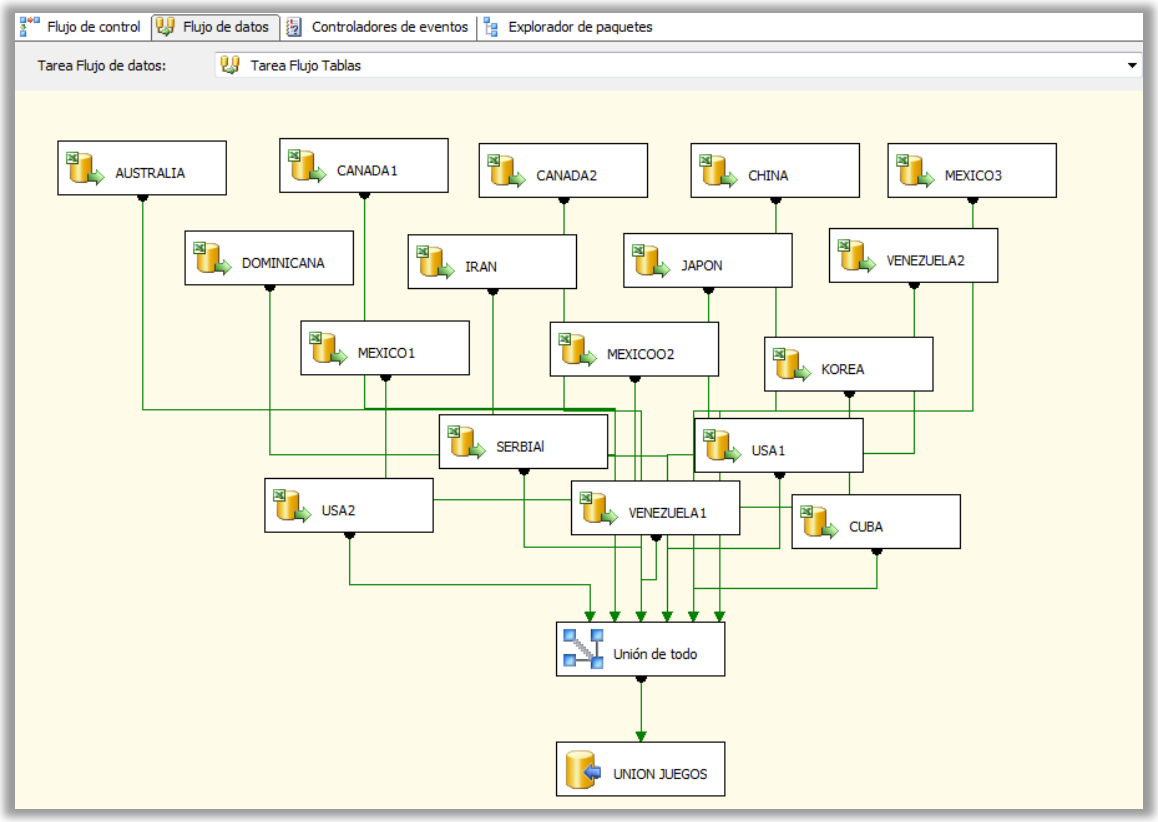


Fig.4-8: Paquete UnirTablas

- Paquete OrdenarValores (figura 4-9).

  Tabla origen: UnionJuegosExcel.

  Tabla destino: JuegosOrdenados.

  Este paquete es uno de los más importantes en cuanto a las transformaciones realizadas.

  La columna Setter fue la que recibió la mayoría de los cambios debido a su importancia como variable target u objetivo. Se unificaron criterios reconocidos en la fase de comprensión de los datos, se redujo la dimensionalidad de la variable y se ordenó el dataset a través de la variable Fase que identifica una situación precisa del levantador en cuanto a su ubicación en el campo de juego.

  Se corrigieron los errores detectados en las tablas de frecuencias, unificando criterios en las definiciones de Se y SE en el campo Setter y de Sx, SX, Dx y DX en el campo RecZone.

  La dimensionalidad de la variable Setter fue redefinida respecto a la propuesta inicial por considerarse el tercer valor posible o tercer tiempo de importancia en la verificación del funcionamiento correcto de los algoritmos. No es necesario

predecir una jugada de tercer tiempo, dado que todo el equipo tiene tiempo de prepararse para la defensa, entonces, si el algoritmo a partir de los datos de entrada responde con un tercer tiempo cuando conocemos que esa es la realidad, estamos cerca de decir que el mismo funciona correctamente. Por lo tanto quedaron agrupadas las jugadas de primer tiempo bajo el nombre FT, las de segundo tiempo bajo el nombre 2T y las de tercer tiempo con el nombre 3T.

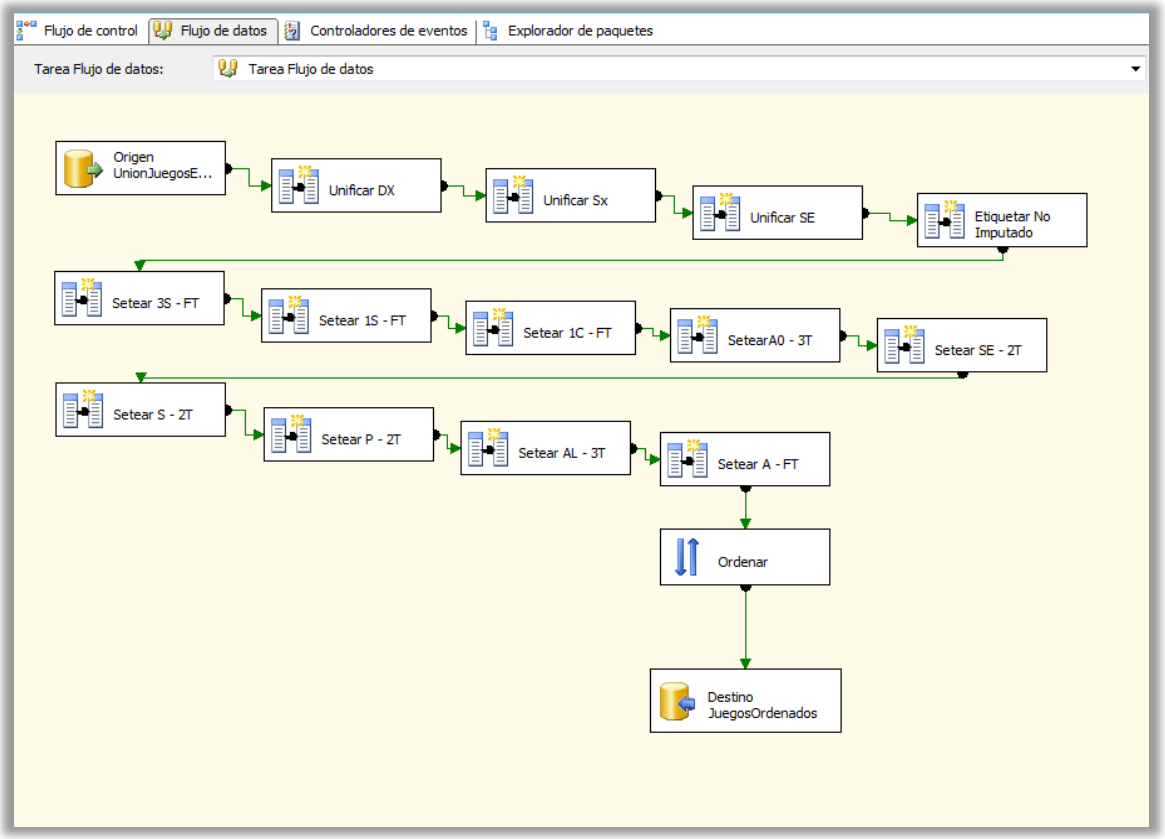


Fig. 4-9: Paquete OrdenarValores

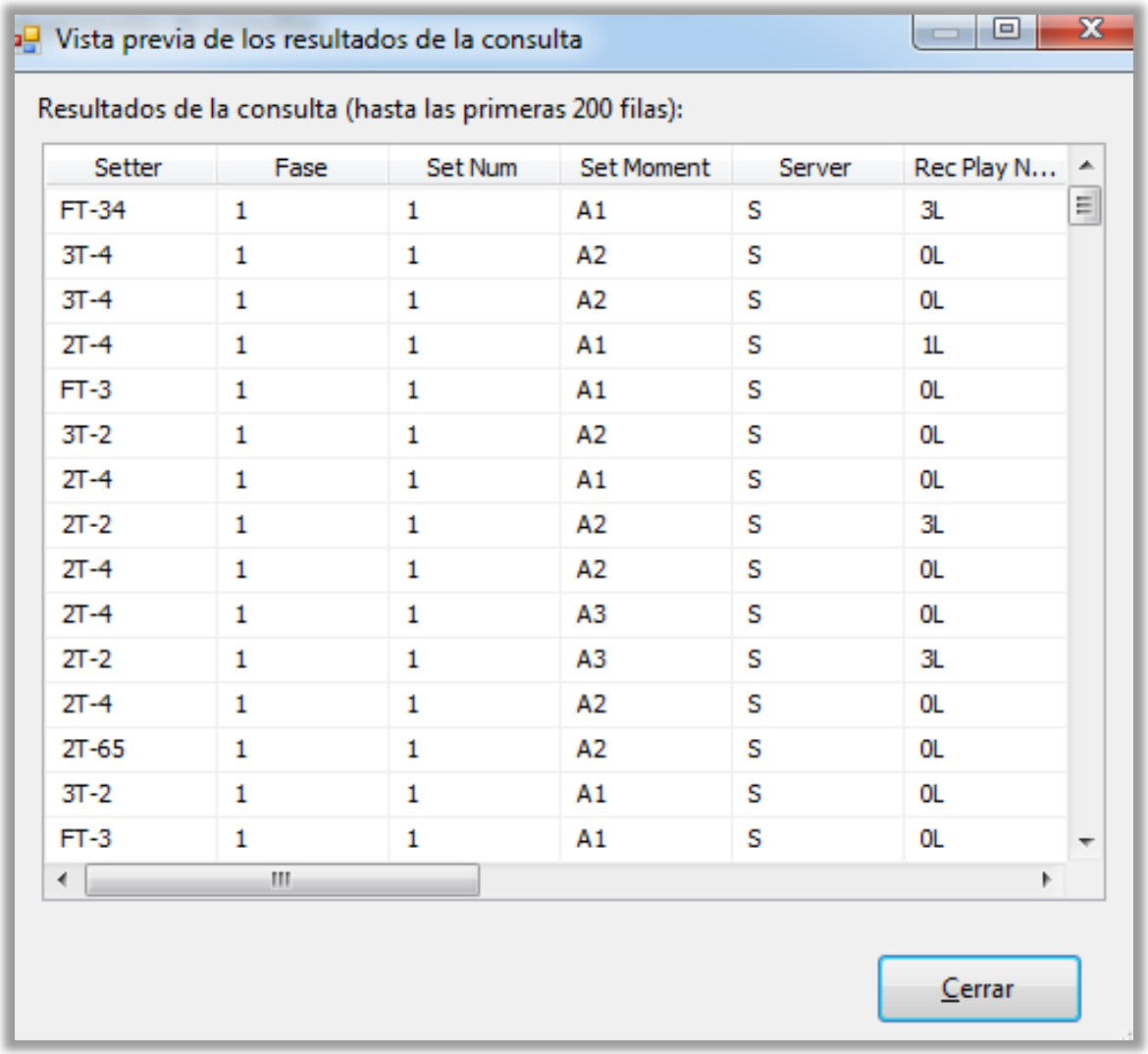
A partir de aquí se probaron modelos y se volvió sobre la limpieza del dataset minable. Después de análisis y reuniones con los expertos se encontró necesario unir los campos Setter y AtkPos, para poder predecir el tipo de levantada y la posición en la redo donde se ejecuta la acción. No serviría de mucho identificar únicamente el tipo de levantada dado que un primer tiempo (FT) puede ser jugado a lo largo de los 9 metros de la red según preferencias o posibilidades del levantador. Dependiendo de la posición donde esta acción se ejecuta es necesario preparar el bloqueo y la defensa adversaria. Estas decisiones se obtuvieron como conclusión a la primera prueba de modelado. Debido a este análisis se decidió confeccionar un nuevo paquete.

- Paquete UnirSetterAtk (figura 4-10).

Tabla Origen: en este caso se trata de una sentencia SQL a partir de la tablaJuegosOrdenados. La misma concatena los campos Setter y AtkPos dentro del campo Setter.

Explicación: Como primera etapa en el análisis del levantador interesa la posición desde donde se ejecutará el ataque. No se pretende estudiar estadísticamente qué jugador está ejecutando la acción pero sí desde qué lugar de la cancha el colocador decide hacer realizar el ataque para penetrar en el campo adversario. Es la decisión del levantador la que se está tratando de anticipar.

Tabla Destino: SetterAtkPos.



Setter	Fase	Set Num	Set Moment	Server	Rec Play N...
FT-34	1	1	A1	S	3L
3T-4	1	1	A2	S	0L
3T-4	1	1	A2	S	0L
2T-4	1	1	A1	S	1L
FT-3	1	1	A1	S	0L
3T-2	1	1	A2	S	0L
2T-4	1	1	A1	S	0L
2T-2	1	1	A2	S	3L
2T-4	1	1	A2	S	0L
2T-4	1	1	A3	S	0L
2T-2	1	1	A3	S	3L
2T-4	1	1	A2	S	0L
2T-65	1	1	A2	S	0L
3T-2	1	1	A1	S	0L
FT-3	1	1	A1	S	0L

Fig. 4-10: Campos Setter y Setter Atk Pos concatenados

La figura 4-10 muestra una vista previa del resultado de la consulta efectuada. La unión de los campos Setter y Atk Pos identifica ahora la acción realizada por el levantador de manera más precisa.

Esto es un paso importante en la comprensión de los datos porque este nuevo campo será para los modelos la variable objetivo o target.

Explicación: El nuevo valor de la variable Setter indica con claridad si el levantador ha elegido una jugada de primero, segundo o tercer tiempo y a continuación la zona en la cancha desde donde se realiza la acción. Ej. FT-3 significa que ocurrió un primer tiempo por desde zona 3 de la cancha.

- Paquete CreateIndice (figura 4-11).

  Tabla Origen: SetterAtkPos

  Tabla Destino: GenerarIndice

  Este paquete ejecuta una sentencia SQL creando un campo numérico que servirá de identificador de registro. Esta condición se requiere solamente a los fines de poder aplicar algoritmos de minería de datos. En este mismo procedimiento se deja excluido del dataset el campo Setter Atk porque a través de las primeras pruebas de modelado se concluyó que es tan insignificante la cantidad de valores registrados en este campo que se podrían distorsionar los resultados hallados. Las asignaciones realizadas muestran que la columna de entrada Setter Atk no tiene columna de destino. En la columna de destino la columna IdRegistro no proviene de la tabla origen sino que ha sido recientemente creada.

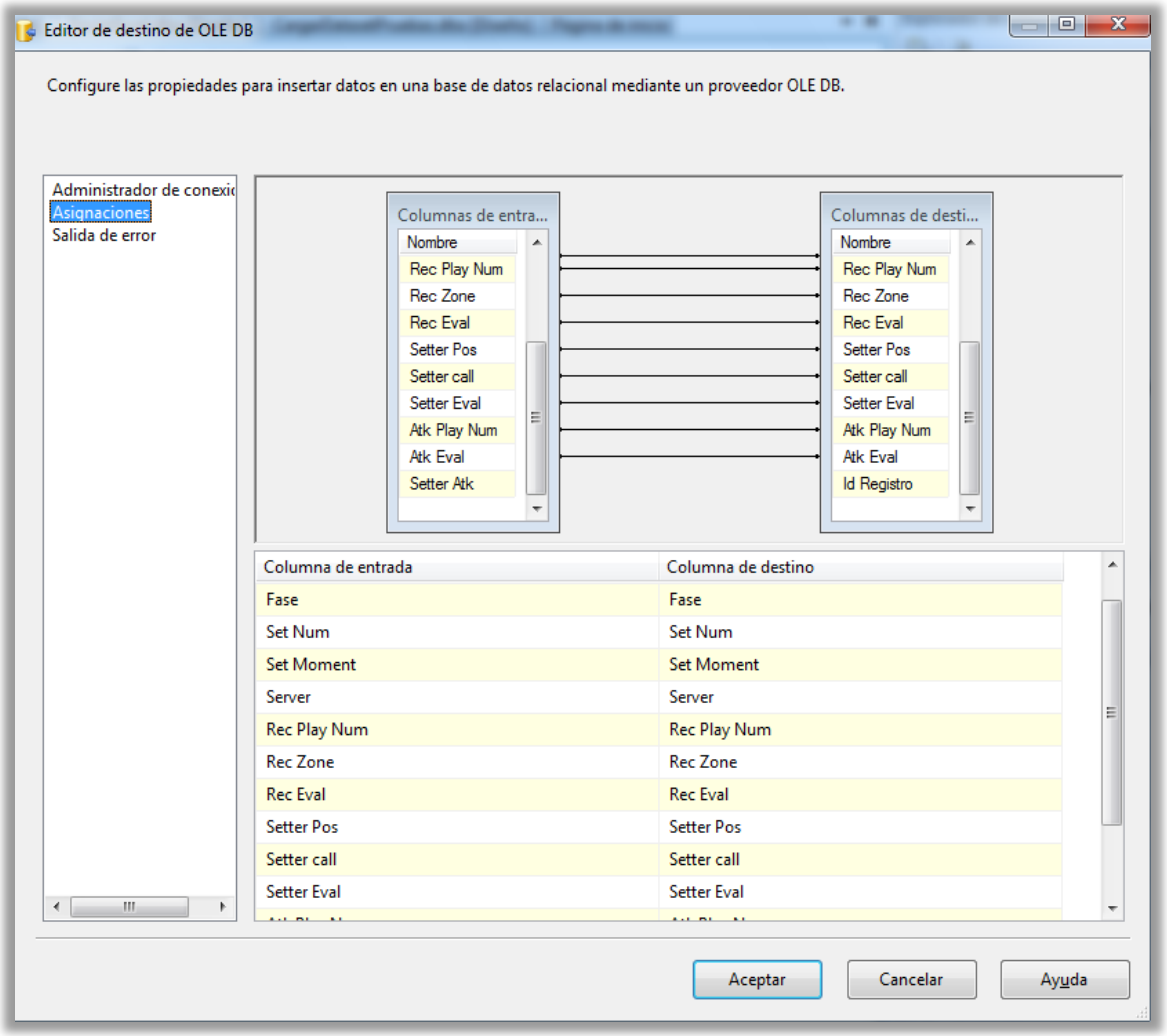


Fig.4-11: Paquete CreateIndice

- Paquete QuitarNulos (figuras 4-12 y 4-13).

  Tabla Origen: es simplemente una consulta SQL que limpia los registros eliminando aquellos cuyo campo Setter es nulo.

  Tabla Destino: DataSetMinar

  Los valores faltantes afectan la construcción del modelo. En un árbol de decisión afectan por ejemplo cómo se computan las medidas de impureza, cómo se distribuyen las instancias con valores faltantes a los nodos hijos y afectan también cómo se clasifica una instancia con valor faltante. Estos inconvenientes y otros hacen esencial el análisis de estos valores y la toma de decisión sobre lo que se debería hacer respecto a los mismos.

  Se analizó detenidamente por qué el campo Setter contenía valores nulos, cuáles eran los posibles motivos y si debían ser considerados o no en el estudio propuesto.

El campo Setter debe estar compuesto por acciones que realiza el levantador porque lo que se pretende anticipar, analizar o predecir es su comportamiento. Los valores nulos en dicho campo se deben a acciones de juego en las cuáles la evaluación de la recepción (en general negativa y sin posibilidades de esquema de juego) hizo que el levantador no haya participado en la acción de levantada. Por lo tanto, los valores nulos en el campo en cuestión no interesan en esta investigación. Cabe mencionar que a raíz de la eliminación de estos campos del dataset preparado para realizar minería de datos, se obtuvo una reducción en la cantidad de registros. Se eliminaron 98 quedando a disposición para el análisis 981. Esta información es importante para el análisis final de los resultados como así también para la manera en que se evaluará la eficiencia de los algoritmos.

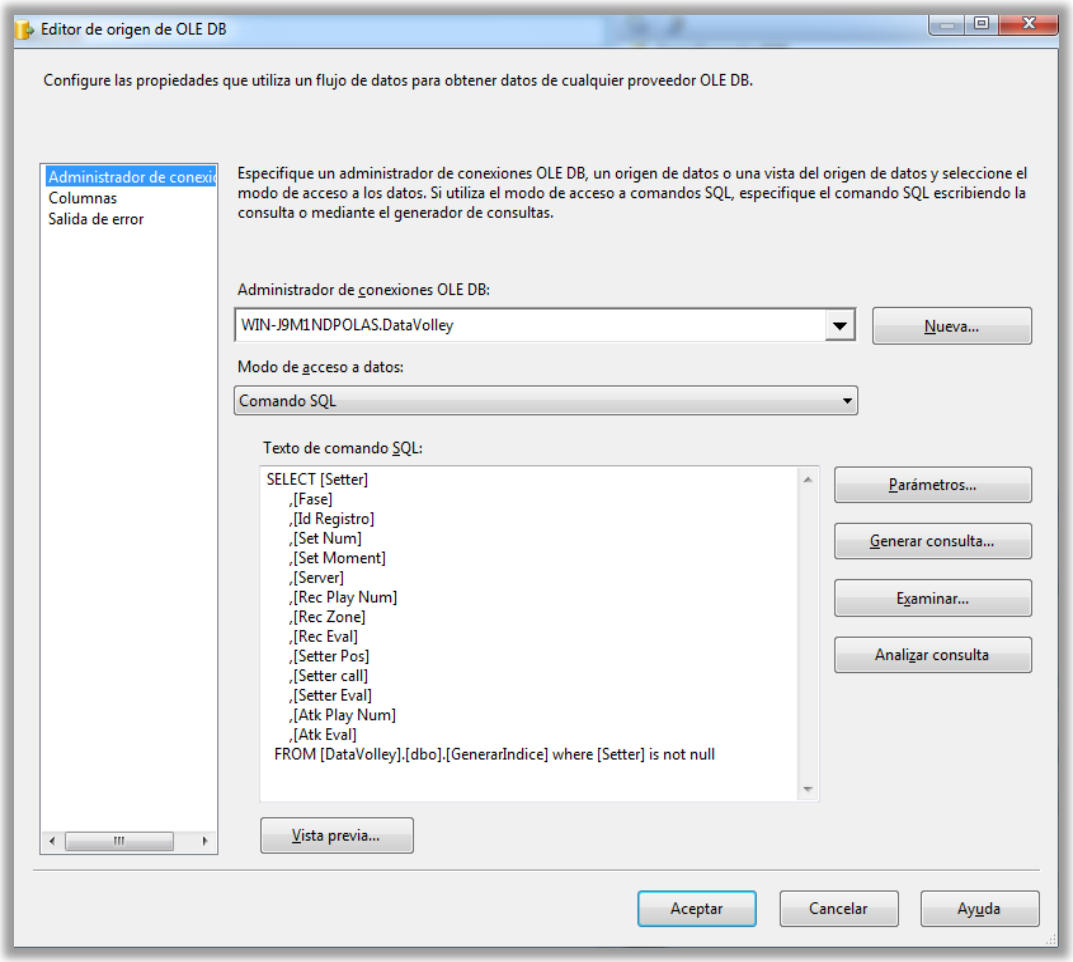


Fig. 4-12: Eliminación valores nulos

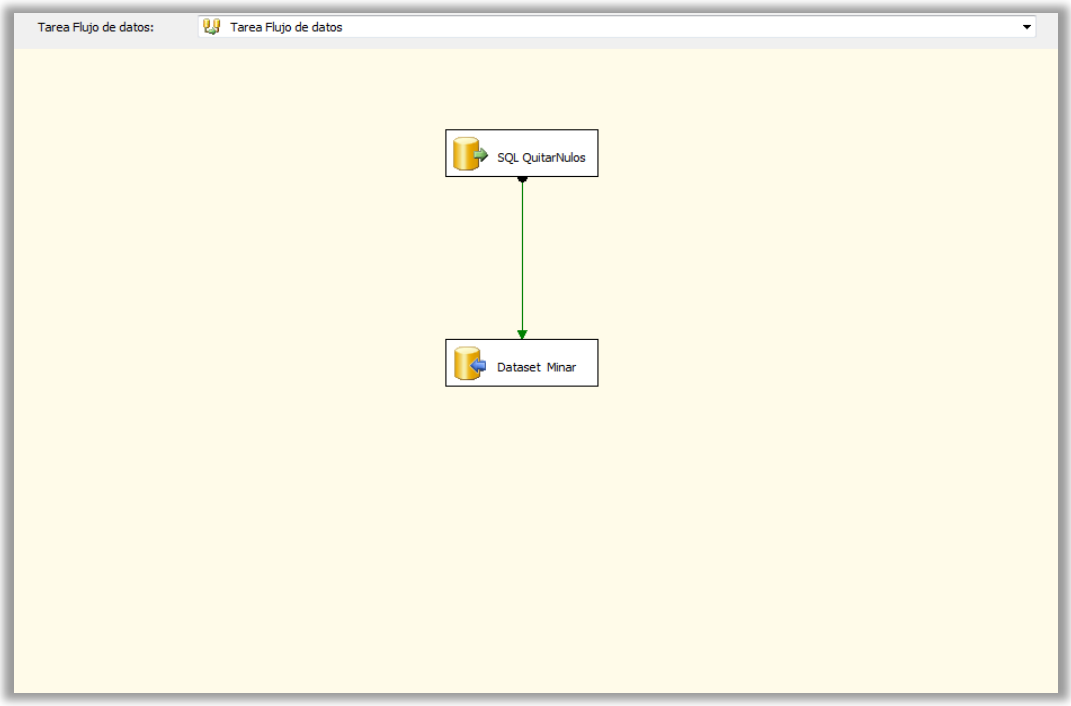


Fig. 4-13: Paquete QuitarNulos

Al finalizar esta fase se obtiene el dataset que será sometido a los algoritmos de minería de datos.

4.4 FASE DE MODELADO

La definición de los modelos está relacionada estrechamente con las tareas de minería de datos, en este caso se aplicarán métodos predictivos como red neuronal y descriptivos como árbol de decisión entre otros.

Se pretende encontrar patrones de comportamiento que reflejen la manera en que un levantador interpreta un partido. El deporte es un juego, el azar y la casualidad también contribuyen a la definición de una jugada, se deberán definir varios modelos, probando diferentes algoritmos para poder analizarlos y llegar entonces a una conclusión que pueda ser presentada ante el responsable del equipo y contribuya a la toma de decisiones. No es posible pensar que se podría hallar un modelo que responda a todas las preguntas que pueda tener un entrenador y que asegure la veracidad de las respuestas. El análisis de las conclusiones será primordial y para ello deberán presentarse herramientas (a través de modelos) que colaboren en dicha tarea.

4.4.1 Seleccionar técnica de modelado

Se estudiarán y analizarán varias técnicas de modelado con el fin de comparar las mismas, identificar patrones y arribar a conclusiones que sean útiles al entrenador del equipo.

4.4.1.1 Modelo de Microsoft Naïve Bayes

Thomas Bayes en 1763 enunció uno de los principios fundamentales que se usan hoy en las máquinas del conocimiento. Su razonamiento ha sido comparado (según Mac Lennan, Tang y Crivat, en el libro Data Mining with Ms SQL Server 2008) al comportamiento de un niño que abre los ojos viendo por primera vez el sol, al percibirlo sonríe y coloca una piedra blanca en un bolso, al día siguiente el sol no asoma, el niño escoge una piedra negra y la coloca en el bolso. Día tras día esta situación se repite y al cabo de un tiempo el pequeño se da cuenta que las piedras negras se pierden entre las blancas, entonces comprende que el sol brilla día y día y espera por él cada vez que despierta y abre los ojos. "Lo sucedido en el pasado va a pasar en el futuro"(Thomas Bayes). La ecuación matemática que enunció Bayes no solo proponía que lo sucedido podría repetirse sino que además era posible calcular lo que pasaría en el futuro a través de sus teorías probabilísticas.

Características del algoritmo

Es uno de los algoritmos de aprendizaje más prácticos junto a árboles de decisión y K-NN (K nearest neighbours).

No es robusto pero es veloz, por eso puede usarse para grandes volúmenes de datos. El algoritmo es una combinación de la probabilidad condicional e incondicional. Las reglas establecen que si existen evidencias sobre la hipótesis E entonces se puede calcular la probabilidad de la hipótesis H a través de la fórmula propuesta en el algoritmo. Trabaja únicamente con atributos discretos. Busca correlaciones entre entradas y salidas pudiendo realizar tareas de clasificación con facilidad.

Supuestos del modelo

El algoritmo no admite valores nulos en el atributo target (esto se cumple en el presente estudio debido a la ejecución del proceso de ETL denominado QuitarNulos) y las entradas deben ser razonablemente independientes entre sí. Significa que si bien la teoría exige que sean independientes en la práctica habitualmente esto no se satisface y aun así es baja la tasa de error en clasificación. Será necesario entonces comparar el comportamiento con otros algoritmos, por ejemplo árboles de decisión.

Definición de parámetros

MAXIMUN_IMPUT_ATTRIBUTES

Cantidad máxima de atributos de entrada, el máximo permitido es 255, lo cual no excede la situación presente. Se mantiene la definición por default.

MAXIMUN_OUTPUT_ATTRIBUTES

Cantidad de atributos que se considerarán para la salida, el máximo permitido es 255, lo cual no excede la situación presente. Se mantiene la definición por default.

MAXIMUN_STATES

Cardinalidad máxima de un atributo. Se mantiene la definición por default que corresponde a 100 y no excede la situación en estudio.

MINIMUN_DEPENDENCY_PROBABILITY

Indica cómo una entrada es predictiva de una salida. Este parámetro no impacta en el modelo de entrenamiento o predicción pero sí en la cantidad de salidas que se obtienen. El valor de default es 0,5, se realizaron pruebas para ver si modificando este valor aumentaba la cantidad de correlaciones encontradas pero eso no ha sucedido.

4.4.1.2 Modelo Árbol de Decisión

La tarea primordial que puede realizar un árbol de decisión es la clasificación.

Clasificar es el acto de asignar una categoría a cada caso presentado. Cada caso contiene un conjunto de atributos, uno de ellos es el atributo clase (target u objetivo). El trabajo consiste en encontrar un modelo (un árbol de decisión) que describa el atributo clase como función de los atributos de entrada.

La meta es que los registros no vistos previamente puedan ser asignados a una clase tan precisamente como sea posible.

La idea principal de los árboles de decisión es dividir los datos en subconjuntos recursivamente. Cada atributo de entrada es evaluado para ver como éste divide al atributo objetivo en subconjuntos. Cuando finaliza el proceso de recursividad el árbol queda completamente formado.

Características del algoritmo

Es la técnica de minería de datos más usada por su velocidad de ejecución en el entrenamiento de los modelos, el alto grado de precisión y la facilidad con la que se comprenden los patrones hallados. Ofrece algunas ventajas respecto a otros algoritmos, rápida construcción y fácil interpretación. Cada nodo se etiqueta en términos de los

atributos de entrada. Cada recorrido del árbol desde la raíz a través de las ramas y hasta las hojas describe una regla acerca del atributo objetivo. Logra una predicción eficiente. En las hojas el valor de predicción está basado en la estadística almacenada en cada nodo. El algoritmo de Microsoft Decisión Trees es híbrido, porque puede trabajar tanto con valores discretos como continuos. Una característica importante es que puede usarse también para análisis de asociación.

Supuestos del modelo

La inducción en la selección del mejor Split[2]puede establecerse según criterios definidos en el algoritmo a través de fórmulas matemáticas. El algoritmo de Microsoft Decision Trees permite la elección entre criterios de Entropía, Bayesian Score, Bayesiano con prioridad K2, Equivalente Dirichlet. Las particularidades de cada uno de ellos quedan fuera del alcance de este estudio, limitándose el mismo a la evaluación de la ejecución del algoritmo frente a las distintas posibilidades y la elección del criterio más adecuado según el análisis de los resultados. También es necesario establecer un límite que determine cuándo es necesario detener el proceso de splitting para evitar sobre entrenamiento (overtraining u overfitting) debido al crecimiento recursivo del árbol cuando el mismo no aporta nueva información al modelo. Los pasos de growning (crecimiento) o pruning (corte) se definen a través de parámetros en base a pruebas debidamente evaluadas. Las reglas de asociación descubiertas con el modelo se encuentran a través del recorrido del árbol por sus distintas ramas.

Definición de parámetros

COMPLEXITY_PENALTY

Define la penalidad que recibirán las ramas largas del árbol. Busca un equilibrio entre el overtraining (ramas largas) y la pérdida de patrones por el corte de dichas ramas. El valor predeterminado de 0,5 es óptimo para este caso.

MINIMUN_SUPPORT

Especifica el tamaño mínimo de cada nodo del árbol, es decir la cantidad de casos mínimos permitidos. Se verificó después de pruebas que 10 es un número aceptable.

SCORE_METHOD

Califica el Split durante el crecimiento del árbol. Se compararon las tres posibilidades: Entropía (definido en el valor 1), Bayesian Dirichlet (valor 4 por default) , Bayesian with

[2] Split es la manera en que el árbol divide las ramas en distintos niveles.

K2 Prior (valor 2). Se concluyó que el óptimo en este estudio es el método Bayesian Dirichlet.

SPLIT_METHOD

Define la manera en que se realizará el Split, si se creará un árbol binario (si el parámetro vale 1), completo (haciendo el Split sobre todas las posibilidades del atributo si el parámetro vale 2) o mixto (si se establece el parámetro en 3). En el caso en estudio se realizaron pruebas entre las tres posibilidades pero la interpretación de los resultados obliga la elección del método completo.

MAXIMUN_IMPUT_ATTRIBUTES

Define la cantidad máxima de atributos de entrada. Si el valor excede el parámetro se ejecutará una rutina interna que seleccionará los atributos a fin de optimizar la performance del algoritmo. No es necesario redefinir un valor en este caso, el valor es inferior a 255 que es el máximo permitido.

MAXIMUN_OUTPUT_ATTRIBUTES

Define la cantidad máxima de atributos de salida. Si el valor excede el parámetro se ejecutaría una rutina interna que seleccionará los atributos a fin de optimizar la performance del algoritmo. No es necesario redefinir un valor en este caso, el valor es inferior a 255 que es el máximo permitido.

FORCE_REGRESOR

Usado para atributos continuos en la selección del atributo regresor en la ecuación de regresión logística.

4.4.1.3 Modelo de Clustering

Clustering es una operación simple, natural que realizan los seres humanos al tratar con varios atributos. Volviendo al ejemplo de las rocas y la bolsa, supongamos que el niño decide de repente vaciarla bolsa y observar lo recogido. Para ello divide en primer lugar las piedras negras de las blancas, un segundo más tarde se da cuenta de que hay algunas con puntas redondeadas y otras tantas con puntas filosas, decide entonces dividirlas según esa característica. Observando más detenidamente puede distinguir entre aquellas transparentes y aquellas que no lo son, entonces comienza a preguntarse si es conveniente agrupar por color, por forma, por transparencia o elegir una cualidad y a su vez subdividirlas. Ese proceso se denomina clustering. Cuando el problema es multidimensional se vuelve complejo y allí es necesaria la ayuda del ordenador. Algunas veces a través de este proceso

se descubren variables o características ocultas. Imaginando la llegada de un avión a un aeropuerto internacional se puede inducir por ejemplo que un determinado grupo de personas se encuentran vestidas con ropa invernal al bajar de un avión porque provienen de una zona de clima frío, mientras que aquellos que bajan del mismo vuelo con ropa ligera probablemente provienen de un lugar de clima cálido. Todas estas clasificaciones y deducciones son posibles a través de algoritmos de clusterización, capaces de manejar múltiples variables para agrupar los datos de manera óptima.

Características del algoritmo

Es un algoritmo muy flexible porque soporta todo tipo de datos, la manera en que los mismos serán presentados al algoritmo puede contribuir a la solución del problema. El algoritmo puede ser usado para predecir pero más comúnmente se utiliza para detectar categorías, etiquetarlas y poder luego armar modelos dentro de un clúster seleccionado. El objetivo es encontrar grupos donde los elementos pertenecientes a ellos sean lo más similares entre sí y lo más diferentes entre los de los otros grupos. La clusterización es parte de un largo proceso de análisis. El algoritmo de Micrsosoft Clustering puede trabajar de dos maneras diferentes, utilizando el método K-means (para cada clúster se elige un punto como centroide, a partir de allí se calculan las distancias Euclideas de los demás elementos y éstos son asignados a un único clúster según dicho valor, la menor distancia determina a qué clúster pertenece cada elemento. Luego se mueve el centro al "centro"de los componentes del clúster y se vuelven a calcular las distancias. Este algoritmo es de tipo "hard clustering" porque cada elemento es asignado a uno y solo un clúster) y a través del método EM cluster-assignment (se calcula una medida probabilística para determinar qué objetos pertenecen a un determinado clúster, se tiene en cuenta la media y el desvío estándar. Los elementos se asignan según una cierta probabilidad. En este caso existe el overlap (solapamiento), es decir un punto puede pertenecer a más de un clúster con una probabilidad asignada en cada uno de ellos).

Uno de los problemas de este tipo de algoritmos es la performance y velocidad de ejecución en las múltiples iteraciones que debe realizar, Microsoft Clustering provee un framework escalable para optimizar ese proceso.

Supuestos del modelo

El algoritmo no tiene en cuenta valores nulos. Si los encuentra los descarta y no los coloca en ningún clúster. La convergencia del modelo estará determinada por el valor de ciertos

parámetros, la selección de los mismos deberá ser el resultado de pruebas y análisis hasta obtener el modelo que más se ajuste a las necesidades.

Definición de parámetros

CLUSTERING_METHOD

Selecciona el tipo de algoritmo a utilizar, EM o K-means, ambos con la posibilidad de que sean escalables si la cantidad de datos es elevada. Se eligió K-means (opción 4) tras realizar suficientes pruebas.

CLUSTER_COUNT

Determina la cantidad de clústeres a formar. Si no existe un número se calculan heurísticamente. Cuando el número de atributos es alto se eleva la cantidad de clústeres a encontrar. Después de varias pruebas y análisis respectivos se consideró que 5 era una cantidad óptima para el presente caso.

MINUMUN_CLUSTER_CASES

Define la cantidad de elementos para el cual el clúster no se considera vacío. Si el número es elevado se podría llegar a resultados incorrectos.

MODELLING_CARDINALITY

Permite mejorar la ejecución del modelo controlando los hilos de ejecución disponibles. No es aplicable en este caso debido a la reducida cantidad de registros disponibles.

STOPPING_TOLERANCE

Usado para determinar la convergencia del modelo. Cuando se trabaja con pocos datos el número no debe ser elevado. En este caso se estableció en 5 después de pruebas y análisis de los modelos obtenidos.

SAMPLE_SIZE

Cuando se utiliza un modelo escalable, este parámetro indica el número de casos en cada paso. El método seleccionado no es escalable.

CLUSTER_SEED

Es un número random usado para inicializar los clústeres. Si al modificar este parámetro el modelo permanece estable, significa que es el modelo correcto. Se han hecho pruebas modificando este valor y se obtuvieron probabilidades similares en los clústeres obtenidos.

MAXIMUN_IMPUT_ATTRIBUTES

Si la cantidad máxima de atributos excede este número el algoritmo invoca el proceso de selección automática donde se elegirán los atributos que más se repiten. Los no elegidos

quedarán fuera del proceso de clusterización. No sucede esta particularidad en el caso en estudio.

MAXIMUN_STATES

Controla la cantidad máxima de estados que puede tener un atributo. Si se excede este límite se agrega la categoría otros que agrupa los estados excedidos. Queda inalterado el valor por default, que corresponde al número 100.

4.4.1.4 Modelo de Reglas de Asociación

Las acciones que realiza una persona, cuando por ejemplo va al supermercado pueden seguir una secuencia de patrones que describan su comportamiento. Se puede observar por ejemplo si las personas que compran el artículo A también compran siempre el artículo B y a su vez las que compran A y B compran o no siempre C.

Encontrar reglas de asociación puede definirse entonces de la siguiente manera: "Dado un conjunto de transacciones, hallar reglas que permitan predecir la ocurrencia de un ítem basado en la ocurrencia de otros ítems en la transacción".

El objetivo de esta tarea es encontrar reglas que gobiernan la distribución de los datos. Se aprenden patrones a través de las transacciones ocurridas.

Características del algoritmo

El algoritmo genera una red entre los estados de los atributos, esta particularidad lo hace diferente a otros algoritmos que encuentran relaciones entre atributos sin importar el estado de los mismos.

El algoritmo no es más que un motor que cuenta la correlación entre atributos.

La ejecución del algoritmo se realiza en dos pasos:

El primero es una fase intensa de cálculo donde el objetivo es encontrar "itemsets" (conjunto de ítems, cada uno formado por el valor de un atributo). Es el corazón del algoritmo.

Cada itemset posee un tamaño según la cantidad de ítems que lo componen y se caracteriza por las siguientes medidas:

- Soporte: es la medida de la popularidad de un itemset. Cuenta la cantidad de ocurrencias que el conjunto de ítems se encuentra en el total de los datos.
- Confianza o Probabilidad: es la probabilidad de ocurrencia de una regla calculada usando el soporte del itemset. Por ejemplo en un itemset formado por los ítems A y

B el soporte de (A, B) dividido el soporte de A me da la confianza para el itemset (A, B).

- Lift o importancia: indica si los itemset son dependientes o independientes. Un valor igual a cero indica que no hay asociación, un valor mayor que uno significa que existe asociación y que siempre que ocurre A seguramente ocurrirá B y un valor menor a uno indica exactamente lo contrario. El cálculo matemático se realiza utilizando el logaritmo de la probabilidad de (B/A) dividido la probabilidad de (B/not A).

El segundo es un paso que genera reglas de asociación basadas en la frecuencia de los itemsets. Estafase consume mucho menos tiempo de ejecución respecto la primera. Si se define un atributo de predicción el mismo aparecerá siempre del lado derecho de la regla, mientras que los atributos de entrada conformarán el lado izquierdo.

El algoritmo de Microsoft Association Rules pertenece a la familia de algoritmos de asociación a priori porque itera en base a la longitud de los itemsets encontrados, iniciando por los de longitud 1 e incrementando la misma hasta no encontrar más conjuntos que se repitan.

Supuestos del modelo

No acepta atributos continuos porque el motor cuenta la cantidad de correlaciones entre atributos discretos. Si bien puede usarse para predecir no es un excelente predictor comparado a otros algoritmos de predicción pero es muy útil en la comparación con otros algoritmos.

Definición de parámetros

Particularmente este algoritmo es muy sensible a la definición del valor de los parámetros.

MINUMUN_SUPPPORT

Es el soporte mínimo (cantidad de ocurrencias) a partir del cual se considerarán útiles las reglas encontradas por el algoritmo. En el presente estudio y luego de pruebas y comparaciones se estableció el valor en 10. Este número indica que para que una regla sea considerada por el algoritmo deberá ocurrir al menos 10 veces en el total de datos.

MAXIMUN_SUPPORT

Indica la cantidad máxima de soporte permitido. Puede expresarse como un número entero o como un porcentaje del total en un rango de 0 a 1. Se definió el valor 1 como valor apropiado.

MINIMUN_PROBABILITY

Define la probabilidad para una regla de asociación hallada. El valor de default es 0,4. Las pruebas realizadas modificando este valor no fueron satisfactorias.

MINUMUN_IMPORTANCE

Filtra las reglas cuya importancia mínima sea inferior a este valor. No se declaró un valor con el fin de poder observar el total de reglas detectadas.

MAXIMUN_ITEMSET_SIZE

Define el tamaño máximo para tener en cuenta un itemset. Afecta el tiempo de ejecución del algoritmo. Cuando el volumen de datos es importante se acota el tamaño del itemset para mejorar la performance. La cantidad reducida de datos hace que este parámetro no sea relevante.

MINIMUN_ITEMSET_SIZE

Se usa cuando es necesario considerar solamente aquellos itemset que contengan una cantidad mínima de ítems. Teniendo en cuenta el acotado conjunto de datos con el que se trabaja no fue necesario redefinirlo.

MAXIMUN_ITEMSET_COUNT

Detiene el algoritmo una vez hallada la cantidad máxima de itemsets definidas por el parámetro. Teniendo en cuenta el acotado conjunto de datos con el que se trabaja no fue necesario redefinirlo.

OPTIMIZED_PREDICTION_COUNT

Optimiza la cantidad máxima de predicciones que al menos debería encontrar el algoritmo en una query. Un valor igual a cero reportará todas las predicciones posibles. No fue oportuno modificarlo.

AUTODETECT_MINIMUN_SUPPORT

Cuando el valor es 1.0 el algoritmo detectará el valor mínimo apropiado para el soporte cuando es 0.0 se utilizará el soporte mínimo definido como parámetro. Se definieron los valores manualmente, no se utilizó esta característica.

4.4.1.5 Modelo de Red Neuronal

Es un modelo matemático óptimo para mejorar o entender las relaciones complejas entre entradas y salidas. El algoritmo combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente estas probabilidades pueden usarse para la

regresión, la clasificación o para predecir el resultado de un atributo de predicción basándose en los atributos de entrada.

El origen del algoritmo se remonta a 1940 con los investigadores Warren McCulloch y Walter Pits intentando encontrar un modelo que simule una red neuronal biológica, centrándose sus trabajos en el comportamiento del cerebro humano. Posteriormente este enfoque sirvió a resolver problemas técnicos fuera del alcance de la biología. En los años 80 maduraron las teorías de las redes neuronales, los ordenadores fueron cada vez más potentes y permitieron que 1982 John Hopfield defina el método de backpropagation o propagación hacia atrás (se ajustan los pesos en la red y se propagan hacia atrás según el cálculo del error).

Un algoritmo de red neuronal puede realizar trabajos de clasificación y regresión. Encuentra relaciones no lineales. Puede ser considerado un instrumento de aprendizaje más sofisticado que los algoritmos de Árboles de Decisión y de Bayes Naïve por su mayor complejidad pero tiene la desventaja del alto tiempo de ejecución que puede requerir y de la dificultad en la interpretación de los resultados.

Se usa cuando el número de datos que se dispone es elevado o también como elemento de comparación y evaluación de otros modelos. Existen distintos tipos de redes neuronales, Microsoft en su algoritmo utiliza el tipo feed forward, es decir inicia con una cantidad de entradas similar a la cantidad de neuronas de la primera capa y la transmisión es hacia adelante.

Características del algoritmo

Nodos o neuronas de entrada: forman la primera capa de la red, cada nodo mapea un atributo (un único estado de un atributo cuando se trata de valores discretos). Los valores deben ser normalizados en una misma escala para que las comparaciones sean posibles.

Nodos o neuronas ocultas: se encuentran en la capa intermedia. Reciben entradas de las neuronas de entrada y proporcionan salidas a las neuronas de salida. En esta capa oculta se asignan pesos a las distintas probabilidades de las entradas. Un peso describe la relevancia o importancia de una entrada determinada para la neurona oculta. Cuanto mayor sea el peso asignado a una entrada, más importante será el valor de dicha entrada. Los pesos pueden ser negativos, lo que significa que la entrada puede desactivar, en lugar de activar, un resultado concreto. Estas neuronas usan la función tangente hiperbólica (tanh) para la función de activación (que describe la relevancia o importancia de una neurona).

Nodos o neuronas de salida: representan valores del atributo de predicción para el modelo de minería de datos. Para atributos discretos una neurona de salida representa un único estado del atributo target.

Combinación y activación: cada neurona de la red es una unidad de procesamiento básico. Para combinar las entradas existen diferentes métodos, el algoritmo de Microsoft Neural Network utiliza una aproximación de los pesos ponderados. Como función de activación en la capa oculta se utiliza la función tangente hiperbólica (tanh) y en las neuronas de salida la función sigmoidea.

Backpropagation: es el corazón del proceso. Al iniciar el algoritmo se asignan los pesos de manera aleatoria a los diferentes nodos, luego el algoritmo calcula las salidas. A continuación se calcula el error para cada salida y neurona de la capa oculta. De esta manera se actualizan los pesos de la red. Este proceso se repite hasta que exista la condición de finalización. La cantidad de iteraciones depende de los datos y de los patrones encontrados. El algoritmo se puede detener porque ha llegado a la cantidad máxima de iteraciones permitidas, debido a la convergencia en cuanto a los pesos porque caen debajo de un umbral permitido o porque el error de exclusión está por debajo del enunciado en la definición de parámetros. Microsoft Neural Network realiza la actualización de los pesos en un proceso batch (epoch) porque es un método robusto, bueno para los modelos de regresión.

Para calcular el error, el algoritmo usa para atributos discretos el método "cross-entropy".

El algoritmo utiliza también un método denominado "conjugate gradient" en el proceso de ajuste de pesos antes de cada iteración. De esta manera logra encontrar la dirección hacia donde debe continuar.

Topología de la red: la topología debe fijarse antes de iniciar el proceso. La cantidad de capas puede generar overtraining y está relacionada a la performance del algoritmo. Debido a estudios que garantizan un óptimo funcionamiento el algoritmo, Microsoft Neural Network no permite más de una capa oculta.

Supuestos del modelo

El algoritmo puede predecir tanto atributos discretos como continuos.

Todas las entradas pueden estar relacionadas a alguna o a todas las salidas y la red considera estas relaciones en el entrenamiento. Las diferentes combinaciones de entrada pueden estar relacionadas diferentemente con las salidas.

Las relaciones detectadas por el algoritmo de Microsoft Neural Network pueden tratarse de dos maneras. En un nivel simple, en cuyo caso particular se trata de una Regresión Logística o en dos niveles cuando se define una capa oculta y las entradas no pasan directamente a la salida como en el caso anterior sino que existe esa capa intermedia.

Debe contener por lo menos una columna de entrada y una columna de salida.

Definición de parámetros

MAXIMUN_IMPUT_ATTRIBUTES

Determina el número máximo de atributos de entrada que se pueden proporcionar al algoritmo antes de emplear la función de selección de características (que elige los más significativos). Esta función se deshabilita cuando el valor se establece en 0. El valor predeterminado es 255. Para el presente estudio no es necesario modificarlo.

MAXIMUN_OUTPUT_ATTRIBUTES

Determina el número máximo de atributos de salida que se pueden proporcionar al algoritmo antes de emplear la función de selección de características (que elige los más significativos). Esta función se deshabilita cuando el valor se establece en 0. El valor predeterminado es 255. Para el presente estudio no es necesario modificarlo.

MAXIMUM_STATES

Especifica el número máximo de estados discretos por atributo que admite el algoritmo. Si para un determinado atributo dicho número es mayor que el número especificado para este parámetro, el algoritmo utiliza los estados más frecuentes de este atributo y trata al resto como estados que faltan. No debe modificarse en este caso el valor 100 que es el que se usa por default.

HOLDOUT_PERCENTAGE

Especifica el porcentaje de escenarios de los datos de entrenamiento utilizados para calcular el error de exclusión, que se utiliza como parte de los criterios de detención durante el entrenamiento del modelo de minería de datos. Se deja el valor por default que corresponde al 30%.

HOLDOUT_SEED

Especifica el número que se utiliza para inicializar el generador pseudoaleatorio cuando el algoritmo determina aleatoriamente los datos de exclusión. Si este parámetro se establece en 0, el algoritmo genera la inicialización basada en el nombre del modelo de minería de datos, para garantizar que el contenido del modelo permanece intacto al volver a realizar el

proceso. Se realizaron pruebas modificando este valor pero no se obtuvieron resultados interesantes, se decide dejarlo en cero.

HIDDEN_NODE_RATIO

Especifica la proporción entre neuronas ocultas y neuronas de entrada y de salida. La fórmula *Total neuronas de entrada * total neuronas de salida* determina el número inicial de la capa oculta. El valor predeterminado es 4. Este parámetro no está disponible en el algoritmo de Regresión Logística. Si se construye un modelo con este valor definido en 0 se obtendrá exactamente el mismo resultado que en la ejecución del algoritmo de Regresión Logística. Para este estudio se deja el valor por default.

SAMPLE_SIZE

Especifica el número de escenarios que se van a utilizar para realizar el entrenamiento del modelo. El algoritmo utiliza el valor menor entre este número o el porcentaje del total de escenarios que no están incluidos en los datos de exclusión, según especifica el parámetro HOLDOUT_PERCENTAGE. Se mantiene el valor por default.

4.4.2 Generar Plan de Pruebas

Los criterios fundamentales para evaluar los modelos enunciados son: la precisión (indica hasta qué punto el modelo pone en correlación un resultado con los atributos de los datos que se han proporcionado), la confiabilidad (evalúa la manera en que se comporta un modelo de minería de datos en conjuntos de datos diferentes) y la utilidad (indica si el modelo proporciona información útil).

Dentro de las técnicas disponibles en Analysis Services, considerando el dominio del problema, el tipo de atributos con los que se trabaja y la cantidad de datos disponibles se opta por realizar las siguientes pruebas:

a) Medir la mejora del modelo respecto al modelo predictivo: se realiza utilizando un gráfico de elevación también llamado en inglés lift chart que representa los resultados de las consultas de predicción de un conjunto de datos de prueba en función de valores conocidos de la columna de predicción que existe en el conjunto de datos. El mismo muestra los resultados del modelo predictivo encontrado, los resultados de una previsión aleatoria y los que generaría un modelo ideal. El gráfico mide el cambio en términos de puntuación respecto al modelo predictivo.

   El modelo encontrado será más efectivo a medida que se aleja por encima del modelo aleatorio. Al comparar las puntuaciones de varios modelos se podrá determinar cuál de

todos es el mejor. En este caso particular se podrá verificar qué tanto mejora una predicción respecto a un modelo aleatorio.

b) Matriz de clasificación (Classification Matrix): es una manera de examinar la precisión del modelo. Es una herramienta valiosa porque no solo muestra la frecuencia con que el modelo predice un valor correctamente, sino que también muestra qué valores predice incorrectamente.

c) Validación cruzada (Cross Validation): sirve para comparar la eficiencia de los algoritmos y la exactitud con la cual pueden interpretarse los datos. Estas pruebas emplean el total de los datos disponibles y utilizados durante el entrenamiento del modelo y los dividen en particiones de tamaño similar. Es fundamentalmente valiosa cuando no se dispone de un alto volumen de datos porque la validación se realiza sobre el mismo conjunto de datos sobre el que se generó el modelo. La técnica consiste en construir un modelo para cada una de las particiones, entonces se validan los mismos contra la partición seleccionada. Se obtienen tantos resultados como particiones se hayan establecido y se comparan los resultados. La similitud entre los resultados obtenidos para las particiones generadas es un indicador de que el modelo trabaja correctamente. Sql Srever 2008crea un clon del modelo para cada partición manteniendo los mismos parámetros del algoritmo, por lo tanto el mismo no debe ser entrenado. Es posible seleccionar el conjunto de datos de la estructura de minería de datos sobre la cual se ejecutará el modelo. De esta manera esta técnica sirve para comparar cuál es el modelo que mejor está trabajando sin tener que entrenar el mismo con el consecuente consumo de tiempo y de recursos.

4.4.3 Construir el modelo

Los modelos se construyen todos sobre la misma estructura de datos configurada a partir del dataset obtenido en el proceso de ETL en la tabla denominada DatasetMinar. Todo se almacena y ejecuta utilizando el motor de Analysis Services.

La figura 4-14 muestra la definición de la estructura de datos y los campos seleccionados para formar parte de la misma. No significa que todos los modelos deban trabajar con todos los campos pertenecientes a dicha estructura. La selección de los mismos puede variar en función del tipo de algoritmo a utilizar. La figura 4-15 muestra los tipos de datos de las columnas de la estructura.

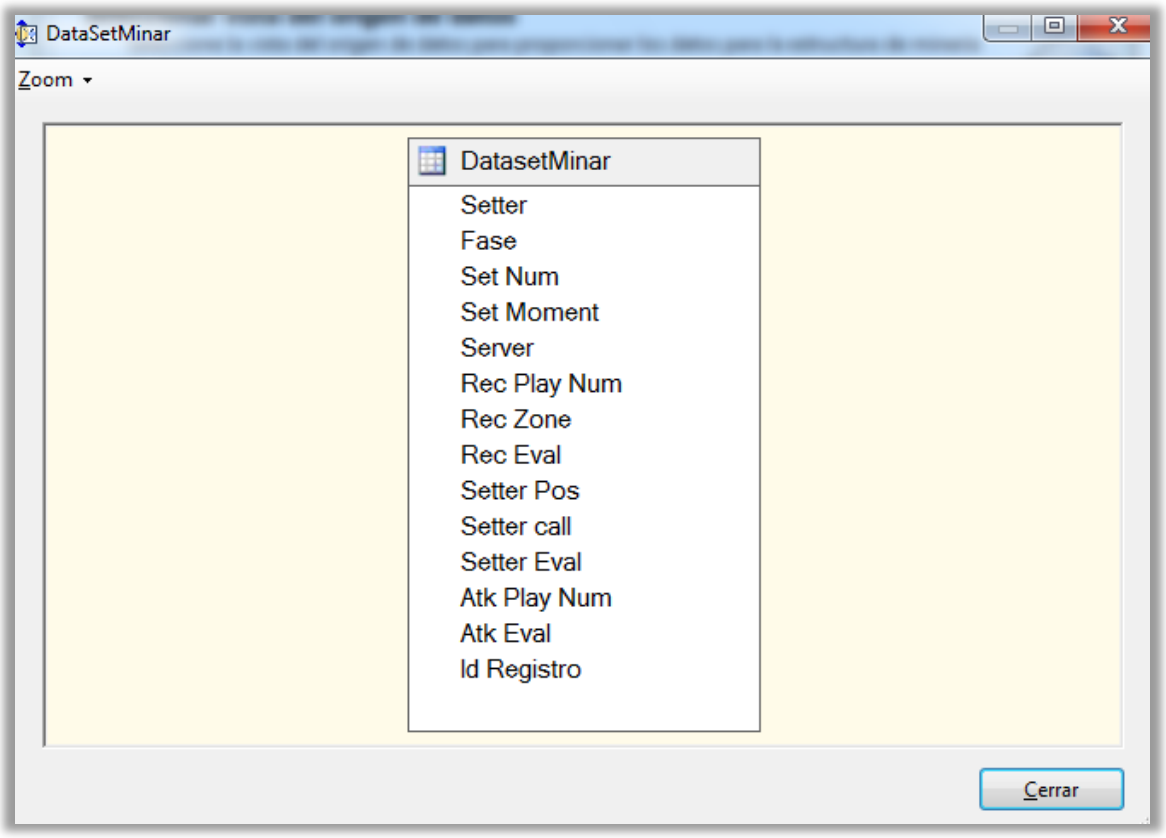


Fig.4-14: Estructura Morales

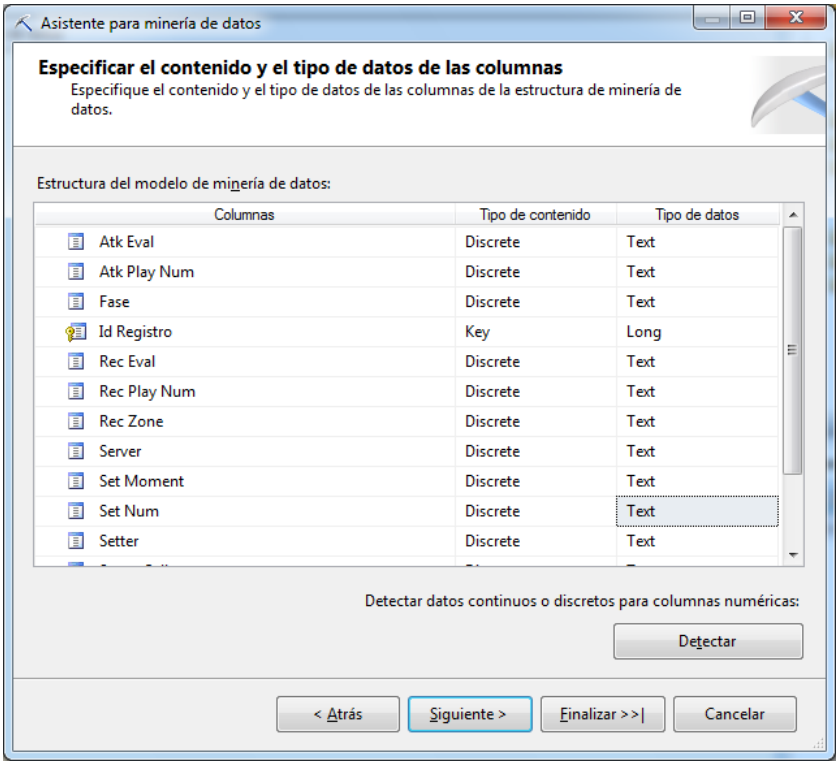


Fig. 4-15: Tipos de Datos "Estructura Morales"

4.4.3.1 Construcción de Bayes Naïve

Parámetros

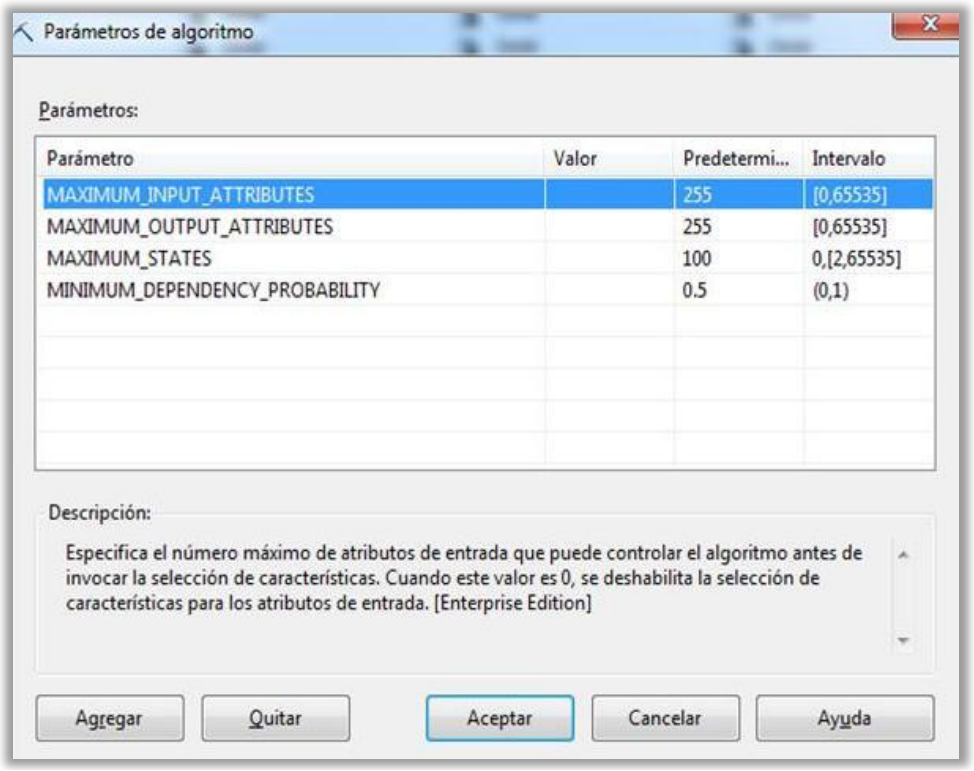


Fig. 4-16: Parámetros Bayes Naïve

Exploración del Modelo

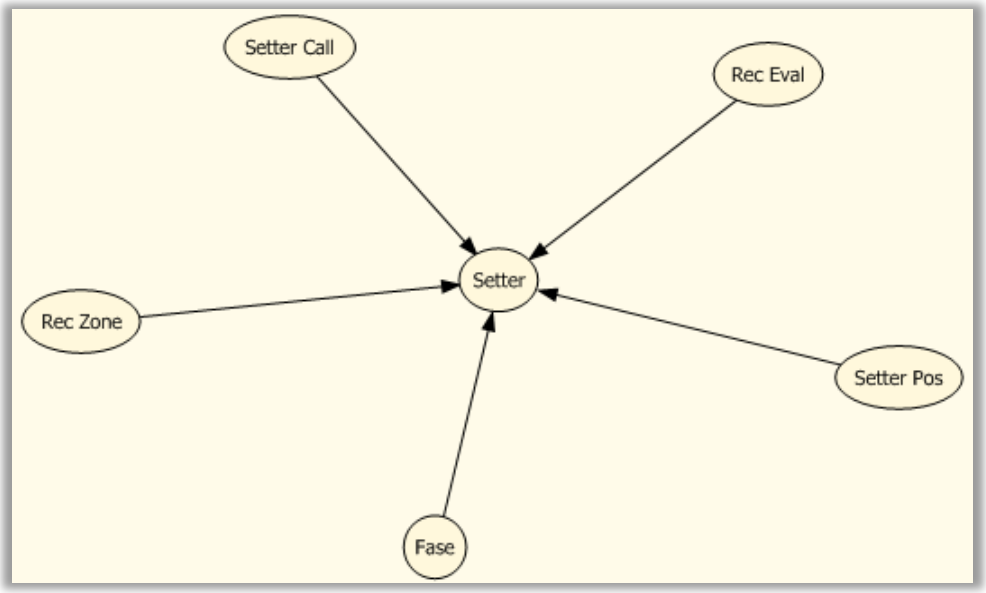


Fig. 4-17: Red de dependencias Bayes Naïve

Se observa que el nodo seleccionado Setter es predecido por Rec Eval, Setter Pos, Setter Call, Fase y Rec Zone. El nodo Set Moment no aparece en el esquema.

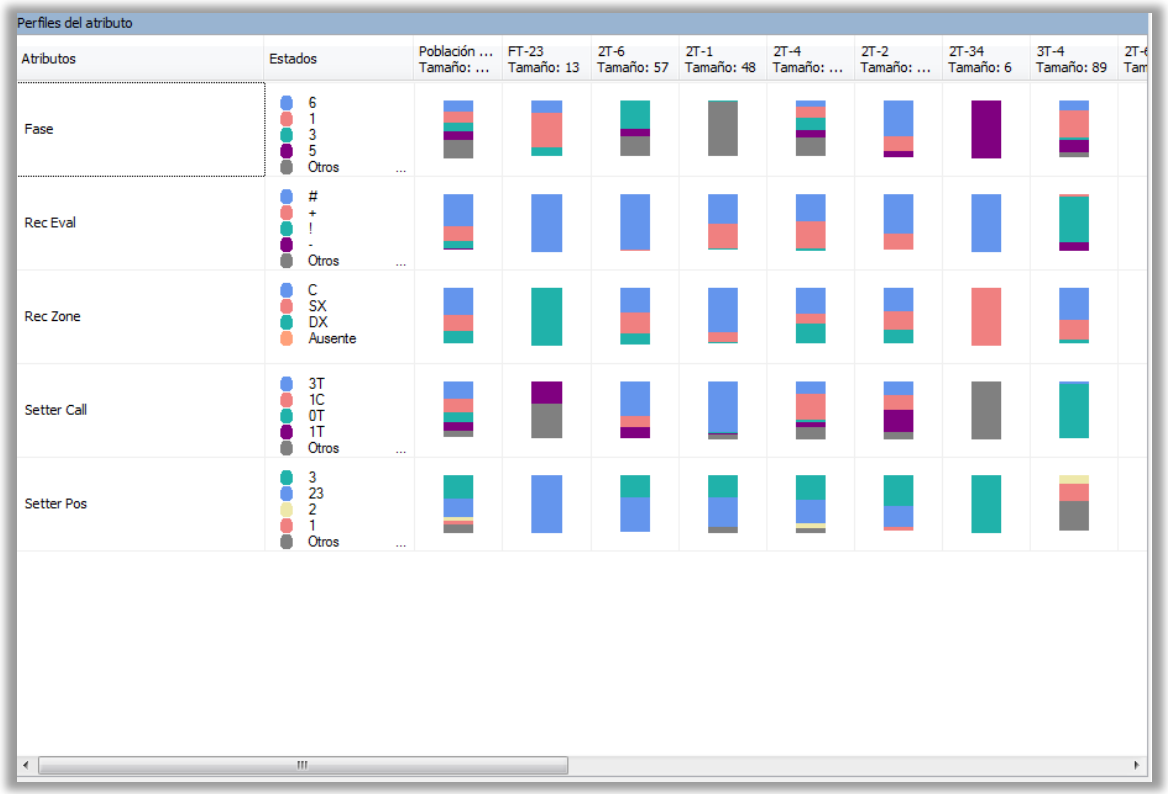


Fig. 4-18: Perfiles del Atributo Bayes Naïve

En la figura 4-18 se observa que para cada valor del atributo Setter (columnas) se obtiene una cantidad o "mezcla" de cada uno de los valores de los atributos de entrada. La tabla 4.13 muestra la caracterización para el valor del atributo Setter en 2T-2.

Atributos	Valores	Probabilidad
Rec Eval	#	69,60%
Fase	6	61,77%
Setter Pos	3	53,92%
Rec Zone	C	42,16%
Setter Call	1T	40,20%
Setter Pos	23	38,24%
Rec Zone	SX	33,33%
Rec Eval	+	29,41%

Setter Call	1C	26,47%
Fase	1	26,47%
Rec Zone	DX	24,51%
Setter Call	3T	24,51%
Fase	5	11,77%
Setter Call	A	7,843%
Setter Pos	1	7,84%
Setter Call	0T	0,98%
Rec Eval	!	0,98%

Tabla 4.13: Características del atributo 2T-2 Bayes Naïve

Esta tabla muestra para el valor del atributo Setter 2T-2 la probabilidad con que cada valor de los atributos de entrada caracteriza al objetivo. Los valores están ordenados de manera decreciente según el valor del campo probabilidad. Se desprecian los inferiores a 0,98% por considerarse insignificantes. Es de notar que si bien la lectura de estos valores podrá ayudar a identificar patrones, los valores que no se encuentran aquí también servirán a extraer conclusiones. Por ejemplo se observa en la tabla 4.14 que las fases 2, 3 y 4 no caracterizan el objetivo Setter 2T-2.

Atributos	Valores	Favorece 2T-2	Favorece todos los otros estados
Fase	6	100,00	
Setter Call	1T	42,64	
Fase	3		31,71
Fase	2		27,50
Setter Call	0T		26,92
Fase	4		23,07

Rec Eval	!		20,63
Setter Pos	2		11,24
Rec Eval	#		6,30
Setter Pos	6	4,86	
Setter Pos	3	3,25	

Tabla 4.14: Distinción del atributo 2T-2 Bayes Naïve

La tabla 4.14 muestra la diferencia entre un estado de un atributo de entrada y los otros, específicamente para el caso del atributo objetivo Setter 2T-2. Los valores reflejados son un número que indica una puntuación o score del 1 al 100. Por ejemplo se lee que la Fase 6 favorece Setter 2T-2, esto no implica que las demás fases favorezcan aquellos valores diferentes a Setter 2T-2, sino simplemente indica que ese valor del atributo Fase favorece el objetivo Setter 2T-2 con una puntuación de 100.

Por otro lado se destaca que las Fases 2, 3 y 4 favorecen los demás valores del atributo Setter, con puntuaciones de 31,71, 27,50 y 23,07 respectivamente. Esto concuerda plenamente con la deducción antes mencionada respecto a la ausencia de algunos valores de los atributos de entrada en la descripción del atributo Setter 2T-2.

4.4.3.2 Construcción del Árbol de Decisión

Parámetros

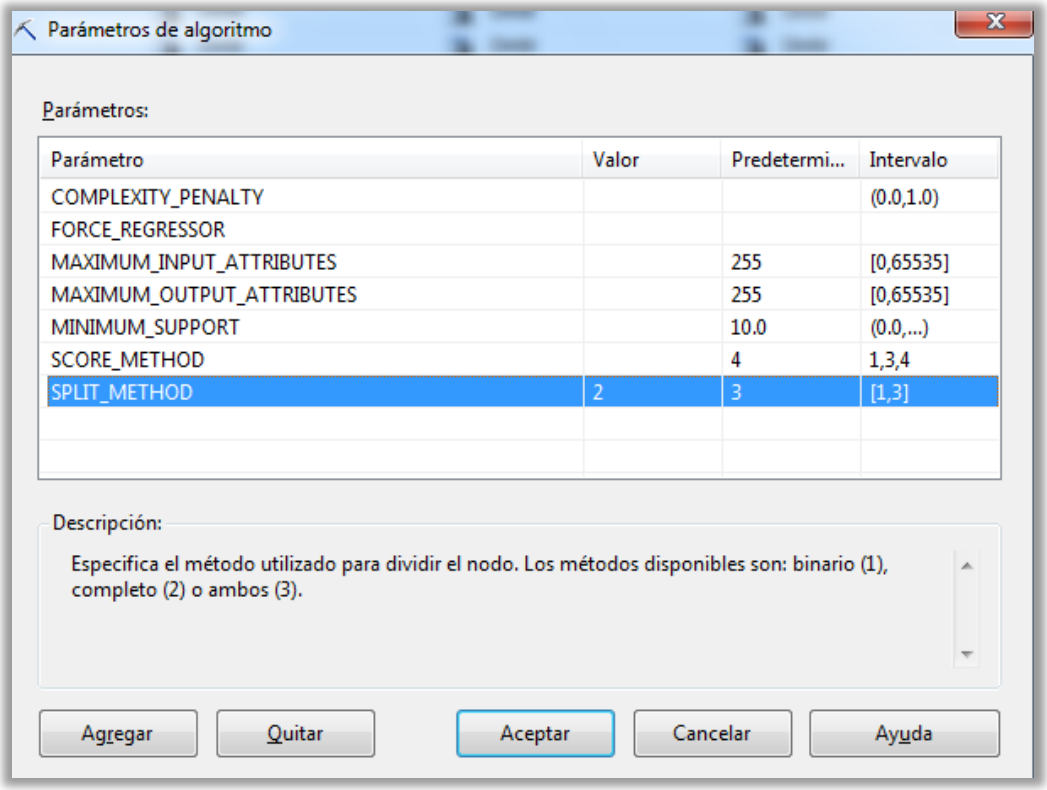


Fig. 4-19: Parámetros Árbol de Decisión

Exploración del modelo

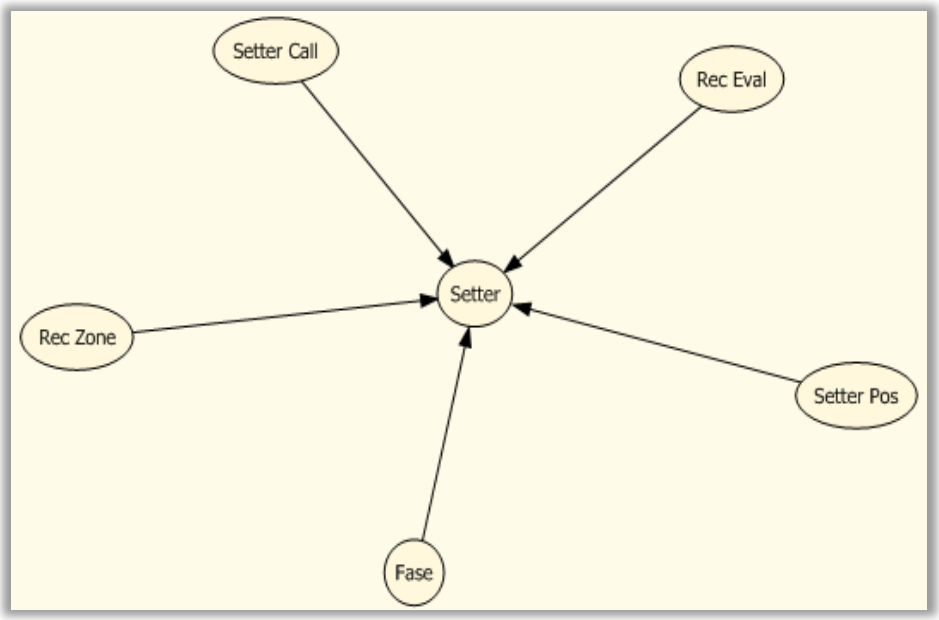


Fig. 4-20: Red de dependencias Árbol de Decisión

La figura 4-20 muestra, al igual que en el modelo Bayes Naïve que el nodo Setter es predicho por los atributos Setter Pos, Rec Eval, Setter Call, Rec Zone y fase.

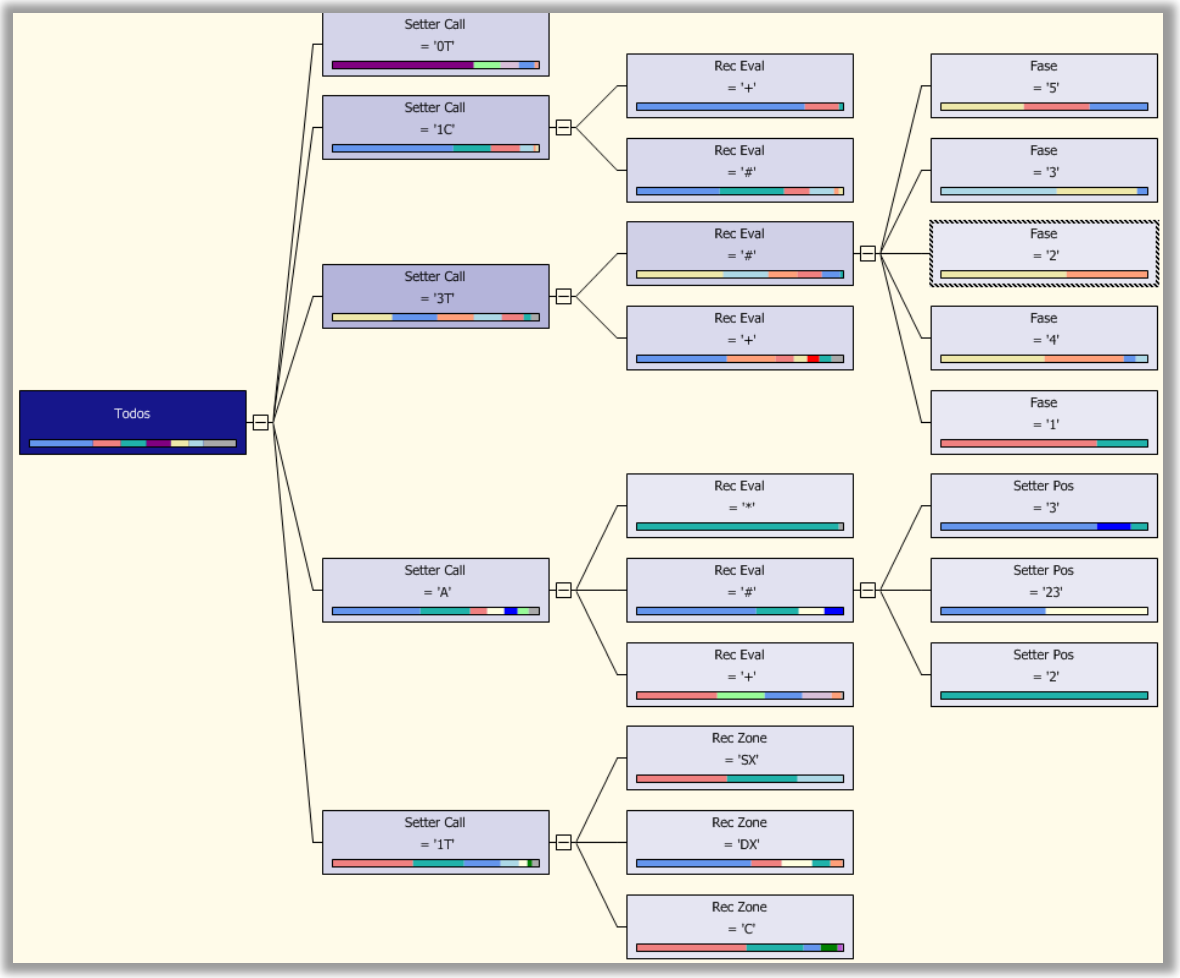


Fig. 4-21: Esquema de visualización Árbol de decisión

La figura 4-21 muestra que se han logrado ramas cuya pureza es importante, es decir las hojas de algunos recorridos del árbol logran predecir aparentemente bien un valor del atributo target.

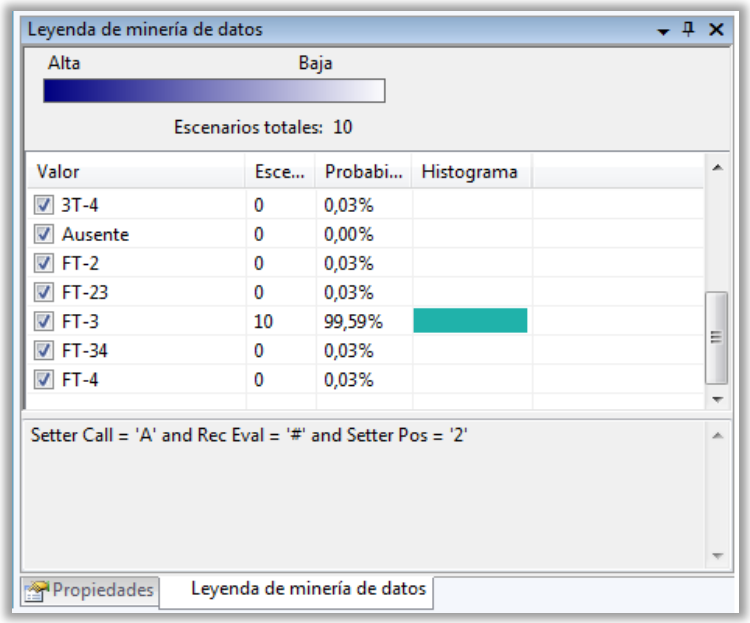


Fig. 4-22: Leyenda visualización Árbol de Decisión

La figura 4-22 describe una regla de asociación que concluye con un 99,59% de probabilidad de que en el caso de que suceda esa secuencia de eventos, el levantador elegirá la opción FT-3.

4.4.3.3 Armado de Clusteres

Parámetros

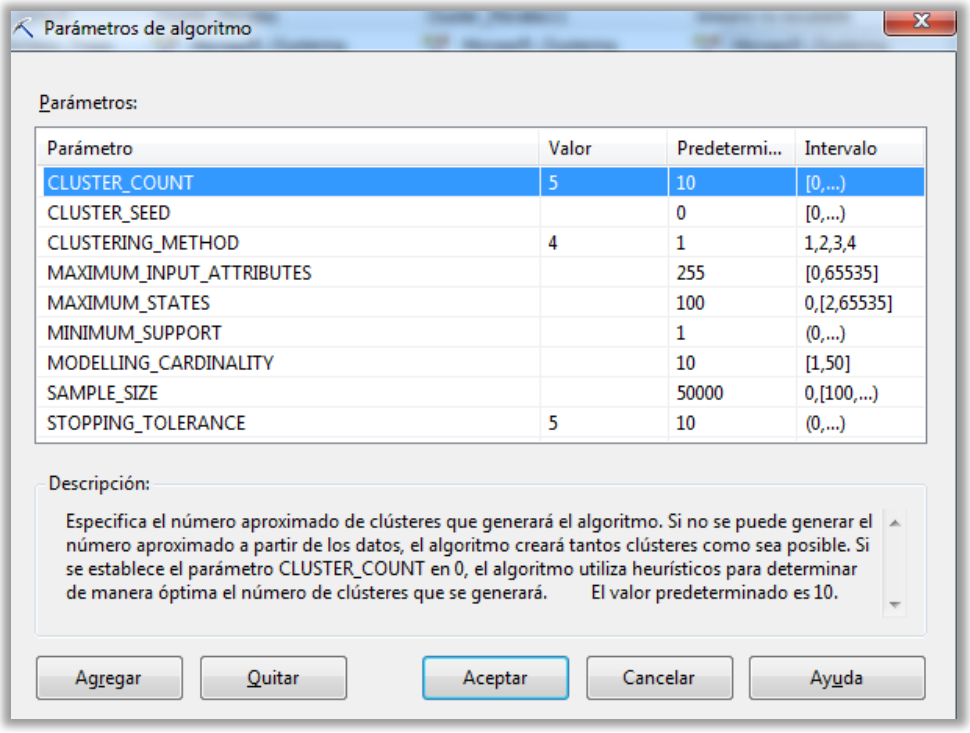


Fig. 4-23: Parámetros Clusterización

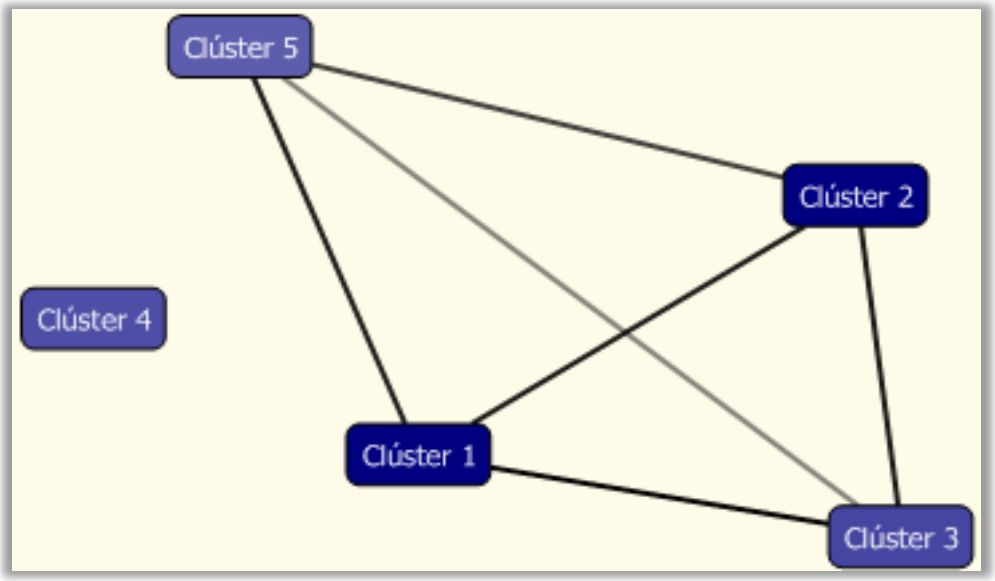


Fig. 4-24: Diagrama de Clústeres

En la figura 4-24 se observa que el clúster 4 es diferente (no se vincula) a los otros cuatro. Seguramente en él se agruparán casos dispares al resto de los clústeres o con alguna particularidad que lo hace disímil a los demás.

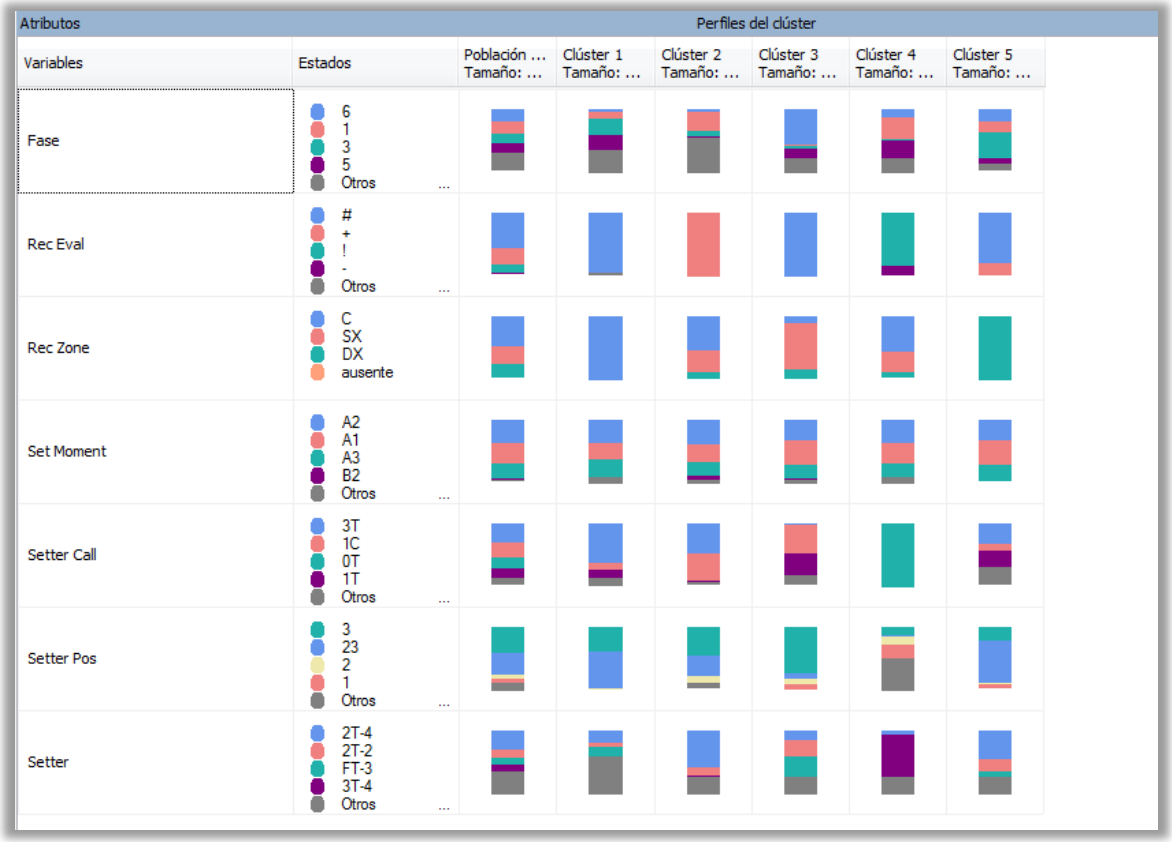


Fig. 4-25: Visualización Perfiles del Clúster

En la figura 4-25 aparecen los 5 clústeres obtenidos con las características de cada uno de ellos de manera gráfica. Esta imagen sirve para observar por ejemplo que el clúster 1 pude definirse por una recepción perfecta proveniente desde el centro del campo. Estas particularidades, se evidenciarán y analizarán aún más en el momento de sacar conclusiones.

A modo de ejemplo la tabla 4.15 muestra las características de los casos que pertenecen al Clúster 1 mostrando los atributos de manera decreciente según los valores de probabilidad de ocurrencia.

Variables	Valores	Probabilidad
Rec Zone	C	100,00%
Rec Eval	#	94,35%
Setter Call	3T	62,14%
Setter Pos	23	58,19%
Setter Pos	3	39,55%
Set Moment	A2	37,29%
Set Moment	A3	29,38%
Setter	FT-34	27,12%
Set Moment	A1	27,12%
Fase	3	27,12%
Fase	4	26,00%
Fase	5	24,30%
Setter	2T-4	20,90%
Setter	FT-3	16,39%
Setter	2T-6	14,12%
Setter Call	A	14,12%
Setter	2T-1	13,00%
Setter Call	1T	13,00%
Setter Call	1C	10,73%
Fase	1	10,73%
Setter	2T-2	8,48%
Fase	2	6,22%
Fase	6	5,65%
Rec Eval	*	5,65%
Set Moment	B1	4,52%
Setter Pos	2	2,26%
Set Moment	B2	1,70%

Tabla 4.15: Características de Clúster 1

Para poder leer con mayor exactitud las características del clúster se pueden realizar comparaciones entre un clúster seleccionado y el resto de la población o bien entre dos clústeres elegidos. La tabla 4.16 compara el Clúster 1 con el resto de la población. Estos números son solo puntuaciones (como en el caso explicado para el algoritmo Bayes Naïve) que indican la influencia de un atributo en la pertenencia o no a un clúster.

Variables	Valores	Favorece Cluster 1	Favorece Complemento
Rec Zone	C	100,00	
Rec Eval	#	53,81	
Rec Zone	SX		43,07
Rec Eval	+		37,79
Rec Zone	DX		31,54
Setter Call	3T	29,28	
Setter	FT-34	26,20	
Setter Call	0T		22,45
Rec Eval	!		18,30
Setter Pos	23	17,12	
Setter	3T-4		14,52
Fase	6		9,80
Fase	4	8,38	
Rec Eval	*	7,73	
Setter Pos	1		6,89
Setter Call	1C		6,38
Setter Pos	6		5,42
Fase	3	4,20	
Fase	2		3,62
Setter	2T-1	3,30	
Setter Pos	2		2,75
Setter	2T-6	2,53	
Fase	1		2,29
Fase	5	1,75	

Tabla 4.16: Distinción del Clúster 1

4.4.3.4 Construcción reglas de Asociación

Debido a la fuerte influencia de los valores de los parámetros en los resultados obtenidos, se trabajó estrechamente con el experto del problema llegando a un acuerdo sobre la definición de los mismos que se refleja en la figura 4-26.

Parámetros

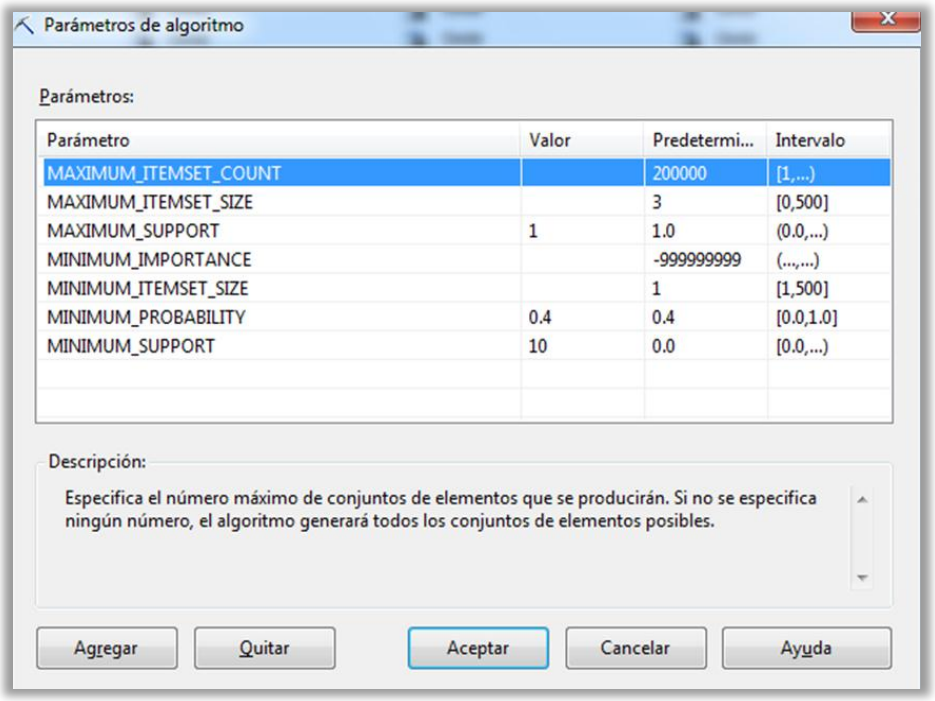
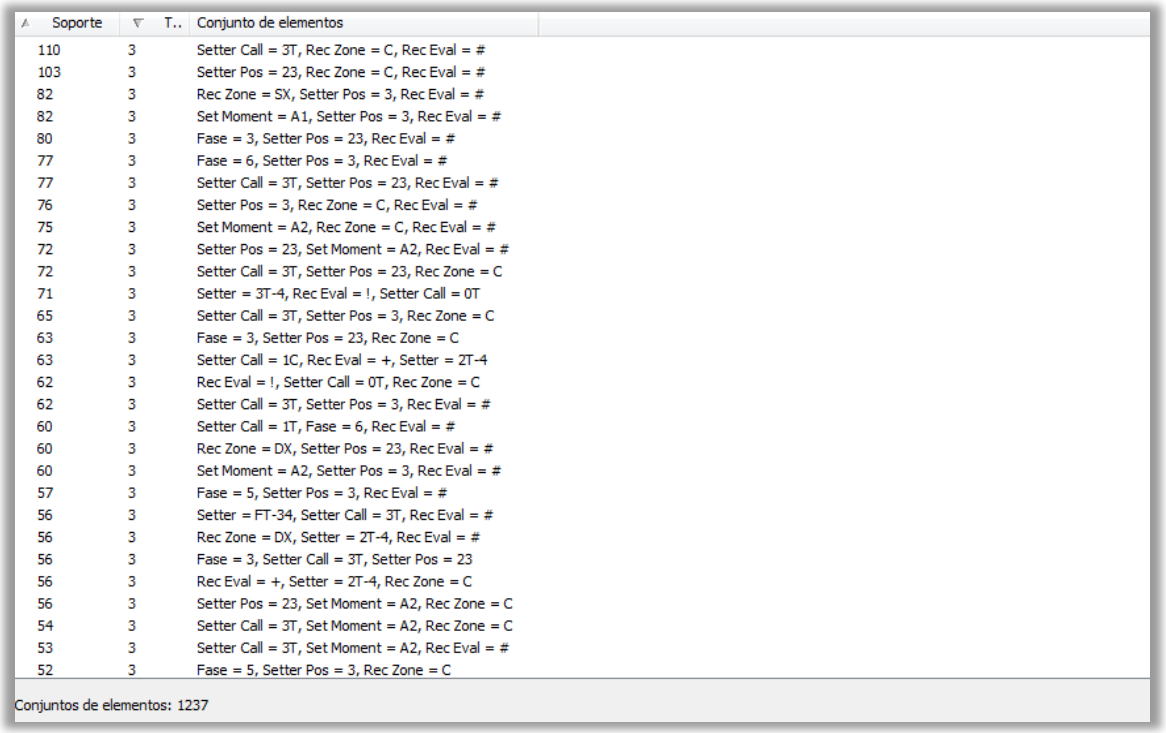


Fig. 4-26: Parámetros Reglas de Asociación

Interpretación del modelo



Soporte	T..	Conjunto de elementos
110	3	Setter Call = 3T, Rec Zone = C, Rec Eval = #
103	3	Setter Pos = 23, Rec Zone = C, Rec Eval = #
82	3	Rec Zone = SX, Setter Pos = 3, Rec Eval = #
82	3	Set Moment = A1, Setter Pos = 3, Rec Eval = #
80	3	Fase = 3, Setter Pos = 23, Rec Eval = #
77	3	Fase = 6, Setter Pos = 3, Rec Eval = #
77	3	Setter Call = 3T, Setter Pos = 23, Rec Eval = #
76	3	Setter Pos = 3, Rec Zone = C, Rec Eval = #
75	3	Set Moment = A2, Rec Zone = C, Rec Eval = #
72	3	Setter Pos = 23, Set Moment = A2, Rec Eval = #
72	3	Setter Call = 3T, Setter Pos = 23, Rec Zone = C
71	3	Setter = 3T-4, Rec Eval = !, Setter Call = 0T
65	3	Setter Call = 3T, Setter Pos = 3, Rec Zone = C
63	3	Fase = 3, Setter Pos = 23, Rec Zone = C
63	3	Setter Call = 1C, Rec Eval = +, Setter = 2T-4
62	3	Rec Eval = !, Setter Call = 0T, Rec Zone = C
62	3	Setter Call = 3T, Setter Pos = 3, Rec Eval = #
60	3	Setter Call = 1T, Fase = 6, Rec Eval = #
60	3	Rec Zone = DX, Setter Pos = 23, Rec Eval = #
60	3	Set Moment = A2, Setter Pos = 3, Rec Eval = #
57	3	Fase = 5, Setter Pos = 3, Rec Eval = #
56	3	Setter = FT-34, Setter Call = 3T, Rec Eval = #
56	3	Rec Zone = DX, Setter = 2T-4, Rec Eval = #
56	3	Fase = 3, Setter Call = 3T, Setter Pos = 23
56	3	Rec Eval = +, Setter = 2T-4, Rec Zone = C
56	3	Setter Pos = 23, Set Moment = A2, Rec Zone = C
54	3	Setter Call = 3T, Set Moment = A2, Rec Zone = C
53	3	Setter Call = 3T, Set Moment = A2, Rec Eval = #
52	3	Fase = 5, Setter Pos = 3, Rec Zone = C

Conjuntos de elementos: 1237

Fig. 4-27: Itemsets Reglas de Asociación

La figura 4-27muestra el resultado de la ejecución de la primera fase del algoritmo, define los itemsets hallados. En este caso en particular no fue necesario filtrar por una cantidad específica de casos debido al objetivo del problema a resolver y a la cantidad de registros disponibles.

Prob...	Importancia	Regla
1,000	0,964	Setter Pos = 6, Fase = 5 -> Setter = 3T-4
1,000	0,890	Rec Eval = *, Setter Call = A -> Setter = FT-3
1,000	0,890	Rec Eval = *, Fase = 5 -> Setter = FT-3
1,000	0,890	Rec Eval = *, Setter Pos = 3 -> Setter = FT-3
1,000	0,890	Rec Eval = *, Rec Zone = C -> Setter = FT-3
1,000	0,521	Setter Pos = 2, Setter Call = 1C -> Setter = 2T-4
1,000	0,890	Setter Pos = 2, Setter Call = A -> Setter = FT-3
1,000	0,527	Fase = 3, Setter Call = 1C -> Setter = 2T-4
1,000	1...	Setter Pos = 6 -> Setter = 3T-4
1,000	1,...	Setter Pos = 6, Rec Eval = ! -> Setter = 3T-4
1,000	1,024	Setter Pos = 1, Fase = 1 -> Setter = 3T-4
1,000	1...	Setter Pos = 6, Setter Call = 0T -> Setter = 3T-4
1,000	0,964	Setter Pos = 6, Fase = 6 -> Setter = 3T-4
1,000	0,971	Setter Pos = 6, Set Moment = A1 -> Setter = 3T-4
1,000	0,937	Setter Pos = 6, Set Moment = A2 -> Setter = 3T-4
1,000	1...	Setter Pos = 6, Rec Zone = C -> Setter = 3T-4
1,000	0,890	Rec Eval = * -> Setter = FT-3
1,000	0,544	Fase = 4, Setter Call = 1C -> Setter = 2T-4
1,000	1,024	Setter Pos = 1, Rec Eval = ! -> Setter = 3T-4
1,000	0,924	Setter Pos = 1, Rec Eval = # -> Setter = FT-3
0,966	1,029	Setter Pos = 1, Setter Call = 0T -> Setter = 3T-4
0,941	0,940	Rec Eval = !, Fase = 6 -> Setter = 3T-4
0,941	0,940	Setter Call = 0T, Fase = 6 -> Setter = 3T-4
0,933	1,...	Setter Call = 0T, Fase = 1 -> Setter = 3T-4
0,933	1,...	Rec Eval = !, Fase = 1 -> Setter = 3T-4
0,917	1,...	Fase = 2, Setter Call = 1T -> Setter = 2T-6
0,909	0,496	Fase = 3, Rec Eval = + -> Setter = 2T-4
0,800	0,425	Setter Call = A, Fase = 3 -> Setter = 2T-4
0,800	0,949	Rec Eval = !, Rec Zone = SX -> Setter = 3T-4

Reglas: 127

Fig. 4-28: Reglas del algoritmo Reglas de Asociación

En la figura 4-28 se califican las reglas halladas en base a la puntuación obtenida en la probabilidad e importancia de la misma. La importancia puede considerarse como una medida de la utilidad de la regla, a mayor importancia mayor calidad de la misma. La máxima probabilidad de ocurrencia no garantiza utilidad, es necesario observar ambos parámetros para obtener el verdadero valor de la regla.

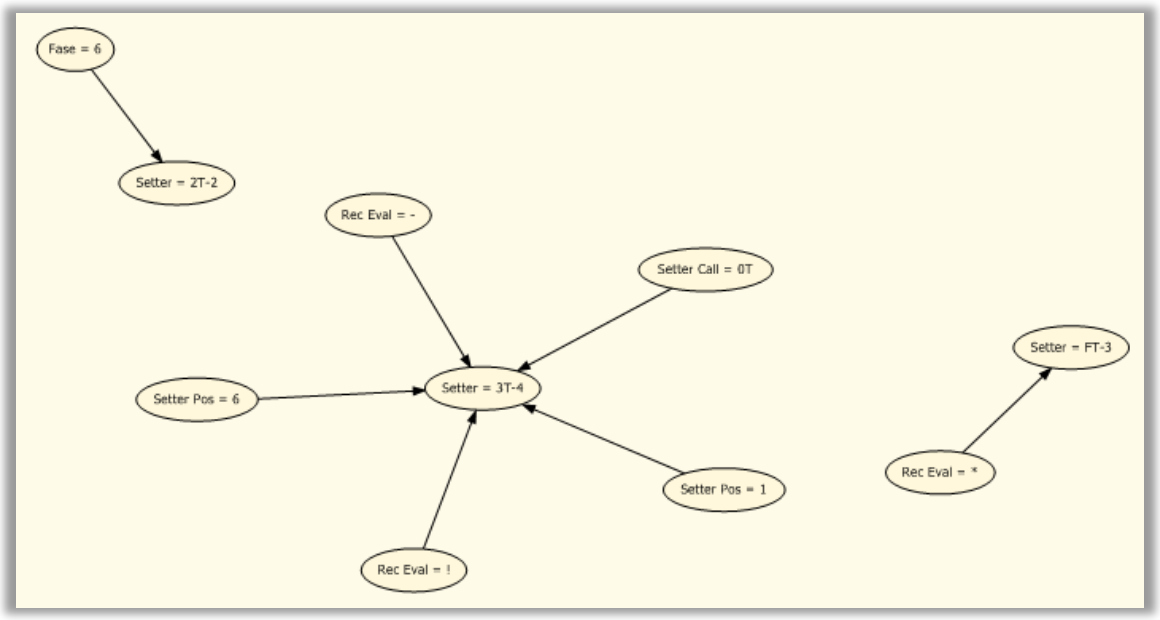


Fig. 4-29: Red de Dependencias Reglas de Asociación

Es oportuno observar en la figura 4-29 que el modelo encuentra una relación ya verificada previamente por el algoritmo Bayes Naïve y es que en la fase 6 existe una fuerte tendencia por la jugada de tipo 2T-2. Estas peculiaridades darán validez a las conclusiones extraídas dado que se repiten más allá del algoritmo seleccionado.

4.4.3.5 Construcción Red Neuronal

Parámetros

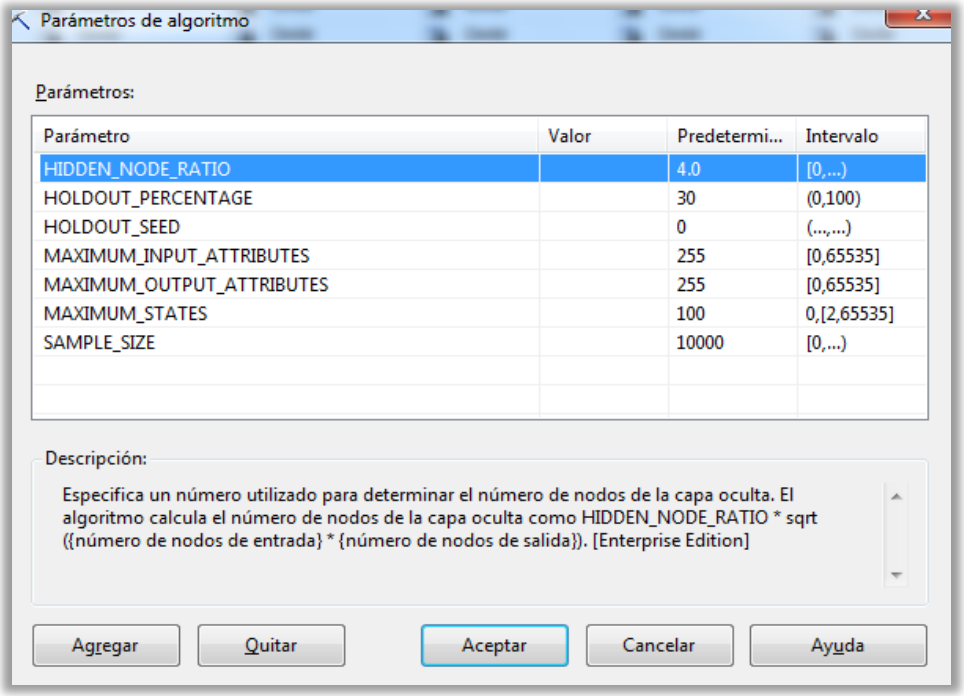


Fig. 4-30: Parámetros Red Neuronal

Interpretación del modelo

La visualización del modelo de Red Neuronal difiere de los algoritmos anteriores. No se muestra la distribución de la red de dependencias sino que se visualiza el impacto de los atributos de entrada respecto el atributo target u objetivo.La tabla se ordena según la puntuación calculada de manera probabilística y se refleja en la figura 4-13.

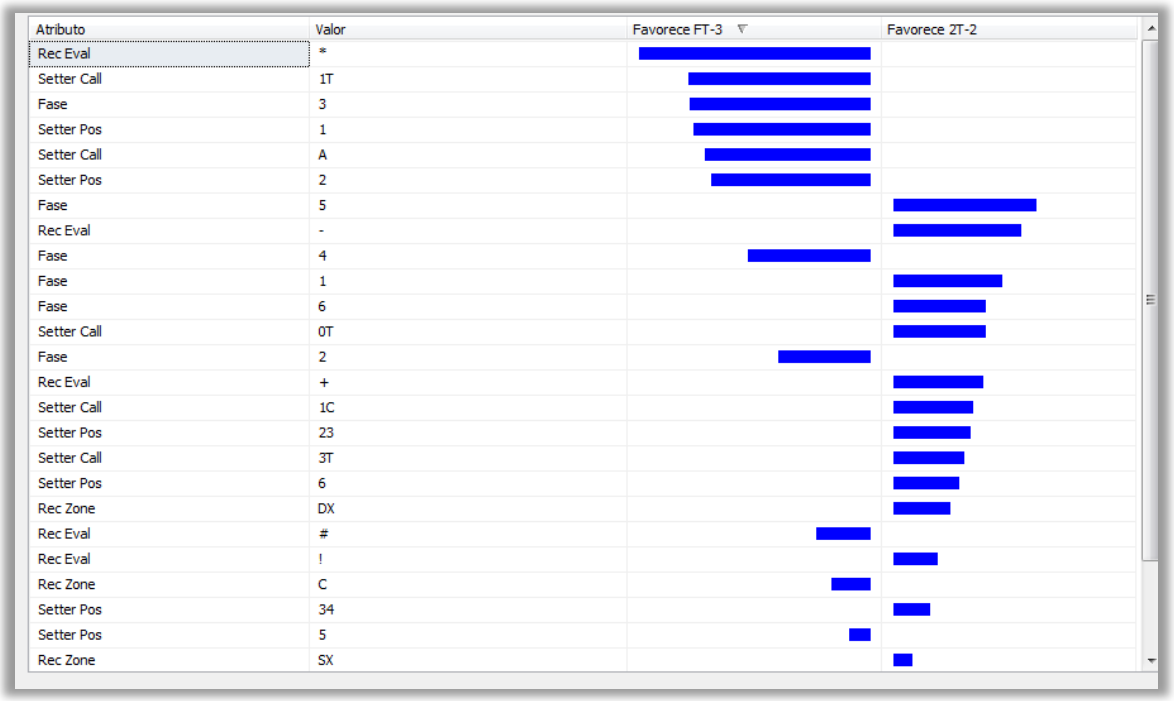


Fig. 4-31: Visor del modelo Red Neuronal

En la tabla 4.17 se muestran las puntuaciones obtenidas al comparar un primer tiempo de tipo FT-3 con un segundo tiempo 2T-2. Se verifica claramente que cuando la evaluación de la recepción permite solo primer tiempo, este atributo favorece con una puntuación de 100 la jugada FT-3 y corresponde efectivamente a un primer tiempo. Si bien estas particularidades podrían parecer obvias confirman el buen funcionamiento del modelo.

Atributo	Valor	Favorece FT-3	Favorece 2T-2
Rec Eval	*	100	
Setter Call	1T	79,16	
Fase	3	78,35	
Setter Pos	1	76,62	
Setter Call	A	71,59	
Setter Pos	2	69,21	

Fase	5		61,94
Rec Eval	-		55,23
Fase	4	53,14	
Fase	1		47,19
Fase	6		40,38
Setter Call	0T		40,24
Fase	2	40,11	
Rec Eval	+		39,34
Setter Call	1C		34,68
Setter Pos	23		33,42
Setter Call	3T		30,72
Setter Pos	6		28,83
Rec Zone	DX		25,11
Rec Eval	#	23,7	
Rec Eval	!		19,6
Rec Zone	C	17,29	
Setter Pos	34		16,16
Setter Pos	5	9,8	
Rec Zone	SX		8,47
Setter Pos	4	5,01	
Setter Pos	61	4,57	
Setter Pos	3	0,15	

Tabla 4.17: Comparación Puntuación FT-3 y 2T-2

4.4.4 Valoración de los modelos

Se realiza teniendo en cuenta la perspectiva de los expertos en el dominio del problema como así también las bondades y limitaciones de las técnicas de minería de datos.

Cada uno de los modelos ha sido probado a través de las iteraciones realizadas. La retroalimentación con el usuario final ha sido constante. Se logró someter el conjunto de datos disponibles a distintos algoritmos, y de cada uno de ellos se extrajo información útil, ya sea como conocimiento ganado o como corroboración cruzada de resultados de otros modelos. Se han descubierto características interesantes y comunes en algunos de los modelos, durante la construcción y exploración de los mismos. Se estima necesario continuar con la fase de evaluación.

4.5 FASE DE EVALUACIÓN

En primer lugar la evaluación se realiza teniendo en cuenta los criterios de éxito del proyecto de minería de datos. Hasta este punto se han vislumbrado relaciones interesantes descubiertas a través del entrenamiento de los modelos. Se concluye entonces que desde esa óptica el proyecto resulta positivo. Es necesario evaluar también la puntuación de los modelos para calificar la eficacia y eficiencia de los mismos siguiendo los criterios de evaluación.

4.5.1 Resultados

Se procede a utilizar los métodos enumerados en el plan de pruebas a cada uno de los modelos generados para poder calificarlos y a su vez compararlos

4.5.1.1 Gráfico de Elevación o Lift Chart

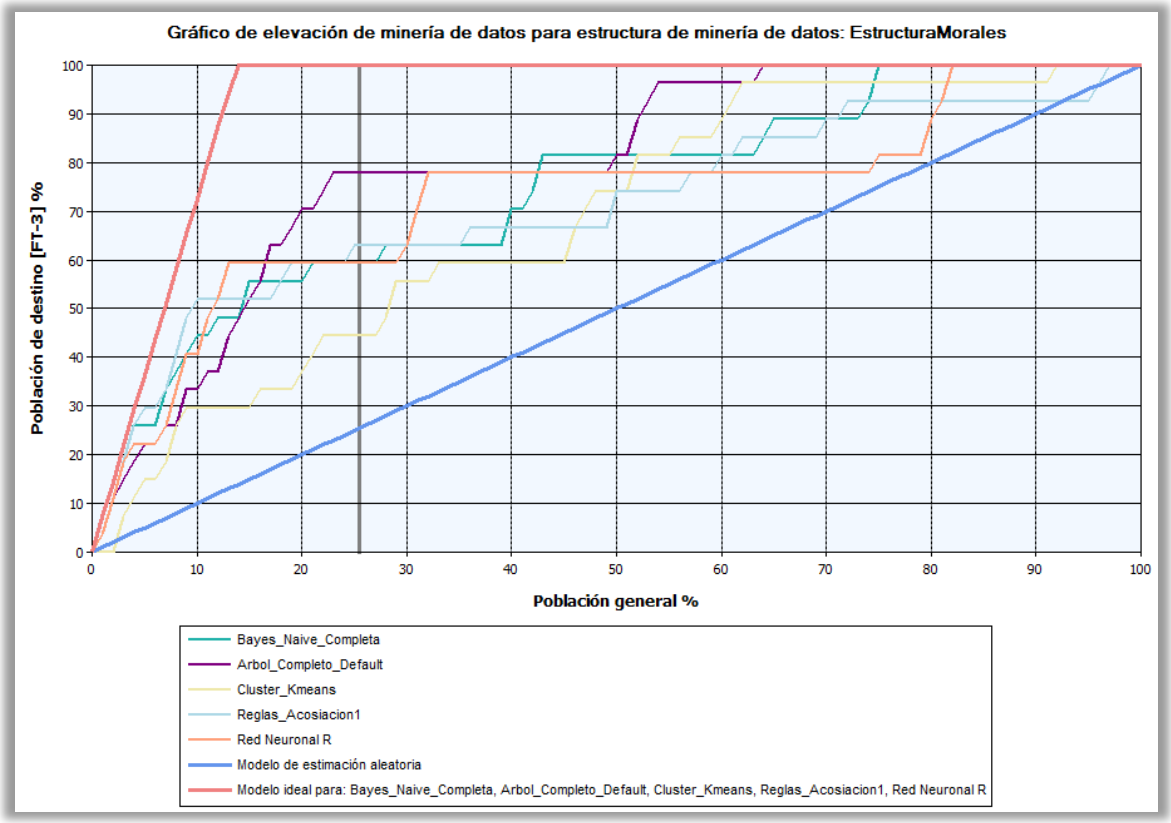


Fig. 4-32: Gráfico de Elevación para FT-3

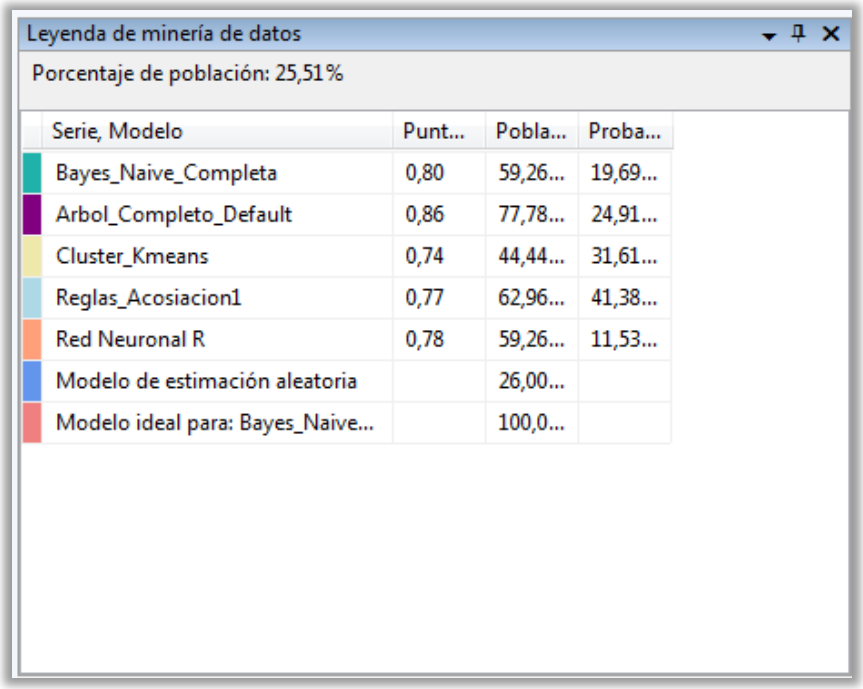


Fig. 4-33: Leyenda gráfico de Elevación

La figura 4-32 muestra que todos los modelos están por encima de la línea de un modelo aleatorio y a su vez por debajo de un modelo ideal. Esto seguramente es positivo porque nos está diciendo que para un planteo estratégico de la defensa ante un posible ataque adversario de tipo FT-3 podemos utilizar los modelos hallados y seguramente la elección será superior a que si dejamos la decisión librada al azar.

En el eje X tenemos la población total, en el eje Y se ha seleccionado la opción "Setter = FT-3". La elección se realiza teniendo en cuenta que los primeros tiempos son los que más atención necesitan a la hora de anticipar una jugada, por eso se vuelve interesante este tipo de análisis.

Para entender claramente el resultado obtenido es necesario comprender las partes del gráfico.

El eje X representa el porcentaje del conjunto de datos de prueba que se usa para comparar las predicciones. El eje Y representa el porcentaje de valores de predicción.

La línea celeste oscura representa los resultados de la estimación aleatoria y en base a ella se evalúa la mejora del modelo. Las líneas celeste, verde, violeta, naranja y amarilla representan cada una un modelo hallado y la línea coral el modelo ideal para cada uno de los modelos entrenados (en este caso particular coinciden los modelos ideales).

La línea gris está posicionada en el punto en el cuál se describe la leyenda.

La puntuación es una fracción de la performance respecto al modelo ideal. Ayuda a comparar la efectividad de los modelos utilizando una población normalizada.

El modelo ideal puede captar el 100% del objetivo utilizando aproximadamente el 14% del total de la población.

El modelo de Árbol de Decisión en cambio puede captar el 77,8% del objetivo a partir del 26% del total de la población. EL umbral de probabilidad necesario para incluir un caso entre los casos con probabilidad de ser FT-3 es 24,91 %.

Imaginando disponer de 1000 casos sé que 140 (el 14% como dice el gráfico) responderán positivamente a la predicción o lo que es lo mismo la predicción será correcta para esos casos. Ordenando los casos disponibles según el modelo ideal los primeros 140 serán los correctos. En la vida real existen predicciones incorrectas, ordenando los casos según el modelo del árbol de decisión encontraré 108 casos exactos (el 77,78% de los 140 que me da el modelo ideal) entre los primeros 260 casos (26% del total de la población).

Interpretando la leyenda vemos que la línea gris dice que en el 25,51% de los casos presentados, el modelo ideal puede predecir correctamente el 100% de ellos, es decir que para ese punto en el eje X predecirá 250 casos perfectos. El Árbol de Decisión puede en cambio predecir correctamente 219 casos, el 77,78% del total de la población en ese punto, Bayes Naïve puede predecir el 59, 26% es decir 148 casos, Clúster 44,44% lo que significa 111 casos predecibles, el algoritmo de Reglas de Asociación el 62,96%, 157 casos y la Red Neuronal 59,26%, es decir 148 casos.

Es importante comprender que para el dominio del que se trata este estudio no es de utilidad conocer la cantidad de casos que se pueden predecir pero sí qué tan bien trabaja el modelo respecto al modelo aleatorio y si se acerca o no a un modelo ideal.

Por lo dicho hasta aquí es válido continuar con la verificación de la exactitud de los resultados ya que se considera positiva la primera prueba, todos los modelos superan la línea aleatoria y podrían estar en condiciones de ayudar a un entrenador en la identificación de patrones y toma de decisiones. Se puede decir entonces que el modelo es de utilidad para el objetivo planteado.

4.5.1.2 Matriz de Clasificación o Matriz de Confusión

A continuación se presenta la matriz hallada para cada uno de los modelos realizados.

		REAL														
		2T-1	2T-2	2T-34	2T-4	2T-6	2T-65	3T-1	3T-2	3T-4	FT-2	FT-23	FT-3	FT-34	FT-4	Total
PREVISTO	2T-1	9	0	0	2	0	0	0	0	0	0	0	0	4	0	15
	2T-2	0	11	0	4	0	0	0	0	0	0	0	2	0	0	17
	2T-34	0	0	5	0	0	0	0	0	0	0	0	0	0	0	5
	2T-4	2	5	0	36	0	1	0	0	1	0	0	11	4	0	60
	2T-6	0	0	0	5	3	0	0	0	0	0	0	1	0	0	9
	2T-65	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
	3T-1	1	0	0	0	0	0	4	0	0	0	0	0	0	0	5
	3T-2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2
	3T-4	0	0	0	2	0	0	0	0	22	0	0	0	0	0	24
	FT-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-23	0	0	0	5	0	0	0	0	0	0	6	0	0	0	11
	FT-3	0	4	0	6	3	0	0	0	0	0	0	13	0	0	26
	FT-34	2	3	0	5	6	0	0	0	0	0	0	0	5	0	21
	FT-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL	14	23	5	65	12	2	4	2	23	0	6	27	13	0	**117**
																59,69%

Tabla 4.18: Recuentos para Bayes Naïve en Setter

		REAL														
		2T-1	2T-2	2T-34	2T-4	2T-6	2T-65	3T-1	3T-2	3T-4	FT-2	FT-23	FT-3	FT-34	FT-4	Total
PREVISTO	2T-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2T-2	0	14	0	1	2	2	0	0	0	0	0	7	0	0	26
	2T-34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2T-4	9	4	5	54	3	0	0	0	1	0	6	18	4	0	104
	2T-6	0	0	0	2	7	0	0	0	0	0	0	0	2	0	11
	2T-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-4	1	0	0	2	0	0	4	2	22	0	0	0	0	0	31
	FT-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-3	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
	FT-34	4	5	0	6	0	0	0	0	0	0	0	0	7	0	22
	FT-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL	14	23	5	65	12	2	4	2	23	0	6	27	13	0	**106**
																54,08%

Tabla 4.19: Recuentos para Árbol de Decisión en Setter

		REAL														
		2T-1	2T-2	2T-34	2T-4	2T-6	2T-65	3T-1	3T-2	3T-4	FT-2	FT-23	FT-3	FT-34	FT-4	Total
PREVISTO	2T-1	13	0	0	9	0	0	0	0	1	0	0	4	6	0	33
	2T-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2T-34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2T-4	0	10	0	38	7	2	0	0	0	0	6	11	7	0	81
	2T-6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2T-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-4	1	0	0	2	0	0	4	2	22	0	0	0	0	0	31
	FT-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-3	0	13	5	16	5	0	0	0	0	0	0	12	0	0	51
	FT-34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL	14	23	5	65	12	2	4	2	23	0	6	27	13	0	**85**
																43,37%

Tabla 4.20: Recuentos para Clustering en Setter

		REAL														
		2T-1	2T-2	2T-34	2T-4	2T-6	2T-65	3T-1	3T-2	3T-4	FT-2	FT-23	FT-3	FT-34	FT-4	Total
PREVISTO	2T-1	1	0	0	0	0	0	0	0	0	0	0	0	2	0	3
	2T-2	0	13	0	3	0	0	0	0	0	0	0	5	0	0	21
	2T-34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2T-4	12	10	5	59	3	2	0	0	1	0	6	14	6	0	118
	2T-6	0	0	0	0	9	0	0	0	0	0	0	0	2	0	11
	2T-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
	3T-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-4	1	0	0	2	0	0	2	2	22	0	0	0	0	0	29
	FT-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-3	0	0	0	0	0	0	0	0	0	0	0	8	1	0	9
	FT-34	0	0	0	1	0	0	0	0	0	0	0	0	2	0	3
	FT-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL	14	23	5	65	12	2	4	2	23	0	6	27	13	0	**116**
																59,18%

Tabla 4.21: Recuentos para Reglas Asociación en Setter

		Real														
		2T-1	2T-2	2T-34	2T-4	2T-6	2T-65	3T-1	3T-2	3T-4	FT-2	FT-23	FT-3	FT-34	FT-4	Total
PREVISTO	2T-1	6	1	0	1	0	0	0	0	0	0	0	0	2	0	10
	2T-2	0	10	0	1	0	0	0	0	0	0	0	3	2	0	16
	2T-34	0	0	5	4	0	0	0	0	0	0	0	0	0	0	9
	2T-4	8	0	0	50	0	2	0	0	0	0	3	7	2	0	72
	2T-6	0	4	0	3	12	0	0	0	0	0	0	7	2	0	28
	2T-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3T-1	0	0	0	0	0	0	4	0	0	0	0	0	0	0	4
	3T-2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2
	3T-4	0	3	0	0	0	0	0	0	23	0	0	0	0	0	26
	FT-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FT-23	0	1	0	2	0	0	0	0	0	0	3	0	0	0	6
	FT-3	0	4	0	1	0	0	0	0	0	0	0	10	0	0	15
	FT-34	0	0	0	3	0	0	0	0	0	0	0	0	5	0	8
	FT-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL	14	23	5	65	12	2	4	2	23	0	6	27	13	0	**130**
																66,33%

Tabla 4.22: Recuentos para Red Neuronal en Setter

Observando la matriz de cada modelo se puede saber rápidamente en cuántas ocasiones ha sido exacta la predicción del mismo.

Las filas de cada matriz representan los valores de predicción del modelo, mientras que las columnas representan los valores reales.

En la diagonal principal de la matriz se encuentran todos los casos que se han predicho con veracidad. De esta manera se lee entonces la exactitud del modelo.

Las matrices se elaboraron con el 25% del total de datos disponibles, según lo especificado en la estructura de minería de datos y en los modelos, no con los que se han utilizado para generar los mismos. Esta particularidad otorga confiabilidad a la prueba realizada.

Las matrices han dado confiabilidad y exactitud a los modelos, por lo tanto se puede considerar satisfactoria la prueba realizada.

4.5.1.3 Validación Cruzada o Cross Validation

Esta prueba es sin dudas la que más información puede aportar. Es la ideal para el presente estudio considerando la reducida cantidad de datos del dataset origen de los modelos. Las pruebas anteriores pudieron asegurar exactitud en la predicción, esta prueba deberá además asegurar si el modelo se ajusta al trabajo para el cuál ha sido creado. Deberá otorgar solidez y permitirá comparar los modelos entre sí desde el punto de vista estadístico.

En primer lugar se analizan todos los modelos menos el de Clustering dado que éste no puede ser sometido a las mismas pruebas por el tipo de resultado que arroja.

Se ha elegido realizar diez particiones. Los resultados obtenidos para los modelos de Bayes Naïve, Árbol de Decisión, Reglas de Asociación y Red Neuronal se muestran a continuación en la tabla 4.23.

Bayes Naïve				
Índice de partición	Tamaño de partición	Prueba	Medida	Valor
1	96	Classification	Pass	57
2	98	Classification	Pass	63
3	100	Classification	Pass	57
4	99	Classification	Pass	65
5	99	Classification	Pass	62
6	99	Classification	Pass	63
7	98	Classification	Pass	65
8	97	Classification	Pass	63
9	98	Classification	Pass	58
10	97	Classification	Pass	68
			Promedio	62,0968
			Desviación estándar	3,5001
1	96	Classification	Fail	39
2	98	Classification	Fail	35
3	100	Classification	Fail	43
4	99	Classification	Fail	34
5	99	Classification	Fail	37
6	99	Classification	Fail	36
7	98	Classification	Fail	33
8	97	Classification	Fail	34
9	98	Classification	Fail	40
10	97	Classification	Fail	29
			Promedio	36,0163
			Desviación estándar	3,7707
1	96	Likelihood	Log Score	-10,929
2	98	Likelihood	Log Score	-10,508
3	100	Likelihood	Log Score	-10,323
4	99	Likelihood	Log Score	-0,9458
5	99	Likelihood	Log Score	-10,171

6	99	Likelihood	Log Score	-0,9801
7	98	Likelihood	Log Score	-0,9196
8	97	Likelihood	Log Score	-0,92
9	98	Likelihood	Log Score	-10,919
10	97	Likelihood	Log Score	-0,7007
			Promedio	-0,9754
			Desviación estándar	0,1093
1	96	Likelihood	Lift	0,964
2	98	Likelihood	Lift	10,345
3	100	Likelihood	Lift	1,052
4	99	Likelihood	Lift	11,379
5	99	Likelihood	Lift	10,556
6	99	Likelihood	Lift	10,762
7	98	Likelihood	Lift	11,664
8	97	Likelihood	Lift	11,791
9	98	Likelihood	Lift	10,285
10	97	Likelihood	Lift	13,782
			Promedio	1,107
			Desviación estándar	0,1098
1	96	Likelihood	Root Mean Square Error	0,3612
2	98	Likelihood	Root Mean Square Error	0,3618
3	100	Likelihood	Root Mean Square Error	0,3741
4	99	Likelihood	Root Mean Square Error	0,3544
5	99	Likelihood	Root Mean Square Error	0,3477
6	99	Likelihood	Root Mean Square Error	0,3531
7	98	Likelihood	Root Mean Square Error	0,3727
8	97	Likelihood	Root Mean Square Error	0,3624
9	98	Likelihood	Root Mean Square Error	0,3475
10	97	Likelihood	Root Mean Square Error	0,363
			Promedio	0,3598
			Desviación estándar	0,0088
Árbol de Decisión				
Índice de partición	Tamaño de partición	Prueba	Medida	Valor
1	96	Classification	Pass	60
2	98	Classification	Pass	58
3	100	Classification	Pass	60
4	99	Classification	Pass	60
5	99	Classification	Pass	57
6	99	Classification	Pass	60

7	98	Classification	Pass	62
8	97	Classification	Pass	60
9	98	Classification	Pass	58
10	97	Classification	Pass	62
			Promedio	59,6952
			Desviación estándar	1,5523
1	96	Classification	Fail	36
2	98	Classification	Fail	40
3	100	Classification	Fail	40
4	99	Classification	Fail	39
5	99	Classification	Fail	42
6	99	Classification	Fail	39
7	98	Classification	Fail	36
8	97	Classification	Fail	37
9	98	Classification	Fail	40
10	97	Classification	Fail	35
			Promedio	38,4179
			Desviación estándar	2,1514
1	96	Likelihood	Log Score	-0,9278
2	98	Likelihood	Log Score	-0,9651
3	100	Likelihood	Log Score	-0,9671
4	99	Likelihood	Log Score	-0,9477
5	99	Likelihood	Log Score	-10,311
6	99	Likelihood	Log Score	-0,9299
7	98	Likelihood	Log Score	-0,885
8	97	Likelihood	Log Score	-0,8475
9	98	Likelihood	Log Score	-0,9436
10	97	Likelihood	Log Score	-0,8161
			Promedio	-0,9265
			Desviación estándar	0,0591
1	96	Likelihood	Lift	11,292
2	98	Likelihood	Lift	11,202
3	100	Likelihood	Lift	11,172
4	99	Likelihood	Lift	1,136
5	99	Likelihood	Lift	10,416
6	99	Likelihood	Lift	11,264
7	98	Likelihood	Lift	1,201
8	97	Likelihood	Lift	12,515
9	98	Likelihood	Lift	11,769
10	97	Likelihood	Lift	12,628
			Promedio	11,559

			Desviación estándar	0,0639
1	96	Likelihood	Root Mean Square Error	0,4091
2	98	Likelihood	Root Mean Square Error	0,4404
3	100	Likelihood	Root Mean Square Error	0,4229
4	99	Likelihood	Root Mean Square Error	0,4075
5	99	Likelihood	Root Mean Square Error	0,4319
6	99	Likelihood	Root Mean Square Error	0,4464
7	98	Likelihood	Root Mean Square Error	0,4145
8	97	Likelihood	Root Mean Square Error	0,4166
9	98	Likelihood	Root Mean Square Error	0,43
10	97	Likelihood	Root Mean Square Error	0,414
			Promedio	0,4234
			Desviación estándar	0,0127
Reglas de Asociación				
Índice de partición	Tamaño de partición	Prueba	Medida	Valor
1	96	Classification	Pass	57
2	98	Classification	Pass	57
3	100	Classification	Pass	59
4	99	Classification	Pass	57
5	99	Classification	Pass	60
6	99	Classification	Pass	57
7	98	Classification	Pass	57
8	97	Classification	Pass	57
9	98	Classification	Pass	55
10	97	Classification	Pass	61
			Promedio	57,7023
			Desviación estándar	1,6746
1	96	Classification	Fail	39
2	98	Classification	Fail	41
3	100	Classification	Fail	41
4	99	Classification	Fail	42
5	99	Classification	Fail	39
6	99	Classification	Fail	42
7	98	Classification	Fail	41
8	97	Classification	Fail	40
9	98	Classification	Fail	43
10	97	Classification	Fail	36
			Promedio	40,4108
			Desviación estándar	1,9027

1	96	Likelihood	Log Score	-0,9415
2	98	Likelihood	Log Score	-1,035
3	100	Likelihood	Log Score	-10,152
4	99	Likelihood	Log Score	-0,9821
5	99	Likelihood	Log Score	-0,9593
6	99	Likelihood	Log Score	-0,9301
7	98	Likelihood	Log Score	-0,9607
8	97	Likelihood	Log Score	-0,8862
9	98	Likelihood	Log Score	-0,9873
10	97	Likelihood	Log Score	-0,9308
			Promedio	-0,9631
			Desviación estándar	0,0416
1	96	Likelihood	Lift	11,154
2	98	Likelihood	Lift	10,503
3	100	Likelihood	Lift	10,691
4	99	Likelihood	Lift	11,016
5	99	Likelihood	Lift	11,134
6	99	Likelihood	Lift	11,261
7	98	Likelihood	Lift	11,254
8	97	Likelihood	Lift	12,128
9	98	Likelihood	Lift	11,332
10	97	Likelihood	Lift	11,481
			Promedio	11,193
			Desviación estándar	0,0418
1	96	Likelihood	Root Mean Square Error	0,3214
2	98	Likelihood	Root Mean Square Error	0,3392
3	100	Likelihood	Root Mean Square Error	0,3438
4	99	Likelihood	Root Mean Square Error	0,3309
5	99	Likelihood	Root Mean Square Error	0,3335
6	99	Likelihood	Root Mean Square Error	0,3164
7	98	Likelihood	Root Mean Square Error	0,3269
8	97	Likelihood	Root Mean Square Error	0,3347
9	98	Likelihood	Root Mean Square Error	0,3336
10	97	Likelihood	Root Mean Square Error	0,3411
			Promedio	0,3322
			Desviación estándar	0,0082

Red Neuronal				
Índice de partición	Tamaño de partición	Prueba	Medida	Valor
1	96	Classification	Pass	72
2	98	Classification	Pass	73
3	100	Classification	Pass	75
4	99	Classification	Pass	72
5	99	Classification	Pass	73
6	99	Classification	Pass	74
7	98	Classification	Pass	69
8	97	Classification	Pass	76
9	98	Classification	Pass	70
10	97	Classification	Pass	87
			Promedio	74,0877
			Desviación estándar	4,7214
1	96	Classification	Fail	24
2	98	Classification	Fail	25
3	100	Classification	Fail	25
4	99	Classification	Fail	27
5	99	Classification	Fail	26
6	99	Classification	Fail	25
7	98	Classification	Fail	29
8	97	Classification	Fail	21
9	98	Classification	Fail	28
10	97	Classification	Fail	10
			Promedio	24,0255
			Desviación estándar	5,0971
1	96	Likelihood	Log Score	-0,763
2	98	Likelihood	Log Score	-0,8185
3	100	Likelihood	Log Score	-0,8171
4	99	Likelihood	Log Score	-0,775
5	99	Likelihood	Log Score	-0,7612
6	99	Likelihood	Log Score	-0,8106
7	98	Likelihood	Log Score	-0,765
8	97	Likelihood	Log Score	-0,6948
9	98	Likelihood	Log Score	-0,7924
10	97	Likelihood	Log Score	-0,4376
			Promedio	-0,7441
			Desviación estándar	0,1073

1	96	Likelihood	Lift	1,294
2	98	Likelihood	Lift	12,669
3	100	Likelihood	Lift	12,672
4	99	Likelihood	Lift	13,087
5	99	Likelihood	Lift	13,115
6	99	Likelihood	Lift	12,456
7	98	Likelihood	Lift	1,321
8	97	Likelihood	Lift	14,043
9	98	Likelihood	Lift	1,328
10	97	Likelihood	Lift	16,413
			Promedio	13,383
			Desviación estándar	0,1086
1	96	Likelihood	Root Mean Square Error	0,4115
2	98	Likelihood	Root Mean Square Error	0,3718
3	100	Likelihood	Root Mean Square Error	0,375
4	99	Likelihood	Root Mean Square Error	0,3768
5	99	Likelihood	Root Mean Square Error	0,3769
6	99	Likelihood	Root Mean Square Error	0,335
7	98	Likelihood	Root Mean Square Error	0,3518
8	97	Likelihood	Root Mean Square Error	0,3412
9	98	Likelihood	Root Mean Square Error	0,3802
10	97	Likelihood	Root Mean Square Error	0,3447
			Promedio	0,3664
			Desviación estándar	0,0219

Tabla 4.23: Validación Cruzada Bayes Naïve, Árbol de Decisión, Reglas de Asociación y Red Neuronal

La investigación de la exactitud de los resultados obtenidos debe leerse teniendo en cuenta dos puntos importantes.

1. Similitud de los resultados entre las particiones

Se observa que existe homogeneidad de resultados entre las distintas particiones. Esto se cumple prácticamente en la totalidad de las pruebas y de las particiones. La única excepción es la probabilidad del LIFT para el árbol de decisión (este valor representa la probabilidad de la mejora del modelo respecto al azar, para que sea válido debe ser mayor a uno y cuanto más grande mejor). Existen dos particiones de las 10 que difieren ampliamente del resto, superan apenas el valor 1 pero las otras 8 se mueven entre 10,416 y 12,628 logrando entre las 10 particiones un promedio de 11,559 con un desvío de 0,0639 que indica homogeneidad a pesar de los valores diferentes encontrados. El Log Score de

algunos modelos (como el de Bayes Naïve) presenta algunos valores distantes entre sí en algunas particiones pero observando los promedios y el desvío estándar estas diferencias se vuelven insignificantes. Por lo tanto el primer análisis dice que el conjunto de datos sobre el cual se ha realizado la validación es bueno para ejecutar la tarea. Hay homogeneidad en las distribuciones de las particiones realizadas.

2. Calidad de los resultados en base a métricas estadísticas

Prueba de clasificación

Se divide en dos opciones "PASS" y "FAIL". La primera muestra la cantidad de clasificaciones correctas para el atributo target (sin determinar un valor específico) en cada una de las particiones. Aquí se observa que los algoritmos Árbol de Decisión y Reglas de Asociación son los más compactos porque las particiones difieren en alrededor del 2% mientras que Bayes Naïve y Red Neuronal muestran diferencias entre el 5% y el 6%. Igualmente son consideradas aceptables las diferencias. La clasificación "FAIL" se comporta de la misma manera que "PASS" en los 4 algoritmos pero con porcentajes ligeramente superiores. De todas maneras lo importante es observar que siempre, todos los algoritmos clasifican más cantidad de valores como "PASS" que como "FAIL", es decir lo que clasifican bien es superior a lo que clasifican mal.

Prueba estadística de logaritmo(Likelihood Log Score)

Denominada también "puntuación del registro", esta métrica es el logaritmo de la probabilidad real de cada caso, sumada y después dividida por el número de filas del conjunto de datos de entrada. Como la probabilidad se representa como una fracción decimal, las puntuaciones del registro son siempre números negativos. Si bien se detectan valores dispares en algunos algoritmos, los promedios de esta métrica son todos valores pequeños, nunca mayores a uno, lo cual es un indicador de que el modelo supera esta prueba.

Prueba estadística de mejora respecto al modelo predictivo (Likelihood Lift)

Es un número calculado utilizando la media del logaritmo para todas las filas con valores para el atributo de destino y las probabilidades actuales y marginales. Puede obtenerse un valor positivo o negativo, pero un valor positivo significa un modelo efectivo que supera la estimación aleatoria. Los resultados para todas las particiones de todos los algoritmos son positivos, por lo tanto todos los modelos superan desde este punto de vista la estimación aleatoria.

Prueba estadística raíz cuadrada del error promedio (Root Mean Square Error)

Denominada también RMSE, es un estimador para los modelos predictivos. La puntuación calcula el promedio de los valores residuales para cada caso con el objeto de producir un único indicador del error del modelo. Cuanto menos sea la variación, más acertado será el modelo. El RMSE promedio mayor encontrado es 0,4234 y corresponde al modelo Árbol de Decisión, con un error de 0,0127. Estos valores y los menores a ellos que corresponden a los demás algoritmos indican que los modelos pueden ser acertados.

La prueba de Validación Cruzada para modelo de Clustering arroja la tabla 4.24.

Clustering				
Índice de partición	Tamaño de partición	Prueba	Medida	Valor
1	99	Clustering	Case Likelihood	0,7704
2	98	Clustering	Case Likelihood	0,715
3	98	Clustering	Case Likelihood	0,7536
4	98	Clustering	Case Likelihood	0,7592
5	98	Clustering	Case Likelihood	0,7136
6	98	Clustering	Case Likelihood	0,7905
7	98	Clustering	Case Likelihood	0,7409
8	98	Clustering	Case Likelihood	0,7626
9	98	Clustering	Case Likelihood	0,7821
10	98	Clustering	Case Likelihood	0,7176
			Promedio	0,7506
			Desviación estándar	0,0265

Tabla 4.24: Validación Cruzada Clustering

Prueba de probabilidad de los casos (Case Likelihood)

Indica la probabilidad de que un caso pertenezca a un clúster determinado. Las puntuaciones se suman y luego se dividen entre el número total de casos de manera que la puntuación es una media de la probabilidad de los casos. En la tabla precedente se observa esta probabilidad media como 0,7506 con un desvío estándar de 0,0265, por lo tanto es una probabilidad bastante cercana a uno lo cual indica que el modelo es aceptable.

4.5.2 Valoración de los resultados

Los resultados obtenidos a lo largo de todas las pruebas superan los criterios de éxito planteados y cumplen con los requisitos mínimos para la solidez, exactitud y confianza. No

hay un modelo que haya pasado todas las pruebas de manera netamente superior al resto. Todos han superado según las particularidades las distintas pruebas en forma diferente. En primer lugar el de Árbol de Decisión en segundo lugar el modelo de Bayes Naïve son los que en general logran un estándar superior pero vale destacar que por ejemplo la red Neuronal si bien mostró "flaquezas"en alguna de las pruebas es óptima en una parte del análisis de la performance. Considerando este estudio como una manera de identificar patrones, descubrir tendencias y mostrar alternativas posibles a un entrenador se consideran todos los modelos estudiados válidos para lograr el objetivo. En el deporte no es posible automatizar las tareas, dejaría de ser deporte si se pudiera, por lo tanto es suficiente y óptimo descubrir tendencias para que en la búsqueda de la anticipación de movimientos adversarios el propio equipo tome ventajas y se encuentre preparado y en condiciones de enfrentar al contrario.

4.5.3 Próximos pasos

Debido a la satisfacción con los resultados encontrados se continuará con el proyecto planteado. Se deberán elaborar informes que muestren el contenido de los modelos entrenados y sirvan de apoyo al entrenador en la toma de decisiones. El ambiente de trabajo seleccionado para llevar a cabo dicha tarea será Reporting Services. Esta herramienta es una plataforma de informes basada en un servidor que proporciona la funcionalidad completa de generación de informes para una gran variedad de orígenes de datos. Incluye una gran cantidad de herramientas para crear, administrar y entregar informes. Funciona en el entorno de Visual Studio y está totalmente integrado con las herramientas de SQL Server. Los informes pueden ser requeridos en distintos formatos, recibidos vía mail o accedidos a través de una URL. En este caso particular el usuario se conectará al servidor de informes y a través de un navegador podrá acceder a los mismos para su posterior análisis. Es importante destacar que se dejará la posibilidad de que el usuario pueda exportar los informes en diferentes formatos según su preferencia. Se confeccionará una carpeta para cada uno de los modelos realizados y allí se guardarán el o los informes necesarios para describir el contenido de cada uno de ellos.

4.6 IMPLANTACIÓN

4.6.1 Plan de Implantación

El plan de implantación que permitirá al usuario final leer los resultados hallados constará de 4 tareas:

1. Generar una carpeta en el Servidor de Informes para cada modelo confeccionado.
2. Crear consultas que muestren resultados útiles al usuario final.
3. Plasmar las consultas de contenido de cada modelo en tablas. Esta presentación será de fácil exportación en el formato preferido por el usuario (Word, Excel, Adobe Acrobat Reader, xml, html)
4. Relacionar variables de entrada con el atributo de predicción o variable objetivo. Identificar la mayor cantidad de ocurrencias de un atributo de predicción o la probabilidad de alcanzar el mismo a través de una condición en los atributos de entrada.

4.6.2 Informe final

Utilizando una conexión al servidor de informes, se pueden obtener reportes en carpetas para cada uno de los modelos estudiados. Los mismos servirán de ayuda al entrenador en las decisiones tácticas de la preparación a un partido. El entrenador puede interpretar los reportes presentados desde diferentes perspectivas. Si su objetivo es encontrar relaciones que detecten patrones de comportamiento en jugadas de primer tiempo deberá leer los mismos observando esos atributos. Si en cambio su objetivo fuera identificar fases de juego en relación a la elección del levantador se podrán analizar los datos buscando estas similitudes. Esta versatilidad muestra una vez más que la comprensión de los resultados finales es válida cuando existe un acabado y completo conocimiento del dominio del problema.

A través de Internet Explorer u otro navegador se puede acceder al servidor de informes y elegir por modelo qué es lo que se pretende observar. A modo de ejemplo se muestra la figura 4-34.

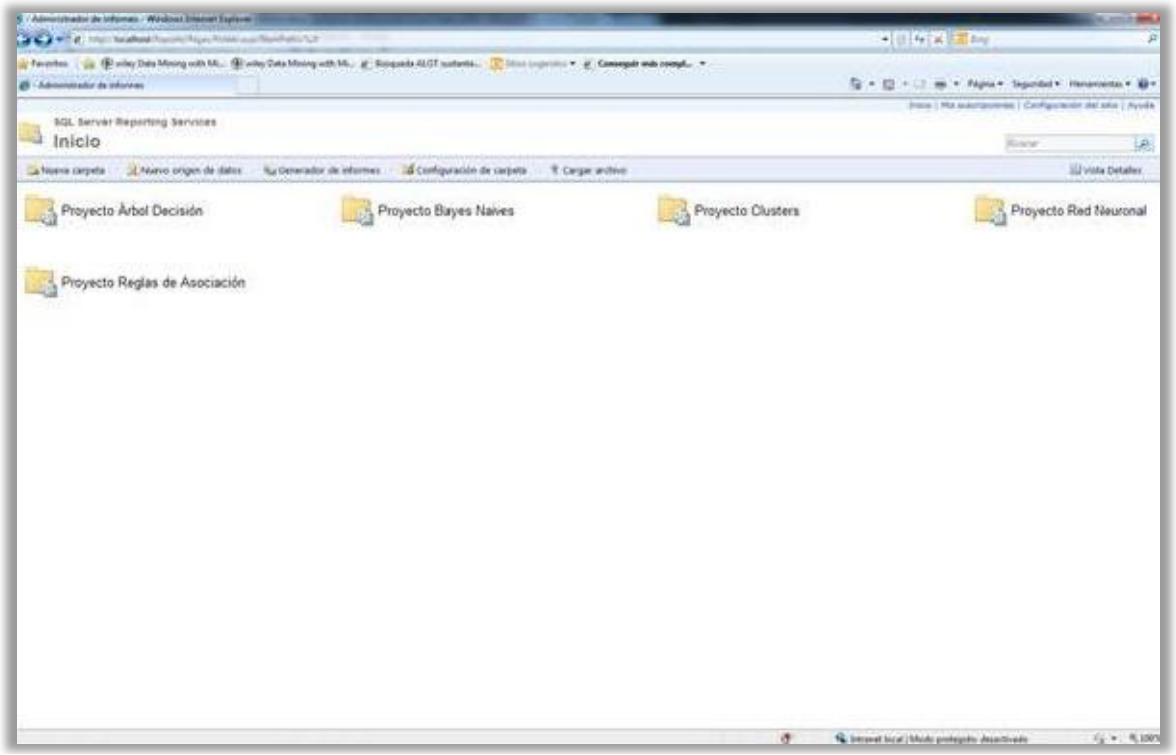


Fig. 4-34: Carpetas Servidor de Informes

4.6.2.1 Carpeta de informes Proyecto Bayes Naïve

Se confeccionó un informe general donde se brinda información sobre los datos de entrada. El mismo refleja cantidad de ocurrencia y probabilidad de cada uno de los posibles valores de entrada y salida del modelo. Se llama Support General.

Se confeccionaron además cuatro informes diferentes para cada uno de los atributos de entrada. Los mismos contienen la influencia en cuanto a cantidad de ocurrencia y probabilidad de cada uno de los posibles valores en relación a un valor del atributo de salida o variable objetivo. Los nombres de losreportes son: Fases Bayes Naïve, Evaluación Recepción, Llamada del Central y Zona Recepción.

La figura 4-35 muestra como ejemplo el informe Fases Bayes Naïve.

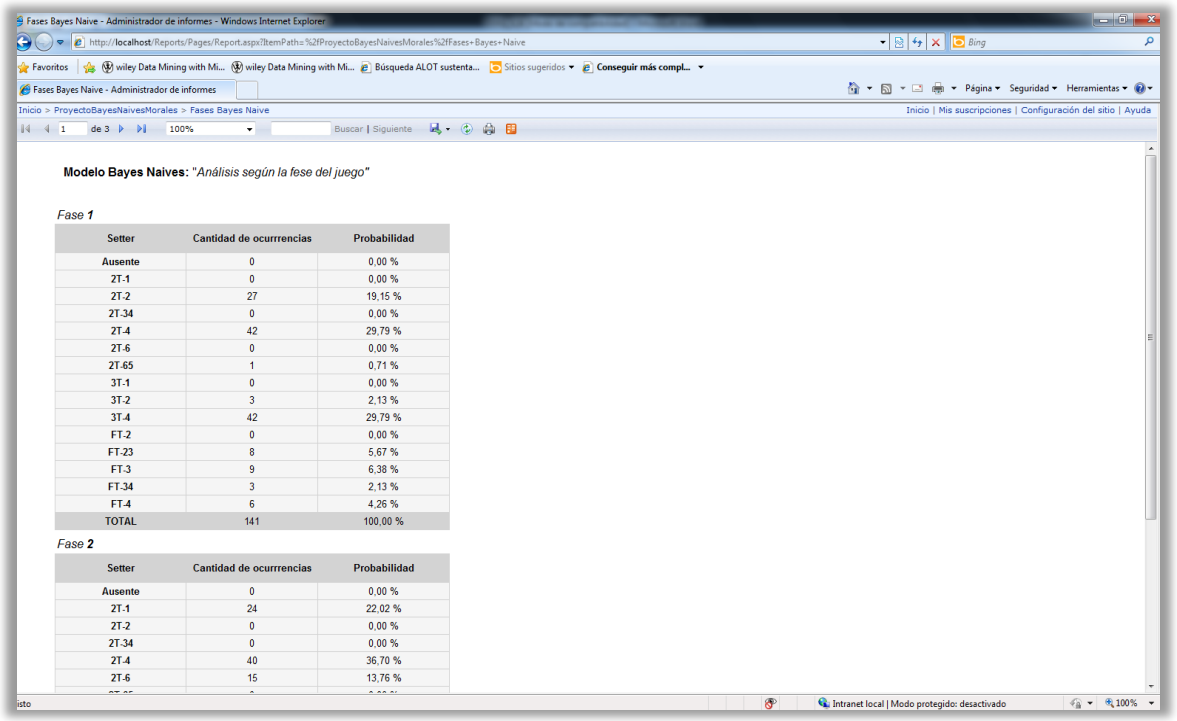
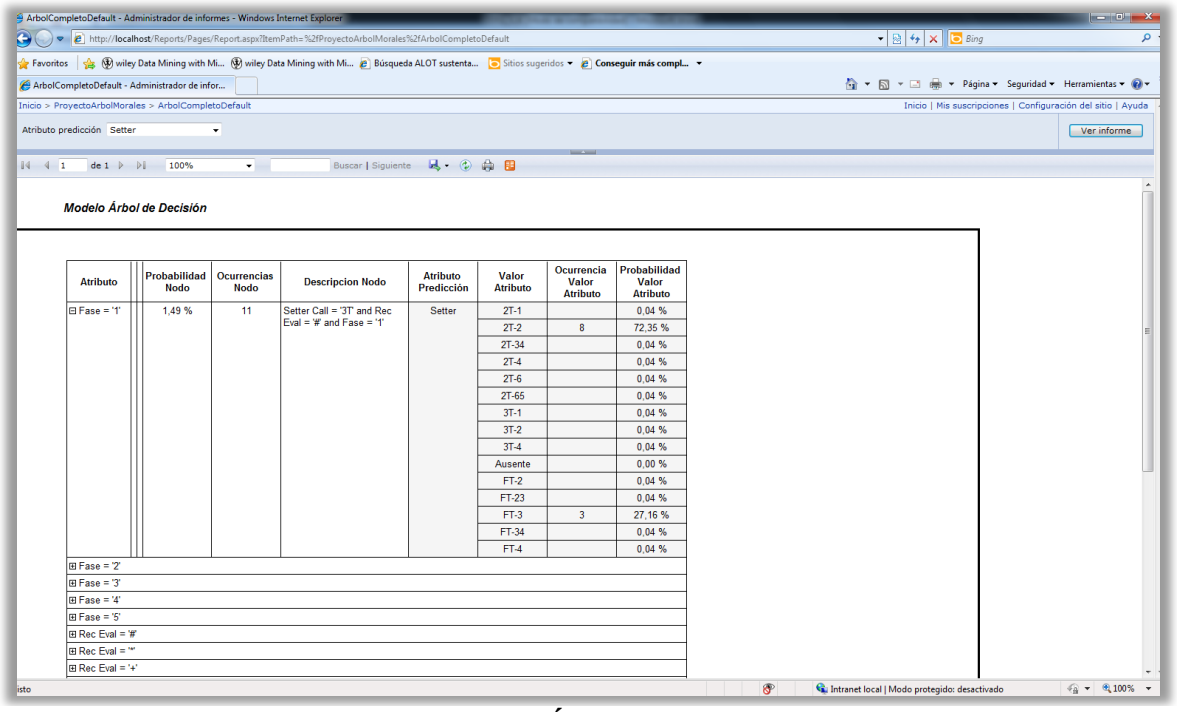


Fig.4-35: Informe Fases Bayes Naïve

4.6.2.2 Carpeta de informes Proyecto Árbol Decisión

Para el modelo Árbol de Decisión se confeccionó un único informe llamado Árbol Completo Default que brinda la posibilidad de elegir un atributo de entrada con su respectivo valor, identificar la regla asociada a ese caso y determinar la ocurrencia y probabilidad de los distintos valores del atributo de salida.

En la figura 4-36 se muestran las características del atributo Setter en la hoja del árbol de decisión identificadapor el atributo “Fase = 1”.



Atributo	Probabilidad Nodo	Ocurrencias Nodo	Descripcion Nodo	Atributo Predicción	Valor Atributo	Ocurrencia Valor Atributo	Probabilidad Valor Atributo
⊟ Fase = '1'	1,49 %	11	Setter Call = '3T' and Rec Eval = '#' and Fase = '1'	Setter	2T-1		0,04 %
					2T-2	8	72,35 %
					2T-34		0,04 %
					2T-4		0,04 %
					2T-6		0,04 %
					2T-65		0,04 %
					3T-1		0,04 %
					3T-2		0,04 %
					3T-4		0,04 %
					Ausente		0,00 %
					FT-2		0,04 %
					FT-23		0,04 %
					FT-3	3	27,16 %
					FT-34		0,04 %
					FT-4		0,04 %
⊞ Fase = '2'							
⊞ Fase = '3'							
⊞ Fase = '4'							
⊞ Fase = '5'							
⊞ Rec Eval = '#'							
⊞ Rec Eval = '*'							
⊞ Rec Eval = '+'							

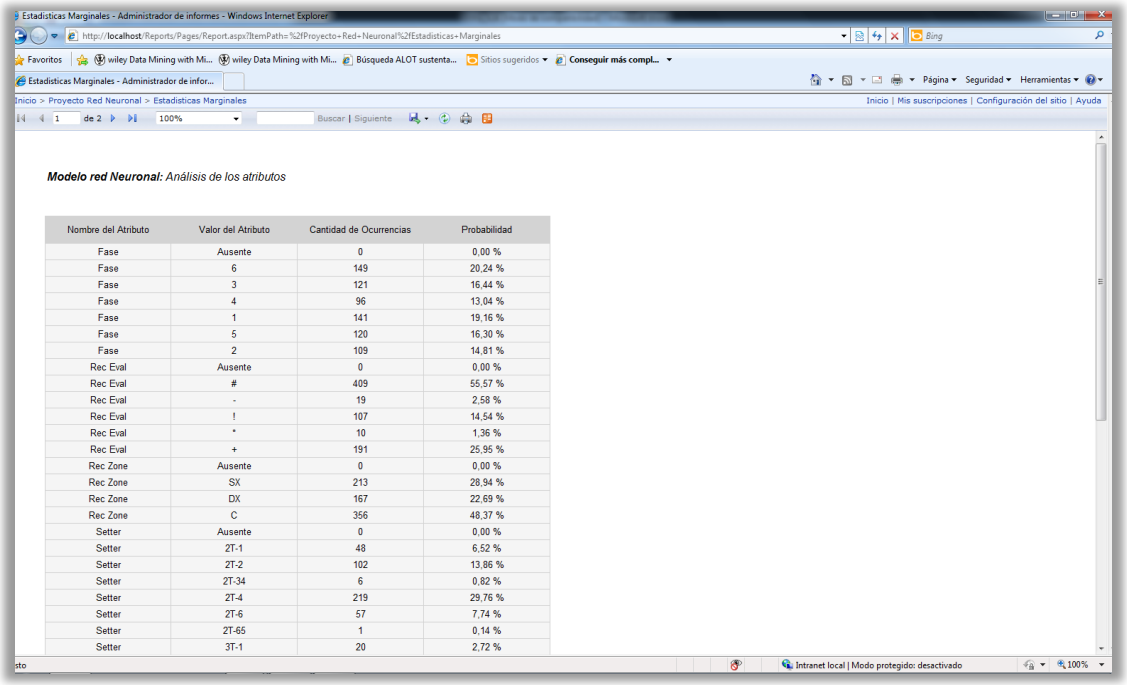
Fig. 4-36: Informe Árbol Completo Default

4.6.2.3 Carpeta de informes Proyecto Red Neuronal

Para el modelo Red Neuronal se realizó un informe denominado Estadísticas Marginales que brinda información general sobre los datos de entrada. Este informe además de mostrar al usuario final la posibilidad de conocer los datos a partir de los cuáles se realizó el estudio sirve a dar consistencia en la comparación de los modelos utilizados. Los valores obtenidos coinciden con los de las estadísticas generales obtenidas en los otros modelos.

Se generaron además cuatro informes mostrando cantidad de ocurrencia y probabilidad para cada atributo de entrada y su relación con los valores del atributo de salida. Los mismos se denominan: Evaluación Recepción, Posición Levantador, Zona Recepción, Fase, Llamada del Central.

La figura 4-37 muestra el informe general sobre las estadísticas marginales del modelo.



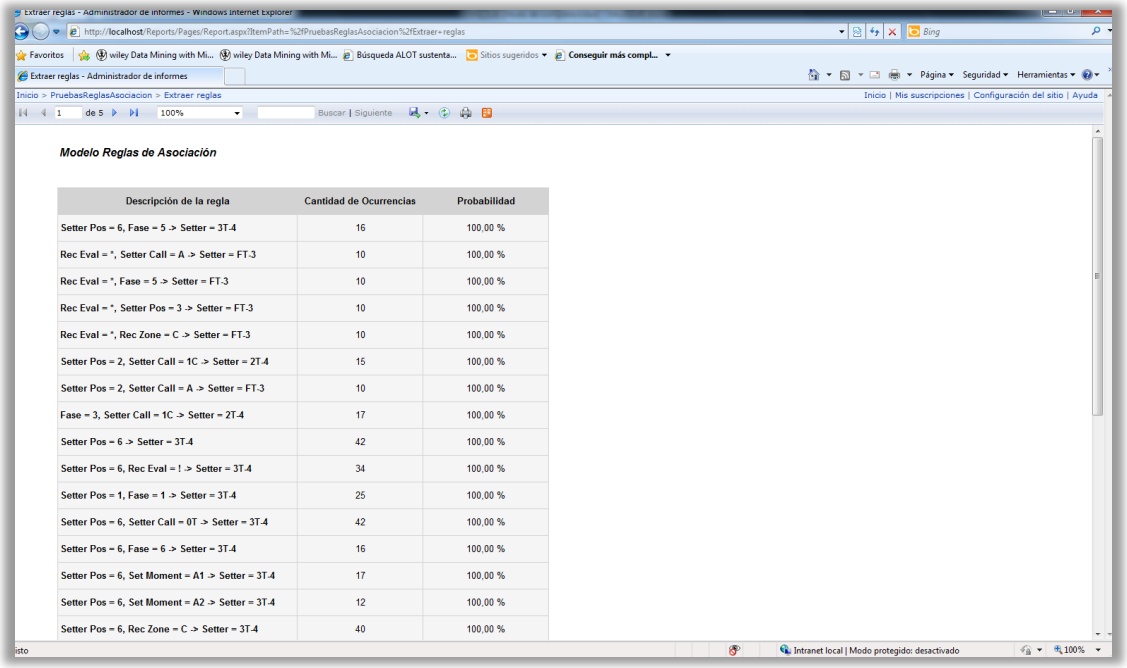
Modelo red Neuronal: Análisis de los atributos

Nombre del Atributo	Valor del Atributo	Cantidad de Ocurrencias	Probabilidad
Fase	Ausente	0	0,00 %
Fase	6	149	20,24 %
Fase	3	121	16,44 %
Fase	4	96	13,04 %
Fase	1	141	19,16 %
Fase	5	120	16,30 %
Fase	2	109	14,81 %
Rec Eval	Ausente	0	0,00 %
Rec Eval	#	409	55,57 %
Rec Eval	-	19	2,58 %
Rec Eval	!	107	14,54 %
Rec Eval	*	10	1,36 %
Rec Eval	+	191	25,95 %
Rec Zone	Ausente	0	0,00 %
Rec Zone	SX	213	28,94 %
Rec Zone	DX	167	22,69 %
Rec Zone	C	356	48,37 %
Setter	Ausente	0	0,00 %
Setter	2T-1	48	6,52 %
Setter	2T-2	102	13,86 %
Setter	2T-34	6	0,82 %
Setter	2T-4	219	29,76 %
Setter	2T-6	57	7,74 %
Setter	2T-65	1	0,14 %
Setter	3T-1	20	2,72 %

Fig.4-37: Informe Estadísticas Marginales

4.6.2.4 Carpeta de informes Proyecto Reglas de Asociación

Para el modelo Reglas de Asociación fue suficiente un único informe que presenta de manera ordenada las reglas de asociación detectadas durante la fase de entrenamiento del modelo. La probabilidad de ocurrencia de las mismas es lo que da soporte al usuario para su aceptación como patrones ciertos de comportamiento del levantador. El informe se denomina Extraer Reglas y se muestra en la figura 4-38.



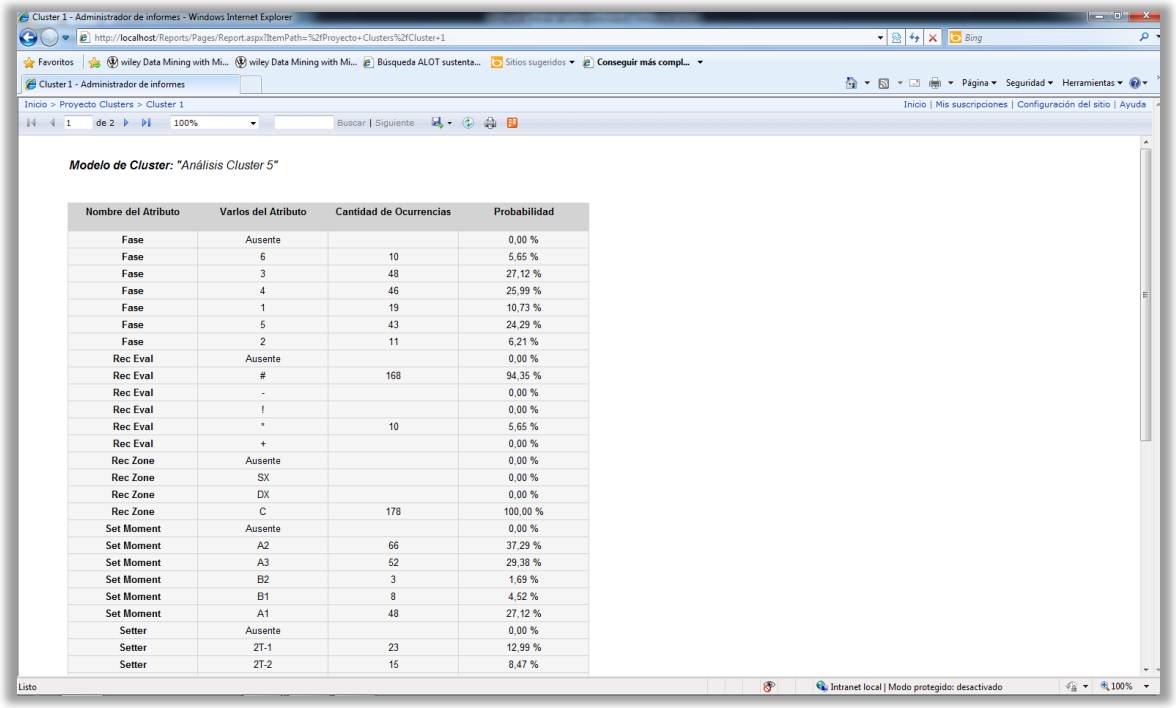
Modelo Reglas de Asociación

Descripción de la regla	Cantidad de Ocurrencias	Probabilidad
Setter Pos = 6, Fase = 5 -> Setter = 3T-4	16	100,00 %
Rec Eval = *, Setter Call = A -> Setter = FT-3	10	100,00 %
Rec Eval = *, Fase = 5 -> Setter = FT-3	10	100,00 %
Rec Eval = *, Setter Pos = 3 -> Setter = FT-3	10	100,00 %
Rec Eval = *, Rec Zone = C -> Setter = FT-3	10	100,00 %
Setter Pos = 2, Setter Call = 1C -> Setter = 2T-4	15	100,00 %
Setter Pos = 2, Setter Call = A -> Setter = FT-3	10	100,00 %
Fase = 3, Setter Call = 1C -> Setter = 2T-4	17	100,00 %
Setter Pos = 6 -> Setter = 3T-4	42	100,00 %
Setter Pos = 6, Rec Eval = ! -> Setter = 3T-4	34	100,00 %
Setter Pos = 1, Fase = 1 -> Setter = 3T-4	25	100,00 %
Setter Pos = 6, Setter Call = 0T -> Setter = 3T-4	42	100,00 %
Setter Pos = 6, Fase = 6 -> Setter = 3T-4	16	100,00 %
Setter Pos = 6, Set Moment = A1 -> Setter = 3T-4	17	100,00 %
Setter Pos = 6, Set Moment = A2 -> Setter = 3T-4	12	100,00 %
Setter Pos = 6, Rec Zone = C -> Setter = 3T-4	40	100,00 %

Fig.4-38: Informe Extraer Reglas

4.6.2.5 Carpeta de informes Proyecto Clústeres

Se confeccionó un informe para cada clúster: Clúster 1, Clúster 2 ,Clúster 3, Clúster 4, Clúster 5. En cada grupo se aglutinan características especiales que identifican la asociación entre distintos valores de los atributos. La figura 4-39 muestra el Clúster 5.



Nombre del Atributo	Varios del Atributo	Cantidad de Ocurrencias	Probabilidad
Fase	Ausente		0,00 %
Fase	6	10	5,65 %
Fase	3	48	27,12 %
Fase	4	46	25,99 %
Fase	1	19	10,73 %
Fase	5	43	24,29 %
Fase	2	11	6,21 %
Rec Eval	Ausente		0,00 %
Rec Eval	#	168	94,35 %
Rec Eval	-		0,00 %
Rec Eval	!		0,00 %
Rec Eval	*	10	5,65 %
Rec Eval	+		0,00 %
Rec Zone	Ausente		0,00 %
Rec Zone	SX		0,00 %
Rec Zone	DX		0,00 %
Rec Zone	C	178	100,00 %
Set Moment	Ausente		0,00 %
Set Moment	A2	66	37,29 %
Set Moment	A3	52	29,38 %
Set Moment	B2	3	1,69 %
Set Moment	B1	8	4,52 %
Set Moment	A1	48	27,12 %
Setter	Ausente		0,00 %
Setter	2T-1	23	12,99 %
Setter	2T-2	15	8,47 %

Fig.4-39: Informe Clúster 5

4.6.2.6 Patrones detectados

- *El variable Set Moment no adquiere importancia en los resultados observados*.

  Los modelos de Bayes Naïve, Árbol de Decisión y Reglas de Asociación no lo incluyen en las redes de dependencia halladas. Red Neuronal no lo destaca con alta puntuación en la comparación de atributos. Solamente el modelo de Clustering lo incluye en las agrupaciones pero esto es debido a la manera de trabajar del algoritmo. No es suficiente para poder ser significativo de una tendencia.

- *En Fase 1 la preferencia es jugar segundos tiempos.*

  El modelo de Bayes Naïve no arroja muestras de jugadas de primer tiempo en esta fase, lo mismo sucede en las Reglas de Asociación. El único clúster con agrupación importante en esta fase tiene como atributo descriptivo en la variable Setter jugadas de segundo y tercer tiempo. Existe una similitud entre la Red Neuronal y el Árbol de Decisión en cuanto al porcentaje hallado, el 72% de jugadas en fase 1 son de

segundo y tercer tiempo. Esto es suficiente para afirmar que son las preferidas en esta posición.

- *En fase 2 la preferencia es jugar segundos tiempos.*

  Tanto los modelos de Bayes Naïve, como Reglas de Asociación y Red Neuronal identifican esta tendencia. El clúster 2 con alta agrupación de fase 2 también define esta preferencia. Solamente el modelo de Árbol de Decisión no logra enunciar esta tendencia pero es oportuno mencionar que la hoja en la cual la Fase final es 2 ocurre solo 10 veces y eso no es suficiente para definir o contradecir un patrón de comportamiento.

- *En las fases 3 y 4 la heterogeneidad de los resultados no permite identificar patrones importantes.*

  Esta conclusión que aparentemente no aporta algo interesante puede ayudar al entrenador a posicionar el equipo en el campo de juego sabiendo que no existe una tendencia. Deberá enfrentar estas fases con jugadores capaces de anticipar la elección del levantador según otras particularidades. En un encuentro de nivel profesional, además del planteo táctico se estudian los movimientos físicos del adversario y muchas veces son esas sutilezas las que permiten anticipar una jugada.

- *En fase 5 se observa que existe una tendencia hacia jugadas de primer tiempo.*

  Esta es una característica importante que puede aportar un valor agregado al entrenador. Una de las decisiones tácticas podría indicar al equipo la anticipación de jugadas rápidas cuando el levantador adversario se encuentra en esta posición. Cabe mencionar que si bien los porcentajes que apoyan la identificación de esta característica no son elevados, variando entre el 26% y el 45% deben considerarse suficientes porque en las otras fases los mismos son nulos o no superan el 10% de las elecciones. Es necesario recordar que para que el levantador pueda jugar primeros tiempos la evaluación de la recepción debe permitírselo, si quitamos de esta fase las evaluaciones de recepción que no permiten esta elección, estamos sí ante la evidencia de que en fase 5 y cuando es posible hacerlo, el levantador prefiere jugar primeros tiempos. Esto se lee por ejemplo en la hoja del árbol con descripción "Setter Call = '3T' and Rec Eval = '#' and Fase = '5'" y un 40% de jugadas de tipo FT- 34.

- *En Fase 6 la preferencia es jugar segundo tiempo por posición dos.*

Esta observación está avalada por dos modelos que identifican claramente la preferencia de Fase 6 con la elección de tipo 2T-2. Ellos son Bayes Naïve y Reglas de Asociación. Clustering apoya la tendencia aunque divide la preferencia con primeros tiempos de tipo FT-3. Red Neuronal y Árbol de Decisión no identifican este patrón de comportamiento. Se puede decir entonces que si bien es una conclusión válida debería ser monitoreada durante el transcurso del juego para poder tomar allí la decisión correcta.

- *La llegada de la recepción por la derecha indica una definición clara de la selección de segundo tiempo por zona 4.*

  Esta tendencia se repite en todos los modelos, pero la Red Neuronal llega a calificar la misma con el 79,21% de probabilidad de ocurrencia, lo cual es realmente significativo.

- *La llegada de la recepción por los lados izquierdo y central muestra diversificación de elecciones.*

  Si bien no existe aquí una preferencia sí se define una no preferencia y es la no elección de segundo tiempo por zona 4. Se repiten las elecciones hacia otras zonas del campo como lo son zona 2, 6 y 1. También aquí aparecen elecciones de primer tiempo de tipo 3, es decir jugada abierta lejos del colocador. De estas observaciones se puede inducir que cuando el levantador recibe la pelota desde el medio del campo o desde la zona izquierda, se intenta buscar una jugada rápida abierta hacia zona 3 / 4 para poder tener la opción de la salida hacia atrás.

- *La jugada de primer tiempo preferida por el colocador Morales es la de tipo FT-3.*

  Sin dudas esta afirmación es importante para un planteo táctico, se obtiene del análisis de las jugadas donde la evaluación de la recepción es de tipo *, es decir permite solo primeros tiempos. Ante esta evidencia se elige la pelota separada del levantador, la flecha. Bayes Naïve, Red Neuronal, Árbol de Decisión y Reglas de Asociación marcan claramente esta tendencia.

- *Cuando la evaluación de la recepción es de tipo ! el juego es alto por posición 4.*

  Este patrón detectado si bien no aporta al entrenador algo desconocido o nuevo sirve a verificar el buen funcionamiento de los algoritmos. La evaluación de la recepción que no permite primeros tiempos necesariamente debe obligar la salida con juego lento y eso mismo dicen los modelos.

- *La evaluación de la recepción como perfecta define en el levantador Morales una preferencia de juego de segundo tiempo.*

  Esta conclusión puede considerarse como una característica distintiva del jugador. Hay armadores de juego arriesgado (juegan primer tiempo aún cuando las condiciones no son ideales), otros conservadores, otros con alta diversidad. Este jugador no arriesga los primeros tiempos, los utiliza cuando puede hacerlo, cuando lo permite la recepción y la ubicación en la cancha pero decididamente no los prefiere.

- *Cuando la llamada del central es de tipo A, es decir por atrás, la jugada se abre hacia zona 4.*

  Todos los modelos identifican esta tendencia, lo cual indica que se intenta llevar al bloqueo adversario hacia atrás para abrir el juego por el lado opuesto.

- *Cuando el central llama el primer tiempo de tipo FT-34 se diversifican las jugadas de esquema alternando el juego con primeros tiempos y segunda línea por posición 1 y 6.*

  Todos los modelos detectan esta característica, incluso el clúster 1 agrupa datos con esta tendencia.

- *La posición del colocador desde zonas no ideales como lo son 4, 5 y 6 inducen jugadas de tipo alto.*

  Este patrón no aporta conocimiento nuevo al entrenador pero sí una vez más la verificación de las conclusiones encontradas. Cuando el levantador está fuera de su posición ideal juega esquema alto hacia adelante. Es decir se prefieren jugadas de tipo 2T-1, 2T-5 y 3-T4 como una manera de asegurar el juego.

  Confección del informe final

  El informe final de la fase de implantación resumirá los resultados obtenidos en cada una de las funcionalidades detalladas inmediatamente antes.

4.6.3 Prueba de campo

A continuación se relata una experiencia realizada con el equipo "Las Lancheras de Cataño" que juega actualmente la Liga de Voleibol Superior Femenina (LVSF) de Puerto Rico.

En el mes de enero del año en curso inició en Puerto Rico el torneo nacional femenino de voleibol. En dicho certamen, que aún se está disputando, se encuentra como entrenador

del equipo "Las Criollas de Caguas" uno de los asesores del presente trabajo, Carlos Cardona. Su equipo se enfrentó el día 31 de enero en la ciudad de Cataño contra Las Lancheras de Cataño, el ex campeón del torneo. Si bien hacía poco que había iniciado el campeonato y no se disponía de un gran volumen de datos, tuve acceso a los mismos. Sometí el dataset entregado a los paquetes de Integration Services diseñados en el presente proyecto. Posteriormente se entrenaron los modelos de Bayes Naïve, Árbol de Decisión, Reglas de Asociación, Clúster y Red Neuronal. Se lograron identificar algunas tendencias de juego que se repetían como resultado del comportamiento de la levantadora del equipo de Cataño. No fueron muchas las conclusiones halladas, se detectó fundamentalmente que en fase 6 se preferían jugadas de primer tiempo, mientras que en fase 4 la elección mayoritaria caía sobre jugadas de segundo tiempo. Considerando el reducido volumen de datos se consideraron suficientes los patrones encontrados. Teniendo la posibilidad de concurrir personalmente a ver el partido pude comprobar favorablemente que la levantadora del equipo de Cataño se comportaba según los patrones detectados. Es decir, cuando se hallaba en fase 6 su preferencia era sistemáticamente el primer tiempo. El equipo de Caguas se posicionó en la cancha tácticamente para anticipar esa jugada. Algunas veces pudo hacerlo otras no. El resultado del partido no fue favorable a Caguas, pero hubo una revancha, en el mes de marzo, hubo tiempo de entrenar y posicionar al equipo para contrarrestar al adversario y finalmente se logró la victoria el día 9 de marzo en la cancha de Caguas. El campeonato está aún en curso, ambos equipos deberán volver a enfrentarse en las fases finales.

Es importante destacar lo siguiente: la herramienta fue creada sobre un dataset perteneciente a dos equipos concretos de varones; una vez puesta a punto, fue utilizada exitosamente sobre un dataset distinto, perteneciente a dos equipos que no eran los originales, y no eran de varones sino de mujeres.

4.6.4 Revisión del proyecto

Con respecto a los objetivos y la delimitación originales del proyecto, en este momento no se considera necesario agregar o modificar algún detalle. Esto seguramente se debe a la colaboración y retroalimentación permanente por parte del usuario final. Se han realizado iteraciones constantes que en este caso particular sirvieron para afianzar las destrezas en la elaboración de proyectos de minería de datos. Los modelos han sido probados, evaluados y modificados tratando de encontrar aquellos que en su definición de parámetros, estructura

y funcionamiento tenían un comportamiento ideal para el caso en estudio. Se da por concluido el proyecto en su especificación actual y se encuentran totalmente satisfactorios los resultados hallados.

5 CONCLUSIÓN

Se ha desarrollado una herramienta que permite detectar patrones en la conducta de los levantadores, y usar esos patrones para diseñar estrategias proactivas. La herramienta se ha probado exitosamente en el campo.

Existen importantes similitudes en los patrones arrojados en los diferentes modelos. Pueden calificarse los resultados hallados como coherentes. Los datos fueron sometidos a distintos algoritmos y todos identifican aproximadamente los mismos patrones. Esto sin dudas está diciendo que se ha alcanzado el objetivo propuesto. Fue posible relacionar atributos de entrada para identificar tendencias en un atributo de salida o target. Se logró verificar que los patrones encontrados ayudan a un entrenador en el planteo táctico del juego.

La evaluación de los modelos expresa además una conclusión que traspasa el objetivo propuesto. Para identificar patrones de comportamiento en el voleibol profesional puede ser suficiente la creación de tres modelos: Bayes Naïve, Árbol de Decisión y Reglas de Asociación. Estos tres superan al resto de los estudiados por su eficiencia y facilidad de uso. Los dos primeros se eligen porque ambos aportan la mayoría de los patrones que luego confirman los otros modelos y porque las pruebas de validación de los mismos son ampliamente superadas. Los dos se alejan notoriamente de predicciones aleatorias acercándose al comportamiento de un modelo ideal. El tercer modelo, Reglas de Asociación, si bien no es tan eficiente como los dos anteriores, es simple de entender y sobre todo práctico en su uso debido a la claridad con que arroja las conclusiones. El mismo necesita sin embargo del apoyo de los dos primeros.

El estudio puede y debe seguir, deben analizarse tendencias en los demás roles del juego, deben relacionarse los mismos y recién allí se estará en presencia de una emulación del conocimiento que permita definir un planteo táctico íntegro y completo. El mismo deberá ser entrenado y posteriormente defendido en la cancha para lograr lo que el deporte profesional busca por sobre todas las cosas “ganar el partido”.

6 Referencias Bibliográficas

(1) O. Kaplan, *Voleibol Actual técnica-táctica Entrenamiento*, Stadium, Buenos Aires, 1982.

(2) J.F. Ferrarese, *El Voleibol*, Emograph S.A., España, 1976.

(3) J. Hegedus, *La planificación del entrenamiento deportivo*, Stadium, Buenos Aires 2009.

(4) A. V. Ivoilov, *Voleibol Técnica-Táctica-Entrenamiento*, Stadium, Buenos Aires 1986.

(5) J. Ramos Leiva, *Táctica en el Voleibol* [online] Recuperado el: 12 abril 2012 Disponible en: http://www.ligavallecaucanadevoleibol.com/PISO/tacticas.htm

(6) O. L. Villamea, *El uso de la estadística en el voleibol* [online] Recuperado el : 12 de abril 2012 disponible en :http://www.efdeportes.com/efd9/voley9.htm

(7) M. de la Puente, *Minería de datos* [online] Recuperado: 12 de abril 2012 disponible en http://profesionalesdecienciasdelainformacion.wordpress.com/2010/05/04/mineria-de-datos/

(8) D. Zaratarelli, *¿In Corpore Sano?*, Lumen, Buenos Aires, 2009.

(9) J.R. López, *Deporte y Ciencia. Teoría de la actividad Física*, INDE Publicaciones, España, 2003.

(10) D. Zavatarelli, *¿In Corpore Sano?*, Lumen, Buenos Aires, 2009.

(11) D. Zavatarelli, *¿In Corpore Sano?*, Lumen, Buenos Aires, 2009.

(12) J.R. López, *Deporte y Ciencia. Teoría de la actividad Física*, INO Reproducciones S.A., Zaragoza, 2003.

(13) Marianne Fiedler, *Voleibol Moderno*, Stadium, Buenos Aires, 1982. Reimpresión alemana.

(14) T. Guterman, *Informática y Deporte*, INO Reproducciones S.A., Zaragoza, 1998.

7 BIBLIOGRAFÍA

A.Rivera, Implementando Minería de datos desde Excel, 2009 [online], Recuperado: 4 abril 2012, Disponible en: http://www.youtube.com/watch?v=43mS9M1ZjU8

Bemagualli, *El Proceso de la Minería de datos* [online] Recuperado: 4 abril 2012, Disponible en: http://www.slideshare.net/bemaguali/mineria-de-datos-1867890

B.Brown, *Volleyball Scores*, 2009 [online] Recuperado: 4 abril 2012, Disponible en: http://www.volleyballscores.co.uk/about-this-site/datavolley.html

Congreso Internacional sobre Entrenamiento Deportivo (8va, 2002 Valladolid, España), *Incidencia del rendimiento de lso complejos de juego por rotaciones sobre la clasificación final de los JJOO de Sidney 2000* [online}, Recuperado: 4 abril 2012, Disponible en: http://www.docstoc.com/docs/33012329/VIII-CONGRESO-INTERNACIONAL-SOBRE-ENTRENAMIENTO

Data Video 2007 Professional(2007) HandBooks Data Project, sport software, Bologna, Italia.

Data Volley 2007 Professional (2007) HandBooks Data Project, sport software, Bologna, Italia.

Fédération Internationale de VolleyBall, Refereeing/Rules [online], Recuperado: 4 abril 2012, Disponible en: http://www.fivb.org/EN/Refereeing-Rules/

Fédération Internationale de VolleyBall (2003) *Top Volley: El Juego masculino: Técnica y Táctica* [CD-ROM] Lausana, Suiza.

Fédération Internationale de VolleyBall (2006) VIS, *Volleyball Information System, Guidelines for the preparation of the VIS personnel.* Lausana, Suiza

Jhon Kessel (2012) *Jim Coleman impact & Immediate Feedback Video System* [online] Recuperado el: 4 abril 2012 Disponible en: http://usavolleyball.org/blogs/growing-the-game-together-blog/posts/3546-jim-coleman-impact-immediate-feedback-video-system#6502

J. MacLennan, Z. Tang, B. Crivat, *Data Mining with Microsoft SQL Server 2008*, Wiley publishing, Inc, Indianápolis, Indiana, 2009.

M. Estévez et al, *La Investigación Científica en la Actividad Física: Su Metodología, Deportes*, La Habana, Cuba, 2004.

M. Pavilik, "La Eficacia del armador y del atacante del 1er tiempo", *Los elementos avanzados del juego*. Asociación de entrenadores de Voleibol de Estados Unidos, 2002.

MSDN Library, *Conceptos de Minería de datos*, 2012 [online], Recuperado 4 abril 2012, Disponible en: http://msdn.microsoft.com/es-es/library/ms174949.aspx

Ms. SQL Server, *Libros en pantalla de SQL Server 2008 R2*, 2012, [online] Recuperado el 4 abril 2012, Disponible en: http://technet.microsoft.com/es-es/library/ms130214(v=sql.105).aspx

R. I. Levin, D. S. Rubbin *Estadística para Administración y Economía* , Pearson Education, México, 2004.

R. S. Pressman *Ingeniería del Software Un enfoque práctico*, McGraw-Hill, México, 2010.

S. Calero Morales *"Las Relaciones entre fundamentos del voleibol. Aspecto Básico que determina la cantidad e influencia de las acciones técnico-tácticas en el rendimiento final"*, EfeDeportes, Año 13, Nro 129, Buenos Aires, Febrero 2009 [online], Recuperado: 4 de abril 2012, Disponible en: http://www.efdeportes.com/efd129/voleibol-influencia-de-las-acciones-tecnico-tacticas-en-el-rendimiento-final.htm

S. Calero Morales, A. Fernandez Lorenzo, R. R. Fernandez Concepción *"Estudio de Variables clave para el análisis del control del rendimiento técnico-táctico del voleibol de alto nivel",* EfeDeportes, Año 13, Nro 121, Buenos Aires, Junio 2008 [online], Recuperado: 4 de abril 2012, Disponible en: http://www.efdeportes.com/efd121/control-del-rendimiento-tecnico-tactico-del-voleibol.htm

8 GLOSARIO

DX, Dx: Derecha. Indica que la recepción se realiza desde la zona derecha del campo.

Aproximadamente desde zona 1.

SX, Sx: Izquierda. Indica que la recepción se realiza desde la zona izquierda del campo.

Aproximadamente desde zona 5.

C: Centro. Indica que la recepción se realiza desde la zona central del campo.

FT: Juego o jugada de primer tiempo. Juego rápido.

2T: Juego o jugada de segundo tiempo.

3T: Juego o jugada de tercer tiempo. Juego Alto.

FT-(número): Posición de la cancha desde donde se realiza el tipo de juego. Ej. FT-23 indica un juego de primer tiempo realizado entre las zonas 2 y 3 del campo.