UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Redes Neuronales y Análisis Multivariado para pronosticar resultados de la NFL

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

MATEMÁTICO

P R E S E N T A:

ISRAEL ALDANA GALVÁN

DIRECTOR DE TESIS:
DRA. RUTH SELENE FUENTES GARCÍA

2010

Redes Neuronales y Análisis Multivariado para pronosticar resultados de la NFL

Israel Aldana Galván

Febrero-2010

Agradecimientos

A mis padres, porque siempre me han apoyado de manera incondicional.
A mi hermana, por sus interesantes puntos de vista.
A mi asesora, por apoyarme y guiarme en la elaboración de este trabajo.
A mis sinodales por sus valiosas aportaciones.
Especialmente agradezco y dedico este trabajo a Gina, porque todo este tiempo ha sido mi pareja, mi compañera, mi inspiración y motor para seguir adelante.

Índice general

Índice de figuras

Índice de cuadros

Introducción

Planteamiento del problema

En la actualidad la información es un recurso imprescindible en cualquier rama de la actividad humana, es importante para poder tomar decisiones adecuadas que nos permitan avanzar y asegurar el cumplimiento de los objetivos. Por supuesto conocer información acerca del futuro sería maravilloso, nos permitiría estar preparados para éste. Considerando diversas restricciones existen herramientas que nos ayudan a tener un aproximado del futuro. Un área en la que son importantes las predicciones sobre el futuro son los deportes profesionales, ya sea como fanático, como analista del mismo, como dueño de un casino o incluso como apostador, como ejemplo podemos mencionar el futbol americano, ya que éste representa un negocio multimillonario no sólo porque genera una gran cantidad de recursos financieros gracias a patrocinios, venta de derechos para la TV, venta de souvenirs, etc. También es un pilar en el mercado de Las Vegas a nivel de apuestas, donde en los casinos y establecimientos parecidos se apuestan grandes cantidades de dinero. En México ésta es un área poco explotada, pero a nivel mundial los casinos en general generan ingresos por alrededor de 450 y 500 billones de dólares. Cifra equivalente a la que genera la industria petrolera, en E.U. (Estados Unidos de América) los casinos generan alrededor de entre 40 y 50 billones de dólares, de los cuales aproximadamente 10 billones son destinados al pago salarios, está de más mencionar que ésta es un área de gran potencial para generar los empleos que tanto requiere México.

En la determinación de los momios[1], es importante utilizar técnicas de diversas áreas, que permitan generar modelos capaces de predecir los resultados de los encuentros entre los equipos, sin embargo el detalle de estos modelos así como de las herramientas que utilizan, generalmente son guardados como secretos de las compañías dedicadas a esto. El futbol americano profesional tiene la característica de ser un deporte que genera un gran número de estadísticas, por lo que representa la oportunidad ideal para aplicar herramientas como SOM (*Self Organized Map*), componenetes principales, análisis de conglomerados, y hacer predicciones sobre los resultados del mismo. De acuerdo con los expertos del área, el futbol americano es uno de los deportes en el que las sorpresas aunque llegan a pasar no son tan frecuentes como en otros deportes, por lo que podríamos preguntarnos si los resultados se

[1]El momio es la probabilidad aplicable para calcular el pago de una apuesta ganadora.

pueden predecir o al menos se puede reducir la incertidumbre de éstos.

En el presente trabajo se pretende construir un modelo capaz de predecir los resultados de encuentros del futbol americano profesional, la idea es mejorar las predicciones que se obtendrían si éstas fueran hechas al azar, aunque sería bueno poder garantizar una probabilidad de éxito mayor, el presente trabajo es de carácter exploratorio por lo que en principio se verá que resultados arroja el mismo, sin garantizar una probabilidad de éxito mínima.

En el presente trabajo se consideraron únicamente cinco temporadas de la NFL (2003 a 2007), se toman cinco temporadas ya que en general cuando un entrenador llega a un equipo, promete que éste alcanzará sus objetivos en un promedio de tres años, la experiencia nos dice que ésto lo consiguen en aproximadamente cinco años, ahora se se consideró a partir de la última temporada completa en ese momento y cinco temporadas hacia atrás.

La idea principal en la construcción del modelo es hacerlo lo mas sencillo posible, de manera que tenga el menor número de restricciones y pensando en que sirva de base a otros modelos más complicados.

Problema

El presente trabajo pretende proponer un modelo que permita predecir el resultado gana, pierde o empatan de los encuentros de post-temporada[2] del futbol americano profesional de la NFL (*National Football League*).

Objetivo general

El obejetivo general del presente trabajo es generar un modelo capaz de predecir el resultado de los encuentros de la post-temporada del futbol americano profesional de la NFL utilizando herramientas de redes neuronales y análisis multivariado.

Obejtivos específicos

Los objetivos específicos del presente trabajo son:

- Generar un modelo de predicción utilizando componentes princiaples[3], SOM[4] y k-medias[5] como herramientas.
- Poder hacer predicciones sobre el resultado de los juegos de post-temporada de las temporadas 2003 a 2007 del futbol americano profesional de la NFL.

[2] Juegos finales, equivalentes a octavos de final.
[3] Herramienta de análisis multivariado.
[4] (*Self Organized Map*) herramienta de agrupación de redes neuronales.
[5] Herramienta de agrupación de análisis multivariado.

- Proponer un método de evaluación del modelo.
- Mostrar la utilidad de las redes neuronales en combinación con el análisis multivariado en un área diferente a aquellas en las que comúnmente son utilizadas.

Descripción

El presente trabajo se desarrolló en varios capítulos que facilitan la comprensión del mismo, a continuación se describen éstos:

- En el capítulo (1) se describen las reglas del futbol americano profesional, también se describe la estructura actual de la NFL y finalmente se indica la teoría de todas las herramientas utilizadas por el modelo.
- En el capítulo (2), se hace una descripción del modelo, en él se presenta su estructura y funcionamiento, así como la forma y objetivo que cumple cada una de las herramientas utilizadas por el mismo.
- En el capítulo (3), se presentan los resultados del modelo aplicados a cinco temporadas de la NFL, cubriendo las temporadas 2003 a 2007, asimismo se analizan los resultados obtenidos por el mismo.
- A continuación se presentan las conclusiones sobre el modelo, así como las propuestas para futuras mejoras del mismo.
- Finalmente los anexos incluyen información variada y valiosa, que va desde los programas en R utilizados por el modelo, hasta los datos utilizados por el mismo.

Capítulo 1

Marco Teórico

En el presente capítulo se hace una descripción del futbol americano, dónde se presentan sus reglas, así como la descripción de la estructura de la NFL, a continuación se describen las herramientas utilizadas a lo largo del trabajo que permitirán al lector conocer o recordar las bases teóricas de las mismas. Comenzando con las **componentes principales**, herramienta de análisis multivariado, utilizada en este trabajo con el propósito de reducir la dimensión del conjunto de variabes utilizadas, para poder hacer el modelo más manejable, a continuación se describen los **mapas auto-organizados (SOM),** herramienta de redes neuronales, utilizada en el presente trabajo con el propósito de generar conglomerados de los equipos de futbol americano de acuerdo a sus características definidas por las varibales previamente seleccionadas, después se describe el **análisis de conglomerados de *k*-medias**, herramienta de análisis multivariado, utilizada también para generar conglomerados de los equipos, a continuación se describe un **método para determinar el número de grupos en un conjunto de datos**, el número que proporciona sirve como parámetro tanto para el análisis de conglomerados como para los mapas auto-organizados, finalmente se decribe el **estimador de máxima verosimilitud** utilizado para estimar la probabilidad de éxito del modelo en la predicción individual del resultado de un encuentro.

1.1. El futbol americano profesional de la NFL (*National Football League*)

En la presente sección se explicarán las principales reglas del futbol americano profesional, las cuales permitirán al lector enteder la filosofía del juego, también serán útiles para conocer el propósito de las estadísticas que se utilizarán en capítulos posteriores.

Un partido de futbol americano empieza con el lanzamiento de una moneda por parte del árbitro principal en presencia de los capitanes de los respectivos equipos. El equipo que gana el lanzamiento de la moneda tiene derecho a elegir entre empezar el partido atacando o defendiendo. Llegados a este punto los dos equipos se colocan sobre el campo y se inicia el

partido, el objetivo del equipo que ataca es ir avanzando yardas hasta la zona de anotación que defiende el equipo contrario. El sistema por el cual la ofensiva consigue mantener la posesión del balón e ir avanzando yardas sobre el campo, es lo que ha convertido a este deporte en una auténtica guerra táctica y mental entre los dos equipos que compiten. Este sistema se llama el *Down System,*[1] consiste en que el equipo que ataca dispone de 4 jugadas o intentos para avanzar un mínimo de 10 yardas, en caso de conseguirlo se le conceden otras 4 oportunidades y así sucesivamente. En el caso de que no consigan avanzar 10 yardas en 4 intentos, la posesión del balón pasa directamente al equipo contrario. Este sistema es lo que hace que el futbol americano sea estratégicamente tan complejo ya que el ataque siempre tiene la posesión del balón antes de cada oportunidad, al finalizar ésta el partido se detiene y los equipos vuelven a formar ordenadamente para iniciar la siguiente oportunidad, lo cuál permite a los entrenadores planificar con total precisión la siguiente acción.

A continuación se explica paso a paso como transcurre cada oportunidad: primero el ataque se reune sobre el terreno de juego y uno de los jugadores explica a sus compañeros la jugada que el entrenador ha decidido poner en práctica, la jugada se inicia con el centro del balón, un jugador llamado centro toma el balón con una mano y lo pasa entre sus piernas a un compañero situado detrás[2], una vez finalizada la jugada[3] el partido se detiene y se vuelve a repetir el mismo proceso desde el punto de máximo avance del balón.

Resumiendo: la ofensiva dispone de 4 jugadas para avanzar un mínimo de 10 yardas, en caso de conseguirlo se les concede otros 4 intentos y así sucesivamente. En el caso de que no logren avanzar estas 10 yardas después de los 4 intentos, la posesión del balón pasa directamente al equipo contrario.

1.1.1. Los jugadores y posiciones

El futbol americano se juega entre dos equipos, cada uno de los cuales tiene once jugadores en el campo que intentan avanzar con la pelota hacia la zona de fondo del equipo contrario. Durante un partido los equipos son designados como equipo ofensivo[4] y equipo defensivo[5].

1.1.2. Equipo ofensivo

Los once jugadores del equipo ofensivo están divididos en dos grupos: siete hombres de línea, que juegan en la línea de golpeo[6] y otro grupo de cuatro jugadores, llamados *backs*,

[1] No tiene traducción directa al castellano así que lo traduciremos como **sistema por intentos o jugadas.**

[2] Todas las jugadas empiezan de esta manera.

[3] Generalmente se da por finalizada la jugada cuando el portador del balón es tacleado o cuando se falla un intento de pase.

[4] El equipo que tiene el balón.

[5] El equipo que defiende la línea de gol contra el equipo ofensivo

[6] Una línea imaginaria que marca la posición del balón.

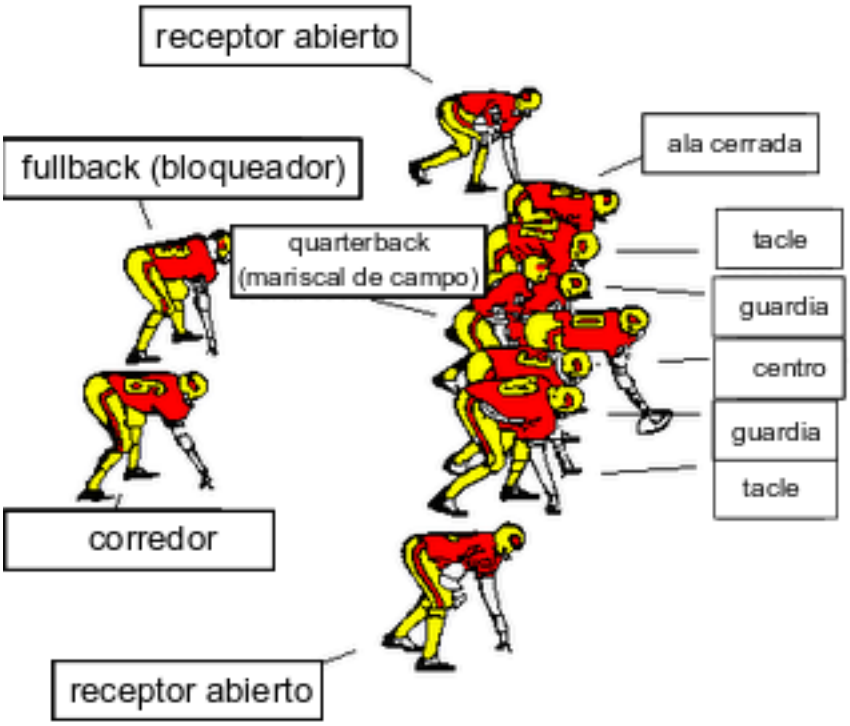

Figura 1.1: Formación de la ofensiva

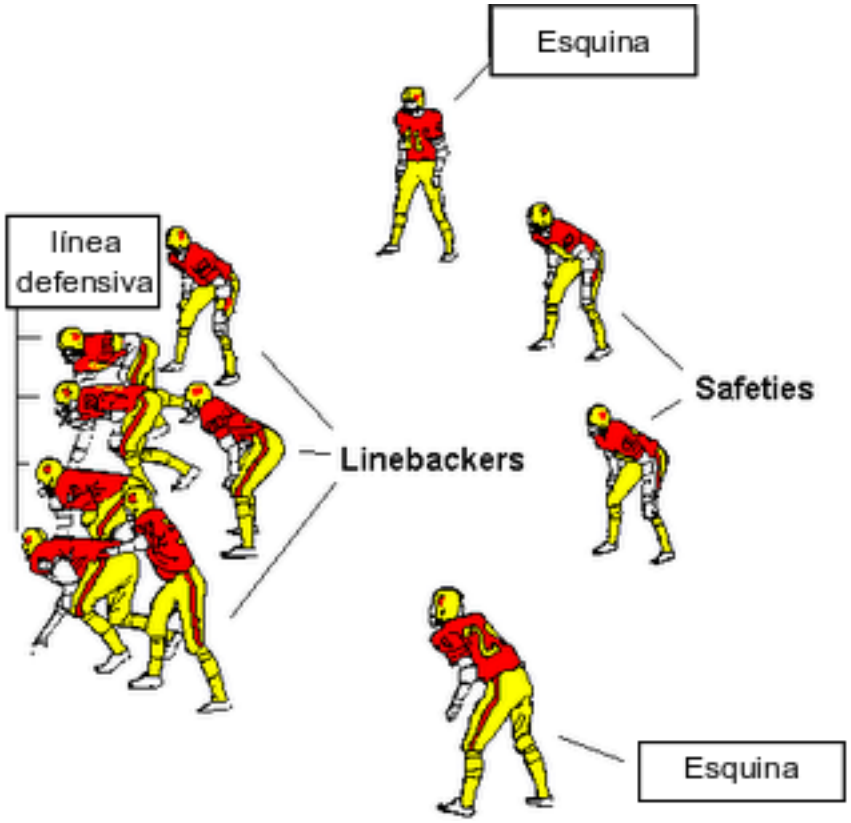

Figura 1.2: Formación de la defensiva

que están situados en diversas posiciones, detrás de los hombres de línea. El jugador situado en el centro de la línea se llama centro. A su izquierda está el guardia izquierdo y a su derecha el guardia derecho. A la izquierda del guardia izquierdo se sitúa el tacle izquierdo y a la derecha del guardia derecho el tacle derecho; de forma similar, en los extremos de la línea se encuentran los defensas laterales. El *back*, que por lo general se sitúa siempre detrás del centro y dirige el juego del equipo ofensivo se le conoce como el *quarterback*. Los equipos a menudo utilizan receptores abiertos en las esquinas, éstos sitúan en la línea de golpeo pero separados del resto de la formación. La formación típica de la ofensiva en un partido de futbol americano se puede ver en la figura (1.1)

1.1.3. Equipo defensivo

El equipo defensivo se compone de una fila de hombres de línea, que comprende la línea defensiva, una fila de *linebackers* y varios *backs* defensivos, conocidos como segunda línea. La línea defensiva puede usar cualquier número de jugadores, aunque la mayoría de los equipos usan tres o cuatro hombres de línea. Los hombres de línea defensivos son

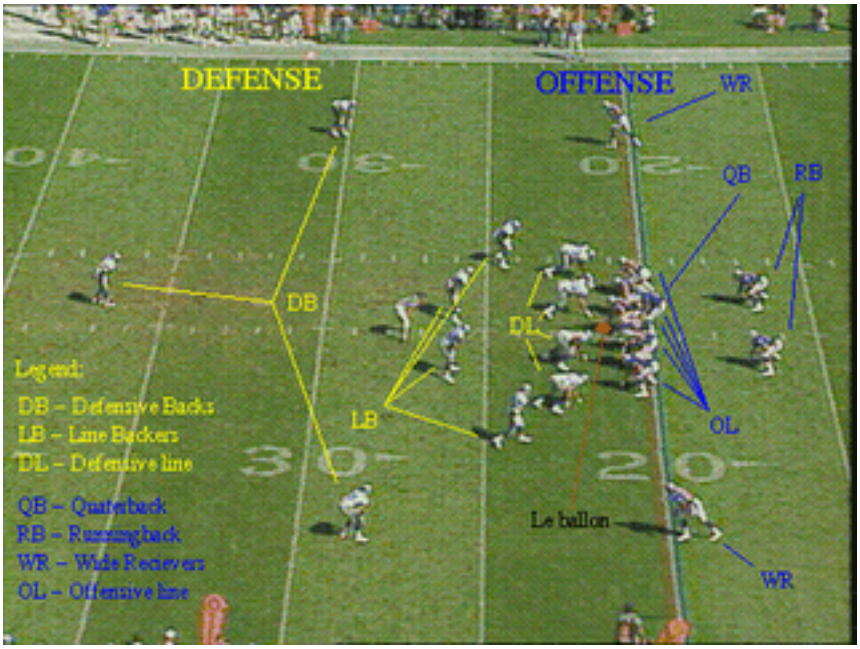

Figura 1.3: Formación típica

responsables de parar el ataque del equipo contrario en situaciones de pases, presionando al *quarterback*. Los *linebackers* se sitúan detrás de la línea defensiva, y dependiendo de la situación, son usados para detener a los corredores, presionar al *quarterback*, o cubrir a los receptores del equipo oponente. Los equipos utilizan como norma tres o cuatro *linebackers*. La segunda línea la componen los esquinas, que cubren a los receptores abiertos, y los *safeties*, que cubren a los receptores, ofrecen apoyo para detener a los atacantes y presionan al *quarterback*. La segunda línea se compone normalmente de dos esquinas y dos *safeties*. La formación típica de la defensiva en un partido de futbol americano se puede ver en la figura (1.2). La formación típica de la ofensiva y defensiva en un partido de futbol Americano se puede ver en la figura (1.3)

1.1.4. Equipos especiales

Los jugadores cuyo trabajo consiste en dar patadas a la pelota son conocidos como equipo especial, se desempeñan en la patada inicial, goles de campo[7] y despejes[8].

[7]Cuando un equipo ofensivo se encuentra en su última oportunidad pero ésta relativamente cerca de la zona de anotación del equipo contrario intenta partear el balón entre los postes del equipo contrario, si lo consigue entonces gana tres puntos para su causa.

[8]Cuando el equipo ofensivo está en su última oportunidad y se encuentra lejos de la marca para otras cuatro oportunidades o se encuentra lejos de la zona de anotación del equipo contrario, entonces patea el balón lo más lejos posible de su propia zona de anotación, de tal forma que el equipo contrario tenga que

1.1.5. Tiempo de juego

Un partido de futbol americano está dividido en cuatro periodos llamados cuartos, de quince minutos de tiempo de juego cada uno. Los dos primeros cuartos constituyen el primer tiempo, los dos segundos, el segundo. Entre los dos tiempos hay un periodo de descanso de quince minutos durante el cual los jugadores pueden abandonar el terreno de juego. Los equipos intercambian las mitades del campo al final de cada cuarto. El reloj se detiene al final de cada cuarto y en algunas ocasiones, cuando ocurren eventos particulares o cuando lo deciden los árbitros. Los equipos disponen de 3 tiempos fuera por mitad, los cuales paran automáticamente el reloj de juego. También el reloj se para cuando un pase es incompleto o cuando el jugador que porta el balón se sale por las bandas laterales.

1.1.6. El campo de juego y el balón

El campo de juego del futbol americano es un rectángulo de 110 m de largo por 48.9 m de ancho. En ambos extremos de su longitud, unas líneas blancas, llamadas líneas de meta o de gol, marcan las entradas a las zonas de fondo, que tienen 9 m y que son defendidas por cada equipo. Un equipo que pretenda anotar debe llevar, pasar o golpear la pelota hacia dentro de la zona de fondo situada en el área del campo del equipo adversario. Varias líneas paralelas a la zona de fondo cruzan el campo a intervalos de 4.5 m y dan al campo el aspecto de una parrilla. Otro juego de líneas, conocidas como líneas de banda, corren a lo largo de ambos lados del terreno de juego. Además, dos franjas de líneas llamadas hash marks, corren paralelas a las líneas de banda y, en la Liga Nacional de futbol (NFL), están a 21.6 m de cada línea de banda. Cada jugada debe comenzar en, o entre, estas líneas. Antes de cada jugada, los árbitros sitúan la pelota, bien entre las *hash marks*, o bien en la *hash mark* más cercana del final de la jugada anterior. Situados en el centro de la línea de atrás de cada zona de fondo están los postes de meta, que se componen de un poste vertical de 3 m, por encima del cual atraviesa una barra horizontal, de cuyos extremos se extienden hacia arriba otros postes, que tienen 5.6 m de separación. La pelota es de goma inflada y está recubierta de cuero o goma. Tiene la forma de un esferoide alargado, con una circunferencia de 72.4 cm alrededor del eje largo y 54 cm alrededor del eje corto; tiene entre 397 y 425 g de peso. En las figuras (1.4) y (1.5) se puede ver un campo de futbol americano clásico.

1.2. La NFL

La NFL es uno de los espectáculos deportivos mas vistos en el planeta, está compuesta por 32 equipos ubicados en ciudades importantes de Estados Unidos, los 32 equipos se dividen en dos conferencias, la Conferencia Nacional (NFC) y la Conferencia Americana (AFC), asimismo cada Conferencia se divide en cuatro divisiones con cuatro equipos cada uno, las divisiones son, para la Conferencia Nacional, la División del Este, la División

avanzar más yardas en su siguiente ofensiva.

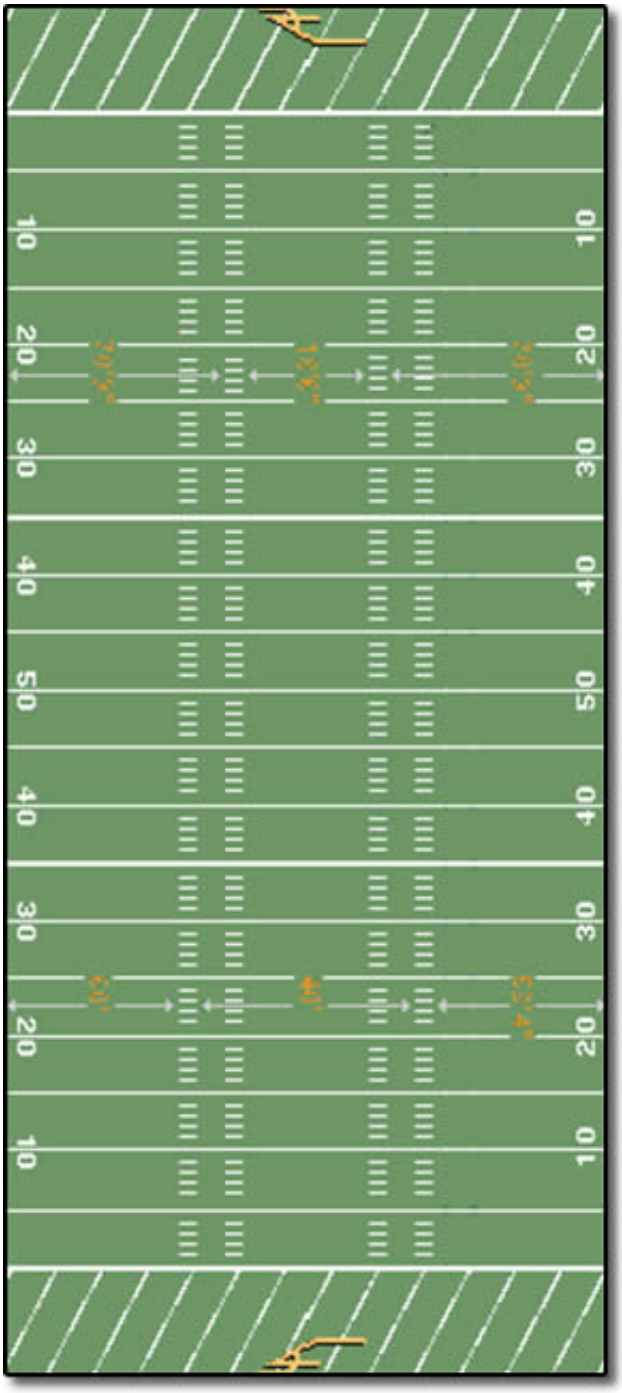

Figura 1.4: Campo de futbol americano

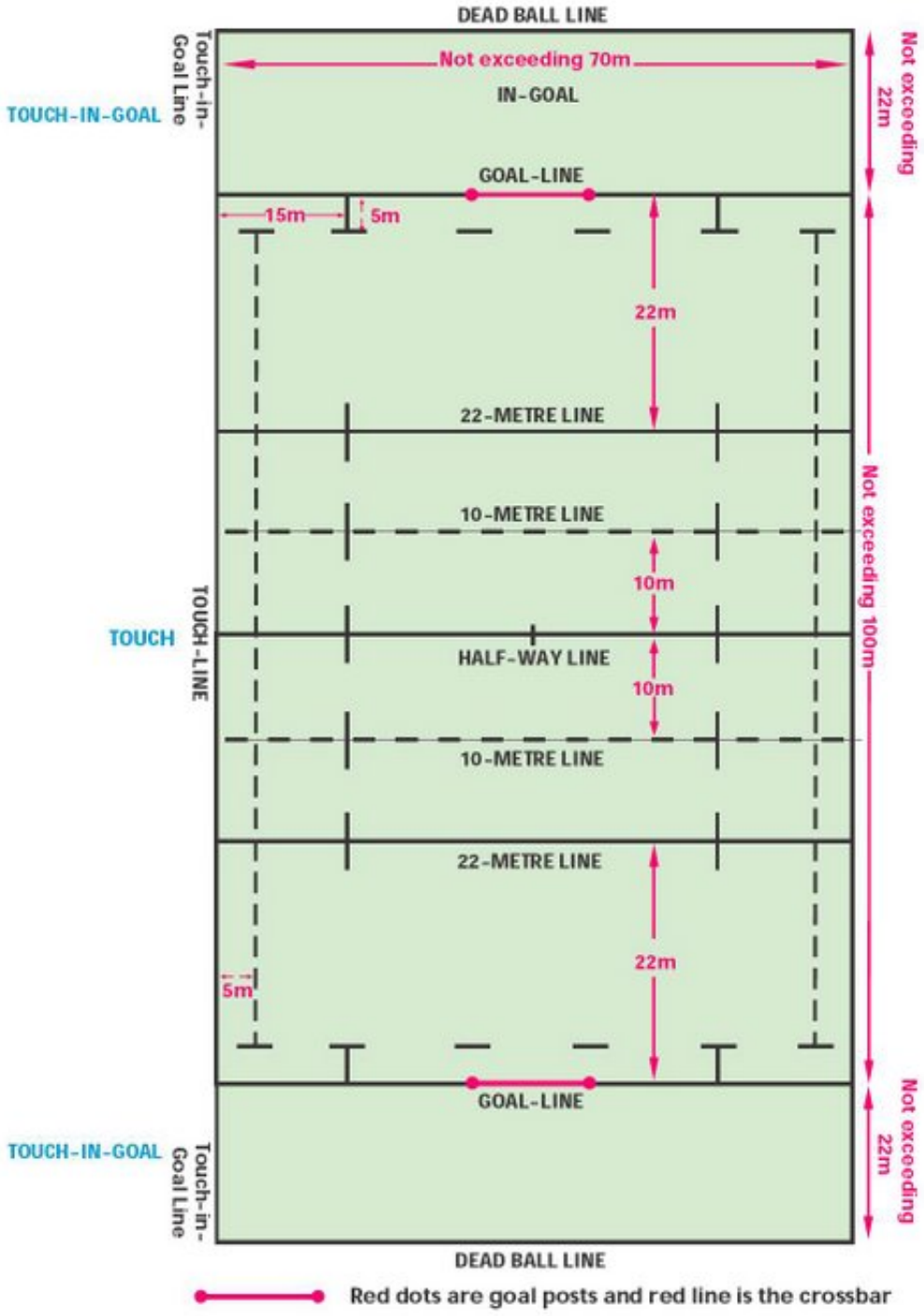

Figura 1.5: Campo de futbol americano

del Norte, la División del Sur y la División del Oeste. Para la Conferencia Americana las divisiones son la División del Este, la División del Norte, la División del Sur y la División del Oeste, a continuación se presentan los equipos por conferencia y división:

Equipos de la Conferencia Americana
División del Este

- Patriotas de Nueva Inglaterra
- Jets de Nueva York
- Bills de Buffalo
- Delfines de Miami

División del Norte

- Acereros de Pittsburgh
- Cuervos de Baltimore
- Bengalíes de Cincinnati
- Cafés de Cleveland

División del Sur

- Potros de Indianapolis
- Jaguares de Jacksonville
- Titanes de Tennessee
- Texanos de Houston

División del Oeste

- Raiders de Oakland
- Broncos de Denver
- Jefes de Kansas City
- Cargadores de San Diego

Equipos de la Conferencia Nacional
División del Este

- Vaqueros de Dallas
- Pieles Rojas de Washington

- Gigantes de Nueva York
- Aguilas de Filadelfia

División del Norte

- Osos de Chicago
- Empacadores de Green Bay
- Leones de Detroit
- Vikingos de Minnesota

División del Sur

- Panteras de Carolina
- Halcones de Atlanta
- Bucaneros de Tampa Bay
- Santos de Nueva Orleans

División del Oeste

- Cuarenta y Nueves de San Francisco
- Carneros de San Luis
- Cardenales de Arizona
- Halcones Marinos de Seattle

En la siguiente sección se describirán las principales herramientas, tanto estadísticas como de redes neuronales que se utilizarán para la creación del modelo.

1.3. Componentes principales

El análisis de componentes principales se encuentra dentro del conjunto de técnicas de análisis multivariado, éste pretende sintetizar un gran conjunto de datos, creando estructuras entre variables cuantitativas para crear unas nuevas variables que son combinaciones lineales de las originales.

El objetivo del análisis de componentes principales es el de reducir la dimensión de un conjunto de p variables a un conjunto menor de k variables para mejorar el manejo y la interpretabilidad de los datos, éste revela relaciones que generalmente no son sospechadas, por lo que permite obtener interpretaciones que de otra forma no se podrían dar.

El análisis de componentes principales es generalmente un paso intermedio en el camino de investigaciones mayores, por ejemplo las componentes principales pueden ser la entrada de un análisis de regresión múltiple, o de un análisis de conglomerados (*clusters*).

Algebráicamente las componentes principales son combinaciones lineales de las p variables aleatorias $X_1, X_2, ..., X_p$. Geométricamente estas combinaciones lineales representan la selección de un nuevo sistema de ejes coordenados obtenido a partir de la rotación del sistema original (con los ejes coordenados $X_1, X_2, ..., X_p$). Los nuevos ejes representan las direcciones con máxima variabilidad y proveen una descripción más simple de la estructura de varianzas y covarianzas.

Un aspecto interesante es que las componentes principales dependen exclusivamente de las matriz de covarianzas Σ o de la matriz de correlación ρ de $X_1, X_2, ..., X_p$. Su desarrollo no requiere ningún supuesto distribucional.

Las nuevas variables, determinan lo esencial de las variables originales y además tienen propiedades interesantes ver [Ric07] Capítulo 8:

1. Son ortogonales (cada componente representa una dirección del espacio de las variables originales)

2. Cada componente es linealmente independiente de la anterior

3. La primera componente es la que más varianza explica y la j-ésima explica menos varianza que la j-1 ésima

De forma geométrica el sub-espacio que se crea con las k primeras componentes, da el mejor ajuste posible al conjunto de datos medido mediante la suma de los cuadrados de las distancias perpendiculares desde cada punto al subespacio. El subespacio de menor dimensionalidad será $k = 1$ componentes, se puede hacer la representación en un sólo eje pero el conjunto inicial se puede distorsionar, así se introduce un nuevo eje para definir un subespacio $m = 2$, donde se pierde menos información. Si $k = p$ se tiene el mismo número de variables, no se reduciría la dimensión, sólo se haría una rotación rígida del conjunto de datos.

Existen dos formas básicas de aplicar el ACP (Análisis de Componente Principales):

1. Método basado en la matriz de covarianzas, que se usa cuando los datos son dimensionalmente homogéneos y presentan valores medios similares.

2. Método basado en la matriz de correlación, se usa cuando los datos no son dimensionalmente homogéneos o el orden de magnitud de las variables aleatorias medidas no es el mismo.

Método basado en correlaciones: Supóngase que existe una muestra con n individuos para cada uno de los cuales se han medido m variables aleatorias $F_j, j = 1, ..., m$, el método parte de la matriz de correlaciones, se considera el valor de cada una de las m variables aleatorias, para cada uno de n individuos se toma el valor de estas variables y

se escribe el conjunto de datos en forma de matriz $(F_{i\times j})_{i=1,...,n\times j=1,...,m}$. Obsérvese que cada conjunto $\mathcal{M}_j = \{\mathcal{F}_{i\times j}; i = 1, ..., n\}$ puede considerarse una muestra aleatoria para la variable F_j. A partir de los $n \times m$ datos correspondientes a las m variables aleatorias, puede construirse la **matriz de correlación muestral** definida por:

$$\mathbf{R} = [r_{ij}] \in M_{m\times m} \qquad \text{con } r_{ij} = \frac{\text{cov}(F_i, F_j)}{\sqrt{\text{var}(F_i)\text{var}(F_j)}} \tag{1.1}$$

Puesto que la matriz de correlaciones es simétrica entonces resulta diagonalizable (ver B) y sus eigenvalores (ver B) λ_i, verifican:

$$\sum_{i=1}^{m} \lambda_i = m \tag{1.2}$$

Debido a la propiedad anterior estos m valores propios reciben el nombre de pesos de cada uno de los m componentes principales. Los factores principales identificados matemáticamente se representan por la base de vectores propios de la matriz $\mathbf{R}$. Entonces cada una de las variables puede ser expresada como combinación lineal de los vectores propios o componentes principales.

Finalmente la varianza total de la población $(\sigma_{11}+\sigma_{22}+...+\sigma_{mm}) = (\lambda_1+\lambda_2+...+\lambda_m)$ y en consecuencia, la proporción de la varianza total que explica la k-ésima componente principal está dada por:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_m} \qquad ; k = 1, 2, ..., m \tag{1.3}$$

ver [Ste82] Capítulo 8.

Si la mayor parte de la varianza total de la población (80 a 90 %) de un número grande m de variables originales se puede explicar con las primeras p componentes, entonces estas p componentes pueden reemplazar a las originales m variables sin mucha pérdida de información.

1.4. Mapas auto-organizados o SOM (*Self Organized Map*)

La corteza es la capa del cerebro más grande en los mamíferos, mucho de su estructura y conectividad depende de la actividad eléctrica durante su desarrollo ver [JR90].

Debido a que varias áreas de la corteza son similares en su estructura anatómica, se ha sugerido un mecanismo común de organización, se puede deducir este mecanismo a través de modelar el desarrollo de la subestructura de la corteza visual primaria.

Existen evidencias que demuestran que en el cerebro hay neuronas que se organizan en muchas zonas, de forma que las informaciones captadas del entorno a través de los órganos sensoriales se representan internamente en forma de mapas bidimensionales. Por ejemplo, en el sistema visual se han detectado mapas del espacio visual en zonas de la corteza (capa externa del cerebro). Aunque en gran medida esta organización neuronal

está predeterminada genéticamente, es probable que parte de ella se origine mediante el aprendizaje, esto sugiere que el cerebro podría poseer la capacidad inherente de formar mapas topológicos de las informaciones recibidas del exterior.

La corteza visual primaria como muchas otras regiones de la corteza, es un mapa topológico y es organizado de tal forma que neuronas adyacentes responden a regiones adyacentes de la retina. Este mapa se forma de manera auto-organizada dependiendo de las entradas (*inputs*) de las conexiones de la corteza. El proceso de auto-organización es dirigido por entradas externas, y aparentemente está correlacionado con la actividad neuronal y el resultado de la cooperación y competencia entre neuronas. Durante el desarrollo estas conexiones se agrupan en grupos bien definidos, la configuración final de grupos corresponde a la distribución de las conexiones en el mapa topológico.

A partir de estas ideas Tuevo Kohonen presentó en 1982 un sistema con un comportamiento semejante, se trataba de un modelo de red neuronal con capacidad para formar mapas de características, de manera similar a como ocurre en el cerebro; el objetivo de Kohonen era demostrar que un estímulo externo por sí solo, suponiendo una estructura propia y una descripción funcional del comportamiento de la red, era suficiente para forzar la formación de los mapas.

Este modelo tiene dos variantes denominadas LVQ (*Learning Vector Quantization*) y TPM (*Topology Preserving Map*) o SOM (*Self Organizing Map*), ambas se basan en el principio de formación de mapas topológicos para establecer características comunes entre las informaciones (vectores) de entrada a la red, aunque difieren en las dimensiones de éstos, siendo de una sola dimensión en el caso de LVQ y bidimensional o tridimensional en la red SOM. El aprendizaje en el modelo de Kohonen es de tipo fuera de línea (*Off-line*), por lo que se distingue una etapa de aprendizaje y otra de funcionamiento. En la etapa de aprendizaje se fijan los valores de las conexiones (*feedforward*[9]) entre la capa de entrada y la salida. Esta red utiliza un aprendizaje no supervisado[10] de tipo competitivo, las neuronas de la capa de salida compiten por activarse y sólo una de ellas permanece activa ante una determinada información de entrada a la red, los pesos de las conexiones se ajustan en función de la neurona que haya resultado vencedora. Durante la etapa de entrenamiento, se presenta a la red un conjunto de datos de entrada (vectores de entrenamiento) para que ésta establezca en función de la semejanza entre los datos, las diferentes categorías (una por neurona de salida), que servirán durante la fase de funcionamiento para realizar clasificaciones de nuevos datos que se presenten a la red. Los valores finales de los pesos de las conexiones entre cada neurona de la capa de salida con las de entrada se corresponderán con los valores de los componentes del vector de aprendizaje que consigue activar la neurona correspondiente. En el caso de existir más patrones de entrenamiento que neuronas

[9]Debido a que las entradas son pasadas a través del modelo para producir la salida, el sistema es conocido como *feedforward*, ver [RT90] Capítulo 3.

[10]El aprendizaje supervisado requiere que la responsabilidad del entrenamiento recaiga en un ente externo, mientras que en aprendizaje no supervisado el entrenamiento es responsabilidad del mismo sistema , ver [RT90] Capítulo 5.

de salida, más de uno deberá asociarse con la misma neurona[11], es decir pertenecerán a la misma clase, en este trabajo necesitamos de antemano una idea de la cantidad de grupos que existen en el conjunto de datos que se desea clasificar, pues es importante determinar el número de neuronas que deberá tener la red neuronal, para determinar este número existen varias técnicas, (la experiencia de los expertos, técnicas estadísticas o incluso se puede hacer de manera heurística).

En este modelo, el aprendizaje no concluye después de presentarle una vez todos los patrones de entrada, sino que habrá que repetir el proceso varias veces para refinar el mapa topológico de salida, de tal forma que cuantas más veces se presenten los datos, tanto más se reducirán las zonas de neuronas que se deben activar ante entradas parecidas, consiguiendo que la red pueda realizar una clasificación más selectiva.

El algoritmo de aprendizaje utilizado para establecer los valores de los pesos de las conexiones entre las n neuronas de entrada y las m de salida, es el siguiente:

1. En primer lugar se define $w_{ij}(t); \quad 0 \leq i \leq n-1 \quad$ como los pesos de la entrada i a la neurona j en el tiempo t. Inicializando los pesos de las n entradas a las neuronas con valores aleatorios pequeños. Se establece un radio grande de la vecindad alrededor de la neurona $j, N_j(0)$.

2. A continuación se presentan a la red, datos de entrada en forma de vector $x(t) = (x_0(t), x_1(t), ..., x_{n-1}(t))$, donde $x_i(t)$ es la entrada a la neurona i en el tiempo t.

3. Puesto que se trata de un aprendizaje competitivo, se determina la neurona vencedora de la capa de salida, será aquella j cuyo vector de pesos w_j[12] sea el más parecido a la información de entrada x (vector de entrada). Para ello se calculan las distancias o diferencias entre ambos vectores, considerando una por una todas las neuronas de salida, suele utilizarse la distancia euclidiana o la siguiente expresión que es similar a aquella, pero eliminando la raíz cuadrada:

$$d_j = (\sum_{i=0}^{n-1}(x_i(t) - w_{ij}(t))^2 \tag{1.4}$$

4. Se selecciona la mínima distancia de las neuronas de salida, designando la neurona vencedora como $j*$

5. Una vez localizada la neurona vencedora $j*$, se actualizan los pesos de las conexiones entre las neuronas de entrada y dicha neurona, así como las de las conexiones entre las neuronas de entrada y las neuronas vecinas $N_{j*}(t)$ de la neurona vencedora, en realidad lo que se consigue con esto es asociar la información de entrada con una cierta zona de la capa de salida. Esto se realiza mediante la siguiente ecuación:

[11] En el presente trabajo cuando hablamos de clasificación, cada neurona representa un grupo, ver [Sch98]

[12] Vector cuyas componentes son los pesos de las conexiones entre esa neurona y cada una de las neuronas de la capa de entrada

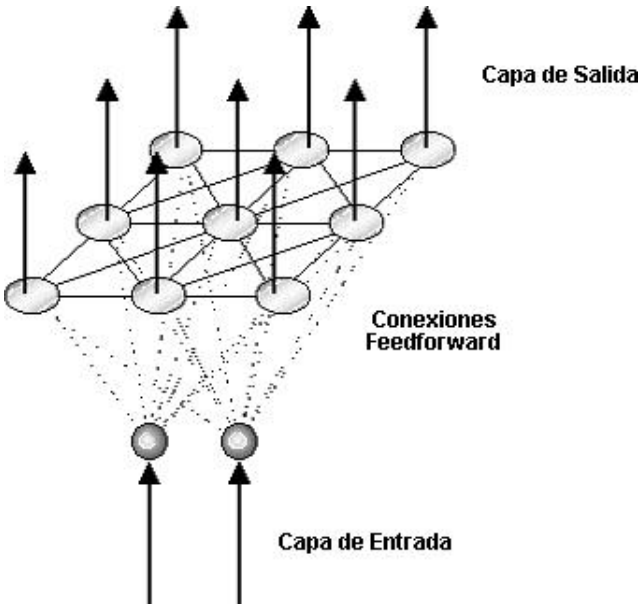

Figura 1.6: Conexiones de una red de Kohonen

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)(x_i(t) - w_{ij}(t)) \quad ; \quad 0 \leq i \leq n-1 \quad ; \quad j \in N_{j*}(t) \quad (1.5)$$

El tamaño de $N_{j*}(t)$ se puede reducir en cada iteración del proceso de ajuste de los pesos, con lo que el conjunto de neuronas que pueden considerarse vecinas cada vez es menor como se observa en la figura 1.7, sin embargo en la práctica es habitual considerar una zona fija en todo el proceso de entrenamiento de la red.

El término $\alpha(t)$ es el coeficiente de aprendizaje o parámetro de ganancia, con un valor entre 0 y 1 el cual decrece con el número de iteraciones (t) del proceso de entrenamiento, de tal forma que cuando se ha presentado un gran número de veces todo el juego de patrones de aprendizaje su valor es prácticamente cero, con lo que la modificación de los pesos es insignificante.

Para hallar α suele utilizarse una de las siguientes expresiones:

$$\alpha(t) = \frac{1}{t} \qquad \alpha(t) = \alpha_1(1 - \frac{t}{\alpha_2}) \quad (1.6)$$

Siendo α_1 un valor de 0.1 ó 0.2 y α_2 un valor próximo al número total de iteraciones del aprendizaje, que por lo general se toma como 10,000 para comenzar.

6. El proceso debe repetirse, volviendo a presentar todo el juego de patrones de aprendizaje $(x_0(t), x_1(t), ..., x_{n-1}(t))$ hasta obtener la salida deseada.

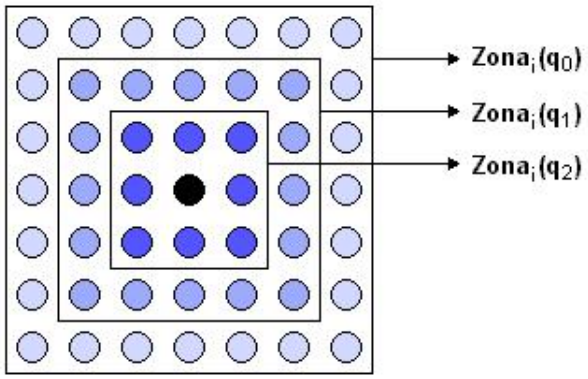

Figura 1.7: Posible evolución de la vecindad en una red de Kohonen

En definitiva lo que hace una red de Kohonen es realizar una tarea de clasificación, puesto que la neurona de salida activada ante una entrada representa la clase a la que pertenece dicha información de entrada, además ante otra entrada parecida se activa la misma neurona de salida, u otra cercana a la anterior debido a la semejanza entre las clases, así se garantiza que las neuronas topológicamente próximas sean sensibles a entradas físicamente similares; por esta causa la red es especialmente útil para establecer relaciones desconocidas previamente entre conjuntos de datos.

1.5. Análisis de conglomerados

El propósito del análisis de conglomerados (*cluster*) es clasificar o agrupar observaciones de forma que los datos sean muy homogéneos dentro de grupos (mínima varianza) y que estos grupos sean lo más heterogéneos posible entre ellos (máxima varianza), sin tener información previa sobre grupos en la población (sin supervisión). De este modo se obtiene una clasificación multivariante de los datos con la que se puede comprender mejor la población de la que proceden. Se puede realizar análisis de conglomerados de casos, un análisis de conglomerados de variables o un análisis de conglomerados por bloques si se agrupan variables y casos. El análisis de conglomerados se puede utilizar para:

La taxonomía, agrupar especies naturales, para la publicidad (*marketing*), clasificar consumidores tipo, medicina, clasificar seres vivos con los mismos síntomas y características patológicas, técnicas de reconocimiento de patrones, formar grupos de pixeles en imágenes digitalizadas enviadas por un satélite desde un planeta, para identificar los terrenos, etc.

Dentro del análisis de conglomerados existen los métodos no jerárquicos ver [Ric07] Capítulo 12, que están diseñados para clasificar elementos independientemente de sus variables, en K grupos. El número de grupos K debe ser especificado por adelantado o en su caso determinado como parte del procedimiento. Dado que no es necesario calcular la matriz de distancias entre los elementos y que los datos básicos no deben ser guardados

durante la ejecución de los cálculos, los métodos no jerárquicos pueden ser utilizados en conjuntos con un gran número de elementos.

Los métodos no jerárquicos comienzan (1) con una partición inicial de elementos o (2) con un conjunto inicial de puntos que formarán el núcleo de los cúmulos.

Uno de los métodos no jerárquicos más comunes, es el de k-medias cuyo algoritmo es sencillo y eficiente. Además, procesa los patrones secuencialmente (por lo que requiere un almacenamiento mínimo), el algoritmo es el siguiente:

Suponga que se tienen n características en vectores $x_1, x_2, ..., x_n$, donde cada x, está representado en un espacio m dimensional se sabe que están agrupados en k cúmulos ($k < n$). Se define m_j como la media del j-ésimo cúmulo. Si los cúmulos están bien separados, se puede usar una mínima distancia de clasificación para separarlos. Esto es, se dice que x_i está en el j-ésimo cúmulo si $||x_i - m_j||$ es el mínimo con respecto a los k cúmulos.

Entonces el algoritmo detallado es:

Se hace una estimación inicial para las k medias $m_1, m_2, ..., m_k$.

Mientras no cambie alguna media:

1. Se usa la media estimada para clasificar los datos en cúmulos. $b(i, j) = 1$ si el i-ésimo dato, es el más cercano a la j-ésima media.

2. Para cada uno de los cúmulos

   - Se calcula la nueva media m_i, utilizando la nueva clasificación

$$m_{i,j} = \frac{\Sigma_{i=1}^{n} b(i,j) * x_i}{\Sigma_{i=1}^{n} b(i,j)} \tag{1.7}$$

Otra forma es la siguiente:

```
;;; -----------------------------------------
;;; K-MEDIAS(K,DATOS,distancia)

;; 1. Inicializar m_i (i=1,...,k)
      (aleatoriamente o con algun criterio
;;    heuristico)
;; 2. REPETIR (hasta que los m_i no cambien):
;;    2.1 PARA j=1,...,N, HACER:
;;        Calcular el cluster correspondiente
          a x_j, escogiendo, de entre
;;        todos los m_i, el m_h tal que
          distancia(x_j,m_h) sea minima
;;    2.2 PARA i=1,...,k HACER:
;;        Asignar a m_i la media aritmetica de
          los datos asignados al
```

```
;;        cluster i-esimo
;; 3. Devolver m_1,...,m_n
;;; ----------------------------------------
```

Para verificar la estabilidad del algoritmo es deseable ejecutarlo más de una vez con nuevas particiones iniciales. Una vez que los cúmulos son determinados, ayuda a la interpretación de los resultados la reorganización de los mismos en el orden generado por el algoritmo.

Existen situaciones que deben considerarse al momento de utilizar k-medias:

1. La existencia de datos fuera de rango (*outliers*) pueden producir un grupo con elementos dispersos.

2. Si dos o más puntos semilla caen en el mismo cúmulo, el resultado serán cúmulos pobremente diferenciados.

3. El algoritmo no cuenta con un método para determinar el número de grupos.

A continuación se muestra un ejemplo en el que se ve porque no siempre se pueden establecer claramente los elementos de un grupo, o incluso si el número de grupos es correcto, en la figura (1.8) se puede distinguir claramente que existen dos grupos bien definidos, en el que incluso a simple vista se puede establecer los integrantes de cada grupo, sin embargo en la figura (1.9) la situación ya no es tan clara, ya que aquí no se sabe por qué los equipos de Buffalo y Cleveland no pertenecen al equipo rojo, o por qué Chicago y Carolina no pertenecen al equipo rojo, o incluso surge la duda de por qué los equipos Cleveland, Chicago, Buffalo y Carolina no forman un tercer grupo, si a esta situación se suman más varibles aumentando la dimensión entonces el problema se vuelve muy complicado, por lo que es necesario confiar en los resultados que arrojan las diferentes herramientas de agrupación.

Para más detalle consultar [Ric07]

1.6. Críterio para determinar el número de grupos en un conjunto de datos

Tanto la técnica de análisis de conglomerados como la de mapas auto-organizados, requieren de antemano el número de grupos en que se agruparán los datos utlizados, este problema se trata en [Ric07], a continuación se presenta un resumen del artículo.

Se plantean 2 preguntas:

1. ¿Cuál es la mejor subdivisión de los elementos de un conjunto en un número g de grupos?

2. ¿Cuál es el mejor valor para g?

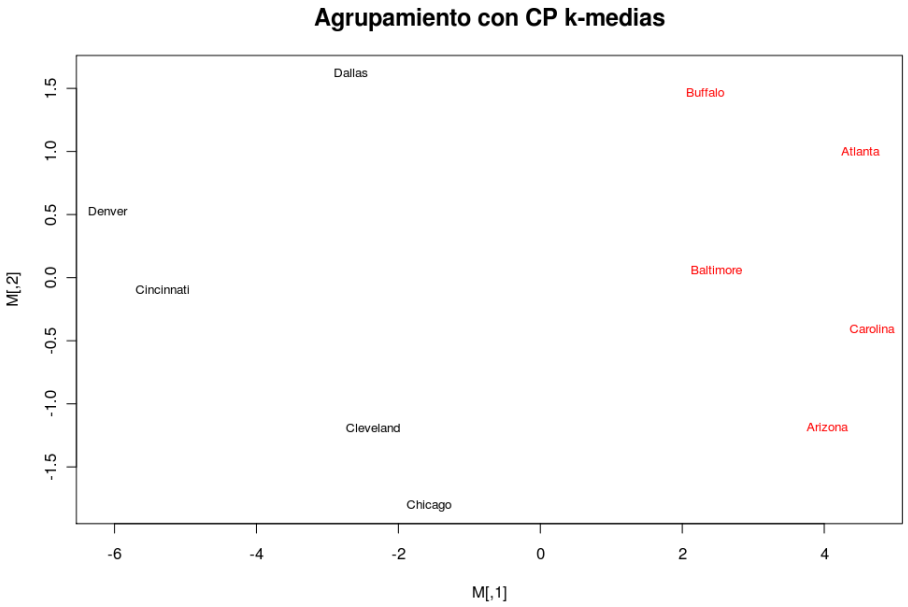

Figura 1.8: Ejemplo de agrupación sin conflictos

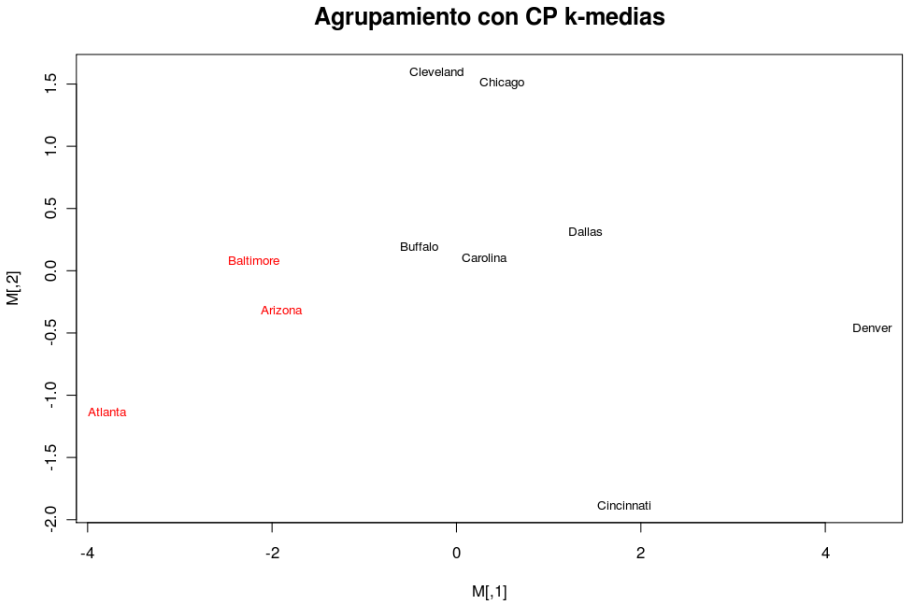

Figura 1.9: Ejemplo de agrupación con conflictos

Para responder a la pregunta (1) es necesario formular una función objetivo que cuantifique la adecuación de una partición del conjunto en g grupos y entonces, encontar la partición que optimice esta función. Al suponer que p variables han sido medidas en cada uno de n miembros de un conjunto, y que W es la matriz de covarianza interna del grupo (*within-group*) (ver A) para alguna partición. Entonces tanto la suma de cuadrados $S = traza(W)$ (ver B) como el determinante $D = |W|$ ver [Ste82] capítulo 4 proveen una función objetivo intuitiva y razonable, pues valores pequeños de S o D indican buenas particiones.

Para responder a la pregunta (2) una opción natural es pensar en calcular particiones del conjunto de datos para diferentes valores sucesivos de g es decir agrupaciones con $g = 2, 3, 4, ...$ y seleccionar el valor de g que optimice la función objetivo antes seleccionada, este proceso es nombrado como regla de paro (*stopping rule*) para la función objetivo asociada. En [JT88] se estudian dos reglas de paro, como es la minimización de $g^2\tilde{D}_g$ donde , donde $\tilde{D}_g$ es el valor óptimo de D asociado a una partición de g grupos. El argumento principal es que al considerar todos las posibles combinaciones en que se puede agrupar un conjunto de elementos en g grupos, donde g es especificado de antemano, alguna hará que $D = |W|$ sea mínimo, esta partición será llamada la subdivisión óptima de los datos en g grupos, y el valor correspondiente de D se denotará $\tilde{D}_g$. Ahora al suponer que las variables medidas son variables aleatorias independientes (ver B) y variables aleatorias uniformente distribuidas (ver B) en un intervalo finito, con la misma variaza σ^2 y el número de grupos es $g = 1$ entonces, $W = T$, y la matriz de dispersión tendrá elementos:

$$t_{ij} = \begin{cases} \sigma^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \quad entonces \quad \tilde{D}_1 = |T| = \sigma^{2p}, \tag{1.8}$$

ahora si se supone que $g > 1$, la subdivisión de una distribución rectangular univariada en g grupos será óptima cuando las secciones son iguales. Esta subdivisión reduce la varianza en un factor de g^2, esta subdivisón en cualquiera de las p variables en un caso multivariado reducirá D también en un factor de g^2. Ahora si g es prima, éstas serán las únicas subdivisiones, si g es compuesta, entonces se tienen varios casos, por ejemplo si $g = g_1 g_2$ entonces la subdivisión de cualesquiera dos variables en subdivisones iguales g_1 y g_2 respectivamente también reducirá D en un factor de g^2. De hecho, todas las posibles subdivisiones en "cuadros" del mismo tamaño, forma y orientación son óptimas para $p \geqslant 2$, existen $v \geqslant p$ subdivisiones óptimas, por lo que para datos uniformes e independientes mientras g varía el valor del criterio $g^2\tilde{D}_g$ permanecerá constante, en particular debería permanecer igual al valor del críterio $g = 1$ llamado $|T|$ de (1.8), por lo tanto $g^2\tilde{D}_g/|T|$ debe ser aproximadamente igual a 1 para toda g.

Ahora para una g dada, considere todas las subdivisiones que optimizan el valor $\tilde{D}_g$, si $\sigma^2_{ij(g)}$ es la varianza interna de la i-ésima variable de la j-ésima subdivisión, entonces $\tilde{D}_g = \Pi_{i=1}^{p}\sigma^2_{ij(g)} \quad ; (j = 1, ..., v)$. La subdivisión que adicionalmente minimiza la suma de cuadrados de la función objetivo $S = traza(W)$ será subdivisión k en la que $\sigma^2_{ik(g)}$ sea lo más parecida para todas las $i = 1, ..., p$, esto es porque para $|W| = \Pi_{i=1}^{p}\sigma^2_{ij(g)}$,

$traza(W) = \Sigma_{i=1}^{p}\sigma_{ij(g)}^2$ es mínima cuando $\sigma_{ij(g)}^2$ son todas iguales. Ahora si $\sigma_{1k(g)}^2 = \sigma_{2k(g)}^2 = ... = \sigma_{pk(g)}^2 = \sigma_{(g)}^2$, entonces $\tilde{D}_g = \sigma_{(g)}^{2p}$, pero $|T| = \sigma^{2p}$ por lo que $g^2\tilde{D}_g/|T| = 1$ implica que $\sigma_{(g)}^{2p} = \sigma^{2p}/g^2$, es decir

$$\sigma_{(g)}^2 = (1/g^{2/p})\sigma^2 \tag{1.9}$$

ahora $traza(T) = p\sigma^2$ y $traza(W) = p\sigma_{(g)}^2$, por 1.9 se tiene $g^{2/p}traza(W) = traza(T)$.

Escribiendo $\tilde{S}_g$ como el valor óptimo de la función objetivo suma de cuadrados para g, se tiene que $g^{2/p}\tilde{S}_g = traza(T)$, por lo tanto si $x_1, x_2, ..., x_p$ son independientes y uniformemente distribuidas, la subdivisión óptima de la población en g grupos reducirá la suma de cuadrados por un factor $g^{2/p}$, por lo que $g^{2/p}\tilde{S}_g$ deberá mantenerse aproximadamente constante sobre g o lo que equivalente $g^{2/p}\tilde{S}_g/traza(T)$ será aproxiamdamente igual a 1 para toda g.

Los argumentos heurísticos anteriores sugieren que si un conjunto de datos con variables aleatorias independientes y uniformente distribuidas, es particionado de manera óptima en g grupos usando la función objetivo determinante, entonces $g^2\tilde{D}_g/|T|$ deberá ser aproximadamente igual a la unidad, mientras que si las particiones son hechas utilizando la función objetivo suma de cuadrados, entonces $g^{2/p}\tilde{S}_g/traza(T)$ deberá ser aproximadamente igual a la unidad.

En [JT88] se realizan diferentes pruebas que verifican las conjeturas anteriores, adicionalmente se argumenta que si el criterio $M_g = g^2\tilde{D}_g$ se mantiene aproximadamente constante sobre g para datos de una población homogénea y uniforme, entonces la subdivisión óptima en g grupos proporcionará una reducción grande en M_g si es que los datos provienen de una población que está fuertemente agrupada alrededor de g nodos, por lo tanto se sugiere utilizar M_g como base para una regla de paro: el valor óptimo de g es el valor que alcanza el mínimo de M_g; graficando M_g contra g se puede observar el valor apropiado de g. El argumento es exactamente el mismo si utilizamos la suma de cuadrados como función objetivo, sin embargo se puede ver que existen dos problemas:

1. En la mayoría de las aplicaciones los datos son sólo una muestra de la población de interés, entonces $g^{2/p}\tilde{S}_g$ tenderá a decrecer mientras g aumenta, incluso si los datos provienen de una población con distribución uniforme ésto se muestra en [JT88] capítulo 2, experimentando con diferentes datos).

2. La obtención de $g^{2/p}\tilde{S}_g$ se sustenta en la hipótesis de que las variables de los datos son independientes y uniformes, lo cual en aplicaciones reales es más que difícil.

Por estas razones si se grafica $g^{2/p}\tilde{S}_g$ contra g en algunas ocasiones, pasará que el valor mínimo no se presenta con $g = 1$ para datos homogéneos, o que $g^{2/p}\tilde{S}_g$ es monótona decreciente con g para datos fuertemente agrupados. Por esto si en lugar de utilizar directamente la función $g^{2/p}\tilde{S}_g$ como regla de paro, es mejor utilizar una regla basada en

diferencias sucesivas. Sea $DIFF(g)$ la diferencia en la función cuando el número de grupos de la partición se incrementa de $(g-1)$ a g es decir:

$$DIFF(g) = (g-1)^{2/p}\tilde{S}_{g-1} - (g)^{2/p}\tilde{S}_g \tag{1.10}$$

Si los datos provienen de una población con distribución uniforme, entonces idealmente los valores de $DIFF(g)$ para $g = 2, 3, 4, ...$ se distribuirán aleatoriamente alrededor del cero. Ahora si se supone que los datos están fuertemente agrupados alrededor de k nodos, entoces se esperará que $\tilde{S}_g$ decrezca dramáticamente mientras g crece y sea menor que k, pero este decrecimiento deberá reducirse después de que $g = k$, entonces:

1. Para $g < k$, ambos $DIFF(g)$ y $DIFF(g+1)$ deberá ser grande y positivo.
2. Para $g > k$, ambos $DIFF(g)$ y $DIFF(g+1)$ deberá ser pequeño y uno o los dos deberán ser negativos.
3. $DIFF(k)$ deberá ser grande y positiva, pero $DIFF(k+1)$ deberá ser relativamente pequeña y podría ser negativa.

Con base en lo anterior **el criterio para la regla de paro será**:

$$C_g = |DIFF(g)/DIFF(g+1)| \tag{1.11}$$

El valor óptimo de g es el valor que maximiza C_g.
Esta sección está basada en su totalidad en [JT88].

1.7. Estimadores de máxima verosimilitud

En estadística, un estimador es un estadístico[13], el valor de un estimador proporciona lo que se denomina estimación puntual del valor del parámetro en estudio. El principio de máxima verosimilitud proporciona un método general el cual, bajo condiciones que seguido se satisfacen en el muestreo aleatorio, genera estimadores con características y propiedades deseables. El rasgo esencial del principio de máxima verosimilitud como se aplica al problema de la estimación, es que se requiere que se escoja como estimación de un parámetro el valor del parámetro para el cual la probabilidad de obtener el punto muestral observado o de obtener un punto muestral cerca de éste, es tan grande como sea posible. Esto es habiendo desarrollado el experimento de muestreo y observado los valores de muestra, se ve hacia atrás y se calcula la probabilidad de que estos valores de muestra sean observados. Esta probabilidad dependerá en general del parámetro al que se le da un valor que hace que esta probabilidad sea tan grande como sea posible.

Al suponer primero que la variable aleatoria de la población x tiene una función de probabilidad que depende del parámetro θ, $f(x;\theta)$. Se supone que la forma de la función

[13]Sea $x_1, x_2, ..., x_n$ una muestra aleatoria de densidad $f(x;\theta)$ un estadístico es una función sólo de las x_i usada para estimar un parámetro desconocido de la población ver [Ale78].

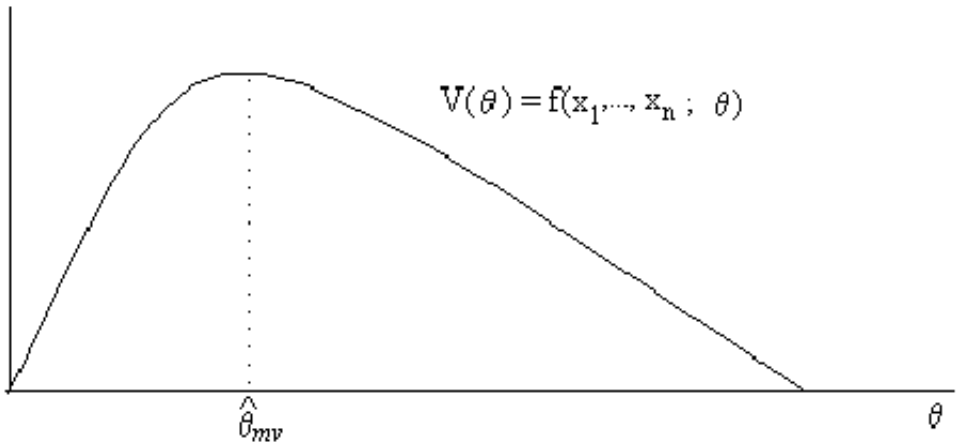

Figura 1.10: Función de verosimilitud

*La **Función de verosimilitud** se obtiene a partir de la función de densidad, intercambiando los papeles entre parámetro y estimador. En una función de verosimilitud consideramos que las observaciones $x_1, ..., x_n$, están fijadas, y se representa la gráfica con el valor de los valores que tomaría la función de densidad para todos los posibles valores del parámetro θ. El **Estimador máximo verosímil** del parámetro buscado, $\hat{\theta}_{\mathcal{MV}}$, es aquel que maximiza su función de verosimilitud, $V(\theta)$ ver [Ale78].*

f es conocida pero no el valor θ. La función de probabilidad conjunta de las variables aleatorias muestrales, evaluada en el punto muestral $(x_1, x_2, ..., x_n)$, es:

$$L(\theta) = f(x_1, x_2, ..., x_n; \theta) = \Pi_{i=1}^{n} f(x_i; \theta) \tag{1.12}$$

Esta función es también conocida como la *función de verosimilitud* de la muestra; se pone especial interés en ésta como función de θ cuando los valores de la muestra $(x_1, x_2, ..., x_n)$ son fijos. El principio de máxima verosimilitud requiere escoger como estimación del parámetro desconocido el valor θ para el cual la función de verosimilitud tome su máximo valor, como se muestran en la figura 1.10.

Si la distribución principal está completamente determinada sólo cuando los valores de dos o más parámetros desconocidos $\theta_1, \theta_2, ...\theta_k$ son especificados, entonces la función de verosimilitud será función de todos ellos:

$$L(\theta_1, \theta_2, ...\theta_k) = f(x_1, x_2, ..., x_n; \theta_1, \theta_2, ...\theta_k) = \Pi_{i=1}^{n} f(x_i; \theta_1, \theta_2, ...\theta_k) \tag{1.13}$$

Los estimadores de máxima verosimilitud de $\theta_1, \theta_2, ...\theta_k$ serán aquellos números que dan a la función de verosimilitud un máximo, para mayores detalles ver [H.D79]. Los esti-

madores de máxima verosimilitud tienen ciertas propiedades en general que a continuación enunciamos: (ver B)

1. Son consistentes
2. Son invariantes frente a transformaciones biunívocas, es decir, si $\hat{\theta}_{\mathcal{MV}}$ es el estimador máximo verosímil de θ y $g(\tilde{\theta})$ es una función biunívoca de $\tilde{\theta}$, entonces $g(\hat{\theta}_{\mathcal{MV}})$es el estimador máximo verosímil de $g(\theta)$.
3. Si $\hat{\theta}$ es un Estadístico suficiente de θ, su estimador máximo verosímil, $\hat{\theta}_{\mathcal{MV}}$ es función de la muestra a través de $\hat{\theta}$.
4. Son asintóticamente normales.
5. Son asintóticamente eficientes, es decir, entre todos los estimadores consistentes de un parámetro θ, los de máxima verosimilitud son los de varianza mínima.
6. No siempre son Estimadores insesgados

El estimador máximo verosimil para el parámetro p de la **distribución binomial** es (ver B):

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{1.14}$$

El detalle para obtener este estimador máximo verosímil de la distribución binomial se puede consultar en [Ale78].

1.8. Diseño de una base de datos relacional

En general, el objetivo del diseño de una base de datos relacional es generar un conjunto de esquemas de relaciones que permitan almacenar la información con un mínimo de redundancia, pero que a la vez faciliten la recuperación de la información. Una de las técnicas para lograrlo consiste en diseñar esquemas que tengan una forma normal adecuada. Para determinar si un esquema de relaciones tiene una de las formas normales, se requiere mayor información sobre el mundo real que se intenta modelar con la base de datos. La información adicional la proporciona una serie de limitantes que se denominan dependencias de los datos.

1.8.1. Fases del diseño de una base de datos

En esta sección se presentan las fases que comprende el diseño de una base de datos.
Recolección y análisis de requerimientos:

Los diseñadores entrevistan a los futuros usuarios de la base de datos para recoger y documentar sus necesidades de información. En paralelo, conviene definir los requerimientos funcionales que consisten en operaciones (transacciones) que se aplicarán a la base de datos, e incluyen la obtención de datos y la actualización.

Diseño conceptual:

Una vez recogidos todos los requerimientos, el siguiente paso es crear un esquema conceptual para la base de datos mediante un modelo de datos conceptual de alto nivel. El esquema conceptual contiene una descripción detallada de los requerimientos de información de los usuarios, y contiene descripciones de los tipos de datos, relaciones entre ellos y restricciones. En el presente trabajo se utilizará para el diseño de esquemas conceptuales el modelo E-R (entidad-relación), que describe los datos como entidades, vínculos (relaciones) y atributos.

Diseño lógico de la base de datos (transformación de modelo de datos):

El siguiente paso en el proceso de diseño consiste en implementar de hecho la base de datos con un sistema manejador de bases de datos (SMBD) comercial, transformando el modelo conceptual al modelo de datos empleados por el SMBD (jerárquico, red o relacional). En el modelo se hará la implementación con un SMBD relacional, por ser el modelo más utilizado.

Diseño físico de la base de datos:

En este paso se especifican las estructuras de almacenamiento internas y la organización de los archivos de la base de datos.

1.8.2. Módelo E-R (entidad-relación)

El modelo E-R (entidad-relación) fue propuesto por Peter P. Chen entre los años 1976 _ 1977. Posteriormente otros muchos autores han investigado y escrito sobre el modelo, proporcionando importantes aportaciones, por lo que realmente no se puede considerar que exista un único modelo E-R. El modelo de datos E-R está basado en una percepción del mundo real que consta de una colección de objetos básicos, llamados entidades, y de relaciones entre estos objetos. Una **entidad** es una cosa u objeto en el mundo real que es distinguible de otros objetos. Por ejemplo, cada persona es una entidad, y las cuentas bancarias pueden ser consideradas entidades. Las entidades se describen en una base de datos mediante un conjunto de atributos. Por ejemplo, los atributos número-cuenta y saldo describen una cuenta particular de un banco y pueden ser atributos del conjunto de entidades cuenta. Análogamente, los atributos nombre-cliente, calle-cliente y ciudad-cliente pueden describir una entidad cliente. Una **relación** es una asociación entre varias entidades. Por ejemplo, una relación impositor asocia un cliente con cada cuenta que tiene. El conjunto de todas las entidades del mismo tipo, y el conjunto de todas las relaciones del mismo tipo, se denominan respectivamente conjunto de entidades y conjunto de relaciones, existen diferenes tipos de entidades relaciones y atributos, éstos se describen a continuación:

Tipos de entidades:

- **Entidades fuertes**(o regulares), son aquellas que tienen existencia por si mismas (Por ejemplo, EMPLEADO). Las entidades fuertes se representan con un rectángulo con trazo simple.

- **Entidades débiles**, cuya existencia depende de otro tipo de entidad (Por ejemplo, FAMILIAR depende de EMPLEADO. La desaparición de un empleado de la base de datos hace que desaparezcan también todos los familiares del mismo). Estos tipos de entidades se representan normalmente con un rectángulo con líneas de doble trazo. Estas entidades normalmente no tienen suficientes atributos para formar una clave primaria.

- Cada entidad tiene **propiedades específicas**, llamadas **atributos**, que la describen. Por ejemplo, una entidad PROVEEDOR puede describirse por su nombre, su teléfono, etc. Los atributos se representan por elipses que están conectadas a su entidad o relación mediante una línea recta, al conjunto de valores que puede tomar un atributo se le llama **dominio del atributo**. Toda entidad debe tener al menos un atributo que permita diferenciar unas entidades particulares de otras, es decir que no toman nunca el mismo valor para dos entidades particulares diferentes. A estos atributos se les llaman **claves**. En el diagrama E-R los atributos clave deben aparecer destacados; por ejemplo, subrayando su nombre.

Tipos de Atributos: Los atributos pueden ser:

- Simples o compuestos: Los compuestos están formados por un conjunto de atributos, mientras que los simples no se pueden dividir.

- Monovaluados o multivaluados: Los monovaluados sólo pueden tener un valor para una entidad particular, mientras que los multivaluados pueden tener más de un valor. Los multivaluados se representan mediante una elipse con trazado doble. (Por ejemplo el atributo color de la entidad COCHE es un atributo multivaluado, pues un coche puede estar pintado de varios colores).

- Almacenados o derivados: Los derivados son atributos cuyo valor para una entidad particular puede obtenerse en función de los valores almacenados en otros atributos. Se representan mediante una elipse con trazo discontinuo.

Vínculo o relación:

Se puede definir como una correspondencia, asociación o conexión entre dos o más entidades. En los diagramas E-R se representa gráficamente como un rombo y sus nombres son verbos. Por ejemplo: VENDE, PERTENECE, etc. Una relación puede tener atributos descriptivos.

El **grado de una relación** es el número de entidades que participan en la relación. Se puede restringir el modelo E-R para incluir sólo conjuntos de relaciones binarias, es decir de grado 2 (es aconsejable).

Correspondencia de cardinalidad, expresa el número máximo de entidades que están relacionadas con una única entidad del otro conjunto de entidades que interviene en la relación. Aunque normalmente interesa sólo la cardinalidad máxima, a veces es útil especificar la cardinalidad mínima. Según su cardinalidad, se pueden clasificar las relaciones en los siguientes tipos:

- 1:1 relación una a una : La cardinalidad máxima en ambas direcciones es 1.
- 1:N relación una a muchas: La cardinalidad máxima en una dirección es 1 y en la otra muchos.
- N:M relación muchas a muchas: La cardinalidad máxima en ambas direcciones es muchos.

Tipos de participación de las entidades en una relación:

- relación opcional (parcial): No todas las ocurrencias de una entidad tienen que estar relacionadas con alguna de la otra entidad. Se representa mediante una línea con trazo sencillo. (Por ejemplo, no toda persona posee animales, y no todo animal es posesión de alguna persona. En este caso ambas entidades participan parcialmente en la relación).
- relación obligatoria (total): Todas las ocurrencias de una entidad deben estar relacionadas con alguna de la entidad con la que esta relacionada. Se dice también, que existen una participación total de ese conjunto de entidades en el conjunto de relaciones, y se representa mediante una línea con trazo doble. (Por ejemplo, todo proveedor tiene que vender algún artículo para serlo, y todo artículo es vendido por algún proveedor. En este caso ambas entidades participan de forma total en la relación).

En la figura (1.11) se presentan los símbolos empleados en el modelo entidad-relación. Para mayor detalle consulte [Abr02].

1.8.3. Reducción de un diagrama E-R a tablas

Tanto el modelo E-R, como el modelo de bases de datos (BD) relacional son representaciones abstractas y lógicas del desarrollo del mundo real. Debido a que los dos modelos emplean principios de diseño similares, se puede convertir un diseño E-R en un diseño relacional, siguiendo una serie de normas que se pueden resumir de la siguiente forma:

Para las ENTIDADES: Se genera una tabla con los atributos de una entidad. La clave primaria de la tabla es la misma que la de la entidad del modelo E-R. En el caso de entidades débiles, se genera una tabla con los atributos de la entidad débil, más la clave primaria de la entidad fuerte. La clave primaria de la tabla generada por la entidad

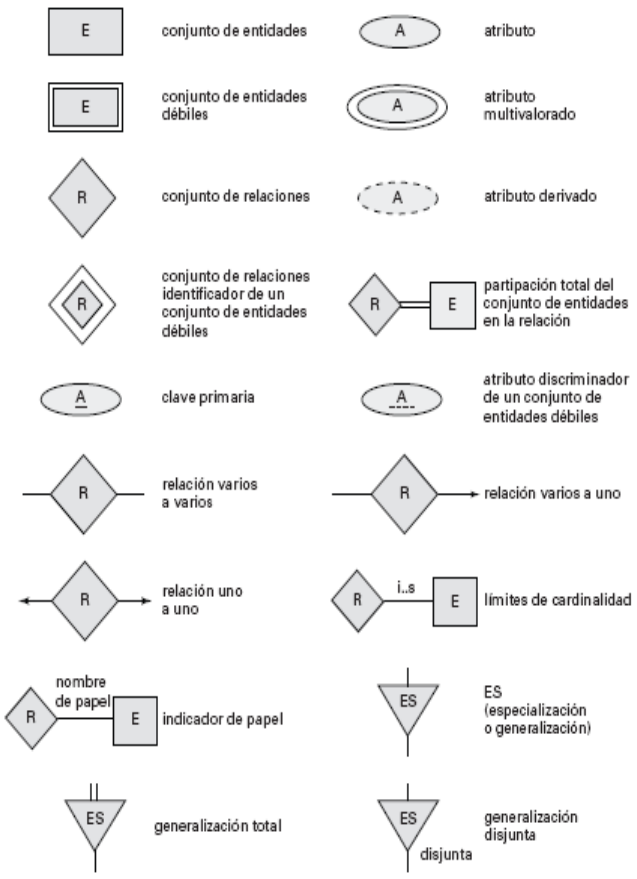

Figura 1.11: Símbolos del modelo entidad-relación

débil estará formada por los atributos clave de la entidad débil en el modelo E-R más los atributos clave de la entidad fuerte en el modelo E-R.

Para las RELACIONES:

- Si la relación es del tipo 1:1 y es obligatorio (total) tipo de participación de ambas entidades, sólo es necesario una tabla con los atributos de las entidades que participan en la relación, como clave primaria se puede tomar cualquiera de las claves de las entidades

- Si la relación es del tipo 1:1 y el tipo de participación de una entidad es obligatoria (total) y el de la otra es opcional (parcial), son necesarias dos tablas, cada una contendrá los atributos de las entidades que participan en la relación, en la tabla correspondiente a la entidad con participación obligatoria se añade una columna que contendrá la clave primaria de la otra entidad (clave ajena).

- Si la relación es del tipo 1:1 y el tipo de participación es opcional (parcial) para las dos entidades, entonces es necesario generar tres tablas, una para cada entidad y otra para la relación que deberá contener como atributos las claves primarias de las entidades que participan en la relación.

- Cuando la relación es del tipo 1:N, y la entidad del lado N es de participación obligatoria (total) se necesita una tabla para cada entidad. A la tabla que representa la entidad N se le añade un atributo que contenga la clave primaria de la entidad con la que se relaciona (clave ajena).

- Cuando la relación es del tipo 1:N, y la entidad del lado N es de participación optativa (parcial) se necesitan tres tablas: una para representar cada entidad y una para representar la relación.

- Si la relación es del tipo N:M, se generan tres tablas, una para cada entidad y otra que contiene los atributos propios de la relación más la claves primarias de las entidades que participan en la relación. En general, cuando la relación es entre una entidad fuerte y una entidad débil, no necesita ser representada en forma de tabla.

- Para atributos multivaluados: se generan tablas separadas, con la clave primaria del conjunto de entidades o relaciones al que pertenecen. Se una tabla para el conjunto de entidades del nivel más alto, para el conjunto de entidades de nivel más bajo, se crea una tabla que incluya una columna para cada uno de los atributos de ese conjunto de entidades, mas una columna que contendrá la clave primaria del conjunto de entidades de nivel superior.

La clave primaria de cada tabla del modelo relacional será la misma que la de las entidades asociadas del modelo E-R. Para mayor detalle consulte [Abr02].

1.8.4. Proceso de normalización de una relación

En el proceso de normalización, según la propuesta original de Codd (1972), se somete un esquema de relación a una serie de pruebas para certificar si pertenece o no a una cierta forma normal. En un principio, Codd propuso tres formas normales, a las cuales llamó primera, segunda y tercera formas normales (1FN, 2FN, 3FN). Posteriormente, Boyce y Codd propusieron una definición más estricta de 3FN, a la que se conoce como forma normal de BoyceCodd (FNBC). Todas estas formas normales se basan en las dependencias funcionales entre los atributos de una relación. Más adelante se propusó una cuarta forma normal (4FN) y una quinta (5FN), con fundamento en los conceptos de dependencias multivaluadas y dependencias de reunión, respectivamente.

La **normalización** de los datos puede considerarse como un proceso durante el cual los esquemas de relación que no cumplen las condiciones se descomponen repartiendo sus atributos entre esquemas de relación más pequeños que cumplen las condiciones establecidas. Un objetivo del proceso de normalización es garantizar que no ocurran anomalías de actualización.

Las formas normales, consideradas aparte de otros factores, no garantizan un buen diseño de BD. En general no basta con comprobar por separado que cada esquema de relación de la BD esté en, FNBC o 3FN. Más bien, el proceso de normalización por descomposición debe confirmar la existencia de propiedades adicionales que los esquemas relacionales, en conjunto, deben poseer dos de estas propiedades son:

- La propiedad de reunión sin pérdida, que garantiza que no se presentará el problema de las tuplas erróneas.
- La propiedad de conservación de las dependencias, que asegura que todas las dependencias funcionales estén representadas en alguna de las relaciones individuales resultantes.

La utilidad práctica de las formas normales queda en entredicho cuando las restricciones en las que se basan son difíciles de entender o de detectar por parte de los diseñadores de BD y usuarios que deben descubrir estas restricciones.

Primera forma normal (1FN):

Una relación está en primera forma normal (1FN) si los valores para cada atributo de la relación son atómicos.

Esto quiere decir simplemente que cada atributo sólo puede pertenecer a un dominio (es indivisible) y que tiene un valor único para cada fila. La primera forma normal se definió para prohibir los atributos multivaluados, compuestos y sus combinaciones. Cuando una relación no está en primera forma normal, se divide en otras relaciones, repartiendo sus atributos entre las resultantes. Normalmente la idea es eliminar el atributo que viola la 1 FN de la relación original y colocarlo en una relación aparte junto con la clave primaria de la relación de partida.

Segunda forma normal (2FN):

Una relación está en segunda a normal si está en la 1 FN y todos los atributos no clave dependen de la clave completa y no sólo de una parte de ésta.

Este paso sólo se aplica a relaciones que tienen claves compuestas, es decir, que están formadas por más de un atributo. Si un esquema de relación no está en 2FN, se le puede normalizar a varias relaciones en 2FN en las que los atributos que dependen de una parte de la clave formarán una nueva relación que tendrá esa parte de la clave como clave primaria.

Tercera forma normal (3FN):

Una relación está en tercera forma normal si y sólo si se cumple:

- La tabla está en la segunda forma normal (2NF)
- Ningún atributo no-primario de la tabla es dependiente transitivamente de una clave primaria

Se puede observar que si una relación está en tercera forma normal, está también en segunda forma normal, sin embargo lo inverso no siempre es cierto. Para mayor detalle consulte [Abr02].

Capítulo 2

Modelo

En el presente capítulo se describe la construcción del modelo. La idea general del modelo es recoger estadísticas de los equipos de futbol americano de la NFL, mismas que se generan partido a partido a lo largo de una temporada[1], después utilizando componentes principales se reduce el número de variables de tal forma que el modelo sea más manejable, a continuación utilizando SOM y k-medias se clasifican los equipos de acuerdo a sus características de juego obtenidas a partir de las componentes principales, la idea es que se puede suponer que existen varios tipos de equipos, por ejemplo: equipos corredores [2], equipos pasadores[3], etc. Las dos clasificaciones anteriores fueron determinadas históricamente por analistas expertos y comentaristas profesionales, sin embargo se puede suponer que existen otros tipos de clasificaciones, algunas imperceptibles al ser humano, aquí es donde ayudan las componentes principales en combinación con SOM y k-medias, pues éstas proporcionan nuevas clasificaciones de los equipos, éstas servirán como herramienta para predecir los resultados de la post-temporada[4] ya que si dos equipos se enfrentarán en la post-temporada por ejemplo digamos el X y el Y ambos pueden pertenecer a grupos diferentes de acurdo con nuestra clasificasción previa, digamos A y B respectivamente, entonces se procede a verificar como le fue al equipo X cuando se efrento a lo largo de la temporada contra equipos que poseen características similares a las del equipos Y, en este caso esos equipos son los pertenecientes al grupo B, y se procede análogamente con el equipo Y, estos resultados permitirán estimar el resultado de este enfrentamiento, para ésto es necesario apoyarse en una base de datos construida para este propósito, finalmente al estimar el resto de los resultados se verá la efectividad[5] del modelo y utilizando un estimador de máxima verosimilitud se estimará la probabilidad de éxito del modelo. En el

[1] En este caso se analizaron varias temporadas 2003 a 2007.

[2] Equipos que basan su juego en la fortaleza de su línea ofensiva y de sus corredores, en promedio la mayoría de sus jugadas son corridas.

[3] Equipos que basan su juego en la protección que su línea ofensiva da al pasador y la habilidad de su pasador y receptores, en promedio la mayoría de sus jugadas son jugadas de pase.

[4] Juegos finales, equivalente octavos de final.

[5] Capacidad para producir el efecto deseado.

modelo se consideran las siguientes etapas:

1. Selección y obtención de los datos.
2. Diseño y construcción de la base de datos.
3. Reducción de variables.
4. Clasificación de los equipos.
5. Predicción de los resultados.
6. Análisis de Resultados y Verificación del modelo.

A continuación se da una descripción detallada de cada etapa, el diagrama de flujo del modelo se presenta en la figura (2.1)

2.1. Selección y obtención de los datos.

Los datos fueron obtenidos directamente del sitio oficial de la NLF www.nfl.com, son las estadísticas básicas de los 32 equipos de la NFL, abarcan 5 temporadas de 16 jornadas cada una, desde la temporada del 2003, hasta la temporada 2007. Se tomaron en cuenta 32 variables en total, así mismo en el cuadro (2.1) se presenta una descripción de las mismas. Estas variables fueron seleccionadas con base en la opinión de los expertos analistas del futbol americano profesional, así como de comentaristas deportivos y fanáticos. Estas estadísticas fueron generadas por los equipos partido a partido a los largo de una temporada. Aunque algunas variables están muy relacionadas con otras, se decidió considerar la mayor cantidad de ellas, para evitar redundancia posteriormente se hizo una reducción de variables utilizando componentes principales.

2.2. Diseño y construcción de la base de datos.

Se generó una base de datos con el propósito de poder contestar preguntas utilizadas por el módelo, preguntas del tipo:

¿Cuál es el marcador de los encuentros que tuvieron el equipo X que pertenece al grupo Y y los equipos pertenencientes al grupo Z?, si es que jugó contra alguno de ellos, tener respuesta a este tipo de preguntas es fundamental en el módelo, pues contituyen de alguna forma la parte medular del mismo.

El diagrama entidad-relación de la base de datos se presenta en la figura (2.2), el modelo de la base de datos se presenta en la figura (2.3) donde se puede observar que ésta cuenta con siete tablas, este. Para diseñar, construir y consultar la base de datos se utilizaron herramientas de dominio público, a continuación se describen:

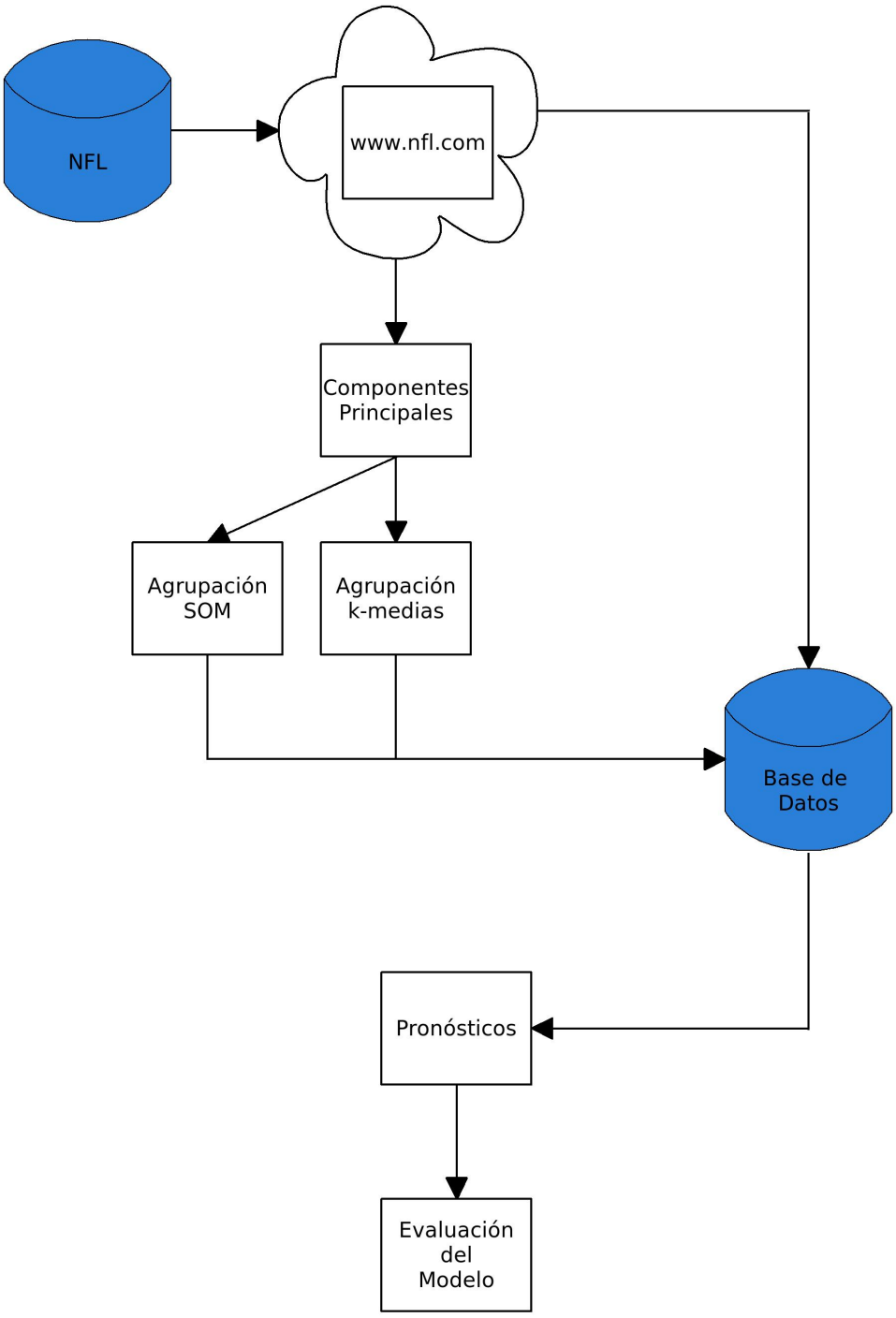

Figura 2.1: Diagrama de flujo del modelo

Variable	Descripción
GO-rk	Ofensiva General Ranking
GO-PtsG	Ofensiva General Puntos anotados por partido
GO-ydsG	Ofensiva General Yardas por partido
GO-ydsP	Ofensiva General Yardas por pase
GO-TopG	Ofensiva General Tiempo de posesion por partido
OP-Comp	Ofensiva Aerea Pases completos
OP-AttG	Ofensiva Aerea Intentos de pase por partido
OP-ydsG	Ofensiva Aerea Yardas por partido
0P-TD	Ofensiva Aerea Anotaciones
OP-Sck	Ofensiva Aerea Capturas del mariscal
OR-PtsG	Ofensiva Terrestre Puntos por partido
OR-AttG	Ofensiva Terrestre Intentos de carreras
OR-YdsG	Ofensiva Terrestre Yardas por partido
OR-TD	Ofensiva Terrestre Anotaciones
OR-FUM	Ofensiva Terrestre Balones sueltos
GD-Rk	Defensiva General Ranking
GD-PtsG	Defensiva General Puntos recibidos por partido
GD-YdsG	Defensiva General Yardas por partido
GD-YdsP	Defensiva General Yardas por pase
GD-TopG	Defensiva General Tiempo de posesión
GD-FUM	Defensiva General balones sueltos
DP-PtsG	Defensiva Aerea Puntos por partido
DP-Comp	Defensiva Aerea Pases completos
DP-AttG	Defensiva Aerea Intentos de Pase
DP-YdsG	Defensiva Aerea Yardas por partido
DP-TD	Defensiva Aerea Anotaciones
DP-Int	Defensiva Aerea Intentos de Pase
DR-PtsG	Defensiva Terrestre Puntos por partido
DR-AttG	Defensiva Terrestre Intentos por partido
DR-YdsG	Defensiva Terrestre Yardas por partido
DR-TD	Defensiva Terrestre Anotaciones
DR-FUM	Defensiva Terrestre Balones sueltos

Cuadro 2.1: Descripción de variables

Script	Tabla	Anexo	Etapa
cargas.sql(E)	nfl.variables	(F)	Selección y obtención de los datos
cargas.sql(E)	nfl.equipos	(G)	Selección y obtención de los datos
cargas.sql(E)	nfl.gposvar	(H)	Selección y obtención de los datos
cargas.sql(E)	nfl.gposeq	(I)	Selección y obtención de los datos
cargas.sql(E)	nfl.asgposvar	(J)	Clasificación de los equipos
cargas.sql(E)	nfl.asgposeq	(K)	Clasificación de los equipos
cargas.sql(E)	nfl.resultados	(M)	Selección y obtención de los datos

Cuadro 2.2: Carga de tablas

- Diagrama entidad-relación.- *kvio 1.6.3 para linux*. Es una herramienta de kOffice para diagramas.

  Diagrama de la base de datos.- *MySQL WorkBench versión 5.1.7 para linux*[6] que es relativamente sencillo de utilizar pues sólo se generan las tablas visualmente, se agregan y definen los campos, así como las relaciones entre tablas y finalmente la herramienta automáticamente genera el script que construirá la base de datos, en el anexo (D), se presenta el script utilizado para la generación de la base de datos.

- Manejador de la base de datos.- *MySQL version 5.0.75 para linux*[7] .

- Herramienta de consulta.- *MySQL Query Browser version 1.2.12 para linux*[8] .

Una vez contruida la base de datos se cargaron los datos en las tablas, mismos que fueron previamente preparados[9] algunos de estos datos se obtuvieron directamente del sitio de la NFL, como es el caso de los resultados, otros sin embargo son el resultado de la etapa de clasificación, con el riesgo de perder un poco la secuencia se presenta en esta sección los scripts utilizados para cargar la base de datos, partiendo del supuesto de que ya se cuenta con ellos, en secciones posteriores se verá el detalle de como se obtuvieron estos datos, a continuación se presenta en la tabla (2.2): el script utilizado para cargar la tabla, la tabla cargada, la refernecia del anexo con el contenido de los datos y finalmente la etapa donde se generaron dichos datos.

El script cargas.sql utilizado para cargar la base de datos se puede ver en el anexo (E), las tablas nfl.asgposeq y nfl.resultados dependen de cada temporada, en el anexo correspondiente se presenta el ejemplo de la temporada 2003.

[6] dev.mysql.com
[7] dev.mysql.com
[8] dev.mysql.com
[9] La preparación consiste en obtener los datos adecuados y guardarlos en un archivo de texto csv; es decir texto separado por comas, para su posterior carga en la base de datos.

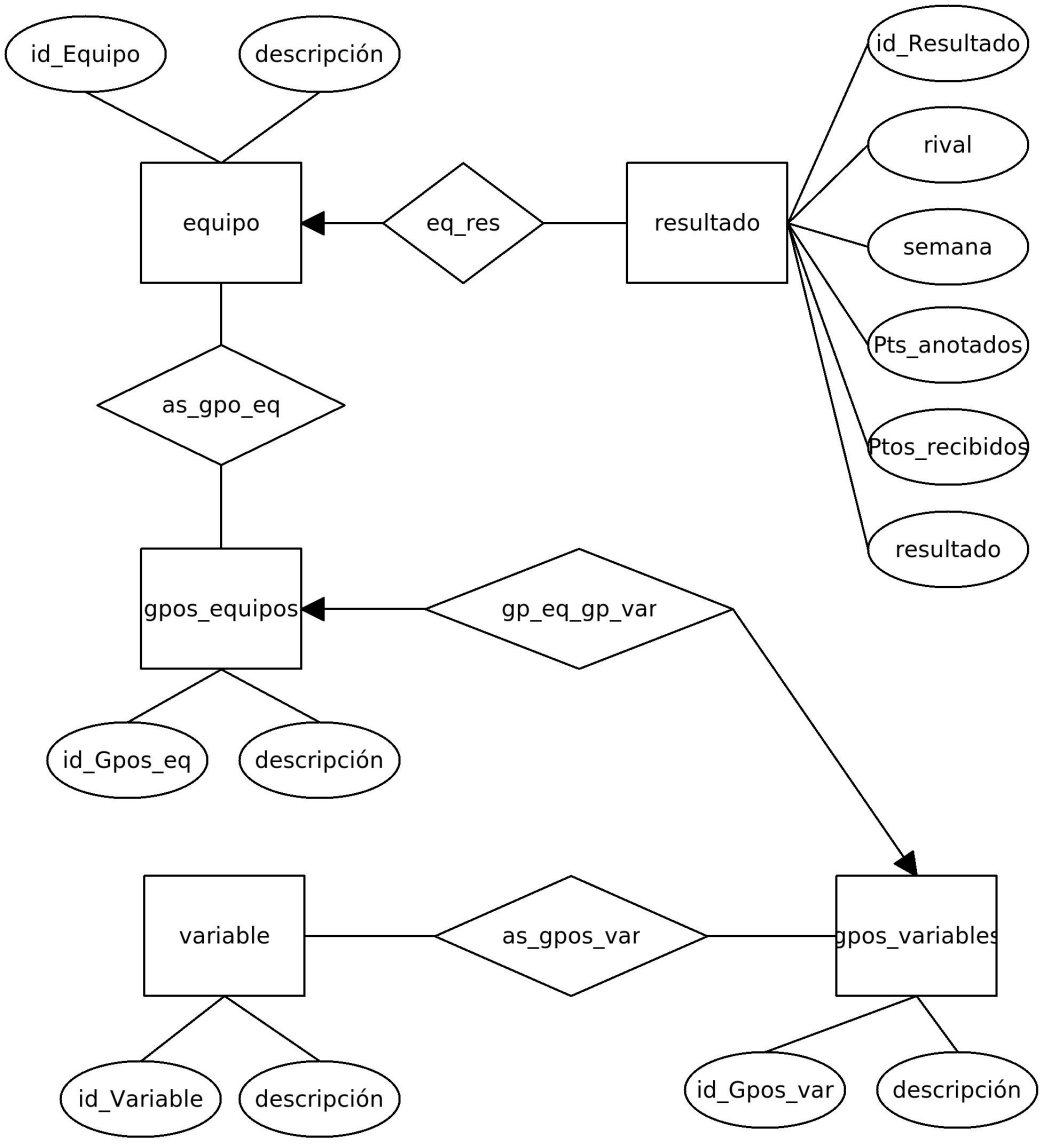

Figura 2.2: Diagrama entidad-relación de la base de datos

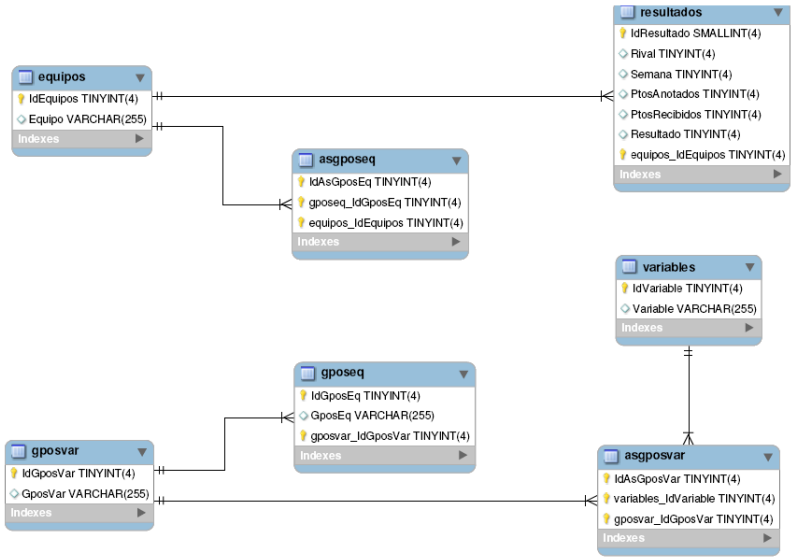

Figura 2.3: Base de datos

2.3. Reducción de variables componentes principales

El análisis de componentes principales se hizo con la herramienta de dominio público *R* version 2.8.1 (2008-12-22) para linux[10].

Se utilizó componentes principales con dos objetivos:

1. Para reducir el número de variables que utiliza el modelo y hacer más sencillo su manejo.

2. Para obtener nuevas clasificaciones de equipos de acuerdo a la información de sus variables.

El código en *R* utilizado para obtener este análisis se presenta en el anexo (N)[11].

Cargas Las cargas de las componentes principales respresentan el segundo objetivo de esta sección, pues indican los coeficientes de la combinación lineal de las variables que integran cada componente principal.

[10] www.r-project.org

[11] Este código se ejecuto para cada temporada cambiando los parámetros adecuados.

2.4. Agrupación SOM

En esta etapa del modelo se agrupan los equipos utilizando SOM. Para generar esta clasificación se utilizó la herramienta R junto la biblioteca (kohonen). El código en R utilizado para obtener éste análisis se presenta en el anexo (N)[12]. Al momento de clasificar los equipos, la biblioteca utilizada necesita como parámetro el número de neuronas, que para este caso cada neurona representa un grupo, para el caso de SOM se utilizaron seis neuronas en todas las temporadas.

2.5. Agrupación k-medias

En esta etapa del modelo se agrupan los equipos utilizando el método de clasificación no supervisada de k-medias. Para generar esta clasificación se utilizó la herramienta R junto la biblioteca (*cluster*), el código en R utilizado para obtener éste análisis se presenta en el anexo (N)[13]. Al momento de clasificar los equipos, k-medias necesita como parámetro el número grupos, para calcular este número se utiliza la metodología presentada en la sección (1.6)[14], la cual se aplica a cada temporada. De acuerdo con la metodología al obtener los resultados el estimador más adecuado es el que muestra un máximo, sin embargo en algunos casos éste no aparece como tal ya que el estimador es siempre creciente, para resolver esto, se tomaron dos críterios que se mencionan a continuación:

- El número máximo de grupos que se tomó es de cinco, ya que al utilizar más provoca que cada grupo quede con muy pocos equipos, lo que dificulta las cosas al momento de buscar encuentros entre equipos, pues en muchas ocasiones éstos no existen.
- En caso de haber más de un máximo se tomó el del menor número de grupos.

2.6. Predicción SOM vs k-medias

El objetivo de esta etapa del modelo es utilizar las agrupaciones realizadas en las secciones (2.4 y 2.5) con el propósito de obtener una predicción de los juegos de post-temporada de las temporadas 2003 a 2007, el mecanismo utilizado es el siguiente:

1. Se selecciona la clasificación hecha por SOM.
2. Se seleccionan los equipos rivales de algún partido de la post-temporada.
3. Se obtienen los resultados que obtuvo el equipo 1 contra todos los integrantes del grupo del equipo 2, se obtienen estadísticas de estos resultados.

[12] Este código se ejecuto para cada temporada cambiando los parámetros adecuados.
[13] Este código se ejecuto para cada temporada cambiando los parámetros adecuados.
[14] El código en R utilizado para obtener éste análisis se presenta en el anexo (N)

4. Se obtienen los resultados que obtuvo el equipo 2 contra todos los integrantes del grupo del equipo 1, se obtienen estadísticas de estos resultados.

5. Con las estadísticas obtenidas se genera una estimación del resultado de la siguiente manera: Para estimar los puntos que anotará el equipo 1.- se promedia el promedio de los puntos que anotó el quipo 1, con el promedio de los puntos que recibió el equipo 2, análogamente se obtiene el estimado de los puntos que anotará el equipo 2, y ambos darán el estimado del marcador.

6. Se selecciona la clasificación hecha por k-medias, y se repiten los pasos 2 al 5 con éstas nuevas agrupaciones.

7. Se comparan los pronósticos obtenidos con ambas agrupaciones contra los resultados reales, con el propósito de verificar cual pronóstico fue mejor y por consiguiente cuál agrupación fue más acertada.

Para ejemplificar se toman los equipos Seattle vs Chicago, en donde se obtuvieron los siguientes encuentros:

En la figura (2.4) se observan los encuentros generados por los equipos Chicago y Seattle.

A continuación en la figura (2.5) se presentan los resultados obtenidos:

Para obtener el pronóstico del resultado se realizó lo siguiente:

1. Se obtuvo el promedio de puntos recibidos y anotados para los 2 equipos.

2. Se obtuvo la varianza de puntos recibidos y anotados para los 2 equipos.

3. Para estimar los puntos que anotará Chicago se promedia, el promedio de los puntos que anotó, con el promedio de los puntos que recibió Seattle, análogamente se obtiene el estimado de los puntos que anotará Seattle, y ambos darán el estimado del marcador.

4. La varianza estimada también se obtuvo con el promedio de las varianzas.

De la misma forma se generó la estimación del marcador con las agrupaciones hechas por k-medias y se compararon ambas estimaciones con el resultado real, los resultados obtenidos se presentan en la figura (2.6).

Para verificar si el resultado estimado es correcto se utilizaron las reglas de pronósticos deportivos (*protouch*), es decir gana por 7 o más, la diferencia es de menos de 6 o pierde por 7 o más.

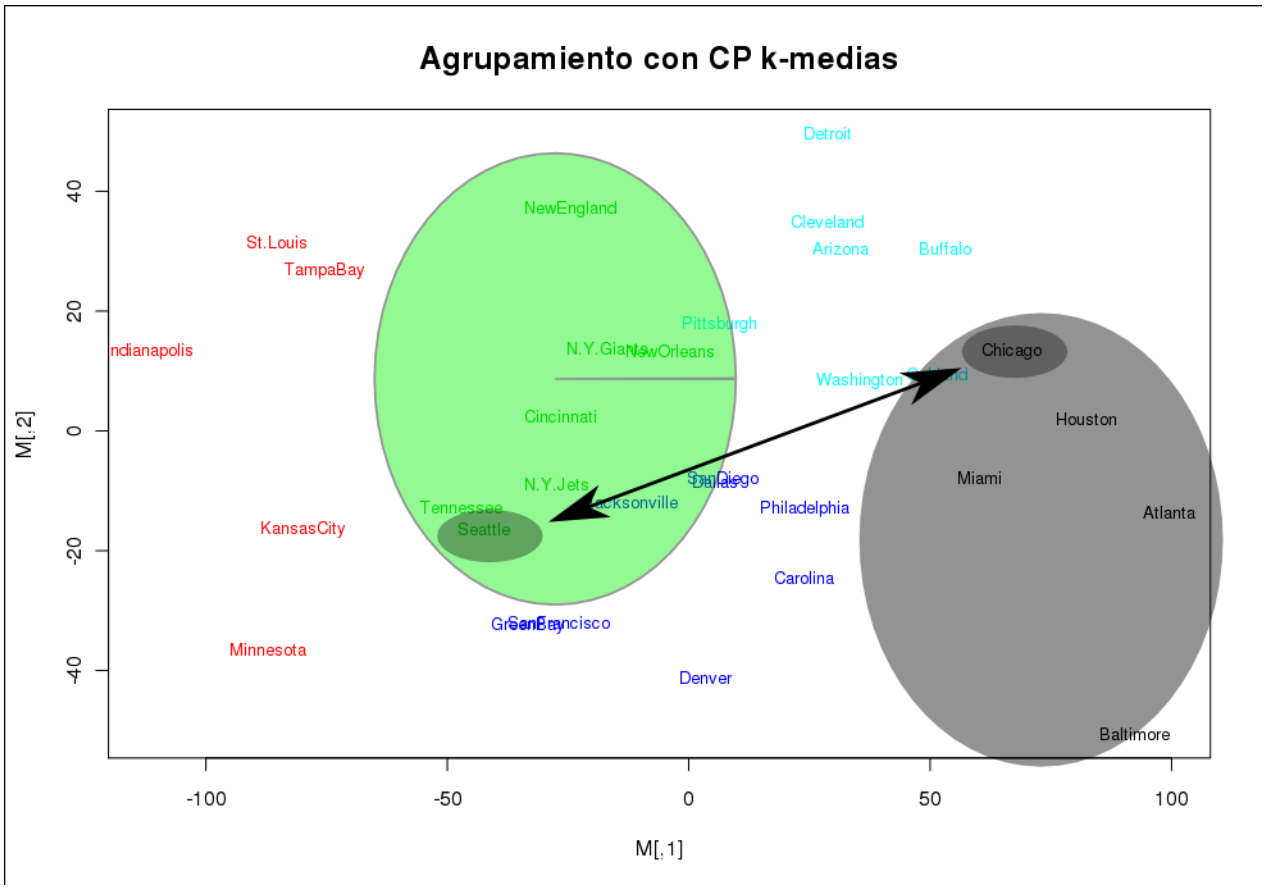

Figura 2.4: Encuentros generados por Chicago y Seattle

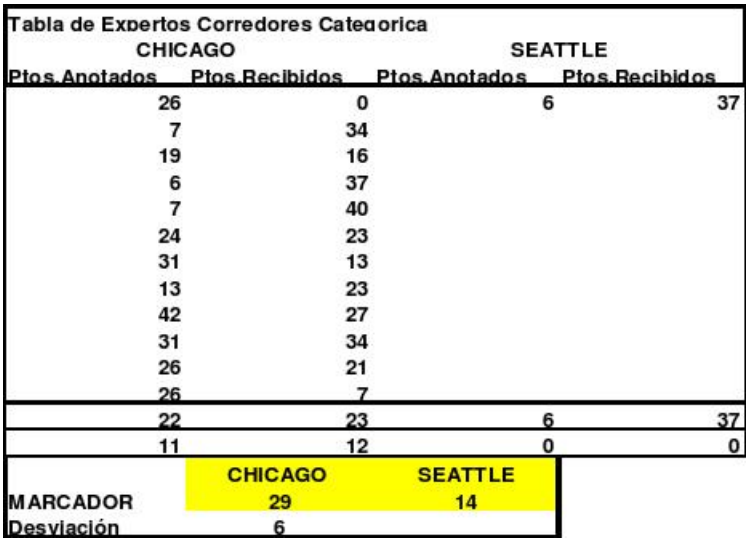

Tabla de Expertos Corredores Categorica

CHICAGO		SEATTLE	
Ptos.Anotados	Ptos.Recibidos	Ptos.Anotados	Ptos.Recibidos
26	0	6	37
7	34		
19	16		
6	37		
7	40		
24	23		
31	13		
13	23		
42	27		
31	34		
26	21		
26	7		
22	23	6	37
11	12	0	0

	CHICAGO	SEATTLE
MARCADOR	29	14
Desviación	6	

Figura 2.5: Resultados generados por Chicago y Seattle

CHICAGO	SEATTLE
27	24

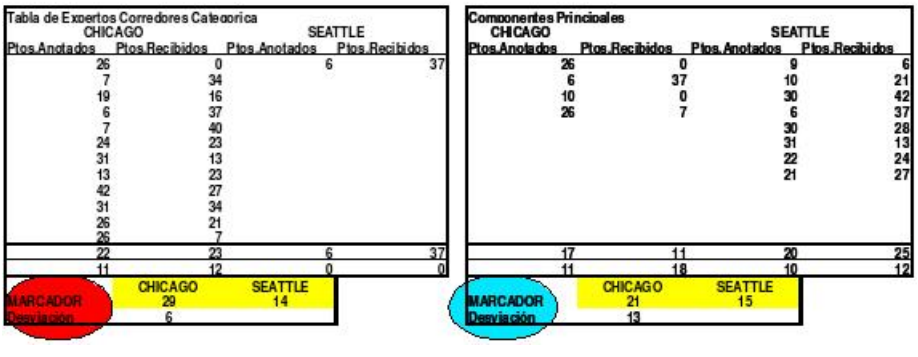

Tabla de Expertos Corredores Categorica

CHICAGO		SEATTLE	
Ptos.Anotados	Ptos.Recibidos	Ptos.Anotados	Ptos.Recibidos
26	0	6	37
7	34		
19	16		
6	37		
7	40		
24	23		
31	13		
13	23		
42	27		
31	34		
26	21		
26	7		
22	23	6	37
11	12	0	0

	CHICAGO	SEATTLE
MARCADOR	29	14
Desviación	6	

Componentes Principales

CHICAGO		SEATTLE	
Ptos.Anotados	Ptos.Recibidos	Ptos.Anotados	Ptos.Recibidos
26	0	9	6
6	37	10	21
10	0	30	42
26	7	6	37
		30	28
		31	13
		22	24
		21	27
17	11	20	25
11	18	10	12

	CHICAGO	SEATTLE
MARCADOR	21	15
Desviación	13	

Figura 2.6: Resultados generados por Chicago y Seattle para SOM y k-medias

2.7. Análisis de resultados y verificación del modelo.

El objetivo de esta etapa del modelo es analizar e interpretar los resultados obtenidos en las secciones anteriores desde el punto de vista estadístico así como desde el punto de vista del futbol americano y obtener un método que permita verificar la efectividad[15] del modelo.

A continuación se plantea un método para verificar la probabilidad de predecir con éxito el resultado de un encuentro que tiene el modelo, esto se hace para SOM, k-medias y de manera conjunta. Entonces si se consideran todos los encuentros realizados durante la predicción se tienen un total de 11 juegos por 5 temporadas, un total de 55 encuentros, de los cuales se sabe cuál es número de aciertos, entonces se tiene un modelo binomial en el que se obtiene la probabilidad de éxito con ayuda del estimador de máxima verosimilitud, de acuerdo con (2.1) y la información obtenida por la predicción hecha en la sección (2.6) el estimador máximo verosimil para estimar la probabilidad de predecir con éxito del modelo utilizando SOM, k-medias es:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{2.1}$$

donde $n = 55$

Para ambos de manera conjunta:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{2.2}$$

donde n es igual al total de encuentros donde ambos coincidieron en la predicción.

[15] Capacidad para producir el efecto deseado.

Capítulo 3

Resultados del modelo

En el presente capítulo se presentan y analizan los resultados obtenidos por el modelo presentado en el capítulo (2), después de aplicarlo a las estadísticas de las temporadas 2003-2007 de la NFL, éstos se presentan en el siguiente órden:

1. Reducción de variables.
2. Clasificación de los equipos.
3. Predicción de los resultados.
4. Análisis de Resultados y Verificación del modelo.

3.1. Reducción de variables componentes principales

A continuación se presenta el análisis de componentes principales, los resultados obtenidos para cada temporada son los siguientes:

En las figuras (3.1),(3.2),(3.3),(3.4),(3.5),(3.6),(3.7),(3.8),(3.9) y (3.10) se presentan los porcentajes explicados de la varianza de acuerdo al número de componentes principales seleccionados para las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente, de acuerdo con estos resultados, se consideraron en todas las temporadas 11 componentes principales ya que éstas explican un porcentaje mayor o igual al 95 % de la varianza, éste es el monto deseado, ya que la intención es reducir el número de variables, pero conservando la mayor cantidad de explicación de la información posible, en este caso la reducción es significativa ya que originalmente el número de variables era de 32.

Cargas Las cargas de las componentes principales indican los coeficientes de la combinación lineal de las variables que integran cada componente principal, en las figuras (3.11),(3.12),(3.13),(3.14) y (3.15) se presentan las cargas de las primeras once componentes principales de las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente.

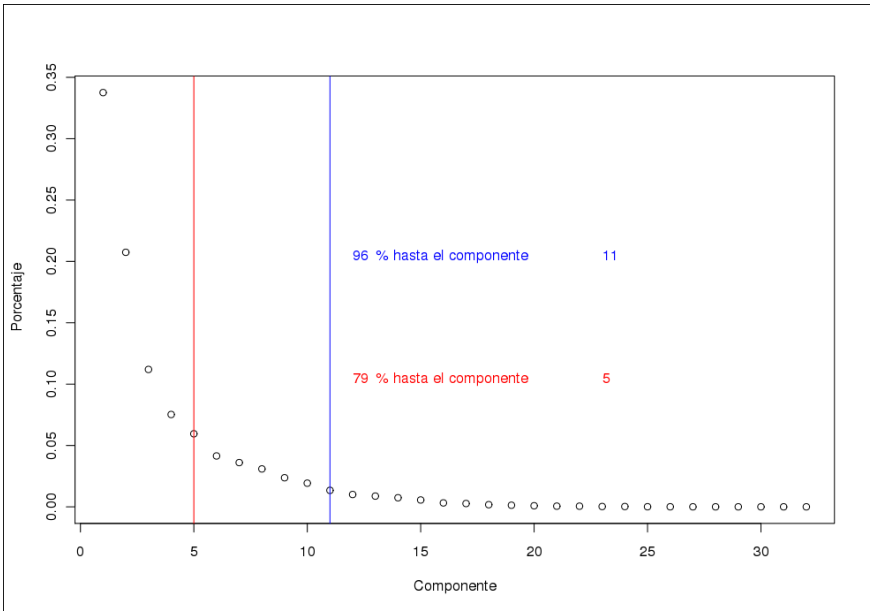

Figura 3.1: Porcentaje explicado por las componentes principales para la temporada 2003

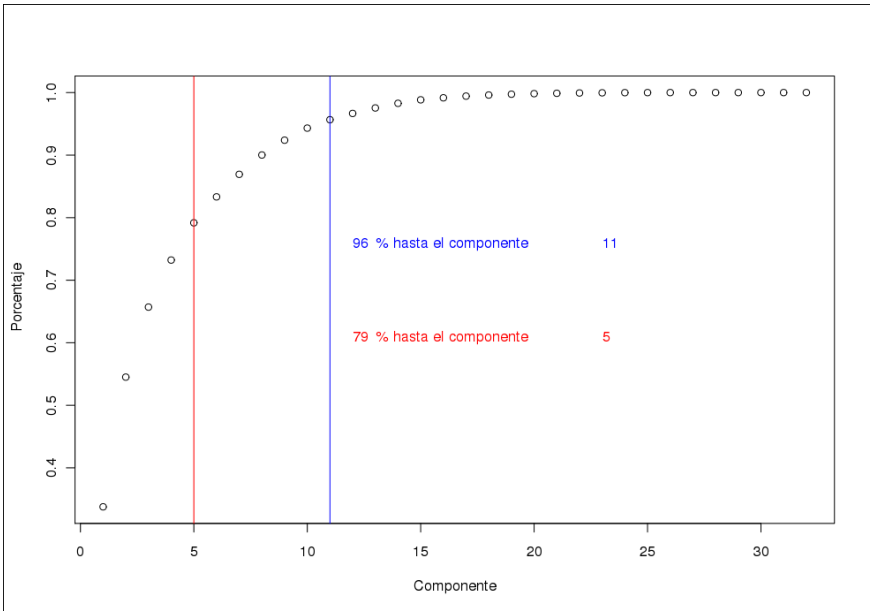

Figura 3.2: Porcentaje acumulado explicado por las componentes principales para la temporada 2003

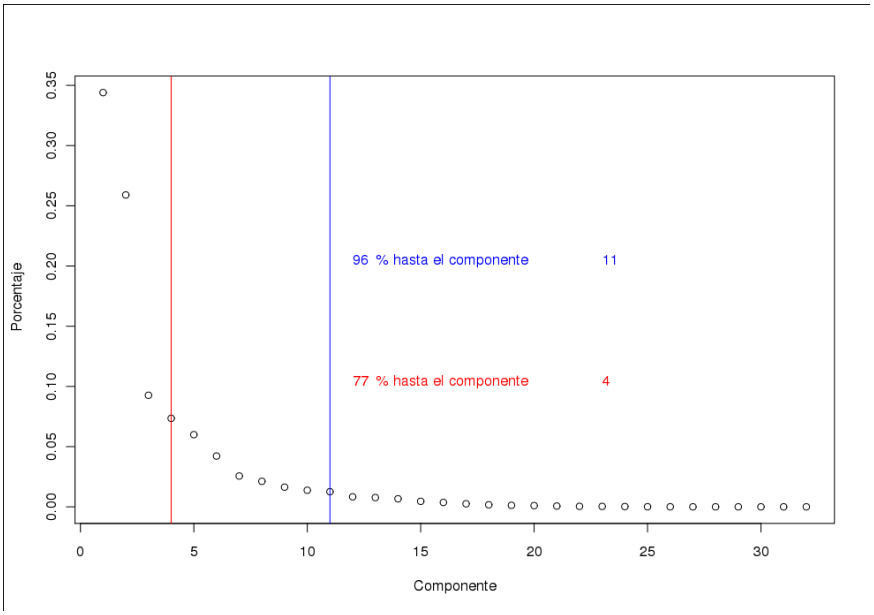

Figura 3.3: Porcentaje explicado por las componentes principales para la temporada 2004

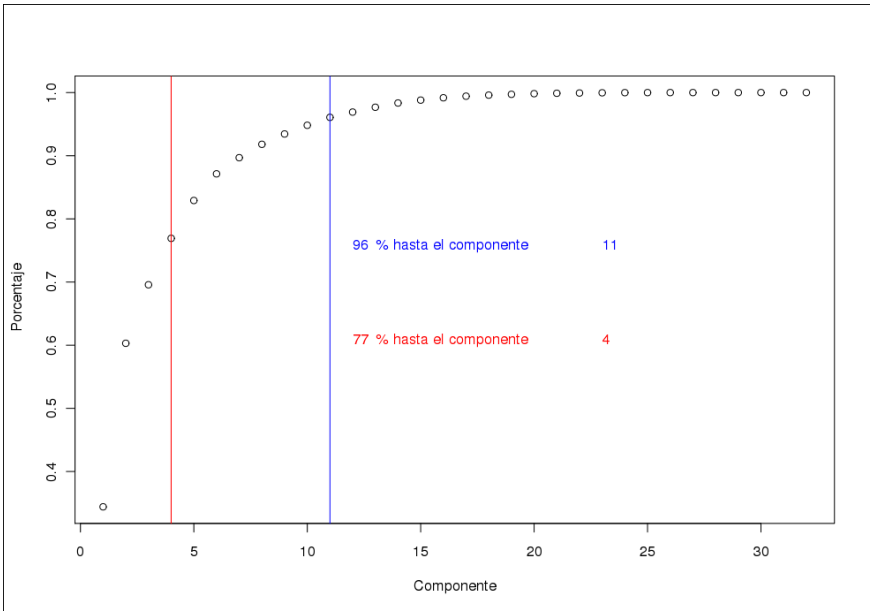

Figura 3.4: Porcentaje acumulado explicado por las componentes principales para la temporada 2004

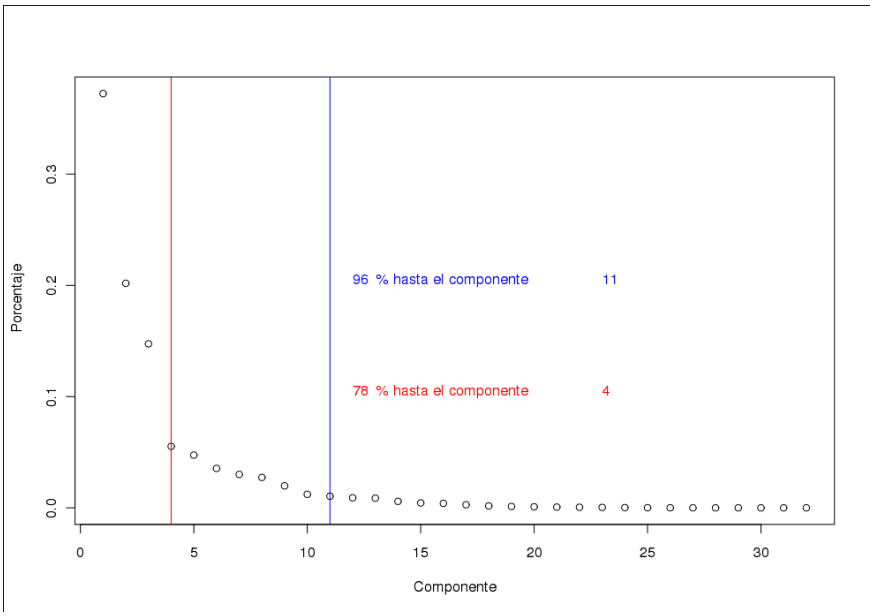

Figura 3.5: Porcentaje explicado por las componentes principales para la temporada 2005

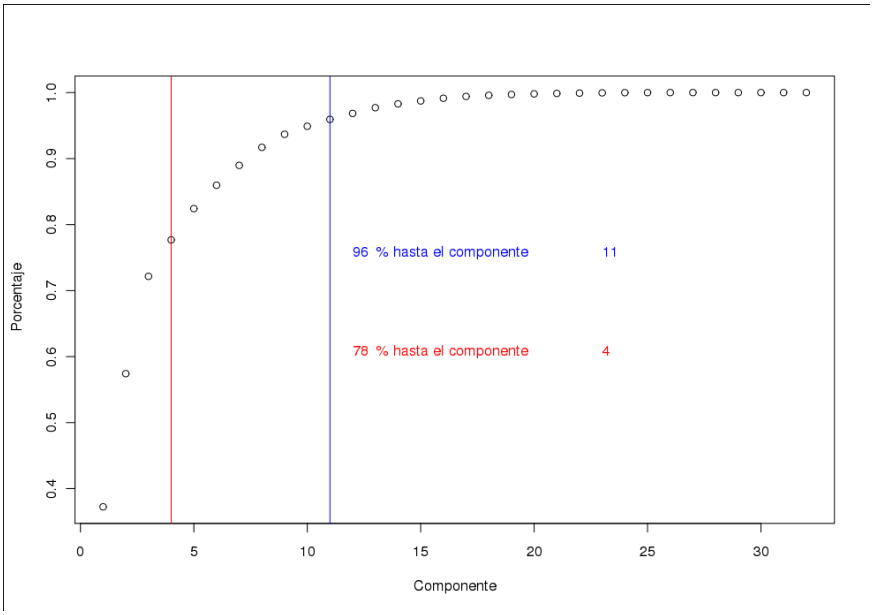

Figura 3.6: Porcentaje acumulado explicado por las componentes principales para la temporada 2005

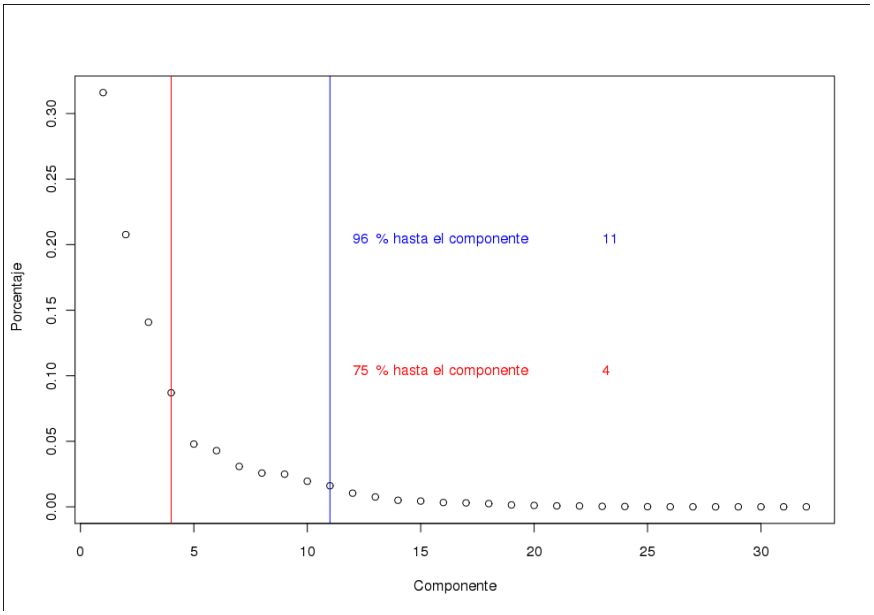

Figura 3.7: Porcentaje explicado por las componentes principales para la temporada 2006

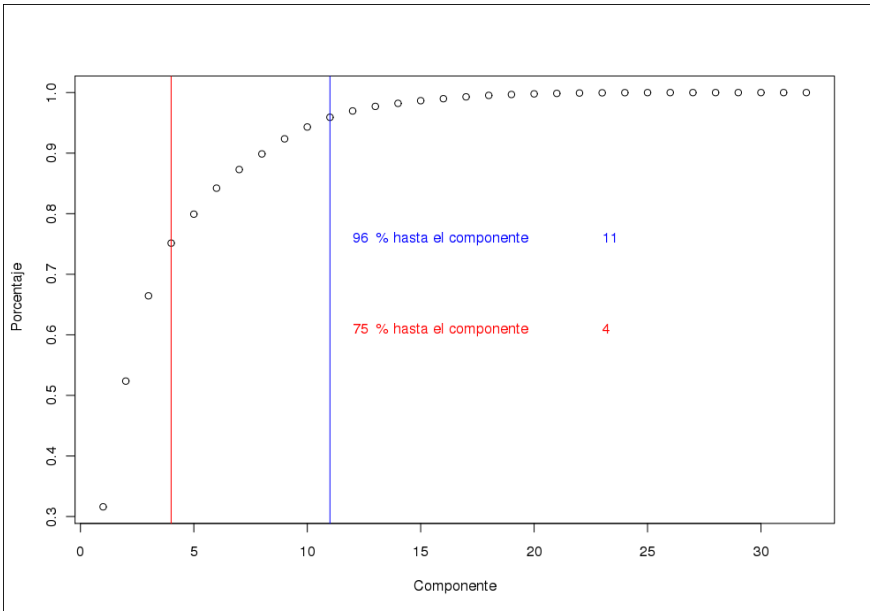

Figura 3.8: Porcentaje acumulado explicado por las componentes principales para la temporada 2006

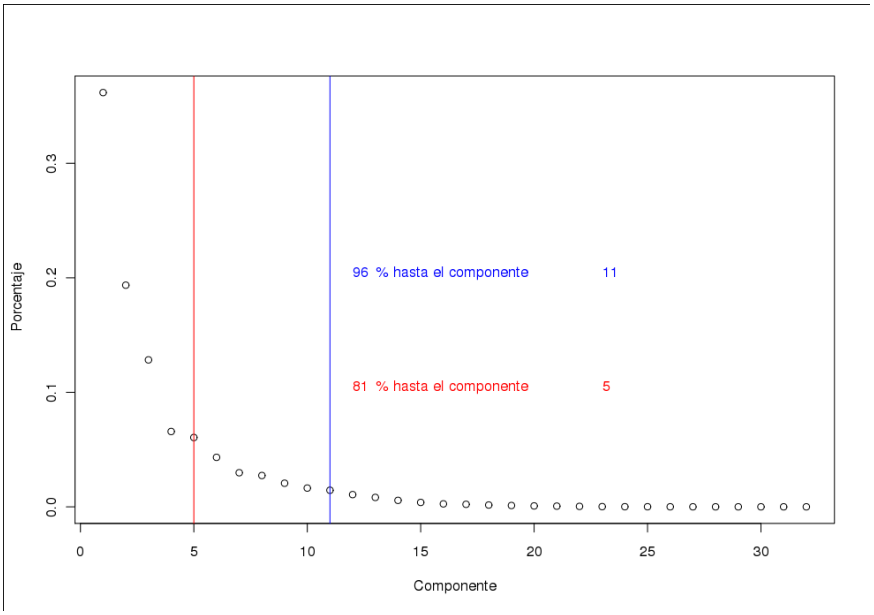

Figura 3.9: Porcentaje explicado por las componentes principales para la temporada 2007

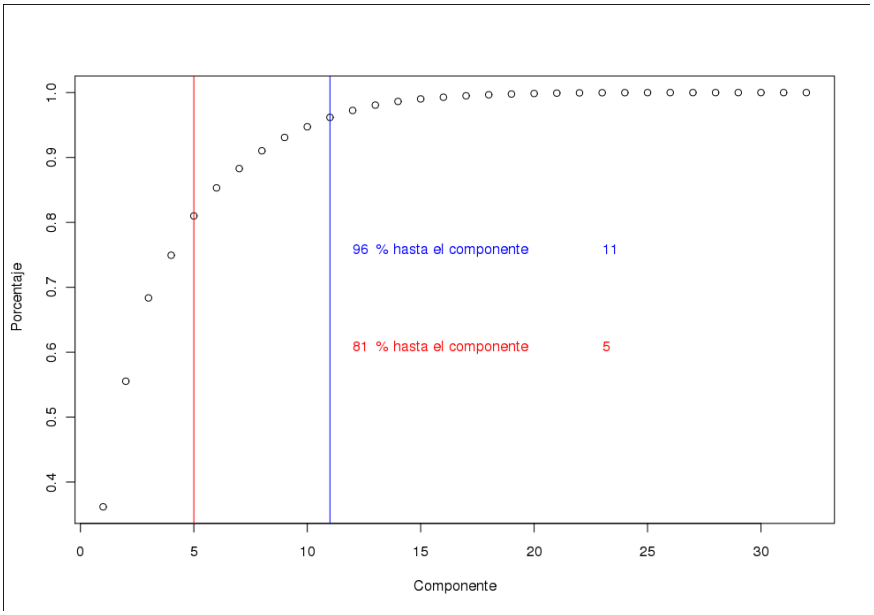

Figura 3.10: Porcentaje acumulado explicado por las componentes principales para la temporada 2007

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	5.4	-0.11	-0.71	-0.55	2.67	-0.97	0.77	1.55	0.57
Atlanta	6.43	-1.54	2.6	-0.35	0.89	2.5	-0.94	-0.98	-0.36
Baltimore	-3.48	3.95	4.6	-0.98	0.67	-0.75	-2.41	0.2	-1.31
Buffalo	-0.01	5.8	-0.4	0.16	-0.53	-1.02	-0.88	0.46	1.14
Carolina	-1.91	1.92	1.33	-0.64	0.2	-0.46	1.54	-0.3	-0.55
Chicago	1.85	2.64	0.98	0.85	0.28	-0.53	0.35	0.69	1.37
Cincinnati	1.63	-2.81	-0.47	-1.17	0.17	1.1	0.57	-0.04	0.63
Cleveland	1.1	2.97	-0.89	0.38	-1.02	-0.62	-0.04	1.62	-0.53
Dallas	-4.02	4.17	-1.29	-1.8	0.19	0.18	0.52	-0.67	1.08
Denver	-4.5	0.69	0.95	-3.87	1.05	1.77	0.33	-1.56	1.04
Detroit	3.98	-0.04	-1.94	2.43	2.42	1.27	0.51	-1.54	-1.35
GreenBay	-4.03	-2.61	2.77	0.45	0.14	-0.83	1.02	-0.36	-1.01
Houston	5.7	0.33	2	-0.4	-1.12	0.69	1.51	0.29	0.04
Indianapolis	-2.69	-3	-3.25	-2.02	-0.9	-0.03	-0.51	-0.21	-0.9
Jacksonville	-1.08	0.86	-0.16	-0.46	1.4	-0.11	0.6	-1.15	0.08
KansasCity	-1.37	-5.76	1.14	1.54	-3.09	-0.15	0.18	-0.81	1.07
Miami	-1.56	2.85	1.69	1.99	-0.62	1.78	0.82	-0.1	-0.32
Minnesota	-2.89	-4.97	0.6	-1.46	0.96	0.95	-0.88	3.12	-0.71
N.Y.Giants	-4.1	1.48	-0.33	4.33	0.27	1.47	-0.48	0.69	-0.7
N.Y.Jets	-0.46	-0.45	-1.34	-1.88	-0.71	-0.33	-0.92	0.14	-0.94
NewEngland	3.56	0.39	-2.67	1.27	1.18	-1.57	-1.43	-0.35	-0.19
NewOrleans	1.9	1.14	-1.37	-0.54	-3.85	-0.79	1.84	0.24	-0.97
Oakland	5.51	0.18	0.31	-1.32	-1.66	0.86	-0.59	0.44	0.14
Philadelphia	0.03	0.09	2.03	1.68	-2.2	-0.17	-1.25	-0.57	0.8
Pittsburgh	0.3	1.68	-1.55	-0.29	0.35	0.57	0.14	0.39	0.97
SanDiego	3.84	-2.28	1.06	-0.49	0.98	-3.25	0.07	-1.54	-0.13
SanFrancisco	-2.55	-1.47	1.12	-0.46	0.69	-1.68	0.48	0.44	-0.37
Seattle	-0.82	-2.45	1.12	1.42	0.25	-0.91	-1.16	-0.92	0.35
St.Louis	-2.22	-2.61	-2.81	0.94	0.02	0.31	-1.95	0.05	1.84
TampaBay	-3.41	1.05	-4.42	0.21	-0.44	0.23	0.19	-1.05	-1.3
Tennessee	-3.8	-2.19	0.13	1.75	1.72	-0.29	2.44	1.11	0.92
Washington	3.67	0.09	-0.82	-0.73	-0.36	0.81	-0.46	0.71	-0.42

Página 1

Figura 3.11: Cargas de las componentes principales temporada 2003

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	2.58	1.23	-0.09	0.18	-2.25	1.66	-0.94	1.29	-0.27
Atlanta	1.89	-1.08	-3.81	0.78	-0.86	-0.56	2.38	0.27	-0.44
Baltimore	4.82	-0.01	-0.35	0.77	1.25	-0.51	0.28	0.21	0.67
Buffalo	4.72	-1.49	0.85	0.72	-1.18	-0.63	-0.82	-1.09	1.53
Carolina	-0.58	-0.03	0.88	1.67	-0.45	0	-0.14	1.94	0.46
Chicago	3.19	4.25	-2.2	0.85	0.81	0.95	-0.48	-0.6	-0.46
Cincinnati	-0.65	0.22	0.25	1.86	-1.59	0.6	-0.92	0.27	0.17
Cleveland	1.55	4.24	-0.78	-0.18	-0.88	-3.85	0.29	-0.06	-1.02
Dallas	-1.19	0.86	-1.09	-2.25	1.04	-0.2	-0.5	-0.51	-0.02
Denver	1.05	-4.01	1.68	-2.62	-1.51	-2.12	0	0.35	-0.62
Detroit	0.13	2.33	-0.13	1.03	2.8	0.66	0.96	-0.63	0.13
GreenBay	-4.89	-1.57	1.97	-2.07	1.24	-0.65	-0.54	-1.15	0.01
Houston	0.05	-0.1	-2.2	0.64	0.66	1.94	-0.96	-1.9	-0.28
Indianapolis	-5.79	-3.14	2.01	3.62	-0.79	-1.02	-0.47	-0.97	-0.59
Jacksonville	2.25	-0.22	0.99	0.35	0.97	1.51	-0.2	0.56	-1.13
KansasCity	-6.35	-3.32	-2.02	-2.42	-2.07	1.47	0.72	0.04	1.52
Miami	2.59	4.95	2.08	-0.14	-2.42	0.35	-0.22	0.01	1.02
Minnesota	-5.52	-0.41	1.15	-0.04	0.01	0.38	-0.1	-0.52	-1.46
N.Y.Giants	1.13	-4.58	0.85	1.27	-0.72	0.42	-0.06	0.14	-0.03
N.Y.Jets	-3.36	2.18	-1.39	1.38	-0.84	-0.06	-0.56	0.24	-0.67
NewEngland	1.51	2.59	-0.96	0.34	-1.64	0.56	1.23	-1.32	0.17
NewOrleans	2.91	-3.3	0.01	-0.36	0	0.82	0.78	-0.36	-0.68
Oakland	-4.45	4.62	0.47	0	1.92	-1.07	1.35	0.2	1.2
Philadelphia	-0.29	-1.32	3.25	2.56	1.56	0.5	1.3	0.97	0.11
Pittsburgh	5.76	-5.44	-0.71	-1.87	-0.16	-0.06	0.74	-0.05	-0.63
SanDiego	-1.25	-5.33	-2.87	0.75	1.79	-1.15	-0.67	0.61	1.06
SanFrancisco	-1.42	4.83	-1.31	-1.38	-0.46	-0.33	-0.57	1.04	-0.4
Seattle	-2.28	-0.57	-1.64	1.16	0.37	-1.04	-0.24	0.17	0.36
St.Louis	-2.34	1.54	2.49	-2.51	0.39	1.74	1.64	0.53	-0.25
TampaBay	2.8	1.81	3.12	-1.03	-0.49	-0.56	-0.27	-1.2	0.54
Tennessee	-3.82	0.54	-1.28	-1.7	0.45	0.31	-1.54	0.87	-0.12
Washington	5.26	-0.29	0.79	-1.37	3.07	-0.06	-1.43	0.64	0.14

Página 1

Figura 3.12: Cargas de las componentes principales temporada 2004

Hoja1

Equipos	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	1.01	-2.59	-4.18	0.25	0.16	-1.94	2.59	-0.07	1.02
Atlanta	-0.4	0.2	2.27	1.12	-1.09	0.79	-0.32	-1.8	0.16
Baltimore	0.52	2.69	-2.14	0.89	0.54	-1.98	0	1.09	0.39
Buffalo	4.04	0.43	1.73	0.16	-0.44	0.53	1.26	-0.64	0.03
Carolina	-3.18	2.82	-0.44	-0.01	1.62	0.42	1.09	-0.44	-0.19
Chicago	-0.84	7.25	1.21	-1.07	-0.21	-0.42	0.56	0.65	-0.04
Cincinnati	-2.64	-3.74	0.78	-2.67	-0.88	1.74	2.01	-0.07	-1.33
Cleveland	3.6	2.39	-1.2	-0.81	0.13	0.09	-1.01	-0.66	-0.7
Dallas	-0.87	0.88	-1.52	2.76	-1.02	-1.28	0.27	-1.93	-0.86
Denver	-5.9	-0.06	2.5	1.22	-0.14	-0.83	0.4	0.6	-0.53
Detroit	2.85	1.14	0.06	-1.06	-0.43	1.45	0.67	1.52	-0.15
GreenBay	1.08	0.03	-4.72	-1.06	-1.31	0.24	-1.42	1.25	-0.34
Houston	6.45	-1.03	2.94	1.15	-1.39	1.1	-0.48	-0.39	1.05
Indianapolis	-4.89	-1.04	-0.93	-1.75	1.12	1.04	0.44	-0.62	0.24
Jacksonville	-2.48	1.93	-0.96	1.67	1.49	2.06	-0.35	0.37	-1.11
KansasCity	-4.42	-2.96	1.49	0.44	-2.51	-0.87	-0.87	0.76	-0.88
Miami	-0.21	0.32	0.35	-2.7	0.11	-1.33	-1.78	0.68	0.17
Minnesota	2.22	0.17	0.7	-1.08	1.06	-0.62	1.07	0	-0.73
N.Y.Giants	-1.03	-3.34	-0.52	1.59	3.32	0.44	-0.65	-0.53	-0.1
N.Y.Jets	3.7	-0.22	-2.53	1.78	-1.01	1.88	-0.55	0.04	-0.05
NewEngland	-3.36	-1.91	1.3	-1.01	0.01	0.17	-0.25	0.32	0.58
NewOrleans	5.42	3.4	-0.34	-1.63	0.07	0.04	-0.06	-1.29	-0.17
Oakland	3.37	-1.57	-0.92	-0.64	0.55	0.77	-0.71	-0.02	2.18
Philadelphia	2.94	-1.57	-1.12	-1.18	1.49	-1.72	-0.82	-1.09	-0.76
Pittsburgh	-4.32	3.04	1.03	0.73	-0.82	-0.7	-0.51	-0.57	0.58
SanDiego	-3.86	-1.75	0.68	0.78	0.76	0.25	0.46	1.04	1.95
SanFrancisco	6.11	-0.61	7.13	0.79	1.14	-1.16	0.38	1.33	-0.38
Seattle	-5.22	-0.44	1.83	-0.88	1.01	0.17	-1.97	-0.94	0.32
St.Louis	1.96	-5.31	-0.65	-0.42	-0.92	-0.8	0.27	-1.02	-0.12
TampaBay	-0.84	3.42	-1.68	1.22	0.35	0.64	0.67	0.53	0.33
Tennessee	2.16	-3.14	-1.8	1.69	0.16	0	-0.67	1.95	-1.14
Washington	-2.99	1.17	-0.34	-0.26	-2.92	-0.22	0.29	-0.06	0.59

Página 1

Figura 3.13: Cargas de las componentes principales temporada 2005

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	2.68	-1.62	-0.75	2.32	-0.21	-0.58	-2.26	-1.17	-0.32
Atlanta	0.42	0.44	3.79	-0.93	-0.27	-2.21	-1.49	0.72	-1.06
Baltimore	-6.26	4.18	-2.06	0.6	-0.04	-0.67	-0.95	-2.29	0.89
Buffalo	2.54	2.37	0.61	-0.7	2.37	-0.34	0.99	0.05	0.04
Carolina	-0.22	2.09	-2.04	0.16	-0.5	-0.5	1.2	0.15	0.1
Chicago	-4.74	1.39	0.59	1.55	2.04	-0.81	1	-0.26	0.01
Cincinnati	0.55	-3.04	0.89	2.46	1.13	-0.05	-0.24	-1.08	1.12
Cleveland	4.39	1.94	-1.55	-0.21	-0.37	1.42	-0.67	-1.67	0.01
Dallas	-2.47	-2.86	0.46	0.26	-1.21	0	-0.51	0.68	0.21
Denver	-0.41	1.67	2.04	-0.15	1.44	-0.79	-0.79	-0.42	-0.17
Detroit	4.75	-2.06	-2.27	2.84	1.2	1.31	0.27	1.65	0.33
GreenBay	-0.27	-1.01	-2.26	1.67	-1.39	0.4	0.64	0.14	0.08
Houston	3.37	0.8	0.84	0.4	-1.14	1.04	0.31	-0.38	-0.83
Indianapolis	-0.64	-5.63	-3.21	-3.25	1.74	0.57	0.69	-0.33	-0.88
Jacksonville	-4.8	2.13	2.16	-2.55	-0.67	2.02	-0.32	0.49	0.13
KansasCity	-0.13	0.5	1.86	-1.07	0.6	-0.59	-0.18	0.4	-0.66
Miami	-0.76	2.81	-3.43	0.87	-0.54	-2.14	1.16	1.54	-0.71
Minnesota	-1.82	2.85	1.35	4.19	-1.19	1.63	-0.1	1.04	-0.53
N.Y.Giants	-5.28	1.22	-0.15	-0.08	0.98	0.91	-0.02	-0.69	-1.02
N.Y.Jets	-3.36	-3.46	-2.65	-1.24	-2.04	-0.58	0.96	0.08	0.5
NewEngland	0.71	-1.82	1.98	1.15	0.45	0.63	0.19	-0.18	-0.16
NewOrleans	-0.59	1.05	1.05	-0.44	-0.57	1.88	2.23	-0.99	-1.76
Oakland	4.42	6.39	-3.2	-3.81	0.54	0.9	-1.24	1.08	0.99
Philadelphia	-1.59	-3.43	-1.3	-0.19	2.3	-0.19	0.3	0.27	1.81
Pittsburgh	-2.57	0.38	0.45	0.44	-1.27	0.81	-0.77	0.84	2.28
SanDiego	-5.19	-1.76	2.54	-1.23	0.11	-0.12	-0.38	1.29	0.02
SanFrancisco	3.85	-0.69	2.86	-0.22	-0.56	0.25	0.19	0.79	0.33
Seattle	1.06	-0.28	-0.27	0.34	0.55	-2.38	-0.66	0.36	-1.15
St.Louis	0.64	-3.87	-3.07	-0.97	-0.86	0.86	-2.17	0.02	-1.55
TampaBay	4.27	2.13	-1.18	0.77	-0.43	-1.57	0.91	-0.36	0.15
Tennessee	4.88	-1.1	4.03	-0.98	0.73	0.75	0.62	-0.57	0.76
Washington	2.58	-1.73	1.89	-1.97	-2.95	-1.87	1.12	-1.18	1.04

Página 1

Figura 3.14: Cargas de las componentes principales temporada 2006

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	1.01	-2.59	-4.18	0.25	0.16	-1.94	2.59	-0.07	1.02
Atlanta	-0.4	0.2	2.27	1.12	-1.09	0.79	-0.32	-1.8	0.16
Baltimore	0.52	2.69	-2.14	0.89	0.54	-1.98	0	1.09	0.39
Buffalo	4.04	0.43	1.73	0.16	-0.44	0.53	1.26	-0.64	0.03
Carolina	-3.18	2.82	-0.44	-0.01	1.62	0.42	1.09	-0.44	-0.19
Chicago	-0.84	7.25	1.21	-1.07	-0.21	-0.42	0.56	0.65	-0.04
Cincinnati	-2.64	-3.74	0.78	-2.67	-0.88	1.74	2.01	-0.07	-1.33
Cleveland	3.6	2.39	-1.2	-0.81	0.13	0.09	-1.01	-0.66	-0.7
Dallas	-0.87	0.88	-1.52	2.76	-1.02	-1.28	0.27	-1.93	-0.86
Denver	-5.9	-0.06	2.5	1.22	-0.14	-0.83	0.4	0.6	-0.53
Detroit	2.85	1.14	0.06	-1.06	-0.43	1.45	0.67	1.52	-0.15
GreenBay	1.08	0.03	-4.72	-1.06	-1.31	0.24	-1.42	1.25	-0.34
Houston	6.45	-1.03	2.94	1.15	-1.39	1.1	-0.48	-0.39	1.05
Indianapolis	-4.89	-1.04	-0.93	-1.75	1.12	1.04	0.44	-0.62	0.24
Jacksonville	-2.48	1.93	-0.96	1.67	1.49	2.06	-0.35	0.37	-1.11
KansasCity	-4.42	-2.96	1.49	0.44	-2.51	-0.87	-0.87	0.76	-0.88
Miami	-0.21	0.32	0.35	-2.7	0.11	-1.33	-1.78	0.68	0.17
Minnesota	2.22	0.17	0.7	-1.08	1.06	-0.62	1.07	0	-0.73
N.Y.Giants	-1.03	-3.34	-0.52	1.59	3.32	0.44	-0.65	-0.53	-0.1
N.Y.Jets	3.7	-0.22	-2.53	1.78	-1.01	1.88	-0.55	0.04	-0.05
NewEngland	-3.36	-1.91	1.3	-1.01	0.01	0.17	-0.25	0.32	0.58
NewOrleans	5.42	3.4	-0.34	-1.63	0.07	0.04	-0.06	-1.29	-0.17
Oakland	3.37	-1.57	-0.92	-0.64	0.55	0.77	-0.71	-0.02	2.18
Philadelphia	2.94	-1.57	-1.12	-1.18	1.49	-1.72	-0.82	-1.09	-0.76
Pittsburgh	-4.32	3.04	1.03	0.73	-0.82	-0.7	-0.51	-0.57	0.58
SanDiego	-3.86	-1.75	0.68	0.78	0.76	0.25	0.46	1.04	1.95
SanFrancisco	6.11	-0.61	7.13	0.79	1.14	-1.16	0.38	1.33	-0.38
Seattle	-5.22	-0.44	1.83	-0.88	1.01	0.17	-1.97	-0.94	0.32
St.Louis	1.96	-5.31	-0.65	-0.42	-0.92	-0.8	0.27	-1.02	-0.12
TampaBay	-0.84	3.42	-1.68	1.22	0.35	0.64	0.67	0.53	0.33
Tennessee	2.16	-3.14	-1.8	1.69	0.16	0	-0.67	1.95	-1.14
Washington	-2.99	1.17	-0.34	-0.26	-2.92	-0.22	0.29	-0.06	0.59

Página 1

Figura 3.15: Cargas de las componentes principales temporada 2007

Intentar hacer observaciones en este punto resulta muy complicado, ya que son demasiados números y equipos, por lo que para analizar las componentes principales utilizamos el siguiente procedimiento para cada temporada:

- Obtenemos la media y la desviación estándar de cada componente principal.
- Comparamos si el coeficiente de la componente principal de cada equipo se aleja de la media más de una desviación estándar, de ser así, este equipo aparecera seleccionado en el resultado del análisis.
- Se genera una tabla con los resultados.

En las en las figuras (3.16),(3.17),(3.18),(3.19) y (3.20) se presenta el resultado de este análisis para las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente. Los equipos que en la matriz tengan un resultado diferente de (NA), serán equipos que sean altamente influenciados por la componente principal respectiva, a partir de aquí se pueden buscar patrones nuevos con ayuda de un experto en el área.

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	5.4	NA	NA	NA	2.67	NA	NA	1.55	NA
Atlanta	6.43	NA	2.6	NA	NA	2.5	NA	NA	NA
Baltimore	-3.48	3.96	4.6	NA	NA	NA	-2.41	NA	-1.31
Buffalo	NA	5.8	NA	NA	NA	NA	NA	NA	1.14
Carolina	NA	NA	NA	NA	NA	NA	1.54	NA	NA
Chicago	NA	2.64	NA	NA	NA	NA	NA	NA	1.37
Cincinnati	NA	-2.81	NA	NA	NA	NA	NA	NA	NA
Cleveland	NA	2.97	NA	NA	NA	NA	NA	1.62	NA
Dallas	-4.02	4.17	NA	-1.8	NA	NA	NA	NA	1.08
Denver	-4.5	NA	NA	-3.87	NA	1.77	NA	-1.56	1.04
Detroit	3.98	NA	-1.94	2.43	2.42	1.27	NA	-1.54	-1.35
GreenBay	-4.03	NA	2.77	NA	NA	NA	NA	NA	-1.01
Houston	5.7	NA	2	NA	NA	NA	1.51	NA	NA
Indianapolis	NA	-3	-3.25	-2.02	NA	NA	NA	NA	-0.9
Jacksonville	NA	NA	NA	NA	NA	NA	NA	-1.15	NA
KansasCity	NA	-5.76	NA	NA	-3.09	NA	NA	NA	1.07
Miami	NA	2.85	NA	1.99	NA	1.78	NA	NA	NA
Minnesota	NA	-4.97	NA	NA	NA	NA	NA	3.12	NA
N.Y.Giants	-4.1	NA	NA	4.33	NA	1.47	NA	NA	NA
N.Y.Jets	NA	NA	NA	-1.88	NA	NA	NA	NA	-0.94
NewEngland	3.56	NA	-2.67	NA	NA	-1.57	-1.43	NA	NA
NewOrleans	NA	NA	NA	NA	-3.85	NA	1.84	NA	-0.97
Oakland	5.51	NA	NA	NA	-1.66	NA	NA	NA	NA
Philadelphia	NA	NA	2.03	1.68	-2.2	NA	-1.25	NA	NA
Pittsburgh	NA	NA	NA	NA	NA	NA	NA	NA	NA
SanDiego	3.84	NA	NA	NA	NA	-3.25	NA	-1.54	NA
SanFrancisco	NA	NA	NA	NA	NA	-1.68	NA	NA	NA
Seattle	NA	NA	NA	NA	NA	NA	-1.16	NA	NA
St.Louis	NA	NA	-2.81	NA	NA	NA	-1.95	NA	1.84
TampaBay	-3.41	NA	-4.42	NA	NA	NA	NA	-1.05	-1.3
Tennessee	-3.8	NA	NA	1.75	1.72	NA	2.44	1.11	NA
Washington	3.67	NA	NA	NA	NA	NA	NA	NA	NA

Página 1

Figura 3.16: Análisis de las componentes principales temporada 2003

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	NA	NA	NA	NA	-2.25	1.66	-0.94	1.29	NA
Atlanta	NA	NA	-3.81	NA	NA	NA	2.38	NA	NA
Baltimore	4.82	NA	NA	NA	NA	NA	NA	NA	NA
Buffalo	4.72	NA	NA	NA	NA	NA	NA	-1.09	1.53
Carolina	NA	NA	NA	1.67	NA	NA	NA	1.94	NA
Chicago	NA	4.25	-2.2	NA	NA	NA	NA	NA	NA
Cincinnati	NA	NA	NA	1.86	-1.59	NA	-0.92	NA	NA
Cleveland	NA	4.24	NA	NA	NA	-3.85	NA	NA	-1.02
Dallas	NA	NA	NA	-2.25	NA	NA	NA	NA	NA
Denver	NA	-4.01	NA	-2.62	-1.51	-2.12	NA	NA	NA
Detroit	NA	NA	NA	NA	2.8	NA	NA	NA	NA
GreenBay	-4.89	NA	1.97	-2.07	NA	NA	NA	-1.15	NA
Houston	NA	NA	-2.21	NA	NA	1.94	-0.96	-1.9	NA
Indianapolis	-5.79	-3.14	2.01	3.62	NA	NA	NA	-0.97	NA
Jacksonville	NA	NA	NA	NA	NA	1.51	NA	NA	-1.13
KansasCity	-6.35	-3.32	-2.02	-2.43	-2.07	1.47	NA	NA	1.52
Miami	NA	4.95	2.08	NA	-2.42	NA	NA	NA	1.02
Minnesota	-5.52	NA	NA	NA	NA	NA	NA	NA	-1.46
N.Y.Giants	NA	-4.58	NA	NA	NA	NA	NA	NA	NA
N.Y.Jets	NA	NA	NA	NA	NA	NA	NA	NA	NA
NewEngland	NA	NA	NA	NA	-1.64	NA	1.23	-1.32	NA
NewOrleans	NA	-3.3	NA	NA	NA	NA	NA	NA	NA
Oakland	-4.45	4.62	NA	NA	1.92	NA	1.35	NA	1.2
Philadelphia	NA	NA	3.25	2.56	1.56	NA	1.3	NA	NA
Pittsburgh	5.76	-5.44	NA	-1.87	NA	NA	NA	NA	NA
SanDiego	NA	-5.33	-2.87	NA	1.79	NA	NA	NA	1.06
SanFrancisco	NA	4.83	NA	NA	NA	NA	NA	1.04	NA
Seattle	NA	NA	NA	NA	NA	NA	NA	NA	NA
St.Louis	NA	NA	2.49	-2.51	NA	1.74	1.64	NA	NA
TampaBay	NA	NA	3.12	NA	NA	NA	NA	-1.2	NA
Tennessee	-3.82	NA	NA	-1.7	NA	NA	-1.54	NA	NA
Washington	5.26	NA	NA	NA	3.07	NA	-1.43	NA	NA

Página 1

Figura 3.17: Análisis de las componentes principales temporada 2004

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	NA	-2.59	-4.18	NA	NA	-1.94	2.59	NA	1.02
Atlanta	NA	NA	2.27	NA	NA	NA	NA	-1.8	NA
Baltimore	NA	2.69	NA	NA	NA	-1.98	NA	1.09	NA
Buffalo	4.04	NA	NA	NA	NA	NA	1.26	NA	NA
Carolina	NA	2.82	NA	NA	1.62	NA	1.09	NA	NA
Chicago	NA	7.25	NA	NA	NA	NA	NA	NA	NA
Cincinnati	NA	-3.74	NA	-2.67	NA	1.75	2.02	NA	-1.33
Cleveland	3.6	NA	NA	NA	NA	NA	-1.02	NA	NA
Dallas	NA	NA	NA	2.76	NA	-1.28	NA	-1.93	-0.86
Denver	-5.91	NA	2.5	NA	NA	NA	NA	NA	NA
Detroit	NA	NA	NA	NA	NA	1.45	NA	1.52	NA
GreenBay	NA	NA	-4.72	NA	-1.31	NA	-1.42	1.25	NA
Houston	6.45	NA	2.94	NA	-1.39	1.1	NA	NA	1.05
Indianapolis	-4.89	NA	NA	-1.75	NA	NA	NA	NA	NA
Jacksonville	NA	NA	NA	1.67	1.49	2.06	NA	NA	-1.11
KansasCity	-4.42	-2.96	NA	NA	-2.51	NA	NA	NA	-0.88
Miami	NA	NA	NA	-2.7	NA	-1.33	-1.78	NA	NA
Minnesota	NA	NA	NA	NA	NA	NA	1.07	NA	NA
N.Y.Giants	NA	-3.34	NA	1.59	3.32	NA	NA	NA	NA
N.Y.Jets	3.7	NA	-2.53	1.78	NA	1.88	NA	NA	NA
NewEngland	NA	NA	NA	NA	NA	NA	NA	NA	NA
NewOrleans	5.42	3.41	NA	-1.63	NA	NA	NA	-1.29	NA
Oakland	NA	NA	NA	NA	NA	NA	NA	NA	2.18
Philadelphia	NA	NA	NA	NA	1.49	-1.72	NA	-1.09	NA
Pittsburgh	-4.32	3.04	NA	NA	NA	NA	NA	NA	NA
SanDiego	-3.86	NA	NA	NA	NA	NA	NA	1.04	1.95
SanFrancisco	6.11	NA	7.13	NA	NA	-1.16	NA	1.33	NA
Seattle	-5.22	NA	NA	NA	NA	NA	-1.97	NA	NA
St.Louis	NA	-5.31	NA	NA	NA	NA	NA	-1.02	NA
TampaBay	NA	3.42	NA	NA	NA	NA	NA	NA	NA
Tennessee	NA	-3.15	NA	1.69	NA	NA	NA	1.95	-1.14
Washington	NA	NA	NA	NA	-2.92	NA	NA	NA	NA

Página 1

Figura 3.18: Análisis de las componentes principales temporada 2005

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	NA	NA	NA	2.32	NA	NA	-2.26	-1.17	NA
Atlanta	NA	NA	3.79	NA	NA	2.21	-1.49	NA	-1.06
Baltimore	-6.26	4.19	NA	NA	NA	NA	NA	-2.29	NA
Buffalo	NA	NA	NA	NA	2.37	NA	NA	NA	NA
Carolina	NA	NA	NA	NA	NA	NA	1.2	NA	NA
Chicago	-4.75	NA	NA	NA	2.05	NA	NA	NA	NA
Cincinnati	NA	-3.04	NA	2.46	NA	NA	NA	-1.08	1.12
Cleveland	4.39	NA	NA	NA	NA	-1.42	NA	-1.67	NA
Dallas	NA	-2.86	NA	NA	NA	NA	NA	NA	NA
Denver	NA	NA	NA	NA	1.44	NA	NA	NA	NA
Detroit	4.75	NA	-2.27	2.84	NA	-1.31	NA	1.65	NA
GreenBay	NA	NA	-2.26	NA	-1.39	NA	NA	NA	NA
Houston	3.37	NA	NA	NA	NA	NA	NA	NA	NA
Indianapolis	NA	-5.63	-3.21	-3.25	1.74	NA	NA	NA	NA
Jacksonville	-4.8	NA	NA	-2.55	NA	-2.02	NA	NA	NA
KansasCity	NA	NA	NA	NA	NA	NA	NA	NA	NA
Miami	NA	2.82	-3.43	NA	NA	2.14	1.16	1.54	NA
Minnesota	NA	2.86	NA	4.19	NA	-1.63	NA	1.04	NA
N.Y.Giants	-5.28	NA	NA	NA	NA	NA	NA	NA	-1.02
N.Y.Jets	-3.36	-3.46	-2.65	NA	-2.05	NA	NA	NA	NA
NewEngland	NA	NA	NA	NA	NA	NA	NA	NA	NA
NewOrleans	NA	NA	NA	NA	NA	-1.88	2.23	-0.99	-1.76
Oakland	4.42	6.39	-3.2	-3.81	NA	NA	-1.24	1.08	NA
Philadelphia	NA	-3.43	NA	NA	2.3	NA	NA	NA	1.81
Pittsburgh	NA	NA	NA	NA	-1.27	NA	NA	NA	2.28
SanDiego	-5.19	NA	2.54	NA	NA	NA	NA	1.3	NA
SanFrancisco	3.85	NA	2.86	NA	NA	NA	NA	NA	NA
Seattle	NA	NA	NA	NA	NA	2.38	NA	NA	-1.15
St.Louis	NA	-3.87	-3.08	NA	NA	NA	-2.17	NA	-1.55
TampaBay	4.27	NA	NA	NA	NA	1.57	NA	NA	NA
Tennessee	4.88	NA	4.03	NA	NA	NA	NA	NA	NA
Washington	NA	NA	NA	-1.97	-2.95	1.87	1.12	-1.18	1.04

Página 1

Figura 3.19: Análisis de las componentes principales temporada 2006

Hoja1

Equipo	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Arizona	NA	-3.31	NA	NA	NA	NA	1.41	NA	NA
Atlanta	4.33	NA	NA	NA	NA	NA	NA	NA	NA
Baltimore	NA	NA	NA	NA	-3.42	NA	1.22	-1.33	NA
Buffalo	3.53	NA	2.66	NA	NA	1.24	NA	1.26	-1.59
Carolina	NA	NA	NA	NA	NA	NA	NA	NA	-1.19
Chicago	NA	NA	NA	1.48	NA	-1.74	1.06	NA	NA
Cincinnati	NA	-3.76	NA	NA	NA	NA	NA	NA	-1.56
Cleveland	NA	-2.8	NA	NA	NA	2.22	-1.49	NA	NA
Dallas	-4	NA	NA	NA	NA	NA	NA	NA	NA
Denver	NA	NA	NA	-1.93	2.19	-1.87	-1.69	-1.22	NA
Detroit	4.55	-5.26	2.96	NA	NA	-1.28	NA	NA	1.84
GreenBay	-3.75	NA	-2.06	NA	NA	NA	NA	NA	NA
Houston	NA	NA	NA	NA	NA	NA	NA	1.07	NA
Indianapolis	-4.83	NA	NA	2.11	2.67	NA	NA	-1.47	NA
Jacksonville	-4	NA	NA	-2.19	NA	1.75	NA	1.34	NA
KansasCity	NA	NA	-3.39	2.22	NA	NA	NA	NA	NA
Miami	5.46	NA	-3.39	-1.94	NA	NA	NA	-1.5	NA
Minnesota	NA	NA	6.96	NA	NA	NA	-1.73	NA	NA
N.Y.Giants	-7.95	-2.76	-2.6	NA	NA	NA	NA	NA	NA
N.Y.Jets	NA	-5.4	NA	NA	-1.62	-1.35	NA	NA	-1.71
NewEngland	NA	NA	NA	-1.72	NA	NA	NA	NA	NA
NewOrleans	NA	2.58	-2.29	NA	NA	NA	NA	NA	NA
Oakland	NA	3.21	NA	-3.81	2.33	NA	NA	1.32	NA
Philadelphia	NA	NA	NA	NA	-1.53	NA	-1.19	1.44	1.27
Pittsburgh	-4.76	3.86	NA	NA	NA	-2.66	-1.36	1.16	-0.84
SanDiego	NA	NA	NA	NA	1.95	2.23	1.81	NA	NA
SanFrancisco	5.53	NA	NA	2.64	-2.04	1.49	NA	NA	NA
Seattle	NA	NA	NA	2.24	NA	NA	1.29	1.16	NA
St.Louis	3.83	NA	NA	NA	NA	NA	1.47	1.11	NA
TampaBay	NA	3.09	NA	NA	NA	-1.53	NA	-1.31	NA
Tennessee	NA	3.75	2.33	NA	NA	NA	1.72	NA	NA
Washington	NA	NA	NA	NA	-1.71	NA	NA	NA	NA

Página 1

Figura 3.20: Análisis de las componentes principales temporada 2007

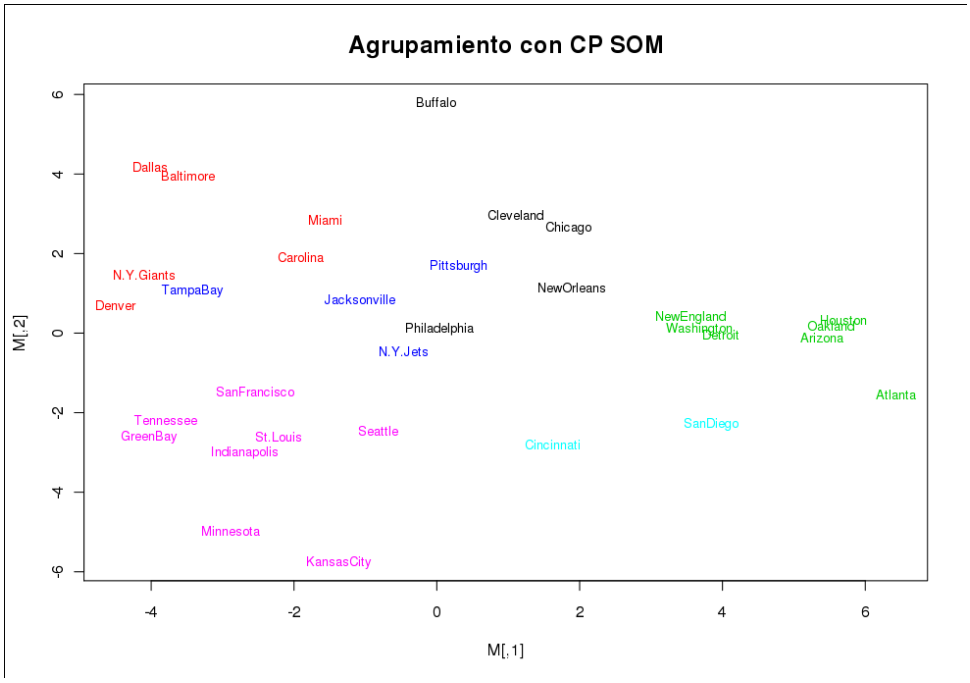

Figura 3.21: Agrupación de equipos utilizando SOM temporada 2003

3.2. Agrupación SOM

En esta sección se presentan los grupos generados utilizando SOM. En las figuras (3.21), (3.22), (3.23), (3.24) y (3.25) se presenta la agrupación por colores de los diferentes equipos de las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente, para poderlos visualizar se utilizaron las 2 primeras componentes principales. En las tablas (3.1), (3.2), (3.3), (3.4) y (3.5) se presenta la agrupación de los diferentes equipos de las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente.

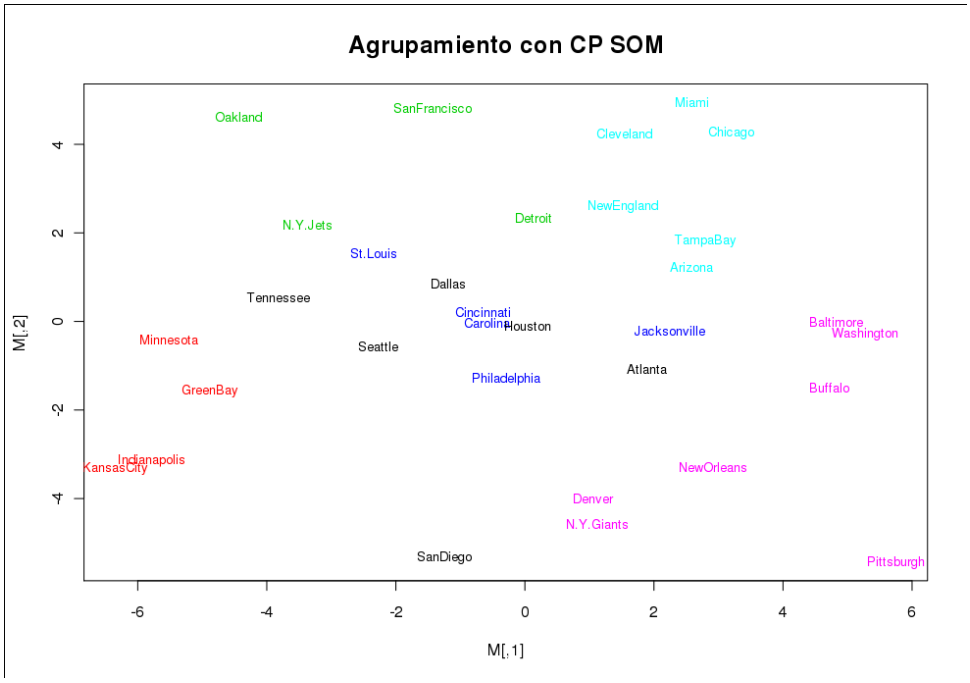

Figura 3.22: Agrupación de equipos utilizando SOM temporada 2004

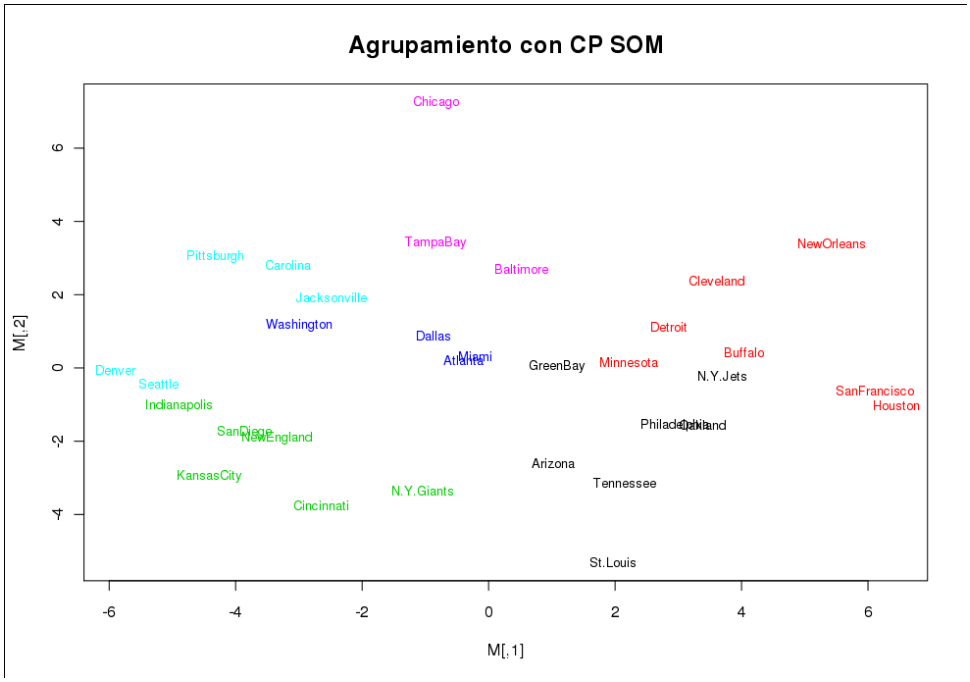

Figura 3.23: Agrupación de equipos utilizando SOM temporada 2005

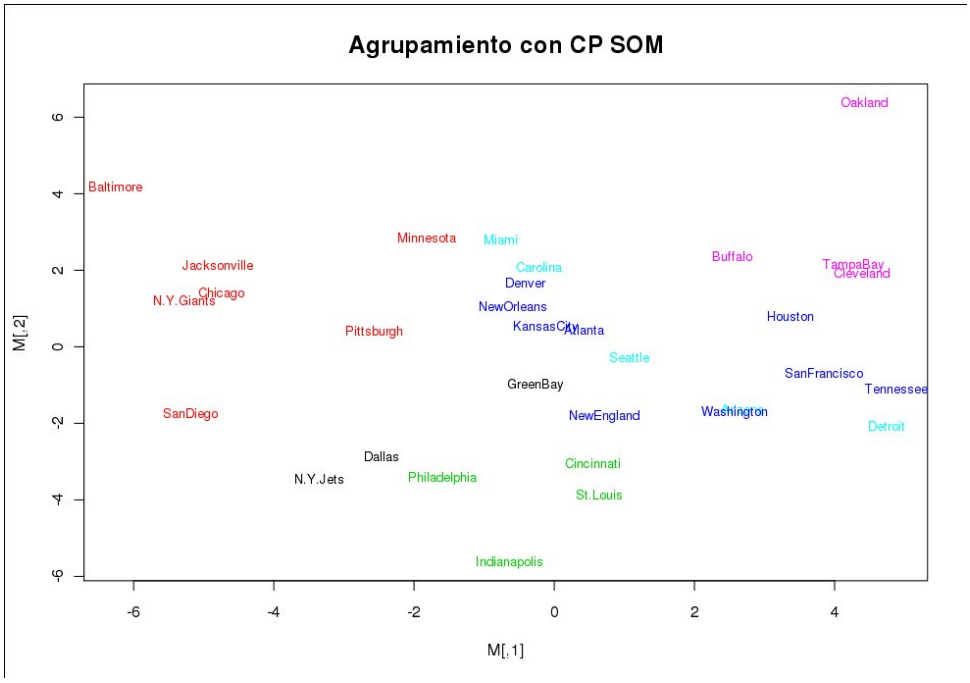

Figura 3.24: Agrupación de equipos utilizando SOM temporada 2006

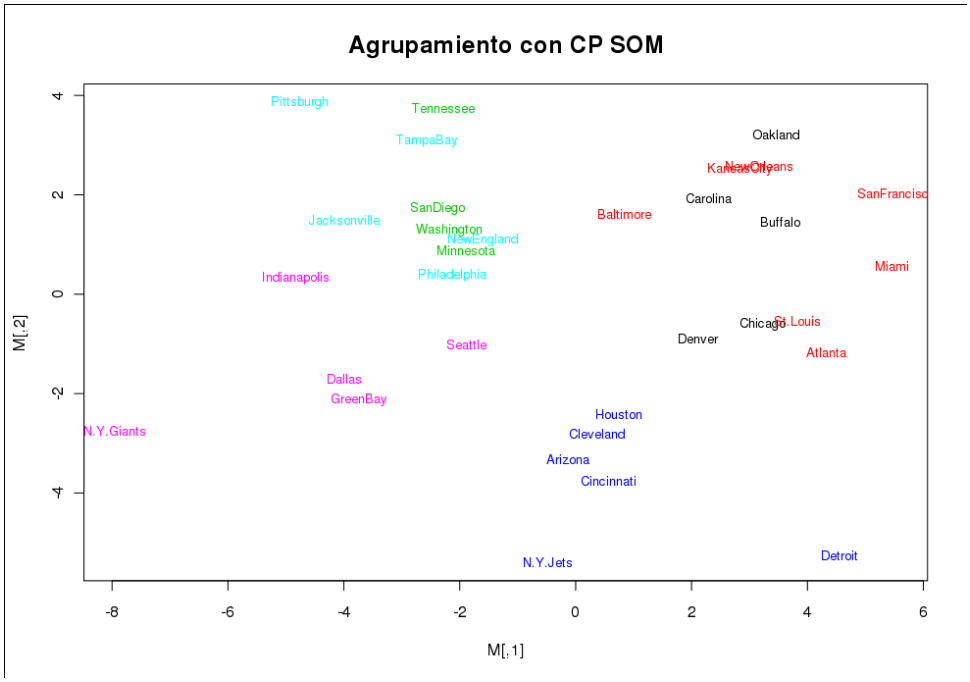

Figura 3.25: Agrupación de equipos utilizando SOM temporada 2007

Cuadro 3.1: Agrupación: equipos SOM temporada 2003

Equipo	Grupo
Dallas	1
Denver	1
Jacksonville	1
TampaBay	1
GreenBay	2
Indianapolis	2
KansasCity	2
Minnesota	2
SanFrancisco	2
St.Louis	2
Tennessee	2
Baltimore	3
Carolina	3
Miami	3
N.Y.Giants	3
Cincinnati	4
N.Y.Jets	4
Philadelphia	4
Seattle	4
Buffalo	5
Chicago	5
Cleveland	5
NewOrleans	5
Pittsburgh	5
Arizona	6
Atlanta	6
Detroit	6
Houston	6
NewEngland	6
Oakland	6
SanDiego	6
Washington	6

Cuadro 3.2: Agrupación: equipos SOM temporada 2004

Equipo	Grupo
Denver	1
N.Y.Giants	1
SanDiego	1
Atlanta	2
Baltimore	2
Buffalo	2
Jacksonville	2
NewOrleans	2
Pittsburgh	2
Washington	2
GreenBay	3
Indianapolis	3
KansasCity	3
Minnesota	3
Carolina	4
Cincinnati	4
Detroit	4
Houston	4
N.Y.Jets	4
Philadelphia	4
Seattle	4
Dallas	5
Oakland	5
SanFrancisco	5
St.Louis	5
Tennessee	5
Arizona	6
Chicago	6
Cleveland	6
Miami	6
NewEngland	6
TampaBay	6

Cuadro 3.3: Agrupación: equipos SOM temporada 2005

Equipo	Grupo
Buffalo	1
Cleveland	1
Detroit	1
GreenBay	1
Miami	1
Minnesota	1
N.Y.Jets	1
NewOrleans	1
Oakland	1
Houston	2
SanFrancisco	2
Baltimore	3
Chicago	3
TampaBay	3
Arizona	4
N.Y.Giants	4
Philadelphia	4
St.Louis	4
Tennessee	4
Atlanta	5
Carolina	5
Dallas	5
Jacksonville	5
Pittsburgh	5
Washington	5
Cincinnati	6
Denver	6
Indianapolis	6
KansasCity	6
NewEngland	6
SanDiego	6
Seattle	6

Cuadro 3.4: Agrupación: equipos SOM temporada 2006

Equipo	Grupo
Carolina	1
GreenBay	1
Miami	1
Minnesota	1
NewOrleans	1
Baltimore	2
Chicago	2
Jacksonville	2
N.Y.Giants	2
Pittsburgh	2
SanDiego	2
Buffalo	3
Cleveland	3
Houston	3
Oakland	3
TampaBay	3
Atlanta	4
Denver	4
KansasCity	4
NewEngland	4
SanFrancisco	4
Seattle	4
Tennessee	4
Washington	4
Arizona	5
Cincinnati	5
Detroit	5
Dallas	6
Indianapolis	6
N.Y.Jets	6
Philadelphia	6
St.Louis	6

Cuadro 3.5: Agrupación: equipos SOM temporada 2007

Equipo	Grupo
Atlanta	1
Buffalo	1
Carolina	1
Chicago	1
SanFrancisco	1
St.Louis	1
Arizona	2
Cincinnati	2
Cleveland	2
Denver	2
Detroit	2
Houston	2
N.Y.Jets	2
KansasCity	3
Miami	3
NewOrleans	3
Oakland	3
Baltimore	4
NewEngland	4
Jacksonville	5
Minnesota	5
Pittsburgh	5
SanDiego	5
TampaBay	5
Tennessee	5
Washington	5
Dallas	6
GreenBay	6
Indianapolis	6
N.Y.Giants	6
Philadelphia	6
Seattle	6

3.3. Agrupación *k*-medias

En esta sección se presentan los grupos generados, utilizando el método de clasificación no supervisada de k-medias, antes de hacerlo se puede observar en la figura (3.26) el número de grupos que se utilizaron para cada temporada de acuerdo con la metodología presentada en la sección (1.6). En las figuras (3.27), (3.28), (3.29), (3.30) y (3.31) se presenta la agrupación por colores de los diferentes equipos de las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente, para poderlos visualizar se utilizaron las 2 primeras componentes principales. En las tablas (3.6), (3.7), (3.8), (3.9) y (3.10) se presenta la agrupación de los diferentes equipos de las temporadas 2003, 2004, 2005, 2006 y 2007 respectivamente.

Hoja1

2003

g	Sg	[g^(2/p)]Sg	Dif	Cg	
1	2	60.89	63.58	0	**0**
2	3	159.33	170.65	-107.07	**0.72**
3	4	291.96	318.38	-147.73	**0.78**
4	5	459.04	507.62	-189.24	**0.77**
5	6	672.86	752.6	-244.97	**0.87**
6	7	914.49	1032.76	-280.16	**0.86**
7	8	1194.46	1360.23	-327.48	**0**

2004

g	Sg	[g^(2/p)]Sg	Dif	Cg	
1	2	67.85	70.86	0	**0**
2	3	169.69	181.75	-110.89	**0.75**
3	4	302.78	330.19	-148.43	**0.82**
4	5	461.81	510.68	-180.49	**0.86**
5	6	643.88	720.18	-209.5	**0.81**
6	7	867.78	980.01	-259.82	**0.83**
7	8	1135.05	1292.58	-312.58	**0**

2005

g	Sg	[g^(2/p)]Sg	Dif	Cg	
1	2	64.46	67.31	0	**0**
2	3	175.07	187.52	-120.2	**0.76**
3	4	317.79	346.55	-159.04	**0.77**
4	5	500.12	553.04	-206.49	**0.93**
5	6	692.83	774.93	-221.89	**0.87**
6	7	912.16	1030.12	-255.19	**0.88**
7	8	1159.12	1320	-289.87	**0**

2006

g	Sg	[g^(2/p)]Sg	Dif	Cg	
1	2	62.17	64.92	0	**0**
2	3	157.43	168.62	-103.69	**0.76**
3	4	280.05	305.4	-136.78	**0.71**
4	5	449.7	497.29	-191.89	**0.84**
5	6	647.79	724.55	-227.26	**0.84**
6	7	880.73	994.63	-270.09	**0.87**
7	8	1146.56	1305.69	-311.05	**0**

2007

g	Sg	[g^(2/p)]Sg	Dif	Cg	
1	2	61.48	64.2	0	**0**
2	3	157.15	168.32	-104.12	**0.76**
3	4	279.36	304.64	-136.32	**0.73**
4	5	444.58	491.63	-186.98	**0.81**
5	6	646	722.55	-230.92	**0.81**
6	7	893.31	1008.84	-286.29	**0.91**
7	8	1162.31	1323.62	-314.79	**0**

Página 1

Figura 3.26: Número de grupos para cada temporada

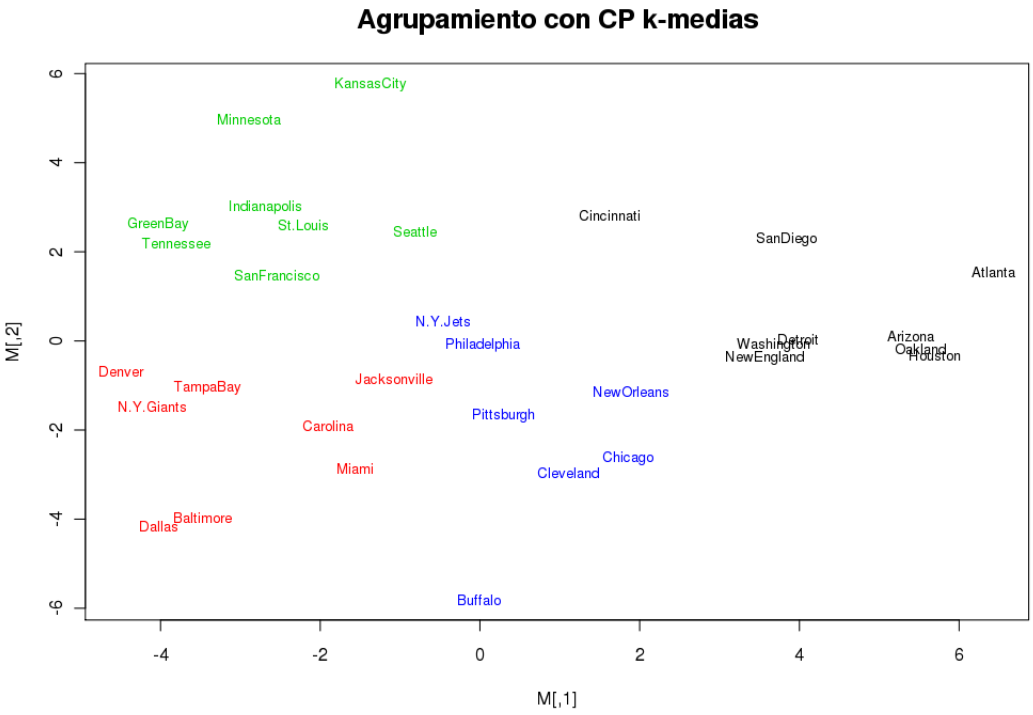

Figura 3.27: Agrupación de equipos utilizando k-medias temporada 2003

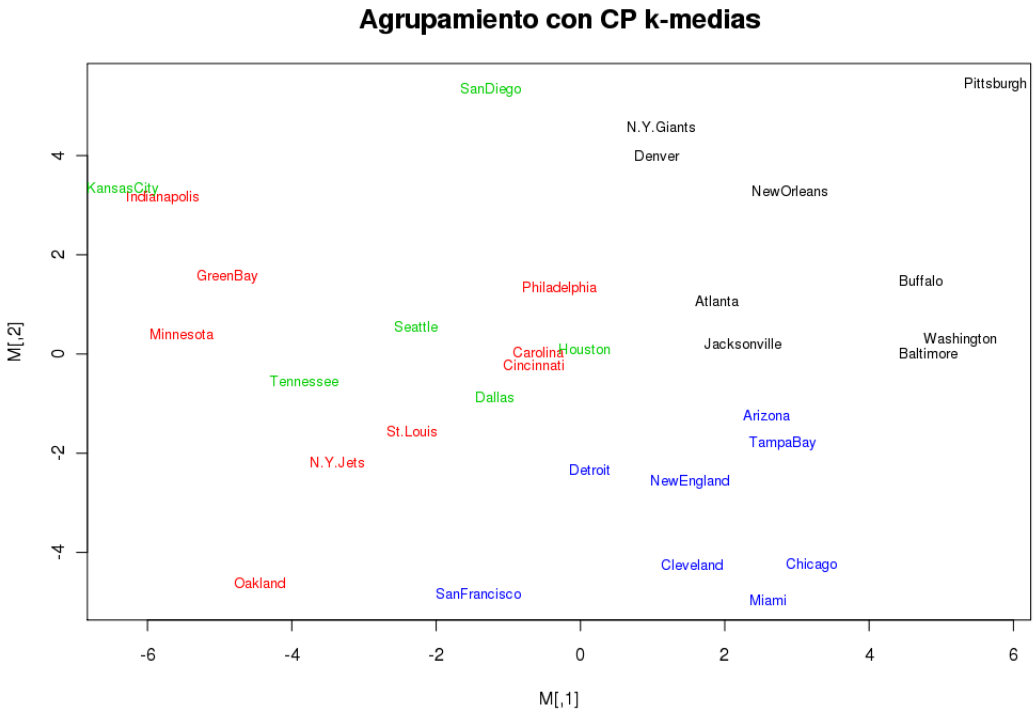

Figura 3.28: Agrupación de equipos utilizando k-medias temporada 2004

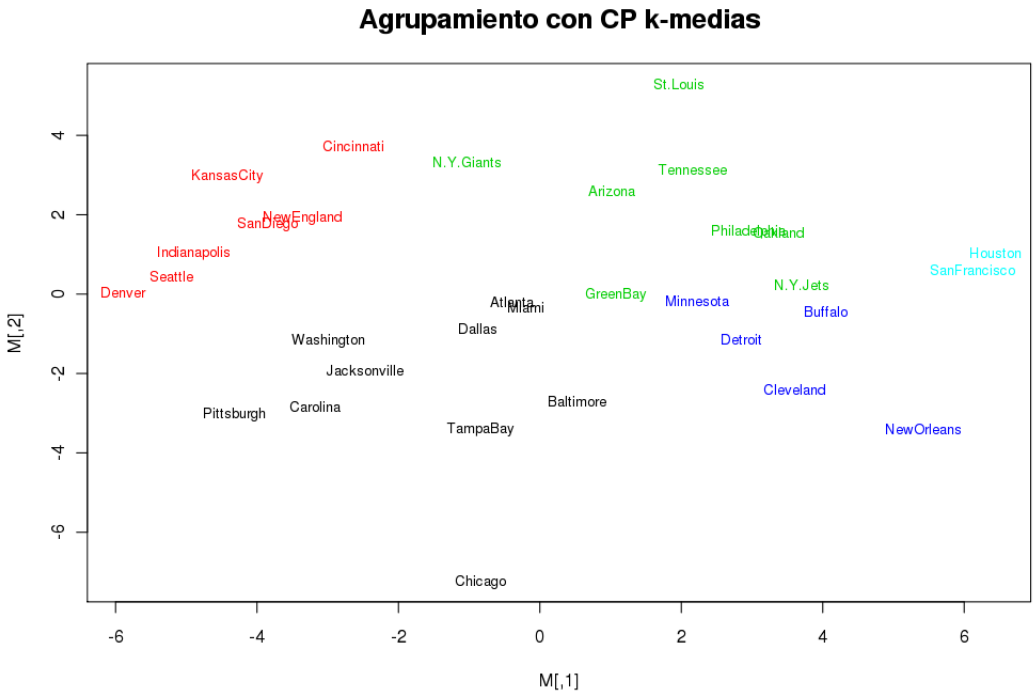

Figura 3.29: Agrupación de equipos utilizando k-medias temporada 2005

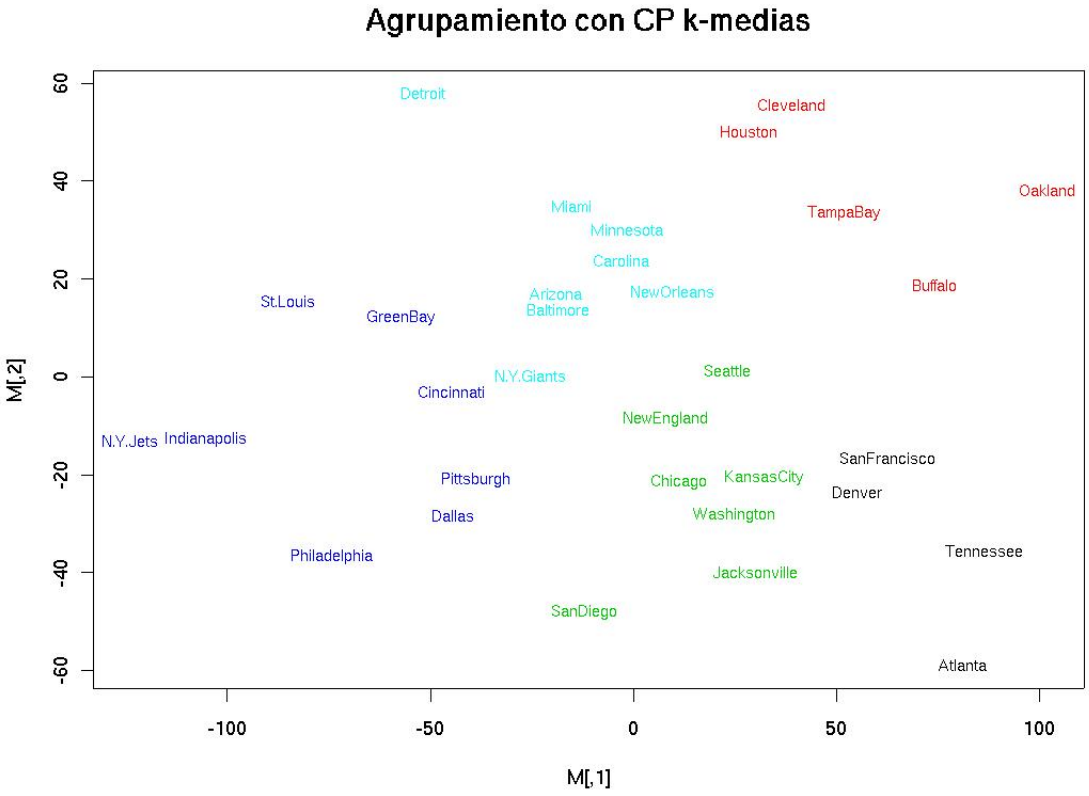

Figura 3.30: Agrupación de equipos utilizando k-medias temporada 2006

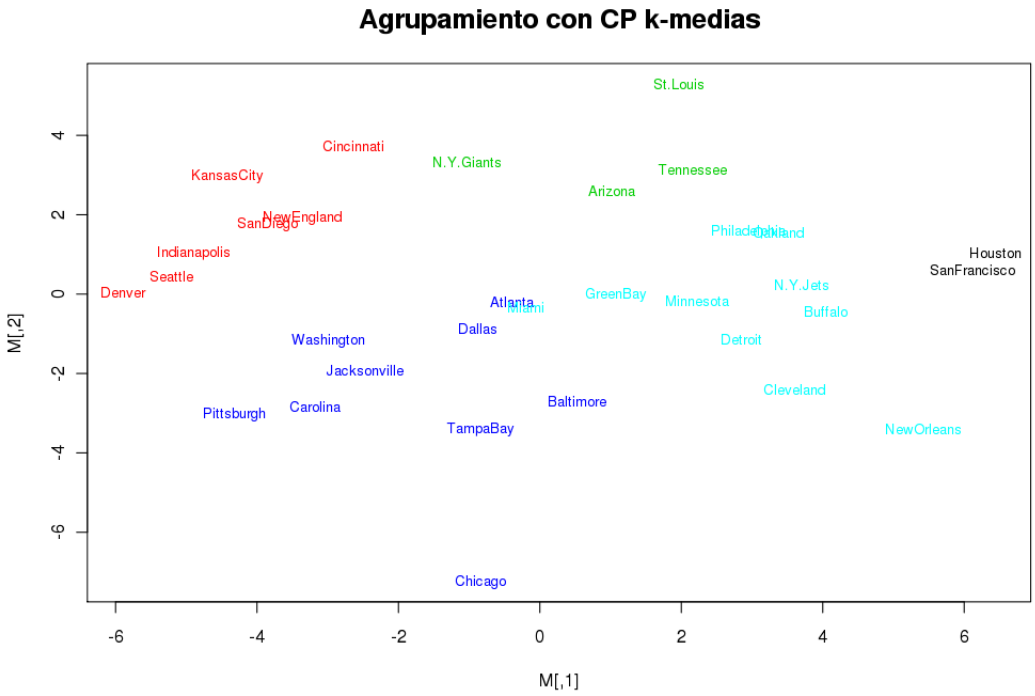

Figura 3.31: Agrupación de equipos utilizando k-medias temporada 2007

Cuadro 3.6: Agrupación: equipos k-medias temporada 2003

Equipo	Grupo
Arizona	1
Atlanta	1
Cincinnati	1
Detroit	1
Houston	1
NewEngland	1
Oakland	1
SanDiego	1
Washington	1
Baltimore	2
Carolina	2
Dallas	2
Denver	2
Jacksonville	2
Miami	2
N.Y.Giants	2
TampaBay	2
GreenBay	3
Indianapolis	3
KansasCity	3
Minnesota	3
SanFrancisco	3
Seattle	3
St.Louis	3
Tennessee	3
Buffalo	4
Chicago	4
Cleveland	4
N.Y.Jets	4
NewOrleans	4
Philadelphia	4
Pittsburgh	4

Cuadro 3.7: Agrupación: equipos k-medias temporada 2004

Equipo	Grupo
Atlanta	1
Baltimore	1
Buffalo	1
Denver	1
Jacksonville	1
N.Y.Giants	1
NewOrleans	1
Pittsburgh	1
Washington	1
Carolina	2
Cincinnati	2
GreenBay	2
Indianapolis	2
Minnesota	2
N.Y.Jets	2
Oakland	2
Philadelphia	2
St.Louis	2
Dallas	3
Houston	3
KansasCity	3
SanDiego	3
Seattle	3
Tennessee	3
Arizona	4
Chicago	4
Cleveland	4
Detroit	4
Miami	4
NewEngland	4
SanFrancisco	4
TampaBay	4

Cuadro 3.8: Agrupación: equipos k-medias temporada 2005

Equipo	Grupo
Atlanta	1
Baltimore	1
Carolina	1
Chicago	1
Dallas	1
Jacksonville	1
Miami	1
Pittsburgh	1
TampaBay	1
Washington	1
Cincinnati	2
Denver	2
Indianapolis	2
KansasCity	2
NewEngland	2
SanDiego	2
Seattle	2
Arizona	3
GreenBay	3
N.Y.Giants	3
N.Y.Jets	3
Oakland	3
Philadelphia	3
St.Louis	3
Tennessee	3
Buffalo	4
Cleveland	4
Detroit	4
Minnesota	4
NewOrleans	4
Houston	5
SanFrancisco	5

Cuadro 3.9: Agrupación: equipos k-medias temporada 2006

Equipo	Grupo
Chicago	1
Jacksonville	1
KansasCity	1
NewEngland	1
SanDiego	1
Seattle	1
Washington	1
Buffalo	2
Denver	2
Oakland	2
Atlanta	2
SanFrancisco	2
Tennessee	2
Indianapolis	3
N.Y.Jets	3
Philadelphia	3
St.Louis	3
Carolina	4
Cleveland	4
Houston	4
Miami	4
Minnesota	4
NewOrleans	4
TampaBay	4
Arizona	5
Baltimore	5
Cincinnati	5
Dallas	5
Detroit	5
GreenBay	5
N.Y.Giants	5
Pittsburgh	5

Cuadro 3.10: Agrupación: equipos k-medias temporada 2007

Equipo	Grupo
Houston	1
SanFrancisco	1
Cincinnati	2
Denver	2
Indianapolis	2
KansasCity	2
NewEngland	2
SanDiego	2
Seattle	2
Arizona	3
N.Y.Giants	3
St.Louis	3
Tennessee	3
Atlanta	4
Baltimore	4
Carolina	4
Chicago	4
Dallas	4
Jacksonville	4
Pittsburgh	4
TampaBay	4
Washington	4
Buffalo	5
Cleveland	5
Detroit	5
GreenBay	5
Miami	5
Minnesota	5
N.Y.Jets	5
NewOrleans	5
Oakland	5
Philadelphia	5

3.4. Predicción SOM vs k-medias

Los resultados obtenidos de las predicciones para los 11 partidos de post-temporada de las temporadas 2003, 2004, 2005, 2006 y 2007, se presentan en las tablas (3.11),(3.12),(3.13),(3.14) y (3.15) respectivamente.

Cuadro 3.11: Pronósticos temporada 2003

Equipos	SOM-Ptos	kmed-Ptos	real-Ptos
DENVER	25	25	10
INDIANAPOLIS	18	24	41
GREEN BAY	37	27	33
SEATTLE	23	22	27
DALLAS	19	20	10
CAROLINA	16	23	29
TENNESSE	26	26	20
BALTIMORE	18	21	17
GREEN BAY	20	26	17
PHILADELPHIA	18	21	20
INDIANAPOLIS	40	35	38
KANSAS CITY	21	24	31
TENNESSE	24	32	14
NEW ENGLAND	31	32	17
CAROLINA	23	21	29
ST. LOUIS	28	26	23
CAROLINA	16	18	14
PHILADELPHIA	25	21	3
INDIANAPOLIS	27	31	14
NEW ENGLAND	32	29	24
CAROLINA	18	14	29
NEW ENGLAND	16	18	32

Cuadro 3.12: Pronósticos temporada 2004

Equipos	SOM-Ptos	kmed-Ptos	real-Ptos
MINNESOTA	28	29	31
GREEN BAY	29	31	17
DENVER	23	23	24
INDIANAPOLIS	24	18	49
NY JETS	27	27	20
SAN DIEGO	25	22	17
ST LOUIS	24	25	27
SEATTLE	21	22	20
INDIANAPOLIS	25	30	3
NEW ENGLAND	26	18	20
MINNESOTA	17	23	14
PHILADELPHIA	25	28	27
ST LOUIS	17	18	17
ATLANTA	19	33	47
NY JETS	10	13	17
PITTSBURG	21	21	20
NEW ENGLAND	19	20	41
PITTSBURG	26	21	27
ATLANTA	16	23	10
PHILADELPHIA	26	19	27
NEW ENGLAND	22	23	24
PHILADELPHIA	16	23	21

Cuadro 3.13: Pronósticos temporada 2005

Equipos	SOM-Ptos	kmed-Ptos	real-Ptos
PITTSBURG	28	23	31
CINCINNATI	21	23	17
CAROLINA	35	22	23
NY GIANTS	21	19	0
JACKSONVILLE	19	18	3
NEW ENGLAND	15	21	28
WASHINGTON	24	23	17
TAMPA BAY	19	20	10
CAROLINA	15	16	29
CHICAGO	16	14	21
PITTSBURG	13	15	21
INDIANAPOLIS	21	23	18
NEW ENGLAND	20	18	13
DENVER	26	30	27
NY JETS	13	11	10
PITTSBURG	5	23	20
CAROLINA	20	23	14
SEATTLE	24	14	34
PITTSBURG	18	20	34
DENVER	24	17	17
SEATTLE	23	21	10
PITTSBURG	9	20	21

Cuadro 3.14: Pronósticos temporada 2006

Equipos	SOM-Ptos	kmed-Ptos	real-Ptos
NY GIANTS	25	21	20
PHILADELPHIA	28	30	23
NY JETS	14	14	16
NEW ENGLAND	20	21	37
DALLAS	24	20	20
SEATTLE	25	22	21
KANSAS CITY	11	28	8
INDIANAPOLIS	13	21	23
NEW ENGLAND	16	20	24
SAN DIEGO	19	20	21
SEATTLE	12	14	24
CHICAGO	31	28	27
PHILADELPHIA	24	24	24
NEW ORLEANS	23	24	27
INDIANAPOLIS	15	12	15
BALTIMORE	25	10	6
NEW ENGLAND	25	22	34
INDIANAPOLIS	25	23	38
NEW ORLEANS	18	16	14
CHICAGO	24	19	39
INDIANAPOLIS	25	19	29
CHICAGO	24	26	17

Cuadro 3.15: Pronósticos temporada 2007

Equipos	SOM-Ptos	kmed-Ptos	real-Ptos
WASHINGTON	13	35	14
SEATTLE	6	14	35
JACKSONVILLE	25	21	31
PITTSBURG	20	23	29
NY GIANTS	17	16	24
TAMPA BAY	10	21	14
TENNESSEE	9	17	6
SAN DIEGO	22	23	16
SEATTLE	25	24	20
GREEN BAY	19	26	42
JACKSONVILLE	8	18	20
NEW ENGLAND	17	29	31
SAN DIEGO	21	22	28
INDIANAPOLIS	25	20	24
NY GIANTS	21	22	21
DALLAS	25	30	17
NY GIANTS	19	17	23
GREEN BAY	29	27	20
SAN DIEGO	17	20	12
NEW ENGLAND	39	25	21
NY GIANTS	24	35	17
NEW ENGLAND	30	38	14

3.5. Análisis de resultados y verificación del modelo.

En esta sección se analizan los resultados obtenidos en las secciones anteriores y se presenta el comportamiento del modelo. Para comenzar se muestra en la tabla (3.17) el resumen de los resultados de la sección anterior, en él se observa que para éste ejercicio la agrupación hecha por SOM fue ligeramente mejor que la de k-medias, ambos tuvieron alrededor de 40 % en la media de efectividad[1] al momento de predecir, aunque si consideramos la desviación estándar entonces son practicamente iguales, en ambos casos el porcentaje no debe ser considerado bajo, dado que los posibles resultados son tres, y esto implicaria que adivinar el resultado de forma aleatoria daría una probabilidad de éxito del 33 %, con el procedimiento presentado se aumenta esa probabilidad en aproximadamente 10 puntos. Por otro lado una pregunta que surge de manera natural es ¿qué sucede si se combinan ambos modelos tanto SOM como k-medias al momento de predecir?, ¿qué reultados se obtendrían?, éstos se presentan en la tabla (3.18), de acuerdo con ella, la media en el procentaje de efectividad es del 35.32 %, parece bajo, pero si se quita la temporada 2005 que claramente se puede ver como un outlier[2], entonces se puede ver que la media de efectividad es del 44.15 %, con una desviación estándar del 13.78 %, lo que representa un incremento en la efectividad de aproximadamente 5 puntos.

A continuación la verificación del modelo permitirá saber cual es la probabilidad que tiene el modelo de predecir un resultado con éxito, se hace para SOM, k-medias y de manera conjunta. Entonces si se consideran todos los encuentros realizados durante la predicción se tienen un total de 11 juegos por 5 temporadas, es decir un total de 55 encuentros, por lo tanto con ayuda del estimador de máxima verosimilitud, de acuerdo con (2.1) y la información de los cuadros (3.17) y (3.18) el estimador máximo verosimil para estimar la probabilidad de éxito del modelo utilizando SOM es:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} = \frac{23}{55} = 0{,}42 \tag{3.1}$$

Para k-medias es:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} = \frac{21}{55} = 0{,}38 \tag{3.2}$$

Para ambos de manera conjunta:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} = \frac{14}{39} = 0{,}36 \tag{3.3}$$

Esto indica que la probabilidad del modelo de predecir un resultado correcto es relativamente buena pues en cada predicción se tiene una probabilidad de éxito de alrededor del 40 %, en la tabla (3.16) se presenta el resumen de esta verificación.

[1] Considerando ésta como el cociente de resultados con aciertos entre el total de resultados.
[2] Punto con datos anormales.

Cuadro 3.16: Evaluación del modelo

modelo	n	p	(1-p)	media	desviación-estándar
SOM	55	0.42	0.58	23.1	13.4
k-medias	55	0.38	0.62	20.9	12.96
conjunta	39	0.36	0.64	14.04	8.99

Cuadro 3.17: Porcentaje de aciertos temporadas 2003-2007

Temporada	SOM	k-medias
2003	$4/11 = 36\,\%$	$7/11 = 64\,\%$
2004	$4/11 = 36\,\%$	$3/11 = 27\,\%$
2005	$2/11 = 18\,\%$	$2/11 = 18\,\%$
2006	$5/11 = 45\,\%$	$5/11 = 45\,\%$
2007	$8/11 = 73\,\%$	$4/11 = 36\,\%$
Media	$41.6\,\%$	$38\,\%$
Desviación Estándar	$20.11\,\%$	$17.68\,\%$

Cuadro 3.18: Porcentaje de aciertos comunes temporadas 2003-2007

Temporada	Resultados Conjuntos	Efectividad
2003	$72.73\,\%$	$4/8 = 50\,\%$
2004	$72.73\,\%$	$2/8 = 25\,\%$
2005	$63.64\,\%$	$0/7 = 0\,\%$
2006	$81.82\,\%$	$4/9 = 44.44\,\%$
2007	$63.64\,\%$	$4/7 = 57.14\,\%$
Media	$35.32\,\%$	
Desviación Estándar	$23.07\,\%$	

Ahora desde el punto de vista de un analista de futbol americano saltan algunas preguntas que deben contestarse, la principal por supuesto es ¿Qué pasó en la temporada 2005?, es posible que haya tenido un comportamiento fuera de los normal, así que se procede a revisar con cuidado cada uno de los resultados, los cuales se presentan en el anexo (C), así buscando explicaciones se mencionan algunos ejemplos de lo que pasó:

- En el partido 1 Pittsburg vs Cincinnati.- en este partido el mariscal de Cincinnati se lesionó en el primer cuarto, lo que cambió radicalmente el partido, situción difícil de preveer.
- En el partido Jacksonville vs NewEngland.- es una sorpresa relativa, se esperaba que ganará NewEngland sin embargo no se esperaba que ganara con tanta facilidad.
- Después vienen una serie de partidos que se esperaban bastante cerrados y que así fueron, sin embargo por alguna razón al final en varios de ellos se abrió el marcador un poco más de lo esperado, en general se considera un efecto psicológico, situación que se da entre dos equipos fuertes en el que por alguna razón uno se empieza a despegar en el marcador y al otro simplemente todo se le complica y muchas veces termina en paliza, no es muy común sin embargo esta temporada en particular estuvo llena de estas situaciones.
- También ocurrió algo curioso, equipos que se consideraban muy fuertes y de acuerdo a sus estadísticas literalmente invensibles comensaron a bajar su nivel de juego hacia el final de la temporada, como es el caso de New England, y otros que no parecían tan fuertes comensaron a jugar muy bien hacia el final de la temporada, como es el caso de Pittsburg que alcanzó a calificar como equipos comodín y que sin embargo llegó a ser campeón, esta situación sólo se ha dado en tres ocasiones a lo largo de la historia.

A la vista de este análisis es prudente considerar a la temporada 2005 como un outlier, ya que la mayoría de las situaciones que aquí se dieron no son muy comunes, sin embargo sería prudente considerar arreglos al modelo de tal forma que puediera considerar la tendencia en el nivel de juego de los equipos al momento de llegar a la post-temporada.

Conclusiones

A lo largo del presente trabajo se ha propuesto un modelo que permite predecir el resultado de partidos de futbol americano profesional, tratando de aumetar el número de resultados correctos bajo la idea de que para ver el poderío de dos equipos que se enfrentarán en la post-temporada, es mejor medir como les fue contra equipos que poseen las características de su rival y no necesariamente como les fue en general contra todos sus rivales, esto hace que se evalúen sus destrezas y debilidades contra rivales muy parecidos al que enfrentarán en la post-temporada. El modelo resultó relativamente bueno pues de acuerdo a los resultados presentados a lo largo del capítulo (2) se puede concluir que:

1. Viendo los resultados de la estadística descriptiva el modelo tiene una media de efectividad[3] buena, pues para SOM es del 41.6 %, para k-medias del 38 % y para ambos del 35.32 %, con desviaciones estándar de 20.11 %, 17.68 % y 23.07 % respectivamente, (nota: si no se considera el outlier (temporada 2005) entonces la media de efectividad del modelo utilizando SOM y k-medias mejora considerablemente, ésta es del 44.15 %, con una desviación estándar del 13.78 %).

2. La probabilidad de éxito que tiene el modelo para predecir un resultado utilizando SOM es del 42 %.

3. La probabilidad de éxito que tiene el modelo para predecir un resultado utilizando k-medias es del 38 %.

4. La probabilidad de éxito que tiene el modelo para predecir un resultado utilizando SOM y k-medias en conjunto es del 36 %.

5. Se tiene que aceptar que las desviaciones estándar son muy grandes, al menos más de lo deseado.

6. En general el módelo es promisorio y puede ser la base para utilizar nuevas herramientas sobre la misma idea que ya funciona aceptablemente bien.

7. Con la construcción del modelo se mostró la utilidad de las redes neuronales en combinación con el análisis multivariado en un área diferente a aquellas en las que comúnmente son utilizadas.

[3] Considerando ésta como el cociente de resultados con aciertos entre el total de resultados.

8. Se logró proponer un método para evaluar el modelo.

De manera general se puede concluir que se cuenta con una primera versión de un modelo que sirve para predecir resultados de la NFL, que éste es perfectible en muchas áreas y que sin embargo, representa una base para hacer nuevas aportaciones al mismo.

Recomendaciones

Al momento de generar el modelo se quizó en principio manternerlo lo más simple posible, y al mismo tiempo lo más general, éste sirve como base y punto de partida para agregar nuevas harramientas y consideraciones al análisis, a continuación se mencionan algunas recomendaciones para mejorar el modelo:

1. Las variables en general no se comportan con una distribución normal, sin embargo están muy cerca de serlo, así que utilizar análisis de factores en vez de componentes principales no es una idea tan descabellada y ésta podría arrojar mejores resultados, pues aunque en general éste se considera una extensión de las componentes principales da información adicional como saber si los datos son consistentes con una estructura.

2. Una de las deficiencias del modelo es que no considera tendencias en las estadísticas, es decir después de agrupar a los equipos al momento de hacer las predicciones en vez de sacar un promedio de los puntos recibidos y anotados para obtener predicción, podríamos hacer un análisis de regresión en los puntos recibidos y anotados, esto permitiría tener una tendencia en las predicciones y así saber además si un equipo va a la alza o a la baja en su nivel de juego.

3. Para evaluar el modelo podríamos además considerar otro tipo de estimadores como el bayesiano, que daría una forma fácil de obtener un intervalo para el estimador.

4. Se podrían considerar todas las temporadas para tener más información al momento de evaluar el modelo y tratar de ver si los resultados se apegan a alguna distribución en particular.

5. Se podrían considerar comparaciones del presente modelo contra el resultado de otras herramientas como análisis de regresión, también podríamos comparar la efectividad del modelo contra otro que no considere agrupación de los equipos.

Apéndice A

Matriz de covarianzas intragrupos (*within-group*)

Para obtener esta matriz se hará con un ejercicio:

x = características (variables independientes) de los datos, donde cada renglón denotado por k representa un objeto, se tiene una columna para cada característica.

y = grupo del objeto (variable dependiente) de los datos, cada renglón representa un objeto.

$$x = \begin{pmatrix} 2{,}95 & 6{,}63 \\ 2{,}53 & 7{,}79 \\ 3{,}57 & 5{,}65 \\ 3{,}16 & 5{,}47 \\ 2{,}58 & 4{,}46 \\ 2{,}16 & 6{,}22 \\ 3{,}27 & 3{,}52 \end{pmatrix}$$

$$y = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{pmatrix}$$

x_k = renglón k por ejemplo $x_3 =$

$$\begin{pmatrix} 3{,}57 & 5{,}65 \end{pmatrix}$$

g = número de grupos en y en nuestro ejemplo 2.

Ahora separamos x en diferentes grupos basados en el número de grupo dado por y:

$x_1 =$

$$\begin{pmatrix} 2{,}95 & 6{,}63 \\ 2{,}53 & 7{,}79 \\ 3{,}57 & 5{,}65 \\ 3{,}16 & 5{,}47 \end{pmatrix}$$

$x_2 =$

$$\begin{pmatrix} 2{,}58 & 4{,}46 \\ 2{,}16 & 6{,}22 \\ 3{,}27 & 3{,}52 \end{pmatrix}$$

$\mu_i =$ es la media de las características en el grupo i para cada x_i

$\mu_1 =$

$$\begin{pmatrix} 3{,}05 & 6{,}38 \end{pmatrix}$$

$\mu_2 =$

$$\begin{pmatrix} 2{,}67 & 4{,}73 \end{pmatrix}$$

$\mu =$ es la media global del vector $\mu =$

$$\begin{pmatrix} 2{,}88 & 5{,}67 \end{pmatrix}$$

$x_i^0 =$ media corregida, es decir $x_i - \mu$

$x_1^0 =$

$$\begin{pmatrix} 0{,}060 & 0{,}951 \\ -0{,}357 & 2{,}109 \\ 0{,}679 & -0{,}025 \\ 0{,}269 & -0{,}209 \end{pmatrix}$$

$x_2^0 =$

$$\begin{pmatrix} -0{,}305 & -1{,}218 \\ -0{,}732 & 0{,}547 \\ 0{,}386 & -2{,}155 \end{pmatrix}$$

$C_1 = \frac{(x_i^0)^T x_i^0}{n_i} =$ matriz de covarianza del grupo i

$C_1 =$

$$\begin{pmatrix} 0{,}166 & -0{,}192 \\ -0{,}192 & 1{,}349 \end{pmatrix}$$

$C_2 =$

$$\begin{pmatrix} 0{,}259 & -0{,}286 \\ -0{,}286 & 2{,}142 \end{pmatrix}$$

$C(r,s) = \frac{1}{n}\sum_{i=1}^{g} n_i c_i(r,s) =$ matriz de covarianza intra-grupo, ésta es calculada para cada entrada (r,s) de la matriz $\frac{3}{7}0{,}166 + \frac{4}{7}0{,}259 = 0{,}206$, $\frac{3}{7}(-0{,}192) + \frac{4}{7}(-0{,}286) = -0{,}233$, $\frac{3}{7}0{,}1349 + \frac{4}{7}2{,}142 = 1{,}689$, por lo tanto:

$C =$

$$\begin{pmatrix} 0{,}206 & -0{,}233 \\ -0{,}233 & 1{,}689 \end{pmatrix}$$

Apéndice B

Algunos conceptos básicos

Definición distribución uniforme.- Se dice que una variable aleatoria X tiene distribución uniforme en el intervalo (a, b) si su función de densidad es: $f_X(x) = 1/(b - a) \quad si \quad x \in (a, b) \quad ; 0 \quad e.o.c.$, ver [Áng05] capítulo VIII.

Definición eigenvalor.- Se dice que una matriz cuadrada A tiene eigenvalor λ con su correspondiente eigenvector $X \neq 0$, si $Ax = \lambda x$, ver [Ric07].

Definición estimador consistente.- Decimos que $\hat{\theta}$ es un estimador consistente con el parámetro θ si:

$$\forall\, \epsilon > 0, \qquad \lim_{n\to\infty} \mathcal{P}[\mid \hat{\theta} - \theta \mid > \epsilon] = 0,$$

o lo que es equivalente:

$$\forall\, \epsilon > 0, \qquad \lim_{n\to\infty} \mathcal{P}[\mid \hat{\theta} - \theta \mid < \epsilon] = 1.$$

Este tipo de propiedades definidas cuando el número de observaciones n, tiende a infinito, es lo que se denomina propiedades asintóticas

Definición estimador insesgado.- Se dice que un estimador $\hat{\theta}$ de un parmetro θ es insesgado si:

$$\mathbf{E}\left[\hat{\theta}\right] = \theta.$$

Definición estadístico sufuciente.- Diremos que $\hat{\theta} \equiv \hat{\theta}(X_1, \ldots, X_n)$ es un estadístico suficiente del parámetro $\hat{\theta}$ si

$$\mathcal{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_{n|\hat{\theta}=a}] \qquad \text{no depende de } \theta$$

para todo posible valor de θ. Esta definición así enunciada tal vez resulte un poco oscura, pero lo que expresa es que un estadístico es suficiente, si agota toda la información existente en la muestra que sirva para estimar el parámetro ver [Ale78].

Definición operador lineal diagonalizable.- Se dice que un operador lineal T sobre un espacio vectorial dimensionalmente finito V, es diagonalizable si existe una base β para V tal que $[T]_\beta$ sea una matriz diagonal, ver [Ste82].

Definición ley binomial.- Se dice que una v.a. X sigue una ley binomial de parámetros n y p, $X \rightsquigarrow \mathbf{B}(n,p)$, si es la suma de n v.a. independientes Bernouilli con el mismo parámetro p; es decir si realizamos n ensayos o repeticiones independientes de Bernoulli, (en idénticas condiciones), y siempre centrados en el suceso A, la variable X que cuenta el número de veces que ha tenido lugar el suceso A define el modelo binomial $B(n,p)$, su función de densidad está definida por:

$$p(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 1, 2, ..., n$$

donde:

la media es np y la varianza es npq.

Definición traza de una matriz.- Sea M de $n \times n$ se denota como $traza(M)$ a la traza de una matriz, que es la suma de los valores de M ubicados en la diagonal, es decir, $traza(M) = M_{11} + M_{22} + ... + M_{nn}$ ver [Ste82] capítulo 1.

Definición variables aleatorias independientes.- Se dice que n variables aleatorias $X_1, X_2, ..., X_n$, son independientes si para cualquier colección de subconjuntos borelianos de números reales $A_1, A_2, ..., A_n$, se tiene que $P[X_1 \in A_1, ..., X_n \in A_n] = P[X_1 \in A_1]...P[X_n \in A_n]$ ver [Áng05] capítulo VI.

Apéndice C

Resultados post-temporada 2005

CP-SOM		25		7	CP-Akmed			
	Pittsburgh		Cincinnati			Pittsburgh		Cincinnati	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
27	13	13	27		20	23	24	7
20	19	38	31		24	22	20	23
34	21				27	13	13	27
13	16				26	7	21	9
31	38				31	38	42	29
21	9						38	31
41	0							
35	21							
28	17	25.5	29		25.6	20.6	26	21
9	11	18	3		4	12	11	10

	Pittsburgh	Cincinnati
MARCADOR	28	21
1 Desviación	10	

	Pittsburgh	Cincinnati
MARCADOR	23	23
Desviación	9	

Pittsburgh	Cincinnati
31	17

CP-SOM		5		19	CP-Akmed			
	Carolina		N.Y.Giants			Carolina		N.Y.Giants	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
20	23	23	45		32	29	13	16
24	27				24	20	36	0
24	20				30	3	17	10
38	13						20	35
34	14							
10	20							
27	10							
25	18	23	45		28.67	17.33	22	15
9	6	N/A	N/A		4	13	10	15

	Carolina	N.Y.Giants
MARCADOR	35	21
1 Desviación	N/A	

	Carolina	N.Y.Giants
MARCADOR	22	19
Desviación	11	

Carolina	N.Y.Giants
23	0

CP-SOM		15	21		CP-Akmed			
Jacksonville		NewEngland			Jacksonville		NewEngland	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
10	9	20	28		26	14	17	27
					3	10	23	20
					7	20	31	28
					23	20	23	16
					18	26	28	0
							26	28
10	9	20	28		15.4	18	25	20
N/A	N/A	N/A	N/A		10	6	5	11

	Jacksonville	NewEngland
MARCADOR	19	15
0 Desviación	N/A	

	Jacksonville	NewEngland
MARCADOR	18	21
Desviación	8	

Jacksonville	NewEngland
3	28

CP-SOM		32	30		CP-Akmed			
Washington		TampaBay			Washington		TampaBay	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
20	17	10	15		9	7	27	13
21	28	36	35		14	13	14	34
0	36	0	28		35	36	36	35
35	36				35	7	30	27
24	9						10	13
17	13						20	10
35	20						27	24
22	23	15.33	26		23.25	15.75	23	22
12	11	19	10		14	14	9	10

	Washington	TampaBay
MARCADOR	24	19
0 Desviación	13	

	Washington	TampaBay
MARCADOR	23	20
Desviación	12	

Washington	TampaBay
17	10

CP-SOM		5		6		CP-Akmed				
	Carolina		Chicago				Carolina		Chicago	
	Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos			Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
	21	20	13	3			24	27	7	9
	3	13	19	7			34	14	10	6
	13	9	24	17			3	13	13	3
	24	6					24	6	13	10
	20	24					10	20	9	21
	44	11					20	24	16	3
							44	11		
	21	14	18.67	9			22.71	16.43	11	9
	14	7	6	7			14	8	3	7
		Carolina	Chicago					Carolina	Chicago	
	MARCADOR	15	16				MARCADOR	16	14	
0	Desviación	8					Desviación	8		

Carolina	Chicago
29	21

CP-SOM		25		14		CP-Akmed				
	Pittsburgh		Indianapolis				Pittsburgh		Indianapolis	
	Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos			Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
	27	7	26	7			20	23	24	7
	24	22					24	22	10	3
	20	10					27	13	26	7
	7	27					7	26	26	18
							31	38		
	20	17	26	7			21.8	24.4	22	9
	9	10	N/A	N/A			9	9	8	6
		Pittsburgh	Indianapolis					Pittsburgh	Indianapolis	
	MARCADOR	13	21				MARCADOR	15	23	
0	Desviación	N/A					Desviación	8		

Pittsburgh	Indianapolis
21	18

CP-SOM		21		10
NewEngland			Denver	
Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos
20	28		21	19
			28	20
20	28		24.5	19.5
N/A	N/A		5	1

	NewEngland	Denver
MARCADOR	20	26
Desviación	N/A	

CP-Akmed			
NewEngland		Denver	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
17	41	20	17
20	28	30	10
21	40	28	20
16	26	27	31
		23	7
18.5	33.75	26	17
2	8	4	9

	NewEngland	Denver
MARCADOR	18	30
Desviación	6	

0

NewEngland	Denver
13	27

CP-SOM		20		25
N.Y.Jets			Pittsburgh	
Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos
26	10		0	0
26	10		0	0
N/A	N/A		N/A	N/A

	N.Y.Jets	Pittsburgh
MARCADOR	13	5
Desviación	N/A	

CP-Akmed		32		28
N.Y.Jets			Pittsburgh	
Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos
17	7		34	7
20	26		20	10
3	13			
14	12			
14	27			
3	30			
20	24			
13	19.86		27	9
7	9		10	2

	N.Y.Jets	Pittsburgh
MARCADOR	11	23
Desviación	7	

0

N.Y.Jets	Pittsburgh
10	20

CP-SOM		5		28
	Carolina		Seattle	
Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos
20	23		42	10
24	27		28	13
24	20		17	23
38	13			
34	14			
10	20			
27	10			
25	18		29	15.33
9	6		13	7

	Carolina	Seattle
MARCADOR	20	24
Desviación	9	

CP-Akmed			
Carolina		Seattle	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
27	10	14	26
		28	18
		17	20
		13	10
27	10	18	19
N/A	N/A	7	7

	Carolina	Seattle
MARCADOR	23	14
Desviación	N/A	

0

Carolina	Seattle
14	34

CP-SOM		25		10
	Pittsburgh		Denver	
Ptos. Anotados	Ptos. Recibidos		Ptos. Anotados	Ptos. Recibidos
34	7		49	21
17	23		31	17
			27	0
			22	3
26	15		32.25	10.25
12	11		12	10

	Pittsburgh	Denver
MARCADOR	18	24
Desviación	11	

CP-Akmed			
Pittsburgh		Denver	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
20	23	10	34
24	22	20	7
27	13	21	19
8	26	24	21
31	3	12	10
22	17.4	17	18
9	9	6	11

	Pittsburgh	Denver
MARCADOR	20	17
Desviación	9	

0

Pittsburgh	Denver
34	17

CP-SOM		28		25
	Seattle		Pittsburgh	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos	
42	0	18	3	
42	0	18	3	
N/A	N/A	N/A	N/A	

	Seattle	Pittsburgh
MARCADOR	23	9
Desviación	N/A	

CP-Akmed			
Seattle		Pittsburgh	
Ptos. Anotados	Ptos. Recibidos	Ptos. Anotados	Ptos. Recibidos
14	26	20	23
28	18	24	22
17	20	27	13
13	10	7	26
		31	38
18	18.5	22	24
7	7	9	9

	Seattle	Pittsburgh
MARCADOR	21	20
Desviación	8	

0

Seattle	Pittsburgh
10	21

SOM	2	18%

Kmedias	2	18%

Apéndice D

Script para la generación de la base de datos

```
CREATE SEQUENCE public.equipos_idequipo_seq;

CREATE TABLE public.equipos (
                idequipo INTEGER NOT NULL DEFAULT nextval
                ('public.equipos_idequipo_seq'),
                equipo VARCHAR(10) NOT NULL,
                CONSTRAINT equipos_pk PRIMARY KEY (idequipo)
);
ALTER SEQUENCE public.equipos_idequipo_seq OWNED BY public.equipos.idequipo;
CREATE SEQUENCE public.gposvar_idgposvar_seq;

CREATE TABLE public.gposvar (
                idgposvar INTEGER NOT NULL DEFAULT nextval
                ('public.gposvar_idgposvar_seq'),
                gposvar VARCHAR(10) NOT NULL,
                temporada VARCHAR(10) NOT NULL,
                CONSTRAINT gposvar_pk PRIMARY KEY (idgposvar)
);
ALTER SEQUENCE public.gposvar_idgposvar_seq OWNED BY public.gposvar.idgposvar;
CREATE SEQUENCE public.gposeq_idgposeq_seq;

CREATE TABLE public.gposeq (
                idgposeq INTEGER NOT NULL DEFAULT nextval
                ('public.gposeq_idgposeq_seq'),
                idgposvar INTEGER NOT NULL,
                gposeq VARCHAR(10) NOT NULL,
```

```
                CONSTRAINT gposeq_pk PRIMARY KEY (idgposeq, idgposvar)
);
ALTER SEQUENCE public.gposeq_idgposeq_seq OWNED BY public.gposeq.idgposeq;
CREATE SEQUENCE public.asgposeq_idasgposeq_seq;

CREATE TABLE public.asgposeq (
                idasgposeq INTEGER NOT NULL DEFAULT nextval
                ('public.asgposeq_idasgposeq_seq'),
                idgposeq INTEGER NOT NULL,
                idequipo INTEGER NOT NULL,
                CONSTRAINT asgposeq_pk PRIMARY KEY
                (idasgposeq, idgposeq, idequipo)
);
ALTER SEQUENCE public.asgposeq_idasgposeq_seq OWNED BY
public.asgposeq.idasgposeq;

CREATE SEQUENCE public.variables_idvariable_seq;

CREATE TABLE public.variables (
                idvariable INTEGER NOT NULL DEFAULT nextval
                ('public.variables_idvariable_seq'),
                variable VARCHAR(10) NOT NULL,
                CONSTRAINT variables_pk PRIMARY KEY (idvariable)
);
ALTER SEQUENCE public.variables_idvariable_seq
OWNED BY public.variables.idvariable;

CREATE SEQUENCE public.asgposvar_idasgposvar_seq;

CREATE TABLE public.asgposvar (
                idasgposvar INTEGER NOT NULL DEFAULT nextval
                ('public.asgposvar_idasgposvar_seq'),
                idvariable INTEGER NOT NULL,
                idgposvar INTEGER NOT NULL,
                CONSTRAINT asgposvar_pk PRIMARY KEY
                (idasgposvar, idvariable, idgposvar)
);
ALTER SEQUENCE public.asgposvar_idasgposvar_seq OWNED BY
public.asgposvar.idasgposvar;

CREATE SEQUENCE public.resultados_idresultado_seq;
```

```
CREATE TABLE public.resultados (
                idresultado INTEGER NOT NULL DEFAULT nextval
                ('public.resultados_idresultado_seq'),
                idequipo INTEGER NOT NULL,
                rival INTEGER NOT NULL,
                ptosanotados INTEGER NOT NULL,
                ptosrecibidos INTEGER NOT NULL,
                semana INTEGER NOT NULL,
                temporada INTEGER NOT NULL,
                resultado INTEGER NOT NULL,
                CONSTRAINT resultados_pk PRIMARY KEY
                (idresultado, idequipo)
);

ALTER SEQUENCE public.resultados_idresultado_seq
OWNED BY public.resultados.idresultado;


ALTER TABLE public.resultados ADD CONSTRAINT equipos_resultados_fk
FOREIGN KEY (idequipo)
REFERENCES public.equipos (idequipo)
NOT DEFERRABLE;


ALTER TABLE public.asgposeq ADD CONSTRAINT equipos_asgposeq_fk
FOREIGN KEY (idequipo)
REFERENCES public.equipos (idequipo)
NOT DEFERRABLE;


ALTER TABLE public.asgposvar ADD CONSTRAINT gposvar_asgposvar_fk
FOREIGN KEY (idgposvar)
REFERENCES public.gposvar (idgposvar)
NOT DEFERRABLE;


ALTER TABLE public.gposeq ADD CONSTRAINT gposvar_gposeq_fk
FOREIGN KEY (idgposvar)
REFERENCES public.gposvar (idgposvar)
NOT DEFERRABLE;
```

```
ALTER TABLE public.asgposeq ADD CONSTRAINT gposeq_asgposeq_fk
FOREIGN KEY (idgposeq)
REFERENCES public.gposeq (idgposeq)
NOT DEFERRABLE;


ALTER TABLE public.asgposvar ADD CONSTRAINT variables_asgposvar_fk
FOREIGN KEY (idvariable)
REFERENCES public.variables (idvariable)
NOT DEFERRABLE;
```

Apéndice E

Script para cargar datos en la base de datos

```
LOAD DATA INFILE
'/home/ISRAEL/Ciencias/Tesis/Tablas/CSV/gral/variables.csv'
INTO TABLE nfl.variables
FIELDS TERMINATED BY ',';

LOAD DATA INFILE
'/home/ISRAEL/Ciencias/Tesis/Tablas/CSV/gral/equipos.csv'
INTO TABLE nfl.equipos
FIELDS TERMINATED BY ',';

LOAD DATA INFILE
'/home/israel/Documentos/Ciencias/Tesis/Tablas/CSV/gral/gposvar.csv'
INTO TABLE nfl.gposvar
FIELDS TERMINATED BY ',';

LOAD DATA INFILE
'/home/israel/Documentos/Ciencias/Tesis/Tablas/CSV/gral/gposeq.csv'
INTO TABLE nfl.gposeq
FIELDS TERMINATED BY ',';

LOAD DATA INFILE
'/home/israel/Documentos/Ciencias/Tesis/Tablas/CSV/gral/asgposvar.csv'
INTO TABLE nfl.asgposvar
FIELDS TERMINATED BY ',';

LOAD DATA INFILE
```

```
'/home/israel/Documentos/Ciencias/Tesis/Tablas/CSV/2005/asgposeq.csv'
INTO TABLE nfl.asgposeq
FIELDS TERMINATED BY ',';

LOAD DATA INFILE
'/home/israel/Documentos/Ciencias/Tesis/Tablas/CSV/2005/resultados.csv'
INTO TABLE nfl.resultados
FIELDS TERMINATED BY ',';
```

Apéndice F

Datos de la tabla nfl.variables de la base de datos

```
1,W
2,Plys
3,Yds.G
4,PenYds
5,TOP
6,Pts.G
7,TDs
8,Run
9,Fum
10,FG
11,SFTY
12,RunAvg
13,RunYds.G
14,RunTDs
15,Cmp
16,PassYds.G
17,PassTDs
18,Int
19,Sck
20,SckYds
21,cp1
22,cp2
23,cp3
24,cp4
25,cp5
26,cp6
```

```
27,cp7
28,cp8
29,cp9
30,cp10
31,cp11
```

Apéndice G

Datos de la tabla nfl.variables de la base de datos

Apéndice H

Datos de la tabla nfl.equipos de la base de datos

```
1,Arizona
2,Atlanta
3,Baltimore
4,Buffalo
5,Carolina
6,Chicago
7,Cincinnati
8,Cleveland
9,Dallas
10,Denver
11,Detroit
12,GreenBay
13,Houston
14,Indianapolis
15,Jacksonville
16,KansasCity
17,Miami
18,Minnesota
19,N.Y.Giants
20,N.Y.Jets
21,NewEngland
22,NewOrleans
23,Oakland
24,Philadelphia
25,Pittsburgh
26,SanDiego
```

```
27,SanFrancisco
28,Seattle
29,St.Louis
30,TampaBay
31,Tennessee
32,Washington
```

Apéndice I

Datos de la tabla nfl.gposvar de la base de datos

```
1,porAEM1
2,porAEM2
3,porAEM3
4,porAEM4
5,porAEM5
6,corredores
7,pasadores
8,errores
9,cp
```

Apéndice J

Datos de la tabla nfl.gposeq de la base de datos

```
1,CPAkemd1,9
2,CPAkemd2,9
3,CPAkemd3,9
4,CPAkemd4,9
5,CPAkemd5,9
6,CPSOM1,9
7,CPSOM2,9
8,CPSOM3,9
9,CPSOM4,9
10,CPSOM5,9
11,CPSOM6,9
```

Apéndice K

Datos de la tabla nfl.asgposvar de la base de datos

```
1,21,9
2,22,9
3,23,9
4,24,9
5,25,9
6,26,9
7,27,9
8,28,9
9,28,9
10,28,9
11,29,9
12,30,9
13,31,9
```

Apéndice L

Datos de la tabla nfl.asgposeq de la base de datos

```
1,1,1
2,1,2
3,2,3
4,4,4
5,2,5
6,4,6
7,1,7
8,4,8
9,2,9
10,2,10
11,1,11
12,3,12
13,1,13
14,3,14
15,2,15
16,3,16
17,2,17
18,3,18
19,2,19
20,4,20
21,1,21
22,4,22
23,1,23
24,4,24
25,4,25
26,1,26
```

```
27,3,27
28,3,28
29,3,29
30,2,30
31,3,31
32,1,32
33,6,1
34,9,2
35,11,3
36,11,4
37,11,5
38,7,6
39,7,7
40,7,8
41,11,9
42,7,10
43,9,11
44,11,12
45,11,13
46,9,14
47,11,15
48,6,16
49,8,17
50,6,18
51,6,19
52,6,20
53,7,21
54,8,22
55,10,23
56,10,24
57,10,25
58,8,26
59,10,27
60,8,28
61,6,29
62,6,30
63,6,31
64,7,32
```

Apéndice M

Datos de la tabla nfl.resultados de la base de datos

```
1,20,1,13,16,0,32
2,21,1,0,31,0,4
3,15,1,23,24,0,5
4,10,1,30,10,1,7
5,14,1,9,6,1,8
6,1,1,24,42,0,11
7,18,1,30,25,1,12
8,26,1,14,27,0,16
9,13,1,21,20,1,17
10,29,1,13,23,0,19
11,3,1,15,34,0,25
12,2,1,27,13,1,9
13,6,1,7,49,0,27
14,22,1,10,27,0,28
15,23,1,20,25,0,31
16,30,1,17,0,1,24
17,32,2,33,31,1,2
18,8,2,13,33,0,3
19,11,2,6,31,0,12
20,31,2,7,33,0,14
21,4,2,38,17,1,15
22,25,2,20,41,0,16
23,13,2,10,31,0,22
24,17,2,21,10,1,20
25,27,2,24,27,0,29
26,28,2,38,0,1,1
```

```
27,5,2,12,9,1,30
28,7,2,20,23,0,23
29,21,2,31,10,1,24
30,10,2,37,13,1,26
31,6,2,13,24,0,18
32,9,2,35,32,1,19
33,30,3,31,10,1,2
34,25,3,17,10,1,7
35,18,3,23,13,1,11
36,16,3,42,14,1,13
37,15,3,13,23,0,14
38,20,3,16,23,0,21
39,22,3,12,27,0,31
40,12,3,13,20,0,1
41,29,3,23,24,0,28
42,19,3,24,21,1,32
43,3,3,24,10,1,26
44,8,3,13,12,1,27
45,4,3,7,17,0,17
46,23,3,10,31,0,10
47,16,4,17,10,1,3
48,24,4,23,13,1,4
49,2,4,3,23,0,5
50,7,4,21,14,1,8
51,15,4,20,24,0,13
52,27,4,7,35,0,18
53,31,4,30,13,1,25
54,1,4,13,37,0,29
55,21,4,17,20,0,32
56,26,4,31,34,0,23
57,11,4,16,20,0,10
58,9,4,17,6,1,20
59,14,4,55,21,1,22
60,12,4,38,23,1,6
61,18,5,39,26,1,2
62,7,5,16,22,0,4
63,22,5,13,19,0,5
64,23,5,21,24,0,6
65,1,5,7,24,0,9
66,28,5,13,35,0,12
67,10,5,23,24,0,16
68,31,5,30,38,0,21
```

```
69,17,5,23,10,1,19
70,26,5,21,27,0,15
71,32,5,25,27,0,24
72,11,5,17,24,0,27
73,8,5,33,13,1,25
74,14,5,38,35,1,30
75,23,6,7,13,0,8
76,24,6,21,23,0,9
77,16,6,40,34,1,12
78,5,6,23,20,1,14
79,17,6,24,10,1,15
80,19,6,6,17,0,21
81,6,6,13,20,0,22
82,13,6,17,38,0,31
83,30,6,35,13,1,32
84,3,6,26,18,1,1
85,25,6,14,17,0,10
86,4,6,3,30,0,20
87,27,6,19,20,0,28
88,2,6,0,36,0,29
89,22,7,45,17,1,2
90,32,7,7,24,0,4
91,31,7,37,17,1,5
92,3,7,26,34,0,7
93,26,7,26,20,1,8
94,9,7,38,7,1,11
95,21,7,19,13,1,17
96,10,7,20,28,0,18
97,24,7,14,10,1,19
98,12,7,24,34,0,29
99,20,7,19,14,1,13
100,30,7,7,24,0,27
101,6,7,17,24,0,28
102,16,7,17,10,1,23
103,10,8,6,26,0,3
104,11,8,16,24,0,6
105,28,8,24,27,0,7
106,31,8,30,17,1,15
107,19,8,29,17,1,18
108,8,8,3,9,0,21
109,5,8,23,20,1,22
110,29,8,33,21,1,25
```

```
111,9,8,0,16,0,30
112,27,8,13,16,0,1
113,13,8,21,30,0,14
114,20,8,17,24,0,24
115,4,8,5,38,0,16
116,17,8,26,10,1,26
117,15,9,17,24,0,3
118,26,9,7,20,0,6
119,32,9,14,21,0,9
120,23,9,13,23,0,11
121,5,9,10,14,0,13
122,14,9,23,17,1,17
123,19,9,31,28,1,20
124,22,9,17,14,1,30
125,7,9,14,17,0,1
126,25,9,16,23,0,28
127,24,9,23,16,1,2
128,29,9,10,30,0,27
129,12,9,30,27,1,18
130,21,9,30,26,1,10
131,30,10,24,27,0,5
132,13,10,27,34,0,7
133,6,10,10,12,0,11
134,14,10,23,28,0,15
135,8,10,20,41,0,16
136,2,10,27,7,1,19
137,1,10,15,28,0,25
138,17,10,7,31,0,31
139,28,10,20,27,0,32
140,18,10,28,42,0,26
141,4,10,6,10,0,9
142,20,10,27,24,1,23
143,3,10,22,33,0,29
144,24,10,17,14,1,12
145,13,11,12,10,1,4
146,32,11,17,20,0,5
147,29,11,23,21,1,6
148,16,11,19,24,0,7
149,1,11,6,44,0,8
150,3,11,6,9,0,17
151,2,11,20,23,0,22
152,19,11,10,28,0,24
```

```
153,15,11,3,10,0,31
154,26,11,8,37,0,10
155,20,11,31,38,0,14
156,18,11,18,28,0,23
157,11,11,14,35,0,28
158,12,11,20,13,1,30
159,9,11,0,12,0,21
160,25,11,14,30,0,27
161,28,12,41,44,0,3
162,14,12,17,14,1,4
163,25,12,13,6,1,8
164,5,12,20,24,0,9
165,27,12,10,20,0,12
166,21,12,23,20,1,13
167,11,12,14,24,0,18
168,15,12,10,13,0,20
169,22,12,20,33,0,24
170,29,12,30,27,1,1
171,6,12,19,10,1,10
172,31,12,38,31,1,2
173,23,12,24,27,0,16
174,7,12,34,27,1,26
175,32,12,23,24,0,17
176,19,12,13,19,0,30
177,12,13,14,22,0,11
178,17,13,40,21,1,9
179,27,13,6,44,0,3
180,24,13,25,16,1,5
181,1,13,3,28,0,6
182,2,13,13,17,0,13
183,21,13,38,34,1,14
184,4,13,24,7,1,19
185,7,13,24,20,1,25
186,18,13,17,48,0,29
187,22,13,24,20,1,32
188,10,13,22,8,1,23
189,16,13,28,24,1,26
190,8,13,7,34,0,28
191,30,13,10,17,0,15
192,31,13,17,24,0,20
193,7,14,13,31,0,3
194,26,14,14,7,1,11
```

```
195,6,14,21,34,0,12
196,13,14,0,27,0,15
197,28,14,7,34,0,18
198,30,14,14,7,1,22
199,32,14,20,7,1,19
200,9,14,10,36,0,24
201,23,14,7,27,0,25
202,14,14,29,27,1,31
203,1,14,14,50,0,27
204,20,14,6,17,0,4
205,16,14,27,45,0,10
206,17,14,0,12,0,21
207,5,14,14,20,0,2
208,29,14,26,20,1,8
209,18,15,10,13,0,6
210,27,15,38,41,0,7
211,2,15,7,38,0,14
212,11,15,17,45,0,16
213,15,15,13,27,0,21
214,25,15,0,6,0,20
215,28,15,22,27,0,29
216,13,15,3,16,0,30
217,4,15,26,28,0,31
218,9,15,27,0,1,32
219,8,15,20,23,0,10
220,3,15,12,20,0,23
221,5,15,20,17,1,1
222,12,15,38,21,1,26
223,19,15,7,45,0,22
224,24,15,34,27,1,17
225,2,16,30,28,1,30
226,16,16,20,45,0,18
227,21,16,21,16,1,20
228,17,16,20,3,1,4
229,11,16,14,20,0,5
230,32,16,24,27,0,6
231,3,16,35,0,1,8
232,19,16,3,19,0,9
233,31,16,27,24,1,13
234,22,16,19,20,0,15
235,7,16,10,27,0,29
236,26,16,24,40,0,25
```

```
237,27,16,31,28,1,24
238,1,16,10,28,0,28
239,10,16,31,17,1,14
240,12,16,41,7,1,23
241,4,17,0,31,0,21
242,28,17,24,17,1,27
243,24,17,31,7,1,32
244,15,17,14,21,0,2
245,8,17,22,14,1,7
246,29,17,20,30,0,11
247,14,17,20,17,1,13
248,6,17,3,31,0,16
249,20,17,21,23,0,17
250,9,17,7,13,0,22
251,30,17,13,33,0,31
252,18,17,17,18,0,1
253,5,17,37,24,1,19
254,10,17,3,31,0,12
255,23,17,14,21,0,26
256,25,17,10,13,0,3
```

Apéndice N

Código en R

```
library(Rcmdr);
################    final     ################
setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/eje-Agrupacion.csv', header=TRUE,row.names=1, sep=","

setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2007.csv', header=TRUE,row.names=1, sep=",");

setwd("/media/ISRAEL/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2005.csv', header=TRUE,row.names=1, sep=",");

setwd("/home/gaviota/ISRAEL/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2006.csv', header=TRUE,row.names=1, sep=",");

#########################################################################
# Funciones  Generales:
#########################################################################
#-------------------------------------------------------------------
# Funcion para obtener las medias y desviaciones estandar de cada variable
# Parametros:
# Matrix: Matrix de datos
# Retorna:
# Matrix: Matrix medias y desviaciones estandar
#-------------------------------------------------------------------
MedDes<-function(matrix){
 D=dim(matrix)[2]
 M=matrix(0,D,2)
```

```
 for (i in 1:D){
  M[i,1]<-mean(matrix[,i])
  M[i,2]<-sqrt(var(matrix[,i]))
  }
M
}
#---------------------------------------------------------------------
# Ejecucion de la funcion MedVar
#---------------------------------------------------------------------
XMV<-MedDes(get)
XMV
#---------------------------------------------------------------------

#######################################################################
# componentes principales:
#######################################################################
#---------------------------------------------------------------------
# Funcion que presenta el procentaje explicado acumulado de las
# componentes principales
# Parametros:
# Matrix: Matriz de datos
# p1:1er porcentaje deseado
# p2:2do porcentaje deseado
# B:0 o 1 Bandera que indica a la funcion que devolver
#   0 -> la funcion devuelve la matriz de porcentajes
#   1 -> la funcion devuelve el numero de componenete donde se
#        alcanzo el 1er porcentaje deseado
# P:1 o 2 Bandera que indica el formato de la grafica
#   1 -> formato con porcentaje individual
#   2 -> formato con porcentaje acumulado
# Retorna:
# Resp: Matriz de porcentajes
# Resp: Numero de componente para el 1er porcentaje deseado
#---------------------------------------------------------------------
PorcentajeCP<-function(matrix,p1,p2,B,P){

 #Para asignar a un device especifico
 x11(width=10,height=7);
 plt2<-dev.cur();
 dev.set(plt2)

 Mcorget<-cor(matrix)
```

```
E<-eigen(Mcorget)
D=dim(matrix)[2]
den=sum(diag(Mcorget))
porcen <- matrix(0,D,2)
colnames(porcen)=c("%Componente","%Acumulado")

for (i in 1:D){
 porcen[i,1]<-E$values[i]/den;
 }

porcen[1,2]<-porcen[1,1]

for (i in 2:D){
 porcen[i,2]<-porcen[i,1]+porcen[i-1,2]
 }

plot(porcen[,P],xlab="Componente",ylab=" Porcentaje")

cp1=0
cp2=0
N=0
for (i in 1:D){
 if (porcen[i,2]>=p1 & cp1==0){
   cp1=i
   N<-i
   }
 }
for (i in 1:D){
 if (porcen[i,2]>=p2 & cp2==0)
   cp2=i
 }

if(P==1){
 abline(v=cp1, col = "blue")
 t1=100*round(porcen[cp1,2],2)
 text(12,0.2, t1, , col = "blue", adj = c(0, -.1))
 text(13,0.2, "% hasta el componente", col = "blue", adj = c(0, -.1))
 text(23,0.2, cp1, , col = "blue", adj = c(0, -.1))

 abline(v=cp2, col = "red")
 t1=100*round(porcen[cp2,2],2)
 text(12,0.1, t1, , col = "red", adj = c(0, -.1))
```

```
 text(13,0.1, "% hasta el componente", col = "red", adj = c(0, -.1))
 text(23,0.1, cp2, , col = "red", adj = c(0, -.1))
}

if (P==2){
 abline(v=cp1, col = "blue")
 t1=100*round(porcen[cp1,2],2)
 text(12,0.75, t1, , col = "blue", adj = c(0, -.1))
 text(13,0.75, "% hasta el componente", col = "blue", adj = c(0, -.1))
 text(23,0.75, cp1, , col = "blue", adj = c(0, -.1))

 abline(v=cp2, col = "red")
 t1=100*round(porcen[cp2,2],2)
 text(12,0.6, t1, , col = "red", adj = c(0, -.1))
 text(13,0.6, "% hasta el componente", col = "red", adj = c(0, -.1))
 text(23,0.6, cp2, , col = "red", adj = c(0, -.1))
}


 if (B==0)
  Resp=porcen
 if (B==1)
  Resp=N
 Resp
}

#-----------------------------------------------------------------------
# Ejecucion de la funcion PorcentajeCP
#-----------------------------------------------------------------------
P<-PorcentajeCP(get,0.95,0.75,1,1);
P
P<-PorcentajeCP(get,0.95,0.75,0,1);
P
P<-PorcentajeCP(get,0.95,0.75,1,2);
P
P<-PorcentajeCP(get,0.95,0.75,0,2);
P
#-----------------------------------------------------------------------
# Hace un analisis de las comoponentes principales de una matriz
# Parametros:
# Matrix: Matrix de Datos
# porc1:  1er porcentage de la variabilidad explicado por las CP deseado
```

```
# porc2: 2o porcentage de la variabilidad explicado por las CP deseado
# Retorna: Matrix de analisis de las componentes principales
#
# Funciones requeridas:
# MedDes
# PorcentajeCP
#-----------------------------------------------------------------------
AnaCP<-function(matrix,porc1,porc2){

  CP<-princomp(matrix,cor=TRUE)
  N<-PorcentajeCP(matrix,porc1,porc2,1,2);
  cp<-CP$scores[,1:N]
  M<-dim(cp)[1]
  medDes<-MedDes(cp)
  resp=matrix(,M,N)
  rownames(resp)=rownames(cp)
  colnames(resp)=colnames(cp)

  for (i in 1:N){
    for (j in 1:M){
      if (abs(cp[j,i])>= max(abs(medDes[i,1]+medDes[i,2]), abs((cp[j,i])>=medDes[i,1]-medDes
        resp[j,i]<-round(cp[j,i],3);
      }
    }
  }
  resp
}
#-----------------------------------------------------------------------
# Ejecucion de la funcion AnaCP
#-----------------------------------------------------------------------
mCP<-AnaCP(get,0.95,0.75);
mCP
#########################################################################
# Numero de grupos:
#########################################################################
#
#-----------------------------------------------------------------------
# Que calcula el numero de grupos a utilizar tanto para k-medias como
# para SOM
# Parametros:
# Matrix: Matrix de Datos
# vars:  vector con las variables a ser consideras (la posisicion)
```

```
# Entero: NG numero de grupos
# Retorna: Matrix de escalamiento y agrupamiento
#
#-----------------------------------------------------------------------
traza<-function(matrix){
 long<-dim(matrix)[2];
 suma<-0;
 for (i in 1:long){
    suma=suma+matrix[i,i];
 }
 suma;
}
#-----------------------------------------------------------------------
# Ejecucion de la funcion traza
#-----------------------------------------------------------------------
covar<-cov(cp)
t<-traza(covar)
t



#-----------------------------------------------------------------------
# Que calcula el numero de grupos a utilizar tanto para k-medias como
# para SOM
# Parametros:
# Matrix: Matrix de Datos
# porc1:  1er porcentage de la variabilidad explicado por las CP deseado
# porc2: 2o porcentage de la variabilidad explicado por las CP deseado
# inicio: 1er Numero de grupos a considerar
# fin: ultimo Numero de grupos a considerar
# Retorna: Matrix de analisis del numero de grupos de acuerdo al
#               estimador "das+dpeqdqd"
#
# Funciones requeridas:
# traza
# Akmedias
# PorcentajeCP
#
#-----------------------------------------------------------------------
Ngrupos<-function(matrix,porc1,porc2,inicio,fin){

  p<-dim(matrix)[1]
```

```
CP<-princomp(matrix,cor=TRUE)
N<-PorcentajeCP(matrix,porc1,porc2,1,2);
cp<-CP$scores[,1:N]
media<-mean(cp)

long<- (fin-inicio+1)
resp <- matrix(0,long,5)
colnames(resp)=c("g","Sg","[gˆ(2/p)]Sg","Dif","Cg")

resp[1,1]<-inicio
contador=(0+inicio)
for (i in 2:long){
  contador=contador+1
  resp[i,1]<-contador
}

tabla <- matrix(0,dim(cp)[2],dim(cp)[2])
for (m in inicio:fin){
  Ak<-Akmed("CP-K-medias",matrix,N,m);
  for (j in 1:m){


    contador=0
    for (i in 1:dim(Ak)[1]){
      if (Ak[i,3]==j){
        contador=contador+1
      }
    }

    tabla1 <- matrix(0,contador,dim(cp)[2])

    contador=0
    for (i in 1:dim(Ak)[1]){
      if (Ak[i,3]==j){
        contador=contador+1
        for (h in 1:dim(cp)[2]){
          tabla1[contador,h]=cp[i,h]
        }
      }
    }
```

```
      tabla1<-tabla1-media
      tabla2<-(t(tabla1) %*% tabla1)/contador
      tabla<-tabla + tabla2
    }

  T<-traza(tabla)
  resp[m-1,2]<- T
  resp[m-1,3]<- (resp[m-1,1]^(2/p))*T
  }

  for (i in 2:long){
    resp[i,4]<- (resp[i-1,3])-(resp[i,3])
  }

  for (i in 1:long-1){
    resp[i,5]<- abs((resp[i,4])/(resp[i+1,4]))
  }

  resp

}
#-----------------------------------------------------------------------
# Ejecucion de la funcion Ngrupos
#-----------------------------------------------------------------------
G<-Ngrupos(get,0.95,0.75,2,8)
G


#########################################################################
# Clusters:
#########################################################################
#-----------------------------------------------------------------
# CLUSTERS UTILIZANDO COMPONENTES PRINCIPALES:
# Clasificacion de los equipos utilizando las componentes principales
#-----------------------------------------------------------------
#
#-----------------------------------------------------------------
# Funcion utiliza agrupamiento y escalamiento multidimensional de
# acuerdo a clasificacion previa de variables utilizando el metodo
# de Kmedias (de los equipos)
# Parametros:
```

```
# nombre: nombre de la clasificacion
# Matrix: Matrix de Datos
# N:  numero de componentes principales a utilizar
# NG: numero de grupos
# Retorna: Matrix de escalamiento y agrupamiento
#
#-----------------------------------------------------------------------
Akmed<-function(nombre,matrix,N,NG){

 x11(width=10,height=7);
 plt3<-dev.cur();

 CP<-princomp(matrix,cor=TRUE)
 cp<-CP$scores[,1:N]
 media<-mean(cp)
 ren=dim(matrix)[1]

 Mmeans<-kmeans(cp,NG)

 M=matrix(0,ren,3)
 rownames(M)=rownames(matrix)
 M[,1]=cp[,1]
 M[,2]=cp[,2]
 M[,3]=Mmeans$cluster

 #Para asignar a un device especifico
 dev.set(plt3)
 plot(M,type="n")
 text(M[,1],M[,2],rownames(M),cex=0.8,col=M[,3])
 #plot3d(M[,1],M[,2],type="s",col=M[,3])

 title(main = list("Agrupamiento con CP k-medias",
       cex=1.5,col="black", font=2))
 M
}

#-----------------------------------------------------------------------
# Ejecucion de la funcion Akmed
#-----------------------------------------------------------------------
N<-PorcentajeCP(get,0.95,0.75,1,2);
N
Ak<-Akmed("CP-K-medias",get,N,5);
```

```
#########################################################################
# SOM:
#########################################################################

#-------------------------------------------------------------------
# Funcion utiliza agrupamiento autoorganizado con el metodo de SOM
# para los (de los equipos)
# Parametros:
# nombre: nombre de la clasificacion
# Matrix: Matrix de Datos
# N:  numero de componentes principales a utilizar
# Entero: R numero de renglones (de grupos)
# Entero: C numero de columnas (de grupos)
# Retorna: Matrix agrupamiento
#
#-------------------------------------------------------------------
Asom<-function(nombre,matrix,N,R,C){

 library("class")
 library("kohonen")

 x11(width=10,height=7);
 plt3<-dev.cur();
 x11(width=10,height=7);
 plt4<-dev.cur();

 CP<-princomp(matrix,cor=TRUE)
 cp<-CP$scores[,1:N]

 ren=dim(matrix)[1]
 colmatrix=dim(matrix)[2]

 get.som <- som(data = cp, rlen=100, grid = somgrid(R, C, "hexagonal"))
 mapping <- map(get.som,get.som)


 M=matrix(0,ren,3)
 rownames(M)=rownames(matrix)
 #P=princomp(Mi)
 M[,1]=cp[,1]
 M[,2]=cp[,2]
```

```
 M[,3]=mapping$unit.classif

 #Para asignar a un device especifico
 dev.set(plt3)
 plot(M,type="n")
 text(M[,1],M[,2],rownames(M),cex=0.8,col=M[,3])
 title(main = list("Agrupamiento con CP SOM",
       cex=1.5,col="black", font=2))

 #Para asignar a un device especifico
 dev.set(plt4)
 plot(get.som, main = "NFL data SOM")

 M
}

#-----------------------------------------------------------------
# Ejecucion de la funcion Asom
#-----------------------------------------------------------------
N<-PorcentajeCP(get,0.95,0.75,1,2);
N
Asom("CP-SOM",get,N,2,3)

#########################################################################
#GENERACION DE ARCHIVOS PARA BD o archivo:
#########################################################################
#   Resultados
####################
#  Temporada 2003
####################
setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2003.csv', header=TRUE,row.names=1, sep=",");

CP<-princomp(get,cor=TRUE)
biplot(CP,cex=0.6,xlim=c(-0.5,0.5))

p<-PorcentajeCP(get,0.95,0.75,1,1);
p<-PorcentajeCP(get,0.95,0.75,1,2);
P<-PorcentajeCP(get,0.95,0.75,0,2);

write.table(P, file = "/home/israel/Documentos/Ciencias/Tesis/res/2003/Porc-CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
```

```
N<-PorcentajeCP(get,0.95,0.75,1,2);
N
Ak<-Akmed("CP-K-medias",get,N,5);
C<-CP$scores[,1:p]
write.table(C, file = "/home/israel/Documentos/Ciencias/Tesis/res/2003/CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

mCP<-AnaCP(get,0.95,0.75);
write.table(mCP, file = "/home/israel/Documentos/Ciencias/Tesis/res/2003/AnaCP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

G<-Ngrupos(get,0.95,0.75,2,8);
write.table(G, file = "/home/israel/Documentos/Ciencias/Tesis/res/2003/Ngrupos.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
g<-4

A<-Akmed("CP-K-medias",get,N,g);
write.table(A, file = "/home/israel/Documentos/Ciencias/Tesis/res/2003/AsGpoKmed.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);


S<-Asom("CP-SOM",get,N,2,3);
write.table(S, file = "/home/israel/Documentos/Ciencias/Tesis/res/2003/AsGpoSOM.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);



####################
#  Temporada 2004
####################
setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2004.csv', header=TRUE,row.names=1, sep=",");

CP<-princomp(get,cor=TRUE)
biplot(CP,cex=0.6,xlim=c(-0.5,0.5))

p<-PorcentajeCP(get,0.95,0.75,1,1);
p<-PorcentajeCP(get,0.95,0.75,1,2);
P<-PorcentajeCP(get,0.95,0.75,0,2);
write.table(P, file = "/home/israel/Documentos/Ciencias/Tesis/res/2004/Porc-CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
```

```
C<-CP$scores[,1:p]
write.table(C, file = "/home/israel/Documentos/Ciencias/Tesis/res/2004/CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

mCP<-AnaCP(get,0.95,0.75);
write.table(mCP, file = "/home/israel/Documentos/Ciencias/Tesis/res/2004/AnaCP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

G<-Ngrupos(get,0.95,0.75,2,8);
write.table(G, file = "/home/israel/Documentos/Ciencias/Tesis/res/2004/Ngrupos.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
g<-5

A<-Akmed("CP-K-medias",get,N,g);
write.table(A, file = "/home/israel/Documentos/Ciencias/Tesis/res/2004/AsGpoKmed.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);


S<-Asom("CP-SOM",get,N,2,3);
write.table(S, file = "/home/israel/Documentos/Ciencias/Tesis/res/2004/AsGpoSOM.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);



####################

####################
#  Temporada 2005
####################
setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2005.csv', header=TRUE,row.names=1, sep=",");

CP<-princomp(get,cor=TRUE)
biplot(CP,cex=0.6,xlim=c(-0.5,0.5))

p<-PorcentajeCP(get,0.95,0.75,1,1);
p<-PorcentajeCP(get,0.95,0.75,1,2);
P<-PorcentajeCP(get,0.95,0.75,0,2);
write.table(P, file = "/home/israel/Documentos/Ciencias/Tesis/res/2005/Porc-CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

C<-CP$scores[,1:p]
```

```
write.table(C, file = "/home/israel/Documentos/Ciencias/Tesis/res/2005/CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

mCP<-AnaCP(get,0.95,0.75);
write.table(mCP, file = "/home/israel/Documentos/Ciencias/Tesis/res/2005/AnaCP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

G<-Ngrupos(get,0.95,0.75,2,8);
write.table(G, file = "/home/israel/Documentos/Ciencias/Tesis/res/2005/Ngrupos.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
g<-5

A<-Akmed("CP-K-medias",get,N,g);
write.table(A, file = "/home/israel/Documentos/Ciencias/Tesis/res/2005/AsGpoKmed.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);


S<-Asom("CP-SOM",get,N,2,3);
write.table(S, file = "/home/israel/Documentos/Ciencias/Tesis/res/2005/AsGpoSOM.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);



####################
#  Temporada 2006
####################
setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2006.csv', header=TRUE,row.names=1, sep=",");

CP<-princomp(get,cor=TRUE)
biplot(CP,cex=0.6,xlim=c(-0.5,0.5))

p<-PorcentajeCP(get,0.95,0.75,1,1);
p<-PorcentajeCP(get,0.95,0.75,1,2);
P<-PorcentajeCP(get,0.95,0.75,0,2);
write.table(P, file = "/home/israel/Documentos/Ciencias/Tesis/res/2006/Porc-CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

C<-CP$scores[,1:p]
write.table(C, file = "/home/israel/Documentos/Ciencias/Tesis/res/2006/CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
```

```
mCP<-AnaCP(get,0.95,0.75);
write.table(mCP, file = "/home/israel/Documentos/Ciencias/Tesis/res/2006/AnaCP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

G<-Ngrupos(get,0.95,0.75,2,8);
write.table(G, file = "/home/israel/Documentos/Ciencias/Tesis/res/2006/Ngrupos.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
g<-5

A<-Akmed("CP-K-medias",get,N,g);
write.table(A, file = "/home/israel/Documentos/Ciencias/Tesis/res/2006/AsGpoKmed.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);


S<-Asom("CP-SOM",get,N,2,3);
write.table(S, file = "/home/israel/Documentos/Ciencias/Tesis/res/2006/AsGpoSOM.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);


####################
#  Temporada 2007
####################
setwd("/home/israel/Documentos/Ciencias/Tesis/R");
get<-read.table('../Tablas/CSV/estad-R/NFL-2007.csv', header=TRUE,row.names=1, sep=",");

CP<-princomp(get,cor=TRUE)
biplot(CP,cex=0.6,xlim=c(-0.5,0.5))

p<-PorcentajeCP(get,0.95,0.75,1,1);
p<-PorcentajeCP(get,0.95,0.75,1,2);
P<-PorcentajeCP(get,0.95,0.75,0,2);
write.table(P, file = "/home/israel/Documentos/Ciencias/Tesis/res/2007/Porc-CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

C<-CP$scores[,1:p]
write.table(C, file = "/home/israel/Documentos/Ciencias/Tesis/res/2007/CP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);

mCP<-AnaCP(get,0.95,0.75);
write.table(mCP, file = "/home/israel/Documentos/Ciencias/Tesis/res/2007/AnaCP.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
```

```
G<-Ngrupos(get,0.95,0.75,2,8);
write.table(G, file = "/home/israel/Documentos/Ciencias/Tesis/res/2007/Ngrupos.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
g<-5

A<-Akmed("CP-K-medias",get,N,g);
write.table(A, file = "/home/israel/Documentos/Ciencias/Tesis/res/2007/AsGpoKmed.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);


S<-Asom("CP-SOM",get,N,2,3);
write.table(S, file = "/home/israel/Documentos/Ciencias/Tesis/res/2007/AsGpoSOM.csv",
        sep = ",", col.names =TRUE, row.names=TRUE);
```

Bibliografía

[Abr02] Silberschatz Abraham. *Fundamentos de Bases de Datos.* Mc. Graw Hill, 4a. edición, España, (2002).

[Ale78] Mood M. Alexander. *Introducción a la teoría de la Estadística.* Aguilar S.A. de ediciones, 4a. edición, España, (1978).

[Áng05] García Álvarez Miguel Ángel. *Introducción a la teoría de la probabilidad (primer curso).* Fondo de Cultura Económica, México, (2005).

[H.D79] Brunk H.D. *Introducción a la estadística matemática.* Trillas, México, (1979).

[JR90] Sirosh Joseph and Miikkulainen Risto. Self-organizing feature maps with lateral connections: Modeling ocular dominance. (1990).

[JT88] Krzanowski W. J. and Lai Y. T. A criterion for determining the number of groups in a data set using sum-of-squares clustering. (1988).

[Ric07] Johnson A. Richard. *Applied Multivariate Statistical Analysis.* Prentice Hall, United States of America, (2007).

[RT90] Beale R. and Jackson T. *Neuronal Computing an introduction.* Taylor and Francis Group, New York, (1990).

[Sch98] Hines O'Hara R. J. Kovacs M. Kit Schreer F. Jason. Classification of dive profiles: A comparison of statistical clustering techniques and unsupervised artifical neural networks. (1998).

[Ste82] Friedberg H. Stephen. *Algebra Lineal.* Publicaciones cultural S.A., México, (1982).

Referencias electrónicas

http://www.nfl.com

Índice alfabético