# PREDICIENDO PARTIDOS DEL MUNDIAL FIFA BRASIL 2014

## PREDICTING MATCHES OF THE 2014 FIFA WORLD CUP BRAZIL

*Pedro Nelson Shiguihara Juárez*[1]
*Víctor Yoel Huamán Asencio*[2]

**Fecha de recepción: 21 junio 2014**
**Fecha de aceptación: 27 junio 2014**

**Resumen**

El objetivo de este artículo es mostrar por qué las herramientas del novel campo de estudio llamado **Big Data**, está teniendo un creciente interés en diversas áreas alrededor del mundo. Para ello, se abordó una de las herramientas de Big Data denominada *Machine Learning*, específicamente un enfoque probabilístico basado en el Teorema de Bayes. Con este enfoque se propone la construcción de un modelo que permita predecir el ganador en un partido de fútbol a partir de información histórica. Esta información histórica contempla los últimos tres años de partidos disputados por cada una de las 32 selecciones participantes en el mundial FIFA Brasil 2014. A pesar de que el modelo es simple, éste obtuvo resultados competitivos en comparación con los resultados obtenidos por una red neuronal y un portal web de predicciones deportivas. Lo resultados proporcionados por las técnicas de Big Data fueron interesantes a pesar de la alta incerteza que presenta un partido de fútbol. Además, para el gobierno y empresas, estas técnicas podrían ser empleadas para salvar vidas, ahorrar tiempo y dinero en tareas críticas.

**Palabras Clave***: Big Data*, Inferencia Probabilística, *Machine Learning,* Predicción de Partidos de Fútbol, Teorema de Bayes.

**Abstract**

Our goal in this paper is to show why important techniques for using Big Data have a growing interest in many areas around the world. In this context, we approach a Machine Learning technique, specifically a probabilistic approach based on Bayes theorem. We propose to construct a probabilistic model in order to predict the winning team of a football match using historical data. The historical data includes football matches of the last three years related to the participating countries of the 2014 FIFA World Cup Brazil. Although we propose a simple model, this model reached better results in comparison with the results of a neural network and a web specialized in sport predictions. We conclude that techniques for using Big Data reached interesting results despite the high uncertainty in a football match. Also, for the government and enterprises, these techniques could be employed to save lives and to save time and money in critical tasks.

**Keywords:** *Bayes theorem, Big Data, Football Predictions, Machine Learning, Probabilistic Inference.*

[1] *Adscrito a la Escuela Académico Profesional de Ingeniería de Sistemas, Ingeniero de Sistemas, Docente de la Universidad Señor de Sipan, Pimentel, Perú, pshiguihara@crece.uss.edu.pe.*
[2] *Adscrito a la Escuela Académico Profesional de Ingeniería de Sistemas, Estudiante de Ingeniería de Sistemas, Estudiante de la Universidad Señor de Sipan, Pimentel, Perú, hasenciov@crece.uss.edu.pe.*

## 1. INTRODUCCIÓN

Big Data está íntimamente ligada a la era en la que actualmente nos encontramos: la era de la información. El nombre no habría podido quedar mejor ya que el internet se está "inundado" de información. En otras palabras, la cantidad de información que se está acumulando diariamente es astronómica. Por ejemplo, a cada 60 segundos alrededor del mundo, en YouTube se suben alrededor de 30 horas en videos, en iTunes se descargan 15,000 canciones, en Facebook se crean 350 GB de información, en Twitter se publican 100,000 nuevos tweets, en Flickr se suben 3000 fotos, en las bandejas de entrada de cuentas de correos electrónicos se reciben 204 millones de emails, en Wikipedia se suben 6 artículos nuevos. Todo eso, solo en 60 segundos, es decir en menos del 0.1% de nuestro día (0.0694% para ser más exactos). En **Intel (2014)** se presenta una infografía que retrata esta avalancha de información generándose cada 60 segundos. De hecho, según **Lohr (2012)**, cada par de años esa información se duplica y la tiene disponible en internet. Pero, ¿Para qué nos sirve toda esa información?

En Big Data, se propone usar esa información, la cual llamaremos información histórica, para predecir eventos futuros y tomar mejores decisiones al respecto **(Gopalkrishnan et al., 2012)**. Hoy en día se tiene un volumen de información como nunca antes en la historia de la humanidad y muchos eventos son registrados digitalmente y están disponibles para todo el mundo a través de internet. Al usar esa información histórica con herramientas de Big Data, es posible salvar vidas y ahorrar tiempo y dinero en la ejecución de diversas tareas. Por ejemplo, se estima que el 97% de compañías con ingresos mayores a $ 100 millones, han usado algún tipo de análisis de negocios que involucra herramientas de Big Data **(Hsinchun, 2012)**. También, fueron analizados datos clínicos a fin de detectar patrones entre los componentes biológicos de enfermedades **(Berman, 2013)**. Asimismo, fue posible la detección de interacciones de proteínas únicamente a partir de grandes volúmenes de artículos biomédicos **(Shiguihara y Andrade, 2013)**. En otro contexto, fue posible el hallazgo de un avión francés perdido en el océano Atlántico en el año 2009. El vuelo 447 de *Air France* desapareció cuando volaba en la ruta París-Rio de Janeiro. Para lidiar con el caso, se reunió información acerca de la ubicación de la caja negra, la velocidad del viento e información acerca de las corrientes marinas entre otros datos. Una herramienta estadística fue usada para determinar cuál erea el lugar más probable donde podía haberse perdido el avión de *Air France* **(Ehab, 2014)**. Esto facilitó enormemente los planes de búsqueda sobre un área que era mucho mayor que el área comprendida por las regiones de Lambayeque, Piura y Tumbres.

En el deporte, también han sido usadas herramientas de Big Data para predecir eventos futuros. Por ejemplo, en el béisbol fueron empleadas para ayudar a las tomas de decisiones con relación a la contratación de jugadores **(Lohr, 2012)**. El objetivo era predecir si los jugadores que potencialmente serían contratados, iban a rendir lo que el club esperaba de ellos. En el fútbol, también se han propuesto técnicas para predecir resultados de partidos. En **Florez (2014)**, fue propuesta una red neuronal para predecir los resultados del mundial FIFA Brasil 2014. A partir de rankings proporcionados por web especializadas en fútbol, la red neuronal "aprende" a identificar las posibilidades de éxito de cada equipo cuando es enfrentado a otro.

El objetivo de este artículo es mostrar el uso de una herramienta de Big Data en la predicción de resultados de partidos del mundial de fútbol FIFA Brasil 2014. Nuestra propuesta se basa en la teoría de modelos gráficos probabilísticos en vez de redes neuronales propuesta en **Florez (2014)**. El modelo emplea tres variables aleatorias independientes que permiten evaluar el partido entre dos equipos: *Local vs. Visitante*. No existe ninguna ventaja para los equipos locales sobre los visitantes, sin embargo, al indicar *Local vs. Visitante*, nos permite generalizar a cualquier par de selecciones adversarias

dentro de la evaluación. Así, un modelo es creado por cada partido que enfrente a un *Local vs. Visitante*. Una vez calculadas las distribuciones de probabilidades correspondientes a cada variable aleatoria, se procede a calcular la probabilidad conjunta. En la probabilidad conjunta se aplica el cálculo de *Maximum a Posteriori (MAP)*. De esa manera, consideramos que la selección ganadora es aquella que tiene la máxima probabilidad dentro de la probabilidad conjunta producidas por el producto de las tres variables aleatorias.

En la sección 2, es presentada la metodología de este trabajo con el fin de explicar los criterios de construcción del modelo probabilístico propuesto para predecir los partidos del mundial FIFA Brasil 2014. En la sección 3, son presentados los resultados de la predicción en la fase de grupos. Tales resultados son comparados con la red neuronal que también predice resultados del mundial FIFA Brasil 2014 en **Florez (2014).** Finalmente, en la sección 4 son presentadas las conclusiones del trabajo y en la última sección es presentada la bibliografía de este artículo.

## 2. METODOLOGÌA

En esta sección se presenta el modelo probabilístico propuesto en este artículo. Este modelo predice resultados de partidos de fútbol basado en información histórica. En la Subsección 2.1 se presenta la construcción del modelo probabilístico, el cual tiene tres variables aleatorias. Asimismo, se presenta el empleo de la estimación MAP para calcular qué equipo es el ganador de un partido. En la Subsección 2.2 se presentan las características de la información histórica que fue utilizada para construir el modelo propuesto. Finalmente, en la Subsección 2.3, como parte de este trabajo, se muestra la aplicación, la cual fue implementada para que el usuario pueda ingresar sus resultados a fin de ser contrastados por el modelo predictor.

### 2.1. RAZONAMIENTO PROBABILÍSTICO

En el caso de la ubicación del avión francés, la herramienta estadística que fue usada para el caso del avión de *Air France* invoucra al Teorema de Bayes. En computación, el teorema de Bayes fue estudiado y posteriormente incorporado a un esquema mucho más amplio que es llamado: modelos gráficos probabilísticos **(Koller y Friedmann, 2009; Russel y Norvig, 2010)**. Los modelos gráficos probabilísticos son programas de computadora que incorporan un razonamiento basado en el Teorema de Bayes para llegar a conclusiones a partir de información histórica. Tales modelos tienen presentan dos componentes: (1) una estructura grafo y (2) un conjunto de distribuciones de probabilidades. Con tales componentes, es posible aplicar el teorema de Bayes sobre el modelo para conseguir nuevo conocimiento acerca de un tema de interés.

En los modelos gráficos probabilísticos, un investigador pionero fue el Dr. Judea Pearl. Como resultado de su estudio, se formaron las bases para el posterior nacimiento de los modelos gráficos probabilísticos. Judea Pearl fue galardonado con el “Premio Turing 2011” por la *Association Computing Machinery (ACM)* debido a “su contribución fundamental a la Inteligencia Artificial a través del desarrollo de un cálculo para razonamiento probabilístico y causal” **(ACM, 2011)**.

Antes de realizar el razonamiento probabilístico de nuestro modelo, es necesario crear las tres variables aleatorias. En la **Figura 1**, se muestra un ejemplo del modelo cuando el partido consiste en un juego “Holanda (local) versus España (visitante)”. Para este partido, se consideran los partidos pre-mundialistas de ambos equipos (ver **Tabla 1**). A partir de tales partidos pre-mundialistas, serán creadas las distribuciones de probabilidades correspondientes a cada variable aleatorias.

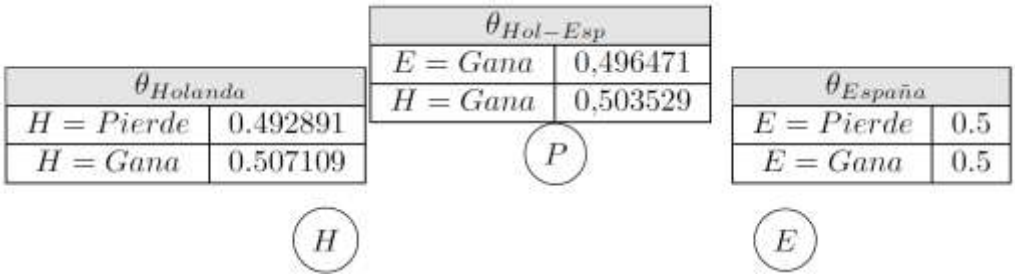



**Figura 1:** *Modelo probabilístico que contiene tres variables aleatorias. Este modelo representa el partido entre Holanda y España. La variable aleatoria "H" representa la probabilidad de que Holanda pierda y gane en cualquier partido (0.49 y 0.51 respectivamente). La variable aleatoria "E" representa la probabilidad de que España pierda y gane en cualquier partido (0.5 y 0.5). La variable aleatoria "P" representa la probabilidad de que Holanda (local) le gane a España (visitante). Este modelo presenta un grafo con tres nodos que representan a las variables aleatorias. Tal grafo es inconexo, indicando la independencia entre todas las variables aleatorias presentes. Cada nodo contiene una distribución de probabilidad que es calculada según el país involucrado en el partido*.

Para el caso "Holanda versus España", la variable "Local" corresponde a "Holanda". En tal sentido, es necesario calcular la distribución correspondiente a la probabilidad que tiene Holanda de ganar cualquier partido. Es decir, se tuvo que hallar la probabilidad marginal: *P(H=Gana)* y *P(H=Pierde)*. No obstante, para calcular estas probabildades, es necesario considerar con qué adversario se está enfrentando. Esto se debe a que en el mundial Brasil 2014 compiten 32 selecciones nacionales que han llegado a estas instancias mundialistas luego de haber superado la fase clasificatoria, la cual dura ente 2 y 3 años. Durante ese tiempo, la exigencia para cada selección participante ha variado según los rivales a los que se han enfrentado. Si consideró que existen 209 selecciones nacionales que están afiliadas a la FIFA (**Wikipedia, 2014a**) y apenas 8 selecciones han logrado ser campeones mundiales en todas las ediciones de copas del mundo, quiere decir que existen diferencias substanciales entre las 209 selecciones.

De hecho, 8 selecciones equivalen al 3.83% del total de selecciones afiliadas a la FIFA mientras que, por otro lado, 64 selecciones nunca han participado en ninguna copa del mundo, es decir, el 22.1% de selecciones aún no han llegado a ningún mundial. En nuestra opinión, esas diferencias son importantes al momento de evaluar a los equipos y a cada partido. Así, evaluar un partido entre España versus Tahití que resultó en 10 a 0, en favor de España, no es lo mismo que evaluar un partido entre España versus Francia que resultó en empate 1 a 1.

Es por ello que el cálculo de probabilidades marginales es calculado basado en la posición dentro del ranking FIFA de los rivales. Los pasos para el cálculo de probabilidades marginales son los siguientes:

1. Calcular la posición promedio de todos los rivales de Holanda y España
2. Calcular la función de frecuencia de victorias y derrotas para cada equipo
3. Calcular las probabilidades marginales de ganar y perder para cada equipo
4. Calcular las probabilidades marginales de que el equipo local gane al equipo visitante.



A continuación se describen los pasos anteriores para calcular las probabilidades marginales en el contexto del partido: "Holanda versus España". El primer paso es calcular la posición promedio de todos los rivales tanto de Holanda como España. Tal posición promedio es: *posProm=18.7436*.

**Tabla 1.**
*En la columna izquierda, se muestran los partidos oficiales jugados por España antes del mundial FIFA Brasil 2014. En la columna derecha, se muestran los partidos oficiales jugados por Holanda. "G1" corresponde al número de goles anotados por el "Equipo 1" mientras que "G2" corresponde al número de goles anotados por el "Equipo 2".*

| Equipo 1 | Equipo 2 | G1 | G2 |
|---|---|---|---|
| Espana | Georgia | 1 | 0 |
| Espana | Bielorrusia | 4 | 0 |
| Espana | Francia | 1 | 1 |
| Espana | Finlandia | 1 | 1 |
| Espana | Francia | 1 | 0 |
| Espana | Uruguay | 2 | 1 |
| Espana | Tahiti | 10 | 0 |
| Espana | Nigeria | 3 | 0 |
| Espana | Italia | 0 | 0 |
| Espana | Brasil | 0 | 3 |
| Espana | Finlandia | 2 | 0 |
| Espana | Bielorrusia | 2 | 1 |
| Espana | Georgia | 2 | 0 |
| Espana | Italia | 1 | 0 |
| Espana | Bolivia | 2 | 0 |
| España | Salvador | 2 | 0 |

| Equipo 1 | Equipo 2 | G1 | G2 |
|---|---|---|---|
| Holanda | Turquia | 2 | 0 |
| Holanda | Hungria | 4 | 1 |
| Holanda | Andorra | 3 | 0 |
| Holanda | Rumania | 4 | 1 |
| Holanda | Estonia | 3 | 0 |
| Holanda | Rumania | 4 | 0 |
| Holanda | Estonia | 2 | 2 |
| Holanda | Andorra | 2 | 0 |
| Holanda | Hungria | 8 | 1 |
| Holanda | Turquia | 2 | 0 |
| Holanda | Francia | 0 | 2 |
| Holanda | Ecuador | 1 | 1 |
| Holanda | Ghana | 1 | 0 |
| Holanda | Gales | 2 | 0 |

Luego se calculó la frecuencia con la que Holanda ganó a sus rivales. La fórmula para calcular la frecuencia es*:*

$$f(H = Gana) = f(H = Gana, posProm) * 0.8 + f(H = Gana, 100)$$

Donde *f(H=Gana,posProm)* es una función que cuenta los partidos en que ganó Holanda si y solo si la posición en el ranking FIFA de sus rivales es menor o igual a *posProm*. La función *f(H=Gana,100)* es una función que cuenta los partidos en que ganó Holanda si y solo si la posición en el ranking FIFA de sus rivales es menor o igual a 100. Puede notarse que los partidos ganados a rivales cuya posición en el ranking es menor o igual a *posProm* tienen un peso adicional de 0.8. Lo que da más mérito cuando vence a rivales complicados según el ranking FIFA. Vale resaltar que el modelo considera un partido como victoria si al menos hay dos goles de diferencia, esto produce un *bias* que tiende a beneficiar a equipos con goleadas a rivales considerados como fuertes.

Asimismo, la fórmula para el cálculo de la frecuencia de las derrotas de Holanda es:

$$f(H = Pierde) = f(H = Pierde, posProm) + f(H = Pierde, 100) * 0.3$$

Donde *f(H=Pierde,posProm)* es una función que cuenta los partidos en que perdió Holanda si y solo si la posición en el ranking FIFA de sus rivales es menor o igual a *posProm*. La función *f(H=Gana,100)* es una función que cuenta los partidos en que perdió Holanda si y solo si la posición en el ranking FIFA de sus rivales es menor o igual a 100. En este caso, los partidos perdidos con rivales débiles tienen un peso adicional de 0.3. Finalmente, se calculó la probabilidad marginal *P(H=Gana) y P(H=Pierde):*

$$P(H = Gana) = \frac{f(H = Gana) + \alpha}{f(H = Gana) + f(H = Pierde) + 2 * \alpha}$$

$$P(H = Pierde) = \frac{f(H = Pierde) + \alpha}{f(H = Gana) + f(H = Pierde) + 2 * \alpha}$$

Donde P*(H=Gana)* y *P(H=Pierde)* son las probabilidades de que Holanda pueda ganar y perder cualquier partido, respectivamente. El valor α representa una constante cuyo valor es 100 y es conocido como hiperparámetro (ver Koller y Friedman, 2009).

Si se toma en cuenta los datos de la **Tabla 1**, se tiene *P(H=Pierde) = 0.492891 y P(H=Gana) = 0.507109.* En el caso de España, es *P(E=Pierde) = 0.5* y *P(E=Gana) = 0.5*.

Con el cálculo de las distribuciones *P(H) y P(E),* se halló la distribución de la variable aleatoria P ("Holanda gana a España"). Para calcular la probabilidad marginal de esta última distribución, se usaron las siguientes fórmulas:

$$P(P = Holanda) = \frac{P(H = Gana) * 0.5}{P(E = Gana) + P(H = Gana)}$$

$$P(P = España) = \frac{P(E = Gana) * 0.5}{P(E = Gana) + P(H = Gana)}$$

Donde P*(P=Holanda)* indica la probabilidad que Holanda gane a España y *P(P=España)* indica la probabilidad que España gane a Holanda.

**Tabla 2.**
*En la columna izquierda, se muestran las distribuciones de probabilidades de las 3 variables aleatorias del modelo probabilístico propuesto. En la columna derecha, se muestra la probabilidad conjunta que es el producto de las 3 distribuciones de la columna izquierda.*

**España (Gana)**

| 1 | Probabilidad |
|---|---|
| Pierde | 0,500000 |
| Gana | 0,500000 |

**Holanda (Gana)**

| 2 | Probabilidad |
|---|---|
| Pierde | 0,492891 |
| Gana | 0,507109 |

**Factor España le gana a Holanda**

| 3 | Probabilidad |
|---|---|
| No | 0,503529 |
| Sí | 0,496471 |

**Probabilidad Conjunta**

| 1 (España) | 2 (Holanda) | 3 (España gana a Holanda) | Probabilidad |
|---|---|---|---|
| Pierde | Pierde | No | 0,124093 |
| Gana | Pierde | No | 0,124093 |
| **Pierde** | **Gana** | **No** | **0,127672** |
| **Gana** | **Gana** | **No** | **0,127672** |
| Pierde | Pierde | Sí | 0,122353 |
| Gana | Pierde | Sí | 0,122353 |
| Pierde | Gana | Sí | 0,125882 |
| Gana | Gana | Sí | 0,125882 |

La distribución de la tercera variable aleatoria es: *P(P=España)= 0,496471* y *P(P=Holanda)= 0,503529*. En síntesis, las distribuciones de las tres variables aleatorias son mostradas en la **Tabla 2**. Una vez creadas las tres variables aleatorias, según el modelo, es necesario calcular la probabilidad conjunta a partir de la cual se pudó estimar qué equipo tendrá mayor probabilidad de ganar el partido. La probabilidad conjunta, en este modelo, se calcula de la siguiente manera:

$$P(H,E,P) = P(H) * P(E) * P(P)$$

Donde *P(H,E,P)* es la probabilidad conjunta. De acuerdo a la notación presentada y según la regla de la cadena, se dedujo que las tres variables aleatorias son independientes entre sí (ver **Koller y Friedman (2009)**).

En la distribución conjunta, el objetivo es hallar el *Maximum a Posteriori (MAP)* el cual consiste en seleccionar la mayor probabilidad (ver **Figura 2**).

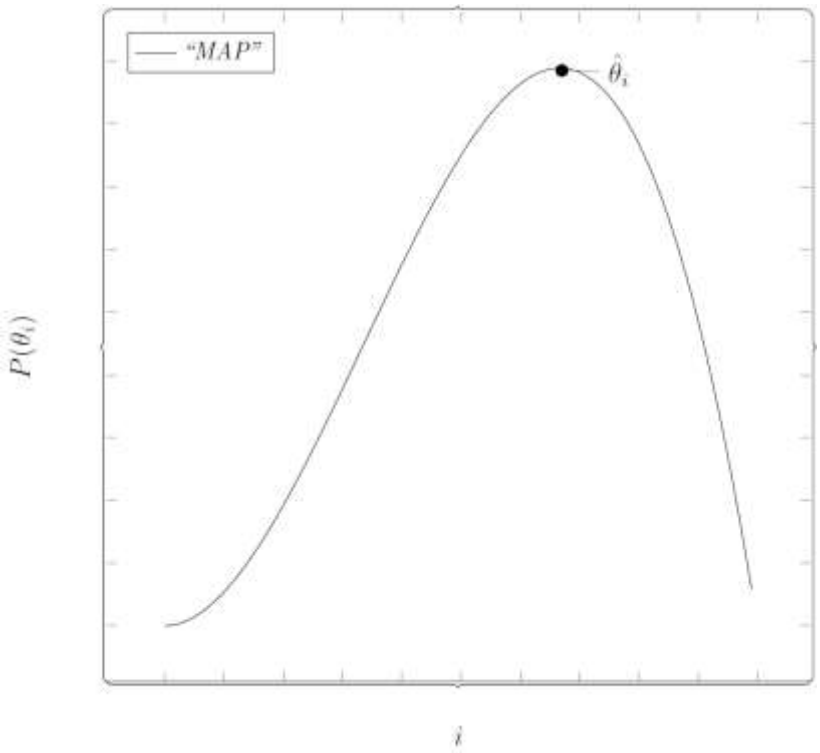



**Figura 2:** Se presenta una distribución donde el *Maximum a Posteriori (MAP)* es el mayor valor, es decir el valor que está en el pico de dicha distribución. En este caso, el pico es denotado por el valor $\theta_i$, donde *i* es el índice del valor dentro de la distribución y *P*($\theta_i$) es la probabilidad en dicha posición *i*.

En la **Tabla 2**, puede notarse que para el partido "Holanda versus España", el MAP es favorable a Holanda con una probabilidad de 0,127672.

Como puede notarse, la construcción del modelo probabilístico propuesto en este artículo es simple y apenas presenta tres variables aleatorias. Cabe recordar que el objetivo de este artículo es mostrar la potencialidad de esta herramienta de Big Data que permite analizar datos históricos a fin de predecir resultados futuros. En ese contexto, aunque el modelo es simple, presenta grandes ventajas antes las limitaciones tales como la

complejidad del fútbol y la poca información disponible de partidos entre selecciones. De hecho, no es posible considerar todos los partidos desde el siglo pasado porque los equipos de fútbol se van renovando periódicamente y eso cambia el rendimiento del equipo. A continuación, se describe la información que fue utilizada para producir este modelo probabilístico.

### 2.2. CONJUNTO DE DATOS

En el presente trabajo, los datos a ser usados corresponden a los partidos oficiales disputados por cada una de las selecciones participantes del mundial de fútbol Brasil 2014. Tales partidos han sido jugados entre el 1 de enero del año 2011 hasta el 11 de junio del 2014 (un día antes del mundial).

El conjunto de datos está dividido en dos partes. La primera parte consiste en los nombres de los 32 países participantes en el mundial Brasil 2014 con sus respectivas posiciones en el Ranking FIFA actualizado hasta junio de 2014 (antes del mundial) (**Wikipedia, 2014b**). La segunda parte consiste en el resultado de los encuentros oficiales disputados para cada una de las 32 selecciones participantes. La fecha de dichos encuentros está comprendida entre los años 2011 y 2014 (antes del inicio del mundial).

### 2.3. APLICACIÓN PARA CONTRASTE DE DATOS

Como parte de este trabajo, se implementó una aplicación a fin de registrar los resultados oficiales de los partidos del mundial. Información usada para contrastar los resultados predichos por el modelo propuesto con los resultados reales.


**Figura 3:** *Interfaz de usuario para el ingreso de los países dentro de la fase de grupo. Esta interfaz permite hacer la simulación de otros posibles grupos diferentes a los del mundial a fin de medir otros equipos como rivales de primera fase.*



La aplicación también puede hacer predicciones sobre grupos ficticios, reagrupando las selecciones a fin de realizar tales simulaciones. En la **Figura 3** se observa el ingreso de los equipos a los grupos de competencia, los cuales suman 8 grupos en total. En la **Figura 4**, se

oberva el ingreso de los resultados dados por el usuario. Estos resultados serán contrastados con los obtenidos por nuestro modelo predictor.

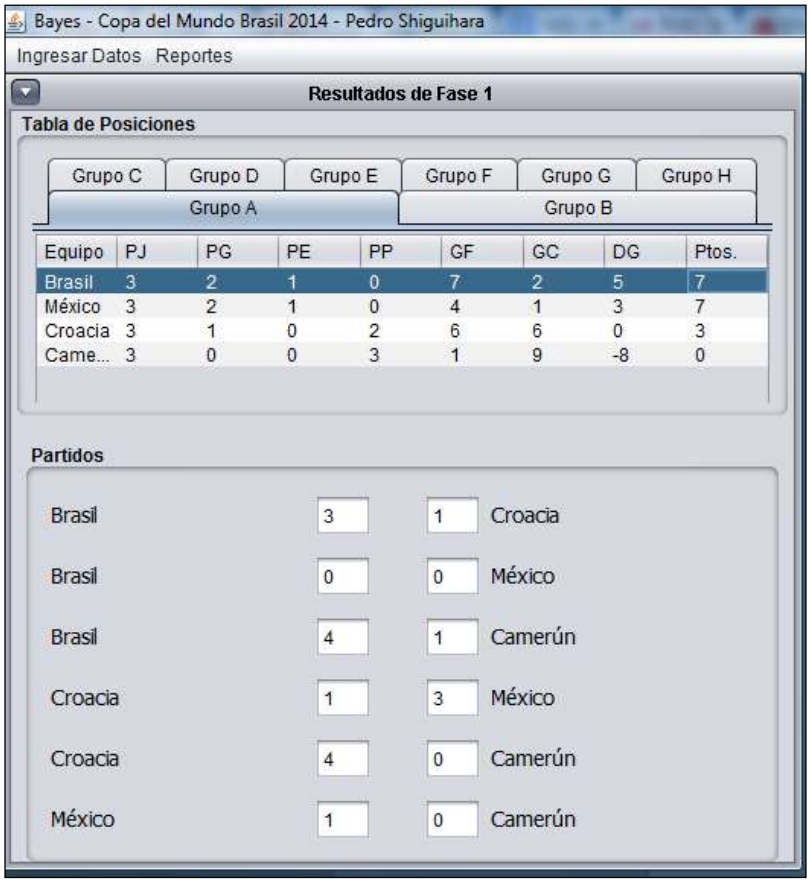



**Figura 4:** *Interfaz de usuario para el ingreso resultados de los grupos formados a partir de la interfaz anterior (ver **Figura 3**).*

En la siguiente sección se presentan los resultados al aplicar el modelo predictor en la primera fase de grupos en el mundial Brasil 2014.

## 3. RESULTADOS

Al momento de publicarse este artículo, los partidos de primera fase del mundial de fútbol FIFA Brasil 2014 recientemente habían terminado. Los resultados sugieren que la tasa de aciertos es considerable. Por tasa de aciertos, referidos al número de predicciones que coincidieron con los resultados reales versus el total de partidos evaluados.

El modelo fue comparado con los modelos propuestos en Florez (2014) y Bloomberg (2014). Este último es un portal web que realiza predicciones deportivas. Dado que tanto en Florez (2014) como en nuestro modelo, la predicción determina siempre ganadores, fue necesario evaluar los empates desde dos puntos de vista. En el primer punto de vista, un empate era considerado como una predicción errada, indicando que el empate no favorecía a los predictores. En la **Figura 5** se muestran los resultados obtenidos al comparar los tres predictores. El predictor propuesto en Bloomberg (2014) obtuvo la menor tasa de predicción con 60.42%, mientras que la red neuronal propuesta en Florez (2014) alcanzó el 64.58%. Por otro lado, la tasa de aciertos de nuestro modelo alcanzó el 66.67%. Eso quiere decir que el modelo probabilístico propuesto en este artículo a pesar de su simplicidad obtuvo 2.09% más aciertos que la red neuronal propuesta y 6.25% más que el predictor propuesto en Bloomberg (2014). Si se considera que los empates juegan a favor de todos los predictores, entonces se podrá observar que la tasa de aciertos aumenta substancialmente. En la **Figura**

**6** se muestran los resultados considerando los empates a favor de los tres predictores. En este caso, se observa que la tasa de aciertos aumenta en las tres técnicas, llegando a obtener más del 80% de aciertos. Además, al hacer esto, se observa algo particular todos los predictores aumentan su desempeño en un 18.75%, indicando que los tres predictores habían realizado exactamente las mismas predicciones en esos empates. Eso hace notar que los partidos de fútbol tienen una alta tasa de incerteza y los empates contribuyen a tal incertidumbre.

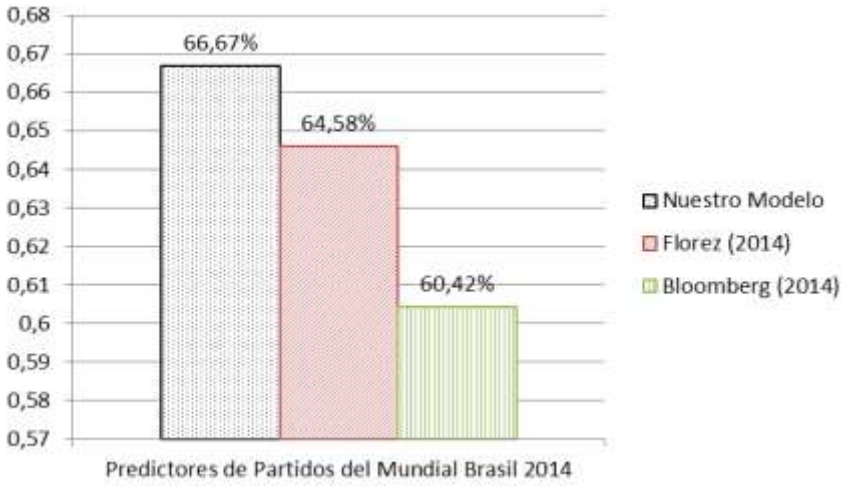



**Figura 5:** *Comparación de tres predictores usados para estimar los ganadores en la fase de grupos (primera fase) del Mundial FIFA Brasil 2014.*

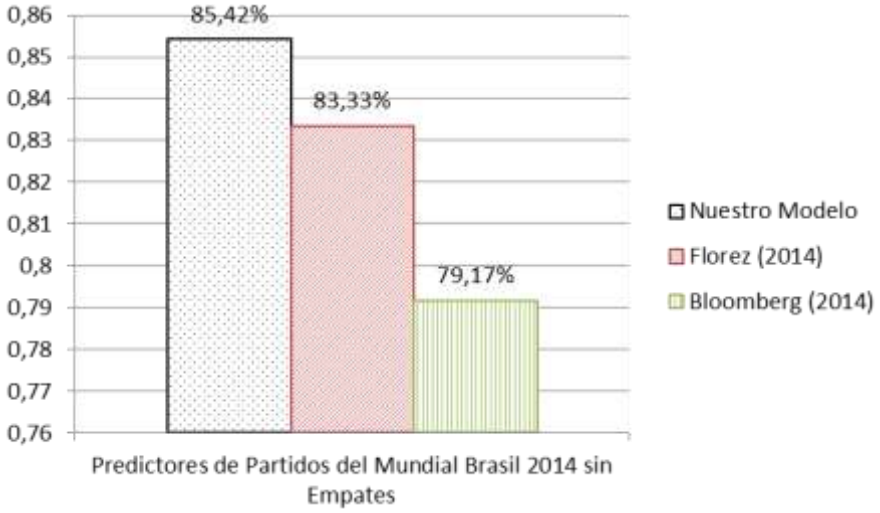
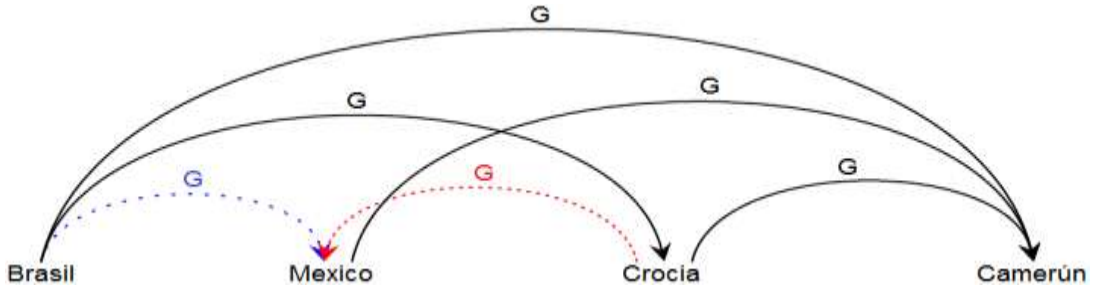


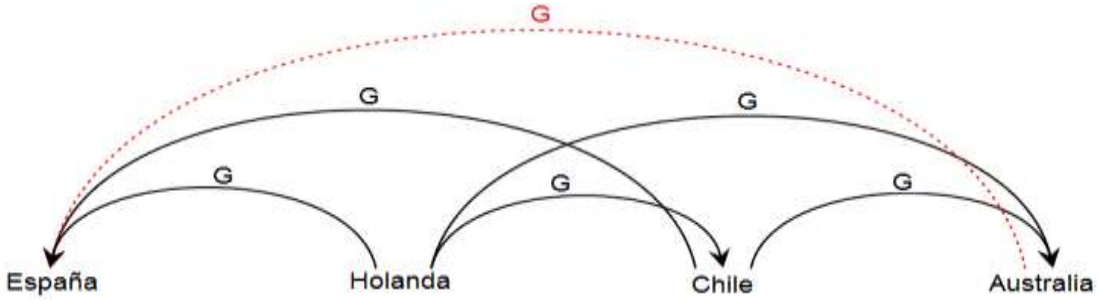

**Figura 6:** *Comparación de tres predictores usados para estimar los ganadores en la fase de grupos (primera fase) del Mundial FIFA Brasil 2014. En este caso, los empates dejan de ser considerados como una predicción errónea.*

De manera general, puede notarse que el modelo probabilístico obtuvo ligeramente una mejor tasa de aciertos en comparación con los otros dos predictores. En la **Figura 7**, son mostrados los aciertos y desaciertos para los partidos del grupo A, B, C y D del mundial. Puede notarse que al considerar los empates acertadamente, el número de aciertos sería mucho mayor al 60%. En la **Tabla 3** son presentadas las predicciones en detalle de nuestra propuesta y son comparados con trabajos previos. Comparando nuestro modelo probabilístico con la red neuronal, ambos obtienen resultados muy parecidos, incluso con una tasa de aciertos regular, lo cual no ocurre con el predictor de **Bloomberg (2014)**.

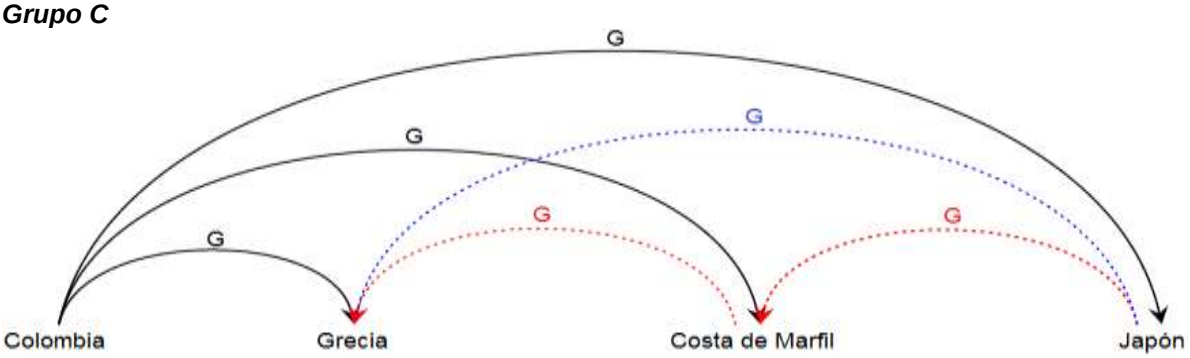
***Grupo A***




***Grupo B***




***Grupo C***




***Grupo D***

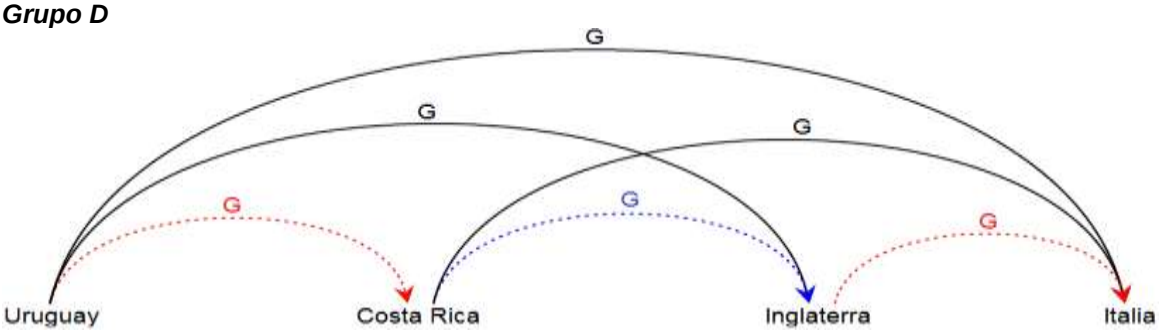



**Figura 7:** *Resultados de predicciones en los grupos A, B, C y D del mundial FIFA Brasil 2014 son presentados. Las aristas dirigidas indican: Ganador → Perdedor. Las líneas oscuras representan los aciertos del modelo probabilístico propuesto mientras que las líneas punteadas representan resultados en los que no acertó. Las líneas punteadas de color azul representan que no acertó y el resultado real fue un empate, mientras que las líneas rojas indican que no acertó y el resultado real fue lo contario a la predicción.*

**Tabla 3.**
*Resultados de las predicciones realizadas por los predictores propuestos en **Florez, (2014), Bloomberg (2014)** y en el presente trabajo. Tales predicciones son contrastadas con los resultados reales de los partidos. Los casilleros en blanco indican que los partidos no fueron realizados aun al momento de la presentación de esta publicación.*

| Local | Visitante | Ganador Real | Florez (2014) | Bloomberg (2014) | Nuestro Modelo |
|---|---|---|---|---|---|
| Brasil | Croacia | Brasil | Brasil | Brasil | Brasil |
| México | Camerun | México | México | México | México |
| Brasil | México | Empate | Brasil | Brasil | Brasil |
| Camerun | Croacia | Croacia | Croacia | Croacia | Croacia |
| Camerún | Brasil | Brasil | Brasil | Brasil | Brasil |
| Croacia | México | México | Croacia | México | Croacia |
| España | Holanda | Holanda | España | España | Holanda |
| Chile | Australia | Chile | Chile | Chile | Chile |
| Australia | Holanda | Holanda | Holanda | Holanda | Holanda |
| España | Chile | Chile | España | España | Chile |
| Australia | España | España | España | España | Australia |
| Holanda | Chile | Holanda | Holanda | Holanda | Holanda |
| Colombia | Grecia | Colombia | Colombia | Colombia | Colombia |
| Costa de Marfil | Japón | Costa de Marfil | Costa de Marfil | Japón | Japón |
| Colombia | Costa de Marfil | Colombia | Colombia | Colombia | Colombia |
| Japón | Grecia | Empate | Grecia | Grecia | Japón |
| Japón | Colombia | Colombia | Colombia | Colombia | Colombia |
| Grecia | Costa de Marfil | Grecia | Costa de Marfil | Grecia | Costa de Marfil |
| Uruguay | Costa Rica | Costa Rica | Uruguay | Uruguay | Uruguay |
| Inglaterra | Italia | Italia | Italia | Inglaterra | Inglaterra |
| Uruguay | Inglaterra | Uruguay | Uruguay | Inglaterra | Uruguay |
| Italia | Costa Rica | Costa Rica | Italia | Italia | Costa Rica |
| Italia | Uruguay | Uruguay | Uruguay | Italia | Uruguay |
| Costa Rica | Inglaterra | Empate | Costa Rica | Inglaterra | Inglaterra |
| Suiza | Ecuador | Suiza | Suiza | Suiza | Suiza |
| Francia | Honduras | Francia | Francia | Francia | Francia |
| Suiza | Francia | Francia | Francia | Francia | Francia |
| Honduras | Ecuador | Ecuador | Ecuador | Ecuador | Ecuador |
| Honduras | Suiza | Suiza | Suiza | Suiza | Suiza |
| Ecuador | Francia | Empate | Francia | Francia | Francia |
| Argentina | Bosnia | Argentina | Argentina | Argentina | Argentina |
| Irán | Nigeria | Empate | Nigeria | Nigeria | Irán |
| Argentina | Irán | Argentina | Argentina | Argentina | Argentina |
| Nigeria | Bosnia | Nigeria | Bosnia | Empate | Bosnia |
| Nigeria | Argentina | Argentina | Argentina | Argentina | Argentina |
| Bosnia | Irán | Bosnia | Bosnia | Bosnia | Bosnia |
| Alemania | Portugal | Alemania | Alemania | Alemania | Alemania |
| Ghana | Estados Unidos | Estados Unidos | Estados Unidos | Estados Unidos | Estados Unidos |
| Alemania | Ghana | Empate | Alemania | Alemania | Alemania |
| Estados Unidos | Portugal | Empate | Portugal | Portugal | Portugal |

| | | | | | |
|---|---|---|---|---|---|
| Estados Unidos | Alemania | Alemania | Alemania | Alemania | Alemania |
| Portugal | Ghana | Portugal | Portugal | Portugal | Portugal |
| Bélgica | Argelia | Bélgica | Bélgica | Bélgica | Bélgica |
| Rusia | Corea del Sur | Empate | Rusia | Rusia | Rusia |
| Bélgica | Rusia | Bélgica | Bélgica | Bélgica | Bélgica |
| Corea del Sur | Argelia | Argelia | Corea | Corea | Argelia |
| Corea del Sur | Bélgica | Bélgica | Bélgica | Bélgica | Bélgica |
| Argelia | Rusia | Empate | Rusia | Rusia | Argelia |

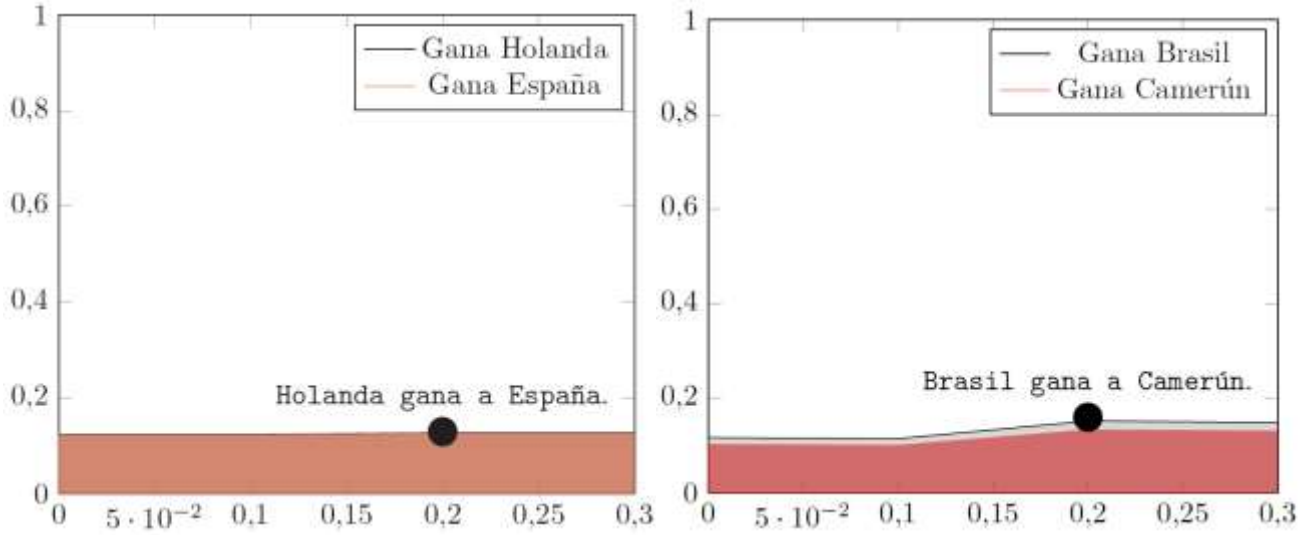



**Figura 8:** *A la izquierda, la distribución conjunta para el partido entre Holanda y España. A la derecha, la distribución conjunta para el partido Brasil y Camerún. En el partido Holanda versus España, el ganador es Holanda por un corto margen de diferencia. En contraste, en el partido de Brasil versus Camerún, el margen de diferencia favoreciendo a Brasil es mayor.*

Las predicciones proporcionadas por nuestro modelo se basan en el cálculo MAP a partir de la distribución conjunta.

En la **Figura 8**, se muestran las distribuciones conjuntas de dos partidos: Holanda versus España y Brasil versus Camerún. En el primer partido, puede notarse que la distribución conjunta que indica a Holanda como ganador tiene valores muy cercanos a la otra parte de esa distribución que indica a España como ganador. Esto resulta ser coherente ya que tanto Holanda como España han demostrado ser equipos muy competitivos en los últimos años. Por otro lado, en el segundo partido, al observar la distribución conjunta que da a Brasil por ganador, ésta se encuentra resaltantemente superior a la parte de esa distribución conjunta que da por ganador a Camerún. Eso resulta ser obvio para algunos expertos en fútbol ya que Camerún es un equipo débil en contraste con Brasil. Por tanto, el análisis MAP de las distribuciones conjuntas es coherente de alguna manera con las realidades que afrontan cada selección. Esto fue posible al incluir únicamente partidos de los últimos tres años para cada una de las selecciones participantes en el mundial Brasil 2014.

## 4. CONCLUSIONES

El modelo probabilístico propuesto fue diseñado para estimar los partidos de primera fase del mundial, obteniendo un desempeño competitivo en comparación con técnicas de redes neuronales y análisis estadístico propuestos por **Florez (2014)** y **Bloomberg (2014)**. Si bien los resultados de nuestro modelo probabilístico fueron ligeramente superiores con las redes neuronales, ambas técnicas demostraron ser importantes en el análisis de datos a partir de información histórica.

La información histórica analizada, para el modelo probabilístico, fueron los partidos oficiales de las 32 selecciones participantes en los últimos 3 años. Aunque esta información es escasa para un modelo probabilístico, este modelo demostró un desempeño coherente con los resultados finales, obteniendo un 67% de efectividad si se consideran los empates como predicciones erróneas y 85% si los consideramos a favor de nuestras predicciones. Un dato interesante con relación a los empates es que las 3 técnicas comparadas estimaron al mismo ganador.

Finalmente, se resalta, aunque en el fútbol hay una alta incertidumbre con relación a los resultados de los partidos, los modelos de redes neuronales y nuestro modelo probabilístico obtuvieron resultados superiores al 80%. Esto sugiere que en otros entornos de alta incertidumbre, estos modelos también podrían ser usados para poder ayudar a la toma de decisiones, principalmente en entidades gubernamentales y empresas. El apoyo en toma de decisiones permite salvar vida y ahorrar tiempo y dinero en la ejecución de tareas. Así, las herramientas de Big Data son actualmente importante dada la variada información que se maneja hoy en día digitalmente para mejorar nuestra vida y economía.

## 5. REFERENCIAS

ACM, J. (2011). For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning. From: http://amturing.acm.org/award_winners/pearl_2658896.cfm

Berman, J. (2013). Principles of BigData: Preparing, Sharing, and Analyzing Complex Information. Ed. Elsevier Inc.

Bloomberg (2014). http://www.bloomberg.com/visual-data/world-cup/#0,0,-1. Fecha de Acceso: 26 de Junio del 2014.

Ehab, Z. (2014). Math equation could help find missing Malaysian plane. From: http://america.aljazeera.com/articles/2014/3/12/mathematical-equationcouldhelpfindmissingmalaysianplane.html

Florez, O. (2014). Predicting the Results of the 2014 FIFA World Cup. https://medium.com/@ouflorez/predicting-the-results-of-the-2014-fifa-world-cup-b8d313138f48. Fecha de Acceso: 26 de Junio del 2014.

Gopalkrishnan, V., Steier, D., Lewis, H. & Guszcza, J. (2012). Big Data, Big Business: Bridging the Gap

Hsinchun, C., Chiang, R. & Storey, V. (2012). Business intelligence and analytics: from big data to big impact. From: http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf

Intel (2014). Internet Minute Infrographics. Fecha de Acceso: 26 de Junio del 2014.

http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html.

Koller, D. & Friedman , N. (2009). Probabilistic Graphical Moldels: Principles And Techniques

Lohr, S. (2012). The Age of Big Data. From: The New York Times.

Russell, S. & Norvig, P. (2010). Artificial Intelligence: A Modern Approach (3ra Edición). Ed. Pearson Educación.

Shiguihara, P. & Andrade, A. (2013). Learning bayesian network using parse trees for extraction of protein-protein interaction. (University of São Paulo, Brazil) From: http://dl.acm.org/citation.cfm?id=2458340

Wikipedia, (2014a). http://es.wikipedia.org/wiki/Anexo:Pa%C3%ADses_afiliados_a_la_FIFA. Fecha de Acceso: 26 de Junio del 2014.

Wikipedia, (2014b). http://es.wikipedia.org/wiki/Copa_Mundial_de_F%C3%BAtbol. Fecha de Acceso: 26 de Junio del 2014.