REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Partidos de Fútbol, ¿Suerte o Causalidad? Introducción a los Modelos Estadísticos de Conteo y RNA para la Predicción de Juegos utilizando la Liga de Fútbol de Chile

Paulo H. Ferreira[1], Leonardo Barrios-Blanco[2], Hugo R. Santana[3], Wilson A. Castillo-Rojas[2], Adriano Suzuki[4], and Francisco Louzada[4]

[1]Universidad Federal de Bahia, Brasil; *paulohenri@ufba.br*
[2]Universidad de Atacama, Chile; *leonardo.barrios.2020@alumnos.uda.cl, wilson.castillo@uda.cl*
[3]Universidad Federal de Rio de Janeiro, Brasil; *hugo.nabity@hotmail.com*
[4]Universidade de Sao Paulo, Brasil; *suzuki@icmc.usp.br, louzada@icmc.usp.br*

RESUMEN

La estocasticidad es un elemento inexorable y frecuente en nuestras vidas. Sin embargo, a la previsibilidad deportiva le importa la aleatoriedad en el rendimiento, aunque, como cualquier otro artefacto humano, también son de esperar patrones ocultos. De esta manera, se puede ver un partido de fútbol a través de la metodología estadística y de aprendizaje automático para desentrañar esos patrones. Este trabajo tuvo como objetivo aplicar técnicas comunes (regresión contable y red neuronal artificial) para explorar la viabilidad de predecir la Premier Soccer League chilena 2020. El enfoque estadístico se centró en la previsibilidad de los goles en cada partido, durante los últimos 5 partidos de la ronda, y luego calculó la probabilidad del resultado (ganador del equipo local, empate o victoria del equipo visitante). Posteriormente, el Brier Score midió la idoneidad del modelo de regresión de Poisson *versus* el modelo de regresión de Bell, y también comparó su desempeño en la predicción del resultado del partido con el modelo ajustado de la Red Neural Artificial (ANN). Los resultados mostraron una argumentación positiva hacia esa metodología probabilística (ya sea estadística o de aprendizaje automático) aplicable en la Liga de Fútbol de Chile y su previsibilidad. La regresión de Poisson predijo bien el resultado del juego en un 60% de las veces, de media, la regresión de Bell en un 62%, y tanto ANN-RNR como ANN-MLP en un 51%.

Palabras claves: Matemáticas aplicadas en deportes, redes neuronales artificiales, regresión contable, ciencia de datos en deportes.

ABSTRACT

Stochasticity is an inexorable element and frequent in our lives. Notwithstanding, sports predictability cares randomness in the performance, though like any other human artifact, hidden patterns are also to be expected. In this manner, a soccer match can be seen through the statistical and machine learning methodology towards unraveling those patterns. This work aimed to apply common techniques (countable regression and artificial neural network) to explore the feasibility of predicting the 2020 Chilean Premier Soccer League. The statistical approach targeted the predictability of the goals in each match, for the last 5 round games, then calculated the probability of the outcome (winning home team, drawing, or winning visiting team). Later, the Brier Score measured the suitability of the Poisson regression model *versus* the Bell regression model, and also compared their performance of predicting the match outcome against the Artificial Neural Network (ANN) adjusted model. Results showed positive argumentation towards that probabilistic methodology (either statistical or machine learning) applicable in the Chilean Soccer League and its predictability. The Poisson regression predicted well the game's outcome in 60% of the time, on average, the Bell regression in 62%, and both ANN-RNR and ANN-MLP in 51%.

Keywords: Applied Mathematics in Sports, Artificial Neural Network, Countable Regression, Data Science in Sports.

1 Introduccion

La aleatoriedad, por mucho tiempo, fue una área de investigación descuidada por la matemática y no apreciada por las ciencias sin tener muchas explicaciones al respecto. Por muchos siglos, en la historia de la humanidad, se relacionó previsiones u ocurrencias de un fenómeno futuro como una cuestión mística (divina) y no mensurable. Sin embargo, los humanos desde mucho tiempo se interesan por "juegos de suerte", por ejemplo, las tabas (juego antiguo que tira astrágalos, huesos de las patas de "ovicápridos"), hasta llegar hoy día a usar herramientas tecnológicas para hacer apuestas en vivo, de juegos que se dan a miles de kilómetros del lugar de donde se da el clic para la apuesta.

El primer relato sobre la palabra posibilidad fue registrada del italiano Gerolamo Cardano en su libro *Liber de ludo aleae* ("Book on Games of Chance"), escrito en 1564, describiendo la relación entre ocurrencia/frecuencia de un evento contra los otros resultados. Posteriormente, Chevalier de Méré contactando dos matemáticos Blaise Pascal y Pierre de Fermat con indagaciones también sobre medidas relacionadas con juegos y apuestas, extendieron algunas aplicaciones de la probabilidad. Pero las ideas sobre probabilidad como una medida formal fue esclarecida mucho tiempo después, solo en el siglo XIX, a modo de ejemplo, Bruno de Finetti, Frank Ramsey y Andrey Kolmogorov definieron con solidez matemática eventos de interés, relacionando a una medida de incertidumbre eventos no-observados o futuros, lo que dio una base sólida y un gran avance para la predicción de eventos.

En la búsqueda de modelar la incertidumbre de un evento, aparecen los modelos de Inteligencia Artificial (IA) que tienen en su base a modelos probabilísticos, asociando la ocurrencia de un evento como una incertidumbre, siendo un número mensurable que pertenezca al intervalo $[0, 1]$. Una sub-área dentro de la IA, conocida como Aprendizaje de Máquinas (*Machine Learning*), permite ejecutar tareas de clasificación o regresión (supervisadas), agrupamiento/anotación (no-supervisadas), aprendizaje reforzado, entre otros, de modelos matemático/estadístico (vía inferencia) como intento de extraer un patrón de un proceso. Estos modelos son un intento de explicar una variable aleatoria, basados en registros históricos y suposiciones, asociando una medida de probabilidad a un evento, buscando generalizar los resultados obtenidos de una muestra para toda la población.

Debido a la gran cantidad de información que maneja el mundo del deporte [1], los registros deportivos son recolectados para poder auxiliar predicciones de desempeño atlético y asociar una medida a posibles resultados [2].

Actualmente, son variados los modelos estadísticos que permiten estimar lo que puede suceder en algún juego deportivo. Por ejemplo para el Béisbol, se han planteado dos modelos predictivos usando métodos del aprendizaje automático [3]. La predicción de resultados y predecir el desempeño de los lanzadores abridores, son ejemplos de tareas que se realizan a través de modelos de IA. De igual manera, se pueden mencionar características que los equipos de baloncesto de la *National Basketball Association* (NBA), que lograron mejorar su estrategia para ganar cada partido [4], también en el ámbito de los juegos de fútbol americano de la *National Football League* (NFL), obtienen sus predicciones, con base en los datos que recolectan [5], tanto de los propios partidos como de los aspectos ambientales que pueden influir en el partido.

El fútbol también forma parte de los deportes que generan millones de datos diariamente en el mundo, debido a que en este deporte se registra desde lo principal, que son los goles, hasta la cantidad de pases errado de cada jugador, lo que genera para todos los equipos las ventajas o desventajas al momento de preparar el próximo encuentro [6, 7]. Este registro de eventos tiene desde principios del siglo XX, cuando un contador fue pionero en la recolección de datos en los partidos de fútbol, logrando registrar las pérdidas de balón, lo cual más adelante sirvió para mejorar las técnicas de posesión de balón y así poder evitar goles en su propia área [8].

Hoy se mide con más detalles cada uno de los momentos del fútbol gracias a la tecnología, y se pueden emplear técnicas contundentes para realizar un mejor análisis de juego, que luego servirán para mejorar las debilidades de los equipos. Esta gran cantidad de datos sirve para poder predecir el ganador de un juego, o mejor aún el marcador final. Santana *et al.* [9] adoptaron modelos de aprendizaje de máquinas para estimar

el posible ganador, perdedor o empate en un juego, considerando las variables como dependiente (*target variable*). Otro tipo de modelo de previsión que ha sido estudiado, está relacionado al número de goles marcados por los equipos [10]. Además, Pedroza [11] analizó la Liga MX del fútbol Mexicano con una propuesta de un modelo de inferencia basado en lógica difusa, para obtener un pronóstico para cada partido, para clasificar por posición predeterminada para jugar en el campo de fútbol aquellas habilidades que deben estar presentes en un jugador.

Con esta motivación, este trabajo analiza el desempeño de tres modelos probabilísticos para previsiones del fútbol de la Primera División de la liga Chilena del año 2020. Adoptando dos modelos estadísticos de distribuciones discretas, que están asociadas a variables de conteo, y posibilitan la predicción de número de goles de cada partido. Adicionalmente, se adoptan dos modelos de Redes Neuronales Artificiales (RNAs). Las RNAs corresponden a modelos computacionales inspirados en el funcionamiento de la red neuronal del cerebro humano, donde la unidad básica de estos modelos es la neurona [12]. Una de las características de las RNAs, es su gran capacidad predictiva que para este trabajo permite inferir el resultado del equipo ganador de un conjunto de partidos, basados en un conjunto de datos históricos de entrenamiento para el modelo.

2 Métodos

La estrategia de predicción de resultados futbolístico de este trabajo está basado en la predicción de victoria, empate y derrota, adoptando dos metodologías (estadística y aprendizaje de máquina). La metodología estadística, primero busca, predecir él numero de goles de cada partido considerando la distinción de ser equipo local o visitante, y en segundo lugar, generar modelos de regresión para datos de conteo [13], y así cuantificar la probabilidad de la tripleta (victoria, empate y derrota). Esta metodología es interesante desde el punto de múltiples tareas logradas de una sola vez como: i) marcador de cada juego individual, ii) probabilidad de victoria de cada partido, y iii) previsión de la puntuación del campeonato (posición de cada equipo).

Además, se adopta la metodología de aprendizaje de máquinas RNA, en particular un modelo de arquitectura *Multi-Layer Perceptron (MLP), y un segundo modelo deep learning*, bajo la arquitectura de una Red Neuronal Recurrente (RNR), ambos modelos predicen directamente el resultado de la tripleta (victoria, empate o derrota) por partido.

2.1 Modelo de Puntuación por Equipo

Lee [13] discute que una manera de calcular la probabilidad de que un equipo de fútbol gane un campeonato podrá ser basada en distintos (independientes) modelos que calculan el resultado de un juego entre dos equipos. Más detalladamente, cada equipo tiene su desempeño (un poder de ataque jugando en casa o como visitante, así como defensa en casa o como visitante), y es posible combinar los desempeños esperados (distribuciones marginales), a fin de obtener previsiones de cada partido como una distribución conjunta.

Para esto, se usará el método de regresión con 4 parámetros por equipo, como una estructura para predecir el número de goles por equipo (en cada partido). Así, cada juego será un evento bivariado (par ordenado), $\mathbf{Y} = (Y_1, Y_2)$, que expresa el resultado del enfrentamiento entre un equipo que juega en casa (Y_1) *versus* un equipo visitante (Y_2), y podrán ser independientes. Es decir, los goles de cada equipo son independientes, por lo tanto las variables de estudio de los goles (desempeño de los poderes de ataque y defensa) de cada equipo se consideran independientes entre sí.

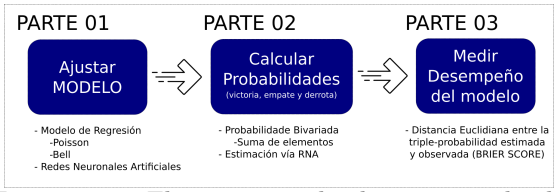

Figure 1: Flujo-grama de los pasos de la metodología adoptada.

La estructura explicativa del número de goles obtenido en un partido podrá considerar como covariables (*features*), por ejemplo, cual equipo juega en casa y cual es visitante. El trabajo se divide en tres partes fundamentales, conforme ilustra la Fig. 1, y que se describen a continuación:

Parte 1: Estimar el desempeño de cada equipo como densidad de probabilidad.

Para realizar las predicciones del número de goles que se podrán anotar en un partido o las cantidades de goles que hace un equipo (denominada una variable aleatoria Y), se usará la distribución de discreta/conteo (por ejemplo, Poisson o Bell), asociando el desempeño de cada equipo a una función de probabilidad de marcar goles.

Para modelar ese desempeño como una variable aleatoria, el parámetro de un modelo puede ser logrado tomando en cuenta una estructura explicativa; por ejemplo, si los números de goles marcados por un equipo, está asociado al desempeño en cada juego como visitante o local. Mientras esa estructura de regresión pueda estimar el parámetro que modela el desempeño de cada equipo.

Al considerar los desempeños de cada i-ésimo equipo como variables independientes (número de goles marcados vía una medida de probabilidad, $\mathrm{P}(Y_i)$), el enfrentamiento de cada juego podrá ser calculado como un producto de variables independientes involucradas ($\mathrm{P}(Y_i,Y_j) = \mathrm{P}(Y_i) \times \mathrm{P}(Y_j)$, para $i \neq j$). Ese proceso es nombrado estimación de la densidad conjunta bivariada de variables aleatorias. De esa manera, cada juego podrá asociar las probabilidades de desempeño de cada equipo (también siendo posible incorporar su dinamismo con el parámetro de la distribución en el cambio del tiempo), como una manera alternativa de asociar al fenómeno de predicciones del número de goles de un partido, también un modelo de IA vía RNAs posibilita estimar el resultado de un partido.

Parte 2: Cálculo de la tripleta de probabilidad (victoria, empate y derrota)

El resultado en un partido determinado, se puede predecir con el vector de probabilidades (tripleta) donde cada una de las componentes es: victoria, empate y derrota. La suma de los tres componentes del vector de probabilidades resulta en uno. Así, un juego es definido como $\mathrm{P}(Y_i,Y_j)$ de equipo i contra j. Según el modelo de Lee [13], la densidad conjunta es el desempeño de las equipos, y podrá ser descrita como una matriz de doble entrada (posibles goles del equipo i en la dimensión de filas y posibles goles del equipo j en las columnas). Luego la distribución bivariada será una matriz, las probabilidades de obtener un empate será la suma de los elementos de la diagonal de la matriz de probabilidades ($\sum_i \sum_j P(Y_i = Y_j)$). La suma de los elementos de la parte triangular superior de la matriz será la probabilidad de victoria del equipo j (o derrota del equipo i, $\sum_i \sum_j P(Y_i < Y_j)$), y la suma de los elementos de la parte triangular inferior de la matriz será la probabilidad de victoria del equipo i (o derrota del equipo j, $\sum_i \sum_j P(Y_i > Y_j)$), descrito visualmente en la Fig. 2.

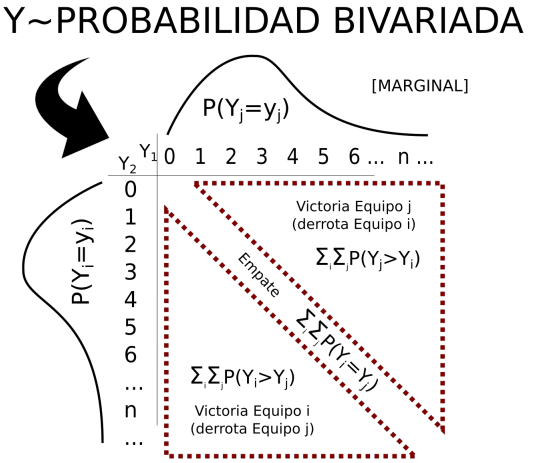

Figure 2: Composición de la función de probabilidad conjunta bivariada (Número de goles del equipo i *versus* del equipo j). Él vector de probabilidades (tripleta) será la suma de los elementos de esa distribución conjunta. Un empate será la suma de los elementos de la diagonal de la matriz de probabilidades ($\sum_i \sum_j P(Y_i = Y_j)$). La suma de los elementos de la parte triangular superior de la matriz será la probabilidad de victoria del equipo j ($\sum_i \sum_j P(Y_i > Y_j)$), y la suma de los elementos de la parte triangular inferior de la matriz será la probabilidad de victoria del equipo i ($\sum_i \sum_j P(Y_i < Y_j)$).

Es importante resaltar que esta metodología de puntuación (número de goles) por equipo, posibilita asociarbvel marcador más probable de cada partido (vía la moda de distribución bivariada discreta/conteo). Lo cual es una diferencia de otros modelos usualmente adoptados por la literatura de aprendizaje de máquina aplicado a predicción deportiva, que tienen el objetivo de clasificación del resultado (por ejemplo, victoria o derrota) sin la obtención del marcador del partido [9].

Parte 3: Calcular la distancia Euclidiana entre el resultado previsto y observado

Una vez obtenida la tripleta de probabilidad (victoria, empate y derrota) estimada de un juego con base a un modelo ajustado, es posible verificar que tan preciso es el modelo adoptado para realizar predicciones (o su bondad de ajuste). Esa adecuación es calculada por la distancia Euclidiana entre los puntos estimados por el modelo y el resultado observado del partido. Un método que resume esas distancias entre valores observados y estimados es el *Brier Score*. Representado como un valor numérico entre 0 (bajo desempeño del modelo) y 1 (alto desempeño), y que tiene como objetivo cuantificar la precisión de las predicciones probabilísticas de los modelos (Poisson, Bell y RNA). Una representación gráfica vía un Simplex de tres dimensiones también es posible.

Según la literatura, un modelo con Brier Score aceptable presentará valor superior a 2/3. En la próxima subsección se explican los modelos estadístico de conteo utilizados para este trabajo (regresión de Poisson y Bell), y también la descripción de los parámetros de los modelos de RNA adoptados.

2.2 Modelos Estadísticos

Él modelo estadístico de regresión univariado para el conteo de datos, se basa en la distribución de Poisson (frecuentemente utilizada en la literatura). Sin embargo, este modelo representa con el mismo parámetro el promedio y la varianza. Esta propiedad, conocida como equi-dispersión, es una peculiaridad del modelo de Poisson, que puede no ser adecuado en diferentes situaciones. Cuando la varianza es mayor que la media, por ejemplo, tenemos el caso de sobre-dispersión. Otra manifestación de escape del supuesto de equi-dispersión es la sub-dispersión, es decir, cuando la varianza es menor que la media. Por estos motivos, es necesario recurrir a otras distribuciones alternativas a la de Poisson (o modificaciones) en casos donde el supuesto de equi-dispersión no sea válido, para ello en este trabajo adoptaremos la distribución Bell.

La estructura básica adoptada por el modelo paramétrico para cada juego es dividida en juego en casa y visitante, que relaciona la característica media de un proceso de una variable de conteo, $\mathbb{E}[Y_i|\mathbf{X}] = \nu_i$, descrito por el modelo logarítmico-lineal, con una estructura de regresión como

$$\log(\nu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

Así, considerando las contribuciones de las variables explicativas (*features*), X's, si mide en la capacidad del equipo local, ν_{home}, como la combinación de una ventaja de casa, β_{home}, añadido por la capacidad de ataque, att_{home}, y reducir el poder de la defensa de los visitantes, def_{away}, es decir,

$$\log(\nu_{home}^{(i)}) = \beta_{home} + \beta_1^{(i)} att_{home} + \beta_2^{(j)} def_{away}$$

y la capacidad del visitante, ν_{away}, es descrita por el modelo que combinación de su poder de ataque, att_{away}, deduciendo la defensa de equipo de casa, def_{home}, es decir,

$$\log(\nu_{away}^{(j)}) = \beta_3^{(j)} att_{away} + \beta_4^{(i)} def_{home}$$

donde, el índice i-th refiere a el equipo de casa y el j-th al equipo visitante. Es importante mencionar que para cada equipo existe una capacidad de ataque en casa, ataque de visita, defensa en casa, y defensa de visita. El parámetro ν puede tomar diferente composición en función de la distribución que se adopte, por ejemplo puede ser explícito de la Poisson ($\nu = \lambda$), o de la Bell ($\nu = \theta e^{\theta}$).

2.2.1 Distribución de POISSON

SY

$$P[Y = y] = \frac{e^{-\lambda}\lambda^y}{y!} \quad \text{con} \quad Y = \{0, 1, 2, \dots\}.$$

X

$$\lambda := \mathbb{E}[Y|\mathbf{X}] = e^{\mathbf{X}\boldsymbol{\beta}}$$

entonces

$$P[Y = y|\mathbf{X}, \boldsymbol{\beta}] = \frac{e^{y\mathbf{X}\boldsymbol{\beta}} e^{-e^{\mathbf{X}\boldsymbol{\beta}}}}{y!}.$$

2.2.2 Distribución de BELL

Sea una variable aleatoria discreta Y que sigue una distribución Bell [14] con parámetro θ, representada por $Y \sim Bell(\theta)$, entonces su función de masa de probabilidad esta representada por:

$$P[Y=y]=\frac{\theta^y e^{-e^\theta+1}B_y}{y!} \quad \text{con} \quad Y=\{0,1,2,\dots\}$$

donde los coeficientes B_y son números Bell [15,16] que viene por

$$B_y=\frac{1}{e}\sum_{k=0}^{\infty}\frac{k^y}{k!}.$$

La esperanza y varianza vienen dadas por $\mathbb{E}[Y]=\theta e^\theta$ y $\mathbb{VAR}[Y]=\theta(1+\theta)e^\theta$.

Sea $\mathbf{X}\in\mathbb{R}^n$ un vector de variables independientes, un modelo de regresión podrá ser escrito como $\log(\mathbb{E}[Y|\mathbf{X}])=\mathbf{X}\boldsymbol{\beta}$. Así, la extensión de ese modelo estadístico, incorporando una estructura explicativa matricial ($\mathbf{X}$) para estimar el parámetro θ sería:

$$\theta e^\theta := \mathbb{E}[Y|\mathbf{X}]=e^{\mathbf{X}\boldsymbol{\beta}}$$

Entonces:

$$P[Y=y|\mathbf{X},\boldsymbol{\beta}]=\frac{W(e^{\mathbf{X}\boldsymbol{\beta}})^y e^{-e^{W(e^{\mathbf{X}\boldsymbol{\beta}})}+1}B_y}{y!}$$

donde $W(.)$ es la función Lambert [17].

Una vez que la distribución sea elegida (Poisson o Bell), y estimado su parámetro, se obtiene un rendimiento esperado para cada equipo usando (como visitante o jugando en casa), descritos por $\{\beta_1^{(i)},\beta_2^{(i)},\beta_3^{(i)},\beta_4^{(i)}\}$. Entonces, cada juego será un evento bivariado, la cual se considera como un enfrentamiento independiente $\mathbf{Y}=(Y_i,Y_j)$, o sea $Y_i\perp Y_j$. Esta distribución conjunta posteriormente irá componer el vector esperado para cada partida (Victoria equipo de casa, Victoria equipo de Visita, o empate) como se muestra en la Fig. 2. Otra metodología para estimar directamente el resultado del vector (como un predicción de clase, victoria equipo de casa, Victoria equipo de Visita, o empate) será descrita a través de modelos de RNAs.

2.3 Redes Neuronales Artificiales (RNAs)

Como se indica en la sección de introducción, las RNAs son modelos computacionales que se basan en el funcionamiento del cerebro humano, cuya unidad básica es la neurona [12]. La arquitectura general de una RNA básica se puede observar en la Fig. 3, la cual requiere un conjunto de variables de entrada representada por el vector $\mathbf{X}=(x_1,x_2,x_3,\dots,x_n)$. Luego, un conjunto de pesos representados por el vector $\mathbf{W}=(W_1,W_2,W_3,\dots,W_n)$ que combinará las variables de entrada, y condicionado a una función de activación transformará en una respuesta (y) como salida.

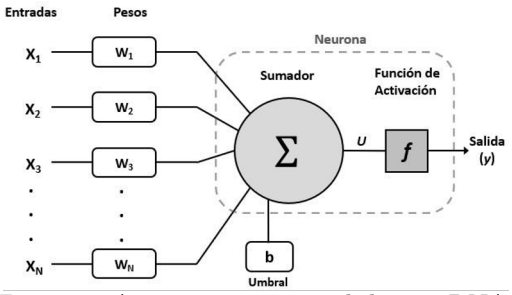

Figure 3: Arquitectura general de una RNA.

Internamente, se debe definir un conjunto de capas ocultas donde se ubican la cantidad de neuronas establecidas para el modelo. En cada neurona, se multiplican los valores contenidos en las variables entradas x_i con sus pesos correspondientes W_i, y posteriormente estos productos son sumados a un valor de umbral b predefinido, y de esta manera se obtiene U, como se puede observar en la siguiente ecuación:

$$U=\sum_{i=1}^{n}(W_i*x_i)+b$$

donde; W_i, representa el peso de interconexión correspondiente a cada una de las entradas x_i. La función de activación f es la que se aplica a U antes de generar la salida representada por y.

Las funciones de activación, son funciones que se encargan de limitar la amplitud de la salida de una neurona a partir de la suma ponderada de la entrada y el umbral, dando como resultado un valor que indica si una neurona puede activarse o no, y con qué magnitud. Estas funciones controlan la salida de cada una de las neuronas y de acuerdo a su expresión matemática puede ser lineal o no-lineal. Existen diferentes funciones de activación que se pueden aplicar dependiendo del tipo de datos con que se están tratando, las más

utilizadas son las funciones; lineales, sigmoide y la tangente hiperbólica [18].

La arquitectura más sencilla de una RNA es el perceptrón, el cual consiste en una neurona con dos entradas, una salida y una función de activación. Para problemas más complejos se utilizan estructuras con múltiples capas ocultas y que son conocidas como perceptrón multicapa o MLP, que además de contar con una capa de entrada y una capa de salida, cuenta con una o más capas ocultas.

El algoritmo de retropropagación (*backpropagation*), es el más común para el entrenamiento de una RNA, el cual tiene en cuenta el valor de salida para ajustar los pesos de las capas anteriores. Una RNA con estas características, es una de las más utilizadas para problemas simples de clasificación y predicción. Las RNA de alta complejidad dan paso a lo que se conoce en la actualidad como Aprendizaje Profundo *(Deep Learning)*. Estos modelos se califican como supervisados, pues buscan pronosticar el comportamiento de un atributo de acuerdo a un aprendizaje previo, a partir de un conjunto de datos conocidos, y son aplicados tanto a problemas de clasificación como de regresión [19].

Dentro de las arquitecturas de RNA de alta complejidad que actualmente se destacan son las conocidas como Redes Neuronales Recurrentes (RNR), que a diferencia de las RNAs clásicas, tratan datos secuenciales de forma eficiente, tomando salidas anteriores como entrada, lo que les da la posibilidad de tratar largas secuencias elemento a elemento. Una RNR simple tiene retroalimentación de activación que incorpora memoria a corto plazo. Una capa de estado se actualiza no solo con la entrada externa, sino también con la activación de la propagación directa del estado anterior [19].

En este trabajo se desarrollan dos modelos de RNAs para predicción del resultado de los partidos, y determinar así, qué equipo ganará si es el equipo A o B, o empate. Fueron consideradas como variables de entradas la cantidad de goles acumulados hasta la ronda observada, y el promedio de goles (sea como juego en casa o visitante). Así, estos modelos tienen nueve variables de entradas (número de ronda, ID del equipo que juega en casa, cantidad de goles acumulados hasta el momento jugando en casa, número partido como local, promedio de goles como local, ID del equipo visitante, la cantidad de goles acumulados hasta el momento jugando como visitante, el número de juego como visita y el promedio de goles como visitante). El primer modelo con una estructura MLP, la arquitectura adoptada es de 3 capas ocultas con 18 neuronas en cada capa, adoptando la función de activación logística, con una tasa de aprendizaje de 0.01 y un *threshold* de 0.5, considerando 10 réplicas.

El segundo modelo corresponde a un estructura de RNR, con cuatro capas densas alternadas por capas de *dropout*, o sea un total de siete capas, totalizando 84,961 parámetros. La función de pérdida adoptada fue el error cuadrático medio (MSE), el optimizador *rmsprop* con un *learning rate* de 0.0001, guiada por las métricas de precisión y de error absoluto medio (MAE).

3 La Liga Chilena de Fútbol

La liga Chilena de fútbol está dividida en tres categorías de manera profesional, que está formada por 45 equipos que forman parte de la Asociación Nacional de Fútbol Profesional (ANFP). La primera categoría llamada "Primera división" está conformada por 18 equipos, la segunda categoría llamada "Primera B" formada por 15 equipos y la tercera categoría llamada "Segunda División Profesional" formada por 12 equipos, este formato se estableció en 2016.

La competición en la "Primera división" se juega en dos rondas de todos contra todos, cada equipo se enfrenta con cada uno de los que compone esta división tanto de local como visitante, por lo tanto como son 18 equipos en esta categoría entonces daría un total al final de la temporada de $18 \times 17 = 306$ juegos y 34 rondas o fechas. Los puntos son obtenidos de la siguiente manera, un equipo por victoria se le asignan 3 puntos, por empate 1 punto y por derrota 0 puntos.

El equipo que obtenga más puntos es el ganador del torneo, en caso de que haya un empate de dos equipos en el primer lugar, al final de la temporada se define el campeón en único partido entre ambos equipos empatados, ahora en el caso que sean más de dos equipos con la misma cantidad de puntos en la cima de la tabla, también se disputará un único partido entre los dos equipos que ocupen los dos primeros lugares según la clasificación en la Tabla Absoluta, que se describe a continuación:

- Mayor cantidad de puntos obtenidos.
- En caso de igualdad, mayor diferencia entre los goles marcados y recibidos.
- En caso de igualdad, mayor cantidad de partidos ganados.
- En caso de igualdad, mayor cantidad de goles marcados.
- En caso de igualdad, mayor cantidad de goles marcados en calidad de visita.
- En caso de igualdad, menor cantidad de tarjetas rojas recibidas.
- En caso de igualdad, menor cantidad de tarjetas amarillas recibidas.
- Sorteo

Los equipos ubicados en la última posición de la Tabla Absoluta y de la tabla de Coeficiente de Rendimiento, desciende a la Primera B. La tabla Absoluta, es la tabla donde se registran los puntos acumulados en los partidos que se disputen durante el Campeonato Nacional Temporada 2020; y la Tabla de Coeficiente de Rendimiento, consiste en un coeficiente de rendimiento calculado mediante la división entre la cantidad de puntos obtenidos y los partidos jugados en la temporada 2019 ponderado por un 60%, y un coeficiente de rendimiento calculado mediante la división entre la cantidad de puntos obtenidos y los partidos jugados en la temporada 2020 ponderado por un 40%. La suma de ambos coeficientes determinará la ubicación final en la Tabla de Coeficiente de Rendimiento 2019-2020.

Y por último, un tercer equipo desciende de la Primera división, y es el perdedor entre un partido único, que se disputará entre los equipos que ocupen el penúltimo lugar de la Tabla Absoluta y el penúltimo lugar la Tabla Coeficiente de Rendimiento, o el antepenúltimo de esta última tabla en el evento que el penúltimo hubiese descendido según el numeral precedente. Si es uno sólo el equipo que reúne las condiciones señaladas en este numeral descenderá automáticamente.

En la temporada de 2020, que terminó en febrero de 2021, una temporada atípica debido a la pandemia por el Covid-19 por lo cual tuvo que extenderse unos meses más y ocupar parte del siguiente año calendario, el equipo ganador fue la Universidad Católica (UCA) que acumuló un total de 65 puntos, y los equipos que descendieron fueron Iquique (DIQ), Coquimbo Unido (COU) y la Universidad de Concepción (UCO), este último equipo descendió por el partido que debe jugar el penúltimo de la tabla Absoluta con el penúltimo de la tabla de coeficiente de Rendimiento.

El inicio de la temporada 2020-2021 fue el día 24 de enero de 2020, la cual inició con el encuentro entre el Everton (EVE) y la Universidad de Concepción (UCO), y culminó el día 15 de febrero de 2021 con el partido entre Cobresal (COB) y Unión Española (UES).

Otra de las cosas que se disputan en la liga Chilena, es la participación en torneos internacionales, por ello también es interesante obtener una buena posición en la Tabla Absoluta. En este caso los cuatro primeros de la Tabla clasifican a la Copa Libertadores y de la quinta a la octava posición clasifican a la Copa Sudamericana, campeonatos donde juegan los mejores equipos de las ligas pertenecientes a la Confederación Sudamericana de Fútbol (CONMEBOL).

4 Resultados

En la temporada 2020, de la primera división del campeonato Chileno de fútbol, participaron 18 equipos, totalizando 306 partidos en 34 rondas o fechas, de las cuales para el desarrollo de esta investigación, las últimas 5 rondas se guardan para la validación de los modelos descritos en la sección de Métodos. Es decir, se consideran 29 rondas (total de 261 juegos) como información previa en los tres modelos propuestos, y las últimas 5 rondas se predicen y se comparan con los marcadores finales de cada juego (total de 45 juegos). Para las RNAs, la base de datos de las 29 rondas fue dividida para el conjunto de datos entrenamiento para evitar problemas de sobre-ajuste en los modelos (*overfitting*), y las 5 últimas rondas para el conjunto de datos de validación.

Para la predicción de las 5 últimas rondas del campeonato, se usaron los modelos descritos anteriormente, que son dos modelos estadísticos y dos modelos de RNAs. Los modelos estadísticos posibilitan la predicción del marcador final del juego (número de goles de cada partido) y luego se calcula la probabilidad de empate o victoria de cada equipo en un juego (vector de probabilidad de los resultados), y de esta manera se puede clasificar el

juego como empate, victoria o derrota. Por otro lado, RNAs predicen el resultado de los partidos como empate o victoria del equipo (resultado como una clase) adoptando el número de goles acumulado del tiempo $(t-1)$ por el equipo hasta el partido (t), según su posición (jugando en Casa o Visitante).

En la Fig. 4, se puede observar el comportamiento inconstante de la mayoría de los equipos con respecto a la posición final de cada temporada (de los últimos 5 años). Los 3 equipos que se mantuvieron en el top 6 de las últimas 5 temporadas, fueron: Universidad Católica (UCA), Universidad de Chile (UCH) y Colo Colo (CCO), sin embargo estos equipos tienen temporadas en las cuales quedaron por debajo las 10 primeras posiciones, como son los casos de la UCH en la temporada 2019 y 2021 quedando en las posiciones 15 y 16 respectivamente, y CCO se posicionó en los puestos 6 y 17 respectivamente. En el campeonato de la primera división de Chile, es posible observar una gran variación de posiciones, sin embargo los equipos de la región central se muestran casi siempre bien posicionadas en el campeonato en las últimas cinco temporadas.

También se puede ver que el equipo que mayormente está en los últimos puestos ha sido Deportes Iquique (DIQ), que a pesar de ocupar estos lugares sólo ha descendido una sola temporada. Observando sólo las posiciones finales de cada temporada, se puede notar que en la mayoría de los campeonatos siempre queda un equipo, que se encontraba en el top 6 o cercano a éste, queda en alguno de los dos últimos lugares del torneo.

Por otro lado, vemos que para las temporadas del 2017 al 2019 participaron 16 equipos y para la temporada 2020 se agregaron dos equipos más, y luego para el año 2021 se tienen 17 equipos, esto debido a los cambios de reglamento de los últimos años y el imprevisto de la pandemia que afectó el desarrollo normal del año calendario.

Otro aspecto relevante que podemos detallar al observar las últimas temporadas, es que no hubo descenso de equipos en los años del 2017 y 2019. Para la temporada de 2017, esto debido a una adaptación al nuevo sistema de competición para otorgar el campeón absoluto del año, ya que antes se jugaba la temporada dividida en dos años calendarios, mientras ahora la temporada se juega en el mismo año calendario.

Como otro dato relevante en el cumplimiento del calendario planteado al inicio de temporada, es que en el año 2019, en Chile se producen grandes movimientos de descontento social que afectaron el normal desarrollo de la liga, y esto complicó el término normal de la temporada, por lo que la ANFP decide finalizar la temporada con los partidos que se habían logrado llevar a cabo hasta ese momento y declarar como ganador al puntero del campeonato, la Universidad Católica (UCA). Para la temporada 2020 se tuvo que jugar una parte en el año 2021 motivado por la pandemia. Al momento de tomar los datos para la comparación de las posiciones de los equipos en el último año (2021), el torneo estaba en la ronda 32.

Para predecir resultados y las posiciones de la liga Chilena de 2020, se utilizaron modelos estadísticos de Poisson y Bell, que se describieron en la sección de métodos. Estos modelos dan la probabilidad de gol de un equipo, luego realizando la composición de la función de la probabilidad conjunta bivariada se calculan las probabilidades de anotaciones en un partido, generando una matriz de las probabilidades que tendría cada marcador. La diagonal de la matriz representa los resultados donde el juego queda empatado y por ello la suma de todas las probabilidades de la diagonal daría en total la probabilidad que el partido quede en empate, y la suma de las probabilidades de la triangular superior o inferior, da la probabilidad de victoria de uno de los equipos.

Por esto, se genera un vector de probabilidades con tres componentes: probabilidad de victoria del equipo local, probabilidad de victoria del equipo visitante y empate. Por ejemplo, en la Fig. 5 se muestra las probabilidades de ganar o empatar de cada equipo, en el encuentro entre la Universidad de Concepción y la Universidad Católica, donde ambos modelos estadísticos dan por ganador a la Universidad Católica con un porcentaje mayor a 50%, acertando la predicción en ambos modelos cuando se compara con el resultado real.

En la Tabla I se presenta el porcentaje de desaciertos en las rondas 30, 31, 32, 33 y 34, de la predicción realizada por los modelos de Poisson y Bell, considerando los datos de validación, es decir, las 5 últimas rondas del campeonato. También se presentan en dicha tabla el Brier Score obtenidos por los modelos estadísticos y las RNAs (MLP y *deep learning*).

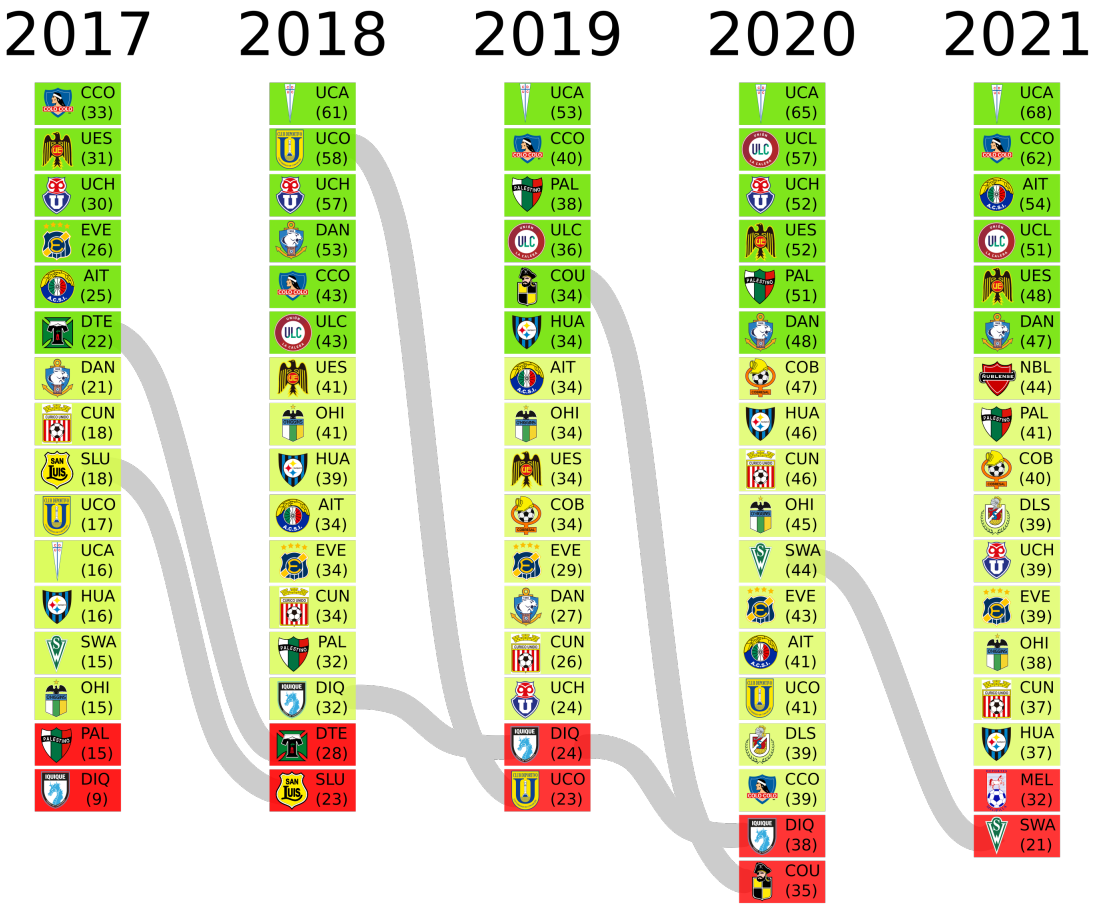

Figure 4: Posiciones de los equipos de las últimas 5 temporadas de la primera división de la liga Chilena. En verde el top 6, en rojo los dos últimos de cada temporada, y en amarillo las equipo con desempeño mediano.

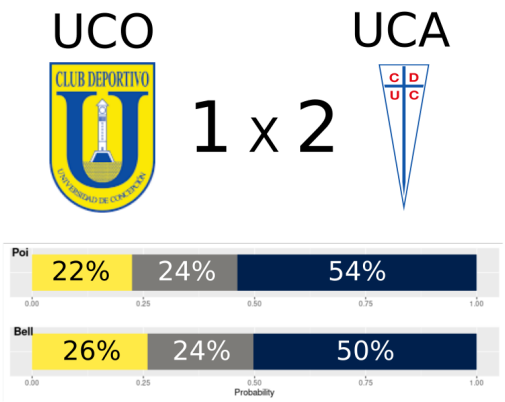

Figure 5: Modelo de conteo estadístico Poisson *versus* Bell, y condicionados a su distribución conjunta bivariada calculado el chanceo de victoria de UCO *versus* UCA (electa al azar). El resultado fue el número de goles observados, mientras la primera barra muestra las estimativas de (victoria de la UCO, empate, y victoria de la UCA) según el modelo de Poisson y luego según el modelo Bell.

La ronda 33 fue la que tuvo más desaciertos en ambos modelos estadísticos con 88.89%, lo cual es un muy alto porcentaje, sin embargo, cabe destacar que en esta jornada prevaleció el empate en el juego. En cambio las fechas 30 y 34, son las mejores estimadas con sólo un 44.44% de desaciertos, es decir, de los 9 partidos que se juegan en cada jornada, no se predijo de forma correcta sólo 4 partidos, de los cuales la mayoría de estos partidos quedaron en empate (6 de 8 partidos no pronosticados correctamente).

Por otro lado, tenemos la comparación de los modelos estadísticos Poisson y Bell por medio del índice Brier Score. Este índice es una forma de verificar la exactitud de una previsión de probabilidad, y además es usada para resultados categóricos siempre que puedan estructurarse como resultados binarios [20], que para este caso nuestra medición es acertar o no el resultado de un juego.

Según la forma de medir este índice, mientras más cerca a 0 tienen mayor precisión, por lo cual vemos que tanto la Poisson como la Bell. Sin embargo, el modelo RNA mostró, con posibilidades de mejoría de desempeño (adicionando variables extras, por ejemplo las adoptadas en [9]), sus predicciones en los datos de validación de las 5 últimas rondas del campeonato. Tabla II muestra las estadísticas de desempeño como: Sensibilidad, especificidad, exactitud balanceada, número de ocurrencias de las clases, entre otras.

La Fig. 6, ilustra la RNA ajustada para el modelo de clasificación de la liga Chilena de 2020 de la primera división, considerando como respuesta la clasificación de una confrontación de un partido (victoria de equipo de Casa o Visitante, o empate).

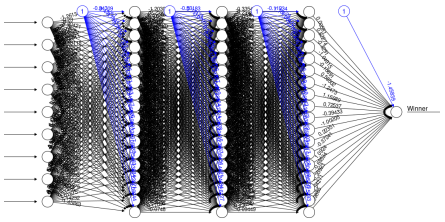

Figure 6: MLP ajustada para la clasificación de una confrontación de la liga Chilena de 2020 de la primera división. Como variables de entrada fueron consideradas: número de ronda, ID del equipo de casa, número del juego en casa, cantidad de goles acumulados del equipo de casa (hasta antes del partido), promedio de goles del equipo de casa y del visitante, ID del equipo visitante, número del juego como visitante y cantidad de goles acumulados del equipo visitante. El resultado es una clase (victoria del equipo casa, victoria del equipo visitante, o empate).

Todos los modelos presentaron desempeños satisfactorios, sin embargo, todavía hay espacio para mejorías en especial para predicciones de empates. Todos los modelos presentaron menores desempeño en la clase de empate, ya discutido en la literatura como gran desafío para estos modelos

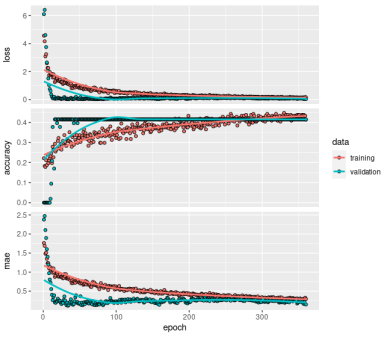

Figure 7: RNR ajustada para la clasificación de una confrontación de la liga Chilena de 2020 de la primera división considerando las mismas variables de la MLP.

TABLA I
DESEMPEÑO DE LOS MODELOS POISSON, BELL Y RNA (MLP Y RNR), SEGÚN DATOS DE VALIDACIÓN DE LAS 5 ÚLTIMAS RONDAS DEL CAMPEONATO DE 2020.

	Ronda 30	**Ronda 31**	**Ronda 32**	**Ronda 33**	**Ronda 34**
Poisson	44.44%	66.67%	55.56%	88.89%	44.44%
(Brier)	0.67	0.70	0.64	0.81	0.66
Bell	55.56%	66.67%	55.56%	88.89%	44.44%
(Brier)	0.68	0.69	0.65	0.81	0.64
MLP	33.33%	66.67%	33.33%	77.78%	44.44%
RNR	55.56%	44.44%	55.55%	66.67%	33.33%

TABLA II
DESEMPEÑO DE PREDICCIÓN DE LA MLP EN LAS 5 ÚLTIMAS RONDAS DEL CAMPEONATO 2020 CHILENO.

	Victoria Casa	**Empate**	**Victoria Visita**
Sensibilidad	0.438	0.700	0.220
Especificidad	0.828	0.560	0.833
Valor Pred Pos	0.583	0.560	0.250
Valor Pred Neg	0.727	0.700	0.811
Prevalencia	0.356	0.444	0.200
Tasa de Detección	0.156	0.311	0.044
Prevalencia de Detección	0.267	0.556	0.178
Precisión Balanceada	0.633	0.630	0.528
Total (n)	16	20	9

predictivos [9].

5 Conclusiones

Los modelos estadísticos aplicados y de RNA, mostraron ser modelos competitivos para hacer predicciones en los partidos de la primera división de la Liga Chilena de fútbol, debido a la paridad de los equipos en este torneo. Los resultados obtenidos por los diversos modelos presentados, corroboran con lo discutido en [10], mostrando la presencia de la estocasticidad en el fútbol, sin embargo es posible expresar la puntuación final de la tabla.

Esta investigación consideró modelos de regresión de conteo para la predicción del resultado de los partidos de fútbol, considerando la distribución de Poisson y Bell. La estructura de predicción se basa en tres etapas: i) marcador del juego, proveniente de las distribuciones marginales, ii) opción de victoria o empate, según la distribución bivariada, y iii) la clasificación de los equipos en el campeonato.

En los modelos de RNA condicionado a la estructura adoptada, es posible obtener el resultado de cada confrontación de equipos (victoria equipo A o equipo B, o empate), como un problema de clasificación. Hay un próximo paso de enriquecimiento de la base de datos para desarrollar la predicción de las RNAs, que será contemplado como trabajo futuro.

El campeonato Chileno demanda un desafío extra por presentar un promedio bajo de goles por partido. Por lo tanto, se hace intuitivo pensar que es por ello que gran cantidad de partidos quedan empatados y consecuentemente, se obtienen puntajes muy estrechos entre los equipos en el campeonato. Futuras investigaciones se podrán dedicar a dar mayor énfasis en la predicción de empates (con mayor peso), por lo cual se deben considerar factores que faciliten la predicción del juego que es más visible en el mundo.

Agradecimientos

Los autores agradecen al fomento brasileño del Consejo Nacional de Desenvolvimiento Científico e Tecnológico (CNPq), y a la Fundación Araucaria y la Coordinación de Mejora del Personal de Nível

Superior (CAPES) por los apoyos financieros parciales para este desarrollo científico.

Referencias

References

[1] Andrew Baerg. Big data, sport, and the digital divide: Theorizing how athletes might respond to big data monitoring. *Journal of Sport Social Issues*, 41, 10 2016.

[2] César Valero and Mabel González. Sabermetría y nuevas tendencias en el analisis estadístico del juego de béisbol. *Retos*, 28:122–127, 2015.

[3] César Soto-Valero. Predicting win-loss outcomes in mlb regular season games – a comparative study using data mining methods. *International Journal of Computer Science in Sport*, 15:91–112, 12 2016.

[4] Fadi Thabtah, Li Zhang, and Neda Abdelhamid. Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1):103–116, 2019.

[5] Bryan L Boulier and Herman O Stekler. Predicting the outcomes of national football league games. *International Journal of forecasting*, 19(2):257–270, 2003.

[6] Adriano K Suzuki, Luis Ernesto Bueno Salasar, JG Leite, and Francisco Louzada-Neto. A bayesian approach for predicting match outcomes: the 2006 (association) football world cup. *Journal of the Operational Research Society*, 61(10):1530–1539, 2010.

[7] Francisco Louzada, Adriano K Suzuki, and Luis EB Salasar. Predicting match outcomes in the english premier league: Which will be the final rank? *Journal of Data Science*, 12(2):235–254, 2014.

[8] David Sally and chris anderson. *The Numbers Game: Why Everything You Know About Soccer Is Wrong*. 06 2013.

[9] Hugo Ribeiro Santana, Francisco Louzada, Paulo Henrique Ferreira, Adriano Kamimura Suzuki, and Anderson Ara. Modelagem estadística e de aprendizado de máquina: Previsao do campeonato brasileiro série a 2017. *Matemática e Estatística em Foco*, 15:42–66, Mayo 2020.

[10] Leonardo Barrios Blanco, Paulo Henrique Ferreira, Francisco Louzada, and Diego Carvalho do Nascimento. Is football/soccer purely stochastic, made out of luck, or maybe predictable? how does bayesian reasoning assess sports? *Axioms*, 10(4):276, 2021.

[11] Enrique Antonio Pedroza Santiago. Modelo de Predicción Difuso para Encuentros de Fútbol Soccer Mexicano, Diciembre 2018.

[12] H. Kukreja, N. Bharath, C. Siddesh, and S. Kuldeep. An introduction to artificial neural network. *Int J Adv Res Innov Ideas Educ.*, (1):27–30, 2016.

[13] Alan J Lee. Modeling scores in the premier league: is manchester united really the best? *Chance*, 10(1):15–19, 1997.

[14] Fredy Castellares, Silvia Ferrari, and Artur Lemonte. On the bell distribution and its associated regression model for count data. *Applied Mathematical Modelling*, 56, 12 2017.

[15] Eric T Bell. Exponential numbers. *The American Mathematical Monthly*, 41(7):411–419, 1934.

[16] Eric Temple Bell. Exponential polynomials. *Annals of Mathematics*, pages 258–277, 1934.

[17] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the lambert w function. *Advances in Computational Mathematics*, (5):329–359, 1996.

[18] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, (1):45–48, 2018.

[19] Feizi H. Sattari M. Colak M. Shamshirband S. Apaydin, H. and K. Chau. Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *MDPI Journal Water*, (1):06–10, 2020.

[20] Mark S Roulston. Performance targets and the brier score. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 14(2):185–194, 2007.