UNIVERSIDAD SEÑOR DE SIPÁN

## FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

### ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS

### TRABAJO DE INVESTIGACIÓN

# ALGORITMOS COMPUTACIONALES PARA LA PREDICCIÓN DE PARTIDOS DE FÚTBOL

### PARA OPTAR POR EL GRADO DE BACHILLER EN INGENIERÍA DE SISTEMAS

**AUTOR**
**Torres Hernández Edwin Bryan**

**ASESOR**
**Mg. Junior Eugenio Cachay Maco**

**LÍNEA DE INVESTIGACIÓN**
**Infraestructura, Tecnología y Medio Ambiente**

**Pimentel – Perú**
**2020**

# ALGORITMOS COMPUTACIONALES PARA LA PREDICCIÓN DE PARTIDOS DE FÚTBOL

**AUTOR**

**Torres Hernández Edwin Bryan**

# RESUMEN

Dentro de los múltiples ámbitos, dónde se han aplicado modernos métodos de aprendizaje automático para conjuntos de datos muy grandes (Big Data), se pueden descubrir ideas que de otro modo podrían permanecer ocultas. No obstante, confiar y basarse solamente en dichos conjuntos de datos para descubrir y predecir conocimiento podría ser equivocado para muchos dominios de este mundo.

En el marco de los deportes, el fútbol es el más representativo, jugado por dos equipos conformado por once jugadores cada uno, consiste en hacer que una pelota ingrese al arco del oponente, a cuya acción se le denominará “gol”.

Los resultados de un partido de fútbol han sido considerados como el esfuerzo científico para mejorar las técnicas de juego y características del equipo. Debido a que, hay una gran cantidad de factores que se deben tener en cuenta y que además, muchas veces no se pueden representar cuantitativamente, predecir dichos resultados es un proceso complejo (Hucaljuk & Rakipovic, 2011).

Dicho de este modo, es un desafío investigar la información y estrategia de clasificación que ayudarían a facilitar la predicción de los resultados de los partidos.

**Palabras clave:** Algoritmos, Predicción.

# ABSTRACT

Within the multiple areas, where modern machine learning methods have been applied for very large data sets (Big Data), you can discover ideas that might otherwise remain hidden. However, relying on and relying solely on such data sets to discover and predict knowledge could be wrong for many domains in this world.

In the framework of sports, football is the most representative, played by two teams made up of eleven players each, is to make a ball enter the opponent's arc, whose action will be called "goal."

The results of a football match have been considered as the scientific effort to improve the game techniques and characteristics of the team. Because, there are a large number of factors that must be taken into account and that in addition, many times cannot be represented quantitatively, predicting these results is a complex process (Hucaljuk & Rakipovic, 2011).

Usually, it is a challenge to investigate the information and classification strategy that would help facilitate the prediction of match results.



**Keywords:** Algorithms, Prediction.

# ÍNDICE

# I. INTRODUCCIÓN

Desde la antigüedad el ser humano ha necesitado imaginar que cosas se pueden predecir, porque nos da mayor tranquilidad descubrir que podemos anticiparnos a ciertas eventualidades; frente a ello y con los alcances tecnológicos de siglo XXI la era donde la información es poder, las ciencias de la computación se han visto en la creativa necesidad de recurrir a la inteligencia artificial, programa computacional que se diseñó para determinar ciertas operaciones que se consideran propias de los seres humanos.

El presente trabajo de investigación tiene como objetivo determinar cuál es el algoritmo computacional para la predicción de resultados de fútbol con mayor precisión, esto debido al interés y pasión de la población latinoamericana con dicho deporte. Además, de aplicar la herramienta de análisis para determinar el nivel de precisión de los algoritmos seleccionados.

# II. MATERIAL Y MÉTODOS

## 2.1. Tipo y Diseño de Investigación

### 2.1.1. Tipo de la Investigación

**a) No experimental**

Según la intervención del investigador, este trabajo de investigación es no experimental, porque se pretende evaluar los resultados en base a las observaciones que se realicen de la aplicación de los algoritmos computacionales para la predicción de resultados de partidos de fútbol sin variar intencionalmente las variables, sin una selección aleatoria de sujetos, ni una manipulación de las condiciones.

**b) Descriptiva**

La presente investigación pretende describir los pasos que se deben seguir para la extracción, limpieza y ordenamiento de datos, así mismo, los pasos a seguir para aplicar en base a estos datos procesados cada uno de los algoritmos computacionales para la predicción de resultados de partidos de fútbol seleccionados con ayuda de la herramienta WEKA y a partir de ello observar ventajas y desventajas entre los mismos.

**c) Deductiva**

Debido a que llegaremos a conclusiones específicas después de la aplicación de los algoritmos computacionales para el pronóstico de resultados de partidos de fútbol los cuales ya se encuentran definidos y establecidos.

**d) Aplicada**

La problemática de la presente investigación está establecida y es conocida por el autor y tiene como finalidad predecir comportamientos específicos en situaciones definidas.

**e) Cuantitativa**

La medición de los indicadores estará en su totalidad en forma numérica.

**f) Longitudinal**

Para la presente investigación se usarán datos históricos de temporadas de fútbol pasadas hasta la actualidad.

### 2.1.2. Diseño de la Investigación

En el presente trabajo de investigación aplicaremos un diseño descriptivo comparativo que se realizara de la siguiente manera:

$$O_i X_1 \neq O_i X_2 \neq O_i X_3$$

$\boldsymbol{O_i}$: Observación

$\boldsymbol{X_1}$: LogitBoost

$\boldsymbol{X_2}$: ClassViaReg

$\boldsymbol{X_3}$: Logistic Regression

## 2.2. Población y Muestra

### 2.2.1. Población

La presente investigación tendrá una población que constará de 18 temporadas de fútbol de la Premier League, específicamente las temporadas 2000/2001 hasta 2017/2018.

### 2.2.2. Muestra

Para la presente investigación se tomaron como muestra 6686 partidos de futbol.

## 2.3. Hipótesis

### 2.3.1. Hipótesis Alterna:

$$\boldsymbol{H_A = X_1 \neq X_2 \neq X_3}$$

Los algoritmos computacionales para la predicción de resultados de partidos de fútbol: LogitBoost, ClassViaReg y Logistic Regression, llegan a resultados diferentes.

### 2.3.2. Hipótesis Nula:

$$\boldsymbol{H_N = X_1 = X_2 = X_3}$$

Los algoritmos computacionales para la predicción de resultados de partidos de fútbol: LogitBoost, ClassViaReg y Logistic Regression, llegan a resultados similares; siendo el algoritmo de Regresión Logística el más preciso.

## 2.4. Operacionalización

| Variables | Indicadores | Formula |
|---|---|---|
| **Variable Dependiente:** Predicción de los resultados de los partidos de fútbol. | **Exactitud** | $\frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{FN} + \mathrm{VN}}$ |
| | **Precisión** | $\frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$ |
| | **Sensibilidad** | $\frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$ |
| | **Especificidad** | $\frac{\mathrm{VN}}{\mathrm{FP} + \mathrm{VN}}$ |
| **Variable Independiente:** Algoritmos computacionales para la predicción. | **Rendimiento** | $\mathrm{T_e} = \text{Tiempo de ejecución}$ |

## 2.5. Métodos, técnicas e instrumentos de recolección de datos

### 2.5.1. Observación

Se denomina así al instrumento visual que nos ayudará a recoger datos de lo que ocurre en alguna situación existente, que permitirá clasificar el contenido o acontecimiento con cierto esquema y dependiendo de la problemática realizar el estudio. Se necesita prestar atención a esta técnica para determinar de manera apropiada los resultados confiables de las predicciones.

### 2.5.2. Análisis documental

En esta técnica se va a extraer información de distintas fuentes: artículos, libros, papers, etc; en los que encontraremos métodos, teorías y técnicas que darán una solución a determinados problemas. Con dicha información se podrá limitar la investigación y estructurar el modelo que se va a estudiar para analizar resultados logrados.

## 2.6. Procedimiento para la recolección de datos

La recolección de información de este trabajo de investigación comienza con la obtención de los archivos ".CSV" de cada temporada jugada en la Premier League, desde la temporada del 2000/2001 hasta la temporada 2017/2018, estos archivos fueron descargados desde la página http://www.football-data.co.uk. Se

descargaron un total de 18 archivos que en conjunto nos dan un total de 6686 partidos de fútbol.

Estos archivos pasaron por una fase de preprocesamiento, en la cual se descartaron datos innecesarios para la investigación. Se generó un único archivo con un total de 6687 filas (1 fila para los encabezados y 6686 filas correspondientes a los partidos de fútbol), después de generado este archivo, se procedió a la elección de los algoritmos a evaluar. Como herramienta para evaluar la precisión de los algoritmos elegidos se utilizó WEKA, aplicación que mostró los resultados esperados, los cuales fueron interpretados en el capítulo 4 de esta investigación.

## 2.7. Criterios éticos

### 2.7.1. Confidencialidad

Toda muestra obtenida en el proceso de la investigación se mantendrá durante el procedimiento y desarrollo en total secretismo.

### 2.7.2. Derechos de Autor

Todos los materiales intelectuales en esta investigación están debidamente citados y referenciados.

### 2.7.3. Objetividad

El análisis del contexto encontrado está basado en criterios imparciales y técnicos.

### 2.7.4. Veracidad

Teniendo en cuenta la confidencialidad, la información que se muestra es verdadera.

## 2.8. Criterios de rigor científico

### 2.8.1. Consistencia

El análisis aplicado a los datos se realizó con supremo profesionalismo, empleando técnicas, habilidades y conocimientos de ingeniería e investigación para mantener la virtud y consistencia de los datos.

### 2.8.2. Validez

El resultado de esta investigación será correctamente evaluado y analizado para poder obtener un resultado válido que nos ayude a resolver la problemática planteada.

### 2.8.3. Fiabilidad

La investigación cumple con este principio al hacer el uso de diferentes técnicas e instrumentos de medición aplicados al proceso de recolección y transformación de la información, para obtener resultados semejantes a lo planteado en el principio.

### 2.8.4. Transferibilidad

La investigación proporcionará información y conocimiento que puede ser transferida a investigadores que se enfoquen en contextos similares.

### 2.8.5. Neutralidad

La manera en cómo se desarrolla la investigación garantiza la seguridad de que los resultados obtenidos, no pueden ser alterados o desviados por intereses, motivaciones y/o perspectivas del investigador.

# III. RESULTADOS

Los resultados de esta investigación se presentaron en fichas de observación de análisis de rendimiento de los algoritmos: LogitBoost, ClassViaReg, Logistic Regression; los cuales fueron sometidos a 4 pruebas cada uno:

- 1 Training Test
- 1 Supplied Test Set
- 1 Cross-validation
- 1 Percentage Split

## 3.1. LogitBoost

### 3.1.1. Training Test

Consiste en utilizar la totalidad de elementos del archivo con el que se entrenó el modelo para realizar las pruebas de rendimiento. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 3104 victorias locales, 1874 victorias visitantes y 1708 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo LogitBoost.

Tabla 1 – Prueba Training Test aplicada al algoritmo LogitBoost

| Ficha de Observación: Evaluación del algoritmo LogitBoost |
|---|

| Resumen de Evaluación | |
|---|---|
| **Número de evaluación** | 1 |
| **Tipo de prueba** | Training Set |
| **Número de instancias** | 6686 |
| **Número de instancias procesadas correctamente** | 6593 |
| **Número de instancias procesadas incorrectamente** | 93 |
| **Tiempo de construcción** | 0.89 segundos |
| **Exactitud promedio** | 99.1% |
| **Precisión promedio** | 98.6% |
| **Sensibilidad promedio** | 98.6% |
| **Especificidad promedio** | 99.1% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 3095 | 9 | 0 |
| **Away** | **b** | 11 | 1863 | 0 |
| **Draw** | **c** | 44 | 29 | 1635 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 99.0% | 98.3% | 99.7% | 98.5% |
| **Away** | 99.3% | 98.0% | 99.4% | 99.2% |
| **Draw** | 98.9% | 100.0% | 95.7% | 100.0% |

Fuente: Elaboración propia

### 3.1.2. Supplied Test Set

Consiste en utilizar un archivo con datos de prueba distintos a los elementos del archivo con el que se entrenó el modelo para realizar las pruebas de rendimiento. La matriz de entrada tuvo 24 columnas y 381 filas (1 fila de encabezados). Se esperaba obtener una salida de 181 victorias locales, 128 victorias visitantes y 71 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo LogitBoost.

Tabla 2 – Prueba Supplied Test Set aplicada al algoritmo LogitBoost

| Ficha de Observación: Evaluación del algoritmo LogitBoost |
|---|

| Resumen de Evaluación | |
|---|---|
| Número de evaluación | 2 |
| Tipo de prueba | Supplied Test Set |
| Número de Instancias | 380 |
| Número de Instancias procesadas correctamente | 376 |
| Número de Instancias procesadas incorrectamente | 4 |
| Tiempo de construcción | 0.48 segundos |
| Exactitud promedio | 99.3% |
| Precisión promedio | 99.0% |
| Sensibilidad promedio | 98.9% |
| Especificidad promedio | 99.4% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| Clasificación | | a | b | c |
| Home | a | 179 | 2 | 0 |
| Away | b | 0 | 128 | 0 |
| Draw | c | 1 | 1 | 69 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| Resultado | Exactitud | Precisión | Sensibilidad | Especificidad |
| Home | 99.2% | 99.4% | 98.9% | 99.5% |
| Away | 99.2% | 97.7% | 100.0% | 98.8% |
| Draw | 99.5% | 100.0% | 97.2% | 100.0% |

Fuente: Elaboración propia

### 3.1.3. Cross-validation

Consiste en particionar la totalidad de elementos del archivo en *K* hojas, *K* - 1 conjuntos se utilizarán para entrenar y 1 conjunto para realizar las pruebas de rendimiento, por omisión *K* = 10. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 3104 victorias locales, 1874 victorias visitantes y 1708 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo LogitBoost.

Tabla 3 – Prueba Cross-validation aplicada al algoritmo LogitBoost

| Ficha de Observación: Evaluación del algoritmo LogitBoost |
|---|

| Resumen de Evaluación | |
|---|---|
| **Número de evaluación** | 3 |
| **Tipo de prueba** | Cross-validation |
| **Número de Instancias** | 6686 |
| **Número de Instancias procesadas correctamente** | 6591 |
| **Número de Instancias procesadas incorrectamente** | 95 |
| **Tiempo de construcción** | 0.42 segundos |
| **Exactitud promedio** | 99.1% |
| **Precisión promedio** | 98.6% |
| **Sensibilidad promedio** | 98.6% |
| **Especificidad promedio** | 99.0% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 3097 | 7 | 0 |
| **Away** | **b** | 15 | 1859 | 0 |
| **Draw** | **c** | 49 | 24 | 1635 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 98.9% | 98.0% | 99.8% | 98.2% |
| **Away** | 99.3% | 98.4% | 99.2% | 99.4% |
| **Draw** | 98.9% | 100.0% | 95.7% | 100.0% |

Fuente: Elaboración propia

### 3.1.4. Percentage Split

Consiste en particionar la totalidad de elementos del archivo en 2 conjuntos, 1 conjunto se utilizará para el entrenamiento y el otro para realizar las pruebas de rendimiento, por omisión se dividen en el 66% para entrenamiento y 34% para pruebas. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 1058 victorias locales, 611 victorias visitantes y 604 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los

resultados de la evaluación de rendimiento del algoritmo LogitBoost.

Tabla 4 – Prueba Percentage Split aplicada al algoritmo LogitBoost

| Ficha de Observación: Evaluación del algoritmo LogitBoost |
|---|

| Resumen de Evaluación | |
|---|---|
| Número de evaluación | 4 |
| Tipo de prueba | Percentage Split |
| Número de Instancias | 2273 |
| Número de Instancias procesadas correctamente | 2229 |
| Número de Instancias procesadas incorrectamente | 44 |
| Tiempo de construcción | 0.39 segundos |
| Exactitud promedio | 98.7% |
| Precisión promedio | 98.2% |
| Sensibilidad promedio | 98.1% |
| Especificidad promedio | 99.3% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| Clasificación | | a | b | c |
| Home | a | 1038 | 20 | 0 |
| Away | b | 0 | 611 | 0 |
| Draw | c | 0 | 24 | 580 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| Resultado | Exactitud | Precisión | Sensibilidad | Especificidad |
| Home | 99.1% | 100.0% | 98.1% | 100.0% |
| Away | 98.1% | 93.3% | 100.0% | 97.4% |
| Draw | 98.9% | 100.0% | 96.0% | 100.0% |

Fuente: Elaboración propia

## 3.2. ClassViaReg

### 3.2.1. Training Test

Consiste en utilizar la totalidad de elementos del archivo con el que se entrenó el modelo para realizar las pruebas de rendimiento. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 3104 victorias locales, 1874 victorias visitantes y 1708 empates. En la siguiente ficha de observación se registró la

prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo ClassificationViaRegression.

Tabla 5 – Prueba Training Test aplicada al algoritmo ClassViaReg

| Ficha de Observación: Evaluación del algoritmo ClassViaReg | | | | |
|---|---|---|---|---|
| **Resumen de Evaluación** | | | | |
| **Número de evaluación** | | | | 1 |
| **Tipo de prueba** | | | | Training Set |
| **Número de instancias** | | | | 6686 |
| **Número de instancias procesadas correctamente** | | | | 6685 |
| **Número de instancias procesadas incorrectamente** | | | | 1 |
| **Tiempo de construcción** | | | | 2.53 segundos |
| **Exactitud promedio** | | | | 99.9% |
| **Precisión promedio** | | | | 99.9% |
| **Sensibilidad promedio** | | | | 99.9% |
| **Especificidad promedio** | | | | 99.9% |
| **Matriz de Confusión** | | | | |
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 3104 | 0 | 0 |
| **Away** | **b** | 0 | 1874 | 0 |
| **Draw** | **c** | 0 | 1 | 1707 |
| **Evaluación de Rendimiento** | | | | |
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Away** | 100.0% | 99.9% | 100.0% | 100.0% |
| **Draw** | 100.0% | 100.0% | 99.9% | 100.0% |

Fuente: Elaboración propia

### 3.2.2. Supplied Test Set

Consiste en utilizar un archivo con datos de prueba distintos a los elementos del archivo con el que se entrenó el modelo para realizar las pruebas de rendimiento. La matriz de entrada tuvo 24 columnas y 381 filas (1 fila de encabezados). Se esperaba obtener una salida de 181 victorias locales, 128 victorias visitantes y 71 empates. En la siguiente ficha de

observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo ClassificationViaRegression.

Tabla 6 – Prueba Supplied Test Set aplicada al algoritmo ClassViaReg

| Ficha de Observación: Evaluación del algoritmo ClassViaReg |
|---|

| Resumen de Evaluación | |
|---|---|
| Número de evaluación | 2 |
| Tipo de prueba | Supplied Test Set |
| Número de Instancias | 380 |
| Número de Instancias procesadas correctamente | 380 |
| Número de Instancias procesadas incorrectamente | 0 |
| Tiempo de construcción | 2.00 segundos |
| Exactitud promedio | 100.0% |
| Precisión promedio | 100.0% |
| Sensibilidad promedio | 100.0% |
| Especificidad promedio | 100.0% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| Clasificación | | a | b | c |
| Home | a | 181 | 0 | 0 |
| Away | b | 0 | 128 | 0 |
| Draw | c | 0 | 0 | 71 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| Resultado | Exactitud | Precisión | Sensibilidad | Especificidad |
| Home | 100.0% | 100.0% | 100.0% | 100.0% |
| Away | 100.0% | 100.0% | 100.0% | 100.0% |
| Draw | 100.0% | 100.0% | 100.0% | 100.0% |

Fuente: Elaboración propia

### 3.2.3. Cross-validation

Consiste en particionar la totalidad de elementos del archivo en *K* hojas, *K* - 1 conjuntos se utilizarán para entrenar y 1 conjunto para realizar las pruebas de rendimiento, por omisión *K* = 10. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de

3104 victorias locales, 1874 victorias visitantes y 1708 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo ClassificationViaRegression.

Tabla 7 – Prueba Cross-validation aplicada al algoritmo ClassViaReg

| Ficha de Observación: Evaluación del algoritmo ClassViaReg |
|---|

| Resumen de Evaluación | |
|---|---|
| Número de evaluación | 3 |
| Tipo de prueba | Cross-validation |
| Número de Instancias | 6686 |
| Número de Instancias procesadas correctamente | 6576 |
| Número de Instancias procesadas incorrectamente | 10 |
| Tiempo de construcción | 1.95 segundos |
| Exactitud promedio | 99.9% |
| Precisión promedio | 99.9% |
| Sensibilidad promedio | 99.9% |
| Especificidad promedio | 99.9% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| Clasificación | | a | b | c |
| Home | a | 3102 | 0 | 2 |
| Away | b | 0 | 1872 | 2 |
| Draw | c | 0 | 6 | 1702 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| Resultado | Exactitud | Precisión | Sensibilidad | Especificidad |
| Home | 100.0% | 100.0% | 99.9% | 100.0% |
| Away | 99.9% | 99.7% | 99.9% | 99.9% |
| Draw | 99.9% | 99.8% | 99.6% | 99.9% |

Fuente: Elaboración propia

### 3.2.4. Percentage Split

Consiste en particionar la totalidad de elementos del archivo en 2 conjuntos, 1 conjunto se utilizará para el entrenamiento y el otro para realizar las pruebas de rendimiento, por omisión se dividen en el 66% para entrenamiento y 34% para pruebas.

La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 1058 victorias locales, 611 victorias visitantes y 604 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo ClassificationViaRegression.

Tabla 8 – Prueba Percentage Split aplicada al algoritmo ClassViaReg

| Ficha de Observación: Evaluación del algoritmo ClassViaReg |
|---|

| Resumen de Evaluación | |
|---|---|
| **Número de evaluación** | 4 |
| **Tipo de prueba** | Percentage Split |
| **Número de Instancias** | 2273 |
| **Número de Instancias procesadas correctamente** | 2270 |
| **Número de Instancias procesadas incorrectamente** | 3 |
| **Tiempo de construcción** | 2.22 segundos |
| **Exactitud promedio** | 99.9% |
| **Precisión promedio** | 99.9% |
| **Sensibilidad promedio** | 99.9% |
| **Especificidad promedio** | 99.9% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 1058 | 0 | 0 |
| **Away** | **b** | 0 | 611 | 0 |
| **Draw** | **c** | 3 | 0 | 601 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 99.9% | 99.7% | 100.0% | 99.8% |
| **Away** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Draw** | 99.9% | 100.0% | 99.5% | 100.0% |

Fuente: Elaboración propia

## 3.3. Logistic Regression

### 3.3.1. Training Test

Consiste en utilizar la totalidad de elementos del archivo con el que se entrenó el modelo para realizar las pruebas de rendimiento. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 3104 victorias locales, 1874 victorias visitantes y 1708 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo Logistic Regression.

Tabla 9 – Prueba Training Test aplicada al algoritmo Logistic Regression

| Ficha de Observación: Evaluación del algoritmo Logistic Regression |
|---|

| Resumen de Evaluación | |
|---|---|
| **Número de evaluación** | 1 |
| **Tipo de prueba** | Training Set |
| **Número de instancias** | 6686 |
| **Número de instancias procesadas correctamente** | 6686 |
| **Número de instancias procesadas incorrectamente** | 0 |
| **Tiempo de construcción** | 5.25 segundos |
| **Exactitud promedio** | 100.0% |
| **Precisión promedio** | 100.0% |
| **Sensibilidad promedio** | 100.0% |
| **Especificidad promedio** | 100.0% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 3104 | 0 | 0 |
| **Away** | **b** | 0 | 1874 | 0 |
| **Draw** | **c** | 0 | 0 | 1708 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Away** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Draw** | 100.0% | 100.0% | 100.0% | 100.0% |

Fuente: Elaboración propia

### 3.3.2. Supplied Test Set

Consiste en utilizar un archivo con datos de prueba distintos a los elementos del archivo con el que se entrenó el modelo para realizar las pruebas de rendimiento. La matriz de entrada tuvo 24 columnas y 381 filas (1 fila de encabezados). Se esperaba obtener una salida de 181 victorias locales, 128 victorias visitantes y 71 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo Logistic Regression.

Tabla 10 – Prueba Supplied Test Set aplicada al algoritmo Logistic Regression

| **Ficha de Observación: Evaluación del algoritmo Logistic Regression** |
|---|

| **Resumen de Evaluación** | |
|---|---|
| **Número de evaluación** | 2 |
| **Tipo de prueba** | Supplied Test Set |
| **Número de Instancias** | 380 |
| **Número de Instancias procesadas correctamente** | 380 |
| **Número de Instancias procesadas incorrectamente** | 0 |
| **Tiempo de construcción** | 4.21 segundos |
| **Exactitud promedio** | 100.0% |
| **Precisión promedio** | 100.0% |
| **Sensibilidad promedio** | 100.0% |
| **Especificidad promedio** | 100.0% |

| **Matriz de Confusión** | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 181 | 0 | 0 |
| **Away** | **b** | 0 | 128 | 0 |
| **Draw** | **c** | 0 | 0 | 71 |

| **Evaluación de Rendimiento** | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Away** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Draw** | 100.0% | 100.0% | 100.0% | 100.0% |

Fuente: Elaboración propia

### 3.3.3. Cross-validation

Consiste en particionar la totalidad de elementos del archivo en *K* hojas, *K* - 1 conjuntos se utilizarán para entrenar y 1 conjunto para realizar las pruebas de rendimiento, por omisión *K* = 10. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 3104 victorias locales, 1874 victorias visitantes y 1708 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo Logistic Regression.

Tabla 11 – Prueba Cross-validation aplicada al algoritmo Logistic Regression

| Ficha de Observación: Evaluación del algoritmo Logistic Regression |
|---|

| Resumen de Evaluación | |
|---|---|
| **Número de evaluación** | 3 |
| **Tipo de prueba** | Cross-validation |
| **Número de Instancias** | 6686 |
| **Número de Instancias procesadas correctamente** | 6681 |
| **Número de Instancias procesadas incorrectamente** | 5 |
| **Tiempo de construcción** | 4.21 segundos |
| **Exactitud promedio** | 99.9% |
| **Precisión promedio** | 99.9% |
| **Sensibilidad promedio** | 99.9% |
| **Especificidad promedio** | 99.9% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 3103 | 0 | 1 |
| **Away** | **b** | 0 | 1871 | 3 |
| **Draw** | **c** | 0 | 1 | 1707 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Away** | 99.9% | 99.9% | 99.8% | 100.0% |
| **Draw** | 99.9% | 99.8% | 99.9% | 99.9% |

Fuente: Elaboración propia

### 3.3.4. Percentage Split

Consiste en particionar la totalidad de elementos del archivo en 2 conjuntos, 1 conjunto se utilizará para el entrenamiento y el otro para realizar las pruebas de rendimiento, por omisión se dividen en el 66% para entrenamiento y 34% para pruebas. La matriz de entrada tuvo 24 columnas y 6687 filas (1 fila de encabezados). Se esperaba obtener una salida de 1058 victorias locales, 611 victorias visitantes y 604 empates. En la siguiente ficha de observación se registró la prueba aplicada, se incluye: El resumen de análisis, la matriz de confusión y los resultados de la evaluación de rendimiento del algoritmo Logistic Regression.

Tabla 12 – Prueba Percentage Split aplicada al algoritmo Logistic Regression

| Ficha de Observación: Evaluación del algoritmo Logistic Regression |
|---|

| Resumen de Evaluación | |
|---|---|
| **Número de evaluación** | 4 |
| **Tipo de prueba** | Percentage Split |
| **Número de Instancias** | 2273 |
| **Número de Instancias procesadas correctamente** | 2272 |
| **Número de Instancias procesadas incorrectamente** | 1 |
| **Tiempo de construcción** | 4.26 segundos |
| **Exactitud promedio** | 99.9% |
| **Precisión promedio** | 99.9% |
| **Sensibilidad promedio** | 99.9% |
| **Especificidad promedio** | 99.9% |

| Matriz de Confusión | | | | |
|---|---|---|---|---|
| **Clasificación** | | **a** | **b** | **c** |
| **Home** | **a** | 1058 | 0 | 0 |
| **Away** | **b** | 0 | 610 | 1 |
| **Draw** | **c** | 0 | 0 | 604 |

| Evaluación de Rendimiento | | | | |
|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% |
| **Away** | 100.0% | 100.0% | 99.8% | 100.0% |
| **Draw** | 100.0% | 99.8% | 100.0% | 99.9% |

Fuente: Elaboración propia

# IV. DISCUCIÓN

A continuación, se presenta un resumen del rendimiento para cada una de las pruebas aplicadas en los algoritmos estudiados: LogitBoost, ClassViaReg, Logistic Regression.

## 4.1. Resumen para LogitBoost

Tabla 13 – Resumen del análisis de rendimiento del algoritmo LogitBoost

| Ficha Resumen: Rendimiento del algoritmo LogitBoost |
|---|

| Según la prueba realizada | | | | | |
|---|---|---|---|---|---|
| **Prueba** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **Training Set** | 99.1% | 98.6% | 98.6% | 99.1% | **98.85%** |
| **Supplied Test Set** | 99.3% | 99.0% | 98.9% | 99.4% | **99.15%** |
| **Cross-validation** | 99.1% | 98.6% | 98.6% | 99.0% | **98.83%** |
| **Percentage Split** | 98.7% | 98.2% | 98.1% | 99.3% | **98.58%** |

| Según el resultado del partido con el tipo de prueba Supplied Test Set | | | | | |
|---|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **Home** | 99.2% | 99.4% | 98.9% | 99.5% | **99.25%** |
| **Away** | 99.2% | 97.7% | 100.0% | 98.8% | **98.93%** |
| **Draw** | 99.5% | 100.0% | 97.2% | 100.0% | **99.18%** |

Fuente: Elaboración propia

Tomando como referencia la Tabla 13, se procedió a elaborar un gráfico para cada clase de análisis, posteriormente se realizó la interpretación de los resultados.

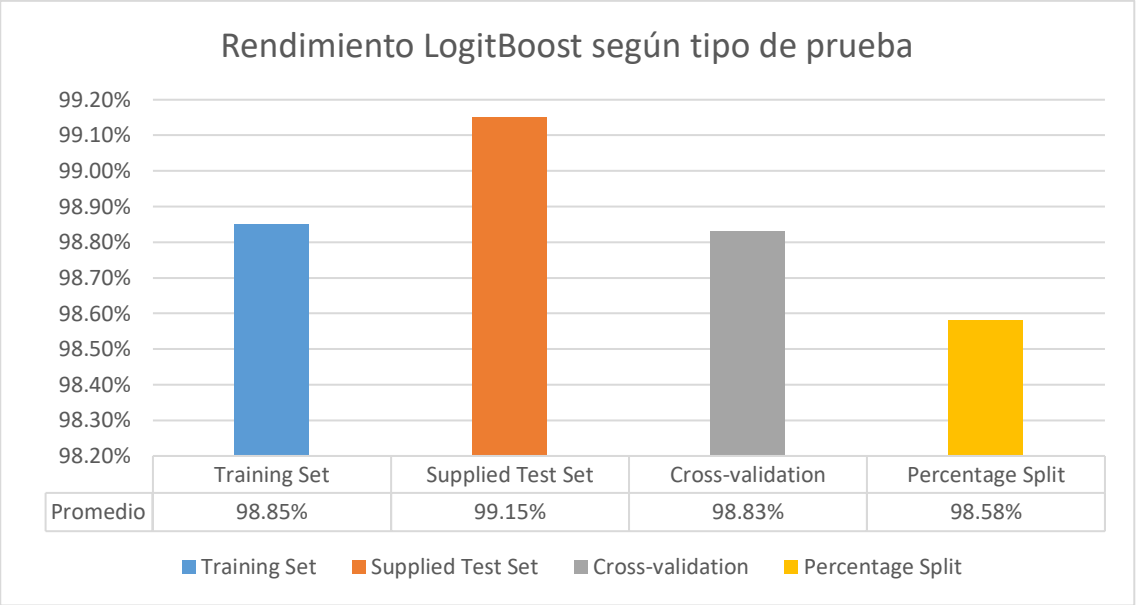


Ilustración 1 - Rendimiento del algoritmo LogitBoost según el tipo de prueba

Fuente: Elaboración propia

La Ilustración 1 resume el rendimiento del algoritmo LogitBoost según el tipo de prueba aplicado, este resultado ha sido obtenido del promedio general de las evaluaciones realizadas por cada tipo de prueba. Se concluye que el algoritmo presenta el mejor rendimiento en las pruebas de Supplied Test Set, alcanzando un 99.15% de efectividad, 0.30%, 0.32% y 0,57% más que en las pruebas de Training Set, Cross-validation y Percentage Split respectivamente.

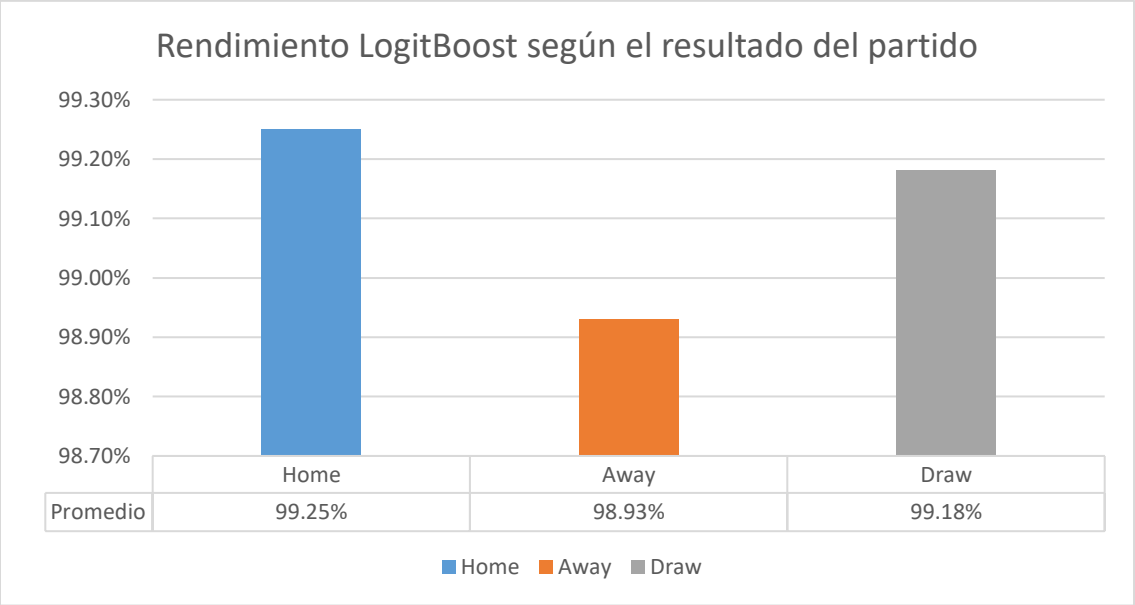


Ilustración 2 - Rendimiento del algoritmo LogitBoost según el resultado del partido

Fuente: Elaboración propia

La Ilustración 2 resume el rendimiento del algoritmo LogitBoost según el resultado del partido, este resultado ha sido obtenido del promedio general de las evaluaciones hechas por cada tipo de resultado mediante la prueba de Supplied Test Set. Se concluye que el algoritmo presenta el mejor rendimiento en la identificación de las victorias locales (Home) alcanzando un 99.25% de efectividad, 0.32% y 0,07% más que prediciendo las victorias visitantes (Away) y los empates (Draw) respectivamente.

## 4.2. Resumen para ClassViaReg

Tabla 14 – Resumen del análisis de rendimiento del algoritmo ClassViaReg

| Ficha Resumen: Rendimiento del algoritmo ClassViaReg |
|---|

| Según la prueba realizada | | | | | |
|---|---|---|---|---|---|
| **Prueba** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **Training Set** | 99.9% | 99.9% | 99.9% | 99.9% | **99.9%** |
| **Supplied Test Set** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Cross-validation** | 99.9% | 99.9% | 99.9% | 99.9% | **99.9%** |
| **Percentage Split** | 99.9% | 99.9% | 99.9% | 99.9% | **99.9%** |

| Según el resultado obtenido con el tipo de prueba Supplied Test Set | | | | | |
|---|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Away** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Draw** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |

Fuente: Elaboración propia

Tomando como referencia la Tabla 14, se procedió a elaborar un gráfico para cada clase de análisis, posteriormente se realizó la interpretación de los resultados.

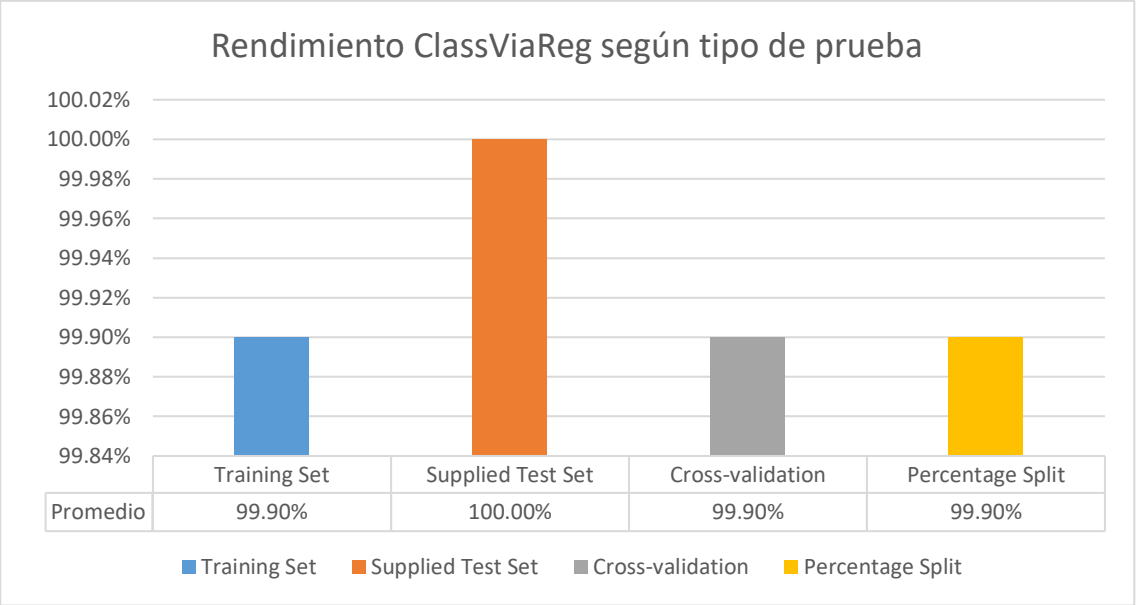


Ilustración 3 - Rendimiento del algoritmo ClassViaReg según el tipo de prueba

Fuente: Elaboración propia

La Ilustración 3 resume el rendimiento del algoritmo ClassViaReg según el tipo de prueba aplicado, este resultado ha sido obtenido del promedio general de las evaluaciones realizadas por cada tipo de prueba. Se concluye que el algoritmo presenta el mejor rendimiento en las pruebas de Supplied Test Set, alcanzando un 100.00% de efectividad, 0.10% más que en las otras 3 pruebas realizadas.

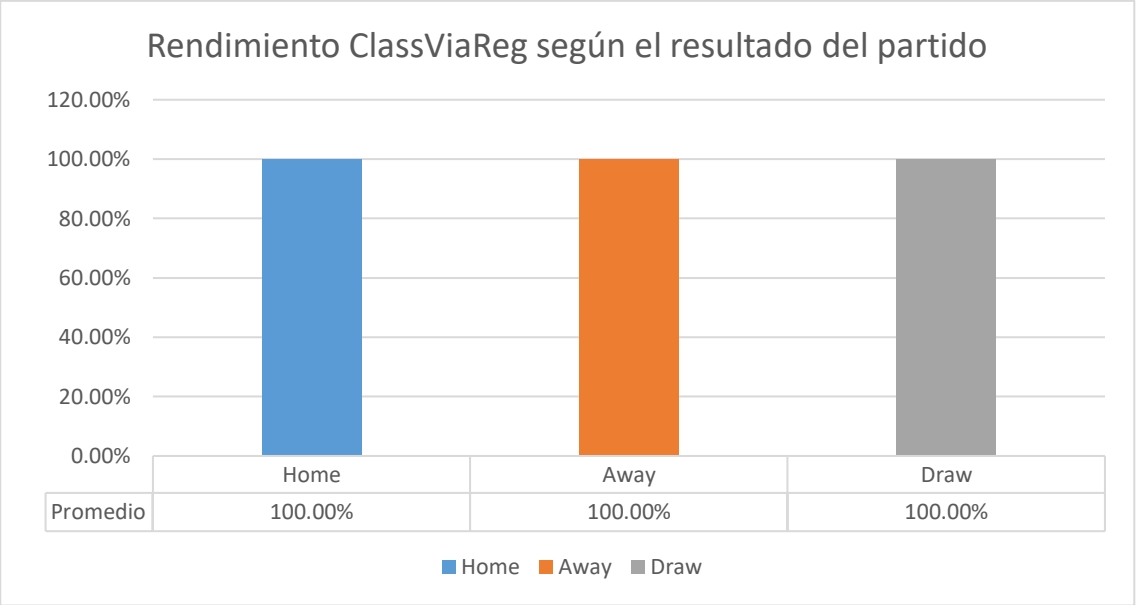


Ilustración 4 - Rendimiento del algoritmo ClassViaReg según el resultado del partido

Fuente: Elaboración propia

La Ilustración 4 resume el rendimiento del algoritmo ClassViaReg según el resultado del partido, este resultado ha sido obtenido del promedio general de las evaluaciones hechas por cada tipo de resultado mediante la prueba de Supplied Test Set. Se concluye que el algoritmo presenta el mismo rendimiento en la identificación de: victorias locales (Home), victorias visitantes (Away) y empates (Draw).

## 4.3. Resumen para Logistic Regression

Tabla 15 – Resumen del análisis de rendimiento del algoritmo Logistic Regression

| Ficha Resumen: Rendimiento del algoritmo Logistic Regression |
|---|

| Según la prueba realizada | | | | | |
|---|---|---|---|---|---|
| **Prueba** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **Training Set** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Supplied Test Set** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Cross-validation** | 99.9% | 99.9% | 99.9% | 99.9% | **99.9%** |
| **Percentage Split** | 99.9% | 99.9% | 99.9% | 99.9% | **99.9%** |

| Según el resultado obtenido con el tipo de prueba Supplied Test Set | | | | | |
|---|---|---|---|---|---|
| **Resultado** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **Home** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Away** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |
| **Draw** | 100.0% | 100.0% | 100.0% | 100.0% | **100.0%** |

Fuente: Elaboración propia

Tomando como referencia la Tabla 15, se procedió a elaborar un gráfico para cada clase de análisis, posteriormente se realizó la interpretación de los resultados.

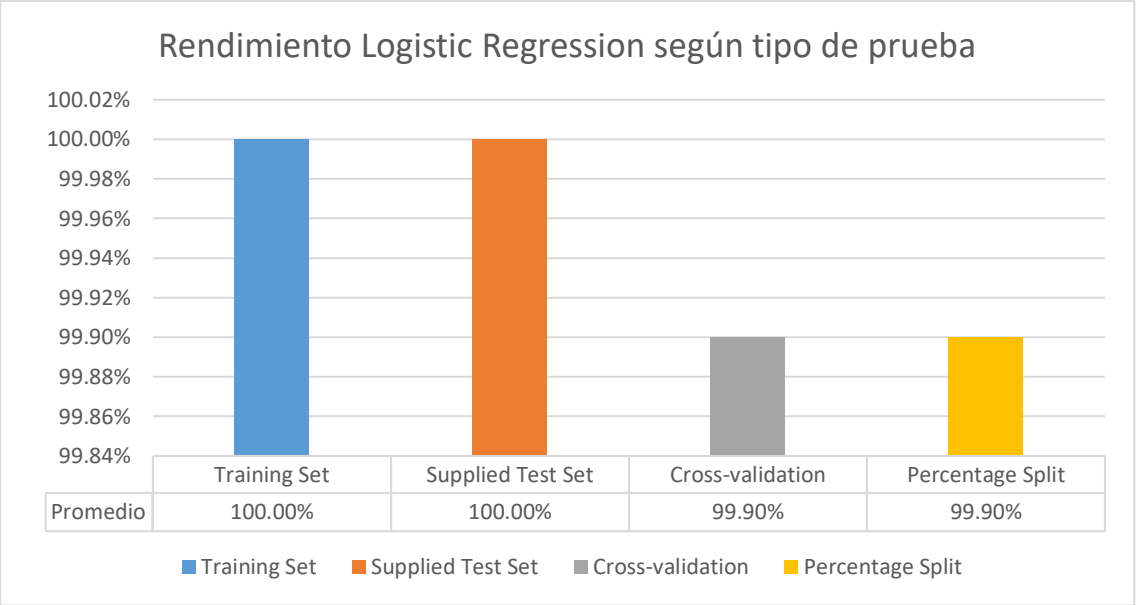


Ilustración 5 - Rendimiento del algoritmo Logistic Regression según el tipo de prueba
Fuente: Elaboración propia

La Ilustración 5 resume el rendimiento del algoritmo Logistic Regression según el tipo de prueba aplicado, este resultado ha sido obtenido del promedio general de las evaluaciones realizadas por cada tipo de prueba. Se concluye que el algoritmo presenta el mejor rendimiento en las pruebas de Training Set y Supplied Test Set, alcanzando un 100.00% de efectividad en ambas, 0.10% más que en las otras 2 pruebas realizadas.

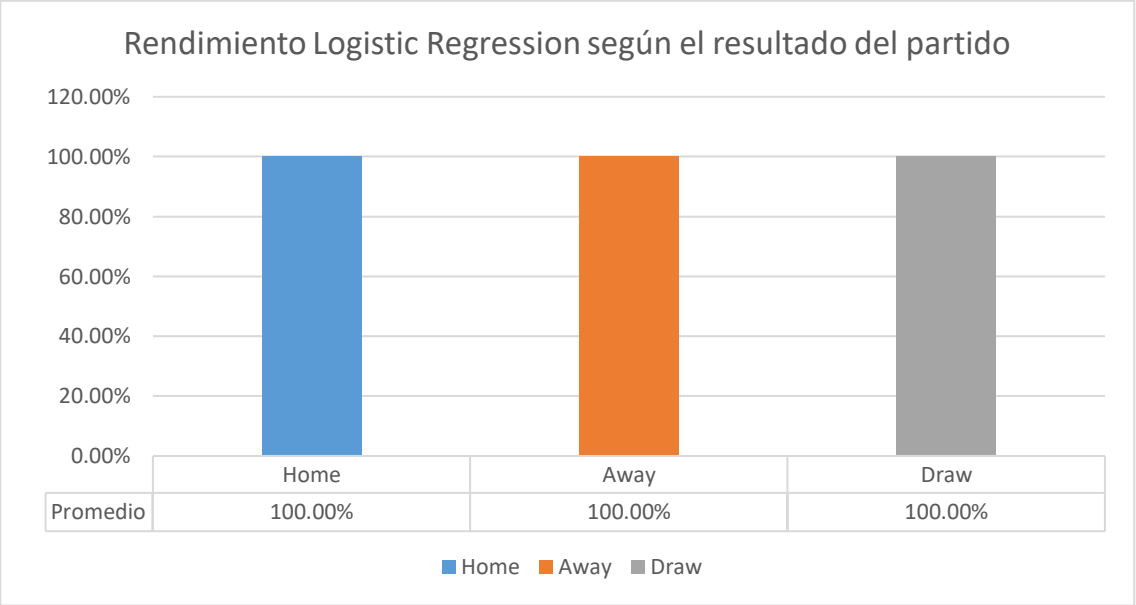


Ilustración 6 - Rendimiento del algoritmo Logistic Regression según el resultado del partido

Fuente: Elaboración propia

La Ilustración 6 resume el rendimiento del algoritmo Logistic Regression según el resultado del partido, este resultado ha sido obtenido del promedio general de las evaluaciones hechas por cada tipo de resultado mediante la prueba de Supplied Test Set. Se concluye que el algoritmo presenta el mismo rendimiento en la identificación de: victorias locales (Home), victorias visitantes (Away) y empates (Draw).

Finalmente se realizó una tabla comparativa del rendimiento de los 3 algoritmos estudiados para definir que algoritmo ofrece el mejor rendimiento.

Tabla 16 – Análisis comparativo del rendimiento de los algoritmos estudiados

| Ficha Resumen<br>Comparación de los algoritmos LogitBoost, ClassViaReg y Logistic Regression<br>Mediante la prueba de Supplied Test Set | | | | | |
|---|---|---|---|---|---|
| **Algoritmo** | **Exactitud** | **Precisión** | **Sensibilidad** | **Especificidad** | **Promedio** |
| **LogitBoost** | 99.30% | 99.03% | 98.70% | 99.43% | 99.12% |
| **ClassViaReg** | 100.0% | 100.0% | 100.0% | 100.0% | 100.00% |
| **Logistic Regression** | 100.0% | 100.0% | 100.0% | 100.0% | 100.00% |

Fuente: Elaboración propia

Según la Tabla 16, se concluye que los algoritmos que ofrecen mejor rendimiento para la predicción de resultados de partidos de fútbol son: ClassificationViaRegression y Logistic Regression, los cuales presentan un promedio equivalente al 100.00%, en comparación con LogitBoost que presentó un rendimiento equivalente al 99.12%, lo que representa una diferencia de 0.88%.

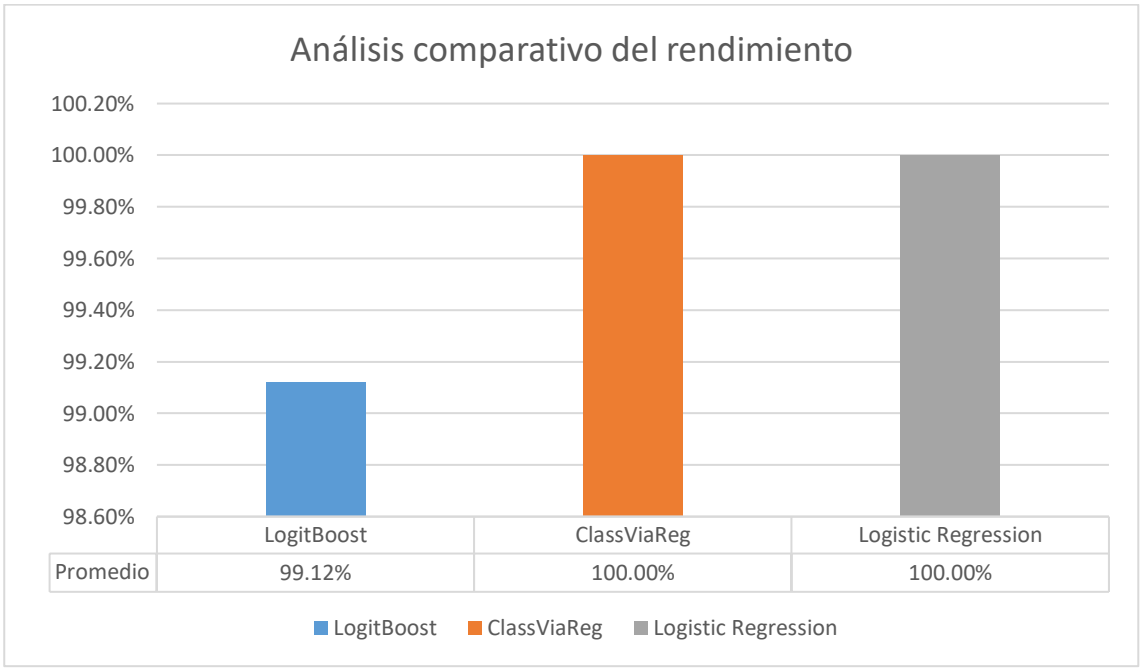


Ilustración 7 - Análisis comparativo del rendimiento de los algoritmos estudiados

Fuente: Elaboración propia

# V. CONCLUSIONES

En la presente investigación se analizaron 3 algoritmos: LogitBoost, ClassViaReg y Logistic Regression, para ello se obtuvieron 18 archivos “CSV” correspondientes cada uno a una temporada de la Premier League, habiendo sido utilizadas las temporadas del 2000/2001 hasta la temporada 2017/2018, estos 18 archivos fueron pre procesados, eliminando información irrelevante, dando como resultado un único archivo con 6687 filas, siendo la primera fila para los encabezados y las siguientes 6686 correspondientes a partidos de fútbol. Haciendo uso de la herramienta WEKA se pudo concluir que los algoritmos de ClassViaReg y Logistic Regression presentan resultados muy similares, entorno al 99.9% y 100.0%, siendo el algoritmo LogitBoost el que presentó un rendimiento ligeramente inferior, en torno al 99.12%.

## VI. REFERENCIAS


Brooks, J., Matthew, K., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(5), 338–349. https://doi.org/10.1002/sam.11318

Bunge, M. (2004). *La Investigacion Cientifica - Su Estrategia Y Su Filosofia*. *Siglo Veintiuno*.

Buursma, D. (2011). Predicting sports events from past results. Towards effective betting on football matches. *14th Twente Student Conference on IT*.

Constantinou, A. C., & Fenton, N. E. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, *124*, 93–104. https://doi.org/10.1016/j.knosys.2017.03.005

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms*. https://doi.org/10.1163/9789004256064_hao_introduction

Duarte, L. (2015). 1X2 – Previsão de Resultados de Jogos de Futebol, 59.

FIFA. (2018). FIFA Official Documents. Retrieved December 1, 2017, from http://www.fifa.com/about-fifa/official-documents/index.html

Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann, Fourth Edition*, 128.

https://doi.org/10.1016/B978-0-12-804291-5.00024-6

Hamido, F., Jun, H., & Kurematsu, M. (2011). *Intelligent Information and Database Systems*.

Hucaljuk, J., & Rakipovic, A. (2011). Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO*, *48*, 1623–1627.

Igiri, C. P. (2015). Support Vector Machine–Based Prediction System for a Football Match Result. https://doi.org/10.9790/0661-17332126

Igiri, C. P., & Nwachukwu, E. O. (2014). An Improved Prediction System for Football a Match Result. *IOSR Journal of Engineerin*, *04*(12), 12–20. Retrieved from www.iosrjen.org

Kohavi, R., & Provost, F. (1998). On Applied Research in Machine Learning. *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, *30*, 178–190. https://doi.org/10.1177/1403494813515131

Lugosi, G., & Cesa-Bianchi, N. (2006). *Prediction, learning, and games*. *Prediction, Learning, and Games* (Vol. 9780521841). https://doi.org/10.1017/CBO9780511546921

Olympic. (2012). All facts London 2012. Retrieved December 1, 2017, from https://www.olympic.org/london-2012

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering*, *5*(5), 812–815. https://doi.org/10.7763/IJCTE.2013.V5.802