Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

# Predicción del comportamiento de los jugadores en el fútbol argentino

**AUTOR: YOEL HERNAN DIAMENT**

**MENTOR: MELISA ELFENBAUM**

**[DICIEMBRE, 2020]**

## Resumen

Los equipos de fútbol compran y venden jugadores todos los años en cada periodo de fichajes, donde gastan mucho dinero en cada transferencia alcanzando cifras millonarias. Dado que en los últimos años se fue incrementando el dinero utilizado en la adquisición de nuevos jugadores, cada equipo se plantea maximizar la eficacia de cada transferencia.

El interrogante que se plantea el presente trabajo consiste en establecer un criterio estadístico para comparar a todos los jugadores del fútbol argentino y así contribuir de una forma objetiva en el mercado de pases.

El objetivo general es analizar el comportamiento de los jugadores en el fútbol argentino durante los últimos 15 años para predecir si van a tener un buen o mal comportamiento futuro. Para ello, se va a analizar las métricas en cada temporada y los fichajes en el mercado de pases mundial. Luego se va a elaborar un modelo para predecir si el jugador va a ser transferido a un equipo de mayor jerarquía en la temporada siguiente.

Para resolver el problema de clasificación, se utilizaron tres técnicas distintas para la creación de los modelos: Regresión logística, *Random Forest* y Redes neuronales. El mejor modelo alcanzado fue la regresión logística, obteniendo la mayor exactitud y la menor varianza.

Este trabajo será de gran utilidad para conocer cuáles son los jugadores del fútbol argentino que tienen mayor probabilidad de ser transferidos a un equipo de mayor jerarquía a nivel internacional. Por lo tanto, a los responsables en la toma de decisiones de los fichajes de los jugadores en el fútbol argentino le pueden ser de gran utilidad los resultados del modelo.



**Palabras clave: Fútbol - Predicción - Argentina - Fichajes - Machine Learning**

# Índice

# 1- Introducción

Los equipos de fútbol compran y venden jugadores todos los años en cada periodo de fichajes, donde gastan mucho dinero en cada transferencia alcanzando cifras millonarias. Dado que en los últimos años se fue incrementando el dinero utilizado en la adquisición de nuevos jugadores, cada equipo se plantea maximizar la eficacia de cada transferencia. Por lo tanto, tomar buenas decisiones en los fichajes puede generar grandes beneficios económicos y deportivos para cada club. Por el contrario, malas decisiones pueden generar grandes pérdidas económicas y mal rendimiento futbolístico del equipo en el campo de juego.

El análisis estadístico aplicado a deportes ha sido exitoso en el baseball y en el basketball, pero la aplicación en el fútbol ha sido limitada (Kumar, 2013). Dado que el fútbol es un deporte en equipo, no es fácilmente identificar si las métricas de cada jugador son por mérito propio o consecuencia del juego en equipo. Las responsabilidades de cada posición son distintas, por lo cual los indicadores del buen comportamiento van a variar por posición. Sin embargo, diversas organizaciones y analistas de diferentes países han estudiado las variables que impactan en el precio de venta de los jugadores de fútbol utilizando modelos predictivos, análisis multivariante y técnicas de *machine learning*.

CIES Football Observatory realizó un estudio en octubre 2018 con el objetivo de predecir el valor de las transferencias de los jugadores. La base de datos utilizada cuenta con 2400 transferencias del top 5 de las ligas europeas entre julio 2011 y agosto 2018. Se consideraron 36 atributos con información de cada jugador, como la duración del contrato, año de transferencia, precio del fichaje, nacionalidad y nivel económico del club. El algoritmo utilizado fue la regresión lineal múltiple en el cual obtuvieron un coeficiente de determinación del 0.86. El estudio concluye en que el modelo es útil para determinar el valor del fichaje y el salario inicial que va a tener el jugador (Poli, Ravenel, & Besson, 2018).

Otro estudio similar fue realizado por Yuan He. La base de datos cuenta con 5 temporadas de historia para cada jugador. La misma contiene variables descriptivas del jugador (nacionalidad, altura, fecha nacimiento y posición, entre otras), del comportamiento de cada temporada en su equipo (partidos jugados, goles y asistencias, entre otras) y del comportamiento en la selección nacional. El algoritmo utilizado fue la regresión Ridge y la métrica para medir la performance fue RMSE (Raíz del error cuadrático medio). El

estudio concluye en que el valor de mercado del último año es el atributo más predictivo. Los goles, asistencias y otros atributos similares también fueron altamente predictivos, mientras que la información personal del jugador y los ratios creados por el autor no fueron variables significativas (Y. He, 2012).

Otro estudio analizado fue "Football player's performance and market value" publicado en 2015 por Ricardo Cachuchino, Miao He y Arno Knobbe. Los autores buscaron entender la relación entre el comportamiento de los jugadores en las temporadas y su valor de mercado. La base de datos cuenta con información de los valores de mercado de TransferMarket y del comportamiento de los jugadores de la liga española en la temporada 2014-2015. Dado el distinto comportamiento que presentan los jugadores de acuerdo con su posición, los autores se enfocaron únicamente en los delanteros. Los atributos considerados fueron las faltas cometidas, tarjetas rojas y amarillas, disparos al arco desde distintas zonas del campo, goles, penales y otras variables relevantes en los delanteros. El algoritmo utilizado fue la regresión Lasso remarcando la importancia de elegir el lambda correcto y la selección de variables adecuada. La conclusión fue que la valuación económica del jugador depende de su comportamiento, pero su comportamiento no puede ser explicado por factores monetarios de fichajes (M. He, Cachucho, & Knobbe, 2015).

El presente trabajo aborda toda la información correspondiente a cada jugador, desde el comienzo de la carrera hasta su último torneo como profesional, para poder analizar y entender diversos patrones y tendencias que presenten los jugadores. Como hipótesis general se plantea que el rendimiento futuro de los jugadores depende de su posición en el campo de juego, sus métricas de cada temporada y las transferencias en el mercado de pases.

El objetivo general es analizar el comportamiento de los jugadores en el fútbol argentino durante los últimos 15 años para predecir si van a tener un buen o mal comportamiento futuro. Para ello, se va a analizar las métricas en cada temporada, las variables descriptivas de cada jugador y los fichajes en el mercado de pases mundial. Luego se va a elaborar un modelo para predecir si el jugador va a ser transferido a un equipo de mayor jerarquía en la temporada siguiente.

Este trabajo será de gran utilidad para conocer cuáles son los jugadores del fútbol argentino que tienen mayor probabilidad de ser transferidos a un equipo de mayor

jerarquía a nivel internacional. Por lo tanto, a los responsables en la toma de decisiones de los fichajes de los jugadores en el fútbol argentino le pueden ser de gran utilidad los resultados del modelo ya que les va a permitir apoyarse en fundamentos estadísticos, además de su conocimiento del deporte. A su vez, dichos resultados van a permitir observar de una forma objetiva a todos los jugadores.

Los interrogantes que se plantea el presente trabajo son: ¿Es posible establecer un criterio estadístico para comparar a todos los jugadores del fútbol argentino y así contribuir de una forma objetiva en el mercado de pases? ¿Cuáles son las principales variables que definen el valor monetario de cada jugador? ¿Estas variables varían para cada posición en el campo de juego?

Para ello, el trabajo estará dividido en varias secciones. Comenzará con el primer apartado donde se describe la gestión de los datos en el contexto de la organización, continúa con el segundo apartado donde se describe la metodología utilizada para recopilar, procesar y analizar la información, luego el tercer apartado donde se define como se va a implementar el modelo de aprendizaje automático y termina con las conclusiones obtenidas, donde quedarán expuestos todos los desarrollos y resultados obtenidos a partir de la investigación del presente trabajo.

# 2- Gestión de datos en contextos organizacionales

En esta sección, se describe el tipo de organización que va a ser desarrollado a lo largo del presente trabajo, indicando su objetivo, su composición y su modelo de negocios. Luego se presenta la gestión de datos que va a atravesar la organización indicando el proceso de toma de decisiones y los datos que utiliza en ese proceso. Por último, se encuentra la problemática de la organización y la gestión de los datos a la que se enfrenta en la actualidad.

## 2.1 – Descripción de la organización

Las predicciones acerca de quienes son los jugadores que tienen más probabilidad de ser transferidos a un equipo de mayor jerarquía, serian de gran utilidad para una organización que asesore a instituciones deportivas indicando quienes son los jugadores ideales para adquirir y/o vender de acuerdo con el presupuesto y necesidades de cada institución. El tipo de organización que le resultaría relevante incorporar dichas predicciones es una consultora deportiva. La misma es una organización de carácter privada que tiene como

objetivo analizar grandes volúmenes de datos del fútbol para asesorar a sus clientes a tomar decisiones basadas en datos.

La consultoría deportiva que brinda la organización es un servicio de asesoría especializada e independiente al que recurren las instituciones deportivas con el fin de encontrar soluciones a uno o más de sus problemas o necesidades deportivas, que se sustenta en la innovación, la experiencia, el conocimiento, las habilidades de los profesionales, los métodos y las herramientas. Los potenciales clientes de esta organización son instituciones deportivas, ligas de fútbol a nivel nacional, federaciones deportivas, representantes de jugadores y cualquier profesional del fútbol que necesite información futbolística para tomar decisiones. A su vez, dichos clientes podrían estar a lo largo de todo el mundo ya que los jugadores pueden ser transferidos tanto a nivel nacional como a nivel internacional.

El objetivo de la organización es obtener grandes volúmenes de datos del fútbol a nivel mundial para luego asesorar a la mayor cantidad de instituciones deportivas. Dicho asesoramiento es variado para cada equipo ya que se adecúa a las necesidades de cada cliente. La organización tiene conocimiento en brindar asesoramiento acerca de transferencias en el mercado de pases, analizar jugadores con gran potencial, analizar el reemplazo ideal para un determinado jugador y análisis de equipos.

El modelo de negocio de la organización es lineal. La misma toma insumos de páginas webs públicas, los procesa y luego los analiza con el fin de conocer a todos los equipos y puntualmente a los jugadores mediante diversas métricas. Luego crea su propio modelo de aprendizaje automático junto a otros análisis de los datos obtenidos y ofrece a sus clientes los resultados de su modelo predictivo junto a otros servicios personalizados.

La organización toma decisiones de negocio basadas en datos definiéndose como "*data driven decisión making*" consistiendo en convertir el análisis de datos en decisiones estratégicas y acciones a futuro. La organización está atravesada por los datos, los cuales influyen en todas las decisiones adoptadas; los métodos empleados son sofisticados, científicos, basados en rigurosidad estadística, econométrica y en conceptos de ciencias de datos. Algunas de las ventajas de es este tipo de organizaciones son la eficiencia de recursos, reducción de costos y mejoras operativas (Penn, 2019). A su vez, Provost & Fawcett (2013) menciona que "*data driven decisión making*" se refiere a que las decisiones deben ser basadas en análisis y no por la intuición. De esta forma, "comprender

este proceso y sus etapas, ayuda a estructurar la resolución de problemas, la hace más sistemática y, por lo tanto, es menos propensa a cometer errores" (Provost & Fawcett, 2013).

Lavalle, Lesser & Shockley (2011) mencionan en su estudio que el rendimiento de una organización depende del valor que brindan sus analíticos. Por lo tanto, si se toman decisiones a nivel gerencial basándose en sus analíticos, van a obtener mejores resultados. A su vez, los autores afirman que obtener los datos correctamente no es el mayor obstáculo de una organización al incorporar analíticos, sino que el nivel gerencial y la cultura organizacional se adapte a estos analíticos (Lavalle, Lesser & Shockley, 2011) Este obstáculo no afecta a la organización que se define en el presente trabajo debido a que la misma nace como una organización atravesada por los datos desde sus comienzos y no tiene que modificar procesos internos para adaptarse a estos cambios.

La organización está compuesta por personas que tienen distintas características y habilidades. A nivel general, se encuentran los que tienen gran conocimiento en la extracción, transformación y análisis de datos. Por otro lado, se encuentran los que tienen gran conocimiento del fútbol. La combinación de estas habilidades logra que la organización sea un gran atractivo para sus clientes.

Entrando en detalle de los profesionales con los que cuenta la organización, la misma está compuesta por *data engineers*, *data scientists, machine learning engineers* y especialistas de fútbol como directores técnicos. Los *data engineers* se encargan de obtener los de datos de distintas fuentes de información y almacenarlos en una base de datos para que los *data scientists* puedan tomar esos datos para sus análisis posteriores. Por otra parte, los *data scientists* se encargan de analizar la información que se encuentra almacenada en la base de datos para brindar soluciones a los distintos problemas que se plantea resolver la propia organización y también a los que plantea cada cliente en particular. A su vez, los especialistas del fútbol trabajan en conjunto con los *data scientist* ya que son los que entienden a la perfección todas las métricas que están almacenadas en la base de datos y aportan su conocimiento en el deporte para obtener mejores análisis de los datos. Por otra parte, los *machine learning engineers* se encargan de buscar cual es el mejor algoritmo de aprendizaje automático para resolver cada problema que plantea la organización y a su vez, calibrar constantemente los parámetros de los modelos que están en funcionamiento.

## 2.2 – Gestión de datos por parte de la organización

La obtención de datos se hace mediante una técnica de programación denominada web scraping debido a que los datos se encuentran en diferentes páginas web de dominio público. La técnica consiste en adquirir datos no estructurados que se encuentren en sitios web y almacenarlos en una base de datos de forma estructurada.

Para que las predicciones a realizar sobre los jugadores de fútbol tengan relevancia estadística es necesario obtener datos de una gran cantidad de años. Por lo tanto, es necesario que el proceso de web scraping sea flexible para obtener un gran volumen de datos.

Luego de que los datos se encuentren estructurados en una base de datos, es necesario realizar un procesamiento y limpieza de los datos. Esta etapa es indispensable y es de las tareas más importantes ya que va a permitir que los datos estén disponibles para utilizar en los análisis posteriores. En esta etapa se pueden detectar errores en el proceso de web scraping realizado anteriormente.

Luego de que los datos ya estén procesados y almacenados correctamente en una base de datos, se realiza un análisis exploratorio de los datos para conocer la distribución y el comportamiento de cada una de las variables. Es necesario identificar que variables tienen valores faltantes y en qué proporción del total, ya que podría ser un inconveniente cuando se usen esos datos en los algoritmos de aprendizaje automático.

Para obtener las mejores predicciones posibles en cada jugador es deseable obtener la mayor cantidad de variables. Por lo tanto, los datos se corresponden con variables del comportamiento en el campo de juego como también variables sociodemográficas. Algunos ejemplos de variables sobre el comportamiento son partidos jugados, goles, asistencias, tarjetas amarillas, tarjetas rojas, partidos ganados, partidos perdidos y partidos empatados, entre otras. Por su parte, las variables que describen al jugador son la nacionalidad, fecha de nacimiento, altura, peso, pie hábil y posición, entre otras.

Como se mencionó anteriormente, la organización brinda asesoramiento personalizado según las necesidades de cada cliente. El principal potencial de la organización es el modelo propio de aprendizaje automático, el cual asigna a cada jugador un score indicando la probabilidad que tiene cada jugador de ser transferido a un equipo de mayor jerarquía en la temporada posterior. Por lo tanto, ofrecer este score a sus clientes es un gran diferencial frente a otras organizaciones que ofrecen los mismos servicios de

asesoramiento. A partir de dicho score, la organización brinda asesoramiento acerca de las transferencias adecuadas para cada equipo, análisis de jugadores jóvenes del mercado con gran potencial y análisis del reemplazo ideal para un determinado jugador.

## 2.3 – Problemática de la organización y la gestión de los datos

Para realizar modelos estadísticos es necesario contar con una gran cantidad de datos para cada jugador para poder entender cuál fue el compartimiento individual a lo largo de su carrera profesional. Para ello, es necesario analizar una ventana temporal de gran cantidad de años por jugador. Sin embargo, la organización se encuentra sumergida en un contexto de grandes volúmenes de datos denominado *big data*, la cual conlleva grandes desafíos.

Mayer-Schönberger y Cukier (2013) sostienen que el *big data* señala que los datos se refieren a gran escala, pero no a una escalara inferior, para extraer nuevas percepciones o crear nuevas formas de valor, de tal forma que transforman los mercados, las organizaciones, las relaciones entre los ciudadanos y los gobiernos, etc. Actualmente las organizaciones se encuentran en un contexto de datos masivos, los cuales se distribuyen caóticamente y pueden ser confusos, y cambiantes, pero logran generar una tendencia general que antes no era posible, perdiendo "exactitud en el nivel micro", pero ganándolo en el "nivel macro" (Mayer-Schönberger y Cukier, 2013).

En junio de 2008 la revista Wired, en el volumen especial The petabyte age, Lev Manovich afirmaba que:

> Nuestra capacidad para capturar, almacenar y comprender enormes cantidades de datos está cambiando la ciencia, la medicina, los negocios y la tecnología. A medida que nuestra colección de datos y cifras crece, también lo hace la oportunidad de encontrar respuestas a preguntas fundamentales.

Este gran volumen de datos que menciona Manovich es parte del fenómeno del *big data*. Por otra parte, Christof Schöch (2013) incorpora la idea de *smart data* (datos inteligentes), obteniendo un nuevo concepto: "*smart big data*". Desde su punto de vista:" solo el big data inteligente permite métodos cuantitativos inteligentes" (Schöch, 2013).

Schmarzo (2013) plantea que la tecnología disponible de la actualidad permite a las organizaciones manipular datos masivos generando una mejor comprensión del comportamiento de sus clientes, productos competencia y el mercado. Impulsado por los *insights* del *big data*, las compañías pueden mejorar la experiencia de sus clientes, agregar

valor y aumentar el retorno de inversión. *Big data* se presenta como un proceso de negocios que facilita la toma de decisiones en tiempo real (Schmarzo, 2013).

A lo largo de la historia del fútbol, fue cambiando la rigurosidad en la captura de datos, por lo que antes del año 2000 los datos que se capturaban de cada partido eran notablemente inferiores a lo que ocurre actualmente. A su vez, la cantidad de años de captura de datos en la historia del fútbol varía según cada país. En determinados sitios web la historia de datos de los principales equipos de Europa comienza hace 30 años, mientras que en Sudamérica solamente hay disponible los últimos 10 años de historia.

Por otro lado, en todos los países no se obtiene la misma cantidad de datos en cada partido de fútbol. En las principales ligas de Europa capturan una mayor cantidad de datos por partido respecto a otras ligas del mundo como las de Sudamérica. Esto es una dificultad en caso de querer comparar una misma métrica en todos los jugadores del mundo y solo cierta parte de la población contiene dicha métrica.

Otra problemática que surge al analizar el comportamiento anual de cada jugador es incluir los datos de los partidos con su selección nacional. Esto se debe a que estos datos no están disponibles con facilidad en los sitios web y, por lo tanto, dificulta obtener una visión completa del jugador. También ocurre algo similar al querer incorporar información de los equipos juveniles de cada club, ya que la mayoría de los datos se corresponden con el plantel superior. Esto dificulta la detección temprana de jugadores con gran potencial para los años futuros.

Dada estas problemáticas detalladas anteriormente, la organización no va a incluir los datos con esas características en los modelos de aprendizaje automático debido a que no va a poder generalizar para la población que no contenga dichos datos. Sin embargo, estos datos van a ser utilizados para un análisis exploratorio de los mismos y van a ser ofrecidos a sus clientes ya que pueden brindar información adicional al modelo de scoring que ofrece la organización.

## 3- Descripción metodológica

En esta sección se describe como fue la recopilación de los datos y como se trataron mediante el procesamiento adecuado. Luego se presenta el análisis exploratorio de los datos y, por último, cuáles fueron los modelos de *machine learning* aplicados junto a las métricas utilizadas para medir su performance.

## 3.1 – Recopilación de los datos

Los datos sobre los jugadores del fútbol argentino fueron obtenidos de la página web de *Football Database* cuyo fin es analizar el comportamiento y transferencias de los futbolistas que jugaron en Argentina. (Football Database, 2020).

El relevamiento de los datos incluye las temporadas desde 2004/2005 hasta 2019/2020. Los datos corresponden a información recopilada online a través de la página web antes mencionada, la cual es susceptible de contener errores provenientes de la carga manual de datos.

La página web contiene información de todos los jugadores del mundo. Sin embargo, a fines de este trabajo, el criterio para definir la población de interés fue obtener los jugadores que participaron en al menos un torneo en el fútbol argentino en los últimos 15 años. Por lo tanto, la base contiene principalmente jugadores argentinos, aunque también jugadores de diversas nacionalidades y diversas ligas a lo largo del mundo.

La base contiene 4533 jugadores únicos. Al tener la información de los jugadores a lo largo de muchas temporadas, cada jugador puede aparecer más de una vez. De esta forma, la base contiene 28544 registros y 29 atributos. Por otra parte, hay varios registros que tienen valores faltantes en varias columnas, lo cual será explicado en el próximo apartado el tratamiento realizado en estos casos.

En el anexo "7.1 - Descripción de la base de datos" se presenta la estructura del conjunto de datos.

## 3.2 – Procesamiento de los datos

La manipulación de la base de datos se ha realizado a través de la librería *dplyr* disponible para el lenguaje de programación R (R Core Team, 2020) en su versión 3.6.3 correspondiente al 29 de febrero de 2020 mediante el Entorno de Desarrollo Integrado *Rstudio*, versión 1.3.1073. Todas las visualizaciones expuestas en el presente documento se han creado utilizando la librería *ggplot2* también disponible para la versión de R recién mencionada.

Las técnicas de regresión logística, *random forest* y redes neuronales fueron implementadas mediante el software *RapidMiner* Studio versión 9.7.002 con licencia educativa activa.

A partir de cada equipo, se incorporó la información del país y continente al que pertenecen. Cabe aclarar que, 493 equipos no fueron asociados a ningún país ya que son equipos de baja jerarquía internacional que no van a influir en el objetivo del presente trabajo. A continuación, se detalla un resumen por continente de la cantidad de países y equipos en los que juegan los futbolistas:

*Tabla 1*

| **Continente** | **Países** | **Equipos** |
|---|---|---|
| Europa | 31 | 321 |
| Sudamérica | 10 | 281 |
| Asia | 15 | 94 |
| Centroamérica | 4 | 49 |
| Norteamérica | 1 | 28 |
| Oceanía | 1 | 5 |
| NA | NA | 493 |

*Fuente: Elaboración Propia*

La variable dependiente que se busca predecir toma los valores 0 o 1. La misma se define dependiendo si el jugador fue transferido en la temporada siguiente a un equipo de mayor jerarquía. Para ello se crea la variable "nivel" haciendo referencia a la jerarquía del equipo, la cual es definida principalmente por la cantidad de campeonatos ganados y el dinero invertido en transferencias de jugadores a lo largo de su historia. A su vez, para cada registro se crea la variable "lead.nivel" indicando el nivel del equipo de la temporada siguiente a la actual. El nivel 1 hace referencia a los equipos fuera de la Argentina de gran importancia mundial, puntualmente los principales equipos de las ligas de Sudamérica, México, Estados Unidos, Europa y Asia. El nivel 2 hace referencia a los 5 equipos más grandes del fútbol argentino, puntualmente son River Plate, Boca Juniors, Independiente, Racing Y San Lorenzo. Entre los niveles 3 y 4 se encuentren el resto de los equipos argentinos que en la mayoría de los últimos 15 años estuvieron participando de la primera categoría del fútbol argentino y se diferencian entre sí por la diferencia de dinero invertido en fichajes de jugadores. En el nivel 5 se encuentran los equipos argentinos que en la mayoría de los últimos 15 años no se encontraron en la primera categoría. Por último, el nivel 6 hace referencia a los equipos extranjeros que no están entre los principales de cada

país, como por ejemplo equipos de la segunda y tercera categoría de las ligas europeas y los equipos que no fueron asociados a ningún país.

Los registros de la base de datos tienen asociados una clasificación binaria en la variable dependiente, siempre y cuando contengan información del equipo al que pertenecen en la temporada inmediata posterior. Esto se hace analizando las variables "lead" y "lead.nivel". Por lo tanto, la variable dependiente va a tomar valor faltante cuando los registros corresponden a la última temporada de cada jugador debido a que todavía no está disponible la información de que sucederá en la siguiente temporada. En los casos que esto ocurra, se van a eliminar dichos registros de la base de datos.

A continuación, se detalla un resumen para cada nivel acerca de la cantidad de equipos, la suma total y el promedio del dinero invertido en fichajes:

*Tabla 2*

| Nivel | Equipos | Suma (M€) | Promedio (M€) |
|---|---|---|---|
| 1 | 327 | 4,479,136,400 | 17,744,351 |
| 2 | 8 | 290,514,200 | 36,314,275 |
| 3 | 10 | 68,944,000 | 7,660,444 |
| 4 | 12 | 52,540,000 | 4,776,364 |
| 5 | 57 | NA | NA |
| 6 | 865 | 31,180,000 | 3,897,500 |

*Fuente: Elaboración Propia*

Por lo tanto, analizando las variables "lead" y "lead.nivel", se crea la variable a predecir denominada "vd". La misma puede ser clasificado como 0 (mantiene o disminuye el nivel del equipo) o 1 (mejora el nivel del equipo).

Cabe aclarar que hay clubes deportivos que figuran en la base de datos como 2 equipos distintos ya que tienen un equipo principal y un equipo alternativo, denominado la reserva del club. A modo de ejemplo, se encuentran los equipos "River Plate" y "River Plate B". Esto puede generar algunos duplicados de equipos en el análisis recién mencionado, como ocurre en el nivel 2 donde se mencionó que son 5 equipos y en el análisis agrupado por nivel figuran 8.

Determinadas variables fueron creadas a partir de las variables originales. La variable edad fue creada a partir de la variable fecha nacimiento ya que varía a lo largo de cada

temporada. La variable edad_debut_inter se crea a partir de la variable primera_vez indicando la edad en que el jugador debutó en un partido internacional con su selección. La variable posición_2 se crea con el fin de sintetizar las principales posiciones en 4 categorías (Arquero-Defensor-Volante-Delantero) a partir de la variable Posición que contiene 34 categorías distintas con mayor detalle de la posición. A su vez, se crea la variable posición_3 con el fin de sintetizar las principales posiciones en 5 categorías (Arquero-Defensor-Volante Defensivo - Volante Ofensivo - Delantero).

Determinadas variables no van a ser consideradas para el desarrollo del modelo ya que presentan información del futuro, son fechas estáticas en el tiempo o no tienen relevancia para predecir el comportamiento futuro. Las mismas son: Lugar_nacimiento, país_origen, Pais_origen_2, numero, ano_nacimiento, país, continente, lead.nivel, posición_actual, fichajes_precio, Partidos_Internacionales, Pie, Palmares.

De los jugadores a considerar, se han encontrado valores faltantes en diversos atributos. En las variables goles, goles_pp, efectividad, asistencias, tarjetas amarillas, tarjetas rojas, presencia, once_inicial_abs, once_inicial_porc y Clean_sheets, serán reemplazados por el valor constante 0 debido a que la ausencia indica que no le corresponde un valor mayor a cero. Por otra parte, las variables peso y altura han sido reemplazadas por la mediana de cada variable, debido a que la ausencia indica que no se encuentra la información disponible en la página web donde se obtuvo la base de datos.

Existen 2 atributos que son específicos de los arqueros, puntualmente Clean_sheets y Goles_recibidos. En consecuencia, estos atributos toman valores faltantes para los jugadores que son defensores, mediocampistas y delanteros. Por lo tanto, se decide dividir la base de datos en dos según la posición del jugador, es decir, arqueros por un lado y el resto de los jugadores por otro.

En la etapa del procesamiento de datos se estandarizarán todas las variables numéricas y se realizará el proceso one-hot encoding en las variables categóricas, ya que algunos algoritmos requieren variables de tipo numéricas exclusivamente para el desarrollo del modelo.

Con respecto a los *outliers*, se realizará el grafico *boxplot* para cada variable con el fin de detectar casos atípicos. No se considera necesario desestimar estos valores extremos ya que los mismos indican características interesantes de los futbolistas. Por ejemplo, se puede observar jugadores con gran cantidad de goles, asistencias o tarjetas rojas.

Dado el procesamiento de datos mencionado previamente en esta sección, todos los registros van a tener una clasificación binaria asociada en la variable dependiente. Para poder realizar una predicción, es preciso contar con una base de datos de entrenamiento y una base de prueba donde se aplicará el mejor modelo obtenido. A estos fines, se utiliza la técnica de validación cruzada (Cross Validation) con 5 carpetas (folds). El rendimiento del modelo se medirá con los criterios de "Accuracy", varianza y el área bajo la curva ROC (AUC ROC).

## 3.3 – Análisis de los datos

El análisis de datos se realizó sobre la población en la que se va a desarrollar el modelo, es decir 9368 registros.

A continuación, se detalla un análisis elaborado con el comando *describe* del software *RStudio* de todas las variables numéricas:

*Tabla 3*

| Clean_Sheets | Cantidad | Promedio | Desvio est | Mediana | Media trunc | Desv med trunc | min | max | rango | Asimetria | Curtosis | Error estandar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ano_temp | 9368 | 2012.62 | 3.93 | 2013 | 2013 | 4.45 | 1992 | 2019 | 27 | -1.05 | 1.54 | 0.04 |
| Goles_pp | 9368 | 0.04 | 0.20 | 0 | 0.00 | 0.00 | 0 | 2 | 2 | 5.36 | 30.18 | 0.00 |
| Asistencias | 9368 | 0.29 | 0.82 | 0 | 0.08 | 0.00 | 0 | 11 | 11 | 4.24 | 25.42 | 0.01 |
| Efectividad | 9368 | 453.13 | 643.12 | 202 | 315.81 | 299.49 | 0 | 3512 | 3512 | 1.88 | 3.50 | 6.64 |
| Minutos_Jugados | 9368 | 1402.48 | 767.11 | 1259 | 1328.70 | 799.12 | 361 | 4080 | 3719 | 0.71 | -0.31 | 7.93 |
| Tarjetas_Amarillas | 9368 | 3.36 | 2.78 | 3 | 3.04 | 2.97 | 0 | 19 | 19 | 1.15 | 1.50 | 0.03 |
| Tarjetas_Rojas | 9368 | 0.26 | 0.54 | 0 | 0.15 | 0.00 | 0 | 4 | 4 | 2.13 | 4.62 | 0.01 |
| Presencia | 9368 | 0.62 | 0.27 | 0.7 | 0.65 | 0.28 | 0 | 1 | 1 | -0.71 | -0.20 | 0.00 |
| Clean_Sheets | 9368 | 0.53 | 2.29 | 0 | 0.00 | 0.00 | 0 | 24 | 24 | 5.14 | 28.97 | 0.02 |
| Goles_concedidos | 672 | 23.06 | 12.44 | 22 | 22.17 | 13.34 | 2 | 66 | 64 | 0.63 | 0.01 | 0.48 |
| Once_inicial_abs | 9368 | 15.81 | 8.94 | 14 | 15.03 | 8.90 | 0 | 45 | 45 | 0.64 | -0.44 | 0.09 |
| Victorias_abs | 9368 | 6.90 | 4.29 | 6 | 6.51 | 4.45 | 0 | 25 | 25 | 0.86 | 0.58 | 0.04 |
| Draws_abs | 9368 | 5.61 | 3.25 | 5 | 5.34 | 2.97 | 0 | 19 | 19 | 0.73 | 0.26 | 0.03 |
| Derrotas_abs | 9368 | 6.31 | 3.62 | 6 | 6.01 | 2.97 | 0 | 26 | 26 | 0.80 | 0.66 | 0.04 |
| PJ_2 | 9368 | 18.82 | 8.49 | 18 | 18.31 | 8.90 | 5 | 45 | 40 | 0.46 | -0.64 | 0.09 |
| Once_inicial_porc | 9368 | 0.82 | 0.21 | 0.9 | 0.86 | 0.15 | 0 | 1 | 1 | -1.21 | 0.67 | 0.00 |
| Victorias_porc | 9368 | 0.36 | 0.15 | 0.4 | 0.36 | 0.15 | 0 | 0.9 | 0.9 | 0.23 | 0.02 | 0.00 |
| Draws_porc | 9368 | 0.30 | 0.12 | 0.3 | 0.30 | 0.11 | 0 | 0.9 | 0.9 | 0.32 | 0.77 | 0.00 |
| Derrotas_porc | 9368 | 0.34 | 0.15 | 0.3 | 0.34 | 0.14 | 0 | 1 | 1 | 0.34 | 0.35 | 0.00 |
| edad | 9368 | 26.22 | 4.49 | 26 | 26.03 | 4.45 | 16 | 40 | 24 | 0.34 | -0.57 | 0.05 |
| Goles_2 | 9368 | 1.69 | 2.66 | 1 | 1.12 | 1.48 | 0 | 29 | 29 | 3.12 | 14.89 | 0.03 |
| nivel | 9368 | 3.70 | 1.11 | 4 | 3.75 | 1.48 | 2 | 5 | 3 | -0.20 | -1.34 | 0.01 |
| vd3 | 9368 | 0.21 | 0.41 | 0 | 0.14 | 0.00 | 0 | 1 | 1 | 1.43 | 0.05 | 0.00 |
| altura | 9368 | 1.79 | 0.06 | 1.8 | 1.79 | 0.06 | 1.55 | 2.0 | 0.4 | -0.07 | 0.22 | 0.00 |
| peso | 9368 | 75.10 | 5.79 | 75 | 75.04 | 4.45 | 50 | 105 | 55 | 0.18 | 1.33 | 0.06 |
| edad_debut_inter | 9368 | 3.44 | 8.55 | 0 | 1.08 | 0.00 | 0 | 36 | 36 | 2.15 | 2.80 | 0.09 |

*Fuente: Elaboración Propia*

A su vez, se grafica un histograma y *boxplot* donde se muestra la distribución de todas las variables numéricas:

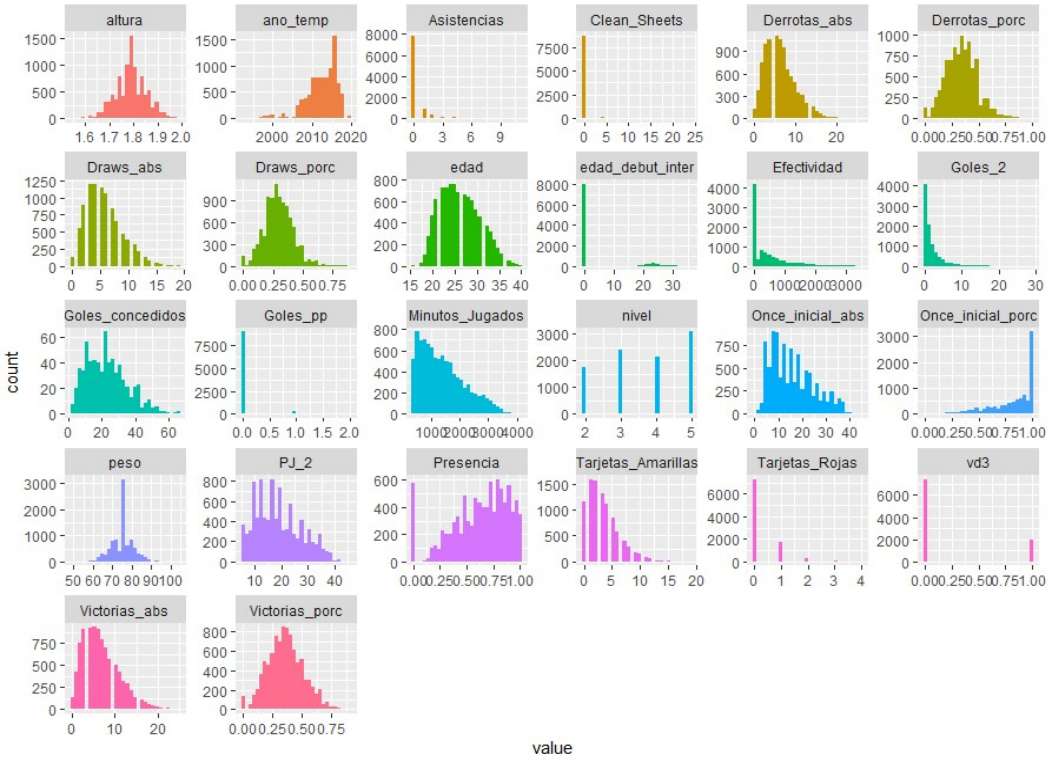



*Ilustración 1*

*Fuente: Elaboración Propia*

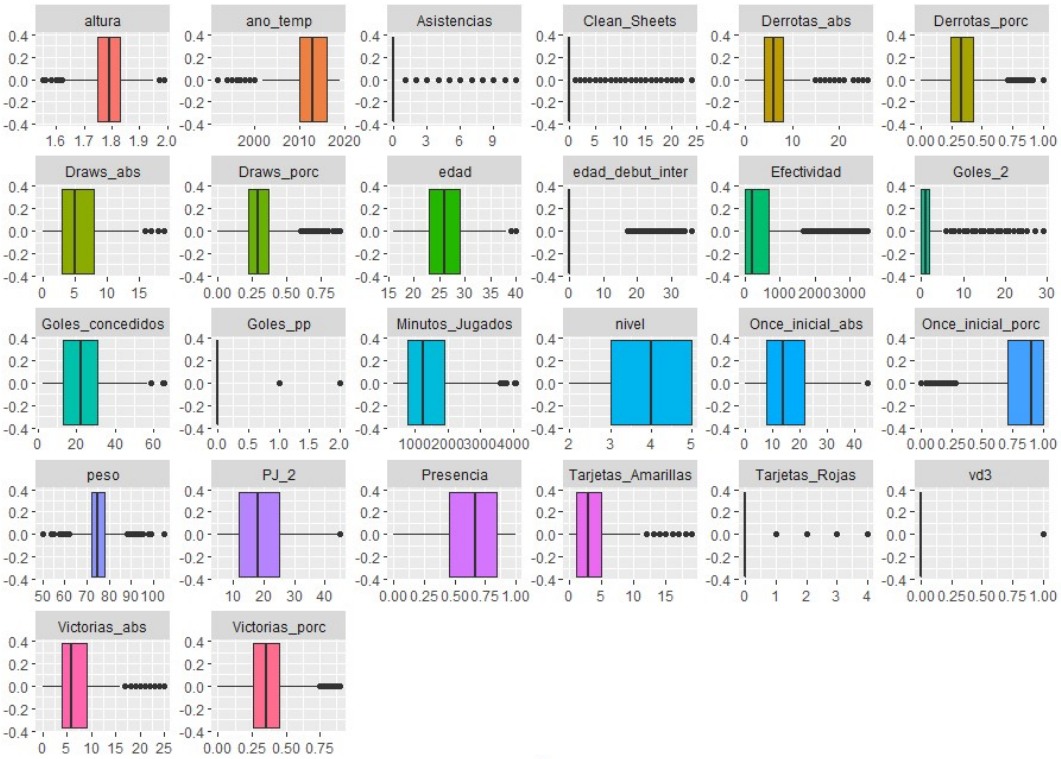



*Ilustración 2*

*Fuente: Elaboración Propia*

En cuanto a las variables categóricas, se realiza el siguiente análisis.

Frecuencia absoluta y relativa de las variables categóricas:

*Tabla 4*

| Posicion_2 | Freq | % Valid | % Valid Cum. |
|---|---|---|---|
| Arq | 672 | 7.17 | 7.17 |
| Def | 3208 | 34.24 | 41.42 |
| Del | 1667 | 17.79 | 59.21 |
| Med | 3821 | 40.79 | 100.00 |
| Total | 9368 | 100.00 | 100.00 |

*Fuente: Elaboración Propia*

*Tabla 5*

| Posicion | Freq | % Valid | % Valid Cum. |
|---|---|---|---|
| Defensor central | 1085 | 11.58 | 11.58 |
| atacante | 973 | 10.39 | 21.97 |
| medio | 973 | 10.39 | 32.35 |
| portero | 672 | 7.17 | 39.53 |
| medio centro | 607 | 6.48 | 46.01 |
| Defensor lateral izquierdo | 590 | 6.30 | 52.31 |
| Defensor lateral derecho | 545 | 5.82 | 58.12 |
| defensa | 509 | 5.43 | 63.56 |
| medio defensivo | 480 | 5.12 | 68.68 |
| Goleador | 463 | 4.94 | 73.62 |
| medio ofensivo | 376 | 4.01 | 77.64 |
| medio izquierdo | 265 | 2.83 | 80.47 |
| medio derecho | 259 | 2.76 | 83.23 |
| Defensor central derecho | 257 | 2.74 | 85.97 |
| medio ofensivo izquierdo | 219 | 2.34 | 88.31 |
| Volante de creación | 163 | 1.74 | 90.05 |
| Centro punta | 156 | 1.67 | 91.72 |
| extremo derecho | 144 | 1.54 | 93.25 |
| Defensor central izquierdo | 135 | 1.44 | 94.69 |
| medio ofensivo derecho | 99 | 1.06 | 95.75 |
| extremo izquierdo | 85 | 0.91 | 96.66 |
| Mediocentro derecho | 62 | 0.66 | 97.32 |
| Defensor lateral | 55 | 0.59 | 97.91 |
| Mediocentro izquierdo | 42 | 0.45 | 98.36 |
| Segunda punta | 41 | 0.44 | 98.79 |
| atacante polivalente | 34 | 0.36 | 99.16 |
| libero | 22 | 0.23 | 99.39 |
| medio polivalente | 17 | 0.18 | 99.57 |
| medio ofensivo polivalente | 13 | 0.14 | 99.71 |
| extremo | 12 | 0.13 | 99.84 |
| defensa polivalente | 10 | 0.11 | 99.95 |
| Volante mixto | 3 | 0.03 | 99.98 |
| medio defensivo derecho | 2 | 0.02 | 100.00 |
| Total | 9368 | 100.00 | 100.00 |

*Fuente: Elaboración Propia*

*Tabla 6*

| Equipo | Freq | % Valid | % Valid Cum. |
|---|---|---|---|
| River Plate | 406 | 4.33 | 4.33 |
| Independiente | 384 | 4.10 | 8.43 |
| Boca Juniors | 368 | 3.93 | 12.36 |
| San Lorenzo | 298 | 3.18 | 15.54 |
| Estudiantes LP | 285 | 3.04 | 18.58 |
| Newell's | 285 | 3.04 | 21.63 |
| Velez Sarsfield | 284 | 3.03 | 24.66 |
| Argentinos | 282 | 3.01 | 27.67 |
| Racing Club | 277 | 2.96 | 30.63 |
| Colon | 276 | 2.95 | 33.57 |
| Gim. La Plata | 271 | 2.89 | 36.46 |
| Lanus | 267 | 2.85 | 39.31 |
| Banfield | 263 | 2.81 | 42.12 |
| Rosario | 260 | 2.78 | 44.90 |
| Tigre | 250 | 2.67 | 47.57 |
| Arsenal | 247 | 2.64 | 50.20 |
| Godoy Cruz | 241 | 2.57 | 52.78 |
| Huracán | 237 | 2.53 | 55.31 |
| Belgrano | 213 | 2.27 | 57.58 |
| Defensa | 213 | 2.27 | 59.85 |
| San Martin SJ | 213 | 2.27 | 62.13 |
| Tucuman | 200 | 2.13 | 64.26 |
| All Boys | 190 | 2.03 | 66.29 |
| Union | 190 | 2.03 | 68.32 |
| Olimpo | 188 | 2.01 | 70.32 |
| Quilmes | 188 | 2.01 | 72.33 |
| Gim. Jujuy | 187 | 2.00 | 74.33 |
| Patronato | 175 | 1.87 | 76.20 |
| Rafaela | 166 | 1.77 | 77.97 |
| Ferro | 157 | 1.68 | 79.64 |
| Aldosivi | 151 | 1.61 | 81.26 |
| Instituto | 151 | 1.61 | 82.87 |
| Chacarita | 128 | 1.37 | 84.23 |
| I. Rivadavia | 115 | 1.23 | 85.46 |
| Sarmiento | 112 | 1.20 | 86.66 |
| Crucero | 101 | 1.08 | 87.73 |
| Talleres | 98 | 1.05 | 88.78 |
| Boca Unidos | 87 | 0.93 | 89.71 |
| Temperley | 83 | 0.89 | 90.60 |
| Douglas Haig | 77 | 0.82 | 91.42 |
| Nueva Chicago | 73 | 0.78 | 92.20 |
| San Martin T | 70 | 0.75 | 92.94 |
| A. Brown IC | 57 | 0.61 | 93.55 |
| Fénix | 55 | 0.59 | 94.14 |
| Santamarina | 50 | 0.53 | 94.67 |
| C. Córdoba SdE | 44 | 0.47 | 95.14 |
| Sp. Belgrano | 41 | 0.44 | 95.58 |
| Los Andes | 40 | 0.43 | 96.01 |
| Guillermo Brown | 39 | 0.42 | 96.42 |
| Estudiantes SL | 38 | 0.41 | 96.83 |
| Merlo | 37 | 0.39 | 97.22 |
| Brown Adrogue | 35 | 0.37 | 97.60 |
| CDJU | 31 | 0.33 | 97.93 |
| Almagro | 27 | 0.29 | 98.22 |
| Dálmine | 27 | 0.29 | 98.51 |
| GAF | 21 | 0.22 | 98.73 |
| Paraná | 18 | 0.19 | 98.92 |
| Flandria | 16 | 0.17 | 99.09 |
| Atlanta | 15 | 0.16 | 99.25 |
| Unión MdP | 11 | 0.12 | 99.37 |
| CAI | 10 | 0.11 | 99.48 |
| Villa San Carlos | 10 | 0.11 | 99.58 |
| Desamparados | 8 | 0.09 | 99.67 |
| Gim. Mendoza | 8 | 0.09 | 99.75 |
| Tiro Federal | 8 | 0.09 | 99.84 |
| CAJUU | 6 | 0.06 | 99.90 |
| Platense | 3 | 0.03 | 99.94 |
| Agropecuario | 2 | 0.02 | 99.96 |
| Morón | 2 | 0.02 | 99.98 |
| Deportivo Riestra | 1 | 0.01 | 99.99 |
| Mitre SdE | 1 | 0.01 | 100.00 |
| Total | 9368 | 100.00 | 100.00 |

*Fuente: Elaboración Propia*

Frecuencia cruzada entre Posicion y Posicion_2:

*Tabla 7*

| | Arq | Def | Del | Med |
|---|---|---|---|---|
| atacante | 0 | 0 | 973 | 0 |
| atacante polivalente | 0 | 0 | 34 | 0 |
| Centro punta | 0 | 0 | 156 | 0 |
| defensa | 0 | 509 | 0 | 0 |
| defensa polivalente | 0 | 10 | 0 | 0 |
| Defensor central | 0 | 1085 | 0 | 0 |
| Defensor central derecho | 0 | 257 | 0 | 0 |
| Defensor central izquierdo | 0 | 135 | 0 | 0 |
| Defensor lateral | 0 | 55 | 0 | 0 |
| Defensor lateral derecho | 0 | 545 | 0 | 0 |
| Defensor lateral izquierdo | 0 | 590 | 0 | 0 |
| extremo | 0 | 0 | 0 | 12 |
| extremo derecho | 0 | 0 | 0 | 144 |
| extremo izquierdo | 0 | 0 | 0 | 85 |
| Goleador | 0 | 0 | 463 | 0 |
| libero | 0 | 22 | 0 | 0 |
| medio | 0 | 0 | 0 | 973 |
| medio centro | 0 | 0 | 0 | 607 |
| medio defensivo | 0 | 0 | 0 | 480 |
| medio defensivo derecho | 0 | 0 | 0 | 2 |
| medio derecho | 0 | 0 | 0 | 259 |
| medio izquierdo | 0 | 0 | 0 | 265 |
| medio ofensivo | 0 | 0 | 0 | 376 |
| medio ofensivo derecho | 0 | 0 | 0 | 99 |
| medio ofensivo izquierdo | 0 | 0 | 0 | 219 |
| medio ofensivo polivalente | 0 | 0 | 0 | 13 |
| medio polivalente | 0 | 0 | 0 | 17 |
| Mediocentro derecho | 0 | 0 | 0 | 62 |
| Mediocentro izquierdo | 0 | 0 | 0 | 42 |
| portero | 672 | 0 | 0 | 0 |
| Segunda punta | 0 | 0 | 41 | 0 |
| Volante de creación | 0 | 0 | 0 | 163 |
| Volante mixto | 0 | 0 | 0 | 3 |

*Fuente: Elaboración Propia*

Frecuencia de la variable dependiente “vd”:

*Tabla 8*

| Categoría | Frecuencia | Proporción |
|---|---|---|
| 0 | 7410 | 0.79 |
| 1 | 1958 | 0.21 |

*Fuente: Elaboración Propia*

## 3.4 - Modelos aplicados

Dado que el presente trabajo tiene como objetivo diferenciar a los jugadores a lo largo de cada temporada en una clasificación binaria, se propone aplicar distintos algoritmos de aprendizaje supervisado para la clasificación. De todas las variables que se encuentran en el anexo "7.1 - Descripción de la base de datos", las utilizadas para predecir la variable dependiente "vd" fueron: goles, goles_pp, asistencias, minutos_jugados, partidos_jugados, tarjetas_amarillas, tarjetas_rojas, presencia, clean_sheets, goles_concedidos, once_inicial_abs, victorias_abs, draws_abs, derrotas_abs, once_inicial_porc, victorias_porc, draws_porc, derrotas_porc, edad, edad_debut_inter, altura, peso, nivel, posición, posición_2, posición_3.

Luego de probar diferentes técnicas de clasificación, se seleccionaron los tres modelos que brindaron la mayor exactitud (accuracy) en sus resultados. El primer modelo desarrollado es la regresión logística, el cual es un algoritmo de aprendizaje supervisado para predecir el resultado de una variable dicotómica en función de las variables independientes. El siguiente modelo desarrollado es *random forest*, el cual utiliza una gran cantidad de árboles, combinando distintas muestras de registros y columnas, teniendo como ventaja que soporta cualquier base de datos, aunque haya datos perdidos. Por último, el algoritmo de Redes Neuronales, el cual es un modelo computacional de clasificación y regresión que está inspirado en las redes neuronales humanas y trabaja utilizando diferentes capas de neuronas conectadas entre sí que aprenden de sí mismas.

## 3.5 – Métricas de evaluación

El procedimiento realizado una vez concluida la manipulación y procesamiento de la base de datos constó de los siguientes pasos: Tal como se mencionó anteriormente, el entrenamiento y la prueba para validar los modelos fueron realizados mediante la utilización de la técnica de validación cruzada con 5 carpetas. A su vez, se utilizó el operador *Optimize Parameters* del software Rapid Miner para optimizar los parámetros de cada modelo.

La métrica elegida para evaluar y comparar los modelos será la exactitud (accuracy). Esta métrica mide la cantidad de predicciones correctamente realizadas sobre la totalidad de los casos predichos. La varianza también será otra métrica utilizada para comparar los

distintos modelos a fin de conocer cuán sobre ajustados se encuentran. Los resultados de cada modelo podrán ser visualizados en la matriz de confusión y a su vez, se podrán visualizar las métricas mencionadas en una tabla comparativa en la sección "3. Resultados". Finalmente, se mostrará la curva ROC (Curva de característica operativa del receptor) y el AUC (Área bajo la curva ROC) del mejor modelo obtenido.

# 4 – Implementación del modelo

En esta sección, se detalla cómo será la implementación del modelo de aprendizaje automático. En primer lugar, se describe como se llevará a cabo la puesta en producción del modelo para que esté disponible para el usuario final como un servicio web. Luego se presentan los resultados que conlleva aplicar cada uno de los distintos algoritmos de aprendizaje automático y se los compara entre sí con el objetivo de definir cual es el modelo más adecuado para resolver el problema de negocio planteado. Luego se presentan diversas visualizaciones donde se exponen los principales resultados del modelo desarrollado e implementado. Por último, se define cual es la metodología de desarrollo elegida para llevar a cabo la implementación de proyectos de aprendizaje automático.

## 4.1 – Puesta en producción del modelo

La implementación en producción del modelo se explica en el siguiente archivo adjunto: "Implementación del modelo.pptx". En el mismo se encuentran todos los pasos elaborados que son necesarios para construir un modelo de aprendizaje automático, desde su concepción hasta su puesta en producción.

Comienza con el entendimiento del problema para luego poder realizar el análisis exploratorio de los datos. Continúa con el proceso de ingeniería de datos en donde se procesan y se manipulan los mismos para poder entrenar diversos modelos de aprendizaje automático. Luego se selecciona el mejor modelo y se realiza la operacionalización del mismo comenzando por el despliegue del modelo, monitoreando las predicciones y, por último, integrándolo.

## 4.2 – Resultados

Como se mencionó anteriormente la base de datos se dividió en dos de acuerdo con la posición del jugador. Dado que, para los arqueros, no fue posible obtener un modelo que supere al *baseline*, se optó por desarrollar un modelo para el resto de los jugadores. En esta última tabla, el *baseline* es de 78.67%, dado que ese es la proporción de la clase mayoritaria en la variable a predecir.

Se utilizaron tres técnicas distintas para la creación de los modelos: Regresión logística, *Random Forest* y Redes neuronales. Todos los modelos superan al *baseline*, aunque no por mucha diferencia.

El modelo que utiliza la regresión logística arroja resultados con una exactitud del 79.21% y una varianza del 0.21%. A continuación, se puede observar la matriz de confusión:

*Tabla 9*

**accuracy: 79.21% +/- 0.21% (micro average: 79.21%)**

| | true 0 | true 1 | class precision |
|---|---|---|---|
| pred. 0 | 6792 | 1759 | 79.43% |
| pred. 1 | 49 | 96 | 66.21% |
| class recall | 99.28% | 5.18% | |

*Fuente: Elaboración Propia*

Utilizando el algoritmo *Random Forest*, se alcanzan dos modelos distintos debido a que uno alcanza una mayor exactitud y el otro, menor varianza. El primer modelo alcanza una exactitud del 78.91% y una varianza del 0.24%, mientras que el segundo modelo, utilizando más árboles y otro criterio de decisión, alcanza una exactitud del 78.99% y una varianza del 0.29%.

Matriz de confusión utilizando *Random Forest* (50 árboles/ Gini index):

*Tabla 10*

**accuracy: 78.91% +/- 0.24% (micro average: 78.91%)**

| | true 0 | true 1 | class precision |
|---|---|---|---|
| pred. 0 | 6767 | 1760 | 79.36% |
| pred. 1 | 74 | 95 | 56.21% |
| class recall | 98.92% | 5.12% | |

*Fuente: Elaboración Propia*

Matriz de confusión utilizando *Random Forest* (70 árboles/ information gain):

*Tabla 11*

**accuracy: 78.99% +/- 0.29% (micro average: 78.99%)**

| | true 0 | true 1 | class precision |
|---|---|---|---|
| pred. 0 | 6789 | 1775 | 79.27% |
| pred. 1 | 52 | 80 | 60.61% |
| class recall | 99.24% | 4.31% | |

*Fuente: Elaboración Propia*

La elección de los dos modelos mencionados utilizando *Random Forest* se obtuvo con la utilización del operador *Optimize Parameters.* A continuación, se observan los resultados de dicho operador:

*Tabla 12*

Optimize Parameters (Grid) (8 rows, 4 columns)

| iteration | Rando... | Rando... | acc... ↓ |
|---|---|---|---|
| 6 | informati... | 70 | 0.790 |
| 3 | gini_index | 50 | 0.789 |
| 2 | informati... | 50 | 0.789 |
| 7 | gini_index | 70 | 0.788 |
| 8 | accuracy | 70 | 0.787 |
| 4 | accuracy | 50 | 0.787 |
| 1 | gain_ratio | 50 | 0.787 |
| 5 | gain_ratio | 70 | 0.786 |

*Fuente: Elaboración Propia*

Utilizando las redes neuronales, también se alcanzan dos modelos distintos debido a que uno alcanza una mayor exactitud y el otro, menor varianza. El modelo con un learning rate de 0.01 alcanza una exactitud del 78.94% y una varianza del 0.44%. El segundo modelo que se desarrolla con un learning rate del 0.007 alcanza una exactitud del 79.20% y una varianza del 0.48%.

Matriz de confusión utilizando redes neuronales (learning rate: 0.01):

*Tabla 13*

**accuracy: 78.94% +/- 0.44% (micro average: 78.94%)**

| | true 0 | true 1 | class precision |
|---|---|---|---|
| pred. 0 | 6740 | 1730 | 79.57% |
| pred. 1 | 101 | 125 | 55.31% |
| class recall | 98.52% | 6.74% | |

*Fuente: Elaboración Propia*

Matriz de confusión utilizando redes neuronales (learning rate: 0.007):

*Tabla 14*

**accuracy: 79.20% +/- 0.48% (micro average: 79.20%)**

| | true 0 | true 1 | class precision |
|---|---|---|---|
| pred. 0 | 6819 | 1787 | 79.24% |
| pred. 1 | 22 | 68 | 75.56% |
| class recall | 99.68% | 3.67% | |

*Fuente: Elaboración Propia*

A continuación, se detalla la comparación de los resultados de los modelos:

*Tabla 15*

| Modelo | Parámetro | Exactitud | Varianza |
|---|---|---|---|
| Regresión Logística | - | 79.21% | 0.21% |
| *Random Forest* | 50 trees/gini index | 78.91% | 0.24% |
| *Random Forest* | 70 trees/ information gain | 78.99% | 0.29% |
| Redes Neuronales | learning rate 0.01 | 78.94% | 0.44% |
| Redes Neuronales | learning rate 0.007 | 79.20% | 0.48% |

*Fuente: Elaboración Propia*

Se puede observar que el mejor modelo que se ajusta al problema planteado en el presente trabajo es la regresión logística, alcanzando la mayor exactitud y a su vez, la menor varianza. Es importante evaluar ambas métricas, ya que obtener una gran varianza implica un sobre ajuste en el desarrollo del modelo con los datos de entrenamiento.

Tal como se planteó en la sección “3.5 – Métricas de evaluación”, se presenta la curva ROC y la métrica de AUC del modelo que presenta mejor exactitud.


*Ilustración 3*

*Fuente: Elaboración Propia*

En la sección “7.2 - Modelos utilizados y parámetros” se presentará con mayor detalle una descripción de los parámetros utilizados para cada modelo.

## 4.3 – Visualización de Insights

En la presente sección, se exponen cuatro gráficos que explican los principales resultados obtenidos de modelo desarrollado y de la puesta en producción. Previo a la exposición de los gráficos, se detallan algunos conceptos claves sobre la variable dependiente a predecir ya que va a permitir entender los resultados alcanzados. Tal como se mencionó anteriormente, la variable dependiente es una variable binaria. Toma valor 1 cuando el jugador es transferido a un equipo de mayor jerarquía en la temporada posterior y si esto no sucede, toma valor 0. A su vez, a los jugadores que tomen el valor 1 se los denominan “Good” (buenos jugadores) y a los que tomen el valor 0 se los denominan “Bad” (malos jugadores).

A continuación, se expone la distribución de la variable dependiente en la muestra de desarrollo segmentado por posición:

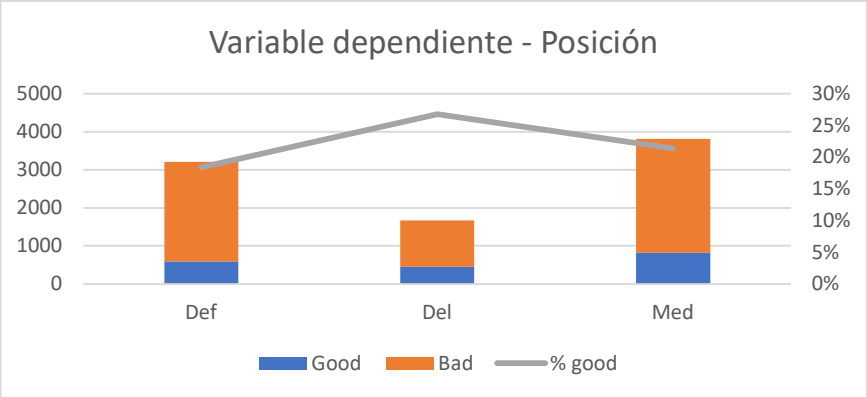



*Ilustración 4*

*Fuente: Elaboración Propia*

Los delanteros representan la menor cantidad de jugadores y son los que tienen la mayor tasa de buenos alcanzando el 25%. Por otro lado, la mayor proporción de jugadores se distribuye entre mediocampistas y defensores con una tasa de buenos promedio del 20%. De esta forma, se obtiene un promedio general de tasa de buenos del 21%.

Por otro lado, se detalla cual es la correlación de las variables independientes frente a la variable dependiente:

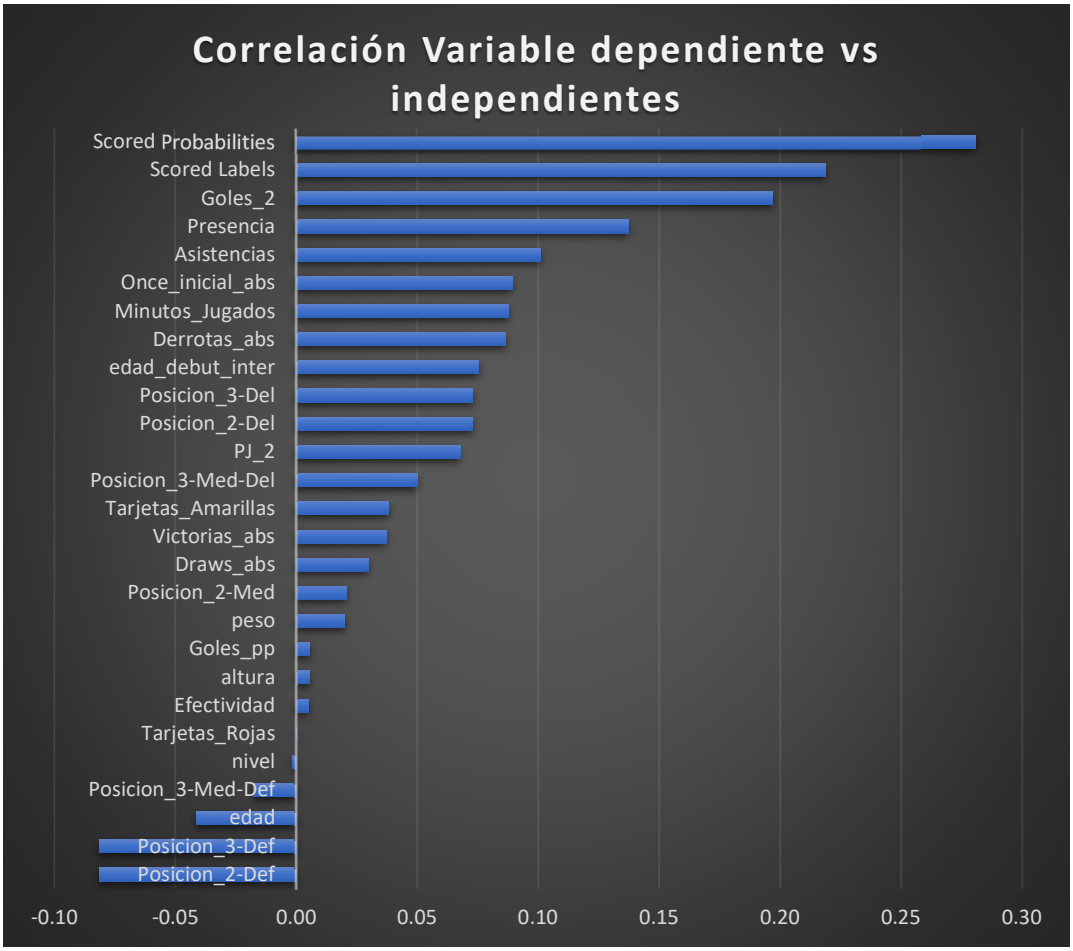



*Ilustración 5*

*Fuente: Elaboración Propia*

Con este gráfico, se puede observar cuales son las variables más correlacionadas con la variable a predecir. El mismo se realizó sobre la muestra de validación donde se encuentran los valores reales de la variable a predecir y también los resultados del modelo. De esta forma se puede observar que la variable "Scored Probabilities" es la variable más correlacionada con la variable dependiente lo cual tiene sentido. Dejando de lado los resultados del modelo, se puede observar que las variables independientes más correlacionadas son los goles, la presencia y las asistencias.

Por otra parte, se expone cual es la tasa de "buenos" para cada rango de score:

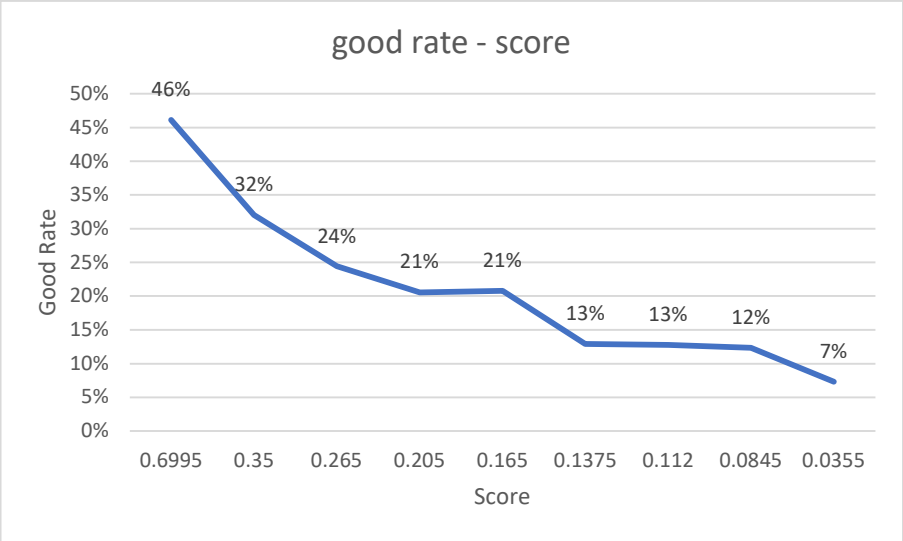



*Ilustración 6*

*Fuente: Elaboración Propia*

Se puede observar que la relación entre el score y tasa de "buenos" es directa. Esto implica que a medida que disminuye el score, también disminuye la tasa de buenos. En contraposición, a medida que aumenta el score, también aumenta la tasa de buenos. Esto tiene sentido ya que a medida que aumenta la probabilidad de encontrar un jugador con etiqueta "1" es esperable que haya mayor cantidad de buenos jugadores respecto a los rangos de probabilidad inferiores.

Por último, se expone cual es la importancia de cada variable independiente en el modelo de clasificación. La misma se calcula a partir del coeficiente que acompaña a cada variable predictora:

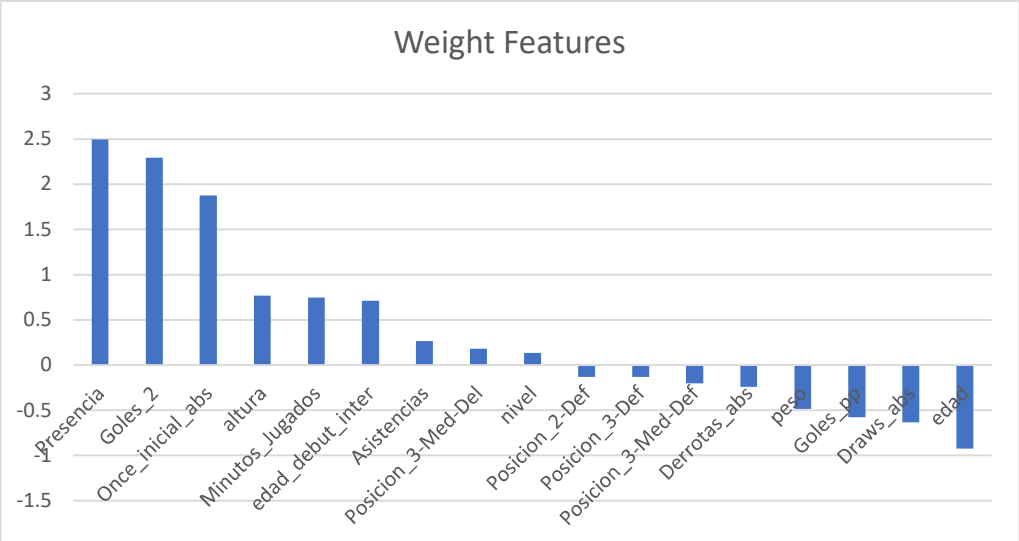



*Ilustración 7*

*Fuente: Elaboración Propia*

Al momento de entender cuáles son las variables más significativas en la construcción del modelo predictivo, resulta imprescindible entender el peso que tiene cada variable independiente en el modelo. Cabe destacar que es relevante observar el valor absoluto del peso para entender cuáles son las variables más significativas. Se puede observar que presencia, goles y once_inicial_abs (cantidad de partidos en el equipo titular) son las principales variables que afectan positivamente en el score, mientras que la edad, draws_abs (cantidad de partidos empatados) y goles_pp (goles en contra) son las principales variables que afectan negativamente en el score.

## 4.4 – Metodologías ágiles

La metodología elegida para la implementación del proyecto que ponga el modelo de aprendizaje automático en producción es una metodología ágil denominada Kanban. La misma va a permitir al equipo trabajar de manera organizada y, por ende, cumplir con sus tareas en tiempo y forma. A su vez, permite visualizar el flujo de trabajo mediante un tablero con tarjetas y columnas. Cada columna en el tablero representa un paso en su flujo de trabajo y cada tarjeta Kanban representa una tarea o elemento de trabajo.

Dicha metodología va a ser implementada mediante la plataforma Trello, la cual estará organizada en cuatro columnas: To Do – In Progress – Review - Done. Se convocará a

todos los integrantes de la organización para explicar cuál es el uso adecuado de la plataforma y cuál es la funcionalidad de cada una de las columnas.

Se realizará una reunión inicial con el equipo para elaborar todas las tareas que se deben realizar para así colocarlas en la primera columna. Una vez realizado este proceso, se le asignará a cada integrante del equipo sus tareas correspondientes donde tendrán una fecha límite para realizarlas. Cada integrante del equipo tendrá ordenadas sus tareas según las prioridades que se definan, por lo tanto, la primera tarea prioritaria a realizar por cada integrante estará colocada en la parte superior de la columna.

Cada integrante pasará su tarea con mayor prioridad a la columna “In Progress”. Una vez finalizada, la tarea pasará a “Review” donde se revisará internamente que los resultados sean los esperados. Luego que se revisen los resultados alcanzados, si los mismos eran los esperados pasará a la siguiente etapa “Done” y seguirá con la tarea posterior en el orden prioritario establecido en un comienzo. En caso de no ser los resultados esperados, la tarea volverá a la etapa “In Progress” hasta que se mejoren los resultados alcanzados y de esta forma, se repetirá el proceso explicado anteriormente.

Kanban es la metodología óptima para la organización debido a que, al descomponer las grandes tareas en tareas cortas y concisas, hace que el equipo de trabajo sea eficiente y conozca con facilidad las tareas prioritarias a realizar. Cada miembro del equipo puede ver y actualizar el estado de cada proyecto o tarea. De esta manera, todas las tareas son visibles, lo que aporta transparencia a todo el proceso de trabajo.

Para el equipo de trabajo es necesario una metodología que sea gráfica y fácil de entender. Esta combinación la tiene Kanban por ser visualmente atractiva en comparación con el resto de las metodologías de trabajo. Por otra parte, la curva de aprendizaje es mínima y permite que los equipos sean flexibles en la producción, sin agregar complejidad innecesaria al proceso.

Kanban se basa en eventos en lugar de segmentos de tiempo, lo que garantiza que se pueda responder a una caída repentina de la demanda de un producto o servicio eliminando tareas de la columna “To Do”. Por otro lado, al ubicarse todo el flujo de trabajo visible en un tablero, se puede ver en qué columna hay más tarjetas, por lo tanto, en qué etapa se lentifica el proceso de entrega.

Para el correcto funcionamiento de la metodología, es necesario que el tablero se actualice constantemente de forma correcta ya que un tablero desactualizado puede bloquear el

desarrollo de un proyecto. Dado que Kanban se basa en eventos, si una tarjeta/tarea no se mueve a la columna/etapa apropiada, las otras tareas que dependen de ella nunca se notifican y, por ende, se quedan bloqueadas.

A la metodología descrita anteriormente, se incorporará algunas características de la metodología ágil Scrum como es la asignación de roles. Se asignará un integrante con el rol de "Product Owner", el cual se encargará de describir y enumerar todas las tareas que irán al "To Do" y un "Kanban Master" que va a asesorar al equipo con todas las dudas e inconvenientes que surjan acerca de la realización de las tareas. Por otra parte, cabe aclarar que, el resto del equipo está compuesto por *data engineers, data scientists* y *machine learning engineers* quienes serán los responsables de ejecutar e implementar todas las tareas.

Para mantener las iteraciones cortas, se utiliza una métrica que define el límite de tareas en progreso (WIP - Work in Progress - por sus siglas en inglés). La misma representa el número máximo de tareas que puede llevar a cabo el equipo y cuando cae por debajo de un nivel predeterminado, se establece un activador de planificación bajo demanda para que el equipo sepa cuándo planificar.

Para medir el éxito del proyecto, se van a utilizar diversos KPIs (Key Performance Indicator): Tiempo promedio de ciclo completo de una tarea, tiempo promedio de tareas en la etapa "In Progress", cociente entre la cantidad de tareas en "Done" y en "To Do", promedio de tareas asignadas por persona, cantidad de tareas "In Progress" respecto al mes anterior. Cuando haya una caída en alguna de estas métricas, la organización va a definir si es necesario tomar una acción correspondiente como puede ser hablar con el equipo involucrado para mejorar a futuro.

Al finalizar cada una de las tareas, se realizará una reunión retrospectiva donde se reunirá todo el equipo involucrado junto al "Product Owner" y al "Kanban Master" para realizar un intercambio de opiniones acerca del proyecto finalizado. Además, se debate sobre cómo mejorar el proceso para futuros proyectos y, por lo tanto, permite al equipo centrarse en su rendimiento general e identificar formas de mejoras continuas.

# 5 - Conclusiones

Para la formulación del presente trabajo se desarrollaron tres apartados donde se expuso la gestión de datos en el contexto de la organización, la descripción metodológica para la

recopilación, procesamiento y análisis de los datos y la implementación del modelo de aprendizaje automático.

Se resolvieron las interrogantes planteadas que propone el presente trabajo ya que se encontró un modelo de *machine learning* que asigna un score a cada jugador del fútbol argentino, permitiendo comparar de una forma objetiva a todos los jugadores bajo un criterio estadístico. A su vez, el modelo encontrado permite diferenciar cuales son las variables independientes que tienen mayor impacto en el resultado del score, ya sea de manera positiva o negativa.

Para resolver el problema de clasificación que plantea el presente trabajo, se utilizaron tres técnicnas distintas para la creación de los modelos: Regresión logística, *Random Forest* y Redes neuronales. Se utilizó el operador *Optimize Parameters* del software Rapid Miner para optimizar los parámetros de cada modelo.

Todos los modelos superaron al *baseline*, aunque no por mucha diferencia. El mejor modelo alcanzado fue la regresión logística, obteniendo la mayor exactitud y la menor varianza de los modelos planteados. El mismo alcanzó una exactitud del 79.21% y una varianza del 0.21%, superando al *baseline* de 78.67%.

A modo de valoración personal, este trabajo será de gran utilidad para conocer cuáles son los jugadores del fútbol argentino que tienen mayor probabilidad de ser transferidos a un equipo de mayor jerarquía a nivel internacional. Por lo tanto, a los responsables en la toma de decisiones de los fichajes de los jugadores le pueden ser de gran utilidad los resultados del modelo ya que les va a permitir apoyarse en fundamentos estadísticos, además de su conocimiento del deporte.

Para un futuro estudio, sería interesante incorporar el atributo “palmares” de la base de datos original como un atributo independiente para el desarrollo del modelo. Por otro lado, analizar con mayor detalle los 493 equipos que no tienen asignado el país al que pertenece, para determinar si alguno es de gran jerarquía internacional. También sería interesante obtener un modelo de predicción para los arqueros que supere al *baseline*. Por último, construir la variable dependiente incorporando nuevos criterios a fines de obtener mayor exactitud en la predicción de los modelos.

# 6 - Referencias bibliográficas


- Cornelius Arndt, Ulf Brefeld (2016). Predicting the performance of soccer players. Retrieved from: https://doi.org/10.1002/sam.11321
- Football Database (2020). Recuperado el 22 de mayo de 2020, de https://www.footballdatabase.eu/
- Gitelman, L. (2013). «Raw data» is an oxymoron. Cambridge, Massachusetts: MIT Press.
- He, M., Cachucho, R., & Knobbe, A. J. (2015). Football player's performance and market value. LIACS, Leiden University, the Netherlands.
- He, Y. (2012). Predicting market value of soccer players using linear modeling techniques. University of Berkeley (working paper).
- Kitchin, R., y McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), doi: 10.1177/2053951716631130.
- Kumar, G. (2013). Machine Learning for Soccer Analytics.
- Lavalle, S., Lesser, E. & Shockley, R. (2011). Big data analytics and the path from insights to value.
- Manovich, L. (2011, junio 20). Trending: The Promises and the Challenges of Big Social Data. Recuperado 6 de mayo de 2019, de Http://manovich.net/ website: http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-socialdata/64-article-2011.pdf
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt.
- Mierswa, I., & Klinkenberg, R. (2020). RapidMiner Studio (9.7) [Data science, machine learning, predictive analytics]. Retrieved from https://rapidminer.com/
- Nikravesh, M. (2016, may). Moneyball: Sports Analytics in Soccer to Predict Performance and Outcomes. Retrieved from https://www.experfy.com/blog/moneyball-some-insights-to-soccer-analytics
- Penn, C. (2019). The Evolution of the Data-Driven Company.

- Poli, R., Ravenel, L., & Besson, R. (2018, Oct). Scientific assessment of football players' transfer value. Retrieved from https://football-observatory.com/IMG/pdf/ note01en.pdf
- Provost, F. & Fawcett, T. (2013). Data Science and its relationship to Big Data and Data driven Decision making.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.
- Rejec, J. (2016, June). How Big Data is Changing the World of Soccer. Retrieved from https://www.smartdatacollective.com/how-big-datachanging-world-soccer/
- Schmarzo, B. (2013). Big Data: Understanding how data powers big business. Indianapolis: John Wiley & Sons.
- Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. Journal of Digital Humanities, 2 (3). Recuperado de http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the humanities/
- Sosa Escudero, W. (2019). Big Data. Buenos Aires: Siglo XXI.
- Weich, B. (2012, June). Why Moneyball Will Not Work in Soccer. Retrieved from http://worldsoccertalk.com/2012/06/06/whymoneyball-will-not-work-in-soccer/


# 7 – Apéndices

## 7.1 - Descripción de la base de datos

Comportamiento histórico de los Jugadores en el fútbol argentino

| Atributo | Tipo de dato | Descripción |
|---|---|---|
| **Nombre** | **Texto (string)** | **Nombre del jugador** |
| **Posicion** | **Texto (string)** | **Posición del jugador en el campo de juego** |
| **Nacimiento** | **Texto (string)** | **Fecha de nacimiento** |
| **Lugar_nacimiento** | **Texto (string)** | **Ciudad de nacimiento** |
| **Altura** | **Número Real** | **Altura del jugador** |
| **Peso** | **Número entero** | **Peso del jugador** |
| **Pais_origen** | **Texto (string)** | **País de nacimiento** |
| **Pais_origen_2** | **Texto (string)** | **País en el caso de estar nacionalizado en otro país** |
| **Pie** | **Texto (string)** | **Pierna hábil** |

| | | |
|---|---|---|
| **Club_Actual** | **Texto (string)** | **Club actual a principios del 2020** |
| **Partidos_Internacionales** | **Número entero** | **Cantidad de partidos jugados con su selección nacional** |
| **Posicion_actual** | **Texto (string)** | **Profesión actual del jugador** |
| **Primera_vez** | **Texto (string)** | **Fecha y rival del debut con su selección nacional** |
| **Temporada** | **Número entero** | **Año de cada temporada** |
| **Equipo** | **Texto (string)** | **Equipo en cada temporada** |
| **Partidos jugados** | **Número entero** | **Partidos jugados en cada temporada** |
| **Goles** | **Número entero** | **Goles convertidos en cada temporada** |
| **Goles pp** | **Número entero** | **Goles en contra en cada temporada** |
| **Pases dc** | **Número entero** | **Asistencias (Pases decisivos) en cada temporada** |
| **Efectividad** | **Número entero** | **Cantidad de minutos que demora el jugador en convertir un gol para cada temporada** |
| **Minutos Jugados** | **Número entero** | **Minutos jugados en cada temporada** |
| **Tarjetas Amarillas** | **Número entero** | **Tarjetas Amarillas en cada temporada** |
| **Tarjetas Rojas** | **Número entero** | **Tarjetas Rojas en cada temporada** |
| **Once_inicial_abs** | **Número entero** | **Cantidad de partidos en el once inicial en cada temporada** |
| **Once_inicial_porc** | **Número Real** | **Porcentaje de partidos en el once inicial en cada temporada** |
| **Presencia** | **Número Real** | **Porcentaje de partidos como titular o suplente en cada temporada** |
| **Numero** | **Número entero** | **Numero de camiseta en cada temporada** |
| **Victorias_abs** | **Número entero** | **Cantidad de partidos ganados en cada temporada** |
| **Victorias_porc** | **Número Real** | **Porcentaje de partidos ganados en cada temporada** |
| **Draws_abs** | **Número entero** | **Cantidad de partidos empatados en cada temporada** |
| **Draws_porc** | **Número Real** | **Porcentaje de partidos empatados en cada temporada** |
| **Derrotas_abs** | **Número entero** | **Cantidad de partidos perdidos en cada temporada** |
| **Derrotas_porc** | **Número Real** | **Porcentaje de partidos perdidos en cada temporada** |
| **Clean Sheets** | **Número entero** | **Cantidad de partidos sin recibir goles en cada temporada (Variable solo para arqueros)** |
| **Goles concedidos** | **Número entero** | **Cantidad de goles recibidos en cada temporada (Variable solo para arqueros)** |
| **Fichajes** | **Texto (string)** | **Todas las transferencias del jugador (Fecha, Equipo y Valor de pase)** |
| **Palmares** | **Texto (string)** | **Todos los campeonatos en que el equipo se consagró en primer puesto o segundo puesto (Tipo torneo y año)** |
| **Posicion_2** | **Texto (string)** | **Agrupación de la variable "Posicion" en 4 categorías** |
| **Posicion_3** | **Texto (string)** | **Agrupación de la variable "Posicion" en 5 categorías** |

| | | |
|---|---|---|
| **Nivel** | **Número entero** | **Jerarquía del equipo de fútbol** |
| **Pais** | **Texto (string)** | **País al que pertenece el equipo de fútbol** |
| **Continente** | **Texto (string)** | **Continente al que pertenece el equipo de fútbol** |
| **Vd** | **Número entero** | **Variable dependiente a predecir** |
| **Edad** | **Número entero** | **Edad del jugador en cada temporada** |
| **Edad_debut_inter** | **Número entero** | **Edad en que el jugador debutó en un partido internacional con su selección** |

## 7.2 - Modelos utilizados y parámetros

En esta sección se presenta la descripción de los modelos, parámetros y métricas usadas. Descripción de los parámetros utilizados en el modelo Regresión Logística:

| **Modelo** | **Regresión Logística** |
|---|---|
| solver | AUTO |
| Reproducible | True |
| Maximum number of threads | 16 |
| Use regularization | False |
| Standarize | False |
| non-negative coefficients | False |
| add intercept | True |
| compute p-values | True |
| remove colinear columns | True |
| missing values handling | MeanImputation |
| max iterations | 0 |

Descripción de la importancia de cada atributo del modelo de regresión logística, ordenados ascendentemente según el p-value:

Warning: Removed collinear columns [PJ_2, Derrotas_porc, Posicion_2_Med, Posicion_3_Def, Posicion_3_Del, Posicion_3_Med-Del]

| Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value ↑ |
|---|---|---|---|---|---|
| Intercept | -1.096 | -1.393 | 0.073 | -15.067 | 0 |
| Presencia | 0.300 | 0.300 | 0.036 | 8.291 | 0.000 |
| edad | -0.224 | -0.224 | 0.029 | -7.694 | 0.000 |
| Goles_2 | 0.245 | 0.245 | 0.033 | 7.419 | 0.000 |
| Posicion_2_Def | -0.505 | -0.244 | 0.098 | -5.138 | 0.000 |
| edad_debut_inter | 0.137 | 0.137 | 0.027 | 5.032 | 0.000 |
| Once_inicial_porc | 0.403 | 0.403 | 0.081 | 4.964 | 0.000 |
| Posicion_3_Med-Def | -0.315 | -0.144 | 0.089 | -3.545 | 0.000 |
| Victorias_abs | -0.333 | -0.333 | 0.107 | -3.118 | 0.002 |
| altura | 0.092 | 0.092 | 0.038 | 2.459 | 0.014 |
| Victorias_porc | 0.170 | 0.170 | 0.070 | 2.445 | 0.014 |
| Once_inicial_abs | -0.582 | -0.582 | 0.286 | -2.036 | 0.042 |
| Minutos_Jugados | 0.473 | 0.473 | 0.257 | 1.845 | 0.065 |
| Derrotas_abs | 0.132 | 0.132 | 0.086 | 1.533 | 0.125 |
| Goles_pp | -0.040 | -0.040 | 0.030 | -1.343 | 0.179 |

| Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value ↑ |
|---|---|---|---|---|---|
| Goles_pp | -0.040 | -0.040 | 0.030 | -1.343 | 0.179 |
| peso | -0.039 | -0.039 | 0.037 | -1.052 | 0.293 |
| Asistencias | 0.027 | 0.027 | 0.026 | 1.045 | 0.296 |
| nivel | 0.030 | 0.030 | 0.030 | 1.000 | 0.318 |
| Posicion_2_Del | -0.093 | -0.037 | 0.097 | -0.965 | 0.334 |
| Efectividad | 0.022 | 0.022 | 0.031 | 0.703 | 0.482 |
| Draws_porc | 0.046 | 0.046 | 0.067 | 0.682 | 0.495 |
| Draws_abs | -0.057 | -0.057 | 0.091 | -0.635 | 0.525 |
| Tarjetas_Rojas | -0.014 | -0.014 | 0.029 | -0.465 | 0.642 |
| Tarjetas_Amarillas | -0.007 | -0.007 | 0.037 | -0.187 | 0.852 |
| PJ_2 | 0 | 0 | ? | ? | ? |
| Derrotas_porc | 0 | 0 | ? | ? | ? |
| Posicion_2_Med | 0 | 0 | ? | ? | ? |
| Posicion_3_Def | 0 | 0 | ? | ? | ? |
| Posicion_3_Del | 0 | 0 | ? | ? | ? |
| Posicion_3_Med-Del | 0 | 0 | ? | ? | ? |

Descripción de los parámetros utilizados en el modelo *Random Forest*:

| **Modelo** | **Random Forest** |
|---|---|
| Number of Trees | 50/70 |
| Criterion | information gain/gini index |
| Maximal Depth | 15 |
| Apply Pruning | False |
| Apply Prepruning | False |
| Random Splits | False |
| Guess subset ratio | True |
| Voting Strategy | Confidence Vote |

| | |
|---|---|
| Use local random seed | False |
| Enable parallel execution | True |

Descripción de los parámetros utilizados en el modelo Redes Neuronales

| **Modelo** | **Redes Neuronales** |
|---|---|
| Hidden Layers | 1 |
| Training Cycles | 20 |
| Learning Rate | 0.01/0.007 |
| Momentum | 0.9 |
| Decay | False |
| Shuffle | True |
| Normalize | True |
| Error epsilon | 1.00E-04 |
| Use local random seed | False |
| Main Criterion | First |
| Accuracy | True |

## 7.3 - Código y/o capturas de pantalla de procedimientos

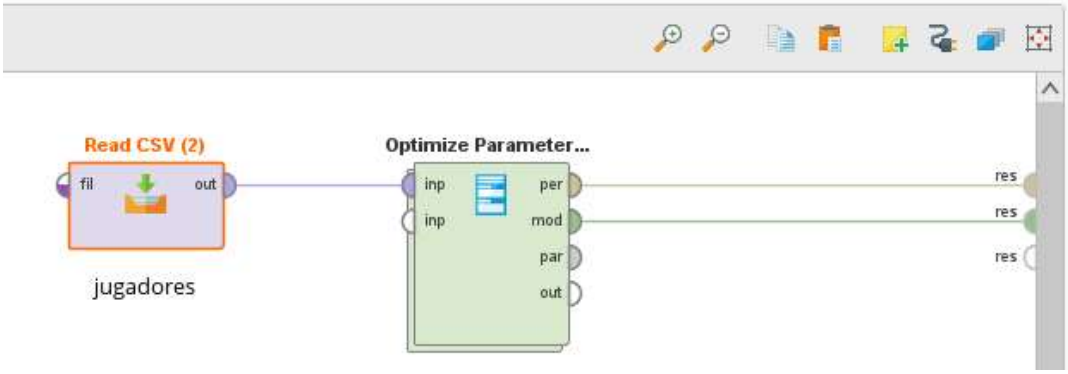
Código utilizado en *RStudio* para realizar la estadística descriptiva:

```
library(dplyr)
library(ggplot2)

#distribucion variables numericas
analisis_datos %>%
   keep(is.numeric) %>%
   gather() %>%
   ggplot(aes(value, fill=key)) +
   facet_wrap(~ key, scales = "free") +
   geom_histogram()

analisis_datos %>%
   keep(is.numeric) %>%
   gather() %>%
   ggplot(aes(value, fill=key)) +
   facet_wrap(~ key, scales = "free") +
   geom_boxplot()
```

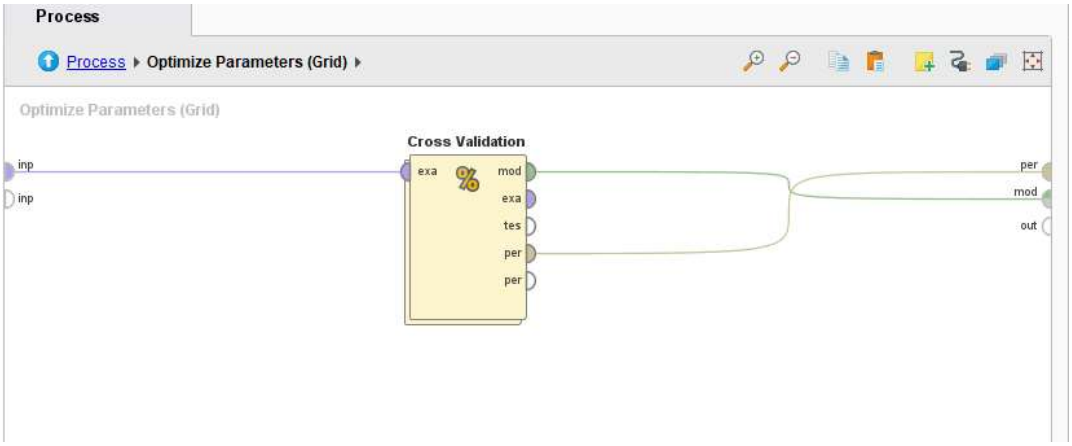
```
#Cálculo simple de estadíticos descriptivos
library(psych)
library(mnormt)
library(modeest)

desvio=sapply(analisis_datos_n, sd) #desvio
var=sapply(analisis_datos_n, var) #varianza
moda=sapply(analisis_datos_n, mfv) #moda
asimetria=sapply(analisis_datos_n, skew) #asimetria
curtosis=sapply(analisis_datos_n, kurtosi) #curtosis
#agrupo la info
cbind(desvio,var,moda,asimetria,curtosis)

#analisis descriptivo
descriptivo = as.data.frame(describe(analisis_datos_n))

write.table(descriptivo, file = "descriptivo.csv",
            sep = ",", col.names = TRUE, row.names = TRUE)
```

```r
#variables categoricas
table(analisis_datos$Equipo)
table(analisis_datos$Posicion)
table(analisis_datos$Posicion_2)

table(analisis_datos$Posicion,analisis_datos$Posicion_2)
prop.table(table(analisis_datos$Posicion,analisis_datos$Posicion_2))*100 #proporcion

library(summarytools)
freq(analisis_datos$Equipo, order = "freq")
freq(analisis_datos$Posicion, order = "freq")
freq(analisis_datos$Posicion_2)

table(analisis_datos$vd3)
table(analisis_datos$vd3)/sum(table(analisis_datos$vd3))
```

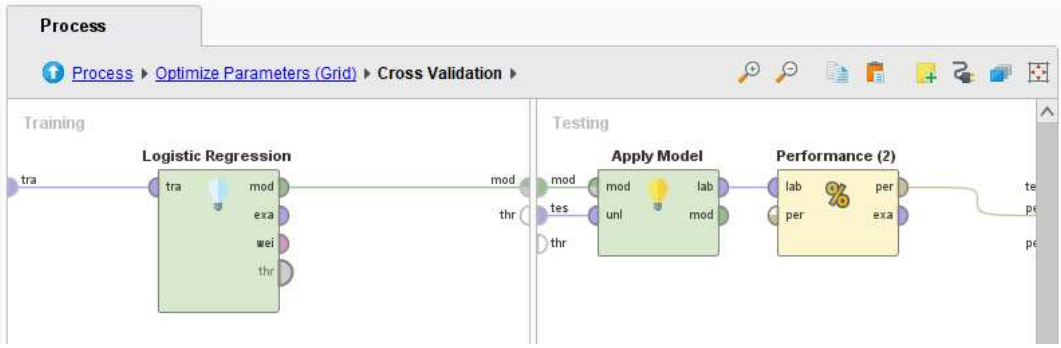
Procedimiento realizado para ejecutar los modelos en *RapidMiner Studio*, optimizando los parámetros de cada modelo:



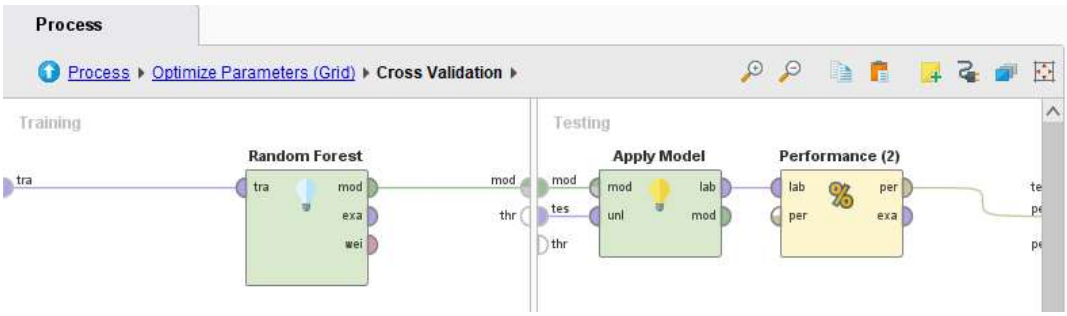

Procedimiento realizado dentro de la optimización de parámetros en *RapidMiner Studio*:



Procedimiento realizado dentro de la Validación Cruzada de Rapid Miner Studio para el modelo Regresión Logística:




Procedimiento realizado dentro de la Validación Cruzada de Rapid Miner Studio para el modelo *Random Forest*:




Procedimiento realizado dentro de la Validación Cruzada de Rapid Miner Studio para el modelo Redes Neuronales: