Universidad del Desarrollo
Facultad de Ingeniería

APROXIMACION A DETERMINANTES SOBRE EL RESULTADO DE UN PARTIDO DE FUTBOL Y RANKING DE JUGADORES DEL CAMPEONATO NACIONAL CHILENO

Predicción, Clasificación y Ranking

POR: MIGUEL ALEJANDRO NINA YANARICO

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Dra. Loreto Bravo

Sr. Hugo Contreras

DICIEMBRE - 2022

SANTIAGO

"… Seguir cuando crees que no puedes más es lo que te hace diferente a los demás…" Dedico este estudio a mi hija Fiorella, cada palabra escrita aquí, fue con ella sentada junto a mi escritorio viéndola crecer.

AGRADECIMIENTO

Al finalizar un trabajo tan arduo y lleno de hermosas dificultades como elaborar esta tesis de magister es inevitable no mencionar a aquellas personas que te prestaron de su tiempo, sin ese aporte hubiese sido imposible poder realizar todo esto, este logro hubiese sido imposible sin la participación de personas e instituciones que han facilitado las cosas para que este trabajo llegue a un feliz término. Por esto y mucho más es que para mí es todo un placer utilizar estas líneas para ser justo y consecuente con todos ellos expresándoles mis agradecimientos. Debo agradecer de manera especial y sincera a mis docentes guías Loreto Bravo y Hugo Contreras por ayudarme a realizar esta tesis bajo su dirección. Por último y no por eso menos importante mi eterna gratitud a mi familia, que durante el desarrollo de esta tesis vimos nacer y crecer a nuestra hija Fiorella, cuya mirada y risas me daban la motivación y el aliento para avanzar en esas duras noches. La perseverancia y sacrificio que poseo fueron heredados de mis padres, quienes día a día con solo una llamada me levantaba para seguir avanzando. Por ellos y para ellos seré cada día un mejor profesional. Gracias.

Tabla de contenido

1. Resumen 1

2. Introducción 1

3. Trabajo Relacionado 4

4. Hipótesis y Objetivos 6

4.1. Objetivo General 6

4.2. Objetivo especifico 6

5. Datos y Metodología 7

5.1. Datos 7

5.2. Metodología 12

Sofware Python 13

Aprendizaje supervisado 13

Support Vector Machine (SVM) 14

Naive Bayes 15

K-Nearest Neighbors (KNN) 16

Regresión Logística 17

Random Forest 17

Análisis de componentes principales (PCA) 18

6. Resultados 19

6.1. Análisis exploratorio de datos 19

Tratamiento de datos – Partidos 20

6.2. Modelos de Machine Learning 28

Preprocesamiento de variables independientes 28

Implementación de Modelos de Machine Learning 29

Análisis de componentes principales (PCA) 31

Extrapolación de coeficientes para ranking de Jugadores. 36

7. Conclusiones 39

8. Bibliografía 41

1. Resumen

En el contexto del fútbol nacional y como cualquier actividad de competición de alto rendimiento el resultado de los encuentros entre 2 rivales depende de varios factores que solo pueden ser identificados al interior de una cancha de fútbol. En un contexto general, cada partido los jugadores, directores técnicos, árbitros, incluso hasta los hinchas forman un escenario de competencia y rivalidad, por otro lado las cualidades únicas de cada jugador representan un factor determinante para la victoria, derrota o empate del equipo. En este trabajo se presentarán tres enfoques usando datos que se generan en cada partido de fútbol y datos de los jugadores de los distintos campeonato en Chile, Primero, se expondrá el origen y detalle de cada variable acompañado de un análisis exploratorio de datos. En segundo lugar se implementaran metodologías de Machine Learning para seleccionar estadísticamente aquellas variables representan la victoria, derrota o empate de los equipos y en tercer lugar, con las variables seleccionadas y compatibles con las variables del jugador se valorizara y creará un ranking de los jugadores de fútbol chileno.

2. Introducción

Es de conocimiento común que la tecnología crece exponencialmente en cada ámbito, y el fútbol no es ajeno a estos avances. El avance tecnológico no solo recae en crear herramientas capaces de desarrollar tareas o actividades más precisas y rápidas, si no que para realizar esas labores de alguna forma esa tecnología aprendió o se programó para realizar una serie de pasos repetitivos con un objetivo en común. Esos pasos repetitivos

no nacieron de la nada, sino que fueron observados, capturados y analizados durante un determinado tiempo para que esa herramienta lograra su objetivo final. Es aquí donde el Big Data y el análisis de datos se volvieron una palabra de moda en el deporte, generando una transformación digital no solo en la sociedad sino que también en el deporte y el fútbol no escapa de este cambio, en el fútbol se volvió tan común el análisis de datos que cada compra o venta de un jugador, el resultado de un partido, los jugadores titulares y las apuestas no se realizaban sin antes tener una visión estadística. Los cambios generados por el uso de la tecnología en el contexto de Big Data y análisis deportivo provocaron una revolución sistemática en el fútbol profesional. Es decir el Big data y el análisis deportivo se han convertido en herramientas complementarias e importantes en el fútbol profesional pudiendo incluso aumentar la competitividad de los clubes de fútbol profesional (Herberger & Litke, 2021).

En el fútbol profesional la responsabilidad del resultado recae principalmente en el entrenador y los jugadores, donde el entrenador influenciado por el análisis de datos selecciona 11 jugadores con características físicas sobresalientes o con algún talento detectado maximizan las probabilidad de ganar enfrentándose a otro grupo de 11 jugadores con el fin de obtener más goles que el otro equipo rival en la cancha (Antón Carranza, 2017). Siguiendo esta misma lógica, formar un equipo competitivo recae en todas aquellas características importantes dentro de una cancha de fútbol, pero esas características deben garantizar la victoria del equipo o en el peor de los escenarios un resultado de empate, entonces ¿Qué variables cuantitativas de los partidos de fútbol

acercan a la victoria? y ¿Qué jugadores pueden aumentar las probabilidad des victoria según su posición y desempeño en la cancha?

Para lograr esto, se obtuvieron datos de la plataforma de estadística deportiva Wyscout, la cual otorga 2 fuentes de datos sobre el fútbol chileno, uno de ellos son los partidos y resultados de los equipos disputados durante los años 2018 y 2022 con 110 variables cuantitativas que ocurrieron en cada partido, la otra fuente de datos es más específica y contiene la información de jugadores chilenos con 115 variables que miden las habilidades del jugador al interior de una cancha de fútbol. Desde el punto de vista como espectador, cada vez que vemos un evento deportivo de alguna liga en el mundo las estadísticas presentadas antes y después de cada partido engloban lo que conocemos como Análisis exploratorio de datos, donde podemos escuchar datos como, goleadores, minutos jugados, cantidad de pases, penaltis, infracciones, entre otros. Este estudio presentara un análisis exploratorio de datos que en esencia ayuda a realizar análisis bivariados sobre las variables de los equipos y permite detectar y planificar un grupo de acciones según los debilidades y fortalezas generales del equipo contrario generando en primera instancia controles observacionales a considerar antes de un partido (González Ramos, Martín Agüero, Montero Quesada, & Rice Nelson, 2021).

Una herramienta interesante que ha ganado popularidad hoy en día en el deporte es el uso de la inteligencia artificial enfocado en métodos de aprendizaje automático, estos ayudan a asimilar y representar los datos de los jugadores o partidos permitiendo clasificar y predecir resultados. Esto combinado con la inteligencia deportiva convierten cualquier estudio estadístico/deportivo en temas recreativos, académicos e incluso científicos

(Albarrán Jardón, 2020). Es aquí donde estas herramientas de predicción y clasificación otorgadas por la estadística tendrán un papel fundamental en el resultado de las competencias entre los equipos y en la performance del jugador, herramientas como Regresión Logística, Maquina de vectores de soporte, Árbol de Decisión y Random Forest son algunos que permiten determinar que variables destacan para que un equipo de fútbol tenga un empate, derrota o victoria.

La información hoy en día es poder, pero poder de decisión y en el mundo del fútbol escoger la mejor combinación de jugadores aumenta la probabilidad de no perder contra el equipo rival, es por ello que a través del análisis de datos ayuda a detectar y seleccionar jugadores de buen desempeño a la hora de competir.

3. Trabajo Relacionado

Con la llegada del mundial Qatar 2022 se usaron nuevas tecnologías y nuevos métodos de medición para que la competencia sea igualitaria y transparente para ambos equipos. El fútbol moderno se ha vuelto muy competitivo, por lo cual enfrenta múltiples desafíos de acuerdo con la naturaleza y dinámica del juego, esto motiva a los investigadores a identificar patrones e indicadores de rendimiento deportivo (Dufour, Phillips, & Ernwein, 2017), el objetivo de estos patrones e indicadores es que puedan brindar a los entrenadores información útil sobre el análisis individual y colectivo de un equipo, no solo en el fútbol, sino que también en deportes como Basketball y Hockey (Gudmundsson & Horton, 2018). Gracias a la recolección masiva de datos la capacidad de poder realizar evaluaciones cuantitativas de los jugadores de fútbol para medir su rendimiento en la cancha resulta una

tarea más rápida y eficiente, ya que por un tiempo prolongado las mediciones solo se se basaron en la cantidad de goles marcados, asistencias realizadas y tiros al arco realizados. (Brooks, Kerr, & Guttag, 2016) elaboraron un modelo clasificador de jugadores basado en el valor de los pases completados. El modelo aplicado es de aprendizaje automático que relaciona los pases en una posición determinada y las oportunidades de tiro generada, usando solo los pases los resultados se asemejan bastante a la percepción de la realidad futbolística, destacando a Messi y Cristiano Ronaldo.

El avance exponencial en tecnologías nombrado anteriormente proporciona flujos de datos de alta fidelidad conseguidos en cada partido de fútbol, esto ayuda que sustentar el enfoque basado en datos y demuestra que existe un potencial para impulsar la compresión cuantitativa del rendimiento del equipo de fútbol, (Cintia, Pappalardo, Pedreschi, & Giannotti, 2015) elaboraron un conjunto de indicadores de rendimiento basado en los pases que realizaba cada equipo, este indicador presento una fuerte correlación con el éxito del equipo, generando que en las simulaciones los resultados de cada equipo fueran cercana a las reales demostrando que el fútbol tiene un potencial de revelar patrones ocultos y comportamientos de calidad.

Con una fuente de datos robusta solo queda aplicar un análisis estadístico que respalde los patrones y el comportamiento de las variables, los métodos estadísticos son variados, según (Barreto, Arruda, & Cunha, 2014), la exploración estadística es de vital importancia en el fútbol durante una competencia, la idea de seleccionar variables representativas para distinguir a los equipos ganadores de los empatados y perdedores, para ello recurrieron al análisis de componentes principales y clusterización, demostrando que la selección y

clasificación se asemejaba bastante a los resultados reales, presentando una forma diferente de usar la estadística en los partidos de fútbol, mostrando diferencias multivariadas entre equipos exitosos y no exitosos.

Fuente de datos, metodologías, selección y clasificación es lo que abarca este estudio, para finalizar según los pronósticos y clasificación se realizara un ranking de los jugadores y sus posiciones. (Pappalardo, y otros, 2018) implementaron Playerank, un método basado en atributos específicos de cada jugador y eventos que ocurren en un partido de fútbol, ofreciendo una evaluación basada en principios multidimensionales, roles y el desempeño del jugador, los resultados de su estudio fueron contrastados con expertos en la materia estando en sintonía con sus conclusiones, demostrando que esta metodología resulta eficiente y flexible para el análisis de fútbol.

4. Hipótesis y Objetivos

Suponemos que las variables para predecir el resultado de un partido son representativos para medir el desempeño de un jugador según sus atributos y posición en la cancha.

4.1. Objetivo General

- Crear modelos de Machine Learning y elaborar un ranking de jugadores de fútbol por posición según su desempeño en los partidos.

4.2. Objetivo especifico

- Elaborar una base de datos con atributos representativos del partido y del jugador.

- Implementar 5 metodología de predicción y clasificación de resultados según atributos del partido de fútbol.
- Analizar el aporte de las variables en la predicción del modelo seleccionado
- Presentar reflexiones obtenidas relacionadas al estudio de los resultados de selección, clasificación y ranking.

5. Datos y Metodología

5.1. Datos

Wyscout es una plataforma futbolera de internet multidispositivo con una trayectoria de más de 20 años con datos de rendimiento de los equipos y jugadores, su plataforma es utilizada por muchos países para la exploración análisis y rendimiento tanto del equipo, como del jugador, sus servicios van desde la digitalización de competencias como videos de partidos o movimientos específicos de un jugador, hasta la captura, procesamiento, análisis y comparación de datos enfocados a las habilidades del jugador. Todo esto con el propósito de que el usuario tome decisiones basadas en datos para alcanzar resultados de la forma más rápida e inteligentemente posible. Wyscout a través de su plataforma otorga un servicio de vanguardia para garantizar la victoria en los partidos de fútbol, su factor diferenciador recae en que se puede navegar, analizar y comparar jugadores y así quizás encontrar la próxima estrella de fútbol.

Los datos que se extrajeron de esta plataforma Wyscout se sintetizan en la tabla Nº1, los cuales poseen un su mayoría datos cuantitativos sobre atributos que ocurren en los partido

del fútbol profesional chileno durante los años 2018 a 2022 y por otro lado los atributos que posee cada jugador en su trayectoria como jugador de fútbol.

El conjunto de datos se encuentro en formato MS Excel (.xlsx) de tipo columnas con datos en su mayoría cuantitativos y pocos datos cualitativos, con énfasis en los jugadores, equipos y partidos disputados.

Tabla Nº1: Conjunto de datos	
Partidos disputados durante 2018-2022	**Jugadores Chilenos**
Cantidad de partidos: 2862	Cantidad de jugadores: 743
Cantidad de variables: 110	Cantidad de variables: 115
Fuente: Wyscout, 2022	

Enfocándose en el conjunto de datos de los partidos un atributo que servirá para crear el modelo de predicción recae en los goles, ya que al no existir una variable objetivo (a predecir) que indique, victoria, empate o derrota, esta tendrá que crearse, cada partido se define por la capacidad de hacer goles y de tener más goles que el rival, por ello para crear esta variable se usara la diferencia entre los goles realizados y los goles recibidos en cada competencia, como lo indica la tabla Nº2.

Tabla Nº2: Categorización resultado del partido	
Goles realizados < Goles recibidos	Derrota (0)
Goles realizados = Goles recibidos	Empate (1)
Goles realizados > Goles recibidos	Victoria (2)
Fuente: Elaboración propia considerando la regla base del fútbol.	

El resultado del partido será considerado como la variable target (independiente o explicada) y en conjunto trabajara con el resto de variables explicativas, de las 109 restantes solo 105 corresponden a variables cuantitativas de tipo discreta y continua, sin

embargo algunas de esas variables son representaciones porcentuales de otras variables en el mismo conjunto de datos como se aprecia en la tabla Nº3.

Tabla Nº3: Conjunto de datos - Partidos			
Fecha (D)	Ataques posicionales con remate (I)	Goles recibidos (I)	Pases en el último tercio (I)
Perfil (S)	Ataques posicionales con remate% (F)	Tiros en contra (I)	Pases en el último tercio logrados (I)
Partido (S)	Contraataques (I)	Tiros en contra a la portería (I)	Pases en el último tercio logrados (I)
Competición (S)	Contraataques con remate (I)	Tiros en contra a la portería% (F)	Pases progresivos (I)
Duración (I)	Contraataques con remate%	Duelos defensivos (I)	Pases progresivos precisos (I)
Seleccionar esquema (S)	Jugadas a balón parado (I)	Duelos defensivos ganados (I)	Pases progresivos precisos%
Goles (I)	Jugadas a balón parado con remate (I)	Duelos defensivos ganados%	Desmarques (I)
xG (F)	Jugadas a balón parado con remate% (F)	Duelos aéreos (I)	Desmarques logrados
Tiros (I)	Córneres (I)	Duelos aéreos ganados (I)	Desmarques logrados% (F)
Tiros a la portería (I)	Córneres con remate (I)	Duelos aéreos ganados% (F)	Saques laterales (I)
Tiros a la portería% (F)	Córneres con remate% (F)	Entradas a ras de suelo	Saques laterales logrados
Pases (I)	Tiros libres (I)	Entradas a ras de suelo logradas (I)	Saques laterales logrados% (F)
Pases logrados (I)	Tiros libres con remate (I)	Entradas a ras de suelo logradas% (F)	Saques de meta (I)
Pases logrados% (F)	Tiros libres con remate% (F)	Interceptaciones	Intensidad de paso
Posesión del balón, % (F)	Penaltis (I)	Despejes (I)	Promedio pases por posesión del balón (I)
Balones perdidos (I)	Penaltis marcados (I)	Faltas (I)	Lanzamiento largo % (F)
Balones perdidos bajos (I)	Penaltis marcados% (F)	Tarjetas amarillas (I)	Distancia media de tiro (I)
Balones perdidos medios	Centros (I)	Tarjetas rojas (I)	Longitud media pases (I)
Balones perdidos altos (I)	Centros precisos (I)	Pases hacia adelante (I)	PPDA
Balones recuperados (I)	Centros precisos% (F)	Pases hacia adelante logrados (I)	Duelos (I)
Balones recuperados bajos (I)	Pases cruzados en profundidad completados (I)	Pases hacia adelante logrados% (F)	
Balones recuperados medios (I)	Pases en profundidad completados (I)	Pases hacia atrás (I)	

Balones recuperados altos (I)	Entradas al área de penalti (I)	Entradas al área de penalti (carreras) (I)	
Fuente: Wyscout. Entre paréntesis tipo de dato I= Integer, F = Float, S= String y D = Date			

Por otro lado, el conjunto de datos de los jugadores presenta una estructura muy parecida a los partidos, hay que considerar que la similitud entre los datos que se recopilan como equipos son conseguidas gracias a los movimientos y acciones que generan los jugadores al interior de la cancha, la tabla Nº4 resume el conjunto de variables otorgadas por Wyscout para los jugadore chilenos.

Tabla Nº4: Conjunto de datos – Jugadores			
Jugador (S)	Posesión conquistada después de una interceptación (F)	Carreras en progresión/90 (F)	Pases en profundidad/90 (F)
Equipo (S)	Faltas/90 (F)	Aceleraciones/90 (F)	Precisión pases en profundidad, % (F)
Equipo durante el período seleccionado (S)	Tarjetas amarillas (I)	Pases recibidos /90 (F)	Ataque en profundidad/90 (F)
Posición específica (S)	Tarjetas amarillas/90 (F)	Pases largos recibidos/90 (F)	Centros desde el último tercio/90 (F)
Edad (I)	Tarjetas rojas (I)	Faltas recibidas/90 (F)	Pases progresivos/90 (F)
Valor de mercado (I)	Tarjetas rojas/90 (F)	Pases/90 (F)	Precisión pases progresivos, % (F)
Vencimiento contrato (D)	Acciones de ataque exitosas/90 (F)	Precisión pases, % (F)	Goles recibidos (I)
Partidos jugados (I)	Goles/90 (F)	Pases hacia adelante/90 (F)	Goles recibidos/90 (F)
Minutos jugados (I)	Goles (excepto los penaltis) (I)	Precisión pases hacia adelante, % (F)	Remates en contra (I)
Goles (I)	Goles, excepto los penaltis/90 (F)	Pases hacia atrás/90 (F)	Remates en contra/90 (F)
xG (F)	xG/90 (F)	Precision pases hacia atrás, % (F)	Porterías imbatidas en los 90 (F)
Asistencias (I)	Goles de cabeza (I)	Pases laterales/90 (F)	Paradas, % (F)
xA (F)	Goles de cabeza/90 (F)	Precisión pases laterales, % (F)	xG en contra (F)
Duelos/90 (F)	Remates (I)	Pases cortos / medios /90 (F)	xG en contra/90 (F)
Duelos ganados, % (F)	Remates/90 (F)	Precisión pases cortos / medios, % (F)	Goles evitados (I)
País de nacimiento (S)	Tiros a la portería, % (F)	Pases largos/90 (F)	Goles evitados/90 (F)

Pasaporte (S)	Goles hechos, % (F)	Precisión pases largos, % (F)	Pases hacía atrás recibidos del arquero/90 (F)
Pie (S)	Asistencias/90 (F)	Longitud media pases, m (F)	Salidas/90 (F)
Altura (I)	Centros/90 (F)	Longitud media pases largos, m (F)	Duelos aéreos en los 90 (F)
Peso (I)	Precisión centros, % (F)	xA/90 (F)	Tiros libres/90 (F)
En préstamo (S)	Centros desde la banda izquierda/90 (F)	Asistencias/90 (F)	Tiros libres directos/90 (F)
Acciones defensivas realizadas/90 (F)	Precisión centros desde la banda izquierda, % (F)	Second assists/90 (F)	Tiros libres directos, % (F)
Duelos defensivos/90 (F)	Centros desde la banda derecha/90 (F)	Third assists/90 (F)	Córneres/90 (F)
Duelos defensivos ganados, % (F)	Precisión centros desde la banda derecha, % (F)	Desmarques/90 (F)	Penaltis a favor (I)
Duelos aéreos en los 90 (F)	Centros al área pequeña/90 (F)	Precisión desmarques, % (F)	Penaltis realizados, % (F)
Duelos aéreos ganados, % (F)	Regates/90 (F)	Jugadas claves/90 (F)	
Entradas/90 (F)	Regates realizados, % (F)	Pases en el último tercio/90 (F)	
Posesión conquistada después de una entrada (F)	Duelos atacantes/90 (F)	Precisión pases en el último tercio, % (F)	
Tiros interceptados/90 (F)	Duelos atacantes ganados, % (F)	Pases al área de penalti/90 (F)	
Interceptaciones/90 (F)	Toques en el área de penalti/90 (F)	Pases hacía el área pequeña, % (F)	
Fuente: Wyscout. Entre paréntesis tipo de dato I= Integer, F = Float, S= String y D = Date			

Es importante destacar que para este estudio no se usaron todas las variables, ya que considerando el objetivo de este estudio es crear un ranking de los jugadores chilenos y para ello solo tenemos aquellas variables que son representativas una con la otra, por ejemplo, la cantidad de pases que ocurren en un partido tiene su variable representativa en el conjunto de datos del jugador que sería la cantidad de pases que realiza el jugador. Entonces nuestro conjunto de datos para predecir el resultado de un partido se acota a 54 variables cuantitativas, las cuales tienen su variable representativa en el conjunto de los

jugadores, la tabla Nº5 detalla las variables explicativas para predecir y que también se ocupara para crear el ranking de jugadores, esta trabajara en conjunto con la variable objetivo, el resultado del partido.

Tabla Nº5: Resumen de variables independientes			
Lanzamiento largo %	Centros/90	Córneres/90	Desmarques/90
Duelos aéreos en los 90	Duelos aéreos ganados %	Duelos atacantes ganados %	Duelos atacantes/90
Duelos defensivos ganados %	Duelos defensivos/90	Duelos ganados %	Duelos/90
Faltas/90	Goles	Goles recibidos/90	Interceptaciones/90
Longitud media pases m	Pases en el último tercio/90	Pases hacia adelante/90	Pases hacia atrás/90
Pases largos/90	Pases laterales/90	Pases progresivos/90	Pases/90
Penaltis a favor	Penaltis realizados %	Precisión centros %	Precisión desmarques %
Precisión pases en el último tercio %	Precisión pases hacia adelante %	Precisión pases hacia atrás %	Precisión pases largos %
Precisión pases laterales %	Precisión pases progresivos %	Precisión pases %	Remates/90
Penaltis marcados	Pases logrados	Duelos ganados	Centros precisos
Duelos ofensivos ganados	Duelos defensivos ganados	Duelos aéreos ganados	Pases hacia adelante logrados
Pases hacia atrás logrados	Pases laterales logrados	Tarjetas amarillas/90	Tarjetas rojas/90
Tiros a la portería %	Tiros libres/90	Toques en el área de penalti/90	xG/90
Pases progresivos precisos	Pases en el último tercio logrados		
Fuente: Wyscout			

5.2. Metodología

En esta sección se describe el enfoque y los procesos metodológicos utilizados en este estudio, para la resolución de los objetivos planteados en los capítulos anteriores se han

utilizado diferentes técnicas de Machine Learning a través del programa informático Python y sus respectivas librerías para análisis.

Sofware Python

Python es un lenguaje de programación ampliamente utilizado en aplicaciones web, el desarrollo de software, la ciencia de datos y Machine Learning. Este lenguaje es multiplataforma, lo que hace sencillo si utilización e interpretación. Para este estudio la sección de ciencia de datos de Python se acopla muy bien, ya que nos permite extraer información importante a partir de los datos y así efectuar predicciones precisas. Con Python y sus librerías para tratamiento de datos como Numpy, Pandas, Seaborn, Matplotlib y SKLearn, entre otros aporta una alta eficiencia en la ciencia de datos ayudando a corregir y limpiar datos, estadístico descriptivos, visualización de datos mediante gráficos y sobre todo para crear y entrenar modelos de Machine Learning para predicción o clasificación.

Aprendizaje supervisado

El presente estudio se enfoca en 5 técnicas de Machine Learning que se distinguen por procesar datos previamente clasificados, por ejemplo el conjunto de datos de los partidos posee una variable objetivo a clasificar Victoria, Derrota y Empate. Es decir este método conoce a priori el grupo o clase a la que pertenecen los datos recolectados, y por tanto la variable respuesta es conocida.

El aprendizaje supervisado es una técnica que utiliza 2 conjunto de datos, uno de entrenamiento y otro que contiene las predicciones, es decir valores de entrada y valores

de salida, de esta forma con un conjunto de datos clasificados enmarcadas en una situación (para este caso el deporte fútbol) se enseña al algoritmo de aprendizaje supervisado crear un modelo que encuentre patrones que puedan aplicarse en un análisis (Mueller & Massaron, 2016). Un aspecto importante a recalcar es que estos algoritmos con frecuencia permiten usar un conjunto de datos de prueba para validar el modelo.

A continuación se mencionaran los aspectos principales de las técnicas de Machine Learning usadas en este estudio.

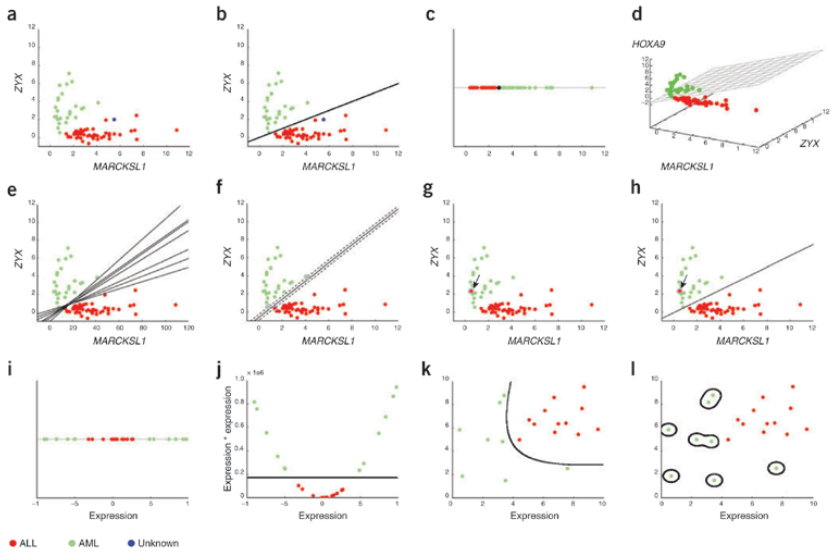
Support Vector Machine (SVM)

SVM es un algoritmo no probabilístico apropiado para resolver problemas de clasificación, este algoritmo se enfoca en clasificar las observaciones considerándolas como vectores de P-dimensiones, siendo este plano en su origen de forma lineal. SVM calcula el hiperplano optimo que divida los puntos cada observación en diferentes hiperplanos, de esta forma clasifica cada categoría en distintos hiperplanos (Zabarte, 2022).

SVM ha ganado su reputación por ser usado en distintas áreas de la ciencia ayudando no solo a clasificar patrones, sino que también ha realizado aportes en estudios de delitos bancarios, reconocimiento de imágenes y clasificación genética como en la imagen Nº1.

La flexibilidad de SVM permite ajustar los hiperparametros para mejorar su predicción en la clasificación haciendo de este algoritmo una herramienta útil para predecir los resultados de un partido de fútbol.

Imagen 1: Clasificación SVM



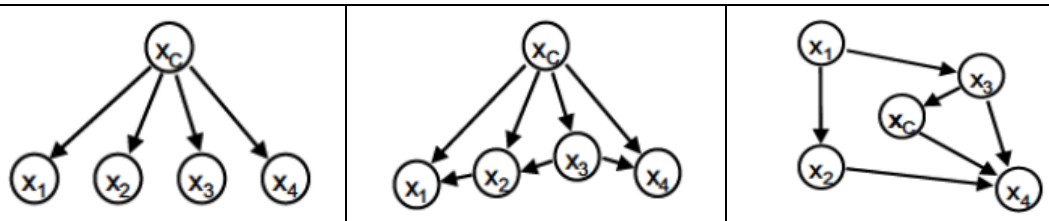
Fuente: SVM para la predicción de leucemia (Noble, 2006)

Naive Bayes

Técnica basada en el teorema de Bayes, que se caracteriza por considerar que cada suceso de eventos está condicionado por la probabilidad de un evento anterior, esta técnica supone que las variables explicativas del modelo son independientes entre ellas (Maron, 1961).

Usar este algoritmo le da una especial importancia a las variables explicativas ya que asume que la presencia o ausencia de otra variable no está relacionada con cualquier otra característica, esto hace que sea bastante precisa con una pequeña cantidad de datos de entrenamiento, posterior a eso realiza la predicción de la clase considerando la mayor probabilidad de ocurrencia del último evento (Imagen Nº2).

Imagen Nº2: Clasificador con redes bayesianas

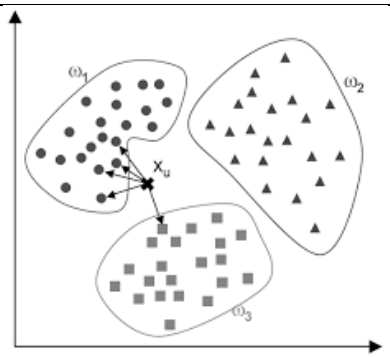


Fuente: Redes bayesianas para clasificación de resultados de fútbol en la Premier Ligue (Abdul , Mustapha, & Fauzi, 2018)

K-Nearest Neighbors (KNN)

El algoritmo KNN utiliza los eventos clasificados para predecir nuevas clasificaciones, su cálculo se basa en la proximidad entre los distintos grupos de clases y las nuevas clasificaciones, no solo es útil para conjunto de datos supervisados sino que también es precisa con conjuntos no supervisados (no poseen una clasificación), todo esto gracias al cálculo de las distancias euclidianas entre los grupos más cercanos y la correlación con la distancia (Imagen Nº3).

Imagen Nº3: Grafico proximidad entre grupos de clases para predicción de clasificadores



Fuente: Evaluación de futbolistas según atributos para determinadas posiciones (Bazmara & Jafari, 2013)

Regresión Logística

El algoritmo de regresión logística es uno de los más utilizado en Machine Learning, esta técnica mide la relación entre la variable dependiente y sus variables independientes, la diferencia con la regresión lineal es que esta metodología usa una función logística que calcula la probabilidad de que ocurra la variable dependiente, logrando así una clasificación.

La ecuación de una regresión logística relaciona la variable dependiente con las independientes donde los β_k son los coeficientes que permite medir efecto que produce cada variables x_i en la variable dependiente. :

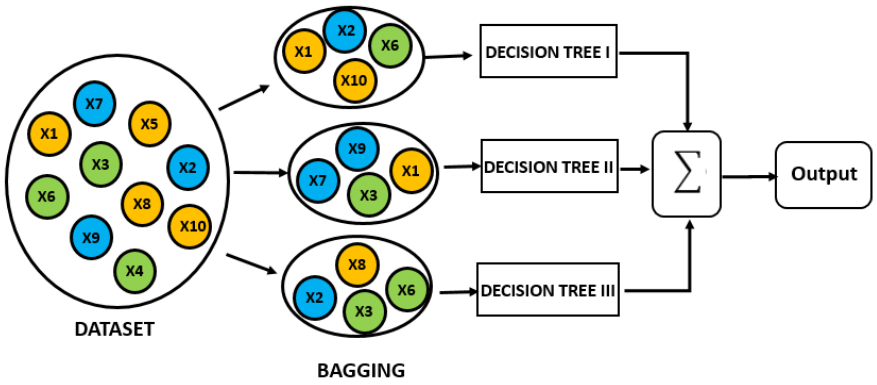
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}$$

Random Forest

El algoritmo Random Forest es una metodología que combina arboles de decisión de forma aleatoria con el objetivo de mejorar la capacidad de predicción (Breiman, 1996), es decir crea múltiples subconjuntos de datos, construye sus modelos y luego los combina calculando un promedio de todas las predicciones, de esta forma considera los mejores coeficientes de clasificación (Imagen Nº4). Una de las ventajas de Random Forest es que teóricamente tiene muy pocas suposiciones para la preparación de datos y es posible usarlo como método no supervisado

Imagen Nº4: Proceso de Bagging en técnica de Random Forest

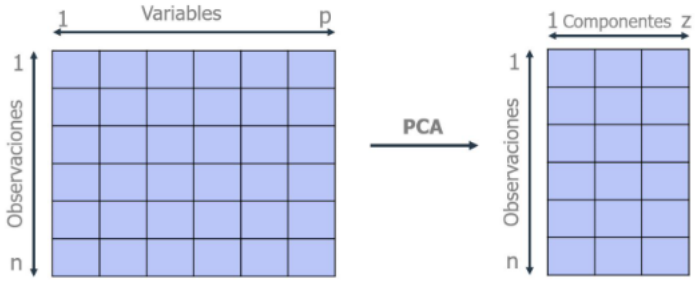


Fuente: Algoritmos basados en arboles de decisión (Orellana, 2018).

Análisis de componentes principales (PCA)

PCA es una técnica para exploración de datos, el cual busca reducir la dimensionalidad de los datos conservando su información, ósea su objetivo es simplificar el modelo (Imagen Nº5) creando una matriz de covarianza y correlación de cada variable generando un nuevo conjunto con menos variables, llamados componentes principales, el principal requisito de esta técnica es normalizar las variables independientes, de esta forma eliminar cualquier efecto que tengan las magnitudes de las variables, al sumar sus varianzas explicadas podemos medir el impacto de la variación según la cantidad de componentes a utilizar para la predicción, algunos documento mencionan que se deben considerar la cantidad de componentes principales que al sumarlos lleguen al 80%, esto nos ayuda a demostrar que en el modelo original pueden existir variables que otorgan poca información a la predicción y eliminarlas incluso puede mantener aumentar o mantener la capacidad de predicción del modelo (Wold, Esbensen, & Geladi, 1987).

Imagen Nº5: Transformación mediante PCA



Fuente: Análisis de componente principales mediante Python (Amat, 2020)

6. Resultados

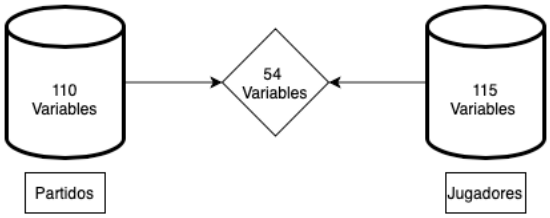
6.1. Análisis exploratorio de datos

Definidas las metodologías usadas en este estudio esta sección corresponde a la definición de los conjunto de datos tratados y un análisis exploratorio.

El presente estudio contempla 2 conjuntos de datos, Partidos y jugadores, cada uno con su conjunto de variables. Como los principales objetivos son predecir el resultado de un partido de fútbol y elaborar un ranking de jugadores según sus atributos generados en cada partido, podemos considerar una relación de contexto entre ambos conjuntos, es decir si un partido de fútbol posee una variable "Pases realizados en el partido", podemos asegurar que esos pases fueron realizados por jugadores de fútbol y no por otro individuo, por ello podemos generar una relación entre algunas variables y así reducir la dimensionalidad de ambos conjuntos de datos.

Al realizar esta relación entre ambos conjuntos de datos la cantidad de variables se reducen a 54 características cuantitativas en común (Imagen Nº6).

Imagen Nº6: Reducción de variables compatibles entre ambos conjuntos de datos



Fuente: Datos extraídos de la plataforma Wyscout, 54 variables compatibles de las cuales 12 fueron homologadas mediante operación de representación entre 2 variables.

Tratamiento de datos – Partidos

La compatibilidad de variables entre el conjunto de datos de partidos y jugador contempla 54 variables, de las cuales no poseían ausencia de datos, evitando el tratamiento de limpieza de datos y mayores supuestos sobre las variables (Tabla Nº6).

Tabla 6: Conjunto de variables para entrenamiento y predicción

Lanzamiento largo %	Goles recibidos	% de Pases en el último tercio logrados	Duelos ofensivos ganados
Centros	Interceptaciones	% de Pases hacia adelante logrados	Duelos defensivos ganados
Córneres	Longitud media pases	% de Pases hacia atrás logrados	Duelos aéreos ganados
Desmarques	Pases en el último tercio	% de Pases largos logrados	Pases hacia adelante logrados
Duelos aéreos	Pases hacia adelante	% de Pases laterales logrados	Pases hacia atrás logrados
% Duelos aéreos ganados	Pases hacia atrás	% de Pases progresivos precisos	Pases laterales logrados
% de Duelos ofensivos ganados	Pases largos	% de pases logrados	Tarjetas amarillas
Duelos ofensivos	Pases laterales	Tiros	Tarjetas rojas
% de Duelos defensivos ganados	Pases progresivos	Penaltis marcados	%Tiros a portería
Duelos defensivos	Pases	Tiros libres	Pases en el último tercio logrados

% de duelos ganados	Penaltis	Toques en el área de penalti	Pases progresivos precisos
Duelos	% de Penaltis marcados	Pases logrados	xG
Faltas	% de Centros precisos	Duelos ganados	
Goles	% de Desmarques logrados	Centros precisos	
Fuente: Datos extraidos de plataforma Wyscout.			

Para definir variable objetivo a predecir se consideró el resultado de cada partido, en cualquier partido de fútbol en el mundo el resultado se define por la cantidad de goles que realiza el equipo, por ello el equipo ganador debe tener más goles que el equipo adversario, sin embargo en un partido de fútbol pueden ocurrir 3 estados, Victoria, derrota y empate. Para obtener una visión panorámica del comportamientos de las 54 variables resulta de utilidad usar Boxplot o diagrama de cajas, la cual contempla todos los estadísticos descriptivos de las variables.

La variable objetivo siempre estará expresada de la misma forma para su clasificación, sin embargo las variables independientes son variadas y expresadas en distintos escalas, un ejemplo es la relación entre la cantidad de goles (media = 1,3) y la cantidad de pases (media = 381) , la comparación entre ambas variables es de un rango muy amplio, y esto ocurre con otras variables independientes. Por ello a continuación se presenta una visualización de las variables a través de un Análisis exploratorio de datos para cuantificar el comportamiento de las variables. Este análisis es útil para considerar como van a ingresar las variables en las distintas técnicas de Machine Learning.

En la imagen Nº7 con la ayuda de Boxplot o diagrama de caja podemos representar gráficamente el comportamiento de las 54 variables independientes, es importante destacar que debido a las dimensiones de las variables para realizar el Boxplot las

variables se estandarizaron y así tener una imagen más compacta sobre la media, mediana y sus respectivos cuartiles, sin embargo los efectos de las magnitudes de las variables no se pueden eliminar por completo, precisamente porque los eventos que ocurren en los partidos como la cantidad de tarjetas rojas muchas veces están ausentes o ocurren pocas veces, por ello es normal ver algunas variables como Penaltis con pocas distribuciones de valores. A través de Boxplot podemos confirmar la presencia de valores atípicos, que si lo llevamos al contexto de los partidos de fútbol se pueden considerar normales, tomando como ejemplo la cantidad de pases, la diferencia entre la cantidad de pases que puede realizar un arquero comparado con la cantidad de pases que realiza un jugador del medio campo es notoria, básicamente porque uno de los objetivos del equipo es que el arquero tenga ninguna actividad de peligro del gol.

Un detalle interesante es que tenemos variables que poseen distribución distintas, algunas con asimetría positiva como lo son los “pases laterales logrados %” y los “pases hacia atrás logrados” y otras negativas como la variable “Gol”, centrándose la información al principio o al final de cada variable. También podemos observar variables que tienen una distribución de datos y bigotes muy parecido como lo son los duelos ofensivos ganados y la cantidad de duelos ofensivos, esto puede ser un pequeño indicio acerca del aporte de aquellas variables en el modelo predictor, tema que se verá en los siguientes capítulos.

Imagen Nº7: Boxplot de las variables independientes estandarizadas

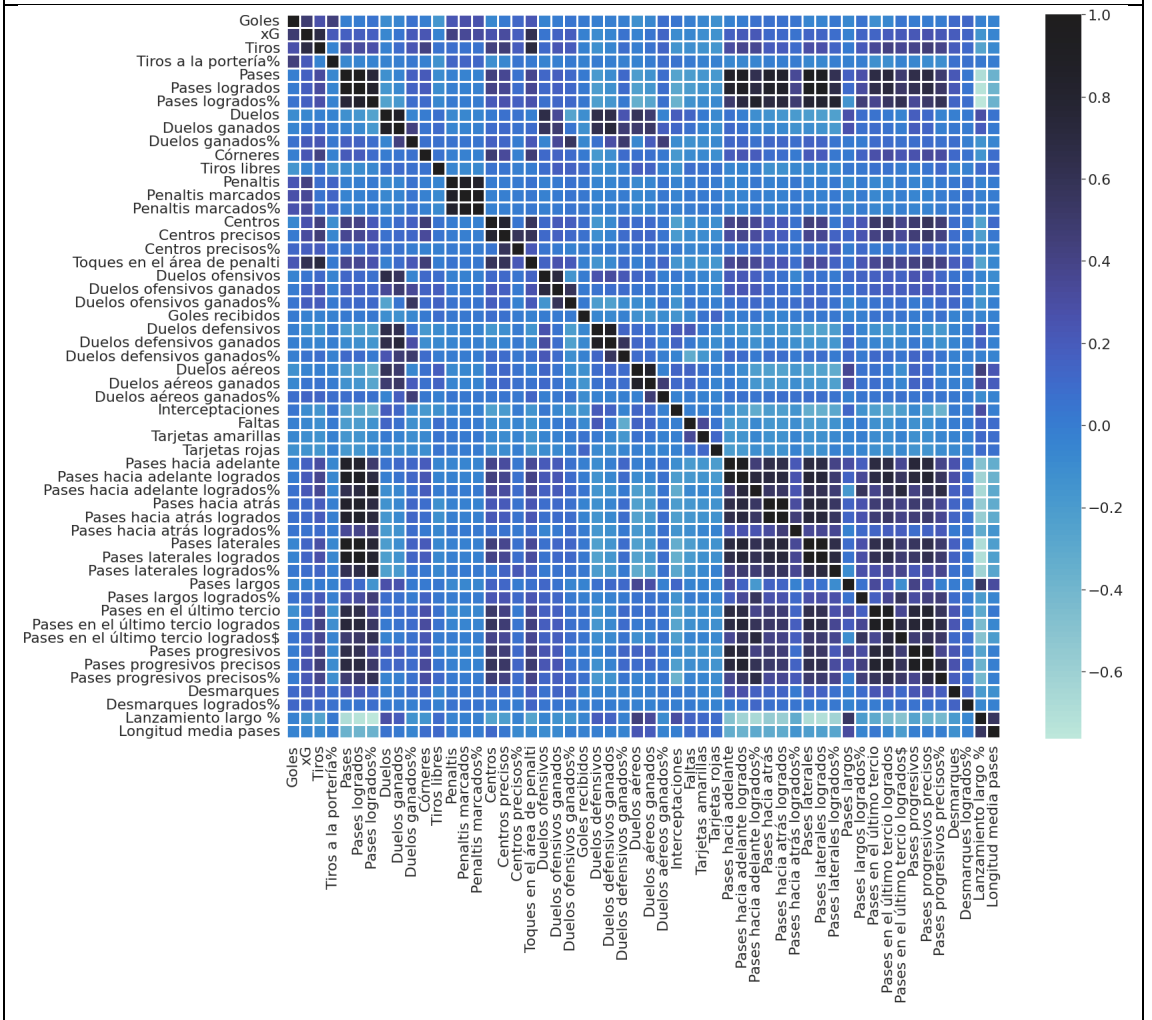
Fuente: Elaboración propia con datos de Wyscout

Previamente antes de modelar es importante revisar la relación que existe entre las variables independientes, un método adecuado es usar un mapa de correlación, la Imagen Nº8 muestra precisamente eso, una representación gráfica sobre la relación positiva, negativa o nula que hay entre las 54 variables independientes usando el coeficiente de Spearman. Esta sección es importante de considerar, ya que con la ayuda de esta imagen podemos estudiar las posibles relaciones que hay entre 2 variables y probar una posible relación causa, efecto o complemento, con el mapa de correlación se identifica y confirma que existe relaciones entre variables y que tan fuerte o débil es su intensidad, con esto se puede tomar importantes decisiones para los modelos de Machine Learning.

Por ejemplo, en la parte inferior izquierda del mapa de correlación podemos notar un grupo considerable de variables con una intensa relación positiva que poseen el nombre "pase", no es extraño pensar que si un jugador realiza muchos pases en un partido a medida que aumentan los pases también aumentaran la cantidad de pases logrados, mismo escenario con la cantidad de duelos, mientras más duelos realice un jugador puede aumentar sus duelos ganados o perdidos. Como contexto general podemos afirmar que tenemos una cantidad considerable de variables tienen una relación intensamente positiva, esto quiere decir que cuando realicemos los modelos de predicción, estas variables aportara una cuota importante, incluso con este mapa podemos verificar que existen variables que pueden estar aportando la misma información para una predicción, como lo son las variables que contienen la palabra pase. Desde otra arista, también podemos decir que aquellos cuadrantes donde el color es relativamente más claro, es decir cercano a cero, son variables que quizás no tienen ninguna relación con otras variables y sean

independientes del comportamiento del resto como lo es la variables “interceptaciones”, estas variables en algunos modelos incluso pueden otorgar casi nada en la predicción. Sin embargo podemos asegurar que tenemos variables muy correlacionadas, esto sin duda será un tema a tratar para entender la relación estadística de la variación con la variable independiente (resultado del partido) y así poder obtener el mejor ajuste al minimizar la desviación entre la predicción y el resultado real.

Imagen Nº8: Mapa de correlación, variables independientes

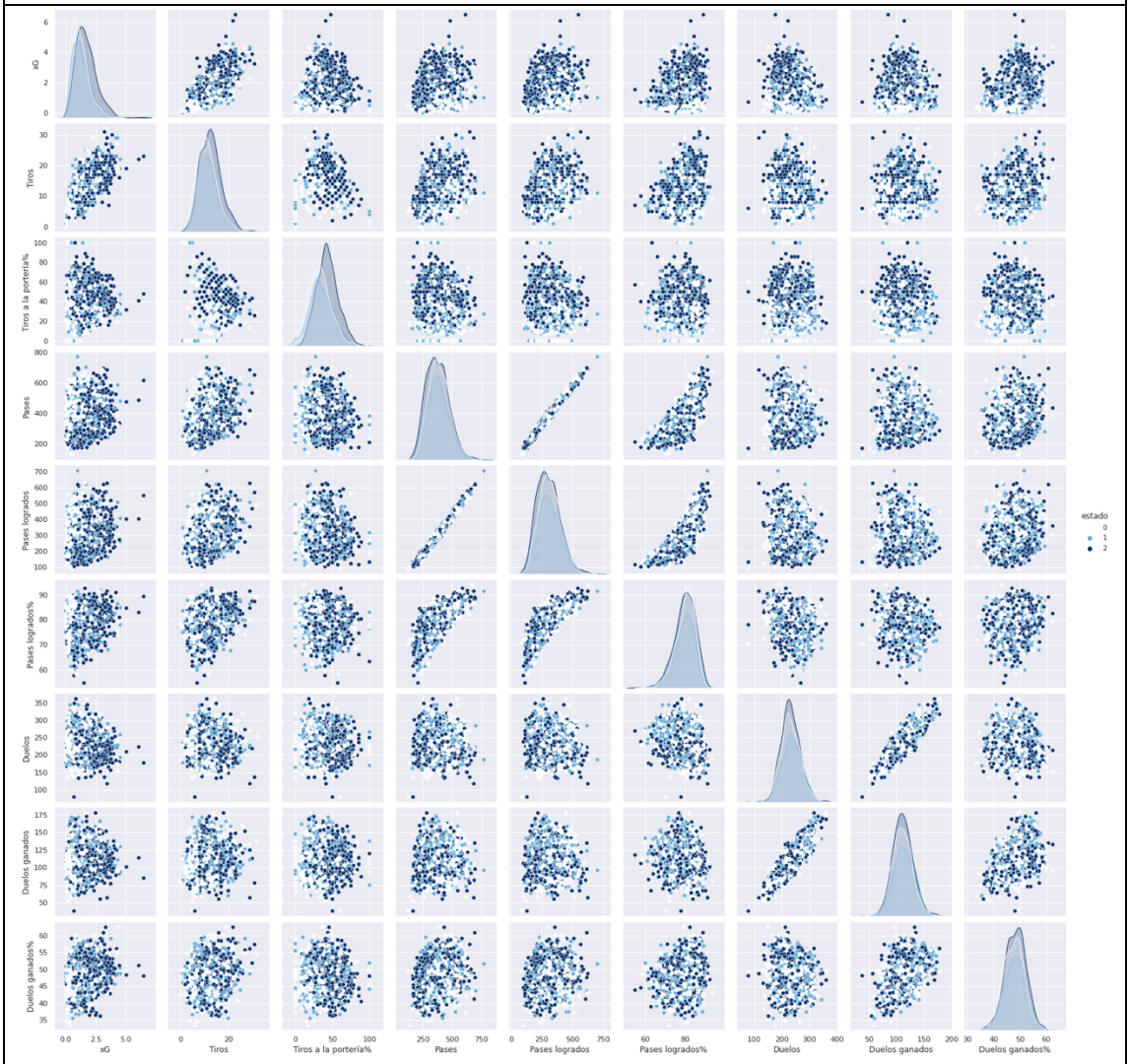


Fuente: Elaboración propia con datos de Wyscout.

Por ultimo en la Imagen Nº9 tenemos la relación que existe entre el comportamiento de las variables independientes con la variable objetivo, a través de este conjunto de gráficos de dispersión permite comprender la relación por pares de variables independientes, además categoriza la dispersión según el resultado del partido, según lo mencionado en capítulos anteriores, “0” indicara derrota, “1” indicara empate y “2” indicara victoria. El grafico 9 solo presenta 10 variables a modo de representación, en esta matriz respaldamos lo que se menciona en el mapa de correlación respecto a las variables que contiene la palabra “pase”, estas variables representan una relación lineal notoria, mientras que otras variables no podemos afirmar su relación, en el eje diagonal podemos encontrar la distribución de ambas variables, visualmente no podemos confirmar si presenta una distribución normal, sin embargo para futuros modelos las variables serán estandarizadas o normalizadas para un mejor tratamiento o interpretación, sin embargo la distribución de las variables independientes solo nos dice la probabilidad de que dicho evento suceda y como se espera que varíen los resultados.

Por otro lado la variable dependiente a predecir, es decir el resultado del partido al ser una variables categórica valorizada podemos ver su distribución en la Imagen Nº10 donde la mayor probabilidad de ocurrencia está en la victoria seguida por las derrotas y en última instancia el empate.

Imagen Nº9: Matriz de gráficos de dispersión de 9 variables independientes y el resultado del partido.

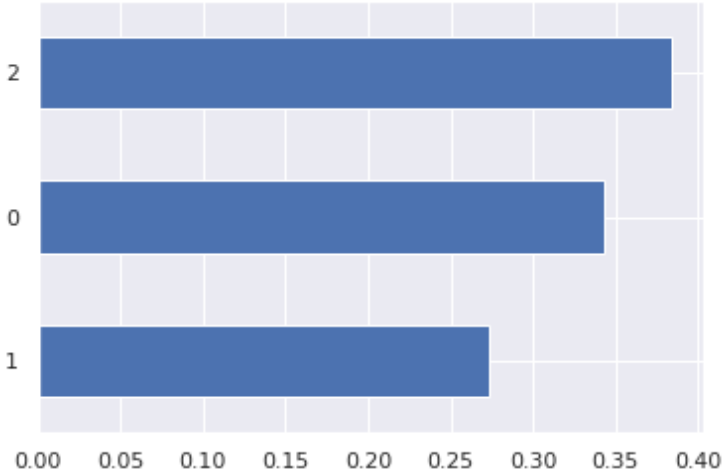


Fuente: Elaboración propia. 0 = Derrota, 1 = Empate y 2 = Victoria

Con la Imagen Nº10 cerramos el análisis exploratorio de datos, el cual nos permitió verificar el comportamiento de las variables independientes con la variable objetivo. Con este análisis podemos tener especial observación en algunas variables que ocuparemos en el siguiente capitulo donde usaremos modelos de predicción.

Correlación, Dispersión y Boxplot confirmaron que el conjunto de datos se encuentra limpios, se confirma la presencia de relación entre variables en conjunto con sus valores atípicos y que las variables están presentes en distintas magnitudes, esto confirma la importancia de estandarizar o normalizar variables para cualquier predicción.

Imagen Nº10: Distribución de la variables dependiente valorizada para el modelo de predicción.



Fuente: Elaboración propia. 0 = Derrota, 1 = Empate y 2 = Victoria

6.2. Modelos de Machine Learning

Preprocesamiento de variables independientes

En los capítulos anteriores se expuso las variables independientes y la variable independiente, más las distintas metodologías que se ocuparan para predecir si un equipo puede perder, empatar o ganar. Antes de iniciar las predicciones es importante destacar que las variables independientes pueden ser estandarizadas o normalizadas, conceptos

muy distintos en estadística a la hora de predecir, por un lado en este estudio normalizamos (MinMaxScaler) según el valor máximo y mínimo de cada variable y para otras predicciones estandarizamos (StandScaler) las variables según su media y desviación estándar, al realizar estas transformaciones eliminamos el efecto que produce los valores atípicos y las varianzas muy altas que hay entre las variables, de esta forma predecir un valor más limpio y preciso.

Implementación de Modelos de Machine Learning

Se realizaron 5 modelos de Machine Learning para predecir la derrota, empate o victoria de un partido de fútbol en el Campeonato Nacional Chileno, ocupando 54 variables independientes y una variable objetivo de tipo categórica, para todas las pruebas se ocupó un 30% de la muestra para pruebas y verificación del nivel de predicción. El primero conjunto de predicciones se realizó con las variables independientes en bruto, es decir sin ningún tipo de transformación, arrojando los resultados en primera columna de la tabla Nº7.

Previamente y gracias al análisis exploratorio de datos, sabemos que el problema de las magnitudes de las variables genera ciertos errores para la predicción y clasificación, es por ello que nos encontramos con niveles de accuracy iguales a 1, es decir una predicción exacta, esto solo presenta que algunos modelos tienen overfitting en la predicción, es decir se sobre ajusto a las observaciones, respaldando que no es una buena idea predecir y clasificar sin antes estandarizar o normalizar las variables independientes.

La segunda columna de la tabla Nº7 posee la predicción usando estandarizadas a través de los mínimos y máximos de cada variable, generando una columna donde los valores fluctúan entre 0 y 1, para eliminar el efecto de los valores atípicos y el efecto que genera las variables que están representadas como % o decimal, de esta forma todas las variables quedan representadas en la misma unidad de medida .

Tabla Nº7: Resumen modelos de predicción y clasificación normalizados entre 0 y 1 de Machine Learning

Modelo	Accuracy (variables en bruto)	Accuracy (variables normalizadas)
Support Vector Machine	1,0000	0,7020
Naive Bayes	0,6884	0,6601
K-Nearest Neighbor	0,4175	0,3288
Regresión Logística	0,6121	0,9692
Random Forest	0,9495	0,9704

Fuente: Elaboración propia con datos de Wyscout.

El efecto que provoca transformar las variables es notorio y a la vez necesario, los métodos que mejoraron en su comportamiento son Regresión Logística y Random Forest. No obstante resulta curioso el cambio que provoco en la precisión de los modelos, pues compilar 54 variables para predecir y clasificar el resultado de un partido de fútbol considera incluso aquellas variables en secciones anteriores estaban altamente correlacionadas o poseían una varianza bastante alta, además considerar que posiblemente estos modelos pueden estar sobre ajustándose a la predicción, para ello la siguiente sección

se enfoca en identificar cuantas y que variables pueden estar aportando mucho o poco para la predicción, este análisis se conoce como Análisis de Componentes Principales, y será evaluado con la metodología de Regresión Logística debido a su nivel de accuracy y por ser considerado uno de los algoritmos más simples y más utilizados para la clasificación multiclase.

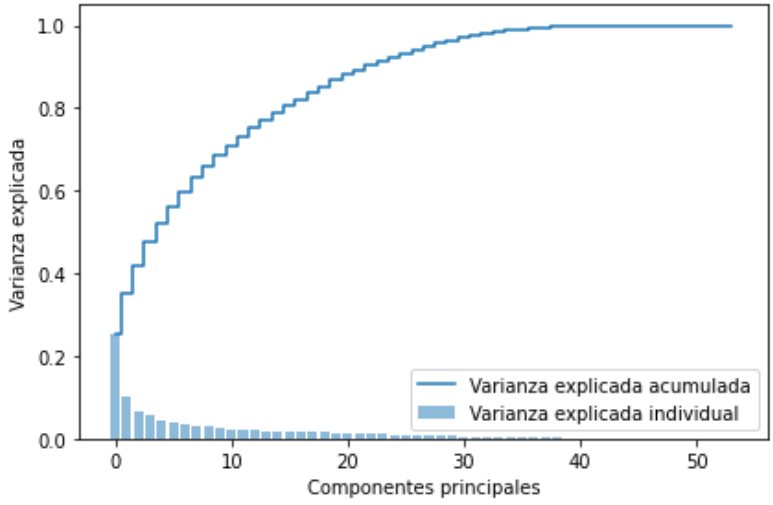
Análisis de componentes principales (PCA)

Anteriormente cada modelo fue elaborado con una dimensión de 54 variables, los mapas de correlación y dispersión indicaban que habían variables que estaba altamente correlacionada y que podían afectar en la predicción de los modelos. Con PCA podemos reducir las dimensiones de las variables independientes, pero el objetivo es reducir la dimensión sin perder información relevante para las predicciones, para ello PCA recurre a la proporción de varianza explicada, de cada una de las variables generando para este caso 54 componentes principales compuestos por la varianza explicada de cada componente, de esta forma genera nuevas variables independientes conocidas como Eigenvector o componentes principales que son una matriz de varianza-covarianza, la suma de cada componente define la cantidad de información que puede aportar al modelo de predicción, algunos documentos aconsejan considerar el 80% del total de la varianza explicada para no perder información de predicción, podríamos decir que nuestra nueva regresión logística quedaría definida de la siguiente forma, donde predecimos el componente principales a través de los coeficientes que ahora son las varianzas de cada variable.

$$\boldsymbol{Componente\ 1} = 0{,}012\ \boldsymbol{Goles} + 0{,}084\ \boldsymbol{xG} + \cdots + (-0{,}079)\boldsymbol{Longitud\ media\ pases}$$

En la imagen Nº11 se muestra la varianza acumulada para cada una de los componentes, esto quiere decir que el primer componente posee un 25% de la varianza observada en los datos, seguida por el componente Nº2 que posee un 10%, al final de grafico podemos ver que existen 14 componentes principales aproximadamente que no superan el 2% de varianza explicada, es decir pueden otorgar muy poca información a la predicción.

Imagen Nº11: Grafico de componentes principales y varianza explicada



Fuente: Elaboración propia con datos de Wyscout.

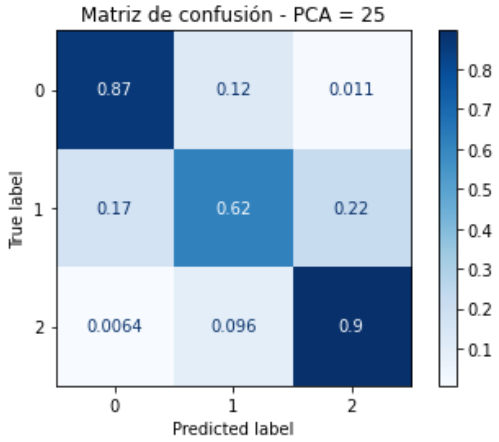
Para respaldar alguna decisión respecto a seleccionar variables importantes para el modelo o eliminar variables que otorgan poca información es importante tener un rango de decisión sobre cuantos componentes principales considerar para la predicción, es importante destacar que con PCA no tenemos las variables originales independientes, si no que tenemos componentes principales. La Imagen Nº11 presenta el porcentaje

acumulado de la varianza explicada a través de los componentes principales donde el 80% y más de la varianza explicada se encuentra seleccionando entre 20 a 25 componentes.

Para respaldar una posible eliminación de componentes y comprobar el nivel de predicción de los componentes seleccionados se procede a realizar un nuevo modelos de Regresión Logística considerando solo 25 componentes principales, la imagen Nº13 representa la matriz de confusión el cual mide el nivel de aciertos para la variable dependiente, es decir el resultado de un partido de fútbol donde las victorias tiene un acierto del 90%. Originalmente considerando las 54 variables estandarizadas el accuracy en la Regresión Logística indicaba un 0,97. Usando componentes principales y reduciendo a más de la mitad los componentes principales obtenemos un accuracy de 0,81.

En términos absolutos descartar 29 componentes principales, perdimos 0,16 de accuracy, esto demuestra que en el contexto original para predecir una victoria, empate o derrota, nos encontramos con variables independientes con alto grado de multicolinealidad, es decir variables que están entregando la misma información a nuestra predicción o que son variables que expresan técnicamente lo mismo, por ello podemos seleccionar algunas variables para eliminarlas, para seleccionar esas variables recurrimos al análisis de correlación y al test Anova.

Imagen Nº13: Matriz de confusión considerando 25 componentes principales



Fuente: Elaboración propia, modelo Regresión Logística normalizada – PCA, Accuracy = 0,81

La pregunta que debemos hacernos es ¿Cuál variable eliminar? y ¿Cómo respaldar esa decisión?, hace unos capítulos atrás mencionábamos que existían variables altamente correlacionadas y mediante PCA descubrimos que aquellas variables con varianza alta podían estar generando multicolinealidad con la variable independiente. A través del mapa de correlación y el test F de Anova se crea un listado de variables que se resumen en la tabla Nº8 una lista de variables con mayor valor de correlación y varianza.

Tabla Nº8: Listado de variables correlacionadas y con varianza alta ordenados de mayor a menor

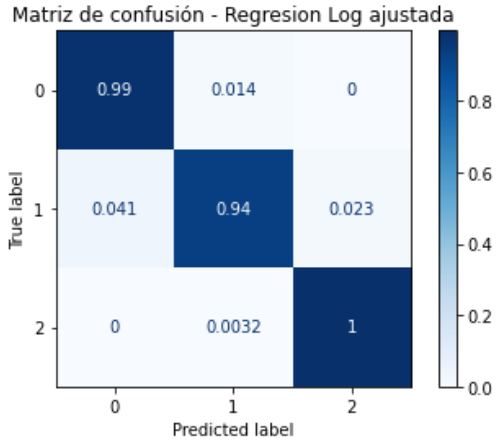
Índice de correlación	Score test F – Anova
Pases en el último tercio logrados	Tiros a la portería%
Toques en el área de penalti	Tiros libres
Duelos aéreos	Pases en el último tercio
Pases laterales	Centros

xG	Interceptaciones
Centros precisos	Penaltis marcados%

Fuente: Elaboración propia con datos de Wyscout.

Al modelar una regresión logística eliminando las variables de la tabla Nº8, es decir predecir la clasificación de victoria, empate o derrota con solo 44 variables arroja un accuracy de 0,9766, esta eliminación de variables aumento su precisión, esto confirma que las variables eliminadas presentaban poca información o simplemente eran variables que eran que representaban lo mismo en el modelo, con esto confirmamos que el conjunto de datos – Partidos posee variables que no representan un aporte en la predicción del resultado del partido, según PCA la cantidad de variables optimas giraban en torno a las 20 a 25 componentes principales, con este análisis ya tenemos un indicio robusto de que podemos seguir eliminado variables y perder nada o muy poca precisión en el modelo de Regresión Logística.

Imagen Nº14: Matriz de confusión Regresión Logística ajustada a 44 variables.



Fuente elaboración propia, Accuracy = 0,9766

Extrapolación de coeficientes para ranking de Jugadores.

Los 2 principales objetivos de este estudio eran crear un modelo de predicción y clasificación de los resultados en los partidos de fútbol en el campeonato chileno y crear una ranking de jugadores según su posición de desempeño. Como vimos en la sección anterior ocupamos 54 variables estandarizadas para predecir el resultado, como esas variables representaban lo ocurrido en un partido de fútbol, podemos realizar el supuesto de que esas mismas 54 variables pueden ser presentadas a un nivel especifico como el jugador, teóricamente la cantidad de pases, goles, asistencias, duelos, tarjetas son creadas y medidas en base al jugador, entonces este supuesto permite realizar la relación entre las variables del partido y las variables de los jugadores, tomaremos todas las variables sin considerar lo generado por PCA, ya que al llevar este supuesto a nivel más especifico como lo es el jugador cada variable de medición es importante para clasificar al jugador y considera una mayor cantidad de atributos del jugador necesarias para cada posición, por ejemplo un jugador que juega como lateral debe tener pases laterales.

Cuando se habla de estadísticas de fútbol, la percepción de las personas se acota a la cantidad de goles, pases, regates, minutos jugados, en el conjunto de datos sobre jugadores podemos encontrar aspectos parecidos, la tabla Nº9 resume algunos indicadores de los jugadores.

Tabla Nº9: Indicadores de desempeño general de jugadores	
Jugador con mayor cantidad de asistencias	B. Yañez (11)
Jugador con mayor cantidad de T. Rojas	M. Fernandez (5)
Jugador con mayor cantidad de T. Amarillas	R. Gonzalez (25)

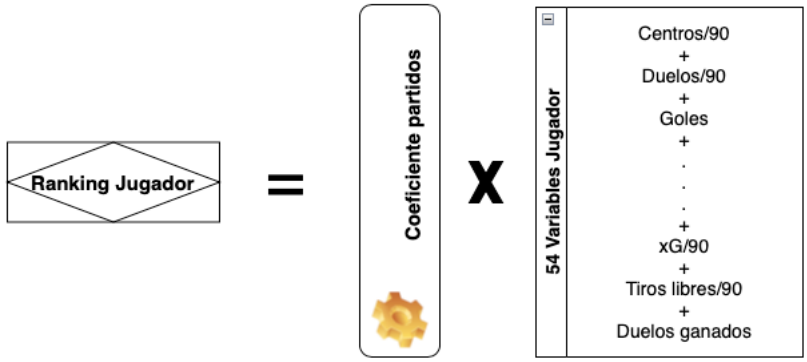
Jugador con mayor cantidad remates	B. Brereton Díaz (130)
Jugador con mayor cantidad de minutos jugados	R. Gonzalez (10.017)
Jugador con mayor cantidad de partidos jugados	R. Gonzalez (141)
Fuente: Elaboración propia con datos de Wyscout	

Para crear el ranking se utilizaran los coeficientes generados por el modelo de Machine Learning – Regresión Logística que considera las 54 variables estandarizadas, cada uno de esos coeficientes se multiplicara por el valor de cada variable de los jugadores, como el entrenamiento del modelo se hizo con las variables independientes estandarizadas entre 0 y 1, entonces las variables de los jugadores también serán estandarizadas.

Además para realizar esta clasificación se asumirá que el jugador solo posee una posición en el campo y será tratada como una posición genérica, de esta forma al multiplicar el coeficiente de predicción de resultado del partido por las variables del jugador se obtendrá una valor que mientras más alto mejor desempeño tendrá en su posición.

La imagen Nº15 representa el cálculo que une los coeficientes del partido y las variables de los jugadores, con esto cada jugador será valorizado y agrupado según la posición única en la que juega.

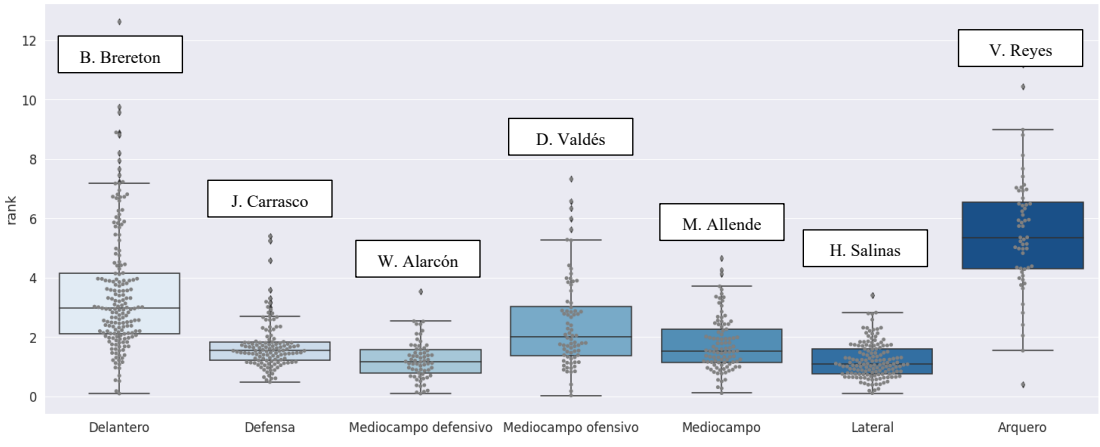
Imagen Nº15: Proceso de ranking jugadores



Fuente: Elaboración propia

El compilar los coeficientes con las variables de los jugadores se genera un ranking de jugadores donde a medida que aumenta su valor, la probabilidad de que el equipo gane aumenta, a través de la Imagen Nº16 se puede ver la distribución de los Jugadores chilenos donde los valores atípicos toman una especial importancia, ya que son los jugadores que con sus atributos atraen mayores posibilidad de clasificar un partido como una victoria.

Imagen Nº 16 : Distribución de ranking de jugadores por posición



Fuente: Elaboración propia

Los equipos están conformados por alrededor de 25 jugadores entre titulares y reservas, y así como en casas de apuestas y videojuegos siempre es importante escoger una formación que garantice una victoria, en la tabla Nº9 se presenta un top 5 de jugadores según si posición de esta forma tratar de acercarse aún más a la percepción de la realidad respecto a jugadores que acercan el resultado a una victoria.

Tabla Nº9: Top 5 de jugadores según su puntuación

Delantero	Ranking	Arquero	Ranking	Defensa	Ranking	Lateral	Ranking
B. Brereton Díaz	12,6	V. Reyes	11,2	J. Carrasco	5,4	H. Salinas	3,4
L. Benegas	9,7	R. Olivares	10,4	N. Vargas	5,2	Y. Gonzalez	2,8
J. Grondona	9,6	G. Collao	9,0	R. Echeverría	4,6	B. Cerezo	2,8
F. Flores	8,9	M. Parra	8,8	B. Gazzolo	3,6	J. Méndez	2,8
J. Castro	8,8	J. González	8,1	C. Meneses	3,3	B. Nieto	2,6

Mediocampo	Ranking	Mediocampo defensivo	Ranking	Mediocampo ofensivo	Ranking
M. Allende	4,6	W. Alarcón	3,5	D. Valdés	7,3
B. Cortés	4,4	J. Fuentes	2,5	J. Abrigo	6,6
M. Dávila	4,3	G. Navarrete	2,5	C. Villanueva	6,5
M. Núñez	4,1	J. Miño	2,5	N. Castro	6,3
S. Leyton	3,7	A. Camargo	2,4	M. Bolados	6,0

Fuente: Elaboración propia

7. Conclusiones

Este trabajo se enfoca en 2 aspectos, implementar un modelo de predicción de resultado de un partido de fútbol y crear un ranking de jugadores, se ha demostrado que las variables recopiladas mediante la plataforma Wyscout representan un aproximación a determinantes para predecir el resultado de un partido, donde es importante considerar las

dimensiones y magnitudes de las variables, el uso de distintas técnicas de Machine Learning permitió experimentar y usar el mejor modelo, esto permiten diversificar las opciones para predecir y clasificar, obteniendo un resultado mas agudo y preciso, para este caso las técnicas de Regresión Logística y Random Forest.

Respecto a las variables independientes podemos afirmar que existe una amplia variedad de medir los eventos que ocurren en un partido de fútbol y que solo algunos otorgan una cuota importante de información para modelar, esto afirma que las variables usadas para predecir un resultado en cualquier modelo genera multicolinealidad, por ello se debe tener especial cuidado y realizar una selección de variables bajo criterios robustos que permitan medir el real efecto de cada variable independiente.

La relación entre la cantidad y tipo de datos que genera en un partido de fútbol está estrechamente relacionada con el jugador, pues es el jugador el que genera el dato con sus habilidades, permitiendo extrapolar y homologar variables dentro del contexto del fútbol. Bajo este supuesto, generar un ranking de jugadores usando los coeficientes de predicción de resultado de partidos genero un resultado final que se cerca a la realidad, y que a más de algún lector asemejara la realidad futbolística con este ranking.

Respecto a las limitaciones que posee este estudio, se enmarca en las variables, la multicolinealidad que hay entre las variables en muchas ocasiones lleva a sobre estimar los resultados, pues podemos tener variables que expresen lo mismo en el modelo, es por ello que una alternativa puede ser generar ratios de desempeño y que pueda unificar variables que contengan la información necesaria para una predicción, otra limitante es que por el momento carece de un grupo de comparación, sería interesante usar la misma

metodología con otra liga de Latinoamérica y comparar los ranking de jugadores que se generan en las otras ligas y comparar con los jugadores chilenos.

8. Bibliografía

Herberger, T., & Litke, C. (2021). The Impact of Big Data and Sports Analytics on Professional Football: A Systematic Literature Review. *Springer Proceedings in Business and Economics*, 147 - 171.

Antón Carranza, M. (2017). *Identificación del talento en la Organización: El Big Data aplicado al fútbol.* Valladolid: Universidad de Valladolid. Facultad de Ciencias .

González Ramos, P., Martín Agüero, O., Montero Quesada, J., & Rice Nelson, D. (2021). Guía para el control observacional de la táctica grupal ofensiva en el fútbol. *Revista de Ciencia y Tecnología en la Cultura Física*, 162 - 176.

Albarrán Jardón, E. (2020). *Inteligencia deportiva: tecnología aplicada al deporte. Pensamiento libre.* Estado de Meximo: Pensamiento Libre.

Dufour, M., Phillips, J., & Ernwein, V. (2017). What makes the difference? Analysis of the 2014 World. *Journal of Human Sport and Exercise*, 616-629.

Gudmundsson, J., & Horton, M. (2018). Spatio-Temporal Analysis of Team Sports . *ACM Computing Surveys (CSUR)*, 1-34.

Brooks, J., Kerr, M., & Guttag, J. (2016). Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 49-55.

Cintia, P., Pappalardo, L., Pedreschi, D., & Giannotti, F. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. *Conference: 2015 IEEE International Conference on Data Science and Advanced Analytics* .

Barreto, L., Arruda, F., & Cunha, S. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences*, 1881-1887.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2018). PlayeRank: Multi-dimensional and role-aware rating of soccer player performance. *ArXiv*.

Mueller, J., & Massaron, L. (2016). *Machine learning for dummies.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Zabarte, G. (2022). *Utilización de técnicas de aprendizaje automático para la predicción del rendimiento de los jugadores de fútbol.* Madrid.

Noble, W. (2006). What is a support vector machine? *Nature Biotechnology*, 1565-1567.

Maron, M. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 404-417.

Abdul , M., Mustapha, A., & Fauzi, R. (2018). Bayesian Approach to Classification of Football Match Outcome. *International Journal of Integrated Engineering*, 155-158.

Bazmara, M., & Jafari, S. (2013). K Nearest Neighbor Algorithm for Finding Soccer Talent.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 123-140.

Orellana, J. (2018). *Bookdown*. Obtenido de Arboles de decision y Random Forest: https://bookdown.org

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 37-52.

Amat, J. (2020). *Ciencia de datos* . Obtenido de PCA con PYTHON: www.cienciadedatos.net