Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

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

TRABAJO FINAL DE ESPECIALIZACIÓN

Soporte algorítmico en la conformación de planteles digitales para competencias oficiales de FIFA EA-SPORTS®.

Técnicas de Machine Learning como propuesta a deportistas electrónicos para predecir el rating de jugadores

AUTOR: MATIAS ANDRES OLEKSIKIW
MENTOR: NÉLIDA MÓNICA CANTONI RABOLINI

DICIEMBRE 2022

Resumen

El videojuego FIFA EA-Sports ® es un simulador virtual de partidos de fútbol, en el cual el usuario puede seleccionar a su equipo de preferencia disponiendo de sus nombres, escudos, estadios y jugadores oficiales fruto de las licencias adquiridas a lo largo del tiempo (Electronic Arts, 2021). Desde sus inicios hasta el presente, el simulador es considerado como una insignia de la industria de los videojuegos y principal impulsor de las competencias de deportes electrónicos ("E-Sports"). El marcado interés por parte de sus consumidores y la creciente popularidad en las competencias virtuales, conocidas como FIFA GLOBAL SERIES – ULTIMATE TEAM ®, (Electronic Arts, EA Games, 2021) han dinamizado el desarrollo de un rentable negocio para gestar un nuevo oficio pago: el deportista electrónico. Como consecuencia de la elevada exigencia y dificultad de los torneos digitales, se dio origen a la conformación de equipos profesionales para potenciar el entrenamiento y resultados esperados. Entre los más exitosos se destacan "DUX", "DIRE WOLVES", "FALCONS", "FOKUS CLAN" y "NEO". Contar con un amplio conocimiento acerca de la mecánica con la que opera el simulador para calificar / valorar a cada componente virtual (jugador - plantel), resulta un factor clave en la estrategia de cada equipo profesional a fin de maximizar su rendimiento. En tal sentido, el objetivo general del presente trabajo final de especialización es analizar la variación de las variables en todas las versiones del videojuego y cómo afectan al rating global de cada jugador para predecir el modo en que dicha métrica se comportará en el siguiente periodo.

Palabras clave: Planteles Digitales – Ultimate Team – Competencias FIFA EA-SPORTS® - Análisis predictivo

ÍNDICE

- **Introducción**

Actualmente los equipos profesionales de deportes electrónicos y usuarios casuales se encuentran inmersos en un entorno digital que propone en forma constante competición entre todos sus partícipes. El sostenido éxito de la franquicia FIFA EA-Sports ® se acopló al proceso de expansión de la industria de los videojuegos, para así también ofrecer su propia competencia digital denominada FIFA GLOBAL SERIES – ULTIMATE TEAM ®. La competición en cuestión ofrece rentables premios nominados en moneda real, constituyendo a tal efecto un considerable estímulo para los profesionales en deportes electrónicos.

Electronic Arts (empresa desarrolladora del videojuego) hasta el día de la fecha no ha brindado detalle alguno de la mecánica con la que opera el simulador para calificar / valorar a cada componente virtual (jugador - plantel), aunque sí publica oficialmente la plantilla consolidada de todos los jugadores indicando las diversas habilidades / cualidades que lo componen. Las habilidades / cualidades, son variables numéricas que presentan un recorrido del 0 al 100 y se encuentran presentes en todos los jugadores. Las mismas abarcan múltiples aspectos como: Habilidad con el balón, Agresión, Velocidad, entre otras. El conjunto de datos utilizado abarca las 106 variables y se extiende desde la versión FIFA 15 hasta FIFA 22, totalizando un total de 231.468 individuos. La completitud, precisión, validez y consistencia de los datos disponibles los vuelven aptos para la aplicación de métodos de análisis multivariado y predictivos.

El objetivo general es analizar la variación de las variables en todas las versiones del videojuego y cómo afectan al rating global de cada jugador para predecir el modo en que dicha métrica se comportará en el siguiente periodo, con la finalidad de brindar herramientas útiles para la conformación de planteles competitivos para profesionales de deportes electrónicos. El objetivo se alcanzará en una primera etapa, se aplicarán métodos de análisis multivariante para reducir la dimensionalidad de los datos y por ende facilitar la comprensión de las variables y su influencia en el rating global de cada jugador. En una segunda etapa, se aplicarán modelos predictivos de regresión ajustados por los resultados obtenidos en la etapa precedente, con la finalidad de obtener el mayor grado de precisión en la predicción.

1 El auge de la industria de los videojuegos; Descripción de las competencias oficiales de FIFA Global Series; estado actual del conocimiento en equipos profesionales de deportes electrónicos y usuarios casuales acerca del cálculo del rating.

El apartado brinda una introducción a la industria de los videojuegos y una explicación detallada de las competencias oficiales de FIFA Global Series (EA-SPORTS ®). Asimismo, exhibe el estado actual del conocimiento en equipos profesionales y usuarios casuales sobre el cálculo del rating en el videojuego FIFA EA-SPORTS ®. Por su parte, se conceptualiza al rating individual de cada jugador con la finalidad de facilitar la compresión de las variables que lo componen y sus variaciones a lo largo de cada versión del videojuego.

1.1 Auge de la industria de los videojuegos, desarrollo de competencias, profesionalización de los usuarios.

De acuerdo con lo comentado por Global Games Market (Newzoo, 2021), se estima que la industria de los videojuegos generará un total de 175.800 millones de dólares en el 2021 luego de haber registrado un incremento interanual del 22% respecto del 2020. Sus principales ingresos están concentrados en 50% en Asia, 24% en Norteamérica, 18% en Europa, y un 8% entre América Latina y África. Las proyecciones de la industria estiman un crecimiento acumulado del 8.7% acumulado al 2024. Hace más de 10 años, las ventas de software empaquetado para consolas domésticas representaban el 64% del mercado mundial de juegos. Desde entonces, dicha proporción ha caído al 30%. Los deportes electrónicos actualmente plantean un cambio de paradigma que les permiten a las empresas desarrolladoras generar ingresos similares a los de las empresas de medios.

En tal sentido, la industria ha experimentado un gran aumento en la inversión por parte de entidades privadas. La cantidad de inversiones en E-SPORTS se duplicó en 2018, pasando de USD 34 millones en 2017 a USD 68 millones en 2018. A medida que los videojuegos de alto nivel competitivo continúan integrándose en la cultura popular, los inversores globales, las marcas y los medios de comunicación comenzaron a prestar mayor atención, como así también sus espectadores. Se estima que la cantidad total de espectadores ascenderá a 26.6 millones en 2021, un 11.4% más que en 2021 (Insider Intelligence, 2021).

Los aficionados promedian alrededor de 100 minutos por transmisión en vivo de eventos digitales. Resulta pertinente remarcar que solo el 50% de los espectadores realmente juega por su cuenta al videojuego involucrado, aspecto que demuestra que la afición por los videojuegos trasciende la propia interacción directa con los mismos.
Por otro lado, el 60% de los fanáticos de los deportes electrónicos están dispuestos a viajar para ver sus juegos, torneos y jugadores favoritos, cuyo comportamiento sólo se vio interrumpido por la pandemia COVID-19 (Total, 2021).

Se indica que el 65% de los espectadores tienen entre 18 y 34 años y, si bien la base está mayormente dada por el género masculino, el 38% son mujeres. Los hombres norteamericanos que componen el grupo mencionado consideran que los deportes electrónicos son tan populares como el béisbol o el hockey sobre hielo. En América del Norte, el deporte más popular de la región, el fútbol, es solo el doble de popular que los deportes *electrónicos.* Para los espectadores masculinos de entre 36 y 50 años, el fútbol es solo el triple de popular (Total, 2021).

El éxito y consecuente auge de los deportes electrónicos no es únicamente atribuible a los espectadores aficionados, sino que existe lógicamente una contraparte que cumple el rol de brindar el espectáculo en cuestión: el deportista electrónico. Estos últimos se los puede clasificar en dos grandes grupos: “Streamers” y “Profesionales” (Total, 2021). Los Streamers son jugadores que se transmiten en vivo a sí mismos mientras juegan videojuegos en modalidad casual (en famosas plataformas como “YouTube” y “Twicht”, entre otras). No todos tienen el nivel apto para adentrarse en competencias profesionales, por lo cual deciden focalizar su esfuerzo en la transmisión, considerando que además brinda ingresos potencialmente elevados: en 2015 el “Youtuber” más famoso conocido como “PewDiePie” registró ingresos aproximados de 8 millones de dólares (Total, 2021).

Los pocos que obtienen el nivel necesario para competir en torneos y competencias oficiales se denominan “Profesionales”, y suelen lanzarse al mercado en forma independiente para luego ser contratados por organizaciones que aprovechan su talento como impulsor de marketing de sus productos. Por ejemplo, FIFA EA-Sports articula con clubes profesionales de fútbol como Manchester United, Barcelona (entre otros), para desarrollar equipos de deportes electrónicos y en tal sentido potenciar sus respectivas marcas. La rentable sinergia generada por las múltiples combinaciones entre organizaciones, desembocan en parte en sueldos para deportistas electrónicos de hasta 100 mil dólares al año (Total, 2021).

Los jugadores profesionales se unen a equipos (modo multijugador) para competir por premios en efectivo. Cada equipo se especializa y compite en un juego específico, representando a la organización de la que forman parte, competirán en la liga respectiva de sus videojuegos, donde hay temporadas regulares, playoffs y campeonatos mundiales (Total, 2021). El marcado interés por parte de sus consumidores y la creciente popularidad en las competencias virtuales, conocidas como FIFA GLOBAL SERIES – ULTIMATE TEAM ®, (Electronic Arts, EA Games, 2021) han dinamizado el proceso de expansión de la industria del videojuego. Como consecuencia de la elevada exigencia y dificultad de los torneos digitales, dio origen a la conformación de equipos profesionales para potenciar el entrenamiento y resultados esperados. Entre los más exitosos se destacan “DUX”, “DIRE WOLVES”, “FALCONS”, “FOKUS CLAN” y “NEO”.

FIFA GLOBAL SERIES 21 es una competencia oficial de celebración anual multirregional (Europa, América, África, Asia, Oceanía) y Multiplataforma (Playstation, Xbox) basado en la modalidad de juego “Ultimate Team” (Electronic Arts, 2021). Dicho modo es el más popular dentro del juego, y basa su mecánica en la creación de plantillas competitivas con cartas digitales coleccionables que incluyen el rating global de los jugadores. Conforme se obtienen puntos se podrá ir accediendo a nuevos artículos que, mediante la aleatoriedad, pueden otorgar grandes recompensas a fin de mejorar la plantilla. El torneo se desarrolla mediante sistema de clasificación, y distribuye un total de premios de 3 millones de dólares a sus jugadores.

El factor económico constituye una fuente significativa de motivación para los competidores y equipos profesionales, que a su vez ha impulsado el desarrollo de variadas estrategias para la formación de una plantilla “Ultimate Team” competitiva en función del rating global de cada jugador. En acotadas entrevistas compartidas en internet (Cadenaser, 2016), Michael Mueller-Moehring (director y responsable de la recolección de datos y arquitectura de bases de Electronic Arts para el producto FIFA) ha brindado detalles generales de la metodología del cálculo de rating. Sin embargo, la consecuente falta de información y precisión sobre este aspecto se justifica en las políticas de confidencialidad practicadas por Electronic Arts.

El interés en este tópico ha evolucionado con el paso del tiempo, para extenderse a páginas web como Futwiz.com que mediante observaciones sobre el desempeño actual del jugador en la vida real intentan predecir el rating futuro. Por su parte, también existen foros conformados por la comunidad de jugadores en los que se debate sobre ello. Los variados impedimentos para acceder a más información, la complejidad de la metodología y la naturaleza de un público objetivo mayormente casual y no tecnificado, configuran un estado del conocimiento actual no desarrollado

1.2 FIFA Ultimate Team, conceptualización del rating y variables que lo componen.

FIFA Ultimate Team (también conocido como FUT) es un modo de juego en FIFA que permite crear y administrar un club / equipo para jugar partidos en línea y fuera de línea y ganar recompensas usando jugadores y gerentes, así como una colección de varios tipos de tarjetas como artículos del club, personal y consumibles (Electronic Arts, EA Games, 2021). Al jugar partidos en el modo FUT, se podrán obtener mejores jugadores y elementos para la calidad y el presupuesto del club / equipo, mediante la acumulación de la moneda virtual "Coin". Dicho dinero constituye un elemento clave para todo tipo de usuario, considerando que su escasez implica sabia administración para la compra de jugadores competitivos frente a un presupuesto determinado (Electronic Arts, EA Games, 2021).

La importancia de contar con jugadores competitivos resulta evidente al analizar la relación directa entre la probabilidad incrementada de ganar un partido y contar con un plantel de valoración significativa. Tal y como se mencionó con anterioridad en el presente trabajo, el público objetivo del videojuego FIFA desconoce la metodología de cálculo del rating a raíz de las políticas de privacidad de la empresa desarrolladora Electronic Arts®.
Este aspecto propone el desafío de que los usuarios analicen las cualidades (variables) que lo componen y en efecto, comprar virtualmente los jugadores más aptos para su plantel. A su vez, se establece una complejidad adicional basada en el rearmado del plantel al inicio de cada temporada (duración: 1 año), motivo que impulsa el interés de predecir el futuro rating de cada jugador. El rating en cuestión contempla un total de 42 cualidades (variables cuantitativas - numéricas), que individualmente califican con un recorrido de 0 a 100 (más alto mejor) las aptitudes de cada jugador.
Las mismas se detallan a continuación:

- Attacking Crossing, Attacking Finishing, Attacking Heading Accuracy, Attacking Short Passing, Attacking Volleys, Defending, Defending Marking Awareness, Defending Sliding Tackle, Defending Standing Tackle, Dribbling, Goalkeeping Diving, Goalkeeping Handling, Goalkeeping Kicking, Goalkeeping Positioning, Goalkeeping Reflexes, Goalkeeping Speed, Mentality Aggression, Mentality Composure, Mentality Interceptions, Mentality Penalties, Mentality Positioning, Mentality Vision, Movement Acceleration, Movement Agility, Movement Balance, Movement Reactions, Movement Sprint Speed, Pace, Passing, Physic, Potential, Power Jumping, Power Long Shots, Power Shot Power, Power Stamina, Power Strength, Shooting, Skill Ball Control, Skill Curve, Skill Dribbling, Skill Fk Accuracy, Skill Long Passing (Anexo I)

Asimismo, existen variables de carácter cualitativo que describen y categorizan a cada individuo:

- Age, Club Position, Club Team, Id, Height Cm, International Reputation, League Level, League Name, Nationality, Id, Preferred Foot, Release Clause Eur, Short Name, Skill Moves, Id, Value Eur, Wage Eur, Weak Foot, Weight Kg (Anexo II)

1.3 Métodos de análisis multivariante aptos para facilitar la compresión de las variables, presentación de la base de datos a utilizar y objetivos de la calidad de datos.

Enfrentarse a un número elevado de variables, profundiza la dificultad para comprender la incidencia de las aptitudes en la conformación del rating global. Por dicho motivo, como propuesta inicial a equipos profesionales de deportes electrónicos, se presentan métodos de análisis multivariante aptos para enfrentar la problemática planteada. Los métodos de análisis multivariado de Análisis de Componentes Principales ("PCA") y Análisis de Clústeres permiten por un lado realizar la reducción de la dimensionalidad, y en función de ello clasificar como grupos a las observaciones de una muestra. Los datos provienen de una base pública en referencia al videojuego FIFA en su versión seleccionando al azar el año 2017. La estructura de esta se basa en las 42 variables anteriormente mencionadas, y a una muestra de 66 observaciones / jugadores (plantel completo de FC Barcelona y Real Madrid FC que representan los equipos más frecuentemente elegidos por los usuarios).

Los datos son de tipo estructurados, dado que cuenta con celdas de formato numérico, texto, categórico, entre otros. Se observa que la base de datos no es muy extensa, cuenta con 17.597 registros únicos (no duplicados) que representan el ID de cada jugador. En cuanto a los objetivos de calidad de datos, se destaca que, respecto a la Completitud, en lo que respecta a ciertos atributos categóricos no se identificaron valores perdidos.

Se considera que la base de datos cumple con los parámetros del objetivo de Unicidad, ya que no se detectaron datos duplicados en las ID. Respecto a la oportunidad, se desconoce la metodología de armado de la base, pero por tratarse de una fuente oficial, es menos probable que haya diferencias temporales entre imputación y registro.
Se considera cumplido el objetivo de validez, dado que no se detectaron valores inusuales dentro de cada atributo, es decir, la sintaxis de los datos es acorde a la definición y tipo de datos que comprende. No se pudo realizar una evaluación respecto a los objetivos de Precisión y Consistencia, dada la especificidad de estos, donde el primero está vinculado a la comparación de valores/datos respecto a un conjunto de datos validado, y en el caso del segundo éste requiere el cruce de datos que representan un mismo aspecto de un tema.

2 Método de componentes principales y clústeres como propuesta para facilitar la comprensión de las variables que mejor definen a la composición del rating.

El apartado en referencia tiene por objeto principal la aplicación de análisis multivariante por método de componentes principales con la finalidad de asignar un peso apropiado a cada variable en el modelo predictivo a aplicar. El método de componentes principales se vale de una muestra tomada de la base de datos seleccionada para llevar adelante el presente trabajo. Como complemento analítico se procede a aplicar análisis de clústeres para clasificar como grupos a las observaciones de la muestra.

2.1 Marco conceptual de los métodos de análisis multivariante seleccionados.

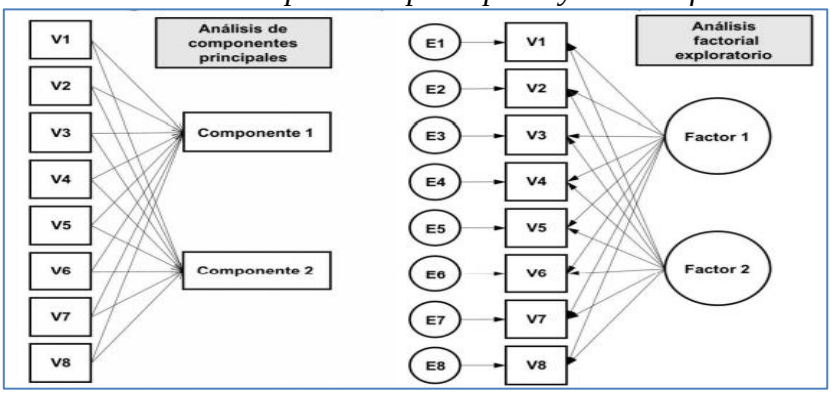
Componentes principales:

Se trata de un método de análisis multivariado que se enfoca en la interdependencia que surge entre variables métricas. Es un método útil para reducir la dimensionalidad de múltiples variables en componentes principales, evitando la pérdida significativa de información (Catena A, Ramos M Y Trujillo H., 2003). Las variables con mayor varianza aportaran una mayor contribución en la generación de un componente principal. Estos últimos se caracterizan por no estar inter correlacionados (es decir, son una combinación lineal de las originales) y presentarse en forma descendente, es decir el primero arrojará la mayor variabilidad y así consecutivamente con los restantes (Catena A, Ramos M Y Trujillo H., 2003).

Las cargas factoriales resultantes de la aplicación del método, indican cuánta influencia ha generado sobre la determinación del componente principal. Si bien el cálculo y posterior lectura de los resultados resultan auto explicativos, la presentación éstos en un gráfico de doble eje permitirá visualizar directamente el grado de asociación de las variables (dada su proximidad entre ellas) e inclusive combinarlas plasmando también los individuos (Catena A, Ramos M Y Trujillo H., 2003). Para diferenciar al método de componentes principales del análisis factorial exploratorio ("EFA"), se puede decir que al primero no le da importancia a la estructura latente de las variables, es decir, si hay o no factores que están provocando que esas variables estén correlacionadas entre sí. Para el PCA las variables son en sí mismas el objeto de interés, y no su estructura subyacente. En la figura 1 se realiza la comparación:

Figura 1

Comparación entre análisis de componentes principales y análisis factorial exploratorio



Nota. Imagen tomada de CATENA A, RAMOS M y TRUJILLO H. (2003). Análisis multivariado. Un Manual para Investigadores. Editorial Biblioteca Nueva.

Como puede observarse, el método PCA asocia directamente las variables a un determinado componente desestimando la causa que define su correlación, mientras que el EFA asigna factores que efectivamente sí lo explican. No obstante, se debe considerar que dichos factores no son explicativos en su totalidad, por lo cual se incorpora al análisis el error (parte de comportamiento no demostrado). Adicionalmente se debe hacer mención, la dirección de

las flechas plasmadas anteriormente: En el PCA las variables constituyen el componente (mediante su combinación lineal), mientras que en el EFA los factores intentan demostrar la estructura subyacente de la relación entre ellas (Catena A, Ramos M Y Trujillo H., 2003).

Previo a la aplicación de los métodos presentados, la obtención de la matriz de correlaciones permite observar el grado de asociación entre todas las variables seleccionadas, acompañando la misma de métodos de visualización como el "Corrplot" (Catena A, Ramos M Y Trujillo H., 2003). Por su parte, resulta importante valerse de diversos indicadores de bondad para determinar si resulta estadísticamente pertinente llevar a cabo el método de análisis mencionado. *Entre ellos se destacan los siguientes:*

Test de esfericidad de Bartlett: donde si el (p-valor) < 0.05 se rechaza H0 (hipótesis nula; la matriz de correlaciones es una matriz de identidad) - se puede aplicar el análisis. De lo contrario se acepta la H0 y no se puede aplicar el análisis (Catena A, Ramos M Y Trujillo H., 2003). Rechazar la hipótesis nula significa que los resultados obtenidos en la prueba estadística son significativos y que es poco probable que se deban al azar. En otras palabras, se concluye que existe evidencia suficiente para respaldar la hipótesis alternativa, que es la afirmación opuesta a la hipótesis nula. Por el contrario, no se puede rechazar la hipótesis nula cuando los resultados de la prueba no son estadísticamente significativos, lo que indica que no hay suficiente evidencia para respaldar la hipótesis alternativa y que los resultados podrían haber ocurrido por casualidad.

KMO (Kaiser, Meyer y Olkin): Es el cociente que se obtiene combinando las correlaciones de las variables originales y las de los factores específicos. Donde, sí dicho valor es próximo a cero, el modelo no será adecuado, mientras que si el mismo se encuentra entre 0.5 y 1, más adecuado lo será (Catena A, Ramos M Y Trujillo H., 2003).

Se busca el "factor principal" siendo éste el que explique la mayor cantidad de la varianza en la matriz de correlación R. En el caso de que las variables estén tipificadas / estandarizadas, la proporción de la variabilidad total de las variables originales captada por un componente principal es igual al autovalor correspondiente dividido por el número de variables originales. Para determinar el número de factores a extraer, se utilizan distintos criterios de elección, estos son: El Criterio del autovalor superior a la unidad, el Criterio del gráfico de sedimentación (Scree plot), y el Parallel Analysis ("método de Horn"), entre otros. (Catena A, Ramos M Y Trujillo H., 2003).

<u>Análisis de Clústeres:</u>

Se trata de una técnica utilizada para clasificar distintos individuos / casos / observaciones en grupos (clústeres). Éstos se caracterizan por su homogeneidad, es decir que las variables que lo conforman presentan similitudes entre sí. A su vez, los grupos deberán presentar marcadas diferencias para facilitar su distinción (Catena A, Ramos M Y Trujillo H., 2003). Se parte de un número determinado de observaciones (n) de las cuales se dispone información sobre un número de variables (k). Posteriormente se deben establecer indicadores de similitud (en que se parecen entre sí cada par de observaciones). En función a este último aspecto se podrán crear grupos mediante dos tipos de análisis:

- Jerárquico:
  - Caracterización:
    - "Aglomerativos": cada observación es un grupo en sí mismo, y sucesivamente se van fusionando para formar menor cantidad de grupos, pero de mayor tamaño. En una instancia final todos los subgrupos son anexados a un único clúster.
    - "Des-aglomerativos": Indica el camino inverso al análisis precedente.
  - Métodos:
    - Del centroide, vecino más cercano, vecino más lejano, vinculación promedio, Ward.
    - Determinación de cantidad óptima de grupos: Utilización de indicadores como Índice Pseudo t2, Índice DB, Índice de Dunn, Estadístico de Hubert, Índice Dindex.

- No Jerárquico:
  - Caracterización:
    - A diferencia del Jerárquico, éste desestima el proceso secuencial de formación de grupos y asigna un número determinado de ellos, sobre los cuales se irán incorporando observaciones que compartan características.
  - Métodos:
    - K-Means (más frecuente).

(Catena A, Ramos M Y Trujillo H., 2003)

2.2 Aplicación de los métodos

Consideraciones generales:

Tal y como se mencionó anteriormente, los "Métodos de Análisis Multivariado" seleccionados pertenecen al grupo de técnicas de interdependencia que buscan, por un lado, analizar la relación entre variables de tipo métricas, y por el otro entre casos. Por dicho motivo, el "dataset" (34 variables métricas x 66 individuos) elegido se considera apropiado para la aplicación de éstas. El desarrollo del presente trabajo se valió de las facilidades de cálculo estadístico que brinda el Software "R" (R Core Team, 2020)

Componentes principales:

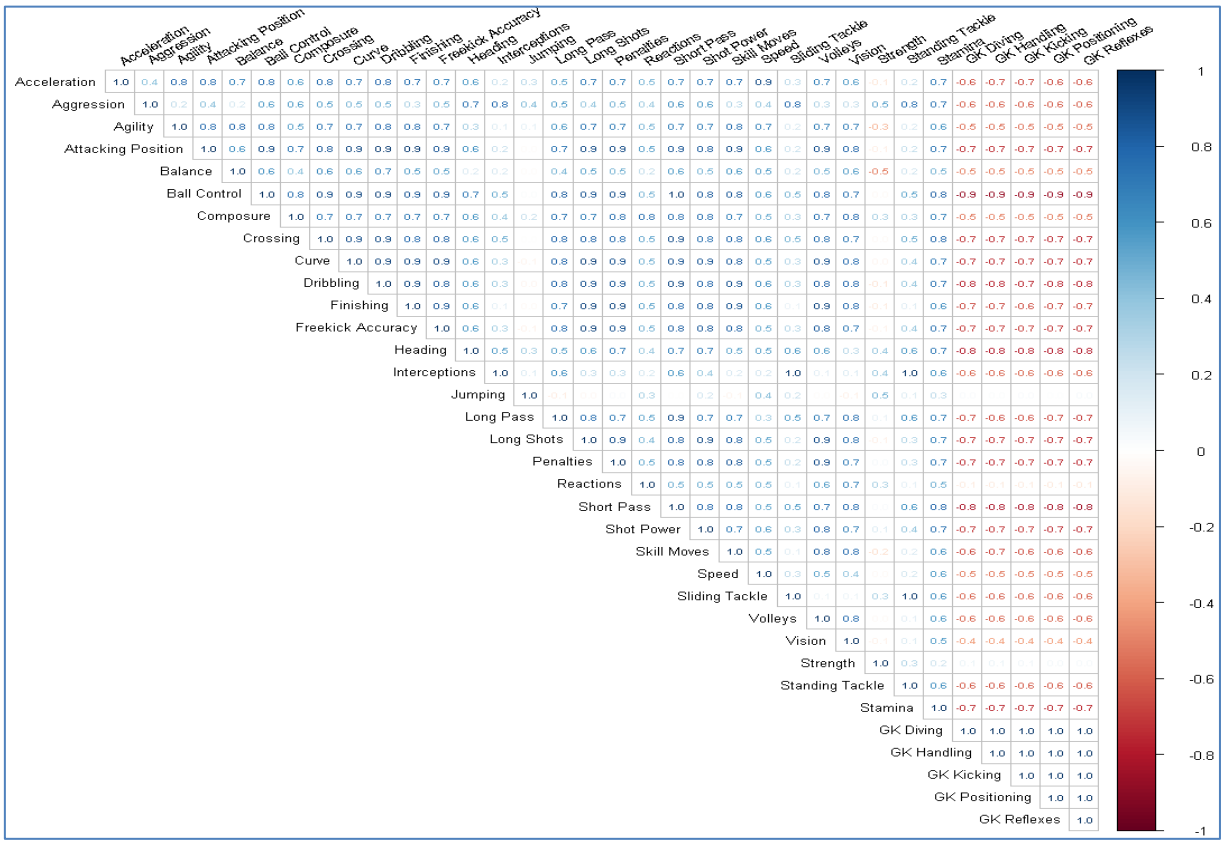
Matriz de correlación e Indicadores de Bondad

La creación de la matriz de correlaciones bajo el método de Pearson está dada por el supuesto de que no existe razón para querer atribuir más importancia a unas variables que a otras, es decir que se trabaja con los datos tipificados y estandarizados para asegurar la variabilidad de cada una de ellas. Un resumen de estadística descriptiva (librería "stat.desc") sobre la base de datos permite apreciar fácilmente la disparidad en los valores de la varianza individual de cada variable. La matriz de correlación (Figura 2) demuestra que gran parte de las variables se encuentran positivamente inter-correlacionadas, a excepción de las asociadas

a GK (“Goalkeeping”, cuyo valor asciende cuando se trata de un jugador arquero, y desciende cuando efectivamente no lo es)

Figura 2

Matriz de correlación



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021)

- Test de esfericidad de Bartlett (librería “cor”): P- Value = 0,000; se rechaza H0 (hipótesis nula; la matriz de correlaciones es una matriz de identidad). Los valores obtenidos fueron de 4112.45 para Chi Cuadrado y 0 para P-Valor.
- Kaiser, Meyer y Olkin (librería “KMO”): dado que el Overall MSA es de 0.89 se determina que la muestra es adecuada. La figura 3 provee un detalle ampliado de lo mencionado:

Determinación de componentes principales a retener

- Criterio del autovalor superior a la unidad: Siendo cada autovalor la varianza extraída por cada componente se seleccionan aquellas superiores a 1.
    - En el presente caso corresponde retener cinco (5) componentes principales, puesto que a partir del número seis (6) la misma adopta valores inferiores a la unidad
- Criterio del gráfico de sedimentación (Scree plot): La Figura 5 (disponible en Apéndices) demuestra cómo el componente principal uno (1) representa el 61.5% de las varianzas, mientras que los subsiguientes recolectan sucesivamente en menor

medida el porcentaje restante. Respaldando el criterio anterior, al retener un total de cinco (5) componentes, la varianza total acumulada alcanzaría un total 90.94% (librería "fviz_screeplot")

- Parallel Analysis ("Horn"): El método sugiere contrastar los "auto-valores" obtenidos mediante un componente principal paralelo aplicado a muestras aleatorias de variables no correlacionadas, con el mismo número de variables y observaciones que la muestra original. La misma generará autovalores que se han ajustado para evitar el error muestral. La aplicación del procedimiento descrito (librería "paran") dio como resultado 3 componentes principales a retener. Visto y considerando que dichas 3 componentes acumulan un 82% de la varianza, y esto no representa una pérdida de información significativa se considera al método como definitivo.

Obtención e interpretación de cargas factoriales y resultados

Mediante la estandarización de los datos originales, se obtuvieron los autovalores que brindan una medida de la longitud del "auto-vector" unitario y estimar así su homogeneidad o heterogeneidad (librería "prcomp"). Se conformaron 34 componentes principales, de los cuales se exhiben los primeros 3 para facilitar la visualización. Se puede apreciar en la Figura 7 que el componente principal 1, indica cargas factoriales cuya máxima magnitud corresponden a Ball Control, Short Pass, Dribbling, Curve, Attacking Position, Crossing, Penalties, entre otras. La interpretación de estos resultados permite anticiparse a que el componente bajo análisis reúne las habilidades más típicas de individuos "mediocampistas ofensivos" y "delanteros". Amerita mencionar que la correlación positiva de las variables típicas de GK ("Goalkeeping") se presenta en forma contrapuesta con el restante. Esto no resulta llamativo, teniendo en cuenta que solo los "arqueros" registran calificaciones elevadas únicamente en dichas variables, y las plantillas de cada equipo solo suelen incorporar entre 3 y 4 de ellos, sobre un total promedio de 23 jugadores. Por su parte, se puede observar que al trasladarse al componente principal 2, la constitución de este está mayormente dada por Sliding Tackle, Standing Tackle, Interceptions, Aggression, Strength, Heading, Jumping, entre otras. Las habilidades destacadas habitualmente son frecuentadas por "defensores" y "mediocampistas defensivos"

En el componente principal 3 su conformación pasa a estar mayormente determinada por variables de carga positiva como Reactions, Strength, Jumping, Composure, GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes, entre otras. Como se indicó con anterioridad, las variables mencionadas se asocian habitualmente a "arqueros" y en menor medida a "defensores" (dado que se los considera el recambio más próximo en caso de ausencia).

Figura 7
Cargas Factoriales

Variable	PC1	PC2	PC3
Acceleration	-0,175	0,090	0,027
Aggression	-0,136	-0,301	0,179
Agility	-0,167	0,181	-0,021
Attacking Position	-0,202	0,135	0,008
Balance	-0,134	0,139	-0,157
Ball Control	-0,214	0,008	-0,040
Composure	-0,169	0,029	0,291
Crossing	-0,200	0,024	0,003
Curve	-0,202	0,095	-0,001
Dribbling	-0,205	0,097	-0,038
Finishing	-0,191	0,187	0,028
Freekick Accuracy	-0,195	0,084	0,001
Heading	-0,163	-0,189	0,080
Interceptions	-0,110	-0,369	0,005
Jumping	-0,018	-0,135	0,387
Long Pass	-0,183	-0,041	0,041
Long Shots	-0,196	0,119	-0,012
Penalties	-0,197	0,085	0,043
Reactions	-0,110	0,098	0,451
Short Pass	-0,207	-0,033	-0,012
Shot Power	-0,194	0,027	0,099
Skill Moves	-0,184	0,174	-0,013
Speed	-0,141	0,048	0,100
Sliding Tackle	-0,108	-0,380	-0,020
Volleys	-0,181	0,180	0,085
Vision	-0,161	0,202	0,140
Strength	-0,004	-0,250	0,440
Standing Tackle	-0,114	-0,370	-0,043
Stamina	-0,185	-0,119	0,084
GK Diving	0,188	0,123	0,226
GK Handling	0,189	0,126	0,225
GK Kicking	0,187	0,128	0,221
GK Positioning	0,188	0,130	0,212
GK Reflexes	0,190	0,127	0,210

Nota. Tabla elaborada mediante la utilización del software Rstudio (PBC, 2021)

Los precedentes resultados logran ser plasmados en un gráfico Plot (librería “fviz_pca_var” – Figuras 9,10 ,11 y 12) de doble eje (para cada dimensión de a pares), donde cada una de las variables originales se presenta como un vector siendo sus coordenadas la carga factorial de este. Asimismo, las puntuaciones en los componentes principales aportan las coordenadas para poder graficar los individuos / observaciones sobre el mismo plano (“fviz_pca_ind”). La combinación de variables y objetos amplía la interpretación de cómo las primeras se asocian entre sí (para formar un componente principal) y a su vez, dada su proximidad, como los individuos se relacionan con estas.

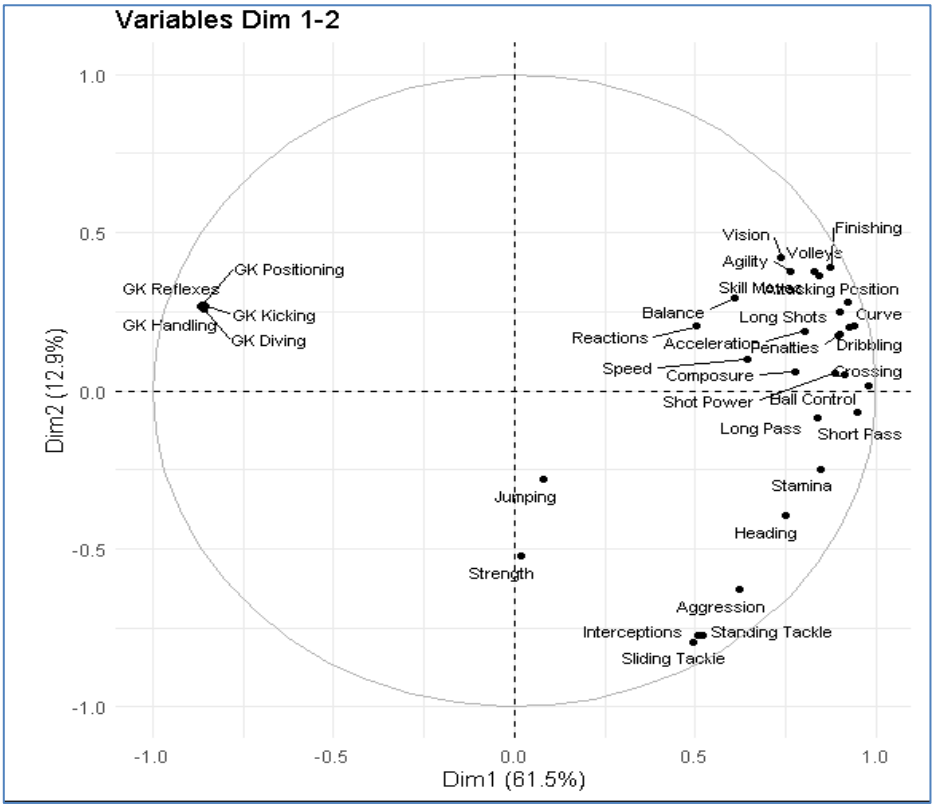
A simple vista se detecta una significativa condensación de variables con contribución elevada tanto en la componente principal 1 y 2. Corresponde reiterar que las mismas podrían ser etiquetadas como “Delanteros y mediocampistas ofensivos”. Al focalizarse en los individuos, se aprecia que “atacantes” como Messi, Neymar, Cristiano Ronaldo, Luis Suarez, Benzema (entre otros), o “mediocampistas ofensivos” como Modric, Turan, Kross

presentan ubicación próxima a las variables antes mencionadas. Dicho análisis se hace extensivo a las variables etiquetadas como "Defensores y mediocampistas defensivos" posicionadas en el cuadrante IV (y en menor medida en el III), correspondiendo individuos "defensores" como Mascherano, Marcelo, Digne, Ramos y a mediocampistas como Sergi y Busquets.

Resulta evidente la agrupación de las habilidades de GK y sus correspondientes individuos "arqueros" en el cuadrante II, por lo descrito con en párrafos precedentes (en resumen).

Figura 9

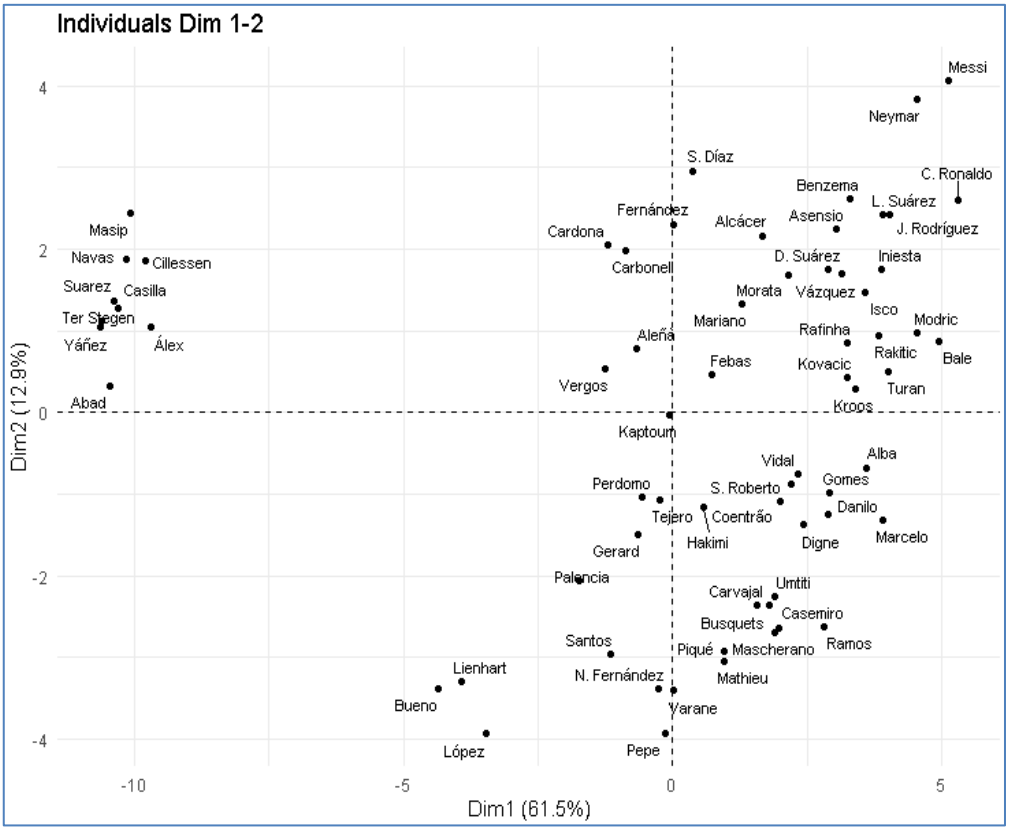
Gráfico de las cargas factoriales como vectores en sus correspondientes coordenadas para los pares de Dimensiones 1 con 2 correspondientes a variables.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021)

Figura 10

Gráfico de las cargas factoriales como vectores en sus correspondientes coordenadas para los pares de Dimensiones 1 con 2 correspondientes a individuos



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Centrándose en el gráfico de a pares de la dimensión 2 y 3 (disponible en apéndices), se distingue una distribución más homogénea entre los 4 cuadrantes, teniendo en cuenta que ambas recolectaron porcentajes menores (y similares) de la variabilidad. A su vez, el etiquetado que estas recibieron fue de "Defensa" y "Portería", habilidades que son compartidas por todo el plantel de jugadores, aunque en menor medida por los "Atacantes" y "Mediocampistas".

Análisis de Clústeres:

Tipificación / estandarización de datos: Con la finalidad de evitar la influencia no deseable de la unidad de medida / valor absoluto de una variable sobre la comparativa de datos, se procede a tipificar / estandarizar los valores. Es decir, se busca el mayor grado de homogeneidad de la muestra.

Test de Hopkins: Se utilizó para medir la tendencia de agrupamiento de un conjunto de datos. Sirve como una prueba de hipótesis estadística en la que la hipótesis nula se define bajo el supuesto de que los datos se generan mediante un proceso de puntos de Poisson y, por lo tanto, se distribuyen de manera uniforme y aleatoria. Un valor cercano a 1 tiende a indicar que los datos están muy agrupados, los datos aleatorios tenderán a dar como resultado

valores alrededor de 0.5 y los datos distribuidos uniformemente tenderán a dar como resultado valores cercanos a 0. En el presente caso, la prueba arrojó un valor favorable de 0.2579, por lo cual se prosiguió con el análisis (Liberia "Hopkins")

Matriz de distancias euclídeas: Representa un proceso esencial en el análisis de clústeres dado que actúa como una medida de similitud (o disimilitud en su defecto) entre las distintas observaciones. En tal sentido, permite establecer un punto de partida para el criterio de agrupación. La matriz se compone de los cálculos matemáticos de la distancia euclídea y distancia euclídea al cuadrado. La figura 13 (disponible en Apéndices) opera como mapa de calor donde la máxima intensidad del color demuestra el valor mínimo de la distancia euclídea, y viceversa (Liberia "dist")

Identificación del número adecuado de clústeres: El análisis se valió de la aplicación inicial de un tipo jerárquico (Método de Ward, nutrido a su vez de los componentes principales anteriormente calculados) con el fin de establecer la cantidad optima de conglomerados. La metodología se extendió al tipo no jerárquico con método K-Means para profundizar el criterio de selección.

- Ward: El método prioriza la maximización de la homogeneidad del grupo, mediante el cálculo de los "centroides" de los grupos resultantes de las posibles funciones. Prosigue con la obtención de la suma de cuadrados total, y aquella que arroje el menor valor será la que representa la maximización mencionada.

    - La aplicación de este procedimiento dio como resultado la creación de 2 grupos (Librería "hclust"). El dendograma correspondiente se expone en la sección de anexos.
    - Como soporte complementario se prosiguió con la incorporación de los estadísticos de Hubert y Dindex (Librería "NbClust"), los cuales sugirieron la creación de 3 grupos (ver Figura 15, disponible en Apéndices)

    - La diferencia de propuestas derivó en la aplicación de la "regla de la mayoría" (propuesta que reúna mayor cantidad de índices). El resultado fue el siguiente:
        - 14 índices propusieron que la cantidad adecuada de clústeres es de 2.
        - 7 índices propusieron que la cantidad adecuada de clústeres es de 3.
        - 2 índices propusieron que la cantidad adecuada de clústeres es de 4.
        - 2 índices propusieron que la cantidad adecuada de clústeres es de 5.

Si bien la regla de la mayoría sugiere la creación de 2 grupos, se considera que a la presente investigación se le adecua de mejor modo la asignación de 3 clústeres, partiendo del estado de conocimiento previo de las variables y observaciones. Esto significa, que agrupar jugadores en 2 grandes grupos, generaría una condensación significativa que impediría la correcta distinción entre ellos. Asimismo, se debe tener en cuenta que el Análisis de Componentes Principales permitió realizar un etiquetado de sus componentes, sobre los cuales se podía apreciar a priori 3 tipos de jugadores. En línea con lo mencionado, se optó por avanzar con un análisis de clústeres no jerárquico con método K-Means.

- K-Means: En el método en referencia se conoce a priori el número de conglomerados a conformar (3), sobre los cuales se asignan cada una de las observaciones para garantizar homogeneidad interna, y a su vez máxima heterogeneidad entre otros grupos.
    - La aplicación del método brindó los siguientes resultados (librería "kmeans" – Figura 16):

Figura 16

Conformación de grupos mediante K-Means

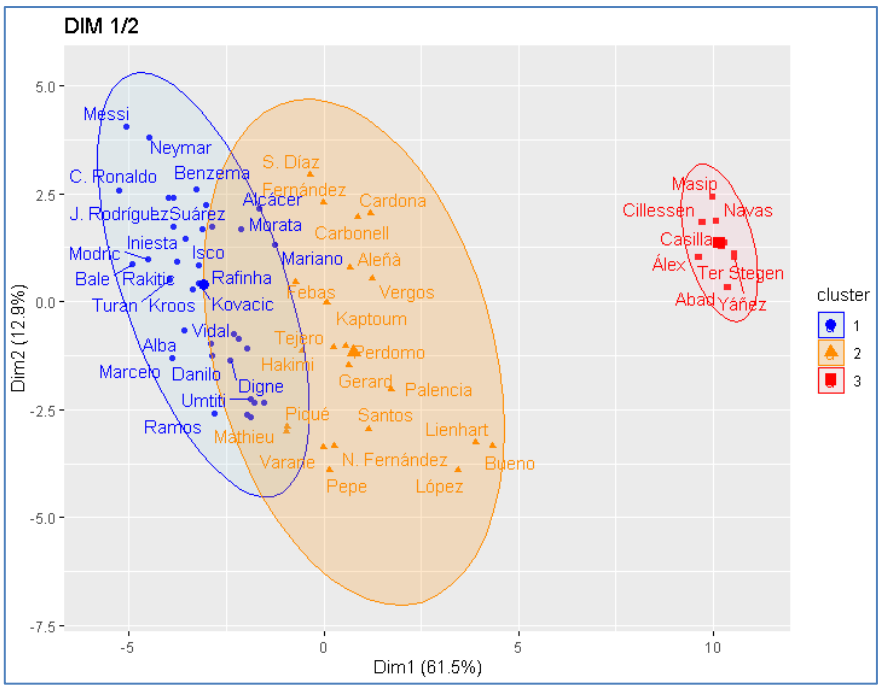
```
K-means clustering with 3 clusters of sizes 9, 22, 35

Cluster means:
  Acceleration  Aggression    Agility Attacking Position    Balance Ball Control  Composure   Crossing      Curve  Dribbling  Finishing
1   -1.6696430 -1.61527111 -1.3863021         -1.8857113 -1.2492707   -2.2257677 -1.3930293 -1.8808499 -1.8684288 -1.9487634 -1.7011134
2   -0.2367634 -0.03014142 -0.3739162         -0.3415200 -0.2147714   -0.1005971 -0.5230589 -0.2798504 -0.3261640 -0.2596446 -0.3519545
3    0.5781595  0.43430147  0.5915107          0.6995669  0.4562402    0.6355727  0.6869874  0.6595531  0.6854705  0.6643158  0.6586577
  Freekick Accuracy    Heading Interceptions     Jumping  Long Pass Long Shots  Penalties  Reactions Short Pass Shot Power Skill Moves      Speed
1        -1.8268053 -2.0284110    -1.4537541 -0.02049876 -1.7044465 -1.8344655 -1.9097551 -0.4112569 -2.1208424 -1.8436477  -1.6554045 -1.3053368
2        -0.2728867  0.2162706     0.1923086 -0.36670008 -0.2066239 -0.2627478 -0.2126544 -0.8690711 -0.1474294 -0.3096333  -0.3369407 -0.1879446
3         0.6412787  0.3856499     0.2529428  0.23576830  0.5681642  0.6368755  0.6247484  0.6520251  0.6380294  0.6687074   0.6374668  0.4537947
  Sliding Tackle    Volleys     Vision    Strength Standing Tackle    Stamina  GK Diving GK Handling GK Kicking GK Positioning GK Reflexes
1     -1.4977255 -1.5192563 -1.1129516  0.06039001      -1.5667867 -1.9150720  2.4458090   2.4393793  2.4326707      2.4183667   2.4234171
2      0.2697981 -0.4675312 -0.6621288 -0.23778568       0.2849761 -0.2272354 -0.4163031  -0.3932339 -0.4102328     -0.3741482  -0.3823955
3      0.2155420  0.6845427  0.7023828  0.13393642       0.2237602  0.6352808 -0.3672461  -0.3800934 -0.3676833     -0.3866868  -0.3828015

Clustering vector:
      Hakimi        Febas        Vidal        Aleñà         Àlex       Tejero        Gomes        Turan        López    Carbonell         Abad
           2            2            3            2            1            2            3            3            2            2            1
    Carvajal     Casemiro   C. Ronaldo       Danilo    D. Suárez    Fernández     Coentrão         Bale       Gerard      Iniesta         Isco
           3            3            3            3            3            2            3            3            2            3            3
     Rakitic J. Rodríguez    Cillessen   Mascherano      Mathieu         Alba        Masip      Suarez      Benzema        Navas      Casilla
           3            3            1            3            2            3            1            1            3            1            1
       Messi        Digne      Vázquez    L. Suárez       Modric      Cardona   Ter Stegen      Marcelo      Asensio      Mariano       Santos
           3            3            3            3            3            2            1            3            3            3            2
     Kovacic       Morata N. Fernández       Neymar       Vergos      Perdomo      Alcácer         Pepe     Lienhart        Piqué      Rafinha
           3            3            2            3            2            2            3            2            2            2            3
      Varane        Yáñez       Umtiti        Bueno     Palencia   S. Roberto     Busquets      S. Díaz        Ramos        Kroos      Kaptoum
           2            1            3            2            2            3            3            2            3            3            2

within cluster sum of squares by cluster:
[1]  63.71725 339.53432 402.00097
 (between_SS / total_SS =  63.6 %)
```

Nota. Tabla elaborada mediante la utilización del software Rstudio (PBC, 2021*)*

Grupos conformados: Las figuras 17 y 18 brindan con claridad visual la conformación de grupos a través del método K-Means. Al observar los mismos con detenimiento se logra apreciar una marcada similitud con la distribución de individuos demostrada por los Componentes Principales previamente descritos. Este aspecto se hace presente tanto en la comparación de la dimensión "1" y "2", donde jugadores como Messi, Neymar y Cristiano Ronaldo son asignados al clúster, recordando que la componente principal que se les asocia correspondía a "Delanteros y Mediocampistas ofensivos". Aplicando la misma interpretación al clúster 2 y 3, nuevamente se replica la distribución aportada sus componentes principales homónimas ("Defensores y mediocampistas defensivos" – "Defensa" y "Portería") agrupando los mismos jugadores observados.

Figura 17

Conformación de clústeres según las dimensiones 1 y 2



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021)

2.3 Interpretación de resultados y su utilidad para optimizar los modelos predictivos

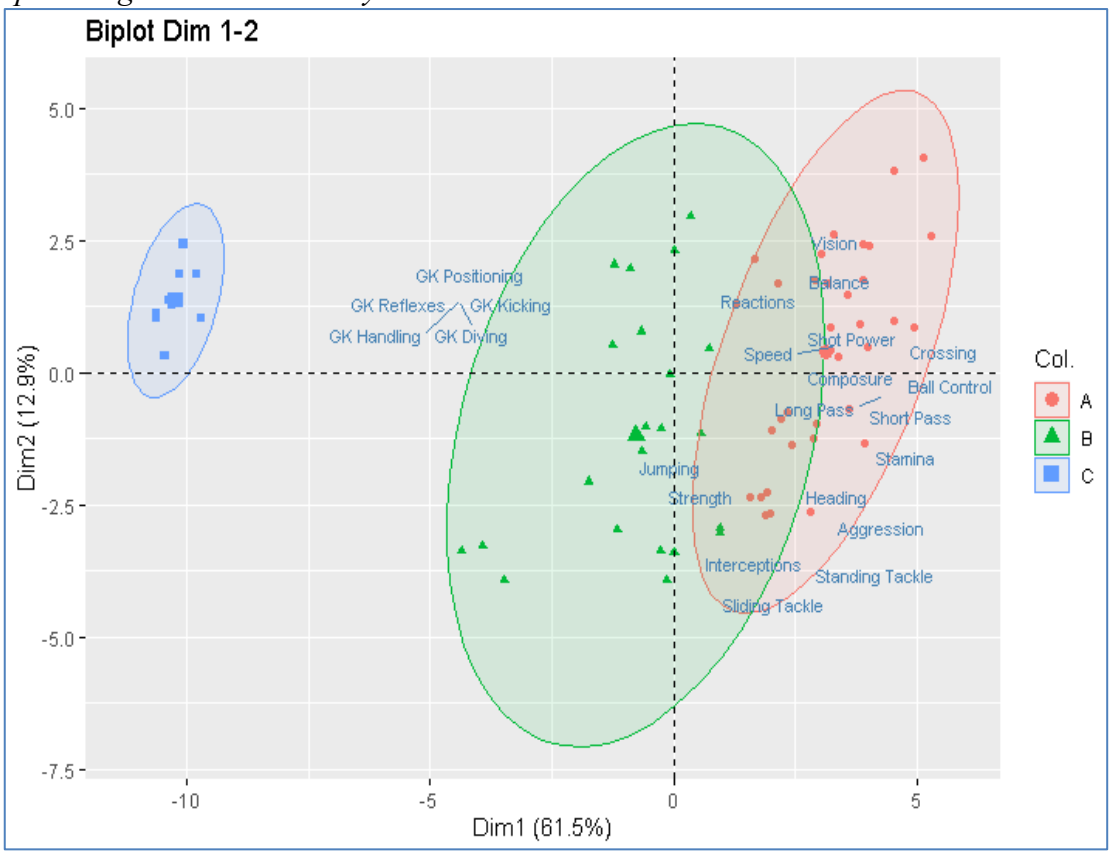
Los métodos de análisis multivariado-presentados tienen objetivos y enfoques independientes. No obstante, la aplicación de ellos manifestó una utilidad clara:

- El Análisis de Componentes Principales" no sólo permitió reducir la dimensionalidad de la información, sino que también anticipó gráficamente la distribución de las variables y la asociación de los individuos con estas.
- Por su parte el "Análisis de Clúster", determinó matemáticamente la conformación apropiada de grupos resultando a tal efecto complementario al etiquetado previamente realizado sobre los componentes principales.

La sinergia existente entre ambos métodos abre la posibilidad de realizar gráficos (Figuras 19 y 20) conjunto de ambos, donde el "Biplot" expuesto en secciones previas incorpora elementos visuales (elipses) que sintetizan fácilmente el desarrollo del presente trabajo:

Figura 19

"Biplot" según dimensiones 1 y 2.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021)

Los métodos de análisis multivariante aplicados actúan como herramienta anticipatoria de las variables a enfocar para la conformación de planteles competitivos para equipos profesionales de E-SPORTS. A su vez, asienta una base sólida para plantear los modelos predictivos a aplicar en el presente trabajo, y efecto iniciar un proceso iterativo de ajuste de estos para alcanzar resultados con la mayor precisión posible. Por dicho motivo, el precedente apartado se focalizará en técnicas de Machine Learning para abordar lo mencionado.

3 Implementación de modelos predictivos para el rating individual de cada jugador

El presente apartado se centra en los métodos analíticos predictivos que mejor se adecuen a la resolución de la problemática planteada, abarcando su marco conceptual, aplicación, análisis de resultados obtenidos, optimización y comparación de resultados obtenidos.

3.1 Marco conceptual de los fundamentos de Métodos Analíticos Predictivos y resumen del desarrollo metodológico.

Al presentar el concepto de métodos analíticos predictivos se establece una relación directa y referencia con la minería de datos. Dicho concepto reúne métodos de análisis empresarial que no se limitan a operaciones matemáticas básicas, estadística descriptiva y/o elaboración de informes gerenciales basados en reglas y preferencias del negocio. En concreto, se refieren a métodos estadísticos y de aprendizaje autónomo, que facilitan la toma de decisiones en forma automatizada (Michael Steinbach, Pang-Ning Tan, Vipin Kumar, 2005). Para ello, la predicción representa comúnmente el componente principal de este concepto. La era del Big Data, ha acelerado y potenciado el uso de la minería de datos, explicándose en un volumen exponencialmente incremental con el paso de los días (Galit Shmueli, Peter C. Bruce, Peter Gedeck, Nitin R. Patel, 2016)

El presente estudio tiene por objeto principal, la aplicación de métodos analíticos predictivos para analizar la variación de las variables en todas las versiones del videojuego y cómo afectan al rating global de cada jugador para predecir el modo en que dicha métrica se comportará en el siguiente periodo. Esto permite brindar herramientas útiles para la conformación de planteles competitivos para profesionales de deportes electrónicos. Para ello, se aplicarán modelos predictivos de regresión ajustados y optimizados en función de una serie de pruebas y así obtener el mayor grado de precisión en la predicción.

El desarrollo se basó inicialmente en la selección de la base de datos oficial de FIFA EA-SPORTS® cuya publicación es de carácter anual por cada entrega del simulador / videojuego, a su vez disponible en múltiples sitios web tales como Kaggle®. Al momento de haberse realizado el presente estudio, la base reunía datos para las versiones comprendidas entre FIFA 15 y FIFA 22. Posteriormente, se prosiguió con la limpieza, transformación y modelado de los datos a fines de obtener una estructura apta para su procesamiento dentro del modelo predictivo. Los modelos aplicados fueron los de “Decision Tree”,” Gradient Tree”, “Random Forest”, “Deep Learning” por tratarse de un problema de regresión. El modelo finalmente seleccionado fue aquel que arrojó, luego de múltiples pruebas, un mejor resultado sobre la métrica para evaluar su eficiencia. Para el caso en cuestión, se tomó como principal métrica al Error Cuadrático Medio, siendo esta acompañada del Coeficiente de Determinación para evaluar la eficiencia de clasificación.

La evaluación de los resultados presentó bases sólidas para la elaboración de una conclusión general, basada en la utilidad del modelo, sus principales fortalezas, debilidades y reflexiones finales acerca de la metodología de cálculo propuesta por FIFA EA-SPORTS® para el rating general de los individuos.

3.2 Análisis e interpretación del comportamiento histórico de la variable a predecir.

El desarrollo e implementación de un método analítico predictivo, no solo consta de un conjunto de técnicas computacionales asociadas a un proceso de automatización en el aprendizaje de datos, sino más bien parte, se basa y fundamenta en un fuerte componente estadístico focalizado en las variables a analizar.

Comprender el comportamiento histórico de la variable a predecir ("Label") constituye un paso esencial en la definición de un método analítico predictivo a sabiendas que en función de ello se comprende el recorrido que puede adoptar ésta (en caso de ser numérica, como el caso en referencia). En tal sentido, comprender los guarismos típicos e históricos que consecuentemente permitirán definir valores esperados y detección de anomalías en su comportamiento.

Para el caso concreto bajo análisis, la variable "Rating General" (originalmente nombrada como "Overall", en la base de datos utilizada), tiene un recorrido de 0 a 99 (más es mejor) y actúa como una calificación general / consolidada de todas las habilidades y características de un individuo. FIFA EA-SPORTS ® representa gráficamente y en forma clara lo mencionado:

Figura 21
"Rating General"

Nota. Imagen tomada de Electronic Arts. (2021). Obtenido de EA Games: https://www.ea.com/es-es/games/fifa/fifa-21

Las expectativas generales de los usuarios presuponen un comportamiento interanual de baja volatilidad en dicha variable, fundamentándose en el paralelismo establecido con la vida real. Esto, en otras palabras, significa que, por tratarse de un videojuego simulador de Fútbol, la calificación global de un individuo se encuentre fuertemente vinculada al rendimiento real de un jugador de fútbol en su vida útil profesional.

Partiendo desde dicha expectativa compartida, se procedió a analizar el comportamiento histórico de la variable a predecir, de acuerdo con los siguientes pasos:

- Carga de la base a un motor de base de datos SQL (Microsoft, 2022)
- Creación de tabla consolidada de Ratings generales para cada versión del videojuego, discriminando a nivel fila cada individuo por su ID.
- Cálculo de diferencia de Rating entre la versión del video juego seleccionado y su precedente inmediato (Ejemplo: FIFA 17 vs FIFA 16; no aplicable a FIFA 15 por ser el primero de la serie histórica)
- Sobre las variaciones obtenidas, se calculó el promedio, mínimo y máximo.

Figura 21

Variación histórica del rating y creación de nuevos atributos.

ID	F15	F16	F17	F18	F19	F20	F20	F222	F16 vs F15	F17 vs F16	F18 vs F17	F19 vs F18	F20 vs F19	F20 vs F20	F22 vs F21
158023	93	94	93	93	94	94	93	93	1	-1	0	1	0	-1	0
158023	93	93	93	93	94	94	93	93	0	0	0	1	0	-1	0
188545	87	87	90	91	90	89	91	92	0	3	1	-1	-1	2	1
188545	87	90	90	91	90	89	91	92	3	0	1	-1	-1	2	1
20801	92	94	94	94	94	93	92	91	2	0	0	0	-1	-1	-1
200389	77	81	87	88	90	91	91	91	4	6	1	2	1	0	0
200389	77	87	87	88	90	91	91	91	10	0	1	2	1	0	0
192985	81	86	88	89	91	91	91	91	5	2	1	2	0	0	0
20801	92	93	94	94	94	93	92	91	1	1	0	0	-1	-1	-1
192985	81	88	88	89	91	91	91	91	7	0	1	2	0	0	0

F16 vs F15	F17 vs F16	F18 vs F17	F19 vs F18	F20 vs F19	F21 vs F20	F22 vs F21	Promedio consolidado
2.86	0.65	0.61	0.57	0.57	0.11	0.62	0.89
-11.00	-8.00	-10.00	-8.00	-10.00	-8.00	-10.00	-9.17
29.00	22.00	21.00	19.00	17.00	15.00	17.00	20.50

Nota. Tablas elaboradas mediante la utilización del software Excel (Microsoft, 2022*)*

Los resultados obtenidos permiten establecer las siguientes conclusiones:

- El videojuego FIFA EA-SPORTS® establece la calificación global de un individuo vinculada al rendimiento real de un jugador de fútbol en su vida útil profesional, visto y considerando que la variación promedio es de tan solo 0.89.
- La menor variación promedio observada se da entre las versiones 21 y 20, siendo esta coherente por el prolongado periodo de inactividad deportiva profesional ocasionada por la pandemia COVID-19.
- Los valores máximos observados en las variaciones son explicados por el efecto "reciente promoción". Dicho concepto hace referencia aquellos individuos recientemente incorporados y promocionados de una categoría inferior / juvenil a primera división. Generalmente, se trata de individuos adolescentes transitando su primer año en competencias de alto nivel, y que transcurrido un año su calificación general es ajustada. Este efecto también es explicado por el hecho de que EA-Sports ® en primera instancia desconoce el rendimiento real del jugador profesional.
- Los valores mínimos se incrementan se explican por aquellos individuos cuyo jugador de futbol se encuentra en el final de su vida útil profesional.
- El promedio consolidado (0.89) permitieron establecer un parámetro de valor mínimo aceptable para el RMSE en el método analítico predictivo.
- Las variaciones analizadas brindaron conocimiento adicional para la creación de nuevos atributos. Los mismos son detallados en el siguiente subapartado del presente informe.

3.3 Limpieza, Transformación, selección y creación de atributos.

Por tratarse de un base de datos de publicación anual, el volumen de datos presenta una tendencia incremental fruto de incorporación de nuevos individuos al videojuego, como así también la repetición de los existentes, pero con sus atributos ajustados.

La base de datos se presenta como un cuadro de doble entrada, donde cada individuo es dispuesto como fila, y sus atributos, variables o características están dadas como columnas. Asimismo, cada versión del videojuego es separado en hojas independientes.

Tal y como se mencionó, en el subapartado 1.2 (“FIFA Ultimate Team, conceptualización del rating y variables que lo componen”), la base de datos cuenta con una serie de atributos numéricos cuyo recorrido es de 0 a 99 (más es mejor) que califican en términos generales la habilidad de cada individuo. Por su parte, se destacan atributos cualitativos que brindan información adicional para la clasificación de dichos individuos.

El proceso de evaluación de la presentación, calidad, cantidad de datos y atributos constituye un proceso clave en la confección de un modelo predictivo, debido a su potencial incidencia en la performance de este. Resulta tentador variados modelos de aprendizaje automático y limitarse a aquel que mejor funcione. Ello, sin embargo, no significa que necesariamente sea el que en un subconjunto de datos de prueba se desempeñó en forma deseable. Este aspecto se encuentra intrínsecamente ligado al problema del sobreajuste (“Overfitting”), donde el modelo se encuentra significativamente ligado a su partición de entrenamiento. Por ende, es incorrecto suponer que el rendimiento representa fielmente el desempeño esperado sobre aquellos a predecir. El sesgo de sobreajuste puede estar siendo generado por un modelo reducido en atributos, dando lugar a una solución sin utilidad en términos de predicción.

En tal sentido, se aplicaron diversos pasos de transformación y limpieza de la base de datos para generar utilidad en los atributos a considerar definitivos. De acuerdo a lo mencionado con Ian Witten, Eibe Frank, Mark Hall, Chris Pal en “Data Mining: Practical Machine Learning Tools and Techniques” (2016), existen múltiples métodos y recomendaciones de transformación de datos para alcanzar un modelo exitoso, entre los cuales se destacan: selección de atributos, discretización numérica de atributos, proyecciones, muestreo, limpieza genérica de caracteres no deseados o registros corrompidos / no deseados, transformación de múltiples clases de atributos a binarios, entre otros. La mayoría de los algoritmos de aprendizaje automático están diseñados para aprender cuáles son los atributos más apropiados para tomar sus decisiones.
Cada método elige el atributo más prometedor para dividir en cada punto, por lo cual se recomiendo eliminar aquellos que resulten irrelevantes o redundantes (Ian Witten, Eibe Frank, Mark Hall, Chris Pal, 2016)

En su estado inicial e inalterado, la base de datos contaba con un total de 106 columnas / atributos distribuidos entre variables cuantitativas y cualitativas. Debido a que gran parte de parte de ellas, desde su comienzo no aportaron utilidad alguna al potencial desarrollo, fueron sustraídas. Su utilidad nula era simplemente explicada por redundancia respecto a otro atributo que aportaba la misma información, o bien datos de tipo URL que contenían la imagen del individuo.

En tal sentido, se listan los pasos de depuración de datos y selección de atributos realizados:

- Transformacion de atributos polinómicos a numéricos: Los atributos Club Position, League Name, Preferred Foot originalmente representados por una cadena de texto fueron transformados en numéricos mediante la asignación de un ID.
- Selección de atributos:
    - Cuantitativos: Attacking Crossing, Attacking Finishing, Attacking Heading Accuracy, Attacking Short Passing, Attacking Volleys, Defending, Defending Marking Awareness, Defending Sliding Tackle, Defending Standing Tackle, Dribbling, Goalkeeping Diving, Goalkeeping Handling, Goalkeeping Kicking, Goalkeeping Positioning, Goalkeeping Reflexes, Goalkeeping Speed, Mentality Aggression, Mentality Composure, Mentality Interceptions, Mentality Penalties, Mentality Positioning, Mentality Vision, Movement Acceleration, Movement Agility, Movement Balance, Movement Reactions, Movement Sprint Speed, Pace, Passing, Physic, Potential, Power Jumping, Power Long Shots, Power Shot Power, Power Stamina, Power Strength, Shooting, Skill Ball Control, Skill Curve, Skill Dribbling, Skill Fk Accuracy, Skill Long Passing
    - Cualitativos: Age, Club Position ID, Club Team Id, Height Cm, International Reputation, League Level, League Name ID, Nationality, Id, Preferred Foot ID, Release Clause Eur, Short Name, Skill Moves, Id, Value Eur, Wage Eur, Weak Foot, Weight Kg.
- Eliminación de atributos:
    - Player Url: Representa la página web de la cual se obtiene la imagen del jugador dentro del videojuego.
    - Long Name: Se trata del nombre completo del jugador (Nombres y Apellidos) y es un valor redundante ya representado por Short Name (Nombre corto).
    - Dob: “Date of Birth”, fecha de nacimiento redundante por contar con el campo “Age” que aporta la edad en valor numérico.
    - Wage Eur: Salario del individuo expresado en EUR. Es un atributo que no es representativo de la habilidad o valoración general, dado que el salario puede definirse por múltiples factores.
    - Player Positions: Cadena de texto que mediante el separador “,” lista todas las posiciones posibles en la cual se puede desempeñar el individuo. La complejidad del atributo para darle un tratamiento específico en términos de modelado de datos, y su redundancia respecto al atributo “Team Position” el cual simplifica en un valor de texto único la posición en la cual se desempeña en su club (más frecuente).
    - Body Type: Código interno del videojuego para asignar el cuerpo virtual al individuo, por lo que se trata de un atributo sin importancia.
    - Real Face: Al igual que atributo precedente, se trata de un valor interno del videojuego para indicar si el individuo cuenta con una representación real de la cara (atributo sin importancia).
    - Player Tags: Cadena de texto que mediante el separador “,” lista y resume características generales del jugador. Atributo complejo para el procesamiento del modelo. A su vez las variables cuantitativas aportan un

valor real a la descripción de las características del jugador tal y como se demostró en el apartado 2 del presente informe.
    - Team Jersey Number: Número estampado en la camiseta del jugador dentro del club al que pertenece (atributo sin importancia)
    - Loaned From: Fecha en la cual fue cedido a préstamo al club en el que se desempeña el individuo (atributo sin importancia)
    - Joined: Fecha en la cual el individuo fue incorporado al club en el que se desempeña (atributo sin importancia)
    - Contract Valid Until: Fecha hasta la cual el individuo se encuentra vinculado al club en el que se desempeña
    - Nation Position: Posición la cual se desempeña el individuo si fuese convocado para su selección nacional. Se trata de un valor mayormente nulo dado que no todos tienen la posibilidad de ser seleccionados. A su vez el modelo se basa en el atributo "Team Position", por lo cual éste es redundante y no representativo.
    - Nation Jersey Number: Número estampado en la camiseta del jugador dentro de la selección nacional a la que podría pertenecer (atributo sin importancia)
    - Player Traits: Descripción ampliada del atributo previamente eliminado "Player_Tags"
- Creación de atributos:
    - "Prev_Overall": Rating Global del periodo anterior
    - "Diff_Prev_Overall": Diferencia respecto al Rating Global del periodo previo
    - "Prev_Potential": Potencial Global del periodo anterior
    - "Potential_Accuracy": Rating Global actual / Potencial del periodo anterior. El atributo busca medir la precisión de la proyección del periodo previo.
    - "Año": Año de la versión del videojuego a la que pertenece el registro
- Muestreo:
    - Se tomó la base de datos comprendida entre los periodos 2015 y 2022.
- Limpieza genérica de caracteres no deseados o registros corrompidos / no deseados:
    - La base contaba con registros íntegros y sin anomalías detectadas en términos de caracteres

3.4 Descripción de los modelos predictivos a implementar

La problemática planteada en el presente estudio se basaba en Regresión, visto y considerando que se pretendía predecir con la mayor precisión posible el Rating Global de un individuo en el siguiente periodo, adoptando la relación entre las variables presentes en la base de datos.

Existe una amplia diversidad de algoritmos predictores posibles y aptos para su consecuente implementación en función de la problemática planteada. Con la finalidad de realizar un experimento consistente, se aplicaron todos los algoritmos disponibles de Regresión que tolerase atributos polinómicos y numéricos conjuntamente, sumado a las recomendaciones realizadas en el Capítulo 8 "Aprendizaje de Ensambles" de "Data Mining: Practical Machine Learning Tools and Techniques" (2016) y la documentación oficial de RapidMinner Studio (software utilizado para la creación del modelo). Los algoritmos utilizados fueron los siguientes:

- Árbol de Decisión: es una colección similar a un árbol de nodos destinados a crear una decisión sobre la afiliación de valores a una clase o una estimación de un valor objetivo numérico. Cada nodo representa una regla de división para un atributo específico. Para la clasificación esta regla separa valores pertenecientes a diferentes clases, para la regresión los separa con el fin de reducir el error de forma óptima para el criterio del parámetro seleccionado. La construcción de nuevos nodos se repite hasta que se cumplen los criterios de parada. Se determina una predicción para el atributo de la etiqueta de clase en función de la mayoría de los ejemplos que llegaron a esa hoja durante la generación, mientras que se obtiene una estimación de un valor numérico promediando los valores de una hoja. (Rapidminer, 2022)
- Random Forest: Es un conjunto de un cierto número de árboles aleatorios que se crean/entrenan en subconjuntos. Cada nodo de un árbol representa una regla de división para un atributo específico. Solo se considera un subconjunto de atributos, especificado con los criterios de relación de subconjunto, para la selección de la regla de división. Esta regla separa los valores de forma óptima para los criterios de parámetros seleccionados. Para la regresión los separa para reducir el error que comete la estimación. La construcción de nuevos nodos se repite hasta que se cumplen los criterios de parada. El modelo resultante es un modelo de votación de todos los árboles aleatorios creados. Dado que todas las predicciones individuales se consideran igualmente importantes y se basan en subconjuntos de ejemplos, la predicción resultante tiende a variar menos que las predicciones individuales. (Rapidminer, 2022)
- Gradient Boosted Trees: Es un conjunto de modelos de árbol de clasificación o de regresión. Ambos son métodos de conjunto de aprendizaje progresivo que obtienen predicciones a través de estimaciones mejoradas gradualmente. “Boosting” es un procedimiento de regresión no lineal flexible que ayuda a mejorar la precisión de los árboles. Mediante la aplicación secuencial de algoritmos de clasificación a los datos modificados de forma incremental, se crean una serie de árboles de decisión que producen un conjunto de modelos de predicción. Si bien el “Boosting” en árboles aumenta su precisión, también disminuye la velocidad y la interpretabilidad humana. El método generaliza la potenciación de árboles para minimizar estos problemas. (Rapidminer, 2022)
- Deep Learning: Se basa en una red neuronal artificial de alimentación hacia adelante de múltiples capas que se entrena con un descenso de gradiente estocástico utilizando propagación hacia atrás. La red puede contener una gran cantidad de capas ocultas que consisten en neuronas con funciones de activación “tanh”, “rectifier” y “maxout”. Las funciones avanzadas, como la tasa de aprendizaje adaptable, el recocido de tasas, el entrenamiento de impulso, el abandono y la regularización L1 o L2 permiten una alta precisión predictiva. Cada nodo de cómputo entrena una copia de los parámetros del modelo global en sus datos locales con subprocesos múltiples (asincrónicamente) y contribuye periódicamente al modelo global a través del promedio del modelo en la red. (Rapidminer, 2022)

3.5 Componentes del proceso de aprendizaje automático

La herramienta de desarrollo elegida fue RapidMiner Studio (Rapidminer, 2022) por potente capacidad de procesamiento a nivel "On-premises" y amplia librería de componentes, operadores y modelos.

El proceso de aprendizaje automático se dividió en dos (2) etapas similares para la determinación del mejor modelo a implementar:

a) Creación de proceso estándar con modelo / algoritmo de aprendizaje automático variable / reemplazable:

De acuerdo con lo mencionado en el subapartado 3.4 del presente estudio, la problemática planteada en el presente estudio se basaba en Regresión, visto y considerando que se pretendía predecir con la mayor precisión posible el Rating Global de un individuo en el siguiente periodo. En virtud de ello, inicialmente se procesaron los datos en forma segmentada, es decir que el modelo captó el dataset individual de cada versión del videojuego con la finalidad de analizar su comportamiento. El proceso en cuestión se mantuvo bajo la misma estructura y parámetros estándar en cada prueba inicial, solo reemplazando el modelo predictor para comparar los resultados obtenidos entre cada versión. Finalmente, bajo la misma metodología se utilizó el mismo modelo, pero con el dataset consolidado, es decir con todos los periodos disponibles.

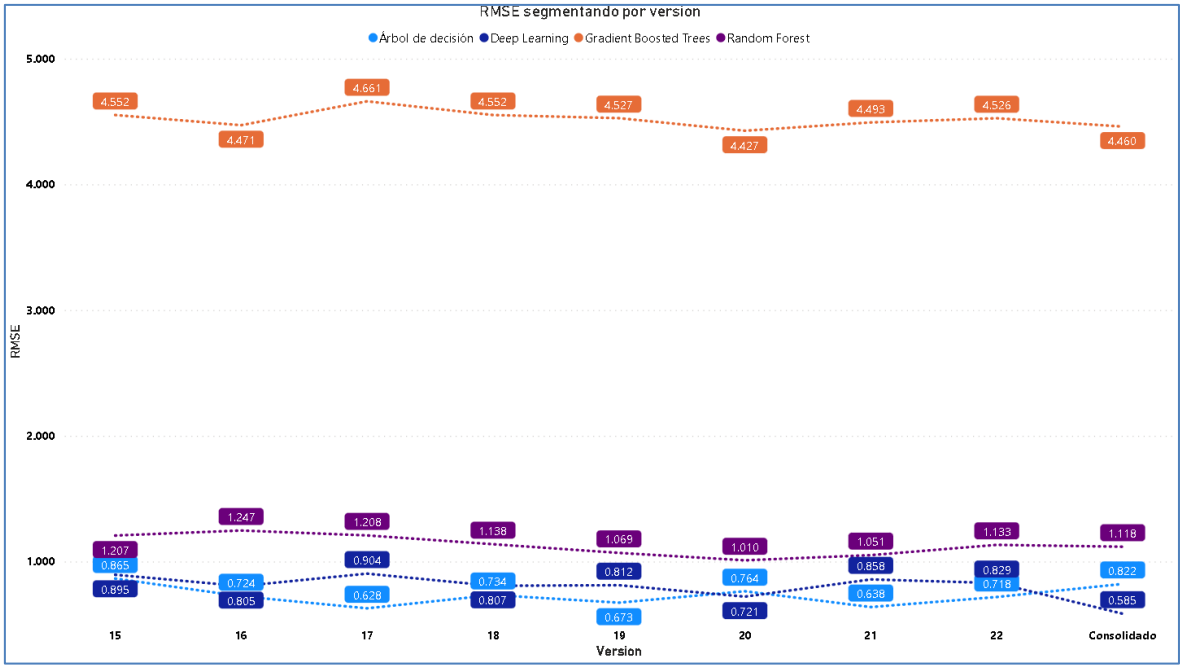
Los componentes y principales características / elementos son los siguientes:

- Lectura de datos: imputación de datos depurados mediante el operador "Read Excel", sin la incorporación de nuevos atributos.
- Definición de roles: la columna "ID" fue definida como identificador de campo, mientras que la columna "Overall" fue definida como variable a predecir ("Label") mediante el operador "Set Role"
- Particionamiento del dataset: Los parámetros aplicados para el particionamiento aleatorio fueron del 70% para entrenamiento y 30% para prueba ("Train vs Test"), logrando de tal modo un proceso de validación cruzada en base a un único dataset.
- Modelo predictivo: Componente variable y reemplazable ("Árbol de decisión"," Random Forest", "Gradient Boosted Trees", "Deep Learning") de acuerdo con cada ejecución realizada por cada dataset segmentado. Sus parámetros fueron definidos por default, es decir sin alteraciones respecto de su configuración estándar.
- Aplicación del modelo y cálculo de performance: En función de los resultados obtenidos del modelo predictivo utilizado en el paso descripto anteriormente, se aplica éste a la participación de prueba ("Test") con la finalidad de procesar y obtener la performance basada en la métrica "RMSE". Los operadores utilizados para dicha subetapa fueron "Apply Model" y "Performance" respectivamente.
- Exportación de resultados: conversión de resultados a un archivo plano ("CSV") para su posterior análisis y validación cruzada con los restantes datasets. Los operadores utilizados para dicha subetapa fueron "Select Attributes" y "Write CSV" respectivamente.

De acuerdo con lo expuesto en la Figura 22, se observa un comportamiento estable sin variaciones significativas en el RMSE entre cada versión del dataset para cada proceso implementado. Por su parte la performance de cada modelo tendió a asemejarse entre los modelos de "Árbol de decisión"," Random Forest" y "Deep Learning", distinguiéndose a su vez en forma notoria de "Gradient Boosted Trees". Sin embargo, cuando se consolidó el dataset se nutrió al modelo de un atributo histórico ("Año") cuya interpretación solo generó una mejora notoria en el modelo "Deep Learning". Dicha mejora permitió elegir al mismo como el modelo a aplicar sobre el proceso y experimento posterior correspondiente a la segunda etapa, en función de un RMSE final de 0.585.

Figura 22

RMSE obtenido por cada modelo a través de las versiones del videojuego.



Nota. Gráfico elaborado mediante la utilización del software Power Bi (Microsoft, 2022*)*

b) Creación de proceso Deep Learning nutrido con nuevos atributos:

La segunda etapa se distinguió de su precedente principalmente en dos aspectos: La utilización del Dataset consolidado con datos hasta la versión 2021 (entrenamiento) para predecir el rating general del Dataset versión 2022 (prueba) y la implementación de los nuevos atributos mencionados en el subapartado 3.3 a ser procesados por el modelo Deep Learning (parámetro de activación: "Rectifier").
Posteriormente, los valores predichos fueron comparados con su par real a fin de realizar una validación cruzada y medir la precisión del modelo.
El proceso arrojó un RMSE de 0.678 tras efectuar la validación cruzada, representando esto una mejora del 23% sobre el objetivo planteado (0.89; variación interanual promedio), y una diferencia de 0.093 respecto de la subetapa 1 (0.585, Dataset consolidado incluyendo 2022)

3.6 Análisis de resultados finales y utilidad del modelo para el público objetivo.

De acuerdo con la figura 23 "Pesos de atributos en la predicción Deep Learning (Top 10)", se observa un comportamiento particular en la ponderación de estos. Al haber predicho la variable Rating (originalmente nombrada "Overall") el nuevo atributo creado "Prev Overall" (que hace referencia al rating antecesor de la versión bajo análisis) cobró relevancia significativa, demostrando una correcta interpretación del modelo Deep Learning sobre la partición de prueba. Referirse a la Figura 23 (disponible en Apéndices) para visualizar el modelo construido y descripto en párrafos anteriores.

Por su parte, se destaca que tres de los 6 nuevos atributos creados forman parte del top 10, y que los atributos cualitativos se destacan por sobre los cuantitativos. Sobre estos últimos mencionados, el que lidera en posicionamiento es Potencial, que hace referencia a la proyección de rating general para el siguiente periodo. En tal sentido, se percibe una asociación entre los atributos directamente derivados del Rating Global con atributos cualitativos que hacen referencia a características reales de cada jugador bajo análisis.

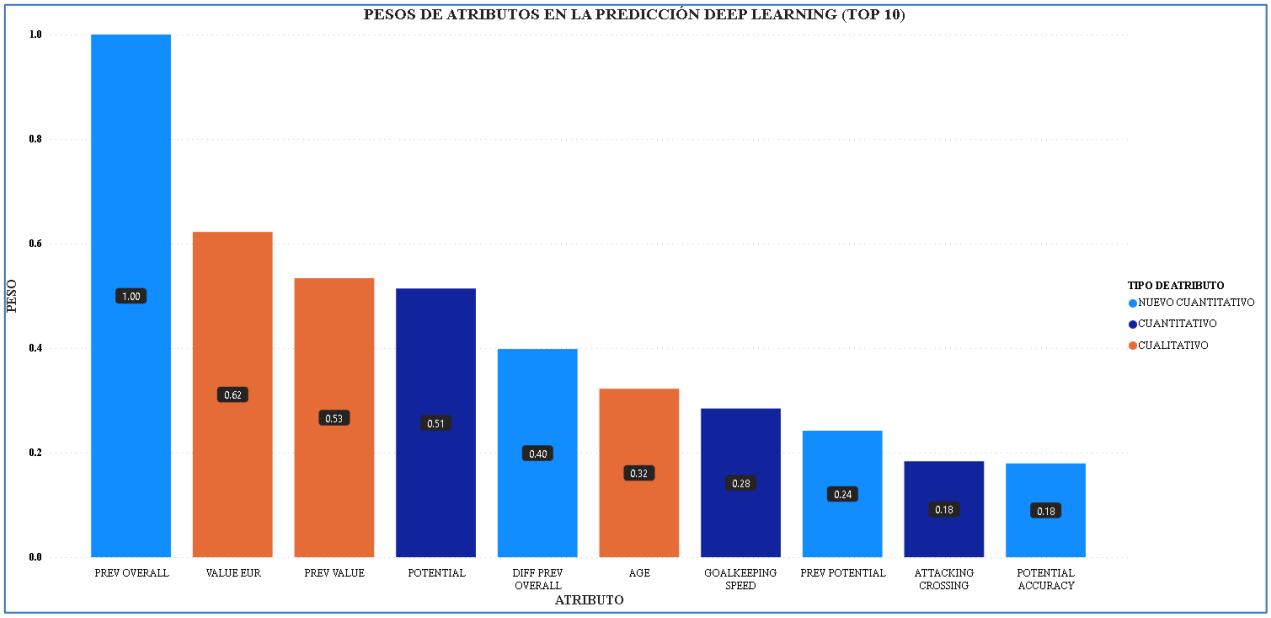
Haciendo referencia nuevamente a los atributos cuantitativos, resulta necesario remarcar que contrario a su interpretación inicial, la influencia sobre el rating global es moderadamente baja. De 43 atributos cuantitativos, solo tres forman parte del top 10, destacando los últimos dos de ellos representan habilidades / cualidades técnicas.

Lo descripto en párrafos precedentes, permitió establecer las siguientes afirmaciones:

- La definición del Rating General mayormente es construida en función de factores exógenos, tales como su valuación en el mercado, edad y liga a la que pertenece.
- La creación de atributos cuantitativos asociados al Rating y/o a factores exógenos mejora la precisión del modelo. Expandir esta metodología a nuevos atributos de igual categorización permite perfeccionar el proceso de predicción.
- La incorporación de un nuevo atributo proveniente de los resultados obtenidos en la etapa de análisis multivariante por componentes principales no resulta útil para perfeccionar el modelo. Esto se debe a que la reducción de la dimensionalidad propone agrupación por puesto de jugador, atributo ya existente en el Dataset original y con baja ponderación en el modelo (0.078).

Figura 24

Pesos de atributos en la predicción Deep Learning (Top 10)



Nota. Gráfico elaborado mediante la utilización del software Power Bi (Microsoft, 2022*)*

4 Conclusiones generales

El objetivo general del presente trabajo final de especialización fue analizar la variación de las variables en todas las versiones del videojuego para determinar cómo afectan al rating global de cada jugador y consecuentemente predecir el modo en que se comportará en el siguiente periodo. Para ello fueron propuestas dos metodologías que en principio fueron pensadas como complementarias, es decir que actuarían con una marcada sinergia para optimización mutua: Métodos de Análisis Multivariante (“PCA” y “Análisis de Clústeres”) y Métodos Analíticos Predictivos. La aplicación de ambas metodologías y análisis de resultados en forma individual brindaron propiciaron conclusiones solidas sobre la cual se fundaron utilidades claras para el público objetivo.

Los métodos de análisis multivariante aplicados actuaron como herramienta anticipatoria de las variables a enfocarse para la conformación de planteles competitivos para equipos profesionales de E-SPORTS. La reducción de la dimensionalidad mediante “PCA” aportó claridad a la asociación de individuos con variables en torno a su posición, mientras que el “Análisis de Clústeres” confirmó lo propio con la creación de 3 grupos. En tal sentido, la utilidad de dicha metodología radica en la facilidad para la conformación de planteles competitivos para profesionales de deportes electrónicos.

Por su parte, la aplicación de Métodos Analíticos Predictivos cumplió con el objetivo planteado de superar el RMSE mínimo (0.89; variación interanual promedio) tras obtener uno de 0.678 (mejora del 23%), pudiendo así aportar una base suficientemente sólida de predicción para el rating del individuo en la próxima versión del videojuego.

Resulta oportuno remarcar, que contrario a la suposición inicial, ambas metodologías (Métodos de Análisis Multivariante y Métodos Analíticos Predictivos), no pudieron ser integradas mutuamente. Esto se debió a que la incorporación de un nuevo atributo proveniente de los resultados obtenidos en la etapa de análisis multivariante no resultó útil para perfeccionar el modelo. La reducción de la dimensionalidad propone agrupación por puesto de jugador, atributo ya existente en el Dataset original y con baja ponderación en el modelo (0.078). Asimismo, se reveló que la construcción del rating mayormente es construida en función de factores exógenos, tales como su valuación en el mercado, edad y liga a la que pertenece, y que la precisión del modelo se incrementa mediante la incorporación de nuevos atributos correlacionados al mismo.
Las utilidades mencionadas a su vez proponen el perfeccionamiento posterior y permanente del modelo predictivo desarrollado, a fin de obtener la mayor precisión posible sobre una variable cuyas unidades de recorrido importan en demasía. La principal dificultad que ello presupone es la creación de nuevos atributos en función de factores exógenos correlacionados al rating global. Esto significa un seguimiento individualizado de cada jugador de su performance real, sabiendo que la base de datos cuenta con al menos 19.239 registros. Por su parte, la incorporación de dichos atributos depende de la disponibilidad de bases datos que cumplan con los principios de calidad de datos.

Conforme a lo expresado en los últimos dos precedentes párrafos, se puede afirmar que la aplicación de Métodos de Análisis Multivariante (“PCA” y “Análisis de Clústeres”) y Métodos Analíticos Predictivos aportan utilidad clara al publico objetivo para la toma de decisiones en su carrera profesional.

5 Referencias Bibliográficas

Cadenaser. (2016). Obtenido de Cadenaser: https://cadenaser.com/ser/2016/09/28/ciencia/1475069136_400673.html

Catena A, Ramos M y Trujillo H. (2003). Análisis multivariado. Un Manual para Investigadores. Editorial Biblioteca Nueva.

Chan, D.Y., Chiu, V. Y Vasarhelyi (2018). Continuous Auditing: Theory and Application. Cap. 1

Electronic arts. (2021). Obtenido de EA Games: https://www.ea.com/es-es/games/fifa/fifa-21

Electronic arts. (2021). Obtenido de EA Games: https://www.ea.com/es-es/games/fifa/compete/fgs-21/teams

Electronic arts. (2021). Obtenido de EA Games: https://www.ea.com/es-mx/games/fifa/fifa-21/news/pitch-notes-fifa21-global-series

Fifa Play (2021). Obtenido de Fifa Play: https://www.fifplay.com/fut-22-starting-guide/

Galit Shmueli, Peter C. Bruce, Peter Gedeck, Nitin R. Patel (2016). Data Mining for Business Analytics.

Ian Witten, Eibe Frank, Mark Hall, Chris Pal (2016). Data Mining: Practical Machine Learning Tools and Techniques.

Insider intelligence. (2021). Insider Intelligence. Obtenido de Insider Intelligence: https://www.insiderintelligence.com/insights/esports-ecosystem-market-report/

Michael Steinbach, Pang-Ning Tan, Vipin Kumar (2005). Introduction to Data Mining.

Newzoo. (2021). Newzoo. Obtenido de Newzoo: https://newzoo.com/insights/trend-reports/newzoo-global-games-market-report-2021-free-version/

RStudio Team (2021). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Toptal. (2021). Toptal. Obtenido de Toptal: https://www.toptal.com/finance/market-research-analysts/esports

Ultimate Team (2021). Obtenido de Ultimate Team: https://ultimateteam.online/

Rapidminer. (2022). Rapidminer. Obtenido de Rapidminer: https://docs.rapidminer.com/10.0/studio/operators/modeling/predictive/trees/parallel_decision_tree.html

Rapidminer. (2022). Rapidminer. Obtenido de Rapidminer: https://docs.rapidminer.com/10.0/studio/operators/modeling/predictive/trees/parallel_random_forest.html

Rapidminer. (2022). Rapidminer. Obtenido de Rapidminer: https://docs.rapidminer.com/10.0/studio/operators/modeling/predictive/trees/gradient_boosted_trees.html

Rapidminer. (2022). Rapidminer. Obtenido de Rapidminer: https://docs.rapidminer.com/10.0/studio/operators/modeling/predictive/neural_nets/deep_learning.html

Microsoft. (2022). Microsoft. Obtenido de Microsoft: https://www.microsoft.com/en-us/microsoft-365/excel

Microsoft. (2022). Microsoft. Obtenido de Microsoft: https://powerbi.microsoft.com/en-au/

Microsoft. (2022). Microsoft. Obtenido de Microsoft: https://learn.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver16

6 Apéndices

Se añaden gráficos de elaboración propia que no fueron incluidos en el cuerpo del Informe para cumplimentar con la extensión permitida del mismo.

Figura 3

KMO: Adecuación factorial

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = correlacion)
Overall MSA =  0.89
MSA for each item =
      Acceleration         Aggression            Agility Attacking Position            Balance       Ball Control          Composure
              0.85               0.91               0.90               0.93               0.91               0.90               0.86
          Crossing              Curve          Dribbling          Finishing  Freekick Accuracy            Heading      Interceptions
              0.88               0.86               0.92               0.95               0.88               0.86               0.82
           Jumping          Long Pass         Long Shots          Penalties          Reactions         Short Pass          Shot Power
              0.33               0.95               0.88               0.92               0.84               0.91               0.93
       Skill Moves              Speed      Sliding Tackle            Volleys             Vision           Strength   Standing Tackle
              0.92               0.81               0.89               0.91               0.91               0.59               0.80
          Stamina         GK Diving        GK Handling         GK Kicking     GK Positioning        GK Reflexes
              0.87               0.94               0.91               0.88               0.95               0.92
```

Nota. Tabla elaborada mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 4

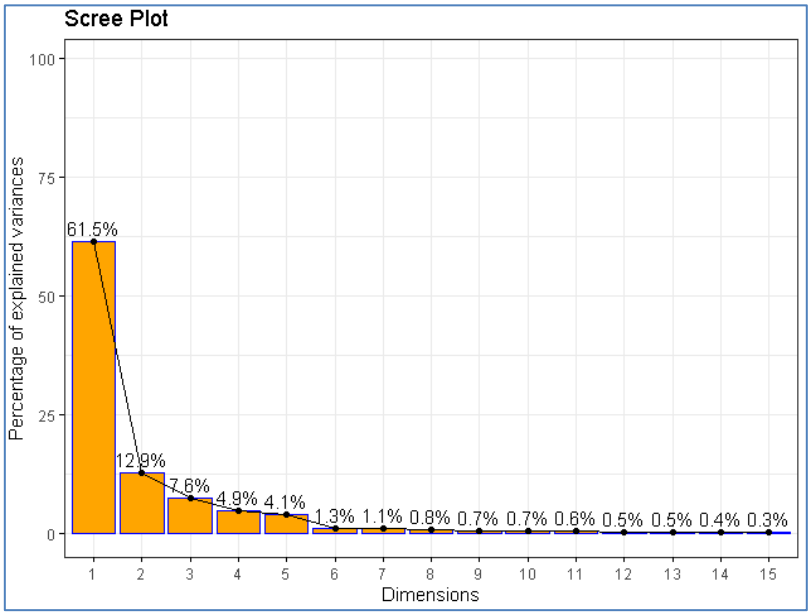
Varianza extraída por cada componente principal

```
> varianza
 [1] 20.922434229  4.371873154  2.575052904  1.671143696  1.380781377  0.436444485  0.383896444  0.283787425  0.248913822  0.221144342
[11]  0.194408237  0.163678330  0.160161740  0.129578284  0.109562102  0.097180286  0.096227265  0.084840668  0.073647106  0.059862166
[21]  0.054366507  0.053164687  0.044522776  0.031987066  0.029920473  0.027513132  0.021419815  0.018076721  0.013887088  0.013084398
[31]  0.009518642  0.008018388  0.005712758  0.004189485
```

Nota. Tabla elaborada mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 5

Gráfico de sedimentación



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 6

Resultados de Parallel Analysis ("Horn")

Component	Adjusted Eigenvalue	Unadjusted Eigenvalue	Estimated Bias
1	19.242889	20.922434	1.679544
2	2.945676	4.371873	1.426197
3	1.338095	2.575052	1.236957

Nota. Tabla elaborada mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 8

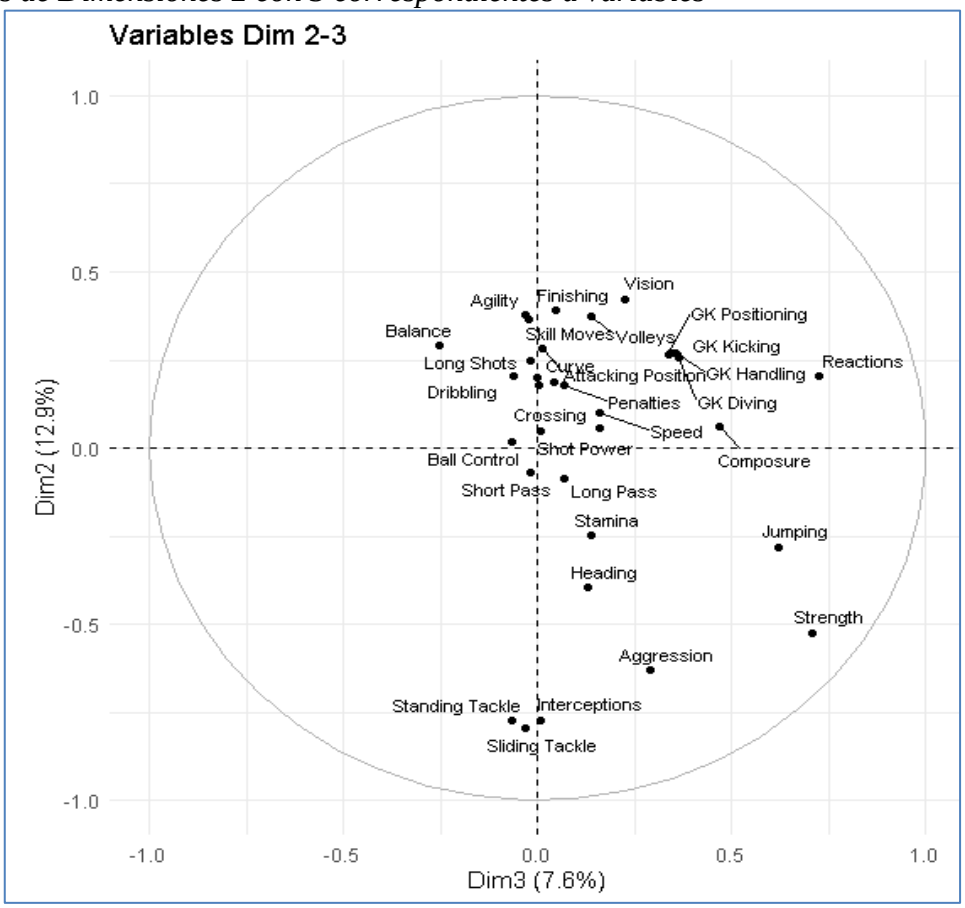
Agrupación de variables según su componente principal

Delanteros Mediocampistas ofensivos	Defensores Mediocampistas defensivos	Arqueros
PC1	PC2	PC3
Ball Control	Sliding Tackle	Reactions
Short Pass	Standing Tackle	Strength
Dribbling	Interceptions	Jumping
Curve	Aggression	Composure
Attacking Position	Strength	GK Diving
Crossing	Heading	GK Handling
Penalties	Jumping	GK Kicking
Long Shots	Stamina	GK Positioning
Freekick Accuracy	Long Pass	GK Reflexes
Shot Power	Short Pass	Aggression
Finishing	Ball Control	Vision

Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 11

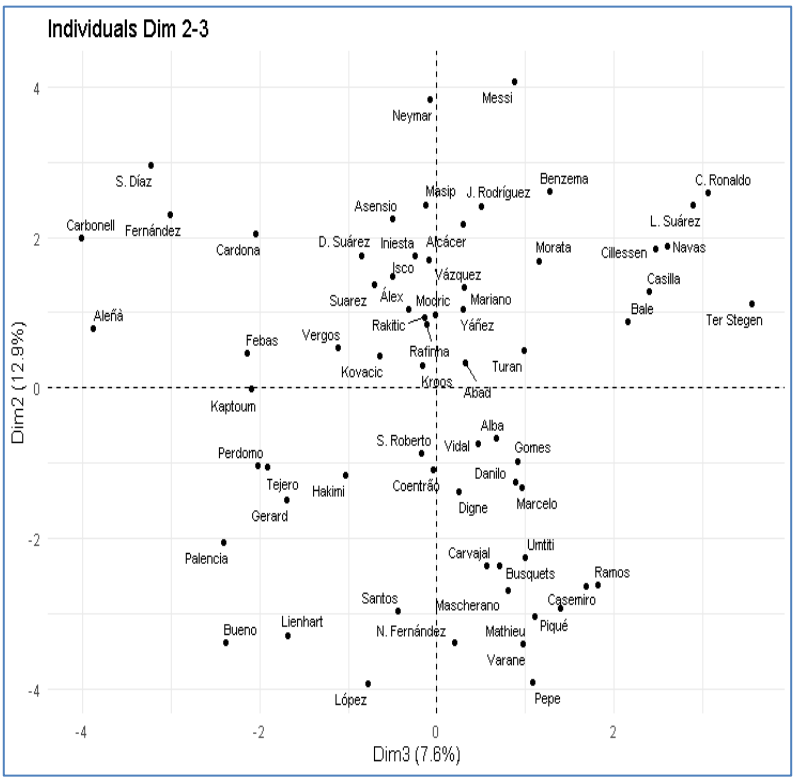
Gráfico de las cargas factoriales como vectores en sus correspondientes coordenadas para los pares de Dimensiones 2 con 3 correspondientes a variables



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 12

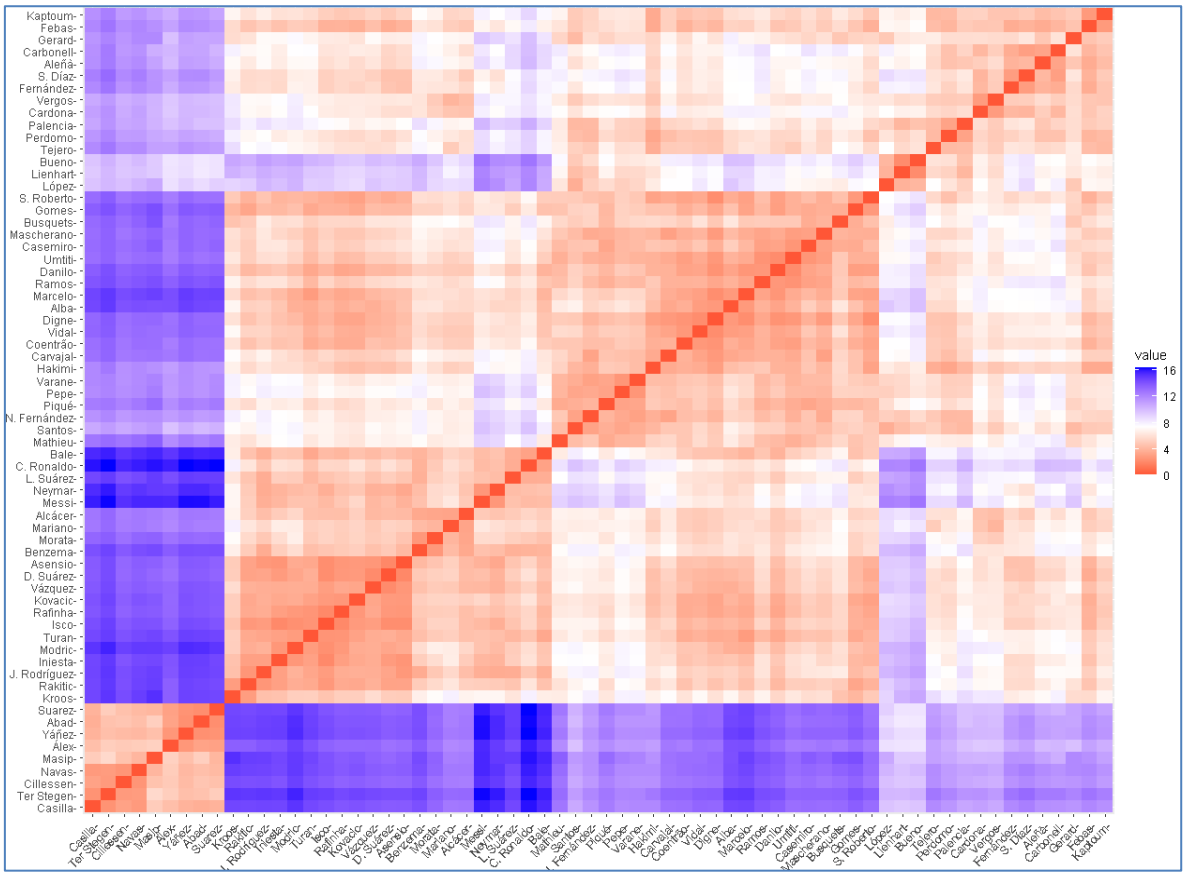
Gráfico de las cargas factoriales como vectores en sus correspondientes coordenadas para los pares de Dimensiones 2 con 3 correspondientes a individuos



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 13

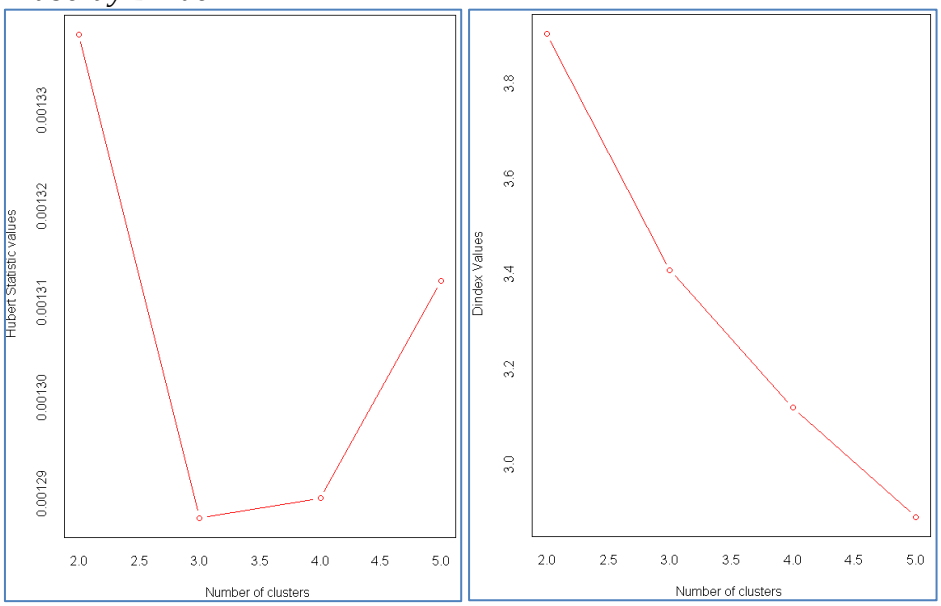
Matriz de distancias euclídeas



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021)

Figura 15

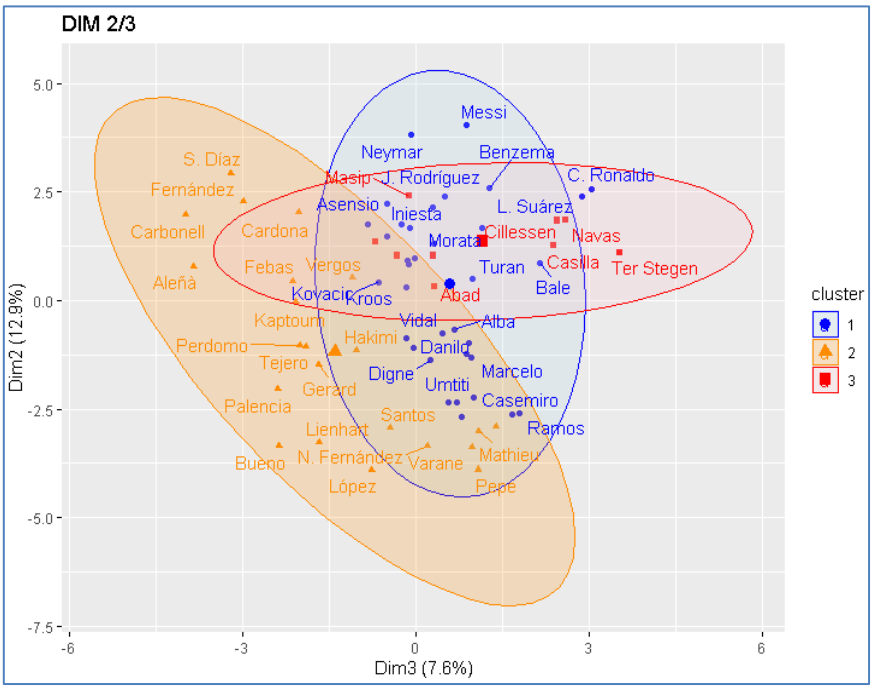
Estadísticos de Hubert y Dindex



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021)

Figura 18

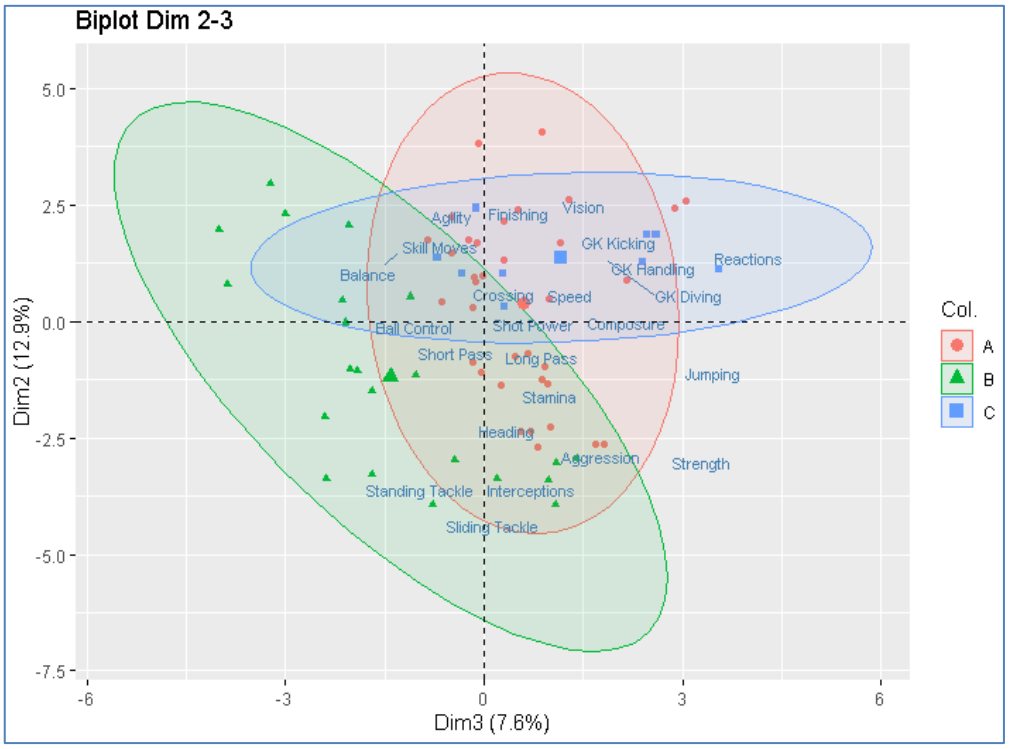
Conformación de clústeres según las dimensiones 2 y 3



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 20

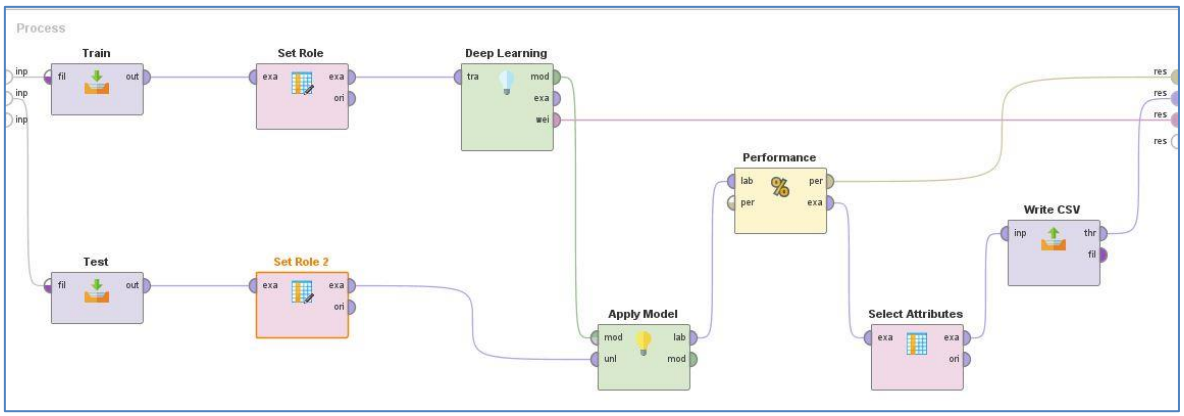
"Biplot" según dimensiones 2 y 3.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (PBC, 2021*)*

Figura 23

Modelo predictivo "Deep Learning"



Nota. Gráfico elaborado mediante la utilización del software Rapidminer (Rapidminer, 2022*)*

Anexo I

Traducción de aptitudes cuantitativas

- Pace: Velocidad
- Shooting: Disparo
- Passing: Pase
- Dribbling: Regate
- Defending: Defensa
- Physical: Físico
- Goalkeeping Diving: Paradas
- Goalkeeping Handling: Trato con el balón
- Goalkeeping Kicking: Despeje
- Goalkeeping Positioning: Colocación
- Goalkeeping Reflexes: Reflejos
- Goalkeeping Speed: Velocidad
- Attacking Crossing: Centros
- Attacking Finishing: Remate
- Attacking Heading Accuracy: Precisión en el juego aéreo
- Attacking Short Passing: Pases cortos
- Attacking Volleys: Voleas
- Defending Marking Awareness: Percepción del marcaje
- Defending Sliding Tackle: Entradas
- Defending Standing Tackle: Tackles
- Skill Ball Control: Control de balón
- Skill Curve: Efecto
- Skill Dribbling: Regate
- Skill Fk Accuracy: Precisión en faltas
- Skill Long Passing: Pases largos
- Mentality Aggression: Agresividad
- Mentality Composure: Compostura
- Mentality Interceptions: Intercepciones
- Mentality Penalties: Penales
- Mentality Positioning: Posicionamiento
- Mentality Vision: Visión
- Power Jumping: Salto
- Power Long Shots: Remates de larga distancia
- Power Shot Power: Potencia de disparo
- Power Stamina: Resistencia
- Power Strength: Fuerza
- Movement Acceleration: Aceleración
- Movement Agility: Agilidad
- Movement Balance: Equilibrio

- Movement Reactions: Reacciones
- Movement Sprint Speed: Velocidad en carrera

Anexo II
Traducción de aptitudes cualitativas

- Age: Edad Club
- Position: Posición en el club
- Club Team: Equipo del club
- ID: ID
- Height Cm: Altura en Cm
- International Reputation: Reputación internacional
- League Level: Nivel de liga
- League Name: Nombre de la liga
- Nationality: Nacionalidad
- Preferred Foot: Pie preferido
- Release Clause Eur: Cláusula de liberación en EUR
- Short Name: Nombre corto
- Skill Moves: Habilidad con el balón
- Value Eur: Valor en EUR
- Wage Eur: Salario en EUR
- Weak Foot: Pie débil
- Weight Kg: Peso en kg

Anexo III
Reporte del mentor

Este trabajo plantea el problema conocer las principales variables que se utilizan para predecir el rating global de cada jugador del videojuego FIFA. Se utilizan técnicas de análisis multivariado y aprendizaje automático. El problema es de interés para los usuarios del videojuego y contribuye al desarrollo del negocio de los creadores de videojuegos con características similares.
En cuanto al planteo del problema, el mismo se encuentra correctamente definido. En la introducción se realiza una descripción de la problemática de los video juegos y del modelo de negocio que se considera en este desarrollo.
El objetivo general del trabajo es analizar la variación de las variables en todas las versiones del videojuego y cómo afectan al rating global de cada jugador para predecir el modo en que dicha métrica se comportará en el siguiente periodo.
Las hipótesis son coherentes con el planteo del problema y los objetivos que se pretenden alcanzar.
El planteo del problema, los objetivos y las hipótesis se encuentran articulados, presentan coherencia interna y corresponden con los contenidos de la especialización que está realizando.

El tema elegido resulta interesante y novedoso para aplicar como trabajo final de la especialización.
El trabajo se realiza con datos reales y actualizados que corresponden a una base de datos abierta.
En el desarrollo del trabajo se observa la fundamentación del problema, el procesamiento de datos, la aplicación de las diferentes metodologías, los resultados obtenidos, la conclusión articulada con el objetivo del trabajo y la bibliografía actualizada.
Se presentan los resultados en forma de tabla y con gráficos que favorecen la interpretación.
En cuanto a los métodos multivariados se utilizó componentes principales y clustering. De aprendizaje automático utilizó árboles de decisión, bosques aleatorios, árboles de clasificación y regresión y redes neuronales.
Para la realización de este trabajo se realizó una etapa de preprocesamiento de datos que incluye el proceso de limpieza, selección y transformación de atributos para poder aplicar los métodos de aprendizaje automático. Se presenta un detalle de todos los procedimientos realizados en el trabajo que son consistentes con los contenidos académicos desarrollados en la especialización. Asimismo, se presenta la aplicación de las metodologías abordadas en las diferentes asignaturas utilizando herramientas informáticas adecuadas a cada tema.
Se considera que se han alcanzado los objetivos propuestos y se presentan los resultados correspondientes acompañados de una fundamentación adecuada y de la bibliografía correspondiente.

Mentora: Nélida Mónica Cantoni Rabolini