PLANTEAMIENTO DE MODELOS DE REGRESIÓN PARA FACILITAR Y MEJORAR LA TOMA DE DECISIONES DE LOS DIRECTIVOS Y CUERPO TÉCNICO EN UN EQUIPO DE FÚTBOL

JUAN PABLO MARTÍNEZ ARANGO
SANTIAGO LONDOÑO GARZÓN

UNIVERSIDAD ICESI
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA INDUSTRIAL
CALI
2019

Planteamiento de modelos de regresión para facilitar y mejorar la toma de decisiones de los directivos y cuerpo técnico en un equipo de fútbol

Juan Pablo Martínez Arango
Santiago Londoño Garzón

Proyecto de Grado para optar el título de Ingeniero Industrial
Director proyecto
Rolando Acosta Amado PH.D

UNIVERSIDAD ICESI
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA INDUSTRIAL
CALI
2019

Contenido

pág.

1. GLOSARIO 7
1.1 Abstract 8
1.2 RESUMEN 9
1 Introducción 10
1.1 Contexto 10
2 Objetivos 15
2.1 Objetivo del Proyecto 15
3 Marco de Referencia 16
3.1 Antecedentes o Estudios Previos 16
3.1.1 Artículo “Fútbol y Big data” [4] 16
3.1.2 “Valoración de futbolistas profesionales mediante un modelo AHP”. . 19
3.2 Marco Teórico 21
3.3 Contribución Intelectual o Impacto del Proyecto 29
4 Metodología 31
5. Resultados 39
6. Conclusiones 48
7. Recomendaciones 50
Bibliografía 52

Lista de Figuras

Figura 3.2: Tipos de variables estadísticas……..……....24

Figura 4.0: Fases de CRIP-DM…….……..32

Figura 4.1.2: Visualización de variables……....33

Figura 5.1.1: Características de los jugadores…….……....39

Figura 5.1.3: Coeficientes de variación……..….41

Figura 5.2.1: Características Modelo 2……..41

Figura 5.2.2: R2 ajustado Acumulado…...…….……..43

Figura 5.2.3: Coeficientes de regresión modelo…….....45

Figura 5.3.1: Características Modelo 3…..…..…….……..….42

Figura 5.3.5 Coeficientes e intercepto segundo modelo…….……..47

Lista de Tablas

Tabla 1.1.A:Jugadores más valiosos ……..……..11

Tabla 1.1.B: Top 20 most valuables soccer teams……..13

Tabla 3.2: Traspasos más costosos del futbol 2019……..…….24

Tabla 5.1.2: Modelos y métricas de regresión…….….40

Tabla 5.2.2: Modelo y métricas del modelo 2……..…….43

Tabla 5.2.3: Lista de las variables más influyentes…….44

Tabla 5.2.3: Tabla de Ratio por jugador…..……..……...45

Tabla 5.3.3: Modelos y métricas del tercer modelo…….47

Lista de Anexos

Anexo 1. Reporte de Cambios y Ajustes 22

Anexo 2. Instrucciones de Entrega 23

1. GLOSARIO

- **Dataset:** Conjunto de datos históricos relativos a un problema o una situación.
- **Overall:** Calificación general de un jugador de fútbol que indica el nivel futbolístico del mismo.
- **Value:** Valor de mercado en unidades monetarias del jugador del fútbol.
- **Big Data:** Conjunto de datos de gran volumen y velocidad.
- **Ratio:** Relación cuantificada entre dos magnitudes que refleja su proporción.

1.1 Abstract

In this work predictive models were made and correlations were found to help decision makers and technical directors in the soccer field. In order to do that we used a database of the FIFA 19 game, it was filtered and through a tool called Jupyter we were able to execute and train the models with the Machine Learning methodology. The results shows three successful predictive models and the most influential variables when determining the goals scored, the Overall and the Value of a soccer player. Therefore, it is concluded that if a coaching staff uses the aforementioned methods, they will have the capacity to determine the Ratio, Overall and the general value of the observed player. All this would help decision makers to take the right choice when they face a crucial decision.

1.2 RESUMEN

En este trabajo se realizaron tres modelos predictivos basados en modelos de regresión con la finalidad de mejorar la toma de decisiones a los directivos y directores técnicos, para esto se utilizó el dataset del juego FIFA 19, el cual se filtró y se le hizo la limpieza de datos necesaria a través de una herramienta de ejecución de código online llamada Jupyter para poder ejecutar y entrenar a los modelos. Los resultados muestran tres modelos predictivos exitosos, y las variables más influyentes a la hora de determinar los goles anotados, el Overall y el Value de un jugador de fútbol. Por lo tanto, se concluye que si un cuerpo técnico utiliza los métodos anteriormente mencionados, está en la capacidad de determinar el Ratio, el Overall y el valor general del jugador observado, lo que le ayudará a la hora de tener que tomar una decisión crucial.

1 Introducción

1.1 Contexto

El fútbol es el deporte más popular del mundo y tuvo sus inicios en 1863 en Inglaterra, el cual consiste en un juego de conjunto donde dos equipos de 11 jugadores cada uno, se enfrentan en busca de la victoria. Este deporte es ampliamente conocido y se estima que unos 270 millones de personas alrededor del mundo lo practican. En la actualidad este deporte está organizado de una manera muy estructurada y casi todos los países del mundo cuentan con un órgano oficial encargado de controlar este juego a nivel profesional y estos a su vez son regidos por la máxima autoridad llamada FIFA.

Todo esto hace que el fútbol sea un negocio muy atractivo para inversionistas y para los países con ligas profesionales muy conocidas, tal es su magnitud que el fútbol por si solo podría ser considerado la decimoséptima economía mundial y que además genera más de 140.000 empleos de jornada completa en el mundo. Para tener un ejemplo claro nos podemos ir a España, donde se disputa la liga más importante del mundo llamada “La liga Santander” que produjo en el 2013 un aporte de 3.000 millones de euros de contribución al PIB. Para ponerlo en contexto, de cada 100 euros que genera la economía española, 2 euros son generados por “La liga” y esto no hace sino aumentar con el paso del tiempo, hasta tal punto que una venta de un jugador de fútbol del F.C Barcelona llego a los 222 millones de euros.

Lista de los jugadores más costosos de Europa (2018)	
Jugador	**Valor (Mill €)**
Neymar Jr	$213,0
Lionel Messi	$202,2
Harry Kane	$194,7
Kylian Mbappe	$192,5
Paulo Dybala	$174,6
Dele Alli	$171,3
Kevin De Bruyne	$167,8
Romelu Lukaku	$164,8
Antoine Griezmann	$150,2
Paul Pogba	$147,5

***Tabla 1.1.A**: Jugadores más valiosos 2018 (Elaboración propia)*

Desde hace muy poco tiempo se ha logrado evidenciar un impacto significativo de la tecnología en el deporte del fútbol y aún hay muchos entrenadores que desconocen qué tipos de herramientas pueden facilitar su trabajo, y aunque hay ejemplos en otros deportes donde los equipos se han apoyado en las matemáticas y las estadísticas para ganar campeonatos, en esta disciplina aún no se ha generalizado su uso en las decisiones críticas. Por esta razón se quiere implementar una herramienta al alcance de los directores técnicos y ayudarlos a la hora de preparar sus partidos para que así logren triunfar y ganar campeonatos, respaldándose y apoyándose en las matemáticas para mejorar la toma de decisiones cruciales en su preparación para afrontar las nuevas temporadas.

Que un club sea exitoso, gane títulos y partidos, no solo le va a generar un prestigio en el mundo del deporte, sino que también les va a ayudar a generar unos ingresos muchos mayores y a participar en torneos exclusivos donde su solo participación les asegura millones de dólares. Alcanzar estos objetivos es solo el comienzo, ya que, al tener mayores ingresos, también pueden adquirir mejores jugadores y esto hace que sea más factible seguir ganando títulos. ¿Pero cómo lograrlo? Las instrucciones son sencillas, ganando los partidos que se le presenten a lo largo de

la temporada. Para esto hay un pilar fundamental dentro del club llamado director técnico que tiene a su disposición los jugadores de fútbol del equipo con los cuales entrena toda la semana, y que se encarga de seleccionar a la hora de los partidos oficiales, además de la táctica de juego y los cambios respectivos que se hacen según avance el juego. Esto lo hace basado en sus conocimientos y en lo visto durante toda la semana y partidos pasados, apoyándose en sus colaboradores. En la actualidad con los grandes avances de la tecnología se pueden recoger innumerables datos sobre cada jugador de fútbol, y de esta manera tener una estadística sobre todas las variables de juego, abriéndonos a un panorama nunca antes utilizado en el fútbol para la toma de decisiones cruciales centrándonos en las matemáticas. Lo que se sugiere es que, con todas estas estadísticas y datos obtenidos, se pueda crear un modelo que ayude a los entrenadores y directivos a tomar las mejores decisiones frente a las nuevas contrataciones y habilidades que necesitan para marcar goles y ganar partidos.

Top 20 de los equipos de fútbol con mayor valor (2016)	
Equipos de fútbol	**Total Valuación (En U.S dólares)**
Real Madrid	$ 3,75 B
Barcelona	$ 3,50 B
Manchester United	$ 3,45 B
Bayern Munich	$ 2,85 B
Arsenal	$ 2,00 B
Manchester City	$ 1,90 B
Chelsea	$ 1,75 B
Liverpool	$ 1,60 B
Juventus	$ 1,40 B
Tottenham	$ 1,00 B
Borussia Dortmund	$ 800 M
Ac Milan	$ 780 M
Paris Saint-Germain	$ 760 M
Schalke 04	$690 M
Atlético de Madrid	$670 M
Inter de -Milán	$ 550 M

***Tabla 1.1.B**: Top 20 Most valuable soccer teams (2016) [1]*

Formulación de la pregunta de investigación u objeto de estudio

¿Existe la posibilidad desarrollar modelos predictivos que ayuden a los directivos y directores técnicos a mejorar la toma de decisiones en un equipo de fútbol?

Justificación o Importancia de la situación objeto de estudio

El poder generar un modelo matemático capaz de ayudar a la toma de decisiones de un director técnico o un directivo, sería innovador en este deporte y un reto bastante grande debido a la gran cantidad de variables que influyen en el fútbol. Las capacidades futbolísticas de un jugador, su proyección, su historia, su estado físico actual, su edad entre muchas otras hacen que a la hora de escoger una formación esto se vuelva muy complejo. Desde el punto de vista ingeniería industrial siempre se busca poder optimizar los procesos que ocurren dentro de una industria,

concepto que queremos aplicar en la industria del fútbol, puesto que el objetivo es determinar todas esas variables que son importantes dentro de un partido y estipular cuales de esas variables son las más importantes de acuerdo a lo que el club está buscando. Poder realizar este modelo no solo pretende cambiar los resultados futbolísticos de un equipo de futbol, sino que también va a marcar sus finanzas y si se ve desde un punto de vista macro, va a poder mejorar la liga en la que juegan y esto a su vez va a poder mejorar la economía del país donde se aplique, al tener equipos con un alto rendimiento en las competiciones internacionales.

2 Objetivos

2.1 Objetivo del Proyecto

Plantear modelos predictivos basado en modelos de regresión, que ayuden a los directivos y cuerpo técnico a tomar mejores decisiones al momento de planificar una temporada o un ciclo.

Objetivos Específicos:

- Crear un modelo capaz de predecir el *Ratio* por jugador (habilidades/costo) para seleccionar a los jugadores más rentables, Identificando a su vez las variables del jugador que más se correlacionan con el *Overall*.
- Crear un modelo capaz de predecir el valor de traspaso de un futbolista, identificando a su vez las variables que más incrementan el *Value* del jugador.
- Crear un modelo capaz de predecir los goles que realizará un jugador en una temporada, identificando a su vez las variables en un jugador que tienen mayor correlación con el número de goles anotados.

Entregables:

- Lista de las variables que tienen mayor correlación con los goles anotados.
- Lista de las variables que incrementan el valor de traspaso de un futbolista.
- Lista de los jugadores con mejor *Ratio*.
- Lista de las variables más influyentes en la predicción del *Overall*.
- Modelo de regresión que predice el *Overall*.
- Modelo de regresión que predice el *Value*.
- Modelo de regresión que predice la cantidad de goles anotados.

3 Marco de Referencia

3.1 Antecedentes o Estudios Previos

3.1.1 Artículo “Fútbol y Big data” [4]

Parte difícil de poder realizar análisis matemático en el fútbol, es poder recolectar de una manera correcta y completa, todos esos distintos datos que nos puede arrojar un partido de fútbol, debido a su duración, la cantidad de jugadores, la cantidad inmensa de distintos tipos de datos, etc. Por esta razón es que hasta hace muy poco tiempo no se tenían datos precisos para realizar modelos matemáticos, ya que gracias a diferentes artefactos tecnológicos se ha logrado recolectar estos datos de manera correcta.

Para poder optar por realizar una investigación en el campo donde el fútbol y las matemáticas se relacionan es necesario saber si ya hay posibilidad de acceder a los datos requeridos para poder hacer un análisis bien fundamentado. Por tal motivo es importante tener en cuenta el artículo [4] que nos cuenta todo ese proceso por el cual ha tenido que pasar el fútbol, para poder procesar toda esa cantidad de datos con extrema exactitud.

El artículo que tiene como eje principal la tecnología, habla sobre como a través de una buena cantidad de cámaras situadas en el campo de juego y una herramienta que digiere los datos captados (Media Coach), se ha logrado digerir toda la información y organizarla de tal forma que esta sea fácil de interpretar y analizar. Utilizando esta herramienta es posible entender cómo se asocian los jugadores, sus movimientos, sus regates, su forma de esprintar, hacia donde se desvían los tiros del delantero, entre muchísimos otros factores que pueden ayudar a mejorar el rendimiento de un equipo. Todo esto sin dejar de tener en cuenta que en este juego

también existe azar y caos, pero a este último es a lo que más apuntan, ya que este tiene un orden y puede predecirse con algún grado de incertidumbre.

Con todo lo anterior se puede deducir que en este momento ya se cuentan con base de datos con la exactitud y organización suficiente para ser utilizados en funciones matemáticas, con el objetivo de intentar predecir con cierto grado de incertidumbre el resultado de un partido de fútbol.

3.1.2 Proyecto de investigación “Sports Predictive Analytics” [5]

La estadística en el fútbol existe desde hace ya unos años, pero hace muy poco tiempo es que se ha comenzado a utilizar para el análisis y el entendimiento del rendimiento de los equipos y los jugadores. Esto se debe a que la estadística en este deporte va de la mano de la tecnología, y como este último ha estado creciendo de manera exponencial, ha facilitado la recolección e interpretación de los datos. Pero el fútbol no es el único deporte que se ha beneficiado de las matemáticas, inclusive hay otros deportes donde la estadística ya ha marcado gran diferencia en los resultados.

Existe un proyecto de investigación [5] donde muestran todo lo logrado por las matemáticas, en dos deportes en específico: NFL y Béisbol. En el proyecto de investigación, se hace referencia a que las tomas de decisiones de un equipo deportivo deberían estar sujetas a la recolección e interpretación de datos del deporte y del rendimiento de los jugadores, a través de los diferentes modelos matemáticos que explican dichos datos.

En el texto se intenta encontrar cuales son las variables que mejor predicen el porcentaje de victorias de un equipo de béisbol, centrando su búsqueda en carreras, homeruns, equipo de bateo, en base, promedio de bateo y carreras ganadas. Todo esto a través de modelos predictivos como lo son la regresión lineal simple, la regresión lineal múltiple, la regresión polinomio y la regresión logística que tienen

como punto de apoyo la distribución de frecuencias, medias, medianas, y modas para su realización.

Los objetivos de la analítica deportiva en este trabajo son:

- ✓ Aplicar modelos estadísticos a los datos deportivos.
- ✓ Realizar ratings y rankings.
- ✓ Crear modelos predictivos.
- ✓ Aplicar los modelos para generar una valoración de los jugadores y equipos.

También se habla de unos objetivos a la hora de analizar a los jugadores:

- ✓ Determinar el talento oculto.
- ✓ Realizar una evaluación general.
- ✓ Realizar una evaluación del rendimiento.
- ✓ Determinar el valor agregado que genera al equipo.

Y por último se plantean unos objetivos cuando se analiza a un equipo:

- ✓ Determinar el ranking del equipo.
- ✓ Evaluar el rendimiento del equipo.
- ✓ Seleccionar los jugadores mejor situados para jugar contra un rival.
- ✓ Determinar la estrategia a usar contra el otro equipo.
- ✓ Poder generar cual va a ser la probabilidad de ganar.

Inclusive en el proyecto de investigación se hace referencia a una película titulada “Moneyball” donde su actor principal es Brad Pitt y se muestra una historia real, en la cual el jefe de un equipo de béisbol de las grandes ligas y un analista de datos, logran realizar una de las mayores hazañas en la historia de este deporte al romper el récord histórico de victorias consecutivas (20), con uno de los equipos de más bajo presupuesto en la liga y todo esto gracias a un modelo matemático hecho por el analista, que buscaba poder fichar a jugadores baratos, que habían sido

ignorados por otros equipos y que se acoplaban perfectamente a sus parámetros de búsqueda.

Todo lo anterior abre una puerta inmensa a la posibilidad de poder crear un modelo parecido al utilizado en béisbol, pero esta vez en el fútbol, guardando la diferencia entre deportes y entendiendo que este proyecto se enfoca en encontrar las variables que más influyan en un partido, y que el criterio de elección y compra de un jugador costoso o económico queda a criterio del director técnico y presidente, aunque nuestros resultados y análisis estén enfocados en parte al aspecto monetario.

3.1.2 "Valoración de futbolistas profesionales mediante un modelo AHP" [7]

Una de las partes fundamentales de [7] es encontrar una lista de variables que incrementen la probabilidad de ganar un partido de fútbol. Para esto se puede revisar el proyecto "Valoración de futbolistas profesionales mediante un modelo AHP" realizado por un profesor de la universidad de Santander UDES.

En este trabajo comparan a los clubes deportivos como empresas futbolísticas que tienen como uno de sus objetivos principales la sostenibilidad financiera, y por esta razón orientan los conocimientos adquiridos a determinar el precio de un futbolista para evitar inversiones sobrevaloradas. Para esto el autor tiene dos grandes grupos de variables: Variables deportivas y variables personales.

Las variables deportivas hacen referencia a: (ejemplo)

- ✓ Asistencias
- ✓ Goles
- ✓ Cortes de balón
- ✓ Posicionamiento táctico

Mientras que las variables personales hacen referencia a: (ejemplo)

- ✓ Edad
- ✓ Disciplina

***** Estas variables están sujetas a cambios, ya sea que se agregue o varíen según los criterios que se quieran evaluar.

Luego de tener claro las variables, el siguiente paso es realizar una ponderación de todas estas, con el fin de poder generar una comparación con otro jugador que tenga un puntaje parecido al futbolista que estamos comparando, para poder asignarle un precio similar.

(Memmert, 2018) “Los pasos que nombran en el documento para la asignación de las habilidades y poder estipular un precio son las siguientes:

1) Selección de expertos.
2) Asignar jugadores comparables.
3) Determinar criterios a evaluar.
4) Aplicación de encuesta al público referente a las habilidades del jugador.
5) Formulación y resolución de matrices ponderadas.
6) Agregación de los vectores obtenidos por cada experto.
7) Ponderación final de los criterios.
8) Vectores propios de las variables de liderazgo y merchandising.
9) Recopilación información de internet.

Y todo esto conlleva a las siguientes estimaciones:

- ✓ Multiplicación de matrices
- ✓ Ponderación del jugador
- ✓ Cálculo de ratio de valoración
- ✓ Valor del jugador”

Con todo esto después de generar la lista de variables que van a tener en cuenta a la hora analizar el precio de un jugador, se obtiene como indica la última estimación, el valor del jugador.

Es por esto que este documento es relevante para el proyecto, puesto que, a través de unas variables y su ponderación, logran determinar el precio justo de un jugador de fútbol, planteamiento parecido al que busca llegar este proyecto de grado, el cual es a través de unas variables inherentes a los jugadores, equipos de fútbol, y una ponderación de los distintos valores, determinar la relación que existe entre las variables y los resultados de los partidos.

3.2 Marco Teórico

3.2.1 Fútbol

Para entender la razón de este proyecto, es fundamental saber que es el fútbol y el interés que genera. Para comenzar, su definición es la siguiente según la RAE: (RAE, 2001).

“Juego entre dos equipos de once jugadores cada uno, cuyo objetivo es hacer entrar en la portería contraria un balón que no puede ser tocado con las manos ni con los brazos, salvo por el portero en su área de meta”.

Además de esto es necesario entender que un equipo no es campeón ganando un solo partido. En el futbol profesional de las grandes ligas del mundo, un equipo es campeón cuando suma más puntos que su rival luego de jugar un promedio de 36 partidos a lo largo del año, donde todos los equipos de la primera división del país se enfrentan en un todo contra todos, para determinar el equipo más regular a lo largo de la temporada. También están las copas de la liga del país, donde compiten los equipos de primera y segunda división, y normalmente se realizan en formatos de eliminación directa, en el cual el equipo campeón juega en promedio 12 partidos, incluyendo semifinales y finales. Por último, también existen competiciones

internacionales (UEFA Champions league, Europa league, Copa Sudamericana, Copa Libertadores etc.) Donde los mejores equipos del continente se enfrentan para determinar el mejor equipo de cada región, y que por lo general tienen un formato mixto donde inician con una fase de grupos y terminan con partidos de eliminación directa con ida y vuelta en los respectivos estadios de los equipos enfrentados, donde el campeón tendrá que jugar en promedio 14 partidos.

3.2.1.1 Mercado de Transferencias en el fútbol

El mercado de transferencias es el periodo del año donde los equipos profesionales de fútbol negocian el adquirir vender un jugador profesional por cierta cantidad de dinero, ya sea con el equipo poseedor del pase del jugador, o con este mismo si no tiene un contrato vigente con ningún equipo. Este periodo de transferencias es regulado por el máximo organismo de este deporte (FIFA) y está restringido a dos periodos del año denominados verano (9 de Junio -31 Agosto) e invierno (1 Enero – 31 Enero).

Dentro del mercado de transferencias los clubes compran los derechos deportivos de los jugadores por cierta cantidad de dinero y firman un contrato para vincularlo oficialmente al club. Los pasos estipulados para contratar a un jugador son los siguientes:

- ✓ Conocer al jugador: Haber realizado un seguimiento previo al jugador para entender que puede aportar, y en qué tipo de situación se encuentra dentro del equipo (Estrella, titular, suplente, etc.).
- ✓ Acercamiento con el agente: Como regla de la organización FIFA, los clubes no tienen permitidos acercarse a los jugadores sin permiso del club, es por este motivo que lo primero que tiene que hacer un equipo es entablar conversaciones con el agente que actúa como intermediario entre el club, el jugador y la parte interesada, esto con el fin de saber si el jugador tiene deseos de salir y posteriormente contactar el club para hablar del negocio.

- ✓ Negociación: Después de contactar el agente y tener la información clara, se procede a negociar con el club de manera oficial la adquisición del jugador. Para esto se debe tener en cuenta la valoración del jugador por parte del club dueño de su pase, la duración del contrato y su cláusula de recisión para determinar su precio final.
- ✓ Verificación de documentos vía TMS: Una vez se haya cerrado el acuerdo entre los dos clubes y el jugador, se procederán a enviar los documentos de trabajo del jugador vía Transfer Matching Sistem a la FIFA, para que puedan comprobar que sus documentos están en regla.
- ✓ Evaluación médica: Al cerrar el acuerdo, el club comprador deberá someter al jugador a una prueba médica, para determinar que el jugador no tiene ninguna condición especial que le impida jugar ya sea por un periodo de tiempo o de forma permanente.
- ✓ Firma del contrato: Realizado el examen médico y con las negociaciones ya finalizadas, el jugador firma con su nuevo equipo y el club comprador está en la obligación de inscribirlo en las competiciones en las cuales van a participar.

Fichajes hasta 2019						
Ranking	**Futbolista**	**Origen**	**Destino**	**Posición**	**Fichaje (M €)**	**Año**
1	**Neymar**	FC Barcelona	Paris Saint-Garmain	Delantero	222	2017
2	**Mbappé**	AS Monaco	Paris Saint-Garmain	Delantero	145	2018
3	**F. Coutinho**	Liverpool FC	FC Barcelona	Centrocampista	120	2018
4	**C. Ronaldo**	Real Madrid	Juventus	Delantero	117	2018
5	**O. Dembélé**	Borussia Dortmund	FC Barcelona	Delantero	105	2017
6	**P. Pogba**	Juventus	Manchester United	Centrocampista	105	2016
7	**G. Bale**	Tottenham FC	Real Madrid	Delantero	100,8	2013
8	**C. Ronaldo**	Manchester United	Real Madrid	Delantero	94,5	2009
9	**G. Higuain**	Napoli	Juventus	Delantero	90	2016
10	**Neymar**	Santos	FC Barcelona	Delantero	88,2	2013

Tabla 3.2*: Traspasos más costosos del fútbol 2019, (elaboración propia)*

3.2.2 Variable Estadística

¿Qué se entiende por variables y a que nos referimos con este término? La definición dice así: (Enciclopedia Económica, 2017)

“La variable estadística se refiere a una característica o cualidad de un individuo que está propenso a adquirir diferentes valores. Estos valores se caracterizan por poder medirse. Por ejemplo, el color de pelo de una persona, las notas de un examen, sexo, estatura de una persona, etc.”

De lo anterior entendemos a que todas las variables trabajadas dentro de este documento estarán ligadas a los jugadores de fútbol o a un club profesional

3.2.2.1 Tipos de variable estadística

(Wiston, 2009) Los tipos de variables se dividen en dos grupos: cualitativas y cuantitativas de acuerdo con las características que la definen. Y estas a su vez se dividen en otros subgrupos:

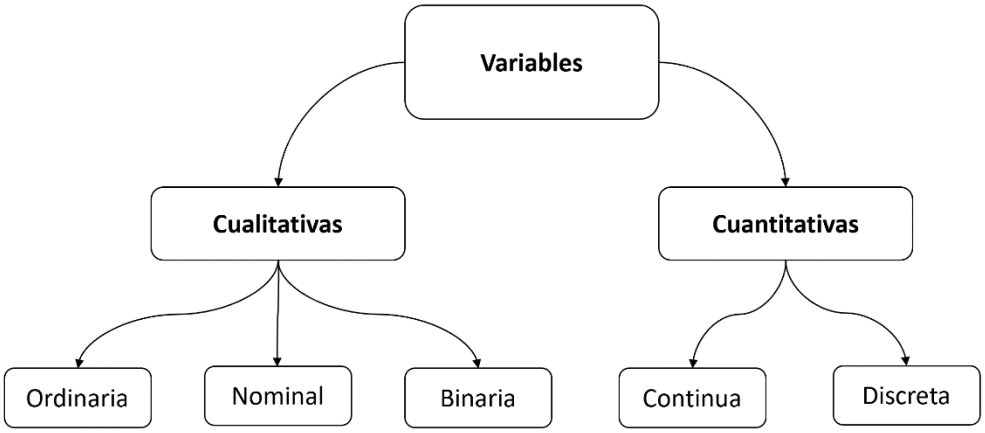

***Figura 3.2**: Tipos de variables estadísticas [2]*

3.2.3 Modelos matemáticos

¿Cuál es la definición de un modelo matemático y para qué sirve? (Roldan, 2018)

“Un modelo matemático es una representación simplificada, a través de ecuaciones, funciones o fórmulas matemáticas, de un fenómeno o de la relación entre dos o más variables. La rama de las matemáticas que se encarga de estudiar las cualidades y estructura de los modelos es la llamada “teoría de los modelos.

Los modelos matemáticos son utilizados para analizar la relación entre dos o más variables. Pueden ser utilizados para entender fenómenos naturales, sociales, físicos, etc. Dependiendo del objetivo buscado y del diseño del mismo modelo pueden servir para predecir el valor de las variables en el futuro, hacer hipótesis, evaluar los efectos de una determinada política o actividad, entre otros objetivos.

3.2.3.1 Regresión lineal simple

Dentro del proyecto se intentará relacionar distintos tipos de variables de los equipos y jugadores de fútbol, con los resultados de un partido. Para esto se usará la regresión lineal simple, que tiene como definición: (EcuRed, 2016)

“La regresión lineal simple se basa en estudiar los cambios en una variable, no aleatoria, afectan a una variable aleatoria, en el caso de existir una relación funcional entre ambas variables que puede ser establecida por una expresión lineal, es decir, su representación gráfica es una línea recta. Es decir, se está en presencia de una regresión lineal simple cuando una variable independiente ejerce influencia sobre otra variable dependiente. Ejemplo: $\hat{y} = b_0 + b_1 x$ “

3.2.3.2 Regresión lineal múltiple

La regresión lineal múltiple es definida de la siguiente manera: (EcuRed, 2016)

“La regresión lineal permite trabajar con una variable a nivel de intervalo o razón, así también se puede comprender la relación de dos o más variables y permitirá relacionar mediante ecuaciones, una variable en relación con otras variables llamándose Regresión múltiple. O sea, la regresión lineal múltiple es cuando dos o más variables independientes influyen sobre una variable dependiente.

$$\hat{y} = f(x_1, x_2, \dots, x_n)$$

3.2.4 Machine Learning

Es una rama de la inteligencia artificial en la que se crean y desarrollan modelos a partir de la estadística y la programación para aprender de los datos, identificando patrones complejos entre los datos. La máquina aprende mediante algoritmos que exploran una gran cantidad de datos y es capaz de predecir comportamientos futuros de manera automática. Este sistema es capaz de mejorar de forma autónoma con el tiempo y sin necesidad de intervención humana, lo que hace que sea una herramienta muy interesante para la ayuda en la toma de decisiones.

3.2.4.1 Evolución de ML

Aunque ML no es una disciplina que se haya creado recientemente, gracias a las nuevas tecnologías, esta ha evolucionado drásticamente en los últimos años. Nació por la necesidad de encontrar herramientas que permitieran reconocer patrones automáticamente, es decir, computadoras que aprenden sin ser programadas para realizar tareas específicas, esto debido a la enorme complejidad de algunas tareas y actividades del mundo cotidiano para ser programadas manualmente. Ahora el ML es un campo muy desarrollado y promulgado, está al alcance de todos. Las compañías tecnológicas más grandes del mundo como IBM, Google y Microsoft han desarrollado herramientas y entornos para facilitar el aprendizaje y aplicación de estas técnicas tan potentes para tomar decisiones. En la actualidad se pueden ver desarrollos de ML presentes en la industria, para mencionar algunos ejemplos:

- Automóviles de conducción autónoma liderados por empresas como Google, Tesla y Uber.
- Recomendación en línea de series o películas – De la empresa Netflix
- Motor de Búsqueda de Google (Cadenas de Márkov).
- Sugerencias de productos de Amazon.

3.2.4.2 Importancia de ML

En los últimos años ha vuelto a salir a flote el interés por entender y procesar grandes volúmenes y variedades de datos, esto debido a la gran cantidad de información que se tiene a disposición, para contextualizar la gran cantidad de información que se produce, el director ejecutivo de Google, Eric Schmidt, afirmó que la Humanidad había creado hasta 2003 una cantidad equivalente a 5 exabytes, y añadió que ahora esta cifra se genera cada dos días. En la actualidad el procesamiento computacional es más económico y poderoso que en el pasado, haciendo que el almacenaje de datos sea más asequible; Con todo lo dicho anteriormente se puede determinar que a través de las computadoras se pueden producir modelos de manera rápida y automática que analizan datos cada vez más grandes y complejos, haciendo que los resultados sean más precisos con el pasar del tiempo y esto conlleva a que las organizaciones cada vez tengan mayor oportunidad de identificar oportunidades rentables en el mercado y de evitar riesgos en la mayoría de sus áreas.

3.2.4.3 Requerimientos para ML

Esta disciplina necesita de parámetros mínimos para funcionar de manera adecuada y correcta, para así poder arrojar resultados óptimos y recomendables. Estos parámetros son los siguientes:

- Recursos de preparación de datos

- Algoritmos – Básicos y Avanzados
- Automatización y procesos iterativos
- Escalabilidad
- Modelado en conjunto

3.2.4.4 Campos en los que se utiliza ML

Existen gran cantidad de áreas que han adoptado este método para mejorar las oportunidades de negocio y optimizar los procesos en cada una de sus especialidades, como es el caso de los siguientes sectores:

Los servicios financieros lo usan para:

- Prevenir los fraudes en transacciones
- Identificar oportunidades de inversión

El gobierno en su área de Seguridad pública y servicio público lo usa para:

- Encontrar formas de incrementar la eficiencia y ahorro de dinero
- Minimizar los robos de identidad

El sector de atención a la salud, que es un campo donde esta tecnología se ha incrementado de manera muy rápida, lo utiliza para:

- Analizar datos para evaluar la salud de los pacientes en tiempo real.
- Identificar diagnósticos y tratamientos mejorados.

En el área de Marketing y Ventas se usa para:

- Se usa para analizar el historial de compras de los clientes y promocionar los artículos según el perfil de cada individuo.
- Analizar datos para entender el comportamiento de las personas e implementar campañas publicitarias.

En el área del petróleo y gas se utiliza para:

- Encontrar nuevas fuentes de energía, a través de la predicción.
- Análisis de minerales en los suelos.
- Predicción de fallos en los sensores.
- Optimización de la distribución del petróleo.

3.2.4.5 Aprendizaje Supervisado

El aprendizaje supervisado, entrena al sistema a través de una gran cantidad de datos etiquetados que sirven o funcionan como entradas, y que son las encargadas de guiar al sistema para entender los patrones que se presentan la estructura de los datos.

3.2.4.6 Aprendizaje no Supervisado

Este tipo de aprendizaje utiliza datos no etiquetados, es decir que los algoritmos no tienen una “respuesta correcta” y este tiene que descubrirlo basándose en el objetivo de encontrar la estructura al interior de los millones de datos que está explorando.

3.3 Contribución Intelectual o Impacto del Proyecto

La contribución al análisis de partidos de fútbol es un desarrollo nuevo y en proceso, donde todavía queda mucho terreno por descubrir. Por esta razón se cree que es pertinente y de gran aporte intelectual a la sociedad, aplicar los modelos matemáticos que se han tratado anteriormente en este trabajo, para facilitar la toma de decisiones en el deporte del fútbol.

Aunque ya se han visto como en otros deportes las matemáticas han conseguido influenciar los criterios de selección de jugadores por parte de los entrenadores, en la actualidad es difícil encontrar un entrenador de fútbol de las grandes ligas que se base en estadísticas y procesamiento de datos para intentar preparar una temporada e ir en busca de resultados positivos. Por tal motivo un proyecto que

busque cambiar la forma de pensar y de ver el fútbol incluso antes de iniciar el partido, es un aporte nuevo y de gran impacto en la comunidad deportiva.

Lo que buscamos al realizar este proyecto es generar un aporte a la comunidad deportiva, generando unos modelos predictivos capaces de facilitar y ayudar a los encargados de planificar una temporada de un equipo de fútbol. Los equipos con menor presupuesto serían los más beneficiados, ya que encontrar jugadores que se ajusten a sus necesidades y que incrementen sus posibilidades de ganar un partido, no serían necesariamente los jugadores más costosos del mercado. Por otro lado, si el proyecto no se llevara a cabo o no se logrará los resultados esperados, el deporte del fútbol seguiría tal cual, como esta, donde los equipos ricos y tradicionales son los que ganan todos los trofeos importantes y los equipos sin mucho poder adquisitivo peleando por mantenerse económicamente estables.

Para poder realizar este proyecto, hubo la necesidad de asistir a un curso de análisis de datos que la Universidad Icesi tiene como electiva para los estudiantes de Ingeniería de Sistemas, y además se realizó un curso online de la Universidad de Stanford acerca de la disciplina Machine Learning, todo esto para poder complementar nuestros conocimientos adquiridos como Ingenieros Industriales y estar en la capacidad de hacer un trabajo completo.

4 Metodología

Se realizó el planteamiento de tres modelos predictivos, pensados en la utilidad que podrían tener en la industria del fútbol a la hora de tomar decisiones.

- El primero pretende obtener una ratio por jugador (habilidad/valor) ayudando así a la selección del jugador que más habilidad me otorga por cada millón de euros invertido, para esto primero se debe predecir el Overall (variable dependiente) del jugador, que es una puntuación totalizante, tiene en cuenta todas las variables (independientes) del jugador presentados por el dataset del FIFA 19 y luego dividirla por la valuación del jugador, dato presente en el dataset.
- El segundo modelo pretende valuar a los jugadores de futbol, ya sea para comprarlos a un precio justo o para asignar el precio a un futbolista que haga parte del plantel. Para esto se seleccionó como variable a predecir el *Value* del jugador, teniendo en cuenta todas las variables del jugador presentes en el dataset de FIFA 19.
- El tercer modelo pretende identificar las variables que más influyen en un delantero al momento de hacer un gol. Para esto se añadió una columna nueva al dataset de FIFA 19 llamada Goles para posteriormente relacionarla con las variables correspondientes a cada delantero.

Para el desarrollo de estos modelos se siguen los lineamientos de CRISP-DM (Cross-industry standard process for data mining). Es una metodología comúnmente utilizada y diseñada para resolver problemas que conciernen a la ciencia de datos. No se abordó la metodología con total especificidad ya que algunos de los lineamientos que esta sigue se sobreentienden o para este caso no son aplicables ya que es una metodología enfocado a los negocios.

Figura 4.0: *Fases de CRIP-DM [4]*

4.1. Modelos predictivos de mejoramiento (CRISP-DM)

4.1.1 Entendimiento del proyecto

Se procedió a hacer un estudio del estado del arte además de tener en cuenta el criterio personal de los realizadores del proyecto por su experticia en el fútbol como espectadores.

4.1.1.1 Evaluación de la situación: Establecidos en la página 10 (Pregunta de investigación u objeto de estudio).

4.1.1.2 Determinar los objetivos del proyecto: Establecidos en la página 12 (Objetivos generales y específicos del proyecto).

4.1.2 Entendimiento de los datos

4.1.2.1 Recolección de los datos: Se tomaron tres dataset, el primero tenía el historial de resultados de los últimos diez años de las ligas más importantes del mundo

4.1.2.2 Descripción de los datos: El dataset cuenta con la información de todos los futbolistas pertenecientes a las ligas afiliadas a FIFA 19. Este cuenta con 18206 registros de jugadores, cada jugador cuenta con 47 variables propias que determinan sus condiciones de juego.

En la siguiente figura se puede visualizar las 35 variables que finalmente se preseleccionan:

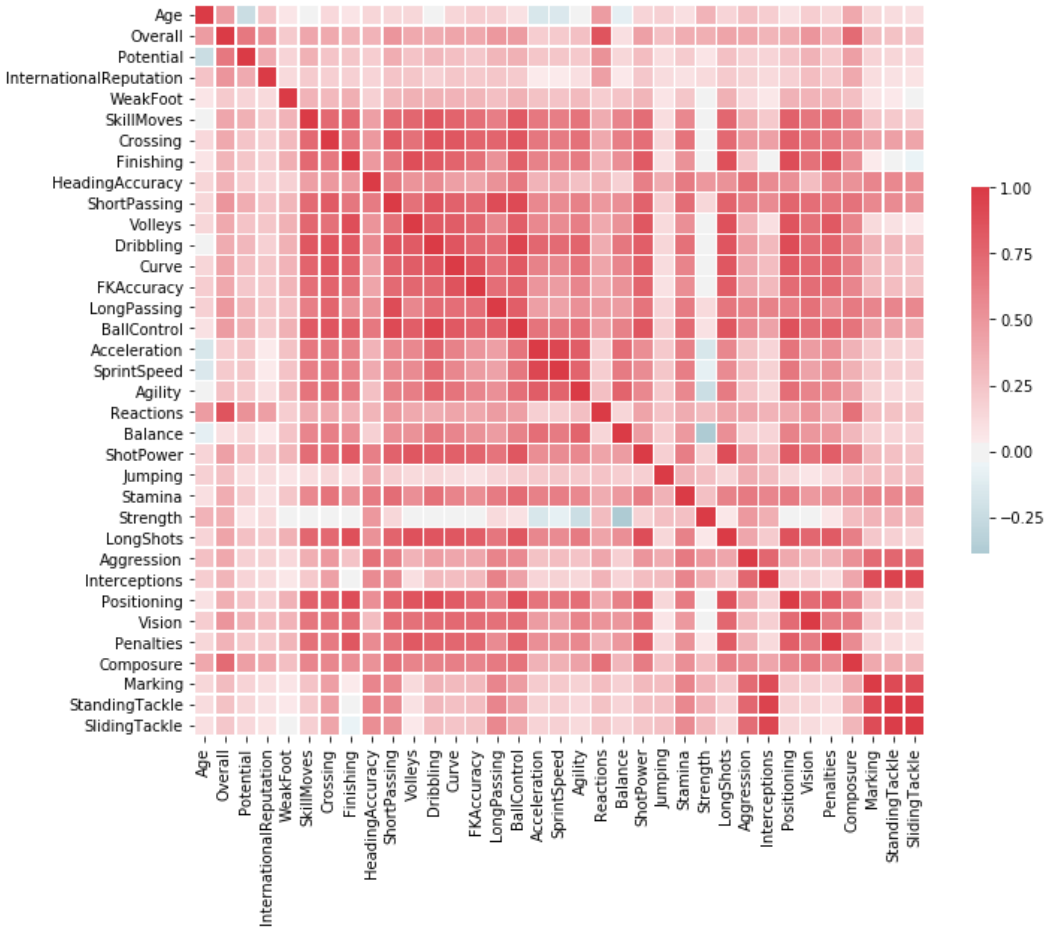

Figura 4.1.2*: Visualización de variables (Elaboración propia)*

Todas las variables del dataset son en su mayoría tipo *int* o *double*. Variables como el club, la nacionalidad y la pierna más hábil son de tipo *String.*

4.1.2.3 Verificación de la calidad de los datos:

Se parte de la fiabilidad de los datos del FIFA 19. El dataset de FIFA es uno de los más completos y gratuitos que se pueden encontrar, consta de un equipo robusto. Mueller-Moehring director de datos en FIFA dirige a 25 productores de EA y 400 contribuyentes externos de datos que recorren constantemente Internet, periódicos y revistas para obtener los datos más actualizados de los equipos y jugadores, además de una comunidad de más de 8.000 entrenadores, exploradores y titulares de boletos de temporada para verificar los datos. Haciendo un análisis descriptivo, no se encuentran datos atípicos por cada columna o variable. Variables como Value y RealClause tienen registros en *null*, además de estar en formato *String*, lo que no es conveniente ya que se pretenden trabajar como variables dependientes (variables a predecir).

4.1.3 Preparación de los datos

4.1.3.1 Selección los datos

Se hace la selección de un subconjunto de datos que cumplan con los requerimientos del modelo, se pasó de 89 columnas a 47.

Antes: ['Unnamed: 0', 'ID', 'Name', 'Age', 'Photo', 'Nationality', 'Flag', 'Overall', 'Potential', 'Club', 'Club Logo', 'Value', 'Wage', 'Special', 'Preferred Foot', 'International Reputation', 'Weak Foot', 'Skill Moves', 'Work Rate', 'Body Type', 'Real Face', 'Position', 'Jersey Number', 'Joined', 'Loaned From', 'Contract Valid Until', 'Height', 'Weight', 'LS', 'ST', 'RS', 'LW', 'LF', 'CF', 'RF', 'RW', 'LAM', 'CAM', 'RAM', 'LM', 'LCM', 'CM', 'RCM', 'RM', 'LWB', 'LDM', 'CDM', 'RDM', 'RWB', 'LB', 'LCB', 'CB', 'RCB', 'RB', 'Crossing', 'Finishing', 'HeadingAccuracy', 'ShortPassing', 'Volleys', 'Dribbling',

'Curve', 'FKAccuracy', 'LongPassing', 'BallControl', 'Acceleration', 'SprintSpeed', 'Agility', 'Reactions', 'Balance', 'ShotPower', 'Jumping', 'Stamina', 'Strength', 'LongShots', 'Aggression', 'Interceptions', 'Positioning', 'Vision', 'Penalties', 'Composure', 'Marking', 'StandingTackle', 'SlidingTackle', 'GKDiving', 'GKHandling', 'GKKicking', 'GKPositioning', 'GKReflexes', 'Release Clause']

Después: ['Name', 'Age', 'Nationality', 'Overall', 'Potential', 'Club', 'Value', 'Wage', 'PreferredFoot', 'InternationalReputation', 'WeakFoot', 'SkillMoves', 'WorkRate', 'Position', 'ContractValidUntil', 'Height', 'Weight', 'Crossing', 'Finishing', 'HeadingAccuracy', 'ShortPassing', 'Volleys', 'Dribbling', 'Curve', 'FKAccuracy', 'LongPassing', 'BallControl', 'Acceleration', 'SprintSpeed', 'Agility', 'Reactions', 'Balance', 'ShotPower', 'Jumping', 'Stamina', 'Strength', 'LongShots', 'Aggression', 'Interceptions', 'Positioning', 'Vision', 'Penalties', 'Composure', 'Marking', 'StandingTackle', 'SlidingTackle', 'ReleaseClause']

4.1.3.2 Limpieza de los datos

- Se eliminaron todas las filas con registros faltantes, quedando así 16643 registros.
- Se cambiaron los valores faltantes en la columna *Value* por el valor de mercado de cada jugador encontrado en la página transfermarket.uk.

4.1.3.3 Estructuración e integración de los datos

- Se añadieron y completaron los datos de las columnas *Goles* y *EfectividadPase* manualmente, el *OverallPrediction* con los valores arrojados por el modelo de regresión lineal múltiple. Por último, se añadió el ratio, como resultado de la división entre el *OverallPrediction* y el *Value.*
- Para el modelo de valuación de jugadores se generó una sub-tabla haciendo una filtración, jugadores con *Overall* mayor a setenta y distintos a la posición de arqueros.

- Para el modelo de selección de jugadores recomendables a partir del ratio se generó una sub-tabla haciendo una filtración, jugadores con *Overall* mayor a setenta y distintos a la posición de arqueros.
- Para el modelo de predicción de goles se generó una sub-tabla haciendo una filtración, jugadores con *Overall* mayor a setenta y delanteros (SF, CT, LF, RF).

4.1.3.4 Formateo de los datos

- Se eliminaron los espacios que tenían los nombres de las columnas, para evitar errores de sintaxis y facilitar el llamado de métodos.
- Se llevaron los datos de la columna *Value* y *RealClause* de tipo String a double haciendo una eliminación de los caracteres € y M ubicados al principio y al final de cada dato y después una conversión de tipo dato por medio de Power Query.
- Variables como *Weight*, y *Height* también estaban como String, también se aplicó una conversión de tipo de dato atreves de Power Query.

4.1.4 Modelado

4.1.4.1 Selección de técnica de modelado

Se seleccionó el modelo de regresión lineal múltiple y arboles regresores.

4.1.4. Plan de prueba

Se seleccionaron los estadísticos o herramientas para hacer la prueba del modelo.

- **Valor P:** Para descartar variables que no influyen o influyen muy poco en el modelo.
- **Matriz de correlación entre variables:** Para determinar independencia de las variables.
- **Lasso y Ridge:** Para la simplificación del modelo.

6.1.4.3 Construcción del modelo

- Modelo de valorización de jugadores: $y \rightarrow Value\ ,\ x \rightarrow Otras\ variables$
- Modelo de selección de jugadores: $y \rightarrow Overall\ ,\ x \rightarrow Otras\ variables$
- Modelo de predicción de goles en delanteros: $y \rightarrow Goles,\ x \rightarrow Otras\ variables$

4.1.4.4 Evaluación del modelo

Se evaluó el modelo a partir de:

- **R cuadrado ajustado de testeo:** Porcentaje de variación de la variable dependiente que es explicado por todas las variables independientes que hacen parte del modelo, con los datos de prueba.
- **RMSE:** Desviación media de la raíz cuadrada, muy utilizado a la hora de evaluar modelos de regresión.

4.1.5 Evaluación

Se hizo un a comparación entre los modelos ejecutados (regresión lineal múltiple, arboles regresores) y se seleccionó el modelo que mejor se ajustó a las pruebas y a la evaluación.

4.1.6 Implementación

La fase de implementación no aplica en este caso, se pretende un planteamiento de tres modelos que mejoren la toma de decisiones.

4.2 Herramientas utilizadas

4.2.1 Jupyter Notebooks: Es un entorno interactivo web de ejecución de código, fue utilizado para facilitar la visualización y ejecutar el código escrito en Python.

4.2.2 Excel y Power Query: Facilito el proceso de transformación de los datos, el manejo de sub-tablas realizadas por cada modelo y la adición de datos faltantes.

Debido a que el proyecto entra en la categoría Small Data, herramientas como estas siguen siendo muy útiles.

4.2.3 Librerías de Python: Python es un lenguaje de programación idóneo para la ciencia de datos, contiene librerías que facilitan el trabajo.

- **Pandas:** Se utilizó pandas para hacer la limpieza de datos. Se eliminaron columnas innecesarias para el modelo y registros incompletos de jugadores, se cambiaron formatos de columnas y se añadieron filas nuevas (Goles anotados, Ratio, Efectividad de pase) para el desarrollo de los modelos.
- **Matplotlib:** Se utilizó para presentar el comportamiento de los datos en visualizaciones.
- **Sk-Learn:** Se utilizó para entrenar el modelo y obtener valores de relevancia para el análisis del modelo como el r2 ajustado y el valor P.
- **StatsModels:** Se utilizó para obtener valores estadísticos que ayudaron a mejorar y simplificar el modelo.

Para facilitar el llamado e importación de estas librerías se utilizó Anaconda, un software libre que recopila todas las librerías importantes de Python.

5. Resultados

Esta sección se divide en tres partes, cada parte consta de una descripción de las condiciones iniciales del modelo, los resultados y experimentos a partir de este. La primera, presenta los resultados encontrados en el modelo de valuación de jugadores a partir de las métricas ya establecidas, la segunda los resultados del modelo de selección de jugadores a partir del ratio, por último, se presentará el modelo de predicción de goles de delanteros. De cada modelo se pretende exponer lo obtenido de una manera gráfica y concreta.

5.1 Modelo de Valuación de jugadores

5.1.1 Condiciones iniciales del modelo

Para realizar este modelo se utilizó la información de 3913 jugadores de fútbol de la base de datos de FIFA 19. Dichos jugadores se seleccionaron luego de filtrar a los jugadores que cumplieran con las condiciones de no ser porteros y tener un Overall mayor a 70, esto último debido a que los jugadores con un Overall menor de 70 carecen de información de calidad y estadísticas detalladas. Para predecir el Value se usaron las siguientes variables que se tuvieron en cuenta para hallar el modelo predictivo:

```
caracteristicas = ['ContractValidUntil','Age','Overall', 'InternationalReputation',
                   'Wage','ReleaseClause', 'Position']
```

Figura 5.1.1: *Características de los jugadores [Elaboración propia]*

5.1.2 Experimentos

- La inclusión de todas las variables tipo habilidad del jugador no produjo un aumento significativo en el modelo**.** Con habilidades: R2_Ajustado_test: 0,9845, sin habilidades: R2_Ajustado_test: 0,9844.

	Metricas		
Modelo	**R2_Ajustado_Test**	**RMSE**	**MSE**
Regresión lineal multiple	0,9844	1,196	1,433
Arboles de regresión	0,9763	1,501	1,648
Regresión Bayesian Ridge	0,9821	1,197	1,482

Tabla 5.1.2: *Modelos y métricas de regresión [Elaboración propia]*

5.1.3 Hallazgos

No hubo necesidad de incluir variables relacionadas a la habilidad del jugador, el Overall considera todas estas variables con una alta correlación como se puede observar en la tabla 5.2.2 del segundo modelo, esto disminuyó en gran proporción el ruido del modelo sin necesidad de perder capacidad de predicción.

Según las métricas arrojadas por el programa se observó que los tres tipos de modelos de regresión obtuvieron un R2 ajustado muy alto, lo que indica que cualquiera de los tres sería un modelo satisfactorio para la predicción de la variable Value. Una vez arrojado el modelo se halló una disminución del RMSE de 1,29 a 1,196 al añadir la variable categórica *Position* al modelo de predicción. Se determinó que el modelo de regresión lineal múltiple es el que mejor se ajusta los datos utilizados, y para comprobarlo se comparó el RMSE con los modelos de árboles de regresión y regresión Bayesian Ridge y sus resultados se muestran en la tabla 5.1.2. La ecuación del modelo Value está definida de la siguiente manera:

```
regreLinear.coef_
```

```
array([[-0.0211694 ,  0.00707877,  0.10685063,  0.26265291,  0.00534421,
         0.4766999 , -0.07597513, -0.30717196, -0.44440862,  0.45737036,
        -0.22103229,  0.3497714 , -0.34042772, -0.2467617 , -0.14163973,
        -0.28279186,  3.00288407, -0.06844583,  0.19078099, -0.43228289,
        -0.29247778, -0.13717188, -0.33128574, -0.25761989, -0.15190308,
        -0.31182955,  0.68813468, -0.29074983,  0.02933094,  0.02606505,
        -0.24165895, -0.1687031 ]])
```

```
regreLinear.intercept_
```

```
array([35.04444558])
```

Figura 5.1.3: *Coeficientes e intercepto del modelo 1 de regresión lineal múltiple [Elaboración propia]*

5.2 Modelo de selección de jugadores a partir del ratio

5.2.1 Condiciones iniciales del modelo

Para realizar este modelo se utilizó la información de 3913 jugadores de fútbol de la base de datos de FIFA 19. Dichos jugadores se seleccionaron luego de filtrar a los jugadores que cumplieran con las condiciones de no ser porteros y tener un Overall mayor a 70, esto último debido a que los jugadores con un Overall menor de 70 carecen de información de calidad y estadísticas detalladas. Para la predicción del Overall se usaron las siguientes variables:

```
caracteristicas = ['Potential','Crossing', 'Finishing', 'HeadingAccuracy',
       'ShortPassing', 'Volleys', 'Dribbling', 'Curve', 'FKAccuracy',
       'LongPassing', 'BallControl', 'Acceleration', 'SprintSpeed',
       'Agility', 'Reactions', 'Balance', 'ShotPower', 'Jumping', 'Stamina',
       'Strength', 'LongShots', 'Aggression', 'Interceptions', 'Positioning',
       'Vision', 'Penalties', 'Marking', 'StandingTackle', 'SlidingTackle']
```

Figura 5.2.1: *Características Modelo 2 [Elaboración propia]*

- Definición del Ratio: Es el resultado de dividir el *OverallPrediction* obtenido del modelo de regresión por el *Value* que es un valor al cual se tiene acceso y conocimiento. El Value es el valor del mercado del jugador de fútbol que aparece en FIFA 19, el cual es extraído de la página Transfermarket

- $Ratio = \frac{OverallPrediction}{Value}$ La función del ratio es determinar cuántos puntos de habilidad (Overall) tiene el jugador por cada millón de euros invertidos en su traspaso. Entre mayor es el ratio, mejor es para el inversor ya que está invirtiendo menos dinero por un punto de habilidad del deportista.

Para determinar las variables más influyentes en el Overall de un jugador, se partió del modelo de regresión lineal usado anteriormente para predecir la característica anteriormente mencionada de cada jugador fútbol.

5.2.2 Experimento

Por el método de selección de variables *Backward* se realizó una a una la eliminación de cada variable para analizar la disminución del R2 Ajustado.

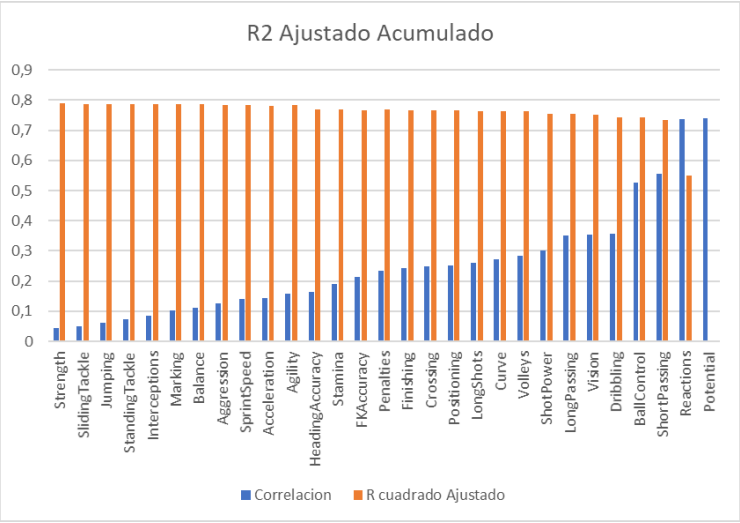

Figura 5.2.2: *Figura R2 ajustado Acumulado*

	Metricas		
Modelo	**R2_Ajustado_Test**	**RMSE**	**MSE**
Regresión lineal multiple	0,803	1,666	2,778
Arboles de regresión	0,789	1,810	2,845
Regresión Bayesian Ridge	0,797	1,745	2,801

Tabla 5.2.2: *Modelo y métricas del modelo 2 [Elaboración propia]*

5.2.3 Hallazgos

Después de que el modelo arrojó las métricas (R2 ajustado, RMSE, MSE) de cada una de las regresiones, se pudo observar que los tres modelos tienen un R2 ajustado muy similar ya que la diferencia máxima entre ellos es de 0,014, comportamiento similar que comparte el RMSE y MSE donde no hay una gran diferencia entre los resultados arrojados de cada modelo.

Al analizar los resultados de las tablas y sus anexos, se puede inferir que este modelo sigue el principio de Pareto, donde las primeras 3 variables equivalentes al 13,8% del total, logran predecir el Overall en un 94,23% con respecto al modelo más complejo de 29 variables. Según los datos las variables que mayor correlación tiene con el Overall de un jugador de fútbol son: *Dribling, LongPassing, ShortPassing, BallControl* y *Reactions,* las cuales se pueden catalogar de tipo técnicas.

Habilidades	Overall	R cuadrado Ajustado
Finishing	0,243294385	0,76681
Crossing	0,247405172	0,76688
Positioning	0,251164478	0,76496
LongShots	0,259567052	0,76273
Curve	0,271695612	0,76274
Volleys	0,284164349	0,76257
ShotPower	0,301492358	0,75414
LongPassing	0,349880652	0,75360
Vision	0,353035126	0,75253
Dribbling	0,356061405	0,74305
BallControl	0,526734412	0,74277
ShortPassing	0,556024645	0,73320
Reactions	0,738359071	0,54871
Potential	0,740867433	0

Tabla 5.2.3*: Lista de las variables más influyentes*

Al realizar las comparaciones entre los distintos modelos y sus métricas, se determinó que el modelo que mejor se ajusta al dataset filtrado es el de Regresión lineal múltiple, y según los datos obtenidos, 11 variables de las seleccionadas previamente no son significativas según el criterio del Valor P. La ecuación del modelo de OverallPrediction es la siguiente:

```
regreLinear.coef_
```

```
array([[ 0.33521711,  0.01017453,  0.00772334,  0.03489563,  0.06480676,
         0.00233205, -0.02780861,  0.00407445,  0.01199074,  0.01361487,
         0.10479304,  0.00890421,  0.01093339, -0.00199387,  0.24670198,
        -0.00573798,  0.02664104,  0.00933496, -0.01100507,  0.02011827,
        -0.02022998,  0.0107184 ,  0.00073267, -0.03519397, -0.0100761 ,
         0.01074824,  0.01593203, -0.01055203, -0.0082418 ]])
```

```
regreLinear.intercept_
```

```
array([14.70718823])
```

Figura 5.2.3: *Coeficientes e intercepto del modelo 2 de regresión lineal múltiple [Elaboración propia]*

- A continuación la tabla con las comparativas de los jugadores que más nivel de Overall otorga por cada millón de Euros invertidos, o en otras palabras el Ratio de los jugadores.

Name	Overall	Value	OverallPrediction	Ratio
L. Messi	94	110,5	94	**0,85**
Cristiano Ronaldo	94	77	89	**1,16**
Neymar Jr	92	118,5	92	**0,78**
K. De Bruyne	91	102	91	**0,89**
E. Hazard	91	93	91	**0,98**
L. Modrić	91	67	91	**1,36**
L. Suárez	91	80	91	**1,14**
Sergio Ramos	91	51	84	**1,65**
R. Lewandowski	90	77	89	**1,16**
T. Kroos	90	76,5	89	**1,16**
D. Godín	90	44	87	**1,98**
David Silva	90	60	90	**1,50**
N. Kanté	89	63	89	**1,41**
P. Dybala	89	89	89	**1,00**
H. Kane	89	83,5	89	**1,07**
A. Griezmann	89	78	91	**1,17**
Sergio Busquets	89	51,5	89	**1,73**
E. Cavani	89	60	89	**1,48**
S. Agüero	89	64,5	89	**1,38**
G. Chiellini	89	27	89	**3,30**
K. Mbappé	88	81	88	**1,09**
M. Salah	88	69,5	91	**1,31**

Tabla 5.2.3: *Tabla de Ratio por jugador [Elaboración propia]*.

En la tabla se pueden observar el Overall estipulado del jugador en FIFA 19, el valor real de traspaso, el ratio y la predicción de nuestro modelo con el método de regresión que mejor se ajusta. Se puede observar el ratio para cada uno de los jugadores mostrados en la tabla, el cual está calculado con el Overallprediction ya que con este valor es que van a trabajar los directores técnicos y directivos a la hora de tomar decisiones. Para hacer una elección más eficiente, se debe saber que el jugador que debe ser seleccionado por su ratio, es aquel que tenga un valor mayor en esa columna.

5.3 Modelo de Predicción de goles por temporada de un delantero

5.3.1 Condiciones iniciales

El modelo consta de 67 registros, esto debido a la filtración de delanteros con Overall mayor a 80, además de la dificultad que implica digitar el número de goles por cada jugador.

Debido a que es un tamaño muestral (67) comparado con todos los jugadores que se desempeñan como delanteros (aprox 5071), el modelo de predicción se tratara con un nivel de confianza de 0,005.

A continuación, las variables que hacen parte del modelo y explican el número de goles realizados por un jugador en una temporada:

```
caracteristicas = ['Potential','Crossing', 'Finishing', 'HeadingAccuracy'
                   , 'Volleys', 'Dribbling', 'Curve', 'FKAccuracy',
       'LongPassing', 'BallControl', 'Acceleration', 'SprintSpeed',
       'Reactions', 'Balance', 'ShotPower', 'Jumping', 'Stamina',
       'Strength', 'LongShots', 'Aggression', 'Positioning',
       'Vision', 'Penalties','InternationalReputation', 'Wage']
```

Figura 5.3.1: *Características Modelo 3 [Elaboración propia]*

5.3.2 Experimentos

Se eliminaron las variables poco significativas del modelo por el criterio del valor P, dando como resultado las características mostradas en la figura 5.3.1.

Se entrenó el modelo con 3 tipos de modelos de regresión lineal múltiple para determinar cuál es el que mejor se ajusta a los datos.

	Metricas		
Modelo	**R2_Ajustado_Test**	**RMSE**	**MSE**
Regresión lineal multiple	0,8385	14,668	215
Arboles de regresión	0,8401	10,685	NA
Regresion Bayesian Ridge	0,8873	5,498	NA

Tabla 5.3.3: Modelos y métricas del tercer modelo

5.3.3 Hallazgos

Pese a que no se tiene intención de ahondar en el funcionamiento y modelo detrás de la Regresión Bayesian Ridge, esta demuestra ser sustancialmente mejor por la métrica de RMSE.

Para este modelo se obtuvo 3 tipos de regresiones con un valor de R2 ajustado mayores a 0,825, pero 2 de estos tuvieron un RMSE alto para la finalidad de predecir los goles de un jugador de fútbol en una temporada regular. Para la métrica MSE no fue posible determinar una comparación entre las regresiones ya que en 2 de ellas no se pudo hallar este tipo de métrica.

Para este tercer modelo, el método que más se ajusta es la Regresión Bayesian Ridge, la cual es capaz de predecir los goles anotados de un futbolista en un año con un error aproximado de 5,5 goles, a continuación el modelo:

```
regreBayesian.coef_
```

```
array([ 0.01269111, -0.01460894,  0.01171688,  0.04452533,  0.02011595,
        0.00479859,  0.00076795,  0.01255964, -0.03405122,  0.01730743,
       -0.01064312, -0.03538714, -0.02129019,  0.02239153,  0.03109004,
        0.01005095,  0.00161142,  0.0426837 ,  0.02608922, -0.00981868,
       -0.00072002, -0.01946434,  0.01306489,  0.00438446,  0.03424045,
        0.00217094,  0.05901309])
```

```
regreBayesian.intercept_
```

```
-5.600884393155493
```

Figura 5.3.5: *Coeficientes e intercepto del modelo 2 de regresión Bayesian [Elaboración propia]*

6. Conclusiones

En el presente proyecto se crearon modelos predictivos y se generaron correlaciones entre variables para ayudar en la toma de decisiones a los directivos y cuerpo técnico de equipos de fútbol. Como partes de las conclusiones se encuentra:

- Los jugadores con un *Overall* superior a 70 son los que tienen información de peso y calidad para lograr sacar resultados.

- En el primer modelo de predicción (*Value*) se obtuvo un R2 ajustado de 0,9844, lo cual indica que es un modelo con una alta exactitud y que no contiene un gran número de variables. Por lo cual este modelo explica en un 98,44% la variable dependiente.

- Para lograr tener un mejor modelo predictivo en la característica *Value* es necesario cambiar la variable Position a una variable binaria, esto con el fin de mejorar el RMSE.
- Se observó en el segundo modelo de predicción, que los jugadores con mayor ratio son aquellos que tienen una edad avanzada (30-35 años) que fueron reconocidos, jugaron en equipos importantes y que todavía tienen un buen nivel futbolístico.
- De acuerdo con los análisis de los modelos de regresión, *Reactions* es la variable que mayor relación tiene con el numerador del Ratio que hace las veces de un Overall. Para este cálculo el numerador se calculó restándole el Overall mínimo (70) al *OverallPrediction* con el objetivo de utilizar los datos de calidad.
- Realizando las relaciones entre todas las variables que tiene un jugador de fútbol, se pudo concluir que dos de las parejas de variables más relacionadas entre sí fueron *BallControl* y *Positioning* junto a *Vision* y *Dribbling.*

- En el segundo modelo de predicción (*Overall*) se obtuvo un R2 ajustado 0,803, lo cual determina que es un modelo satisfactorio que predice con exactitud el Overall de los jugadores con una poca cantidad de variables.

  Según el criterio del Valor P, las siguientes variables no son significativas a la hora de predecir el *Overallprediction*: *Strength, SlidingTackle, Jumping, StandingTackle, Interceptions, Marking, Balance, Agression, SprintSpeed, Acceleration.*

- Entre mayor la diferencia (Positivamente) entre el *Overallprediction* y el Value, mejor es el Ratio del jugador. En la tabla se logra determinar que el jugador que tiene el mejor Ratio es Chiellini.

- Las variables que más influyentes en la determinación del Overall de un jugador de fútbol son: *Dribling, LongPassing, ShortPassing, BallControl* y *Reactions.*

- Las características que menos se correlacionan con el Overall en términos de r2 ajustado se relacionan con las habilidades físicas. Habilidad que pueden se trabajadas a partir del entreno arduo, esto sugiere que las habilidades a buscar son las innatas, propias del jugador y difícilmente mejorables con el entreno.

7. Recomendaciones

- Se recomienda poder realizar un análisis de los goles recibidos y su correlación con los jugadores, además de descifrar la forma correcta de realizarlo, puesto que los goles recibidos dependen en gran medida de los 4 defensores y el arquero.

- Para un mejor método de predicción con los goles anotados, se recomienda una mayor base de datos con los goles de los delanteros, esto con el fin de tener más datos para un mejor resultado de la metodología Machine Learning.

- Si es posible tener una base de datos que tenga un buen análisis estadístico de los jugadores de fútbol no tan conocidos. Esto serviría para poder tener una base de datos aún más robusta y así poder hallar mejores modelos predictivos

- Para el uso adecuado de estos modelos se debe realizar una interpolación que relacione las variables del dataset de FIFA 19 que comúnmente van de 0 hasta 100 con los datos cuantitativos que se pueden extraer de una página web. Para ejemplificar, se puede desarrollar una relación a través de una interpolación entre el porcentaje de acierto en pases de un jugador y el puntaje del jugador en variables presentes en el dataset de FIFA 19 como ShortPassisng y LongPassing.

- Siempre una data más grande y detallada permitirá desarrollar modelos más robustos y específicos según la necesidad de los directivos y directores técnicos de fútbol, es por esto que si se quiere implementar en totalidad las herramientas de machine Learning para el desarrollo de mejores modelos se deben conseguir bases de datos de pago que tienen mayor informan de la que puede brindar el dataset de FIFA 19.

- Se recomienda a la Universidad Icesi implementar cursos para los Ingenieros Industriales que tengan como eje principal el Big Data y su análisis, ya que puede ser muy beneficioso para el intelecto de los estudiantes y además es muy valorado en las empresas hoy en día.

- Si se tiene la intención de desarrollar estos modelos en un ámbito local, el dataset de FIFA 19 resulta insuficiente, los datos no son los suficientemente exactos, además carecen de información real en este tipo de ligas secundarias como la colombiana donde la recolección de datos y posterior puesta en base de datos no se realiza.

- se puede desarrollar una relación a través de una interpolación entre el porcentaje de acierto en pases de un jugador y el puntaje del jugador en variables presentes en el dataset de FIFA 19 como ShortPassisng y LongPassing.

- Modelos como el predictor de goles a partir de las características de las características de los jugadores carecen de un análisis de inferencia estadística, estos deben ser tratados desde este enfoque por la carencia de datos de goles anotados de los futbolistas, es decir, el tamaño de la muestra se sigue considerando muestral en comparación con todos los delanteros que realizaron goles en todas las ligas del mundo.

Bibliografía

[1] Ecured. (23 de Febrero de 2016). *Ecured*. Recuperado el 27 de Marzo de 2019, de Ecured: https://www.ecured.cu/Regresi%C3%B3n_lineal

[2] Enciclopedia Económica. (1 de Junio de 2017). *Enciclopedia Económica*. Recuperado el 27 de Marzo de 2019, de Enciclopedia Económica: https://enciclopediaeconomica.com/variable-estadistica/

[3] Link, D. (2018). *Data analytics in profesional soccer.* Springer.

[6] Matplotlib. (2015). *Matplotlib*. Recuperado el 27 de Marzo de 2019, de Matplotlib: www.matplotlib.org

[4] Memmert, D. (2018). *Data Analytics in Football.* Routledge.

[5] Pawa, D. A. (2017). *Sports Predictive Analytics.* San Diego: IEE Society.

[8] RAE. (2001). *Diccionario de la lengua española.* Madrid: Diccionario de la lengua española.

[7] Rojas, D. A. (2017). *Aplicación del modelo AHP en la valoración de futbolistas profesionales.* Bucaramanga: Universidad de Santander UDES.

[9] Roldan, P. N. (14 de Septiembre de 2018). *Economipedia*. Recuperado el 27 de Marzo de 2019, de Economipedia: https://economipedia.com/definiciones/modelo-matematico.html

[10]W. Hasmer, D. (Regression). *Applied Logistic.* Massachusetts: Wiley.

[11] Wiston, W. (2009). *Mathletics.* New Jersey: Princeton.