DESARROLLO DEL CASO DE ESTUDIO DE LA AGENCIA ANALYTICS FÚTBOL S.A

Analítica para la toma de decisiones

SUSANA BARRIENTOS
JAVIER BURGOS
FABIO ANDRES GONZALEZ

Caso de estudio:

Analytics Fútbol S.A. Se dedica al análisis de datos en el mundo del fútbol y busca desarrollar un sistema que permita agrupar automáticamente a los jugadores en las posiciones típicas de delantero, portero, defensa y medio. Para ello, han recopilado datos de jugadores de fútbol que incluyen atributos como el pase, la pegada, el salto, los reflejos y otros. El objetivo es descubrir patrones ocultos y agrupar a los jugadores en estas posiciones de manera eficiente.

Problema:

El problema radica en la necesidad de agrupar automáticamente a los jugadores de fútbol en las posiciones típicas (delantero, portero, defensa y medio) a partir de datos de habilidades y características.

Diseño de solución propuesto:

La solución implica utilizar técnicas de aprendizaje no supervisado, como el clustering, para agrupar a los jugadores en clusters. Luego, se pueden identificar los clusters que representan a las posiciones de delantero, portero, defensa y medio. Esto proporcionará una forma objetiva de agrupar a los jugadores en sus posiciones.

Objetivos:

- ➢ Determinar la agrupación de los jugadores de acuerdo a las características presentadas en la base de datos.
- ➢ Proporcionar información útil para los equipos de fútbol y cazatalentos al determinar la posición más adecuada para un jugador en función de sus habilidades.
- ➢ Utilizar técnicas de aprendizaje no supervisado, como el clustering, para agrupar a los jugadores en grupos basados en sus habilidades y características, con el objetivo de identificar las posiciones comunes: delantero, portero, defensa y medio.

En referencia a los objetivos se consideró trabajar con las variables que se enmarquen dentro de los siguientes atributos:

Estadísticas físicas: Altura, peso, velocidad, fuerza, resistencia, agilidad, equilibrio, etc. Estas variables pueden ayudar a identificar jugadores con características físicas que se adapten mejor a determinadas posiciones. Por ejemplo, los jugadores más altos y fuertes pueden ser más adecuados para posiciones defensivas, mientras que los jugadores más rápidos y ágiles pueden ser más adecuados para posiciones ofensivas.

Estadísticas técnicas: Tiro, pases, regate, visión de juego, etc. Estas variables pueden ayudar a identificar jugadores con habilidades técnicas que se adapten mejor a determinadas posiciones. Por ejemplo, los jugadores con una buena habilidad de tiro pueden ser más adecuados para posiciones ofensivas, mientras que los jugadores con una buena habilidad de pase pueden ser más adecuados para posiciones de mediocampo.

Estadísticas mentales: Agresión, concentración, decisión, etc. Estas variables pueden ayudar a identificar jugadores con características mentales que se adapten mejor a determinadas posiciones. Por ejemplo, los jugadores más agresivos pueden ser más adecuados para posiciones defensivas, mientras que los jugadores más decididos pueden ser más adecuados para posiciones de ataque.

PRE-SELECCIÓN DE VARIABLES

Las variables que descartamos son aquellas que presentan un alto porcentaje de valores nulos (> 50%) o que no aportan información relevante al modelo. En el caso de las variables con valores nulos, no es posible imputarlos de manera confiable, ya que no se dispone de información suficiente para hacerlo. Por otro lado, las variables que no aportan información relevante al modelo son aquellas que no están relacionadas con el objetivo del estudio, que es identificar perfiles de jugadores según su posición.

Por ejemplo, la variable **club_loaned_from** indica si el jugador está cedido a otro club. Esta información no es relevante para el objetivo del estudio, ya que no está relacionada con la posición del jugador. De igual manera, la variable **nation_team_id** indica el ID del equipo nacional del jugador. Esta información tampoco es relevante para el objetivo del estudio, ya que no está relacionada con la posición del jugador en el club.

PROCESAMIENTO DE DATOS

Posteriormente se analizó los datos faltantes, donde se identificó que algunas variables numéricas carecían de datos; estos fueron rellenados con la media de la variable. Ya con estos procedimos hacer un análisis exploratorio por medio de una matriz de correlación donde se determinó que las variables más correlacionadas eran las subcategorías del defender, atacar, entre otras. Con base en esta información, se tomó la decisión de agrupar las variables relacionadas con la defensa, consolidando así las características de esta categoría en una sola variable representativa; de manera análoga, se agruparon las demás variables relacionadas con sus respectivas categorías. Este proceso de agrupación resultó en una simplificación de las dimensiones del conjunto de datos.

Una vez completado el proceso de agrupación, se procedió a refinar la matriz de correlación, donde se observó que se solucionó la correlación alta de las variables mencionadas anteriormente, sin embargo, se encontró una variable con alta correlación la que fue eliminada. En el análisis exploratorio también se encontraron variables con demasiados atípicos; a través de diagramas de bigotes, las cuales fueron eliminadas.
Después del análisis exploratorio pasamos al pre-procesamiento de datos donde fuimos estrictos en el tratamiento de atípicos, eliminando todos y quedando solo con 6251 filas y 17 columnas. Ya con nuestra base de datos limpia pasamos a la aplicación de modelos.

De acuerdo a las consideraciones mencionadas, las variables que quedaron para realizar el estudio son las siguientes:

Variable	*Traducción*	*Descripción*
overall	Valoración general	Valoración general del jugador.
age	Edad	Edad del jugador.
height_cm	Altura en centímetros	Altura del jugador en centímetros.
weight_kg	Peso en kilogramos	Peso del jugador en kilogramos.
club_position	Posición en el club	Posición del jugador juega en su club actual.
skill_moves	Movimientos técnicos	Número de movimientos técnicos
pace	Ritmo	Ritmo del jugador.
shooting	Disparo	Disparo del jugador.
passing	Pase	Pase del jugador.
dribbling	Regate	Regate del jugador.
defending	Defensa	Defensa del jugador.
physic	Físico	Físico del jugador.
attacking_crossing	Centro	AGRUPADA EN UNA SOLA VARIABLE
attacking_finishing	Definición	
attacking_heading_accuracy	Precisión de cabeza	
attacking_short_passing	Pase corto	
attacking_volleys	Voleas	
skill_dribbling	Regate	AGRUPADA EN UNA SOLA VARIABLE
skill_curve	Curva	
skill_fk_accuracy	Precisión de tiros libres	
skill_long_passing	Pase largo	
skill_ball_control	Control del balón	
movement_acceleration	Aceleración	AGRUPADA EN UNA SOLA VARIABLE
movement_sprint_speed	Velocidad de sprint	
movement_agility	Agilidad	
movement_reactions	Reacciones	
movement_balance	Equilibrio	
power_shot_power	Potencia de disparo	AGRUPADA EN UNA SOLA VARIABLE
power_jumping	Salto	
power_stamina	Resistencia	
power_strength	Fuerza	
power_long_shots	Disparo de larga distancia	
mentality_aggression	Agresividad	AGRUPADA EN UNA SOLA VARIABLE
mentality_interceptions	Intercepciones	
mentality_positioning	Posicionamiento	
mentality_vision	Visión	
mentality_penalties	Penaltis	
mentality_composure	Compostura	

Tabla 1: Variables finales

DESARROLLO

FASE 1: Implementación de modelo K-means con todas las variables.

Resultado prueba “elbow“ para determinar el # de cluster: 3.

K-MEANS
Inertia: 4218161.141257677
Silhouette Score: 0.2952984873311604 Calinski harabasz score: 3511.4521799457143

FASE 2: Implementación del modelo K-means con aplicación de PCA (Reducción de dimensionalidad)

Aplicación de PCA:

Variable	***PC 1***	***PC 2***	***PC 3***
overall	0.02	0.20	0.05
age	-0.00	0.05	0.11
height_cm	-0.07	-0.06	0.32
weight_kg	-0.05	-0.04	0.34
club_position	0.02	-0.02	0.02
pace	0.18	0.08	-0.54
shooting	0.42	0.34	0.25
passing	0.10	0.37	-0.05
dribbling	0.18	0.28	-0.13
defending	-0.53	0.34	-0.06
physic	-0.12	0.12	0.33
defending1	-0.59	0.36	-0.14
attacking1	0.20	0.28	0.14
skills1	0.15	0.31	-0.02
mentality1	0.03	0.32	0.11
movement1	0.15	0.16	-0.41
power	0.09	0.22	0.24

Tabla 2: Cargas de las variables en cada componente

	Varianza explicada (%)
0	46,94%
1	28,20%
2	11,36%

Tabla 3: Porcentajes que explican el 85% de las variables del modelo

El 85% de las variables se ven explicadas en 3 componentes PCA.

ALGORITMO K-MEANS

Resultado prueba “elbow“ para determinar el # de cluster: 3

K-MEANS
Inertia: 3023652.2984972713
Silhouette Score: 0.36524639699250255
Calinski harabasz score: 4883.444567366748

FASE 3: Implementación de modelo Gaussian Mixture con aplicación de PCA (Reducción de dimensionalidad)

Tras la aplicación de este modelo, y obtener un **Silhouette Score: 0.3552561376002395** se puede inferir hasta el modelo, que no realiza una adecuada perfilación respecto a cada cluster.

Interpretación de los clusters

La distribución de los datos entre los clusters no fue significativa, no hubo una diferencia tan marcada, por el contrario si bien el cluster 2 contó con mayor número de datos y el cluster 1 con la menor cantidad, de forma global los tres rondaron la misma cantidad, lo que nos demuestra que no hay un grupo minoritario de jugadores que comparten características.

Cluster 1

- El 79 % de los jugadores en este cluster son defensas.
- Seleccionó la mayor parte de jugadores con una calificación 'overall' igual a 65.
- Los jugadores tienen una puntuación de 'defensa' entre 50 y 80 puntos, prevaleciendo los jugadores con puntuación de 60 puntos.
- Contiene a los jugadores con menor puntuación en habilidades generales.

Cluster 2

- El 59 % de los jugadores son delanteros
- Contiene a los jugadores mejor calificados en habilidades defensivas y también de ataque, en este custer también están los jugadores con mayor puntuación en habilidades mentales, por ejemplo la destreza de tomar una decisión al momento de realizar un pase o lanzar a la cancha.
- Los jugadores tienen una puntuación de 'defensa' entre 25 y 50 puntos.

Cluster 3

- El 59 % de los jugadores son mediocampistas
- No tomo a porteros dentro de los jugadores que seleccionó.
- Contiene a jugadores con mayor puntuación en el parámetro 'overall' No seleccionó jugadores con 'overall menor a 60.

Evaluación de los algoritmos

A pesar de que los dos algoritmos evaluados emitieron una baja puntuación respecto al parámetro **Silhouette Score**de, el algoritmo con mejores métricas fue el de **K-MEANS,** y se evidencia en los gráficos de dispersión donde hace un agrupamiento de los clusters sin solapamientos o superposiciones.

Indicadores:	FASE 1 - K-MEANS DataSet Original	FASE 2 - K-MEANS DataSet PCA	FASE 3 - GaussianMixture PCA
Inertia:	4.218.161,14126	3.023.652,29850	
Silhouette Score:	0,295298487	0,365246397	0,355256138
Calinski harabasz score:	3511,45218	4883,444567	4706,089042

Conclusiones

Luego de la comparación final, se observa que en las tres fases del proceso los modelos aplicados cumplen con la finalidad: agrupar los jugadores. Sin embargo, el modelo con mejores métricas corresponde al K-Means con aplicación de reducción de dimensionalidad. A pesar de la reducción drástica de los datos, los algoritmos utilizados lograron mostrar un buen desempeño.

Este ejercicio ayuda a entender mejor la clasificación de una gran cantidad de jugadores y así permitir a un club deportivo generar una elección adecuada de jugadores respecto a la posición y habilidades registradas.

Por último, se entiende que la flexibilidad frente a la eliminación de valores atípicos debe considerarse y ser moderada, en este caso se realizó una eliminación drástica que beneficio el desarrollo del proceso y culmino en resultados fáciles de interpretar.