Introducción a la ciencia de datos

FING 2024

Tarea Final

Segmentación de Equipos en la Liga Profesional de Fútbol Argentino a través del Clustering de estadísticas de pases.

Autor
Guido Pereyra

1. Introducción

En el ámbito del fútbol moderno, el análisis de datos se ha convertido en una herramienta esencial para comprender y mejorar el rendimiento de los equipos. Este proyecto se centra en la aplicación de métodos de aprendizaje no supervisado para segmentar a los equipos de la primera división del fútbol argentino, utilizando estadísticas de pases. Los datos se extraen de la plataforma FBref, conocida por su amplia cobertura de estadísticas de ligas de fútbol de todo el mundo y los valores de los planteles se extraen de la web transfermarkt.

2. Descripción del Juego de Datos

En esta sección, se describe el conjunto de datos utilizado en el análisis, incluyendo su origen, características principales y posibles problemas de calidad.

Fuente del Conjunto de Datos:

- **Origen:** Los datos se extraen de dos principales plataformas:
  - **FBref [(1)]:** Una reconocida fuente de estadísticas de fútbol que abarca una amplia gama de ligas de todo el mundo, tanto masculinas como femeninas. Los datos están disponibles en formato de tablas HTML, lo que facilita su extracción mediante el uso de la biblioteca de pandas en Python.
  - **Transfermarkt [(2)]:** Utilizada para obtener información sobre el valor de mercado de los jugadores y el costo total de cada plantel. Los datos se presentan en formato de tablas HTML y también pueden ser fácilmente extraídos usando pandas.

Características del Conjunto de Datos:

- **Variables Incluidas en FBref:** Las estadísticas utilizadas en este análisis se centran exclusivamente en los pases. Las variables específicas incluyen:
  - **Short_Att (Short Passes Attempted):** Número de pases cortos intentados.
  - **Medium_Att (Medium Passes Attempted):** Número de pases medios intentados.
  - **Long_Att (Long Passes Attempted):** Número de pases largos intentados.
  - **Short_Eff (Short Passes Effectiveness):** Efectividad de los pases cortos dados.

(1) https://fbref.com/ **(2)** https://www.transfermarkt.com/

    - **Medium_Eff (Medium Passes Effectiveness):** Efectividad de los pases medios dados.
    - **Long_Eff (Long Passes Effectiveness):** Efectividad de los pases largos dados.
    - **KP (Key Passes):** Pases clave que crean una oportunidad de gol.
    - **1/3 (Final Third Passes):** Pases en el tercio final del campo.
    - **PPA (Passes into the Penalty Area):** Pases dentro del área de penal.
    - **CrsPA (Crosses into the Penalty Area):** Centros dentro del área de penal.
    - **PrgP (Progressive Passes):** Pases progresivos que avanzan el balón significativamente hacia el arco rival.
- **Variables Incluidas en Transfermarkt:** Información sobre los jugadores y el valor de mercado de los equipos. Las variables específicas incluyen:
    - **Team Name (Nombre del Equipo):** Nombre del equipo.
    - **Market Value (Valor de Mercado):** Valor total de mercado de todos los jugadores del plantel en euros.

Posibles Problemas de Calidad:

- **Datos faltantes:** Es posible que algunas entradas tengan valores faltantes debido a errores de registro o inconsistencias en la recopilación de datos.
- **Valores atípicos:** Puede haber valores atípicos que no representen con precisión el rendimiento típico de un equipo.
- **Inconsistencias:** Diferencias en la definición y registro de las estadísticas entre diferentes partidos o temporadas.

Preprocesamiento de Datos: Para abordar los posibles problemas de calidad, se realizarán las siguientes acciones:

- **Imputación de Datos Faltantes:** Utilizando métodos estadísticos para llenar los valores faltantes.
- **Detección y Tratamiento de Valores Atípicos:** Identificación y, si es necesario, tratamiento de valores que se desvíen significativamente de la media.
- **Normalización:** Estandarización de las variables para asegurar que todas contribuyen equitativamente al análisis de clustering.

3. Preguntas / Problemas a Resolver

En esta sección, se plantea la pregunta o problema que se pretende resolver utilizando el conjunto de datos descrito anteriormente y las herramientas de análisis aprendidas en el curso.

Pregunta Principal: ¿Cómo se pueden segmentar los equipos de la primera división del fútbol argentino según su estilo de juego utilizando estadísticas de pases y análisis de clustering?

Sub Preguntas Específicas:

1. **Identificación de Estilos de Juego:**
    - ¿Qué estilos de juego predominan en una liga basados en las estadísticas de pases?
    - ¿Cómo se pueden agrupar los equipos en clusters que reflejan diferentes formas de intentar llegar al arco rival, como posesión, juego directo y contraataque?
2. **Análisis de Presupuestos y Rendimientos:**
    - ¿Cuál es la relación entre el presupuesto de los clubes y sus estilos de juego?
    - ¿Cómo se compara el rendimiento de un equipo con otros equipos dentro de su cluster y entre clusters?

Relevancia del Problema:

El análisis de los estilos de juego y su relación con el presupuesto y el rendimiento de los equipos es fundamental para comprender las dinámicas de los equipos y económicas de la liga. Este conocimiento puede ser valioso para entrenadores, analistas y gestores de clubes, ayudándolos a tomar decisiones informadas sobre estrategias de juego y planificación a largo plazo.

4. Proceso de Análisis Propuesto

1. **Extracción y Preparación de Datos:**
    - Obtener las estadísticas de pases desde FBref y los valores de mercado de Transfermarkt.
    - Realizar el preprocesamiento de los datos, incluyendo imputación de datos faltantes, tratamiento de valores atípicos y normalización.
2. **Análisis Exploratorio de Datos:**
    - Visualizar las distribuciones de las variables clave.
    - Identificar patrones y correlaciones entre las variables.
3. **Aplicación de Técnicas de Clustering:**
    - Utilizar técnicas de clustering como K-means, PAM y métodos jerárquicos para agrupar a los equipos según sus estadísticas de pases.
    - Evaluar los resultados de clustering utilizando el método del codo y el coeficiente de silueta para determinar el número óptimo de clusters.
    - Calcular los índices de comparación de particiones, como el Adjusted Rand Index (ARI) y la Normalized Mutual Information (NMI), para medir la robustez de los modelos.
4. **Análisis de Clusters:**
    - Describir las características de cada cluster en términos de estadísticas de pases, presupuesto y rendimiento.
    - Comparar los equipos dentro de cada cluster y entre clusters para identificar diferencias y similitudes.
5. **Visualización de Resultados:**
    - Crear gráficos que muestran la distribución de los equipos en los clusters y las relaciones entre las variables clave.
    - Utilizar gráficos de dispersión, diagramas de radar y otras visualizaciones relevantes para presentar los hallazgos.