CALEISEIS
FUNDACIÓN

Módulos de Habilidades Técnicas
13 de noviembre, 2019

ANÁLISIS DE DATOS DE MUNDIALES DE FÚTBOL

POR RICARDO DEL RÍO GUZMÁN
Estudiante de Especialidad en Machine Learning

Profesores:
- Álvaro Fuentes
- Esteban Navarro
- Javier Gonzalez

Índice

Introducción

En este trabajo utilizamos una base de datos compuesta por 3 tablas con información de los mundiales de fútbol de la FIFA. Cada tabla respectivamente contiene datos de los partidos, datos de los jugadores que han participado e información general de cada uno de los mundiales.

El objetivo del proyecto es poder analizar los datos para determinar características comunes de los equipos, jugadores y partidos, seleccionar a los equipos y jugadores con mejores rendimientos, analizar las ciudades y países sede de los encuentros en relación con los años de cada mundial. Y, establecer relaciones entre las características y desempeño de equipos y jugadores.

En una etapa inicial de este proyecto se creó un script que permite almacenar las descripciones y nombres de las columnas, el tipo de dato que almacenan. Y, además, los números, nombres y descripciones de las categorías de las columnas con datos categóricos. Toda esta información queda almacenada en archivos de texto en una carpeta de Google Drive, para no tener que cargar los datos cada vez que se inicia el programa y para que toda esta información no se encuentre mezclada con el código, aumentando considerablemente la extensión y dificultando su legibilidad.

Luego se realizó la importación de todas las librerías Python necesarias. Se definieron funciones y variables para ayudar a reducir la repetición de algoritmos que se deben utilizar varias veces a lo largo de todo el proyecto. Por ejemplo, para la creación y almacenamiento de los gráficos y el cálculo de medidas descriptivas. Se realizan configuraciones del formato de los gráficos y se cargan los datos desde los archivos, lo cuales son convertidos en tablas DataFrames de la librería pandas y a los que se les añade toda la información de las categorías y columnas.

Al cargar los datos se aprecia que hay columnas que no tiene bien definido el tipo de dato que almacenan, por lo que se realizó el cambio correspondiente. La cantidad de asistentes se configuró para poder ser interpretada como datos enteros y la fecha y hora que era interpretada como texto fue configurada para que se almacenara en un formato de fecha y hora entendible por el programa.

A continuación, presentaremos el desarrollo y los análisis de las distintas partes de este proyecto.

Desarrollo

1) Análisis Exploratorio Inicial

Primeramente, se calcularon las medidas descriptivas de todas las columnas de las tablas para poder observar el comportamiento de los datos y los rangos en los que estos se encontraban. Luego se analizaron los datos nulos, donde se observó que cada tabla tenía una particular cantidad y distribución de valores nulos. La tabla de partidos tenía 852 filas con todos sus valores nulos, las que fueron eliminadas. La de jugadores tenía una gran cantidad de nulos en las columnas de posición del jugador y de eventos en los que el jugador había estado involucrado. La eliminación tanto de las filas como de las columnas con dichos valores habría generado la perdida de información valiosa para los análisis posteriores, por lo que se

conservaron y su tratamiento se realizó en casos particulares. La tercera base de datos con la información general de las copas mundiales no tenía valores nulos, por lo que no se realizó ningún tratamiento.

Se encontraron una gran cantidad de columnas idénticas en los datos de las partidas y de los jugadores, por lo que se eliminaron los duplicados.

2) Tratamiento de Outliers

La limpieza de outliers solo se realizó para las columnas con datos numéricos continuos o discretos, dejando de lado todos los datos categóricos. Todos los datos que estaban alejados de la media más de 3 veces la desviación estándar fueron reemplazados por la media de los datos truncados. En el caso de los datos discretos los valores fueron reemplazados por el entero más cercano al resultado del cálculo anterior. Se realizaron gráficos de tipo boxplot para los valores antes y después de la limpieza para tener una representación más visual de los valores más alejados.

3) Visualizaciones

Con el objetivo de comprender la distribución de los valores (rangos y valores más repetidos), para estas primeras visualizaciones se optó por la generación de histogramas de todas las columnas con valores numéricos.

4) Análisis de Concurrencia de Público

Este análisis se realizó para comprender factores que afecten en la asistencia de público a los partidos de fútbol y a los mundiales en general. Para esto se generaron clasificaciones de la cantidad de asistentes en los mundiales cada año y también de asistentes dependiendo de horario del partido.

Con respecto al análisis por año, se observa una clara tendencia al aumento de personas que van a los mundiales desde 1934 al 2014. El mundial con más asistencia fue el de Estados Unidos en 1994 con más de 3 millones y medio de asistentes, seguido por los mundiales de Alemanias (2006) y Brasil (2014). Por otro lado, la copa mundial con menos asistencia han sido la de Italia (1934) y la de Francia (1938), que no lograron superar los 500 mil asistentes.

Al observar la afluencia por horarios se observa que los encuentros con más asistentes se han realizado a las 11:30, 12:30, 14:15 y a las 20:00 horas. Mientras que los partidos con menos público en promedio son aquellos que se han realizado a las 12:50, a las 15:40 y a las 18:10.

5) Visualización y Análisis de Goles Anotados

Se pretende establecer cuáles han sido los mundiales con mayor cantidad de goles y analizar si existe diferencias entre la cantidad de goles que han sido anotados por los equipos de local y los quipos de visita. Para lograr lo primero, se agruparon y sumaron los datos con respecto a los años de realización de cada mundial. Luego se sumaron los goles anotados en el primer tiempo, con aquellos que se marcaron en el

segundo tiempo. Para lo segundo se realizaron histogramas para apreciar la distribución de los goles de los equipos de local y visitas.

Se aprecia una clara tendencia en el aumento de goles por cada mundial a lo largo de los años. Los mundiales en los que se han metido más goles fueron los de 1998 y 2006 con más de 60 goles, seguidos por el mundial de 1958 con más de 50. El mundial con menos goles fue el de 1938, con poco más de 20 goles.

Tanto para los equipos de local como de visita, la cantidad de goles que más se repite en los partidos es de un único gol. En cuanto a los goles de los equipos de local, la moda está seguida por los 2 y 0 goles en un partido. El máximo de goles que ha metido un equipo de local es de 6. Los equipos de visita han tenido tantos partidos sin meter goles como en los cuales han logrado un solo gol. Eso si el máximo de goles que ha apuntado un equipo visita son 4.

6) Resultados por Países

Se busca encontrar los países que han tenido más partidos ganados, perdidos y empatados a lo largo de la historia de los mundiales de fútbol. Para esto se crean nuevas columnas a la tabla de los partidos en las que quedan más evidenciados los equipos ganadores, perdedores y si hubo o no empate en los primeros 90 minutos del partido. Luego se crea una nueva matriz en la que se crea una columna con todos los países y otras columnas con la cantidad de partidos ganados, perdidos y empatados. Como hay países que han participado en más partidos y en más mundiales, para poder comparar la información con otros países se calcularon los porcentajes de partidos ganados, perdidos y empatados.

El país con un mayor porcentaje de partidos ganados es Brasil, seguido por Alemania, Turquía, la República Federal Alemana y Dinamarca. Hay varios países que han perdido todos los partidos que han jugado en un mundial, entre los que podemos mencionar a lo Emiratos Árabes Unidos, la República de Zaire, las Islas Orientales Neerlandesas, Iraq y Togo. Los países con mayor porcentaje de partidos empatados son Israel, Angola, República de Irlanda, Wales y Egipto.

7) Análisis de Año y Ciudad de Eventos Mundiales

Con el objetivo de encontrar las ciudades y años en los que se ha realizado más partidos se utilizó la estructura “pivot table” para filtrar y agrupar los valores por los parámetros mencionados. Se llega a la conclusión de que las ciudades en las que se juegan más partidos son Montevideo en 1930, Johannesburgo en 2010, Ciudad de México en 1986 y en Buenos Aires en 1978.

8) Análisis de Jugadores

En esta sección se intentan generar categorías de jugadores basado en su rendimiento y en su participación para poder establecer en base a distintos criterios los mejores y peores jugadores. Para esto se realizan diversas agrupaciones y filtrajes que luego son graficados para favorecer el análisis.

Primeramente, se analizó la participación de cada jugar en todos los mundiales a lo largo de la historia, sin importar si participaron como titulares o como suplentes. En esta categoría se logra establecer que Ronaldo, Klose, Cafu Sepp Maier y Dida son quienes han estado presentes en más instancias de la copa. Luego se determinó a lo jugadores que han sido más veces titulares en mundiales, donde se repiten los nombres de Ronaldo y Klose, pero el tercer lugar pasa a ser ocupado por Diego Maradona, seguido de Uwe Seeler y Wladyslaw Zmuda.

Para el análisis de rendimiento se consideraron 3 factores principales, la cantidad de anotaciones (contando goles y penales), la cantidad de puntos perdidos por autogoles o fallos en tiros penales y por último, según las tarjetas que ha recibido el jugador.

Quienes más anotaciones han logrado en los mundiales son Ronaldo, Klose y Gerd Mueller. Hay una gran cantidad de jugadores que han perdido un penal, pero nadie a perdido más de uno en un mundial. Con respecto a las tarjetas, quienes más advertencias y expulsiones han recibido son Paredes, V. Bronckhorst y Boulahrouz.

9) Dependencia Lineal Entre Variables

Aquí se intenta establecer relaciones entre elementos de una misma tabla y entre elementos de tablas distintas, con el objetivo de determinar las características de los equipos ganadores de partidos y la relación entre los jugadores goleadores y equipos ganadores. Se realizan varias combinaciones entre tablas y se generan matrices de correlación para ver de forma gráfica aquellos elementos con mayor correlación lineal.

Conclusiones

Este proyecto permitió reforzar los contenidos de tratamiento de nulos, tratamiento de outliers, la importancia de los análisis exploratorios, transformación y mapeo de datos. Además, generó aprendizajes sobre la importancia de corroborar la duplicidad de datos dentro de una tabla.

Respecto a los resultados se puede concluir que hay una clara tendencia al aumento de personas que van a los mundiales desde 1934 al 2014. El mundial con más asistencia fue el de Estados Unidos en 1994 con más de 3 millones y medio de asistentes, seguido por los mundiales de Alemanias (2006) y Brasil (2014). Por otro lado, la copa mundial con menos asistencia han sido la de Italia (1934) y la de Francia (1938), que no lograron superar los 500 mil asistentes.

Se observa que los encuentros con más asistentes se han realizado por la mañana o en las primeras horas de la tarde. No se logra establecer una regla para los partidos de la tarde y tarde-noche pues aquellos horarios con más público se encuentran intercalados con aquellos con menor afluencia.

Se aprecia una clara tendencia en el aumento de goles por cada mundial a lo largo de los años. Los mundiales en los que se han metido más goles fueron los de 1998 y 2006 con más de 60 goles, seguidos por el mundial de 1958 con más de 50. El mundial con menos goles fue el de 1938, con poco más de 20 goles.

Se aprecia una pequeña ventaja para los equipos que juegan de local, porque estos han logrado más goles a lo largo de la historia que aquellos que juegan de visita. Pero en ambos casos la cantidad de goles que se ha anotado en más partidos es 1.

El país con un mayor porcentaje de partidos ganados es Brasil, seguido por Alemania. Hay varios países que han perdido todos los partidos que han jugado en un mundial. Los países con mayor porcentaje de partidos empatados son Israel, Angola y República de Irlanda.

Las ciudades en las que se juegan más partidos son Montevideo en 1930, Johannesburgo en 2010, Ciudad de México en 1986 y en Buenos Aires en 1978.