UNIVERSIDAD DEL VALLE DE GUATEMALA
CC3066 - Data Science
Sección 10
Ing. Lynette Garcia

Proyecto 2. Análisis Exploratorio

Fernando Garavito, 18071
Oliver de León, 19270
Daniela Batz, 19214

GUATEMALA, 19 de septiembre de 2022

SITUACIÓN PROBLEMÁTICA

¿Qué se necesita para ser un jugador de fútbol profesional?

Fue lo que la Bundesliga se preguntó hace unos cuantos meses. Probablemente el dinero, una edad juvenil y un entrenamiento exhaustivo sean suficientes para cumplir los requisitos para llegar a formar parte de una liga profesional. Sin embargo, hemos de notar que la próxima superestrella del fútbol ciertamente empezará su camino a través de las ligas juveniles o semi-profesionales. Estas ligas con normalidad mantienen una cantidad de recursos limitados dada su naturaleza, incluyendo el material/data que potencialmente brindará un insight significativo en el desempeño de los nuevos jugadores.

En la actualidad esta data es obtenida de forma manual, lo que lleva a que su recolección implique una gran cantidad de pasos y personas implicadas. Los altos costos de estas desgastantes prácticas limitan su uso solamente para competencias profesionales, lo que potencialmente perjudica el desarrollo de muchas nuevas estrellas del fútbol.

PROBLEMA CIENTÍFICO

Dada la problemática planteada, la Deutsche Fußball Liga (DFL) ha decidido lanzar un reto en torno a la visión artificial, analizando y detectando data de las bases de video más grandes del mundo, con el fin de proveer un algoritmo de gran rapidez y profundidad que permita analizar diversas condiciones de juego (pases, saques de banda, centros, etc.) y poner bajo el foco a jugadores con un gran potencial.

OBJETIVOS

- **Objetivo General**
  - Desarrollar un modelo de visión artificial computarizado, mediante el cuál se puedan analizar, detectar y clasificar de manera automatizada condiciones de juego a través del uso de datasets de videos de fútbol.
- **Objetivo Específico**
  - Detectar pases de fútbol, saques de banda, centros y desafíos en partidos originales.
  - Comunicar los resultados de la detección de condiciones por medio de recursos estadísticos visuales, a través del análisis y clasificación del dataset.

DESCRIPCIÓN DE LOS DATOS

	video_id	time	event	event_attributes
0	1606b0e6_0	200.265822	start	NaN
1	1606b0e6_0	201.150000	challenge	['ball_action_forced']
2	1606b0e6_0	202.765822	end	NaN
3	1606b0e6_0	210.124111	start	NaN
4	1606b0e6_0	210.870000	challenge	['opponent_dispossessed']
...	...	...	...	...
11213	ecf251d4_0	3056.587000	challenge	['opponent_dispossessed']
11214	ecf251d4_0	3058.072895	end	NaN
11215	ecf251d4_0	3068.280519	start	NaN
11216	ecf251d4_0	3069.547000	throwin	['pass']
11217	ecf251d4_0	3070.780519	end	NaN

imagen 1: previsualización del dataset train.csv

El dataset estudiado denominado bajo el nombre **train.csv** cuenta con un total de **11,218** filas y **4** columnas.

Cuenta con las siguientes **variables**:

- **video_id:** ID del video al que se le analizan eventos, tiempo y atributos.
- **time:** Tiempo en segundos en el que el evento ocurre.
- **event:** Tipo de evento de ocurrencia.
- **event_attributes:** Información adicional para la ocurrencia de eventos.

Operaciones de limpieza
No fueron necesarias operaciones de limpieza.

Nombre de la variable	Cualitativa		Cuantitativa	
	Ordinal	Nominal	Discreta	Continua
video_id		✓		
time				✓
event		✓		
event_attributes		✓		

ANÁLISIS EXPLORATORIO

Para realizar un análisis exitoso de nuestro dataset, hemos de estudiarlo desde los grandes rasgos y buscar una culminación introspectiva del mismo. Como primer paso hemos de indicar el número de vídeos estudiados y su respectiva cantidad de entradas, dichas entradas se describen a continuación:

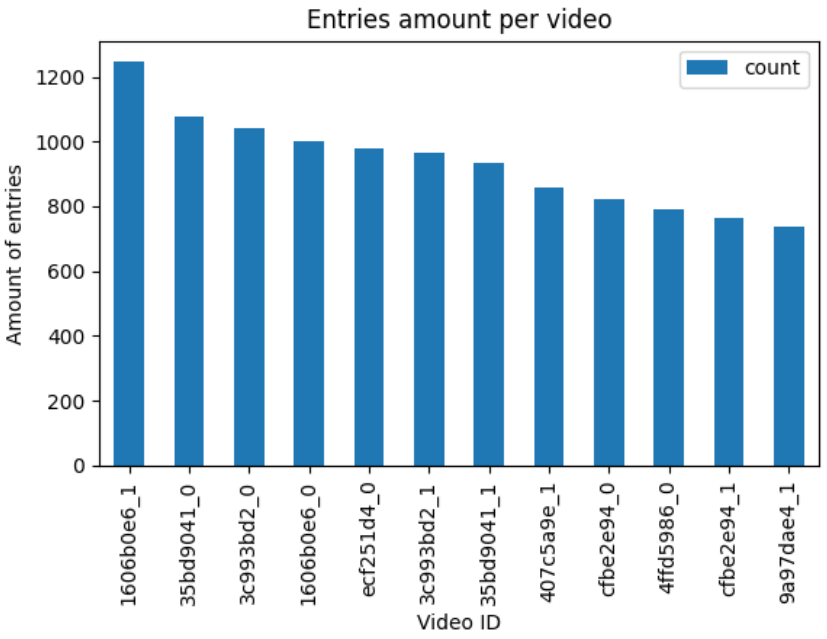

Gráfico 1: cantidad de entradas por video

El gráfico 1 nos muestra una cantidad de **12 videos** y su respectiva cantidad de entradas asociadas. El **número promedio de entradas** asciende a aproximadamente **935** por cada video del dataset.

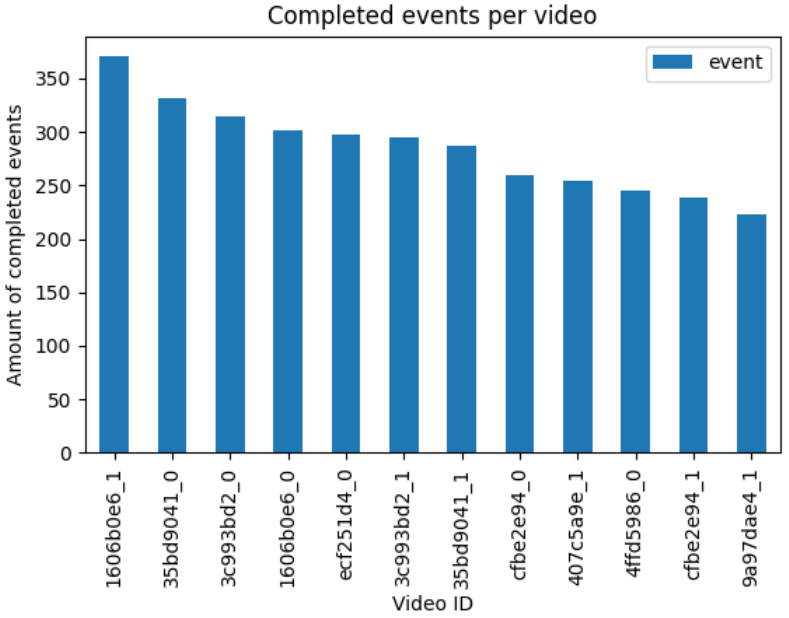

Gráfico 2: cantidad de eventos completados por video

Una vez analizadas las entradas, se identificó cuántas de estas representaban **eventos completos**, por lo que se inspeccionó la columna de eventos con el fin de delimitar la completación de los mismos bajo el tag de “end”. Una vez delimitados, estos se describen en el **gráfico 2.** El número de eventos completos promedio por video asciende a aproximadamente **285** eventos.

Con la intención de mensurar temporalmente la cantidad de eventos, se procedió a analizar el tiempo que cada video registra en materia de eventos, tomando como inicio los múltiples eventos de apertura y como conclusión los eventos de clausura. Los tiempos de eventualidad registrados por cada video se describen a continuación, con un promedio de **870 segundos** por grabación (aproximadamente unos **14 minutos y medio** por video).

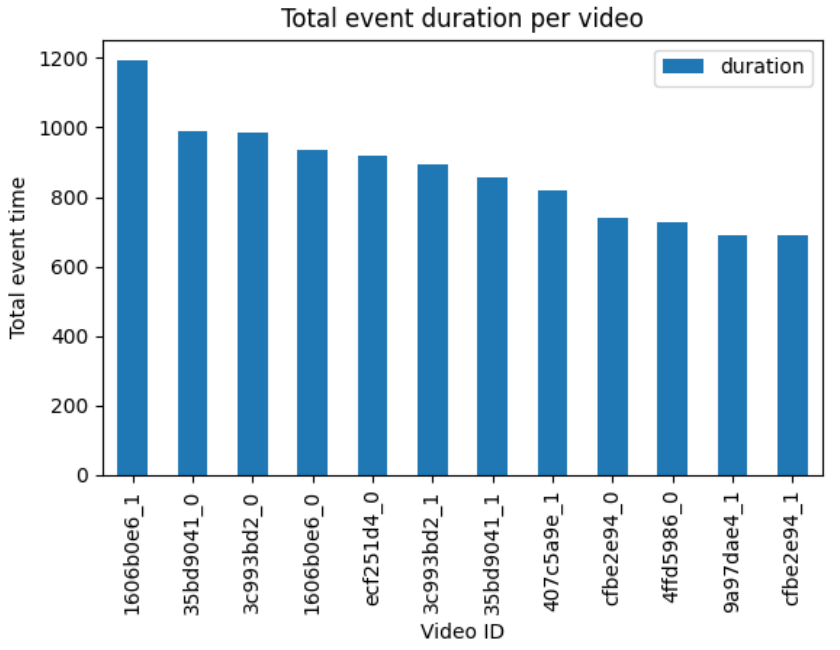

Gráfico 3: Duración total de los eventos por video en segundos

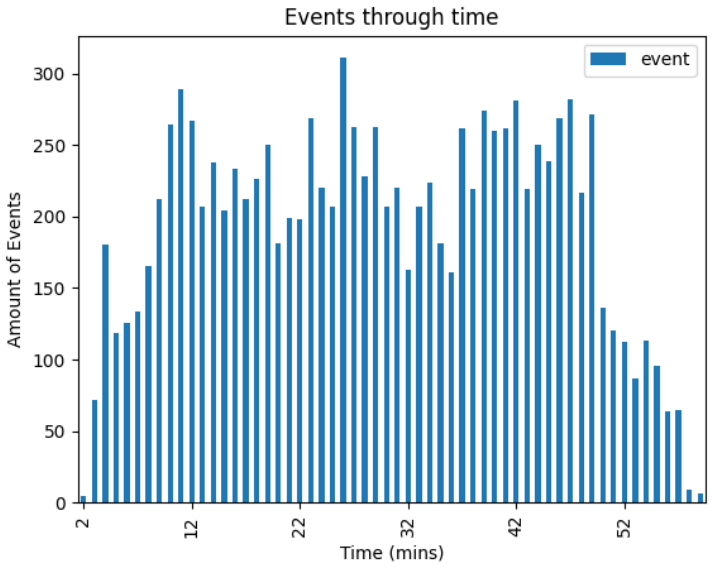

Gráfico 4: Eventos a través del partido

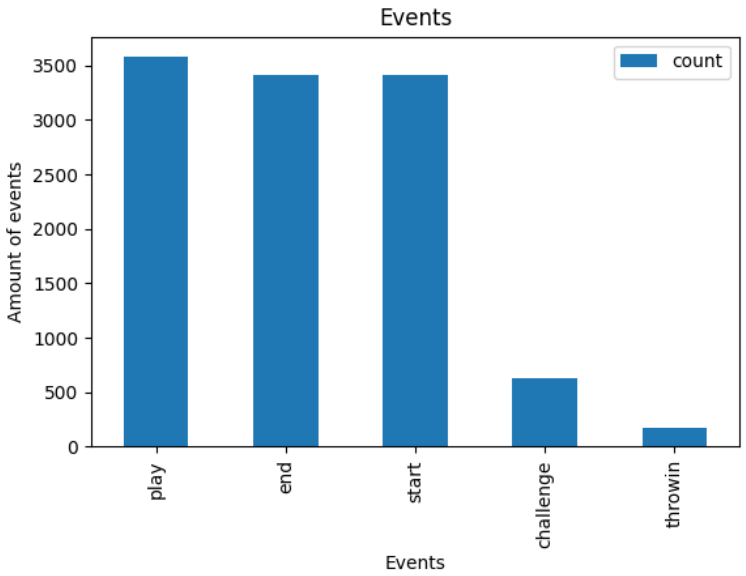

Gráfico 5: Eventos del dataset

Finalmente, se estudiaron la cantidad de eventos producidos a lo largo de todo el dataset. Esto con el fin de fundamentar las mensuraciones en torno a la proporción de distribución de ocurrencia. Se identificó que se realizan en promedio **187 eventos** por minuto.

HALLAZGOS Y CONCLUSIONES

- Se determinó que solamente el 25% del tiempo de un partido de fútbol representa datos significativos.
- Se determinó que la actividad u ocurrencia de eventos decae a medida que el tiempo transcurre, sugiriendo un posible agotamiento por parte de los participantes
- Se concluye que la cantidad de tiempo transcurrido en un partido es proporcional al número de eventos realizados.

Github: https://github.com/Jos260400/Proyecto-2-Data-Science
Dataset: https://www.kaggle.com/competitions/dfl-bundesliga-data-shootout
Presentación:https://docs.google.com/presentation/d/1tBohHCkWXE8JwZpNvCo8M_ew1KFTpFfsYfOARPOY9Co/edit?usp=sharing