Predicción del Torneo de Baloncesto NCAA 2017

Integrantes del grupo:
Andrés Darío Peña
Andrés Ramirez Aristizabal
Leider Caicedo Palacios

Materia:
Modelos y Simulación 1

Docente:
Raúl Ramos Pollán

Universidad de Antioquia
Facultad de ingeniería
Octubre 2023

Planteamiento del problema.

Predecir los resultados de un torneo de baloncesto universitario masculino en Estados Unidos en el año 2017. Utilizaremos los resultados recopilados de torneos pasados para poder construir y probar diferentes modelos de predicción. Después pronosticamos los resultados de todos los enfrentamientos posibles en el torneo de 2017.

Simulación del dataset para requerimientos:

Se realizaron las simulaciones de los requerimientos pedidos para la base de datos seleccionada, los cuales fueron:

El 5% de los datos de al menos 3 columnas deben ser faltantes: Con este ciclo se ejecutó dicha simulación de datos faltantes en las columnas Lftm, Ldr y Lfga3

```
[ ]  for i in range(int(len(cdb1) * 0.05)):
         x = np.random.randint(len(cdb1))
         cdb1.loc[x, "Lftm"] = np.nan
         cdb1.loc[x, "Ldr"] = np.nan
         cdb1.loc[x, "Lfga3"] = np.nan
```

Este es el resultado presenciado donde el 3% de los datos de las columnas mencionadas (3744) se presentan como nulos gracias a la información proporcionada por la función isna de pandas dataframe.

```
[ ] k = cdb1.isna().sum()

[ ] k

Season       0
Daynum       0
Wteam        0
Wscore       0
Lteam        0
Lscore       0
Wloc         0
Numot        0
Wfgm         0
Wfga         0
Wfgm3        0
Wfga3        0
Wftm         0
Wfta         0
Wor          0
Wdr          0
Wast         0
Wto          0
Wstl         0
Wblk         0
Wpf          0
Lfgm         0
Lfga         0
Lfgm3        0
Lfga3     3744
Lftm      3744
Lfta         0
Lor          0
Ldr       3744
Last         0
Lto          0
Lstl         0
Lblk         0
Lpf          0
dtype: int64
```

Al menos un 10% de las columnas han de ser categóricas: El 10% de las 34 columnas pertenecientes al data frame equivalen a 3 columnas redondeadas a entero (3,4 originalmente). Las columnas simuladas a ser establecidas como categóricas han sido: “Numot”, “Daynum” y “Wloc”.

```
[17] cdb1["Numot"] = pd.Categorical(cdb1.Numot)
     cdb1["Daynum"] = pd.Categorical(cdb1.Daynum)
     cdb1["Wloc"] = pd.Categorical(cdb1.Wloc)
```

```
[21] cdb1.columns
```

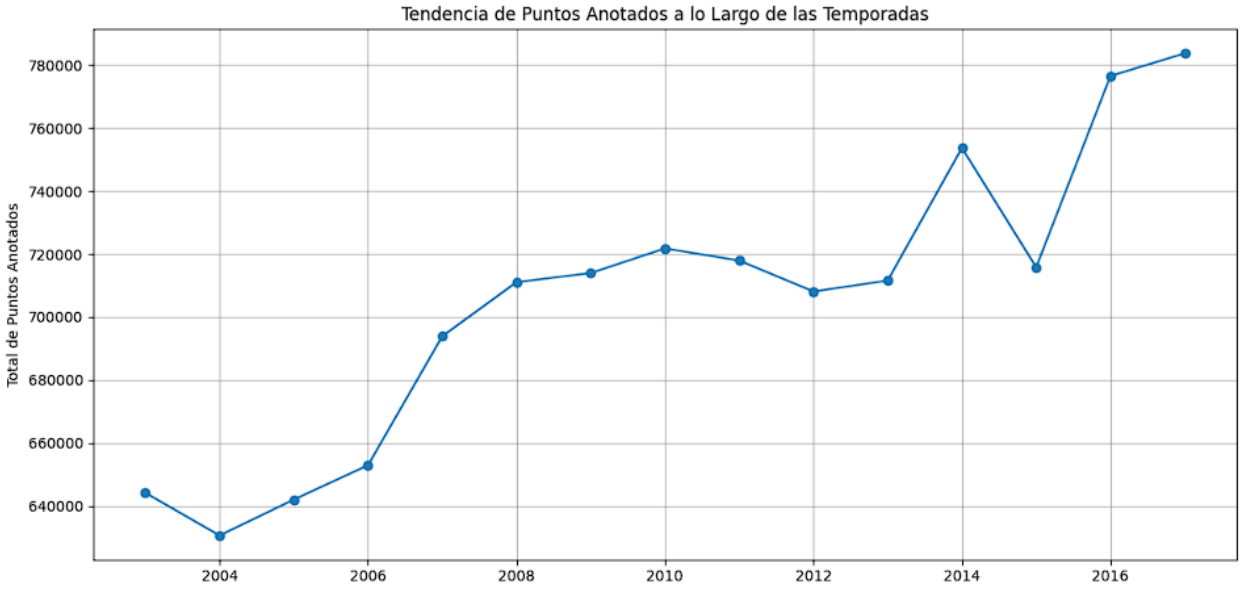
```
Index(['Season', 'Daynum', 'Wteam', 'Wscore', 'Lteam', 'Lscore', 'Wloc',
       'Numot', 'Wfgm', 'Wfga', 'Wfgm3', 'Wfga3', 'Wftm', 'Wfta', 'Wor', 'Wdr',
       'Wast', 'Wto', 'Wstl', 'Wblk', 'Wpf', 'Lfgm', 'Lfga', 'Lfgm3', 'Lfga3',
       'Lftm', 'Lfta', 'Lor', 'Ldr', 'Last', 'Lto', 'Lstl', 'Lblk', 'Lpf'],
      dtype='object')
```

```
[22] cdb1.dtypes
```

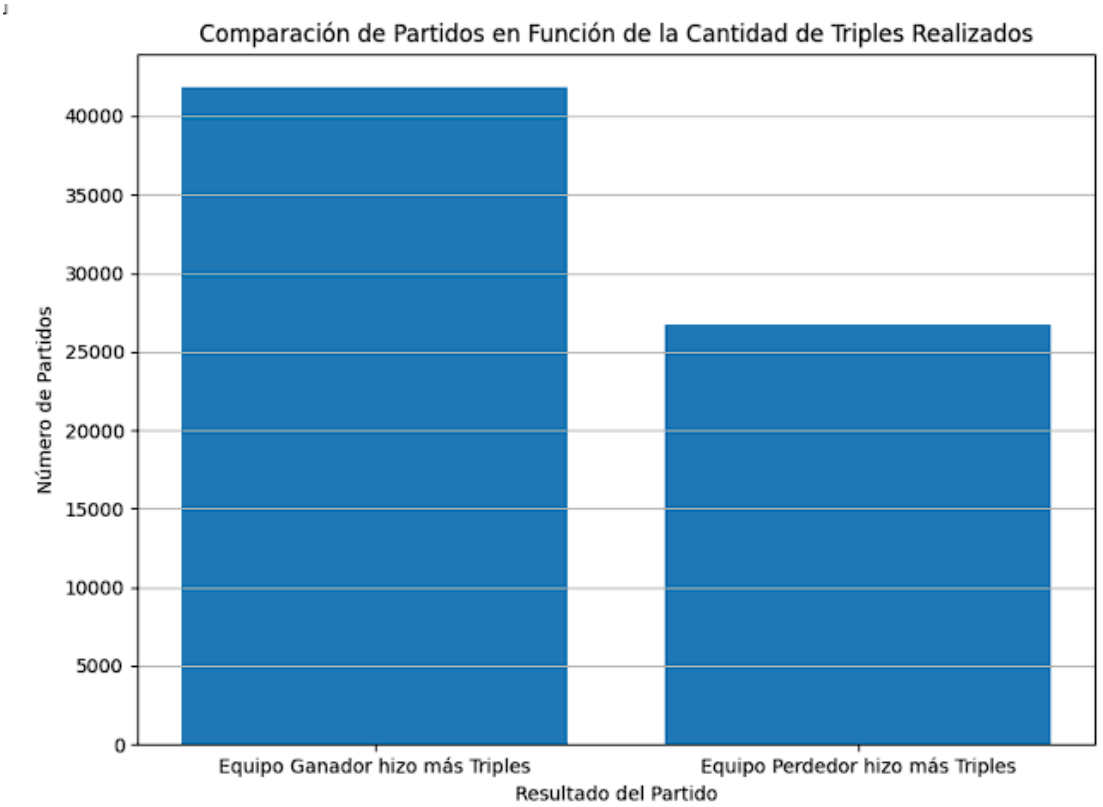
```
Season       int64
Daynum    category
Wteam        int64
Wscore       int64
Lteam        int64
Lscore       int64
Wloc      category
Numot     category
Wfgm         int64
Wfga         int64
Wfgm3        int64
Wfga3        int64
Wftm         int64
Wfta         int64
Wor          int64
Wdr          int64
Wast         int64
Wto          int64
Wstl         int64
Wblk         int64
Wpf          int64
Lfgm         int64
Lfga         int64
Lfgm3        int64
Lfga3      float64
Lftm       float64
Lfta         int64
Lor          int64
Ldr        float64
Last         int64
Lto          int64
Lstl         int64
Lblk         int64
Lpf          int64
dtype: object
```

Exploración de datos. Inicialmente tomamos el conjunto de datos que seleccionamos para analizarlo, revisamos las distintas variables que lo conforman, los datos asociados a estas variables y la relación entre estos datos, todo esto para definir nuestro modelo de forma eficiente y acertada.

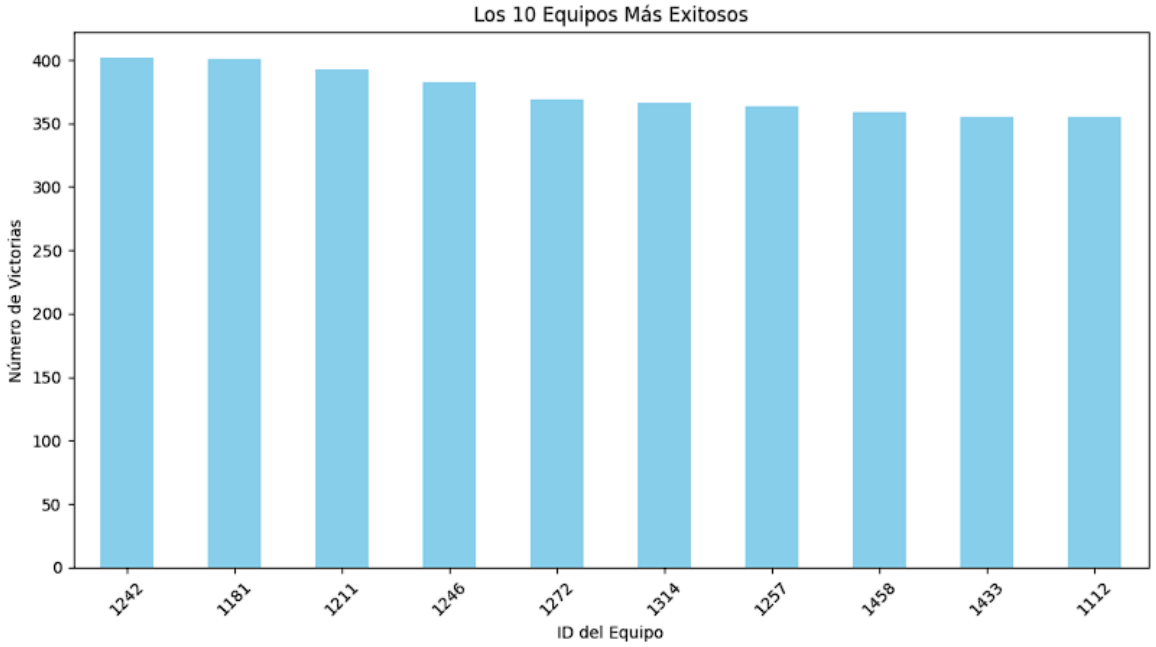
Gráficos obtenidos. Con el análisis y el proceso anterior realizado en los datos hemos obtenido algunos gráficos en los que se reflejan estadísticas comparativas entre partidos como la cantidad de puntos anotados, los triples anotados, faltas, robos, etc. También se observan otras estadísticas relacionadas a los resultados de los encuentros.

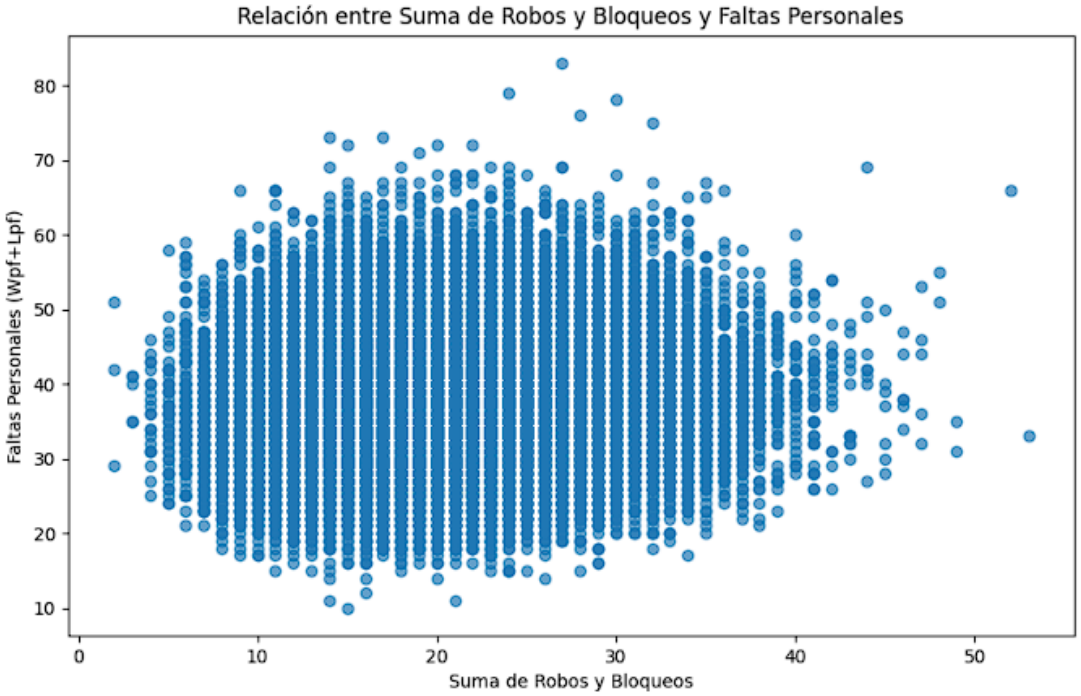


Observaciones: A medida que transcurren los años el baloncesto se vuelve un deporte más ofensivo, un deporte en el que se convierten cada vez más puntos , aunque existe una caída en los puntos en el año 2015 en comparación con el año anterior(2014).



Este gráfico muestra que los equipos ganadores en general convierten más triples que los equipos perdedores.





En este gráfico de dispersión, los bloqueos (wblk+lblq) más los robos (wstl+lstl) se representan en el eje x, y las faltas personales (wpf+lpf) en el eje y. Cada punto en el gráfico representa un partido ganado por un equipo. Si se observa una tendencia en la que equipos con más faltas personales también tienen más robos y bloqueos, podría sugerir que están dispuestos a asumir más riesgos en defensa para intentar obtener ventajas en el juego.

Se realiza la **limpieza de datos** al eliminar las columnas innecesarias “Daynum” y “Numot” las cuales contienen el día en que se jugó el partido y Numot el cual indica el número de períodos de tiempo extra en el juego.

```
#Limpieza de datos
RegularSeasDetai = RegularSeasDetai.drop(["Daynum", "Numot"], axis=1)
RegularSeasDetai
```

	Season	Wteam	Wscore	Lteam	Lscore	Wloc	Wfgm	Wfga	Wfgm3	Wfga3	...	Lfga3	Lftm	Lfta	Lor	Ldr	Last	Lto	Lstl	Lblk	Lpf
0	2003	1104	68	1328	62	N	27	58	3	14	...	10.0	16.0	22	10	22.0	8	18	9	2	20
1	2003	1272	70	1393	63	N	26	62	8	20	...	24.0	9.0	20	20	25.0	7	12	8	6	16
2	2003	1266	73	1437	61	N	24	58	8	18	...	26.0	14.0	23	31	22.0	9	12	2	5	23
3	2003	1296	56	1457	50	N	18	38	3	9	...	NaN	NaN	15	17	NaN	9	19	4	3	23
4	2003	1400	77	1208	71	N	30	61	6	14	...	16.0	17.0	27	21	15.0	12	10	7	1	14
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
76631	2017	1276	71	1458	56	N	27	48	10	23	...	NaN	NaN	8	14	NaN	10	15	4	3	13
76632	2017	1343	71	1463	59	N	25	52	11	26	...	20.0	13.0	19	14	20.0	12	7	4	5	13
76633	2017	1348	70	1433	63	N	24	54	8	20	...	14.0	17.0	22	23	24.0	8	5	4	1	16
76634	2017	1374	71	1153	56	N	26	52	10	19	...	NaN	NaN	18	17	NaN	7	7	7	1	13
76635	2017	1407	59	1402	53	N	21	60	1	17	...	17.0	7.0	8	9	27.0	10	17	1	7	18

Este gráfico te ayuda a visualizar la relación entre estas variables y evaluar si existe una correlación entre las faltas personales, los robos y los bloqueos en los equipos ganadores.

Preprocesado de datos. Después realizaremos un preprocesado de datos, limpiando y puliendo un poco el dataset, estableciendo un conjunto de datos sólido y completo, teniendo en cuenta la usabilidad de variables categóricas, la eliminación de columnas innecesarias y el filtrado de datos valiosos para las predicciones; entre otras cosas a tomar en cuenta.