Proyecto predicción de partidos Rocket League

Diego Soto Curilén

https://github.com/disotoc/RLCSAnalysis21-22

CODER HOUSE

Introducción

Rocket League es un videojuego que combina elementos de fútbol y carreras de autos, en pocas palabras es un juego donde se debe jugar fútbol (metiendo goles en el arco contrario) con autos que al ser propulsados por nitro puede incluso volar. Hay distintos modos de juego online y offline, pero este proyecto se centrará en el modo principal en e-sport que es en equipos de 3 vs. 3

Este proyecto está enfocado en realizar un análisis predictivo para saber los ganadores de partidos y resolver un problema de clasificación del dataset que tiene datos de los *encuentros de los equipos extraídos del siguiente link

https://www.kaggle.com/datasets/dylanmonfret/rlcs-202122

(*) Cada encuentro tiene de 5 a 7 partidos, el ganador del encuentro es el primero en ganar 2 o 3 partidos respectivamente

Encuentros por equipo

El primer set de datos analizado corresponde a los encuentros por equipo, inicialmente tenía 54 columnas y luego de una limpieza quedó solo con 25 (5 categóricas; 19 numéricas; 1 booleana), la más importante de todas es la variable llamada “winner” que identifica si el encuentro se ganó o perdió.

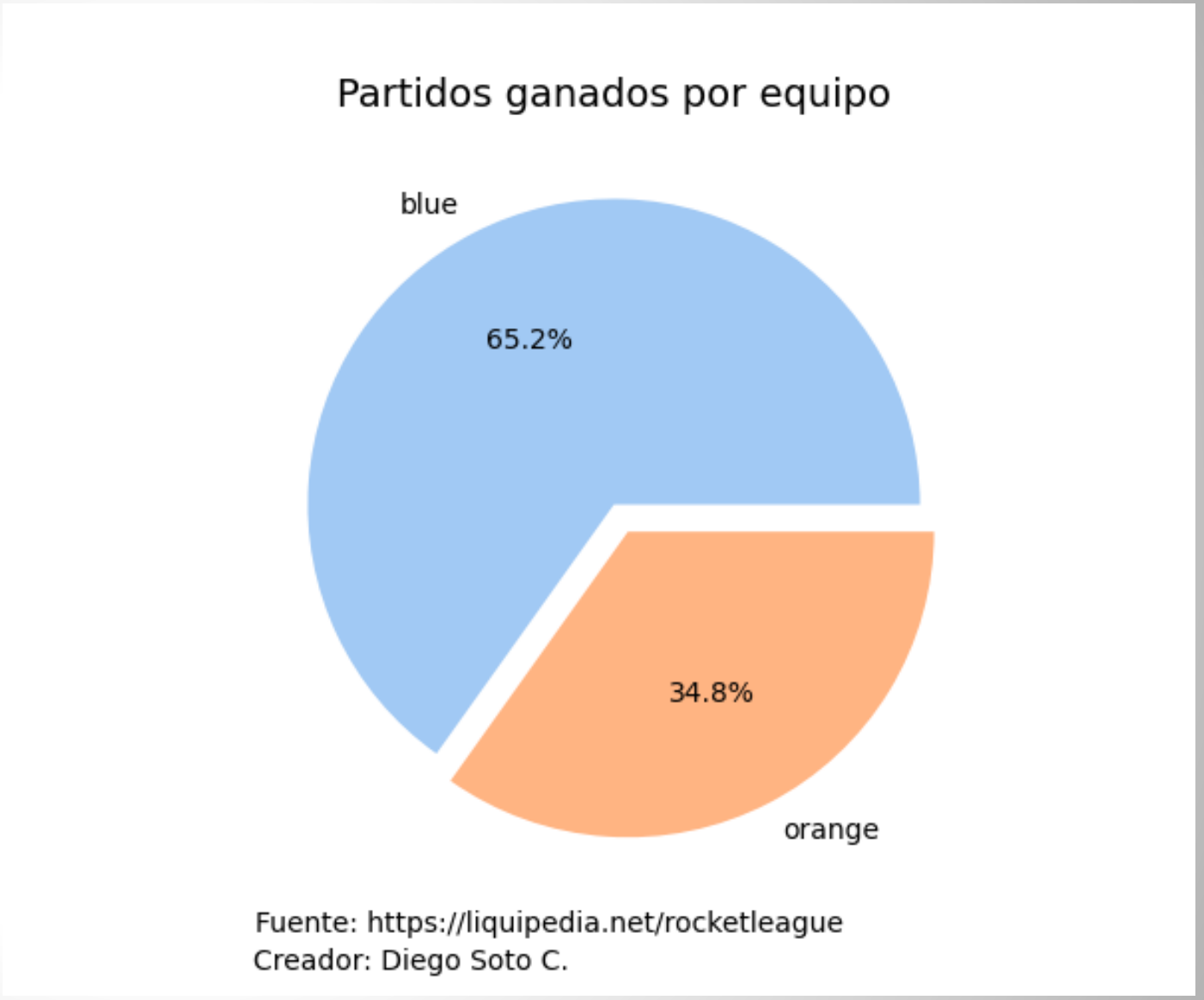
Además, se verificó que hay algunas columnas con datos nulos que representan en general un **18.2%**

Hipótesis

Inicialmente, hay algunas hipótesis para las que podemos buscar respuestas y se responderán con algunos gráficos en las siguientes diapositivas

¿Hay alguna inclinación en cuanto a los partidos ganados por esquina (azul/naranja) dada?

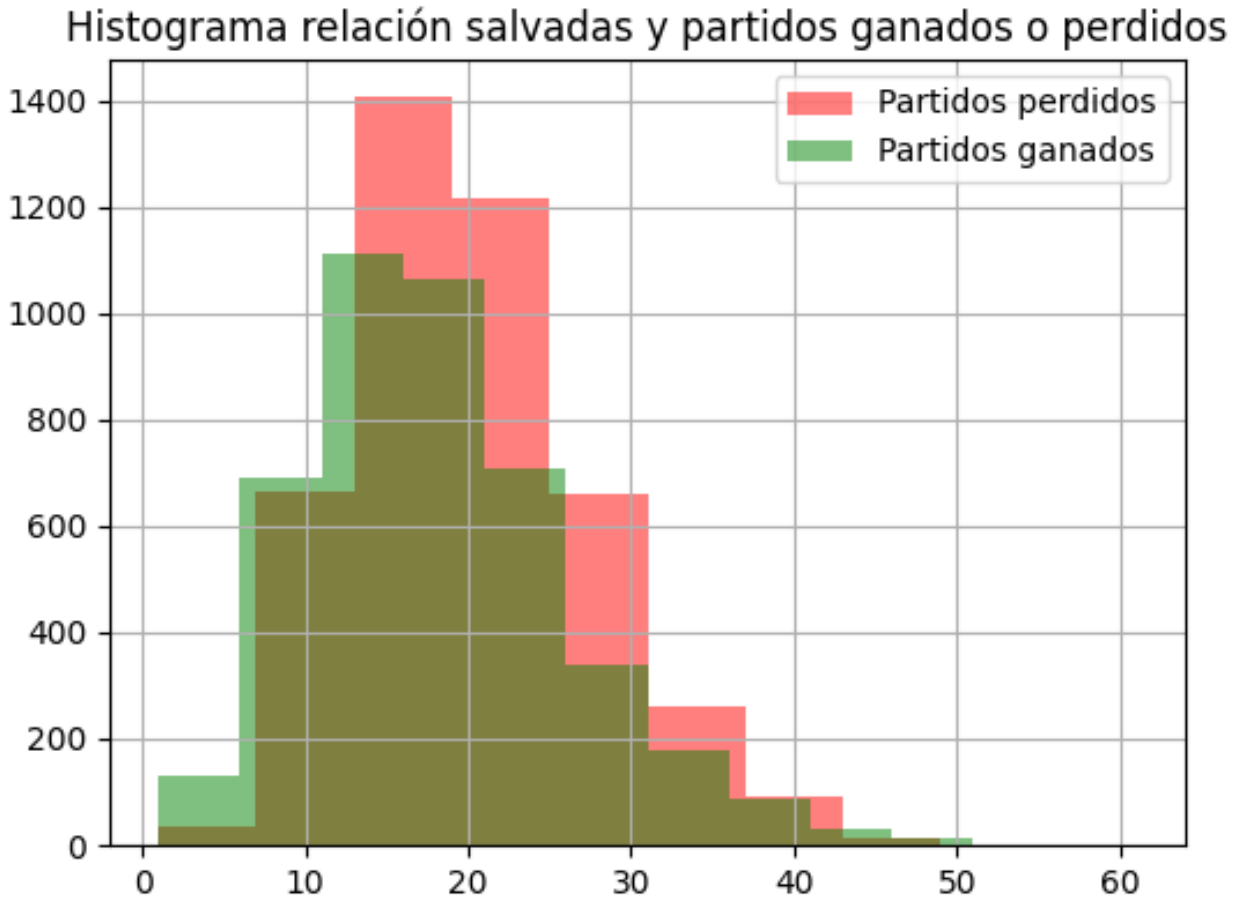
Podemos concluir que hay una clara inclinación que gane la esquina azul, sin embargo, no podemos asegurar que no se deba a factores externos, de todas maneras, se podría utilizar este parámetro para el modelo de datos que se creará.



¿Hay una relación entre las salvadas o salvadas épicas realizadas y los partidos ganados?

Por ahora, no podemos concluir que hay una relación directa o indirecta en cuanto a los partidos ganados o perdidos versus la cantidad de salvadas.

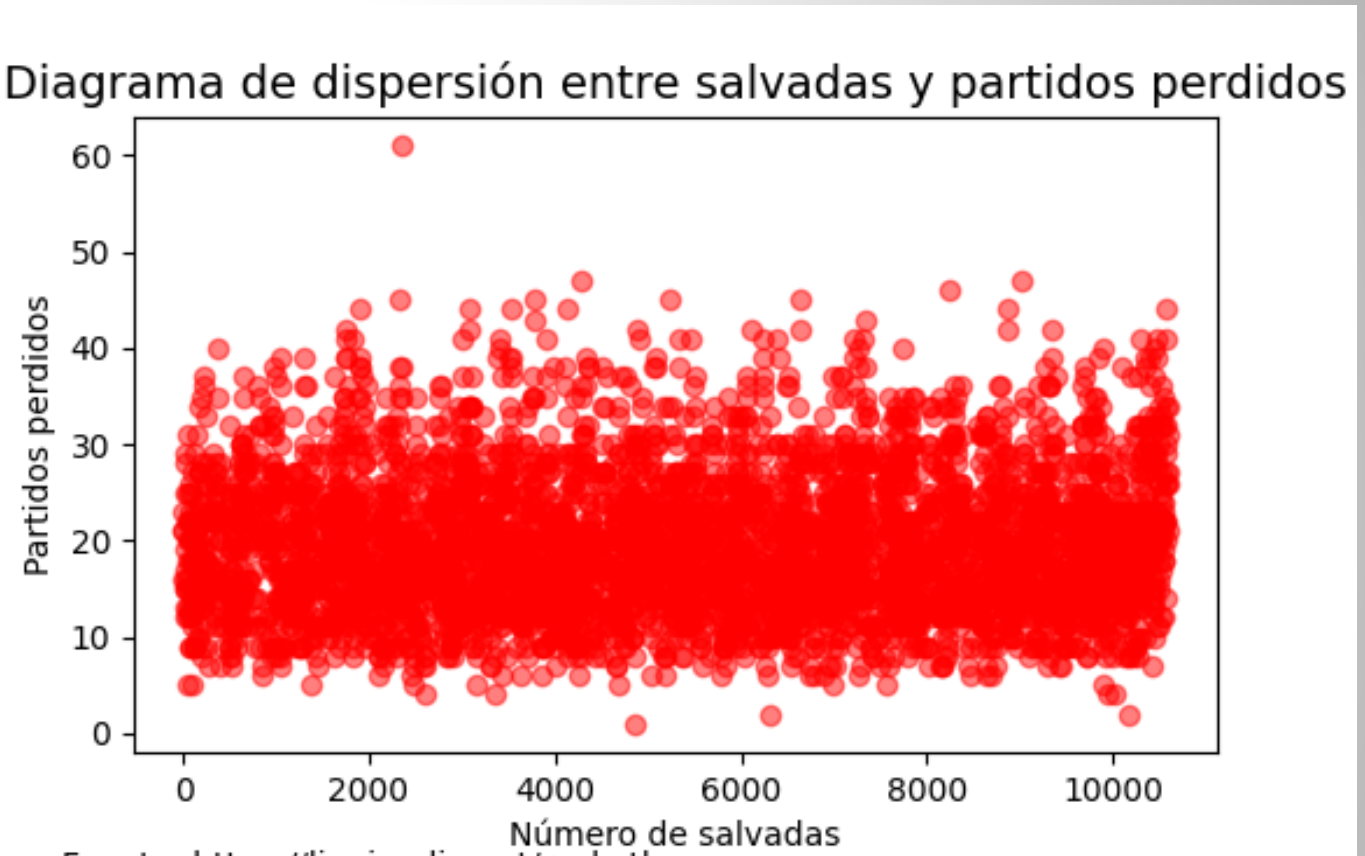
De todas maneras, hay frecuencias más altas en los partidos perdidos, esto puede deberse a que realmente hubo más salvadas porque también hubo más defensa, por lo tanto, se podría decir que, la cantidad de partidos perdidos aumenta mientras más salvadas haya porque la mayor parte del tiempo se defendieron de los ataques del oponente.



¿Hay una relación entre las salvadas o salvadas épicas realizadas y los partidos ganados?

Para comprobar lo que pasó en el punto anterior se optó por hacer un diagrama de dispersión que muestre si hay una relación entre los partidos perdidos y las salvadas.

Gracias a este gráfico podemos decir que no hay correlación entre los partidos perdidos y las salvadas realizadas, por lo tanto, se puede descartar.

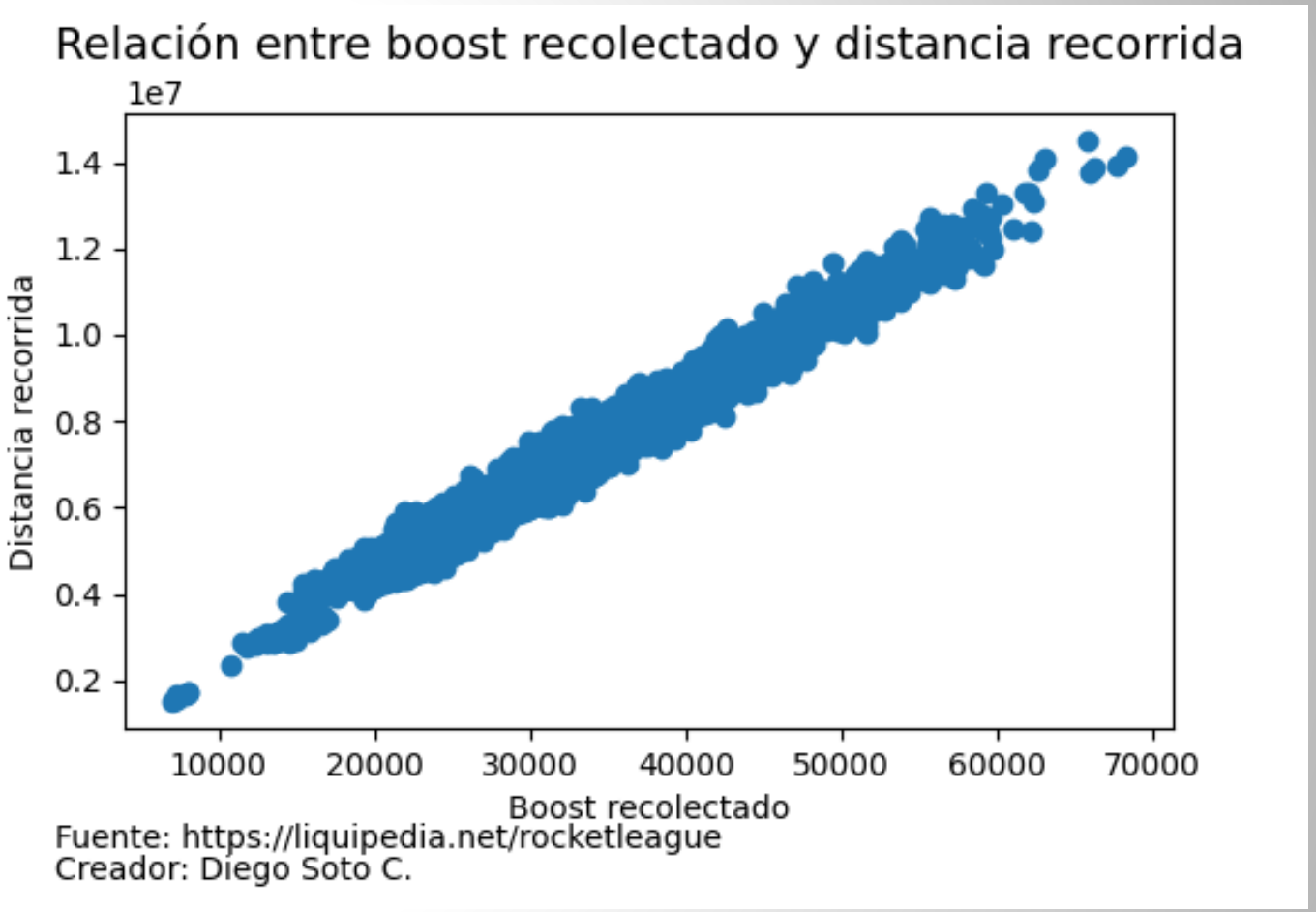


Fuente: https://liquipedia.net/rocketleague
Creador: Diego Soto C.

¿Se podría decir que mientras más distancia se recorre, más boost se utiliza?

Podemos ver una clara correlación entre el *boost recolectado y la distancia recorrida en el campo.

(*) Dentro del juego hay ciertos espacios de la cacha en la cual hay esferas de nitro para poder ir más rápido o tener la opción de volar, al nitro entregado se le llama comúnmente boost.



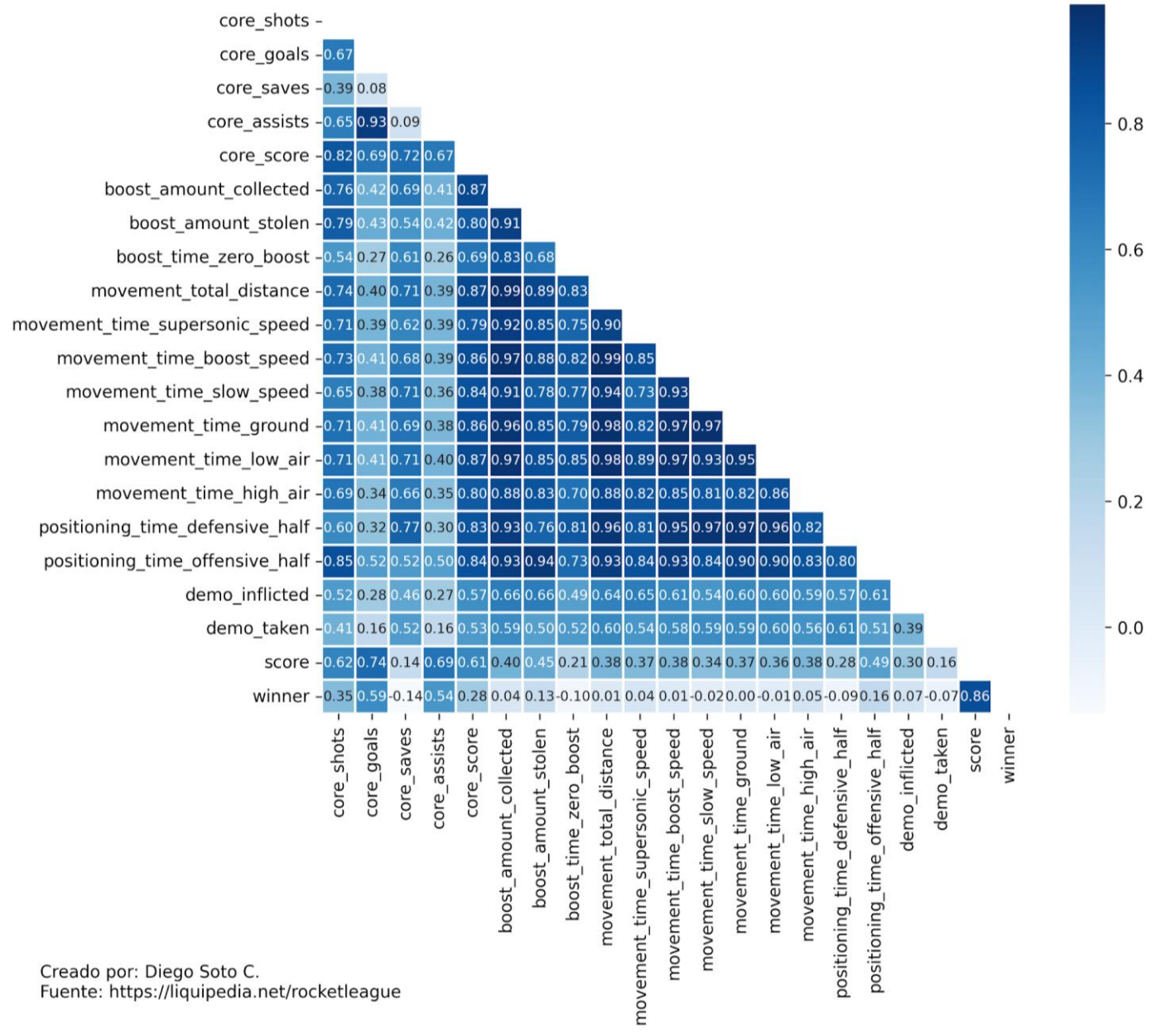
Equipos más ganadores

Luego de analizar lo anterior, el siguiente paso fue verificar los equipos con más partidos ganados.

Los resultados dan a pensar que los primeros 2 equipos pueden ser los ganadores del torneo, sin embargo, ya que es un torneo pasado, ya se saben los resultados, el ganador fue "TEAM BDS" que está en tercer lugar con más partidos ganados, y el que está en segundo lugar "TOKYO VERDY ESPORTS", en el general quedó solo entre los 20 mejores, por lo tanto, puede ser que no tenga mucha relación si lo vemos desde ese punto de vista.

Ahora bien, se podría hacer el mismo análisis separado por región. Ambos son líderes en su región, y realmente podemos validar que los resultados si son distintos, y probablemente se deba a la cantidad de equipos que hay en cada región sea distinta

team_name	winner	team_region
G2 ESPORTS	62	North America
TOKYO VERDY ESPORTS	60	Asia-Pacific North
TEAM BDS	56	Europe
FURIA ESPORTS	56	South America
PIRATES EXDEE	53	Sub-Saharan Africa
NRG ESPORTS	51	North America
TEAM FALCONS	50	Middle East & North Africa
RENEGADES	49	Oceania
DETONATOR	48	Asia-Pacific North
FAZE CLAN	47	North America



Mapa de correlaciones

Con el fin de ayudar un poco en la búsqueda de relaciones se generó un mapa de calor, en primer lugar, se eliminaron todas las columnas con variables string y luego se tomaron todas las otras variables.

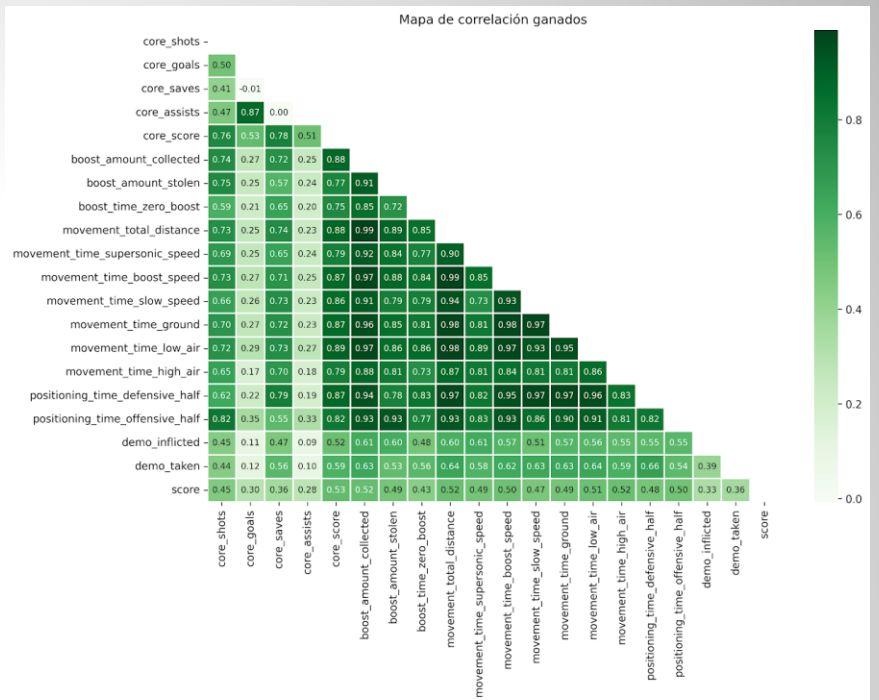
Como el proyecto trata sobre una predicción de partidos ganados, la variable más importante para verificar es “winner”, y se alcanza a ver que tiene una gran relación con la variable “score”, esto tiene sentido ya que esta última mide la cantidad de partidos ganados en el encuentro, fuera de eso, también “core_goals” y “core_assists” también tienen una relación positiva que también hace mucho sentido al ser tanto los goles, como las asistencias, un factor determinante para ganar encuentros.

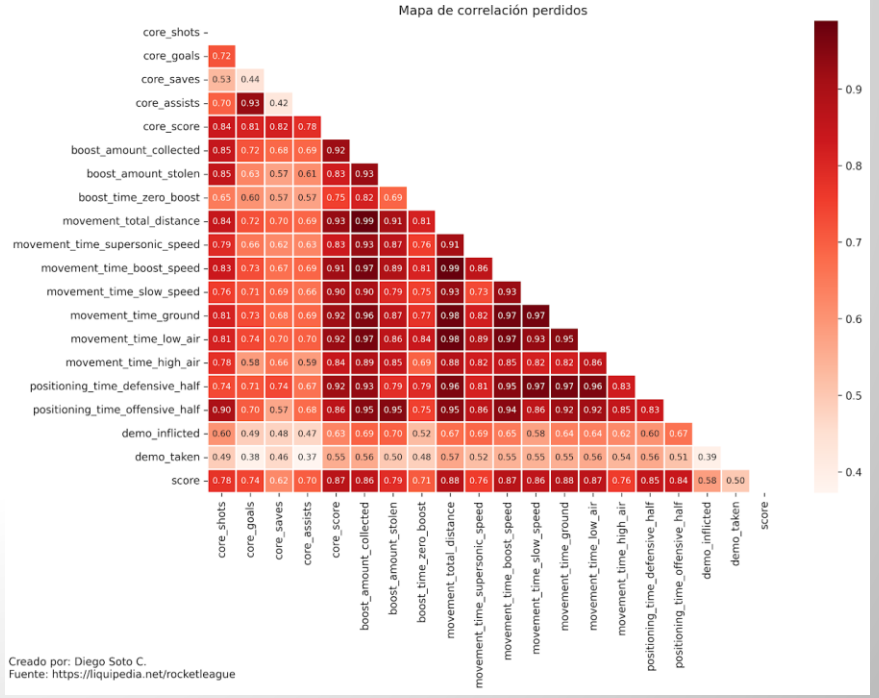
Mapa de correlaciones

Independiente de lo anterior, la variable principal tiene pocas correlaciones, esto se puede deber a que es una variable booleana, debido a esta razón se generó otro gráfico a parte donde se separan los partidos ganados y los partidos perdidos para ver las mismas correlaciones, los resultados están en el gráfico de la izquierda:

A simple viste se logran apreciar unos fenómenos que deberían ser investigados más en profundidad, y son:

- Los goles realizados tienen mayores correlaciones en los partidos perdidos, esto por lógica debería ser al revés.
- Al igual que el punto anterior, también las correlaciones en las asistencias son mayores en los partidos perdidos.
- Tanto las demoliciones realizadas tienen más correlaciones en los partidos ganados, esto tiene más sentido, en especial en las demoliciones realizadas, pero puede ser un factor determinante para el modelo de aprendizaje.





Elección de modelo

Para elegir el modelo, en primer lugar, debemos recordar que esto se realizará con un modelo de **clasificación** en base a esto, se realizaron pruebas con diferentes modelos, entre ellos:

- KNC
- Regresión logística
- Random Forest
- SVC

Elección de modelo

Se evaluaron diferentes métricas con todos los modelos.

En base a los resultados, podemos tener las primeras conclusiones:

- En primer lugar, podemos descartar el modelo SVC ya que tiene los indicadores más bajos.
- El modelo de Random Forest en todas las evaluaciones tiene cálculos cercanos al 100%, por ende, podemos decir que tiende al overfitting, y también se descartará.
- Entre los 2 que quedan, el que tiene mejores números es **LogisticRegression**, por lo tanto, se trabajará con él.

Modelo: **KNeighborsClassifier**
Accuracy: 0.6418121755545069
Precisión: 0.6389396709323584
Recall: 0.6575729068673566
F1: 0.6481223922114048
Varianza: 0.2499142054106734
Sesgo: 0.20778761061946902

Modelo: **LogisticRegression**
Accuracy: 0.727017461066541
Precisión: 0.6940894568690096
Recall: 0.8174976481655691
F1: 0.7507559395248379
Varianza: 0.24217897164138835
Sesgo: 0.2671386430678466

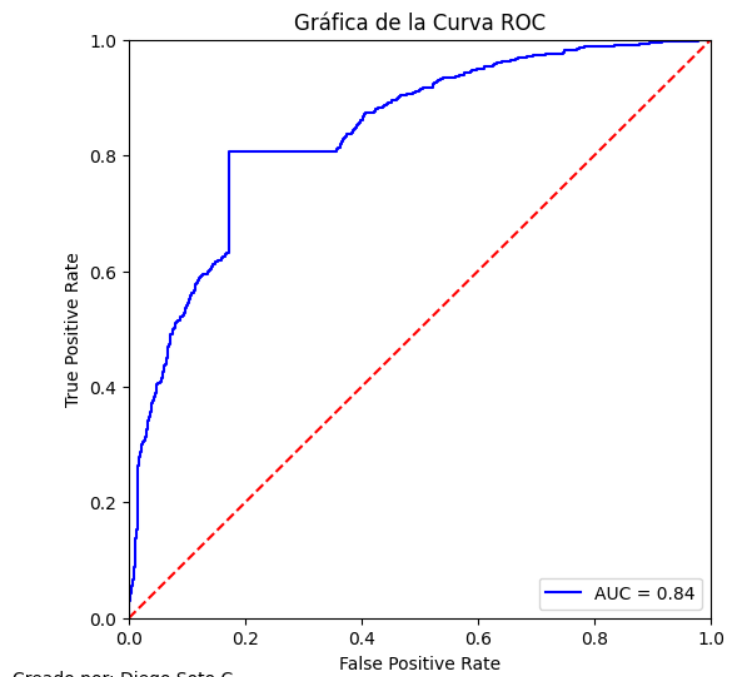
Modelo: **RandomForestClassifier**
Accuracy: 0.997168475696083
Precisión: 0.9953139643861293
Recall: 0.9990592662276576
F1: 0.9971830985915493
Varianza: 0.24999957884111687
Sesgo: 0.00011799410029498526

Modelo: **SVC**
Accuracy: 0.4789995280792827
Precisión: 0.4861205145565335
Recall: 0.6754468485418627
F1: 0.5653543307086615
Varianza: 0.2040599577100791
Sesgo: 0.49262536873156343

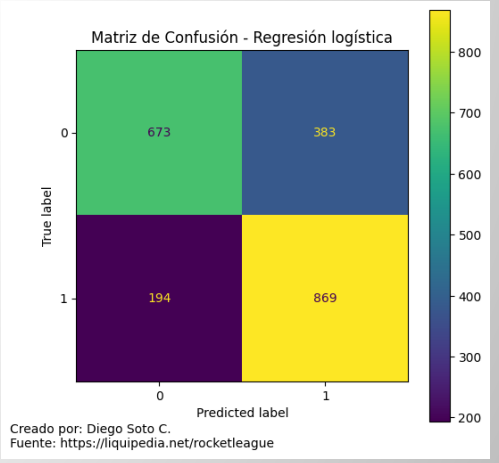
Logistic Regression

En base a sus resultados, podemos decir lo siguiente:

- Tiene una buena efectividad, pero debemos recodar que este es solo la punta del iceberg y es un indicador que tiende a engañar un poco.
- La precisión indica el porcentaje de las predicciones positivas sean correctas, por lo tanto, llevado a este caso indica que de los partidos que el modelo predijo como "partidos ganados" tuvo un **69,4%** de acierto.
- El recall indica el porcentaje en que el modelo puede predecir resultados positivos, es decir, en este caso se indica que, de todos los partidos ganados, acertó un **81,7%**
- El valor F1 es una combinación entre precisión y recall, ya que, en rigor, la precisión busca reducir los falsos positivos y el recall los falsos negativos. En base a esto, podemos decir que el porcentaje de **75,1%** de esta métrica sugiere que hay un buen equilibrio entre ambos porcentajes.
- La varianza tiene como resultado un **24.2%** lo que sugiere que los valores predichos están relativamente cerca del valor medio.
- El Bias con un porcentaje de **26.7%** indica que los resultados están relativamente cerca del valor real.
- El gráfico de la curva ROC indica que el modelo tiene un buen rendimiento.



Creado por: Diego Soto C.
Fuente: https://liquipedia.net/rocketleague



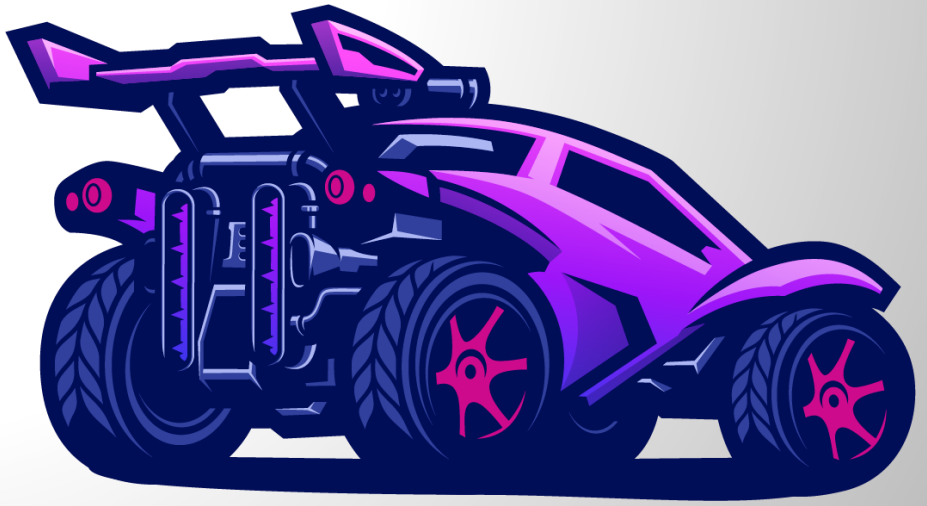
Creado por: Diego Soto C.
Fuente: https://liquipedia.net/rocketleague

Cross Validation - Leave One Out Cross-Validation

Para realizar una validación cruzada en este modelo, se estará ejecutando con Leave One Out Cross-Validation con base en las métricas anteriores, además considerando que la demora es mayor, con ayuda de **%%time** calculamos el tiempo que demora en medir. Los resultados.

Gracias a esto podemos decir que la validación con LOOCV se demora más de 18 minutos por la cantidad de datos. Ahora bien, hay un aumento en todas las métricas exceptuando F1-score, donde realmente es bastante mínimo el cambio, debido a los resultados podemos decir que el modelo funciona de muy buena manera y nos serviría para realizar una predicción de los partidos ganados.

Modelo: **LogisticRegression**
Accuracy promedio: 0.744289220313385
Precisión promedio: 0.8294317538229187
Recall promedio: 0.9148574664904663
F1 promedio: 0.744289220313385
CPU times: total: 2min 18s
Wall time: 18min 22s



CODER HOUSE