MACC
Matemáticas Aplicadas y
Ciencias de la Computación

Análisis de estadísticas y predicción de resultados del juego League of Legends

Ana María Garzón Sánchez , Gabriela Linares Chávez , Kiara Nicole Velásquez y Juan Manuel Dávila Rivera
Estudiantes de Matemáticas Aplicadas y Ciencias de la Computación - Universidad del Rosario - Bogotá D.C, Colombia

RESUMEN El presente proyecto pretende ofrecer una herramienta que pueda encontrar qué campeones representan una ventaja a la hora de competir, haciendo uso de datos de partidas de alto nivel y de los atributos de los personajes elegidos para jugar en cada posición. Por otro lado, buscamos predecir los resultados de una partida a partir de las estadísticas de sus jugadores.

INTRODUCCIÓN

League of Legends (LoL) es un juego del género MOBA, desarrollado por Riot Games para Microsoft Windows y OS X. Este videojuego se enfoca en la experiencia multijugador, donde el modo principal de juego consiste en un combate entre dos equipos, cada uno compuesto por cinco jugadores. Es un juego de alta competitividad siendo uno de los juegos más populares de los Deportes electrónicos (e-sports), por lo que ofrece una amplia variedad de datos para analizar. Dada su naturaleza, tiene varios factores y variables que pueden afectar el resultado de una partida, entre estos se encuentran los 160 campeones con diversas habilidades que a su vez caen en seis categorías generales, estas siendo asesino, luchador, mago, soporte, tanque y tirador dependiendo del estilo de juego del personaje, por lo que pueden pertenecer a más de una a la vez. También se juegan posiciones establecidas en el mapa, estas siendo top, mid, bottom, soporte y jungla, donde ciertas categorías son más beneficiosas.

OBJETIVOS

- Predecir la victoria en una partida a partir de los datos de los equipos participantes.
- Clasificar los campeones de acuerdo a sus características.
- Identificar si existen diferencias significativas entre equipos y campeones dadas sus estadísticas.
- Encontrar si hay campeones que representen una ventaja a la hora de competir.

IMPLEMENTACIÓN

Preparación de los datos

La base de datos tomada estaba distribuida en 86 archivos csv, cada uno con 1000 observaciones. Para poder trabajar con todo el conjunto de datos , lo primero que se hizo fue concatenar estos archivos y transformarlos de un formato .json, a formato csv mediante el uso de un script de Python. Lo siguiente fue hacer una limpieza de los datos, eliminando columnas que no servirían para el análisis posterior; reduciendo el dataset primario de 87 variables a 52. Además de esto se agregó la columna de winrate (tasa de victorias) por campeón ya que esta era requerida. Para cada uno de los análisis se modificó el dataset dependiendo de las necesidades de las pruebas.

Es importante mencionar también que esta base de datos sólo contiene datos de partidas jugadas recientemente (un máximo de seis meses), en la última versión del juego, donde se encuentra la adición del último personaje agregado al juego y estas partidas son de nivel profesional.

Clasificación y análisis de categorías de campeón

Manova Como se había mencionado previamente, el juego cuenta con seis categorías de personajes: Asesino, Luchador, Mago, Soporte, Tanque y Tirador. Un campeón pertenece a una categoría según sus estadísticas y el estilo de juego que favorece.

El primer test realizado fue un MANOVA para indagar si existen diferencias significativas entre las diferentes categorías ya mencionadas, y en caso de que existan indagaremos cuales son.

```
                  Df Pillai approx F num Df den Df    Pr(>F)
independent_var    5 2.2643   6.0986     95    700 < 2.2e-16 ***
Residuals        154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dado que

$$valor - p < 2.2e - 16$$

se dice que existe una diferencia estadísticamente significativa entre las diferentes categorías. Ya que se ha comprobado que si

hay diferencias, se procede a realizar el análisis ANOVA de cada una de las variables para identificar en qué variables específicas se encuentran diferencias o semejanzas.

Anova La prueba ANOVA se realizó con cada una de las variables que corresponden a las estadísticas de los campeones, a partir de esta se demostró que existen tres variables que no presentan alguna diferencia significativa:

- Puntos de maná (mp)

```
 Response 3 :
                Df   Sum Sq  Mean Sq F value    Pr(>F)
(Intercept)      1 22266698 22266698 37.7222 6.655e-09 ***
independent_var  5  4618075   923615  1.5647    0.1733
Residuals      154 90903359   590282
```

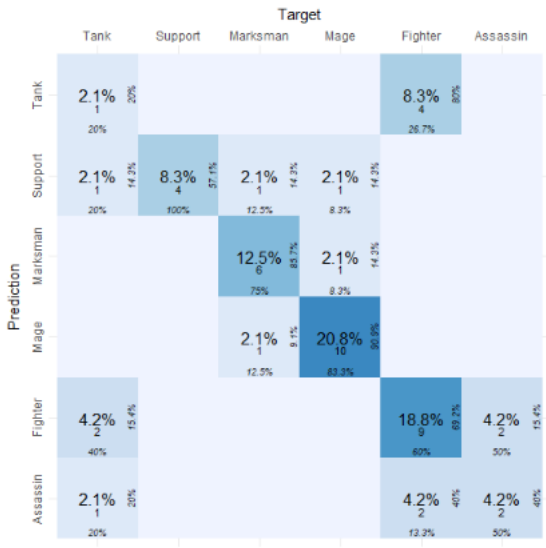
- Armadura por nivel (armorperlevel)

```
 Response 7 :
                Df Sum Sq Mean Sq    F value Pr(>F)
(Intercept)      1 3456.3  3456.3 12540.2775 <2e-16 ***
independent_var  5    1.6     0.3     1.1293 0.3472
Residuals      154   42.4     0.3
```

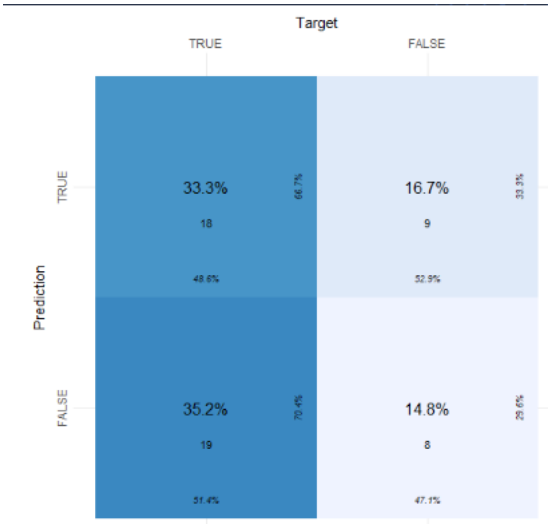
- Regeneración de maná (mpregen)

```
 Response 13 :
                Df  Sum Sq Mean Sq  F value Pr(>F)
(Intercept)      1 10745.7 10745.7 163.9270 <2e-16 ***
independent_var  5   363.9    72.8   1.1102 0.3573
Residuals      154 10095.0    65.6
```

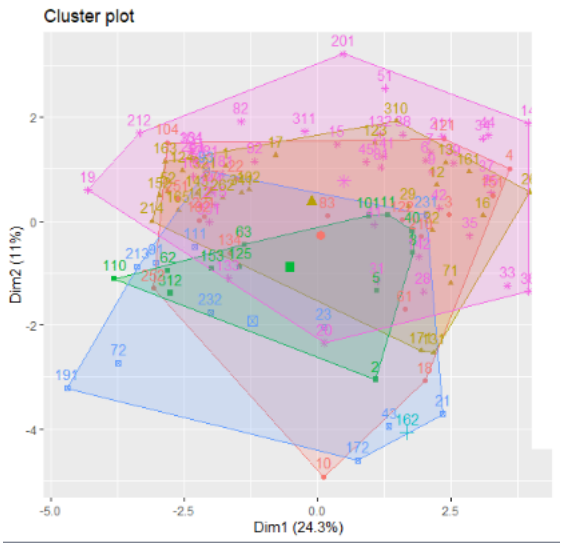
Clasificación por categorías A continuación se realizó la clasificación por categorías, donde se implementó LDA y se graficó su respectiva matriz de confusión como se muestra en la siguiente imagen:



Se tiene que hacer la salvedad de que se contaba con apenas 160 campeones, por lo que la clasificación donde la cuenta era menor se dio de mejor manera. La categoría que se clasificó por completo correctamente fue aquella de los supports, con el 100 porciento clasificado de cuatro sujetos tomados. Los más difíciles de clasificar fueron tanto los "Tanks" como los "Fighter", esto se debe a su similitud en habilidades donde un personaje puede caer en ambas categorías. Seguido a esto se identificaron los personajes que tenían una tasa de victoria (partidas ganadas / partidas jugadas) mayor a 0.5 y los que tenían una tasa de victoria menor. Se dividieron en dos grupos, y se realizó una clasificación para ver si un jugador ganaba más o menos de la mitad de las veces. Realizada una prueba ANOVA, se identificó que no hay diferencias significativas en estos grupos. Así, se puede deducir objetivamente que no hay personaje que sea mejor o peor a la hora de buscar una victoria, por lo que se dice que el juego está bastante balanceado.



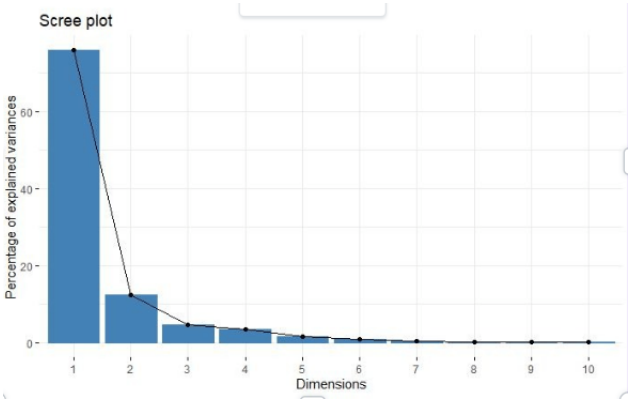
Finalmente, se buscó realizar una agrupación por categorías mediante clustering, y ver si se obtenían resultados similares o diferentes a las categorías propuestas por el juego. Las agrupaciones presentadas no corresponden en alguna medida a las propuestas por el juego, ni son buenos indicadores para proponer una separación diferente.



	1	2	3	4	5	6
Assassin	5	3	0	1	1	3
Fighter	4	5	5	0	3	14
Mage	3	5	1	0	3	12
Marksman	4	5	4	0	4	3
Support	1	3	1	0	1	3
Tank	1	5	1	0	0	8

Clasificación entre equipos que ganan y pierden

PCA Para dar inicio a este análisis, se realizó un PCA para delimitar qué variables eran significativas. Este PCA arrojó que la primera componente ya explicaría el 75 porciento de los datos, en esta componente podemos encontrar el oro y la experencia ganados por el equipo.



Manova Con base en el PCA ya realizado, se procede a realizar varios test de MANOVA, cuatro en específico, y se muestra que el valor-p es extremadamente pequeño en todos los test. Esto indica que existe una diferencia estadísticamente significativa en los datos individuales de los jugadores según la victoria o derrota del su equipo.

```
> summary(manova_model, test = "Pillai", intercept = TRUE)
                    Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)          1 0.97297   239863     26 173245 < 2.2e-16 ***
independent_var      1 0.82826    32136     26 173245 < 2.2e-16 ***
Residuals       173270
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(manova_model, test = "Wilks", intercept = TRUE)
                    Df    Wilks approx F num Df den Df    Pr(>F)
(Intercept)          1 0.027029   239863     26 173245 < 2.2e-16 ***
independent_var      1 0.171736    32136     26 173245 < 2.2e-16 ***
Residuals       173270
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(manova_model, test = "Hotelling-Lawley", intercept = TRUE)
                    Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
(Intercept)          1           35.998   239863     26 173245 < 2.2e-16 ***
independent_var      1            4.823    32136     26 173245 < 2.2e-16 ***
Residuals       173270
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(manova_model, test = "Roy", intercept = TRUE)
                    Df    Roy approx F num Df den Df    Pr(>F)
(Intercept)          1 35.998   239863     26 173245 < 2.2e-16 ***
independent_var      1  4.823    32136     26 173245 < 2.2e-16 ***
Residuals       173270
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

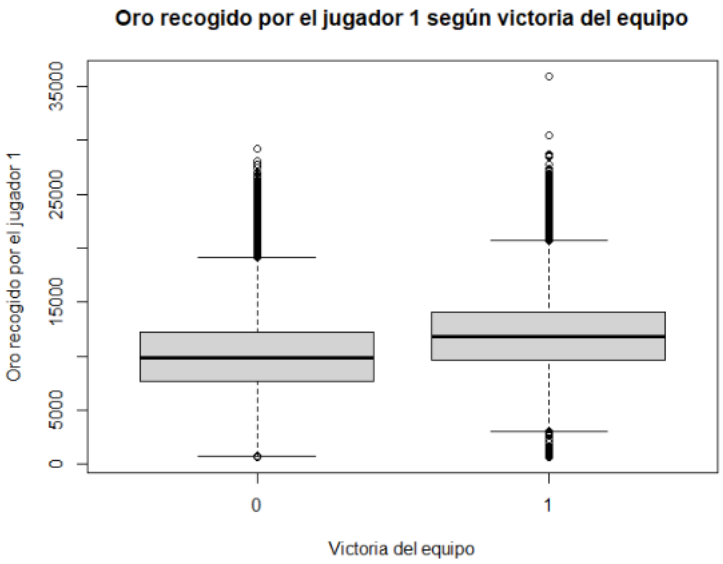
Anova Dado que existen diferencias significativas, se procede a realizar un test ANOVA a cada variable siginificativa para hallar las similitudes y diferencias que sean de importancia para dictar la victoria o derrota del equipo para cada una de las variables.

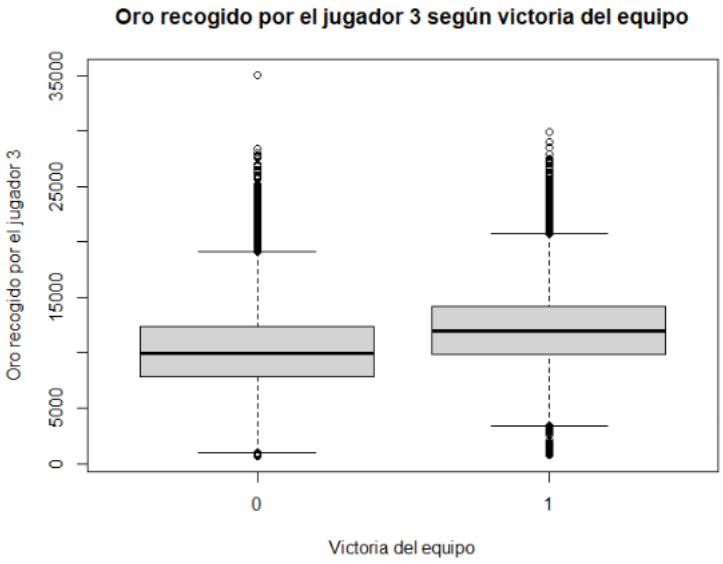
```
> summary.aov(manova_model, test = "Pillai", intercept = TRUE)
 Response 1 :
                    Df     Sum Sq    Mean Sq F value    Pr(>F)
(Intercept)          1 4.9225e+14 4.9225e+14 2496017 < 2.2e-16 ***
independent_var      1 3.9167e+12 3.9167e+12   19860 < 2.2e-16 ***
Residuals       173270 3.4171e+13 1.9721e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 2 :
                    Df     Sum Sq    Mean Sq F value    Pr(>F)
(Intercept)          1 2.1088e+13 2.1088e+13 1740874 < 2.2e-16 ***
independent_var      1 1.4679e+11 1.4679e+11   12118 < 2.2e-16 ***
Residuals       173270 2.0989e+12 1.2113e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 3 :
                    Df     Sum Sq    Mean Sq F value    Pr(>F)
(Intercept)          1 3.1126e+13 3.1126e+13 1999725 < 2.2e-16 ***
independent_var      1 1.5927e+11 1.5927e+11   10232 < 2.2e-16 ***
Residuals       173270 2.6970e+12 1.5565e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 4 :
                    Df     Sum Sq    Mean Sq  F value    Pr(>F)
(Intercept)          1 5.7991e+13 5.7991e+13 607177.5 < 2.2e-16 ***
independent_var      1 3.7351e+11 3.7351e+11   3910.7 < 2.2e-16 ***
Residuals       173270 1.6549e+13 9.5510e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 5 :
                    Df     Sum Sq    Mean Sq   F value    Pr(>F)
(Intercept)          1 1.1131e+14 1.1131e+14 843096.81 < 2.2e-16 ***
independent_var      1 9.4866e+10 9.4866e+10    718.53 < 2.2e-16 ***
Residuals       173270 2.2876e+13 1.3203e+08
```

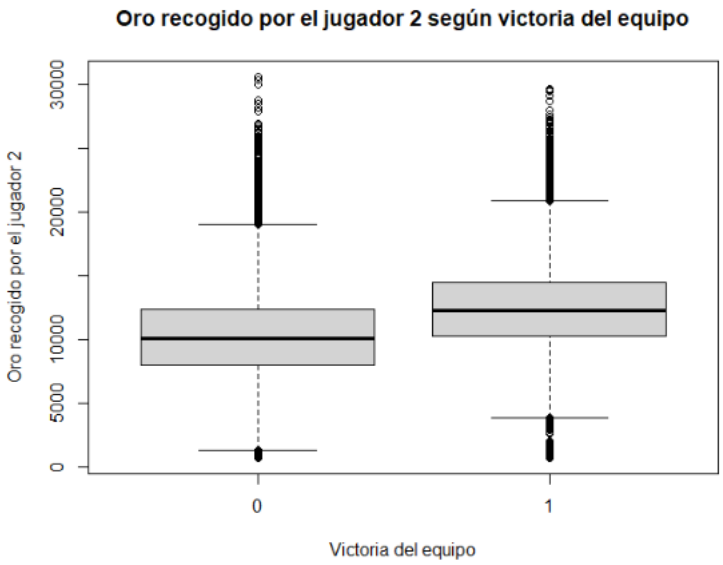
En la imagen es posible notar que cuando se hace una prueba ANOVA por cada una de las variables independientes también se puede concluir que existen diferencias estadísticamente significativas según la derrota o victoria del equipo por variable.

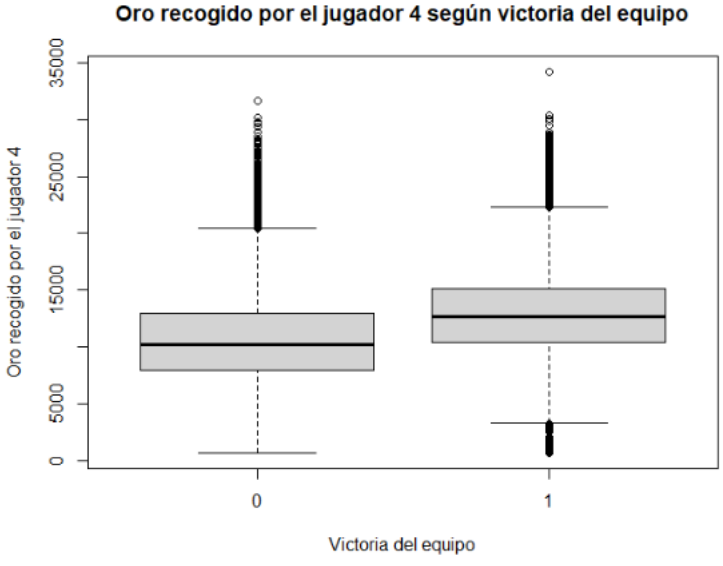
Dado que se muestran diferencias significativas para aquellos que son victoriosos y que el oro era parte de nuestra primera componente principal que explicaría la mayoría de los datos, se realizó una visualización con boxplots del oro obtenido por cada jugador de un equipo frente al jugador de su mismo rol en el equipo enemigo.

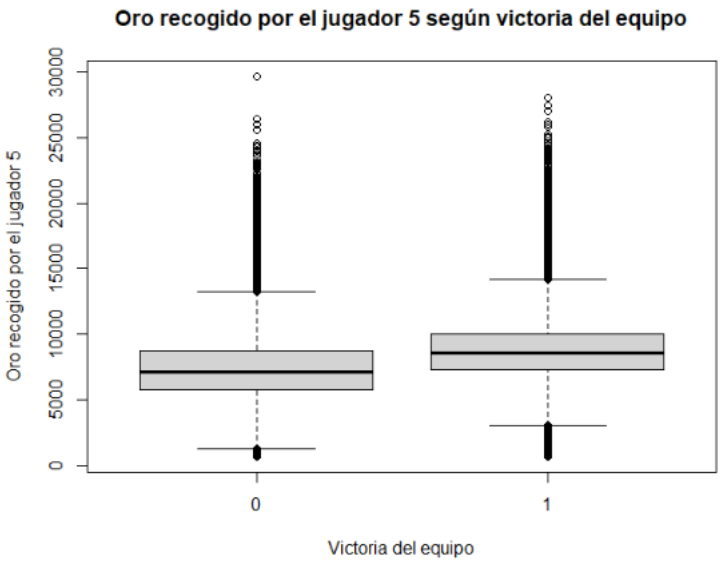
Universidad del Rosario


Oro recogido por el jugador 1 según victoria del equipo
Oro recogido por el jugador 1
0
5000
15000
25000
35000
0
1
Victoria del equipo


Oro recogido por el jugador 3 según victoria del equipo
Oro recogido por el jugador 3
0
5000
15000
25000
35000
0
1
Victoria del equipo


Oro recogido por el jugador 2 según victoria del equipo
Oro recogido por el jugador 2
0
5000
10000
20000
30000
0
1
Victoria del equipo


Oro recogido por el jugador 4 según victoria del equipo
Oro recogido por el jugador 4
0
5000
15000
25000
35000
0
1
Victoria del equipo



De esta manera se puede apreciar de manera visual que aquellos integrantes del equipo ganador tenían en promedio una mayor cantidad de oro en comparación con sus contrapartes, en todas las posiciones.

Clasificación Mediante este método se busco predecir los equipos que ganan y pierden utilizando el dataset que contiene las estadísticas de los jugadores por equipo, este dataset fue modificado previamente en el paso de PCA. En primer lugar se dividieron los datos entre entrenamiento y prueba con un 70% y 30% respectivamente. Además se escogieron 4 modelos para compararlos entre si y de esa manera identificar que tipo de modelo se adapta mejor a nuestros datos si a uno lineal o a uno no lineal.

- Modelos lineales
  - **Regresión logística**: Con este método es posible predecir variables categóricas (en nuestro caso "gana" o "pierde") a partir del uso de la función logística. Mediante este se realizaron 2 modelos, uno con las variables obtenidas en el paso de PCA (4 variables) y otro con las variables resultantes de la limpieza de datos.
    - Regresión logística sin PCA:

```
          Reference
Prediction     0     1
         0 28725   634
         1   545 30741

               Accuracy : 0.9806
```

    - Regresión logística con PCA:

```
          Reference
Prediction     0     1
         0 28712   647
         1   539 30747

               Accuracy : 0.9804
```

    Note que ambos modelos obtuvieron resultados muy buenos con una diferencia del 0.0002 lo que consideramos no es relevante, por lo tanto a partir de este momento se utilizará el PCA para el entrenamiento de los demás modelos.

  - **Discriminante de Fisher**:A este metodo de clasificacion supervisado se le entreno con las variables mencionadas previamente.

```
          Reference
Prediction FALSE  TRUE
     FALSE 28710   576
     TRUE    649 30710

               Accuracy : 0.9798
```

- Modelos no lineales
  - **Naives Bayes**: Fundamentalmente probabilistico y basado en el teorema de Bayes.

```
          Reference
Prediction FALSE  TRUE
     FALSE 27675   303
     TRUE   1684 30983

               Accuracy : 0.9672
```

  - **K-vecinos cercanos**: Basa su entrenamiento en el calculo de las distancias de una nueva observación a los demás datos

```
          ytest
modelknn FALSE  TRUE
   FALSE 28197  1066
   TRUE   1162 30220

               Accuracy : 0.9633
```

De estos modelos se tomaron los estadísticos de accuracy, Kappa y balanced accuracy para realizar la comparación y determinar así cual es el que más se adapta a los datos. De estos resultados, todos mayores a 0.9 es posible afirmar que los datos tienen una alta concordancia con los resultados obtenidos además de estar bien balanceados, lo que nos permitió obtener un buen modelo sen importar el método.

CONCLUSIONES

- Existe una diferencia estadísticamente significativa en los datos individuales de los jugadores según la victoria o derrota del su equipo.
- Aquellos integrantes del equipo ganador tenían en promedio una mayor cantidad de oro en comparación con sus contra partes en el equipo enemigo. Además, esta variable pertenecía a la primera componente principal, lo que significa que el oro es una factor absolutamente decisivo en la victoria de un equipo.
- La clasificación y agrupación de personajes en sus debidas categorías apoya el hecho de que la clasificación de estos personajes no está dado de manera estadística para ser agregados al juegos, en cambio estos son modelados alrededor de su rol en la arena.
- A partir de las pruebas realizadas, se concluye que el juego está bastante balanceado y que objetivamente no hay un personaje que signifique obtener una victoria segura, esto se debe recalcar que es a nivel competitivo profesional.
- En una futura ocasión, se puede analizar la sinergia entre personajes y la ventaja que esta pueda ofrecer en una partida competitiva de alto nivel.

REFERENCIAS

[1] LoL: predicting victory before the game starts. (2022, 12 septiembre). Kaggle. https://www.kaggle.com/datasets/ezalos/lol-victory-prediction-from-champion-selection