

# Tópicos I

Ciencia de Datos

Romel Zavaleta Bueno

Facultad de Ciencias

15 de diciembre 2018

# Introducción

El objetivo del estudio es el de encontrar características del juego que nos permitirán poder predecir el resultado de una partida.

## Clasificación

El enfoque principal de clasificación es la binaria (predecir si un equipo ganará). Aunque también se clasificará cuales héroes son mas efectivos en un rol( su desempeño en la partida) .

# Fase de Pick

El primer enfoque, se obtendrá resultados a partir de la elección del héroe.  
Las implementaciones se muestran en los cuadernos

Tablas utilizadas:

- `hero_names.csv`
- `test_player.csv`
- `test_labels.csv`

# Preparando los datos

Tenemos los datos de los slots de los jugadores y sus id de heroes separados de los resultados del match. Las convertiremos en una solo tabla.

|       | match_id | radiant_win | 0  | 1  | 2   | 3  | 4   | 128 | 129 | 130 | 131 | 132 |
|-------|----------|-------------|----|----|-----|----|-----|-----|-----|-----|-----|-----|
| 50000 | 100000   | 0           | 96 | 84 | 46  | 85 | 39  | 9   | 75  | 106 | 74  | 62  |
| 50001 | 100001   | 1           | 44 | 10 | 57  | 2  | 106 | 58  | 61  | 21  | 18  | 14  |
| 50002 | 100002   | 0           | 74 | 7  | 42  | 99 | 88  | 69  | 8   | 25  | 26  | 79  |
| 50003 | 100003   | 0           | 44 | 15 | 110 | 56 | 94  | 2   | 101 | 32  | 7   | 72  |
| 50004 | 100004   | 0           | 98 | 26 | 73  | 51 | 46  | 2   | 106 | 50  | 65  | 21  |

# Regresión Logística

La matriz de confusión obtenida es:

|             |           | Predicción |           |
|-------------|-----------|------------|-----------|
|             |           | Positivos  | Negativos |
| Observación | Positivos | 5189       | 2         |
|             | Negativos | 4807       | 2         |

# Naive Bayes classifier

Se hizo uso los parámetros GaussianNB. La matriz de confusión obtenida es:

|             |           | Predicción |           |
|-------------|-----------|------------|-----------|
|             |           | Positivos  | Negativos |
| Observación | Positivos | 5181       | 10        |
|             | Negativos | 4801       | 8         |

# k-Nearest Neighbors

Se hizo uso los parámetros GaussianNB. La matriz de confusión obtenida es:

|                    |                  | <b>Predicción</b> |                  |
|--------------------|------------------|-------------------|------------------|
|                    |                  | <b>Positivos</b>  | <b>Negativos</b> |
| <b>Observación</b> | <b>Positivos</b> | <b>2813</b>       | <b>2378</b>      |
|                    | <b>Negativos</b> | <b>2606</b>       | <b>2203</b>      |

# Random Forest Classifier

El número de estimadores es igual a 10. La matriz de confusión obtenida es:

|             |           | Predicción |           |
|-------------|-----------|------------|-----------|
|             |           | Positivos  | Negativos |
| Observación | Positivos | 2335       | 2862      |
|             | Negativos | 2047       | 2756      |



# Resultados de Entrenamientos

| Modelos     | Exactitud | Precisión | Exhaustividad | F1-Score |
|-------------|-----------|-----------|---------------|----------|
| Logistica   | 0.5191    | 0.5191    | 0.8772        | 0.6543   |
| Naive Bayes | 0.5189    | 0.5190    | 0.9980        | 0.6829   |
| KNN         | 0.5016    | 0.5190    | 0.5418        | 0.5302   |
| Forest      | 0.5091    | 0.5328    | 0.4492        | 0.4875   |
| SVC         | 0.5191    | 0.5191    | 1.0           | 0.6834   |

# Conclusiones sobre el Pick

No se logra una buena clasificación solo enfocándonos en el estudio de la selección del pick. Uno de los problemas se debe a que solo dispones del nombre del Héroe ( no nos ofrecen características básicas del Héroe como fuerza , inteligencia o agilidad ), además que el número de identificación del héroe no tiene ninguna relación a sus características propias , solo están ordenados alfabéticamente por lo que en el entrenamiento se toma solo como etiquetas y no como una magnitud de estudio ( En el modelo de SVM esta peculiaridad hace que su entrenamiento demore horas ). Resultados en otros estudios similares (bibliografía entregable 2), nos muestra que la precisión no llega al 60%.

# Promedio de victorias

Los códigos se encuentra en el archivo `taza_entrenamiento`

Tablas utilizadas:

- `player_ratings.csv`
- `test_player.csv`
- `test_labels.csv`

# Promedio de victorias

Este enfoque es muy similar al del solo pick, pero aquí filtramos las partidas atípicas para los estudios. Primero comenzamos creando la columna `ratio_v`, que es la cantidad de partidas ganadas sobre el total de jugadas, algunos de estos resultados son 0 y 1 (resultados atípicos) y se deben a jugadores con pocas partidas que pudieron perder todas o ganar todas. Debido a esto escogí solo los partidos donde ningún jugador superase el 0.8 ó se menor del 0.2 (la gran cantidad de jugadores ronda entre 0.4 y 0.6).

# Resultados de Entrenamiento

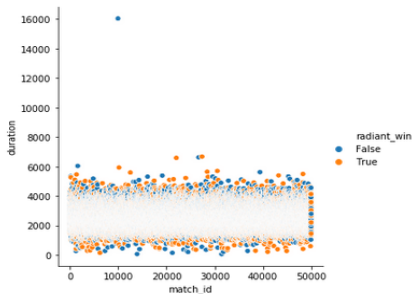
| Modelos     | Exactitud | Precisión | Exhaustividad | F1-Score |
|-------------|-----------|-----------|---------------|----------|
| Logistica   | 0.5149    | 0.5217    | 0.9996        | 0.6833   |
| Naive Bayes | 0.5110    | 0.5234    | 0.7317        | 0.6103   |
| KNN         | 0.5013    | 0.5229    | 0.5374        | 0.5300   |
| Forest      | 0.5013    | 0.5281    | 0.4415        | 0.4809   |

# Conclusion sobre el Promedio de victorias

Se noto una ligera mejoría de la **precisión** en la mayoría de modelos entrenados antes con el enfoque del pick. Aun así el porcentaje no es elevado para determinar una buena clasificación. Demuestra que el sistema de emparejamiento de **Valve** es bueno, pues los jugadores no pueden sacar alguna ventaja con respecto a otros por su historial de victorias.

# Tiempo de partida y Equipo ganador

En este pequeño estudio se vio que tanto influye la duracion de una partida para el equipo ganador. Se obtuvo una gráfica respecto a si el equipo Radiant era el ganador.



# Conclusion sobre el tiempo de Partida

Se dividió el conjunto de partido en 3, los de mayor duración (mayor a 4000s) ,los de duración media (entre 2000s a 4000s) y los de menos duración (entre 600 a 2000). No se tomo en cuenta los partidos menores a 10 minutos pues muchos de ellos son debido a fallos en conexión entre usuarios.

## Resultado

En la partida de menor duración se ve una clara ventaja para el equipo Radiant, mientras que en los restantes grupos de duración la probabilidad de un equipo es cercana al 50% .



# Caracterizando Rol de Héroe

En este estudio se medirá el desempeño del héroe en la partida, podremos determinar el rol que toma. La implementación se encuentra en el cuaderno `Heroes_desarrollo`.

Tablas utilizadas:

- `players.csv`
- `hero_names.csv`

# Caracterizando Rol de Héroe

Seleccionando las principales estadísticas de los héroes , que se obtuvieron del promedio de los jugadores podremos identificar , podemos clasificar el rol que cada heroe desarrolla en la partida. Utilizando Kmeans identificaremos 6 tipos de héroes ( roles que desempeñan en la partida).

# Caracterizando Rol de Héroe

Los roles definidos son:

- **Fighter / Charger** : va a la línea de medio solo (Mid Laner) , conocido como posición 2 .
- **Tank / Support**: Generalmente solo, va a la línea bottom (caso dire) o línea top (Radiant), conocido como posición 3(offlane).
- **Laner Solo / Squishy**: No tiene una posición fija(roaming), es conocido como el 4 hombre.
- **Healer**: El héroe apoyo del equipo, acompaña al Carry. Conocido como el sacrificado posición 5.
- **Assassin / OP**: Todos los heroés trabajan para su desarrollo, es el asesino del equipo(Carry), el posición 1.
- **Average**: No especifica un rol fijo.

# Conclusiones Caracterizando Rol de Héroe

En cada parche nuevo algunos héroes pueden cambiar de posición habitual, un estudio de acuerdo a los datos del parche actual puede darnos un punto de vista general del rol que cumplira estos héroes.