

Chess Mining

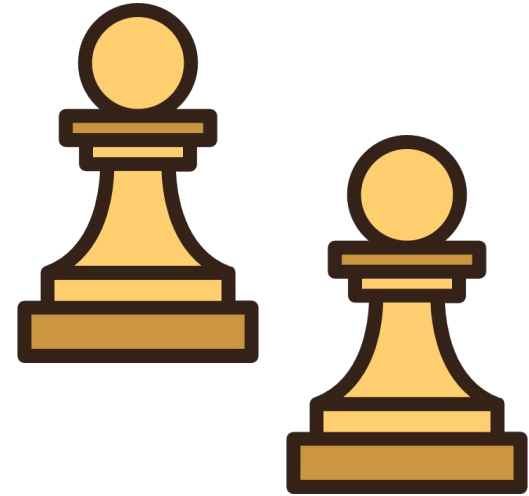
*Applying **KDD** on chess games*

Alberto Velasco Mata, Diego Pedregal Hidalgo
Rubén Márquez Villalta, Antonio Manjavacas Lucas



ÍNDICE DE CONTENIDOS

1. Introducción
2. Obtención de los datos
3. Recopilación de datos objetivo
4. Visualización y clustering
5. Modelo de predicción
6. Resultados obtenidos
7. Trabajo futuro



1. Introducción

- *DeepBlue* (fuerza bruta)
- *AlphaZero* (red neuronal profunda y aprendizaje por refuerzo)
- *Stockfish*

Diferentes fines:

- ❖ Jugar contra humanos
- ❖ Jugar contra otras máquinas
- ❖ Predecir resultados
- ❖ Analizar partidas



1. Introducción

HIPÓTESIS

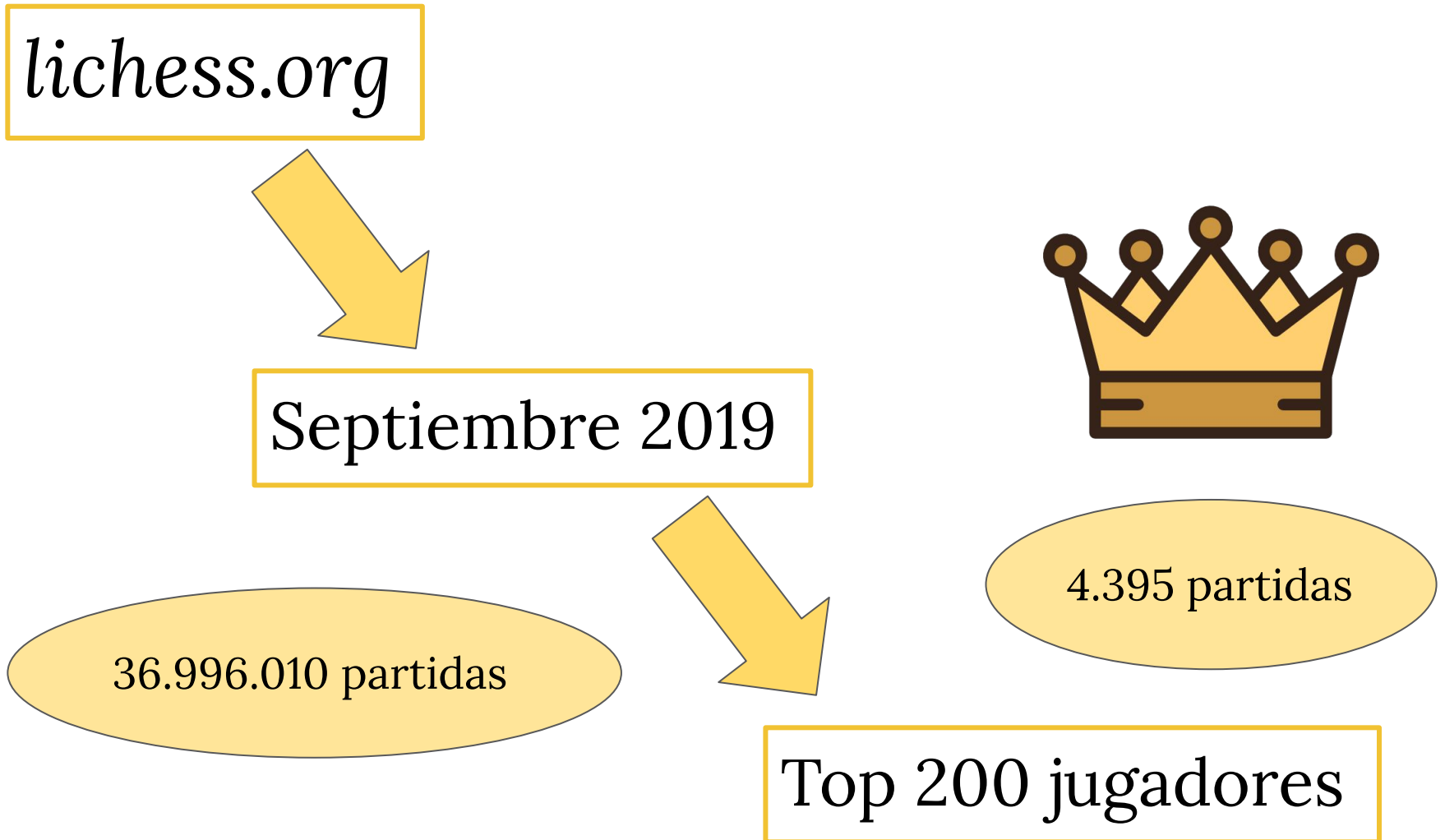
- H1 Es posible pronosticar el **resultado de un torneo**
- H2 Es posible definir los **factores de éxito** que influyen en el devenir de una partida
- H3 Es posible definir el **estilo de juego de un jugador**

OBJETIVOS

- O1 Establecer un **perfil de juego a cada jugador** en base a sus partidas
- O2 Pronosticar el **resultado de diferentes partidas individuales**



2. Obtención de los datos



2. Obtención de los datos

LIBRERÍAS UTILIZADAS

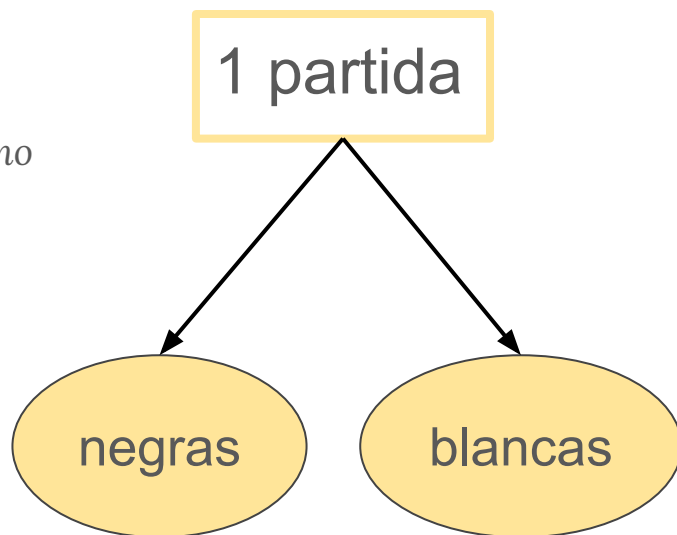
- API **berserk**
 - Obtención de la tabla de clasificación:
 - Partidas Clásicas
 - 200 mejores jugadores
 - Obtención de las partidas:
 - Anteriores jugadores
 - Septiembre 2019
- PGN (Portable Game Notation)
- **Python-chess**
 - Lectura del archivo PGN con las partidas
- **json, datetime, numpy**



3. Recopilación de datos objetivo

DATOS OBJETIVO

1. ELO
2. Color
3. Apertura
4. Movimientos
5. Tiempo:
 - a. Tiempo total
 - b. Tiempo por jugador
 - c. Early, middle, end → Media, mediana, varianza, máximo y mínimo
6. Balance de puntos
7. Balance de piezas
8. Agresividad:
 - a. Piezas eliminadas → Early game
 - b. Apertura agresiva
 - c. Castling
9. Result ← **variable objetivo**



3. Recopilación de datos objetivo

elo	int64
colour	category
opening	category
result	int64
movements	int64
total_time_player	float64
total_time	float64
early_times_mean	float64
early_times_median	float64
early_times_variance	float64
early_times_max	float64
early_times_min	float64
mid_times_mean	float64
mid_times_median	float64
mid_times_variance	float64

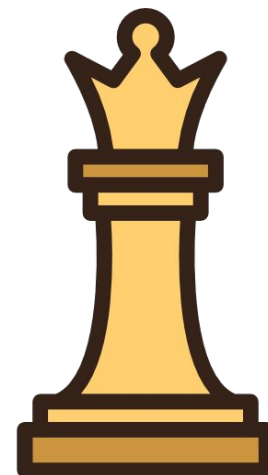
mid_times_max	float64
mid_times_min	float64
end_times_mean	float64
end_times_median	float64
end_times_variance	float64
end_times_max	float64
end_times_min	float64
points_balance	int64
taken_balance	int64
aggressiveness	float64
colour_enc	int8
opening_enc	int16



4. Visualización y clustering

SELECTED FEATURES

1. *elo*
2. *opening_enc*
3. *result*
4. *movements*
5. *total_time_per_player*
6. *early_times_median*
7. *early_times_max*
8. *early_times_min*
9. *mid_times_median*
10. *mid_times_max*
11. *mid_times_min*
12. *end_times_median*
13. *end_times_max*
14. *end_times_min*
15. *points_balance*



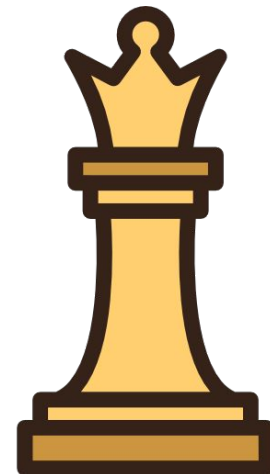
MinMaxScaler

4. Visualización y clustering

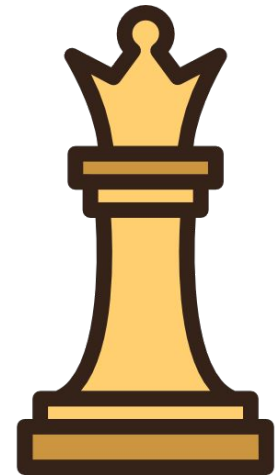
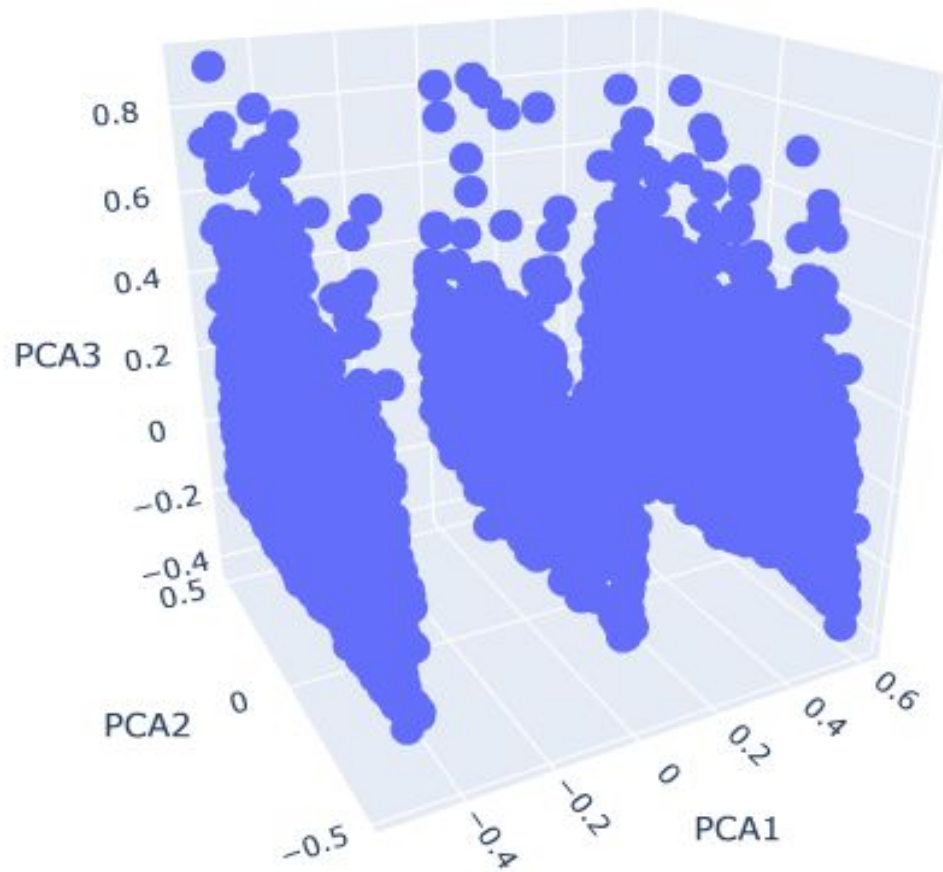
Principal component analysis (PCA)

- Num. componentes = 3
 - [0.62322601, 0.22088364, 0.06225892]
 - 0.9063685719687516 → **90%**

	PC-0	PC-1	PC-2
elo	-0.110880	-0.006919	0.288907
opening_enc	-0.000983	-0.999766	0.014340
result	-0.989784	0.001753	-0.033313
movements	-0.000406	0.000115	0.443249
total_time_player	0.001063	0.015091	0.730605
early times median	-0.001556	0.002331	0.090972
early times max	-0.000368	0.004323	0.229377
early times min	-0.001495	0.000307	0.014759
mid times median	-0.001311	0.011402	0.302232
mid times max	0.000450	0.003064	0.076533
mid times min	-0.005953	0.004159	0.075692
end times median	0.001603	0.002938	0.055031
end times max	0.003802	-0.000395	0.132851
end times min	-0.001926	0.000521	0.010731
points_balance	-0.089260	-0.000120	0.011578

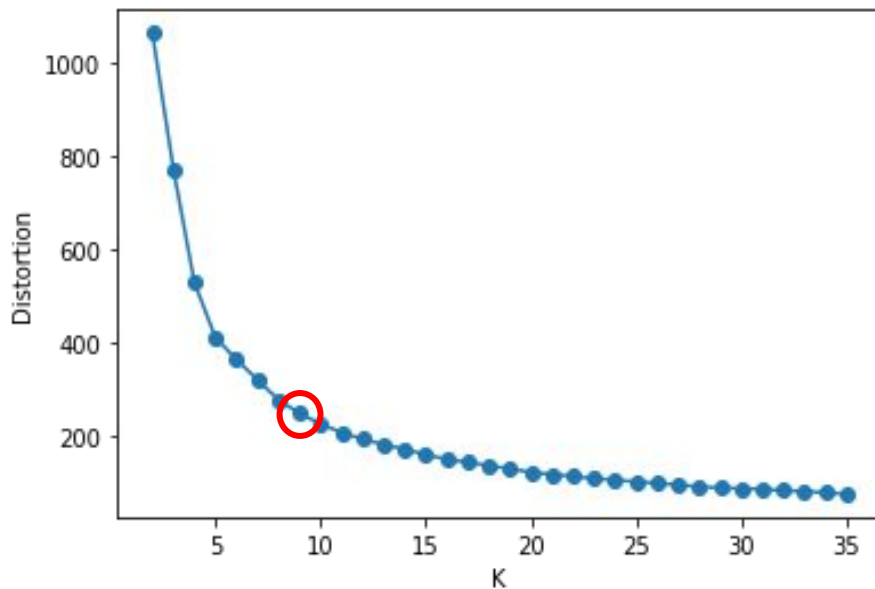


4. Visualización y clustering

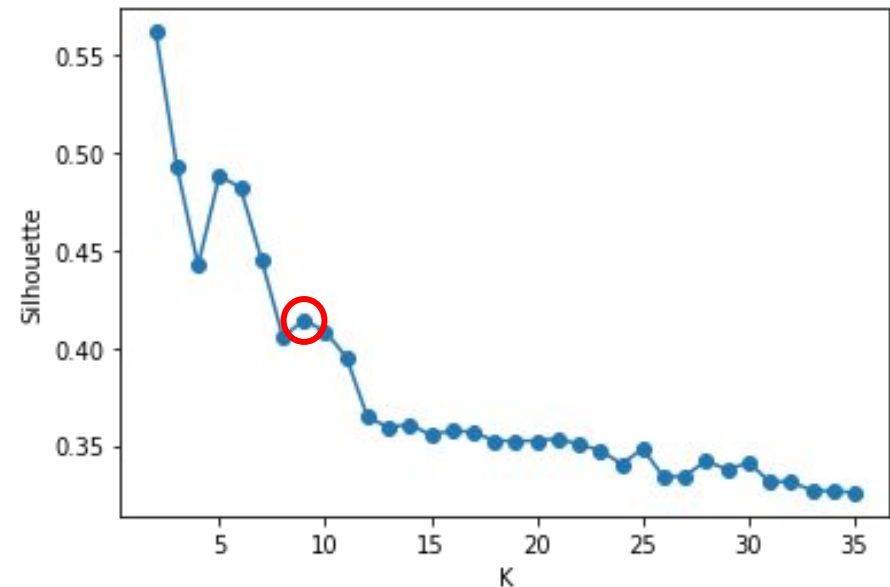


4. Visualización y clustering

Distortion



Silhouette

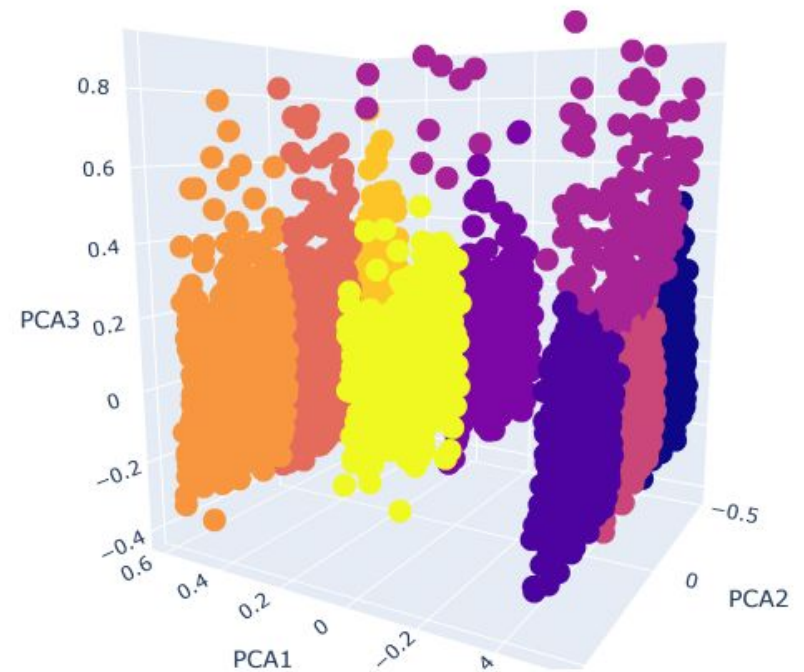
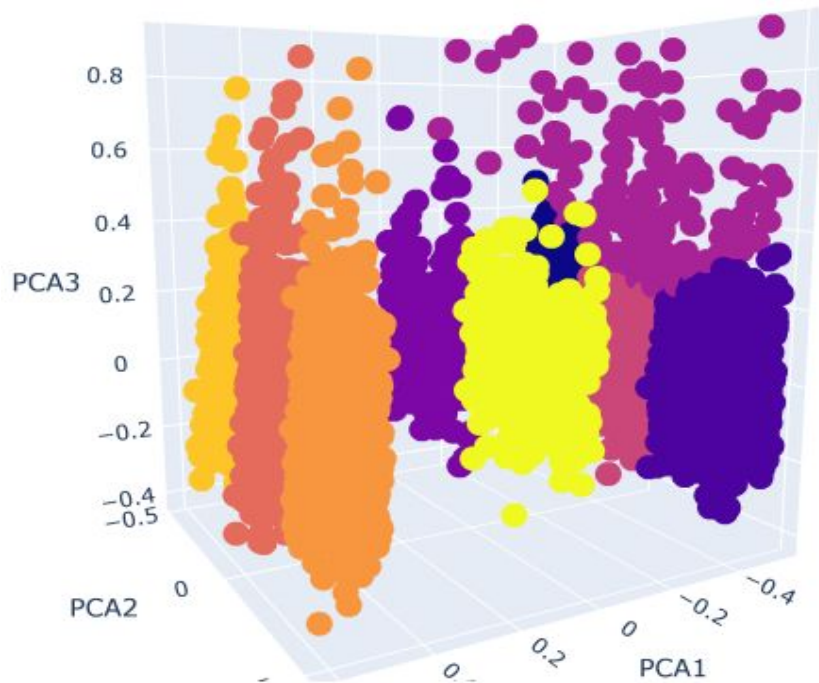


K = 9

Silhouette Coefficient: 0.414

Distortion: 247.02

4. Visualización y clustering



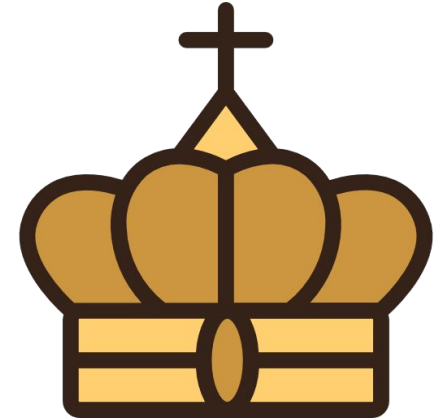
5. Modelo de predicción

PROCESO

1. Carga de datos (**source**)
2. Generación del **dataset**
3. Filtrado de variables

Descartamos IDs de los usuarios, nombre de la apertura e ID de cada uno de los registros.

4. Construcción del **modelo**



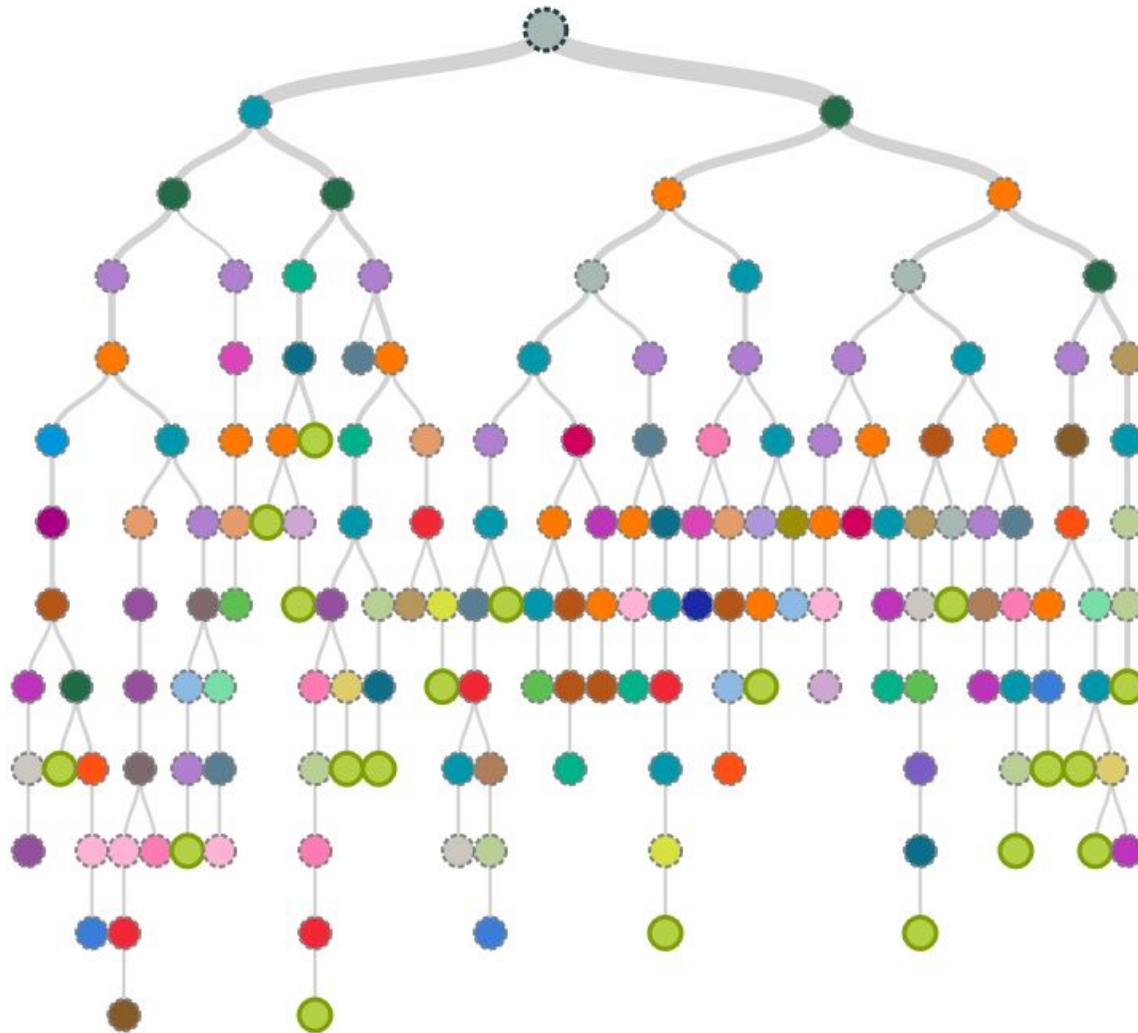
Generamos el modelo...



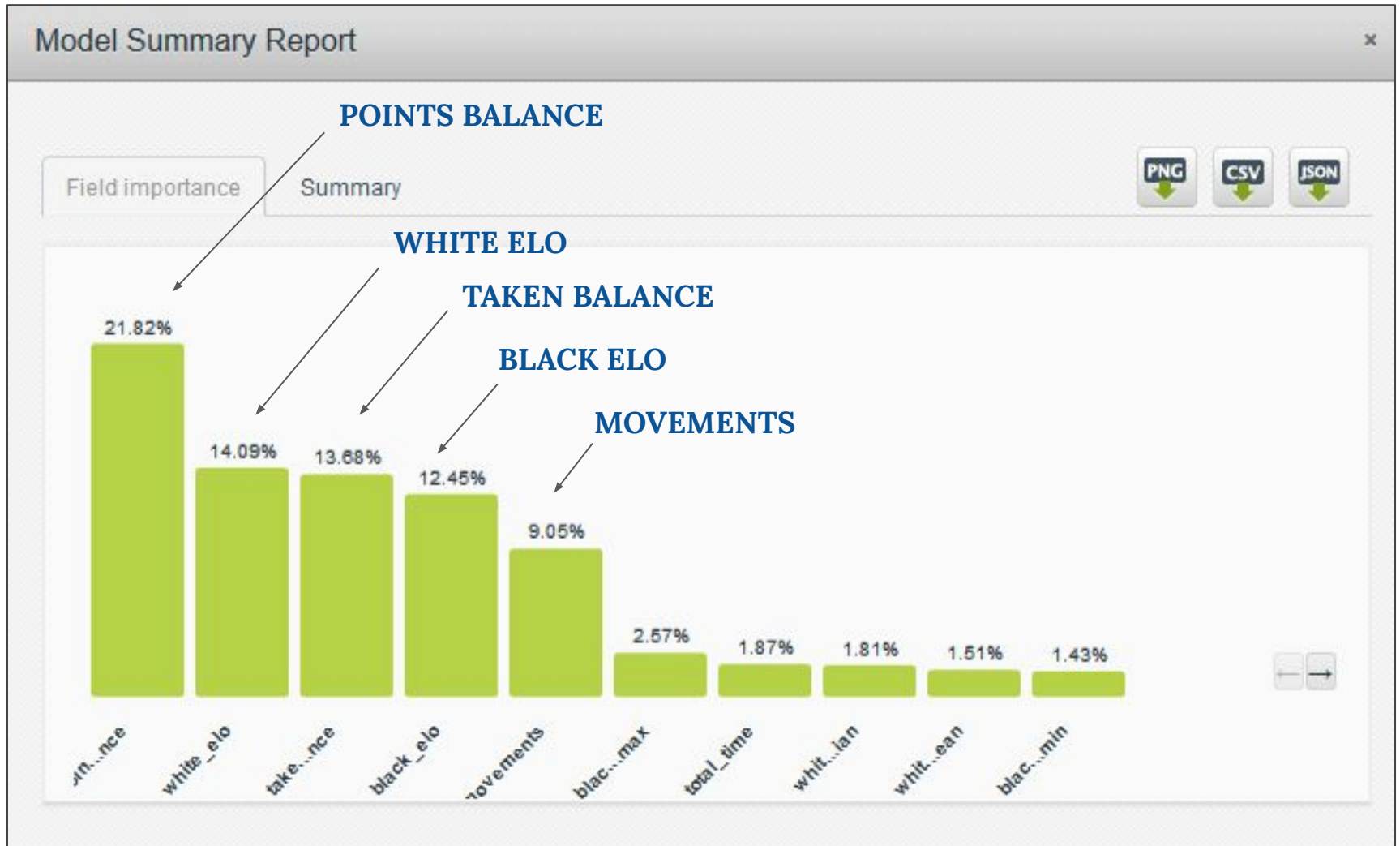
Processing...

Your model is being created

5. Modelo de predicción

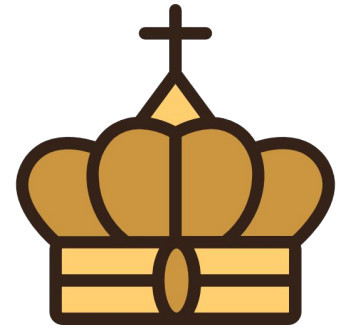


5. Modelo de predicción



5. Modelo de predicción

Matriz de confusión



GANAN NEGRAS (result=0)

TP

FN

FP

TN

ACTUAL VS. PREDICTED				ACTUAL	RECALL	F	Phi
	0	1	2				
0	859	39	300	1,198	71.70%	0.76	0.56
1	59	18	114	191	9.42%	0.13	0.10
2	148	27	886	1,061	83.51%	0.75	0.53
PREDICTED	1,066	84	1,300	2,450	54.88% AVG.RECALL	0.55 AVG. F	0.40 AVG. Phi
PRECISION	80.58%	21.43%	68.15%	56.72% AVG.PRECISION	71.96% ACCURACY		

77.7%
Accuracy

0.7588
F-measure

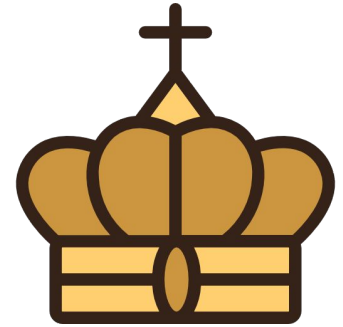
80.6%
Precision

71.7%
Recall

0.5563
Phi coefficient

5. Modelo de predicción

Matriz de confusión



TABLAS (result=1)

TP

FN

FP

TN

ACTUAL VS. PREDICTED

	0	1	2	ACTUAL	RECALL	F	Phi
0	859	39	300	1,198	71.70%	0.76	0.56
1	59	18	114	191	9.42%	0.13	0.10
2	148	27	886	1,061	83.51%	0.75	0.53
PREDICTED	1,066	84	1,300	2,450	54.88% AVG.RECALL	0.55 AVG. F	0.40 AVG. Phi
PRECISION	80.58%	21.43%	68.15%	56.72% AVG. PRECISION	71.96% ACCURACY		

90.2%
Accuracy

0.1309
F-measure

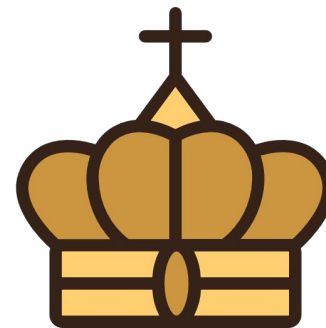
21.4%
Precision

9.4%
Recall

0.0958
Phi coefficient

5. Modelo de predicción

Matriz de confusión



GANAN BLANCAS (result=2)

				<div><div></div>TP</div>		<div><div></div>FN</div>		<div><div></div>FP</div>		<div><div></div>TN</div>		<div><div></div></div>	
ACTUAL VS. PREDICTED													
	0	1	2	ACTUAL	RECALL	F		Phi					
0	859	39	300	1,198	71.70%	0.76		0.56					
1	59	18	114	191	9.42%	0.13		0.10					
2	148	27	886	1,061	83.51%	0.75		0.53					
PREDICTED	1,066	84	1,300	2,450	54.88% AVG.RECALL	0.55 AVG. F		0.40 AVG. Phi					
PRECISION	80.58%	21.43%	68.15%	56.72% AVG. PRECISION	71.96% ACCURACY								

76.0%
Accuracy

0.7505
F-measure

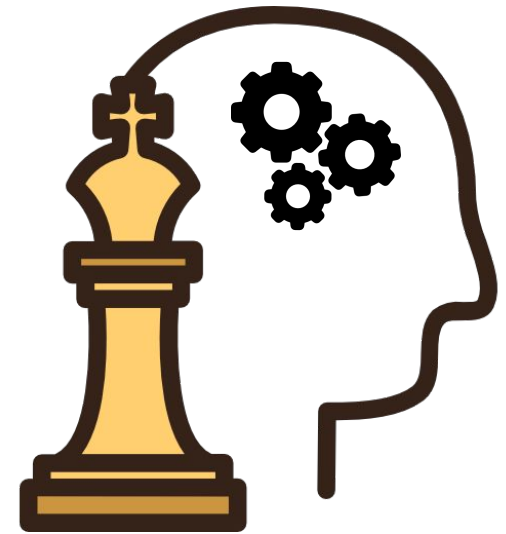
68.2%
Precision

83.5%
Recall

0.5332
Phi coefficient

7. Resultados obtenidos

- H1 Es posible pronosticar el **resultado de un torneo**
- **Pronóstico de partidas mediante el modelo**
 - **Alta precisión**
- H2 Es posible definir los **factores de éxito** que influyen en el devenir de una partida
- **Variables más influyentes en el modelo**
- H3 Es posible definir el **estilo de juego de un jugador**
- **Clustering → perfiles de los jugadores**

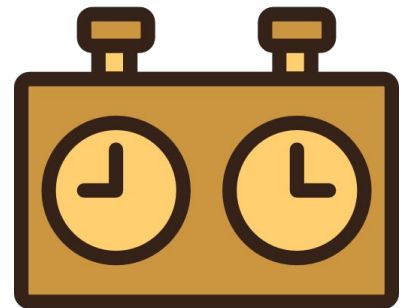


7. Trabajo futuro

- Un modelo que no requiera del balance de **piezas comidas** / **puntos** para llevar a cabo la predicción.

Aunque...

- El ajedrez no es similar a otros juegos competitivos donde acumular más puntos siempre significa ganar.
 - **Especialmente en niveles profesionales.**
- ¿Búsqueda de otros posibles factores de predicción?



¡A jugar!

