Influencia de las aperturas de ajedrez en el resultado de la partida

S. Colorado, *Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia*

***Resumen*— Este artículo presenta un enfoque basado en la predicción de la tasa de victoria en el ajedrez según la apertura elegida al inicio del juego.**
Se recopilaron datos históricos de partidas de ajedrez y se utilizó aprendizaje automático para desarrollar un modelo predictivo. Se realizó un análisis exploratorio de los datos y se seleccionaron las características relevantes para alimentar el modelo. Se emplearon algoritmos de aprendizaje automático, como árboles de decisión y para entrenar y ajustar el modelo. Los resultados experimentales demuestran que el modelo desarrollado no es tan preciso en la predicción de la tasa de victoria basada en la apertura seleccionada.

***Abstract*-- This article presents an approach based on predicting the victory rate in chess based on the opening chosen at the beginning of the game. Historical chess game data was collected, and machine learning was used to develop a predictive model. An exploratory data analysis was performed, and relevant features were selected to feed the model. Machine learning algorithms, such as decision trees, were employed to train and fine-tune the model. The experimental results demonstrate that the developed model is not as accurate in predicting the victory rate based on the selected opening..**

I. Introducción

La elección de la apertura es una de las decisiones más importantes que un jugador de ajedrez debe tomar al comienzo de la partida, ya que puede influir significativamente en el resultado final del juego. Por lo tanto, es importante para los jugadores de ajedrez y los entrenadores conocer si la apertura que se está realizando favorece o no al jugador.

Dentro de los trabajos de referencia se buscó priorizar artículos y códigos que trabajaran con el mismo data set y el mismo objetivo de predecir el resultado de las partidas, sin embargo no se encontró ningún trabajo centrado en predecir cómo afecta la apertura de la partida en el resultado de esta. Se describe brevemente cada uno de los artículos de referencia:

A. Predict Chess Winner after 1st move [1]

El modelo se enfoca en predecir el resultado de la partida en base al primer movimiento, empleando una técnica de regresión y validación cruzada. Su principal métrica para la validación del modelo fue la prueba RMSE, con el que valido que el modelo desarrollado era bastante débil, al tener un valor RMSE relativamente alto.

B. Chess Win Prediction with RNNs [2]

Este modelo, similar al nuestro, busca predecir el ganador el juego, a partir de una lista de movimientos, sin embargo no enfocada únicamente en las aperturas. Emplea la técnica de redes neuronales recurrentes, y sus principales métricas para evaluar el modelo fueron la precisión y el análisis ROC, donde consiguió valores de 0.82 y 0.91 respectivamente, lo que demuestra que es un modelo bastante bueno para predecir el resultado de las partidas de ajedrez.

C. Chess Winner Predictor [3]

Este modelo, busca predecir el ganador del juego, basándose en diferentes cantidades de movimientos, tomando como referencia diferentes puntos de la partida. Emplea el método de Random Forest y hace la validación del modelo con datos de prueba que separo al inicio. Su métrica principal es la precisión, y en los resultados del modelo son bastantes satisfactorios cuando se toman todos los movimientos de la partida, alcanzando una precisión del 90%.

D. Predicting results of matches - Random Forest [4]

Este modelo, busca definir el resultado de una partida a partir de parámetros de antes de que este empiece, cómo lo son los puntajes de cada jugador y sus tasas de victorias y derrotas. Emplea la técnica de Random Forest con validación cruzada. Este modelo resulta interesante por la partición que realiza de 50/50. Su principal métrica es la precisión, en donde alcanza un resultado satisfactorio del 80%

II. DATA SET

A. Descripción

El dataset utilizado en este proyecto proviene de la plataforma de ajedrez en línea Lichess.org y se recopiló utilizando la API de Lichess. La URL para acceder a la página de descarga del dataset es la siguiente: https://www.kaggle.com/datasnaek/chess.
El dataset consta de un solo archivo CSV que contiene más de 20,000 partidas de ajedrez. Cada fila representa una partida de ajedrez y contiene información como el ID de la partida, si fue clasificada o no, la hora de inicio y finalización de la partida, el número de turnos, el estado de la partida, el ganador, el incremento de tiempo, la ID y la clasificación de los jugadores blancos y negros, todos los movimientos en la notación estándar de ajedrez, la apertura (identificada por su código ECO y su nombre) y el número de jugadas en la apertura.
Además, hay columnas para el rating de los jugadores, la fecha de la partida y la velocidad del juego. El archivo CSV tiene un tamaño de aproximadamente 6,5 MB y consta de 15 columnas y más de 20,000 filas. Este dataset proporciona una gran cantidad de información valiosa para analizar las aperturas de ajedrez y su relación con la tasa de victoria de los jugadores.

B. *Limpieza de datos*

Para la limpieza del dataset, no se emplearon métodos de detección de atípicos, sino que se hará la detección de estos datos por medio de aplicar conocimientos técnicos de las partidas, eliminando datos se las siguientes formas

- Se eliminaron aquellas partidas con menos de 4 movimientos, puesto que es imposible hacer un checkmate en 3 turnos
- Se eliminaron partidas con duración superior a 6 horas.
- Se eliminaron las partidas con una duración de 0 segundos, al ser considerados datos erróneos.
- Se eliminaron las partidas que terminaron en empate.
- Debido a la gran cantidad de aperturas diferentes (más de 400), se eliminaron las partidas en las que la apertura no se repita al menos unas 100 veces.

Luego de esta limpieza, quedaron solamente 6708 partidas para analizar, casi el 50% de los datos iniciales.

C. *Balanceo del dataset*

En el tema de balanceo nos son relevantes dos variables importantes: el ganador, el cual pueden ser las blancas, las negras o que la partida termine en un empate, y la cantidad de aperturas diferentes.

Para el primer caso cómo se observa en la figura 1, el ganador esta bien distribuido en cuanto a blancos y negros, denotando que el empate es algo poco común en las partidas. Para este caso particular se optará por no hacer balanceo, sino que se procederá a eliminar las partidas terminadas en empate, pues el objetivo del modelo es mejorar la tasa de victorias, por lo que podemos prescindir de estas partidas.

Fig. 1. Distribución de resultados

Para la distribución del tipo de apertura, quedaron 28 aperturas diferentes, la más utilizada con 510 partidas y la menos utilizada con 107 partidas. Se considera aceptable el balanceo dentro de la cantidad de aperturas y su distribución, por lo que no se incurrirá en métodos de submuestro al ya haber perdido una gran cantidad de datos, ni tampoco de sobre muestreó al no saber cómo afectara esto el balanceo de la distribución de victorias y las otras variables.

III. MODELO IMPLEMENTADO

A. *Variables utilizadas*

Para nuestro modelo utilizaremos las siguientes variables al considerarlas las variables de mayor interés:

- Estado de la victoria (Codificada)
- Rango del jugador blanco.
- Rango del jugador negro.
- Apertura utilizada (Codificada)
- Cantidad de jugadas de la apertura

B. *Técnica utilizada*

La técnica que se decidió utilizar será una clasificación empleando Random Forest, para la validación del modelo se utilizará una validación cruzada.

Dentro de la búsqueda de mejores hiperparametros, se consiguieron los siguientes: una profundidad máxima de 10, con 14 características y con 200 estimadores.

Cómo métricas se utilizarán la precisión, la exhaustividad y el valor F1.

IV. RESULTADOS

Se obtuvieron los siguientes resultados:

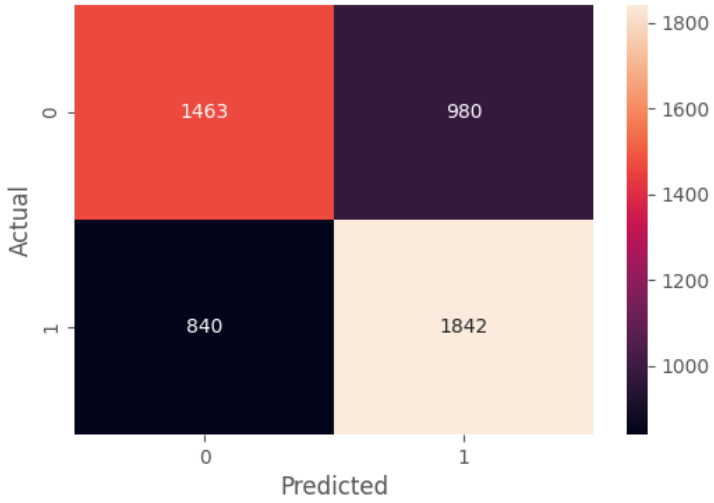

Fig. 2. Matriz de confusión

TABLE I
MÉTRICAS DEL MODELO

Ganador	Métricas		
	Exactitud	Exaustividad	F1
Negras	0.61	0.64	0.62
Blancas	0.66	0.64	0.65

Ciertamente los resultados del modelo no son los deseados, al tener una exactitud apenas del 65%, sin embargo, la exhaustividad es considerablemente alta para las blancas, en parte porque son estas quienes marcan el inicio de la apertura. El modelo con una exactitud tan baja no podría cumplir con el objetivo de predecir si el jugador o no podría o no ganar de acuerdo con la apertura que este utilizando en su respectivo rango y con su respectivo color.

En comparación con los otros modelos, queda en claro que la importancia de la apertura en el ajedrez, aunque si bien es decisiva, no termina demarcando tanto el resultado de la partida. Esto se puede corroborar en parte con el primer modelo [1] que también consiguió resultados bastante malos prediciendo únicamente con el primer movimiento.

En conclusión el resultado de las partidas no puede ser predicho únicamente con la apertura de la partida.

REFERENCIAS

[1] M. Bogacz, “Predict Chess Winner after 1st move” Junio 2022.
[2] G. Atkin, “Chess Win Prediction with RNNs” Marzo 2020.
[3] I. Seyam, “Chess Winner Predictor” Novembre 2022.
[4] M. P. Moretti, “Predicting results of matches - Random Forest”