Documentación del análisis de la base de datos Chess Games

ÍNDICE

Descripción de la base de datos

Esta base de datos fue obtenida de la plataforma ***Kaggle*** (https://www.kaggle.com/datasets/datasnaek/chess) y está constituida por 16 columnas que contienen datos de 20.058 partidas de ajedrez jugadas en la plataforma Lichess.

El nombre y contenido de las columnas de nuestra base de datos es el siguiente:

- **Game_id**: Identificador único de la partida
- **Rated**: Indica si la partida es válida para ELO (True) o no (False)
- **Created_at**: Fecha y hora de la partida
- **Last_move_at**: Tiempo final de la partida
- **Turns**: Número total de movimientos en la partida
- **Victory_status**: Forma en la que terminó la partida *(Mate, Resign, Draw, Outoftime)*
- **Winner**: Ganador de la partida *(White, Black o Draw)*
- **Increment_code:** Ritmo de juego de la partida (en formato n + m, donde n representa la cantidad de minutos por jugador y m el incremento en segundos por jugada realizada)
- **White_Id:** Nombre de usuario del jugador de blancas
- **White_Rating:** ELO del jugador blanco
- **Black_Id:** Nombre de usuario del jugador de negras
- **Black_Rating:** ELO del jugador negro
- **Moves:** Todos los movimientos de la partida en notación estándar de ajedrez
- **Opening_eco:** Código de la apertura utilizada
- **Opening_name**: Nombre de la apertura
- **Opening_ply**: Número de movimientos en la apertura

Limpieza y transformación de los datos

Para la ejecución de los pasos realizados en este apartado se ha utilizado el editor *PowerQuery* ya que facilita mucho la limpieza, transformación y creación de nuevas columnas a partir de los datos originales.

Eliminación de columnas

En primer lugar, vamos a eliminamos algunas variables que por criterio propio las considero innecesarias para incluirlas en el análisis debido a su poco aporte de información, estas son:

Gamed_Id, Created_at, Last_move_at, White_Id, Black_Id, Opening_eco, Opening_ply

Transformación de las columnas

Inspeccionando los datos nos damos cuenta de que la columna *Opening_name* sigue siempre el mismo formato, *opening: variation* y procedemos a separar estos dos valores en 2 columnas diferentes, una columna *opening* y otra columna *variation*.

Revisando las columnas de nuestro dataset encontramos que la columna recién creada *variation* es la única que posee **valores faltantes** y procedemos a reemplazarlos por el término *'No identificada'* para hacer alusión a que no sabemos qué tiempo de variante de la apertura principal han jugado.

La última transformación que hacemos en los datos originales es en la columna *rated*, en la que reemplazamos los valores True por RATED y False por UNRATED simplemente por mejorar el entendimiento de la base de datos.

Creación de nuevas columnas

Una vez finalizadas las transformaciones procedemos con la creación de nuevas columnas a partir de las columnas originales con el propósito de agregar y agrupar información.

La primera columna que creamos es una columna llamada '*tipo de partida'* en la que le agregamos la etiqueta *Blitz, Rápida o Lenta* (que son los tipos de partidas que hay en ajedrez en función del ritmo de juego), el criterio utilizado para asignar estas etiquetas es la suma de los elementos de la columna Increment_code (minutos que tiene cada jugador + segundos de incrementos por jugada) si la suma de estos valores es menor o igual a 10 la partida será *Blitz*, si la suma está entre 11 y 40 la partida será Rápida y si es más de 40 se catalogará como partida Lenta.

Una vez creada esta columna eliminamos la columna inicial Increment_code ya que nos interesa más el tipo de ritmo al que estén jugando que el tiempo exacto con el que parte cada jugador.

Las siguientes columnas que creamos son las de '*ELO medio'* (promedio de ELO entre el jugador de blancas y el de negras) y '*diferencia de ELO'* (valor absoluto de la diferencia entre el ELO del jugador blanco y el negro)

A partir de la columna 'ELO medio' creamos otra columna a la que llamamos '*Nivel Partida'* donde etiquetamos los datos de la columna *'ELO medio'* como Principiante, Intermedio o Avanzado según el nivel de los jugadores que conforman la partida. El criterio elegido para la asignación de las etiquetas es el siguiente:

- Principiante→ 0 ≤ ELO medio ≤ 1300
- Intermedio→ 1300 ≤ ELO medio ≤ 1900
- Avanzado→ 1900 < ELO medio

Finalmente, la columna '*diferencia de ELO'* la utilizamos para crear una columna llamada *'Sorpresa'* que tomará los valores *Si/No* según se haya habido un resultado sorprendente en la partida.

Catalogaremos de resultado sorpresa aquel en el que el ganador de la partida tenga mínimo 200 puntos de ELO menos que el jugador que pierde.

Como último paso del preprocesamiento de los datos eliminamos las instancias duplicadas, obteniendo un total de **947 valores duplicados**, haciendo que nuestra base de datos pase de 20.058 partidas a 19.111 y de 16 a 13 columnas (habiendo transformado 1, eliminado 7 y creado 4 a partir de las originales)

Columnas Originales

Game_Id	White_Id
Rated	White_Rating
Created_at	Black_Id
Last_move_at	Black_Rating
Turns	Moves
Victory_status	Opening_eco
Winner	Opening_name
Increment_code	Opening_ply

Columnas Nuevas

Rated	Nivel Partida
Winner	ELO Medio
Victory_status	Diferencia ELO
Turns	Sorpresa
Tipo de partida	Opening
White_Rating	Variation
Black_Rating	

Análisis Exploratorio de los datos

Como primera parte del análisis exploratorio analizaremos la distribución de algunas de las variables categóricas de nuestra base de datos (aquellas que toman valores no numéricos) y después realizaremos distintos análisis en función de los valores que tomen las características más importantes de nuestro dataset.

Estadísticas Generales

En primer lugar, observamos que el 81% de las partidas totales son RATED, es decir, la mayoría de las partidas que analizaremos son válidas para el sistema de clasificación. Esto hace que los jugadores se tomen más en serio la partida concentrándose más y evitando movimientos de prueba o experimentales asemejándose así a lo que sería una partida de torneo oficial.

La distribución de las partidas según el ritmo de juego es 47,9% Blitz, 47,5% Rápida y 4,6% Lentas. Este resultado parece natural ya que las plataformas de ajedrez han popularizado los ritmos rápidos de ajedrez mediante formatos dinámicos y adaptados al entorno online, esto sumado a las posibles distracciones que puede tener una persona estando en casa (teléfono, familia, entorno...) dificulta la práctica de partidas lentas.

Examinando el nivel de las partidas vemos que la amplia mayoría son de catalogadas como intermedias (80%), mientras que las partidas de nivel avanzado son las más escasas (7%).

Centrándonos ahora en la columna vinculada al resultado de la partida, observamos que solamente el 5% de las partidas acaban en empate, esto puede ser por la dominancia en la base de datos de las partidas de ritmo rápido, las cuales suelen ser mucho más caóticas, provocando que estas se decidan por uno de los dos bandos y no por un empate.

Por último, nos fijamos en la distribución de partidas con un resultado sorprendente y vemos que solo el 6% de las partidas se produce una sorpresa. Este resultado ensalza la calidad del sistema de clasificación por ELO demostrando que es un claro indicador del nivel ajedrecístico de cada jugador.

Análisis ELO

Analizando estadísticas sobre el ELO de los jugadores, observamos que existen algunos patrones marcados según las partidas sean RATED o UNRATED.

Para empezar, vemos que la diferencia media de ELO entre los jugadores es de 154 para partidas RATED y 251 para partidas UNRATED. Esto es debido a que las partidas RATED suelen ser aleatorias (tu no eliges al jugador al que te enfrentas) con un sistema de emparejamientos basado en encontrar jugadores con un nivel similar, mientras que las partidas UNRATED suelen ser partidas entre amigos (los cuales pueden tener mucha diferencia de nivel) en las que se busca la diversión más que la competición.

También vemos que los valores de ELO máximo son mayores para las partidas UNRATED lo que nos puede hacer pensar que al alcanzar un nivel muy muy alto de ELO algunos jugadores dejan de jugar partidas RATED ya que se complica mucho el seguir subiendo ELO.

Análisis Aperturas

En esta sección realizaremos un análisis sobre las 5 aperturas más utilizadas (de las 128 distintas que aparecen en la base de datos)

Las aperturas más usadas son las siguientes:

- Defensa Siciliana (13,22%)
- Defensa Francesa (7,13%)
- Aperturas de peón de Dama (6,13%)
- Apertura Italiana (4,85%)
- Gambito de dama (4,59%)

Porcentaje de Victoria

% Win Rate	Ganador		
Aperturas	Black	Draw	White
Sicilian Defense	49,37%	4,91%	45,72%
French Defense	46,29%	4,99%	48,72%
Queen's Pawn Game	48,89%	4,69%	46,42%
Italian Game	46,00%	4,97%	49,03%
Queen's Gambit	39,91%	3,76%	56,33%

Porcentaje de Sorpresa

% Sorpresa	Sorpresa	
Aperturas	No	Si
Sicilian Defense	94,30%	5,70%
French Defense	94,79%	5,21%
Queen's Pawn Game	93,94%	6,06%
Italian Game	93,95%	6,05%
Queen's Gambit	93,16%	6,84%

Observamos que el porcentaje de sorpresa para cada apertura es similar entre ellos (y respecto al % de sorpresas global) por lo que de manera general no podemos recomendar ninguna apertura en concreto cuando nuestro rival es mejor (veremos más adelante diferenciando según el ritmo de juego)

Sin embargo, observando el porcentaje de victorias si vemos que el porcentaje de victorias del blanco es superior a la media (50%) cuando se juega el gambito de dama. En el caso del jugador negro, el porcentaje de victoria global es del 45% no obstante, cuando se usa la defensa siciliana el porcentaje de victoria del negro es casi del 50%.

Por tanto, basándonos en este análisis se recomienda jugar el gambito de dama cuando juguemos de blancas y la defensa siciliana cuando juguemos de negras.

Análisis según Ritmo de juego

En esta sección nos centraremos en las diferencias existentes en las partidas basadas en los distintos ritmos de juego.

La primera diferencia notable que encontramos es que paradójicamente las partidas con mayor duración de tiempo son las que menos movimientos tienen, esto puede ser debido a que en una partida lenta los movimientos son más precisos ya que se tiene más tiempo para pensar, por tanto, no se desaprovechan las oportunidades de victoria que se presentan en la partida, no como en las partidas Blitz o Rápidas, en las que un error del rival puede pasar desapercibido debido a los apuros de tiempo.

También observamos que la manera en la que acaba la partida es muy parecida en los 3 ritmos de juego, al igual que el porcentaje de victorias del jugador blanco y el negro, aunque como razonamos en la sección de estadísticas generales, en las partidas lentas el porcentaje de tablas es algo superior que el de los ritmos rápidos.

Analizando las aperturas más usadas en los catalogados como ritmos rápidos (partidas Blitz y Rápidas) vemos que se suelen usar las mismas aperturas en todos los ritmos y que suelen tener el mismo porcentaje victorias y de sorpresas.

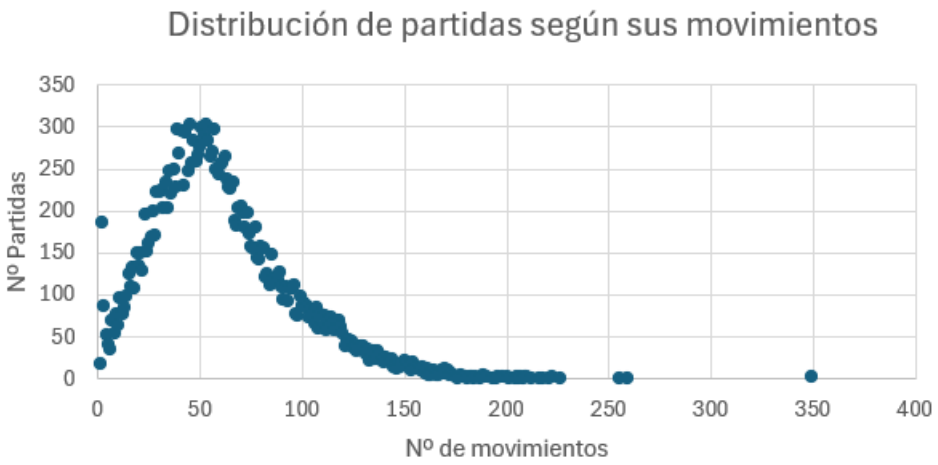
Sin embargo, en partidas lentas nos percatamos que la segunda apertura más usada es la apertura inglesa, una apertura prácticamente inusual en partidas de ritmo rápido y que además su porcentaje de sorpresa es extremadamente alto (27,45%). La apertura inglesa es una apertura que empieza con c4 (movimiento raro a nivel amateur ya que la gran mayoría de partidas se empiezan con e4 o d4) lo que permite a las blancas desarrollar sus piezas de manera sólida y posicional

siendo una muy buena opción para quienes buscan partidas estratégicas a largo plazo.

Análisis Sorpresas y Turnos

En el ámbito de las sorpresas vemos que el porcentaje de estas se mantiene constante a niveles generales acentuándose un poco más en las partidas lentas y si miramos este mismo resultado según el nivel de juego de los jugadores, notamos que a medida que aumenta el nivel el porcentaje de sorpresas disminuye.

En cuanto al número de movimientos de las partidas la distribución es la siguiente:



Aquí podemos ver que lo más común es que una partida tenga entre 40 y 60 movimientos (el promedio total es 61) y que las partidas de más de 125 movimientos son prácticamente inexistentes.

Como último análisis de esta sección observamos que las partidas son más largas en cuestión de movimientos a medida que sube el nivel de los jugadores, este resultado puede ser debido a las altas capacidades defensivas que tienen los jugadores de alto nivel, lo que les permite resistir posiciones críticas de ataque del rival las cuales un jugador principiante o amateur no serían capaz de aguantar.

Conclusiones

A partir del análisis de las partidas de la base de datos, se han extraído varios hallazgos relevantes que ayudan a entender patrones en el ajedrez en línea, permitiendo obtener insights prácticos para jugadores de distintos niveles:

Efectividad de Aperturas:

Las aperturas más utilizadas, como la Defensa Siciliana y el Gambito de Dama, presentan patrones interesantes de victoria según el color de las piezas. El Gambito de Dama es especialmente efectivo para las blancas, en cambio, la Defensa Siciliana ofrece a las negras una buena posibilidad de igualar el porcentaje de victorias con las blancas.

La Apertura Inglesa, aunque rara en partidas rápidas, muestra un alto porcentaje de sorpresas en partidas lentas, lo que la convierte en una opción estratégica para quienes buscan explorar configuraciones menos comunes.

Impacto del Ritmo de Juego:

Se observa una preferencia por ritmos rápidos (Blitz y Rápidas), representando el 95% de las partidas. Esto refleja una tendencia en el ajedrez en línea hacia formatos dinámicos y rápidos.

El porcentaje de partidas que terminan en empate es mayor en ritmos lentos, lo que sugiere que estos tiempos permiten un juego más equilibrado y menos caótico, favoreciendo desenlaces menos abruptos.

Distribución de Sorpresas y ELO:

Las sorpresas representan solo un 6% del total de las partidas. Esta cifra demuestra que el sistema de clasificación por ELO es en general un indicador preciso del nivel de juego.

Sin embargo, las sorpresas son ligeramente más frecuentes en partidas lentas, lo que podría explicarse por la posibilidad de aplicar estrategias menos convencionales, como la Apertura Inglesa, que puede desconcertar a jugadores menos experimentados en este tipo de ritmo.

Limitaciones y Perspectivas:

Aunque los datos analizados provienen de partidas en línea y reflejan una muestra significativa, pueden existir diferencias en el comportamiento y resultados respecto a partidas presenciales o en torneos formales.