TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi d'obertures dels jugadors d'escacs catalans

Autor: Pau Toquero Gracia

Director: Isaac Subirana Cachinero

Departament: Genètica, microbiologia i estadística

Convocatòria: Juny 2024

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM I PARAULES CLAU

Els escacs és un esport de fa més de mil anys el qual ha anat evolucionant al llarg dels anys fins a arribar a tenir un total de 180.000 jugadors federats i més de 20 milions de jugadors en línia. Per a poder classificar a tots aquests jugadors, es va crear un sistema de puntuació denominat Elo on els millors jugadors tenen la puntuació més alta i els jugadors més fluixos la més baixa. Per això, l'objectiu d'aquest TFG consisteix a veure si a partir de les característiques d'una partida d'escacs, especialment la diferència d'Elo entre els dos jugadors, es pot predir el resultat d'una partida d'Escacs. Així que, hem agafat una base de dades de 5.014 partides descarregades d'internet de jugadors de la Federació Catalana d'Escacs. Després de classificar les diferents obertures, crear la diferència i la mitjana d'Elo de cada partida, creem dos models lineals multinomial, amb la variable resultat de la partida com a variable dependent. Un d'ells només amb la variable de diferència d'Elo i l'altre amb la diferència d'Elo, la mitjana d'Elo i el grup d'obertura que es juga a la partida. Amb els models, podem veure que no és possible predir al 100% el resultat d'una partida sobre la base de les seves característiques però ens dona una precisió de gairebé el 60% d'encertar el resultat.

Paraules clau: escacs, Elo, resultat, predicció, model multinomial, precisió

ABSTRACT AND KEYWORDS

Chess is a sport that dates back more than a thousand years and has evolved over the years to have a total of 180,000 federated players and more than 20 million players online. To be able to classify all these players, a rating system called Elo was created where the best players have the highest rating and the weakest players the lowest. Therefore, the aim of this dissertation is to see if the characteristics of a chess game, especially the Elo difference between the two players, can predict the outcome of a game of chess. So, we have taken a database of 5014 games downloaded from the internet of players of the Catalan chess federation. After classifying the different openings, creating the difference and the average Elo of each game, we created two multinomial linear models, with the game outcome variable as the dependent variable. One of them only with the Elo difference variable and the other with the Elo difference, the Elo average and the opening group played in the game. With the models, we can see that it is not possible to predict 100% the result of a game based on its characteristics, but it gives us an accuracy of almost 60% of getting the result right.

Keywords: chess, Elo, result, prediction, multinomial model, precision

CLASSIFICACIÓ AMS

62J12 Generalized lineal models

AGRAÏMENTS

Al meu tutor Isaac Subirana, per acceptar la meva proposta d'un treball basat en els escacs i ajudar-me durant aquests mesos i tots els seus consells.

A la meva família, per tots aquests anys de competicions, classes i viatges dedicats als escacs que han creat aquesta passió que li tinc a aquest esport.

ÍNDEX DE CONTINGUTS

1.- INTRODUCCIÓ

1.1.- Evolució dels escacs

Els escacs, tal com els coneixem avui, van néixer del joc indi "chaturanga" abans de l'any 600. El joc es va estendre per Àsia i Europa al llarg dels segles següents i va acabar evolucionant fins a convertir-se en el que coneixem com a escacs al voltant del segle XVI. Les peces representaven originalment les unitats militars habituals en la guerra de l'època: infanteria, cavalleria, elefants, carros, un general i un rei.

Amb el pas dels anys, els escacs han anat evolucionant, des dels primers mestres com Ruy López, Lucena o Greco, que amb els seus estudis i els seus llibres van ajudar al desenvolupament dels escacs analitzant per primera vegada les obertures i alguns finals senzills.

A mesura que passaven els segles, els escacs es van anar desenvolupant en quant a l'estratègia. Trobarem, per exemple, que en els segles XVIII i XIX es desenvolupa els escacs romàntics, caracteritzat per les partides d'atac amb tàctiques de sacrifici amb l'únic objectiu de donar escac i mat. En aquesta època destaquem jugadors com Philidor o Morphy.

El 1886, trobem el primer campió del món, Steinitz, i l'inici d'un nou sistema de joc caracteritzat pel joc posicional, on s'abandona el joc de sacrificis i es comencen a utilitzar les posicions estratègiques basades en obertures caracteritzades per dominar el centre, intentant acumular petites avantatges en espai, caselles clau, diagonals i files.

Un cop exposada la base teòrica de les partides, on, com hem comentat, existien les posicions més tàctiques i les més estratègiques, arribem a un domini soviètic dels escacs durant tot el segle XX, amb l'excepció del mestre americà Bobby Fischer.

Finalment, en els darrers anys, hem vist un desenvolupament dels ordinadors en el món dels escacs, destacant sobretot la victòria de l'ordinador Deep Blue contra el campió del món del moment, Garry Kasparov, l'any 1997.

Avui dia, els millors "jugadors" són una combinació d'equips humans i ordinadors, en una variant dels escacs anomenada escacs avançats, on un jugador humà tria un programa per ajudar-lo a prendre les decisions que finalment pren. Els escacs centaure, com se li diu, superen tant els jugadors d'escacs humans com els ordinadors individualment (Hector Apolo Rosales Pulido 2016).

Tot i que els escacs és un esport que ha estat dominat pels ordinadors en els darrers anys, trobem que arreu del món, hi ha un total de 183207 jugadors actius actualment registrats a la FIDE, la qual es la Federació Internacional d'Escacs, dels quals, 16739 són espanyols, sent Espanya el tercer país del món amb més jugadors federats superats per Índia i Rússia. A part dels jugadors federats, tenim més de 150 milions de jugadors registrats a la pàgina chess.com,

la qual es la pàgina web més famosa del món per jugar partides online el que ens fa entendre que cada cop més triomfa els escacs online per sobre dels escacs presencials.

De la mateixa manera que ha passat als escacs, en els darrers anys ha hagut un desenvolupament dels ordinadors que han ajudat a millorar l'anàlisi de les dades que obtenim dels esports. És per la importància de la disciplina Sports Statistics, la qual ens ajuda a analitzar-ho tot a partir del model Sports Analytics, que coneixem com la cerca de millores en el rendiment esportiu a través de l'anàlisi de dades (Benjamin Baumer).

La funció d'una persona especialista en estadística esportiva és recopilar, analitzar i interpretar les dades dels esdeveniments esportius de manera que s'obtingui informació valuosa, per exemple, per als jugadors, equips, lligues i organitzacions de la indústria de l'esport. Amb aquesta informació també es pretén, per exemple, millorar les estratègies del joc i la presa de decisions.

Els escacs han estat cèlebres per la seva profunditat i complexitat estratègiques. Exigeixen dels seus jugadors previsió, planificació i la capacitat de navegar a través d'un ampli ventall de possibles jugades i contra jugades. Això fa dels escacs no només un joc d'intel·lecte i intuïció, sinó també un terreny fèrtil per a l'aplicació de la ciència de dades. A la inversa, la ciència de les dades, que es nodreix de la selecció i el sentit de densos conjunts de dades per extreure idees processables, troba en els escacs un camp de proves ideal per al desenvolupament i perfeccionament de models analítics i algorismes. Aquesta relació simbiòtica entre els escacs i la ciència de dades està obrint noves àrees d'innovació i transformant la manera de jugar, analitzar i aprendre (Rohan Whitehead).

Quan parlem de Sports Analytics en els escacs, trobem especialment l'estudi sobre el concepte Elo (Arthur Berg, 2020; Mark E. Glickman, 1995), el qual és el sistema de mesura del nivell dels jugadors utilitzada per separar-los per ranking. També trobem estudis amb la finalitat de deduir el resultat a partir d'una posició (Hector Apolo Rosales Pulido, 2016) amb l'ús de Machine Learning i, últimament, per detectar si els jugadors estan fent trampes en les partides en línia (Kenneth Regan, 2013).

1.2.- Objectiu del treball

L'objectiu principal d'aquest treball és, a partir d'una una base de dades amb diferents partides d'escacs, veure si existeix alguna relació entre la diferència d'Elo dels jugadors i el resultat de la partida.

Addicionalment, el treball planteja altres objectius a analitzar, com veure si les obertures influeixen en el resultat i quines són les més jugades, avaluar la relació entre les obertures jugades i l'Elo dels jugadors.

2.- METODOLOGIA

2.1.- Procés i creació de les dades

2.1.1.- Definició d'Elo i sistema de càlcul

El sistema de puntuació Elo, utilitzat en escacs, mesura la força relativa d'un jugador en comparació a uns altres. El seu creador, Arpad Elo, va ser un professor de física nord-americà i un mestre d'escacs que va centrar els seus esforços a millorar la forma en la qual la Federació Estatunidenca d'Escacs mesurava el nivell de joc dels jugadors d'escacs.

El sistema de puntuació Elo va ser adoptat oficialment per la Federació Estatunidenca d'Escacs en 1960 i per la FIDE en 1970. Avui dia, existeixen una gran varietat d'organitzacions i portals d'escacs que utilitzen també aquest mètode per a avaluar als jugadors. Chess.com empra una versió modificada del sistema de puntuació Elo, coneguda com a sistema Glicko, que té en consideració més variables per a determinar la força d'un jugador.

La puntuació Elo d'un jugador d'escacs està representada mitjançant un número basat en l'actuació d'aquest jugador en partides avaluades disputades prèviament. Després de cada partida avaluada, l'Elo de tots dos jugadors varia en funció del resultat de la partida. Encara que és una creença generalitzada que el sistema de puntuació Elo mesura la força absoluta d'un jugador d'escacs, convé observar que això no és realment així. El que calcula aquest sistema és el resultat probable de l'enfrontament entre un jugador i un altre.

Quan un nou jugador comença a jugar, se li atorga un rating de 1000 punts d'Elo, que es el mínim de Elo estipulat per la FIDE, i per tant, a la nostra base de dades, quan no tenim l'Elo del jugador, tant si juga amb peces blanques com amb peces negres, se'ls hi assigna directament un Elo de 1000 punts, ja que aquest cas es dona quan no s'han jugat un número mínim de partides però el jugador està federat.

Un cop els jugadors ja tenen el Elo assignat, cada partida que juguen contra jugadors amb Elo fa que variï, ja sigui augmentant o disminuint segons els seu resultats.

La fórmula que es fa servir per poder canviar aquest Elo es la següent:

$$R'_A = \begin{cases} R_A + K_A(1 - E_{AB}), & Si\ A\ guanya \\ R_A + K_A\left(\frac{1}{2} - E_{AB}\right), & Si\ A\ fa\ taules \\ R_A + K_A(0 - E_{AB}), & Si\ A\ perd \end{cases}$$

De la que les diferents variables serien:

La primera que ens surt es $\boldsymbol{R_A}$ la qual fa referencia al Elo del jugador quan comença la partida.

Després tenim la variable $\boldsymbol{K_A}$ la qual depèn del jugador que està jugant la partida. Aquesta K, fa referència a la quantitat de Elo que pot pujar o baixar un jugador a cada partida, a continuació detallo els tipus de K que existeixen:

1. $K = 40$

Aquesta K es la que ens correspon quan comencem a jugar o quan som jugadors menors de 18 anys amb un Elo inferior a 2300. Al guanyar a un jugador que posseeix 400 o més de 400 punts de Elo més que nosaltres, podem pujar fins a 36 punts d'Elo en cas de guanyar i restar 36 en cas contrari. Quan diem que fins a 400 es perquè es el màxim que es comptabilitza a la hora de fer la variació de l'Elo de cada partida, un cop es supera els 400 punts, s'agafa com si el rival tingues 400 més o 400 menys mes que nosaltres per fer el càlcul del nou Elo.

2. $K = 20$

És el més habitual, la que tenen la majoria dels jugadors amateurs o semi-professionals. Com podem observar, el número K ha baixat a la meitat, per tant, el límit que podem pujar o baixar es redueix a la meitat, en aquest cas, 18.

3. $K = 10$

La K 10 es coneguda com la K dels mestres. Sol s'aconsegueix quan arribem a 2400 d'Elo algun cop a la vida. A partir d'aquell moment, es més complicat augmentar l'Elo, ja que el màxim d'Elo que es pot pujar o restar en una partida passa a ser 9.

Seguidament, trobem la variable $\boldsymbol{E_{AB}}$. Aquesta variable fa referencia al resultat esperat del jugador A respecte al jugador B i per poder trobar aquest resultat, s'utilitza la fórmula següent:

$$Prob(A\ guanya\ a\ B) = E_{AB} = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}}$$

A la fórmula trobem $R_A\ i\ R_B$ que fa referencia al Elo que té el jugador A, que seria el jugador al qual li volem variar el seu Elo i el jugador B el qual es el seu rival a la partida.

Per tant, un cop tenim totes les incògnites, ens fixem que segons el resultat obtingut a la partida, es farà una formula o una altra.

Quan parlem de resultats, trobem que tenim tres resultats possibles:

En cas de Victoria del jugador A, es farà servir $1 - E_{AB}$ ja que el resultat obtingut quan guanya el jugador A es 1.

En cas de taules, és a dir empat, la fórmula que es fa servir és $1/2 - E_{AB}$, ja que el resultat obtingut és taules i les taules es posen com a ½.

Finalment, en cas de derrota, la fórmula que es farà servir es $0 - E_{AB}$ ja que el resultat que s'obté en cas de derrota és 0.

2.1.2.- Identificació de les fonts d'informació

Per poder trobar unes dades amb criteri per poder analitzar l'objectiu exposat, el primer que necessitava decidir era els jugadors dels que es faria l'estudi de les seves partides. Llavors, com que no existia cap base de dades creada, es va decidir agafar els jugadors amb llicència catalana que estiguessin federats a la Federació Catalana d'Escacs el més de febrer de 2024.

Un cop es va obtenir la llista de jugadors, a través de la pròpia pàgina de la Federació Catalana d'Escacs (www.escacs.cat), es va decidir fer una segona selecció de jugadors, ja que havia un total de 7.517 jugadors amb fitxa a la federació catalana, per tant, es seleccionar els jugadors titulats.

Aquesta titulació l'atorga la FIDE (Federació Internacional d'Escacs), que regula les diferents titulacions que poden assolir els jugadors un cop aconsegueixen cert nivell o Elo. En aquest cas, vam decidir que seleccionaríem jugadors amb les titulacions de Gran Mestre, Mestre internacional, Mestre FIDE i Candidate Master, tant el títol absolut com el femení. Per a completar la selecció, a més d'aquest títols de la FIDE, també vam seleccionar la titulació que atorga la Federació Catalana d'Escacs com a Mestre Català i només vigent a Catalunya.

En resum, els jugadors seleccionats són els que han aconseguit una d'aquestes 8 titulacions de la FIDE i aquells que han aconseguit la titulació detallada de la Federació Catalana d'Escacs.

Un cop feta la selecció de jugadors, vam obtenir un total de 504 jugadors i jugadores per formar la base de dades a analitzar. D'aquests jugadors, el següent pas va ser consultar a la pàgina Chess-Results (https://chess-results.com/) i descarregar les seves partides.

El Programa ChessBase està especialitzat en escacs i en bases de dades de partides d'escacs, donant un total de 24.827 partides descarregades. Un cop descarregades les partides i emmagatzemades a la base, es seleccionaven les partides que tinguessin el nom dels jugadors, el seu Elo, l'obertura jugada, la data i el resultat per passar-la a una nova taula per poder analitzar posteriorment aquesta base de dades.

Aquesta selecció de jugadors, la recerca de les seves partides i la posterior neteja de la base de dades en el Programa ChessBase va requerir molta dedicació, aproximadament dos mesos des de el 29 de febrer fins a principis de maig, per poder completar una base de dades de jugadors i partides, que complien els requisits que vam decidir per poder fer l'anàlisi de les qüestions plantejades a l'objectiu d'aquest treball.

2.1.3.- Selecció de les variables d'estudi

Com s'ha explicat fins ara, el punt de partida era d'una única base de dades introduïda al programa ChessBase, amb les partides descarregades d'internet segons la Taula 2.1

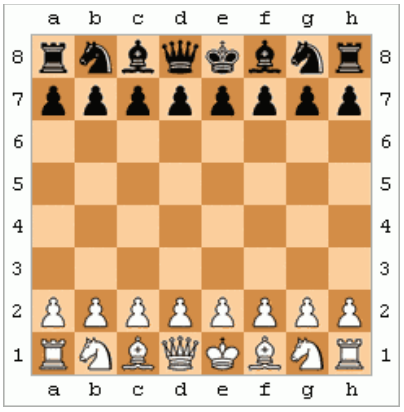
Taula 2.1.- TAULA 1

VARIABLE	DESCRIPCIÓ	TIPUS
Blancas	Nom jugador de Blanques	Qualitativa
Elo Blanco	Elo Jugador de Blanques	Numèrica
Negras	Nom Jugador Negres	Qualitativa
Elo Negro	Elo Jugador de Negres	Numèrica
Resultado	Resultat de la partida	Qualitativa Dicotòmica
Jugadas	Jugades de la partida	Numèrica
ECO	Codi de la Obertura	Qualitativa
Torneo	Torneig on s'havia disputat la partida	Qualitativa
Fecha	Data de la partida	Qualitativa

2.1.3.1.- Creació de taules i variables

A la figura 2.1 es mostra un tauler d'escacs el qual està representat per vuit fileres i vuit columnes fent un total de 64 caselles, amb 16 peces, vuit blanques i vuit negres. Quan es comença una partida de escacs, el primer en jugar és el blanc i la jugada que fa es una obertura o una defensa.

Figura 2.1 Tauler d'escacs



La obertura, és el conjunt de moviments amb què comença una partida del joc dels escacs. Les seqüències estàndard de jugades inicials són anomenades obertures quan són jugades per les blanques, o bé defenses, quan són jugades per les negres. Hi ha unes quantes dotzenes d'obertures diferents, amb centenars o milers de variants i subvariants.

A partir del tipus d'obertura o de defensa que es juga, obtenim la següent taula:

Taula 2.2.- TAULA 2

VARIABLE	DESCRIPCIÓ	TIPUS
Apertura	Codi de la Obertura	Qualitativa
Nombre	Nom de la Obertura	Qualitativa
GRUPO	Tipus de Obertura segons el tipus de estructura	Qualitativa

El nombre d'obertures existents és molt gran però es poden agrupar en 5 grans grups etiquetats per les primeres lletres de l'abecedari. A continuació es descriu cadascun dels grups.

A: són les anomenades obertures de flanc (aquelles en què la jugada inicial del blanc no és 1.e4 ni 1.d4), així com les línies antigues de la defensa índia (1.d4 Cf6).

Un exemple, seria la obertura anglesa, la qual es comença amb la jugada 1.c4.

Figura 2.2 obertura anglesa

B: són aquelles que es comença amb el peó de rei fent la jugada 1.e4 però no es respon la jugada 1...e5, entre elles, tenim la defensa siciliana.

Figura 2.3 defensa siciliana

C: Dels diferents grups d'obertures que fa referencia a les obertures obertes, que son aquelles on contra la jugada 1.e4, si que es juga la jugada 1...e5, una d'elles, es la obertura Ruy Lopez també coneguda com a obertura espanyola.

Figura 2.4 obertura espanyola

D: són aquelles, on hi ha les obertures tancades i semitancades. La primera jugada de la partida es el peó de dama a 1.d4, entre elles, trobem el gambit de dama.

Figura 2.5 gambit de dama

E: hi ha les defenses índies que son aquelles on la partida es comença amb el peó de dama a 1.d4 i es segueix amb moviment de cavall a 1...Cf6, entre les més reconegudes, destacar la obertura catalana.

Figura 2.6 Obertura Catalana

Per codificar les obertures utilitzem les lletres que acabem d'explicar pels 5 tipus d'obertures (A, B, C, D, E) i afegim una sots-classificació en funció de cada tipus d'esquema que s'utilitza a cada obertura i que s'anomena Nombre, codificat del 0 al 99. Per tant, cada obertura té un

codi forma per una lletra de la A a la E + un dígit del 0 al 99. Per exemple, el codi B55 correspon a l'esquema de la defensa siciliana per una obertura semioberta.

La taula amb les dades de les partides (Taula 2.1) obtinguda amb el programa ChessBase es descarrega en un Excel. El nom i el tipus de les variables o columnes d'aquesta taula es descriu Taula 2.3:

Taula 2.3.- TAULA 3

VARIABLE	DESCRIPCIÓ	TIPUS
Blancas	Nom jugador de Blanques	Qualitativa
Elo Blanco	Elo Jugador de Blanques	Numèrica
Negras	Nom Jugador Negres	Qualitativa
Elo Negro	Elo Jugador de Negres	Numèrica
Apertura	Codi de la Obertura	Qualitativa
Fecha	Data de la partida	Qualitativa
Resultado	Resultat de la partida	Qualitativa Dicotòmica

2.1.3.2.- Taula Final

A partir de les Taules 2.2 i 2.3 es fan els següents passos per obtenir la taula final:

1.- El primer s'uneix la Taula 2.2 i la Taula 2.3, ja que tenen en comú la variable APERTURA, fent que les variables NOMBRE i GRUPO s'afegeixin a la nova Taula.

2.- En segon lloc, es creen tres noves variable:

a) la primera s'anomena Elo_Difference, la qual surt de restar la variable Elo Blancas amb la variable Elo Negras de la taula 2.3.

b) la segona se li diu Media_ELO on es calcula la mitjana de Elo dels 2 jugadors junts a la partida.

c) I finalment, una tercera variable s'anomena Year i que fa referencia a l'any el qual s'ha jugat la partida, extreta de la variable Fecha.

Un cop completat els dos passos, la Taula Final resultant amb la que es fa l'anàlisi és la següent:

Taula 2.4.- TAULA FINAL

VARIABLE	DESCRIPCIÓ	TIPUS
Blancas	Nom jugador de Blanques	Qualitativa
Elo Blanco	Elo Jugador de Blanques	Numèrica
Negras	Nom Jugador Negres	Qualitativa
Elo Negro	Elo Jugador de Negres	Numèrica
Apertura	Codi de la Obertura	Qualitativa
Fecha	Data de la partida	Qualitativa
Resultado	Resultat de la partida	Qualitativa Dicotòmica
Nombre	Nom de la Obertura	Qualitativa
GRUPO	Tipus de Obertura segons el tipus de estructura	Qualitativa
Elo_Difference	Diferencia entre Elo blanc i Elo Negre	Numèrica
Media_ELO	Mitjana de Elo entre el jugador de blanques y el de negres	Numèrica
Year	Any el qual es va jugar la partida	Numèrica

2.1.3.3.- Tractament de les dades de la Taula Final

Un cop es té la Taula Final, observem que hi ha *missings* i dades duplicades.

Les dades duplicades venen donades degut a que a l'hora d'extreure les dades, s'ha anat buscant jugador per jugador les partides, per tant, a la base de dades inicial, les partides que eren fetes per dos jugadors titulats federats a la Federació Catalana d'Escacs, sortien dos cops. Pel que de les 5.913 partides que hi ha al principi, 489 estan repetides

$$\% \text{ de duplicades} = \frac{489}{5913} = 0.082699137$$

Per tant, de la taula final (Taula 2.4), hi ha un 8.83% de les dades son dades duplicades.

A part, les dades buides, com no es pot deduir la dada, s'elimina. En total, s'obté una base de dades amb un total de 5.401 partides d'escacs dels jugadors seleccionats.

2.2.- Mètodes per analitzar la base de dades

2.2.1.- Anàlisi Univariant i Bivariant

Una anàlisi Univariant proporciona una sèrie d'eines per descriure, tabular, representar i treure gràfics de variables de la manera més útil i eficaç possible per poder obtenir informació rellevant.

Per poder fer l'anàlisi Univariant de les variables categòriques, es fa a partir de la funció *summary()* la qual, tornarà les freqüències de la variable i algunes mesures de centralitat i localització d'una variable numèrica.

L'altre forma de representar les dades, és mitjançant el paquet *ggplot2.* Segons el tipus de variable i la manera en la que es decideixi representar les dades, les principals formes de representació gràfica que farem servir, són, l'histograma, el diagrama de línies, el diagrama de barres, el diagrama de caixes i el diagrama de dispersió.

A més, es faran gràfics de normalitat per comprovar si són normals o no, com és el cas del test de Shapiro-Wilk o Kolmogorov-Smirnov.

L'anàlisi Bivariant, s'encarrega de l'anàlisi estadístic de dues variables. L'anàlisi bivariant busca principalment explicar com dues variables es relacionen entre si. Quan es parla, doncs, d'associacions o efectes entre variables, només es fa mitjançant les tècniques de l'anàlisi bivariant, que permetrà suggerir des d'un punt de vista estadístic si el comportament d'una variable està parcialment determinat per l'altra.

Per visualitzar i quantificar estadísticament una associació bivariant, s'utilitza una tècnica diferent segons si les variables són numèriques o categòriques.

La ***taula de contingència*** és la manera de representar una relació bivariant quan les variables independent i dependent són categòriques. Normalment es representen els valors de la variable independent a l'eix horitzontal i els valors de la dependent a l'eix vertical, de manera que cada cel·la de dins de la taula representa el recompte total d'observacions que corresponen a cada combinació entre nivells de les dues variables. Un cop es té a cada cel·la el nombre de freqüències, es calcula els percentatges de cada freqüència sobre el total de la columna.

Quan es vol observar l'associació entre una variable independent categòrica i una dependent numèrica s'utilitza la **diferència de mitjanes**. Hi ha diverses maneres de visualitzar una diferència de mitjanes. Es pot, per exemple, utilitzar un diagrama de barres en el qual l'alçada de les barres representi la mitjana de cada categoria. Dues eines també comunes per representar visualment la diferència de mitjanes són el diagrama de caixes i el diagrama de dispersió.

La eina que es fa servir per fer la diferència de mitjanes es la Prova Tukey HSD, que és una tècnica estadística utilitzada per realitzar comparacions múltiples entre mitjanes de grups després d'una anàlisi de variància (ANOVA). L'objectiu és determinar quines mitjanes són significativament diferents entre elles. Per fer-la, es realitza una ANOVA per determinar si hi ha diferències significatives entre les mitjanes de diversos grups. Si l'ANOVA indica que hi ha diferències significatives (p-valor < nivell de significança, típicament 0,05), es pot procedir amb la prova Tukey HSD. Es calcula la diferència entre cada parella de mitjanes i es compara amb un valor crític basat en la distribució de Tukey. La diferència significativa (HSD) es calcula com:

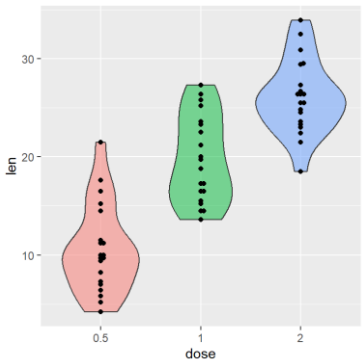
$$HSD = q\sqrt{\frac{MS_w}{n}}$$

On q es el valor crític de la distribució de Tukey, MS_w es la mitjana dels quadrats obtinguda a l'ANOVA i n es el nombre d'observacions del grup.

Un cop s'obté el Tukey fet, es fa servir el paquet *agricolae* la qual ajuda a separar les diferents mitjanes en grups a partir de lletres. Quan les diferents respostes tenen la mateixa lletra, vol dir que no són significatives entre elles mentre que els grups amb lletres diferents sí que ho són.

Per explicar-ho de forma gràfica es fa servir els Violin Plot. Els violin plots son semblants als Box-plots, que proporcionen informació resumida de la distribució, com ara la mediana, els quartils i possibles valors atípics. Però amb la particularitat de que es combina amb el Kernel Density Plot, una tècnica que estima la densitat de probabilitat d'una variable de manera suau i contínua, pel es fa servir per mostrar la distribució de una variable numèrica al llarg de diferents categories o grups aquest gràfic proporciona una representació visual de la densitat de una variable una tècnica que estima la densitat de probabilitat d'una variable de manera suau i contínua. Els Violin Plots són simètrics respecte a la línia central vertical, mostrant la mateixa densitat a ambdós costats. Això facilita la comparació visual de la distribució de dades en diferents categories o grups.

Figura 2.7 Violin Plot



2.2.2.- GLM (regressió logística i multinomial)

El model lineal generalitzat o, més breument, GLM (en anglès: *General Linear Model*) és una classe bastant àmplia de models introduïts per Nelder i Wedderburn (1972). Aquests models assumeixen que les respostes provenen d'una distribució que pertany a una família de distribucions anomenada família exponencial de distribucions o família de models de dispersió exponencial (EDM, que ve de l'anglès: *Exponential Dispersion Model*), concepte va ser presentat per primera vegada per Jorgensen (1987).

Els EDMs continus inclouen les distribucions Normal, Normal inversa, Exponencial, Weibull i Gamma. Els EDMs discrets inclouen les ditribuciones Binomial, Poisson i Binomial negativa.

La família de distribucions EDM permet que els GLMs siguin ajustats a un rang ampli de tipus de dades, com a dades binàries, proporcions, recomptes, dades contínues positives i dades contínues amb zeros exactes.

Un GLM multinomial és una extensió dels models lineals generalitzats que permeten modelar variables de resposta categòriques amb més de dues categories. A diferència de la regressió logística binària, que es limita a dues categories, el GLM multinomial es pot aplicar quan la variable de resposta té tres o més categories. És a dir, és un model que s'utilitza per a predir les probabilitats dels diferents resultats possibles d'una variable dependent distribuïda categòricament, donat un conjunt de variables independents (que poden ser de valor real, de valor binari, de valor categòric, etc.).

Existeixen múltiples formes equivalents de descriure el model matemàtic subjacent a la regressió logística multinomial. Això pot dificultar la comparació dels diferents tractaments del tema en diferents textos.

La idea que es troba en totes elles, com en moltes altres tècniques de classificació estadística, és construir una funció de predicció lineal que construeixi una puntuació a partir d'un conjunt de ponderacions que es combinen linealment amb les variables explicatives (característiques) d'una observació donada utilitzant un producte punt:

$$f(X_i, k) = B_k \cdot X_i,$$

On X_i és el vector de variables explicatives que descriuen l'observació i, B_k és un vector de ponderacions (o coeficients de regressió) corresponents al resultat k, i puntuació (X_i, k) és la puntuació associada a l'assignació de l'observació i a la categoria k.

·Predicció Lineal

Com en altres formes de regressió lineal, la regressió logística multinomial utilitza una funció predictiva lineal $f(k, i)$ per a predir la probabilitat que l'observació i tingui el resultat k, de la manera següent:

$$f(k, i) = B_{0,k} + B_{1,k}x_{1,i} + B_{2,k}x_{2,i} + \cdots + B_{M,k}x_{M,i}$$

on $B_{m,k}$ és un coeficient de regressió associat amb la m-ésima variable explicativa i el k-ésimo resultat. Com s'explica en l'article sobre regressió logística, els coeficients de regressió i les variables explicatives s'agrupen normalment en vectors de grandària M+1, de manera que la funció predictora pugui escriure's de forma més compacta:

$$f(k, i) = B_k \cdot x_i$$

on $B_k$és el conjunt de coeficients de regressió associats al resultat k, i x_i és el conjunt de variables explicatives associades a l'observació i.

Per a arribar al model logit multinomial, es pot imaginar, per a K resultats possibles, l'execució de K models de regressió logística binària independents, en els quals es tria un resultat com a "pivot" i, a continuació, els altres K-1 resultats es tornen per separat contra el resultat pivoti. Si el resultat K (l'últim resultat) es tria com a pivot, les equacions de regressió de K-1 són:

$$ln\frac{\Pr\,(Yi=k)}{\Pr\,(Yi=K)}=B_k\cdot x_i\,,k<K$$

Per tant, es fa servir,

$$ln\frac{\Pr(Resultado)=1/2}{\Pr\,(Resultado)=1}$$

$$ln\frac{\Pr(Resultado)=0}{\Pr\,(Resultado)=1}$$

Per poder predir el nostre model on Resultado=1 quan guanya el blanc, Resultado=1/2 quan el resultat acaba en taules i finalment Resultado=0 quan la victòria es del jugador de negres.

· Matriu de confusió

Una matriu de confusió és una eina que s'utilitza per a avaluar el rendiment d'un model de classificació. Permet visualitzar l'acompliment del model en mostrar les prediccions correctes i incorrectes comparades amb els valors reals de les classes en el conjunt de dades.

Per a problemes de classificació amb múltiples categories, com el cas en el meu anàlisis on hi ha tres possibles resultats (0, ½, 1), la matriu de confusió s'estén per a incloure totes les combinacions possibles de prediccions i valors reals.

Taula 2.5 Matriu de Confusió

Actual/Predit	**0**	½	**1**
0	00	0½	01
½	½1	½½	½1
1	10	1½	11

On:

00 és el nombre d'observacions predites com a Classe 0 i que realment són Classe 0 (Veritables Positius per a Classe 0).

0½ és el nombre d'observacions predites com a Classe 0 però que realment són Classe ½ (Falsos Positius per a Classe 0).

01 és el nombre d'observacions predites com a Classe 0 però que realment són Classe 1 (Falsos Positius per a Classe 0).

½0 és el nombre d'observacions predites com a Classe ½ però que realment són Classe 0 (Falsos Negatius per a Classe 0).

½½ és el nombre d'observacions predites com a Classe ½ i que realment són Classe ½ (Veritables Positius per a Classe ½).

I així successivament per a cada combinació.

Per a problemes multi classe, les mètriques es calculen per a cada classe i després es fan una mitjana de. Les mètriques clau inclouen:

Exactitud Global (Accuracy): Proporció de prediccions correctes.

$$Exactitud = \frac{Suma\ de\ tots\ els\ elements\ diagonals}{Total\ de\ elements}$$

Un valor d' *accuracy* més pròxim a 1 indica un millor rendiment del model, ja que significa que el model està predint correctament una major proporció dels casos.

Kappa: El coeficient Kappa de Cohen és una mesura que compara la categoria predita pel model amb la categoria real. S'utilitza per a avaluar la qualitat d'un model de classificació

$$k = \frac{p_o - p_e}{1 - p_e}$$

On: p_o es la proporció de cops que els avaluadors estan d'acord (observada)

p_e es la proporció de vegades que s'esperaria que els avaluadors estiguessin d'acord per atzar (esperada)

Quan k=1, es diu que tenia una concordança perfecta, mentre que quan k=0, diem que la concordança es la esperada per l'atzar.

Es pot donar el cas on k<0 i en aquests casos, es troba que la concordança es pitjor que fet al atzar.

Aquests conceptes són essencials per a avaluar l'eficàcia d'un model de classificació, proporcionant una idea de que bé el model està fent les seves prediccions i quant millor l'està fent en comparació amb l'atzar

3. RESULTATS

3.1.- Anàlisi Univariant:

3.1.1.- ELO

Un cop tenim la Taula Final (n=5401) detallada a la Taula 2.4, es descriuen totes les dades relacionades amb el concepte Elo a la Taula 3.1, a partir de diferents *summary.*

Taula 3.1 Taula resum de les variables Elo Blanco, Elo Negro, Elo_Difference, Media_ELO

Min.	1st Qu	Median	Mean	3rd qu	Max
Elo Blanco					
1000	2182	2329	2300	2458	2800
Elo Negro					
1000	2177	2324	2296	2455	2909
Elo_Difference					
-1462	-150	1	4.728	163	1426
Media_Elo					
1000	2170	2318	2298	2443	2808

Observem que les dades referents al Elo, estan censurades per la esquerra, ja que totes aquelles dades de les variables Elo Blanc i Elo negre que eren 0 o no tenien cap número associat, se li assigna directament un Elo de 1000 punts, ja que es el mínim que estipula la FIDE.

A continuació, fem una representació gràfica de les dades a partir de diferents histogrames.

Figura 3.1. Histograma conjunt de Elo Blanco i Elo Negro

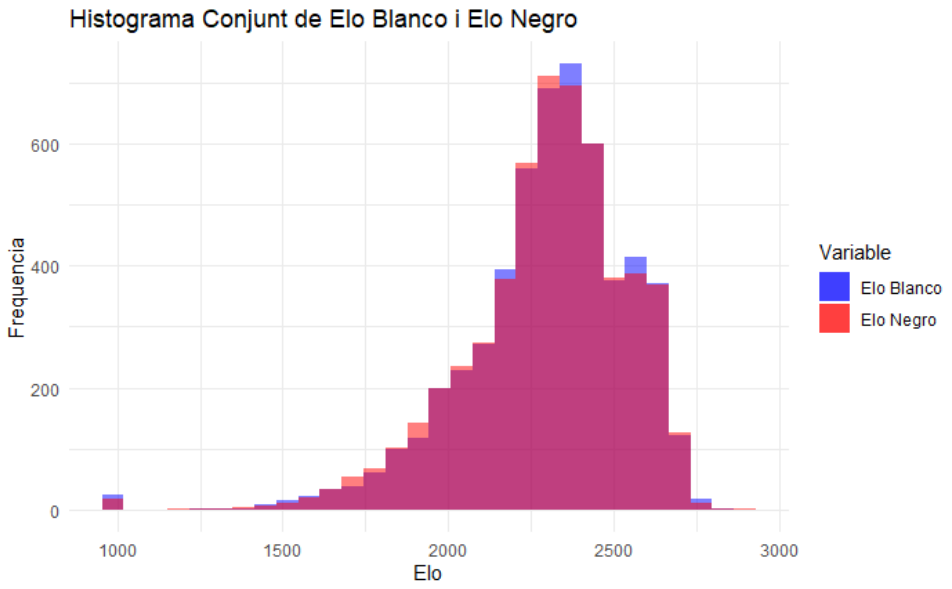


Figura 3.2. Histograma de Media_ELO

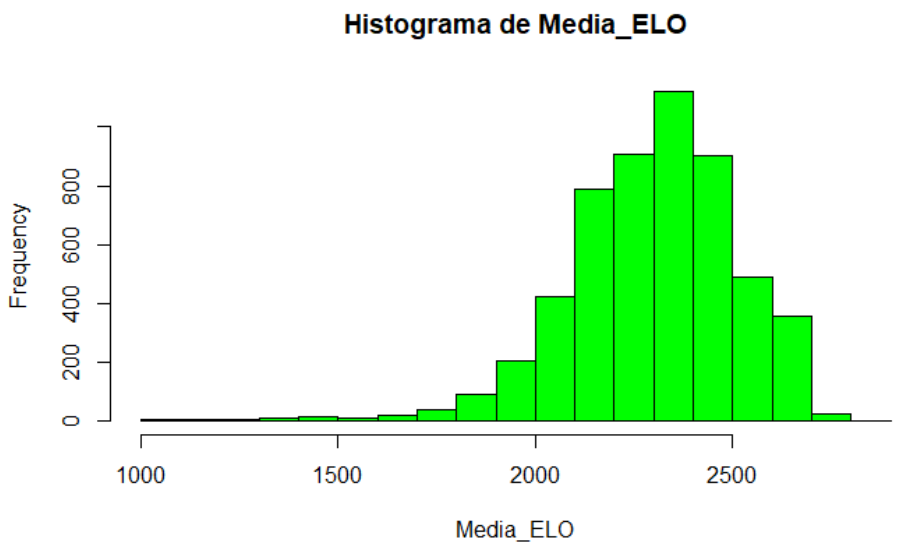
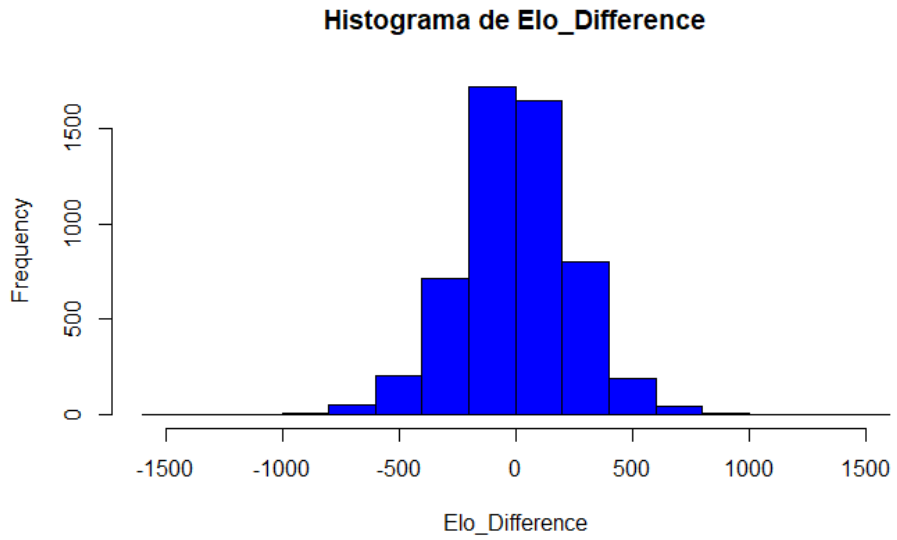


Figura 3.3. Histograma de Elo_Difference



Observant els gràfics anteriors i les taules, ens fixem que tant Elo Blanco com Elo Negro son molt similars durant tots els rangs d'Elo, això també ho veiem en els resultats de la variable Media_ELO, on es fa la mitja entre el jugador blanc i el negre, i s'observa que tant el gràfic com les taules són resultats molt similars al Elo blanc i Elo negre. Això al final ho veiem representat a l'Histograma de Elo_Difference on es veu que les dades estan al voltant del 0, demostrant que les partides són bastant igualades en base a l'Elo.

3.1.2.- GRUPO

Hi ha un total de cinc grups d'obertura en base a l'esquema d'obertura que es jugui a la partida, per això, fem una taula de freqüències per veure quin grup es el que més es juga.

Taula 3.2. Taula de freqüències de la variable GRUPO

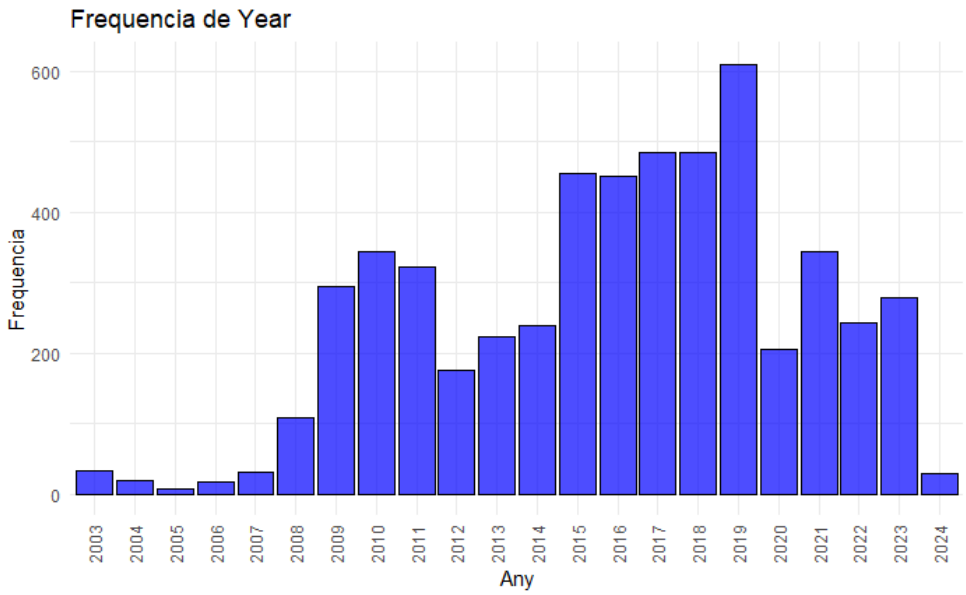
GRUPO	FREQÜÈNCIA	FREQ. RELATIVA
Semiabiertas	1750	0.32
De flanco	1083	0.20
Abiertas	896	0.17
Cerradas	844	0.16
Indias	828	0.15

Ens fixem que el grup d'obertures que més partides tenim són de les obertures semiobertes que representen el 32% de les obertures jugades. Són aquelles on es comença amb el Peó de rei fent la jugada 1. e4 i no es continua amb la jugada 1... e5. Destacar d'aquest tipus, la obertura Siciliana que contra la jugada 1. e4, respon amb la jugada 1... c5.

3.1.3.- YEAR

Aquesta variable fa referència a l'any el qual es va jugar la partida, ens serveix per veure si hi ha les mateixes partides cada any o hi ha anys en els quals es juguen més partides que altres.

Figura 3.4. Diagrama de Barres de Freqüència de Year



En afegit, hem fet un test X^2 per poder veure si hi ha diferencies significatives entre els diferents anys.

Taula 3.3. Test X^2 de la variable year

Mètric	Value
Test Statistic	2803.9
Degrees of Freedom	21
P-Value	< 2.2e-16

Amb el resultat del test X^2, veiem que si, que hi ha diferencia a la freqüència de nivells de year, on observem que destaquen des de els anys 2015 fins al 2019, que són els anys on tenim més partides jugades.

3.1.4.- RESULTADO

La variable principal del nostre treball, és la variable Resultado, per això, es fa una taula de freqüències per veure quin resultat és el que més hem obtingut.

Taula 3.4. Taula de freqüències de la variable Resultado

RESULTADO	FREQÜÈNCIA	FREQ. RELATIVA
0-1	1557	0.29
1-0	2159	0.40
1/2-1/2	1685	0.31

Observem que el resultat més repetit es el 1-0 que fa referència a la victòria de les blanques i representa el 40% del total de resultats obtinguts. Durant el treball s'analitza si aquest fet es dona per atzar o està relacionat amb les altres característiques de la partida.

3.2.- Anàlisi Bivariant

3.2.1.- Year i Grupo

El primer estudi bivariant que fem, tracta de trobar si hi ha una relació al llarg del temps de les obertures que es juguen, per això, fem un anàlisi i ho exposem amb dos diagrames de barres que ens demostrin si la freqüència d'obertures que es juguen cada any varia o es manté constant, ja que com bé hem vist al anàlisi Univariant, sabem que hi ha algun grup d'obertura que s'ha jugat més que d'altres i també trobem que hi ha anys on hi ha més partides jugades.

Figura 3.5. Diagrama de barres de freqüències de GRUPO per year

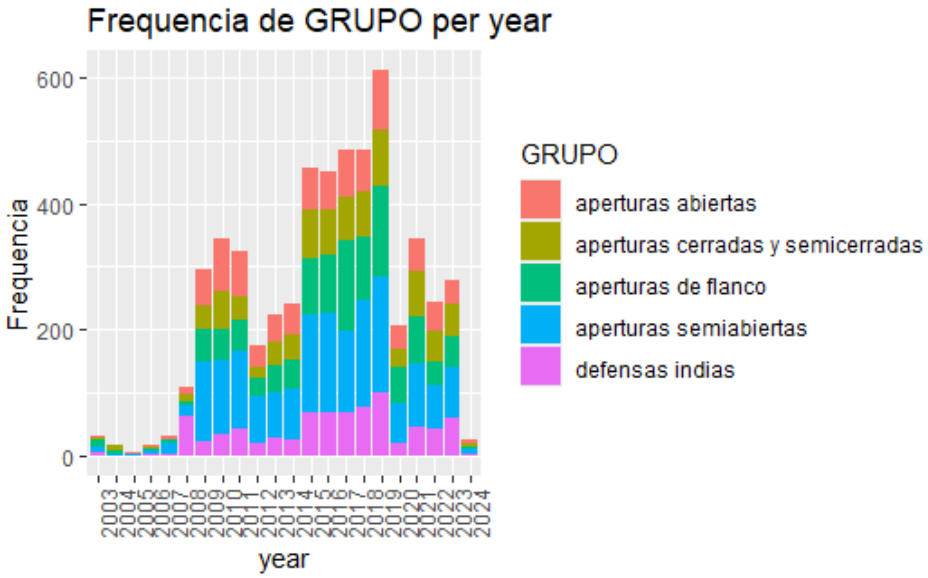
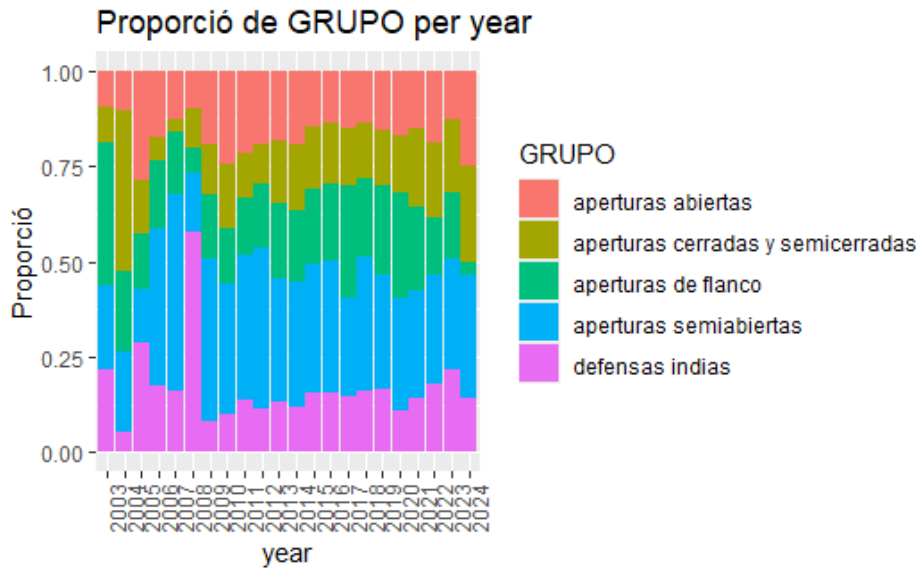


Figura 3.6. Diagrama de barres de la proporció de GRUPO per year



Quan fem aquest gràfics ens dona una visió clara si es mantenen el GRUPO d'obertures al llarg dels anys. Ens fixem que les obertures Semiobertes son les que més destaquen al llarg dels anys, sent l'estil d'obertures que més es juga, per tant, l'estil d'obertures semiobertes es el més popular entre els jugadors. A part, tot i que hi ha anys on altre tipus d'obertura es juga més, com podria ser el 2008 on trobem que la defensa Índia va ser bastant popular, no manté aquesta popularitat durant la resta d'anys.

3.2.2.- Elo_Difference i Resultado

El nostre objectiu principal consisteix en observar si la diferencia d'Elo dels jugadors, afecta a l'hora de predir una partida. Com hem pogut veure a l'anàlisi Univariant, trobem que la diferència d'Elo, està al voltant de 0, el que vol dir que els jugadors tenen un nivell similar, per això, fem una taula de freqüències per veure si es manté als tres resultats.

Taula 3.5. Taula de freqüències de Resultado per Elo_Difference

Resultado	count	mean_Elo_Difference	sd_Elo_Difference	min_Elo_Difference	max_Elo_Difference
0-1	1557	-190,286	203,8054	-1462	646
1-0	2159	154,1718	218,28	-981	1426
1/2-1/2	1685	-6,55549	171,4882	-955	658

Observant els resultats de la Taula 3.5, ens fixem que quan el resultat és victòria negra, la diferencia d'Elo sol ser negativa pel qual els jugadors negres son els que tenen més Elo, en canvi, quan el resultat és victòria blanca, ens fixem que el jugador de blanques es el que té un Elo superior. Finalment, quan el resultat és taules, observem, que la diferencia d'Elo és semblant a 0. Per tant, fem un comparador de mitjanes per demostrar-ho, fem una Taula ANOVA, el qual ens dona els següent resultats:

Taula 3.6. Taula ANOVA entre les variables Resultado i Elo_Difference

Variable	Df	Sum sq	Mean sq	F value	Pr(>F)
Resultado	2	107645780	53822890.16	1338.718	0
Residuals	5398	217025549	40204.81		

Seguidament, fem un test de Tukey HSD amb els següents resultats:

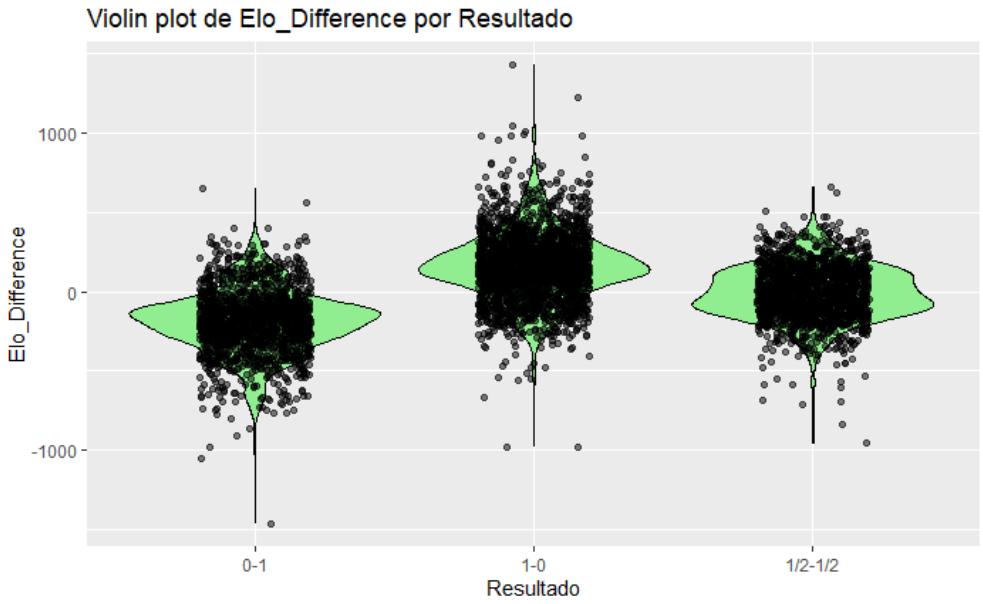
Taula 3.7. Test de Tukey HSD de les variables Resultado i Elo_Difference

Resultado	Diff	lwr	upr	P adj
1-0 i 0-1	344.4576	328.8287	360.0865	0
½-½ i 0-1	183.7303	167.2060	200.2546	0
½-½ i 1-0	-160.7273	-176.0074	-145.4472	0

Un cop hem fet l'ANOVA destaquem que el p-valor de l'estadístic F es 0. Aquest resultat vol dir que la diferencia entre els resultats és altament significativa. Per aquest motiu realitzem el Test de Tukey per veure si es produeix a tots els resultats i observem que els tres resultats tenen un p-valor de 0, el que ens fa reafirmar-nos de que les diferències son significatives i que per tant, hi ha diferències de resultat segons la diferencia d'Elo que hi hagi entre els dos jugadors.

Per poder veure-ho de forma visual, fem un Violin plot per veure com es produeixen aquestes diferències de grups en base a la diferencia d'Elo entre jugadors.

Figura 3.7. Gràfic Violin Plot de les variables Elo_Difference i Resultado



Un cop exposat el Violin Plot, es confirma que hi ha partides de victòria negra on es el jugador blanc el que té més Elo i hi ha victòries de les blanques on és el jugador de negres els que tenen més Elo, però encara així el gràfic respecta els resultats de la taula anterior.

3.2.3.- Media_Elo i Grupo

En l'apartat 3.1. hem comparat els diferents grups d'obertures per veure si durant els anys es respecta la proporció d'obertures de cada tipus que es juguen, però en aquest apartat volem analitzar si segons la mitja d'Elo dels jugadors, és més comú que juguin un tipus d'obertura o un altre. Per això, el primer que fem és separar les obertures en grups i fer una taula de freqüències per veure les dades dels diferents grups d'obertures.

Taula 3.8. Taula de freqüències de Media_ELO i GRUPO

GRUPO	COUNT	Mean_Media__ELO	Sd_Media_Elo	Min_Media_Elo	Max_Media_Elo
Semiabiertas	1750	2285	216	1000	2766
De flanco	1083	2311	189	1526	2717
Abiertas	896	2274	249	1000	2733
Cerradas	844	2305	225	1000	2808
Indias	828	2327	189	1503	2750

En base als resultats de la Taula 3.8, observem que hi ha algun tipus d'obertura que la mitjana dels jugadors que la juguen és més alta que la mitjana dels jugadors que juguen un altra estil d'obertura com pot ser el cas de les obertures Obertes amb les defenses índies. Per tant, igual que a l'apartat anterior, fem una taula ANOVA amb els següents resultats:

Taula 3.9. Taula ANOVA entre les variables Media_ELO i GRUPO

Variable	Df	Sum sq	Mean sq	F value	Pr(>F)
GRUPO	4	1739233	434808	9,466	1.27e-07 ***
Residuals	5396	247847922	45932		

Observem a la Taula 3.9. que el p-valor associat al test és molt petit, el que entenem que hi ha una diferència entre els diferents grups d'obertures. Per aquest motiu, fem un test de Tukey per analitzar les diferències entre grups i després fem un segon test de Tukey amb *agricolae* per poder analitzar si alguns dels diferents grups d'obertures s'assemblen i com els podem classificar.

Taula 3.10. Test de Tukey HSD amb agricolae de les variables Media_ELO i GRUPO

agricolae	Media_ELO	groups
defensas indias	2.327.033	a
aperturas de flanco	2.311.235	a
aperturas cerradas y semicerradas	2.305.027	ab
aperturas semiabiertas	2.285.407	bc
aperturas abiertas	2.273.623	c

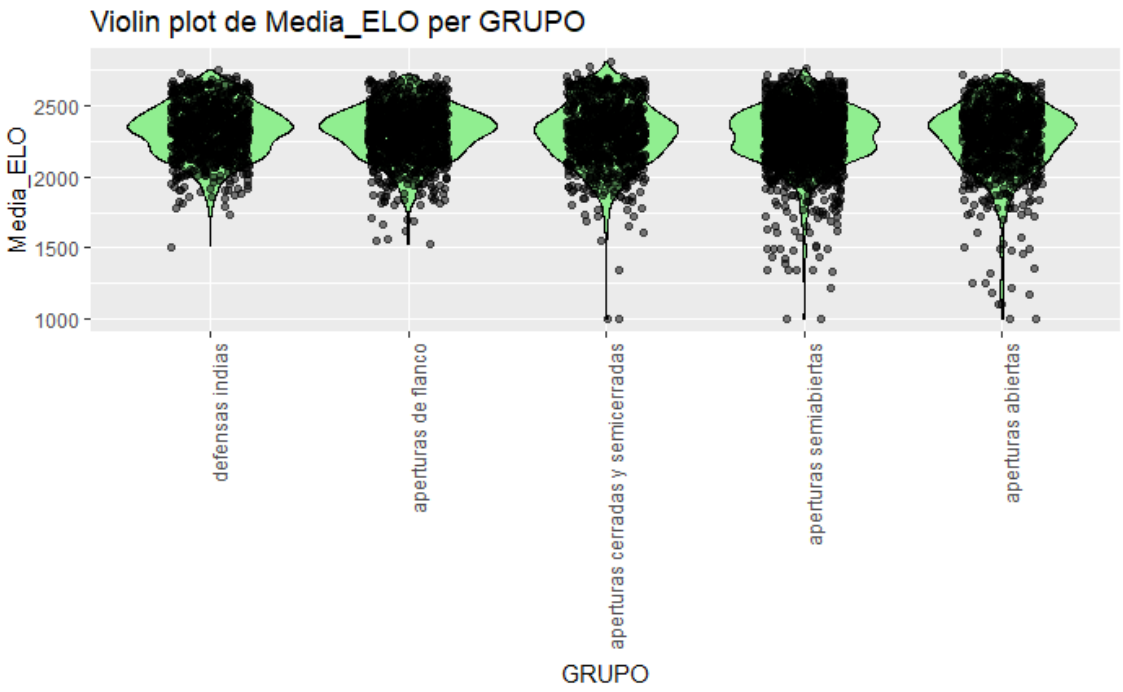
En base als resultats dels test, obtenim que la defensa índia i la obertura de flanc, no tenen gaire diferència entre elles, de la mateixa manera que la obertura tancada i semitancada no hi ha diferències significatives amb les defenses índies i la obertura de flanc.

Amb un cas similar tenim les obertures semiobertes, que trobem que no hi ha diferencies significatives amb les obertures tancades i semitancades.

Finalment, tenim les obertures obertes que és significativament diferent de les defenses índies, les obertures de flanc i les obertures tancades i semitancades, però en canvi amb les obertures semiobertes si que trobem que no és significativament diferent.

Amb aquestes premisses fem un Violin Plot per comprovar aquests resultats i veure-ho de manera gràfica.

Figura 3.8. Gràfic Violin Plot de les variables Media_ELO i GRUPO



En base als resultats veiem que la majoria de partides tenen una mitjana semblant on ronden els 2300 d'Elo. Però el més destacat és que les partides on els jugadors tenen menys Elo juguen obertures semiobertes i obertes, mentre que les altres tres obertures es juguen sobretot quan els jugadors tenen ja un Elo més alt superior als 1500 de rating.

3.3.- MLG

Per continuar l'estudi, decidim fer un model lineal de regressió lineal multinomial, amb l'objectiu d'analitzar la variable Resultado i estudiar si les diferents variables que hem anat analitzant durant els anàlisis anteriors, poden ajudar a predir el resultat de la partida.

Fem dos tipus de MLG, el primer d'ells es només comparant amb la variable Elo_Difference i així resoldre la qüestió que ens hem plantejat a l'objectiu de concloure si podem esbrinar el resultat de la partida només amb la diferencia d'Elo.

3.3.1.- Model1:Model tenint en compte Elo_difference

El primer que fem és passar els resultats de la partida de 1-0, ½-½ i 0-1 a 1, 0.5 i 0, sent 1 la victòria del blanc, 0.5 les taules i 0 la victòria del negre. Un cop tenim això, ajustem el model amb la següent formulació:

$$model <- multinom(Resultado \sim Elo\_Difference)$$

I llavors, farem un summary del model per veure els seus resultats.

Taula 3.11. Taula de coeficients, error, estàndard i valors P respecte al model

Coeficients	**0.5**	**0**
Intercept	0,510647	0,565838
Elo_Difference	0,009334	0,00507
Error Estándar	**0.5**	**0**
Intercept	0,046153	0,044739
Elo_Difference	0,000206	0,000224
Valores p	**0.5**	**0**
Intercept	0	0
Elo_Difference	0	0

Observant el p-valor del model, ens fa entendre que els coeficients són significatius, pel que, ens suggereix que ambdós resultats tenen un impacte significatiu a la variable dependent. Un cop tenim els coeficients, volem crear una matriu de confusió a partir del paquet *caret*, per poder avaluar el rendiment del model i veure el nombre de prediccions correctes.

Comencem creant una matriu de confusió dels diferents resultats possibles i un cop obtenim la matriu de confusió, calculem la precisió del model a partir del seu *accuracy* i finalment calculem el coeficient de Kappa.

Taula 3.12 Matriu de Confusió del MLG de Elo_Difference

Reference	Prediction_1	Prediction_1/2	Prediction_0
1	1053	215	289
1/2	235	1661	263
0	450	759	476

Taula 3.13.- Càlculs de la precisió del model i altres mètriques

Overall Statistics	Value
Accuracy	0.5906
95% CI	(0.5774, 0.6038)
No Information Rate	0.3997
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.3729
Mcnemar's Test P-Value	< 2.2e-16

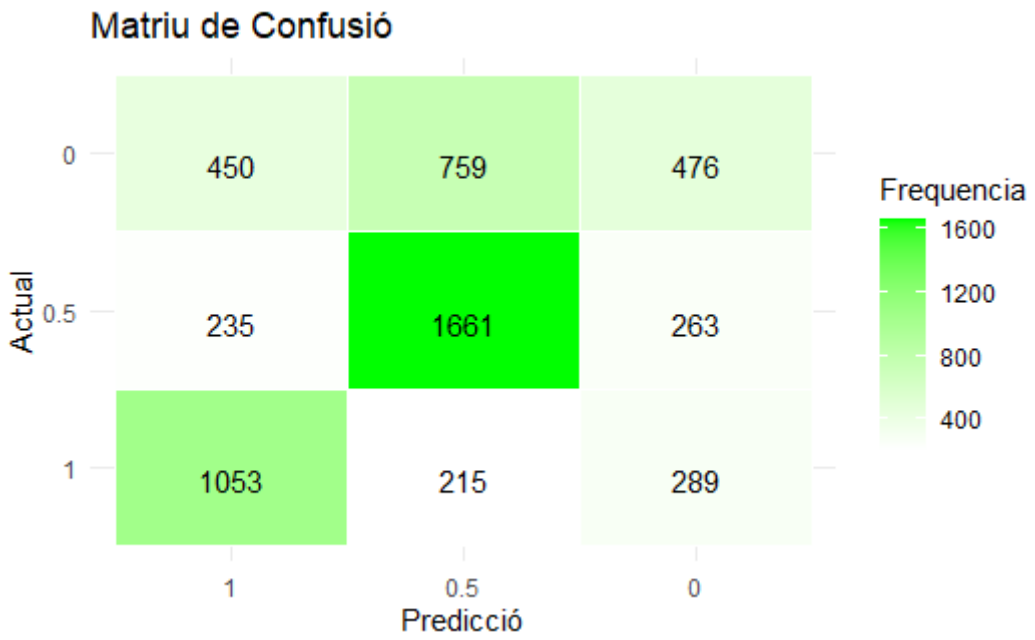
A la taula 3.13 observem que el model té una precisió del 59.06%, veiem que surt un rendiment moderat, el qual prediu correctament més de la meitat de les partides però, tenim marge d'error.

Tot i així, segons la Taula 3.13 on es detalla el NIR, que seria la taxa de Precisió si es predis sempre la classe més freqüent a les dades, i que en aquest cas, seria la victòria blanca, observem que es del 39.97% que vol dir que el model està fent una bona feina respecte a una predicció ingènua.

Finalment, a la Taula 3.13 es detalla el coeficient de Kappa, amb un valor de 0.3729, que significa que el model té l'habilitat de classificar correctament els resultat final més enllà de l'atzar.

Per continuar l'anàlisi, creem un gràfic amb la Matriu de Confusió per veure de forma visual la predicció del model envers els resultats reals.

Figura 3.9 Matriu de confusió del MLG de la variable Elo_Difference



3.3.2.- Model2: Model tenint en compte Elo_Difference, Media_Elo i GRUPOS

Creem un segon model lineal de regressió multinomial, però afegint les variables Media_Elo i GRUPOS per veure si podem predir el model d'una forma més correcte.

Per això, formulem el model de la següent manera:

$$model <- multinom(Resultado \sim GRUPO + Elo\_Difference + Media\_ELO)$$

En primer lloc observem els resultats del *summary* que detallem en la següent taula:

Taula 3.14 Taula de coeficients, error estàndard i p-valors

Term	Coefficient_0.5	Coefficient_0	Std_Error_0.5	Std_Error_0	P_Value_0.5	P_Value_0
(Intercept)	-0,3800444	-0,5060892	0,00044227	0,00044227	1,1769E-06	1,2589E-06
GRUPOaperturas cerradas y semicerradas	-0,3482624	-0,2048665	0,00046496	0,00046496	0,05390652	9,1732E-06
GRUPOaperturas de flanco	-0,219756	0,00981458	0,0005273	0,0005273	0	0
GRUPOaperturas semiabiertas	-0,141826	0,00476946	0,00048467	0,00048467	0	0
GRUPOdefensas indias	-0,2912412	-0,00573715	0,00047267	0,00047267	0	0
Elo_Difference	-0,00605145	0,00981458	0,00044227	0,00046496	0	0
Media_ELO	-0,1392229	0,00476946	0,00046496	0,0005273	0	0

La majoria de les variables són estadísticament significatives, el que ens fa pensar que tenen un impacte important a la variable dependent. En base a aquest resultat, creem una nova matriu de confusió dels diferents resultats possibles i a continuació, calculem la precisió del model a partir del seu *accuracy* i finalment calculem el coeficient de Kappa

Taula 3.15. Matriu de Confusió del MLG de Elo_Difference

Reference	Prediction_1	Prediction_1/2	Prediction_0
1	1026	222	309
½	227	1601	331
0	394	701	590

Taula 3.16 Càlculs de la precisió del model i altres mètriques

Overall Statistics	Value
Accuracy	0.5956
95% CI	(0.5824, 0.6088)
No Information Rate	0.3997
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.3819
Mcnemar's Test P-Value	< 2.2e-16

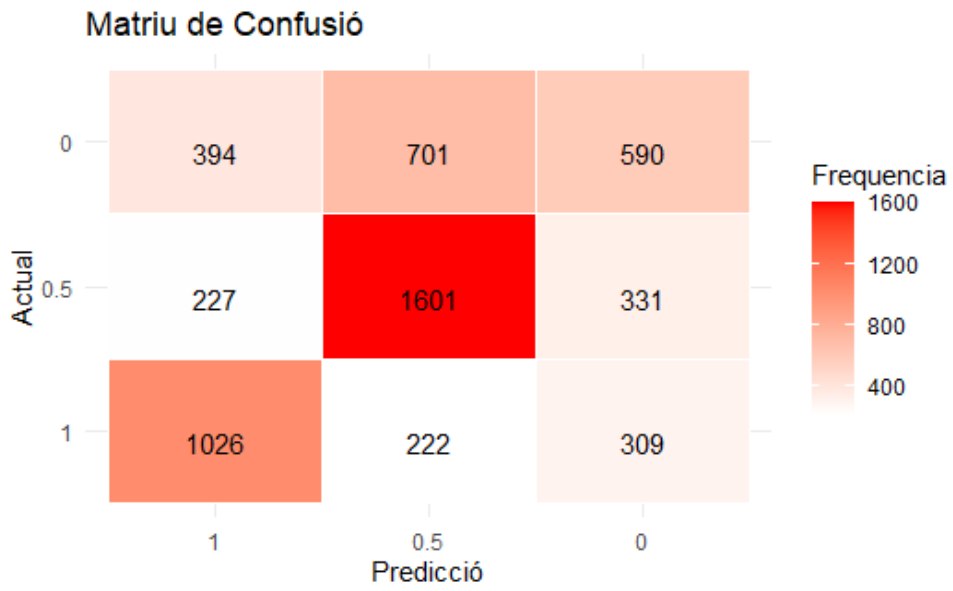
Observem que el model té una precisió del 59.56%, que ens fa pensar que té un rendiment moderat, igual que al model anterior i encara ens deixa marge d'error.

Segons la Taula 3.16 i ens fixem al resultat del NIR, que seria la taxa de Precisió si es predis sempre la classe més freqüent a les dades, que en aquest cas, seria la victòria blanca, observem que es del 39.97% igual que al model anterior, donat que tenim les mateixes dades.

Finalment en la mateixa taula obtenim el coeficient de Kappa per un valor de 0.3819, i en conclusió es una mica millor que el model anterior, això vol dir que existeix una concordança una mica millor entre les prediccions i les dades reals.

Amb tots els resultats obtinguts, creem un gràfic amb la Matriu de Confusió per veure de forma visual la predicció del model envers als resultats reals.

Figura 3.10 Matriu de confusió del Model MLG complet



4.- DISCUSSIONS I CONCLUSIONS

4.1.- Discussions

Durant els últims anys, els escacs s'han anat perfeccionant per intentar assolir el nivell màxim de joc i millorar els resultats. Actualment els millors jugadors d'escacs del món no són humans, en realitat són ordinadors amb una capacitat de càlcul molt superior a la dels humans que han assolit un nivell gairebé de perfecció escaquística, on els errors que fan a les partides es van corregint i això permet que el seu Elo sigui molt superior a la de qualsevol humà. Aquesta circumstància fa que la probabilitat de perdre una partida sigui molt baixa (Arthur Berg, 2020).

Amb totes les eines que es disposa actualment, s'ha aconseguit que un cop acabada una partida d'escacs es pugui calcular el nivell de perfecció de la mateixa, amb un rati de 100 equivalent a jugar tots els moviment la millor jugada que faria la màquina. Per altra banda, si partim de que la fórmula de l'Elo, que et diu el resultat esperat de la partida només tenint en compte l'Elo dels dos jugadors, vaig entendre que es podien relacionar aquests factors per predir el resultat d'una partida.

Si partim de l'Elo dels jugadors, existeix un resultat esperat, però en quants casos aquest resultat esperat es el que acaba succeint?. Per respondre aquesta pregunta vaig decidir seleccionar diferents factors que son propis dels jugadors o que es fan servir durant la partida, com són la diferencia d'Elo, l'obertura jugada o la mitjana d'Elo dels dos jugadors, i si tots ells afecten al resultat de la partida o poden fer variar el resultat.

Inicialment, vaig recopilar un total de 24.827 dades a analitzar, però el fet de que moltes d'elles estiguessin repetides, d'altres no s'identifiqués l'obertura que es va jugar o que només surtís l'Elo i el resultat, va provocar que es seleccionessin un total de 5.401 partides a analitzar, que representen un 21.75% de les partides inicials.

Respecte dels resultats de l'anàlisi, hem calculat que el nombre de partides guanyades pel jugador blanc, es superior al nombre de taules i al nombre de victòries del jugador amb peces negres. Això es degut a que la mitjana d'Elo de les partides es de 2298 punts d'Elo, el que vol dir que les partides estudiades son partides d'alt nivell. Els jugadors que s'enfronten son jugadors titulats, per tant, el seu nivell de precisió a les partides es més alt. Tot això, provoca que el jugador blanc comenci amb una petita avantatja al ser el que fa el primer moviment, i, com el marge d'error dels jugadors titulats és més baix, ja que tenen un nivell de coneixement i de càlcul superior al dels jugadors amateurs, aconsegueixen guanyar la partida en un percentatge més elevat el fet de jugar amb blanques el primer moviment.

Com a limitació a l'anàlisi cal tenir present que hi ha altres variables que poden influir al resultat, trobem que per futurs estudis s'hauria de tenir en compte noves variables per poder calcular la probabilitat de predir el resultat, per exemple, es pot analitzar variables com l'estat físic dels jugadors, el tipus de partida (ritme ràpid o clàssic), o fins i tot, les vegades que ha jugat un jugador un tipus d'obertura en anteriors partides per l'experiència que això suposa, tot això ajudaria a crear un model més complet per intentar predir el resultat final de la partida.

4.2.- Conclusions

Els escacs, es basen en la capacitat mental de cada jugador i en el seu nivell de joc, el que fa variar les decisions provocant que un mateix jugador jugui de forma diferent en cada partida. Per aquest motiu i segons el model creat, ens ha permès deduir que en més de la meitat de les partides, el fet de que els dos jugadors tinguin el mateix Elo o que un d'ells tingui més Elo que un altre, pot influir al resultat, ja que el jugador de més nivell normalment juga millor i fa menys errors que el de menys nivell. Però hi ha casos on no es dona aquest resultat i, per tant, el jugador de menys nivell, acaba aconseguint un resultat superior a l'esperat o en cas del jugador de més Elo, inferior a l'esperat. En conclusió, no es pot predir amb total seguretat el resultat d'una partida només amb l'Elo dels jugadors.

Per altra banda, també hem pogut analitzar les diferents obertures, entre les que unes són més comuns que altres i, per tant, les trobem en més partides de la base analitzada. El resultat que obtenim és que jugadors de més nivell (amb més Elo), utilitzen més en les seves partides un tipus d'obertura que jugadors de menys nivell que prefereixen jugar un tipus d'obertura diferent.

Només afegir que estic satisfet de l'estudi realitzat, en primer lloc, per que penso que els escacs tenen molt de recorregut en l'anàlisi estadístic, en primer lloc per la seva dispersió de variables i resultats i en segon lloc per que les partides queden enregistrades (cada cop es fan més partides en retransmissió amb registre de totes les jugades) i això ens permet tenir moltes dades per analitzar. En afegit, el fet que cada cop hi hagi més programes d'anàlisi i participin de forma activa en aquest esport, sent capaços de calcular totes les millors jugades, fa necessari analitzar les dades de forma continuada. Tot i això, als escacs també tenen un factor humà que provoca que totes les partides d'escacs tinguin un resultat incert però queda demostrat que encara que sigui més probable que guanyin el jugadors amb més Elo (nivell de joc), no tenen garantida la victòria.

5.- BIBLIOGRAFIA

1.- Bathelemy, Marc. (2023, April). RESEARCH GATE. *Statistical analysis of chess games: space control and tipping points.*

https://www.researchgate.net/publication/370226973_Statistical_analysis_of_chess_games_space_control_and_tipping_points

2.- Baumer Benjamin, Gregory J. Matthews, Quang Nguyen. (2023, Jan 10). *Big Ideas in Sports Analytics and Statistical Tools for their Investigation*. https://arxiv.org/abs/2301.04001

3.- Berg, Arthur (2020). CHANCE. *Statistical Analysis of the Elo Rating System in Chess*. https://chance.amstat.org/2020/09/chess/

4.- Blissett, Richard *(2017, Noviembre).* Logistic Regression in R.

https://rpubs.com/rslbliss/r_logistic_ws

5.- Chess-Results.com. https://chess-results.com/TurnierDB.aspx?lan=2

6.- Federació Catalana d'Escacs. https://escacs.cat/index.php/component/fce?op=2

7.- Fredes, Daniel (2020, Diciembre). RPUBS. *Ejemplo práctico de análisis bivariado en R*. https://rpubs.com/fredesdanyel/698502

8.- Herrera Reyes, José Antonio. (2018, Julio). CHESS.COM. *¿Qué es el ELO?* https://www.chess.com/es/blog/stewiiegriffin/que-es-el-elo

9.- International Chess Federation. https://www.fide.com/

10.- Llinás Solano, Humberto. (2022, Julio). RPUBS. Modelo lineal generalizado. https://rpubs.com/hllinas/R_GLM_Teoria

11.- Rosales Pulido, Hector Apolo. Universitat Politècnica de Catalunya Barcelonatech Facultat D'Informatica de Barcelona. (2016, October). *Predicting the Outcome of a Chess Game by Statistical and Machine Learning techniques.* https://upcommons.upc.edu/bitstream/handle/2117/106389/119749.pdf?sequence=1&isAllowed=y

12.- Saavedra Rodríguez, Mario Gregorio (2020, Julio). RPUBS. Análisis descriptivo univariado. https://rpubs.com/mgsaavedraro/EDU

13.- Stapczynski, Colin (2023, Mar 29). CHESS.COM. *History of Chess | From Early Stages to Magnus*.

https://www.chess.com/article/view/history-of-chess#:~:text=Chess%2C%20as%20we%20know%20it,Spanish%20priest%20named%20Ruy%20Lopez

14.- STHDA. Statistical tools for high-throughput data análisis. *Ggplot2 violin plot: Quick start guide – R software and data visualization.* http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization

15.- The House of Stauton. *History of chess*. https://www.houseofstaunton.com/history-of-chess

16.- Whitehead Rohan (2024, Mar 21). INSTITUT OF ANALYTICS. Chess Engines: The Integration of Chess and Data Science Techniques.

https://ioaglobal.org/blog/integration-of-chess-and-data-science-techniques/#

17.- Wikipedia. https://en.wikipedia.org/wiki/Chess

18.- Zuñiga Guamán P. RPUBS. *Estadística descriptive Bivariante.* https://rpubs.com/PabloZ/1165808

6.- ANNEX

6.1 Codi R

Els codis i les bases de dades de R que s'han utilitzat per realitzar el treball, es poden trobar al següent enllaç https://github.com/pautoquero/TFG

6.2 Exemple de partida d'escacs descarregades de Chess-Results

Les dades de la Taula 2.1 que s'obtenen de cada partida descarregada són les següents:

Martinez Fernandez,Adrian (2317) - Toquero Gracia,Pau (2080) [B80] European Youth Chess Championships 2019 Bratislava, Incheba Expo (4.30), 05.08.2019
1.e4 c5 2.Cf3 d6 3.d4 cxd4 4.Cxd4 Cf6 5.Cc3 a6 6.Ae3 e6 7.f3 b5 8.Dd2 Cbd7 9.0-0-0 Ab7 10.Rb1 Ae7 11.g4 Cb6 12.Ad3 Cfd7 13.g5 Ce5 14.f4 Cec4 15.Df2 Cxe3 16.Dxe3 b4 17.Cce2 d5 18.e5 g6 19.h4 Dc7 20.h5 0-0-0 21.Dh3 Thg8 22.hxg6 hxg6 23.Cxe6 fxe6 24.Dxe6+ Rb8 25.Th7 Tde8 26.Txe7 Dxe7 27.Dxb6 Tc8 28.Axa6 Tc6 29.Dxb7+ Dxb7 30.Axb7 Rxb7 31.Txd5 Tc4 32.Rc1 Tgc8 33.Td7+ T8c7 34.Txc7+ Rxc7 35.Rd2 Tc5 36.Cc1 Ta5 37.Cd3 Txa2 38.b3 Ta1 39.Cxb4 Th1 40.c4 Tb1 41.Rc2 Th1 42.Cd3 Rc6 43.Rc3 Th8 44.b4 Td8 45.Cc5 Tf8 46.Ce6 Te8 47.Cd4+ Rd7 48.Rd3 Tf8 49.Re4 Tb8 50.b5 Tc8 51.Rd5 Tf8 52.Ce6
1-0

A la primera filera, es detallen les següents variables:

Blancas: Martinez Fernandez,Adrian

Elo Blanco: (2317)

Negras: Toquero Gracia, Pau

Elo Negro: 2080

ECO: [B80]

Torneo: European Youth Chess Championships 2019 Bratislava, Incheba Expo (4.30)

Fecha: 05.08.2019

A la segona filera, es detallen les jugades de la partida amb la notació utilitzada als escacs

Finalment, a la ultima filera, es detalla el resultat de la partida