Universitat Autònoma de Barcelona

Facultat de ciències

Predicció en un partit de tenis utilitzant l'expectativa pitagòrica i una simulació

Treball Final de Grau

Autor:
Jan Moreno - 1531382

Tutor:
Toni Lozano

Juny 2022

Agraïments

Voldria transmetre el meu sincer agraïment a en Toni Lozano, per la seva ajuda durant tot aquest procés i per deixar-me desenvolupar aquest treball amb total llibertat cap a la direcció on jo volia. També a la meva família i amics per ajudar-me en qualsevol moment amb qualsevol aspecte d'aquest treball.

Resum

En aquest treball adaptarem l'expectativa pitagòrica, un mètode d'anàlisis esportiu ideat per Bill James pensat pel beisbol primerament, al tenis. Començarem trobant un exponent que s'adapti al joc utilitzant dos mètodes diferents, i posteriorment testejarem la fórmula en diversos partits. A més, també provarem predir el guanyador simulant el partit, així posarem en comparativa els dos mètodes.
Per desenvolupar i utilitzar els dos models utilitzarem una base de dades gratuïta.

Summary

In this project we will adapt the Pythagorean expectation, a method of sports analysis created by Bill James firstly designed for baseball, to tennis. We will start by finding an exponent that fits the game using two different methods, and then we will test the formula in several games. In addition, we will also try to predict the winner by simulating the match, so we will compare the two methods.
We will use a free database to develop and use both models.

Índex

Introducció

A finals de la dècada dels 70, Bill James, en aquell moment un aficionat al beisbol, va escriure un llibre d'estadística aplicada al beisbol on presentava unes idees i uns nous valors estadístics no vistos fins aquell moment, entre les quals destacava l'expectativa pitagòrica. 20 anys més tard, Billy Bean (General Manager dels Oakland Athletics), va decidir aplicar les idees de Bill James en la seva política esportiva degut al baix pressupost de l'equip. Aquesta estratègia va resultar molt reeixida i a partir d'aquell any molts equips van començar a utilitzar l'estadística per analitzar jugadors, anys més tard al bàsquet i a l'hoquei sobre gel.

A l'hora d'escollir un tema sabia que volia que fos relacionat amb l'anàlisi estadístic i l'esport. Llavors necessitava un esport amb les següents condicions:

- Suficient popular per a trobar una base de dades completa i gratuïta.
- Un esport de caràcter estadístic i que no s'hagués estudiat molt l'aplicació d'aquesta expectativa a l'esport.

El tenis és un esport que compleix aquests requisits perquè és suficient popular per tenir una base de dades gratuïta molt completa i alhora no hi ha molta recerca en aquest àmbit.

Amb la temàtica general del treball ja definida i amb una base dades que cobreix totes les possibles necessitats pel treball, falta per definir cap a on enfocar aquesta anàlisi estadística.
Entre totes les idees que va proposar Bill James, la més senzilla i extrapolada a altres esports és l'expectativa pitagòrica. Una funció molt simple per descriure el percentatge de victòries d'un equip en una temporada i també per atorgar probabilitats en un enfrontament entre dos equips. La fórmula és la següent:

$$WP = \frac{Runs\ Scored^2}{Runs\ Scored^2 + Runs\ Allowed^2}$$

Doncs en aquest treball adaptarem aquesta fórmula al tenis com ja s'ha fet en altres esports, buscant un exponent òptim i llavors observar quin percentatge d'encert obté.
A més, també utilitzarem una simulació considerant moltes més variables per comparar com rendeixen aquests dos mètodes en afrontar una predicció.

El procediment d'aquest treball és molt similar a qualsevol altre projecte d'anàlisis de dades. Primerament, explorarem la base de dades sobre la qual treballem. Seguidament, definirem i realitzem una sèrie de testos. Finalment, en traurem conclusions i durem a terme alguna prova més en funció dels resultats assolits.

Després de realitzar diversos testos, en línies generals observarem com l'expectativa pitagòrica fa una millor predicció envers la simulació, que necessitava més dades, més temps i obtenia un pitjor percentatge d'encert. En general, obtenim un 60% màxim d'encert quan tenim en compte tots els partits predits, però, també veurem que obtenim un 80% d'encert en aquells partits els quals estarem segurs de la nostra predicció.
Evidentment no són uns resultats excel·lents, però sí els podem considerar bons si notem que hem utilitzat molt poques dades i que no utilitzem alts coneixements del tenis, doncs sí que els podríem considerar bons.

Base de Dades

Origen

Per la realització d'aquest treball necessitem una base de dades suficient gran, coherent i amb les dades necessàries per realitzar tant l'expectativa categòrica com la simulació.
A més, necessitem que sigui gratuïta, ja que no comptem amb cap recurs econòmic. Gràcies al fet que la **ATP** (*Assossicació de professionals del tenis*) publica les dades de cada partit, només necessitem trobar una publicació per internet amb aquestes dades recopilades.

Finalment considerem el següent enllaç (*DataHub*) on hi ha publicats diversos fitxers amb les dades necessàries.

Dades

Entre els diversos fitxers disponibles seleccionem els següents:

- Marcador partits 1991-2016.
- Estadístiques partits 1991-2016.
- Marcador partits 2017.
- Estadístiques partits 2017.

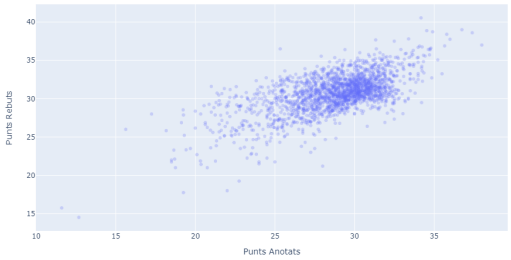
Amb aquests 4 fitxers tenim totes les dades necessàries: Noms dels tenistes, quantitat de punts, jocs i sets guanyats per cada jugador, es podia calcular el percentatge d'encert al primer i segon sac, entre altres.

Addicionalment ja tenim feta la divisió de la base de dades feta: Les dades del 1991-2016 seran utilitzades per optimitzar l'expectativa pitagòrica i per llavors predir els resultats dels partits del 2017. Del fitxer del 2017 només necessitem els noms dels jugadors i el guanyador del partit.

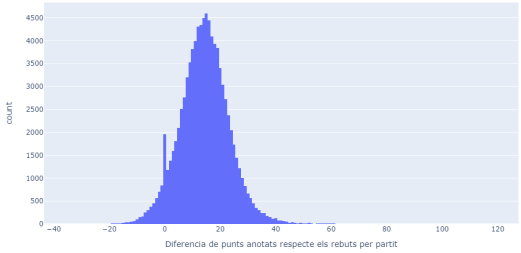
En aquesta ocasió no serveix realitzar un *cross-validation*, ja que estem intentant predir el guanyador d'un partit futur del qual no es tenen dades. Per això és important mantenir la linealitat temporal en tot moment.

Anàlisis descriptives de les dades

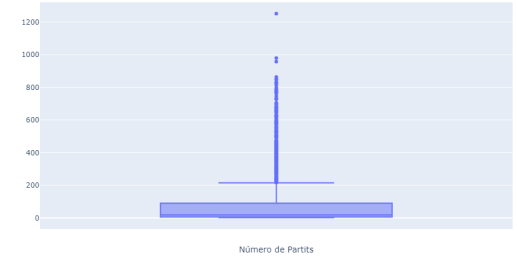
Per mostrar en més detall de com les dades de les quals estem parlant, el fitxer referent a les dades entre 1991 i 2016 hi figuren 1986 tenistes diferents amb 91806 partits entre 89 tornejos diferents.
A continuació podem veure com es relacionen els punts anotats per un jugador i els que rep.



Observem que les diferències són molt petites i que estan molt correlacionats entre ells. Això ja ens dona una pista que les prediccions no seran molt acurades perquè el guanyador ve determinat per una petita diferència de punts, la qual és difícil de predir.
També és important mencionar que és normal que els jugadors rebin més punts dels que anoten, si considerem que normalment el guanyador fa més punts que el rival (Com veiem al següent gràfic) i que la majoria dels jugadors queden eliminats en les primeres rondes, llavors és lògic que hi hagi més casos on els jugadors rebin més punts que els anotats.



A continuació hi ha representat la quantitat de partits jugats per cada jugador i com es distribueix el percentatge de victòries entre els jugadors.



Observem que la majoria dels tenistes no arriben als 100 partits, fet prou rellevant, ja que indica que només tindrem pocs jugadors per realitzar un estudi històric de les seves dades.

Tractament de les dades

Per treballar amb dos fitxers diferents utilitzant dades provinents dels dos hem de llegir les dades i ajuntar-les en cada execució. Tot aquest procés és costós computacionalment i requereix molt temps.

Per estalviar tot aquest procés, ajuntem els dos fitxers en un, i llavors el guardem en format **csv**. Aquesta fusió no és tan senzilla com posar un fitxer a l'esquerra de l'altre, ja que l'ordre dels partits no coincideix en els dos fitxers. Per solucionar aquesta petita problemàtica, ajuntem els dos fitxers usant funcions de la llibreria **Pandas** i les variables que es repeteixen en els dos fitxers per identificar els partits.

Una vegada tenim els dos fitxers ajuntats en un, els guardem en format **csv** i a partir d'aquest moment només utilitzarem aquest únic nou fitxer.

Pel fitxer referint-se els partits del 2017, no cal que repetim aquest procés perquè només utilitzarem els noms dels tenistes i guanyadors per realitzar i testejar les prediccions.

Expectativa Pitagòrica

L'expectativa pitagòrica és una fórmula creada per Bill James primerament pensada pel Baseball que descriu el percentatge de partits guanyats d'una temporada segons les carreres (*puntuació al beisbol*) rebudes i fetes al llarg d'aquesta. Obté aquest nom a causa de la versemblança que té amb el teorema de Pitàgores.

Fòrmula

En un inici Bill James va publicar la fórmula de la següent manera:

$$WP = \frac{Runs\ Scored^2}{Runs\ Scored^2 + Runs\ Allowed^2}$$

Amb aquesta fórmula Bill James era capaç d'explicar el percentatge de partits guanyats de qualsevol equip amb un marge d'error petit.
A partir d'aquesta base, diversos estadístics van començar a buscar un exponent ideal per reduir al mínim l'error. Després de diversos estudis van veure que el millor exponent és 1.83.

Més tard, altres estadístics van començar a utilitzar aquesta mateixa fórmula per altres esports: bàsquet, futbol americà, hoquei sobre gel... amb un exponent diferent, ja que el sistema de puntuació és diferent per cada esport, per exemple en el bàsquet es fan 80 o més punts, en canvi, al beisbol poques vegades s'arriba a les 5 carreres, llavors s'ha de buscar un exponent per cada esport.

Al ser una fórmula senzilla, la provarem al tenis de la següent manera:

$$WP = \frac{PS^\gamma}{PS^\gamma + PA^\gamma}$$

On

- PS: Punts per Set anotats.
- PA: Punts per Set Rebuts.
- WP: Percentatge de victòries.
- γ: Exponent.

Utilitzem els punts anotats per set en lloc del total de punts perquè en alguns tornejos es juguen 5 sets i en altres 3, llavors aquesta diferència pot afectar en els resultats.

Ara hem de buscar el millor exponent.

Cerca de l'exponent

Per trobar l'exponent podem utilitzar dos mètodes diferents, l'error dels mínims quadrats, i la regressió lineal.

Mètodes

Error dels Mínims Quadrats

Per utilitzar aquest mètode necessitem una fórmula amb una variable desconeguda la qual representi l'error de la predicció al quadrat per llavors intentar-la igualar a 0, o aproximar-la al màxim.
Finalment queda de la següent manera:

$$(WP - \frac{PS^\gamma}{PS^\gamma + PA^\gamma})^2 = 0$$

Seguidament realitzem un descens del gradient per igualar la fórmula anterior a 0, o aconseguir el valor que més s'hi aproxima.

Com que cada jugador té un exponent diferent per la seva manera de jugar, calculem un exponent per cada jugador i utilitzem la mitjana aritmètica per calcular la γ general.

Regressió Lineal

L'altre mètode és la regressió lineal, per això necessitem expressar la fórmula anterior d'una manera que una variable sigui explicada per l'exponent i una altra variable explicativa.
Per aconseguir una expressió com la descrita anteriorment hem de reformular l'expressió inicial:

$$\begin{aligned}
WP &= \frac{PS^\gamma}{PS^\gamma + PA^\gamma} \\
WP \cdot \frac{PS^\gamma + PA^\gamma}{PA^\gamma} &= \frac{PS^\gamma}{PS^\gamma + PA^\gamma} \cdot \cdot \frac{PS^\gamma + PA^\gamma}{PA^\gamma} \\
On : (1 - WP)^{-1} &= \frac{PS^\gamma + PA^\gamma}{PA^\gamma} \\
Llavors : \frac{WP}{(1 - WP)} &= \frac{PS^\gamma}{PA^\gamma} \\
ln(\frac{WP}{(1 - WP)}) &= \gamma \cdot ln(\frac{PS}{PA})
\end{aligned}$$

Finalment s'obté una expressió semblant a una recta de regressió.

Per trobar el pendent de la recta, que en aquest cas equival a l'exponent, utilitzem el paquet de *python* **sklearn**, on li passem totes les dades necessàries i retorna el valor de l'exponent.

Cerca de diversos exponents

Una vegada ja tenim dues maneres per calcular l'expectativa pitagòrica, falta utilitzar les dades de manera lògica per obtenir el millor exponent.

Per aquest motiu definim les següents idees per posteriorment utilitzar la que millor percentatge d'encert tingui.

Totes les Dades

La primera opció és utilitzar totes les dades disponibles per poder buscar l'exponent. D'aquesta manera tenim en compte tots els jugadors des del 1991 fins al 2016.

Mínim nombre de Partits

Una altra manera per trobar l'exponent és utilitzar només els jugadors que hagin jugat una quantitat mínima de partits. La raó d'aquesta selecció de les dades es deu que hi ha molts jugadors que han fet participacions esporàdiques als tornejos que figuren al fitxer. Llavors suposem que no cal tenir-los en compte, ja que no són una mostra representativa dels jugadors habituals en aquests tornejos.

Per Anys

Encara que el meu nivell de tenis no sigui molt elevat, sé per altres esports que la manera que es juguen evoluciona i potser les dades del 1991 no tenen gaire utilitat per predir partits de l'actualitat (2017). Per exemple, el futbol ha canviat molt respecte al 1991, tant en tècnica com plantejament tàctic... igual que el bàsquet, llavors malgrat la meva ignorància per aquest aspecte al tenis, no podia evitar posar-ho en dubte. Per aquest motiu decidim estudiar l'evolució del valor de l'exponent en períodes de 5 anys.

Nivell de Forma del Jugador al Torneig

Un altre aspecte a tenir en compte és el nivell físic dels jugadors al qual es presenten en un torneig. Com és lògic un jugador no rendirà de la mateixa manera al sortir d'una lesió que havent entrenat durant mesos, i aquest aspecte és prou important en el nivell d'un jugador per passar-ho per alt.

Simulació

Les simulacions són molt útils per predir el futur, ja que una vegada coneixem i podem modelar un medi/situació matemàticament podem per simulacions per veure com es pot esdevenir el futur. Aquest mètode és utilitzat en diversos àmbits, fins i tot a la fórmula 1.

Doncs com que aquest és un mètode molt utilitzat i el tenis és un esport fàcil de modelitzar, fer una simulació per predir un partit de tenis és prou interessant per nosaltres poder veure com rendeix i també comparar-ho amb l'expectativa pitagòrica.

Estructura

L'estructura de la simulació amb la qual hem treballat és similar al joc de tenis mateix, hi ha diverses funcions que criden unes a les altres i així es va construint el marcador:

normalize_probs(prob_1, prob_2)

Aquesta funció serveix per normalitzar les probabilitats de cada jugador envers a l'altre. Per exemple, si el jugador A guanya un 70% dels punts, i el jugador B un 65%, doncs aquesta funció les normalitza aquests percentatges per un partit entre A i B, en aquest exemple retornaria 51.8% i 48.1% respectivament.

point_sim(server_pct, receiver_pct)

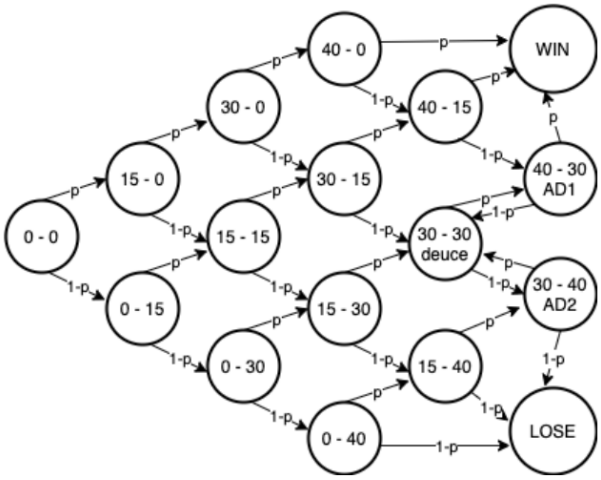
Aquesta funció rep el percentatge de punts guanyats, o l'estimació del percentatge de punts guanyats, els normalitza utilitzant la funció **normalize_probs** i retorna el guanyador seguint la distribució de **Bernoulli**.

tiebreak_sim(player_a, player_b)

Amb aquesta funció simulem un *tiebreak*, utilitzant la funció **point_sim** per simular els punts, tenint en compte el canvi d'ordre en la sacada i la diferència necessària per guanyar un cop s'ha arribat als 7 punts, podem determinar el guanyador d'un set.
Com que la probabilitat de guanyar el punt canvia per cada jugador si realitza o rep la passada, la funció rep dos vectors amb un conjunt de valors estadístics de cada jugador i així calcular les probabilitats en cada situació

game_sim(server,receiver)

En aquesta funció simulem el desenvolupament d'un joc de tenis emprant també la funció **point_sim** per determinar el guanyador de cada punt. Evidentment, també es té en compte la diferència necessària per guanyar el joc.



La simulació del joc segueix la mateixa estructura que la foto anterior. També rep un vector amb valors estadístics per calcular la probabilitat de fer punt pel servidor i receptor.

set_sim(player_a, player_b)

Per simular el desenvolupament d'un *set* utilitzem aquesta funció, que té una estructura similar a **game_sim**, respectant l'estructura del *set*, i en cas d'empat crida a **tiebreak_sim** per desempatar el set. Evidentment, es canvia el jugador que fa la sacada en cada joc, per això rep dos vectors amb valors estadístics per cada jugador.

match_sim(player_a, player_b, num_set)

Finalment aquesta funció crida tantes vegades com nombres de *sets* tinguem en els paràmetres de la funció.

Com calcular percentatge d'encert

Com hem comentat al principi, en algunes ocasions el percentatge de punts anotats pot variar segons si el jugador rep la pilota o realitza la sacada.

$$server\ pct = \frac{(server\ first\ serves\ in)\cdot(server\ first\ serves\ total)}{(server\ first\ serves\ total)\cdot(server\ first\ serves\ points\ total)} + \frac{(1\ -\ server\ first\ serves\ in)\cdot(1\ -\ server\ double\ faults)\cdot(server\ second\ serve\ points\ won)}{(server\ first\ serves\ total)\cdot(server\ second\ serve\ points\ total)\cdot(server\ second\ serve\ points\ total)}$$

$$receiver\ pct = (server\ first\ serves\ total)\cdot\frac{receiver\ first\ serve\ return\ won}{receiver\ first\ serve\ return\ total} + \frac{(server\ double\ foults)}{(server\ second\ server\ return\ total)} + \frac{(1\ -\ server\ player\ first\ serve\ in)\cdot(receiver\ second\ serve\ return\ won)}{(server\ first\ serves\ total)\cdot(receiver\ second\ serve\ return\ total)}$$

Utilitzant els valors estadístics d'aquesta manera tenim una millor aproximació de la proporció d'encerts dels tenistes a l'hora de sacar o rebre la sacada.

A més, també tenim en compte la capacitat dels jugadors en moments difícils, és a dir en *tie breaks*. Tenim en consideració aquesta facultat perquè alguns tenistes, com Rafa Nadal per exemple, són molt poderosos en aquest aspecte.

Variacions a la simulació original

A part d'utilitzar els percentatges d'encert d'una manera tan precisa, també és interessant fer la mateixa simulació descrivint el percentatge d'encert d'una manera més simple. Per aquest motiu també realitzem la simulació utilitzant els percentatges de punts realitzats globals.

Així, podrem comparar com rendeixen les dues simulacions les quals una és més complexa que l'altra.

Altres Consideracions

A l'hora de predir el guanyador d'un partit qualsevol, és necessari repetir el procés més d'una vegada per fer una regressió a la mitja i obtenir el guanyador d'una manera més segura.

Altres Mètodes

A part d'intentar predir amb els mètodes explicats anteriorment, també és important que provem el mateix objectiu amb mètodes molt més simples. D'aquesta manera podrem veure si és útil utilitzar models complicats, o si utilitzant models tan simples són suficients per obtenir un bon percentatge d'encert.
Per aquest motiu també hem utilitzat els següents mètodes:

Percentatge de victòries

Hem utilitzat el percentatge de victòries dels jugadors com a estadístic per determinar el guanyador de qualsevol partit. Evidentment, és un estadístic fàcil de calcular i utilitzar, però també pot ser enganyós en certes ocasions. Per exemple, un jugador s'estrena en algun dels tornejos que figuren a la base de dades i guanya, el seu percentatge de victòries serà del 100%, i quan jugui contra un rival amb certa experiència, al tenir un percentatge millor, el model el donarà per guanyador. Evidentment, és un cas molt poc probable, però a tenir en compte.

Nombre total de partits guanyats

Per compensar una mica el possible desavantatge del mètode anterior, també utilitzem com a estadístic el nombre total de partits guanyats, d'aquesta manera podem contrarestar els alts percentatges per la manca de partits. Per contra partida, aquest estadístic dona molta importància al passat dels jugadors, de manera que si un jugador molt veterà té com a rival a un jugador menys experimentat, però amb un bon nivell i estat de forma, aquest estadístic donarà per favorit al veterà, quan seria lògic pensar que el jove en aquest cas guanyaria el partit.

Modificació de la fórmula

Com que la sacada al tenis és un factor molt important al desenvolupament del joc, era lògic intentar reflectir aquesta rellevància també a la fórmula, d'una manera que ponderés positivament els jugadors amb una bona sacada, i negativament els jugadors amb una mala.

Aquesta modificació consisteix a dividir els punts que fa el jugador en la sacada pels punts que guanyen els seus rivals en la sacada. D'aquesta manera obtenim un factor multiplicatiu a prop d'1 per incrementar les probabilitats dels jugadors amb bones sacades.

Resultats

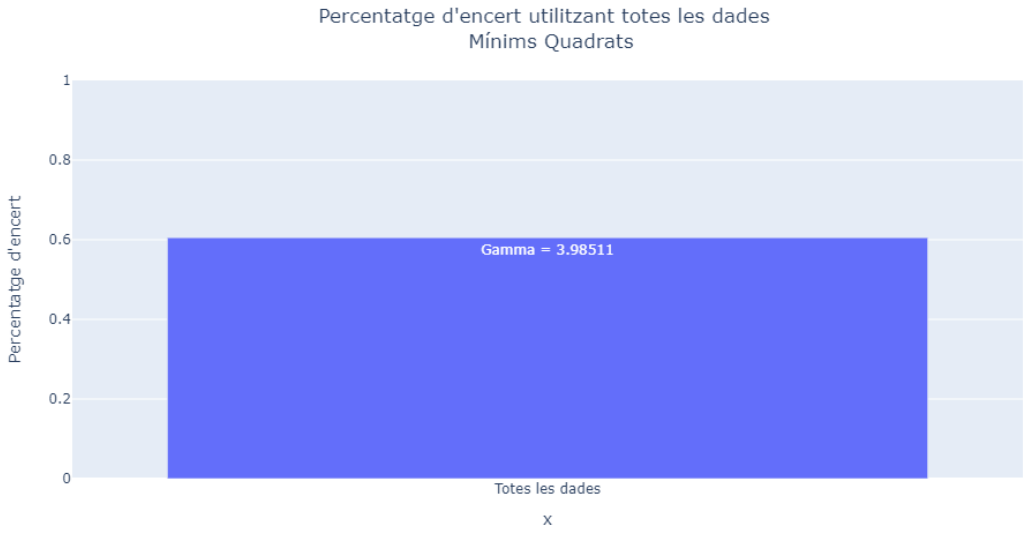
Per posar a provar els mètodes utilitzats hem utilitzat el fitxer amb les dades dels partits en l'any 2017. Aquest fitxer ens proporciona un ampli llistat de partits els quals tenim les dades històriques de la majoria de jugadors. Llavors aquesta situació és perfecta per simular futurs partits amb la peculiaritat que ja sabem el guanyador, així no cal esperar per avaluar els mètodes anteriorment creats.

Amb la multitud d'experiments que hem realitzat, hem obtingut resultats molt diversos, els quals la majoria estan entre 50%-60% quan fem servir tots els partits del 2017. A continuació els desglossarem més en detall:

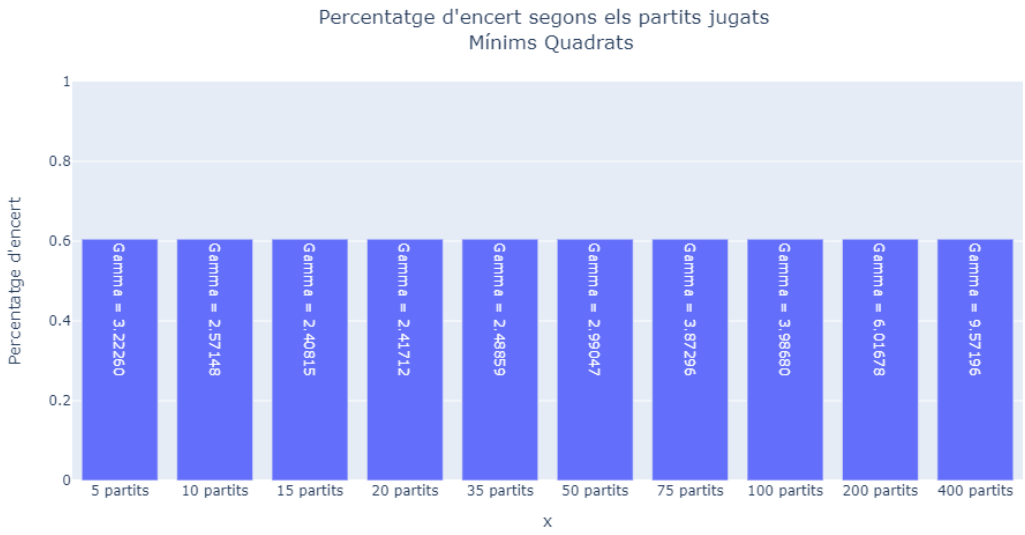
Expectativa Pitagòrica

Mínims quadrats

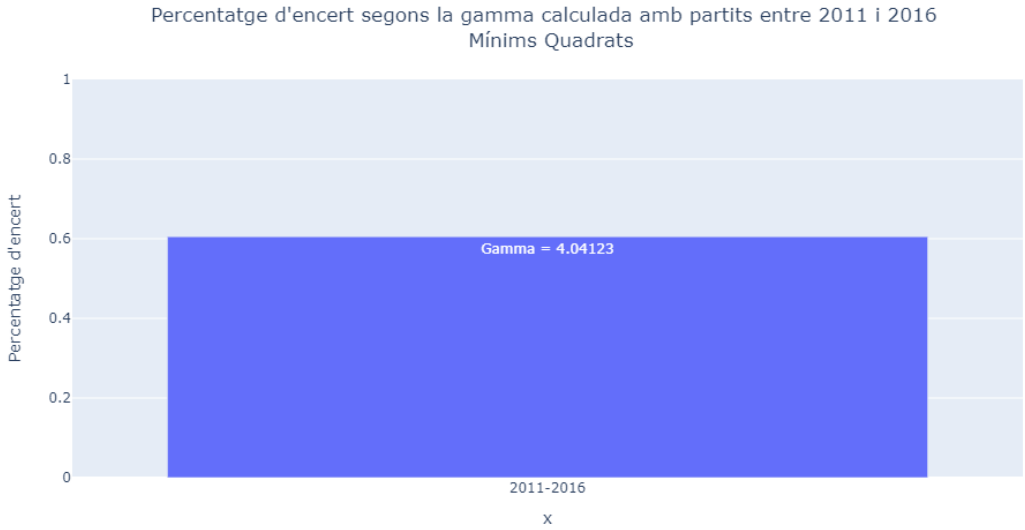
Totes les dades



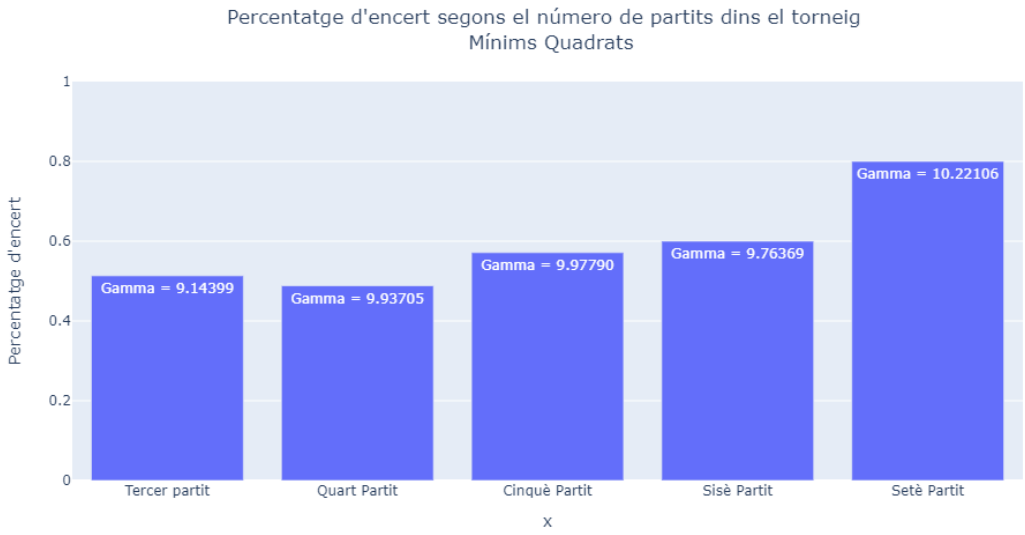
Mínim de partits jugats



Per anys



Nivell de forma del jugador al torneig



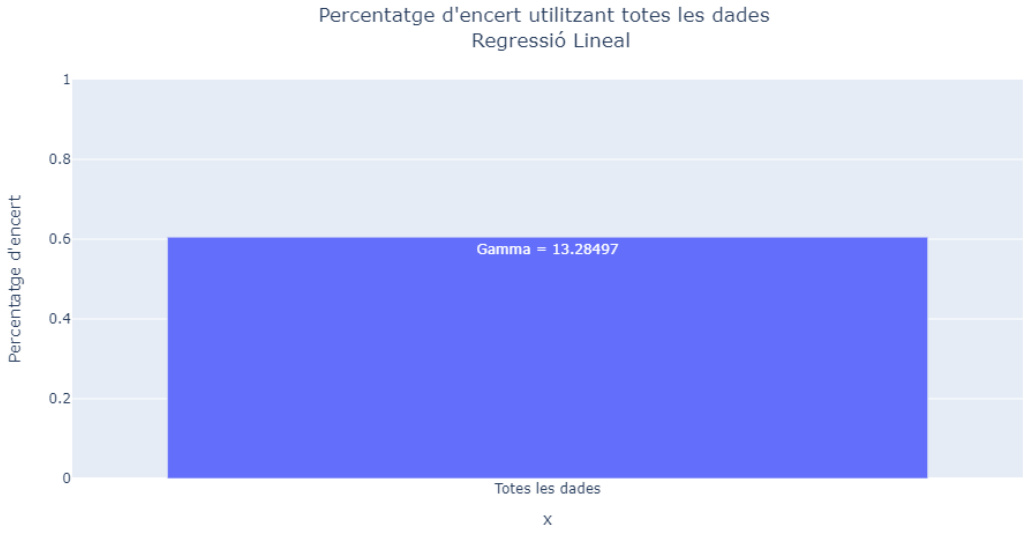
Amb els gràfics de barres anteriors podem veure que malgrat la variació de la γ el percentatge d'encert es manté constant. Només apreciem variacions en el percentatge quan tenim en compte el nivell del jugador durant el torneig.

Amb aquesta constància del model, malgrat la variació en el paràmetre γ, podem pensar que hem estat tractant amb un model robust, que no és gens sensible, llavors algunes petites variacions en aquest paràmetre no tenen conseqüències al resultat. Una altra explicació no tan probable és que en tots els partits del 2017 hi ha una diferència de nivell tan gran que malgrat les variacions en la γ, el jugador predit com a guanyador continua sent el mateix.

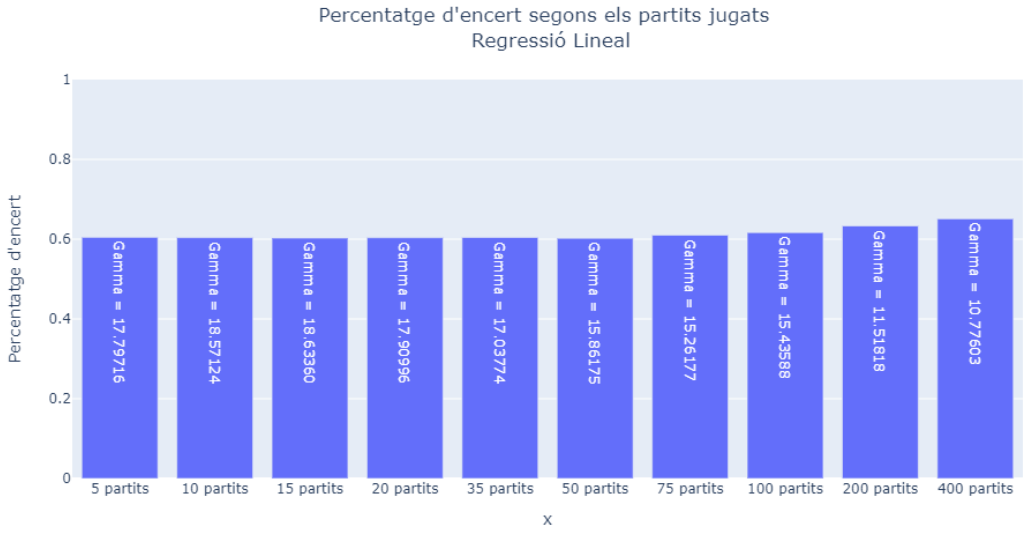
Pel que fa a les γ on tenim en compte el nivell del jugador al torneig, malgrat que algunes γ s'aproxim a la dels altres experiments, tenen un percentatge d'encert diferent perquè només s'aplica en els partits on es compleixin les condicions en les quals s'ha calculat la γ. És a dir, la γ del setè partit només s'aplica als partits els quals els jugadors ja portin 7 partits jugats al torneig.

Regressió lineal

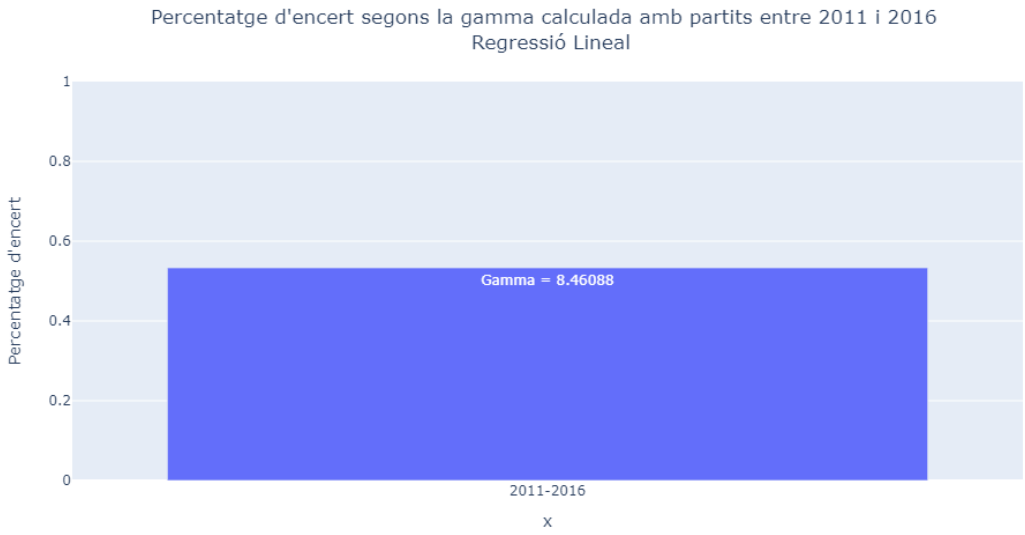
Totes les dades



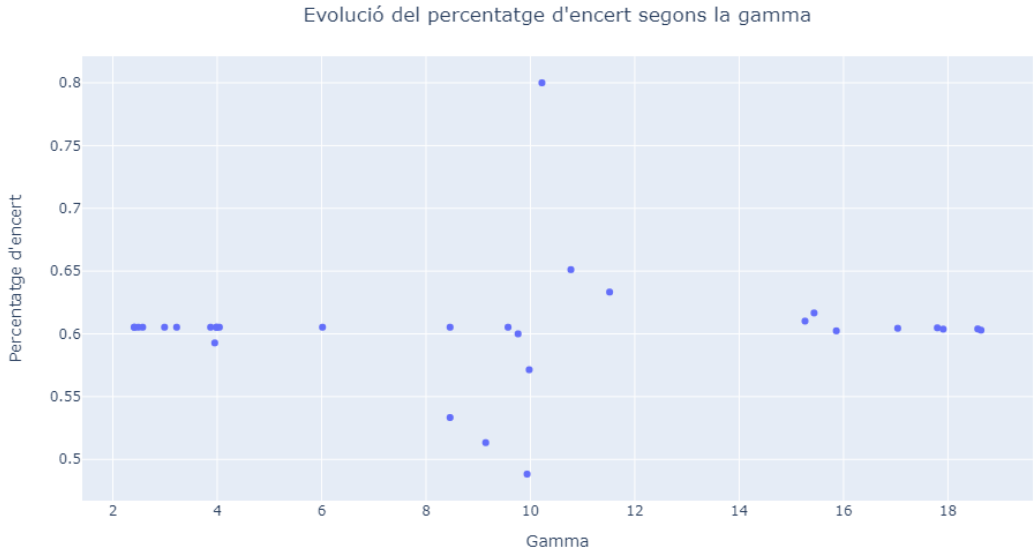
Mínim de partits jugats



Per anys

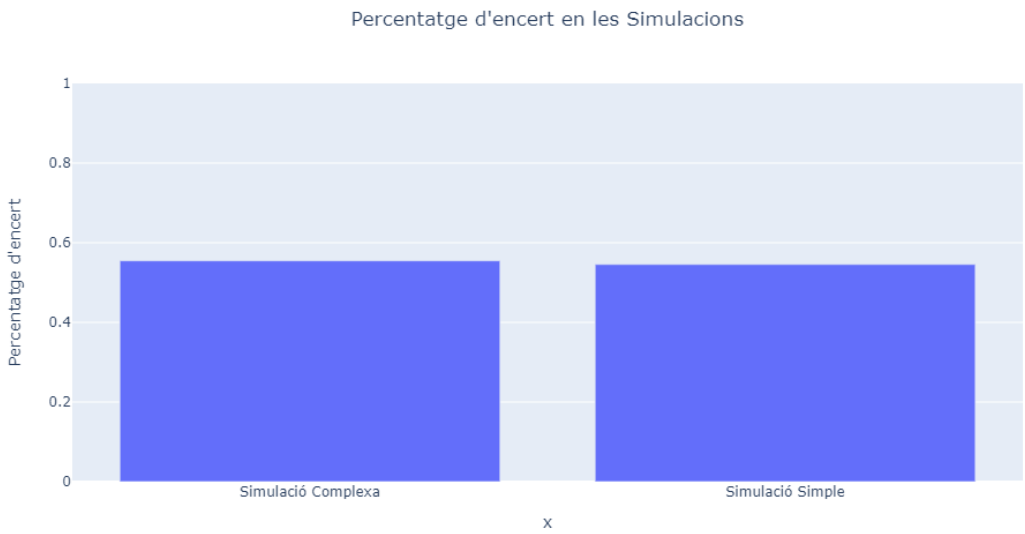


Pel que fa a les γs calculades utilitzant la regressió lineal són molt diferents respecte a les calculades utilitzant els mínims quadrats. Malgrat aquestes diferències, no podem apreciar una gran millora o empitjorament dels resultats.



En una imatge global podem observar que les γ baixes ($\gamma \in (3,\ 5)$) i altes ($\gamma \in (14,\ 19)$) són més estables que les γ que es troben pel mig. Amb aquesta última imatge podem apreciar la solidesa que hem comentat anteriorment, excepte per γ a prop de 10. També hem de mencionar que els valors de γ a prop de 10 la majoria corresponen a les γ obtingudes als models on teníem en compte el nivell del tenista dins del torneig.

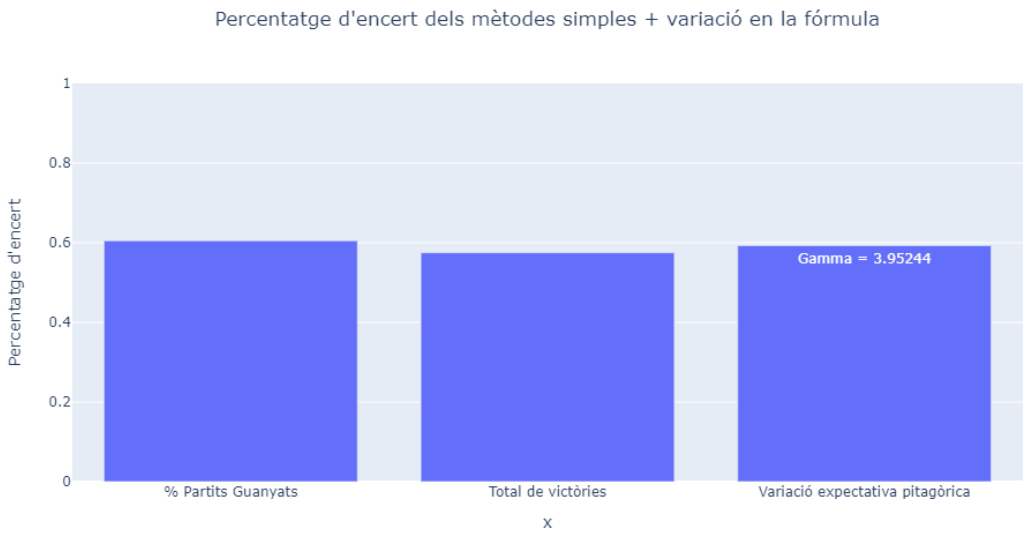
Simulació



Respecte a la simulació, amb la imatge superior veiem la poca diferència d'encert quan calculem el percentatge de punts d'una manera explícita i quan els calculem d'una manera molt més simple.

Malgrat la simulació complexa és una mica millor que la simple, el percentatge d'encert sí que és considerablement inferior als encerts realitzats per l'expectativa pitagòrica. Llavors amb aquests percentatges, ja podem dir que l'expectativa pitagòrica funciona millor.

Altres mètodes



Amb aquest últim gràfic veiem el bon rendiment que han tingut els models alternatius malgrat la seva simplicitat, amb un percentatge d'encert molt similar a l'obtingut amb l'expectativa pitagòrica.

Respecte al model amb la fórmula modificada observem que ha rendit igual que qualsevol altre model de l'expectativa pitagòrica. Això es deu al fet que incorporar una petita modificació al valor no canvia en la predicció final degut a la solidesa del model. Per aquest motiu és lògic que no veiem algun tipus de variació important al resultat.

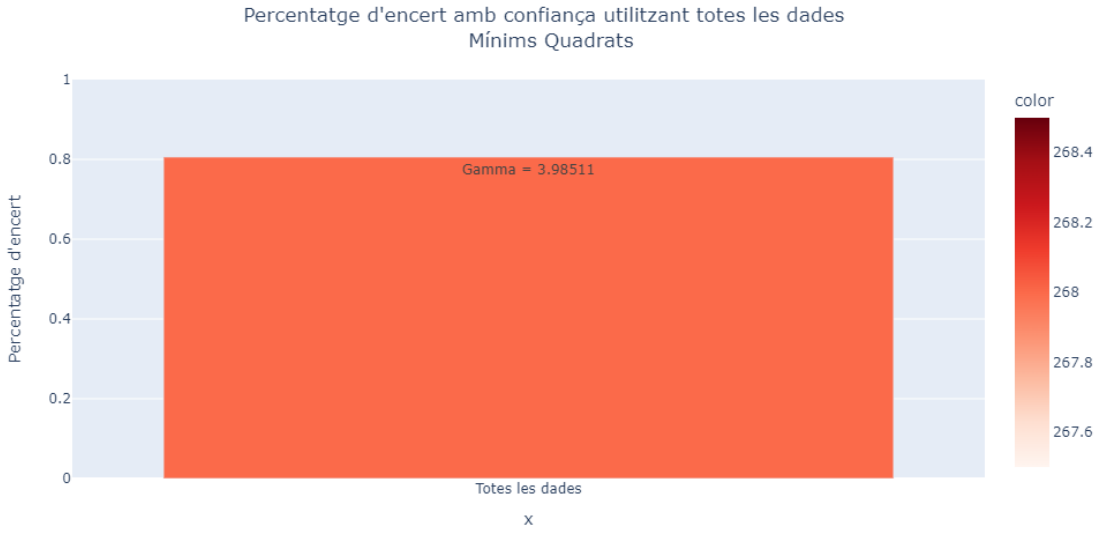
Expectativa Pitagòrica amb Confiança

En les anteriors prediccions teníem en compte tots els partits que figuren al fitxer del 2017, però no sabíem amb quina confiança assignàvem el guanyador del partit. Per exemple, si el jugador A té un 51% de probabilitats de guanyar el partit i el jugador B un 49%, direm que guanya el jugador A, però amb molta poca confiança.

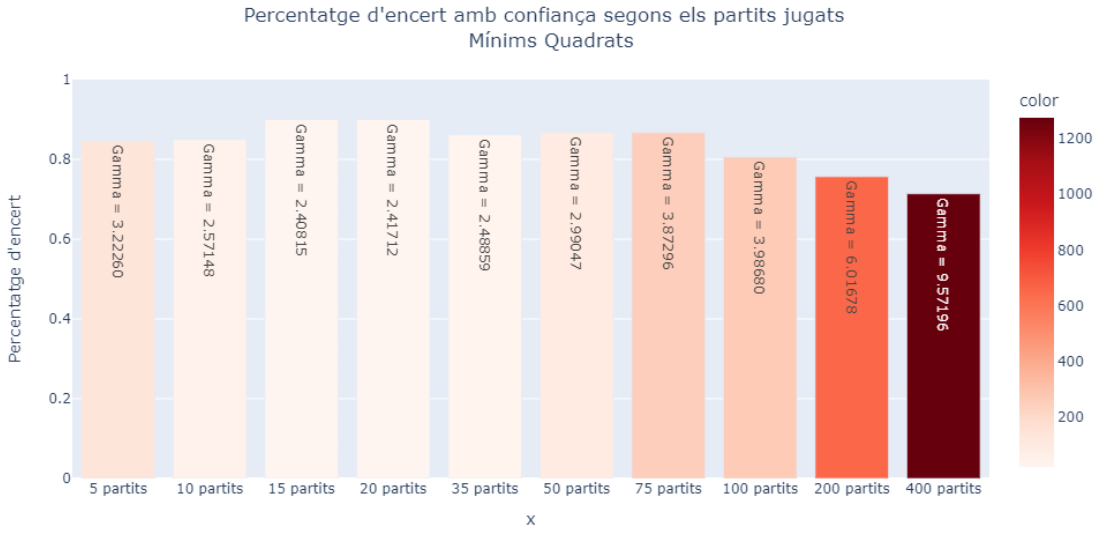
Per evitar males prediccions, tornarem a aplicar l'expectativa pitagòrica, només aquest mètode perquè és el que ha rendit millor, i només realitzarem prediccions en partits on la diferència de probabilitats sigui superior o igual al 25%.

Mínims quadrats

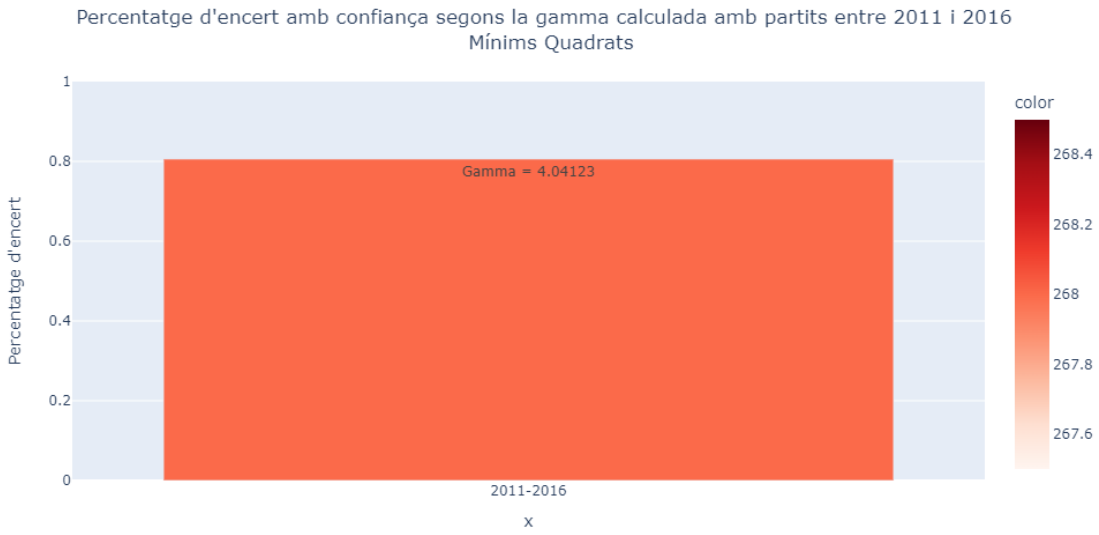
Totes les dades



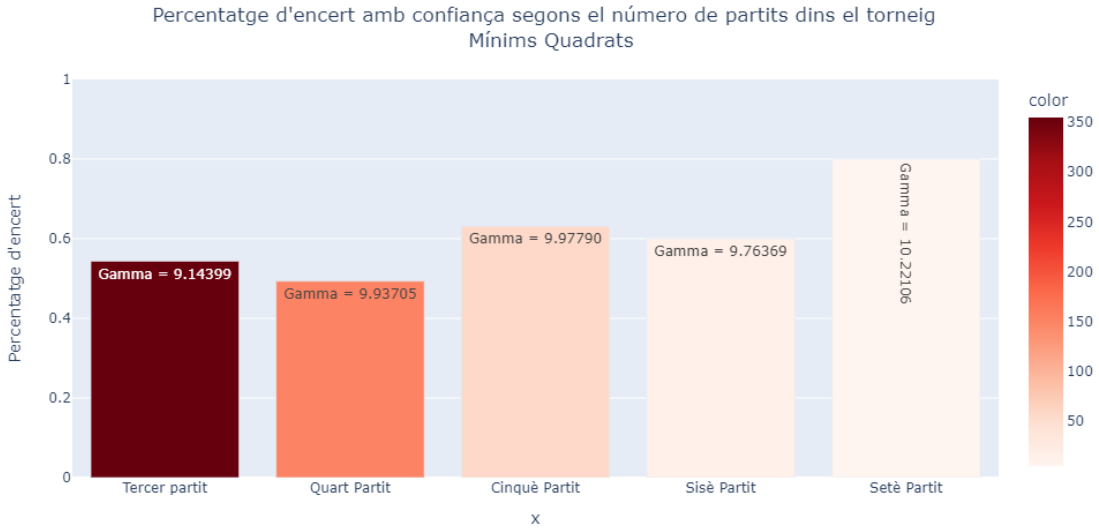
Mínim de partits jugats



Per anys



Nivell de forma del jugador al torneig

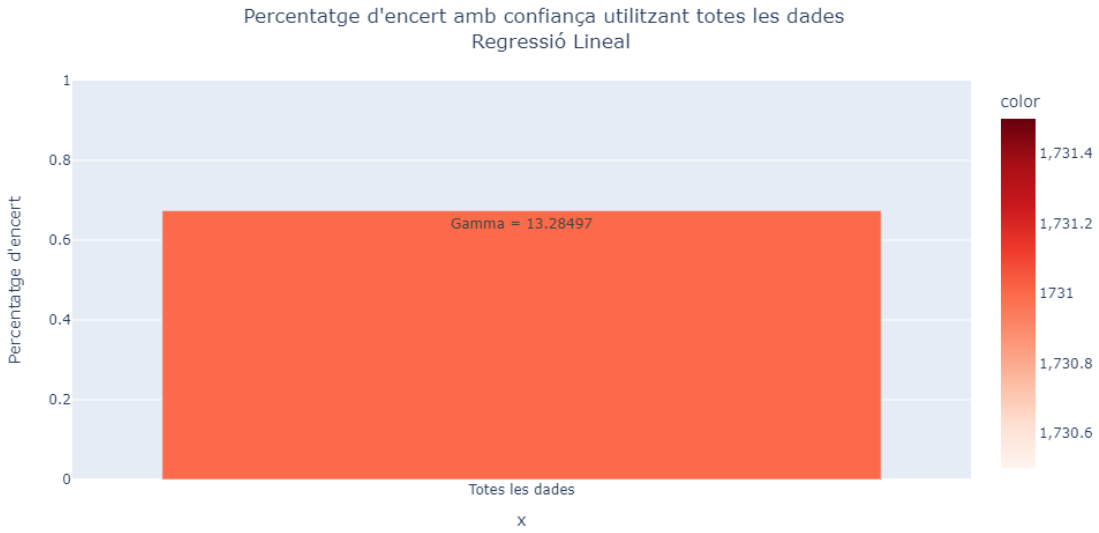


Assignant un guanyador als partits els quals estem més segurs del guanyador veiem que millorem el percentatge d'encert fins a un 85% aproximadament. Evidentment, la quantitat de partits en la qual fem una predicció redueix considerablement, anteriorment treballàvem en gairebé 3800 partits, i filtrant segons el nivell de confiança, utilitzem entre 100 i 1000 partits generalment **(La variable color al gràfic representa el nombre de partits.)**.

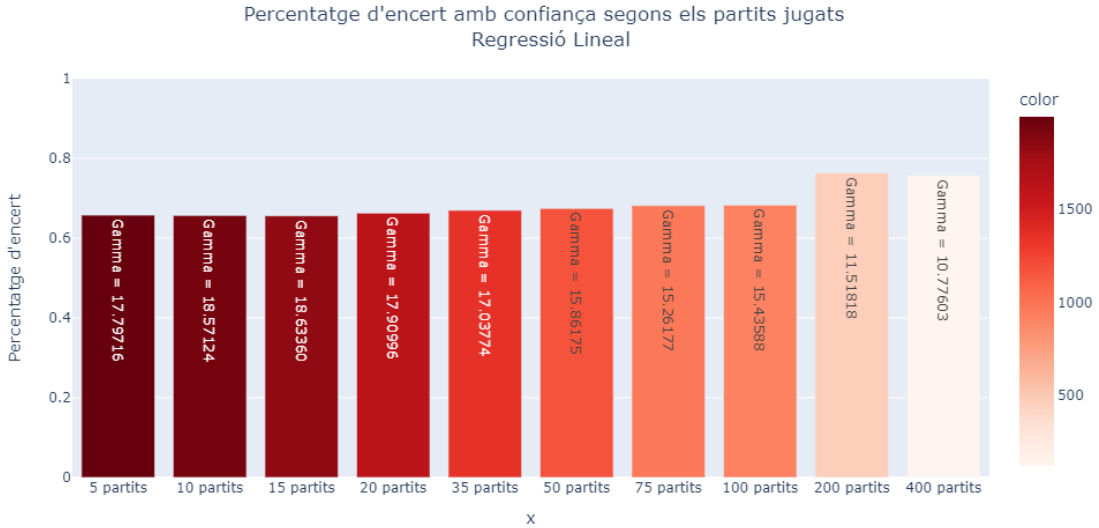
Veiem aquesta millora en els casos quan utilitzem totes les dades, o només utilitzem els jugadors que han jugat un mínim de partits, però també s'ha de tenir en compte el nombre de partits els quals prediuen. Llavors hem de ser nosaltres qui decidim què prioritzar, si un al percentatge d'encert i pocs partits, o un percentatge d'encert més baix amb més partits.

Regressió Lineal

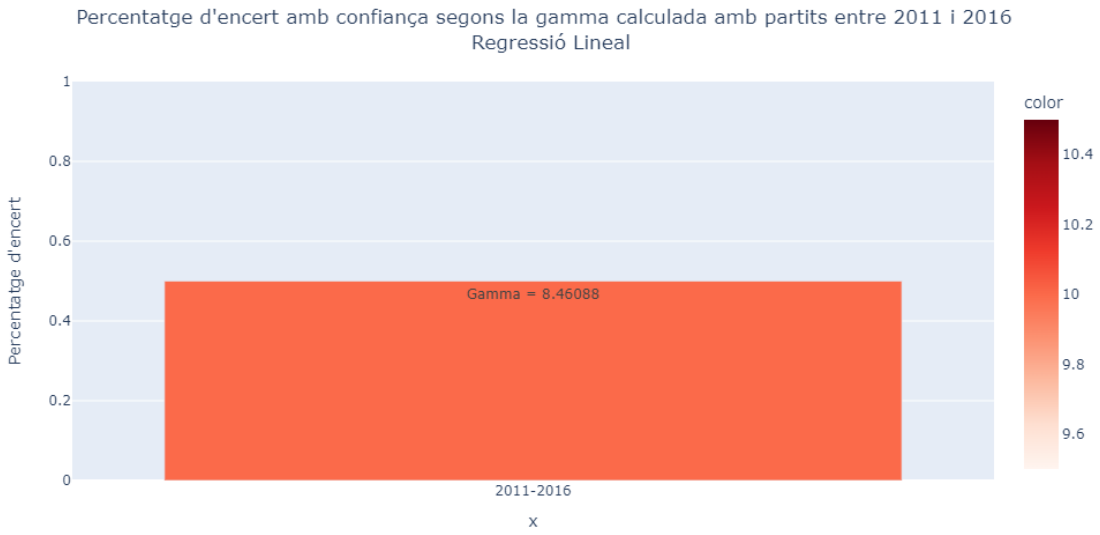
Totes les dades



Mínim de partits jugats



Per anys



Pel que fa a la predicció amb confiança utilitzant la regressió lineal, podem veure que no obté tants bons percentatges com mínim quadrats, però també utilitza molts més partits.

Aquesta diferència tan gran amb la quantitat de partits la podem justificar amb el valor de la γ. En la majoria de casos de la regressió lineal, obtenim un valor de γ molt elevat, en canvi, per mínims quadrats molt baix. Llavors podem pensar que com més elevat sigui el valor de γ, més augmenta les diferències entre les probabilitats dels jugadors, i com més baix el valor, més igualades seran les probabilitats.

Conclusions

Abans de començar amb les conclusions cal recalcar que el tenis és un esport molt complex on no tots els punts tenen el mateix valor que altres, llavors al no tenir un valor constant dificulta molt la nostra tasca, ja que un jugador pot puntuar més que el rival i perdre el partit, i tenir tots els possibles casos en consideració dificulta la nostra tasca de predicció.

Generals

Després de realitzar diverses proves de predicció amb mètodes variats podem concloure que l'expectativa pitagòrica és més bona que una simulació a l'hora de predir el guanyador d'un partit de tenis. A part de ser més ràpida computacionalment una vegada ja s'ha calculat l'exponent, també obté una millor qualificació.

Malgrat l'expectativa pitagòrica sigui millor que la simulació, potser no és el millor mètode per predir el guanyador d'un partit de tenis, ja que els mètodes de decisió simple que també hem provat han aconseguit uns resultats similars a l'expectativa pitagòrica. Aquest fet deixa la porta oberta a futura investigació en utilitzar altres mètodes de predicció pel mateix objectiu.

Simulació

Respecte al mètode de simulació, malgrat que la simulació complexa fos més precís a l'hora de determinar el guanyador, la poca diferència que hi ha entre aquesta i la simulació simple fa que ens plantegem que la simulació simple sigui òptima. Tot i tenir un pitjor percentatge, la simplicitat en calcular les probabilitats fa que el no haver de necessitar tants components estadístics fa que en la imatge global sigui millor utilitzar-la.

Expectativa Pitagòrica

Pel que fa a l'expectativa pitagòrica hem vist que hi ha dos intervals provinents de dos mètodes diferents on els resultats són estables: $\gamma \in (3,\ 5)$ correspon al mètode de mínims quadrats i l'interval $\gamma \in (14,\ 19)$ correspon a la regressió lineal. Aquests valors de γ obtenen uns valors d'encert a prop del 60%.

També hem vist com es comportaven les diferents γ quan només realitzàvem prediccions en aquells partits on estàvem segurs del resultat, i aquí sí que hi havia més divergència en els percentatges d'encerts i en la quantitat de partits utilitzats.
Les γ de valor baix obtenien un percentatge d'encert molt elevat, però també utilitzaven pocs partits (100 aprox.) a diferència de les γ amb valors alts que sí que milloraven el percentatge d'encert respecte a la predicció amb tots els partits, però utilitzaven molts més partits i obtenien un percentatge d'encert menor respecte als valors baixos de γ.

A partir d'aquí nosaltres hem de decidir quina γ utilitzar segons el nostre objectiu. Si volguéssim prioritzar un alt percentatge d'encert hauríem d'utilitzar valor de γ baixos. En canvi, si volguéssim prioritzar la quantitat de partits, doncs empraríem valors de γ alts. I en cas que optéssim entre un equilibri de quantitat de partits

i qualitat de la predicció, llavors faríem servir un valor de γ intermedi.

En línies generals podem estar contents dels resultats obtinguts amb l'expectativa pitagòrica, ja que sense utilitzar grans coneixements de tenis hem pogut predir el guanyador en diversos partits en un esport que és bastant imprevisible, i més utilitzant una variable que no té per què ser determinant en el guanyador del partit.

Altres mètodes

Amb els mètodes encara més simples hem obtingut una quantitat d'encerts similar a la de l'expectativa pitagòrica. Aquests resultats ens fan plantejar si hem escollit el millor mètode per intentar predir el tenista guanyador.

Treball Futur

Hem vist que l'expectativa pitagòrica és un bon mètode per predir partits de tenis, però durant aquest treball hem descartat diverses opcions tant per falta de dades com per falta de temps.

Per exemple, en les cases d'apostes per determinar les probabilitats d'un guanyador en un partit de tenis tenen en compte els següents factors: l'estat de forma, terreny de joc, i enfrontaments passats en un 70%, 20% i 10% respectivament, segons un treballador de **Winamax**, una casa d'apostes localitzada a París. Nosaltres sí que hem pensat el nivell de forma, però no l'hem pogut descriure tan bé com les cases d'apostes, ja que ells treballen amb molts més valors estadístics, el terreny de joc no l'hem pogut utilitzar, ja que tampoc teníem una variable que ens descrigués la superfície del terreny. Respecte a els enfrontaments passats és un aspecte que no hem elaborat, però que ho haguéssim pogut fer.

A més, en altres esports també s'ha enfocat un estudi similar, però estudiant molt més la distribució estadística de les dades, un enfocament que encara es podria aplicar al tenis.

Deixant de banda l'expectativa pitagòrica, hi ha molts més mètodes els quals es podrien aplicar fer aconseguir el mateix objectiu. Des dels arbres de decisió fins a les xarxes neuronals, amb totes les possibilitats que aquestes impliquen. Llavors encara tenim moltes possibilitats per millorar les prediccions fetes fins al moment.

Bibliografia

Mark Jamison (2021). "Modelling a Game of Tennis". Recuperat de: https://towardsdatascience.com/building-a-tennis-match-simulator-in-python-3add9af6bebe

John Chen & Tengfei Li (2016), "The Shrinkage of the Pythagorean exponents". Journal of Sports Analytics. 2 Vol., pàg. 37-48

Steven J. Miller (2006), "A derivation of the pythagorean won-loss formula in baseball".

Joe Peta (2014), "Trading Bases: How a Wall Street Trader Made a Fortune Betting on Baseball". (1r Edició). New American Library.

Tom M. Mango & Mitchel G. Lishtman & Andrew E. Dolphin (2014), "The Book: Playing the percentages in baseball". (1r Edició). Independent.

Plotly Express Documentation. Recuperat de: https://plotly.com/python-api-reference/plotly.express.html