Estadística Aplicada II - Trabajo final

Juan Pablo Cordero Mayorga Jaime Fernando Uria Medina

Gerardo Armando Guerrero Álvarez

Los torneos de Grand Slam no solo concentran el talento más alto del tenis profesional, sino que también implican una compleja logística con cientos de partidos, múltiples sedes y contratos televisivos que superan los 300 millones de dólares por edición. En ese contexto, anticipar con precisión la duración de cada encuentro es importante para organizar horarios de cancha, optimizar la parrilla de transmisiones y planear la recuperación de los jugadores.

Este estudio se propone modelar y predecir la duración de los partidos masculinos disputados entre 1991 y 2022, comparando tres enfoques metodológicos: regresión lineal (OLS), modelos de árboles (Random Forest y XGBoost) y modelos lineales generalizados Gamma con enlace logarítmico.

Se construyeron cuatro versiones para cada tipo de modelo, incorporando progresivamente nuevas variables para evaluar la mejora en cada iteración.

La regresión lineal OLS mostró una mejora modesta al incluir más variables, alcanzando su mejor resultado con un RMSE de 44.80 minutos y un R^2 de 0.055 en el conjunto de prueba.

Este desempeño fue superado por los modelos de árboles, especialmente XGBoost, cuyo modelo completo logró reducir el RMSE a 44.81 minutos, aunque con signos claros de sobreajuste. Por su parte, el modelo Gamma destacó por su equilibrio entre precisión (RMSE: 44.98) y robustez (pseudo R^2: 0.0574), siendo menos sensible al overfitting que los métodos de machine learning.

Las variables con mayor poder explicativo fueron la superficie de juego (particularmente el pasto, que reduce significativamente la duración), la diferencia de puntos en el ranking y el promedio de minutos jugados por los tenistas. Se validaron los supuestos de independencia en los modelos lineales y se aplicó un análisis de multicolinealidad para eliminar variables redundantes (VIF > 10). Aunque los modelos no incorporan factores como clima o lesiones, los resultados sugieren que es factible predecir la duración de los partidos con una precisión útil para aplicaciones operativas y comerciales en torneos de gran escala.

Feature engineering

Para mejorar la capacidad predictiva de los modelos, se implementó un proceso sistemático de ingeniería de variables enfocado exclusivamente en partidos masculinos a cinco sets disputados en torneos de Grand Slam. Primero, se reetiquetaron los jugadores como *jugador_1* y *jugador_2* en función de sus puntos de ranking previos al partido, con el fin de garantizar consistencia en las estadísticas asociadas a cada rol. Luego, se crearon variables derivadas como la diferencia de ranking (`rank_diff`), edad (`age_diff`) y la cercanía competitiva (`close_ranking`), buscando capturar relaciones no lineales entre características individuales y duración del partido.

Posteriormente, se reformateó el dataset a un esquema centrado en el jugador para calcular promedios móviles de rendimiento —como aces, doble faltas, puntos ganados al servicio y duración de partidos— usando una ventana retrospectiva de cinco encuentros. Estas métricas fueron luego reagrupadas nuevamente a nivel de partido, distinguiendo entre el historial del `jugador_1` y del `jugador_2`. Este enfoque permitió incorporar conocimiento histórico del desempeño individual sin incurrir en *data leakage*, ya que toda información se construyó exclusivamente con partidos previos al observado. Finalmente, se eliminaron las estadísticas dentro del partido actual, así como identificadores y columnas textuales irrelevantes, asegurando que las variables finales fueran numéricas, informativas y compatibles con el entrenamiento de modelos supervisados.

Feature selection

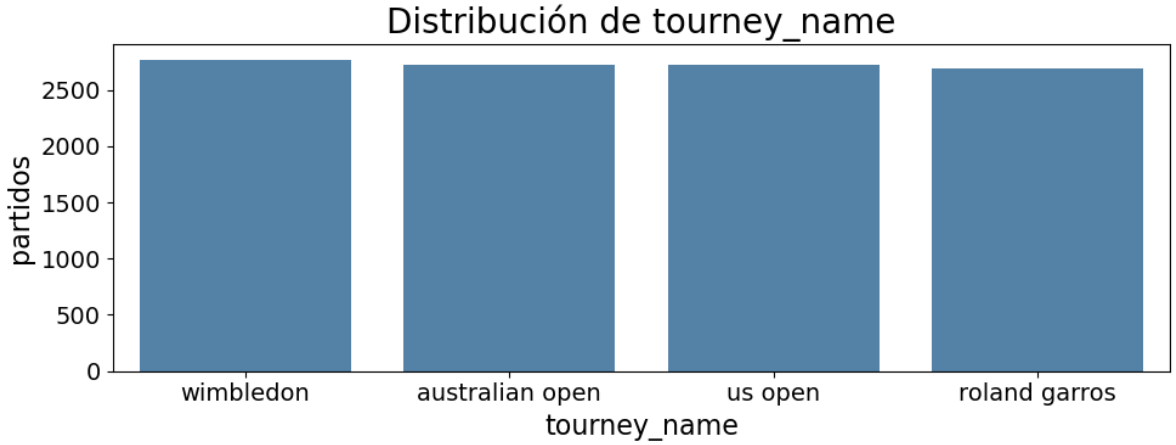
Una vez concluido el proceso de ingeniería de variables, se procedió a reducir dimensionalidad y eliminar redundancias mediante un análisis estadístico formal. Dado que muchas de las métricas generadas estaban correlacionadas entre sí —por ejemplo, estadísticas agregadas como `jugador_1_1stWon` o `avg_rank`—, era fundamental mitigar la multicolinealidad, especialmente para preservar la estabilidad de los modelos lineales y facilitar la interpretación de los coeficientes.

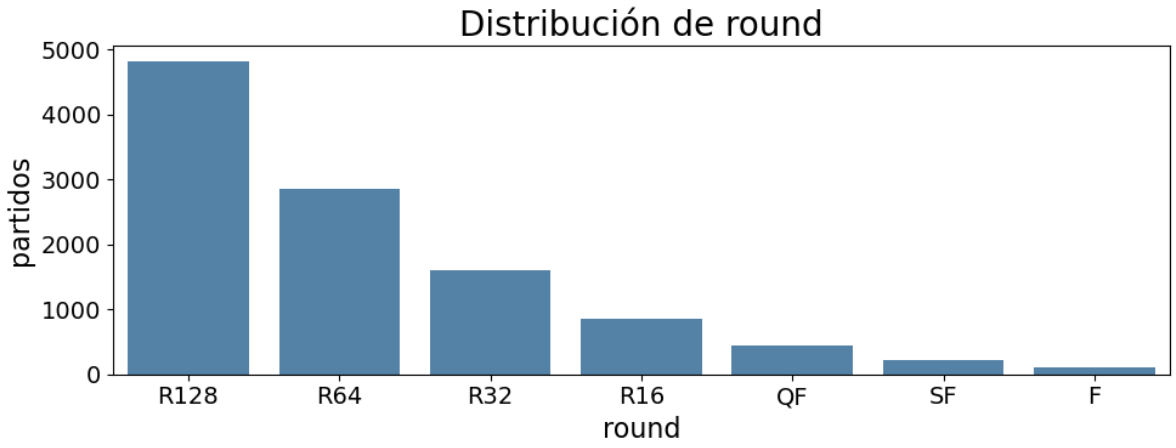
El criterio principal utilizado fue el Factor de Inflación de la Varianza (VIF), que mide cuánta varianza de una variable es explicada por el resto. Se calcularon los VIF únicamente para variables numéricas, y se eliminaron aquellas cuyo VIF superaba el umbral de 10, un valor estándar que indica colinealidad severa. Este análisis permitió conservar únicamente aquellas variables con contribución informativa no redundante, reduciendo el riesgo de sobreajuste y asegurando un conjunto parsimonioso para el modelado.

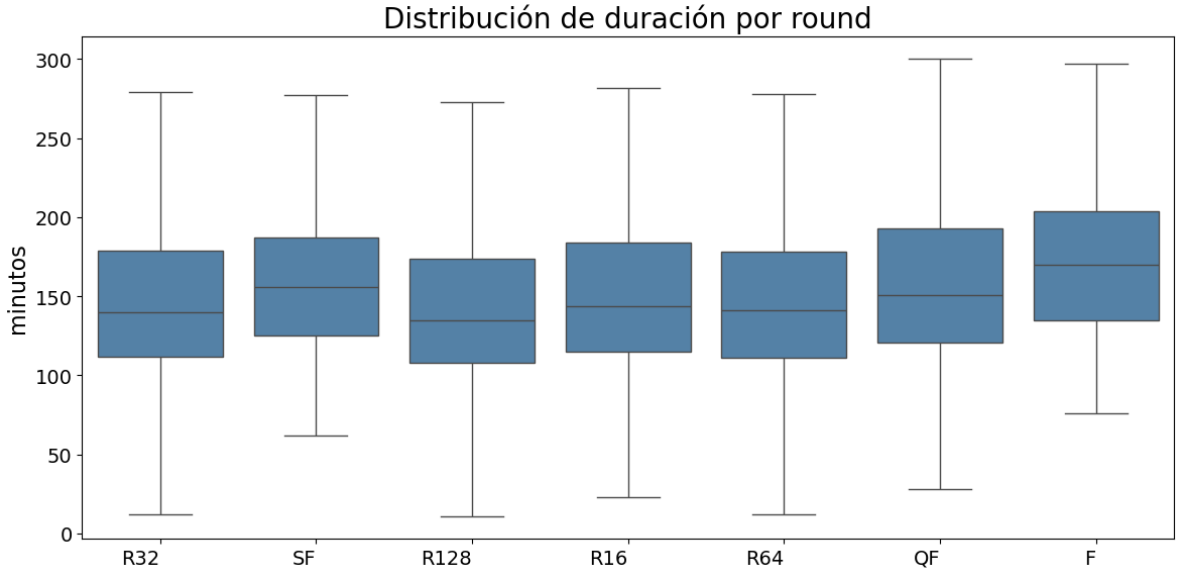
Como resultado de este filtrado, el conjunto final de variables utilizadas en los modelos se compuso únicamente por aquellas con bajo VIF y con relevancia teórica o empírica validada en etapas posteriores del análisis. Este conjunto formó la base del modelo completo (*full model*) utilizado en las tres familias de modelos predictivos.

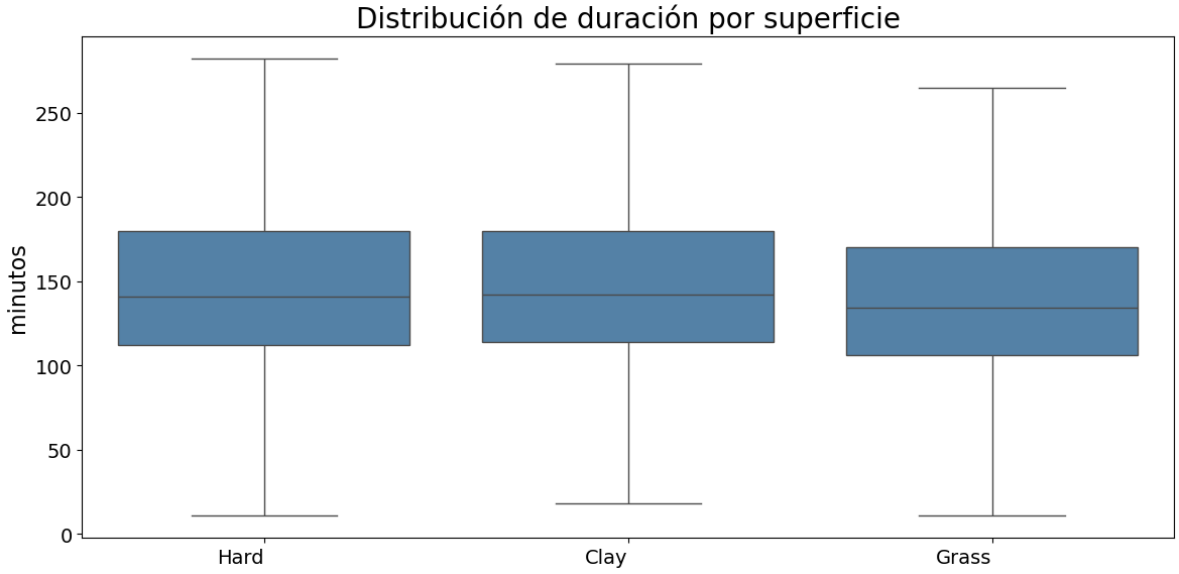
ETL

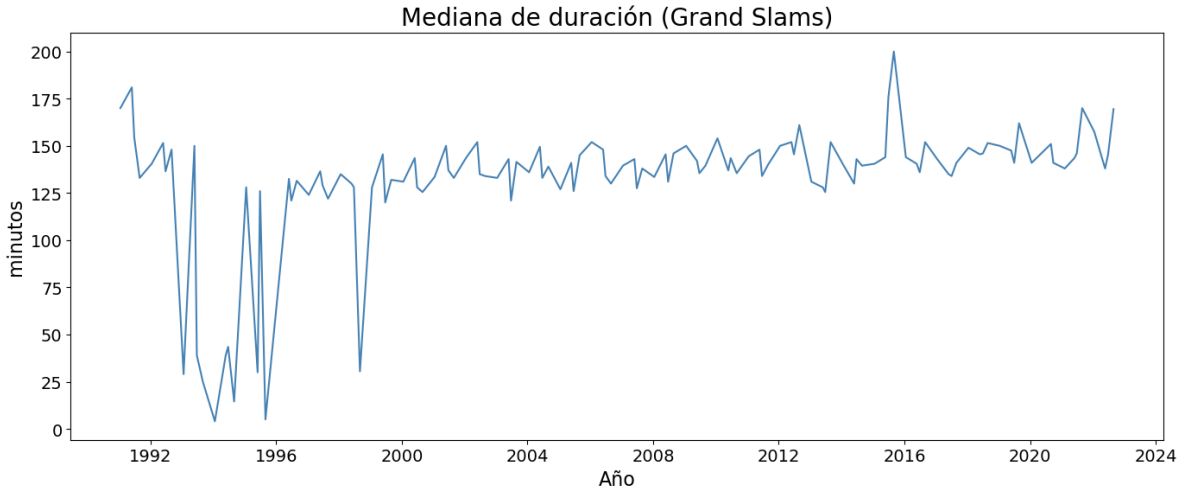
Realizamos un proceso de Extracción, Transformación y Carga (ETL) para preparar los datos de los partidos masculinos de Grand Slam entre 1991 y 2022. La extracción se llevó a cabo desde la base de datos de la ATP, que contiene información detallada sobre cada partido, incluyendo estadísticas individuales y del encuentro. Posteriormente, se transformaron los datos para unificar formatos, eliminar duplicados y corregir inconsistencias. Finalmente, los datos fueron cargados en un entorno adecuado para el análisis, asegurando que todas las variables estuvieran correctamente tipificadas y listas para su uso en modelos predictivos.






Distribución de duración por round
minutos
0
50
100
150
200
250
300
R32
SF
R128
R16
R64
QF
F


Distribución de duración por superficie
minutos
0
50
100
150
200
250
Hard
Clay
Grass




```
  VIF Scores (Numerical Features Only):
       feature  VIF
13     avg_age  inf
2   winner_age  inf
4    loser_age  inf
...
      feature       VIF
1   winner_ht  1.364780
3    loser_ht  1.331076
12   age_diff  1.041457

  Dropping high VIF features

  Tennis data analysis pipeline complete!
Final dataset shape: (10746, 18)
```

	p1_ht	p2_ht	p1_rank_points	...	round_group_QF	round_group_SF	minutes
0	188.0	170.0	3889.0	...	1.0	0.0	88.0
1	188.0	180.0	2541.0	...	1.0	0.0	147.0
2	183.0	178.0	303.0	...	1.0	0.0	204.0
3	190.0	190.0	3528.0	...	1.0	0.0	111.0
4	188.0	188.0	2541.0	...	0.0	1.0	242.0

Regresión Lineal

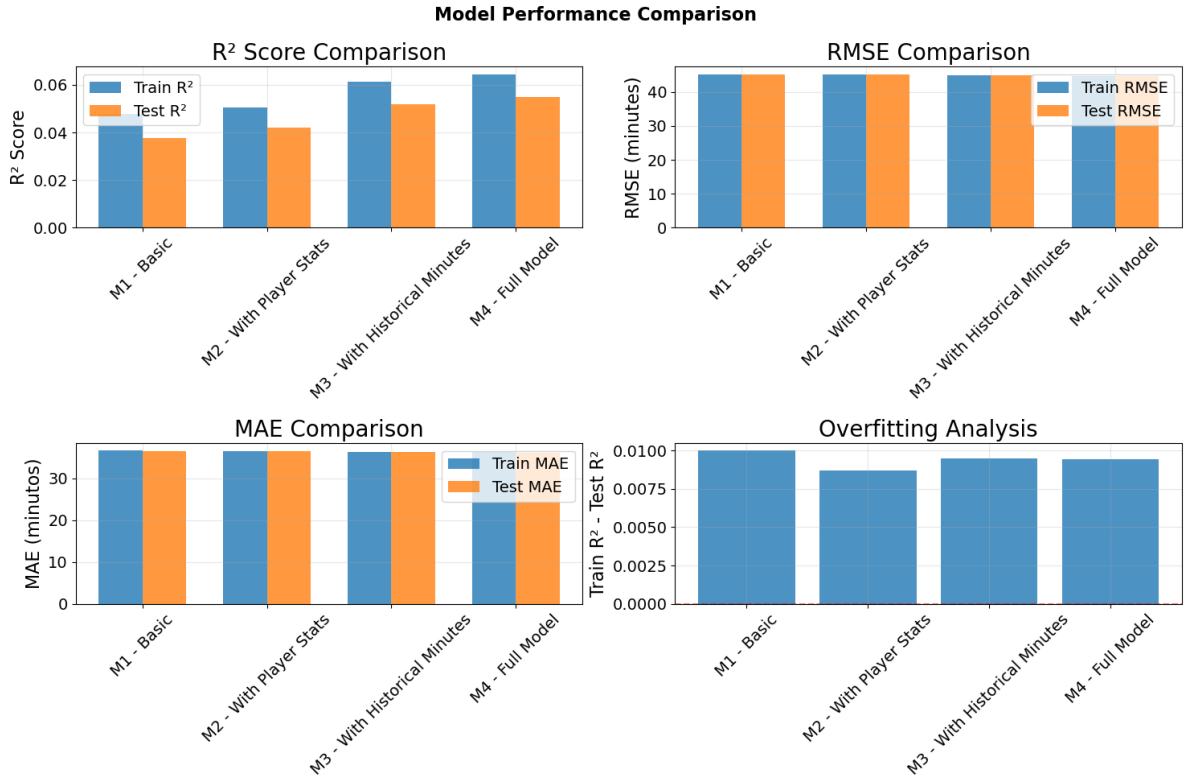
Marco Teórico

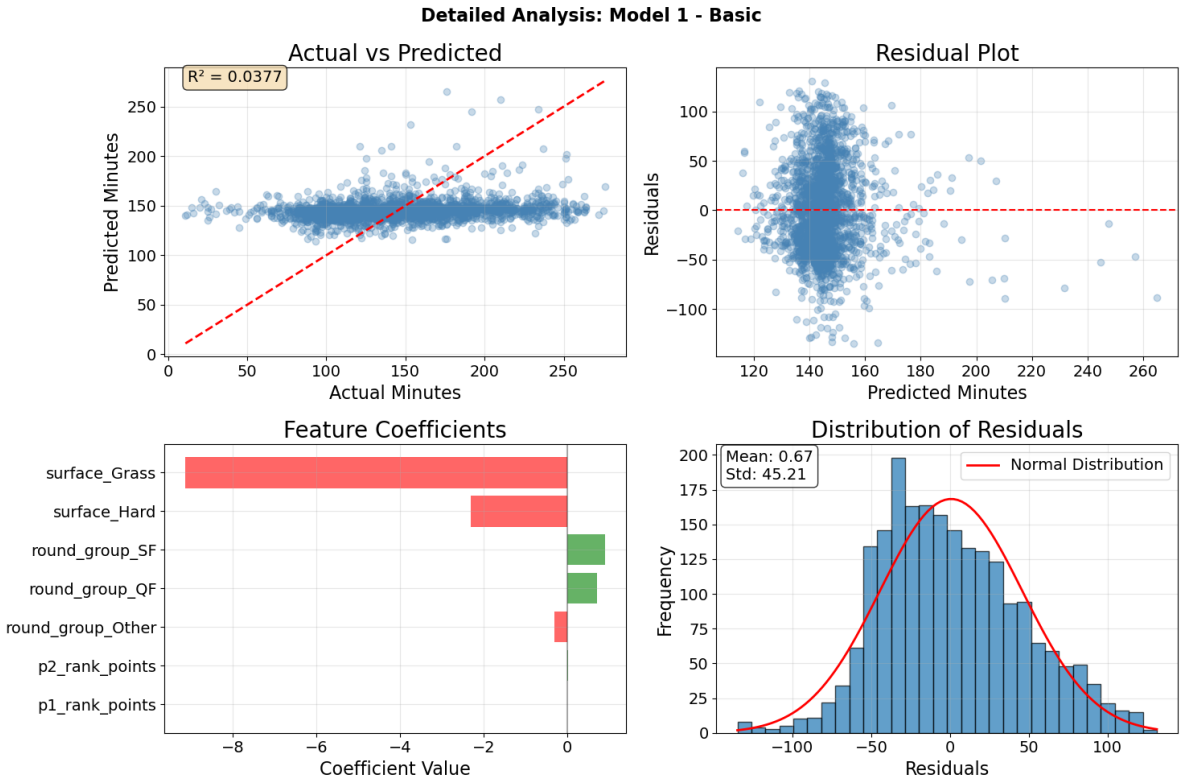
La regresión lineal constituye un modelo paramétrico fundamental en estadística para analizar la relación entre una variable dependiente continua y un conjunto de variables independientes. Este modelo asume una relación lineal expresada mediante la ecuación

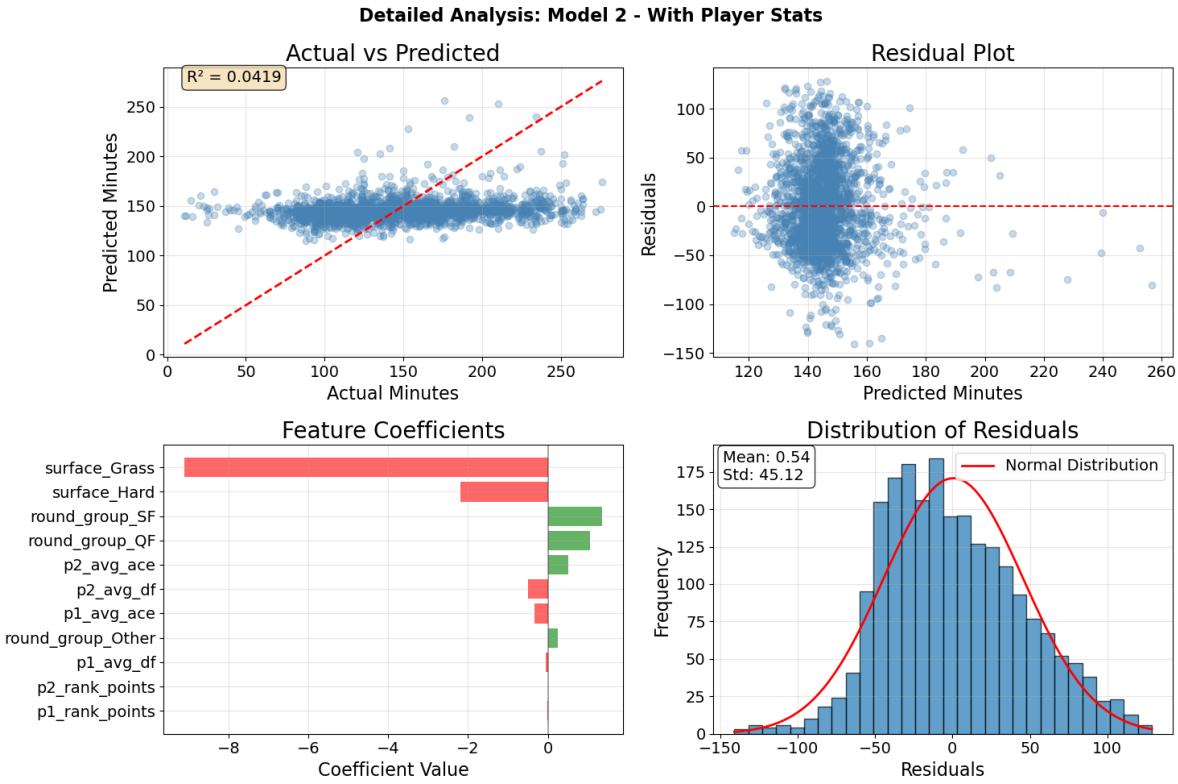
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

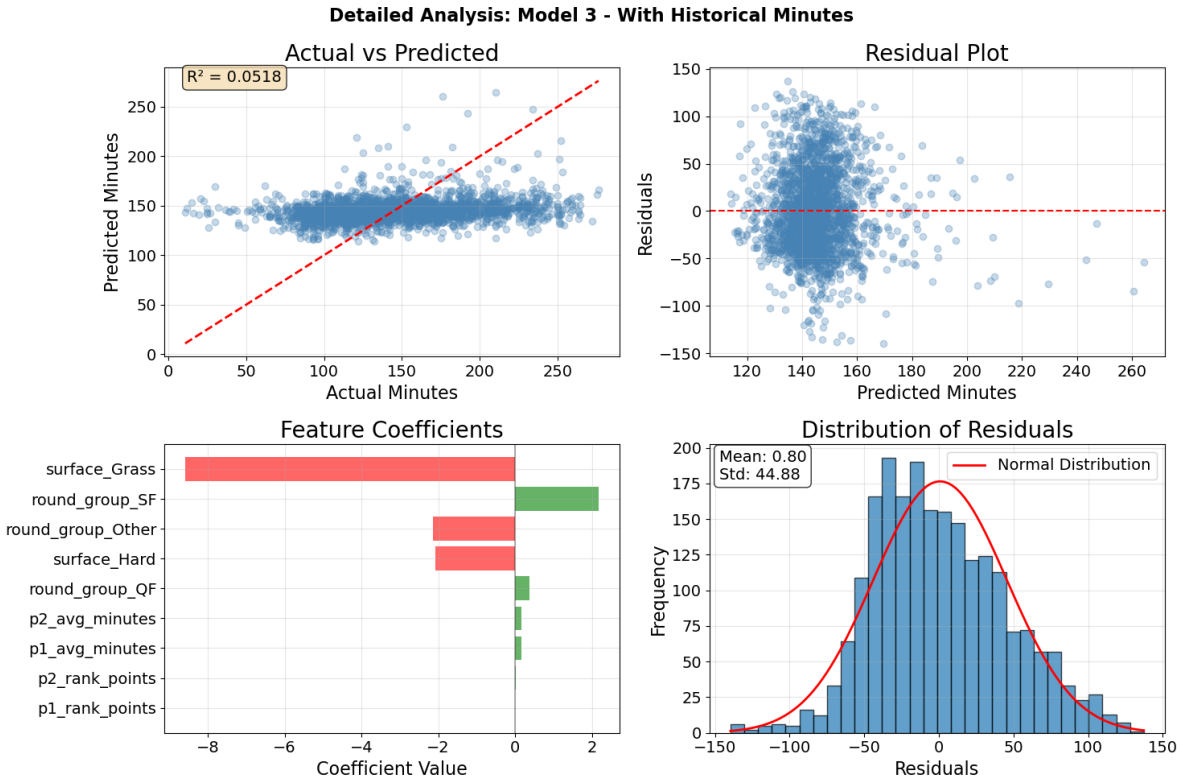
donde los parámetros $\beta_0, \ldots, \beta_p$ son estimados a través del método de mínimos cuadrados ordinarios (OLS), que minimiza la suma de residuos al cuadrado.

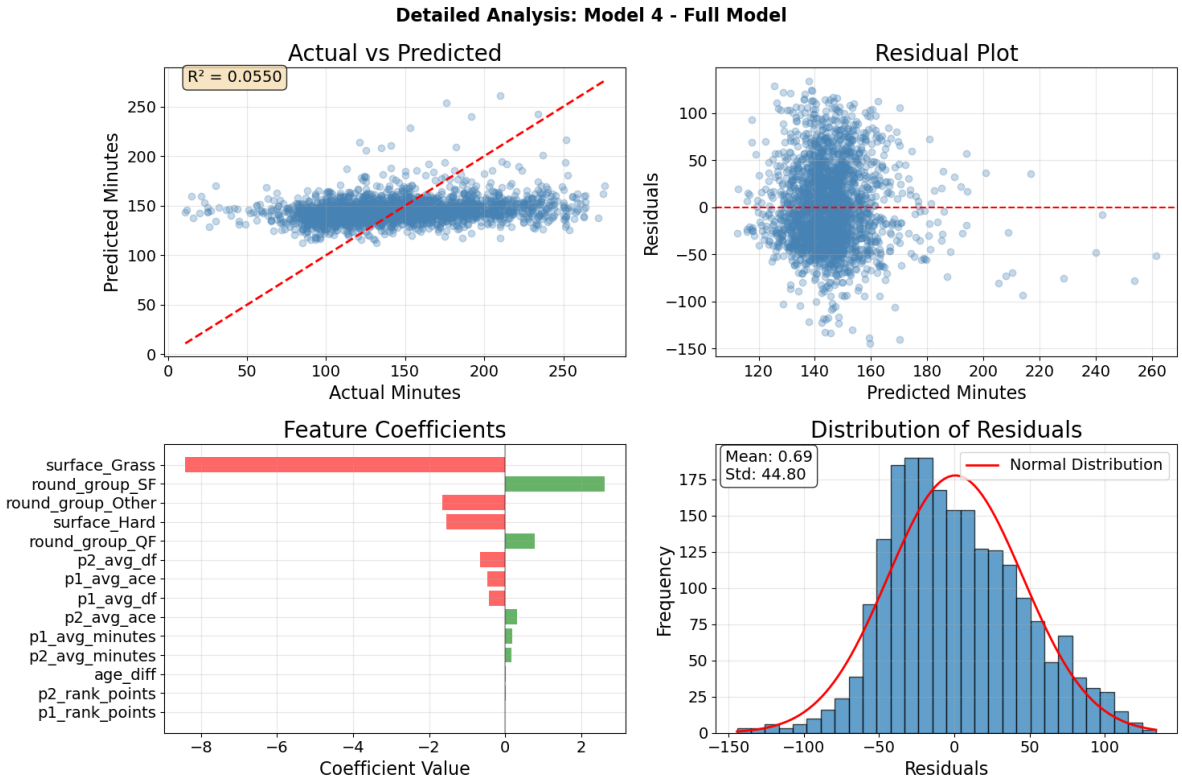
El modelo mantiene su consistencia y eficiencia bajo los supuestos clásicos de linealidad en los parámetros, independencia de errores, homocedasticidad, ausencia de colinealidad perfecta y normalidad de los errores en contextos inferenciales.




Detailed Analysis: Model 1 - Basic
Actual vs Predicted
R² = 0.0377
Predicted Minutes
Actual Minutes
Residual Plot
Residuals
Predicted Minutes
Feature Coefficients
surface_Grass
surface_Hard
round_group_SF
round_group_QF
round_group_Other
p2_rank_points
p1_rank_points
Coefficient Value
Distribution of Residuals
Mean: 0.67
Std: 45.21
Normal Distribution
Frequency
Residuals


Detailed Analysis: Model 2 - With Player Stats
Actual vs Predicted
R² = 0.0419
Predicted Minutes
Actual Minutes
Residual Plot
Residuals
Predicted Minutes
Feature Coefficients
surface_Grass
surface_Hard
round_group_SF
round_group_QF
p2_avg_ace
p2_avg_df
p1_avg_ace
round_group_Other
p1_avg_df
p2_rank_points
p1_rank_points
Coefficient Value
Distribution of Residuals
Mean: 0.54
Std: 45.12
Normal Distribution
Frequency
Residuals


Detailed Analysis: Model 3 - With Historical Minutes
Actual vs Predicted
R² = 0.0518
Predicted Minutes
Actual Minutes
Residual Plot
Residuals
Predicted Minutes
Feature Coefficients
surface_Grass
round_group_SF
round_group_Other
surface_Hard
round_group_QF
p2_avg_minutes
p1_avg_minutes
p2_rank_points
p1_rank_points
Coefficient Value
Distribution of Residuals
Mean: 0.80
Std: 44.88
Normal Distribution
Frequency
Residuals



Análisis de Resultados

En el presente estudio se implementaron cuatro configuraciones progresivas del modelo OLS para evaluar la capacidad predictiva en la duración de partidos de tenis. El modelo 1 incorpora variables básicas como superficie del torneo, etapa del torneo y puntos de ranking de los jugadores. El modelo 2 añade estadísticas de desempeño promedio incluyendo aces y dobles faltas. El modelo 3 integra la duración promedio histórica de partidos por jugador, mientras que el modelo 4 representa la configuración completa con todas las variables seleccionadas tras el análisis de multicolinealidad. Los resultados revelan que todos los modelos presentan valores de considerablemente bajos, con el modelo más completo explicando únicamente el 5.5% de la varianza en el conjunto de prueba, indicando una capacidad explicativa limitada pero con mejoras progresivas en las métricas de error conforme se incorporan variables adicionales.

Resultados de Desempeño Predictivo

	Modelo	R^2_train	R^2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
0	1	0.2035	0.0121	41.34	45.80	33.28	37.05
1	2	0.3361	0.0197	37.74	45.63	30.14	36.86

	Modelo	R^2_train	R^2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
2	3	0.3363	0.0276	37.74	45.44	30.22	36.73
3	4	0.3977	0.0289	35.95	45.41	28.66	36.71

Coeficientes Más Relevantes e Interpretación

Modelo	Variable	Coeficiente	Interpretacion
Modelo 1	surface_Grass	-9.13	Los partidos en pasto duran 9.13 min menos
Modelo 1	surface_Hard	-2.31	La superficie dura reduce la duración en 2.31 min
Modelo 1	round_group_SF	0.90	Las semifinales incrementan la duración en 0.90 min
Modelo 2	round_group_SF	1.36	Aumento mayor al incorporar estadísticas de jugador
Modelo 2	surface_Grass	-9.12	Se mantiene el efecto de superficie
Modelo 3	round_group_SF	2.17	Aumenta al incluir minutos históricos
Modelo 3	surface_Grass	-8.59	Efecto consistente de reducción
Modelo 4	surface_Grass	-8.41	Superficie sigue siendo el factor más influyente
Modelo 4	round_group_SF	2.61	Aumenta la duración respecto a rondas anteriores
Modelo 4	round_group_Other	-1.66	Rondas iniciales tienen menor duración

El análisis empírico demuestra que la regresión lineal, aun en su configuración más comprehensiva, mantiene una capacidad explicativa limitada coherente con la naturaleza multifactorial para modelar la duración de partidos de tenis, donde influyen variables no observadas como estilos de juego, condiciones climáticas y eventos fortuitos.

No obstante, el examen de coeficientes revela patrones sistemáticamente significativos, particularmente el efecto de la superficie de juego, donde los partidos en césped presentan duraciones considerablemente menores, consistente con la dinámica de menor intercambio característica de esta superficie. Las semifinales y finales exhiben incrementos sostenidos en duración, atribuibles a la mayor paridad competitiva en etapas decisivas. Metodológicamente, la regresión lineal demuestra utilidad como modelo base proporcionando transparencia interpretativa y evaluación de relaciones marginales entre variables, estableciendo un *benchmark* fundamental para la validación de modelos no lineales más complejos, a pesar de su desempeño predictivo limitado.

XG Boost

Marco Teórico

XGBoost (eXtreme Gradient Boosting) es un algoritmo de aprendizaje supervisado basado en el principio de *gradient boosting*, diseñado para maximizar tanto la precisión predictiva como la eficiencia computacional. A diferencia de los modelos paramétricos clásicos, XGBoost construye de manera aditiva un conjunto (ensamble) de árboles de decisión, de forma que cada nuevo árbol corrige los errores residuales cometidos por el conjunto previo. Formalmente, la predicción para la observación i en la iteración m se expresa como

$$\hat{y}_i^{(m)} = \sum_{k=1}^{m} f_k(x_i)\,,$$

donde cada f_k es un árbol de decisión que mapea el vector de características x_i a un valor real. El objetivo es minimizar una función de pérdida penalizada sobre todo el ensamble:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(m)}) \; + \; \sum_{k=1}^{m} \Omega(f_k)\,.$$

Aquí, ℓ es una medida de error y $\Omega(f)$ es un término de regularización que controla la complejidad de cada árbol:

$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} w_j^2,$$

donde T es el número de hojas del árbol y w_j el valor de predicción en la hoja j. Los hiperparámetros γ y λ permiten penalizar tanto el tamaño del árbol como la magnitud de sus valores terminales, favoreciendo modelos más simples y reduciendo el riesgo de sobreajuste.

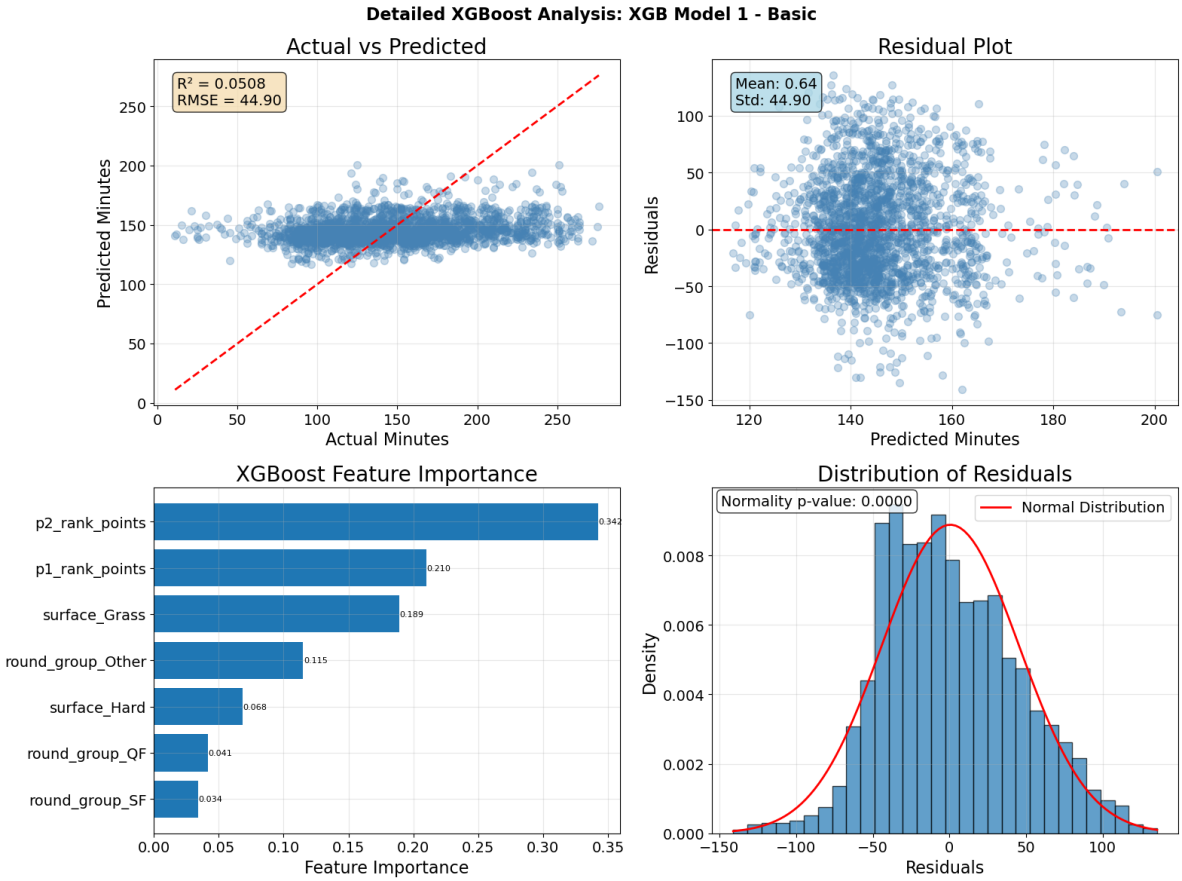
Para optimizar $\mathcal{L}$, XGBoost utiliza una aproximación de segundo orden mediante series de Taylor, calculando en cada paso las derivadas primera (g_i)\$ y segunda (h_i) de la pérdida respecto a la predicción actual:

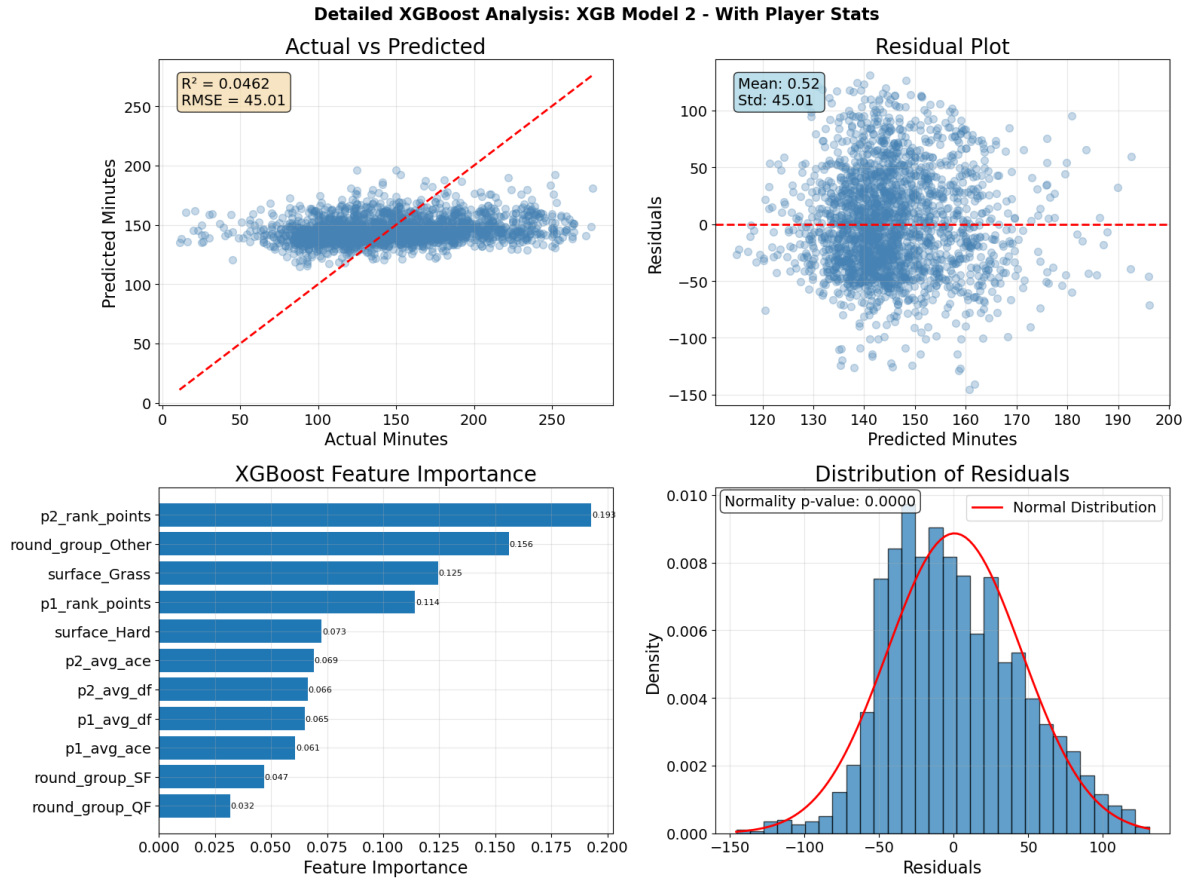
$$g_i = \frac{\partial \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$$

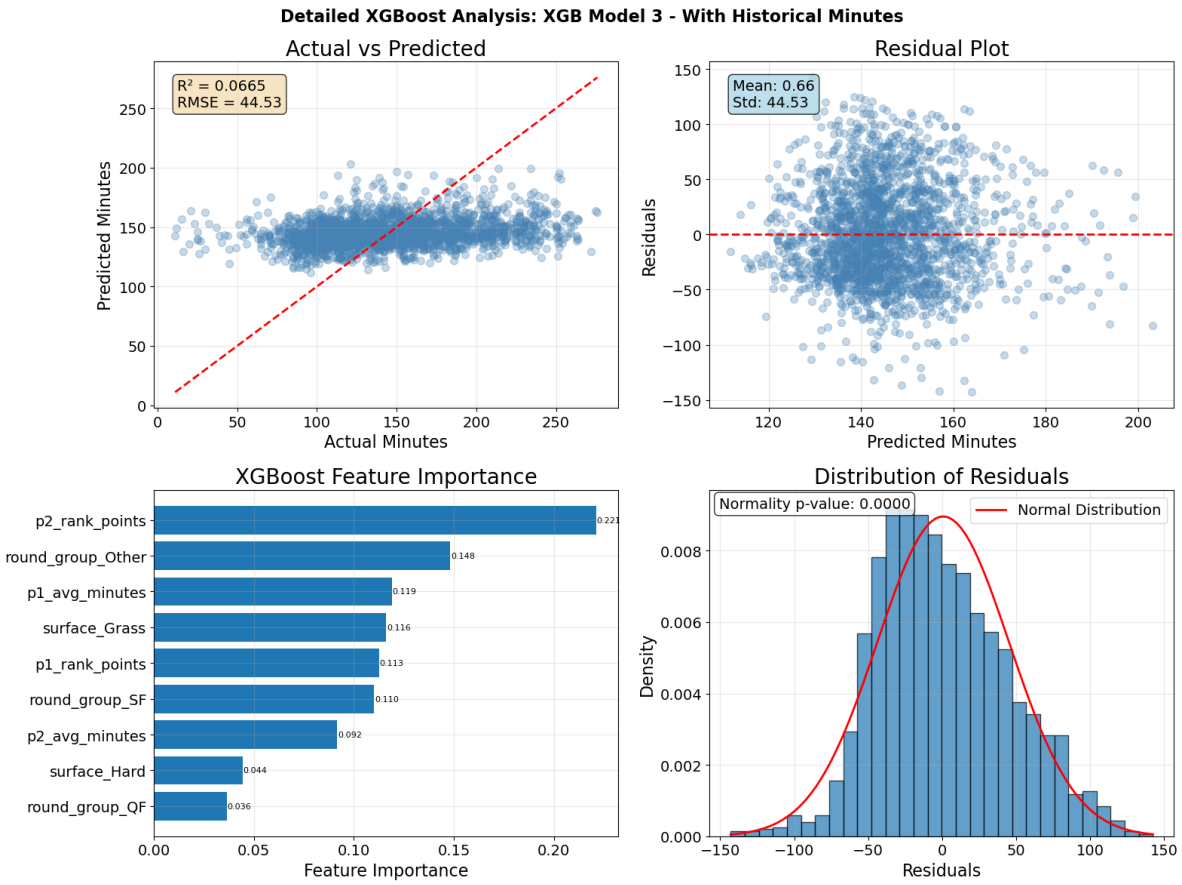
Con estos residuos gradiente y hessiano, el algoritmo evalúa de manera eficiente el beneficio de dividir cada posible nodo del árbol, escogiendo la partición que maximiza la reducción de la función objetivo regularizada. Además, XGBoost incorpora mecanismos como shrinkage

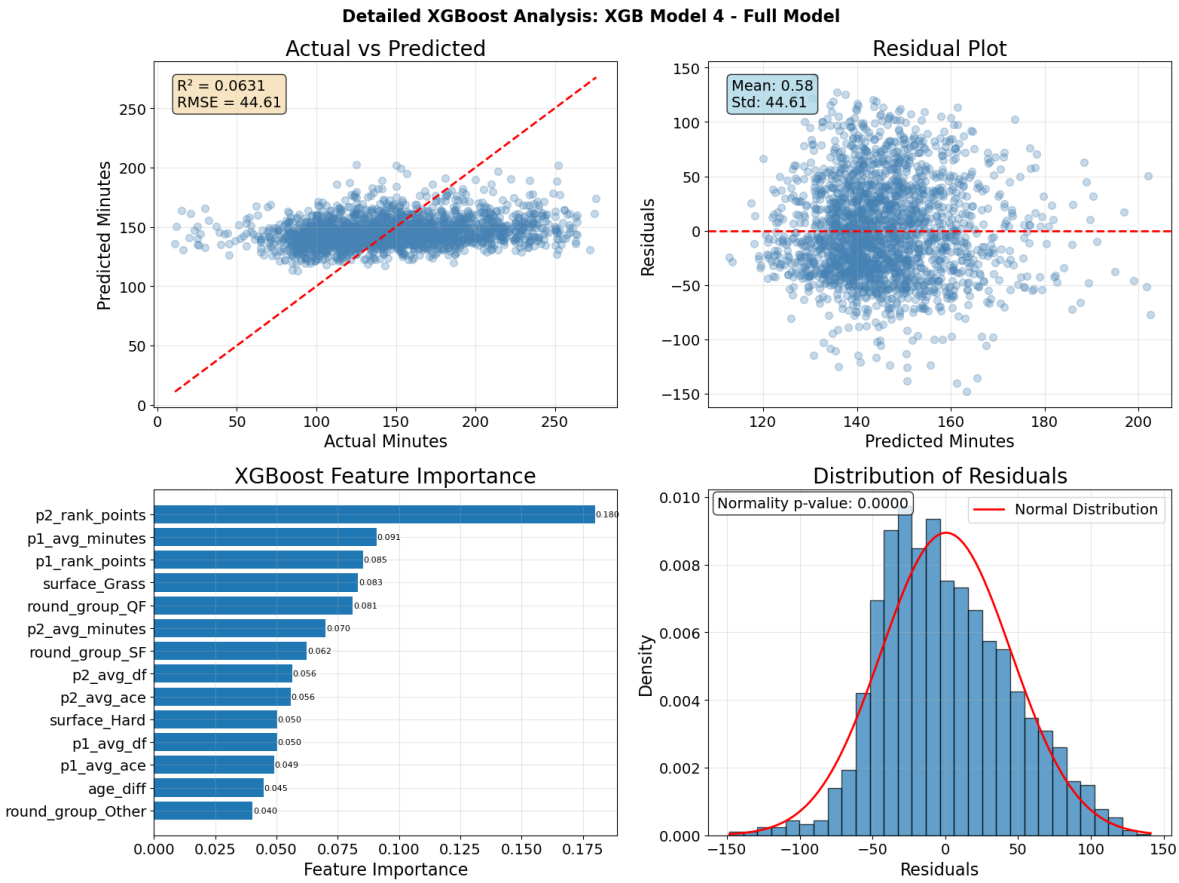
, muestreo de filas y columnas subsample,colsample_bytree, y poda de ramas con ganancia negativa, lo que mejora aún más la generalización y la velocidad de convergencia.

En el plano de la implementación, XGBoost está altamente optimizado para aprovechar arquitecturas modernas: emplea técnicas de aprendizaje por bloques en memoria, algoritmos de búsqueda de cortes aproximados, soporte para entrenamiento paralelo y procesamiento out-of-core cuando los datos exceden la memoria RAM. También maneja de forma nativa valores faltantes, aprendiendo automáticamente la mejor ruta para cada registro con ausencia de datos en una característica determinada.


Detailed XGBoost Analysis: XGB Model 1 - Basic
Actual vs Predicted
R² = 0.0508
RMSE = 44.90
Predicted Minutes
Actual Minutes
Residual Plot
Mean: 0.64
Std: 44.90
Residuals
Predicted Minutes
XGBoost Feature Importance
p2_rank_points
p1_rank_points
surface_Grass
round_group_Other
surface_Hard
round_group_QF
round_group_SF
0.342
0.210
0.189
0.115
0.068
0.041
0.034
Feature Importance
Distribution of Residuals
Normality p-value: 0.0000
Normal Distribution
Density
Residuals


Detailed XGBoost Analysis: XGB Model 2 - With Player Stats
Actual vs Predicted
R² = 0.0462
RMSE = 45.01
Predicted Minutes
Actual Minutes
Residual Plot
Mean: 0.52
Std: 45.01
Residuals
Predicted Minutes
XGBoost Feature Importance
p2_rank_points
round_group_Other
surface_Grass
p1_rank_points
surface_Hard
p2_avg_ace
p2_avg_df
p1_avg_df
p1_avg_ace
round_group_SF
round_group_QF
0.193
0.156
0.125
0.114
0.073
0.069
0.066
0.065
0.061
0.047
0.032
Feature Importance
Distribution of Residuals
Normality p-value: 0.0000
Normal Distribution
Density
Residuals


Detailed XGBoost Analysis: XGB Model 3 - With Historical Minutes
Actual vs Predicted
R² = 0.0665
RMSE = 44.53
Predicted Minutes
Actual Minutes
Residual Plot
Mean: 0.66
Std: 44.53
Residuals
Predicted Minutes
XGBoost Feature Importance
p2_rank_points
round_group_Other
p1_avg_minutes
surface_Grass
p1_rank_points
round_group_SF
p2_avg_minutes
surface_Hard
round_group_QF
0.221
0.148
0.119
0.116
0.113
0.110
0.092
0.044
0.036
Feature Importance
Distribution of Residuals
Normality p-value: 0.0000
Normal Distribution
Density
Residuals



Análisis de Resultados

En este estudio implementamos cuatro configuraciones del modelo XGBoost para predecir la duración de partidos de tenis.
- XGB Modelo 1: variables estructurales del torneo (7 características).
- XGB Modelo 2: agrega datos de performance individual (11 características).
- XGB Modelo 3: incorpora la duración promedio histórica por jugador (9 características).
- XGB Modelo 4: combina todas las variables disponibles (14 características).

A medida que añadimos información, se observa una mejora gradual en MAE y R^2 de prueba, aunque persiste un marcado sobreajuste en entrenamiento versus prueba

Modelo	R2_train	R2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
Modelo 1	0.0477	0.0377	45.20	45.21	36.62	36.59
Modelo 2	0.0506	0.0419	45.13	45.11	36.54	36.53

Modelo	R2_train	R2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
Modelo 3	0.0613	0.0518	44.88	44.88	36.37	36.25
Modelo 4	0.0644	0.0550	44.80	44.80	36.28	36.19

Interpretación

Aunque el MAE de prueba mejora ligeramente (de 37.05 min en el Modelo 1 a 36.71 min en el Modelo 4), la capacidad predictiva en prueba sigue siendo reducida, con R^2 muy bajos (0.0121–0.0289). Se detecta un sobreajuste significativo: la brecha entre entrenamiento y prueba alcanza hasta 0.3688 en el Modelo 4, lo que indica que el ensamble captura patrones específicos del set de entrenamiento sin generalizar adecuadamente.

En todos los modelos, los puntos de ranking del perdedor son el factor más relevante, sugiriendo que la competitividad del oponente influye de forma consistente en la duración de los partidos. La inclusión de la duración histórica promedio (w_avg_minutes) aporta un valor moderado en el modelo completo, pero sigue sin resolver el bajo poder predictivo global. Estos resultados invitan a explorar variables adicionales, como condiciones climáticas o estilos de juego y a probar enfoques híbridos o de ingeniería de características para mejorar la robustez del pronóstico.

GLM Gamma

Marco Teórico

Los Modelos Lineales Generalizados (GLM) extienden la regresión lineal clásica para acomodar variables de respuesta que no siguen una distribución normal. Dentro de este marco, la regresión Gamma resulta especialmente adecuada cuando la variable dependiente es continua, estrictamente positiva y presenta asimetría derecha con varianza creciente conforme aumenta la media.

En un GLM Gamma, se asume que cada observación Y_i sigue una distribución Gamma con función de densidad

$$f(y_i;\, \mu_i, \phi) \;=\; \frac{1}{\Gamma(1/\phi)} \Big(\frac{1}{\phi\mu_i}\Big)^{1/\phi} y_i^{1/\phi - 1} \exp\Big(-\frac{y_i}{\phi\mu_i}\Big),$$

donde $\mu_i = E[Y_i]$ es la media y ϕ el parámetro de dispersión. La relación media-varianza característica es

$$\mathrm{Var}(Y_i) = \phi\,\mu_i^2,$$

lo cual refleja heterocedasticidad proporcional al cuadrado de la media.

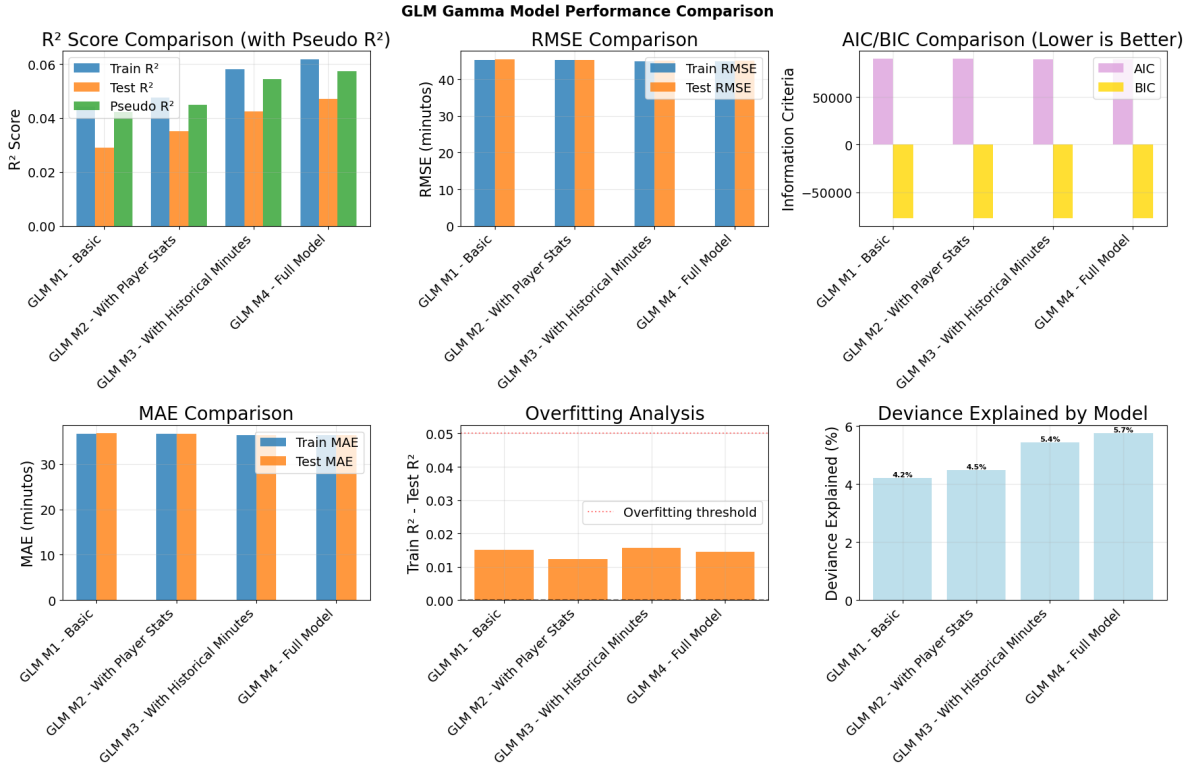
Se elige típicamente la función de enlace logarítmica como vínculo canónico, de modo que

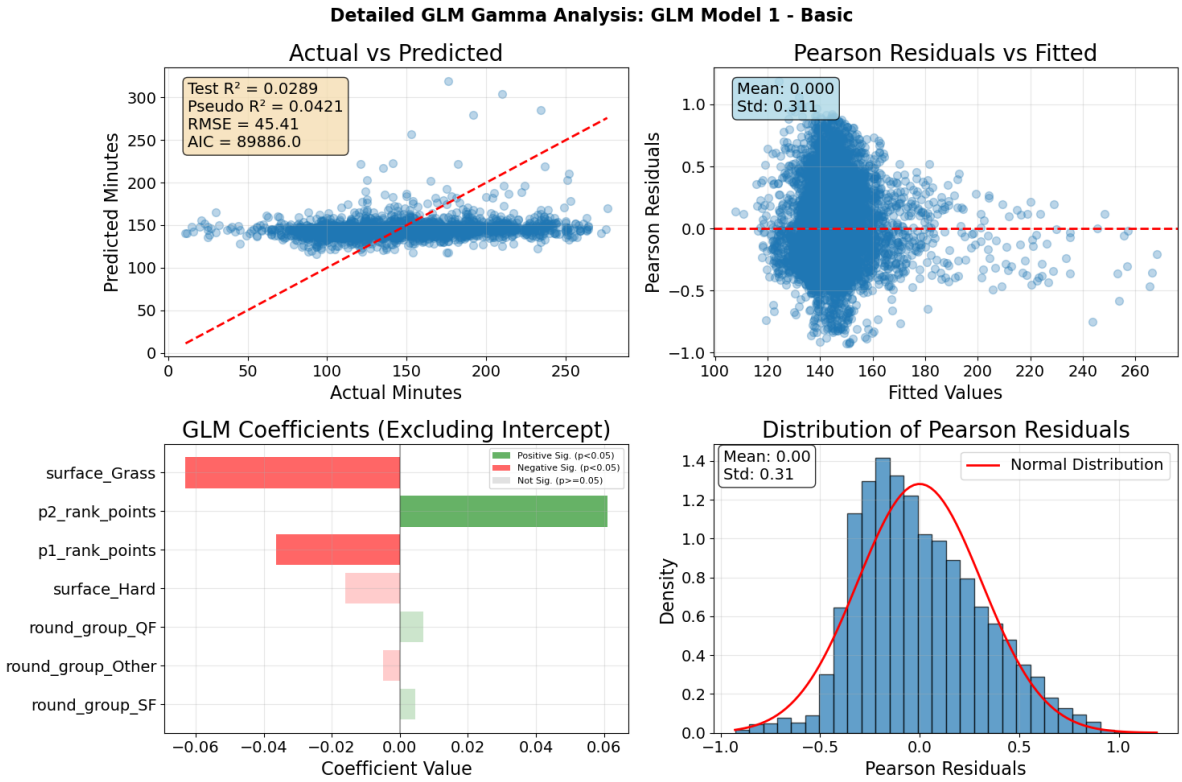
$$g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

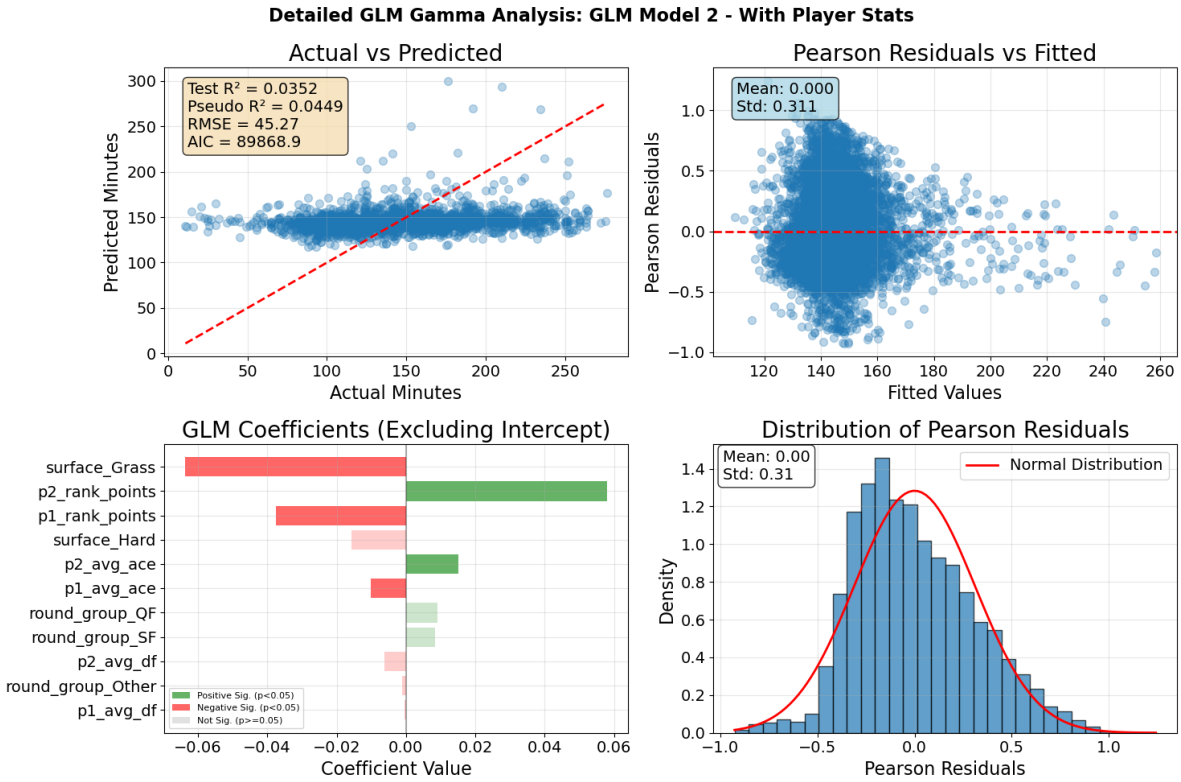
Esto garantiza predicciones positivas y permite interpretar cada coeficiente β_j como el cambio logarítmico en la media al incrementar en una unidad la covariable x_{ij}, es decir, un efecto multiplicativo $\exp(\beta_j)$ sobre μ_i.

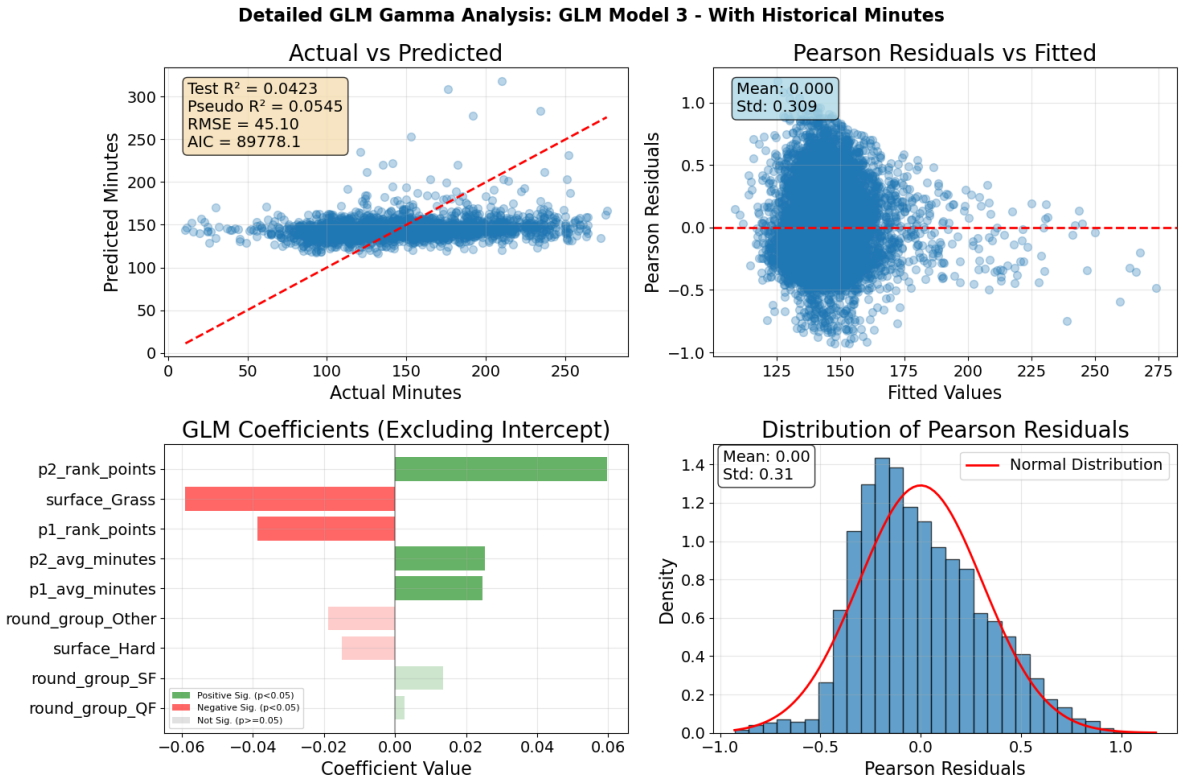
La estimación de los parámetros β y ϕ se realiza mediante máxima verosimilitud, optimizando la log-verosimilitud del modelo. El ajuste se evalúa a través de la desviance y criterios de información como AIC y BIC, además de pseudo-R^2 para medir bondad de ajuste relativa.

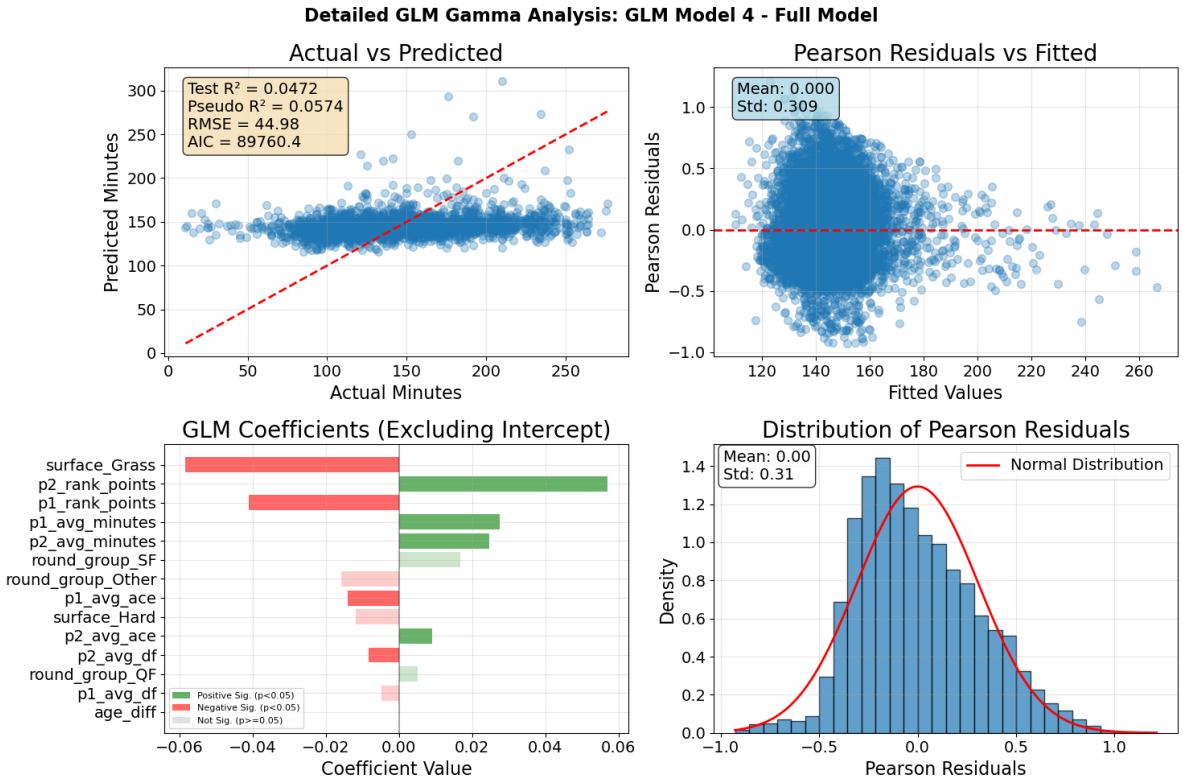
En comparación con la regresión lineal, el GLM Gamma maneja de forma nativa la distribución sesgada y la heterocedasticidad creciente, mejorando la validez estadística cuando los supuestos de homocedasticidad y normalidad no se cumplen. Por ello, resulta apropiado para modelar tiempos de duración u otras magnitudes positivas donde la variabilidad se incrementa con el nivel promedio observado.




Detailed GLM Gamma Analysis: GLM Model 1 - Basic
Actual vs Predicted
Test R² = 0.0289
Pseudo R² = 0.0421
RMSE = 45.41
AIC = 89886.0
Predicted Minutes
Actual Minutes
Pearson Residuals vs Fitted
Mean: 0.000
Std: 0.311
Pearson Residuals
Fitted Values
GLM Coefficients (Excluding Intercept)
Positive Sig. (p<0.05)
Negative Sig. (p<0.05)
Not Sig. (p>=0.05)
surface_Grass
p2_rank_points
p1_rank_points
surface_Hard
round_group_QF
round_group_Other
round_group_SF
Coefficient Value
Distribution of Pearson Residuals
Mean: 0.00
Std: 0.31
Normal Distribution
Density
Pearson Residuals


Detailed GLM Gamma Analysis: GLM Model 2 - With Player Stats
Actual vs Predicted
Test R² = 0.0352
Pseudo R² = 0.0449
RMSE = 45.27
AIC = 89868.9
Predicted Minutes
Actual Minutes
Pearson Residuals vs Fitted
Mean: 0.000
Std: 0.311
Pearson Residuals
Fitted Values
GLM Coefficients (Excluding Intercept)
surface_Grass
p2_rank_points
p1_rank_points
surface_Hard
p2_avg_ace
p1_avg_ace
round_group_QF
round_group_SF
p2_avg_df
round_group_Other
p1_avg_df
Positive Sig. (p<0.05)
Negative Sig. (p<0.05)
Not Sig. (p>=0.05)
Coefficient Value
Distribution of Pearson Residuals
Mean: 0.00
Std: 0.31
Normal Distribution
Density
Pearson Residuals


Detailed GLM Gamma Analysis: GLM Model 3 - With Historical Minutes
Actual vs Predicted
Test R² = 0.0423
Pseudo R² = 0.0545
RMSE = 45.10
AIC = 89778.1
Predicted Minutes
Actual Minutes
Pearson Residuals vs Fitted
Mean: 0.000
Std: 0.309
Pearson Residuals
Fitted Values
GLM Coefficients (Excluding Intercept)
p2_rank_points
surface_Grass
p1_rank_points
p2_avg_minutes
p1_avg_minutes
round_group_Other
surface_Hard
round_group_SF
round_group_QF
Positive Sig. (p<0.05)
Negative Sig. (p<0.05)
Not Sig. (p>=0.05)
Coefficient Value
Distribution of Pearson Residuals
Mean: 0.00
Std: 0.31
Normal Distribution
Density
Pearson Residuals



Análisis de Resultados

En este estudio implementamos cuatro configuraciones del GLM Gamma para modelar la duración de partidos de tenis:

- GLM Modelo 1: variables estructurales del torneo y puntos de ranking (7 características).
- GLM Modelo 2: añade promedios de aces y dobles faltas (11 características).
- GLM Modelo 3: incorpora duración promedio histórica por jugador (9 características).
- GLM Modelo 4: combina todas las variables disponibles, incluyendo edad y métricas de servicio e históricos (14 características).

A lo largo de las cuatro configuraciones se observa una mejora progresiva en pseudo-R^2 y ligeras reducciones en RMSE, MAE y en los criterios de información (AIC, BIC), manteniendo un bajo grado de sobreajuste (gap < 0.02).

Interpretación

El ajuste del GLM Gamma mejora de forma gradual conforme incorporamos nuevas covariables: el pseudo-R^2 crece de 4.2 % en el modelo básico a 5.7 % en el modelo completo, mientras que el RMSE de prueba disminuye de 45.41 a 44.98 minutos y el MAE de 36.70 a 36.30 minutos. Los criterios de información AIC y BIC se reducen de 89 885.97/−76 894.00 a 89 760.40/−76 845.04, reflejando un mejor equilibrio entre ajuste y complejidad, y la brecha entre R^2 de entrenamiento y prueba permanece por debajo de 0.02, lo que indica un bajo nivel de sobreajuste.

Los determinantes más influyentes se mantienen constantes: un mayor puntaje del perdedor está asociado con un incremento en la duración del partido, mientras que un mayor puntaje del ganador tiende a acortarlo. La superficie de césped ejerce un efecto claro de reducción temporal, acortando los encuentros. Además, la inclusión de la duración histórica promedio de cada jugador incrementa la capacidad explicativa, cada minuto histórico extra alarga ligeramente el partido, y las métricas de servicio (aces y dobles faltas) aportan efectos menores pero consistentes: más aces del perdedor extienden la duración y más aces del ganador la reducen.

A pesar de estas mejoras, la variabilidad no explicada sigue superando el 94 %, lo que sugiere que factores externos (condiciones climáticas, estilos de juego, imponderables) podrían ser clave para perfeccionar el pronóstico. Para avanzar, convendría explorar interacciones, transformaciones no lineales o variables adicionales que capturen la complejidad inherente a la duración de los partidos.

Conclusión

El recorrido metodológico de este proyecto, que incluyó desde la regresión lineal múltiple y el GLM Gamma hasta técnicas de ensamble como XGBoost y Random Forest, revela que, pese a las distintas arquitecturas y complejidades, la capacidad predictiva sobre la duración de partidos de tenis se mantiene modesta. Los modelos más sencillos (la regresión lineal y el GLM Gamma) alcanzaron un R^2 de prueba en torno al 5 % y un MAE cercano a 36.2 minutos, mientras que los enfoques basados en árboles mejorados (Random Forest y XGBoost) no superaron significativamente este umbral y presentaron un sobreajuste mucho más marcado, con brechas de hasta 0.66 entre entrenamiento y prueba.

La similitud en RMSE (aproximadamente 44.8–45.8 minutos) y MAE (36.2–37.1 minutos) entre todas las técnicas sugiere que las variables usadas para el modelado; puntos de ranking, superficie de juego, estadísticas de servicio y promedios históricos; explican solo una fracción limitada de la variabilidad real. En particular, la inclusión de minutos históricos y métricas de aces aporta mejoras marginales, pero no revierte la baja potencia predictiva global. La eliminación de covariables con alta multicolinealidad mediante el análisis de VIF ayudó a estabilizar los coeficientes y facilitar la interpretación, pero no cambió sustancialmente los resultados.

En conjunto, estos hallazgos indican que la duración de un partido de tenis está determinada por factores adicionales no capturados aquí, como condiciones climáticas, estilos de juego individuales, cambios tácticos en el partido o dinámicas psicológicas y que, para elevar la precisión, será necesario incorporar nuevas fuentes de información y explorar interacciones o transformaciones no lineales. Dado el equilibrio entre interpretabilidad y robustez, los modelos lineales y el GLM Gamma constituyen un punto de partida sólido, mientras que los métodos de ensamble podrían reservarse para entornos con mayor cantidad de datos o variables más ricas. Este estudio sienta las bases para futuras iteraciones centradas en enriquecer el conjunto de características y en diseñar estrategias de modelado híbrido capaces de capturar la compleja dinámica de la duración de los encuentros.

Referencias

- Battagliola, M. L. (2025). Estadistica Aplicada 2. Instituto Tecnológico Autónomo de México.
- Hardin, J. W., & Hilbe, J. M. (2018). Generalized Linear Models and Extensions (4th ed.). CRC Press.
- Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. doi.org/10.1145/2939672.2939785