Universidad Bernardo O'Higgins

Facultad de Ingeniería, Ciencia y Tecnología

Análisis de Datos | Sección: TEO 1

Profesor Eliecer Peña Ancavil

Colinealidad y desempeño de modelos

3 de noviembre de 2025

Aldo Hernández
aldo.hernandezt@uanl.edu.mx
Universidad Autónoma de Nuevo León
San Nicolás de los Garza, Nuevo León, MX

1. Introducción

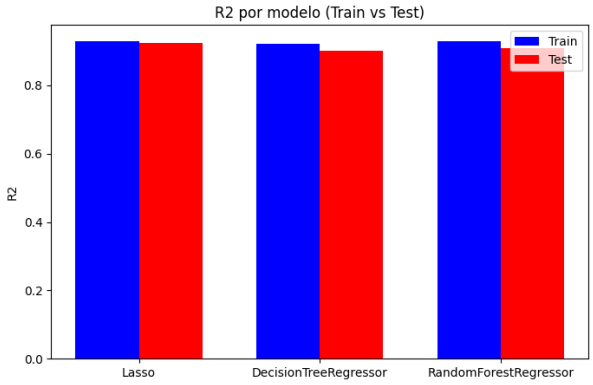
Este trabajo busca comparar y analizar tres modelos distintos entrenados con un conjunto de datos acerca del videojuego *League of Legends* que predigan de forma precisa la cantidad de oro del equipo azul. Se toma como base el análisis realizado con anterioridad sobre el mismo conjunto pero con menos columnas, concretamente sin las columnas **blueTeamTotalKills** y **blueTeamTotalDamageToChamps**.

Los modelos entrenados fueron un árbol de decisión, un bosque aleatorio y una regresión Lasso. Antes de entrenar los modelos, se tomaron en cuenta las siguientes consideraciones:

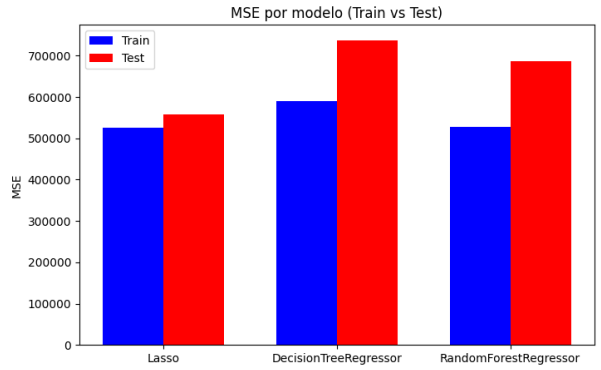
- La mejor profundidad (relativamente) del árbol se encontró mediante *GridSearchCV*, dando como resultado una profundidad de 8 niveles.
- Los datos fueron escalados antes de entrenar el modelo de regresión Lasso.
- Para el bosque aleatorio se usaron 300 árboles con profundidad máxima de 10 niveles, con un mínimo de 5 muestras para dividir y un mínimo de 4 hojas por nodo.

2. Comparación de modelos

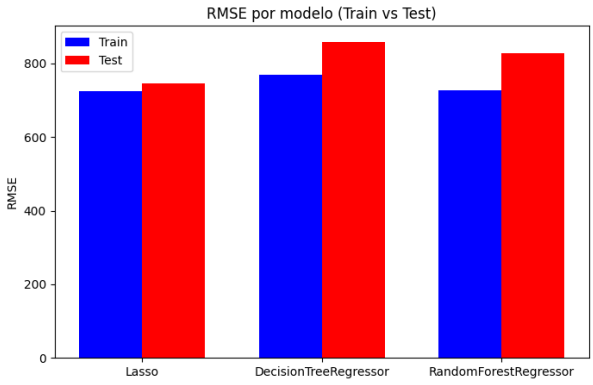
En la figura 1 se observa el rendimiento de cada modelo en distintas métricas. En la subfigura 1a se puede notar un valor cercano a 0,9 para R^2 en los tres modelos tanto en el conjunto de entrenamiento como de prueba; por otro lado, en las subfiguras 1b, 1c y 1d se puede ver un mismo comportamiento: los tres modelos logran un desempeño parecido en el conjunto de entrenamiento pero los modelos basados en árboles aumentan su error significativamente en los conjuntos de prueba, indicando un posible pequeño sobreajuste en dichos modelos (ligeramente mayor en el árbol de regresión).



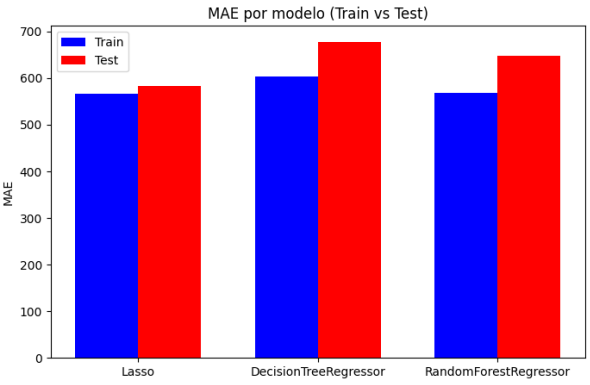
(a) Valor de la métrica R^2 en cada conjunto de datos.



(b) Valor de la métrica MSE en cada conjunto de datos.



(c) Valor de la métrica RMSE en cada conjunto de datos.



(d) Valor de la métrica MAE en cada conjunto de datos.

Figura 1: Valor de las métricas de desempeño en los tres modelos en los datos de entrenamiento y validación.

3. Análisis de modelos

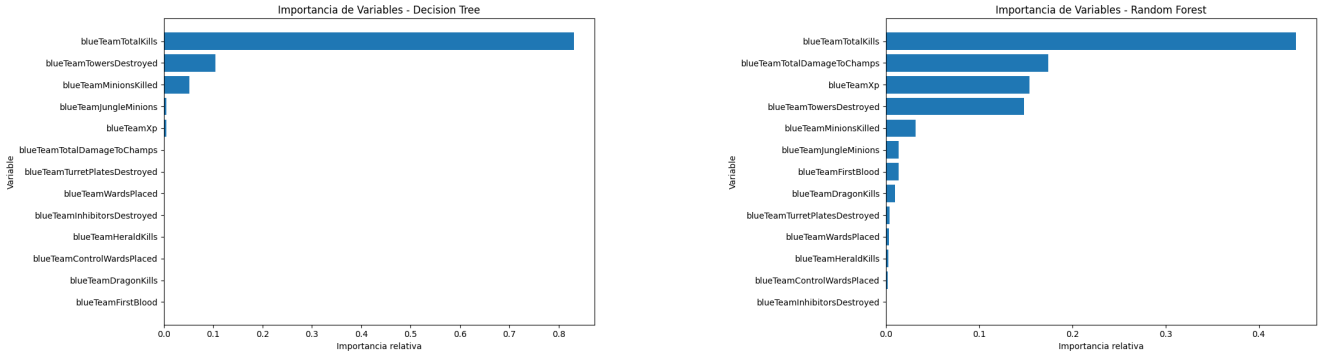
Además de las métricas, se puede rescatar más información valiosa de los tres modelos que pueda mostrar qué tanto impacto tiene una variable en la predicción final; en el caso de los modelos basados en árboles existe la *importancia* de las variables, mientras que en la regresión están los valores de los *coeficientes*.

3.1. Importancia de variables

Observando la figura 2 se puede notar rápidamente un gran problema: más de la mitad de las variables han resultado insignificantes para la predicción final. Sin embargo, esto no puede ser posible ya que **todas las variables otorgan oro al equipo azul**, por lo que su impacto no puede ser prácticamente cero.

Se puede notar que tanto en el árbol de decisión (subfigura 2a) como en el bosque aleatorio (subfigura 2b) las variables que se han decidido incluir justamente en este trabajo se encuentran en los primeros puestos. Estas dos variables han alterado completamente la capacidad explicativa de los modelos: ahora no es posible saber el verdadero efecto individual de cada variable en la cantidad de oro del equipo azul.

Es importante mencionar que el bosque aleatorio consiguió evitar en cierta medida este sesgo inducido debido a su naturaleza de generar árboles menos correlacionados entre sí, mejorando la capacidad de generalización.



(a) Importancia de las variables en el árbol de decisión.

(b) Importancia de las variables en el bosque aleatorio.

Figura 2: Importancia de las variables en los modelos basados en árboles.

3.2. Coeficientes Lasso

En la figura 3 podemos observar que pasa lo mismo en el modelo de regresión Lasso que en los otros dos modelos: determinadas variables **toman el control** de la predicción final. Sin embargo, este modelo en particular logra una mejor generalización que los otros dos debido a su capacidad de *suavizar* los coeficientes, evitando en cierta medida el control absoluto de una sola variable.

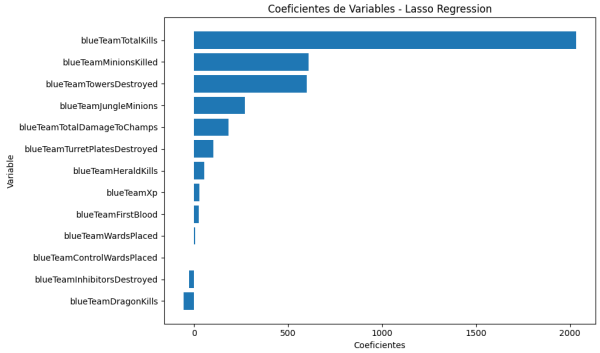


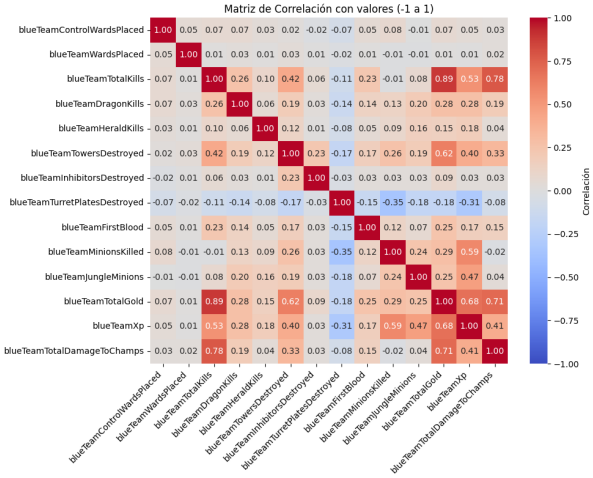
Figura 3: Coeficientes de la regresión Lasso.

4. Interpretación de resultados

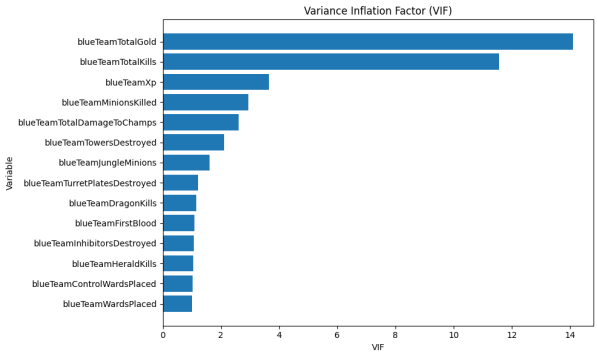
A pesar de que estos tres modelos ofrecen en cierta parte mejores capacidades predictivas de acuerdo a las métricas, pierden casi toda su explicabilidad debido a que no se sabe el impacto de cada variable en la variable objetivo.

Para encontrar la razón detrás de este problema, es fundamental interpretar la figura 4. En la matriz de correlación (subfigura 4a) se puede notar que ambas variables "nuevas"tienen una **alta correlación** con otras variables, lo cual tiene sentido bajo el funcionamiento del juego ya que, por ejemplo, conseguir *kills* implica hacer más daño a campeones enemigos.

Sin embargo, ese no es el único problema. Observando la subfigura 4b en la que se encuentra el VIF de cada variable podemos notar algo interesante: **el verdadero problema es la variable *blueTeamTotalKills*** y no particularmente ambas como se especulaba al inicio, esto tiene sentido ya que esta variable induce una correlación muy alta con la otra variable añadida **blueTeamTotalDamageToChamps**, siendo que esta última no tiene una correlación significativamente alta con otras variables; además, se puede observar que otra variable da indicios de *multicolinealidad*: **blueTeamXp**.



(a) Matriz de correlación.



(b) Factor de inflación en la varianza.

Figura 4: Indicadores de colinealidad y multicolinealidad.

Todas estas relaciones comprobadas analíticamente tienen sentido dentro del juego:

- Más *kills* te da más libertad en el mapa del juego, otorgando mayor control sobre el equipo enemigo y facilitando la toma de decisiones (implicando conseguir más objetivos como dragones, más experiencia, destruir más torres, etcétera).
- Se puede obtener más experiencia (XP) matando esbirros, destruyendo torres y consiguiendo objetivos.
- El daño a campeones enemigos no está tan fuertemente correlacionado con las demás variables ya que su valor no tiene un impacto garantizado en ellas (por ejemplo, el tener alto daño no garantiza que consigas más torres o mates más esbirros).

Finalmente, podemos concluir que estos tres modelos tienen un mejor desempeño predictivo que los anteriores debido a que prácticamente **hacen trampa** gracias a la multicolinealidad existente en este conjunto de datos: tienen variables que contienen información de otras, por lo que le estamos diciendo a los modelos *en qué se deben fijar*.

Una solución a este problema sería **eliminar las dos columnas** *blueTeamTotalKills* y *blueTeamXp* del conjunto de datos, esto no representa problema alguno porque gran parte de la información que proporcionan se encuentra distribuida en las demás variables; esta eliminación nos permitiría conocer el verdadero impacto de cada una de las demás variables en la cantidad de oro total del equipo azul.