Aprenentatge Automàtic

Practica: Predición del Mundial de Qatar 2022

Bachelor of Technology
in
COMPUTER SCIENCE AND ENGINEERING

Ignasi Fibla Figuerola (ignasi.fibla)
Mark Smithson Rivas (mark.smithson)

DEPARTMENT OF COMPUTER SCIENCE
Universitat Politècnica de Catalunya
Gener, 2023

Índice

1. Introdución

El Mundial de Fútbol es un torneo internacional de fútbol que se celebra cada cuatro años y es organizado por la Federación Internacional de Fútbol Asociación (FIFA). Es considerado el evento deportivo más importante del mundo y cuenta con la participación de las selecciones nacionales de fútbol de todos los países afiliados a la FIFA.

El primer Mundial de Fútbol se celebró en 1930 en Uruguay y desde entonces se ha realizado cada cuatro años, excepto durante la Segunda Guerra Mundial, cuando no se disputó el torneo de 1942 y 1946. Actualmente, el Mundial de Fútbol consta de una fase de grupos y una fase eliminatoria, y culmina con la final, que determina al campeón del mundo. La selección que resulte campeona del Mundial de Fútbol recibe la Copa del Mundo, un trofeo que se entrega al ganador del torneo.

Este año se celebrará en Qatar y será la 22ª edición de este importante evento deportivo. En este trabajo, mediante el uso del aprendizaje automático, se analizarán diversos factores que pueden influir en el rendimiento de las selecciones participantes y se realizarán predicciones sobre quiénes podrían ser los favoritos para ganar el torneo.

Nuestros modelos de aprendizaje automático, van a ajustar una recta de regresión que nos va a permitir simular y predecir los resultados de todos los encuentros del Mundial. Un partido de futbol tiene muchos resultados posibles, pero se podrían resumir en Victoria, que puntúa 3 puntos, Derrota, que no aporta puntos, o Empate, que aporta un punto a cada equipo. Como el Mundial consta de una fase de grupos con un enfrentamiento por pareja de equipos, se ha visto que la posibilidad de que dos equipos empaten a puntos en esta fase es bastante elevada, por eso, se ha decidido tomar como variables objetivos los goles marcados por ambos equipos, una decisión, que nos permitirá mantener la diferencia de goles y así, clasificar a los equipos correctamente.

2. Conjunto de datos

El conjunto de datos que se utilizará en este trabajo ha sido descargado de la plataforma [Kaggle.com], una comunidad en línea de científicos de datos que comparten y colaboran en proyectos de análisis de datos. Además, es una de las principales plataformas de recursos de datos en la industria, y ofrece una amplia gama de datos públicos y privados. Por lo tanto, elegir descargar el conjunto de datos de Kaggle.com es una opción lógica para obtener datos de alta calidad y relevantes para nuestro trabajo.

El [Conjunto de datos] es conocido como [FIFA World Cup Qatar 2022] y en los meses previos al evento ha tenido bastantes actividades fallidas de otros analistas de datos. La mayoría de proyectos, se han quedado a medio camino de predecir los resultados, haciendo solo un pequeño precocinado de los datos, que, en muchas ocasiones, era erróneo. Si entramos en la descripción del conjunto de datos se puede ver como se actualizó por última vez el 28 de agosto de 2022.

Al cargar por primera vez el conjunto de datos podemos ver como cuenta con un total de 23921 entradas, en las cuales existen valores perdidos, anómalos e incoherentes. Después de analizar y corregir estas entradas se ha logrado obtener un conjunto de datos con 4303 entradas. Se ha visto, que las columnas más afectadas por valores perdidos son las puntuaciones de los distintos equipos, un valor que creemos importante para predecir el resultado de los partidos. Por este motivo, se ha decidido eliminar esas entradas del conjunto de datos, antes que aplicar otros métodos, como el uso de la mediana muestral.

A continuación, se ha hecho limpieza del conjunto de datos, eliminando aquellas columnas que no tendrían relevancia ni sentido al predecir las variables objetivos. Nuestros datos tienen una variable temporal, que contiene la fecha en la que se disputó el partido, esta columna no tendrá un desempeño relevante en las predicciones de los partidos. Nuestros modelos simularán analíticamente el resultado de un partido sin importar variables temporales, debido a que esta componente añadiría complejidad innecesaria al problema. Las variables categóricas, que almacenan los nombres de los países tampoco tendrán demasiada relevancia en el modelo final, debido a que para muchas selecciones esta es su primera participación en un Mundial. Otras columnas que se ha creído oportuno eliminar son tournament, city, country y neutral_location, debido a que para este mundial tendrán todas el mismo valor y, al eliminarlas, sacamos ruido del modelo. Para finalizar, se ha eliminado la columna home_team_result la cual indica de manera categórica el resultado, diciendo si el resultado para el equipo local equivale a una Victoria, a un Empate o a una Derrota.

En este punto, se ha hecho gran parte del trabajo de la fase de precocinado de los datos, se han solucionado los problemas relacionados con valores perdidos o incoherentes, se han elimando aquellas variables poco útiles para la predicción de los resultados y ahora queda la estandarización de los datos y convertir

las variables categóricas en formato de texto a variables categóricas en formato numérico.

De ahora en adelante, empezaremos haciendo un estudio de los datos, empezando por las correlaciones y finalizando las variancias de las distintas variables del conjunto de datos. Solo como recordatorio, si la correlación entre dos variables es cercana a -1 indica que existe una correlación lineal negativa entre ambas, por contra, si el valor es cercano a 1 quiere decir que existe una correlación lineal entre ambas y, por último, si es cercano a 0 indica que no existe una correlación entre las variables. En el conjunto de datos con el que vamos a entrenar los distintos modelos, hemos podido crear la matriz de correlación que se muestra en la Figura 1.

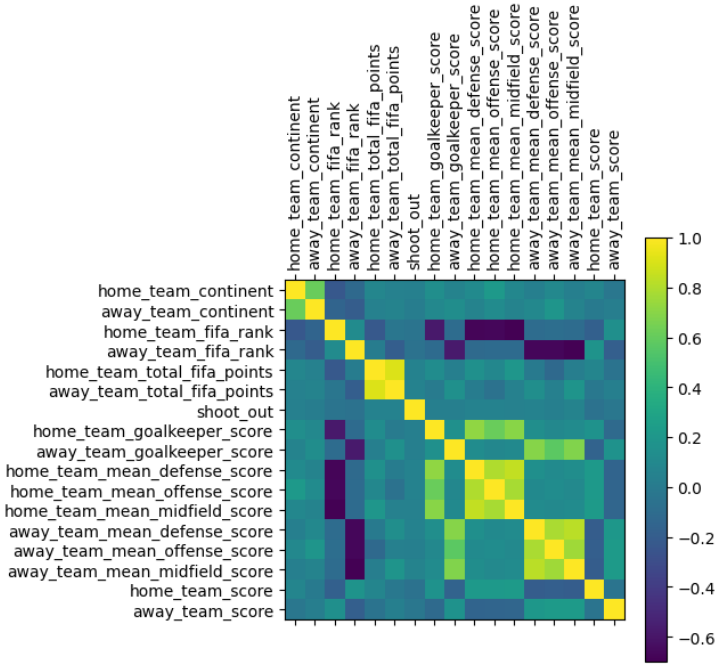

Figura 1: Matriz de correlaciones del conjunto de datos

Se puede ver en la figura anterior como no existen relaciones lineales muy destacables en el conjunto de datos y, que en la gran mayoría el valor de correlación se encuentra cercano a cero. Las únicas variables entre las que existe una correlación lineal son las puntuaciones otorgadas por la FIFA en las distintas líneas de juego con la puntuación final del equipo. Después de este breve estudio se ha querido comparar ambas variables objetivas con los datos del conjunto, buscando la existencia de alguna relación lineal. Este estudio se encuentra en las Figuras 2 y 3.

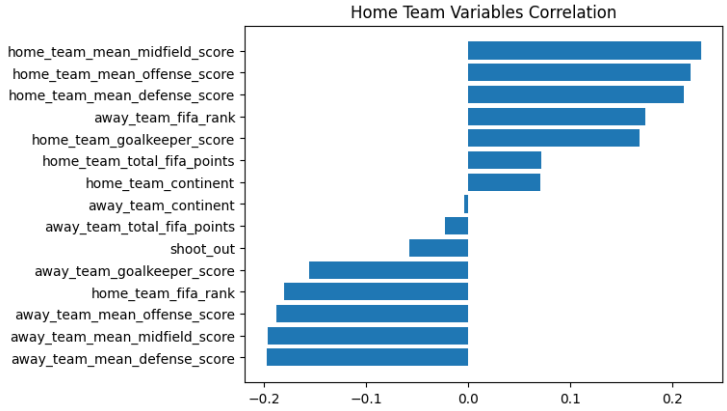

Figura 2: Correlaciones de la variable home_team_score

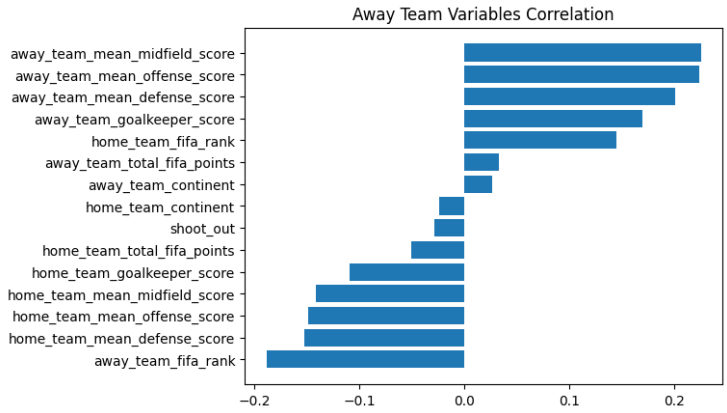

Figura 3: Correlaciones de la variable away_team_score

Como se venía comentando en la matriz de correlaciones, no existe ninguna pareja de variables que destaque por una correlación lineal positiva o negativa, y menos que una des estas variables sea del conjunto de variables objetivos. Esto se podría dar por muchos factores, pero, después de muchas horas de tratamiento de los datos, creemos que la componente de aleatoriedad de los datos va a tener más repercusión en los resultados finales de lo esperado.

2.1. Modificación de la Variable Objetivo

Con el objetivo de modificar las variancias de las variables objetivo y, intentar adquirir mejores resultados, se ha creído que operar con el valor de la diferencia de las variables objetivos podría ayudar a los modelos a obtener un

mejor resultado. Al calcular la variable resultado mediante la resta de las variables home_team_score y away_team_score, podemos afirmar que un valor positivo en la variable resultado equivale a decir que el equipo local ha salido victorioso del partido, por otra parte, un resultado negativo equivale a la victoria del equipo visitante. Por último, un resultado equivalente a cero, significa empate en el partido. Este estudio ha empezado analizando la matriz de correlación, que se encuentra en la Figura 4

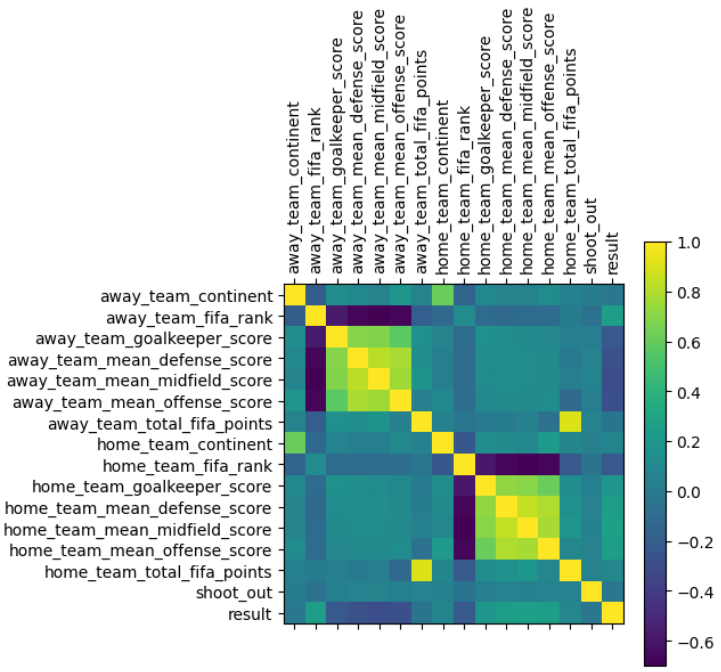

Figura 4: Matriz de correlaciones del conjunto de datos con una variable objetivo

Aparentemente, las correlaciones de las variables no se han modificado demasiado, aunque para obtener unos resultados exhaustivos se compararan las Figuras 2, 3 y 5.

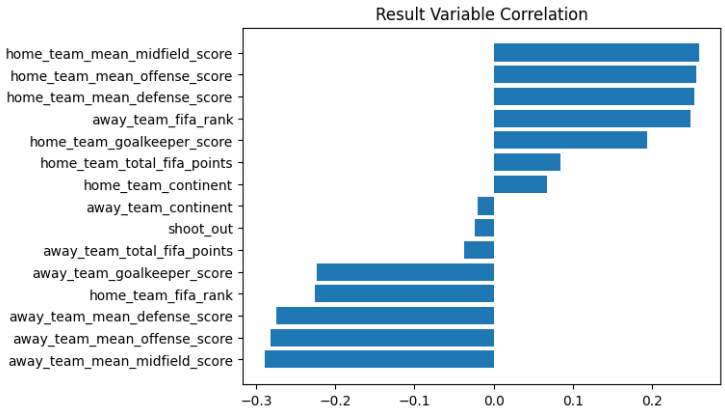

Figura 5: Correlaciones de la variable resultado

De la gráfica anterior se puede ver la existencia de una correlación positiva entre la variable resultado y todos los datos relacionados con el equipo local. Esta correlación se puede dar a raíz de que una victoria del equipo local equivale a un resultado positivo. Por otra parte, se puede ver como el intervalo de valores de las correlaciones se ha mantenido en relación a las correlaciones del conjunto de datos con dos variables objetivo.

2.2. Reducción de la dimensionalidad

Como se ha podido observar, nuestro conjunto de datos no destaca por la existencia de componentes con alta correlación con las variables objetivo, un hecho que dificultará la obtención de resultados destacables. Mediante la reducción de dimensionalidad y el Análisis de Componentes Principales (PCA) se buscará una combinación lineal de las variables originales que capture la mayor cantidad de variabilidad en los datos posible. Esto se logra a través del uso de componentes principales.

Para no implementar manualmente el Análisis de Componentes Principales se ha optado por usar una funcionalidad de la librería de [Scikit-learn]. Se empezará el estudio con el análisis de los autovalores, también conocidos como eigenvalues. Para realizar el estudio se ha generado la representación gráfica de la cantidad de varianza explicada por cada componente principal. Esta representación, también conocida como Scree Plot se puede visualizar en la Figura 6.

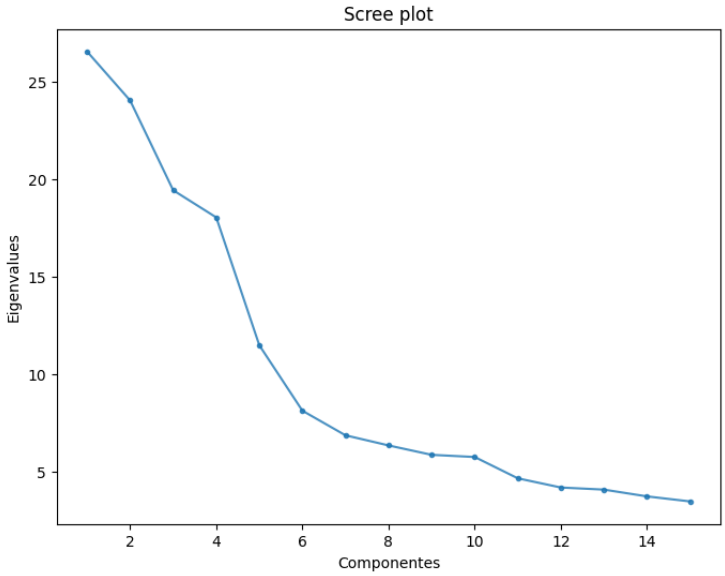

Figura 6: Análisis de Componentes Principales: Scree Plot

A simple vista se puede evaluar la bondad de ajuste del modelo de PCA. Podemos ver como la mayor parte de la varianza de los datos originales se explica en las primeras componentes, lo que nos permite deducir que el modelo de PCA tiene un buen ajuste. Además, si entramos un poco más en detalle, podemos observar como el número óptimo de componentes principales es 8, ya que en ese punto, encontramos una estabilización de la gráfica, hecho que nos permite afirmar, que usando más de 8 componentes principales estamos perdiendo varianza explicada. Estas conclusiones se han podido comprobar con el gráfico de la Figura 7. Se puede ver como a partir de la tercera componente, la varianza explicada se mantiene y, por lo tanto, añadir más componentes nos generara ruido en el resultado final.

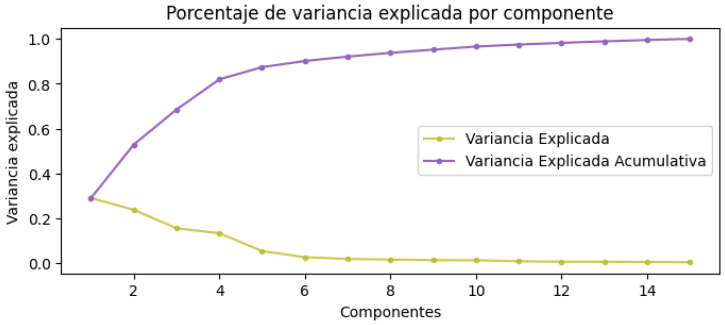

Figura 7: Análisis de Componentes Principales: Varianza Explicada

El estudio ha finalizado buscando patrones con 2 y 3 componentes. Para realizar el estudio se han generado las Figuras 8 y 9. De ambas visualizaciones podemos sacar las mismas conclusiones, no existen patrones en los datos que nos permitan ajustar fácilmente una recta de regresión, un hecho que va en sintonía con nuestras hipótesis, será altamente complicado obtener resultados satisfactorios debido a la gran variabilidad de los datos.

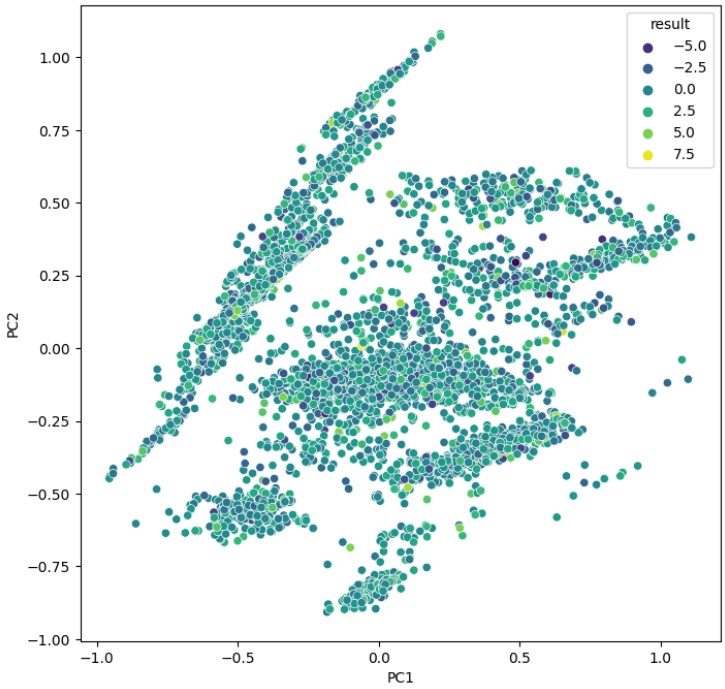

Figura 8: Análisis de Componentes Principales: Dispersión 2 Componentes

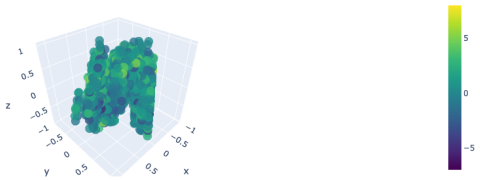

Figura 9: Análisis de Componentes Principales: Dispersión 3 Componentes

3. Modelos lineales

Los modelos lineales son un tipo de modelo matemático que utilizan una función lineal para describir la relación entre dos o más variables. Estos modelos son ampliamente utilizados en muchas áreas, como la economía, la ingeniería, la física y la psicología, debido a su simplicidad y facilidad de interpretación. Además, son útiles para hacer predicciones sobre el comportamiento de un sistema en el futuro.

Un modelo lineal se compone de una ecuación matemática que describe la relación entre las variables independientes y la variable dependiente. Es importante tener en cuenta que los modelos lineales tienen algunas limitaciones, como la supuesta linealidad de la relación entre las variables y la supuesta independencia y homogeneidad de los errores. Estas limitaciones deben tenerse en cuenta al utilizar este tipo de modelos.

3.1. Linear Regression

El [Sklearn Linear Regression] es un modelo lineal que permite ajustar a un conjunto de datos y utilizarlo para hacer predicciones. Este modelo se basa en la minimización de la suma de los cuadrados de los residuos entre las predicciones del modelo y los datos observados, y se puede utilizar tanto para resolver problemas de regresión como de clasificación binaria. En este modelo no se pueden modificar los parametros, por lo que la experimentación es bastante sencilla. Se ha entrenado el modelo con los datos estandarizados y con la resta de las variables objetivo, hecho que nos ha permitido obtener los siguientes resultados.

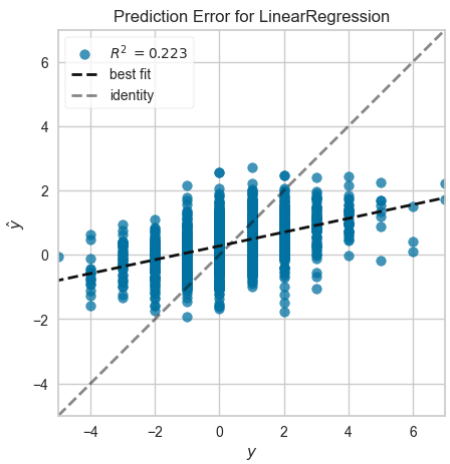

Figura 10: Linear Regression: Prediccion del error

En la Figura 10 podemos ver como los resultados obtenidos no son dema-

siado satisfactorios. Esta conclusion se puede obtener tanto graficamente como analiticamente, consultando el valor R^2 que equivale a 0,223.

Además, se realizará con todos los modelos es el calculo del tiempo medio de entrenamiento, en este caso, el modelo de regresión lineal tiene un tiempo de entrenamiento medio de 0,013 segundos. Por otra parte, al ejecutar la Validación cruzada al modelo entrenado se han obtenido, de media, un valor de 0,216. Si nos fijamos en los valores obtenidos por cada uno de los pliegues que se ha realizado al conjunto de datos encontramos en el Cuadro 1.

Pliegue	1	2	3	4	5
Resultado	0.25609482	0.1231758	0.25783003	0.16857474	0.27594974

Cuadro 1: Linear Regression: Resultados Validación Cruzada

3.2. Ridge Regression

La [Sklearn Ridge] es una variante de la [Sklearn Linear Regression] que se utiliza para controlar el sobreajuste. A diferencia de la Regresión Lineal, que solo minimiza la suma de los cuadrados de los residuos entre las predicciones del modelo y los datos observados, la Ridge Regression también agrega una penalidad por la magnitud de los coeficientes del modelo. Esta penalidad se controla mediante el parámetro alpha, que determina qué tanto se penalizan los coeficientes.

La Ridge Regression es útil en situaciones en las que el conjunto de datos tiene una alta dimensionalidad y existe el riesgo de que el modelo se sobreajuste. Al penalizar los coeficientes del modelo, la Ridge Regression puede ayudar a evitar el sobreajuste y mejorar la generalización del modelo a datos nuevos.

En este caso, se ha decidido empezar la experimentación encontrando el valor óptimo para el parámetro alpha. Se ha entrenado el modelo por una lista finita de valores comprendidos entre 0, que anula los pesos y por lo tanto es equivalente a la Linear Regression, y 10. Después de comprovar los resultados de la validación cruzada y el valor de R^2, se ha podido observar como el resultado es el mismo por todos los valores possibles de alpha, hecho que nos lleva a pensar que no existe sobreajuste.

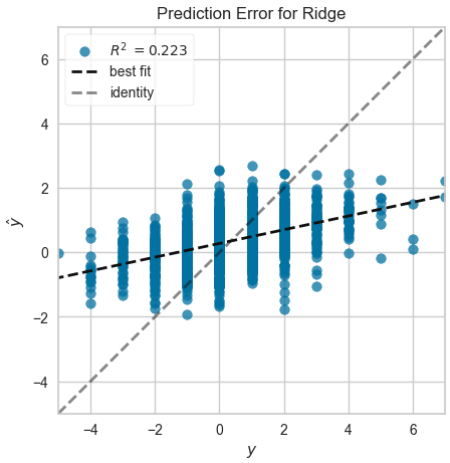

Figura 11: Ridge Regression: Prediccion del error

De la misma forma que en el modelo de Regresión Lineal, en la Figura 11 podemos ver como los resultados obtenidos no son demasiado satisfactorios. El valor de R^2, es el mismo que en la Regresión Lineal 0,223, pero el valor de la validación cruzada ha mejorado levemente. En este caso, la media de las validaciónes cruzada es 0,218. Y si queremos fijarnos en los valores individuales de cada validación, lo podemos encontrar en el Cuadro 2.

Pliegue	1	2	3	4	5
Resultado	0.25630708	0.12472019	0.25798172	0.17091906	0.28008048

Cuadro 2: Ridge Regression: Resultados Validación Cruzada

Para acabar, se ha visto que en este caso, el tiempo de entrenamiento es inferior al de la regression lineal, obteniendo un valor cercano a 0,002 segundos de media.

3.3. Lasso Regression

3.4. KNeighbors Regressor

	R^2	**Validación Cruzada**	**Tiempo de Entrenamiento**
Linear Regression	0.223	0.216	0.013
Ridge	0.223	0.218	0.002
Lasso	-0.000	-0.003	
KNeighbors Regressor	0.202	0.181	

Cuadro 3: Resultados Modelos Lineales

Referencias

[FIFA World Cup Qatar 2022] FIFA World Cup Qatar 2022. (s. f.). Recuperado 8 de enero de 2023, de https://www.fifa.com/fifaplus/en/tournaments/mens/worldcup/qatar2022

[Kaggle.com] Kaggle: Your Machine Learning and Data Science Community. (s. f.). Recuperado 7 de diciembre de 2022, de https://www.kaggle.com

[Conjunto de datos] FIFA World Cup 2022. (2022, 28 agosto). Kaggle. Recuperado 4 de diciembre de 2022, de https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022

[Scikit-learn] Scikit-learn: Machine learning in Python — Scikit-learn 1.2.0 documentation. (s. f.). Recuperado 8 de enero de 2023, de https://scikit-learn.org/stable/index.html

[Sklearn Decomposition PCA] Sklearn Decomposition PCA. (s. f.). scikit-learn. Recuperado 11 de diciembre de 2022, de https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[Sklearn Linear Regression] Sklearn Linear Regression. (s. f.). scikit-learn. Recuperado 16 de diciembre de 2022, de https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[Sklearn Ridge] Sklearn Ridge. (s. f.). scikit-learn. Recuperado 16 de diciembre de 2022, de https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html