Aplicación de ML para la estimación del porcentaje de grasa corporal

Santiago Rivera Montoya
Leider Felipe Caicedo
Emanuel López Higuita
Universidad de Antioquia, Colombia
santiago.riveram@udea.edu.co
felipe.caicedo@udea.edu.co
emanuel.lopezh@udea.edu.co

Resumen

Objetivo. En este informe se abordó el problema de regresión Bodyfat el cual busca predecir el porcentaje de grasa corporal para una muestra de 252 hombres. **Metodología.** Se empleó una validación cruzada con K-fold y se utilizaron el MSE (error cuadrático medio) y R^2 (coeficiente de determinación) como métricas de desempeño. Se implementaron tres técnicas predictivas para el dataset de BodyFat: regresión lineal, regresión polinomial y KNN. **Análisis y Resultados.** Los resultados obtenidos fueron los siguientes: Para la regresión lineal, se logró un MSE de 1.96 y un R^2 de 0.97. En cuanto a la regresión polinomial, se obtuvo un MSE de 2.02 y un R^2 de 0.956. Para el método KNN, el MSE del mejor modelo fue de 26.00 y el R^2 de 0.66. **Conclusión.** Tras analizar cada valor de las métricas de evaluación, se llegó a la conclusión de que el modelo de regresión lineal es óptimo para este conjunto de datos. Además, se destacó la idea de que una mayor complejidad en los modelos no necesariamente se traduce en mejores resultados, como se evidenció en este informe.

Palabras clave: grasa corporal; aprendizaje automático; regresión lineal; regresión polinomial; KNN; KFold; MSE.

I. Introducción

La estimación de grasa corporal es un componente importante en la evaluación de la salud y el bienestar físico de una persona, sin embargo, poder obtener mediciones sobre este parámetro que sean precisas puede ser difícil y costoso.

El uso de técnicas de aprendizaje automático ofrece una solución eficiente y accesible para estimar el porcentaje de grasa corporal sin los inconvenientes asociados con métodos tradicionales.

En este proyecto, se explora cómo aplicar estas técnicas para desarrollar sistemas de predicción precisos y útiles en la evaluación de la composición corporal.

II. Metodología

Se usará la base de datos BodyFat [1], la cual busca resolver un problema de regresión relacionado a estimaciones del porcentaje de grasa corporal determinadas mediante pesaje subacuático y diversas mediciones de la circunferencia corporal de 252 hombres. Entre las columnas está la edad, el peso, diámetro de cintura, de cuello, de brazos, etc. Contando con un total de 252 filas y 14 características sin contar la variable objetivo(class).

Para la validación se usará K-fold cross, el cuál contará con una relación entre el porcentaje de entrenamiento y validación de 80-20% respectivamente. Este algoritmo inicialmente mezcla de forma aleatoria los datos e inicializa el parámetro k, que determina el número de particiones de los datos y las iteraciones de entrenamiento y validación. Durante cada una de las k iteraciones, se oculta una partición y se entrena el modelo con las k-1 particiones restantes. Se evalúa el desempeño del modelo tanto en el conjunto de entrenamiento como en el conjunto de validación. Después de completar las iteraciones, se promedian las medidas de desempeño para obtener el desempeño final del modelo.

Para la implementación de los modelos se implementarán las siguientes técnicas de regresión:

- **Regresión Lineal:**

Esta técnica predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal [2].

- **Regresión Polinomial:**

Esta técnica es utilizada para modelar la relación entre una o muchas variables independientes (o predictor) y una variable dependiente (o respuesta) que no es lineal; utiliza una función polinomial para realizar el ajuste [3].

- **KNN (K-Nearest Neighbors):**

Es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual [5].

III. Métricas de Evaluación

Cómo métricas de evaluación [4] se usarán las siguientes:

Mean Squared Error (MSE): Mide el error cuadrado promedio de nuestras predicciones. Para cada punto, calcula la diferencia cuadrada entre las predicciones y el objetivo y luego promedia esos valores.

R^2: Evalúa la capacidad del modelo para explicar la variabilidad total de los datos.

Se tomará como mejor o mejores modelos los que tengan la mejor combinación de ambos parámetros.

IV. Análisis y Resultados

Regresión Lineal

En una primera parte se entrenó el modelo sin utilizar validación cruzada para ver cómo se comportaba sin realizar iteraciones sobre los datos, para lo cual arrojó un $MSE = 0.380$ y un $R^2 = 0.991$, lo que indica que el modelo de regresión lineal es óptimo para este conjunto de datos, ya que como se puede observar en el EDA, todas las características siguen el mismo patrón lineal

Para la segunda parte con la validación cruzada se decidió utilizar un k = 5 y se evaluó el error en cada uno de los fold, para luego obtener el promedio de todas las iteraciones y se registró en la siguiente tabla (Tabla 1).

Tabla 1. Número de folds con su respectivo error (MSE) y R^2

Fold	MSE	R^2
1	1.85	0.98
2	6.45	0.90
3	0.21	0.99
4	0.60	0.99
5	0.71	0.99

Se obtuvo así un promedio del $MSE = 1.96$, y un promedio del $R^2 = 0.97$, lo que sugiere que el modelo tiene una muy buena capacidad para predecir y una buena capacidad para explicar la variabilidad de los datos. Este valor indica que, en promedio, los valores nuevos para el modelo están 1.96 unidades alejadas de los valores reales

Cómo se puede observar en la Figura 1, los valores reales, el modelo se comporta muy bien alrededor de los datos y además se puede intuir por esos pocos valores atípicos de donde sale ese error de la segunda iteración, que es probable que se hayan juntado esos valores diferentes y, por consiguiente, esa iteración es la que más variabilidad tiene respecto a las demás

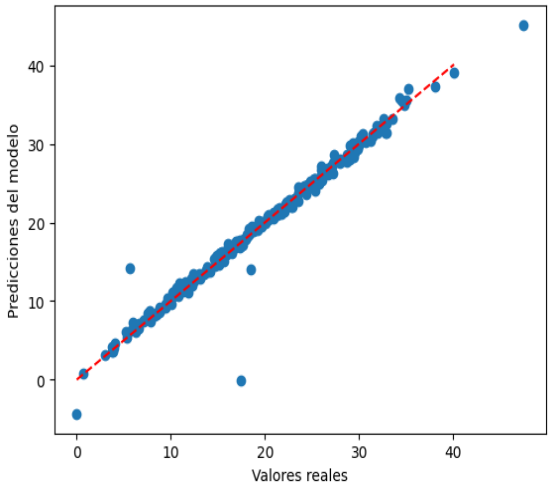

Figura 1. Predicciones del modelo vs valores reales

Regresión Polinomial

En este modelo se realizaron dos iteraciones, en la primera iteración se aplicó el modelo sin implementar la validación cruzada y en la segunda iteración se implementó la validación (k-fold).

Para este modelo se estableció el grado del polinomio en **2**, esto para obtener resultados óptimos evitando sobreajustes o subajustes de los datos en el modelo.

En la **primera iteración** no se implementó la validación cruzada con KFold. Los resultados del error cuadrático medio **(MSE)** y el **R^2** fueron:

Error cuadrático medio (MSE) = 2.02
R^2 = 0.956

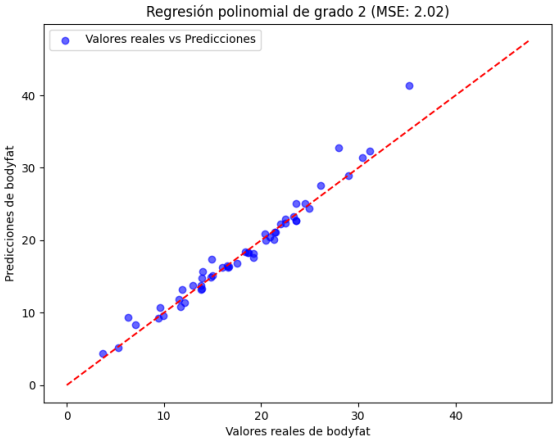

Figura 2. Predicciones del modelo vs valores reales sin implementar KFold.

Al ser un valor bajo, el resultado del error cuadrático medio indica que el modelo está prediciendo valores cercanos a los valores reales, mientras que el resultado del R^2 indica que el modelo tiene un buen ajuste en los datos.

Estos resultados dan a entender que el modelo es válido y tiene un buen rendimiento en los datos de entrenamiento.

En la **segunda iteración** se implementó la validación cruzada con KFold.

Los resultados fueron los siguientes (Tabla 2):

Tabla 2. Número de folds con su respectivo error (MSE) y R^2

Fold	MSE	R^2
1	9.35	0.98
2	5.52	0.90
3	2.45	0.99
4	2.13	0.99
5	2.48	0.99

Error cuadrático medio (MSE) = 4.39
R^2 = 0.97

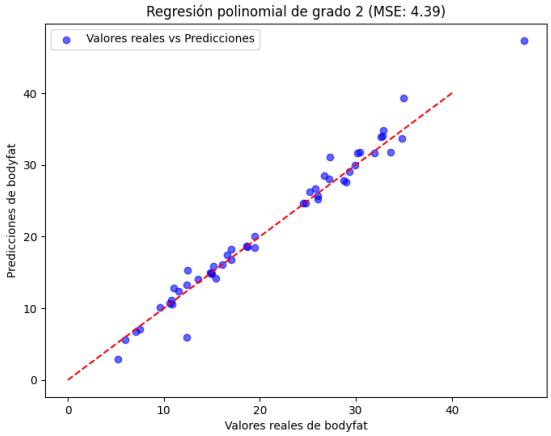

Figura 3. Predicciones del modelo vs valores reales implementando KFold.

En los resultados se puede observar cierta variabilidad en los errores de los folds, con valores que van desde **2.14** hasta **9.36.**

También se observa que el valor del error cuadrático medio es mayor al que se obtuvo sin usar la validación cruzada, esto sugiere que el modelo tiene un rendimiento más deficiente cuando se evalúa con diferentes conjuntos de datos.

Po otro lado, el valor del R^2 es muy alto y ligeramente superior al obtenido sin la validación cruzada, esto indica que el modelo tiene un buen ajuste y explica la mayor parte de la variabilidad de los datos.

K-Nearest Neighbors

Para este modelo se realizaron dos tipos de iteraciones, la primera sin implementar la validación cruzada y la segunda con la validación (k-fold) implementada.

Para la selección del número de vecinos (k), se consideró el tamaño del conjunto de datos (252) y se decidió realizar un análisis experimental dentro del rango [1, 13].

Ya teniendo el rango experimental definido se realizó la primera implementación donde fueron arrojados los siguientes resultados (Tabla 3):

Tabla 3. K vecinos, error (MSE), y R^2.

K	MSE	R^2
1	39.71	0.48
3	26.00	0.66
5	26.17	0.66
7	26.28	0.65
9	29.08	0.62
11	28.75	0.62
13	28.92	0.62

Según los resultados obtenidos anteriormente **podemos** concluir que los mejores modelos fueron aquellos en los que **el número de vecinos k** fue igual a 3, 5 y 7 en donde tanto su MSE como R^2 cuentan con unos valores muy similares.

Se usaron dos gráficas (Figura 4 y 5) para visualizar mejor los resultados y donde se ven mejor las diferencias de forma más intuitiva:

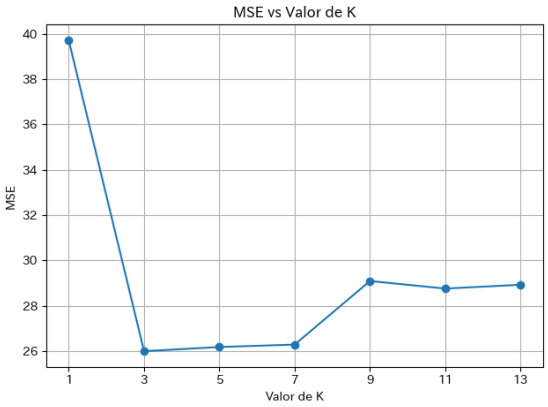

Figura 4. Error cuadrático medio vs valores de K

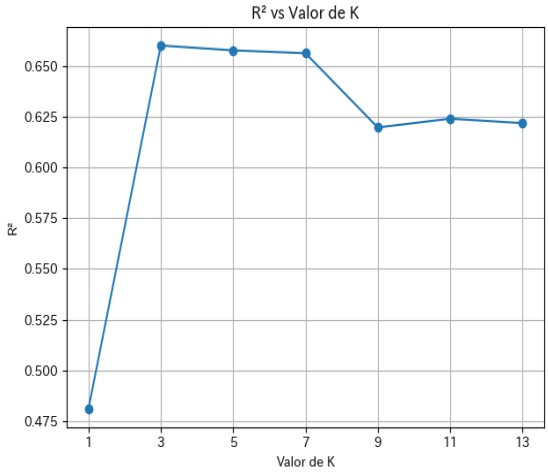

Figura 5. R^2 vs valores de K

Después de tener una visión más clara de los resultados obtenidos mediante la gráfica, se puede concluir que, independientemente del número de vecinos (k) utilizados, la técnica de KNN no logra alcanzar resultados óptimos. Se observa un MSE considerablemente alto y un coeficiente de determinación (R^2) que no alcanza el umbral del 70%, indicando que el modelo no logra explicar la variabilidad de los datos de manera satisfactoria. En consecuencia, se obtienen modelos con una capacidad predictiva baja.

Para la segunda iteración fue implementada la validación cruzada (k-fold) con kf = 5 y se eligieron candidatos a mejores modelos el número de vecinos (k) igual a 3, 5 y 7 dado que fueron los mejores modelos en la primera implementación.

Una vez definido los hiperparámetros a utilizar se obtuvieron los siguientes resultados (Tabla 4):

Tabla 4. Número de folds con su respectivo error (MSE), K vecinos y R^2.

K	MSE	R^2	Fold
3	24.36	0.47	1
3	35.00	0.59	2
3	33.37	0.40	3
3	33.36	0.58	4
3	26.00	0.66	5
5	24.31	0.48	1
5	32.88	0.62	2
5	31.14	0.45	3
5	30.97	0.61	4
5	26.17	0.66	5
7	24.14	0.48	1

7	33.51	0.61	2
7	33.54	0.40	3
7	33.31	0.58	4
7	26.28	0.65	5

Analizando los resultados se puede concluir con seguridad que la técnica de KNN no es la más adecuada para este tipo de problema ya que incluso implementado la validación cruzada y tomando el mejor modelo donde el MSE = 26.00 y el R^2 = 0.66 sigue teniendo carencias para predecir el porcentaje de grasa.

V. Conclusiones

Un paso vital en la creación de un modelo es conocer el trasfondo y la naturaleza del dataset para elegir cuál modelo se ajusta mejor a las necesidades ya que en este informe se demostró que no necesariamente más complejidad implica mejores resultados.

VI. Referencias

[1] Roger W. Johnson. BodyFat. 2014-10-03. Recuperado de: https://www.openml.org/search?type=data&status=active&qualities.NumberOfClasses=lte_1&id=560&sort=runs

[2] AWS. (s.f.). Recuperado de: https://aws.amazon.com/es/what-is/linear-regression/

[3] Machine Learning Basics: Polynomial Regression. Recuperado de: https://towardsdatascience.com/machine-learning-basics-polynomial-regression-3f9dd30223d1

[4] Aprendizaje Automático ml: Métricas de regresión. Recuperado de: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/

[5] K-Nearest Neighbors (KNN) — Explained.

Recuperado de: https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3