ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.520

Preventing Overfitting and Underfitting in Machine Learning Model Development: A Practical Analysis

Dr. José Iván Calderón Carrillo[1]
[1]Universidad Privada del Norte, Perú, *jose.calderon@upn.edu.pe*

***Abstract*– The main objective of this document is to develop a guide, supported by a comparative case study, to help future researchers identify fit problems in predictive models used in machine learning. To this end, theoretical aspects were addressed, and, most importantly, a practical case study based on synthetic data was presented. Three predictive models were constructed using synthetic data: an underfitted model, an overfitted model, and an ideal model. This analysis allowed for the identification of the characteristics of each type of fit. Finally, a simple, practical guide was proposed outlining the steps to follow to obtain a model with an adequate fit, that is, one with good generalization capabilities..**

Prevención del Sobreajuste y Subajuste en el Desarrollo de Modelos de Machine Learning: Un Análisis Práctico

Dr. José Iván Calderón Carrillo[1]
[1]Universidad Privada del Norte, Perú, *jose.calderon@upn.edu.pe*

***Resumen*– El presente documento tiene como objetivo principal desarrollar una guía, apoyada en un caso comparativo, para ayudar a futuros investigadores en la detección de problemas de ajuste en modelos de predicción utilizados en Machine Learning. Para ello, se abordaron aspectos teóricos y, principalmente, se presentó un casó práctico basado en datos sintéticos, con los cuales se construyeron tres modelos de predicción: un modelo subajustado, un modelo sobreajustado y un modelo ideal. A partir de este análisis, se pudo identificar las características de cada tipo de ajuste. Finalmente, se propuso una guía práctica y sencilla con los pasos a seguir para obtener un modelo con un ajuste adecuado, es decir, con una buena capacidad de generalización.**

Palabras clave—Machine Learning, modelos, sobreajuste, subajuste.

I. Introducción

La aplicación de Machine Learning (aprendizaje automático o aprendizaje de máquina) en diferentes industrias ya no es una novedad, se viene realizando durante varios años trayendo excelentes resultados para las organizaciones.

El Machine Learning es uno de los campos más influyentes de la inteligencia artificial, sus aplicaciones impactan en diferentes sectores como la medicina, economía, ciberseguridad, educación, etc. Los modelos de predicción utilizados en este campo, juegan un papel muy importante ya que permiten que los sistemas puedan anticipar comportamientos, categorizar datos, o estimar valores futuros a partir del aprendizaje obtenido de los datos [1].

La capacidad de obtener información y tomar decisiones acertadas mediante el procesamiento de los datos, es sinónimo de competitividad organizacional.

Sin embargo, esto no es tan sencillo como parece. No basta con tener un conjunto de datos, conocimientos en la industria o sector, y conocimientos en algoritmos de Machine Learning para desarrollar un modelo, implementarlo y esperar resultados favorables. Se debe de tener en cuenta un riguroso proceso de evaluación del modelo, el cual consiste en determinar si el modelo desarrollado podrá desempeñarse de manera correcta con datos que no ha visto antes, es decir, tener capacidad de generalización.

Existen errores que se cometen en el desarrollo e implementación de estos modelos, y caen en una situación el cual se denomina mal ajuste del modelo. El mal ajuste pude clasificarse en subajuste y sobreajuste.

El subajuste ocurre cuando el modelo que se ha desarrollado sigue de forma escasa o débil, el patrón o comportamiento de los datos.

Por el contrario, el sobreajuste ocurre cuando el modelo que se ha desarrollado sigue casi a la perfección el patrón o comportamiento de los datos con el cual ha sido desarrollado.

El Sobreajuste genera problemas de generalización y algunos algoritmos de Machine Learning pueden caer en esto si es que se usa un conjunto de datos pequeño [2].

Se preguntarán, ¿Pero que tiene esto de malo? El problema es que este modelo se ha ajustado tanto a los datos con los cuales ha sido entrenado, que ahora le cuesta realizar predicciones de forma más precisa cuando le presentan nuevos datos, es decir, pierde la capacidad para poder generalizar.

Se redacta una metáfora para comprender mejor estos conceptos. Es como estar sentado en una mesa y tener dos platos de sopa. El primer plato contiene sopa fría (modelo subajustado) y el segundo plato contiene sopa muy caliente (modelo sobreajustado).

Lo explicado anteriormente se resume a continuación. Ver Figura 1.

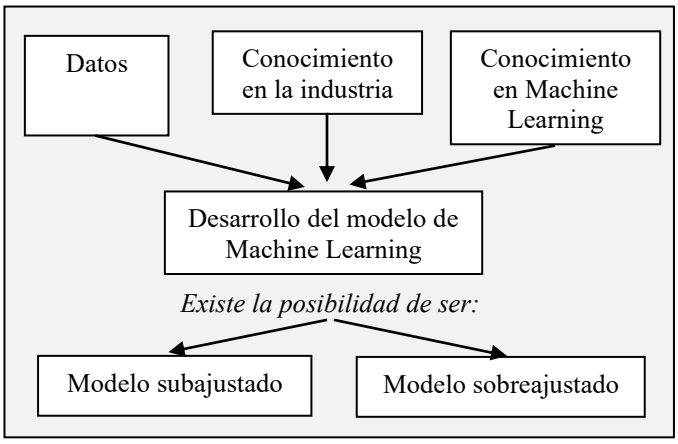

Fig 1. Tipos de mal ajuste del modelo de Machine Learning.

Teniendo en cuenta los escenarios de subajuste y sobreajuste en los modelos de predicción, debemos de tener el enfoque de encontrar ese modelo ideal, un modelo adecuado, un plato de sopa a la temperatura perfecta.

Una gran interrogante en este contexto es, ¿Que tan frecuente ocurren escenarios de subajuste o sobreajuste?

¿Los autores de diferentes publicaciones relacionadas a las aplicaciones de Machine Learning, tuvieron estos conceptos en consideración en sus investigaciones?

A continuación, se muestra un breve conjunto de investigaciones relacionadas con el tema para poder responder las preguntas planteadas.

A. Revisión de algunas investigaciones

En la investigación titulada "Clasificador de residuos sólidos para la i.e. Juan XXIII del municipio de Algeciras con aplicación de Machine Learning", los autores desarrollaron una red neuronal capaz de clasificar los residuos sólidos para promover una mejor experiencia y capacitar a la comunidad de la institución sobre el reciclaje. Desarrollaron la red neuronal en el laboratorio de Google Teachable Machine y usando aproximadamente 2400 datos, imágenes descargadas de Google.

Los autores concluyen mencionando la importancia de este modelo de Machine Learning, indican que el uso de esta tecnología puede generar soluciones disruptivas a problemas reales y que al ser de bajo costo, puede replicarse en otros centros educativos o sectores para replicar su impacto [3].

Sin embargo, no se puede apreciar en el documento un análisis o revisión de un posible sobreajuste o subajuste del modelo desarrollado, ya que aparentemente usaron todos los datos disponibles para entrenar el modelo por lo que existe la posibilidad que haya un sobreajuste.

En la investigación titulada "Aplicación de Machine Learning a un modelo tradicional de prevención y detección de fraude en entidad financiera proyectado a periodos trimestrales", la autora tuvo como objetivo implementar un modelo de Machine Learning para complementar y fortalecer la detección de fraudes en dicha organización.

La autora concluye que el modelo de Machine Learning logró mejorar la capacidad de detección de fraude en comparación con el método tradicional, basado en reglas. Además, indica que el modelo de regresión logística y árbol de decisión poseen mejor capacidad predictiva [4].

En el documento en mención, tampoco se detalla o describe una evaluación de sobreajuste o subajuste del modelo de Machine Learning.

En la investigación titulada "Aplicación de Machine Learning y metodología CRISP-DM para la clasificación precisa de severidad en casos de dengue", los autores tuvieron el objetivo de clasificar los casos sin alarma y con alarma, y para ello probaron varios modelos como la regresión logística, K-vecinos más cercanos y bosque aleatorio, resultando el más preciso el algoritmo de bosque aleatorio.

Los autores utilizaron un conjunto de datos que contenía 2010 filas y 54 columnas. Luego de la preparación y limpieza de datos, el conjunto de datos sr redujo a 1994 filas y 12 columnas.

Los autores afirman que su modelo logró tasa de clasificación correcta del 100% en la gravedad del dengue [5].

En el documento, los autores mencionan que separaron los datos en prueba y entrenamiento, realizaron ajuste de hiperparámetros y prepararon los datos. Etas últimas son acciones que contribuyen a prevenir el sobreajsute en un modelo de Machine Learning.

En la investigación titulada "Aplicación de Machine Learning a la eficiencia energética en edificios. Caso práctico: edificio ETOPIA", el autor como tuvo como objetivo analizar los datos relacionados con las condiciones térmicas y el consumo energético del edificio.

La autora desarrolló un modelo de regresión, indica que se puede predecir el comportamiento térmico del edificio en su mayoría con exactitud. Además de ello, menciona la problemática del desajuste (subajuste) y sobreajuste como un problema al cual se debe prestar atención luego que el modelo está entrenado [6].

Para evaluar el modelo y reducir la probabilidad que caiga en un mal ajuste, el conjunto de datos de entrenamiento se divide en subconjuntos: el subconjunto de entrenamiento, subconjunto de validación cruzada y subconjunto de prueba.

El subconjunto de entrenamiento, aproximadamente el 60% de los datos, se utiliza para crear el modelo. El subconjunto de validación cruzada, aproximadamente el 20% de los datos, sirve para visualizar el rendimiento del modelo y su grado de ajuste en datos que no ha visto. Finalmente, el subconjunto de prueba, aproximadamente el 20% de los datos, una vez seleccionado el modelo, se utiliza para conocer su nivel de generalización [6].

En la investigación titulada "Aplicaciones de Machine Learning en el campo de variabilidad estelar", el autor tuvo como principal objetivo desarrollar un modelo capaz de identificar de forma confiable posibles planetas dentro de una base de datos, con la intención de automatizar esta actividad y permitir a los astrónomos dedicar más tiempo de estudio a estos objetos de interés.

El autor desarrolló un modelo de bosque aleatorio y máquinas vectoriales de soporte. El autor indica que es posible detectar señales planetarias en datos de velocidad radial usando modelos de Machine Learning relativamente simples [7].

El autor no menciona de manera explícita temas referentes al subajuste o sobreajuste de los modelos desarrollados, tampoco no se menciona la separación de datos en entrenamiento y prueba.

Se han mencionado cinco investigaciones aleatorias referentes a aplicaciones o desarrollo de algún modelo de Machine Learning aplicado en diferentes campos o industrias.

Se puede apreciar que no en todas las investigaciones se hace mención o se explica la verificación de un correcto ajuste del modelo a los datos, para evitar el subajuste o sobreajuste del modelo desarrollado.

No se pretende indicar que la ausencia de este apartado desacredita el resultado o la rigurosidad de las investigaciones que se han mencionado. Pero se considera que debe estar presente ya que incrementa la confiabilidad y complementa el desarrollo del modelo de Machine Learning.

Por ello, el objetivo principal de esta investigación es desarrollar un procedimiento o un manual que sirva como guía para futuros investigadores, de modo que puedan identificar casos de subajuste o sobreajuste y determinar los pasos y validaciones necesarios para obtener un modelo de Machine Learning correctamente ajustado.

II. Materiales y Métodos

Para desarrollar este manual o guía, se realizó el siguiente procedimiento:

Se utilizó la plataforma de Google Colab para trabajar con lenguaje de programación Python para la creación de datos y el desarrollo de modelos de predicción usados en Machine Learning.

Primero, se creó un conjunto de datos sintéticos, una variable dependiente y una variable independiente. En total se crearon 500 registros.

Posterior a ello, se procedió a crear el primer modelo de Machine Learning, un modelo subajustado, se explicó sus características para identificar este tipo de error de ajuste.

Luego, se procedió a crear el segundo modelo de Machine Learning, un modelo sobreajustado, también se explicó sus características y cómo identificar cuando nuestro modelo ha caído en este tipo de ajuste.

Por último, se desarrolló un modelo ajustado correctamente, explicando sus características y cómo se diferencia de los modelos anteriores.

Finalmente, se compararon los resultados de los tres modelos de Machine Learning y su rendimiento frente a datos nunca vistos, con ello se creó un instructivo práctico y fácil de entender, indicando los pasos que debemos seguir y señales que debemos tener en cuenta para asegurar un correcto ajuste de nuestro modelo de Machine Learning.

III. Resultados

A. Creación de datos sintéticos

Se crearon dos variables, una dependiente y la otra independiente. Las variables se desarrollaron de la siguiente forma:

```
vind = np.random.normal(15.5,1.5,500)
ruido = np.random.normal(0,400,500)
vdep = (vind – 20)**4 + Ruido
```

Datos fuera de rango:

```
vind_extra = np.linspace(x['vind'].min() - 5,x['vind'].max() + 5, 200).reshape(-1,1)
vdep_extra = (vind_extra.flatten() - 20)**4
```

La variable independiente sigue una distribución normal con media de 15.5 y una desviación estándar de 1.5 unidades. El ruido, sigue una distribución normal con valores que van desde 0 hasta 400 y la variable dependiente está en función de la variable independiente sumado el ruido.

Adicional a ello, se creó un conjunto de 200 registros de datos fuera de rango, es decir, con valores más extremos pero que siguen el mismo patrón de los datos sintéticos.

Los datos creados se pueden apreciar a continuación. Ver Tabla 1.

TABLA I
DATOS SINTÉTICOS

Index	vind	vdep
0	13.58406	2418.89183
1	15.953674	512.409354
2	16.963596	-316.216539
3	15.264758	600.592997
4	13.717445	1185.5539
...	...	...
495	13.164184	2333.23151
496	16.023625	361.585302
497	14.035563	417.241252
498	15.396065	108.597795
499	15.388809	332.150873

Posterior a ello, se procedió a realizar la gráfica de dispersión con los datos sintéticos. Ver Figura 2.

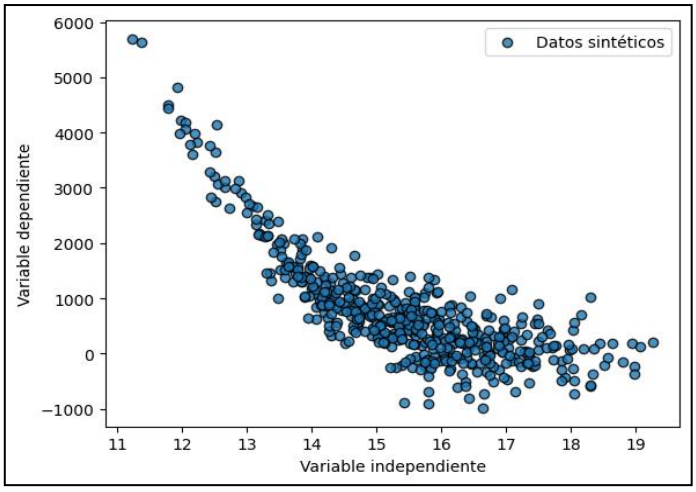

Fig. 2 Gráfico de dispersión de los datos sintéticos

En la Figura 2 se aprecia cómo se comporta la variable dependiente en función de la variable independiente. Se observa la tendencia de una curva marcada.

El coeficiente de correlación de Pearson entre ambas variables es -0.801676, indica una correlación negativa fuerte, es decir, si la variable independiente aumenta, la variable dependiente tiende a disminuir.

Con el valor del coeficiente de correlación y según lo que se observó en la Figura 1, se puede afirmar que contamos con dos variables asociadas entre sí, listas para el análisis y desarrollo de los diferentes modelos.

B. Desarrollo del modelo 1 (subajustado)

Para el desarrollo del modelo 1, del total de 500 datos, el 80% se tomaron como datos de entrenamiento y el 20% como datos de prueba. Se entrenó un modelo de regresión lineal simple. A continuación, se muestra el modelo de regresión lineal simple con los datos de prueba. Ver Figura 3.

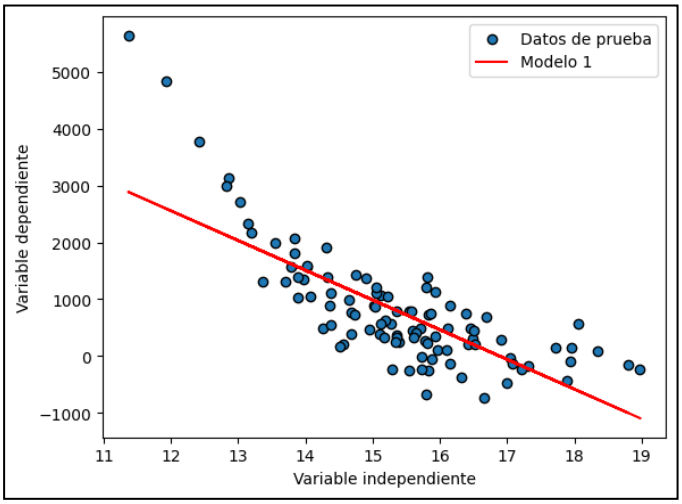

Fig. 3 Ajuste del modelo 1 sobre los datos de prueba

El modelo 1 es el siguiente:

Y = 8478.3141 - 501.5599 X

El valor del R^2 (coeficiente de determinación) con los datos de prueba es de 0.6155. Esto se puede interpretar que aproximadamente, en los datos de prueba, el 61.55% de la variación de la variable dependiente se ve explicada por la variable independiente, según el modelo desarrollado.

Otro indicador en el cual debemos fijarnos en la evaluación de un modelo de predicción es en el error cuadrático medio o MSE, por sus siglas en inglés (mean squared error).

El MSE del modelo 1 con los datos de prueba es de 411866.5783. Es un valor muy grande, significa que el error (diferencia entre el valor real y la predicción del modelo) también es grande. Posterior a ello, se realizó la evaluación con los datos fuera de rango pero que siguen el mismo comportamiento de los datos sintéticos. Ver Figura 4

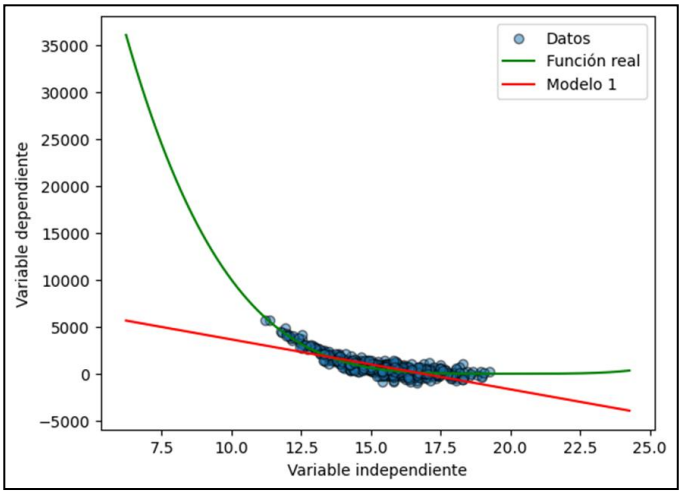

Fig. 4 Ajuste del modelo 1 con la función real de los datos

Como se aprecia en la Figura 4, el modelo 1 (regresión lineal simple) presenta mucha diferencia con la tendencia y función real de los datos sintéticos (línea color verde). El MSE es 71955821.063 y el R^2 es 0.1199. En pocas palabras, el modelo 1 al ser puesto a prueba con datos que no ha visto antes, presenta una gran diferencia entre los datos y las predicciones. Además de ello, se aprecia que el nivel de explicación de la variación de la variable dependiente en función de la independiente es del 11.99%, un valor bastante bajo.

Resumiendo, visualmente se aprecia que el modelo 1 sigue de forma débil y no se adecua del todo al comportamiento de los datos, las métricas del modelo, el R^2 y el MSE obtuvieron valores regulares y al ponerlo a prueba con otros valores, los resultados fueron bastante bajos. Estas son características de un modelo subajustado.

C. Desarrollo del modelo 2 (sobreajustado)

Para el desarrollo de este segundo modelo, se tomó la misma proporción de datos, es decir, 80% para entrenamiento y 20% de datos para prueba.

En este segundo modelo, se creó un polinomio de grado 30 para tratar de que el modelo pueda seguir de forma más certera el comportamiento de los datos.

Debemos recordar lo que caracteriza a un modelo sobreajustado. Intenta seguir prácticamente a la perfección la tendencia de los datos con los que fueron entrenados, trayendo como consecuencia que aprende el ruido de los datos, es decir, componentes aleatorios presentantes en los datos que no contienen información útil para explicar o predecir la variable de interés.

A continuación, se aprecia el modelo 2, con los datos de prueba. Ver Figura 5.

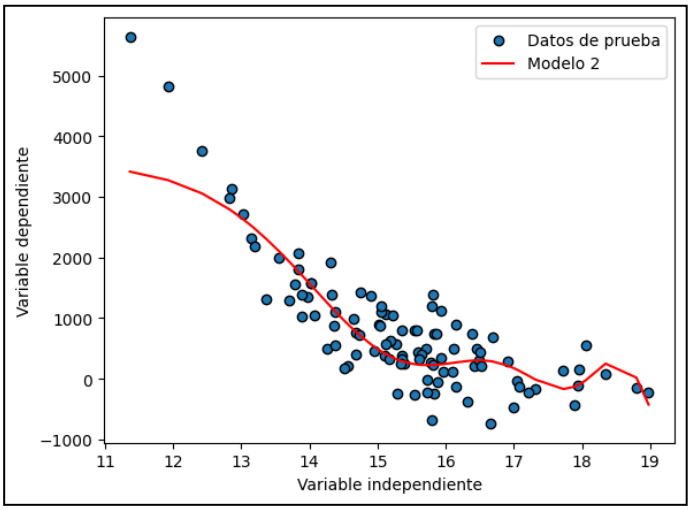

Fig. 5 Ajuste del modelo 2 sobre los datos de prueba

La Figura 5 muestra de forma visual como se ajusta el modelo 2 a los datos sintéticos.

El modelo 2 es el siguiente:

$$Y = 2954.0696 - 4.207085E{-}27X - 2.506466E{-}27X^{2} - 9.94341915E{-}30X^{3} + 1.532E{-}33X^{4} + 5.61883238E{-}35X^{5} - 6.71317477E{-}38X^{6} - 5.87747407E{-}39X^{7} - 3.24760284E{-}40X^{8} - 4.61855906E{-}39X^{9} - 6.44465075E{-}38X^{10} - 8.79907391E{-}37X^{11} - 1.17508438E{-}35X^{12} - 1.53318661E{-}34X^{13} - 1.95060225E{-}33X^{14} - 2.41304402E{-}32X^{15} - 2.89138980E{-}31X^{16} - 3.33836649E{-}30X^{17} - 3.68806704E{-}29X^{18} - 3.86118296E{-}28X^{19} - 3.77922321E{-}27X^{20} - 3.38982021E{-}26X^{21} - 2.70089731E{-}25X^{22} - 1.81244450E{-}24X^{23} - 9.21526292E{-}24X^{24} - 2.66510345E{-}23X^{25} + 7.54215688E{-}24X^{26} - 8.22282481E{-}25X^{27} + 4.44068485E{-}26X^{28} - 1.19432369E{-}27X^{29} + 1.28197483E{-}29X^{30}$$

El valor del R^2 del modelo 2 con los datos de prueba es de 0.74445. Y el valor del MSE es 273759.

Estos valores son mucho mejores que el modelo 1. El modelo 2, aparentemente explica mucho mejor el porcentaje de variación de la variable dependiente en función de la variable independiente, además de ser más preciso (menor valor del MSE).

Visualmente se aprecia que el modelo 2 es más complicado y aparentemente representa mejor el comportamiento de los datos. La primera impresión es que podría ser un buen modelo de predicción, debido a que pudo haber capturado la complejidad del comportamiento de los datos, es por ello que presenta varias curvas y puntos de inflexión.

Ante estos primeros resultados, surgen las siguientes incógnitas: ¿El modelo 2 podría ser un modelo ideal? y ¿el modelo 2 tendrá buena capacidad de generalización?

A continuación, se realizó la evaluación del modelo 2 con los datos fuera de rango. Ver Figura 6.

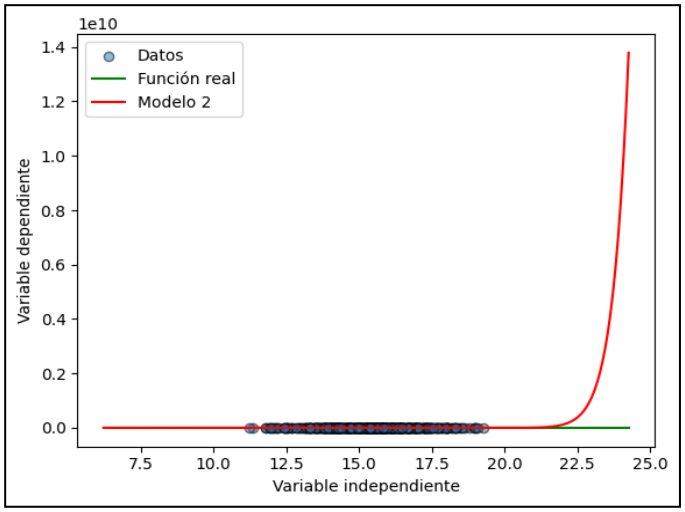

Fig. 6 Ajuste del modelo 2 con la función real de los datos

En la Figura 6 se puede apreciar claramente un problema, el cual se explica en los siguientes párrafos.

El gráfico se ve de esa forma debido a que el polinomio de grado 30 aprendió a interpolar los datos dentro del rango observado, pero al extrapolar fuera de ese rango los términos de mayor grado dominan el comportamiento de la función. Aunque los coeficientes son valores muy pequeños, al multiplicarse por potencias altas de la variable independiente, producen valores muy grandes, lo que genera una explosión en la curva, evidenciando inestabilidad numérica y evidentemente una falta de capacidad de generalización.

El valor del R^2 del modelo 2 es -39925432417.48 y su MSE es de 3264532832084790000. Estos valores simplemente son absurdos, un coeficiente de determinación negativo, indica un nivel de ajuste muy malo, y el valor del error es muy elevado. Son el resultado de intentar desarrollar un modelo que se ajuste a la perfección a los datos recolectados.

Esto demuestra que el modelo 2 no ha aprendido la relación real entre las dos variables, solo aprendió patrones de ruido de los datos de entrenamiento, por lo que este modelo resulta inútil para extrapolar, a pesar de tener buenos valores de R^2 y MSE con los datos de entrenamiento, se evidenció que no puede generalizar. Simplemente es un modelo sobreajustado.

D. Desarrollo del modelo 3 (adecuado)

El desarrollo de este modelo tuvo un procedimiento fue similar al de los anteriores. Se dividió los datos en entrenamiento y prueba, pero la diferencia en el desarrollo de este modelo es que se utilizó la técnica de validación cruzada.

La validación cruzada es una técnica importante en la evaluación y mejora de modelos, ya que supera algunas limitaciones de la simple división en datos de entrenamiento y prueba. Esta técnica permite obtener generalizaciones más

confiables. Usa diferentes particiones de datos para lograr maximizar la información obtenida y robustecer al modelo [8].

En el desarrollo de este modelo se utilizó la validación cruzada para dividir los datos en 10 subconjuntos. Primero, se consideró un modelo de grado 2, al probar este modelo en los 10 subconjuntos de datos, en promedio, dio un R^2 de 0.8167. Por lo tanto, se consideró desarrollar el modelo de grado 3, y al ser probado con los 10 subconjuntos de datos, en promedio el R^2 fue de 0.8302.

No se hizo la prueba con un modelo de grado 4, debido a que probablemente se podría estar cayendo en el error del sobreajuste, por lo que se mantuvo en grado 3.

A continuación, se aprecia el modelo 3 con los datos de prueba. Ver Figura 7.

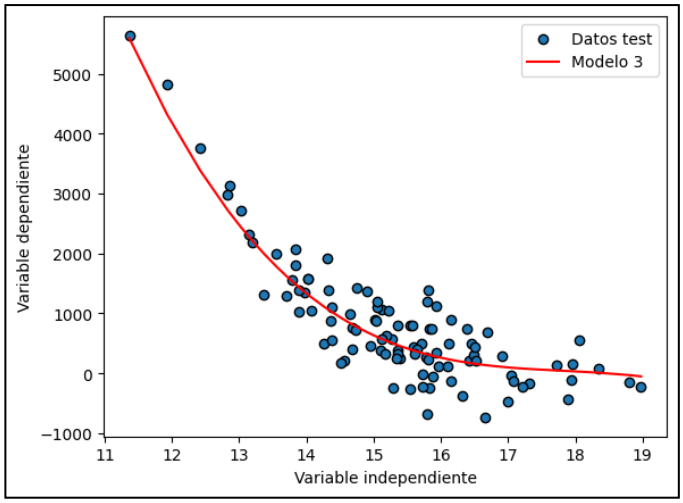

Fig. 7 Ajuste del modelo 3 sobre los datos de prueba

El modelo 3 se aprecia continuación:

$$Y = 105511.4735 - 17605.8090\,X + 980.1367\,X^2 - 18.1875\,X^3$$

El R^2 del modelo 3 con los datos de prueba tiene un valor de 0.8142 y su MSE es de 133063.79. Son valores muy buenos. Este modelo explica en un 81.42% la variación de la variable dependiente en función de la variable independiente y el error es un valor bajo comparado al de los otros modelos.

El modelo 3 sigue la tendencia de una curva ligera, aparentemente capturando el patrón de los datos.

No es un modelo tan simple ni subajustado como el modelo 1, pero tampoco es un modelo tan complicado como el modelo 2.

Ahora entra la duda si es que el modelo 3 puede estar sobreajustado o no. Para ello, se realizó la prueba del modelo 3 con los datos fuera de rango.

Los resultados se muestran a continuación. Ver Figura 8.

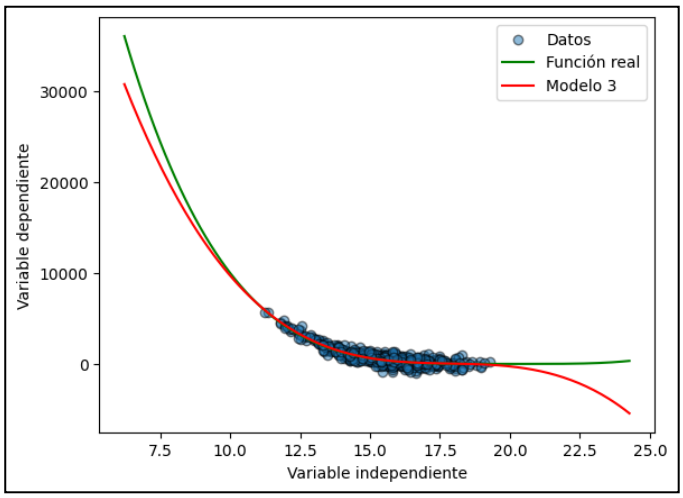

Fig. 8 Ajuste del modelo 3 con la función real de los datos

Como se observa, el modelo 3 tiene una gran semejanza con la función real de los datos sintéticos. En términos numéricos, el valor del R^2 es de 0.9610 y el MSE es 3186927.87.

A pesar de no ser un modelo tan complicado, ha logrado capturar correctamente el patrón y comportamiento de las variables. Se hace esta afirmación debido al buen resultado obtenido en los datos de prueba y con los datos fuera de rango.

Este modelo tiene una buena capacidad de generalización. No es tan simple, ni tan complicado. Recordemos la metáfora, no es una sopa muy fría ni muy caliente, simplemente está a la temperatura ideal.

Debemos recordar que parte de este resultado se debe al uso de la validación cruzada. Esta técnica ayuda a evitar el sobreajuste ya que el modelo se evalúa en varios conjuntos de datos. Esto permite evaluar si el modelo es capaz de generalizar de forma correcta en vez de ajustarse demasiado a un solo conjunto de datos [9].

Ya desarrollado y explicado los tres modelos, se puede resumir los resultados de la siguiente manera. Ver Tabla 2.

TABLA II
RESUMEN DE LOS RESULTADOS DE LOS MODELOS DESARROLLADOS

Tipo de datos	Métricas	Modelo 1	Modelo 2	Modelo 3
Prueba	R^2	Regular	Aceptable	Aceptable
	MSE	Regular	Aceptable	Aceptable
Nuevos	R^2	Pésimo	Pésimo	Aceptable
	MSE	Pésimo	Pésimo	Aceptable

Como se aprecia en lata Tabla 2, el modelo 1 (modelo de regresión lineal simple) al ser evaluado con los datos de prueba tuvo resultados regulares y al ser evaluado con los datos fuera de rango o nuevo conjunto de datos, tuvo pésimos resultados.

El modelo 2 (modelo polinomial de grado 30), al ser evaluado con los datos de prueba, obtuvo resultados aceptables o conformes. Pero al ser evaluado con el nuevo conjunto de datos, tuvo pésimos resultados.

Por último, el modelo 3 (modelo polinomial de grado 3), al ser evaluado con los datos de prueba, tuvo resultados aceptables al igual que al ser evaluado con el nuevo conjunto de datos o datos fuera de rango.

De acuerdo con los resultados del análisis práctico, se ha propuesto una breve guía para el correcto desarrollo de un modelo de Machine Learning y como prevenir el subajuste y sobreajuste, el cual se muestra a continuación. Ver Figura 9.

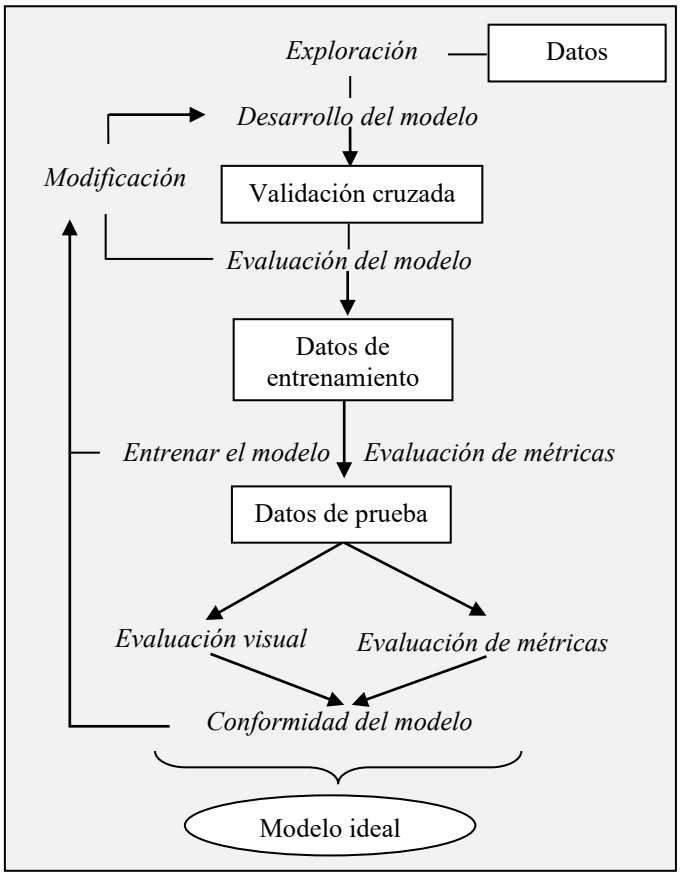

Fig. 9 Guía para la prevención del subajuste y sobreajuste en el desarrollo de modelos de Machine Learning

La guía propuesta inicia con una exploración de los datos. Esto quiere decir que debemos entender primero el comportamiento de los datos, como se relaciona una variable sobre la otra, para ello, dependiendo del tipo de datos, en el caso sean variables numéricas, se recomienda un gráfico de dispersión. Con ello nos podríamos dar una primera idea de la complejidad del modelo que debemos de desarrollar.

El segundo paso es el desarrollo del modelo propiamente dicho. En función de las primeras impresiones del comportamiento de los datos, se buscará un modelo que pueda aproximarse al comportamiento de los datos, puede ser quizás una regresión lineal, polinomial, etc.

El tercer paso es realizar la técnica de validación cruzada. Es decir, dividir nuestro conjunto de datos en varios grupos para probar el rendimiento de nuestro primer modelo, si en promedio los resultados son aceptables, entonces se procederá a entrenar el modelo con los datos de entrenamiento. Pero si en promedio los resultados no son aceptables, entonces se realizará las modificaciones necesarias al modelo para luego ser verificadas nuevamente con la validación cruzada.

El cuarto paso es realizar el entrenamiento del modelo y realizar la evaluación de las métricas, como el coeficiente de determinación (R^2), el MSE, entre otros. Si estos resultados son aceptables, se procede a ejecutar el modelo con los datos de prueba, caso contrario se realizarán las modificaciones correspondientes al modelo.

El quinto paso es ejecutar el modelo con los datos de prueba y realizar dos evaluaciones, visual y de métricas. La evaluación visual consiste en verificar que el modelo desarrollado siga el comportamiento de los datos, un comportamiento que no debe ser ni muy simple y ni muy complicado, como se vieron en los modelos 1 y 2 del caso práctico. Se recomienda a nivel visual, que el modelo siga suavemente la tendencia de los datos. Por otro lado, la evaluación de métricas consiste en analizar los valores del R^2, MSE, entro otros, con el fin de determinar si el modelo es aceptable.

Hasta este punto, si el modelo presentó resultados favorables con los datos de entrenamiento, pero no pasó lo mismo con los datos de prueba, se debe realizar las modificaciones del modelo y seguir nuevamente los pasos indicados.

Una vez evaluados los resultados del modelo con la validación cruzada, datos de entrenamiento y datos de prueba, se puede dar conformidad de manera teórica a este modelo de predicción. La probabilidad de que el modelo desarrollado esté subajustado o sobreajustado es baja. En otras palabras, estamos ante un potencial modelo ideal.

IV. Conclusiones y Recomendaciones

El uso de modelos de Machine Learning está presentes y cumplen un rol muy importante en diferentes industrias, tales como salud, finanzas, manufactura, gestión ambiental, y la astronomía, facilitando y agilizando actividades de detección o predicción, lo cual contribuye en la toma de decisiones.

Los modelos de Machine Learning se construyen a través de datos, aprenden de los datos. Se debe de tener en cuenta un correcto procedimiento para evitar que nuestro modelo caiga en un subajuste o sobreajuste, lo que trae consigo problemas para generalizar.

Mediante la revisión de una breve lista de investigaciones relacionadas con el tema, se pudo observar que no todos los autores tuvieron presente o mencionaron la problemática de un posible subajuste o sobreajuste en sus modelos de Machine Learning.

Del caso práctico desarrollado, se pudo determinar cuales son las principales características de un modelo subajustado,

sobreajustado y un modelo con el ajuste ideal. Y lo más importante, como aprender a identificarlos.

Es importante tener en cuenta un correcto procedimiento para el desarrollo de modelos de predicción utilizados en Machine Learning para evitar que dichos modelos caigan en un mal ajuste.

Se pudo realizar una propuesta de una guía práctica para que futuros investigadores puedan utilizarla como referencia y tener en cuenta los pasos a seguir para lograr desarrollar un modelo ideal con buena capacidad de generalización. Sin embargo, es necesario aclarar que la técnica de la validación cruzada no es la única utilizada para evitar un mal ajuste de nuestros modelos.

Para evitar un mal ajuste del modelo se puede utilizar otras técnicas como el ajuste de hiperparámetros y el procesamiento de los datos [10].

Así mismo, es necesario mencionar que existen más modelos en Machine Learning como el árbol de decisión, redes neuronales, máquinas de soporte vectorial, bosques aleatorios, entre otros. Todos ellos también presentan el riesgo de un mal ajuste, el cual debe mitigarse mediante las técnicas que se mencionaron anteriormente.

Una de las limitaciones de esta investigación, por motivos de extensión del documento, solo se desarrolló y explicó la técnica de validación cruzada como un método para la prevenir el sobreajuste. También por el mismo motivo, no se pudo realizar una muestra más extensa de revisión de investigaciones anteriores.

Otra limitación fue el uso de datos sintéticos. Habría sido más beneficioso emplear datos provenientes de un proceso real para el desarrollo de los distintos modelos de Machine Learning.

Finalmente, debemos tener en cuenta que, hasta este punto, todo es teórico, pero, potencialmente funcione en la práctica. El último paso que debemos realizar es ejecutar el modelo de predicción y darle seguimiento a una muestra de resultados, si todo está conforme, entonces podremos afirmar que hemos desarrollado un modelo con el ajuste ideal. Hemos encontrado nuestra sopa a la temperatura perfecta.

REFERENCIAS

[1] S. Melián, “Modelos de Predicción en Machine Learning : Una Revisión,” Universidad de La Laguna, 2025. [Online]. Available: https://riull.ull.es/xmlui/bitstream/handle/915/42994/Modelos de prediccion en Machine Learning Una revision.pdf?isAllowed=y&sequence=1&utm_source=chatgpt.com

[2] J. Hermitaño, “Aplicación de machine learning en la gestión de riesgo de crédito financiero : Una revisión sistemática,” *Interfases*, 2022.

[3] D. Gómez, E. Cerquera, A. Tamayo, O. Ortiz, and C. Pérez, “Clasificador de residuos sólidos para la i . e . Juan XXIII del municipio de Algeciras con aplicación de Machine Learning,” *Rev. del Sist. Ciencia, Tecnol. e Innovación*, vol. 6, no. 1, [Online]. Available: https://revistas.sena.edu.co/index.php/sennova/article/view/5409

[4] Lady Quintero, “Aplicación de Machine Learning a un modelo tradicional de Prevención y detección de fraude en entidad financiera proyectado a periodos trimestrales,” Universidad de La Salle, 2023. [Online]. Available: https://ciencia.lasalle.edu.co/server/api/core/bitstreams/41dc6b1e-6220-40a9-8239-952787597ba5/content

[5] C. Mejía, M. Rincón, L. Palmera, and L. Arévalo, “Aplicación de machine learning y metodología CRISP-DM para la clasificación precisa de severidad en casos de dengue,” *Rev. Colomb. Tecnol. Av.*, pp. 78–85, 2024.

[6] E. Van der, “Aplicación de Machine Learning a la eficiencia energética en edificios. Caso práctico: edificio ETOPIA.,” Universidad Zaragoza, 2023. [Online]. Available: https://zaguan.unizar.es/record/133565/files/TAZ-TFM-2023-1631.pdf

[7] F. Burgos, “Aplicaciones de Machine Learning en el campo de variabilidad estelar,” Universidad de Concepción, 2022. [Online]. Available: https://repositorio.udec.cl/server/api/core/bitstreams/003f576b-1b31-432c-9809-35db16c3db5c/content

[8] D. Zuleta, O. Bru, and K. Pastor, “Validación Cruzada: una herramienta crucial para mejorar la eficiencia de modelos de clasificaci´on con datos biomédicos,” *Comun. en Estadística*, vol. 18, no. 1, pp. 51–85, 2025, [Online]. Available: https://revistas.usantotomas.edu.co/index.php/estadistica/article/view/11214/9030

[9] J. Corrales, C. Barreno, A. Salvatierra, and N. Pastuña, “Toma de decisiones de riego inteligente para cultivos de café con la utilización de sensores IoT,” *Rev. Científica Mundo la Investig. y el Conoc.*, vol. 9, pp. 380–394, 2025, doi: 10.26820/recimundo/9.(esp).mayo.2025.378-394.

[10] L. García and M. Vallejo, “Máquinas de aprendizaje y soft sensores de tipo nariz y lengua electrónica para la detección de cáncer,” *Tecnológicas*, vol. 28, pp. 1–21, 2025.