Predicción de la Temperatura Crítica de Materiales Superconductores Mediante Modelos de Aprendizaje de Máquina

Gabriela Rashuamán Arce
Sección de Física
Departamento de Ciencias e Ingeniería, PUCP
Lima, Perú
g.rashuaman@pucp.edu.pe

Jhairt Vega Quino
Sección de Ingeniería Informática
Departamento de Ciencias e Ingeniería, PUCP
Lima, Perú
j.vega@pucp.edu.pe

Ricardo Uribe Bejarano
Sección de Electricidad y Electrónica
Departamento de Ciencias e Ingeniería, PUCP
Lima, Perú
ruribebejarano@gmail.com

Sofía Escajadillo Bazán
Sección de Ingeniería Informática
Departamento de Ciencias e Ingeniería, PUCP
Lima, Perú
s.escajadillo@pucp.edu.pe

***Abstract*—El presente trabajo se enfoca en la predicción de la temperatura crítica (Tc) de materiales superconductores a partir de sus fórmulas químicas. Esta capacidad es clave para acelerar el desarrollo de nuevos superconductores con aplicaciones en medicina, energía y física de partículas.**

Se utilizó un conjunto de datos reales y públicos de 21,263 superconductores, extraído de la base de datos de la NIMS (National Institute for Materials Science, Japón). A partir de las fórmulas químicas se extrajeron 81 características basadas en propiedades elementales (masa atómica, energía de ionización, conductividad térmica, etc.).

Se implementaron cinco enfoques de aprendizaje automático: *Random Forest Regressor*, *XGBoost Regressor*, *Support Vector Regression (SVR)*, *modelo de Stacking* y *regresión polinomial con regularización Ridge*. Los resultados fueron comparados mediante análisis de precisión (R^2) y RMSE. Toda la implementación se realizó en Python 3.0 usando Google Colab.

I. INTRODUCCIÓN

La búsqueda de materiales superconductores que operen a temperaturas críticas más elevadas constituye uno de los retos centrales en la ciencia de materiales, debido a su enorme potencial en aplicaciones médicas, energéticas y de transporte de alta eficiencia. Tradicionalmente, la predicción de la temperatura crítica (Tc) de un compuesto se ha basado en modelos físicos teóricos y en ensayos experimentales costosos y prolongados. Sin embargo, el creciente volumen de datos experimentales disponibles permite abordar este problema desde la perspectiva del aprendizaje automático, donde modelos estadísticos detectan patrones en la composición química y propiedades elementales para estimar Tc con alta precisión.

El objetivo de este estudio es tomar como punto de partida el trabajo de Hamidieh *et al.* [1], quienes extrajeron un amplio conjunto de características a partir de la fórmula química de cada superconductor y aplicaron XGBoost sobre más de 21 000 superconductores para estimar directamente su temperatura crítica. Partiendo de esa metodología, aquí comparamos cuatro enfoques de regresión —Random Forest Regressor, XG Boost, Multiple Linear Regression y Redes Neuronales— evaluando su desempeño mediante validación cruzada y analizando cómo la selección de características y los parámetros de regularización afectan la capacidad predictiva.

Este trabajo está organizado de la siguiente manera. En la sección II se presentan los trabajos previos en la predicción de la temperatura crítica mediante aprendizaje automático. La sección III describe el conjunto de datos, la extracción de características y las técnicas de preprocesamiento aplicadas. A continuación, la sección IV aborda el análisis de correlaciones entre variables y las estrategias de reducción de dimensionalidad. En la sección V se detallan los cinco modelos de regresión evaluados —Random Forest Regressor, XG Boost Regressor, Support Vector Regression, un modelo de Stacking y Regresión Polinomial (Polynomial Ridge)— así como el procedimiento de ajuste de sus hiperparámetros. La sección VI discute los resultados obtenidos, comparando fortalezas y limitaciones de cada enfoque, y, finalmente, la sección VII expone las conclusiones principales y las líneas de trabajo futuro.

II. ESTADO DEL ARTE

La predicción de la temperatura crítica (T_c) de superconductores ha pasado de reglas empíricas basadas en la valencia de los elementos [2], [3] a modernas técnicas de aprendizaje automático. Por ejemplo, Owolabi et al. [4] aplicaron redes neuronales para estimar T_c en superconductores basados en hierro, alcanzando un error cuadrático medio (ECM) de aproximadamente 5 K. De manera similar, Owolabi y Olatunji [5] entrenaron máquinas de vectores de soporte (SVM) para MgB_2, obteniendo una precisión global cercana al 85

Sin embargo, estos primeros estudios se centraron en familias concretas de materiales. Stanev et al. [6] fueron de los primeros en explotar la base de datos SuperCon (NIMS), con más de 12 000 compuestos. Dividieron el problema en una clasificación binaria ($T_c > 10$ K frente a $T_c \leq 10$ K) y lograron una precisión fuera de muestra de $\approx 92\%$. Además, construyeron modelos de regresión especializados

para cupratos, hierro-basados y superconductores de baja T_c, con coeficientes de determinación (R^2) superiores a 0.80.

Más recientemente, Hamidieh [1] extrajo 81 características a partir de propiedades elementales y empleó el algoritmo XGBoost para predecir T_c directamente de la composición química. Su modelo alcanzó un RMSE fuera de muestra de 9.5 K y un R^2 de 0.92, mejorando notablemente un modelo de regresión múltiple convencional (RMSE de 17.6 K, R^2 de 0.74). Estos avances ilustran el gran potencial de las técnicas de de aprendizaje de máquina y la riqueza del conjunto de datos NIMS para generar predictores fiables de la temperatura crítica.

III. PREPROCESAMIENTO DE LA DATA

A. Descripción del conjunto de datos

El conjunto de datos final consta de 21 263 muestras de materiales superconductores, cada una caracterizada por 81 características y una variable dependiente continua (T_c). Todas las características son numéricas, obtenidas a partir de estadísticas (media, desviación típica, rango, entropía, etc.) de ocho propiedades elementales (masa atómica, energía de ionización, afinidad electrónica, conductividad térmica, radio atómico, densidad, calor de fusión y valencia) extraídas de la fórmula química de cada compuesto. Tras un proceso de limpieza inicial —eliminación de registros con T_c nulo o faltante, fórmulas químicas inválidas, materiales con elementos de número atómico mayor que 86 y duplicados— no se detectaron valores faltantes en ninguna de las 82 columnas finales. La distribución de la variable objetivo T_c es altamente asimétrica, abarcando desde aproximadamente 0.0002K hasta 185K, con un pequeño grupo de superconductores de muy alta temperatura que actúan como valores atípicos extremos. Además, varias de las características presentan rangos intercuartílicos amplios y correlaciones significativas entre sí, lo que sugiere la conveniencia de un análisis exhaustivo de outliers y la aplicación de técnicas de escalado robusto antes del entrenamiento de los modelos de regresión.

B. Balanceo y preprocesamientos

La presencia de valores atípicos en las variables continuas puede distorsionar los estimadores y comprometer la capacidad de generalización de los modelos de regresión, por lo que es necesario detectar y tratar estos outliers antes de proceder al escalado y entrenamiento. Para este fin, en el notebook se implementó la *winsorización*, una técnica estadística que sustituye los valores extremos por los percentiles predefinidos, atenuando su influencia sin eliminar observaciones completas. Concretamente, para cada feature se calcularon los cuartiles Q_1 y Q_3, se determinó el rango intercuartílico

$$IQR = Q_3 - Q_1$$

y se establecieron umbrales en $Q_1 - 1.5\,IQR$ y $Q_3 + 1.5\,IQR$. A continuación, se evaluó la asimetría (skewness) de cada variable para decidir el truncamiento: si skewness > 1 solo se limitó la cola alta, si skewness < -1 solo la cola baja, y cuando existían outliers en ambos extremos se aplicaron límites simétricos basados en el porcentaje de valores extremos detectados. Estos umbrales dinámicos se aplicaron de forma idéntica sobre X_{train} y X_{test}, garantizando que el modelo nunca se enfrentara a valores fuera de los límites definidos. Así, la winsorización estabiliza la varianza de las variables y las prepara para la estandarización posterior con `StandardScaler`, mejorando la robustez y el rendimiento de los algoritmos de regresión.

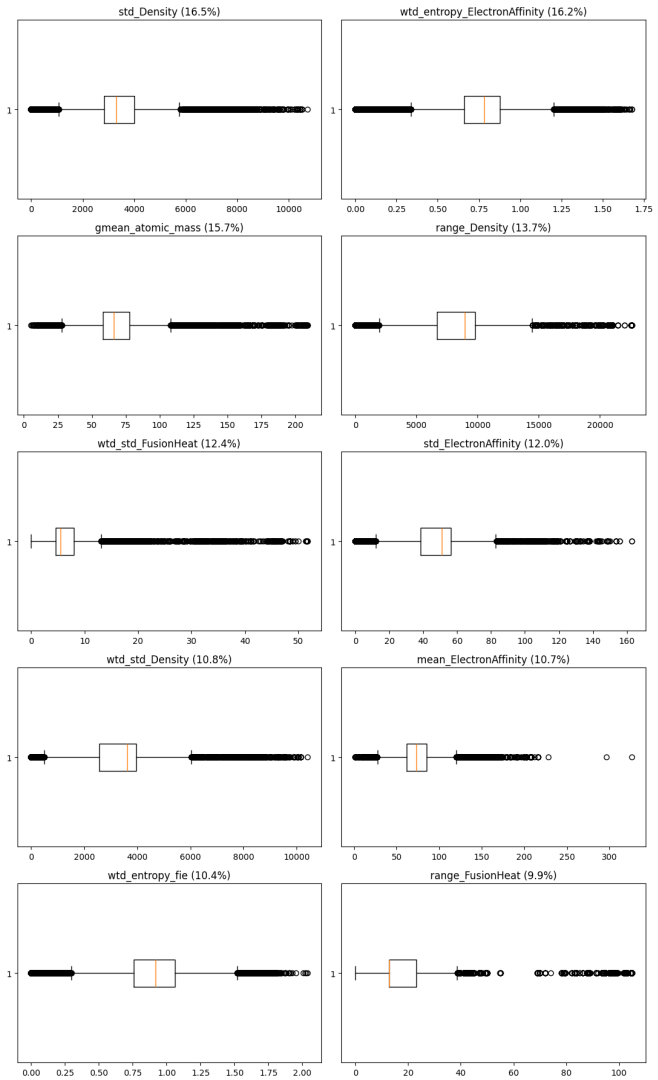

Fig. 1. Diagramas de caja para las 10 variables con mayor proporción de outliers.

IV. ANÁLISIS DE CORRELACIONES Y REDUCCIÓN DE DIMENSIONALIDAD

Para optimizar el desempeño de los modelos y reducir el riesgo de sobreajuste, en el notebook se aplicaron dos pasos sucesivos sobre el conjunto de entrenamiento (y se replicaron sobre el conjunto de prueba): eliminación de features de baja variabilidad y descarte de features muy correlacionadas.

A. *Eliminación de características de baja variabilidad*

En primer lugar, se empleó la clase `VarianceThreshold` de Scikit-learn para descartar aquellas variables que aportan poca o ninguna información. Con un umbral inicial de 0 se eliminaron las columnas completamente constantes; a continuación, se ajustó un segundo filtro con umbral 0.01 para eliminar también las variables cuya varianza era prácticamente nula (casi constantes). Esta operación se entrenó exclusivamente sobre `X_train` y luego se aplicó la transformación resultante a `X_test`, garantizando coherencia entre ambos conjuntos.

B. *Eliminación de características altamente correlacionadas*

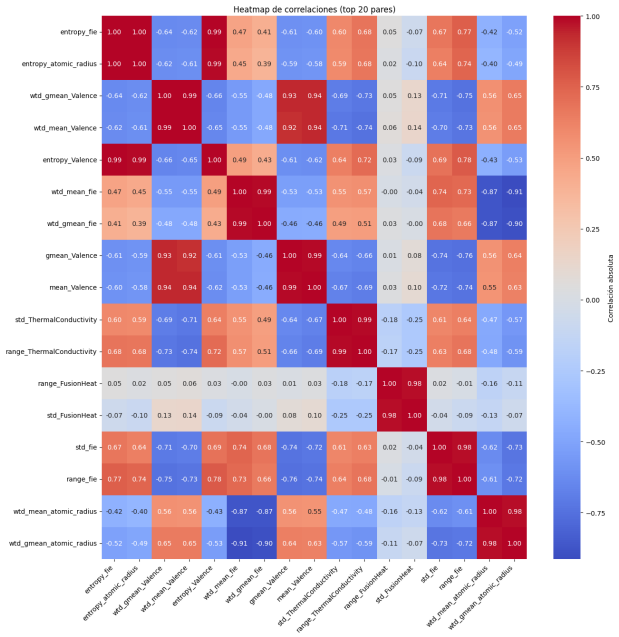

Fig. 2. Matriz de correlaciones absolutas entre las 20 variables más correlacionadas.

A continuación, se calculó la matriz de correlación de Pearson sobre `X_train` mediante `pandas.DataFrame.corr()`, y se extrajo la zona superior de dicha matriz (sin incluir la diagonal) para identificar pares de variables con correlación absoluta mayor a 0.7. Se generó así una lista de columnas redundantes y se eliminaron de `X_train`; la misma lista se utilizó para filtrar `X_test`. Con este paso se reduce la multicolinealidad, se simplifica el espacio de características y se mejora la estabilidad de los coeficientes en los modelos de regresión.

V. AJUSTES DE LOS MODELOS DE REGRESIÓN

A. *Random Forest Regressor*

El modelo *Random Forest Regressor* es un algoritmo de ensamble que construye múltiples árboles de decisión de forma aleatoria y combina sus predicciones mediante el promedio, lo que mejora la estabilidad y reduce el sobreajuste de un único árbol. Este modelo se optimizó mediante `RandomizedSearchCV` con 5 iteraciones y validación cruzada de 3 pliegues, obteniéndose como mejor configuración `n_estimators=73`, `max_depth=20`, `max_features=0.5`, `min_samples_leaf=1` y `min_samples_split=3`. Con estos parámetros, el modelo alcanzó en el conjunto de prueba un RMSE $= 9.3013$ y un $R^2 = 0.9265$,

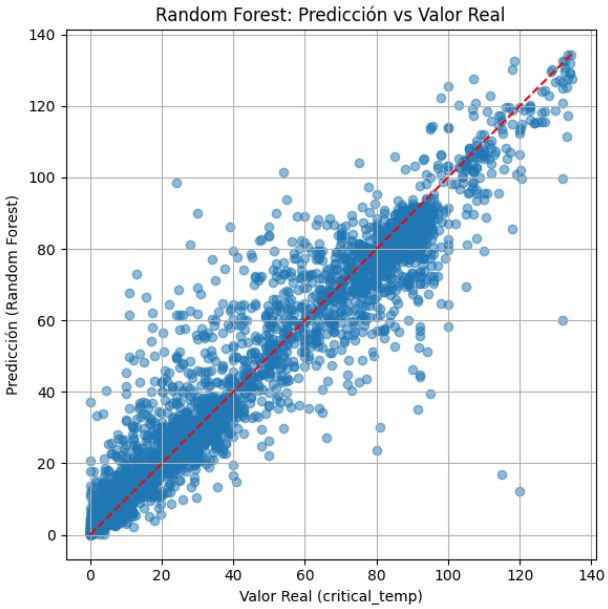

Fig. 3. Dispersión de predicciones frente a valores reales de T_c en el conjunto de prueba.

evidenciando un ajuste sólido y buena capacidad de generalización (frente a un $R^2_{\text{train}} = 0.9764$). Se observó además una fuerte capacidad del modelo para identificar relaciones no lineales entre las variables. Se graficó la importancia de las características y se detectó una alta dependencia de un subconjunto reducido de atributos.

B. *XGBoost Regressor*

El modelo *XGBoost Regressor* es un algoritmo de boosting de gradiente que construye árboles de forma secuencial, corrigiendo en cada paso los errores de los árboles anteriores, lo que le permite capturar relaciones complejas y no lineales en los datos. El modelo se optimizó mediante `RandomizedSearchCV` (10 iteraciones, CV=3), obteniéndose como mejor configuración `n_estimators=199`, `learning_rate=0.1189` y `max_depth=7`. Con estos parámetros, el modelo alcanzó en el conjunto de prueba un RMSE $= 9.3080$ y un $R^2 = 0.9264$, muy similar al rendimiento reportado por el Random Forest. Además, para facilitar la interpretación, se incluyó un gráfico de barras con las 15 características más importantes según el atributo `feature_importances_` de XGBoost, lo que permitió visualizar de forma clara qué variables aportaron con mayor peso a la predicción de la temperatura crítica.

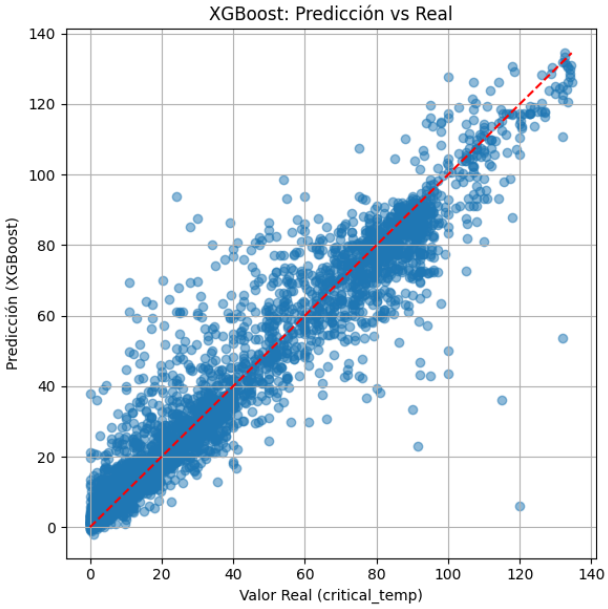

Fig. 4. Dispersión de predicciones frente a valores reales de T_c en el conjunto de prueba.

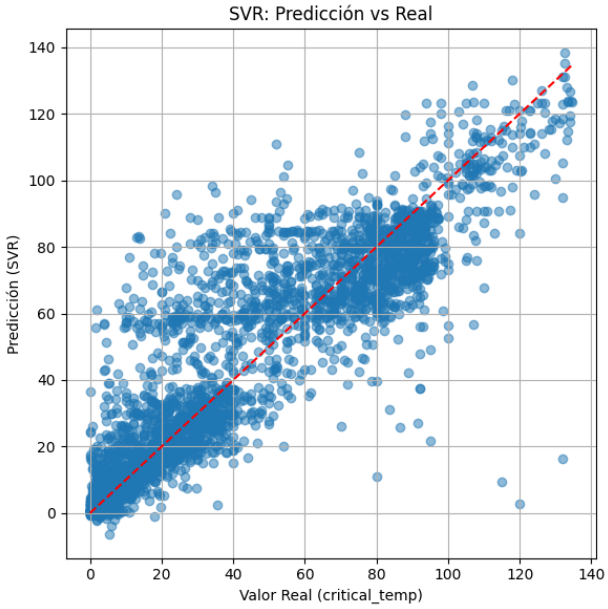

Fig. 5. Dispersión de predicciones frente a valores reales de T_c en el conjunto de prueba.

C. Support Vector Regression (SVR)

El modelo *Support Vector Regressor* es un método de regresión basado en máquinas de vectores de soporte que emplea un *kernel* para proyectar los datos a un espacio de mayor dimensión y ajustar una función que se mantenga dentro de un margen de tolerancia ϵ, al tiempo que maximiza el margen entre los vectores de soporte. En el notebook se optimizaron los hiperparámetros `C`, ϵ y el tipo de kernel mediante `RandomizedSearchCV` (10 iteraciones, validación cruzada de 3 pliegues). La mejor configuración resultó ser `C` $= 38.08$, $\epsilon = 0.4803$ y `kernel` $=' rbf'$. Con estos valores, el modelo alcanzó en el conjunto de prueba un RMSE $= 13.7864$ y un $R^2 = 0.8385$ (frente a un $R^2_{\text{train}} = 0.8470$). Estos indicadores, sensiblemente inferiores a los de Random Forest y XGBoost, se deben principalmente a la gran heterogeneidad de la distribución de la temperatura crítica y al elevado número de variables de entrada. SVR, aunque muy efectivo en espacios de menor dimensión y con kernels adecuados, puede requerir un ajuste más fino de γ o transformaciones adicionales del target (p. e. logarítmicas) para capturar correctamente la no linealidad y el rango dinámico presente en los datos.

D. Stacking

El *Stacking Regressor* es un método de ensamble que combina las predicciones de varios modelos base para obtener un estimador final más robusto. En nuestro caso, los modelos base fueron XGBoost, SVR y Random Forest, y como meta-estimador se utilizó `RidgeCV`, aprovechando la opción `passthrough=True` para incluir directamente las predicciones de los bases junto con las variables originales.

Tras ajustar el `StackingRegressor` con validación cruzada de 3 pliegues, se obtuvo en el conjunto de prueba un RMSE $= 9.2572$ y un $R^2 = 0.9281$, mejorando ligeramente los resultados individuales de los modelos base. Este ligero

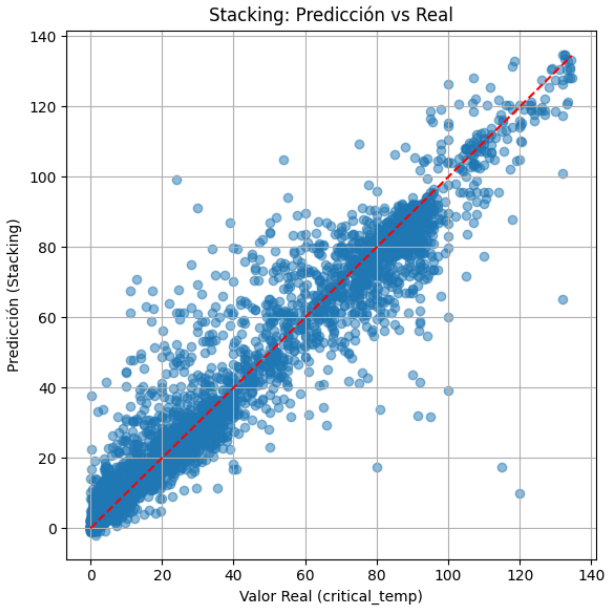

Fig. 6. Dispersión de predicciones frente a valores reales de T_c en el conjunto de prueba.

incremento en la capacidad predictiva se explica porque el stacking logra capturar distintos tipos de sesgos y varianzas: XGBoost aporta flexibilidad no lineal, SVR proporciona regularización en regiones de densidad baja, y Random Forest estabiliza variaciones locales. El meta-estimador RidgeCV, al reequilibrar linealmente estas predicciones, refina aún más la

salida conjunta, mejorando la generalización sin incurrir en sobreajuste.

E. Polynomial Regression (Ridge)

El *Polynomial Regressor* es una extensión de la regresión lineal que permite capturar relaciones no lineales al incluir términos de interacción y potencias de las variables originales. Para evitar el sobreajuste que puede surgir al expandir dimensionalmente las features, se combinó con una regularización tipo Ridge. El modelo se implementó como un pipeline (PolynomialFeatures + Ridge), y se seleccionaron las variables base usando RFECV con regresión lineal como estimador inicial. Se consideraron términos hasta grado 2 e hiperparámetros ajustados mediante RandomizedSearchCV (alpha [0.01, 10.0]).

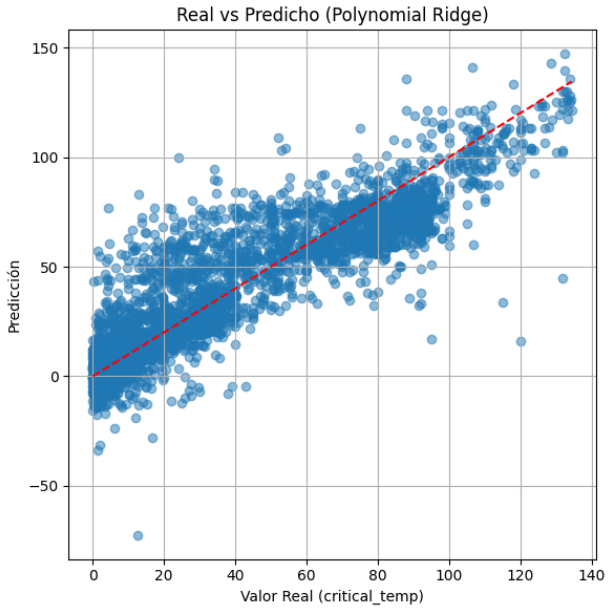

Fig. 7. Dispersión de predicciones frente a valores reales de T_c en el conjunto de prueba.

El modelo final obtuvo un RMSE = 14.21 y R² = 0.87 en el conjunto de prueba, superando ampliamente a la regresión lineal múltiple y mostrando resultados competitivos frente a los modelos de ensamble, pero con mejor interpretabilidad. Además, se visualizaron los coeficientes más relevantes del modelo polinomial para identificar interacciones significativas entre variables predictoras.

VI. DISCUSIÓN

Los métodos de ensamble, como Random Forest, XGBoost y Stacking, demostraron una alta capacidad para capturar relaciones no lineales e interacciones complejas entre las variables, logrando ajustarse de manera robusta a la heterogeneidad de los datos de temperatura crítica. En contraste, el Support Vector Regressor ofrece un equilibrio intermedio, con mayor flexibilidad que la regresión lineal pero limitado cuando la dimensionalidad y el rango dinámico del target son elevados. Por su parte, la regresión lineal múltiple resultó insuficiente para reflejar la complejidad del problema, evidenciando un sesgo considerable pese a su transparencia y facilidad de interpretación. Sin embargo, la elevada capacidad de los modelos de ensamble conlleva un riesgo mayor de sobreajuste y una menor interpretabilidad que los métodos lineales.

Para cuantificar estas diferencias, en la Tabla I se presentan las métricas de RMSE y R^2 obtenidas en entrenamiento y prueba para cada modelo:

Modelo	$\text{RMSE}_{\text{test}}$	R^2_{test}	R^2_{train}
Polynomial Regressor	14.21	0.87	0.82
SVR	13.79	0.84	0.85
Random Forest Regressor	9.30	0.93	0.98
XGBoost Regressor	9.31	0.93	0.97
Stacking Regressor	9.21	0.93	0.98

TABLE I
COMPARACIÓN DE MÉTRICAS DE DESEMPEÑO DE LOS DISTINTOS MODELOS

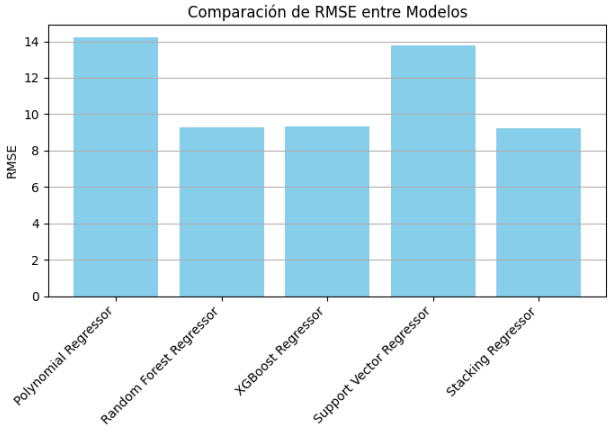

Fig. 8. Comparación del error cuadrático medio (RMSE) entre los modelos evaluados.

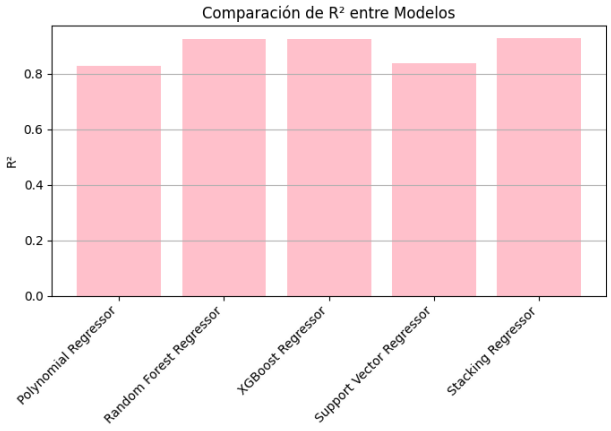

Fig. 9. Comparación del coeficiente de determinación R^2 entre los modelos evaluados.

La estrategia de selección de features basada en eliminación de baja varianza y de alta correlación simplificó el espacio de características y aceleró el entrenamiento, pero introduce varias limitaciones. En primer lugar, pueden haberse descartado señales no lineales relevantes al eliminar variables de varianza casi nula. En segundo lugar, la supresión de variables altamente correlacionadas puede obviar interacciones útiles entre ellas que un modelo más complejo podría aprovechar. Además, los umbrales elegidos (varianza < 0.01 y correlación > 0.7) son en gran medida arbitrarios y podrían requerir ajuste para equilibrar mejor la retención de información y la reducción de redundancia.

El tratamiento de outliers mediante winsorización mostró ser eficaz para atenuar la influencia de valores extremos y estabilizar la varianza previa al escalado, pero también implica un riesgo de distorsionar la distribución original de propiedades físicas críticas. Al truncar valores fuera de los percentiles definidos, es posible suavizar fenómenos significativos que se manifiestan precisamente en esas colas de la distribución. Por ello, conviene validar que las observaciones extremas no contengan información esencial antes de aplicar esta técnica y considerar transformaciones alternativas (por ejemplo, logarítmicas o de Yeo–Johnson) que preserven mejor la forma de la distribución.

Como líneas de trabajo futuro, resulta recomendable implementar validación cruzada anidada para obtener estimaciones de rendimiento más robustas y libres de sesgo, así como explorar transformaciones del target (log o Box–Cox) para homogenizar la varianza de la variable objetivo. También sería valioso generar nuevas características mediante interacciones entre variables clave o indicadores binarios de elementos químicos, y evaluar modelos alternativos como LightGBM o arquitecturas de redes neuronales profundas. Finalmente, la incorporación de datos experimentales adicionales a la base original podría mejorar la validez externa y la capacidad de generalización de los modelos.

VII. CONCLUSIONES

En este trabajo se ha presentado un flujo de trabajo completo para la predicción de la temperatura crítica de superconductores a partir de propiedades físicas y químicas derivadas, abarcando desde la limpieza y preprocesamiento de datos hasta la optimización y comparación de varios algoritmos de regresión.

Los métodos de ensamble (*Random Forest*, *XGBoost* y *Stacking*) demostraron un desempeño superior, con un RMSE de prueba en torno a 9.26 K y coeficientes de determinación cercanos a 0.93. Se destacó el *Stacking Regressor* como el modelo más preciso gracias a la combinación sinérgica de sus estimadores base.

A pesar de la solidez de los ensambles, la regresión polinomial regularizada demostró que es posible capturar relaciones complejas con alta precisión sin recurrir a árboles de decisión o métodos de ensamble más costosos, intrínsecos al fenómeno de la superconductividad, y el SVR obtuvo resultados intermedios, lo que pone de manifiesto la importancia de seleccionar algoritmos adecuados al rango dinámico y la complejidad del target. La estrategia de reducción de dimensionalidad y el tratamiento de outliers contribuyeron de manera significativa a la estabilidad del entrenamiento, pero es posible que se hubieran descartado señales sutiles; por ello, la selección de características y la winsorización deben aplicarse con cuidado y complementarse con otras técnicas en trabajos posteriores.

Finalmente, el modelo desarrollado ofrece una herramienta prometedora para acelerar la identificación de nuevos materiales superconductores de alto T_c en fase de diseño; sin embargo, su aplicación práctica requerirá la validación con datos experimentales adicionales y la exploración de arquitecturas más avanzadas. Entre las líneas de trabajo futuro se destacan la implementación de validación cruzada anidada, la ingeniería de features basada en conocimiento específico del dominio y la integración de datos experimentales con simulaciones para construir modelos híbridos más precisos y generalizables.

REFERENCES

[1] K. Hamidieh, "A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor," *arXiv:1803.10260*, 2018.

[2] B. T. Matthias, "Empirical rules for superconductivity," *Phys. Rev.*, vol. 97, pp. 74–76, 1955.

[3] K. Conder, "Status of empirical rules guiding the search for new superconductors," *J. Supercond. Nov. Magn.*, vol. 29, pp. 3–12, 2016.

[4] J. Owolabi and O. Olatunji, "Neural-network prediction of critical temperature for Fe-based superconductors," *Physica C*, vol. 501, pp. 1–5, 2014.

[5] J. Owolabi and O. Olatunji, "Support vector machine model for predicting T_c of MgB_2 superconductors," *Supercond. Sci. Technol.*, vol. 28, p. 115007, 2015.

[6] V. Stanev *et al.*, "Machine learning modeling of superconducting critical temperature," *npj Comput. Mater.*, vol. 4, art. 29, 2018.