Universidad Técnica Federico Santa María
Campus Santiago San Joaquín
Introducción a Machine Learning para
Astronomía

Predicción del Diámetro de Asteroides por Medio de Algoritmos de Machine Learning

Autores. José Lizana[1] Rafael Reveco [1]

[1]*Departamento de Física, Universidad Técnica Federico Santa María, San Joaquín, Chile*

1. Links github

1.1. José Lizana

ML Jose Lizana Proyecto Final

1.2. Rafael Reveco

ML Rafael Reveco Proyecto Final

2. Resultados y Análisis

El objetivo de este trabajo fue predecir el díametro de asteroides (*Diametro_km*) a partir de sus parámetros fotométricos y orbitales, utilizando técnicas de regresión en Machine Learning.

2.1. Preparación y Limpieza del Dataset

Primero se cargó el dataset original (*dataset.csv*) y se realizó un proceso de estandarización de nombres de columnas.
Mediante dos diccionarios de mapeo (*rename_column y trad*) se tradujeron y normalizaron los nombres de las variables, pasando de etiquetas crudas como H, a, per a nombres más descriptivos, como :

- $H \rightarrow Magnitud\_Absoluta$
- $a \rightarrow Eje\_Semi\_Mayor\_au$
- $diameter \rightarrow Diametro\_km$
- $albedo \rightarrow Albedo\_Geometrico$

Esto facilita la interpretación posterior de los resultados y la lectura del código.

Posteriormente, se eliminaron columnas claramente irrelevantes para la predicción, principalmente identificadores, cadenas de texto y metadatos temporales, tales como:

- `ID`, `SPK_ID`, `Nombre_Completo`, `Des_prim`,
- `ID_Solucion_Orbital`, `Epoca_Calendario`, `Tiempo_Paso_Perihelio_Calendario`, entre otras.

Luego se abordó el problema de los valores faltantes. En primer lugar, se eliminó cualquier fila donde el objetivo Diametro_km fuera nulo, ya que no tiene sentido intentar predecir un valor desconocido. Para las variables numéricas se utilizó un imputador de mediana SimpleImputer(strategy='median'), lo que resulta robusto frente a valores atípicos.

Para las variables categóricas, los valores nulos fueron reemplazados por la moda (valor más frecuente), asegurando que el modelo no encuentre NaN durante el entrenamiento.

Posteriormente, se aplicó una función de tratamiento de los outliers handle_outliers_iqr basada en el rango intercuartílico (IQR). En lugar de eliminar filas completas, se utilizó el método 'cap', que consiste en recortar los valores extremos al límite inferior o superior permitido. Esto reduce el impacto de valores atípicos sin perder información valiosa del conjunto de datos.

Finalmente, obtenemos nuestro mapa de calor para ver si la relación que tienen nuestras variables Magnitud_Absoluta y Albedo_Geometrico importantes de estudio es coherente con la columna Diametro_km, tambien notar multicolinealidad en variables orbitales (Movimiento_Medio y Eje_Semi_Mayor_au practicamente le dan la misma información a nuestro dataset).

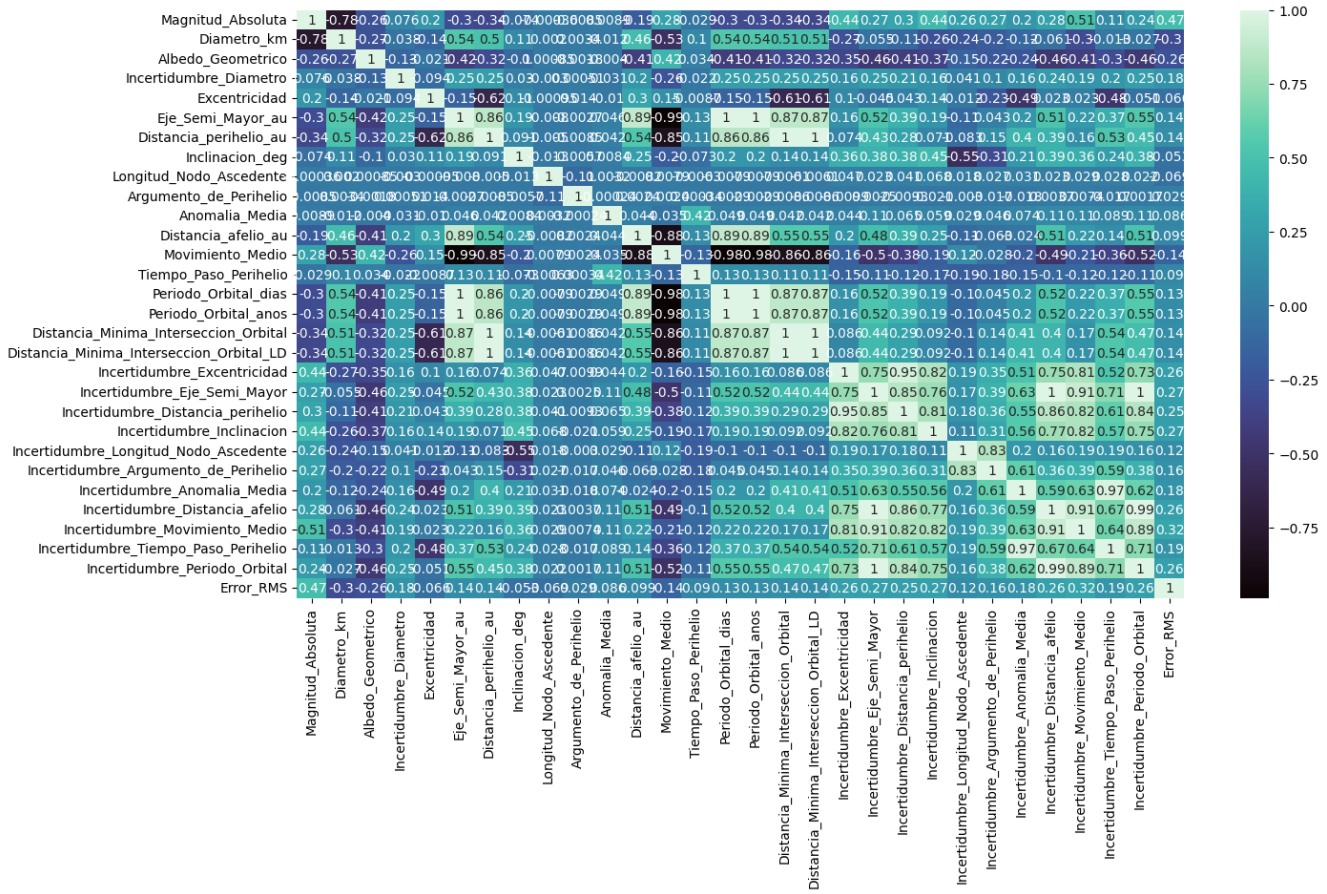

Figura 1: Mapa de calor

2.2. Selección de características Feature Importance

Una vez limpio el conjunto de datos ast_cleaned, se entrenó un modelo RandomForestRegressor ligero (50 árboles) exclusivamente con el propósito de estimar la importancia de cada variable mediante el atributo feature_importances .

El análisis reveló que la predicción del diámetro está dominada por un conjunto reducido de variables:

- `Magnitud_Absoluta`,
- `Albedo_Geometrico`,
- en menor medida: `Movimiento_Medio`, `Eje_Semi_Mayor_au`, `Periodo_Orbital_anos`, `Tiempo_Paso_Perihelio` y `Periodo_Orbital_dias`.

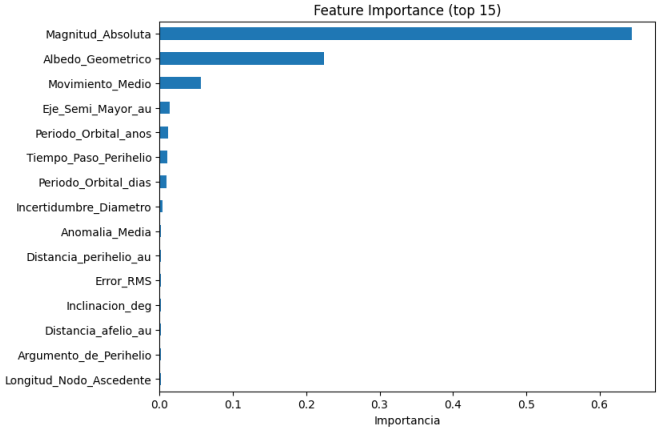

Figura 2: Características importantes con respecto a la variable objetivo diametro_km

Las dos primeras variables, Magnitud_Absoluta y Albedo_Geometrico, concentraron la mayor parte de la importancia del modelo, lo cual es coherente con la física del problema: el diámetro de un asteroide se relaciona directamente con su brillo intrínseco y su albedo.

A partir de esta información, se construyó un conjunto de datos reducido ast_top7, conservando únicamente las siete variables más relevantes junto con la variable objetivo Diametro_km. Este proceso permitió disminuir el ruido y la complejidad del modelo, haciendo el entrenamiento más eficiente sin sacrificar capacidad predictiva.

2.3. Modelos de Regresión

Para evaluar el desempeño de los modelos se trabajó directamente con ast_top7, definiendo :

$$X = ast\_top7.drop('Diametro\_km', axis = 1)$$
$$y = ast\_top7['Diametro\_km']$$

En lugar de usar un único split entrenamiento prueba, se optó por una validación cruzada 5-fold, lo que proporciona métricas más estables y menos dependientes de una partición específica del dataset.

Las métricas utilizadas fueron:

- MSE (Error cuadrático Medio): Mide el error promedio al cuadrado entre los diámetros predichos y los reales
- R^2(coeficiente de determinación): Indica que proporción de la viriabilidad de Diametro_km es explicada por el modelo.

2.3.1. Decision Tree Regressor

Se entrenó un árbol de decisión de profundidad máxima 25:

- MSE promedio = 0.39
- R^2 promedio = 0.93

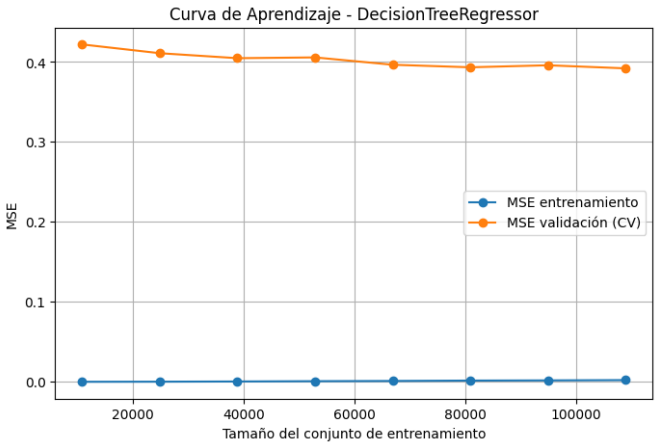

Figura 3: Curva de Aprendizaje del Algoritmo Decisión Tree Regresor

La curva de aprendizado mostró:

- MSE de entrenamiento $\tilde{0}$ en todos los tamaños de muestra, por lo que el árbol memoriza los datos de entrenamiento.
- MSE de validación bastante mayor, lo que indica sobreajuste, por lo que hay varianza alta.
- Decision Tree no generaliza bien los datos nuevos.

2.3.2. Random Forest Regressor

Luego se entrenó un RandomForestRegressor con 100 árboles, lo que nos dió:

- MSE promedio = 0.22
- R^2 promedio = 0.96

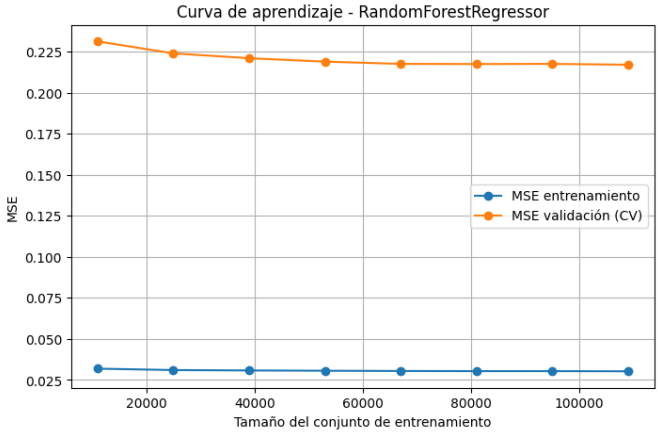

Figura 4: Curva de Aprendizaje del Algoritmo Random Forest Regressor

La curva de aprendizaje mostró:

- Un MSE de entranemiento bajo, pero no nulo, debido al promedio entre muchos árboles.
- Un MSE de validación más bajo y estable que en el árbol individual, disminuyendo ligeramente al aumentar el tamaño de entrenamiento.
- Presenta una brecha moderada y constante entre ambas curvas, indicando un buen equilibrio entre bias y varianza.

En conjunto, estos resultados confirman que el Random Forest reduce la varianza del árbol individual, y ofrece la mejor capacidad de generalización entre los modelos probados.