Universidad del Rosario

Analisis Avanzado de Datos

W8. Métodos de Suavización: Splines

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Profesor

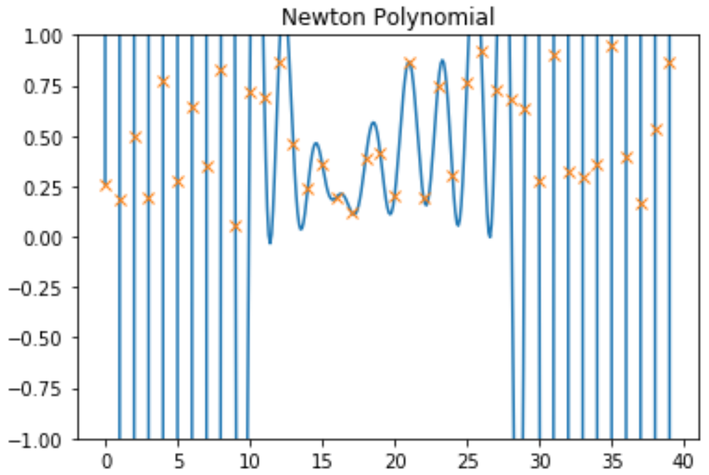
FERNEY ALBERTO BELTRAN MOLINA
ferney.beltran@urosario.edu.co

Ingeniero Electrónico.
Magister en TIC
Candidato Doctor en TIC

Director del Centro de investigación e innovación CEINTECCI.
Miembro de la junta directiva Avanciencia

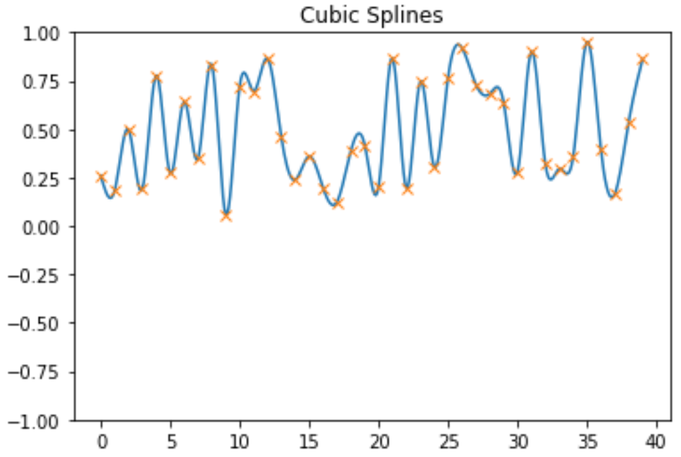
Procesamiento y análisis de datos basadas en IA.
Simulación y modelado por computación,
Optimizan Sistemas de procesamiento en hardware y software
Diseño de sistemas electrónicos reconfigurables

Interpolación



La interpolación **polinómica** cuando aumenta el número de puntos de datos:

- Los polinomios tienden a oscilar entre puntos, lo que aumenta el error entre puntos de datos.
- La extrapolación es muy peligrosa, ya que los polinomios de alto orden varían rápidamente.



Splines cúbicos aborda el problema del sobreajuste de polinomios de alto orden cuando hay muchos puntos de datos.

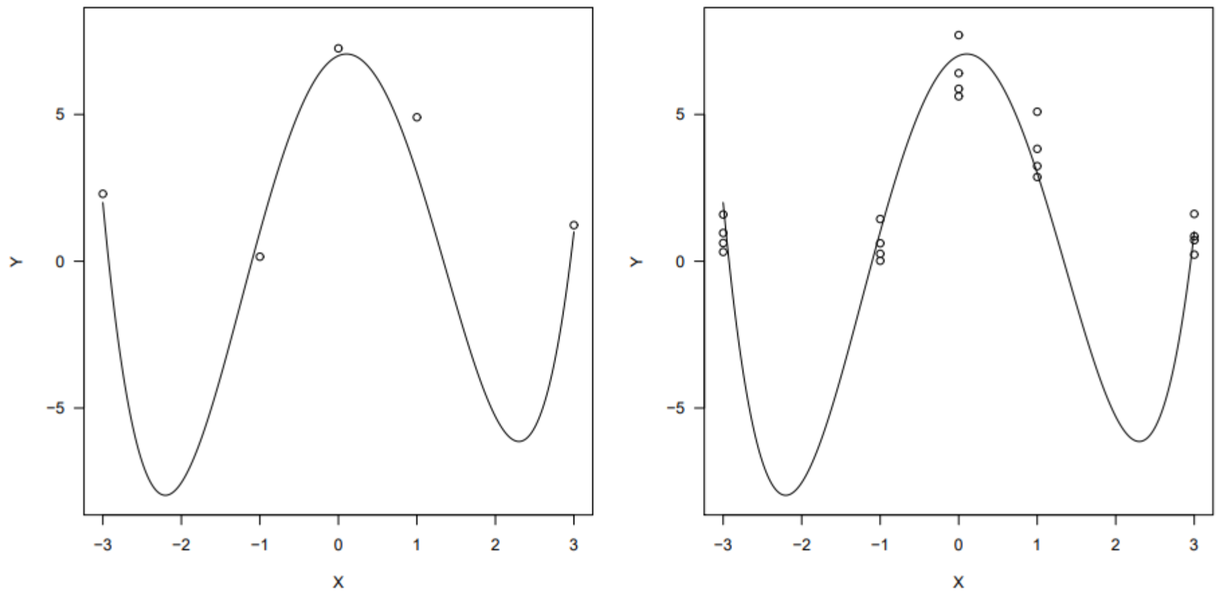
una spline cúbica modela una viga elástica doblada por pasadores ubicados en puntos de datos, también conocidos como nudos

$$f(x) = \begin{cases} f_0(x), & \forall x \in [x_0, x_1] \\ f_1(x), & \forall x \in [x_1, x_2] \\ \dots \\ f_{n-1}(x), & \forall x \in [x_{n-1}, x_n] \end{cases}$$

La interpolación es un tema de estudio del análisis numérico, la suavización spline es un tema estadístico. Lo que haremos será tomar prestadas algunas ideas tomadas de la interpolación par el estudio de la suavización spline.

Estimación

Se sabe que los datos disponibles son mediciones sujetas a error y que aunque se cree que la relación entre X y Y es funcional, no necesariamente los puntos observados están sobre la curva que relaciona las dos variables



El propósito de la estimación es proponer una función f ˆ que permita aproximar f.

Ajuste de mínimos cuadrados para la estimación de spline

Técnica de ajuste de curvas que combina elementos de interpolación y regresión para crear una función que se ajusta de forma óptima a un conjunto de datos.

A diferencia de la interpolación pura, que requiere que la función pase por cada punto de datos, la estimación spline por mínimos cuadrados busca una función que sea un "**buen ajuste**" en términos de minimizar la suma de los cuadrados de las diferencias entre los valores observados y los predichos.

La estimación Spline es un método de regresión no paramétrica que se utiliza para modelar relaciones complejas entre variables. Se basa en la combinación de funciones polinómicas para crear una función más flexible.

Ventajas de la estimación de spline

- **Flexibilidad**: Los splines son muy flexibles y pueden modelar una amplia variedad de formas funcionales.

- **Evita el sobreajuste**: Al utilizar polinomios de bajo grado en cada segmento, se reduce el riesgo de sobreajuste, especialmente en comparación con la interpolación pura o la regresión polinomial de alto grado.

- **Eficiencia computacional**: La optimización generalmente implica un número menor de parámetros que otros métodos, como la regresión polinomial, lo que puede hacer que los cálculos sean más eficientes

Problema guiado

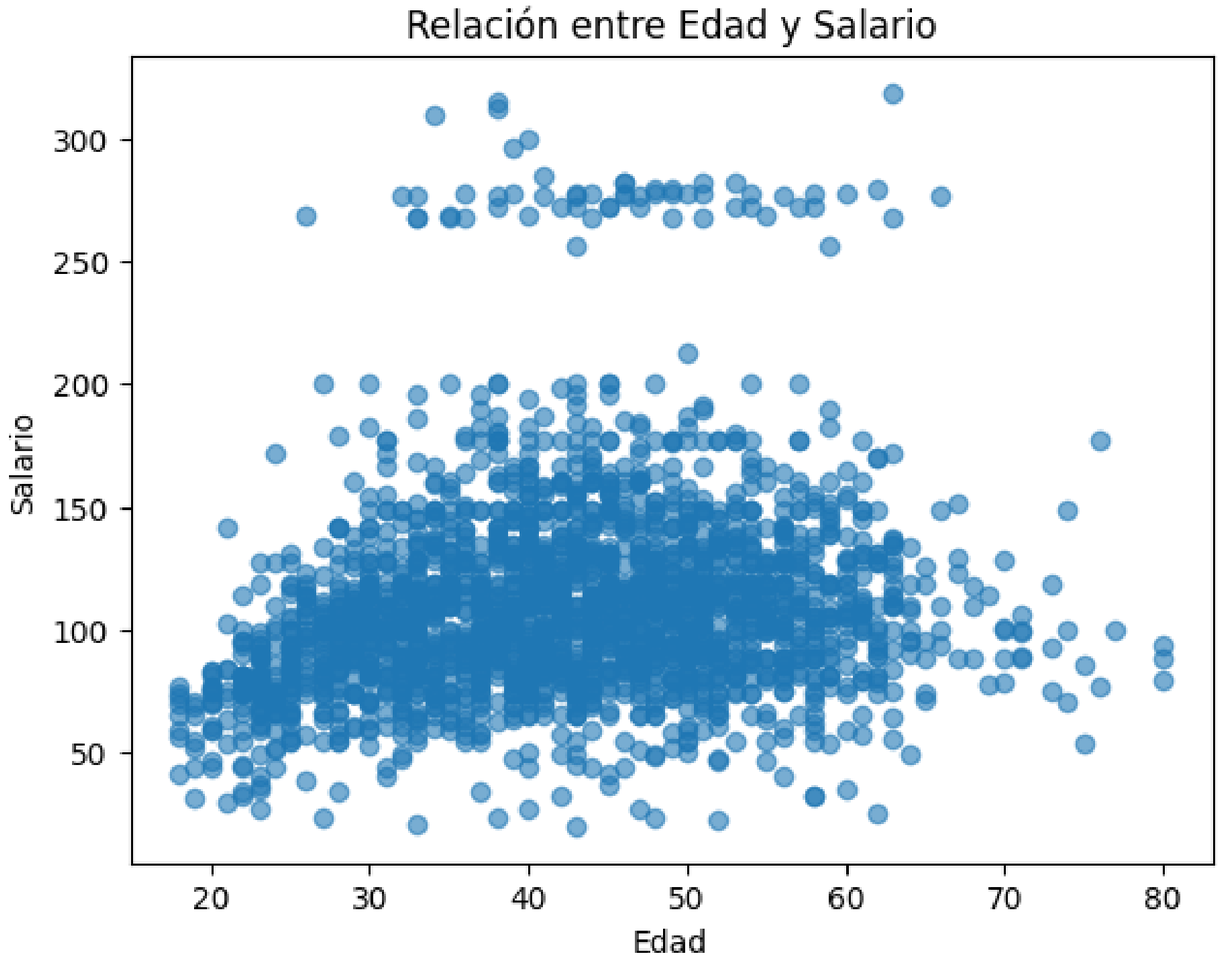
¿Cómo se relaciona la edad de un individuo con su salario en el conjunto de datos proporcionado? Específicamente, ¿es posible predecir el salario de un individuo en función de su edad utilizando técnicas de ajuste de curva como la estimación spline?

1 Calcula con regresión lineal
2 Calcula con spline

Para ellos usa el notebook AAD_lab_spline

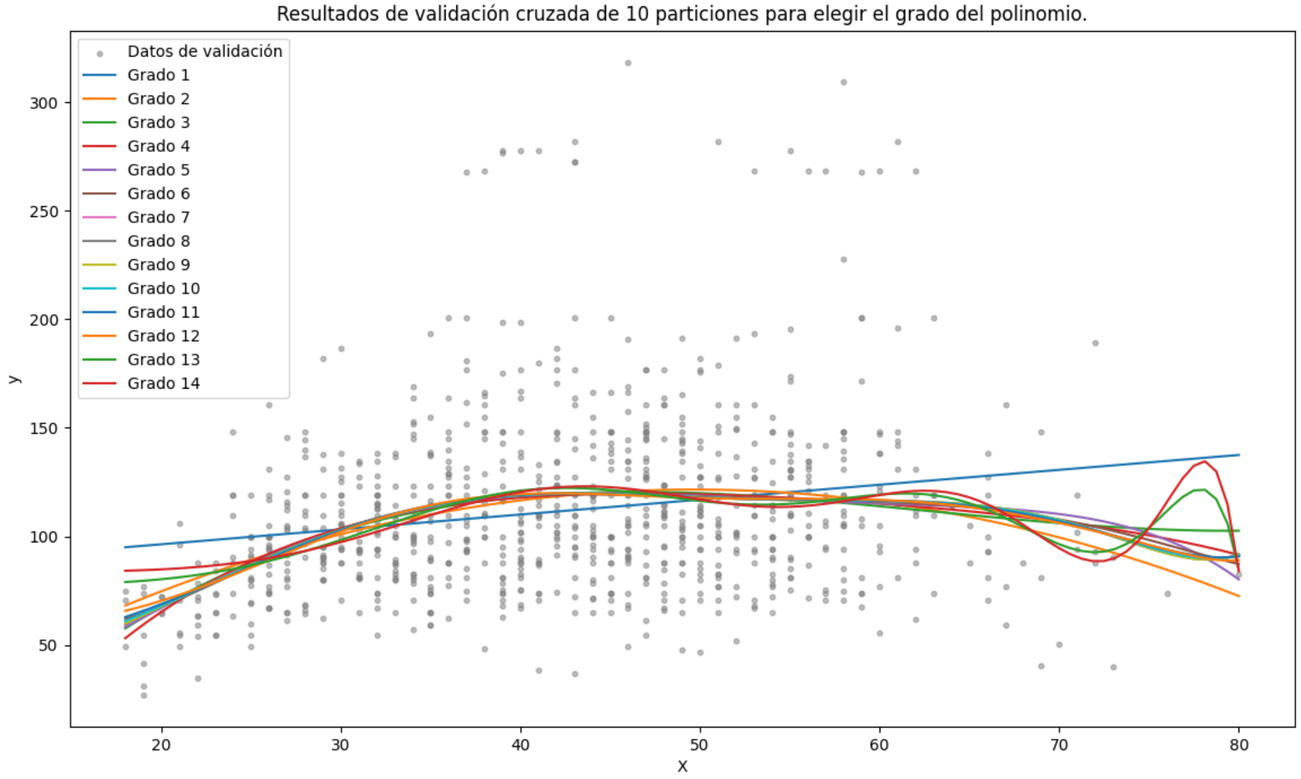
Problema guiado

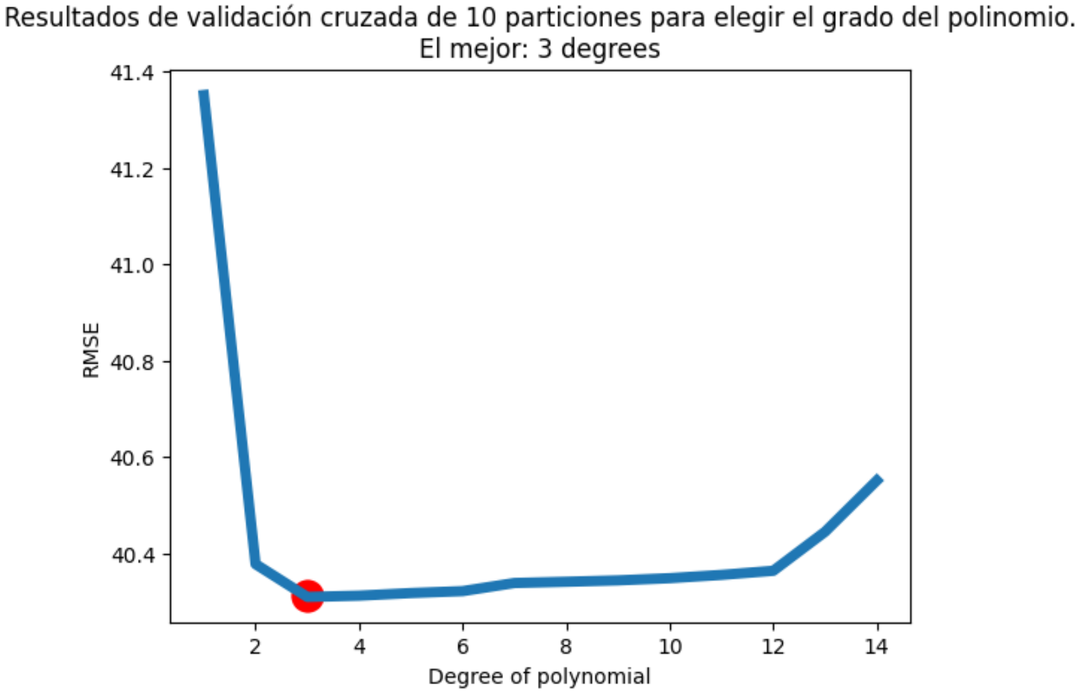
¿Cómo se relaciona la edad de un individuo con su salario en el conjunto de datos proporcionado? Específicamente, ¿es posible predecir el salario de un individuo en función de su edad utilizando técnicas de ajuste de curva como la estimación spline?



Problema guiado

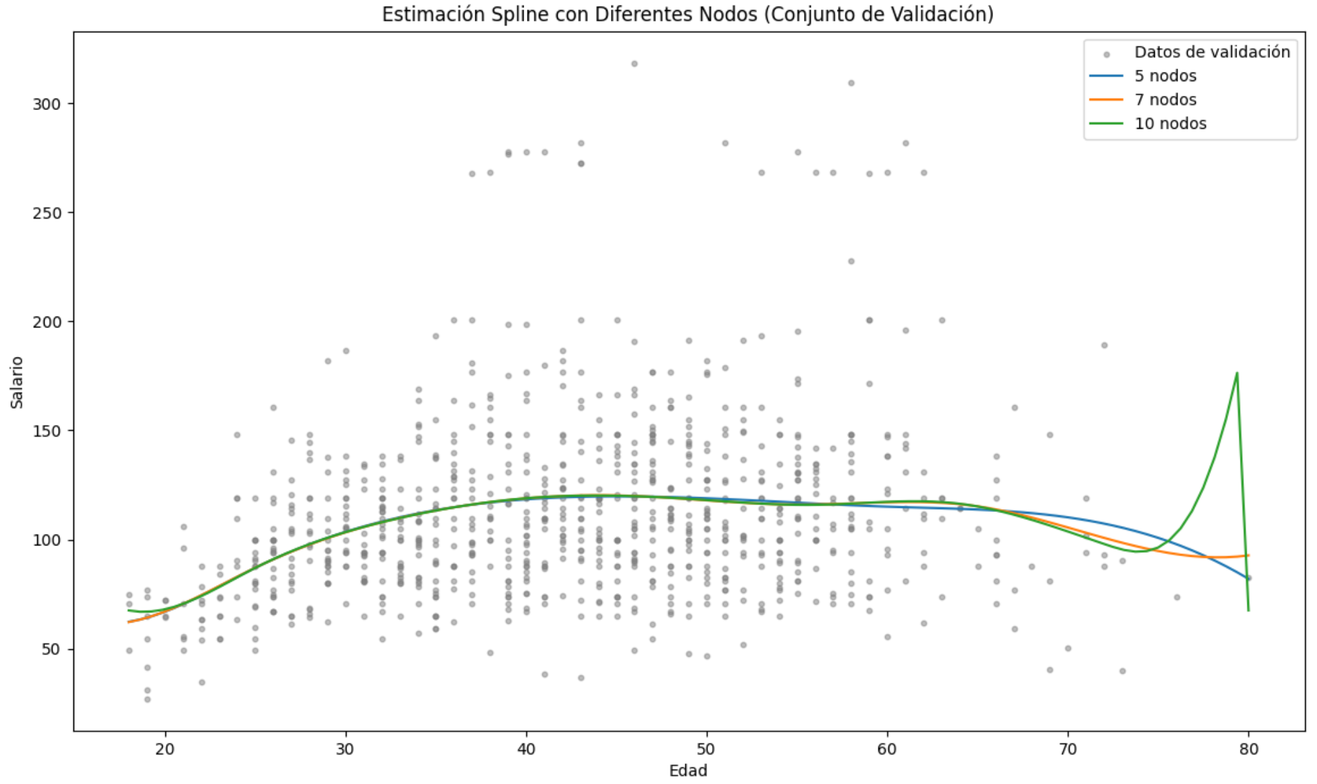
Realizaremos una regresión polinomial, variando el grado del polinomio de 1 a 15 Esto nos permitirá comparar cómo cada modelo se ajusta a los datos y qué grado del polinomio podría ser el más adecuado para describir la relación.





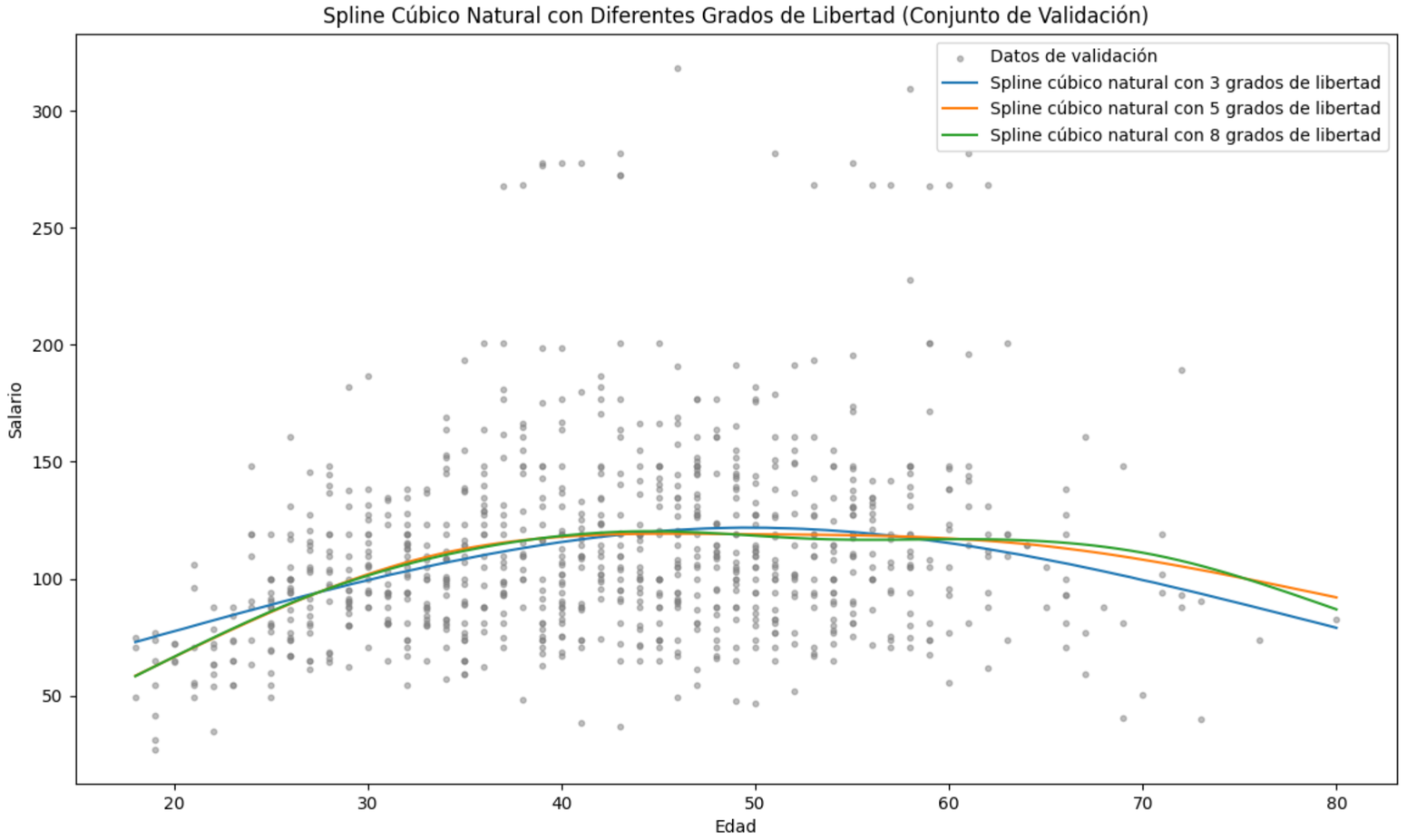
Problema guiado

vamos a aplicar un enfoque diferente: la estimación spline. Este método nos permite tener más flexibilidad en el ajuste del modelo, especialmente en conjuntos de datos que presentan cambios no lineales y complejidades que los modelos polinomiales pueden no capturar de manera eficiente.



Problema guiado

vamos a aplicar un enfoque diferente: la estimación spline. Este método nos permite tener más flexibilidad en el ajuste del modelo, especialmente en conjuntos de datos que presentan cambios no lineales y complejidades que los modelos polinomiales pueden no capturar de manera eficiente.



Ejercicio

En el estudio de los factores que afectan los salarios, es esencial considerar múltiples variables que podrían tener un impacto en los ingresos de un individuo. Aunque hemos examinado previamente el efecto de la edad, otro factor crucial es el nivel de educación. En nuestro conjunto de datos, la variable "educación" se categoriza en cinco niveles distintos, desde menos que un graduado de secundaria hasta un grado avanzado. Dado que tanto la edad como la educación son variables significativas que podrían influir en el salario, sería interesante explorar un modelo de regresión con splines que incorpore ambas variables.

```
from patsy import dmatrix
from sklearn.linear_model import LinearRegression

# Generar matriz de diseño para la primera variable X1 usando splines
X1_spline = dmatrix("bs(X1, degree=3, include_intercept=False)", {"X1": df['X1']}, return_type='dataframe')

# Generar matriz de diseño para la segunda variable X2 usando splines
X2_spline = dmatrix("bs(X2, degree=3, include_intercept=False)", {"X2": df['X2']}, return_type='dataframe')

# Concatenar las dos matrices de diseño
X_spline_combined = pd.concat([X1_spline, X2_spline], axis=1)

# Ajustar el modelo de regresión lineal
model = LinearRegression()
model.fit(X_spline_combined, df['y'])
```