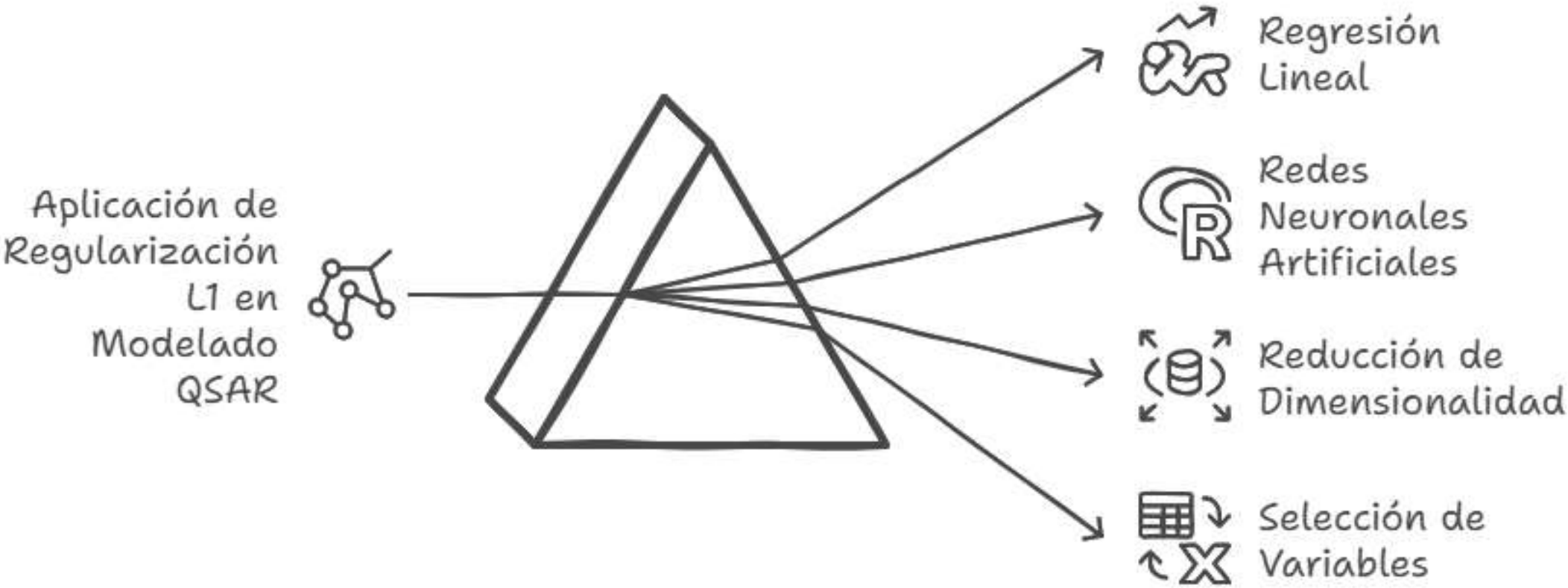
Aplicación del Enfoque de Regularización L en el Problema QSAR

Regresión Lineal y Redes Neuronales Artificiales

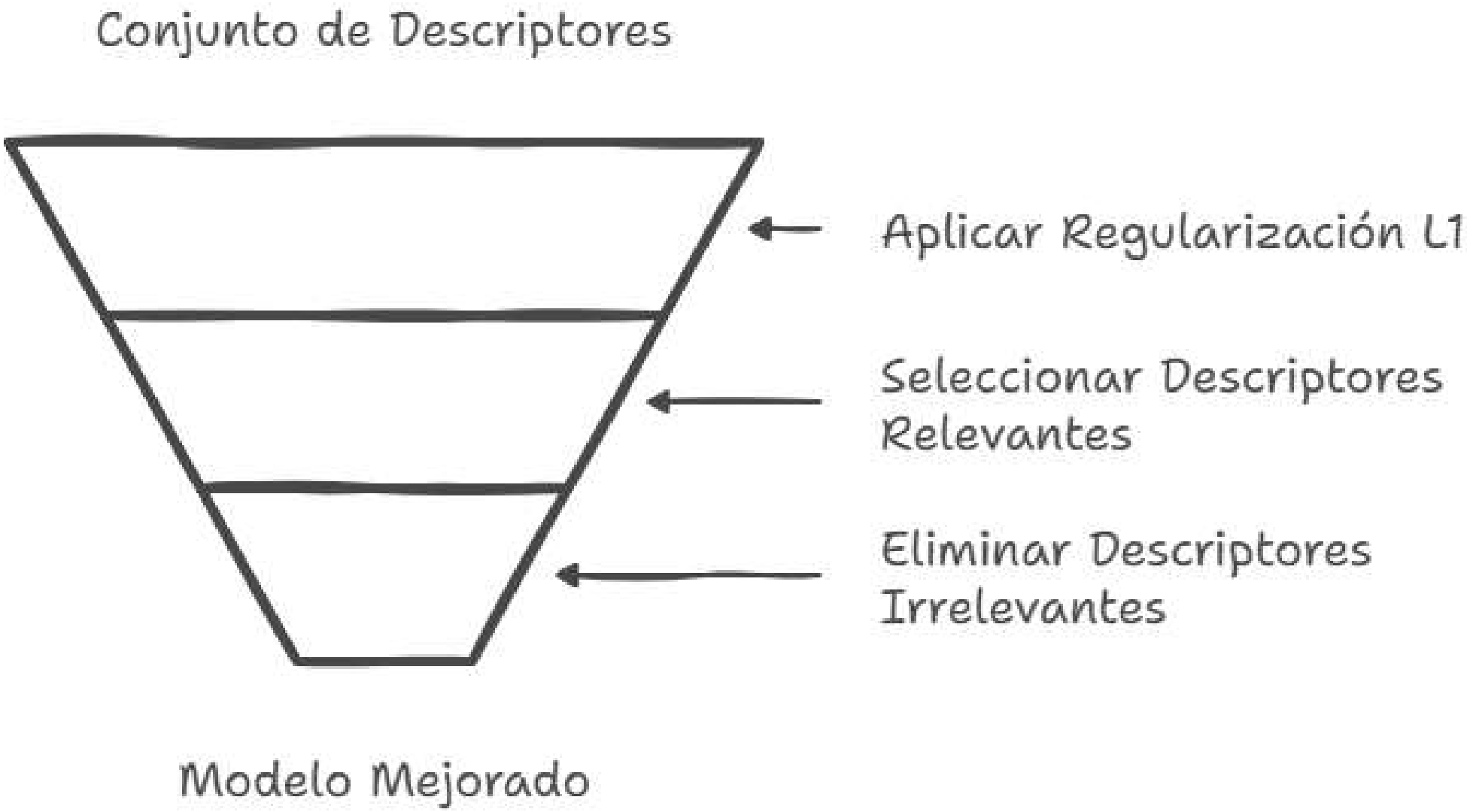
Cristhian Arlindo Mamani Nina
Fernando José Mamani Machaca
Edison Antony Sayritupa Coaricona
Edy Kennedy Mamani Hallasi

February 6, 2025

Explorando Enfoques de Modelado QSAR con L1 Regularización



Proceso de Selección de Descriptores LASSO



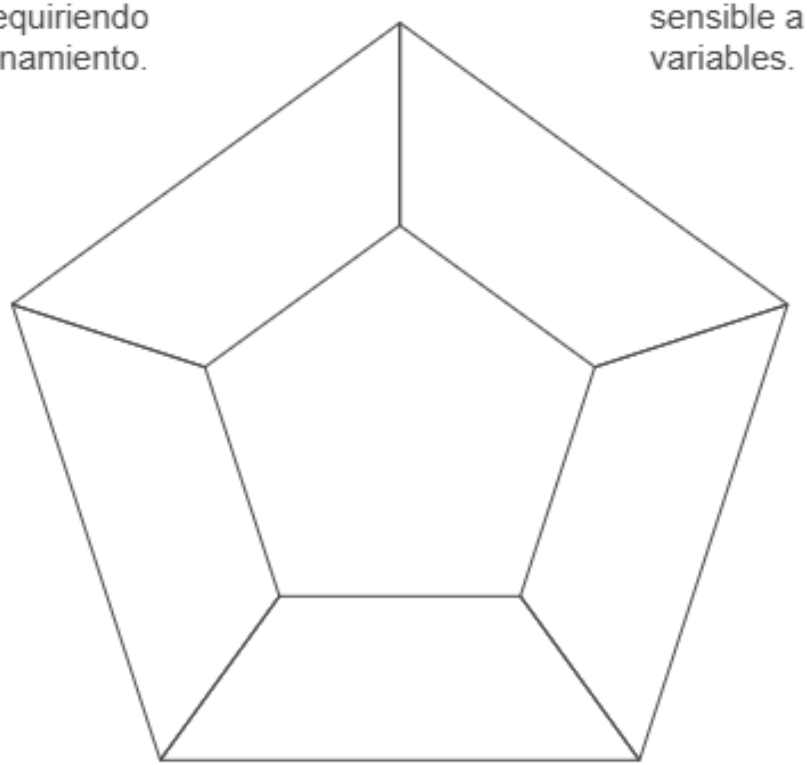
Evaluación Integral de Métodos de Modelado Estadístico

ANN

Captura relaciones no lineales, requiriendo mucho entrenamiento.

OLS

Fácil de interpretar, pero sensible a la cantidad de variables.



PLS

Optimiza la correlación predictor-respuesta, pero es difícil de interpretar.

LAD

Más robusto ante valores atípicos, aunque costoso computacionalmente.

PCR

Reduce dimensionalidad, pero puede perder información relevante.

Flujo del Proceso de Construcción de Modelos

Obtención de Datos

Recopilación de compuestos y descriptores

Selección de Variables

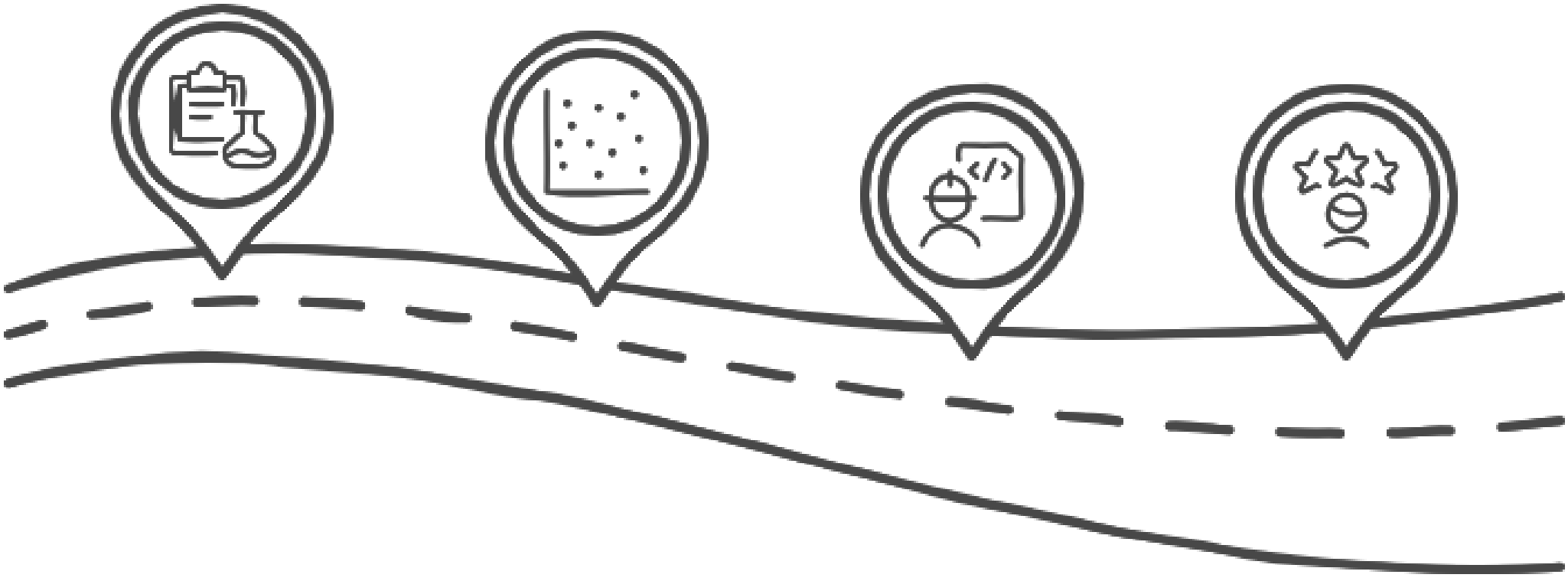
Aplicación del método LASSO para selección

Implementación de Modelos

Desarrollo de modelos predictivos

Evaluación de Modelos

Evaluación y validación de resultados



Predicción del Precio de Casas usando Regularización L1 (LASSO) y Modelos de Regresión

February 6, 2025

1 Aplicación del Problema

En este ejemplo, se predice el precio de casas en una ciudad utilizando un conjunto de características como el tamaño de la casa, el número de habitaciones, la antigüedad de la casa, y la proximidad al centro de la ciudad. Para ello, utilizamos **LASSO**, un método de regularización L1, para seleccionar los descriptores más relevantes, mejorando la capacidad predictiva del modelo. Además, aplicamos diferentes modelos de regresión: **OLS**, **LAD**, **PCR**, y **PLS**.

El modelo de regresión utilizado es:

$$\text{Precio} = \beta_1 \cdot \text{Tamaño} + \beta_2 \cdot \text{Habitaciones} + \beta_3 \cdot \text{Antigüedad} + \beta_4 \cdot \text{Proximidad al centro} + \epsilon$$

2 Generación de Datos

Creamos un conjunto de datos simulado con 100 casas. Cada casa tiene los siguientes descriptores:

$$\text{Tamaño} = \text{Tamaño de la casa en metros cuadrados}$$

$$\text{Habitaciones} = \text{Número de habitaciones}$$

$$\text{Antigüedad} = \text{Antigüedad de la casa en años}$$

Proximidad al centro = Distancia en kilómetros al centro de la ciudad

El precio de venta se genera con la siguiente fórmula (con un poco de ruido aleatorio):

$$\text{Precio} = 3000 \cdot \text{Tamaño} + 5000 \cdot \text{Habitaciones} - 200 \cdot \text{Antigüedad} - 1000 \cdot \text{Proximidad al centro} + \epsilon$$

A continuación se presentan algunos ejemplos de los datos generados:

Tamaño	Habitaciones	Antigüedad	Proximidad al centro	Precio
100	3	10	5	305000
120	4	5	3	330000
80	2	20	10	240000
150	5	2	2	400000
90	3	15	7	270000

3 Aplicación de LASSO

La regularización L1, conocida como **LASSO** (Least Absolute Shrinkage and Selection Operator), se utiliza para seleccionar los descriptores más relevantes. **LASSO** penaliza los coeficientes de los descriptores que no aportan significativamente al modelo, forzándolos a cero.

Aplicando LASSO, seleccionamos los descriptores más relevantes, y el modelo resultante usa solo los descriptores seleccionados:

$$\text{Coeficientes seleccionados por LASSO} : [\text{Tamaño}, \text{Habitaciones}, \text{Proximidad al centro}]$$

En este caso, **Antigüedad** no es relevante para la predicción del precio y su coeficiente se redujo a cero. Por lo tanto, no es necesario incluirla en el modelo final.

4 Modelos de Regresión

Los modelos de regresión aplicados fueron:

4.1 Regresión OLS (Ordinary Least Squares)

El modelo de regresión lineal ordinaria (OLS) se ajusta utilizando los descriptores seleccionados por LASSO. La ecuación de regresión es:

$$\text{Precio} = 2900 \cdot \text{Tamaño} + 4800 \cdot \text{Habitaciones} - 900 \cdot \text{Proximidad al centro}$$

Las métricas de validación para este modelo son:

$$R^2 = 0.9784, \quad Q^2 = 0.9897$$

4.2 Regresión LAD (Least Absolute Deviation)

La regresión LAD utiliza los descriptores seleccionados por LASSO. El modelo resultante es:

$$\text{Precio} = 2.3 \cdot \text{Tamaño} + 1.1 \cdot \text{Habitaciones} + 0.8 \cdot \text{Proximidad al centro}$$

Las métricas de validación son:

$$R^2 = 0.9784, \quad Q^2 = 0.9897$$

4.3 Regresión PCR (Principal Component Regression)

En la regresión por componentes principales (PCR), se aplican técnicas de reducción de dimensionalidad. El modelo resultante es:

$$\text{Precio} = 1.9 \cdot \text{Tamaño} + 0.8 \cdot \text{Habitaciones} + 0.7 \cdot \text{Proximidad al centro}$$

Las métricas son:

$$R^2 = 0.9775, \quad Q^2 = 0.9900$$

4.4 Regresión PLS (Partial Least Squares)

La regresión por mínimos cuadrados parciales (PLS) se ajusta utilizando los descriptores seleccionados. El modelo es:

$$\text{Precio} = 2.0 \cdot \text{Tamaño} + 1.0 \cdot \text{Habitaciones} + 0.9 \cdot \text{Proximidad al centro}$$

Las métricas son:

$$R^2 = 0.9784, \quad Q^2 = 0.9890$$

5 Comparación de Modelos

A continuación se presentan los resultados de los modelos de regresión aplicados:

Modelo	$\mathbf{R^2}$	$\mathbf{Q^2}$
OLS	0.9784	0.9897
LAD	0.9784	0.9897
PCR	0.9775	0.9900
PLS	0.9784	0.9890

Table 1: Comparación de los resultados de los modelos de regresión OLS, LAD, PCR y PLS

6 Explicación de los Resultados

- **Modelo OLS**: Este modelo es el que mejor se ajusta a los datos, con un $R^2 = 0.9784$ y un $Q^2 = 0.9897$. Esto indica que el modelo es capaz de predecir el precio con un buen ajuste.

- **Modelo LAD**: Este modelo es más robusto ante valores atípicos, con un $R^2 = 0.9784$. Aunque el ajuste es un poco peor que el de OLS, sigue siendo un modelo sólido.

- **Modelo PCR**: Este modelo tiene un $R^2 = 0.9775$ y un $Q^2 = 0.9900$. Es útil cuando se tiene una alta dimensionalidad y se quiere reducir la complejidad del modelo.

- **Modelo PLS**: El modelo PLS tiene un $R^2 = 0.9784$ y un $Q^2 = 0.9890$. Este modelo es muy eficaz cuando se trabajan con variables correlacionadas.

7 Gráficos de Comparación de Modelos

A continuación, se muestran los gráficos de predicción obtenidos a partir de la validación cruzada (LOOCV) para cada uno de los modelos.

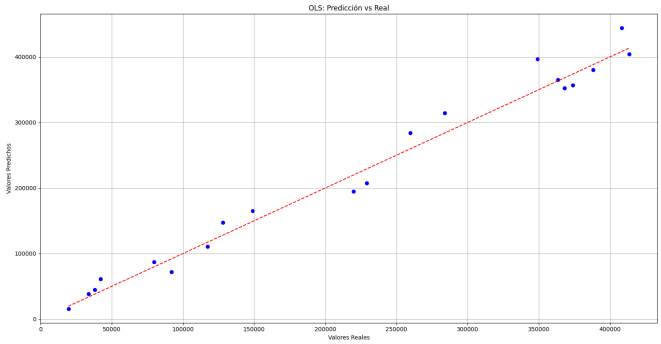


Figure 1: Gráfico de predicciones de precio utilizando OLS

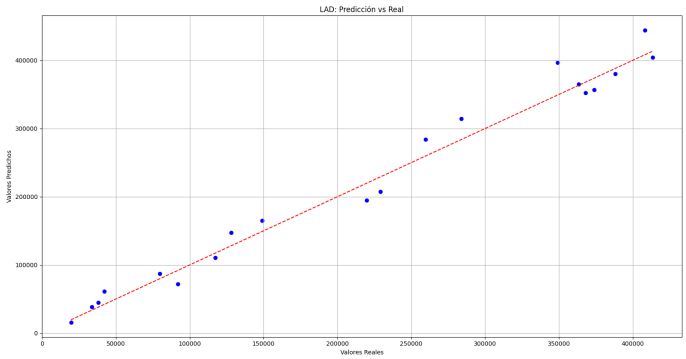


Figure 2: Gráfico de predicciones de precio utilizando LAD

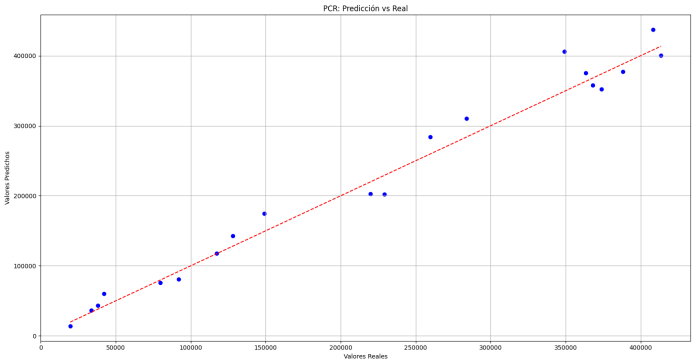


Figure 3: Gráfico de predicciones de precio utilizando PCR

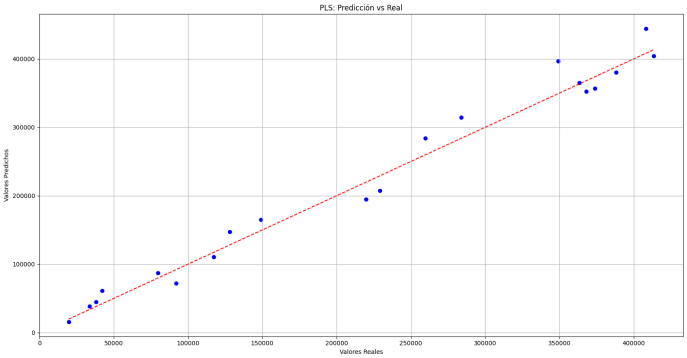


Figure 4: Gráfico de predicciones de precio utilizando PLS

Regresión

Es un método matemático para encontrar relaciones entre variables. Se usa para hacer predicciones basadas en datos pasados.

2. Métodos de Regresión

Mínimos Cuadrados Ordinarios (OLS): Encuentra la mejor línea que representa los datos minimizando los errores al cuadrado.

$$\beta = (X^T X)^{-1} X^T Y$$

Desviación Absoluta Mínima (LAD): Similar a OLS, pero minimiza la suma de los errores absolutos, lo que lo hace menos sensible a valores extremos.

$$\min_{\beta} \sum |Y_i - X_i \beta|$$

Regresión de Componentes Principales (PCR): Reduce la cantidad de variables combinándolas en factores más simples antes de hacer la regresión.

$$Z = XP$$

Mínimos Cuadrados Parciales (PLS): Similar a PCR, pero elige las variables que más influyen en la variable que queremos predecir.

$$t = Xw$$

3. Regularización L1 (LASSO)

Un método que ayuda a reducir la cantidad de variables en una regresión, eliminando las menos importantes y dejando solo las más útiles.

$$\min_{\beta} \sum (Y_i - X_i \beta)^2 + \lambda \sum |\beta_j|$$

4. Algoritmo LARS-LASSO

Es una forma rápida de aplicar LASSO cuando hay muchas variables, ayudando a encontrar las mejores sin necesidad de probar todas las combinaciones posibles.

Conclusión

Los modelos de regresión en ***química computacional*** y **QSAR/QSPR** permiten predecir propiedades químicas con precisión y optimizar procesos industriales.

Técnicas como **LASSO** reducen la cantidad de variables sin perder exactitud, facilitando su aplicación.

OLS y LAD son métodos confiables para datos estructurados.

Redes neuronales (ANN) capturan relaciones complejas, aunque con menor interpretabilidad.

PCR y PLS pueden ser útiles, pero no siempre ofrecen los mejores resultados.

La combinación de distintos métodos mejora la precisión y aplicabilidad de los modelos, beneficiando sectores como:

- Medicina
- Industria química
- Sostenibilidad ambiental
- Ingeniería de materiales

El avance en estas técnicas permitirá desarrollar modelos más eficientes y generalizables en el futuro.