Regresión Lineal y Bayesiana

INTRODUCCIÓN A CIENCIA DE DATOS

13 de Octubre de 2025

Jessica Rubí Lara Rosales jessica.lara@cimat.mx
Rodrigo Gonzaga Sierra rodrigo.gonzaga@cimat.mx

1. Introducción

En este trabajo se presentara un análisis comparativo y complementario a dos artículos que usan regresión lineal y regresión logística respectivamente. Debido a nuestro principal interés en las áreas de biología y ciencias ambientales. Hemos seleccionado el articulo 'Development and Validation of Multiple Linear Regression Models for Predicting Chronic Zinc Toxicity to Freshwater Microalgae' [Price et al., 2023] para hacer un análisis de regresión lineal y 'Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy' [Rimal et al., 2025] para hacer un análisis de regresión logística.

2. Regresión lineal

El articulo 'Development and Validation of Multiple Linear Regression Models for Predicting Chronic Zinc Toxicity to Freshwater Microalgae' [Price et al., 2023] desarrolla modelos de regresión lineal multiple para predecir la toxicidad crónica del zinc a una microalga de agua dulce llamada Chlorella sp.

La toxicidad del zinc para los organismos acuático depende de su biodisponibilidad (cantidad disponible que el organismo puede absorber) la cual esta influenciada por parámetros de calidad del agua. Este articulo utiliza los factores de ph, dureza y carbono orgánico disuelto, para determinar la toxicidad del zinc. El articulo realiza un modelo de regresión lineal múltiple considerando estás variables para poder predecir la toxicidad del zinc en función de estos parámetros los datos los obtuvo de otros estudios realizados

2.1. Datos

Estos datos fueron obtenidos de [Price et al., 2023]. Todas las pruebas se llevaron a cabo utilizando aguas sintéticas preparadas en laboratorio con niveles de pH que varían de 6.7 a 8.3, concentraciones de dureza de 5 a 400 mg/L CaCO_3 , y concentraciones de DOC de 0 – 15 mg/L C . Las pruebas se realizaron a una alcalinidad constante y una relación de calcio a magnesio ($\tilde{0}.7$). Los datos de toxicidad utilizados cumplieron con los criterios de aceptabilidad establecidos para su uso en el desarrollo de pautas de calidad del agua en Australia.

- **Authors**: Fuente de datos (Prince y equipo)
- **Test ID**: Identificador único de prueba
- **Number of tests**: Número de test (1 o 2)
- **Species**: Chlorella sp. (tipo de alga)
- **Phylum**: Nivel de clasificación biológica
- **Taxonomic Group**: Grupo taxonómico
- **Life stage**: EX (fase crecimiento exponencial)
- **Endpoint**: -

- **Endpoint Measurement**: Tasa de crecimiento poblacional
- **Toxicity Value**: EC10, EC20, EC50 (reducción del crecimiento)
- **Exposure Duration**: 3 días
- **Conc**: Concentración de zinc (mg/L)
- **Standard Deviation**: Desviación estándar
- **Measured or nominal**: Tipo de medición
- **pH**: pH de la muestra
- **Ca**: Concentración de calcio (mg/L)
- **Mg**: Concentración de magnesio (mg/L)
- **Hardness**: Dureza total (mg/L $CaCO_3$)
- **DOC source**: SI/NO (fuente externa ácido húmico)
- **DOC**: Carbono orgánico disuelto
- **Temp**: Temperatura constante 27°C

De las cuales **Conc, pH, Hardness, DOC, Ca_mgL, Mg_mgL**, **Standard Deviation**, son variables numéricas y el resto son variables categóricas. La variable **Temp**, se podría clasificar como una variable numérica, por representar la temperatura, pero solo tiene un valor constante, así que será tomada como categórica.

El valor de toxicidad se clasifica en las siguientes clases

- **EC10** : Concentración Efectiva del 10 %
- **EC20** : Concentración Efectiva del 20 %
- **EC50** : Concentración Efectiva del 50 %

Esto representa la concentración de una sustancia química (en este caso el zinc), que produce un efecto adverso especifico en un porcentaje dado de la población de organismos que en este caso es Chlorella sp. Nos centraremos en el análisis de las columnas **Conc, pH, Hardness** y **DOC**. Se dispone de 90 observaciones en total, con 30 muestras por cada clase ECX. Debido a las diferencias en el comportamiento observadas entre las clases, implementaremos un modelo de regresión lineal múltiple por separado para cada clase, utilizando los 30 datos correspondientes.

2.2. Análisis

2.2.1. Modelos

En el articulo menciona que se consideraron dos maneras de regresión lineal múltiple, una sin interacción considerando solo variables independientes $\ln$(hardness), $\ln$(DOC) y pH) y otra con interacciones del tipo producto de dos variables independientes, solo se incluyeron dos tipos de interacción $\ln$(hardness)$\cdot$ pH y $\ln$(DOC)$\cdot$ pH, esto debido a que ningún experimento disponible varió simultáneamente estas variables, lo que impide estimar de forma confiable este efecto conjunto. Se realizo un análisis para detectar cual es el modelo de regresión lineal múltiple con mayor desempeño para cada nivel de toxicidad usando varios criterios de ajuste como el de AIC, el BIC, R^2 y R^2-ajustada.

2.2.2. Multicolinealidad

Para detectar multicolinealidad en los datos se usa la técnica de VIF el cual es un análisis de factores de inflación de varianza para detectar multicolinealidad en un modelo de regresión múltiple. Para obtener este valor se selecciona

una variable independiente X_i y se estima una regresión lineal simple con las variables restantes, se calcula R_i^2 y se sustituye en la siguiente expresión

$$VIF = \frac{1}{1 - R_i^2}.$$

Si $VIF < 5$ indica multicolinealidad baja, VIF entre $5 - 10$ multicolinealidad moderada y $VIF > 10$ indica multicolinealidad alta.

2.2.3. Validación

Para realizar la validación del modelo se graficó la toxicidad observada frente a la toxicidad predicha. Esto para visualizar qué tan bien el modelo predice los datos observados, el desempeño se evaluó según el porcentaje de datos observados que caían dentro de un factor 2 o 3 de los valores de toxicidad predichos. Un dato predicho se considera de **Factor 2** si el valor predicho está entra la mitad y el doble del valor observado, el modelo se considera confiable, análogamente para el **Factor 3**.

2.3. Resultado y comparación

2.3.1. Modelo

EL articulo concluyo que los mejores modelos para EC10 y EC20 son:

$$\ln(\text{toxicity}) = \beta_0 + \beta_1 \cdot \ln(\text{hardness}) + \beta_3 \ln(\text{DOC}) + \beta_4 \cdot \text{pH} + \beta_5 \cdot \ln(\text{DOC}) \cdot \text{pH}$$

EC50:

$$\ln(\text{toxicity}) = \beta_0 + \beta_1 \cdot \ln(\text{hardness}) + \beta_2 \cdot \ln(\text{DOC}) + \beta_3 \cdot \text{pH}$$

En nuestro caso se obtuvieron

	sin interacción	ln(DOC) · pH	ln(hardness) · pH	ambos
AIC	56.42	55.42	57.93	57.25
BIC	62.02	62.43	64.94	65.66
R^2	0.370	0.430	0.381	0.433
R^2-ajustada	0.298	0.339	0.281	0.315

Cuadro 1: Comparación de ajuste para EC10

Así que para EC10 del cuadro 1 podemos notar que AIC apunta a que el mejor modelo es considerando la interacción $\ln(\text{DOC}) \cdot pH$ y el BIC me dice que el mejor es sin interacción, pero el R^2 ajustada indica que si es mejor considerar la interacción $\ln(\text{DOC}) \cdot pH$.

	sin interacción	ln(DOC) · pH	ln(hardness) · pH	ambos
AIC	59.99	59.58	61.60	61.45
BIC	65.59	66.58	68.61	69.86
R^2	0.456	0.498	0.463	0.500
R^2-ajustada	0.393	0.417	0.377	0.396

Cuadro 2: Comparación de ajuste para EC20

En el caso de EC20 de cuadro 2 tenemos nuevamente que AIC apunta a que el mejor es con $\ln(\text{DOC}) \cdot pH$ y BIC al de sin interacción, pero el R^2-ajustada nos dice que el mejor en efecto es con interacción.

	sin interacción	ln(DOC) · pH	ln(hardness) · pH	ambos
AIC	74.08	74.74	75.94	76.36
BIC	79.69	81.75	82.94	84.77
R^2	0.497	0.519	0.500	0.525
R^2-ajustada	0.439	0.443	0.420	0.427

Cuadro 3: Comparación de ajuste para EC50

Del cuadro 3 tenemos que para EC50 el AIC y BIC indican que el mejor modelo es sin interacción y el R^2-ajustada reafirma eso pues en ese caso es el segundo mayor.

De ello podemos concluir que en efecto los mejores modelo para predicir son los mencionados en el articulo. Aunque los resultados obtenido de R^2 y R^2-ajustada, apuntan a que no es un buen modelo para estos datos.

Implementando estos modelos obtenemos los siguientes coeficientes.

	Coeficiente	Error Estándar	p-valor
const	0.1651	1.4715	0.9116
pH	-0.0557	0.1938	0.7761
ln(hardness)	0.2884	0.0931	0.0048
ln(DOC)	-2.1429	1.4335	0.1475
pH· ln(DOC)	0.3025	0.1867	0.1176

Cuadro 4: EC10 modelo con interacción seleccionado

	Coeficiente	Error Estándar	p-valor
const	0.1891	1.5770	0.9055
pH	-0.0095	0.2077	0.9640
ln(hardness)	0.4321	0.0998	0.0002
ln(DOC)	-2.1135	1.5363	0.1811
pH· ln(DOC)	0.2891	0.2000	0.1608

Cuadro 5: EC20 modelo con interacción seleccionado

	Coeficiente	Error Estándar	p-valor
const	2.1753	1.9981	0.2863
pH	-0.1377	0.2615	0.6029
ln(hardness)	0.6431	0.1284	0.0000
ln(DOC)	-0.0291	0.1242	0.8168

Cuadro 6: EC50 modelo sin interacción seleccionado

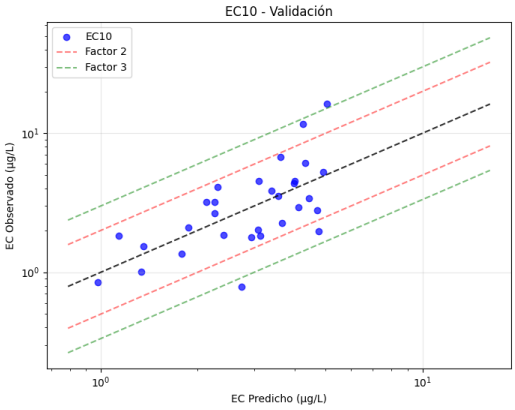
Los cuales coinciden justamente con los valores obtenidos en el articulo.

2.3.2. Multicolinealidad

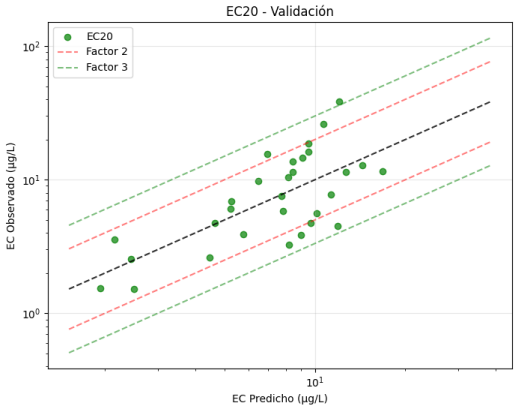
Los valores del VIF (Factor de Inflación de la Varianza) para el pH, ln(hardness) y ln(DOC) en todos los modelos cuando no se considera interacción es un valor alto lo que sugiere que existe cierta multicolinealidad entre las variables. De ello que una posible propuesta de modelo sea la regresión Ridge.

2.3.3. Criterio de desempeño predictivo

Realizando predicciones y comparandolos con los valores observados. Graficamos en el eje X los valorse que el modelo calcula y en el eje Y los valores reales medidos en los experimentos y graficamos las lineas que representan el Factor 2 y 3. Se obtienen los siguinetes resultados que son muy similares al articulo.



(a) EC10 comparación de predicciones



(b) EC20 comparación de predicciones

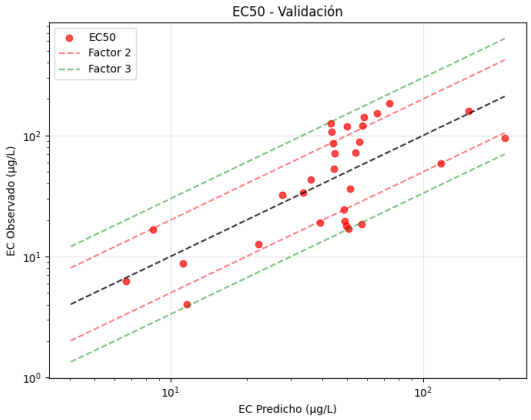


Figura 2: EC50 comparación de predicciones

De las gráficas podemos notar como en negral las predicciones no son muy buenas pues hay algunos puntos que están muy cerca de las bandas o se salen como es el caso para EC10 y EC20. Aunque en EC50 no se salgan, los puntos no se encuentran tan cercanos a la recta identidad.

2.4. Propuesta y análisis extra

En general los resultado obtenidos nos dicen que este modelo no resulta ser bueno para predecir el nivel de toxicidad de zinc en el agua. Esto puede ser porque se consideran muy pocas variables predictivas (independientes) que expliquen la variabilidad de la variable de respuesta. O bien los datos no tienen patrones lineales.
De igual manera realizaremos un análisis extra con modelos robustos de regresion lineal como los son Regresion Ridge, Lasso, Hube y un modelo robusto M-estimadores. Proponemos estos nuevos modelos debido a que el VIF

obtenido para los modelos que no tienen interacción resultó estar algo elevado para algunas variables. Entonces penalizaremos la multicolinealidad con la Regresión Ridge. Además implementaremos regresión huber (scikit-learn) el cual combina M-estimadores con regularización L2, usa la función de pérdida de Huber la cual resulta ser menos sensible a los outliers.
Graficamos los residuales contra los observados de cada modelo y clase de toxicidad para verificar la homocedasticidad de los datos, también para saber si no existe un patron no lineal evidente.

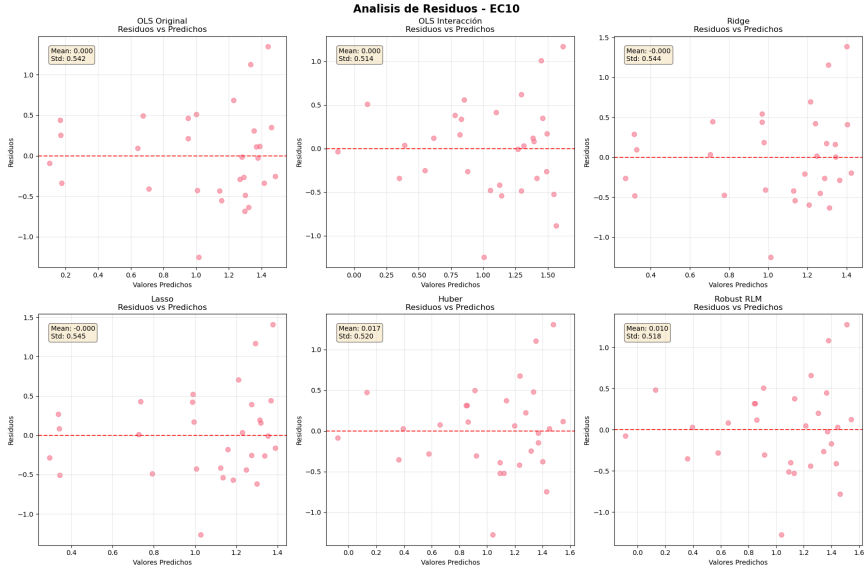


Figura 3: Análisis de residuos de distintos modelos implementados para los datos de toxicidad EC10

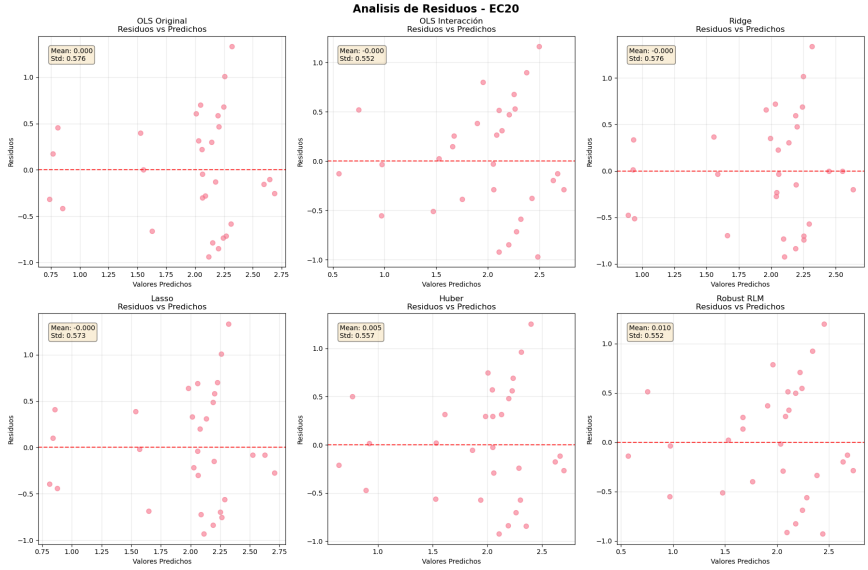


Figura 4: Análisis de residuos de distintos modelos implementados para los datos de toxicidad EC20

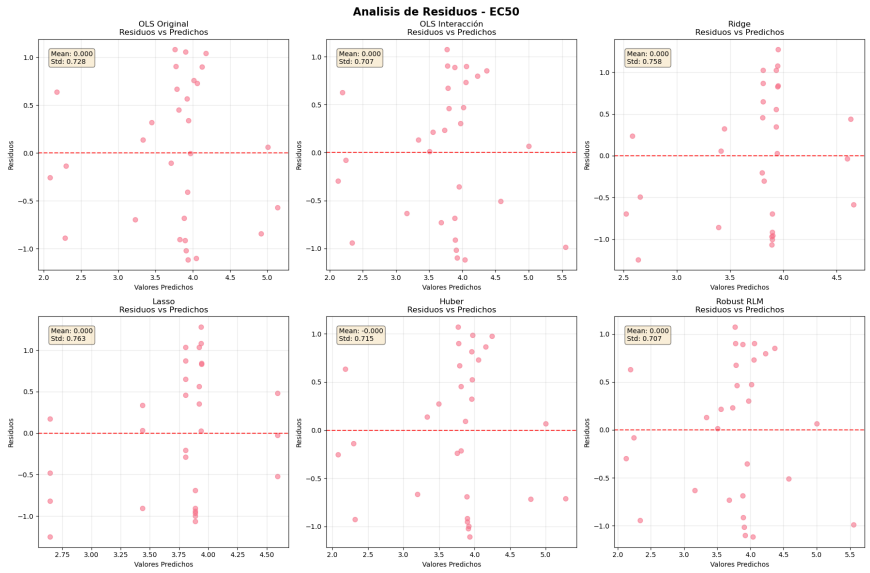


Figura 5: Análisis de residuos de distintos modelos implementados para los datos de toxicidad EC50

En general, los residuos de todos los modelos se distribuyen de forma aproximadamente simétrica alrededor de cero con ciertas agrupaciones en algunas partes pero sin patrones evidentes, lo que sugiere una especificación adecuada del modelo lineal. Sin embargo, la dispersión observada y el R^2 ajustado de los modelos que se encuentra entre 0.3 y 0.5 es relativamente bajo lo que indican que existe una cantidad considerable de variabilidad no explicada. Los métodos robustos (Huber y RLM) presentan ligeras mejoras en la estabilidad de los residuos, lo que sugiere que podrían ser más apropiados en presencia de valores atípicos, aunque las diferencias globales respecto a OLS son pequeñas.

2.5. Conclusiones

El análisis de regresión lineal múltiple realizado replicando el estudio de [Price et al., 2023] permitió evaluar la capacidad predictiva de modelos lineales para estimar la toxicidad crónica del zinc en *Chlorella sp.* en función de pH, dureza y carbono orgánico disuelto (DOC). Los resultados obtenidos fueron consistentes con los reportados en el artículo original, tanto en la selección de modelos como en la estimación de coeficientes.

Como se mencionó anteriormente los valores de R^2-ajustado fueron bajos (entre 0.3 y 0.5). El R^2 de predicción varió entre 0.3 y 0.7. Tanto en el articulo como el los resultados replicados se tiene que la inclusión de términos de interacción no mejoró significativamente los ajustes. Los modelos no se sobreajustaron (ya que los R^2 ajustado y predicho son similares); Además, los valores de R^2-ajustada entre 0.3 y 0.5 revelan que los modelos explican solo una parte moderada de la variabilidad en la respuesta toxicológica. Esto sugiere que, si bien las variables consideradas son relevantes, existen otros factores no incluidos en el modelo que influyen en la toxicidad del zinc, o bien que la relación entre las variables no es completamente lineal.

La validación predictiva mostró que un porcentaje significativo de las observaciones se encuentra dentro de los factores 2 y 3 de predicción, lo que respalda la utilidad del modelo en contextos aplicados, aunque con márgenes de error considerables. Adicionalmente, la presencia de multicolinealidad moderada detectada mediante VIF motivó la exploración de métodos robustos como Ridge, Lasso y Huber, los cuales mostraron ligeras mejoras en el manejo de valores atípicos, pero sin incrementar sustancialmente el poder predictivo.

En conclusión, los modelos de regresión lineal propuestos ofrecen una base útil para la predicción de toxicidad de zinc en condiciones controladas, pero su rendimiento limitado recomienda complementarlos con técnicas más flexibles o con la inclusión de variables adicionales que capturen la complejidad bioquímica y ambiental del fenómeno.

3. Regresión Logística

El articulo 'Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy' [Rimal et al., 2025] discute e implementa tecnicas de validación para aplicaciones del aprendizaje automático en la predicción de enfermedades cardíacas. En general su objetivo es mejorar la precisión del diagnostico de enfermedades cardíacas usando modelos de aprendizaje automático. Para eso se plantea comparar la eficiencia de cuatro modelos, Regresión Logística, SVM, K -NN y random forest usando validación cruzada.

3.1. Datos

Se trabajara con un conjunto de datos público que se encuentra en el articulo [Rimal et al., 2025] el cual tienen 303 registros (pacientes) con 14 variables relacionadas con enfermedades cardíacas. Las columnas son las siguientes

- **age**: Edad del paciente en años
- **sex**: Sexo biológico (0=Femenino, 1=Masculino)
- **cp**: Tipo de dolor torácico
- **trestbps**: Presión arterial en reposo (mm Hg)
- **chol**: Nivel de colesterol LDL o HDL (mg/dl)
- **fbs**: Azúcar en ayunas $>$ 120 mg/dl (0=No, 1=Sí)
- **restecg**: Resultado electrocardiográfico en reposo
- **thalach**: Frecuencia cardíaca máxima alcanzada
- **exang**: Angina inducida por ejercicio (0=No, 1=Sí)
- **oldpeak**: Depresión ST inducida por ejercicio
- **slope**: Pendiente del segmento ST en ejercicio máximo
- **ca**: Número de vasos principales coloreados por fluroscopia (0-3)
- **thal**: Tipo de prueba de talio (1,2,3)
- **target**: Diagnóstico de enfermedad cardíaca (0=No, 1=Sí)

Los valores detallados de la variables categóricas son:

- **cp:** 0 = Angina típica, 1 = Angina atípica, 2 = Dolor no anginoso, 3 = Asintomático
- **restecg:** 0 = Normal, 1 = Anormalidad de Onda ST-T, 2 = Hipertrofia ventricular izquierda
- **slope:** 0 = Ascendente, 1 = Plana, 2 = Descendente
- **thal:** 1 = Normal, 2 = Defecto fijo, 3 = Defecto reversible

Lo datos numéricos son age, trestbps, chol, thalach, oldpeak y los categóricos son el complemento.

Realizando primero un preprocesamiento haciendo codificación One-Hot Encoding resulta en un total de 23 columnas de datos. Se realiza una normalización con StandardScaler para estandarizar las variables numéricas. Esta

normalización es crucial en los modelos de Regresión Logística SVM. Normalizando se mide que variables con valores grandes como lo es el colesterol, domine el calculo en las distancias.

Además dividiremos los datos en 80 % train y 20 % test. Se usa stratify para que la proporción de clases objetivo (0 o 1) se mantenga 50:50, en el conjunto de entrenamiento y el de prueba.

3.2. Resultados y comparación

3.2.1. Correlación

Primeramente realizamos un mapa de calor figura 6 para describir las correlaciones entre las variables dependientes e independientes. Esto nos permite visualizar cuales variables están significativamente relacionadas entre si Es de

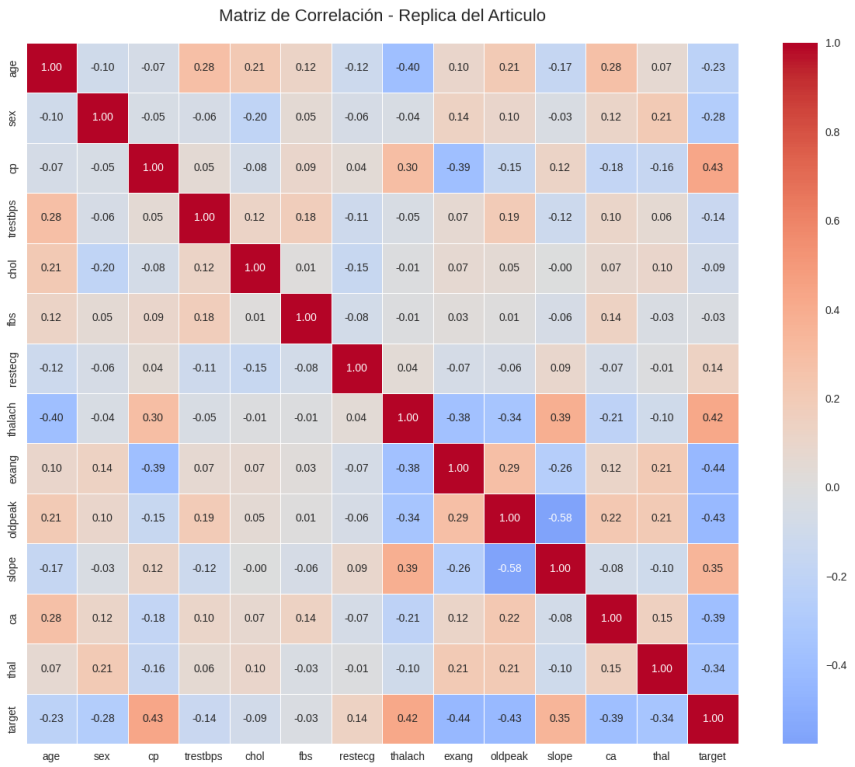


Figura 6: Matriz de correlación entre las variables

principal interés verificar la correlación con la ultima fila/columna. Las correlaciones más fuertes son las de **cp** con 0.43 lo cual implica que a medida que el tipo de dolor de pecho se acerca a la categoría se acerca a valores altos de cp la probabilidad de enfermedad cardíaca aumenta significativamente. Análogamente para thalach con 0.42, exang con -0.44 y oldpeak con -0.43. En estos últimos dos carros la correlación es negativa, lo que indica que para valores menores es más probable que el target sea 1. Como por ejemplo en exang si un paciente no experimenta angina inducida por el ejercicio exang = 0, hay mayor probabilidad de tener target=1.

3.2.2. Implementación de métodos

Implementando los métodos de regresión logística, K-NN, SVM (Maquina de Vectores de Soporte) y Random Forest que fueron realizados en el articulo, usando 80 % train y 20 % test y con semilla 42 como se menciona en el articulo obtenemos los siguientes resultados

	Accuracy	Sensibilidad	Especificidad	F1
Regresión Logística	0.869	0.909	0.821	0.882
K-NN	0.787	0.818	0.750	0.806
SVM	0.820	0.939	0.679	0.849
Random Forest	0.820	0.879	0.750	0.841

Cuadro 7: Ajuste de los modelos

Aunque usamos la misma semilla y los datos proporcionados en el github del articulo, obtuvimos resultados diferentes. Hay una parte del articulo que menciona que los datos después de codificar son 76 lo cual no coincide con las 23 columnas que nosotros obtuvimos. Entonces suponemos que por eso es el error. Igual replicaremos los análisis extras que se implementaron en el articulo.
De esta tabla podemos notar como la regresión logística resulta tener el mejor desempeño en general. Pero como nosotros estamos interesando en la sensibilidad pues un falso negativo (paciente enfermo diagnosticado como sano) conlleva un riego significativo, ya que puede retrasar tratamientos necesarios y potencialmente llevar a complicaciones graves. Con esta perspectiva concluimos que el modelo más factible resulta ser entonces el SVM con una sensibilidad de 93.9 %, aunque su especificidad resulta ser algo mala pues genera un numero considerable de falsos positivos. Aunque la regresión logística está muy cercano a los valores de SVM.

3.2.3. Validación cruzada

El artículo implementa una validación cruzada para asegurar que los modelos (Random Forest, Regresión Logística, SVM y k-NN) generalicen adecuadamente y no presenten sobreajuste, permitiendo una evaluación más estable y confiable del rendimiento. Usando

Modelo	**Scores CV (5 folds)**	**Exactitud media (± desviación)**
Regresión Logística	[0.951, 0.869, 0.738, 0.883, 0.867]	0.861 (± 0.138)
SVM	[0.918, 0.852, 0.770, 0.817, 0.833]	0.838 (± 0.097)
K-Nearest Neighbors	[0.852, 0.820, 0.738, 0.783, 0.767]	0.792 (± 0.080)
Random Forest	[0.852, 0.803, 0.721, 0.783, 0.817]	0.795 (± 0.087)

Cuadro 8: Resultados de Validación Cruzada por Modelo

Además, realizamos una comparación de la exactitud con validación cruzada y sin ella

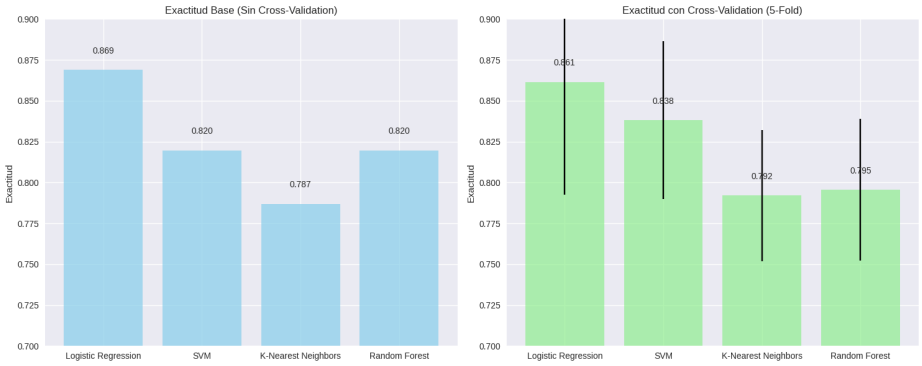


Figura 7: Comparación de exactitud con validación cruzada y sin ella

A simple vista podemos creer que el mejor método de clasificación resulta ser la regresión logística pues tiene una exactitud mayor, pero su desviación nos hace dudar si realmente es el mejor método o tal vez el SVM es mejor pues tiene un buen valor de exactitud y tiene menor desviación.

En los métodos de clasificación algo que nos interesa es averiguar cuales son las variables de mayor importancia para la clasificación. Es por ello que usamos el método de random forest, el cual nos permite justo hacer esto por la naturaleza de su algoritmo.

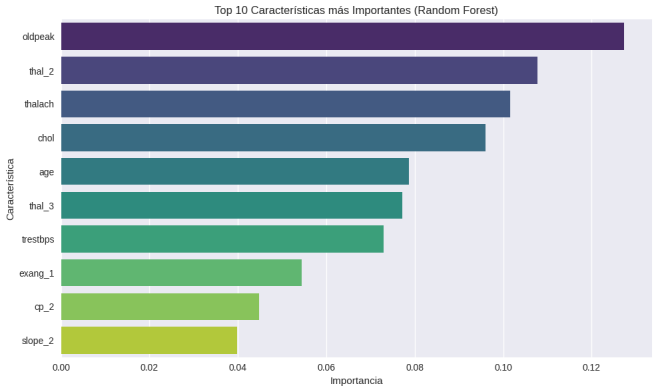
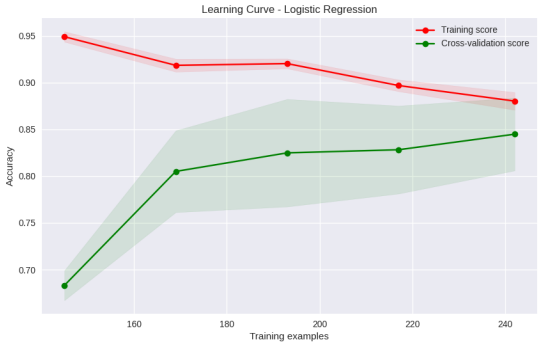


Figura 8: Nivel de importancia en la clasificación de random forest

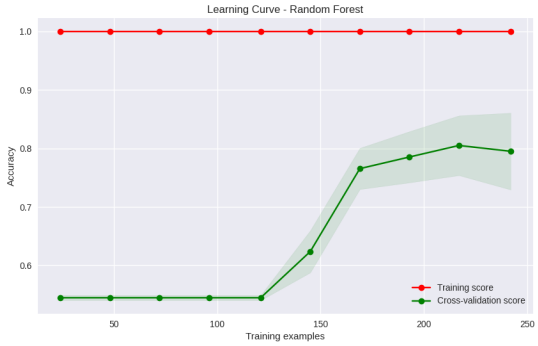
De esta manera sabemos cuales estudios son más importantes para el diagnostico de alguna afección cardíaca.

Finalmente el artículo realiza curvas de aprendizaje para la regresión logística y el random forest para verificar si los modelos están aprendiendo patrones generalizables o solo esta memorizado los datos de entrenamiento.

En ambas gráficas parece presentar un sobre ajuste a los datos de entrenamiento. De ello que justo en la validación cruzada se tenga mayor desviación estándar. Pero en el gráfico 9 resulta tener menor rendimiento en los datos que no ha visto.



(a) Curva de aprendizaje para el modelo de Regresión Log



(b) Curva de aprendizaje para el modelo de Random Forest

Figura 9: Curvas de aprendizaje de los modelos

3.3. Propuesta y análisis extra

Nuestra propuesta es realizar regresión logística L1 y L2 para mejorar la generalización del modelo y controlar el sobre ajuste. En particular proponemos la regresión logística L2 para reducir magnitudes de los coeficientes y de esta manera minimiza cierta multicolinealidad que pudiese existir en las variables. La regresión logística L1 la proponemos para favorecer la selección de variables, pues algunos coeficientes se vuelven cero. Ajustando estos modelos obtenemos los siguientes resultados.

	Accuracy
Regresion Logística L1	0.861 (+/- 0.111)
Regresión Logística L2	0.861 (+/- 0.138)

Cuadro 9: Ajuste obtenido para los modelos RL L1, RL L2

De aquí podemos notar como ambas obtuvieron el mismo accuracy, pero para la Regresión Logística L1 la desviación estándar disminuyo, lo que habla de una mejora en el ajuste de modelo y una mejor clasificación.

3.4. Conclusión

En este caso la clasificaciones realizadas con los modelos propuestos fueron bastante buenas teniendo muy buena sensibilidad. Destacamos que gracias a la validación cruzada pudimos obtener modelos más robustos para evitar el sobreajuste en nuestros datos. Además el modelo de regresión logística L1 propuestos aparte del articulo, resultaron mejorar el desempeño de la clasificación.

4. Conclusiones finales

La replicación de los artículos [Price et al., 2023] y [Rimal et al., 2025] ha permitido evaluar la aplicabilidad y limitaciones de los modelos de regresión lineal y logística en contextos reales de toxicología ambiental y diagnóstico médico.

En el caso de la regresión lineal para predicción de toxicidad de zinc, si bien se logró replicar los modelos propuestos por [Price et al., 2023] con coeficientes consistentes, los bajos valores de R^2-ajustado (0.3-0.5) indican una capacidad explicativa limitada. Esto sugiere que las variables consideradas (pH, hardness y DOC) no capturan completamente la complejidad del fenómeno toxicológico, dejando en evidencia la necesidad de incorporar variables adicionales o explorar modelos no lineales para mejorar la predictibilidad.

Por otro lado, la replicación del análisis de regresión logística para predicción de enfermedades cardíacas demostró el robusto desempeño de este modelo en clasificación médica, alcanzando una exactitud del 86.9 % y una sensibilidad del 90.9 %. Estos resultados, consistentes con los reportados por [Rimal et al., 2025], confirman la utilidad de la regresión logística como herramienta confiable en diagnóstico, en particular cuando la detección de verdaderos positivos (sensibilidad) es importante para la salud del paciente.

La implementación de técnicas de validación cruzada y análisis de residuos en ambos casos refuerza la importancia de la validación rigurosa de modelos, asegurando que las conclusiones derivadas sean estadísticamente sólidas.

Referencias

[Lopez, 2024] Lopez, M. (2024). Presentation_10 - diapositivas de ciencia de datos. Recuperado de GitHub. Diapositivas de clase del curso de Ciencia de Datos.

[Price et al., 2023] Price, G. A., Stauber, J. L., Jolley, D. F., Koppel, D. J., Van Genderen, E. J., Ryan, A. C., and Holland, A. (2023). Development and validation of multiple linear regression models for predicting chronic zinc toxicity to freshwater microalgae. *Environmental Toxicology and Chemistry*, 42(12):2630–2641.

[Rimal et al., 2025] Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., and Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, svm, knn, and random forest with cross-validation for improved accuracy. *Scientific Reports*, 15(1):13444.