UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS
Revista BOLETÍN DE CIENCIAS DE LA TIERRA

Documento: Plantilla para la presentación de un manuscrito
Revista: Boletín de Ciencias de la Tierra

Generalidades
Fuente: Times New Roman, 10.
Una sola columna.
Interlineado 1,5.
Máximo 20 páginas.
Márgenes derecha e izquierda: 2 cm.
Figuras incluidas en el texto.
Citación bibliográfica: norma APA.

CONTENIDO DEL MANUSCRITO

Título: Aplicación de métodos de Machine Learning aplicado al cálculo del transporte de sedimentos
Title: Application of Machine Learning methods applied to the calculation of sediment transport

Autores:
Juan Felipe Ochoa, Universidad Nacional de Colombia, Sede Medellín, jufochoa@unal.edu.co

1. Introducción

El transporte de sedimentos es un fenómeno físico muy complejo en el cual interactúan diferentes variables entre las que se incluyen: información del fondo del cauce en términos de la granulometría del lecho en función de la gradación de los sedimentos (G) y un tamaño representativo, normalmente asociado al tamaño medio (D50), además de información geométrica del canal, como el ancho (B) y la pendiente (S). Adicionalmente, se consideran otras variables propias del campo de flujo, las cuales definen la capacidad motriz, como el caudal (Q), la velocidad(V) y la profundidad del flujo.

En términos generales, se han desarrollado más de 200 modelos de transporte de sedimentos derivados bajo diferentes conceptualizaciones: modelos analíticos, semi-analíticos, experimentales y empíricos (Merrit *et al*.,2003; Dey,2014).

En términos experimentales, diferentes autores han realizado mediciones en canales de laboratorio, así como en canales naturales con el objeto de derivar modelo directamente de los datos a partir de técnicas de ajuste convencional y poder ajustar diferentes parámetros de los modelos matemáticos.

Sobresale en este sentido, la base de datos conformada por Brownlie (1981) quien conformó una base de datos con mediciones de transporte de sedimento de canales de laboratorio y de canales abiertos que incluye 7027 registros (1815 mediciones en canales naturales y 5212 mediciones en canales de laboratorio) a partir de 79 autores.

Las variables compiladas en dicha base de datos corresponden a las siguientes variables: caudal, Q(L/s), ancho del canal, B(m), profundidad del flujo, y (m), pendiente, S (m/m*1000), tamaño medio del sedimento del lecho, D_{50}(mm), coeficiente de gradación del lecho, G (adimensional), concentración de sedimentos, C(ppm), temperatura, T, (°C) y forma del lecho, FL (0 - no se observa, 1 – Lecho plano previo al inicio del movimiento, 2 – Rizos, 3 – Dunas, 4 – Lecho en transición, 5 – Lecho plano, 6 – Ondas permanentes, 7 – Antidunas, 8 – Rápidos y pozos). Se observa que todas las variables son de tipo numérico, a excepción de la variable FL que es una variable categórica.

En función de esta información surgen diferentes preguntas en función de los datos experimentales, para lo cual se propone dar una aproximación a la respuesta bajo un esquema de implementación de modelos de Machine Learning. Específicamente, las inquietudes son:

P1. ¿Es posible obtener un modelo lo suficiente mente robusto para estimar la concentración de sedimentos mediante modelos de machine learning? ¿Hay diferencias en el mejor modelo ajustable entre los datos de ríos y los datos de canales de laboratorio para el cálculo de la concentración?

P2. ¿Cuál es el mejor modelo para clasificar la diferencia entre transporte de sedimentos de baja intensidad y de alta intensidad?

P3. ¿ Cuál es el mejor modelo para clasificar las formas del lecho?

Para efectos de proponer una respuesta a los anteriores interrogantes, se implementan diferentes modelos de Machine Learning en Python. El enfoque metodológico y las bases teóricas de los métodos consultados pueden verse en mayor detalle en (Aristizabal,2022; Hastie *el al*.,2009)

2. Materiales y métodos

Para efectos de realizar los análisis propuestos en Python, se partió de digitalizar en excel la base de datos de Brownlie(1981), de forma tal de disponer en un único archivo de la información completade todas la variables.

Para efectos de dar respuesta a los planteamientos propuestos, el esquema general de desarrollo consistió en:

- Desarrollar el análisis exploratorio con el objeto de identificar que variables deben ser consideradas y que elementos deben removerse de la base de datos, bien porque se consideren inconsistentes (valores negativos) o porque se consideren fuera de tendencia (outliers).

- Aplicar los modelos de machine learning para todas las variables y luego para las variables reducidas tras la implementación de algoritmos de Feature Selection, con el objeto de disminuir el número de variables en el modelo.
- Evaluar diferentes procedimientos de escalamiento de variables con el objeto de reducir el efecto de las escalas en las variables independientes

Análisis exploratorio

P1. Diferencia entre modelos de clasificación supervisada para ríos y canales naturales para el cálculo de la concentración de sedimento

Una vez realizado el análisis exploratorio se realiza una comparación de los siguientes modelos de regresión: LinearRegression, Ridge, Lasso, ElasticNet, KNeigborsRegressor, DecisionTreeRegressor y SupportVectorRegressor(SVR) con una comparativa de tres métricas: Coeficiente de correlación (R^2), Error medio cuadrático (MSE), Error medio absoluto (MSE).

Estos modelos se evalúan para el conjunto de datos completo, así como para otros conjuntos reducidos de variables tras definir las variables de mayor importancia. En función de los resultados obtenidos, los análisis se realizaron con el conjunto de variables originales, así como con las variables agrupadas en variables adimensionales.

Este análisis se desarrolla con los tres conjuntos de datos: la base de datos completa con todas las mediciones, la base de datos con las mediciones del laboratorio y la base de datos con las mediciones de ríos naturales. De esta forma, en función de la aplicación del mismo procedimiento a los tres conjuntos de datos, se concluye respecto al planteamiento inicial.

Finalmente se realiza una optimización de los hiperparámetros del mejor modelo implementado

P2. Diferencia entre modelos de clasificación(binaria) no supervisada para ríos y canales naturales para la intensidad del transporte de sedimentos

Este problema corresponde a un problema de clasificación binaria, para lo cual se creó dentro del conjunto de datos una variable adicional denominada Tipo, la cual toma valores de uno (1) para registros considerados de Transporte ilimitado y de cero (0) para registros considerados de Transporte limitado.

Posterior al análisis exploratorio se realiza una comparación de los siguientes modelos de regresión: LogisticRegressor, ADL, KNN, DecisionTreeClassifier, Naive Bayes y SupportVectorClassifier (SVC) con una comparativa de tres métricas: Coeficiente de correlación (R^2),Error medio cuadrático (MSE), Error medio absoluto (MSE).

Estos modelos se evalúan únicamente para el conjunto de datos completo, en función de los resultados obtenidos en la implementación de los modelos de regresión.

P3. Diferencia entre modelos de clasificación(multiclase) no supervisada para ríos y canales naturales para la definición de la forma del lecho

Este problema corresponde a un problema de clasificación multiclase, en el cual cada forma de lecho tiene una codificación de 0 a 7, según se ha indicado previamente en la introducción del documento.

Posterior al análisis exploratorio se realiza una comparación de los siguientes modelos de regresión: LogisticRegressor, ADL, KNN, DecisionTreeClassifier, Naive Bayes y SupportVectorClassifier (SVC) con una comparativa de tres métricas: Coeficiente de correlación (R^2), Error medio cuadrático (MSE), Error medio absoluto (MSE).

Estos modelos se evalúan únicamente para el conjunto de datos completo, en función de los resultados obtenidos en la implementación de los modelos de regresión.

3. Resultados

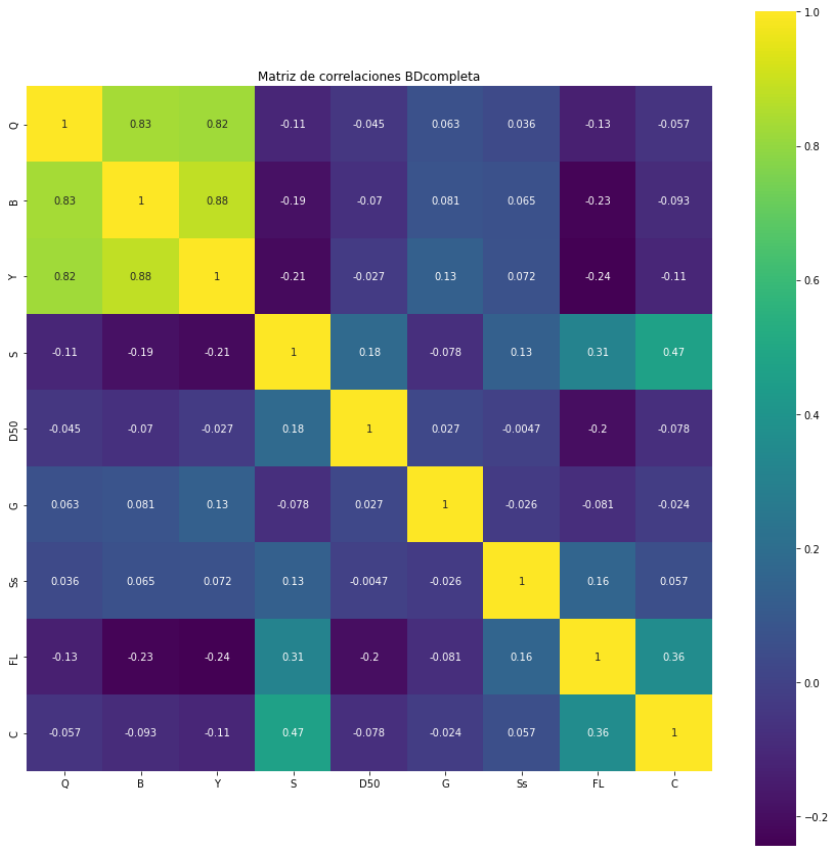
Cálculo de la concentración de sedimentos en canales naturales y de laboratorio

Luego de la limpieza de la base de datos se conformó una base de datos con 6641 variables, 7 variables independientes (Q, B, Y, S, D50, G, Ss) y la variable concentración (C) como variable dependiente.

Análisis Exploratorio

En lo referente al análisis exploratorio, se obtuvo para cada conjunto de datos la matriz de correlación de la base de datos completa (ver Figura 1).

Figura 1. Matriz de correlación para la base de datos completa (ríos+laboratorio)



En cuanto a la estructura de correlación se observa que:

- En la base de datos completa, las variables de caudal, ancho y profundidad tienen una alta correlación entre sí. Por otro lado, la variable de pendiente y formas del lecho tiene una correlación moderada con la concentración.
- En la base de datos de canales de laboratorio, se observa la misma estructura de correlación observada en la base de datos completa. Se explica esto en función de que el 75% de los datos totales se conforman con datos de laboratorio.
- En la base de datos de ríos, se observa que las variables de caudal, ancho y profundidad tienen una alta correlación entre sí. Sin embargo, no hay ninguna variable que presente una correlación fuerte con la concentración. En este caso la pendiente y las formas del lecho tienen una correlación baja respecto a la concentración.

A partir de la aplicación de los modelos previamente descritos con los datos originales con una validación cruzada en 10 particiones se calcularon las métricas indicadas (r2,MSE,MAE). Los resultados obtenidos se presentan en el Anexo 2 (Resultados 2a).

En vista de los altos errores obtenidos se procedió a formular una simplificación en su estructura, tras una identificación de las variables que añadan más información al modelo mediante técnicas de Feature Selection.

Feature Selection

Se implementaron tres técnicas para identificar su importancia en añadir información al modelo: Arboles de decisión (Decision Tree), Eliminación univariable (Elimination), Eliminación recursiva (Backward elimitation). La Tabla 1 presenta los resultados del Feature Selection.

Tabla 1. Variables seleccionadas mediante técnicas de feature selection para las variables originales

Base de datos	Decision Tree	Eliminación Backward	Eliminación recursiva	Variables Seleccionadas
Laboratorio + Ríos	B,S,D50,G	S,D50	B,S,D50,G	**Q,B,S,D50**
Laboratorio	Q,Y,S,D50	Q,S,D50,G	B,S,D50,G	**Q,S,D50,G**
Ríos	Q,Y,S.D50,G	Q,Y,S,D50	Q,B,D50,G	**Q,S,D50,G**

A partir de las variables seleccionadas, se aplicaron nuevamente los diferentes modelos de regresión comparados con las métricas definidas. Los errores obtenidos se presentan en el Anexo 2 (Resultados 2b).

Escalamiento de variables originales

Dado que persisten errores altos en las métricas implementadas, se procedió a realizar una transformación de las variables que definen el conjunto de variables dependientes. Se evaluaron 3 tipos de transformaciones: MinMaxScaler, Standarized y Normalized con la métrica R2 como criterio de desempeño del escalamiento (Resultados 2c). A partir de los valores obtenidos, se seleccionó la transformación Normalizer con las bases de datos total y de ríos para proseguir con la evaluación de los modelos y se usó la transformación StandardScaler con la base de datos de laboratorio. Los errores obtenidos se presentan en el Anexo 2 (Resultados 2d).

Regresión con variables adimensionales

En vista de la permanencia de errores muy altos en los modelos implementados, se procedió a formular una estructura diferente de variables mediante el uso de parámetros adimensionales en función de las variables originales, el cual es un procedimiento ampliamente utilizado en la obtención de modelos de transporte de sedimentos.

Partiendo de que la concentración de sedimentos es una función de las siguientes variables (Yalin, 1977; Dogan,2008):

$$C = f(u_*, q, D_{50}, \rho, \rho_s, h, B, \nu, \sigma_g, S, u_m, g)$$

Donde: C= es la concentración de sedimentos (ppm), u_* es la velocidad cortante (LT^{-1}), D_{50} es el tamaño de partículas tal que el 50% de tamaños es más fino(L), es la densidad del agua (ML^{-3}), es la densidad del sedimento (ML^{-3}), B es el ancho del canal (L), es la viscosidad cinemática (L^2T^{-1}) y g es la aceleración de la gravedad (LT^{-2}).

Mediante el análisis dimensional, Dogan (2008) configuró los siguientes parámetros adimensionales en función de las variables que determinan la tasa de transporte de sedimentos

$$C = f\left(\begin{array}{c} S, \sigma_g, \frac{R}{D_{50}}, \frac{h}{D_{50}}, \frac{B}{D_{50}}, \frac{\rho}{\rho_s}, \frac{u_m h}{\nu}, \frac{u_* D_{50}}{\nu}, \frac{u_m h}{\nu}, \frac{hS}{D_{50}(G_s-1)}, \frac{u_m}{u_*}, \frac{B}{h}, \frac{q}{\sqrt{gh^3}}, \frac{q}{u_* D_{50}}, \\ \frac{\nu u_*}{D_{50}^2(G_s-1)g}, \frac{\nu^2}{D_{50}^3(G_s-1)g}, \frac{q^2}{D_{50}^3(G_s-1)g}, \frac{\rho_s u_*^2}{\gamma_s D_{50}}, \frac{u_m}{\sqrt{gD_{50}(G_s-1)}} \end{array}\right)$$

En el Anexo 1 se resumen los parámetros adimensionales, los cuales se utilizan para efectos de realizar la evaluación de modelos predictivos.

Con las variables adimensionales, se realizó la evaluación de los modelos de regresión con sus respectivas métricas. Los resultados se presentan en el Anexo 2(Resultados 2e). En vista de los altos errores obtenidos se procedió a formular una simplificación en su estructura, tras una identificación de las variables que añadan más información al modelo mediante técnicas de Feature Selection.

Feature Selection variables adimensionales

Se implementaron las técnicas indicadas previamente para identificar la importancia de las variables). La Tabla 2 presenta los resultados del Feature Selection.

Tabla 2. Variables seleccionadas mediante técnicas de feature selection para las variables originales

Base de datos	Decision Tree	Eliminación Backward	Eliminación recursiva	Variables Seleccionadas
Laboratorio + Ríos	Ref, EA, Fr, VCA	AA, PA, Ref, Res, Fr, Ff, VCA, TPA, CUA2, PM, Frp	AA, Ref, CUA, CUA2	**Ref, EA, Fr, VCA**
Laboratorio	AA, EA, Fr, CUA, PM	AA, PA, EA, Fr, Ff, CUA, VCA, TPA, CUA2, Frp	AA, Ref, CUA, CUA2	**AA, EA, Fr, CUA, PM**
Ríos	EA, Fr, CUA, TPA, PMA	AA, PA, Ref, Res, EA, Fr, CUA, VCA, CUA2, Frp	AA, Ref, CUA, CUA2	**EA, Fr, CUA, TPA**

Con estos modelos se aplicaron nuevamente los diferentes modelos de regresión comparados con las métricas definidas. Los errores obtenidos se presentan en el Anexo 2 (Resultados 2f).

Escalamiento de variables adimensionales

Con estos modelos se aplicaron nuevamente los diferentes modelos de regresión comparados con las métricas definidas (Resultados 2g). Los errores obtenidos se presentan en el Anexo 2 (Resultados 2h).

Selección del mejor modelo y optimización de hiperparámetros

A partir de los resultados obtenidos, no se encontró un modelo lo suficientemente robusto para predecir adecuadamente la concentración de sedimentos.

En términos de los resultados obtenidos, la Tabla 3 resume los mejores modelos obtenidos para las diferentes métricas y los diferentes conjuntos de datos.

Tabla 3. Mejor modelo en función del procedo de análisis y el conjunto de datos

Análisis con variables:	Ríos + Laboratorio			Laboratorio			Ríos		
	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE
Originales	SVR	KNR	KNR	SVR	KNR	KNR	SVR	SVR	SVR
Originales reducidas	SVR	KNR	KNR	SVR	KNR	KNR	SVR	KNR	SVR
Originales reducidas y escaladas	SVR	KNR	KNR	SVR	KNR	KNR	SVR	KNR	KNR
Adimensionales	SVR	SVR	KNR	SVR	SVR	KNR	SVR	SVR	KNR
Adimensionales reducidas	SVR	KNR	DTR	SVR	KNR	DTR	SVR	SVR	DTR
Adimensionales reducidas y escaladas	SVR	DTR	DTR	SVR	KNR	KNR	SVR	KNR	KNR

A partir de los datos presentados se observa que:

- El modelo SVR es el modelo con el menor error para la métrica R^2 en los tres conjuntos de datos.
- El modelo KNR es el modelo con el menor error para la métrica MSE en los tres conjuntos de datos.
- El modelo DTR es una alternativa para la evaluación de los modelos y una búsqueda de optimización de hiperparámetros.

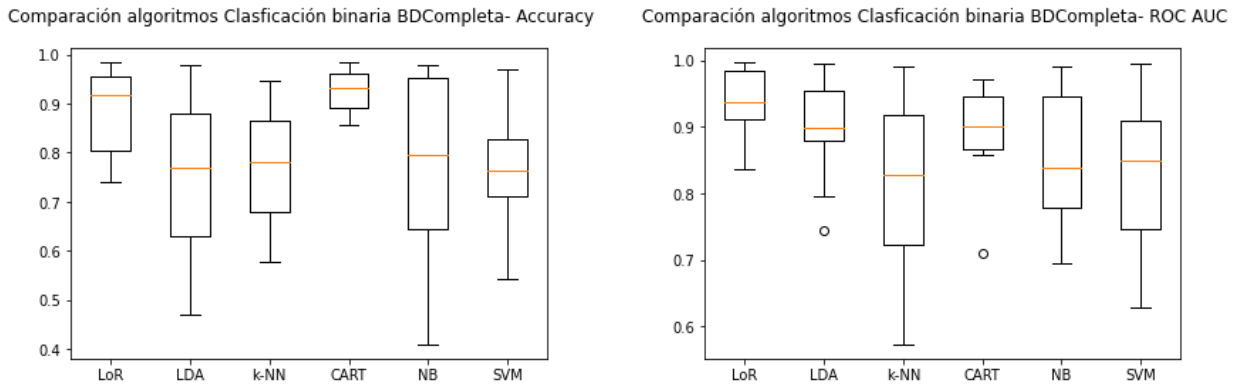
Clasificación de la intensidad del transporte de sedimento

La evaluación de la capacidad de los modelos de ML en un problema de clasificación binaria, se utilizaron las variables de la base de datos de Brownlie (1981) creando una variable auxiliar denominada intensidad del transporte de sedimento, mediante la cual se cataloga cada registro como de condición limitada (0) o condición ilimitada (1). La formulación utilizada para realizar esta clasificación puede verse con más detalle en Gomez & Soar (2020).

La base de datos utilizada para este análisis corresponde al conjunto de datos de mediciones de ríos y laboratorio, la cual luego de ser depurada tiene un total de 6641 registros en los cuales 3526 corresponden a la condición de transporte limitado(Tipo 0) y 3115 mediciones de transporte ilimitado (Tipo 1), lo cual permite identificar que la base se encuentra balanceada.

Se realizó la evaluación de las métricas de Accuracy y ROC para evaluar el desempeño de los diferentes modelos implementados (Figura 2).

Figura 2. Métricas de comparación de modelos para el análisis de clasificación binaria



En términos de los resultados obtenidos se concluye que:

- A partir del Accuracy, el modelo de árboles de decisión (CART) proporciona los mejores resultados en términos del valor promedio y la deviación de los resultados.
- Mediante la métrica ROC, el modelo de regresión logística es el modelo de mejor desempeño.

Una evaluación más detallada en términos del reporte de clasificación permite ver como el modelo de árboles de decisión (CART) es el de mejores resultados en términos de los valores obtenidos para la precisión, el Recall y el F1 score(Tabla 4).

Tabla 4. Reportes de clasificación para para el análisis de clasificación binaria

Regresión Logística		Precision	Recall	F-1 score	support
	0	0.84	0.95	0.89	1033
	1	0.93	0.81	0.86	960
	accuracy			0.88	1993
	macro avg	0.89	0.88	0.88	1993
	wieghted avg	0.88	0.88	0.88	1993

CART		Precision	Recall	F-1 score	support
	0	0.98	0.97	0.98	1033
	1	0.97	0.97	0.97	960
	accuracy			0.97	1993
	macro avg	0.97	0.97	0.97	1993
	wieghted avg	0.97	0.97	0.97	1993

Discriminante Lineal		Precision	Recall	F-1 score	support
	0	0.68	0.98	0.8	1033
	1	0.96	0.5	0.66	960
	accuracy			0.75	1993
	macro avg	0.82	0.74	0.73	1993
	wieghted avg	0.81	0.75	0.73	1993

NB		Precision	Recall	F-1 score	support
	0	0.67	0.99	0.8	1033
	1	0.98	0.48	0.64	960
	accuracy			0.74	1993
	macro avg	0.82	0.73	0.72	1993
	wieghted avg	0.82	0.74	0.72	1993

KNN		Precision	Recall	F-1 score	support
	0	0.78	0.85	0.81	1033
	1	0.82	0.74	0.78	960
	accuracy			0.8	1993
	macro avg	0.8	0.79	0.79	1993
	wieghted avg	0.8	0.8	0.79	1993

SVM		Precision	Recall	F-1 score	support
	0	0.85	0.78	0.81	1033
	1	0.78	0.85	0.81	960
	accuracy			0.81	1993
	macro avg	0.81	0.81	0.81	1993
	wieghted avg	0.81	0.81	0.81	1993

Clasificación de la forma del lecho

La evaluación de la capacidad de los modelos de ML en un problema de clasificación multiclase, se realizó mediante el uso directo de la variable de formas de lecho disponible dentro de la base de datos de Brownlie (1981) acorde a la clasificación de la forma de lecho indicada previamente.

La base de datos utilizada para este análisis corresponde al conjunto de datos de mediciones de ríos y laboratorio, la cual luego de ser depurada tiene un total de 6641 registros, correspondiendo a una base de datos desbalanceada, como se presenta en la Tabla 5.

Tabla 5. Conjunto multiclase para la clasificación multiclase

Tipo	Cantidad	Porcentaje
0 – sin forma de lecho	3514	52.9%
1 - Lecho plano imbricado	44	0.7%
2 - Rizos	702	10.6%
3 - Dunas	1103	16.6%
4 - Lecho en Transición	338	5.1%
5 - Lecho plano	630	9.5%
6 - Ondas permanentes	34	0.5%
7 - Antidunas	269	4.1%
8 - Rápidos y pozos	7	0.1%

Se realizó la evaluación de las métricas de Accuracy y el coeficiente Cohen-Kappa dado los desbalanceado de la base de datos ROC para evaluar el desempeño de los diferentes modelos implementados (Figura 3Figura 2).

Figura 3. Variación del accuracy para la comparación de modelos para el análisis de clasificación multiclase

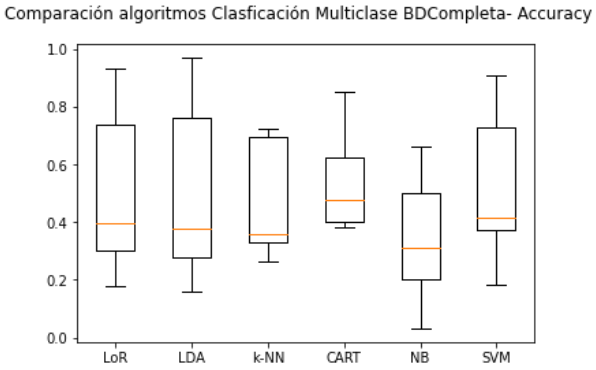


Tabla 6. Métricas de desempeño para la comparación de modelos para el análisis de clasificación multiclase

Método	Accuracy		Coeficiente Cohen-Kappa
	Media	Desviación	
Regresión Logística	51.71	27.37	15.71
Discriminante Lineal	50.49	29.35	8.5
KNN	47.34	19.03	33.63
CART	52.4	15.69	78.73
NB	34.36	20.02	22.51
SVM	52.97	24.37	28.9

Una evaluación más detallada en términos del reporte de clasificación permite ver como el modelo de árboles de decisión (CART) es el de mejores resultados en términos de los valores obtenidos para la precisión, el Recall y el F1 score().

Tabla 7. Reportes de clasificación para para el análisis de clasificación multiclase

Regresión Logística		Precision	Recall	F-1 score	support
	0	0.58	0.96	0.73	1053
	1	0	0	0	12
	2	0	0	0	197
	3	0.29	0.1	0.15	324
	4	0	0	0	114
	5	0.47	0.23	0.31	203
	6	0	0	0	12
	7	0.45	0.22	0.3	76
	8	0	0	0	2
	accuracy			0.56	1993
	macro avg	0.2	0.17	0.17	1993
	wieghted avg	0.42	0.56	0.45	1993

CART		Precision	Recall	F-1 score	support
	0	0.96	0.94	0.95	1053
	1	0.88	0.58	0.7	12
	2	0.77	0.84	0.81	197
	3	0.8	0.85	0.83	324
	4	0.51	0.5	0.51	114
	5	0.75	0.71	0.73	203
	6	0.62	0.67	0.64	12
	7	0.75	0.71	0.73	76
	8	1	1		2
	accuracy			0.86	1993
	macro avg	0.78	0.76	0.77	1993
	wieghted avg	0.86	0.86	0.8	1993

Discriminante Lineal		Precision	Recall	F-1 score	support
	0	0.55	0.98	0.7	1053
	1	0	0	0	12
	2	0	0	0	197
	3	0	0	0	324
	4	0	0	0	114
	5	0.55	0.09	0.15	203
	6	0	0	0	12
	7	0.37	0.38	0.38	76
	8	0.2	1	0.33	2
	accuracy			0.54	1993
	macro avg	0.19	0.27	0.17	1993
	wieghted avg	0.36	0.54	0.4	1993

NB		Precision	Recall	F-1 score	support
	0	0.98	0.41	0.58	1053
	1	0.02	0.5	0.03	12
	2	0.25	0.9	0.4	197
	3	0.15	0.05	0.08	324
	4	0.05	0.01	0.01	114
	5	0.16	0.04	0.07	203
	6	0.3	0.67	0.41	12
	7	0.22	0.74	0.34	76
	8	0.33	1	0.5	2
	accuracy			0.36	1993
	macro avg	0.27	0.48	0.27	1993
	wieghted avg	0.6	0.36	0.38	1993

KNN		Precision	Recall	F-1 score	support
	0	0.7	0.8	0.75	1053
	1	1	0.33	0.5	12
	2	0.38	0.38	0.38	197
	3	0.39	0.39	0.39	324
	4	0.33	0.26	0.29	114
	5	0.41	0.26	0.32	203
	6	0	0	0	12
	7	0.45	0.3	0.36	76
	8	0.5	0.5	0.5	2
	accuracy			0.58	1993
	macro avg	0.46	0.36	0.39	1993
	wieghted avg	0.56	0.58	0.56	1993

SVM		Precision	Recall	F-1 score	support
	0	0.63	0.92	0.75	1053
	1	1	0.33	0.5	12
	2	0.47	0.27	0.35	197
	3	0.51	0.4	0.45	324
	4	0.45	0.13	0.2	114
	5	0.61	0.14	0.22	203
	6	0	0	0	12
	7	0	0	0	76
	8	0	0	0	2
	accuracy			0.6	1993
	macro avg	0.41	0.24	0.27	1993
	wieghted avg	0.55	0.6	0.54	1993

En términos de los resultados obtenidos, se observa como el método de árboles de decisión permite obtener los mejores resultados en la clasificación multiclase para predecir la forma del lecho en función de las variables descriptivas.

4. Conclusiones

La aplicación de los métodos de Machine Learning en el cálculo del transporte de sedimentos permitió encontrar que:

- No se encontró un método lo suficientemente robusto para el cálculo de la concentración de sedimentos unificando las bases de datos con mediciones de laboratorio y ríos. El método con el menor error obtenido para la determinación de la concentración de sedimento fue el modelo de Support Vector Regressor (SVR)
- Los resultados del análisis exploratorio, la selección de variables relevantes y los resultados de las métricas de los análisis realizados por separado, indica que la estructura de correlación de los datos de laboratorio es diferente a la de los datos de ríos. Por lo cual los modelos matemáticos obtenidos son diferentes y por tanto hay limitaciones en l transferencia de esta información.
- En los métodos de clasificación binaria y multiclase, analizados para predecir la condición del tipo d canal en cuanto a limitación del transporte de sedimentos como a la formación de la forma de lecho, el método de árboles de decisión mediante CART fue el mejor modelo implementado con métricas satisfactorias.

5. Datos suplementarios

Se presentan al final del documento los resultados comparativos de las métricas de desempeño citadas en el artículo.

6. Declaración de Conflicto de interés

El autor no presenta conflicto de interés

7. Referencias

Aristizabal, E. (2022). Inteligencia artificial y aprendizaje aplicado en Geociencias. https://edieraristizabal.github.io/MachineLearning/

Bhattacharya, B., Price, R. K., & Solomatine, D. P. (2007). Machine learning approach to modeling sediment transport. *Journal of Hydraulic Engineering*, *133*(4), 440-450.

Brownlie, W. R. (1981). Compilation of alluvial channel data: laboratory and field (p. 209). California Institute of Technology, WM Keck Laboratory of Hydraulics and Water Resources.

Dey, S. (2014). *Fluvial hydrodynamics* (Vol. 818). Berlin: Springer.

Dogan, E., Tripathi, S., Lyn, D. A., & Govindaraju, R. S. (2009). From flumes to rivers: Can sediment transport in natural alluvial channels be predicted from observations at the laboratory scale?. Water resources research, 45(8).

Goldstein, Evan B., Giovanni Coco, and Nathaniel G. Plant. "A review of machine learning applications to coastal sediment transport and morphodynamics." *Earth-science reviews* 194 (2019): 97-108.

Gomez, B., & Soar, P. J. (2022). Bedload transport: beyond intractability. *Royal Society Open Science*, *9*(3), 211932.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Kitsikoudis, V., Sidiropoulos, E., & Hrissanthou, V. (2015). Assessment of sediment transport approaches for sand-bed rivers by means of machine learning. *Hydrological sciences journal*, *60*(9), 1566-1586.

Merritt, W. S., Letcher, R. A., & Jakeman, A. J. (2003). A review of erosion and sediment transport models. *Environmental modelling & software*, *18*(8-9), 761-799.

Sahraei, Shahram, et al. "Bed material load estimation in channels using machine learning and meta-heuristic methods." *Journal of Hydroinformatics* 20.1 (2018): 100-116.

Tayfur, G., Karimi, Y., & Singh, V. P. (2013). Principle component analysis in conjuction with data driven methods for sediment load prediction. *Water resources management*, *27*(7), 2541-2554.

Anexo 1 – Parámetros adimensionales

$AA = \frac{B}{D_{50}}$, Ancho adimensional

$PA = \frac{h}{D_{50}}$, profundidad de flujo adimensional

$Ref = \frac{u_m h}{\nu}$, número de Reynolds del flujo

$Res = \frac{u_* D_{50}}{\nu}$, número de Reynolds de la partícula de sedimento

$EA = \frac{hS}{D_{50}(G_s-1)} = \tau_*$, esfuerzo cortante adimensional

$Ff = \frac{u_m}{u_*}$, factor de fricción

$Fr = \frac{Q}{B\sqrt{gh^3}}$, número de Froude

$CUA = \frac{Q}{Bu_* D_{50}}$, caudal unitario adimensional

$VCA = \frac{\nu u_*}{D_{50}^2(G_s-1)g}$, velocidad cortante adimensional

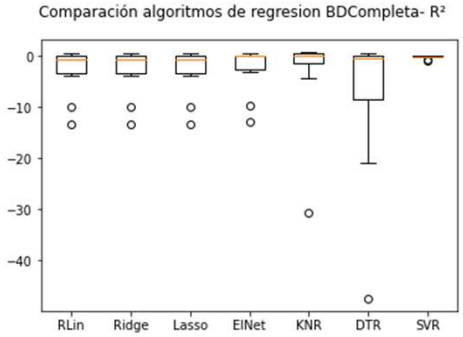
$TPA = \frac{\nu^2}{D_{50}^3(G_s-1)g} = \frac{1}{D_*^3}$, tamaño de partícula adimensional

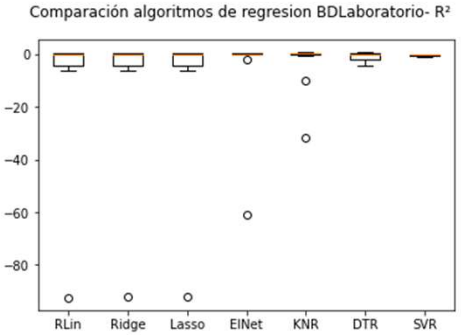
$CUA2 = \frac{q^2}{D_{50}^3(G_s-1)g}$, caudal unitario adimensional2

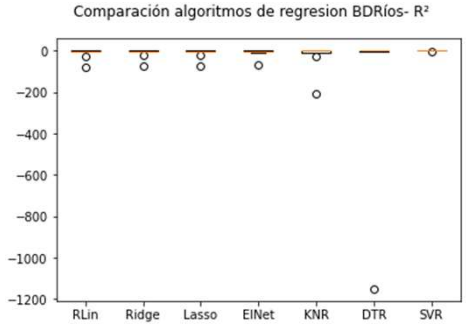
$PM = \frac{\rho_s u_*^2}{\gamma_s D_{50}} = \frac{hS}{D_{50}}$, número de movilidad (relacionado al tamaño de partícula)

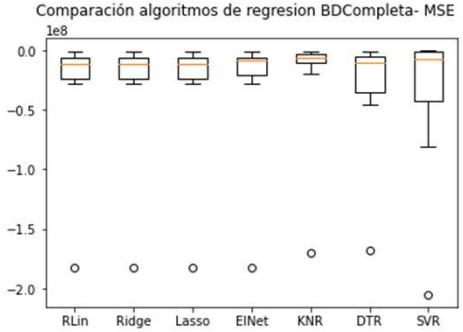
$Frp = \frac{u_m}{\sqrt{gD_{50}(G_s-1)}}$, número de Froude (relacionado al tamaño de partícula)

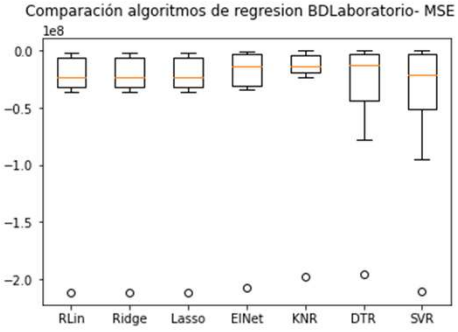
Resultados 2a- Métricas con variables iniciales (R^2,MSE, MAE)

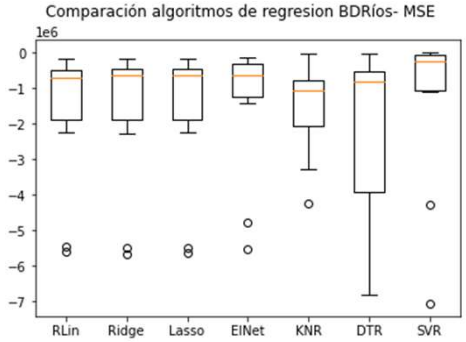


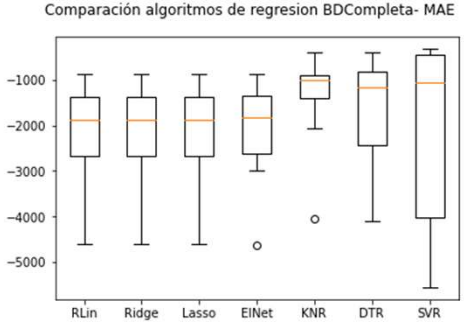


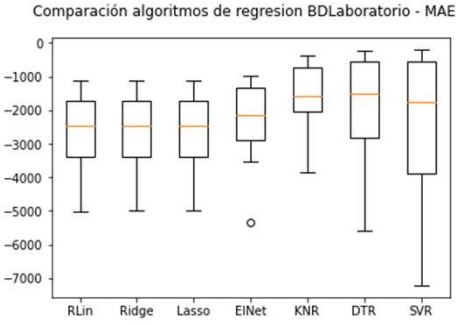


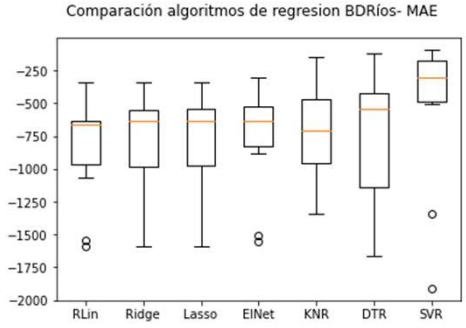




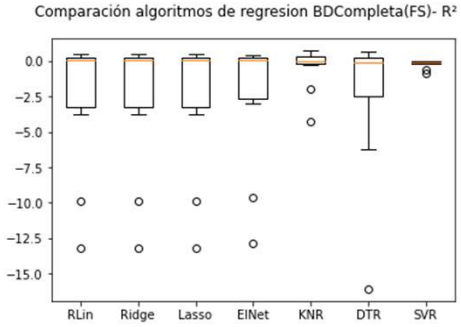


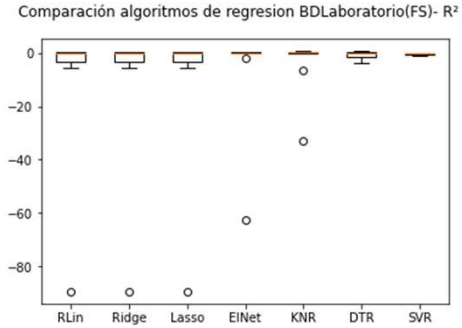


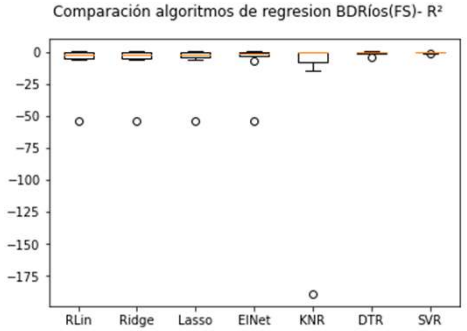


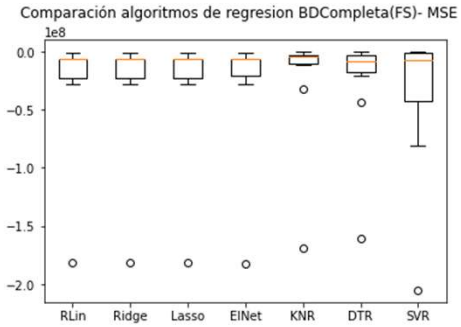


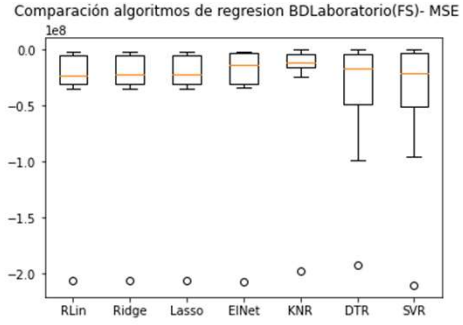
Resultados 2b- Métricas con variables reducidas (R^2,MSE, MAE).

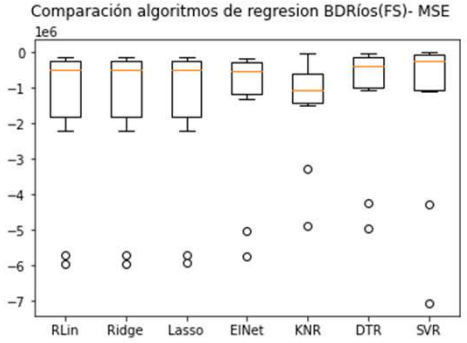


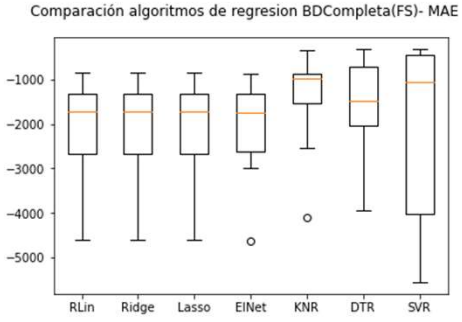


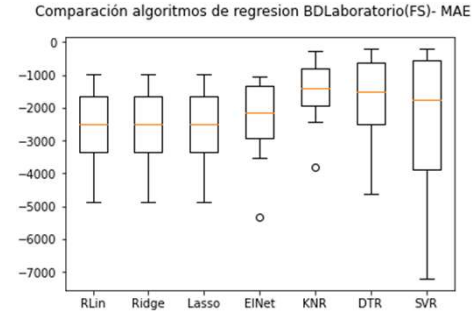


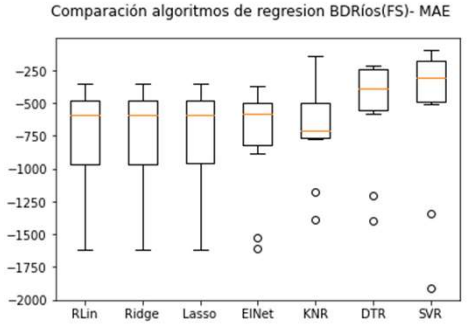




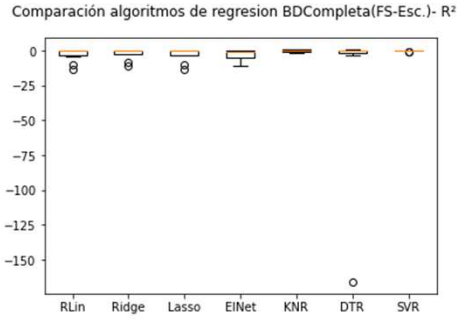


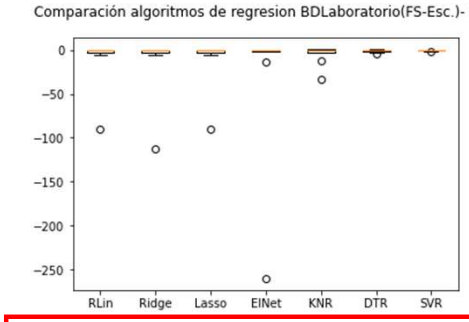


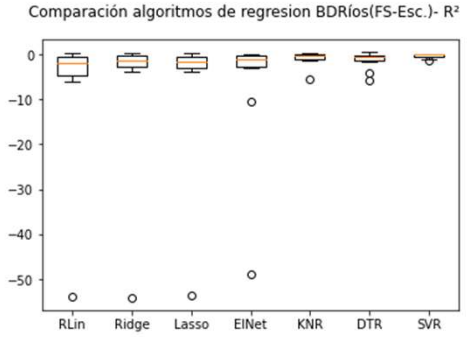




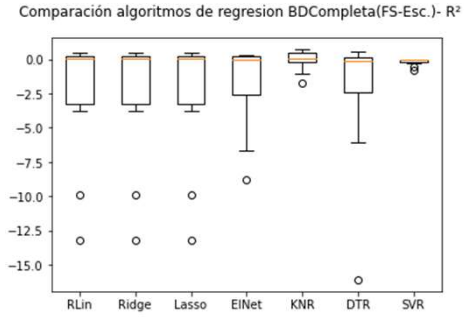
Resultados 2c -Escalamiento variables originales (MinMaxScaler, StandardScaler, Normalizer)

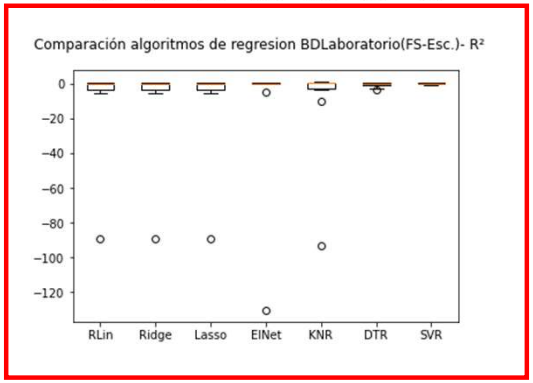


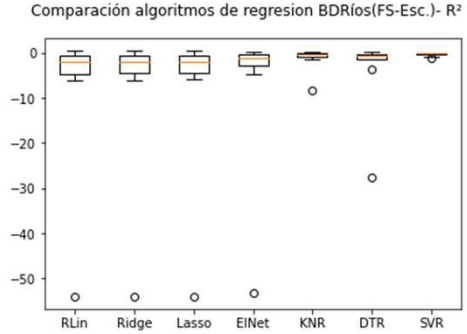




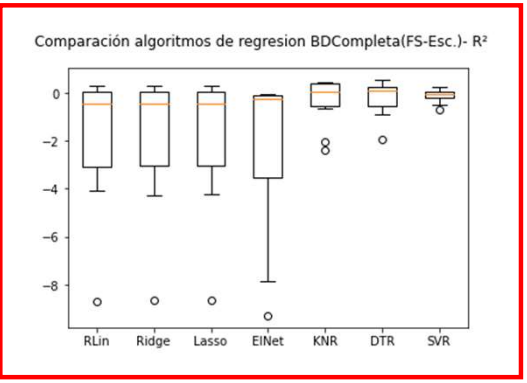
MinMaxScaler

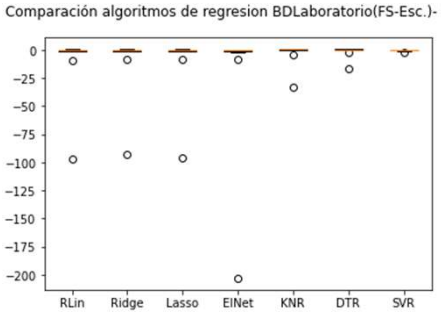


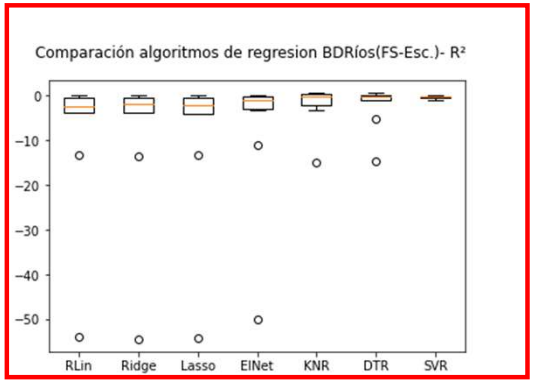




StandardScaler

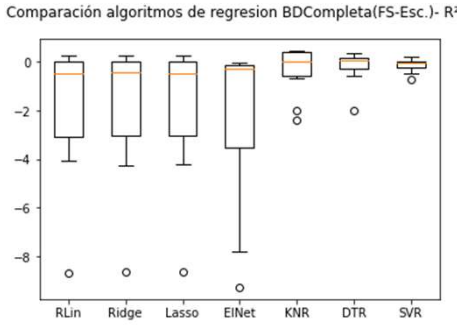


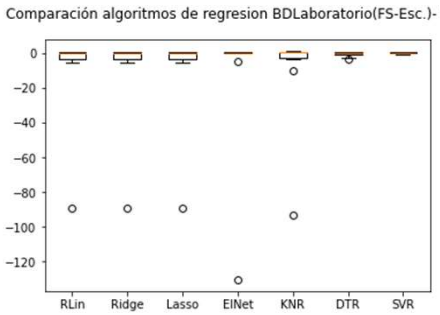


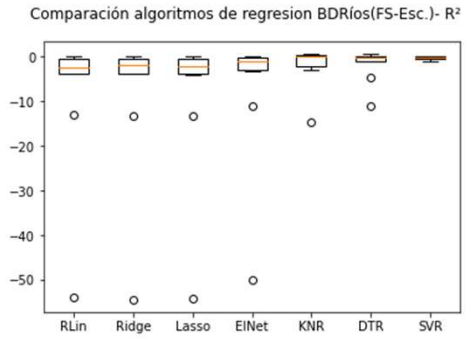


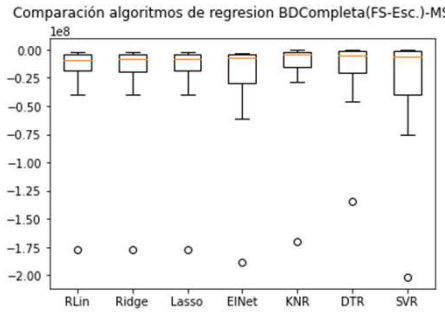
Normalizer

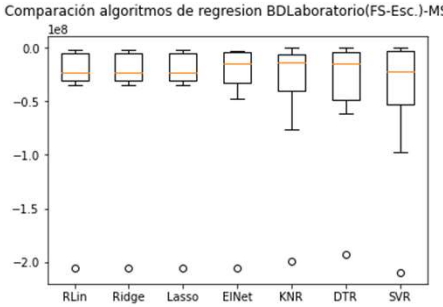
Resultados 2d -Métricas con variables reducidas y escaladas(R^2,MSE, MAE).

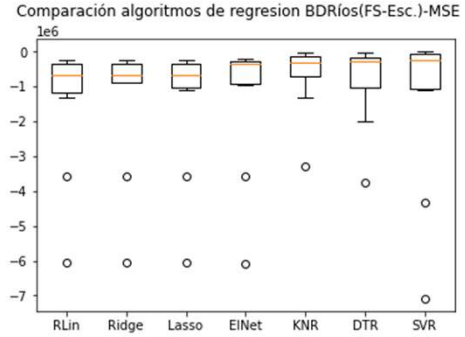


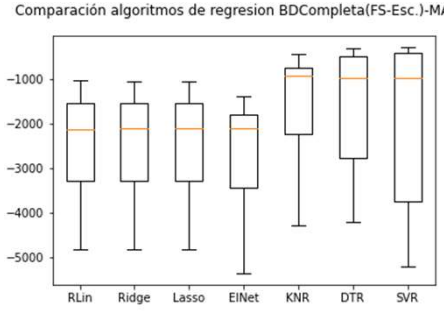


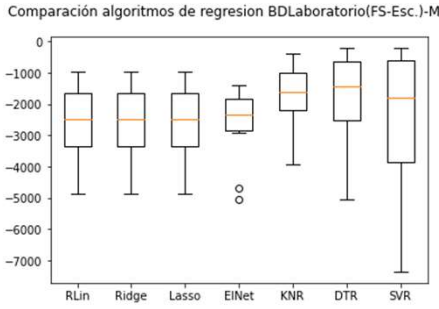


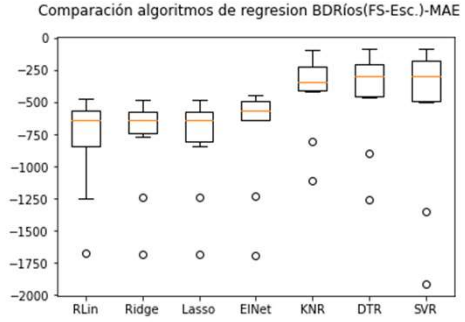




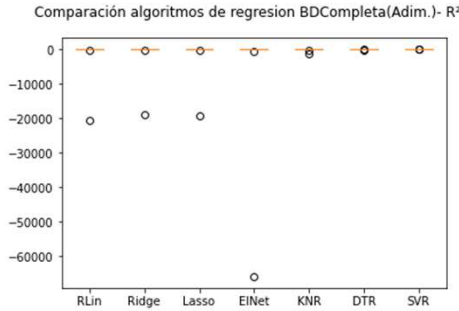


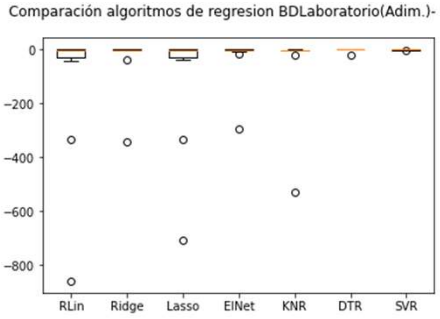


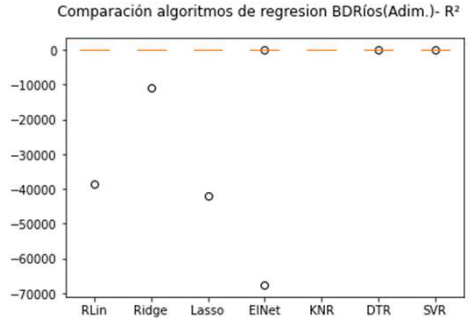


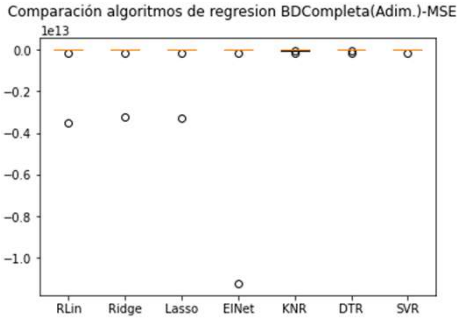


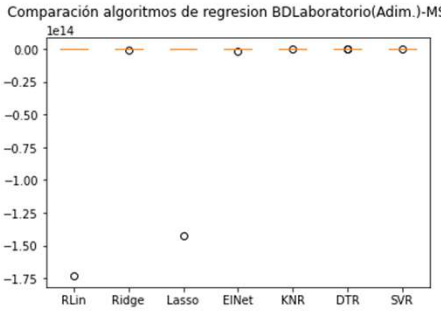
Resultados 2e -Métricas con variables adimensionales (R^2,MSE, MAE).

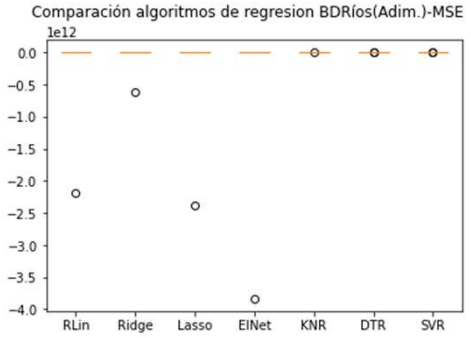


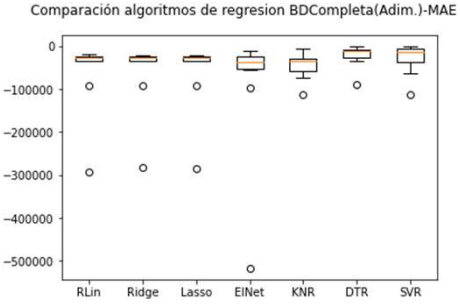


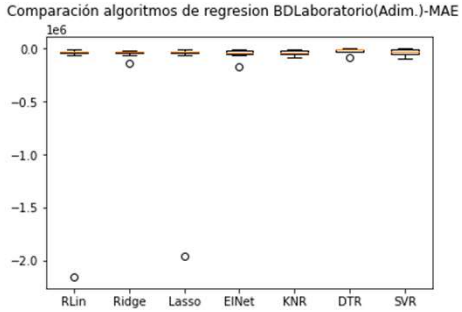


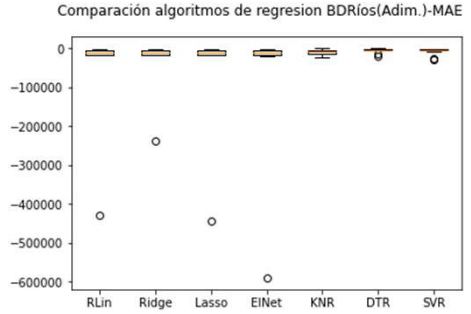




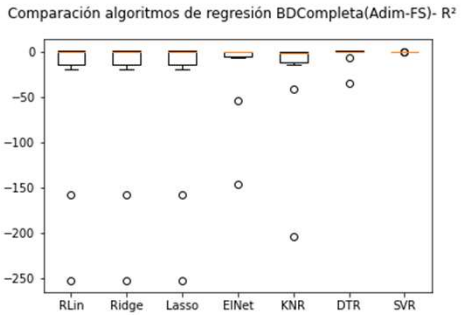


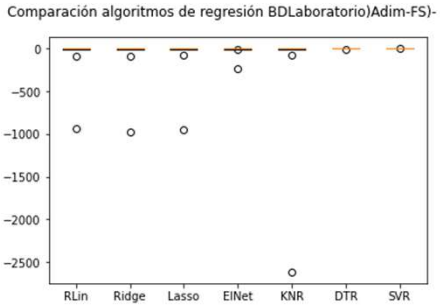


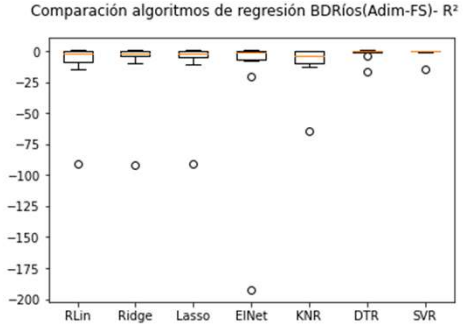


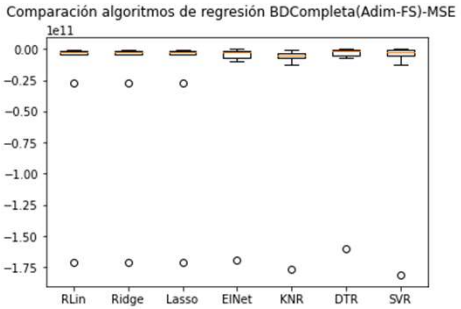


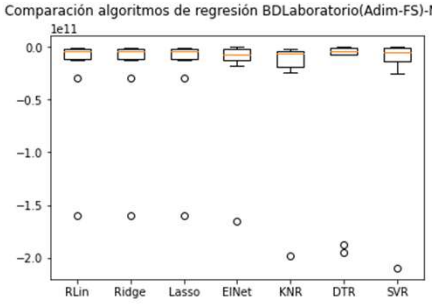
Resultados 2f- Métricas con variables adimensionales, reducidas (R^2,MSE, MAE)

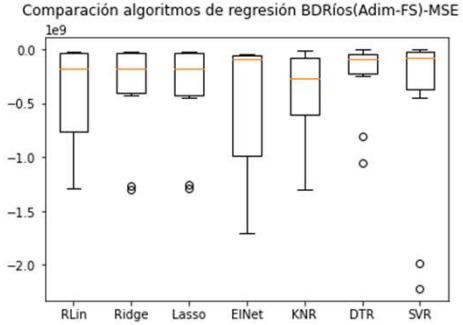


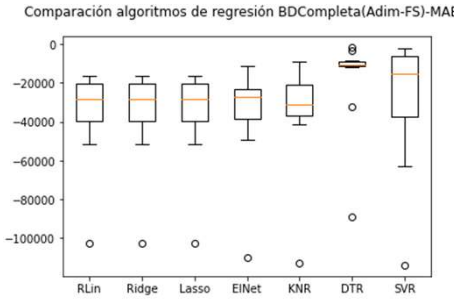


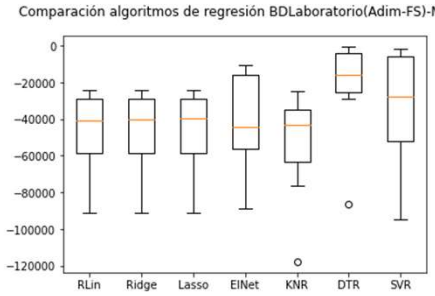


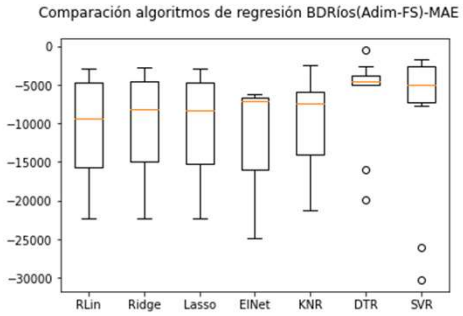




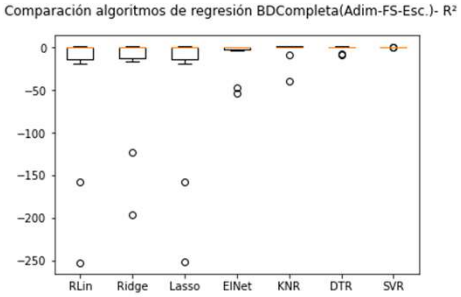


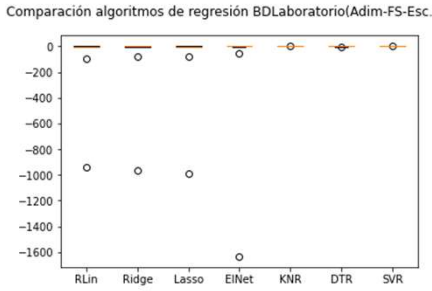


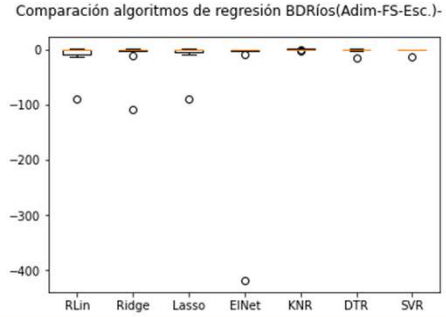




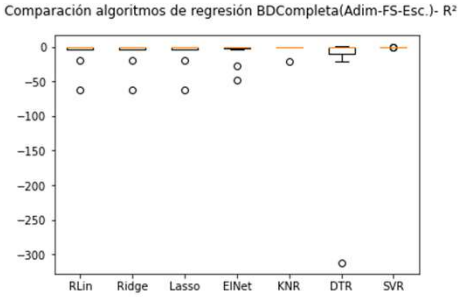
Resultados 2g- Escalamiento variables adimensionales (MinMaxScaler, StandardScaler, Normalizer)

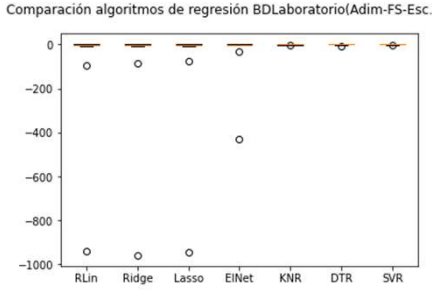


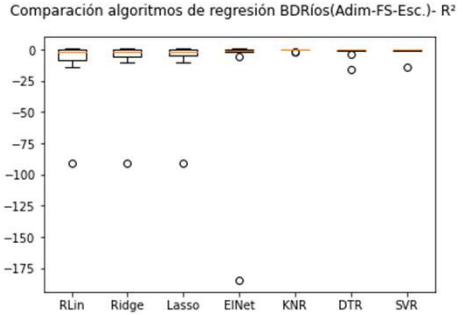




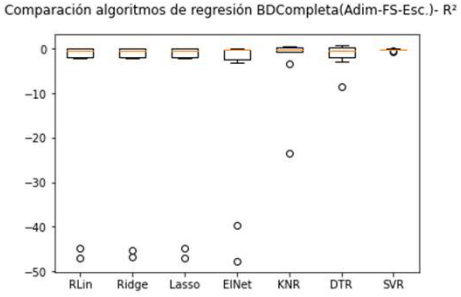
MinMaxScaler

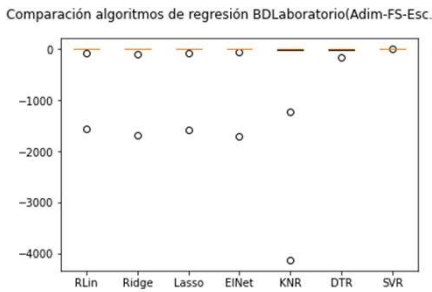


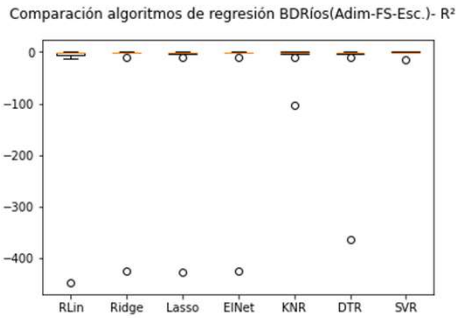




StandardScaler







Normalizer

Resultados 2h- Métricas con variables adimensionales, reducidas y escaladas (R^2,MSE, MAE)