Aplicación de la técnica validación cruzada para determinar el número óptimo de datos utilizados en modelos predictivos.

Rodríguez Murillo, Natalia Arcedalia; Murillo Macera, Margarita Arcedalia; Ríos Lira, Armando Javier
Departamento de Ingeniería Industrial, Tecnológico Nacional de México en Celaya,
Av. Tecnológico y A. García Cubas S/N, A.P.57, C.P. 38010, Celaya, Gto, México
natrdgz.murillo@outlook.com*, armando.rios@itcelaya.edu.mx

Recibido: *20 de febrero de 2019*
Aceptado: *28 de febrero de 2019*

RESUMEN

Al someter información a analizar mediante un modelo de regresión lineal, existe la duda sobre la capacidad predictiva del modelo al predecir información nueva. Por ello es necesario validar la capacidad predictiva del modelo mediante algún método adecuado. La técnica de Validación cruzada es el método más usado, ya que divide los datos en dos partes: un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El primer conjunto es utilizado para estimar los coeficientes de regresión lineal, mientras que el segundo es utilizado para medir la capacidad predictiva del modelo. La literatura existente propone guardar el 10 por ciento de los datos, pero podría no ser el apropiado en las diferentes situaciones presentadas dentro de una investigación, por lo tanto, el número óptimo de datos de predicción a guardar es incierto al realizar la técnica de validación cruzada bajo un nivel de ruido y un tamaño de información del conjunto. La investigación presentada plantea los diferentes escenarios para identificar el porcentaje óptimo a ser guardado al utilizar la técnica anteriormente mencionada.

Palabras clave: Regresión lineal, capacidad predictiva, validación cruzada, nivel de ruido, PRESS.

ABSTRACT

When submitting information to analyze using a linear regression model, there is doubt about the predictive capacity of the model when predicting new information. For this reason it is necessary to validate the predictive capacity of the model by means of some suitable method. The Cross Validation technique is the most used method, since it divides the data into two parts: a set of training data and a set of test data. The first set is used to estimate the linear regression coefficients, while the second is used to measure the predictive capacity of the model. The existing literature proposes to save 10 percent of the data, but it may not be appropriate in the different situations presented within a research, therefore the optimal number of prediction data to keep is uncertain when performing the cross-validation technique under a noise level and a set information size. The presented research raises the different scenarios to identify the optimal percentage to be saved when using the aforementioned technique.

Keywords: Linear regression, predictive ability, cross-validation, noise level, PRESS.

1. INTRODUCCIÓN.

La estadística propone encontrar patrones en un río de información confusa. Para ello, surgen diversas cuestiones que deben ser analizadas: ¿Cómo y qué información o datos elegir? ¿Cómo analizar y resumir el análisis efectuado? (Kerner, 1015).

Durante la etapa de construcción de modelos hay factores que pueden afectar significativamente las nuevas observaciones, lo que hace las predicciones menos exactas (García & Lara, 1998). Además, la estructura correlativa entre los regresores puede diferir el rendimiento para el modelo; esto puede resultar en un rendimiento predictivo pobre para el modelo. La validación adecuada de un modelo desarrollado para predecir nuevas observaciones debe involucrar y probar el modelo en el medio a desarrollarse antes de que sea entregado al usuario (Montgomery et al, 2014).

Con la finalidad de poder analizar algunos métodos, se realiza una validación del modelo obtenido. Para ello, existen diversas técnicas, entre las cuales destacamos los métodos de validación cruzada. La validación cruzada hace uso de distintos subconjuntos de los datos disponibles para realizar el entrenamiento del modelo y su posterior validación (Pérez-Planells et al, 2015).

Muchos casos los nuevos modelos son representados sin una adecuada validación, por ello la validación propia de un modelo de regresión incluye un estudio de coeficientes para determinar si sus signos y magnitudes son razonables, así como el análisis en la estabilidad de los coeficientes de regresión (Montgomery et al, 2011).

El Análisis de Regresión Lineal Múltiple permite establecer la relación que se produce entre una variable dependiente Y y un conjunto de variables independientes (X_1, X_2, ... X_K). El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción.

En el análisis de regresión lineal múltiple la construcción de su correspondiente ecuación se realiza seleccionando las variables una a una, "paso a paso". La finalidad perseguida es buscar de entre todas las posibles variables explicativas aquellas que más y mejor expliquen a la variable dependiente sin que ninguna de ellas sea combinación lineal de las restantes.

Dentro del análisis existen dos estadísticos a analizar: R^2 ya que mide la capacidad explicativa de la variable X sobre la variable Y. al introducir en el modelo otra variable regresoras el nivel explicativo será mayor entre las dos que solo con la primera o, en todo caso, no disminuirá, pues la primera variable continúa como explicativa, dado esto la interpretación de R^2 no solo debe considerar la muestra, sino también el número de variables explicativas incluidas en el modelo (Martínez Rodríguez, 2005).

Y el estadístico PRESS el cual mide la calidad del modelo de regresión y es definida como la suma de cuadrados de los errores de la predicción para la variable que represente a la desviación al cuadrado entre el valor observado y estimado (Valencia Delfa et al, 2003).

1.1 Marco teórico

Como el ajuste del modelo a los datos disponibles forma la base de muchas de las técnicas que se usaron en el proceso de desarrollo del modelo (por ejemplo, la selección de las variables), es atractiva la conclusión de que un modelo que se ajusta bien a los datos también funcionará bien en la aplicación final; eso no es necesariamente cierto; por ejemplo, un modelo puede haber sido desarrollado principalmente para predecir observaciones nuevas. No hay la seguridad de que la ecuación que proporciona el mejor ajuste a los datos observados sea un buen predictor (Montgomery et al., 2014).

1.1.1 Técnicas de validación

El uso final pretendido del modelo indica, frecuentemente, cuál es la metodología adecuada de validación, así, la validación de un modelo para usarse como ecuación de predicción se debe concentrar en la determinación de la exactitud del modelo. Sin embargo, como el desarrollador consiguientemente no controla el uso del modelo, se recomienda que, siempre que sea posible, se usen todas las técnicas de validación que se acaban de mencionar (Snee, 1977) (Detteet al, 1998).

1.1.1.1 La estadística PRESS y $R^2_{predicción}$ basado en PRESS

Por sus siglas en inglés Predicción Error Sum of Squares, es una medida de cómo podrá el modelo predecir nueva información y se calcula a partir de los errores de predicción obtenidos al predecir el i-ésimo punto de información con un modelo que incluye todas las observaciones excepto la i-ésima observación (Allen, 1971) (Liu et al., 1999). Un modelo con un valor PRESS pequeño indica que el modelo es un buen predictor (Montgomery et al., 2014).

Este estadístico $R^2_{predicción}$ basado en PRESS da cierta indicación de la capacidad predictiva del modelo de regresión y cabe esperar que explique el porcentaje de la variabilidad cuando se predigan nuevas observaciones, en comparación con el porcentaje de variabilidad en los datos originales, explicado por el ajuste de mínimos cuadrados (Montgomery et al., 2007).

1.1.1.2 Recolección de nuevos datos – Corridas de confirmación

La recopilación de nuevos datos proporciona un control total sobre el proceso de construcción de todo el modelo (Snee, 1977). Si el modelo produce predicciones exactas para nuevos datos, el usuario tendrá más confianza tanto en el modelo como en el proceso de su obtención. A veces, a esas nuevas observaciones se les llama corridas de confirmación y es preferible contar con al menos de 15 a 20 observaciones nuevas, para tener una evaluación fiable de la eficiencia de predicción del modelo (Montgomery et al., 2007).

1.1.1.3 División de datos (validación cruzada)

La idea de validación cruzada se originó en la década de 1930´s (Larson, 1931), donde una muestra se utiliza para la regresión y una segunda para la predicción; Una declaración clara de validación cruzada, que es similar a la actual versión de k-fold, apareció por primera vez en (Mosteller et al, 1968); en la década de 1970´s, se empleó la validación cruzada como medio para elegir los parámetros del modelo adecuado, en lugar de emplear dicha técnica puramente para estimar el rendimiento del modelo. Actualmente, la validación cruzada sirve como un procedimiento estándar para la selección de modelo y estimación de rendimiento (Stone, 1974) (Geisser, 1975).

2. METODOLOGÍA.

Se selecciona un modelo verdadero de primer orden (Ecuación 1) en base a las necesidades de la investigación. Se seleccionará un modelo lineal múltiple, con tres variables regresoras (x_1, x_2, x_3), tres interacciones de dos factores (x_1x_2, x_1x_3, x_2x_3) y un error aleatorio U (0,1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon \quad (1)$$

Se simularán conjuntos de n (n=30, 50, 100, 500 y 1000) cantidad de datos y un nivel de ruido (bajo ($3\sigma^2$), medio ($2\sigma^2$) y alto (σ^2)) según los requisitos del experimento, se usará la simulación Monte Carlo para generarlos ya que utiliza una secuencia de números aleatorios para llevar a cabo una simulación estadística con el fin de conocer algunas propiedades estadísticas del sistema. Se diseñaron diferentes escenarios (tamaño de conjunto de información vs. Nivel de ruido), combinando diferentes valores (Figura 1), que permitirán analizar las variables que influyen sobre el número óptimo de datos que deben ser guardados al momento de realizar la validación del modelo.

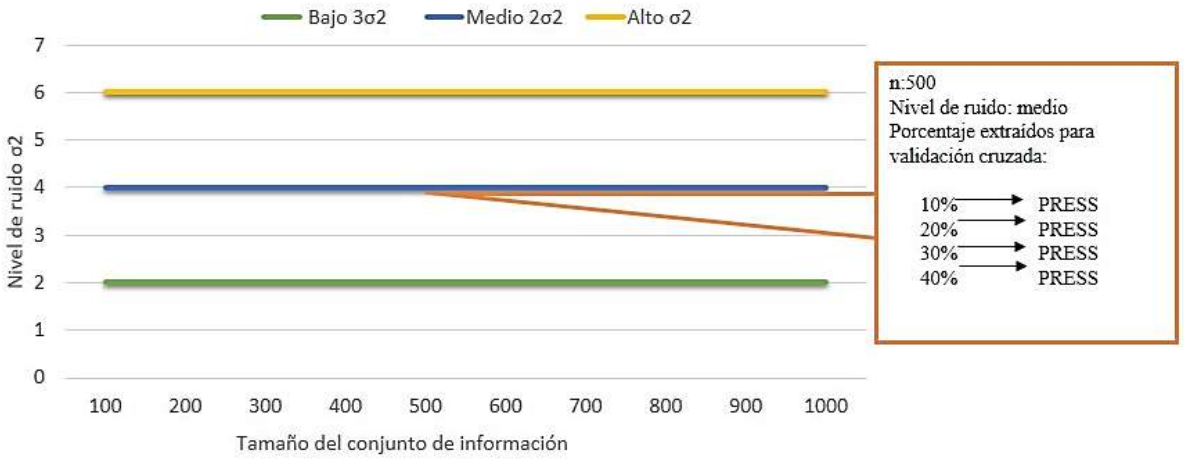

Figura 1: Planteamiento para realizar validación cruzada en los diferentes escenarios posibles (Fuente: Elaboración propia).

Al aplicar la técnica de validación cruzada se toma una muestra de la población n, con un nivel de ruido determinado, con un porcentaje diferente de separación de los datos de prueba (10%, 20%, 30% y 40%), a partir de esos datos se determina la suma de cuadrados de los errores de la predicción y se observa gráficamente el comportamiento de cada porcentaje de datos que se determinó para cada interacción (Arriaga Balderas, 2017).

Mediante el uso del software MATLAB 2012,se programa un algoritmo (Figura 3) en base a una distribución uniforme (considerando la misma probabilidad de ocurrencia para cada evento) se simula un conjunto de información con una matriz (*1000 X 7*) tabla 1, se divide en dos grupos una que se mencionó es de estimación (entrenamiento) y otra de predicción (prueba) de manera aleatoria con un determinado porcentaje determinado de datos a separar y se almacenan esos datos en dos matrices; de la matriz de estimación se obtienen los coeficientes de regresión y valores involucrados de la matriz de predicción se obtiene el estadístico PRESS, éstos a su vez se almacenan en otra matriz, pues se realizan n *iteraciones* para tal porcentaje de datos de predicción y se observa gráficamente el comportamiento de la capacidad predictiva del modelo.

Tabla 1: Matriz del conjunto de n=1000 nivel alto de ruido con 10% de separación.

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	yo
1	-0.59250605	-0.71072966	0.98467817	0.42111163	-0.58342777	-0.69983998	1.86267701
2	0.30510748	-0.55171385	0.05480294	-0.16833202	0.01672079	-0.03023554	4.39184905
3	-0.00640889	0.15150905	-0.65217538	-0.00097101	0.00417972	-0.09881047	1.83090065
4	0.95333339	0.32066831	0.37124604	0.30570381	0.35392125	0.11904684	8.56763243
5	-0.3776412	-0.09379163	-0.69424008	0.03541958	0.26217366	0.06511391	3.78660145
6	-0.9100206	-0.46986628	0.69004136	0.42758799	-0.62795185	-0.32422717	0.03982999
7	0.63708913	-0.8219517	-0.56486263	-0.52365649	-0.35986784	0.4642898	5.64996746
8	0.2330195	0.87210716	0.10537817	0.20321798	0.02455517	0.09190106	2.51866728
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

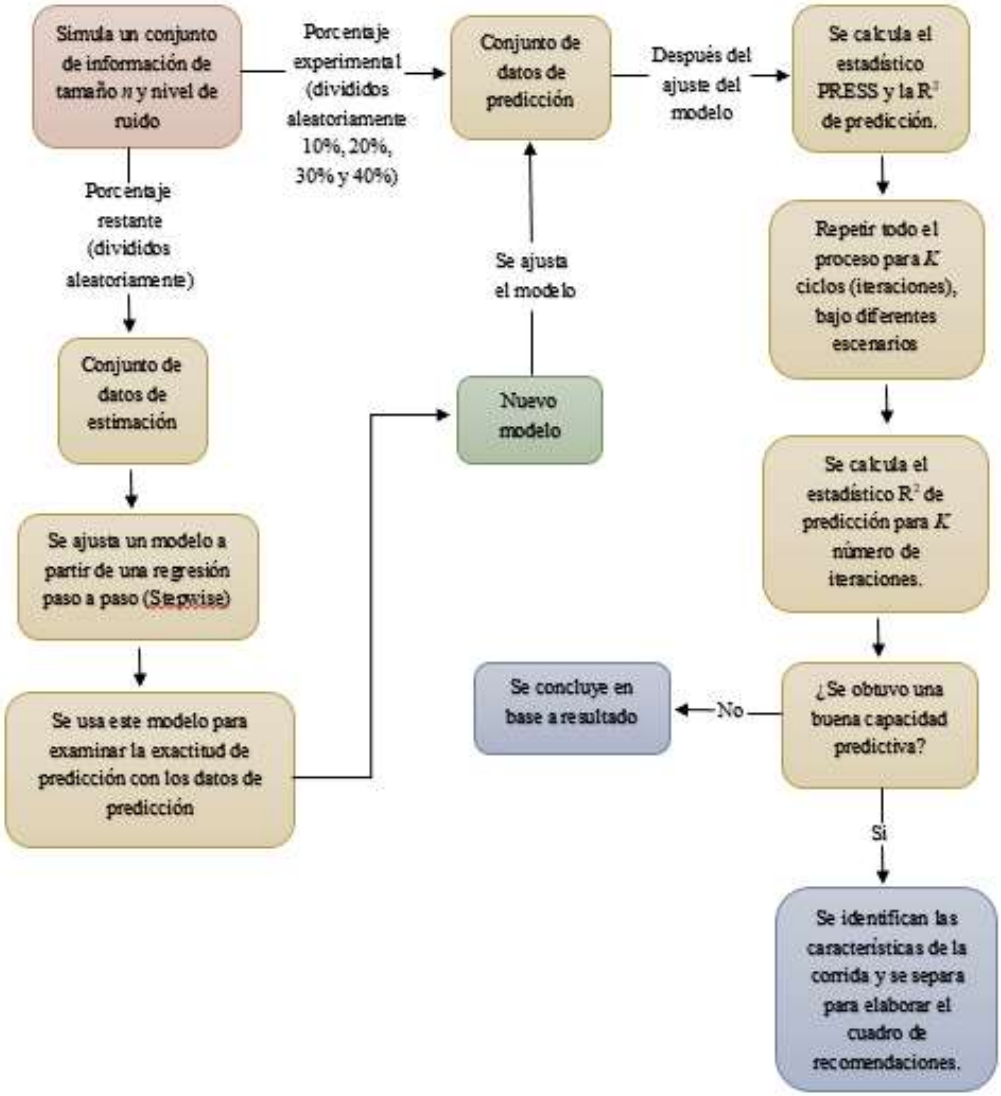

Figura 2: Planteamiento de la secuencia de pasos a realizar en el software MATLAB (Fuente: Elaboración propia).

Una vez obteniendo el valor PRESS y R^2 (tabla 2) de cada escenario planteado se realiza el análisis de los resultados obtenidos y posteriormente se realiza una comparativa con el fin de determinar el escenario más favorable para la investigación que corresponde a seleccionar el porcentaje a guardar óptimo generando la menor variación en el modelo.

Tabla 2: PRESS y $R^2_{\text{predicción}}$ del conjunto n=1000 con nivel de ruido alto.

	PRESS	$R^2_{\text{predicción}}$
1	118.645254	0.87843881
2	159.525367	0.82111443
3	94.8646656	0.8668562
4	100.351502	0.88335974
5	132.899034	0.86890237
6	121.596882	0.84470567

7	117.080373	0.85280013
8	116.744942	0.8952075
:	:	:

Los escenarios que se analizrán serán los siguientes mostrados en la figura 3.

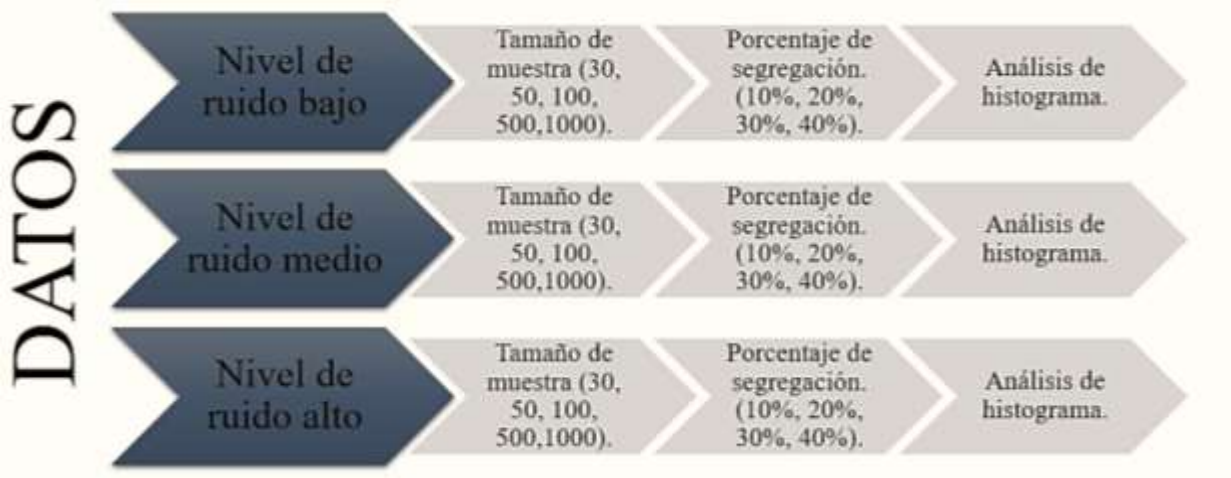

Figura 3: Escenarios planteados para la investigación (Elaboración propia).

De los escenarios propuestos se determinó un modelo de verdadero para cada nivel de ruido, tomando los lineamientos mencionados en la ecuación 1, las ecuaciones a utilizar son las siguientes (Ecuaciones 2-4):

Nivel de ruido bajo

$$y = 10 + 6x_1 - 5x_2 + 7x_3 - 6.8x_1x_2 + 6.3x_1x_3 + 0.6x_2x_3 + \varepsilon(0,2.1) \quad (2)$$

Nivel de ruido medio

$$y = 6 + 4x_1 - 2x_2 + 5x_3 - 4.8x_1x_2 + 4.3x_1x_3 + 0.9x_2x_3 + \varepsilon(0,2.1) \quad (3)$$

Nivel de ruido alto

$$y = 4 + 2x_1 - 1 + 3x_3 - 2.8x_1x_2 + 3.3x_1x_3 + 2.9x_2x_3 + \varepsilon(0,2.1) \quad (4)$$

3. RESULTADOS.

A continuación se dan a conocer los resultados de la experimentación realizada con el objetivo de encontrar el número óptimo de datos a guardar para llevar a cabo el método de Validación Cruzada aplicado en un modelo de regresión lineal múltiple y de igual manera verificar la capacidad predictiva del modelo bajo ciertos argumentos estadísticos como un valor bajo de PRESS (*Predicción Error Sum of Squares)* y el valor más alto de $R^2_{\text{predicción}}$, el cual muestra cuánto cabe esperar que el modelo explique la variabilidad de las nuevas observaciones.

Los resultados obtenidos se graficaron en histogramas, los cuales muestra de manera visual el comportamiento de los datos como se muestra en la figura 4 y 5respectivamente para el PRESS y $R^2_{\text{predicción}}$ para un conjunto de muestra tamaño 1000 con un nivel de ruido alto.

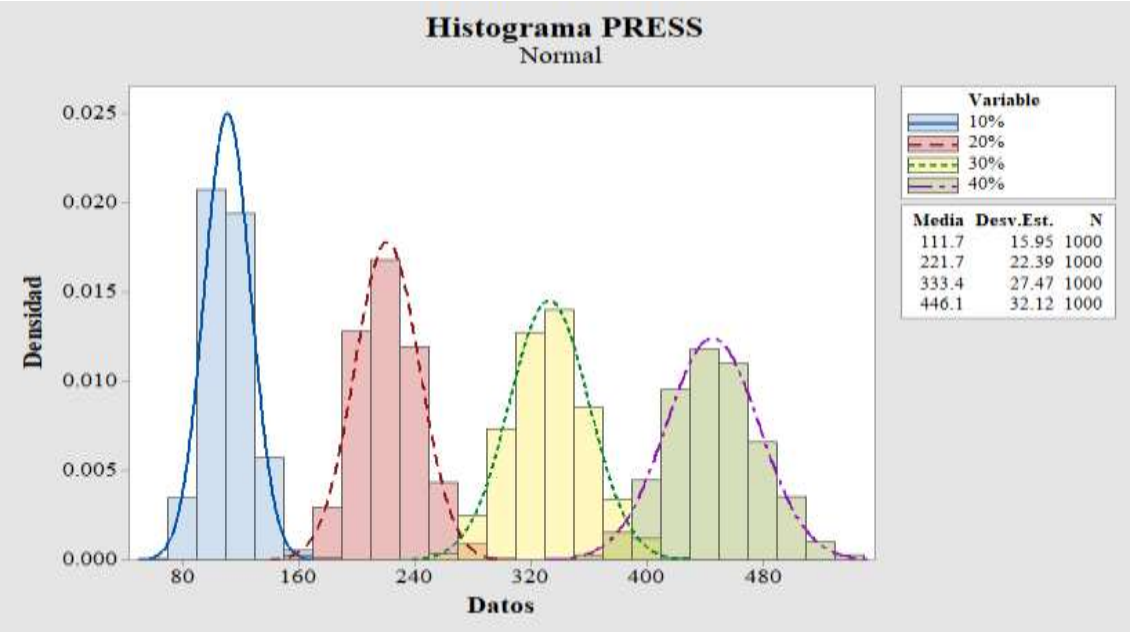

Figura 4: Histograma PRESS nivel de ruido alto con n=1000 (Minitab).

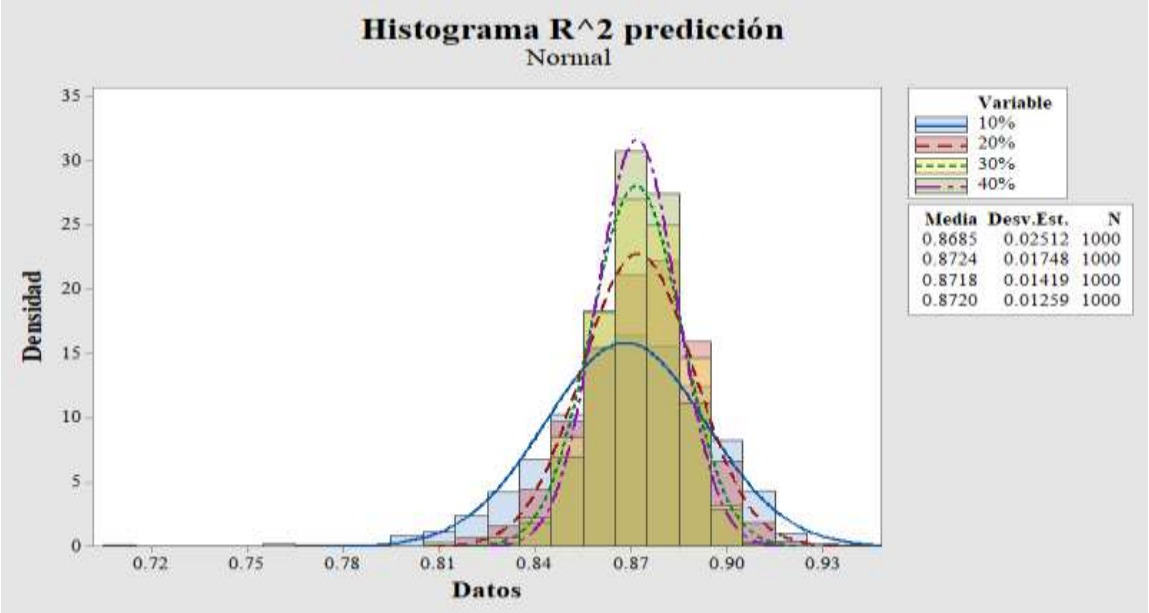

Figura 5: Histograma $R^2_{\text{predicción}}$ nivel de ruido alto con n=1000 (Minitab).

Tabla 3: Valores para la $R^2_{\text{predicción}}$ en nivel de ruido bajo para los cuatro porcentajes de separación y *n* diferentes

$R^2_{predicción}$		Tamaño del conjunto de información (nivel de ruido bajo)				
		30	**50**	**100**	**500**	**1000**
Porcentaje de datos (conju	**10**	98.04%	97.89%	96.24%	97.01%	97.34%
	20	97.07%	96.95%	97.38%	97.34%	97.54%

	30	97.93%	97.67%	97.56%	96.61%	97.04%
	40	97.81%	97.81%	97.68%	96.88%	96.84%

Tabla 4*:* Valores para la $R^2_{predicción}$ en nivel de ruido medio para los cuatro porcentajes de separación y *n* diferentes

$R^2_{predicción}$		Tamaño del conjunto de información (nivel de ruido medio)				
		30	**50**	**100**	**500**	**1000**
Porcentaje de datos (conjunto de predicción)	**10**	95.58%	94.64%	94.89%	93.72%	89.76%
	20	93.35%	94.87%	95.21%	94.87%	94.77%
	30	95.41%	93.52%	94.73%	93.52%	94.53%
	40	94.89%	94.92%	94.50%	94.21%	94.02%

Tabla 5: Valores para la $R^2_{predicción}$ en nivel de ruido alto para los cuatro porcentajes de separación y *n* diferentes.

$R^2_{predicción}$		Tamaño del conjunto de información (nivel de ruido alto)				
		30	**50**	**100**	**500**	**1000**
Porcentaje de datos (conjunto de predicción)	**10**	88.84%	88.67%	81.32%	84.40%	86.79%
	20	84.32%	84.18%	87.88%	86.78%	84.12%
	30	89.38%	87.64%	87.02%	86.44%	89.01%
	40	87.49%	87.47%	87.13%	88.08%	89.48%

Tabla 6: Valores para el PRESS en nivel de ruido bajo para los cuatro porcentajes de separación y *n* diferentes.

PRESS		Tamaño del conjunto de información (nivel de ruido bajo)				
		30	**50**	**100**	**500**	**1000**
Porcentaje	**10**	109.455137	137.336065	154.631401	116.813634	127.716982

	20	263.791744	287.020585	220.807645	216.394257	220.988674
	30	302.238539	308.390769	381.186692	357.719843	347.590536
	40	458.835454	411.642399	453.658005	509.22016	417.544415

Tabla 7: Valores para el PRESS en nivel de ruido medio para los cuatro porcentajes de separación y *n* diferentes.

PRESS		**Tamaño del conjunto de información (nivel de ruido medio)**				
		30	**50**	**100**	**500**	**1000**
Porcentaje de datos (conjunto de predicción)	**10**	109.455137	96.1998951	105.708743	100.798171	172.384255
	20	263.791744	188.070845	211.636562	216.394257	220.988674
	30	302.238539	377.073672	331.601402	357.719843	352.411825
	40	458.835454	410.860445	438.883688	500.846209	417.544415

Tabla 8: Valores para el PRESS en nivel de ruido alto para los cuatro porcentajes de separación y *n* diferentes.

PRESS		**Tamaño del conjunto de información (nivel de ruido alto)**				
		30	**50**	**100**	**500**	**1000**
Porcentaje de datos (conjunto de predicción)	**10**	109.455137	137.336065	154.631401	117.031057	131.332295
	20	263.791744	287.020585	220.807645	230.673495	241.579281
	30	302.238539	308.390769	381.186692	349.080184	311.169041
	40	458.835454	411.642399	453.658005	443.440227	364.469413

4. CONCLUSIONES Y PROYECTOS FUTURTOS.

En investigaciones posteriores a la realizada se encuentra que la manera más común de aplicar la técnica de Validación Cruzada es dejar el 10% de las muestras para realizar la evaluación y entrenar el 90% restante (Cox & Gaudard, 2013).

Dados los resultados en la investigación realizada es notorio que dependiendo el nivel de ruido el porcentaje de datos para evaluar no necesariamente será del 10%. Mostrando un panorama en el cual el investigador determinara de primera instancia el nivel de ruido para posteriormente segregar los datos con un porcentaje adecuado.

Con lo anterior se puede concluir que un modelo de regresión lineal múltiple permite manipular un grupo de datos que darán como resultado por medio de la aplicación de la técnica de Validación cruzada bajo diferentes condiciones de ruido y conjuntos de *n* tamaño obtener la cantidad de datos a guardad óptima.

Se determina que el estadístico PRESS en los tres niveles de ruido, el porcentaje óptimo es el 10 por ciento en todos los casos. Por ello es necesario validar con el segundo estadístico $\boldsymbol{R^2_{predicción}}$ donde los resultado obtenidos muestran valores que propician un panorama diferente a lo esperado, ya que para un nivel bajo el tamaño del conjunto de los analizados es el de n=30 con un porcentaje del 10 por ciento de separación; en un nivel de ruido medio obtenemos el mismo resultado donde el mejor es un tamaño del conjunto n=30 con un porcentaje de separación del 10%; mientras tanto para un nivel de ruido alto el tamaño del conjunto óptimo es de n=1000 con un porcentaje del 40 por ciento de separación.

Con los resultados obtenidos se determina que el tamaño de muestra afecta el estadístico $\boldsymbol{R^2_{predicción}}$ de manera significativa para seleccionar el tamaño del conjunto óptimo para la aplicación de la técnica de Validación cruzada.

Por lo tanto, se concluye que mientras el nivel de ruido sea bajo o medio el porcentaje a segregar óptimo será del 10 por ciento sin que afecte la capacidad predictiva; mientras que en un nivel de ruido alto el porcentaje será del 40 por ciento.

5. AGRADECIMIENTOS.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por otorgarme el apoyo financiero necesario para llevar a cabo mi desarrollo profesional. Al Dr. Armando Javier Ríos Lira por su dedicación y constante orientación a lo largo de la investigación realizada. Al Departamento del Posgrado de Ingeniería Industrial del Tecnológico Nacional de México en Celaya por su asistencia y soporte. A mi familia por el apoyo incondicional para lograr cada meta profesional y personal a lo largo de mi vida.

6. REFERENCIAS BIBLIOGRÁFICAS.

Allen, D. (1971). *The prediction sum of squares as a criterion for selecting predictor variables.* Kentuchy: Department of statistics. University of Kentuchy.

Arriaga Balderas, A. V. (2017). *Determinación número óptimo de datos para realizar una validación cruzada.* Celaya: Instituto Tecnológico de Celaya.

Cox, L., & Gaudard, M. (2013). *Discovering Partial Least Squares with JMP*. North Carolina, USA: SAS Institute, Inc.

Dette, H., & Munk, A. (1998). Validation of linear regression models. *The annals of statistics; Institute of mathematical statistics*, 778-800.

García, L., & Lara, P. (1998). *Diseño estadístico de experimentos. Análsisi de la Varianza.* España: Grupo Editorial Universitario.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Am. Stat. Assoc*, 70, 320-328.

Kerner, G. (1015). *Análisis Estadístico con el Método Bootstrap: Aplicaciones en Problemas de Regresión.* Argentina: Universidad de Buenos Aires.

Larson, S. (1931). The shrinkage of the coefficient of multiple correlation. *J. Educat. Psychol*, 22, 45-55.

Liu, H., Weiss, R., Jenrich, R., & Wenger, N. (1999). PRESS model selection in repeated measures data. *Computational statistics & daya analysis*, 16.

Martínez Rodríguez, E. (2005). Errores frecuentes en la interpretación del coeficiente de determinacion lineal. *Anuario Jurídico y Económico Escurialense*, 315-332.

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probalility for Engineers.* United Satates of America.: John Wiley & Sons, Inc.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2011). *Introducción al análisis de regresión lineal.* México. D.F.: Grupo Editorial Patria.

Montgomery, D., Peck, E., & Vining, G. (2007). *Introduction to linear regression analysis.* México: Continental Publishing Company.

Mosteller, F., & Tukey, J. (1968). *Data analysis, including statistics. In Handbook of social psychology.* M.A.: Addison-Wesley.

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *REVISTA DE TELEDETECCIÓN*, 44, 55-65.

Snee, R. (1977). Validation of regression models: methods and examples. *Technometrics C*, 15.

Stone, M. (1974). Cross-Validatory and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B*, 111-147.

Valencia Delfa, J., Díaz-LLanos, F., & Calleja, S. (2003). *Regresión PLS en las ciencias Esperimentales.* Madrid, España: Editorial Complutense.