Dra. Daniela Dianda; Dra. Marta Quaglino; Dr. José Pagura

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.

METODOS PREDICTIVOS DE DATA MINING EN EL CONTROL DE PROCESOS INDUSTRIALES.

Resumen.

Entre los objetivos principales del análisis de datos en contextos industriales, aparece la predicción, es decir, identificar una función que permita predecir el valor de una respuesta de interés a partir de los valores que toman otras variables consideradas como potenciales predictores de esa respuesta. Los grandes volúmenes de datos que la tecnología actual permite generar y almacenar han hecho necesario el desarrollo de técnicas de análisis alternativas a las tradicionales para lograr este objetivo, que permitan procesar y predecir la respuesta en tiempo real. Englobados bajo la denominación de Data Mining, muchos de estos nuevos métodos están basados en algoritmos automáticos originados mayormente en el ámbito informático. No obstante, la calidad de la información que alimenta a estos procedimientos sigue siendo un factor clave para asegurar la confiabilidad de los resultados. Con esta premisa es que en este trabajo se aborda el estudio del efecto que la presencia de fallas en los dispositivos de medición que originan la información, pueden causar sobre la capacidad predictiva de uno de los métodos disponibles, los árboles de decisión. Las medidas de eficiencia se definen a partir de la comparación con una técnica estadística tradicional, la regresión lineal múltiple. Los resultados señalan que la existencia de errores de medida tiene un efecto diferenciado sobre la capacidad predictiva de los árboles de decisión, según la naturaleza del error.

Palabras claves: Árboles de decisión CART; Regresión lineal; Error de medición; Error de predicción.

Abstract.

One of the main objectives of data analysis in industrial contexts is prediction, that is, to identify a function that allows predicting the value of a response from the values of other variables considered as potential predictors of this outcome. The large volumes of data that current technology allows to generate and store have made it necessary to develop methods of analysis alternative to the traditional ones to achieve this objective, which allow mainly to process these large amounts of information and to predict the response in real time. Enclosed under the name of Data Mining, many of these new methods are based on automatic algorithms mostly originated in the computer field. However, the quality of the information that feeds these procedures remains a key factor in ensuring the reliability of the results. With this premise, in this work we deal with the study of the effect that the presence of faults in the measurement devices that originate the information to be analyzed, can cause on the predictive ability of one of the predictive methods of data mining, the decision trees. The results are compared with those obtained using one of the traditional statistical techniques: multiple linear regression. The results obtained indicate that the effect of measurement related errors on the predictive ability of decision trees, compared to traditional regression models, depends on the nature of the measurement error.

Keywords: CART decision trees; Linear regression; Measurement error; Prediction Error.

INTRODUCCIÓN

La mejora de la calidad de procesos industriales requiere recolectar y analizar datos de los procesos, con el objetivo de resolver los problemas que se presenten y/o identificar oportunidades de mejora. El avance de la tecnología experimentado en las últimas décadas ha facilitado y automatizado la toma y registro de información sobre los procesos, permitiendo disponer en poco tiempo de grandes cantidades de datos que deben ser analizados. En este contexto, nuevos métodos de análisis fueron desarrollados para permitir procesar estas grandes cantidades de información con la rapidez necesaria para poder accionar en forma dinámica sobre los procesos. Englobados bajo la denominación de Data Mining, haciendo alusión a las actividades mineras, capaces de extraer materiales con valor económico de yacimientos donde están escondidos, muchos de estos nuevos métodos de análisis tuvieron su origen fuera del ámbito de la estadística y están basados en algoritmos automáticos desarrollados mayormente en el ámbito de la informática (Frawley et al. (1992), Apté (1997), Fayyad et al. (1996), Nisbet et al. (2009), Larose y Larose (2015)).

Mientras que en muchos campos de aplicación esta automatización se ha interpretado como un sustituto de la supervisión humana, la búsqueda de soluciones a un problema sigue siendo un proceso que requiere la intervención de personal calificado tanto en las técnicas que se implementan como en el conocimiento de las características propias del proceso analizado. La minería de datos debe ser y usarse como una herramienta más en el proceso de búsqueda de conocimiento.

Como sucede con cualquier técnica de análisis, la calidad de los datos es uno de los factores clave para asegurar la bondad de los resultados. Aunque el proceso de descubrimiento de conocimiento (KDD, *knowledge discovery in databases*) involucra pasos específicamente dedicados a la preparación y depuración de las bases de datos, existe una fuente de variabilidad adicional que, de ser controlada, puede ser evitada en forma previa a la toma y almacenamiento de la información: los errores de medición.

La mayoría de los datos disponibles, principalmente en el ámbito de procesos industriales, proviene de mediciones. Los sistemas de medición se han visto envueltos también en la vorágine del avance tecnológico y han evolucionado a la par. Hoy en día, las industrias cuentan con sofisticados equipos de medición incorporados a la maquinaria propia del proceso, que permiten evaluar continuamente diversas características de los productos en diferentes fases del proceso de producción y registrar automáticamente toda la información recolectada. Bajo este contexto, cualquier pequeña distorsión en alguno de los sensores o dispositivos de medición, podría resultar en una gran cantidad de información “contaminada” en un período de tiempo muy pequeño.

Existen varios trabajos en los que se ha demostrado que los errores de medición pueden ejercer un gran efecto en los resultados de varias de las técnicas tradicionales de análisis estadístico de uso frecuente en el campo de la mejora de procesos (Mittag (1995), Mittag (1997), Bordignon y Scagliarini (2002), Shishebori y Hamadami (2009), Dianda (2015)). Frente a tales antecedentes, es natural preguntarse si la presencia de fallas en algunos de los dispositivos de medición que produzcan mediciones imprecisas o sesgadas puede afectar el comportamiento de los algoritmos de data mining, en cuanto a su habilidad para reconocer los patrones de relación entre las variables existentes en los datos para proveer descripciones y/o predicciones confiables según sea el interés.

En este trabajo, se hace foco en uno de los métodos de data mining aplicables a problemas en los que el interés es predecir los valores de alguna respuesta de interés, los árboles de

decisión. Esta técnica es empleada para modelar datos reales de un proceso de producción de mezcla para cemento de construcción, y su desempeño, en términos de capacidad predictiva, es analizado frente a la presencia de errores de medición tanto aleatorios como estocásticos.

Los resultados obtenidos son posteriormente comparados con los obtenidos al emplear regresión lineal múltiple como técnica de análisis para predecir los valores de la respuesta de interés.

DISEÑO DEL ESTUDIO.

1. Técnicas empleadas.

El ***análisis de regresión lineal*** ha sido ampliamente utilizado en los más diversos ámbitos para modelar la relación estructural entre una variable respuesta o dependiente y una o un conjunto de predictores o variables independientes o explicativas. En general, el objetivo final es utilizar dicho modelo para predecir, tan precisamente como sea posible, los valores de la variable respuesta frente a observaciones futuras de las variables explicativas. Estos modelos resultan de fácil construcción y, por sobre todo, de fácil interpretación, siendo quizás estas características las que han popularizado el análisis. Sin embargo, la bondad del modelo obtenido depende del cumplimiento de una serie de supuestos o condiciones sobre los que se basa su construcción, limitando su correcto uso a situaciones en las que tales condiciones se satisfagan. Hoy en día, no es difícil encontrar situaciones en las que las variables de interés se relacionan a través de complejas funciones no lineales que raramente puedan ser advertidas o sospechadas a priori para ser tenidas en cuenta en la postulación de los modelos. De allí surge la necesidad de contar con métodos que permitan modelar estas relaciones de manera más flexible en cuanto a los requerimientos que deben cumplirse.

Los ***árboles de decisión*** son un método no paramétrico de aprendizaje supervisado, cuyo objetivo es crear un modelo para predecir los valores de una variable respuesta de interés, basado en reglas de decisión sencillas inferidas de los datos. En términos generales, un árbol de decisión tiene la estructura de un árbol invertido, con el tronco o raíz en el extremo superior y las ramas u hojas hacia abajo. Los algoritmos funcionan particionando reiteradamente los datos, con el objetivo de que cada partición genere grupos más homogéneos respecto de la variable de interés. Comenzando por el tronco o raíz (llamado “nodo raíz”) el árbol se divide en dos o más ramas, donde cada rama a su vez puede nuevamente dividirse en dos o más ramas. El proceso continúa hasta que se arriba a un “nodo terminal”, es decir, un nodo o rama que no puede ser nuevamente dividido (Nisbet et al. (2009), Larose y Larose (2015), Williams (2011)).

Los árboles de clasificación y regresión CART (Breinman et al., 1984) poseen una serie de propiedades que los hacen sumamente atractivos, siendo por ello uno de los algoritmos más utilizados en las aplicaciones prácticas. El algoritmo CART realiza particiones binarias en cada paso, de modo que la estructura final del árbol queda determinada por una red de preguntas simples; comenzando por el nodo raíz, la respuesta a cada pregunta determina qué rama del árbol seguir y cuál es la pregunta siguiente, hasta que se arriba a un nodo terminal, que representa la “decisión” final para esa observación.

2. Los datos.

El concreto u hormigón es un material esencial en la ingeniería civil, puesto que es utilizado para la construcción de las más diversas estructuras, desde una casa particular hasta puentes o grandes edificios. Una de las propiedades más importantes de este material es su

resistencia, existiendo una variedad específica denominada concreto de alta resistencia (HSC, por sus siglas en inglés, *high strength concrete*), el cual ofrece una resistencia superior a la del concreto normal, permitiendo en ciertas construcciones utilizar menos material para aliviar la carga de la estructura. El HSC es producto de una sofisticada mezcla de cemento, agua, agregados finos y gruesos y varios otros ingredientes. La resistencia del concreto depende de las proporciones de cada ingrediente que se agreguen a la mezcla, siendo la razón de agua a cemento uno de los factores clave. Cuando el cemento se mezcla con el agua crea un medio de cementación, si la mezcla es adecuada esta pasta cubre cada partícula del resto de los ingredientes y todos los espacios libres ente ellas. Cuando la pasta cuaja y endurece, todos los ingredientes quedan adheridos y transformados en masa sólida.

Las reacciones químicas entre el agua y el cemento que hacen que la pasta cementada endurezca se producen de manera rápida al principio y más lenta después, de modo que bajo condiciones normales, el concreto alcanza su máxima resistencia con el tiempo (Yeh, (1998), Deepa et al. 2010).

En este trabajo, se utiliza una base de datos de 1030 registros, sin datos faltantes, correspondientes a las composiciones de mezclas de concreto con los siguientes ingredientes: cemento, escoria, ceniza volante, agua, plastificante, agregado grueso y agregado fino. Cada uno de estos componentes de la mezcla constituye una variable de entrada para el proceso (X_1 a X_7, respectivamente), todas medidas en kg por metro cúbico de mezcla. Se tiene además una variable adicional que mide la "edad" del concreto, registrada en días (X_8). Por último, la base contiene para cada mezcla el valor de resistencia del concreto formado (Y), medida en megapascales (Mpa).

3. Errores de medición.

Un sistema de medición puede definirse como el conjunto de dispositivos, herramientas, procedimientos, personas y ambientes usados para asignar un número a una característica que está siendo medida (AIAG (2002)). Por lo tanto, una de las posibles causas de variabilidad en la salida del proceso es la forma en que se mide tal salida, con qué instrumento, en qué lugar, con qué operarios y bajo qué condiciones ambientales. Dado que el sistema de medición es también un proceso en sí mismo, como tal se verá afectado por causas de variación del mismo tipo de las que influyen sobre el proceso pudiendo originar mediciones inexactas o poco precisas. Un sistema de medición ideal es aquel que produce siempre mediciones correctas, esto es, precisas y exactas.

La precisión de un sistema de medición hace referencia a la variabilidad que se observa ante mediciones repetidas de la misma unidad bajo las mismas condiciones. Un sistema de medición será preciso si es capaz de producir los mismos resultados cuando se mide repetidamente una misma unidad, bajo condiciones uniformes. La exactitud del sistema se refiere a la diferencia que se observa entre el verdadero valor de la característica que se mide y el promedio de las mediciones que se obtienen al aplicar el procedimiento de medición. Un sistema de medición será exacto si posee la habilidad de producir mediciones que, en promedio, coincidan con el verdadero valor de la característica que se está midiendo.

El error de medición puede, a partir de esto, descomponerse en dos elementos: una componente sistemática, haciendo referencia a la exactitud de la medición, y una componente aleatoria, que se corresponde con la precisión de la medición.

Cuando se asume que en un proceso existen errores de medición, es necesario tener en cuenta que los valores observados de la o las características analizadas estarán afectadas por tales errores, de modo que lo que se observará no es la variable real o latente, X, sino una variable empírica, X^e (Mittag (1997)). La presencia de errores sistemáticos genera un sesgo constante en las mediciones, de modo que la variable empírica resulta: $X^e = \mathrm{X} + \mathrm{c}$,

donde c es una constante real que representa la magnitud del sesgo del sistema de medición. En cambio, la presencia de una componente estocástica de error implica la existencia de una variable aleatoria V que la representa, de modo que la variable observable resulta $X^e = \mathrm{X} + \mathrm{V}$.

4. Escenarios analizados.

En primer lugar, el conjunto de datos original, libre de errores de medición, es analizado mediante las dos técnicas mencionadas. Los modelos ajustados en cada caso son utilizados para establecer una medida de la capacidad predictiva de los mismos, la cual es utilizada como medida de referencia para las posteriores comparaciones.

En segunda instancia, los datos reales son modificados incorporándoles diferentes esquemas de errores de medición. Los datos modificados son utilizados para predecir la respuesta de interés usando los modelos construidos, y las predicciones resultantes son utilizadas para calcular una medida del error de predicción, que será comparada con la lograda por los modelos basados en datos reales, sin errores.

En cuanto a los esquemas de errores de medición introducidos, se consideran tanto errores de tipo aleatorio como sistemático. Dado que siete de los ocho predictores considerados corresponden a mediciones de peso de diferentes ingredientes, es razonable suponer que tales mediciones podrían ser susceptibles de error, introducidos por el instrumento o el operario. Dado que cada ingrediente tiene un rango de variación diferente, se supone que cada uno es medido con un dispositivo diferente, y por lo tanto los errores se introducen de manera independiente a cada una de las variables

Para el caso de errores de medición aleatorios, se asume que los mismos están generados mediante un modelo normal de media cero y variancia constante para cada variable $\sigma^2_{e_i}$. Bajo este supuesto, los valores reales de las variables X_i $(i = 1, \dots 7)$ contenidos en la base de datos original, son reemplazados por valores aleatorios de variables nuevas X_i^e, obtenidas sumando a las originales una cantidad generada aleatoriamente por el modelo asumido para los errores.

La incorporación de errores de tipo sistemáticos se lleva a cabo adicionando a los valores reales de las variables X_i, $i = 1, \dots, 7$, una cantidad constante c_i.

En ambos casos, errores aleatorios y sistemáticos, el estudio considera diferentes valores de $\sigma^2_{e_i}$ y c_i, con el objetivo de evaluar también el efecto del cambio en la magnitud de los errores de medición. Para $\sigma^2_{e_i}$ se considera una grilla de valores proporcionales a la magnitud de la variabilidad que presenta la variable X_i: $\sigma^2_{e_i} \in \{0.01j\ \sigma^2_{X_i},\ j = 1, \dots, 100, \}$, para $i = 1, \dots, 7$.

Similarmente, para la constante c se asume una grilla de valores que en este caso son proporcionales a la media de la variable que será afectada, contemplando valores entre el 1% y el 100% de la media de las variables originales, es decir, $c_i \in \{0.01j\ \bar{X}_i,\ j = 1, \dots, 100\}$ para $i = 1, \dots, 7$.

5. Criterios para el análisis.

Todos los análisis se implementan usando el software R-project. El conjunto de datos original es dividido aleatoriamente en un conjunto de entrenamiento y un conjunto de prueba, con una razón de 70 a 30 respectivamente. Los datos de entrenamiento son usados para construir y seleccionar los modelos, y una vez elegidos, éstos son utilizados para predecir la respuesta sobre los registros incluidos en el conjunto de datos de prueba. Los errores de medición se incorporan a los registros incluidos en este conjunto de prueba.

La bondad de los modelos ajustados se evalúa a través de dos medidas: la raíz del error cuadrático medio (RECM) y la desviación media absoluta (DMA), ambas calculadas a partir de las predicciones realizadas sobre el conjunto de datos de prueba.

Previo a los análisis de regresión lineal múltiple se realiza una selección de variables, aplicando el procedimiento paso a paso, con los parámetros que establece por defecto la función *step* de paquete *stats.*

Por su parte, el árbol de decisión es construido haciendo uso de la función *rpart* del paquete del mismo nombre. Los parámetros de control son establecidos de modo que el árbol crezca hasta su máxima expresión, es decir, hasta que todas las observaciones estén correctamente clasificadas, para luego proceder a la poda de los mismos en función del aporte que cada división de los nodos brinda a la reducción del error de predicción (calculado internamente por R mediante validación cruzada). En este sentido, una vez construido el árbol se procede a la poda del mismo seleccionando el primer valor de complejidad para el cual la próxima partición genera una reducción del error menor al 0,1%.

RESULTADOS.

1. Análisis de los datos libres de errores de medición.

Los 1030 registros contenidos en la base de datos analizada se dividieron aleatoriamente en un grupo de datos de entrenamiento, conteniendo el 70% de los registros, y el 30% restante fue asignado al conjunto de datos de prueba. La Tabla 1 muestra las medidas descriptivas de las variables explicativas y respuesta globales y por subgrupo, cuyos resultados corroboran el balance de los grupos.

Tabla 1: Medidas descriptivas de las variables analizadas.

	Datos completos			Datos de entrenamiento			Datos de prueba		
Variable	**Rango**	**Media**	**DE**	**Rango**	**Media**	**DE**	**Rango**	**Media**	**DE**
Cemento	102; 540	281,17	104,51	102; 540	286,20	105,21	102; 540	269,40	102,06
Escoria	0; 359,4	73,90	86,28	0; 359,4	74,86	87,59	0; 359,4	71,65	83,22
Ceniza volante	0; 200,1	54,19	64,00	0; 200	52,02	63,16	0; 200,1	59,25	65,74
Agua	121,8; 247	181,57	21,35	121,8; 247	181,90	20,83	121,8; 246,9	180,90	22,56
Plastificante	0; 32,2	6,20	5,97	0; 32,2	6,16	5,89	0; 32,2	6,30	6,17
Agregado grueso	801; 1145	972,92	77,75	801; 1145	972,90	78,22	801; 1134,3	972,90	76,79
Agregado fino	594; 992,6	773,58	80,18	594; 992,6	771,80	81,63	594; 992,6	777,70	76,63
Edad	1; 365	45,66	63,17	1; 365	44,47	59,66	3; 365	48,43	70,71
Resistencia	2,33; 82,60	35,82	16,71	2,33; 82,60	36,32	16,90	4,83; 81,75	34,64	16,21

El procedimiento de selección de variables aplicado previo al ajuste del modelo de regresión lineal múltiple, identifica como variables relevantes a las cantidades de cemento, plastificante, escoria, agua y cenizas con que se prepara la mezcla, así como también a la edad del cemento (Tabla 2). El análisis de la variancia para el modelo ajustado con los predictores seleccionados arroja un p-valor menor a 0.0001 y el coeficiente de determinación resulta 0.62.

Las estimaciones de los coeficientes del modelo indican que la resistencia del hormigón aumenta con la edad y al aumentar, en diferentes proporciones, las cantidades de todos los ingredientes a excepción del agua, cuyo aumento en la mezcla produce el efecto contrario. Específicamente, el modelo estimado resulta:

$$\hat{y} = 32{,}156 + 0{,}104\,X_1 + 0{,}086\,X_2 + 0{,}070\,X_3 - 0{,}235\,X_4 + 0{,}209\,X_5 + 0{,}125\,X_8 \qquad (1)$$

Tabla 2: Resultados del proceso de selección de variables.

Coeficiente	Estimación	Error std.	Valor t	p-valor
Intercepto	32,156	5,175	6,214	<0,001
Cemento (X_1)	0,104	0,005	20,723	<0,001
Plastificante (X_5)	0,209	0,105	2,003	0,045
Edad (X_8)	0,125	0,007	18,244	<0,001
Escoria (X_2)	0,086	0,006	14,451	<0,001
Agua (X_4)	-0,235	0,026	-9,013	<0,001
Cenizas (X_3)	0,070	0,010	7,232	<0,001

El modelo en (1) es usado para predecir la resistencia del concreto que se obtendría bajo cada una de las condiciones dadas por los registros del conjunto de datos de prueba. La comparación de los valores de resistencia predichos por el modelo con los valores reales observados bajo cada condición, se resumen en las dos medidas mencionadas, obteniéndose: $\boldsymbol{RECM_{RLM} = 10,483}$ y $\boldsymbol{DMA_{RLM} = 8,234}$.

En lo que respecta a la construcción del árbol de decisión, el árbol completo posee 460 nodos. La Tabla 3 muestra las complejidades y errores asociados a cada división sucesiva del árbol, para las primeras 17 particiones. La aplicación del criterio de selección del valor de complejidad establecido, indica que el árbol debe podarse a la altura de la partición número 12, pues la partición subsiguiente genera una reducción absoluta del error basado en validación cruzada menor al 0.1% elegido como punto de corte.

Tabla 3: Tabla de complejidad del árbol de decisión construido.

Complejidad	Número de particiones	Error relativo en entrenamiento	Error relativo en validación	Error estándar
2,41E-01	0	1,0000000	1,00935	0,048339
1,87E-01	1	0,7586929	0,76530	0,037851
6,59E-02	2	0,5720161	0,57773	0,029817
6,42E-02	3	0,5060859	0,54496	0,028787
4,07E-02	4	0,4418724	0,46804	0,025410
3,82E-02	5	0,4011390	0,45844	0,023822
3,22E-02	6	0,3629869	0,41659	0,021231
1,95E-02	7	0,3308067	0,37844	0,019360
1,89E-02	8	0,3113339	0,36729	0,017996
1,83E-02	9	0,2924311	0,36368	0,017950
1,73E-02	10	0,2740841	0,34855	0,017643
1,01E-02	11	0,2568203	0,29966	0,015162
1,00E-02	12	0,2467083	**0,28522**	0,014215
9,37E-03	13	0,2367019	**0,28619**	0,014347
8,80E-03	14	0,2273370	0,30085	0,019211
8,36E-03	15	0,2185320	0,29626	0,019093
7,86E-03	16	0,2101720	0,29640	0,019186
7,84E-03	17	0,2023080	0,29166	0,019148
...	...	...	...	...

Sólo se reproducen los primeros 17 registros de la tabla completa.

El árbol resultante posee entonces 13 nodos terminales, y las variables utilizadas en su construcción consideran las seleccionadas por el procedimiento de selección paso a paso aplicado previo al ajuste por regresión lineal y agregan una variable adicional. Es decir, la decisión se basa en criterios que consideran: cantidad de agua (X_4), cantidad de cemento(X_1), edad (X_8), cantidad de escoria (X_2), cantidad de plastificante (X_5) y cantidad

de agregado grueso (X_6). Las reglas que componen el árbol bajo estas variables se esquematizan en la Figura 1. La Figura 2 muestra las curvas del error relativo en función del número de particiones, para las primeras 100 particiones del árbol. La línea de puntos vertical señala el número de particiones en el que se ha podado el árbol.

Figura 1: Árbol de decisión elegido.

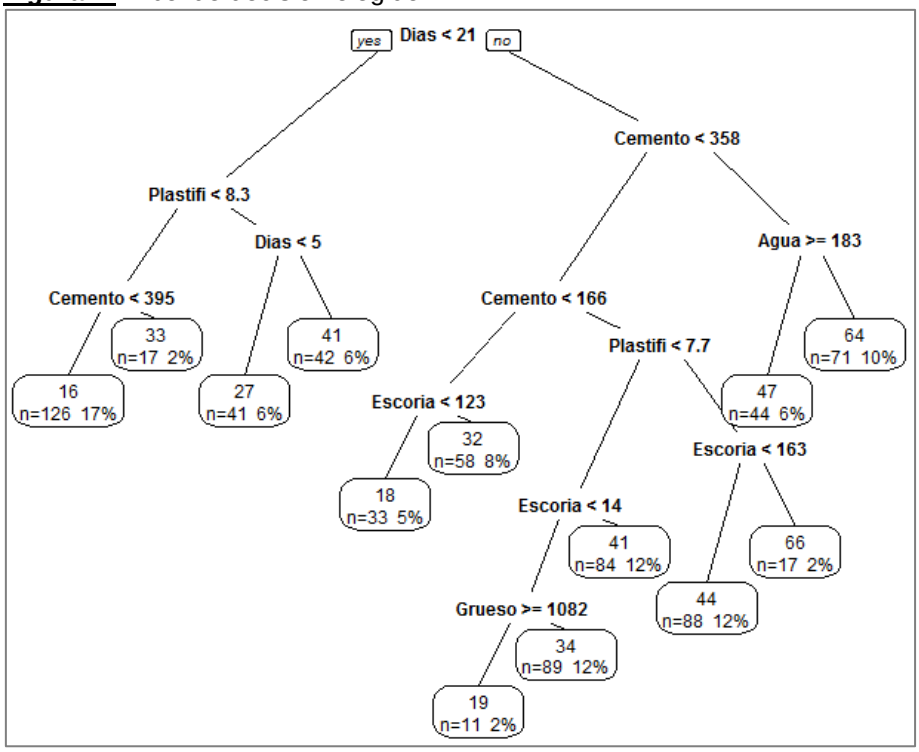
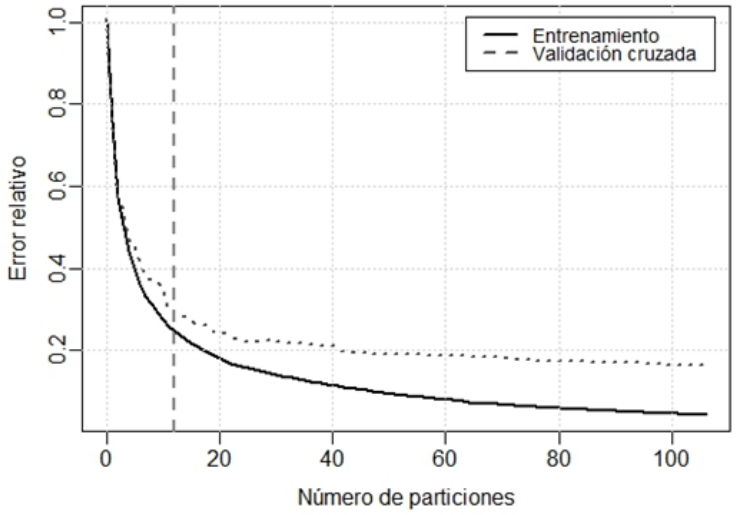


Figura 2: Curvas de error relativo para los datos de entrenamiento completos y por validación cruzada. La línea vertical señala el número de particiones del árbol podado.



Las reglas que componen el árbol seleccionado son aplicadas al conjunto de datos de validación para predecir la respuesta bajo cada condición. La comparación de los valores de resistencia predichos por el árbol con los valores reales observados, arroja los siguientes resultados: $\boldsymbol{RECM_{CART} = 9,359}$ y $\boldsymbol{DMA_{CART} = 7,390}$**.**

2. Predicción a partir de datos afectados por errores de medición aleatorios.

Para cada una de las magnitudes de error de medición consideradas ($j = 1, \dots, 100$) y para cada una de las variables X_i, $i = 1, \dots, 7$, se simulan 500 muestras de tamaño 309 (número de datos en el conjunto de prueba) de valores aleatorios normales de media cero y variancia dada por la magnitud del error en cada caso ($\sigma^2_{e_i}$). Estos valores se adicionan a las observaciones reales de cada variable en el conjunto de prueba, generando así, para cada magnitud de error de medición, 500 conjuntos de datos "contaminados", con los cuales se predice la respuesta utilizando el modelo de regresión en (1) y el árbol de decisión seleccionado (Figura 1). Los valores de RECM y DMA obtenidos en cada conjunto de datos son resumidos en media a través de las 500 simulaciones.

La Figura 3 muestra las medias de RECM de ambos análisis, en función de la magnitud del error de medición, este último expresado como porcentaje de la variancia de cada variable. El gráfico muestra que la capacidad predictiva tanto del modelo de regresión lineal como del árbol de decisión se ve afectada por la presencia de errores de medición aleatorios, evidenciado por un aumento en los valores del error cuadrático medio, respecto a los logrados para los datos sin error. En ambas técnicas este efecto se acentúa a medida que aumenta la magnitud de los errores de medición.

En términos relativos, el efecto resulta, en general, más nocivo para las predicciones obtenidas mediante los árboles CART, a menos que la magnitud de los errores de medición sea elevada (variabilidad introducida por el sistema de medición mayor al 70% de la variabilidad de la variable real). Esto puede verse en la Figura 4 donde se grafican, para cada método, las curvas de RECM promedio en datos con error de medición relativo al RECM obtenido a partir de los datos sin error de medida, en función de la magnitud del error de medición.

En lo que respecta a la comparación entre ambas técnicas, en la Figura 5 se grafica el cociente de RECMs entre CART y regresión lineal, encontrándose que para errores de medición de pequeña magnitud (menos del 30% de la variabilidad original) el modelo de regresión lineal tiene mejor comportamiento predictivo que CART, mientras que para errores de mayor magnitud CART es definitivamente mejor.

Figura 3: RECM promedio de los modelos ajustados sobre datos con error de medición aleatorio, en función de la magnitud del error de medición. Las líneas de puntos representan los valores de RECM obtenidos para los datos sin error de medición.

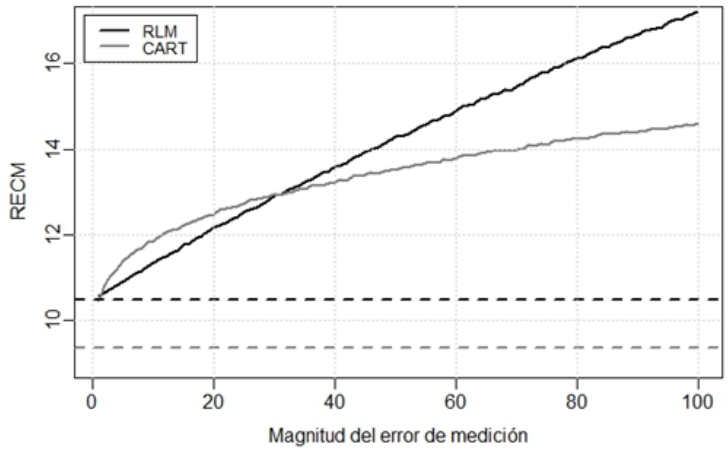


Figura 4: RECM relativo de los modelos ajustados sobre datos con y sin error de medición aleatorio, en función de la magnitud del error de medición. La línea de puntos señala el valor unitario para referencia.

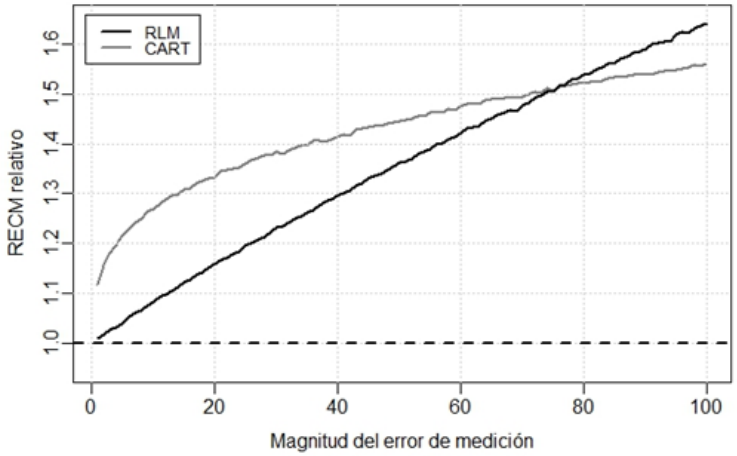
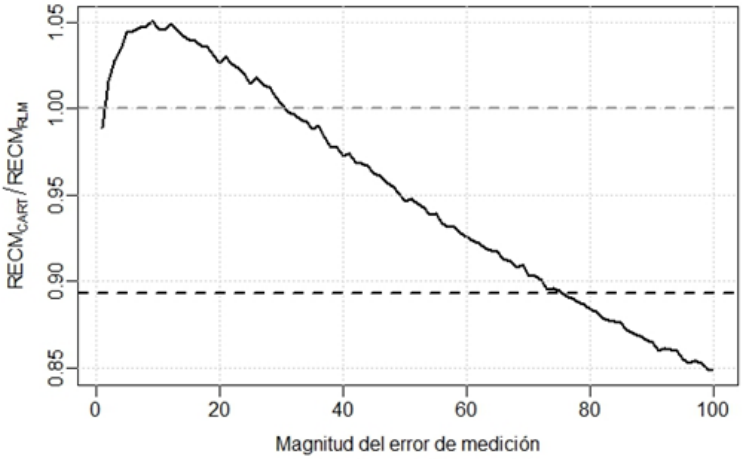


Figura 5: Cociente de RECMs entre CART y regresión lineal, en función de la magnitud del error de medición aleatorio. La línea punteada oscura corresponde a la razón de RECMs en datos libres de error. La línea de puntos más clara señala el valor unitario para referencia.



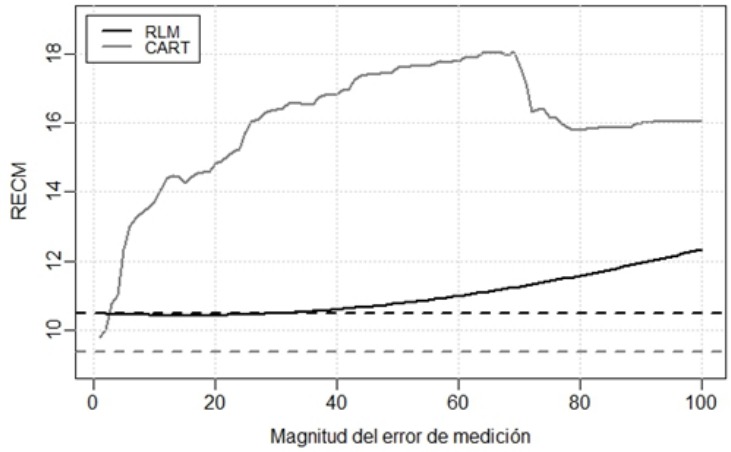
3. Predicción a partir de datos afectados por errores de medición sistemáticos.

En el caso de errores sistemáticos cada uno de los datos del conjunto de prueba es modificado mediante la adición de un error constante, proporcional a la media de la variable que corresponda en cada caso. De este modo, para cada una de las magnitudes de error consideradas se cuenta con un conjunto de datos "contaminado" asociado. Nuevamente estos datos son usados para predecir la respuesta de interés mediante las dos técnicas bajo estudio, y los valores obtenidos son utilizados para calcular el RECM y el DMA.

Las curvas de RECM para CART y regresión lineal se muestran en la Figura 6, junto con los valores de RECM de ambas técnicas logrados a partir de los datos libres de error. Es posible observar que ante la presencia de este tipo de errores, la capacidad predictiva de los árboles de decisión se ve seriamente afectada. No ocurre lo mismo con el modelo de regresión lineal, para el que se observa que su error de predicción se incrementa ligeramente a medida que aumenta la magnitud de los errores de medida, siendo este

incremento casi nulo para errores de magnitud inferior al 40% de la media de las variables.

Figura 6: RECM de los modelos ajustados sobre datos con error de medición sistemático, en función de la magnitud del error de medición. Las líneas de puntos representan los valores de RECM obtenidos para los datos sin error de medición.



La misma información pero en términos relativos se muestra en la Figura 7, donde cada RECM proveniente de datos con error de medición es dividido por el RECM obtenido a partir de los datos sin errores de medida. Las curvas muestran el serio deterioro en la capacidad predictiva de los árboles CART, ya que ante la presencia de errores de medición sistemáticos en los datos el RECM puede llegar a ser casi el doble de lo que en realidad es para los mismos datos sin error de medida.

En cuanto a la comparación entre ambos procedimientos, para datos libres de error CART mostró un mejor comportamiento que regresión lineal en términos del error de predicción. Sin embargo, la presencia de errores sistemáticos en las mediciones genera el comportamiento opuesto (Figura 8). El RECM obtenido por CART resulta superior al logrado mediante el modelo de regresión lineal. Dependiendo de la magnitud del error de medición, el error de predicción en CART puede llegar a ser hasta un 60% mayor que el que genera el modelo de regresión.

Figura 7: RECM relativo de los modelos ajustados sobre datos con y sin error de medición sistemático, en función de la magnitud del error de medición. La línea de puntos señala el valor unitario para referencia.

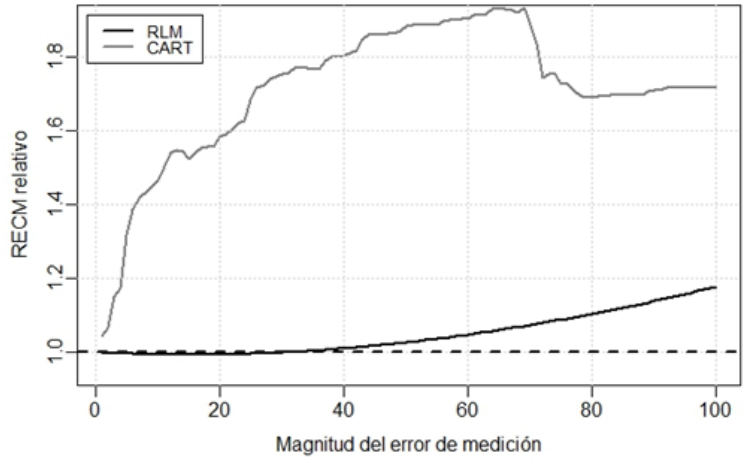
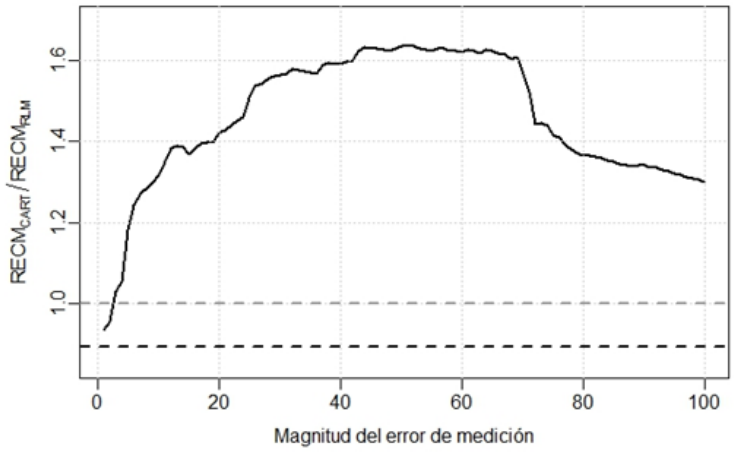


Figura 8: Cociente de RECMs entre CART y regresión lineal, en función de la magnitud del error de medición sistemático. La línea punteada oscura corresponde a la razón de RECMs en datos libres de error. La línea de puntos más clara señala el valor unitario para referencia.



CONCLUSIONES.

Los árboles de decisión han surgido como una alternativa no paramétrica para identificar o descubrir las relaciones existentes en un conjunto de variables, con el objetivo de crear un modelo que permita predecir el valor de una respuesta en función de las variables de entrada y sus relaciones. Hoy en día es quizás una de las herramientas de data mining más utilizadas en diversas disciplinas, en virtud de sus múltiples ventajas: es fácil de entender e interpretar, requiere poca preparación previa de los datos, puede incorporar variables tanto numéricas como categóricas, entre otras.

Desde el punto de vista práctico, sin embargo, el interrogante que ha motivado este trabajo es cuán robustos son estos algoritmos para realizar predicciones cuando los datos comienzan a incorporar "ruido" externo, en particular, el proveniente de los errores de medición.

Se aplicó el algoritmo CART sobre un conjunto de datos reales, libres de error de medición, para construir un árbol de regresión para predecir la resistencia del hormigón en función de una serie de variables que representan los componentes de la mezcla y el tiempo de "curado" del material. Este modelo fue utilizado luego para predecir la respuesta sobre un conjunto de datos a los que intencionalmente se les incorporaron diferentes esquemas de errores de medición, analizando el efecto que esto genera sobre el error de la predicción (RECM).

Los resultados mostraron que el error de predicción se ve siempre afectado por la presencia de errores de medida, ya sean aleatorios o sistemáticos, y aun cuando éstos sean de pequeña magnitud. No obstante, la presencia de errores sistemáticos se mostró mucho más nociva para la capacidad predictiva de los árboles que la de errores aleatorios, generando aumentos en el RECM de hasta un 60% aproximadamente. Adicionalmente, la comparación del desempeño predictivo del algoritmo CART con el de un modelo de regresión lineal múltiple, para datos libres de error, señaló al primero como de mejor comportamiento predictivo. Este resultado se mantiene ante la presencia de errores de medición aleatorios, aun cuando CART se ve más perjudicado por los errores de medición. Sin embargo, lo contrario ocurre ante la presencia de un sesgo constante en las mediciones. En este caso, el

deterioro en la capacidad predictiva de CART supera ampliamente al de regresión lineal múltiple.

Estos resultados, aunque corresponden al estudio sobre un caso particular, ponen de manifiesto la importancia de asegurar constantemente la calidad de la información recolectada, adquiriendo esto mayor relevancia aun en los contextos actuales en los que la automatización permite generar y almacenar una gran cantidad de datos en períodos muy cortos de tiempo, y su análisis en tiempo real direcciona acciones correctivas de manera casi inmediata.

REFERENCIAS BIBLIOGRÁFICAS

AIAG-Automotive Industry Action Group (2002). *Measurement System Analysis*, 3era Ed. Detroit, MI.

Apté, C. (1997). Data Mining: An Industrial Research Perspective. *IEEE Computational Science & Engineering*, 4(2): 6-9.

Bordignon, S., Scagliarini, M. (2002). Statistical Analysis of Process Capability Indices with Measurement Errors. *Quality and Reliability Engineering International*, 18: 321-332.

Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Deepa, C.; SathiyaKumari, K.; Pream Sudha, V. (2010). Prediction of the Compressive Strength of High Performance Concrete Mix Using Tree Based Modeling. *International Journal of Computer Applications*, 6(5): 18-24.

Dianda, D. (2015). Estudio estadístico de sistemas de medida e indicadores de capacidad de procesos multivariados, en contextos de mejora de la calidad y la productividad. Tesis Doctoral. Universidad Nacional de Rosario.

Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, 13(3): 57-70.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3): 37-54.

Larose, D.; Larose, C. (2015). *Data mining and predictive analytics*, 2da. Ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

Mittag, H-J. (1995). Measurement Error Effect on Control Chart Performance. *Annual Quality Congress*, 49(0): 66-73.

Mittag, H-J. (1997). Measurement Error Effects on the Performance of Process Capability Indices. *Frontiers in Statistical Quality Control*, 5: 195-206.

Nisbet, R.; Elder, J.; Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Elsevier Inc., Amsterdam.

Shishebori, D., Hamadami, A.Z. (2009). The Effect of Gauge Measurement Capability and Dependency Measure of Process Variables on the MC_p. *Journal of Industrial and Systems Engineering*, 4(1): 59-76.

Williams, G. (2011). *Data Mining with Rattle and R*. Springer, New York.

Yeh, I.C. (1998). Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12): 1797-1808.

FUENTE

La base de datos utilizada pertenece al Prof. I-Cheng Yeh(*), fue donada por él al UCI Machine Learning Repository (Universidad de California, Irvine), siendo de libre acceso desde https://archive.ics.uci.edu/ml/datasets.html, bajo aviso de copyright para el mencionado Prof. y el artículo de su autoría incluido en las referencias bibliográficas.

(*) Prof. I-Cheng Yeh (Original Owner and Donor). Department of Information Management, Chung-Hua University. Hsin Chu, Taiwan 30067, R.O.C. E-mail:icyeh@chu.edu.tw