TÉCNICAS DE SELECCIÓN DE VARIABLES EN REGRESIÓN LINEAL MÚLTIPLE

por

Jennifer Roque López

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalusia

December 2021

Técnicas de Selección de Variables en Regresión Lineal Múltiple

Jennifer Roque López

Máster en Economía, Finanzas y Computación
Antonio J. Tallón Ballesteros

Universidad de Huelva y Universidad Internacional de Andalucía

2021

Abstract

Variable selection is a crucial data preprocessing activity. Various techniques have been proposed, mainly under four approaches: filter, embedded, semi-wrapper and wrapper models irrelevant information. The objective of this Master dissertation is to apply variable selection methods to a test-bed to reduce irrelevant information and propose an action plan to be able to tackle tasks in the field of dimensionality reduction. For this purpose, two Stata commands are used, vselect and gvselect, proposed by Lindsey and Sheather, which allow the selection of variables after performing a linear regression. The CFS and ReliefF methods provided by the Weka tool are also used. For subsequent measurement and predictions, the Multiple Linear Regression model will be used.

Key words: machine learning, regression, filter, wrapper, variable selection, best subset.

Resumen

La selección de variables es una actividad crucial del preprocesamiento de datos. Diversas técnicas han sido propuestas, bajo cuatros enfoques: métodos filtro, embebido, semi-wrapper y wrapper. El objetivo del presente TFM es aplicar métodos de selección de variables a un conjunto datos para reducir la información irrelevante y proponer un plan de actuación para poder abordar tareas en el ámbito de reducción de dimensionalidad. Para tal fin se utilizan dos comandos de Stata, vselect y gvselect, propuestos por Lindsey y Sheather, los que permiten realizar la selección de variables tras realizar una regresión lineal. También se hace uso de los métodos CFS y ReliefF que brinda la herramienta Weka. Para la posterior medición y predicciones se utilizará el modelo de Regresión Lineal Múltiple.

Palabras claves: aprendizaje automático, regresión, filtro, wrapper, selección de variables, mejor subconjunto.

Contenido

Lista de tablas

Lista de figuras

Agradecimientos

Quiero agradecer en primer lugar a mi tutor, Antonio Tallón, por su gran ayuda y asesoramiento. A la Universidad Internacional de Andalucía, por permitirme transitar esta gran aventura. Y no podría olvidarme de mi familia, por su comprensión y cariño en todo momento.

1. Introducción

La información es actualmente uno de los recursos de mayor valor dentro de cualquier campo de investigación ya que ha crecido exponencialmente en los últimos años gracias a la aparición de internet, las redes sociales, etc. Actualmente, existen estudios en diferentes campos, dedicados a extraer información valiosa sobre tendencias, desafíos y oportunidades a partir de los datos. El uso de grandes volúmenes de datos es cada vez más frecuente, pero estos datos son complejos y provienen de diversas fuentes, de ahí la necesidad de su preprocesamiento. Existen diferentes mecanismos para el preprocesamiento y análisis de datos, entre ellos, técnicas estocásticas, estadísticas y las basadas en el aprendizaje automático.

En muchas ocasiones, la información que se desea examinar cuenta con atributos que son irrelevantes para el análisis, otros son redundantes o la base de datos tiene un elevado número de variables explicativas (más conocido como la maldición de la dimensionalidad), lo que trae aparejado determinadas consecuencias. La primera puede producir problemas de sobreaprendizaje además de hacer más confusos los modelos resultantes; la segunda es nociva para ciertos algoritmos de aprendizaje; la tercera es una cuestión a tener en cuenta cuando hay pocos datos en relación a la presencia de muchos atributos. ¿Qué variables hay que seleccionar en un modelo?

El problema de la selección de variables puede plantearse del siguiente modo: dado un conjunto de atributos, seleccionar un subconjunto de ellos que optimice alguna medida de evaluación y que tenga el menor tamaño posible. De esta forma, mediante la aplicación de un algoritmo, se seleccionan las variables más apropiadas para la estimación de modelos, omitiendo aquellas que sean irrelevantes y/o redundantes.

Una vez que se ha aceptado el supuesto de que algunos datos no son relevantes, se obtienen algunos beneficios:

• Simplificación de modelos.

• Tiempos de formación más reducidos.

• Mejor generalización.

El objetivo del presente trabajo es aplicar métodos de selección de variables a un conjunto datos para reducir la información irrelevante y proponer un plan de actuación para poder abordar tareas en el ámbito de reducción de dimensionalidad.

Cada muestra del conjunto de datos se describe como un vector de características. El tamaño del conjunto de datos tiene dos dimensiones, el número de instancias o muestras (N, filas) y el número de características o variables (P, columnas). Una de las opciones más sencillas para reducir el problema en cuestión es eliminar muestras y características al azar. Pero si el número de muestras, N, se reduce sin cuidado, simplemente se eliminan los datos sin procesar y esto puede ser perjudicial para la tarea de aprendizaje, debido a la pérdida de información que pudiera ser relevante. Esto no suele ser una buena idea cuando se trata de un gran número de características.

La hipótesis que permite la simplificación es la presencia de datos redundantes. Hay que señalar que existen dos fuentes claras de datos no relevantes, una es la falta de cualquier tipo de relación entre un atributo y la variable a predecir y otra es la correlación o relación funcional de un conjunto de variables, que agrega información repetida.

Para afrontar este problema, han aparecido diferentes técnicas durante la última década, la selección de variables es una de ellas. Esta técnica selecciona un subconjunto de características, clasificando los subconjuntos con un valor de utilidad. Si el número de columnas, P, se reduce, el espacio de datos es simplificado y probablemente también el modelo objeto de estudio. Si se selecciona con cuidado las funciones a eliminar, casi no se pueden obtener pérdidas. Otra opción para simplificar el problema es utilizar métodos de reducción de dimensionalidad, como Análisis de Componentes Principales, (PCA), o Descomposición de Valores Singulares, (SVD).

Dependiendo del número de variables involucradas en el proceso de selección de un buen subconjunto y la forma en que se utilizan, se pueden clasificar los métodos en cuatro grupos:

- Métodos de filtro: se calcula la relevancia de los atributos y se eliminan los menos relevantes [3].
- Métodos embedidos: realizan la selección de variables dentro del propio proceso de entrenamiento del clasificador, incluyéndola en la función objetivo a optimizar.
- Método semi-wrapper: se lleva a cabo la evaluación mediante un método de aprendizaje automático distinto del algoritmo que se utilizará en predicción [25].
- Métodos de wrapper: la evaluación del subconjunto de variables se obtiene mediante aprendizaje y evaluación de un clasificador [11].

Hay muchos software útiles para trabajar el análisis de selección de variables, algunos de ellos son:

- Stata: implementado por StataCorp. comandos vselect y gvselect desarrollados por los investigadores Lindsey, C., & Sheather, S.
- Weka: en Java, desarrollado por la Universidad de Waikato. Licencia GNU GPL.
- Scikit-Learn: en Python, por INRIA, Google y otros. Licencia BSD.
- Paquetes R: Lenguaje con buena comunidad de desarrolladores. Licencia GNU GPL.

La presente investigación está organizada de la siguiente manera: En el apartado 2 se hace una breve introducción sobre los principales aspectos teóricos que deben tomarse en consideración en este trabajo. En el apartado 3 se exponen las técnicas principales de selección de variables. En el apartado 4 se comenta todo el proceso de experiementación llevado a cabo, para posteriormente en el apartado 5, mostrar los resultados obtenidos.

En el análisis se han empleado ocho bases de datos, que corresponden a problemas de regresión. Han sido extraídas del repositorio Machine Learning Database de la Universidad Central de California en Irvine (UCI) y fueron procesadas usando las herramientas Stata y Weka [7].

2. Aspectos teóricos claves

En este apartado se exponen los principales aspectos teóricos. Se define qué se entiende por aprendizaje automático, regresión lineal simple y múltiple y se comentan las principales métricas a tener en cuenta para seleccionar el mejor modelo de regresión. No pretende ser una revisión completa de todos los puntos relevantes relacionados con estas técnicas, pero solo un resumen de teoría básica y puntos útiles que son necesarios para comprender la investigación en cuestión.

2.1. Aprendizaje Automático

El Aprendizaje Automático (AA, o *Machine Learning*, por su nombre en inglés) es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente a partir de datos, para identificar patrones y tomar decisiones con la mínima intervención humana [17][18]. El desarrollador no tendrá que sentarse a programar por horas tomando en cuenta todos los escenarios posibles ni todas las excepciones posibles. Lo único que hay que hacer es alimentar el algoritmo con un volumen gigantesco de datos para que el algoritmo aprenda y sepa qué hacer en cada uno de estos casos [23]. Infiere conocimientos a partir de datos con el uso de técnicas que provienen de diferentes campos dentro de la Informática y las Matemáticas mediante diferentes enfoques o estilos, como tablas de decisión, árboles de decisión, modelos basados en reglas [5], enfoques conexionistas (redes neuronales artificiales, máquinas de soporte vectorial entre otros, cada uno con sus propias ventajas y limitaciones. Estas familias de algoritmos dan solución a problemas de regresión lineal y clasificación, según el tipo de situación que se presente.

Hay dos tipos de aprendizaje:

- Aprendizaje supervisado: Es cuando se entrena un algoritmo de aprendizaje automático dándole las preguntas (características) y las respuestas (etiquetas). Así en un futuro, el algoritmo puede hacer una predicción conociendo las características.

- Aprendizaje no supervisado: el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formados únicamente por entradas al sistema, sin conocer su clasificación correcta, por lo que se busca que el sistema sea capaz de reconocer patrones para poder etiquetar las nuevas entradas.

Dentro del aprendizaje supervisado, existen algoritmos de clasificación y de regresión. En los problemas de clasificación se espera que el algoritmo diga a qué grupo pertenece el elemento en estudio. El algoritmo encuentra patrones en los datos dados y los clasifica en grupos. Luego compara los nuevos datos y los ubica en uno de los grupos y es así como puede predecir de qué se trata. Por otra parte, en los problemas de regresión lo que se espera es un número. No lo ubica en un grupo, sino que devuelve un valor específico.

Este trabajo se centra en el análisis de bases de datos que se corresponden a problemas de regresión.

2.2. Nociones básicas sobre los modelos de regresión

En este apartado se explican de forma muy resumida los aspectos claves de los modelos de regresión usados más adelante. El apartado se divide en dos secciones: Regresión Lineal Simple y Regresión Lineal Múltiple.

2.2.1. Regresión Lineal Simple.

La regresión lineal como forma de aprendizaje supervisado tiene como objetivo predecir valores continuos a partir de datos históricos etiquetados. Para aprender este tipo de modelos es necesario establecer la relación entre un cierto número de características X continuas y una variable objetivo Y continua, donde el comportamiento de una variable dependiente Y, se puede explicar a través de al menos una variable independiente X, lo que se representa mediante una recta $Y = f(X)$ [19], o sea, una o varias variables dependientes pueden ser escritas en términos de una combinación lineal de las variables independientes [27]. Según

sea el número de variables independientes xi estamos en presencia de una regresión lineal simple o múltiple.

Un modelo de Regresión Lineal Simple (RLS) consiste en expresar la dependencia lineal de la variable objetivo o dependiente, y, respecto a la variable independiente o explicativa, x, y al término error o perturbación del modelo, ε.

Consideremos una relación del tipo:

$y_t = \alpha + \beta x_t + \varepsilon_t$ con t = 1,...,T

Donde:

y_t es la variable dependiente, explicada o endógena.
x_t es la variable independiente, explicativa o exógena.

A este modelo se le llama modelo de regresión lineal. Sus supuestos serán los siguientes [6]:

1.- La forma funcional es lineal: $y_t = \alpha + \beta x_t + \varepsilon_t$

2.- El error esperado es 0.

$E(\varepsilon_t) = 0$ (nota: si el error fuera igual a una constante, distinta de cero, el modelo sería determinista)

Este supuesto significa que la media de los errores respecto al modelo sea cero y que los errores se promedien entre sí, es decir, que se compensen de forma que no distorsionen sistemáticamente nuestra relación a estimar.

3.- Homoscedasticidad (la varianza del error es constante).

$Var(\varepsilon_t) = \sigma^2, \forall_t$

4.- Ausencia de correlación de los residuos.

$Cov(\varepsilon_t, \varepsilon_k) = 0, \forall_t \neq k$

Los supuestos 3 y 4 se pueden resumir diciendo que la matriz de varianzas y covarianzas es una matriz del tipo: Var (u)= σ^2_uI, lo que resume que los términos de error están incorrelados entre sí y que la matriz es escalar o esférica, siendo los elementos de la diagonal principal constantes. (Dicho de otra forma: no existe heteroscedasticidad ni autocorrelación).

5.- Ausencia de correlación entre el regresor y la perturbación.

$Cov(x_t, \varepsilon_k) = 0, \ \forall t,_k$

6.- Los coeficientes son constantes en el tiempo.

7.- Las variables explicativas no son linealmente dependientes (no existe multicolinealidad exacta)

8.- Las variables explicativas son deterministas.

9.- Normalidad de las perturbaciones.

Para hacer una estimación del modelo de regresión lineal simple, se trata de buscar una recta de la forma:

$$\hat{Y} = \hat{a} + \hat{B}X = a + bX$$

de modo que se ajuste a la nube de puntos. Para esto se utiliza el método de mínimos cuadrados. Este método consiste en minimizar la suma de los cuadrados de los errores, es decir, la suma de los cuadrados de las diferencias entre los valores reales observados (yi) y los valores estimados ($\hat{y}_i$) de la siguiente forma:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Sin embargo, hay que tener en cuenta que a la hora de representar este análisis no todos los puntos se encontrarán dentro de la línea de regresión. En los casos reales, los ajustes absolutos del modelo a la realidad no se dan. Es por eso es que existe una medida que describe

cómo de precisa es la predicción de Y en función de X. O, al revés, lo inexacta que puede ser la estimación. A esta medida se le llama error estándar de estimación. Se utiliza en el análisis de regresión lineal para medir la dispersión alrededor de la línea de regresión.

El coeficiente de regresión da información sobre el comportamiento de la variable Y frente a la variable X, de manera que:

a) Si b = 0, para cualquier valor de X la variable Y es constante (es decir, no cambia).
b) Si b > 0, esto indica que al aumentar el valor de X, también aumenta el valor de Y.
c) Si b <0, esto indica que al aumentar el valor de X, el valor de Y disminuye.

Una vez que se ha calculado la recta de regresión y la bondad de ajuste que se ha conseguido con el modelo de regresión lineal, el siguiente paso es realizar un contraste de hipótesis en el que la hipótesis nula se corresponderá con la ausencia de relación y el rechazo de la hipótesis nula con la presencia de una relación significativa.

Para ello, se debe contrastar si la correlación entre ambas variables es distinta de cero o si el modelo de regresión es válido en el sentido de contrastar si el análisis de la variable endógena (Y) es válido a través de la influencia de la variable explicativa (X).

En resumen, el análisis de regresión lineal se aplica a innumerables aspectos de la vida real. Se utiliza tanto en el ámbito social como el ámbito científico y es clave para entender algunas relaciones entre variables en estadística.

2.2.2. Regresión Lineal Múltiple

El modelo de Regresión Lineal Múltiple (RLM) es la extensión del modelo de Regresión Lineal Simple cuando consideramos k variables explicativas. En general, la variable objetivo, y, depende de muchas otras variables x1, ..., xk, aunque algunas de estas pueden no ser observables o desconocidas. El modelo de regresión incluye las que más efecto tienen y las

restantes las representa como una variable aleatoria que denominaremos perturbación o error del modelo.

Para realizar un análisis de regresión lineal múltiple se hacen las siguientes consideraciones sobre los datos [1]:

1.- Linealidad: los valores de la variable dependiente están generados por el siguiente modelo lineal: $Y = X * \beta + \mu$

Cada predictor numérico tiene que estar linealmente relacionado con la variable respuesta Y mientras los demás predictores se mantienen constantes, de lo contrario no se puede introducir en el modelo. La forma más recomendable de comprobarlo es representando los residuos del modelo frente a cada uno de los predictores. Si la relación es lineal, los residuos se distribuyen de forma aleatoria entorno a cero. Estos análisis son solo aproximados, ya que no hay forma de saber si realmente la relación es lineal cuando el resto de predictores se mantienen constantes.

2.- Homocedasticidad: todas las perturbaciones tienen la misma varianza: $V(u_i) = \sigma^2$

La varianza de los residuos debe de ser constante en todo el rango de observaciones. Para comprobarlo se representan los residuos. Si la varianza es constante, se distribuyen de forma aleatoria manteniendo una misma dispersión y sin ningún patrón específico. Una distribución cónica es un claro identificador de falta de homocedasticidad. También se puede recurrir a contrastes de homocedasticidad como el test de Breusch-Pagan.

3.-Independencia: las perturbaciones aleatorias son independientes entre sí: $E(u_i, u_j) = 0, \forall i \neq j$

Los valores de cada observación son independientes de los otros, esto es especialmente importante de comprobar cuando se trabaja con mediciones temporales. Se recomienda

representar los residuos ordenados acorde al tiempo de registro de las observaciones, si existe un cierto patrón hay indicios de autocorrelación. También se puede emplear el test de hipótesis de Durbin-Watson.

4.- Normalidad: la distribución de la perturbación aleatoria tiene distribución normal: $U \approx N(0, \sigma^2)$.

Los residuos se deben distribuir de forma normal con media cero. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a test de hipótesis de normalidad.

5.-Las variables explicativas X_k se obtienen sin errores de medida.

Si se admite que los datos presentan estas hipótesis entonces el teorema de Gauss-Markov establece que el método de estimación de mínimos cuadrados va a producir estimadores óptimos, en el sentido que los parámetros estimados van a estar centrados y van a ser de mínima varianza.

Para realizar una regresión lineal múltiple se deben seguir una serie de pasos. Los dos primeros pasos hacen referencia a la bondad del modelo, es decir, si el conjunto de variables independientes (explicativas) se relacionan con la variable dependiente (resultado).

1 – Significación de F-test: si es menor de 0,05 es que el modelo es estadísticamente significativo y por tanto las variables independientes explican algo la variable dependiente, cuánto algo es el R-cuadrado.

2 – R cuadrado (R^2): es cuánto las variables independientes explican la variable dependiente, indica el porcentaje de la varianza de la variable dependiente explicado por el conjunto de variables independientes. Cuanto mayor sea la R-cuadrado más explicativo y mejor será el modelo explicativo.

Los dos siguientes pasos hacen referencia a la influencia de cada una de las variables independientes:

3 – Significación de t-test: si es menor de 0,05 es que esa variable independiente se relaciona de forma significativa con la variable dependiente, por tanto, influye sobre ella, es explicativa, ayuda a predecirla

4 – Coeficiente beta (β): indica la intensidad y la dirección de la relación entre esa variable independiente y la variable dependiente. Cuanto más se aleja de 0 más fuerte es la relación.

El signo indica la dirección (signo + indica que al aumentar los valores de la variable independiente aumentan los valores de la variable dependiente; signo – indica que al aumentar los valores de la variable independiente, los valores de la variable dependiente descienden).

En los modelos lineales múltiples los predictores deben ser independientes, no debe de haber colinialidad entre ellos. La colinialidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores.

La Tolerancia (TOL) y Factor de Inflación de la Varianza (VIF) se tratan de dos parámetros que vienen a cuantificar lo mismo (uno es el inverso del otro). El VIF de cada predictor se calcula según la siguiente fórmula:

$$VIF_{\hat{\beta}_j} = \frac{1}{1 - R^2}$$

$$Tolerancia_{\hat{\beta}_j} = \frac{1}{VIF_{\hat{\beta}_j}}$$

Los límites de referencia que se suelen emplear son:

- VIF = 1: Ausencia total de colinealidad
- 1 < VIF < 5: La regresión puede verse afectada por cierta colinealidad.
- 5 < VIF < 10: Causa de preocupación
- El término tolerancia es 1/VIF por lo que los límites recomendables están entre 1 y 0.1.

En caso de encontrar colinealidad entre predictores, hay dos posibles soluciones. La primera es excluir uno de los predictores problemáticos intentando conservar el que, a juicio del investigador, está influyendo realmente en la variable respuesta. Esta medida no suele tener mucho impacto en el modelo en cuanto a su capacidad predictiva ya que, al existir colinealidad, la información que aporta uno de los predictores es redundante en presencia del otro. La segunda opción consiste en combinar las variables colineales en un único predictor, aunque con el riesgo de perder su interpretación [21].

Cuando se intenta establecer relaciones causa-efecto, la colinealidad puede llevar a conclusiones muy erróneas, haciendo creer que una variable es la causa cuando en realidad es otra la que está influenciando sobre ese predictor.

2.3. Métricas para seleccionar el mejor modelo de regresión

El mejor modelo de regresión es aquel que proporciona los mejores valores predichos. Por ello se debe utilizar un estadístico que permita comparar modelos. Existen muchísimos estadísticos, los estudiados en este trabajo son:

- Coeficiente de Determinación (R^2):

El Coeficiente de Determinación R^2 da la proporción de variación de la variable Y que es explicada por la variable X (variable predictora o explicativa). Si la proporción es igual a 0, significa que la variable predictora no tiene ninguna capacidad predictiva sobre la variable a predecir (Y). Cuanto mayor sea R^2, mejor será la predicción. Si llegara a ser igual a 1 la

variable predictora explicaría perfectamente la variación de Y, y las predicciones no tendrían error[22].

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

- Coeficiente de Determinación Corregido o Ajustado (R^2 ajustado):

El Coeficiente de Determinación Ajustado (R^2 ajustado) es la medida que define el porcentaje explicado por la varianza de la regresión en relación con la varianza de la variable explicada. Es decir, lo mismo que el R^2, pero con una diferencia: el coeficiente de determinación ajustado penaliza la inclusión de variables.

El coeficiente de determinación de un modelo aumenta aunque las variables que incluyamos no sean relevantes, ya que esto supone un problema, para intentar solventarlo, el R cuadrado ajustado queda tal que:

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1 - R^2]$$

En la fórmula, N es el tamaño de la muestra y k el número de variables explicativas. Por deducción matemática, a valores más altos de k, más alejado estará el R^2 ajustado del R^2 normal. Al revés, a valores más bajos de k, más cerca estará de 1 la fracción central y, por tanto, más parecidos serán el R^2 ajustado y el R^2 normal.

Decir que k es el número de variables explicativas y que no puede ser cero. Si fuese cero, no existiría modelo. Como mínimo se tiene que explicar una variable en función de otra variable. Dado que k debe ser como mínimo 1, el R^2 ajustado y el R^2 normal no pueden tener el mismo valor. Es más, el R^2 ajustado será siempre inferior al R^2 normal.

- Error Cuadrático Medio (MSE):

El error cuadrático medio (MSE o mean squared error en inglés) es una forma de evaluar la diferencia entre un estimador y el valor real de la cantidad que se quiere calcular. El MSE mide el promedio del cuadrado del error, siendo el error el valor en la que el estimador difiere de la cantidad a ser estimada. En otras palabras, se construye es estimador muestral de $E((y - X\beta)^2)$ como:

$$MSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}i)^2}{N}$$

Cuanto mayor sea este valor, peor es el modelo. Nunca es negativo, ya que se están cuadrando los errores de predicción individuales antes de sumarlos, pero sería cero para un modelo perfecto.

- Raíz del Error Cuadrático Medio (RMSE):

Es la raíz cuadrada del promedio de diferencias cuadradas entre la predicción y la observación real.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

El uso de RMSE es muy común y se considera una excelente métrica de error de propósito general para predicciones numéricas.

- Criterio de Información de Akaike (AIC):

El criterio de información de Akaike (An Information Criterion, AIC) proporciona un método simple y objetivo que selecciona el modelo más adecuado para caracterizar los datos experimentales. Este criterio, que se enmarca en el campo de la teoría de la información, se define como:

$$AIC = -2\ log(L(\theta b)) + 2K\ (1)$$

Donde log(L(θb)) es el logaritmo de la máxima verosimilitud, que permite determinar los valores de los parámetros libres de un modelo estadístico y K es el número de parámetros libres del modelo. Esta expresión proporciona una estimación de la distancia entre el modelo y el mecanismo que realmente genera los datos observados, que es desconocido y en algunos casos imposible de caracterizar. Como la estimación se hace en función de los datos experimentales, esta distancia es siempre relativa y dependiente del conjunto de datos experimentales. Por tanto, un valor individual de AIC no es interpretable por sí solo, y los valores AIC sólo tienen sentido cuando se realizan comparaciones utilizando los mismos datos experimentales.

- Criterio de Información de Akaike Corregido (AIC C):

Hurvich & Tsai en 1989 proponen una corrección del criterio de Akaike para el caso de modelos de regresión cuando el tamaño muestral no es elevado o el número de parámetros es relativamente grande, lo que palía el problema del sesgo y tiende a proporcionar modelos más adecuados[9]. Para ello, modifican el término de penalización, utilizando:

$$AICC = n \ln \hat{\sigma}_k + 2\frac{(k+1)n}{nk2}$$

- Criterio de Información Bayesiano (BIC):

Otro criterio de selección de modelos se basa en el denominado criterio de información bayesiano, comúnmente referido como BIC (por su abreviación en inglés). La expresión matemática que estima el BIC se define como:

$$BIC = (-ln\ L \times 2) + [\theta \times ln\ (n)]$$

Donde θ es el número de parámetros estimados y n es el número de observaciones. Ya con el valor estimado de BIC, el modelo candidato que resulta ganador se define con la menor estimación de BIC [26].

De acuerdo con esto, la elección de modelos candidatos de crecimiento individual no deben basarse en criterios como valores de coeficientes de determinación (R^2) o coeficientes de variación (CV), pues no son criterios adecuados para la elección de modelos, y en su lugar se sugieren criterios robustos como la teoría de la información.

- Coeficiente Cp de Mallows (Cp):

Este estadístico fue ideado por Colin Mallows y Cuthbert Daniel en 1973, de forma que tuviera pequeños valores como resultado:

La definición original de Cp es la siguiente:

$$c_p = \frac{RSS_p}{\hat{\sigma}_c^2} + 2p - n$$

Siendo RSS la suma de residuos al cuadrado del modelo con p regresores, $\hat{\sigma}^2$ es el estimador de la varianza del término de error del modelo completo (el modelo con todos los posibles regresores), p el número de regresores del modelo en cuestión y n el número de observaciones por predictor.

Si un modelo predice bien, entonces Cp ≈ p. Si un modelo predice mal, C será mucho mayor que p. En definitiva, el modelo a tener en cuenta es el que tenga un p pequeño y un valor p en torno o por debajo de p [16].

Por lo general, se debe buscar modelos donde el valor del Cp de Mallows sea pequeño y esté cercano al número de predictores del modelo más la constante (p). Un valor del Cp de Mallows pequeño indica que el modelo es relativamente preciso (tiene una varianza pequeña) para estimar los coeficientes de regresión verdaderos y pronosticar futuras respuestas. Un valor del Cp de Mallows que esté cerca del número de predictores más la constante indica que, relativamente, el modelo no presenta sesgo en la estimación de los verdaderos coeficientes de regresión y el pronóstico de respuestas futuras. Los modelos con falta de ajuste y sesgo tienen valores de Cp de Mallows más grandes que p.

3. Métodos de selección de variables

Dentro del aprendizaje automático, el área de la selección de variables se especializa en la tarea de seleccionar el subconjunto de atributos que pueden resultar de máxima utilidad para las tareas de clasificación o regresión posteriores.

Uno de los desafíos más importantes del análisis estadístico en grandes volúmenes de datos es identificar aquellas variables que provean información valiosa. Cuando el interés radica en ajustar un modelo de regresión, la elección del mejor subconjunto de variables predictoras es una de las cuestiones clave en la formulación del mismo [2].

La selección de variables puede aportar varios beneficios: reducción de overfitting, mejora de la precisión de las predicciones, eliminación de atributos redundantes, simplificación de los modelos y reducción del tiempo de proceso. El objetivo que persiguen las técnicas de selección de variables es obtener la lista de atributos más relevantes de un conjunto de datos dado. Si dicho proceso tiene éxito, un subconjunto de atributos aportará la misma información que el conjunto original, dejando por el camino, a los atributos irrelevantes o redundantes que pudieran existir en los datos originales.

Una cuestión importante en la selección de variables es que en algunos problemas puede ocurrir que algunos atributos no estén correlacionados con la clase por separado, pero si cuando actúan juntos, por lo que el objetivo último de la selección es encontrar el subconjunto de atributos más apropiado.

Dado que la bondad del subconjunto de atributos seleccionados se mide según un criterio específico, comúnmente dependiente de la aplicación que se quiera dar al mismo, no siempre será posible obtener un subconjunto de atributos que optimice cualquier criterio dado. Así, los atributos que compongan el subconjunto optimo según un criterio no necesariamente serán los óptimos para optimizar otro criterio distinto.

Los métodos de selección de variables pueden agruparse, atendiendo al tipo de evaluación en cuatro categorías: métodos filtro, embebido, semi-wrapper y wrapper, aunque este trabajo se centra en las técnicas filtro y wrapper, las cuales pueden aplicarse en ranking (atributos individuales) o para la selección de subconjuntos.

Los métodos de filtro seleccionan variables dependiendo del modelo utilizado. Se basan únicamente en características generales como la correlación con la variable que se va a pronosticar. Normalmente se eliminan las variables de menor interés, las demás formarán parte del modelo de regresión utilizado para clasificar/predecir los datos. Estos métodos permiten ahorrar tiempo y son especialmente robustos para el sobreaprendizaje (sobre-fitting). Sin embargo, como no tienen en cuenta las relaciones entre las variables, tendrán que seleccionar variables redundantes y, por lo tanto, se utilizarán generalmente en el preprocesamiento [3].

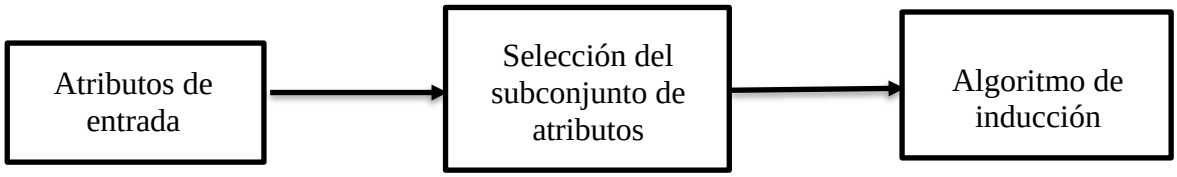

Figura 1: Esquema genérico de algoritmo tipo filtro

En las técnicas de wrapper, la evaluación del subconjunto de variables se obtiene mediante aprendizaje y evaluación de un clasificador. El clasificador es una caja negra, no es consciente de que se está realizando la selección. Hay cierto riesgo de sobreaprendizaje.

Los métodos wrapper, aunque son eficaces para eliminar atributos irrelevantes y redundantes, son muy lentos pues aplican numerosas veces el algoritmo minero, variando en cada ejecución el número de atributos, siguiendo algún criterio de búsqueda y paro[11]. Los métodos tipo filtro emplean algún tipo de medida de ganancia de información, distancia o consistencia, entre el atributo y la clase, siendo éstos mucho más eficientes que los wrapper; sin embargo, debido a que miden la importancia de cada atributo en forma aislada, no pueden detectar si existen atributos redundantes, y tampoco son capaces de determinar si la

combinación de dos o más atributos, aparentemente irrelevantes en forma aislada, se pueden transformar en relevantes.

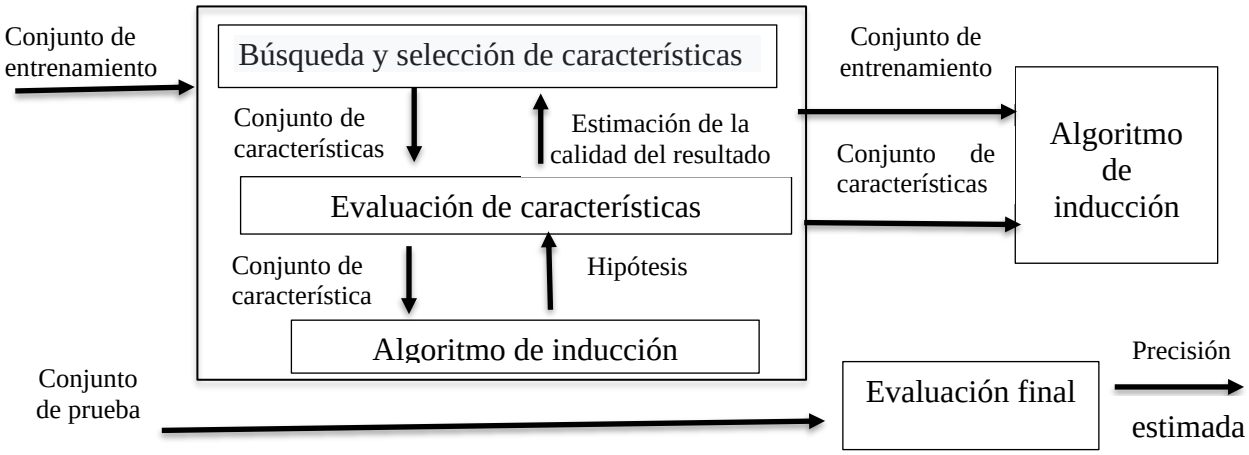

Figura 2: Esquema genérico de algoritmo tipo wrapper[11]

Para definir un método de selección de variables es necesario definir un espacio de búsqueda y un método de evaluación de la calidad de los subconjuntos, de ahí a que se clasifiquen en dos tipos principales: ranking y selección de subconjuntos [12][28].

- Tipo ranking: evaluación y ordenación de las características de manera individual eliminándose las menos valorada [4]. Existen varios criterios para evaluar estas características como son la entropía, chi-square, información mutua, etc. Una vez evaluadas y ordenadas, se eliminan las k peores. La principal ventaja es la rapidez y sus principales inconvenientes son que no elimina características redundantes y no detecta características que funcionen bien de manera conjunta.

- Selección de subconjuntos: no es más que la búsqueda del subconjunto de características más relevante. Estos métodos recorren un espacio de búsqueda de subconjuntos de características, evaluando subconjuntos completos de características. No se recorre el espacio entero, sino sólo aquellos subconjuntos más prometedores, además, evalúan el subconjunto de manera conjunta

Cuantos menos atributos, más fácil de interpretar es el modelo. Para ello, es necesario definir una manera de moverse por el espacio de búsqueda de subconjuntos de atributos, que puede ser mediante la selección hacia adelante o de la eliminación hacia atrás.

En la selección hacia adelante se comienza con el conjunto vacío de atributos y se van añadiendo variables según un criterio que distingue el mejor atributo del resto. En la selección hacia atrás, se comienza con el conjunto de todos los atributos e iterativamente se van eliminando uno a uno [5].

- Selección hacia adelante:

Paso 0: Se parte del modelo vacío de regresión. $Y = \beta 0 + \varepsilon$.

Paso 1: Entra la variable más significativa al modelo según el criterio que se esté utilizando, es decir, el modelo queda de la forma: $Y = \beta 0 + \beta 1X1 + \varepsilon$ y se realizan los siguientes contrastes:

$H0 : \beta 0 = 0$ y $H0 : \beta 1 = 0$
$H1 : \beta 0 \neq 0$ $H1 : \beta 1 \neq 0$

Paso 2: Entra la segunda variable más significativa. El modelo queda de la forma: $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$. Y se realizan los siguientes contrastes nuevamente:

$H0 : \beta 0 = 0$ y $H0 : \beta 1 = \beta 2 = 0$
$H1 : \beta 0 \neq 0$ H1 : Otro caso

Se va realizando este proceso tantas veces hasta que se encuentre una variable que no es significativa.

- Selección hacia atrás:

Paso 0: Este modelo parte con todas las variables a estudiar, es decir, $Y = \beta 0X0 + \beta 1X1 + \beta 2X2 + ... + \beta kXk + \varepsilon$.

Paso 1: Se identifica la variable menos significativa en el modelo. Se realiza un contraste de hipótesis, si la variable es no significativa, se elimina. El modelo quedaría de la siguiente manera: $Y = \beta 0X0 + \beta 1X1 + \beta 2X2 + ... + \beta k{-}1Xk{-}1 + \varepsilon$.

Se va realizando este proceso tantas veces hasta que todas las variables del modelo sean significativas.

- Selección paso a paso:

Este procedimiento es una combinación de los dos anteriores.
Paso 0: El modelo empieza vacío, $Y = \beta 0 + \varepsilon$.

Paso 1: Introduce la primera variable en el modelo y comprueba si es significativa. Si lo es entra en el modelo. $Y = \beta 0 + \beta 1X1 + \varepsilon$.

Paso 2: Se elimina o añaden más variables según los test de hipótesis. En cada paso se va comprobando si el modelo es significativo y las variables que no lo sean se eliminan. El proceso termina cuando no haya mejoras significativas a la hora de añadir o eliminar alguna variable.

Además de definir una manera de moverse por el espacio de búsqueda de subconjuntos de atributos, también es importante definir una medida para evaluar el subconjunto de atributos. Uno de los métodos usados en este trabajo es el CFS, conocido por sus siglas en inglés como Correlation Feature Selection.

El método CFS, clasificado como un método de filtro, evalúa un subconjunto de atributos calculando la media de las correlaciones (o similar) de cada atributo con la clase y las correlaciones por redundancia entre atributos.

$$\text{Evaluación (Ai)} = \frac{\text{correlación con la clase}}{\text{correlaciones entre atributos}} = \sum_j U(A_j, C) \Big/ \sqrt{\sum_i \sum_j U(A_i, A_j)}$$

Dentro de las ventajas que proporciona este método se encuentran: la rapidez y la eliminación de atributos redundantes, aunque también puede eliminar atributos que por sí solos no están correlacionados con la clase, pero que si lo están con otro atributo.

Otro método usado fue el ReliefF. Kira y Rendell formularon el algoritmo original de ReliefF inspirado en el aprendizaje basado en Instancias. Como método de selección de variables de filtrado de evaluación individual, ReliefF calcula una estadística aproximada para cada atributo que puede utilizarse para estimar la calidad o relevancia del atributo con respecto al concepto objetivo (es decir, predecir el valor del punto final) [10].

Según el método empleado, existen diferentes criterios para realizar la selección de los mejores modelos.

4. Experimentación

En esta sección se describen los conjuntos de datos y métodos (algoritmos) utilizados para realizar las simulaciones, asistidos por las herramientas Stata y Weka. Se utilizaron 8 bases de datos, las cuales fueron extraídas del repositorio de la Unversidad de California en Irvine (UCI) y se corresponden a problemas de regresión [7]. Tienen como promedio 9908 instancias y 15 atributos y se muestran en la tabla 1 a continuación:

Tabla 1: Lista de bases de datos

Base de Datos	Instancias totales	Cantidad de atributos	Variable a predecir
Abalone	4177	8	class_reg
Bodyfat	252	14	class
Cpu.small	8192	12	usr
Elevators	16599	18	goal
Fried	40768	10	Y
Housing	506	13	class
Puma32H	8192	32	theatadd6
Sensory	576	11	score

Las herramientas usadas para el preprocesamiento de los datos fueron:

- Stata, software de estadística completo e integrado que provee todo lo que se necesita para el análisis de datos. Fue creado en 1985 por StataCorp. Cuenta con una suite amplia de características y cientos de herramientas estadísticas.

- Weka, plataforma de software libre para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato.

Primeramente se realiza un análisis exploratorio de los datos, lo que permite analizar las variables contenidas en las bases de datos, verificando observaciones duplicadas, posibles problemas de la base o las características de las variables presentes. Para ello se utilizan en Stata algunos comandos de verificación de datos como:

- *codebook,* informa sobre las variables: estadísticas descriptivas, etiquetas, detalle de los valores missing, etc.
- *describe*, enumera las variables, sus etiquetas y formatos. También da información sobre la memoria utilizada por los datos y el recuento de observaciones de los datos.
- *summarize*, brinda estadísticas resumidas de las variables especificadas, muestra el total de observaciones, la desviación estándar de cada variable de la base de datos así como el máximo y mínimo.
- *correlate*, expresa la correlación existente entre cada variable de la base de datos.

Se utilizó Weka para realizar un hold-out 80-20, donde se divide al conjunto de observaciones en dos grupos (datos de entrenamiento y datos de prueba).

Al 80% que representa al grupo de entrenamiento se le aplicó validación cruzada en Weka con 5-folds mediante el método StratifiedRemoveFolds. Cada fold se analizó de manera independiente, por lo que para cada procedimiento metaheurístico se ejecutó el algoritmo cinco veces con diferentes semillas para obtener estadísticas confiables para luego llegar a una conclusion final sobre cada base de datos analizada. Todas las particiones de las bases de datos generadas en Weka se exportaron a Stata para realizar la Regresión Lineal Múltiple sin

realizar la selección de variables y posteriormente aplicando la selección de variables. También se realiza selección de variables en Weka, utilizando el método CFS y ReliefF y se comparan los resultados obtenidos en ambas herramientas.

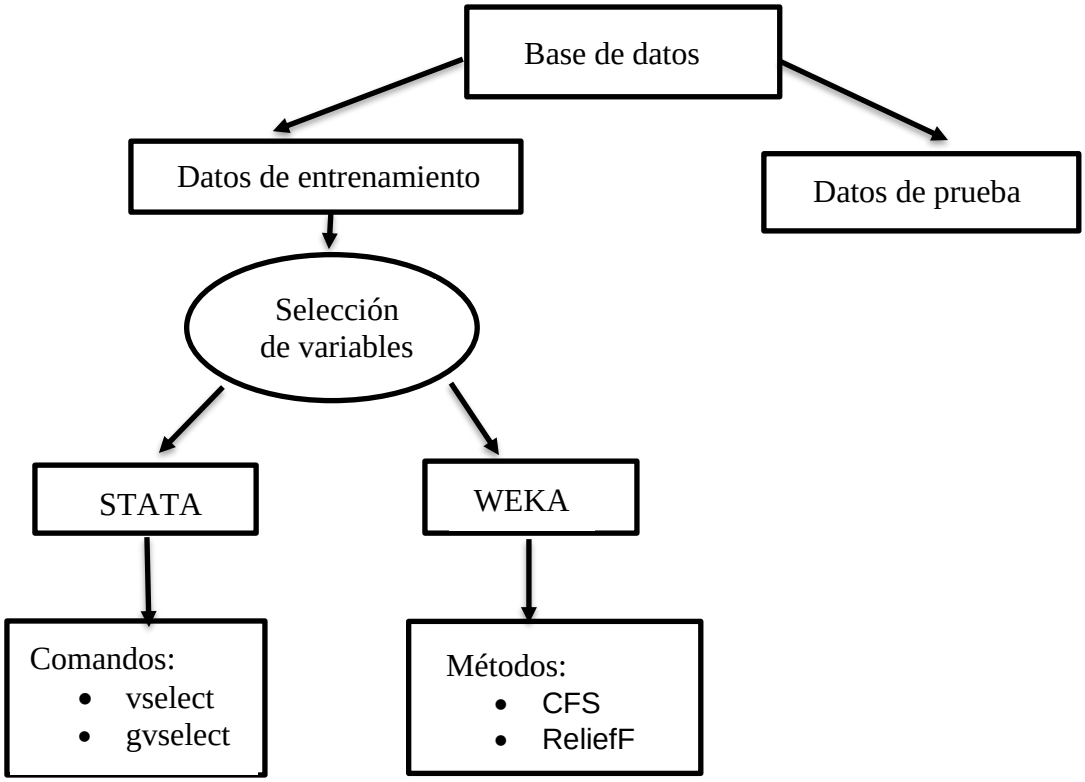

Figura 3: Diseño de la propuesta

Los métodos de selección de subconjuntos de atributos se aplican solo con los datos de entrenamiento mientras que los datos de prueba permanecen ocultos durante la fase de selección de variables. Los atributos seleccionados son los que se utilizan posteriormente en la evaluación de los datos de prueba [24].

Para la posterior medición y predicciones se utilizará el modelo de Regresión Lineal Múltiple. Las métricas elegidas para evaluar el clasificador fueron el R^2 y la RMSE.

Para realizar la regresión en Stata, se usa el comando *regress* y se realiza una regresión de la variable dependiente con respecto al resto de las variables de la base de datos. Se observa el

histograma de los errores, para ver si su distribución se aproxima a una normal. Para hacerlo con un test, se aplica el de Shapiro-Wilk y el de Breusch-Pagan. Con el comando *hettest* se comprueba si la varianza del error es constante. Si la probabilidad resultante es menor a 0,05 se rechaza la hipótesis nula de homocedasticidad, lo cual indica que los residuos son heterocedásticos, por lo tanto, se debe incluir *robust* en el comando de la regresión, si es mayor, no se debe incluir la opción robust.

Además, se analiza la Tolerancia (TOL=1/VIF) y el Factor de Inflación de la Varianza (VIF) usando la sentencia *estat vif.* Si los factores de inflación de la varianza exceden de 5, se demuestra que existen problemas de multicolinealidad. La eliminación de predictores redundantes debería resolver este problema.

Para la selección de atributos en Stata, se utilizó el comando *vselect,* elaborado por los investigadores Lindsey y Sheather. No es un comando oficial de Stata por lo que para su uso debe descargarse previamente.

Se utiliza el código *vselect*, se introduce posteriormente la variable dependiente y luego todas las variables independientes del modelo. Al final se le añade la opción, *forward aic*, *forward bic*, *backward aic o backward bic* dependiendo del criterio de selección que se quiera tomar, que en este caso fue AIC. Este comando ayuda a los usuarios a realizar selección de variables después de realizar una regresión lineal.

El método procede de la siguiente manera: el modelo inicial solo contiene el término de intersección. Luego, un procedimiento iterativo incluye secuencialmente las variables proporcionando la mayor mejora en el criterio de información seleccionado por el usuario. El algoritmo termina cuando no hay una variable adicional para incluir que mejora el criterio de información. En el segundo caso, la eliminación progresiva hacia atrás comienza con un modelo con el conjunto completo de las variables y la iteración funcionan hacia abajo, secuencialmente (una por paso) selecciona la variable a eliminar para obtener la mayor mejora en el criterio de información seleccionado por el usuario. Como en el caso de la selección

hacia adelante, el algoritmo de eliminación hacia atrás finaliza cuando no hay variable disponible para mejorar el criterio de información [8].

Para determinar los mejores subconjuntos se utiliza el algoritmo de Furnival-Wilson Leaps-And-Bounds, donde se especifican las variables de la regresión, y la mejor opción. Con el comando *vselect*, luego de introducir las variables, se le añade la opción *best* al final de la sentencia. Este comando genera los criterios de información para el mejor modelo según cada tamaño del subconjunto y devuelve las n mejores combinaciones de predictores para cada uno de los modelos así como los criterios de información para cada uno [13]. El usuario puede especificar el Cp de Mallows (Cp), el Criterio de Información de Akaike (AIC), el Criterio de Información Corregido de Akaike (AIC C), el Criterio de Información Bayesiano (BIC) o R-cuadrado ajustado (R^2 ajustado) como el criterio de información para la selección [14].

Otro comando elaborado por lo mismos investigadores y que también realiza la selección de variables en Stata es el comando *gvselect*[15]. A diferencia del comando *vselect* que da la opción de realizar la selección hacia adelante o hacia atrás, este comando muestra las n combinaciones obtenidas siguiendo los criterios de AIC y BIC, señalando el conjunto que optimiza cada criterio. Otra diferencia es que incluye una nueva columna en los resultados de la regresión, la de probabilidades logarítmicas(LL). Es un comando que tarda un poco más de tiempo en ejecutarse, por el cual no se pude aplicar a la base de datos puma32H pues la ejecución del proceso tardó más de cinco horas sin mostrar los resultados obtenidos.

En Weka, para realizar la regresión, en la pestaña *Classify*, se pulsa sobre el botón choose. Una vez pulsado se desplegará un árbol que permitirá seleccionar el clasificador deseado, en este caso sería LinearRegression.

Para realizar la selección de variables, la pestaña *Select Atributes* permite acceder al área de selección variables. El objetivo de estos métodos es identificar, mediante un conjunto de datos que poseen unos ciertos atributos, aquellos atributos que tienen más peso a la hora de determinar si los datos son de una clase u otra.

Se utilizó el algoritmo evaluador de subconjuntos de atributos CfsSubsetEval, clasificado como método de filtro. Se ejecutó en combinación con el método de búsqueda BestFirst, el cual busca en el espacio de los subconjuntos de atributos utilizando la estrategia greedy hillclimbing con backtracking. La dirección de la búsqueda realizada por BestFirst fue hacia adelante partiendo del conjunto vacío de atributos y posteriormente hacia atrás.

También se hizo uso del algoritmo ReliefFAtributeEval combinado con el método Ranker, seleccionando el mismo número de variables que automáticamente selecciona el CFS.

5. Resultados

Esta sección informa los resultados de las pruebas realizadas en las distintas bases de datos. El preprocesamiento de los datos se ha llevado a cabo usando las herramientas Stata y Weka, primeramente sin realizar la selección de variables y posteriormente aplicando la selección de variables.

5.1.Análisis en Stata

En este apartado se muestran los resultados obtenidos en Stata.

5.1.1 Sin realizar selección de variables en las bases de datos.

Se lleva a cabo Regresión Lineal Múltiple primeramente con los datos de entrenamiento y posteriormente con los datos de prueba en cada una de las bases de datos objeto de estudio, obteniéndose los resultados que se muestran en la tabla 2.

En la tabla 2 se evidencia que la base de datos bodyfat es la que presenta el mayor valor de R^2 donde el modelo explica aproximadamente el 99% de toda la variabilidad de los datos de respuesta en torno a su media, seguida por la base de datos housing. En algunos campos, se espera completamente que los valores del R^2 sean bajos. Este es el caso de la base de datos sensory que por la natualeza de los datos se obtiene un valor de R^2 muy bajo. Con respecto a la RMSE, se obtienen valores bajos en la mayoría de las bases de datos a excepción de la base

de datos cpu.small, donde la desviación estándar de la varianza inexplicada es de 10 aproximadamente. Los valores más bajos de RMSE indican un mejor ajuste del modelo.

Tabla 2: Resultados sin realizar la selección de variables en Stata.

Base de datos	**Instancias**			**Número de variables**	**Resultados en los datos de entrenamiento**		**Resultados en los datos de prueba**	
					R^2	RMSE	R^2	RMSE
	Totales	**Datos de entrenamiento**	**Datos de prueba**					
Abalone	4177	3341	836	8	0.5300	1.9456	0.6152	2.5802
Bodyfat	252	201	51	14	0.9764	1.2667	0.9881	1.2608
Cpu.small	8192	6553	1639	12	0.7208	9.7695	0.6956	10.033
Elevators	16599	13279	3320	18	0.8177	0.0028	0.8046	0.0032
Fried	40768	32614	8154	10	0.7244	2.6302	0.7171	2.6340
Housing	506	404	102	13	0.7465	5.0470	0.8810	2.1703
Puma32H	8192	6553	1639	32	0.2282	0.0268	0.2157	0.0264
Sensory	576	460	116	11	0.0274	0.7997	0.1701	0.7352

Existen bases de datos con un alto número de atributos por lo que se considera necesario realizar la selección de variables. Además, en algunas bases de datos hay presencia de multicolinealidad y al realizar la regresión existen variables que son no significativas (que tienen un p-valor superior a 0.05) y que deberían eliminarse. Una de las opciones más sencillas para reducir el problema en cuestión es eliminar características al azar. Pero si el número de variables se reduce sin cuidado y simplemente se eliminan los datos sin procesar, esto puede ser peligroso para la tarea de aprendizaje, debido a la pérdida de información que pudiera ser relevante. Es por ello que se utilizan métodos de selección de variables.

5.1.2. Aplicando selección de variables en las bases de datos.

En la investigación, en el caso del comando vselect, solo se realiza la selección hacia adelante y hacia atrás siguiendo el criterio de AIC, dejando fuera el criterio de BIC. Para la selección de los mejores subconjuntos con el propio comando, se estableció como criterio de selección, aquel subconjunto que optimice el valor de AIC, que en casi la totalidad de los casos aportaba el mejor valor también siguiendo los criterios de AIC C y Cp de Mallows y en algunos

optimizaba los cinco criterios (mejor valor para BIC y R^2 ajustado también). Con respecto al comando gvselect, para seguir con la línea anterior, se seleccionó el subconjunto con mejor valor para AIC. Cabe mencionar que se tomó como referencia de comparación al valor de AIC, dado que es una métrica común en todos los métodos aplicados y además incluye en la selección de variables, mayor cantidad de atributos que si se siguiera el criteriode BIC, aunque en muchas ocasiones la selección realizada es la misma en ambos criterios.

Se utiliza para evaluar la selección de atributos un método de regresión, por ello, los datos de entrenamiento se han particionado utilizando la técnica de validación cruzada con 5-folds.

Para seleccionar el mejor modelo para cada método, dado que en la mayoría de bases de datos se obtienen subconjuntos diferentes para cada fold, se estableció como criterio de selección, tomar las variables que aparecieran en al menos 3 de los 5 folds establecidos. En la base de datos sensory, excepto en la seleción hacia atrás, en los restantes métodos no se pudo aplicar este criterio debido a los resultados obtenidos, por lo que se tomaron las variables que aparecieran al menos dos veces.

Luego del análisis anterior, y aplicando los criterios de selección fijados, las variables seleccionadas en cada una de las bases de datos y para cada método aplicado se muestran en la tabla 3. Los resultados obtenidos en cada fold de cada base de datos, para cada método aplicado se muestran en los anexos A1, A2, A3, A4, A5, A6, A7 y A8.

Tabla 3: Modelos finales por métodos tras la selección de variables en Stata.

Base de datos	**Número de variables**	**Selección de variables**			
		Selección hacia adelante, AIC (vselect, forward aic)	Selección hacia atrás, AIC (vselect, backward aic)	Mejor subconjunto (vselect, best)	Mejor subconjunto (gvselect)
Abalone	8	shell_weight shucked_weight diameter sex whole_weight height viscera_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	shucked_weight whole_weight shell_weight sex height viscera_weight diameter	sex diameter height whole_weight shucked_weight viscera_weight shell_weight
Bodyfat	14	density abdomen	density abdomen	density ankle weight	density ankle weight
Cpu.small	12	freeswap runqsz fork rchar freemem wchar lread scall	lread scall fork rchar wchar runqsz freemem freeswap	runqsz freeswap fork freemem rchar lread wchar scall	lread scall fork rchar wchar runqsz freemem freeswap
Elevators	18	satime1 diffrollrate absroll satime4 p climbrate diffclb sgz diffsatime1 diffsatime3 diffdiffclb satime3	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime3 diffsatime1 diffsatime3 satime2 diffsatime2 sa	diffrollrate absroll p climbrate sgz diffclb satime1 diffsatime3 diffsatime1 diffdiffclb satime4 satime3	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime3 satime4 diffsatime1 diffsatime3
Fried	10	x4 x2 x1 x5 x3 x10	x1 x2 x3 x4 x5 x10	x4 x2 x1 x5 x3 x10	x1 x2 x3 x4 x5 x10

Housing	13	lstat ptratio dis rm nox chas rad crim b zn tax	crim zn indus chas nox rm dis rad ptratio b lstat tax	lstat dis rm ptratio nox rad crim chas zn b tax	crim zn chas nox rm dis rad ptratio b lstat tax
Puma32H	32	tau4 tau5 dm1 thetad5 db2 theta1 da5	theta1 thetad5 tau4 tau5 dm1 db2 da5	tau4 tau5 dm1 thetad5 theta1 db2 da5	Tiempo de procesamiento muy largo
Sensory	11	occasion rows judges	occasion halfplot	occasion rows judges trellis	occasion rows judges trellis

Al aplicar la selección de variables en Stata, la base de datos que logra una mayor reducción de los atributos es *puma32H*, seguida de *bodyfat*.

5.1.3. Comparación de los resultados sin hacer la selección de variables y aplicando la selección de variables en las bases de datos.

A partir del análisis realizado en los dos apartados anteriores, al aplicar el clasificador Regresión Lineal Múltiple se obtienen los resultados que se muestran en la tablas 4 y 5.

Para evaluar la eficacia de las diferentes técnicas de selección se utiliza el R^2 y la RMSE. En este caso se va a analizar la media y la desviación típica del conjunto de bases de datos para tener una visión general del cambio de los resultados aplicando diferentes técnicas de selección.

Tabla 4: Resultados sin selección de variables y con selección de variables en los datos de entrenamiento (Stata).

Datos de entrenamiento										
Base de datos	**Sin selección de variables**		**Selección hacia adelante (forward, aic)**		**Selección hacia atrás (backward, aic)**		**Mejores subconjuntos (vselect, best)**		**Mejores subconjuntos (gvselect)**	
	R^2	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Abalone	0.5300	1.9456	0.5299	1.9456	0.5299	1.9456	0.5299	1.9456	0.5299	1.9456
Bodyfat	0.9764	1.2667	0.9751	1.2618	0.9751	1.2618	0.9750	1.2656	0.9750	1.2656
Cpu.small	0.7208	9.7695	0.6956	0.0330	0.6956	10.0330	0.6956	10.0330	0.6956	10.0330
Elevators	0.8177	0.0028	0.8177	0.0028	0.8177	0.0028	0.8177	0.0028	0.8177	0.0028
Fried	0.7244	2.6302	0.7244	2.6301	0.7244	2.6301	0.7244	2.6301	0.7244	2.6301
Housing	0.7465	5.0470	0.7461	5.0383	0.7465	5.0409	0.7461	5.0383	0.7461	5.0383
Puma32H	0.2282	0.0268	0.2268	0.0268	0.2268	0.0268	0.2268	0.0268	0.2268	0.0268
Sensory	0.0274	0.7997	0.0182	0.7982	0.0099	0.8007	0.0224	0.7974	0.0224	0.7974
Promedio	0.5964	2.6860	0.5917	2.7171	0.5907	2.7177	0.5922	2.7175	0.5922	2.7175
Desviación típica	0.3188	3.2982	0.3199	3.3789	0.3220	3.3789	0.3188	3.3787	0.3188	3.3787

Tabla 5: Resultados sin selección de variables y con selección de variables en los datos de prueba (Stata).

Datos de test										
Base de datos	**Sin selección de variables**		**Selección hacia adelante (forward, aic)**		**Selección hacia atrás (backward, aic)**		**Mejores subconjuntos (vselect, best)**		**Mejores subconjuntos (gvselect)**	
	R^2	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Abalone	0.6152	2.5802	0.6145	2.5811	0.6145	2.5811	0.6145	2.5811	0.6145	2.5811
Bodyfat	0.9881	1.2608	0.9848	1.2348	0.9848	1.2348	0.9845	1.2586	0.9845	1.2586
Cpu.small	0.6956	10.033	0.6944	10.0410	0.6944	10.0410	0.6944	10.0410	0.6944	10.0410
Elevators	0.8046	0.0032	0.8040	0.0032	0.8040	0.0032	0.8040	0.0032	0.8040	0.0032
Fried	0.7171	2.6340	0.7171	2.6337	0.7171	2.6337	0.7171	2.6337	0.7171	2.6337
Housing	0.8810	2.1703	0.8426	2.4684	0.8431	2.4779	0.8426	2.4684	0.8426	2.4684
Puma32H	0.2157	0.0264	0.2026	0.0264	0.2026	0.0264	0.2026	0.0264	0.2026	0.0264
Sensory	0.1701	0.7352	0.0745	0.7555	0.0015	0.7813	0.0916	0.7785	0.0916	0.7785
Promedio	0.6359	2.4304	0.6168	2.4680	0.6078	2.4724	0.6189	2.4739	0.6189	2.4739
Desviación típica	0.2968	3.2479	0.3169	3.2488	0.3354	3.2468	0.3127	3.2457	0.3127	3.2457

En los datos de entrenamiento, sin aplicar ninguna técnica de selección de variables, el promedio de la RMSE es de 2.6860 y la desviación típica de 3.2982, con un R^2 promedio de 0.5964 y una desviación típica de 0.3188. Aplicando selección de variables, la media de la RMSE aumenta ligeramente al igual que la desviación típica y se reduce el promedio y la desviación típica del R^2. En algunos casos, el resultado tras la selección de variables coincide con los obtenidos sin hacer la selección de variables. Los datos de prueba siguen un comportamiento similar.

En las representaciones gráficas de las figuras 4 y 5 se pueden apreciar los resultados conseguidos para cada base de datos.

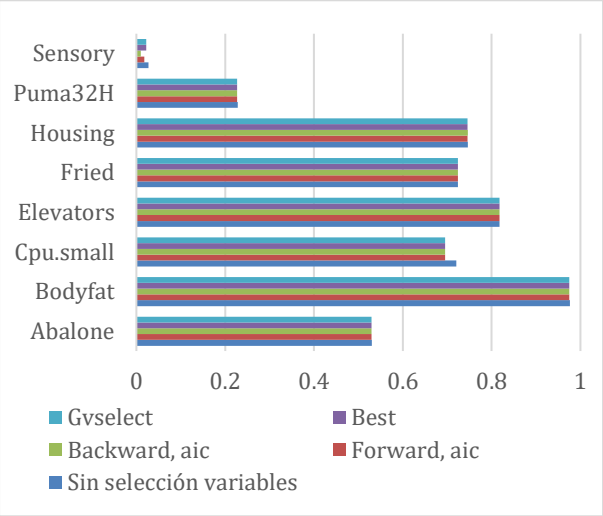

Figure 4: R^2. RLM en Stata.

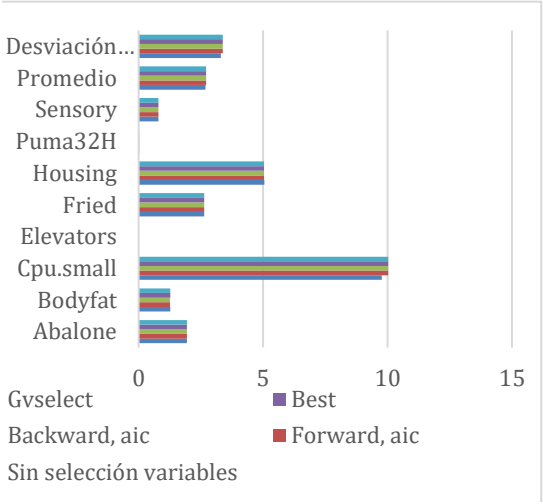

Figure 5: RMSE. RLM en Stata.

Como se puede observar, aplicando el clasificador Regresión Lineal Múltiple tras realizar la selección de variables se obtienen resultados muy similares que a los conseguidos aplicando el mismo clasificador pero sin la selección de variables. No obstante, una vez que se ha comprobado que algunos datos no son relevantes, se obtienen modelos más simplicados, y se necesita menos tiempo para su preprocesamiento, de ahí que se recomienda el uso de técnicas de selección de variables.

5.2.Análisis en Weka

Al igual que se hace, en su momento, en el apartado 5.1, en esta sección se estima el modelo de Regresión Lineal Múltiple sin realizar selección de variables y aplicando selección de variables en las ocho de bases de datos seleccionadas, pero esta vez usando la herramienta Weka.

5.2.1. Sin realizar selección de variables en las bases de datos.

Los resultados correspondientes al realizar la Regresión Lineal Múltiple en Weka para los datos de entrenamiento y aplicarlos a los datos de prueba se muestran en la tabla 6.

Tabla 6: Resultados sin realizar la selección de variables en Weka.

Base de datos	**Instancias**			**Número de atributos**	**Resultados en los datos de entrenamiento**		**Resultados en los datos de prueba**	
	Totales	**Datos de entrenamiento**	**Datos de prueba**		**R^2**	**RMSE**	**R^2**	**RMSE**
Abalone	4177	3341	836	8	0.7291	1.9396	0.7663	2.6582
Bodyfat	252	201	51	14	0.9879	1.2277	0.9934	1.1167
Cpu.small	8192	6553	1639	12	0.8489	9.7613	0.8338	10.0001
Elevators	16599	13279	3320	18	0.9043	0.0028	0.8969	0.0032
Fried	40768	32614	8154	10	0.8511	2.6298	0.8468	2.6326
Housing	506	404	102	13	0.8638	4.9629	0.9365	2.0601
Puma32H	8192	6553	1639	32	0.4763	0.0268	0.4558	0.0262
Sensory	576	460	116	11	0.4688	0.7085	0.6043	0.6176

Al igual que se demostró en Stata, queda evidenciado en Weka que la base de datos bodyfat es la que presenta el mayor valor de R^2 donde el modelo explica aproximadamente el 99% de toda la variabilidad de los datos de respuesta en torno a su media, seguida por la base de datos housing.

5.2.2. Aplicando la selección de variables en las bases de datos.

Se realiza la selección de variables en los datos de entrenamiento mediante la aplicación de dos métodos:

- CFS: se ejecutó en combinación con el método de búsqueda BestFirst. La dirección de la búsqueda realizada por BestFirst fue hacia adelante partiendo del conjunto vacío de atributos.

- ReliefF: Se aplicó combinado con el método Ranker, seleccionando el mismo número de variables que automáticamente selecciona el CFS.

Al realizar la selección de variables se obtuvieron los resultados que se muestran en la tabla 7 a continuación.

Al igual que en Stata, aplicando selección de variables en Weka, la base de datos que logra una mayor reducción de los atributos es *puma32H*, seguida de *bodyfat*.

Tabla 7: Resultados tras selección de variables en Weka.

Base de datos	**Instancias**		**Número de variables totales**	**Número de variables seleccionadas**	**CFS**	**ReliefF**
	Totales	Datos de entrenamiento				
Abalone	4177	3341	8	2	height shell_weight	shell_weight shucked_weight
Bodyfat	252	201	14	3	density age abdomen	density abdomen chest

Cpu.small	8192	6553	12	9	lread lwrite scall sread swrite exec rchar runqsz freeswap	fork freeswap scall rchar wchar exec lread freemen sread
Elevators	16599	13279	18	4	climbrate satime1 diffsatime2 diffsatime4	diffrollrate absroll curroll q
Fried	40768	32614	10	5	x1 x2 x4 x5 x10	x4 x2 x1 x3 x5
Housing	506	404	13	5	chas nox rm ptratio lstat	rm lstat nox dis ptratio
Puma32H	8192	6553	32	13	theta1 theta4 thetad3 thetad5 tau3 tau4 dm1 dm2 dm5 da3 da5 db2 db4	tau4 theta5 tau3 db3 theta4 theta6 da1 da3 tau1 dm2 thetad1 thetad5 dm5
Sensory	576	460	11	6	occasion judges interval position rows trellis	judges interval rows sittings occasion squares

5.2.3. Comparación de los resultados sin hacer la selección de variables y aplicando la selección de variables en las bases de datos.

A partir del análisis realizado en los dos apartados anteriores, se obtienen los resultados que se muestran en las tablas 8 y 9.

En la base de datos elevator, la aplicación de cualquiera de los dos métodos de selección de variables aplicados en Weka y la posterior aplicación del clasificador Regresión Lineal Múltiple, conlleva a una reducción considerable del R^2. A pesar de esto y del aumento de la RMSE al aplicar CFS o ReliefF, en algunos casos el cambio es poco significativo (base de datos bodyfat, fried, housing, puma32H), en otros la diferencia es más notable cuando se compara con los resultados obtenidos sin realizar selección de variables (bases de datos cpu.small, housing).Al realizar la selección de variables se obtienen modelos más simplicados, y se necesita menos tiempo para su preprocesamiento, de ahí que se recomienda su uso.

Tabla 8: Resultados sin selección de variables y con selección de variables en los datos de entrenamiento (Weka).

Datos de entrenamiento						
Base de datos	**Sin selección de variables**		**CFS**		**ReliefF**	
	R^2	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Abalone	0.7291	1.9396	0.6496	2.1547	0.6956	2.0362
Bodyfat	0.9879	1.2277	0.9875	1.2523	0.9875	1.2523
Cpu.small	0.8489	9.7613	0.8273	10.3774	0.7699	11.7862
Elevators	0.9043	0.0028	0.6575	0.0049	0.4128	0.0060
Fried	0.8511	2.6298	0.8510	2.6303	0.8511	2.6299
Housing	0.8638	4.9629	0.8268	5.5402	0.8438	5.2856
Puma32H	0.4763	0.0268	0.4759	0.0268	0.4743	0.0268
Sensory	0.4688	0.7085	0.4688	0.7085	0.4087	0.7320
Promedio	0.7663	2.6574	0.7181	2.8369	0.6805	2.9694
Desviación típica	0.1948	3.2973	0.1864	3.5382	0.2224	3.9548

Tabla 9: Resultados sin selección de variables y con selección de variables en los datos de prueba (Weka).

Datos de prueba						
Base de datos	**Sin selección de variables**		**CFS**		**ReliefF**	
	R^2	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Abalone	0.7663	2.6582	0.7498	2.7373	0.7261	2.8448
Bodyfat	0.9934	1.1167	0.9921	1.2196	0.9921	0.3716
Cpu.small	0.8338	10.0001	0.8148	10.5021	0.7551	11.8754
Elevators	0.8969	0.0032	0.6473	0.0056	0.3808	0.0068
Fried	0.8468	2.6326	0.8467	2.6332	0.8468	2.6326
Housing	0.9365	2.0601	0.9101	2.4349	0.9101	2.4349
Puma32H	0.4558	0.0262	0.4487	0.0263	0.4558	0.0262
Sensory	0.6043	0.6176	0.5241	0.6601	0.5447	0.6501
Promedio	0.7917	2.3893	0.7417	2.5274	0.7014	2.6053
Desviación típica	0.1797	3.2558	0.1888	3.4116	0.2205	3.9322

En las representaciones gráficas de las figuras 6 y 7 se pueden apreciar los resultados conseguidos para cada base de datos.

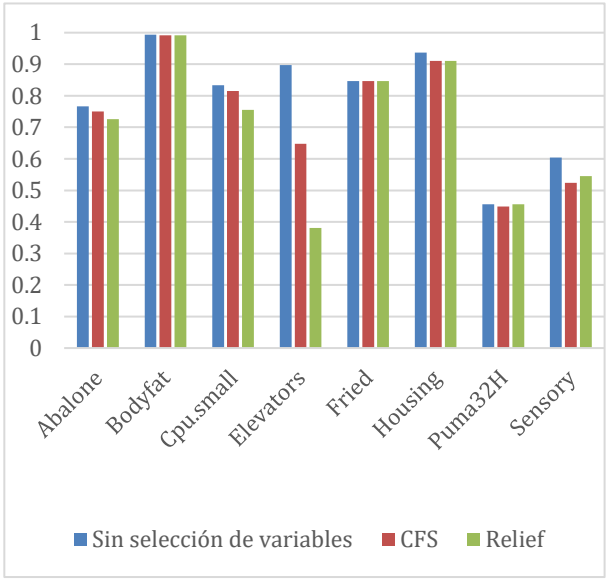

Figure 6: R^2. RLM en Weka

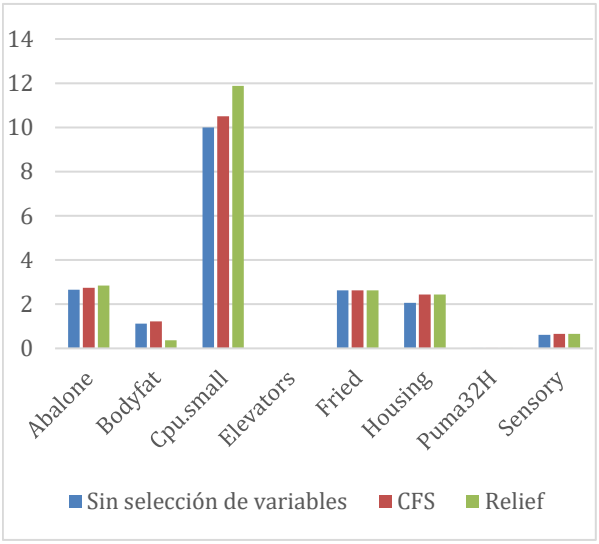

Figure 7: RMSE. RLM en Weka.

5.3. Comparación de los resultados obtenidos en Stata y Weka.

En la sección 5.1 y 5.2 se proponen diferentes métodos para realizar la selección de variables. Se utilizan los comandos *vselect* y *gvslect* de Stata, específicamente las opciones de selección hacia adelante (forward) y hacia atrás (backward) bajo el criterio de AIC y mejores subconjuntos. Otros métodos usados fueron el CFS y ReliefF en Weka.

Los análisis realizados en cada herramienta son diferentes y al comparar los resultados obtenidos aplicando una u otra cabe destacar que:

- Las variables seleccionadas, si bien difieren en cada herramienta, la mayoría son comunes en ambas.

- Se produce una reducción significativa de las variables en las bases de datos lo que demuestra la importancia de aplicar la selección de variables, pues se obtienen modelos más fáciles de interpretar y con un período de ejecución menor.

6. Conclusiones

En este trabajo se ha expuesto tanto la teoría como la práctica de la selección de variables y mejores subconjuntos. Se han utilizado ocho bases de datos correspondientes a problemas de regresión, extraídas del repositorio de la Universidad de California en Irvine. Se ha ilustrado diferentes formas de realizar selección de variables: comandos vselect y gvselect de Stata y métodos CFS y ReliefF que proporciona Weka.

La principal contribución de este trabajo ha sido mostrar el éxito del proceso de selección de variables en las bases de datos seleccionadas ya que se obtuvieron subconjuntos de atributos que aportan la misma información que el conjunto original, dejando por el camino, a los atributos irrelevantes o redundantes que pudieran existir en los datos originales.

A modo de resumen, ha quedado evidenciado que a veces no existe un modelo único que optimice todos los criterios. En este caso lo que se puede hacer es reducir las opciones a los pocos modelos que estén cerca de la optimización y establecer alguna regla de decisión o realizar una selección arbitraria de alguno de ellos.

A partir de los resultados de la presente investigación, se abren una serie de oportunidades de desarrollo que permitirán explotar y ampliar los resultados obtenidos.

Bibliografía

[1] Abuín, J. M. (2007). *Regresión lineal múltiple.* Madrid: Instituto de Economía y Geografía.

[2] Allasia, M. B., Branco, M. D., & Quaglino, M. B. (noviembre de 2016). *Regresión Lasso Bayesiana. Ajuste de modelos lineales penalizados mediante la asignación priores normales con mezcla de escala.* Instituto de Investigaciones Teóricas y Aplicadas en Estadística. Vigesimoprimeras Jornadas Investigaciones en la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario.

[3] Ansari, G., Doja, T., & Ahmad, M. (noviembre de 2019). Hybrid Filter–Wrapper Feature Selection Method for Sentiment Classification. *Arabian Journal for Science and Engineering*, vol. 44, no. 11, págs. 9191–9208.

[4] Brown, G. (2009). *A new perspective for information theoretic feature selection. Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR vol. 5* págs.*49-56.*

[5] Brown, G., Pocock, A., Zhao, M.-J., & Lujan, M. (2012). Conditional Likelihood Maximisation: A Unifying Framework for. *Journal of Machine Learning Research*, vol. 13, págs.27-66.

[6] Congregado, E. (2020). *El modelo clásico de regresión lineal. Apuntes de la asignatura Minería de Datos I. Máster en Economía, Finanzas y Computación UNIA-UHU.*

[7] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA.*

[8] Gluzmann, P., & Panigo, D. (1 de junio de 2015). Global Search Regression: A New Automatic Model-selection Technique for Cross-section, Time-series, and Panel-data Regressions. *The Stata Journal, vol.15, No 2,* págs.. *325–349.*

[9] Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, Vol. 76, No.2, págs. 297-307.*

[10] Kira, K., & A.Rendell, L. (1992). A practical approach to feature selection. *Machine Learning Proceedings 1992, págs. 249-256.*

[11] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence Journal, vol. 97, págs. 273-324.*

[12] Ladha, L., & Deepa, T. (mayo de 2011). Feature Selection Methods And Algorithms. *International Journal on Computer Science and Engineering*. *ISSN : 0975-3397, vol. 3, No. 5.*

[13] Lindsey, C. (2009). A Modern Approach to Regression with R. *Stata Premier*.

[14] Lindsey, C., & Sheather, S. (2010). Variable selection in linear regression. *The Stata Journal, vol.10 No.4, págs. 650-669.*

[15] Lindsey, C., & Sheather, S. (2015). Best subsets variable selection in nonnormal regression models. *The Stata Journal, vol. 15, No.4, págs. 1046–1059.*

[16] Mallows, C. L. (noviembre de 1973). Some Comments on Cp. *Technometrics , vol 15, págs. 661-675.*

[17] Mello, R. F., & Ponti, M. A. (2018). *Machine Learning.* Cham: Springer International Publishing.

[18] Mello, R. F., Ponti, M. A., Mello, R. F., & Antonelli, M. (2018). A Brief Review on Machine Learning," in Machine Learning. *Springer International Publishing*, págs. 1–74.

[19] Olive, D. J. (2017). Multiple linear regression. In Linear regression . *Springer, Cham,* págs. 17-83.

[20] Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research, vol. 156, págs.483–494.*

[21] Rodrigo, J. A. (2016). *cienciadedatos.net.* Obtenido de Introducción a la Regresión Lineal Múltiple.

[22] Rodríguez, E. M. (2005). Errores frecuentes en la interpretación del coeficiente de determinación lineal. *Anuario jurídico y económico escurialense*, XXXVIII *págs*.315-332/ISNN:1133-3677.

[23] Sandoval, L. J. (2018). Algoritmos de aprendizaje automático para el análisis y predicción de datos. *Revista Tecnológica*, No.11.

[24] Tallón-Ballesteros, A. J., Correia, L., & Leal-Díaz, R. (2021). *Attribute Subset Selection for Image Recognition. Random Forest Under Assessment.* 16th International Conference on Soft Computing Models in Industrial and Environmental Applications, *págs*.821-827.

[25] Tallón-Ballesteros, A. J., Riquelme, J. C., & Ruiz, R. (2019). *Semi-wrapper feature subset selector for feed-forward neural networks: Applications to binary and multi-class classification.*

[26] Wang, Y., & Liu, Q. (2006). Comparison of Akaike information criterion (AIC). *Fisheries Research, vol. 77, págs. 220-225*.

[27] Young, D. S. (2018). Manual de métodos de regresión. *Prensa CRC*.

[28] Yu, L., & Liu, H. (2003). *Feature Selection for High-Dimensional Data: A Fast Correlation Based Filter Solution.*

Anexos

Anexo A: Resultados detallados de selección de atributos.

Anexo A1: Selección de variables en la base de datos abalone

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	shell_weight shucked_weight diameter sex whole_weight height viscera_weight	shell_weight hucked_weight height sex diameter whole_weight viscera_weight	shell_weight shucked_weight diameter sex whole_weight viscera_weight height	shell_weight shucked_weight diameter sex height whole_weight viscera_weight	shell_weight shucked_weight diameter sex whole_weiht viscera_weight height
Selección hacia atrás, AIC (vselect, backward aic)	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight
Mejor subconjunto (vselect, best)	shucked_weight whole_weight shell_weight sex height viscera_weight diameter	shucked_weight shell_weight height whole_weight sex viscera_weight diameter	shucked_weight shell_weight whole_weight sex viscera_weight height diameter	shucked_weight shell_weight whole_weight sex viscera_weight height diameter	shucked_weight shell_weight whole_weight sex viscera_weight height diameter
Mejor subconjunto (gvselect)	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight	sex diameter height whole_weight shucked_weight viscera_weight shell_weight

Anexo A2: Selección de variables en la base de datos bodyfat

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	density abdomen ankle weight thingh	density height	density abdomen	density chest	density abdomen age
Selección hacia atrás, AIC (vselect, backward aic)	density abdomen	density	density abdomen	density chest	density abdomen
Mejor subconjunto (vselect, best)	density ankle weight abdomen thigh	density height thigh weight chest	density abdomen	density ankle biceps age weight	density ankle age biceps weight
Mejor subconjunto (gvselect)	density weight abdomen thigh ankle	density weight height chest thigh	density abdomen	density age weight ankle biceps	density age weight ankle biceps

Anexo A3: Selección de variables en la base de datos cpu.small

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	freeswap runqsz fork rchar freemem wchar lread scall	freeswap runqsz fork freemem wchar lread lwrite scall	freeswap runqsz fork wchar freemem rchar lread swrite	freeswap runqsz fork rchar freemem whar lread scall lwrite swrite	freeswap runqsz fork rchar freemem whar lread scall
Selección hacia atrás, AIC (vselect, backward aic)	lread scall fork rchar wchar runqsz freemem freeswap	lread lwrite scall fork rchar wchar runqsz freemem freeswap	lread swrite fork rchar wchar runqsz freemem freeswap	lread lwrite scall swrite fork rchar wchar runqsz freemem freeswap	lread scall fork rchar wchar runqsz freemem freeswap
Mejor subconjunto (vselect, best)	runqsz freeswap fork freemem rchar lread wchar scall	runqsz freeswap fork freemem rchar wchar lread lwrite scall	runqsz freeswap fork freemem wchar lread rchar swrite	runqsz freeswap fork freemem rchar lread wchar scall lwrite swrite	freeswap runqsz fork freemem wchar lread rchar scall
Mejor subconjunto (gvselect)	lread scall fork rchar wchar runqsz freemem freeswap	lread lwrite scall fork rchar wchar runqsz freemem freeswap	lread swrite fork rchar wchar runqsz freemem freeswap	lread lwrite scall swrite fork rchar wchar runqsz freemem freeswap	lread scall fork rchar wchar runqsz freemem freeswap

Anexo A4: Selección de variables en la base de datos elevator

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	satime1 diffrollrate absroll satime4 p climbrate diffclb sgz diffsatime1 diffdiffclb diffsatime3 satime3 curroll	satime1 diffrollrate absroll satime4 p climbrate diffclb sgz diffsatime1 diffsatime3 diffdiffclb satime3	satime1 diffrollrate absroll satime4 p climbrate diffclb diffdiffclb diffsatime1 sgz diffsatime3 satime2	satime1 diffrollrate absroll satime4 p climbrate diffclb diffsatime1 diffdiffclb diffsatime3 satime3 sgz	satime1 diffrollrate absroll satime4 p climbrate diffclb sgz q diffsatime1 diffsatime3 curroll satime3 diffdiffclb
Selección hacia atrás, AIC (vselect, backward aic)	climbrate sgz p curroll absroll diffclb diffrollrate diffdiffclb satime1 satime2 satime3 diffsatime1 diffsatime2 diffsatime3 sa	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime3 satime4 diffsatime1 diffsatime3	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime2 satime3 satime4 diffsatime1 diffsatime2 diffsatime3 diffsatime4 sa	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime3 diffsatime1 diffsatime3 sa	climbrate sgz p curroll absroll diffclb diffrollrate diffdiffclb satime1 satime2 satime3 diffsatime1 diffsatime2 diffsatime3 sa
Mejor subconjunto (vselect, best)	diffrollrate absroll p climbrate sgz diffclb satime1 diffsatime3 diffsatime1 curroll	diffrollrate absroll p climbrate diffclb sgz satime1 diffsatime3 diffsatime1 diffdiffclb	diffrollrate absroll p climbrate sa diffclb satime1 diffsatime1 sgz diffsatime3	diffrollrate absroll p climbrate diffclb satime1 diffsatime3 diffsatime1 sgz diffdiffclb	diffrollrate absroll p climbrate satime1 diffclb sgz diffsatime3 diffsatime1 curroll

	diffdiffclb satime4 satime3	satime4 satime3	diffdiffclb satime2	sa diffsatime2 satime2	diffdiffclb satime4 satime3
Mejor subconjunto (gvselect)	climbrate sgz p curroll absroll diffclb diffrollrate diffdiffclb satime1 satime3 satime4 diffsatime1 diffsatime3	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime3 diffsatime1 diffsatime3 sa	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime2 satime4 diffsatime1 diffsatime3	climbrate sgz p absroll diffclb diffrollrate diffdiffclb satime1 satime2 diffsatime1 diffsatime3 sa	climbrate sgz p curroll absroll diffclb diffrollrate diffdiffclb satime1 satime3 satime4 diffsatime1 diffsatime3

Anexo A5: Selección de variables en la base de datos fried

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	x4 x2 x1 x5 x3	x4 x2 x1 x5 x3	x4 x2 x1 x5 x3 x10	x4 x1 x2 x5 x3 x10	x4 x2 x1 x5 x3 x10
Selección hacia atrás, AIC (vselect, backward aic)	x1 x2 x3 x4 x5	x1 x2 x3 x4 x5	x1 x2 x3 x4 x5 x10	x1 x2 x3 x4 x5 x10	x1 x2 x3 x4 x5 x10
Mejor subconjunto (vselect, best)	x4 x2 x1 x5 x3	x4 x1 x2 x5 x3	x4 x2 x1 x5 x3 x10	x4 x1 x2 x5 x3 x10	x4 x2 x1 x5 x3 x10
Mejor subconjunto (gvselect)	x1 x2 x3 x4 x5	x1 x2 x3 x4 x5	x1 x2 x3 x4 x5 x10	x1 x2 x3 x4 x5 x10	x1 x2 x3 x4 x5 x10

Anexo A6: Selección de variables en la base de datos housing

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	lstat ptratio dis rm nox chas rad crim indus b zn	lstat ptratio chas dis nox rm zn b rad crim tax age	lstat ptratio rm b dis nox chas zn crim rad tax	rm tax ptratio b lstat dis age nox indus rad crim	lstat rm ptratio dis nox zn chas crim rad tax
Selección hacia atrás, AIC (vselect, backward aic)	crim zn indus chas nox rm dis rad ptratio b lstat	crim zn indus chas nox rm age dis rad ptratio b lstat	crim zn chas nox rm dis rad tax ptratio b lstat	crim indus nox rm age dis rad tax ptratio b lstat	crim zn chas nox rm dis rad tax ptratio lstat
Mejor subconjunto (vselect, best)	lstat dis rm ptratio nox rad crim chas zn b indus	lstat ptratio nox dis chas rad crim zn b tax rm age	lstat dis rm nox ptratio rad zn tax b crim chas	rm ptratio dis b age nox tax indus rad lstat crim	lstat rm dis ptratio rad nox crim zn tax chas
Mejor subconjunto (gvselect)	crim zn indus chas	crim zn chas nox	crim zn chas nox	crim indus nox rm	crim zn chas nox

	nox rm dis rad ptratio b lstat	rm age dis rad tax ptratio b lstat	rm dis rad tax ptratio b lstat	age dis rad tax ptratio b lstat	rm dis rad tax ptratio lstat

Anexo A7: Selección de variables en la base de datos puma32H

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	tau4 tau5 dm1 thetad5 tau1 theta1 db2	tau4 da5 dm1 thetad5 tau5 theta1 db2	tau4 tau5 dm2 da5 db3 dm1	tau4 da5 theta1 dm1 tau5	tau4 da5 theta6 db2 dm1 thetad5 thetad3
Selección hacia atrás, AIC (vselect, backward aic)	theta1 thetad5 tau1 tau4 tau5 dm1 db2	theta1 thetad5 tau4 tau5 dm1 da5 db2	tau4 tau5 dm1 dm2 da5 db3	theta1 tau4 tau5 dm1 da5	theta6 thetad3 thetad5 tau4 dm1 da5 db2
Mejor subconjunto (vselect, best)	tau4 tau5 dm1 thetad5 tau1 theta1 db2	tau4 da5 dm1 thetad5 theta1 tau5 db2	tau4 dm2 db3 tau5 da5 dm1	tau4 da5 theta1 dm1 tau5	tau4 da5 theta6 db2 dm1 thetad5 thetad3
Mejor subconjunto (gvselect)	Tiempo de procesamiento muy largo	Tiempo de procesamiento muy largo	Tiempo de procesamiento muy largo	Tiempo de procesamiento muy largo	Tiempo de procesamiento muy largo

Anexo A8: Selección de variables en la base de datos sensory

Método	**Train**				
	F1	**F2**	**F3**	**F4**	**F5**
Selección hacia adelante, AIC (vselect, forward aic)	occasion interval trellis	-	occasion	rows judges	judges rows
Selección hacia atrás, AIC (vselect, backward aic)	occasion interval halfplot trellis	occasion halfplot	occasion halfplot	occasion judges rows halfplot	occasion judges rows halfplot
Mejor subconjunto (vselect, best)	occasion interval trellis	trellis	occasion	rows judges	judges rows
Mejor subconjunto (gvselect)	occasion interval trellis	trellis	occasion	rows judges	judges rows