Predicción y selección de variables con bosques aleatorios en presencia de variables correlacionadas

Tesis para optar al título de:
Maestría en Ciencias - Estadística

Presentada por:
Néstor Iván Cardona Alzate
neicardonaal@unal.edu.co

Asesor:
Juan David Ospina Arango, PhD.
jdospina@unal.edu.co
Co-asesor:
Juan Carlos Correa Morales, Ph.D.
jccorrea@unal.edu.co

UNIVERSIDAD NACIONAL DE COLOMBIA
Sede Medellín
Escuela de Estadística
Facultad de Ciencias

Septiembre de 2019

Contenidos

Nomenclatura

ABREVIATURAS

AID	Automatic Interaction Detection.
Bagging	Bootstrap Aggregation.
CART	Classification and Regression Trees.
ID3	Induction of Decision Trees.
LASSO	Least Absolute Shrinkage and Selection Operator.
MSE	Mean Squared Error.
NRFE	Non Recursive Feature Elimination.
OOB	Out-Of-Bag.
RFE	Recursive Feature Elimination.
RF	Random Forests.
RSS	Residual Sum of Squares.
THAID	Theta Automatic Interaction Detection.

SÍMBOLOS

ϵ	Error del modelo.
$\hat{y}_i^{(t)}$	Respuesta o clase predicha por el árbol t al emplear la observación de prueba i, siendo esta última OOB al construirse dicho árbol.
π_1, π_2	Proporción de observaciones en los nodos generados por una partición.
ρ_{ij}	Correlación entre las variables predictoras X_i y X_j.

σ^2	Varianza de una variable aleatoria.
τ_j	Correlación entre la variable predictora X_j y la variable respuesta Y.
I	Impureza de las observaciones al realizar una partición.
I_1, I_2	Impureza de las observaciones en los nodos generados por una partición.
I_a	Impureza de las observaciones antes de realizarce una partición.
k	Número de observaciones presentes en un nodo terminal.
$MSE_{OOB}^{(t)}$	MSE promedio generado por el árbol t al emplear como datos de prueba las observaciones OOB que resultaron de la construcción de dicho árbol.
$mtry$	Número de variables predictoras preseleccionadas aleatoriamente como candidatas para hacer una partición.
n	Número de observaciones disponibles para entrenar (o ajustar) un método.
n_1, n_2	Número de observaciones en los nodos generados por una partición.
$n_{OOB}^{(t)}$	Número de observaciones OOB resultantes en la construcción del árbol t.
$nodesize$	Número de observaciones que contiene cada nodo terminal de un árbol.
$ntree$	Número de árboles que conforman el bosque aleatorio.
p	Número de variables predictoras en el conjunto de datos.
q	Número de variables predictoras en un grupo.
VI	Importancia de una variable.
X	Variable predictora.
Y	Variable respuesta.

Lista de tablas

Lista de figuras

List of Algorithms

Resumen

El presente trabajo aborda el problema de selección de variables empleando el método de bosques aleatorios cuando el modelo subyacente para la variable respuesta es de tipo lineal. Para ello se configuran conjuntos de datos simulados con diferentes características, sobre los cuales se aplica la metodología y se mide el error de predicción al eliminar cada variable. Con esto se realiza en primera instancia, una evaluación del algoritmo de selección en la que se identifica que este es eficiente cuando los conjuntos de datos contienen grupos de variables predictoras con tamaño inferior a 8 y en segunda instancia, una evaluación del método de bosques aleatorios en la que se identifica que el número total de variables predictoras es el factor que mas fuertemente impacta su desempeño.

Abstract

This thesis addresses the problem of variable selection using the random forest method when the underlying model for the response variable is linear. To this end, simulated data sets with different characteristics are configured and then, the methodology applied, and the prediction error measured each time a variable is eliminated. This is done to evaluate the selection algorithm, which leads to identifying that it is efficient when data sets contain groups of predictor variables with a size less than 8. Also, this is done to evaluate the random forest method, which leads to identifying that the total number of predictor variables is the factor that most strongly impacts its performance.

Capítulo 1

Introducción

Gracias a los avances tecnológicos disponibles para capturar información, es común encontrar hoy en día conjuntos de datos caracterizados por un número creciente de variables predictoras, lo cual favorece en ciertas ocasiones el desarrollo de modelos mas completos y precisos. Sin embargo, esta adición de variables a los modelos con frecuencia tiene efectos desfavorables que se ven reflejados tanto en la perdida de precisión de los pronósticos como en la perdida de interpretabilidad del modelo.

Estos efectos adversos han hecho que la selección de variables se convierta en una necesidad tangible, especialmente en el mundo aplicado, para abordar diferentes problemas de regresión y clasificación, ya que por medio de esta es posible incorporar a los modelos solo las variables predictoras que cuentan con niveles importantes de relevancia para predecir la respuesta.

El problema de selección de variables puede ser abordado de diferentes formas, una de ellas es por medio de bosques aleatorios debido a que es una metodología versátil que proporciona a la vez estimación para la variable respuesta y medidas de importancia para las variables predictoras, siendo esto último un elemento clave que permite abordar la selección de variables.

Ya que bosques aleatorios cuenta con estas características y además es un método no paramétrico alternativo a los tradicionales, resulta interesante poder evaluar su desempeño para predecir y seleccionar variables. Es por esto que acá esta metodología será puesta a prueba bajo diversos escenario de simulaciones en los que se controlará la cantidad de variables predictoras, las correlación existentes entre grupos de variables predictoras y la cantidad de observaciones.

Capítulo 2

Árboles, bosques y medidas de importancia

2.1. Árboles de clasificación y regresión

Los árboles de decisión para clasificación y regresión son una metodología no paramétrica de la cual hacen parte múltiples propuestas que han evolucionado a partir del trabajo de Morgan y Sonquist (1963), quienes desarrollaron el algoritmo *Automatic Interaction Detection (AID)* para abordar problemas de regresión. Así mismo, dentro de ésta etapa inicial de propuestas es importante mencionar el trabajo de Messenger y Mandell (1972), quienes extendieron la idea a los problemas de clasificación por medio del algoritmo *Theta Automatic Interaction Detection (THAID)*.
Posteriormente, en la década de los ochentas, surgieron varias propuestas que enriquecieron el concepto de árboles. Entre dichos planteamientos, los que tuvieron mayor acogida y comenzaron a despertar interés por el tema dentro de la comunidad estadística fueron *Classification and Regression Trees (CART)* (Breiman, Friedman, Olshen, y Stone, 1984) e *Induction of Decision Trees (ID3)* (Quinlan, 1986) – luego conocido como *C4.5* (Quinlan, 1993).
En el presente trabajo todos los conceptos expuestos e implementaciones realizadas estarán basadas en el algoritmo *CART*. Sin embargo, para una rápida revisión conceptual de otras propuestas relacionadas puede referirse a Loh (2011, 2014).

2.1.1. Descripción, usos y ventajas

Los árboles de clasificación y regresión particionan recursivamente el espacio conformado por las p variables predictoras X (también conocidas como variables independientes, variables explicativas o covariables), de forma tal que las regiones

generadas por dichas particiones contienen observaciones cuyos valores respecto a la variable respuesta Y son lo más homogéneos posibles. Bajo el algoritmo *CART* todas las particiones realizadas son binarias, lo cual produce subconjuntos disjuntos p-dimensionales con forma *"rectangular"*.
Al igual que los métodos paramétricos clásicos de regresión lineal y regresión logística, los árboles pueden ser empleados para realizar las mismas tareas de predicción y clasificación. Además, tal como lo afirman James, Witten, Hastie, y Tibshirani (2013): "El desempeño de los árboles puede ser superior al de los métodos clásicos cuando la verdadera forma funcional que relaciona las variables predictoras con la variable respuesta es altamente no lineal y compleja".
Esto último se debe a que las técnicas de regresión combinan la información de las diferentes variables predictoras de forma lineal, mientras que los árboles lo hacen de una forma no especificada de antemano y determinada por lo datos de entrenamiento que en particular se estén considerando para el ajuste (*data driven recursive partitioning*). De este modo los árboles logran capturar con bajo sesgo las relaciones existentes entre las predictoras.
Otra situación bajo la cual el desempeño de los árboles puede ser superior al de su contraparte paramétrica clásica es en el manejo de los efectos de interacción entre las variables predictoras. Ya que como lo señala Strobl, Malley, y Tutz (2009): "Cuando se presentan interacciones de orden alto o cuando la combinación de todos los efectos principales y de interacción es un numero alto, las metodologías clásicas pueden tener limitaciones para su aplicación".
El uso de los árboles también comprende escenarios donde la cantidad de variables predictoras supera la cantidad de observaciones $(n < p)$, la cual es una situación que se presenta cada vez con mayor frecuencia en diferentes áreas (e.g. biología, medicina) y que no puede ser abordada por los métodos tradicionales de regresión, ya que bajo estos escenarios no es posible obtener estimaciones para los parámetros del modelo.
Entre las ventajas que les han dado popularidad y acogida a los árboles de clasificación y regresión pueden mencionarse las siguientes: fácil interpretación y explicación de los resultados generados, posibilidad de representar gráficamente un problema multidimensional, procesamiento de variables cualitativas y cuantitativas sin necesidad de crear variables indicadoras (*dummy*) y manejo implícito de datos faltantes haciendo uso de variables sustitutas (*surrogate splits* en el algoritmo *CART*).

2.1.2. Construcción

Los árboles se construyen realizando recursivamente particiones binarias del espacio conformado por las p variables predictoras. Para ello, inicialmente se selecciona una de las variables y sobre ésta se determina un punto de corte óptimo

que segmente en dos su rango. Luego, sobre cada una de las dos particiones generadas (también conocidas como nodos), se aplica iterativamente el mismo procedimiento hasta que el número de observaciones contenidas en una partición cumpla algún criterio preestablecido, usualmente denominado criterio de parada. El proceso de optimización y los criterios empleados para la construcción de un árbol varían dependiendo del tipo de variable respuesta que se esté considerando. Estos son descritos con mayor detalle por Hastie, Tibshirani, y Friedman (2009). Así, para el caso de árboles de regresión cuya respuesta es numérica, la variable predictora y el punto de corte se seleccionan de forma tal que se minimice el *Residual Sum of Squares (RSS)* calculado sobre las observaciones contenidas en cada partición.

Para el caso de árboles de clasificación donde la respuesta es categórica, la variable predictora y el punto de corte se establecen minimizando la reducción de impureza, siendo esta última entendida como la fracción, tasa o proporción de observaciones que no pertenecen a la clase más común dentro de cada partición. Para el cálculo de esta impureza debe establecerse la diferencia entre la impureza de las observaciones antes de realizar la partición (I_a) y el promedio de las impurezas de las observaciones contenidas en cada uno de los dos nodos (I_1, I_2) que se generan en la partición, tal como se muestra en la ecuación (2.1),

$$I = I_a - (\pi_1 I_1 + \pi_2 I_2) \tag{2.1}$$

Esta impureza I queda definida entonces en función de la proporción de observaciones contenidas en cada partición $(\pi_1 = \frac{n_1}{n}, \pi_2 = \frac{n_2}{n})$ y las impurezas I_a, I_1, I_2, donde estas últimas pueden ser medidas de diferentes formas, bien sea haciendo uso del error de clasificación, el índice Gini o la devianza. Sin embargo, también pueden ser consideradas otras medidas de impureza tal como lo presentan Rokach y Maimon (2005).

Una interpretación de la reducción de la impureza que plantean Strobl, Malley, y Tutz (2009) es que ésta puede pensarse como *"uno de los muchos medios posibles para medir la fortaleza de la asociación entre la variable que particiona y la variable respuesta"*, luego la variable que está más fuertemente asociada con la respuesta es la variable que se selecciona para hacer la partición.

Respecto al criterio de parada del algoritmo se debe tener en cuenta que éste es un parámetro que define hasta qué punto se realizan las particiones recursivas y por ende determina la cantidad de observaciones que contendrá un nodo terminal (k). Este criterio puede establecerse de diferentes formas, para el caso de regresión es usual fijar un número mínimo de observaciones contenidas en los nodos terminales y para el caso de clasificación establecer que todas las observaciones de los nodos terminales pertenezcan a una misma clase o que la reducción de impureza no sobrepase un valor fijado.

Una vez el algoritmo para, los nodos terminales del árbol quedan conformados y sobre cada uno de ellos se establece un valor de predicción o clasificación (según sea el caso) para la variable respuesta. En un árbol de regresión la predicción se genera al promediar las observaciones contenidas en una partición, mientras que en un árbol de clasificación la predicción se establece como la clase más frecuente entre las observaciones contenidas en la partición. Así, al consolidar los valores de predicción generados por cada uno de los nodos terminales del árbol, se logra formar una función de predicción irregular que es constante por regiones.

2.1.3. Desventajas

El problema más notorio que presentan los árboles de clasificación y regresión es la alta varianza que puede tener la respuesta ante cambios pequeños en los datos de entrenamiento (Xiang, Minghui, y Heping, 2011), por lo cual el método tiende a exhibir un pobre desempeño bajo datos de prueba. Esto suele afectar la interpretabilidad de los resultados, ya que una alta variabilidad no permite realizar análisis consistentes del fenómeno subyacente.
Para tratar de mitigar esta inestabilidad se emplea el concepto de poda (*pruning*), el cual consiste en generar primero un gran árbol y luego reducir su tamaño (profundidad del árbol) al eliminar nodos internos. Este procedimiento llamado *cost-complexity pruning* es presentado con mayor detallen en Gey y Nedelec (2005).
También es problemático el hecho de que el proceso de generación de particiones de los árboles *CART* presenta sesgos en favor de variables predictoras que contengan mayor número de categorías y/o mayor cantidad de valores faltantes (Kim y Loh, 2001).
Por último, los árboles no son el método más apropiado para abordar problemas donde la estructura que relaciona las variables predictoras con la variable respuesta es aditiva, esto debido a que por medio de particiones binarias es difícil capturar este tipo de comportamiento (James et al., 2013). Lo anterior sugiere la necesidad de valorar cuidadosamente qué tan apropiado o inapropiado puede llegar a ser el uso de la metodología en el problema que se esté considerando.

2.2. Bosques aleatorios

Bosques aleatorios o *Random Forests* (Breiman, 2001) es una generalización de la metodología *Bagging* (Breiman, 1996), la cual construye múltiples árboles basados en el algoritmo *CART* y posteriormente combina las predicciones o clasificaciones (según sea el caso) entregadas por cada árbol. Con esto lo que se pretende es reducir la alta varianza que tiene la respuesta de un árbol individual

ante datos de prueba y por tanto mejorar el desempeño del método.
Bosques aleatorios, al igual que *Bootstrap Aggregation (Bagging)*, hacen parte de un grupo de métodos conocidos como métodos de ensamble, los cuales según Zhou (2012) "son atractivos principalmente porque son capaces de impulsar métodos débiles y convertirlos en métodos fuertes, con los cuales pueden hacerse predicciones muy precisas", siendo esto último justamente lo que se requiere mejorar o potenciar en los árboles.

2.2.1. Descripción, usos y ventajas

Los bosques aleatorios están compuestos por un conjunto de árboles sin podar, de los cuales se extrae un valor global de predicción o clasificación para la respuesta. En el caso de la regresión este valor se obtiene al promediar las predicciones generadas por cada árbol del bosque, mientras que en el caso de la clasificación se obtiene al seleccionar la clase que logra la mayoría de votos entre todos los árboles del bosque (respuesta más frecuente).
El desempeño de bosques aleatorios, al igual que el de los árboles, puede ser superior al de los métodos paramétricos clásicos cuando hay presencia de efectos de interacción complejos entre las variables y cuando la forma funcional que aproxima el verdadero modelo es no lineal.
Otra situación que favorece el uso de bosques aleatorios es cuando se requiere identificar variables predictoras informativas, ya que la metodología genera medidas de importancia para las variables, lo que hace atractivo al método.
Otras ventajas que heredan los bosques aleatorios directamente de los árboles son: el manejo implícito de datos faltantes y el procesamiento de variables cualitativas y cuantitativas sin la necesidad de crear variables indicadoras. Por otra parte, desde el punto de vista computacional, el algoritmo de bosques aleatorios posee la ventaja de poder ser ejecutado en paralelo, ya que los árboles que conforman el bosque pueden ser construidos de forma independiente y esto, en conjunto con el uso de técnicas de computación paralela y distribuida, puede mejorar ostensiblemente el tiempo de ejecución y el escalamiento del algoritmo.

2.2.2. Construcción

Los bosques aleatorios se construyen generando a partir de los datos de entrenamiento un número preestablecido de remuestras aleatorias $(ntree)$, las cuales son tomadas con reemplazo de forma similar a como lo hace *Bootstrap* (Efron, 1979a, 1979b). Por lo tanto, aunque las remuestras incorporan ligeras variaciones aleatorias, estas continúan reflejando el comportamiento que tiene el proceso que generó los datos.

Posteriormente, sobre cada remuestra se construye un árbol en el cual el procedimiento de generación de particiones recursivas difiere respecto al del algoritmo *CART*. La diferencia radica en que previo a la selección de una variable y la determinación de su respectivo punto de corte, se debe generar un subconjunto de variables predictoras $(mtry)$ y de este realizase la selección de la variable que generará la partición. Este subconjunto usualmente tiene tamaño $\frac{p}{3}$ para los problemas de predicción y $\sqrt{p}$ para los problemas de clasificación y puede obtenerse por medio de muestreo aleatorio con o sin reemplazo.
Dicho en otras palabras, el algoritmo de bosques aleatorios agrega un paso previo al algoritmo de construcción de árboles *CART*. Así que, una vez se genera $mtry$, lo siguiente es seleccionar de éste una variable predictora y establecer (de acuerdo a un criterio de minimización) un punto de corte que segmente en dos su rango, tal como se describe en la sección 2.1.2 o en las revisiones del tema que presentan Ziegler y König (2014); Fawagreh, Gaber, y Elyan (2014).
Respecto a la restricción $mtry$ que se impone sobre el espacio de las predictoras, puede decirse que lo que esta busca es construir árboles con combinaciones mas diversas de variables, de forma tal que el modelo global del bosque logre capturar una mayor cantidad de comportamientos conjuntos debidos a combinaciones diferentes de efectos principales y de interacción.
Lo anterior facilita que variables con poder predictivo débil tengan la oportunidad de revelar comportamientos que de otra forma pueden ser omitidos, ya que al momento de realizarce una partición, estas variables no siempre tendrán que competir contra variables de poder predictivo mayor que generalmente las relegan. De esta forma bosques aleatorios logra mitigar el hecho de que las particiones de los árboles no son realizadas bajo un proceso de optimización global sino local y en consecuencia no garantiza que los árboles resultantes sean los mas óptimos respecto a todas las combinaciones posibles de variables.
Sobre estos y otros parámetros que influyen en la estabilidad de las estimaciones, Boulesteix, Janitza, Kruppa, y König (2012) plantean algunas recomendaciones para la elección de sus valores y a la vez describen otros aspectos prácticos del algoritmo.

2.2.3. Precisión y error

La precisión en las predicciones obtenida con bosques aleatorios y *Bagging* por lo general supera por amplio margen a la de los árboles. Para el caso de *Bagging*, Bühlmann y Yu (2002) muestran que matemáticamente esto se debe a que la función de predicción generada es más suave que la de los árboles, ya que en estos últimos dicha función es constantes por regiones (*piecewise function*) y por tanto la respuesta puede pasar rápidamente de un valor a otro ante cambios pequeños en las variables predictoras.

En el caso de bosques aleatorios no es fácil mostrar matemáticamente la conexión existente entre los diferentes elementos que conforman el método y la precisión obtenida con éste. Biau y Scornet (2016) presentan una revisión de varias propuestas que emplean simplificaciones del algoritmo para tratar de explicar su comportamiento teórico, las cuales hasta el momento solo generan resultados parciales, que según estos autores, "son insuficientes para explicar con total generalidad el notable comportamiento que presenta bosques aleatorios".
Un par de propiedades estadísticas del método que ayudan a comprender su relación sesgo-varianza y por ende su bondad para predecir son presentadas por Breiman (2001). Una de ellas determina la convergencia de bosques aleatorios, la cual explica por qué el método no sobreajusta los datos al agregar más árboles al bosque y la otra determina una cota superior para el error, la cual depende de la precisión de los árboles individuales y la correlación presente entre éstos.
Hastie et al. (2009) relacionan estas propiedades con el comportamiento que tiene la media y la varianza muestral de variables aleatorias idénticamente distribuidas y dependientes, ya que bosques aleatorios construye árboles a partir de muestras con estas características. Así, el sesgo del método conserva el mismo sesgo que tienen los árboles debido a que la esperanza de las predicciones de cada árbol es la misma esperanza que la del agregado de todos los árboles, mientras que la varianza del método (ecuación (2.2)) disminuye cuando se logra disminuir la correlación ρ entre los árboles.

$$V[\bar{X}] = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2 \tag{2.2}$$

En el algoritmo de bosques aleatorios el parámetro *mtry* es el encargado de controlar la correlación entre cualquier par de árboles del bosque y por ende, es el que ayuda a disminuir la varianza del método, con lo cual se logra mejorar su desempeño. Es oportuno aclarar que $mtry < p$, ya que cuando $mtry = p$ bosques aleatorios es equivalente a *Bagging* y en este último caso no hay control sobre la correlación.
Otro parámetro que también ayuda a reducir la varianza de bosques aleatorios es *nodesize*, aunque en menor medida. Ya que al igual que en los árboles, con este se controla el número de datos que deben ser agregados para generar la predicción o clasificación del ensamble.
Al igual que en muchas otras metodologías, el desempeño de bosques aleatorios puede determinarse estimando su error sobre un conjunto de datos de prueba o empleando una técnica de remuestreo como validación cruzada sobre el conjunto de datos de entrenamiento. Sin embargo, una característica útil que presenta bosques aleatorios (y *Bagging*) es que realiza validación al mismo tiempo que construye los árboles, generando así implícitamente una estimación del error de prueba conocido como error *out-of-bag (OOB)*.

Para realizar dicha estimación *OOB* se aprovecha una característica del muestreo aleatorio con reemplazo, en el cual la probabilidad de que una observación sea seleccionada al menos una vez para conformar la muestra es de aproximadamente $0.6321 \approx 1 - 1/e$. De esta forma, durante el proceso de construcción de bosques aleatorios pueden dejarse por fuera del ajuste de cada árbol (*OOB*) aproximadamente 36.78 % de las observaciones de entrenamiento y sobre estas estimarse el error de prueba.
Para los problemas de regresión el error de prueba sobre una de estas observaciones *OOB* se calcula tomando el promedio de las predicciones generadas por los árboles que no usaron dicha observación para hacer el ajuste. De manera similar se procede en los problemas de clasificación, pero en este caso el error se calcula tomando la mayoría de votos.

2.3. Importancia de variables

Los bosques aleatorios permiten generar implícitamente medidas de importancia para las variables del modelo, las cuales pueden construirse empleando el índice Gini o la permutación aleatoria de variables (Breiman, 2003). Con estas medidas lo que se busca es cuantificar el impacto que tiene cada variable predictora sobre la variable respuesta y por ende determinar su relevancia o poder predictivo, lo cual facilita el planteamiento de algoritmos de selección de variables como el que se describe en la sección 3.2.1.

2.3.1. Medida de importancia por permutación

La medida de importancia por permutación aleatoria de variables está basada en la precisión que tiene la estimación de la respuesta generada por bosques aleatorios al emplear datos de prueba. Lo que cuantifica esta medida es qué tanto decrementa el error del modelo cuando se rompe la asociación existente entre la respuesta Y y una variable predictora de interés X_j.
Para un árbol *t* del bosque, el valor de esta medida se obtiene al tomar cada una de las observaciones que resultaron *OOB* durante el proceso de construcción del árbol y realizar lo siguiente:

1) Estimar su respuesta o clase $\hat{y}_i^{(t)}$.
2) Calcular su error cuadrático en el caso de predicción (o determinar si la clase predicha fue correcta en el caso de clasificación).
3) Calcular el promedio de los errores *Mean Squared Error (MSE)* en caso de predicción (o la proporción de errores en el caso de clasificación), como se indica en la ecuación (2.3),

$$MSE_{OOB}^{(t)} = \frac{1}{n_{OOB}^{(t)}} \sum_{i=1}^{n_{OOB}^{(t)}} (y_i^{(t)} - \hat{y}_i^{(t)})^2 \tag{2.3}$$

4) Tomar los valores de la variable predictora X_j y permutarlos aleatoriamente con los valores OOB correspondientes a la misma variable.
5) Sobre el mismo árbol *t* y sobre cada una de las nuevas observaciones OOB que resultaron en el paso anterior, repetir los pasos 1), 2) y 3). Esto generará una nueva medición del error, tal como lo indica la ecuación (2.4),

$$MSE_{OOB,permutado}^{(t)} = \frac{1}{n_{OOB}^{(t)}} \sum_{i=1}^{n_{OOB}^{(t)}} (y_i^{(t)} - \hat{y}_{i,permutado}^{(t)})^2 \tag{2.4}$$

6) Calcular la diferencia de los *MSE* obtenidos después y antes de la permutación, como se muestra en la ecuación (2.5),

$$VI_{X_j}^{(t)} = MSE_{OOB,permutado}^{(t)} - MSE_{OOB}^{(t)} \tag{2.5}$$

El valor $VI_{X_j}^{(t)}$ será entonces la medida de importancia de la variable predictora X_j generada por el árbol *t*.
Finalmente, para cada uno de los *ntree* árboles que conforman el bosque se repite el proceso anteriormente descrito en las ecuaciones (2.3), (2.4), (2.5) y las medidas de importancia generadas se promedian para obtener una medida de importancia global para la variable X_j, lo cual es expresado matemáticamente en la ecuación (2.6),

$$VI_{X_j} = \frac{1}{ntree} \sum_{t=1}^{ntree} VI_{X_j}^{(t)} \tag{2.6}$$

Entre mayor sea el valor de la métrica VI_{X_j} mas fuerte sera la asociación que tendrá la variable X_j con la respuesta y por tanto mas importante sera para generar predicciones. Por el contrario, valores cercanos a cero (o incluso negativos) indican que la variable en cuestión carece de poder predictivo o lo que es lo mismo, es una variable irrelevante.

2.3.2. Medida de importancia Gini

La medida de importancia Gini, como su nombre lo indica, está basada en el índice Gini calculado en los árboles al momento de realizar una partición (ver sección 2.1.2). Lo que cuantifica esta medida es la reducción que genera una variable predictora de interés X_j en la heterogeneidad o impureza de la variable respuesta *Y*.

Para un árbol *t* del bosque, el valor de esta medida se obtiene al sumar los decrementos del índice Gini (reducción de impureza) proporcionados por los nodos donde la variable predictora X_j realizó una partición del espacio. Este procedimiento se ejecuta sobre cada uno de los *ntree* árboles que conforman el bosque y posteriormente se promedian (o simplemente suman) todas las medidas de importancia que se generan para la variable X_j, esto con el fin de obtener una medida global de la importancia Gini.

Un inconveniente que presenta esta medida de importancia está relacionado con el hecho de que el proceso de generación de particiones de los árboles *CART* es sesgado cuando alguna de las variables predictoras es categórica y/o tiene valores faltantes (Strobl, Boulesteix, Zeileis, y Hothorn, 2007), lo que hace que las medidas de importancia calculadas sobre árboles sesgados también sean sesgadas. No obstante, Sandri y Zuccolotto (2008) proponen una solución para este problema que puede ser considerada cuando los datos presentan dichas características.

Capítulo 3

Selección de variables

La selección de variables, también conocida como *feature selection*, es un tema que por largo tiempo ha sido estudiado dentro y fuera de la comunidad estadística debido a la relevancia que tiene en el mundo aplicado, tanto así que en algunas áreas (e.g. genética) prácticamente se ha convertido en un requerimiento para poder analizar la información.

Dentro de este tópico existen una amplia gama de propuestas que se han desarrollado con el fin de establecer, a partir del conjunto de datos originales, un subconjunto de variables predictoras con las cuales pueda construirse un modelo estadístico mas simple (principio de la *cuchilla de Occam* – ver: Kelly (2007)), que genere un menor error de predicción o clasificación (según sea el caso) y/o que ayude a mejorar la interpretación de los resultados.

En el contexto de bosques aleatorios, para su implementación es necesario recurrir al concepto de importancia de una variable, ya que con base en este, los métodos de selección pueden construir medidas con las cuales es posible establecer relaciones de orden entre las predictoras y de esta forma lograr identificar y eliminar las variables que se consideran irrelevantes. Esto debido a que cuando se incluyen variables no informativas en un modelo, la precisión de las estimaciones de la variable respuesta usualmente disminuye.

Desafortunadamente las medidas de importancia son un elemento controversial que genera discrepancias entre las metodologías existentes para seleccionar variables, esto debido a que tal como lo menciona Grömping (2009), "teóricamente no hay definida una métrica de importancia de variable en el sentido de una cantidad paramétrica que un estimador deba tratar de calcular". Esta falta de consenso alrededor del tema se ha manifestado en la diversidad de propuestas que han surgido para tratar de evaluar empíricamente la medida de importancia de una variable (Wei, Lu, y Song, 2015) y en la implementación de estas que se realizan en los algoritmos de selección.

3.1. Clasificación de los métodos de selección

En la literatura las diferentes metodologías para seleccionar variables suelen clasificarse de acuerdo a la forma en que usan el algoritmo de selección. Varios autores (e.g. Chandrashekar y Sahin (2014); Saeys, Inza, y Larrañaga (2007); Guyon y Elisseeff (2003); Blum y Langley (1997)) emplean tres categorías para describir este comportamiento, las cual se presentan a continuación:

- Metodologías filtro: el algoritmo de selección calcula primero una métrica o *score* para cada variable predictora y luego construye un *ranking* sobre el cual se establece un umbral o punto de corte que permite realizar la eliminación de las variable. Todo lo anterior se realiza sin recurrir a un método de estimación y por lo tanto este tipo de algoritmos dependen fuertemente de las características intrínsecas que tengan los datos de entrenamiento.

- Metodologías embebidas (o *embedded* por su nombre en inglés): el algoritmo de selección realiza la eliminación de las variables al ejecutarse un método de estimación particular y por lo tanto son algoritmos que dependen completamente del método empleado. Dos metodologías que han sido ampliamente utilizadas para este fin son *Least Absolute Shrinkage and Selection Operator (LASSO)* y bosques aleatorios.

- Metodologías empaquetadas (*wrapper*): el algoritmo de selección aplica primero un método de estimación y mide el desempeño que tienen las predicciones que este genera, luego en base a estas predicciones se determina un subconjunto de variables sobre el cual se aplica nuevamente el método de estimación. El procedimiento anterior se repite iterativamente hasta obtener un subconjunto óptimo de variables.
  Este tipo de algoritmos dependen entonces del método de estimación usado y del tipo de error empleado para cuantificar el desempeño.

Los métodos tipo *wrapper* y *embedded* son los que se emplean con mayor frecuencia para abordar el problema de selección de variables, ya que tal como lo afirman Guyon y Elisseeff (2003), estos "mejoran el desempeño predictivo comparado con los métodos de ranqueo de variables", aunque también hacen la salvedad de que las mejoras no siempre son significativas.
Una de las fortalezas que presentan los algoritmos tipo *wrapper* es la flexibilidad que tienen para incorporar diferentes métodos paramétricos y no paramétricos en la estimación de la respuesta (e.g. modelos lineales generalizados, bosques aleatorios), permitiendo esto que de acuerdo a las características que presenten los datos pueda seleccionarse el método mas apropiado para su estimación.

3.2. Selección de variables con bosques aleatorios

Una alternativa común son las técnicas tipo *wrapper* que emplean las medidas de importancia generadas por bosques aleatorios como criterio base para realizar selección de las variables predictoras. Como suele ocurrir con las diferentes clases de metodologías de selección de variables, el universo de propuestas existentes es amplio y variado, por lo que en la literatura pueden encontrarse diferente trabajos relacionados. Algunos de estos trabajos son los de Gregorutti, Michel, y Saint-Pierre (2017), Hapfelmeier y Ulm (2013), Genuer, Poggi, y Tuleau-Malot (2010), Díaz-Uriarte y Alvarez de Andrés (2006).
Gran parte de la diversidad de propuestas para seleccionar variables con bosques aleatorios se debe a los diferentes criterios que estas emplean, e.g. "el número de pasos usados para rechazar o incluir variables, la fracción de variables rechazadas por paso, el (re-)cálculo de la importancia de las variables, el tipo de medida de importancia empleada, el método usado para evaluar la precisión de las predicciones, la aplicación de métodos de muestreo, la selección *forward* o *backward* utilizada y el criterio de parada del algoritmo"(Hapfelmeier y Ulm, 2013).
Esta variedad de criterios puede verse reflejada en los múltiples algoritmos desarrollados para implementar selección de variables con bosques aleatorios. Por ejemplo, en el software estadístico `R` (R Core Team, 2018), algunos de los paquetes que se encuentran disponibles para este fin son: `vita` (Janitza, Celik, y Boulesteix, 2018), `ranger` (Wright y Ziegler, 2017), `r2VIM` (Szymczak et al., 2016), `VSURF` (Genuer, Poggi, y Tuleau-Malot, 2015), `Boruta` (Kursa y Rudnicki, 2010), `varSelRF` (Diaz-Uriarte, 2007), los cuales – exceptuando el paquete `VSURF` – son evaluados por Degenhardt, Seifert, y Szymczak (2017) es una aplicación de datos genéticos.
Independientemente de los efectos que puedan tener los criterios anteriormente mencionados sobre el proceso de selección, existen otros elementos que afectan directamente las medidas de importancia de las variables (criterio base) y por ende los resultados que se obtienen con los diversos métodos de selección que emplean bosques aleatorios. Para el caso particular en el que se hace uso de las medidas de importancia por permutación (ver sección 2.3.1), estos elementos son:

- La correlación existente entre pares de variables predictoras (ρ_{ij}).
- La correlación entre cada variable predictora X_j y la respuesta Y (τ_j).
- El numero de variables predictoras correlacionadas (q).

Gregorutti et al. (2017) muestran como para datos cuyo modelo generador es el modelo lineal clásico, el valor de la importancia por permutación que tiene una

variable X_j puede ser determinado por medio de la ecuación (3.1),

$$VI_{X_j} = 2\left(\frac{\tau_j}{1-\rho_{ij}+q\cdot\rho_{ij}}\right)^2 \tag{3.1}$$

Un par de hechos que evidencia la ecuación (3.1) es que **i)** cuando la correlación entre X_i y X_j ($\rho_{ij} > 0$) aumenta, la importancia VI de ambas variables disminuye y **ii)** cuando el numero de variables correlacionadas (q) aumenta, la importancia (VI) disminuye y lo hace mas rápidamente. Otro comportamiento interesante, aunque un tanto sorprendente, que también revela esta ecuación es que **iii)** puede darse el caso en el que variables predictoras independientes ($\rho_{ij} = 0$) con poder predictivo (τ_j) bajo tengan mayor importancia ($VI_{X_j} = 2\tau_j^2$) que variables predictoras correlacionadas ($\rho_{ij} > 0$) con poder predictivo (τ_j) alto. Desafortunadamente es difícil extender estos hallazgos a otros modelos (e.g. el modelo de regresión logística), por lo cual es importante tener presente que los valores de ρ_{ij}, τ_j y q que tengan los datos que se estén considerando pueden influir notoriamente sobre los resultados obtenidos al aplicar un procedimiento de selección de variables con bosques aleatorios.

3.2.1. Algoritmo de eliminación recursiva de variables

Muchos de los algoritmos tipo *wrapper* que existen para seleccionar variables hacen uso de la técnica de eliminación hacia atrás (*backward*) para generar iterativamente subconjuntos de variables cada vez menores. Cuando dicha técnica se emplea conjuntamente con bosques aleatorios, las medidas de importancia basadas en permutación aleatoria (VI_{X_j}) que se obtienen para cada variable, permiten configurar algoritmos de selección con los cuales puede mejorarse la predicción de un modelo.

Estos algoritmos, conocidos como algoritmos de eliminación recursiva de variables (*RFE*), calculan en cada iteración las medidas de importancia de todas las variables predictoras y las ordenan de mayor a menor (o a la inversa) de acuerdo al valor obtenido. Esto con el fin de eliminar aquellas variables que obtuvieron el valor mas bajo de importancia y así lograr construir un subconjunto de variables menor en cada paso del algoritmo (ver pseudocódigo 1), siendo óptimo aquel subconjunto en el que el error de predicción se minimiza.

Otro enfoque existente para aplicar selección es el de eliminación no recursiva de variables (*NRFE*), el cual difiere del *RFE* en cuanto a que el ranking de importancia para las variables solo es calculado una vez sobre el modelo completo con todas las predictoras y es mantenido inalterado durante todas las iteraciones del algoritmo.

Un problema que puede presentarse con la selección *NRFE* tiene que ver con las relaciones que se evidencian en la ecuación (3.1), ya que en cada paso del

algoritmo el valor de q disminuye y posiblemente el valor muestral de ρ_{ij} también sufra modificaciones, lo que hace probable que tanto las medidas de importancias como el ranking de las variables cambie. Sobre esto, el estudio realizado por Gregorutti et al. (2017) concluye que el algoritmo *RFE* es mas eficiente que el *NRFE*, además permite obtener modelos con un menor número de variables predictoras y con un error de predicción menor, esto debido a que las variables más informativas quedan mejor posicionadas en los últimos pasos del procedimiento *backward*, incluso cuando hay presencia de correlación entre las variables.

Pseudocódigo 1: *Selección de variables RFE con bosques aleatorios.*

```
while p > 1 do
    Aplicar bosques aleatorios sobre los datos de entrenamiento.
    Calcular el error de bosques aleatorios sobre datos de prueba.
    Tomar los valores de las medidas de importancia de cada variable.
    Ranquear las variables de acuerdo a las importancias obtenidas.
    Eliminar la variable (o grupo de variables) con la menor importancia
end
Seleccionar el modelo compuesto por el subconjunto de variables donde se
  obtuvo el menor error de prueba.
```

Capítulo 4

Estudios de simulación

En esta sección se realizarán estudios de simulación en los que se evaluará el comportamiento que tiene la metodología de selección de variables con bosques aleatorios en problemas de regresión cuando los conjuntos de datos empleados cuentan con **a)** diferentes niveles de correlación entre las variables predictoras, **b)** diferentes niveles de correlación entre las variables predictoras y la variable respuesta, **c)** diferente cantidad de variables predictoras y **d)** diferente cantidad de observaciones.

Para esto se empleará el error cuadrático medio como medida de desempeño y los resultados obtenidos se compararán contra conjuntos de datos que tienen características similares pero que cuentan $n = 12800$ observaciones, sirviendo estos últimos como referencia del comportamiento asintótico que puede llegar a tener el error.

Con los escenarios y comparaciones planteadas lo que se espera es tratar de dilucidar pautas o criterios que ayuden a hacer un uso mas informado de la metodología.

La implementación de los estudios se realizará usando el software estadístico `R`, haciendo uso de los paquetes `MASS` (Venables y Ripley, 2002), `randomForest` (Liaw y Wiener, 2002) y `tidyverse` (Wickham, 2017).

4.1. Configuración de los escenarios de simulación

Los diferentes conjuntos de datos simulados que serán considerados contarán con una variable respuesta Y cuantitativa generada por medio del modelo lineal sin

intercepto presentado en la ecuación (4.1),

$$Y_i = \sum_{i=1}^{p} \beta_i X_i + \epsilon_i \quad con \quad \epsilon_i \sim \mathcal{N}(0, 0.1) \tag{4.1}$$

La cantidad de variables predictoras X_i que harán parte de este modelo serán $p = \{8,\ 16,\ 32\}$ y dichas variables procederán de una distribución normal multivariada ($X \sim N_p(\mu, \Sigma)$) con vector de medias $\mu = 0$ y matriz de covarianzas Σ que cambiará de acuerdo a valores particulares dados a las correlaciones entre las variables.

Estas variables predictoras conformarán cuatro grupos de $q = \{2,\ 4,\ 8\}$ variables y cada grupo tendrá un nivel de correlación diferente **i)** entre sus variables (ρ_{ij}) y **ii)** entre sus variables y la variable respuesta (τ_j)[1].
La idea de considerar un diseño con diferentes valores de q, ρ_{ij} y τ_j es poder configurar conjuntos de datos en los que las variables predictoras tengan diferentes valores de importancia por permutación (ver sección 2.3.1), ya que los resultados generados por los algoritmos de selección de variables con bosques aleatorios dependen directamente de las medidas de importancia calculadas para cada variable (tal como se mencionó en la sección 3.2).

Los valores aproximados de relevancia τ_j que tendrán las variables predictoras de cada grupo se muestran en la Tabla 4.1, estos valores se mantendrán fijos durante todos los escenarios de simulación y lo que se pretende con ellos es incorporar casos en los que la relevancia de las variables sea nula, baja, media y alta.

Tabla 4.1: *Correlaciones muestrales promedio entre la variable respuesta y las variables predictoras de cada grupo.*

		Grupo 1	*Grupo 2*	*Grupo 3*	*Grupo 4*
p	q	$\widehat{\tau_j}$	$\widehat{\tau_j}$	$\widehat{\tau_j}$	$\widehat{\tau_j}$
8	2	0	0.24	0.48	0.74
16	4	0	0.21	0.43	0.74
32	8	0	0.16	0.42	0.74

Para obtener estos niveles de relevancia $\widehat{\tau_j}$ es necesario asignar previamente valores a los coeficientes β_i del modelo que genera los datos de la variable respuesta (ecuación (4.1)), por lo cual las combinaciones de valores que serán empleadas

[1] A esta correlación usualmente se le conoce con el nombre de relevancia, ya que precisamente indica la relevancia que tiene una variable en el modelo.

acá son las que se presentan en el Apéndice A.

Para las correlaciones intra-grupo ρ_{ij} de las variables predictoras, se considerará el caso en el que el grupo de variables irrelevantes ($\tau_j = 0$) es independiente ($\rho_{ij} = 0$) y los grupos de variables relevantes ($\tau_j \neq 0$) son bajamente dependientes ($\rho_{ij} = 0.1$), medianamente dependientes ($\rho_{ij} = 0.5$) y altamente dependientes ($\rho_{ij} = 0.9$).

También, como puede verse en la Tabla 4.2, serán considerados los casos en los que el grupo de variables irrelevantes (*Grupo 1*) es medianamente dependiente y altamente dependiente, mientras que los grupos de variables relevantes (*Grupo 2, 3 y 4*) se mantendrán constantes en los valores de correlación establecidos previamente.

Tabla 4.2: *Correlaciones entre las variables de cada grupo.*

Caso	*Grupo1* ρ_{ij}	*Grupo2* ρ_{ij}	*Grupo3* ρ_{ij}	*Grupo4* ρ_{ij}
Irrelevantes – independientes	0.0	0.1	0.5	0.9
Irrelevantes – medianamente dependientes	0.5	0.1	0.5	0.9
Irrelevantes – altamente dependientes	0.9	0.1	0.5	0.9

Como ilustración, a continuación se presentan la tres matrices de covarianzas Σ que serán empleadas para generar las variables predictoras cuando $q = 2$ ($p = 8$), así como las respectivas relevancias τ_j que tendrán dichas variables.

$$\Sigma = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{matrix} \\ \begin{bmatrix} 1 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 1 \end{bmatrix} \end{array} \qquad \tau_j = \begin{array}{c} \begin{matrix} Y \end{matrix} \\ \begin{bmatrix} 0 \\ 0 \\ 0.24 \\ 0.24 \\ 0.48 \\ 0.48 \\ 0.74 \\ 0.74 \end{bmatrix} \end{array} \begin{array}{l} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \end{array}$$

$$\Sigma = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{matrix} \\ \begin{bmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 1 \end{bmatrix} \end{array} \quad \tau_j = \begin{array}{c} \begin{matrix} Y \end{matrix} \\ \begin{bmatrix} 0 \\ 0 \\ 0.24 \\ 0.24 \\ 0.48 \\ 0.48 \\ 0.74 \\ 0.74 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \end{matrix}$$

$$\Sigma = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{matrix} \\ \begin{bmatrix} 1 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 1 \end{bmatrix} \end{array} \quad \tau_j = \begin{array}{c} \begin{matrix} Y \end{matrix} \\ \begin{bmatrix} 0 \\ 0 \\ 0.24 \\ 0.24 \\ 0.48 \\ 0.48 \\ 0.74 \\ 0.74 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \end{matrix}$$

Con la configuración de valores establecida para q, ρ_{ij} y τ_j, podrán generarse 9 escenarios (ver Tabla 4.3) en los que se examinarán casos de gran interés desde el punto de vista teórico (cuando $\rho_{ij} = 0$ y $\widehat{\tau_j} = 0$ en los *Escenarios 1, 4, 7*), así como casos de gran interés desde el punto de vista aplicado (cuando $\rho_{ij} \neq 0$ y $\tau_j \neq 0$ en los *Escenarios 2, 3, 5, 6, 8, 9*).

Tabla 4.3: *Escenarios de simulación.*

Escenario	p	q	ρ_{ij} *Grupo1*	ρ_{ij} *Grupo2*	ρ_{ij} *Grupo3*	ρ_{ij} *Grupo4*	$\widehat{\tau_j}$ *Grupo1*	$\widehat{\tau_j}$ *Grupo2*	$\widehat{\tau_j}$ *Grupo3*	$\widehat{\tau_j}$ *Grupo4*
1	8	2	0	0.1	0.5	0.9	0	0.24	0.48	0.74
2	8	2	0.5	0.1	0.5	0.9	0	0.24	0.48	0.74
3	8	2	0.9	0.1	0.5	0.9	0	0.24	0.48	0.74
4	16	4	0	0.1	0.5	0.9	0	0.21	0.43	0.74
5	16	4	0.5	0.1	0.5	0.9	0	0.21	0.43	0.74
6	16	4	0.9	0.1	0.5	0.9	0	0.21	0.43	0.74
7	32	8	0	0.1	0.5	0.9	0	0.16	0.42	0.74
8	32	8	0.5	0.1	0.5	0.9	0	0.16	0.42	0.74
9	32	8	0.9	0.1	0.5	0.9	0	0.16	0.42	0.74

Inicialmente lo que se espera determinar en todos los escenarios es si el algoritmo de selección de variables con bosques aleatorios permite obtener modelos

mas parsimoniosos y con menor error de predicción, ya que al haber presentes variables irrelevantes en todos los conjuntos de datos, lo deseable es que dichas variables puedan ser eliminadas correctamente y ello termine repercutiendo favorablemente sobre el error.

En segunda instancia lo que se quiere examinar es el comportamiento que tiene el algoritmo de selección al aumentar el número de variables predictoras en el modelo (4.1), dado que según lo que plantea Gregorutti et al. (2017) en la ecuación (3.1), las medidas de importancia por permutación para un modelo lineal pueden llegar a disminuir rápidamente al incrementar al mismo tiempo q y ρ_{ij}.
Luego, lo que inquieta conocer es si al aumentar q puede llegarse a presentar el caso en el que las medidas de importancia por permutación generadas por los bosques aleatorios para las variables relevantes altamente correlacionadas del *Grupo 4* empiezan a ser tan cercanas a 0 que terminan pareciéndose a las medidas de importancia que tienen las variables irrelevantes independientes del *Grupo 1*, ya que magnitudes similares quizás puedan "confundir" al algoritmo al momento de eliminar las variables y esto provocar resultados no satisfactorios. En nuestro caso será importante prestar atención a lo ocurra en los *Escenarios 7, 8, 9* que es donde los valores de q son mayores.

Otro parámetro que será examinado en cada uno de los 9 escenarios es el número de observaciones, por lo cual se considerarán 3 tamaños de muestra diferentes $n = \{200,\ 800,\ 3200\}$ con los que se conformarán un total de 27 conjuntos de datos sobre los cuales podrá evaluarse tanto el desempeño que tiene la metodología de bosques aleatorios (error de predicción), así como la precisión que tiene la metodología de selección de variables con bosques aleatorios para eliminar las variables irrelevantes.

Para bosques aleatorios la configuración de parámetros que será empleada para realizar la selección será la que por defecto tiene el algoritmo en el caso de regresión, que es: $ntree = 500$, $mtry = p/3$ y $nodesize = 5$.

Configuraciones de escenarios como los descritos anteriormente, son comúnmente empleados en el desarrollo y evaluación de propuestas metodológicas para seleccionar variables en bosques aleatorios. Es por eso que en diferentes estudios, e.g., Archer y Kimes (2008), Altmann, Toloşi, Sander, y Lengauer (2010), Genuer et al. (2010), Toloşi y Lengauer (2011), Gregorutti et al. (2017), pueden encontrarse diseños de experimentos similares con los que además usualmente se abordan problemas de clasificación.

4.2. Resultados

A continuación se presentan los resultados obtenidos al medir el error de predicción MSE sobre datos de prueba en cada uno de los conjuntos considerados. Se iniciará presentando los resultados para los *Escenarios 1-3* y luego se continuará con los resultados de los *Escenarios 4-6* y *7-9*.

Selección de variables partiendo de un modelo lineal con $p = 8$ variables

Para este caso la evolución que presentaron los errores al eliminar de forma iterativa una variable del modelo (4.1) por medio del algoritmo de selección de variables con bosques aleatorios son los que se ilustran en la Figura 4.1.

En ella se aprecia que independientemente de la cantidad de observaciones que tienen los conjuntos de datos ($n = \{200,\ 800,\ 3200\}$) e independientemente del nivel de correlación intra-grupo que tienen las variables, el error mínimo (señalado con puntos sólidos en las gráficas) se alcanza cuando el algoritmo elimina el primer grupo de variables ($p = 6$).
Lo anterior indica que el método de selección es capaz de detectar y eliminar correctamente las variables irrelevantes que se encuentran presentes en los diferentes conjuntos de datos, ya que de acuerdo a como fue configurado el *Grupo 1* de variables, el aporte que estas hacen a la respuesta es nulo o casi nulo (ver valores de los coeficientes del modelo en el Apéndice A) y por lo tanto solo introducen "ruido" que el algoritmo es capaz de filtrar.

En esta figura también se aprecia que independientemente del nivel de correlación que tienen las variables irrelevantes del *Grupo 1* ($\rho_{ij} = \{0.0,\ 0.5,\ 0.9\}$), las curvas de los errores van disminuyendo de nivel al aumentar la cantidad de observaciones en los conjuntos de datos y se van acercando al nivel que tiene la curva de referencia (linea roja punteada).
Este comportamiento revela que el ajuste de los datos por medio de bosques aleatorios requieren tamaños muestrales grandes (alrededor de $n = 3200$) para lograr predicciones razonables cuando el modelo subyacente en los datos es de tipo lineal como el que se empleo acá. Por ejemplo, puede verse que antes de iniciarse la selección de variables ($p = 8$) la diferencia entre las curvas de los errores es amplia y solo disminuye al aumentar n.

Selección de variables partiendo de un modelo lineal con $p = 16$ variables

Para este caso los errores obtenidos al realizar la selección de variables con bosques aleatorios se muestran en la Figura 4.2.

Allí se observa que los errores mínimos y los niveles de las curvas de los errores presentan el mismo comportamiento descrito en la Figura 4.1 (*Escenarios 1-3*). Sin embargo, puede verse que el error mínimo aumenta alrededor de 45 veces al duplicarse la cantidad inicial de variables predictoras en el modelo, ya que se pasa de un $MSE \approx 0.02$ en la Figura 4.1 a un $MSE \approx 0.90$ en la figura 4.2 al incrementar p de 8 a 16.
Este aumento en el error es considerable y es explicado en parte por la *"maldición de la dimensionalidad"*, la cual hace que al aumentar p se reduzca la densidad del conjunto de datos en $n^{1/p}$ (Hastie et al., 2009). Lo que sorprende acá es el crecimiento acelerado que tiene el error, siendo algo que pone en evidencia la alta sensibilidad que tiene el desempeño de la metodología de bosque aleatorios ante cambios en el número de variables predictoras. Esto cuando los datos subyacentes tienen el comportamiento lineal (4.1).

Selección de variables partiendo de un modelo lineal con $p = 32$ variables

En este caso los errores de predicción que se registraron al aplicar el algoritmo de selección se ilustran en la Figura 4.3.

En ella se aprecia que al igual que en el caso anterior (*Escenarios 4-6*) los errores mínimos y los niveles que tienen las curvas de los errores vuelven a ser sistemáticamente mayores (alrededor de 50 veces mas) debido al efecto que tiene el aumento de p. Pero ahora, al observar detenidamente los niveles que tienen las curvas de referencia (lineas rojas punteadas) puede verse que empieza a presentarse un comportamiento diferente cuando la correlación entre las variables irrelevantes del *Grupo 1* es alta ($\rho_{ij} = 0.9$), ya que allí (*Escenario 9*) el nivel de la curva del error es superior al de los *Escenario 7* y *8* donde las variables irrelevantes del *Grupo 1* no están correlacionadas o tienen correlación media.
Este comportamiento lo que sugiere es que el método de bosques aleatorios es sensible a la correlación entre las variables predictoras (colinealidad) a partir de ciertos valores de q (8 en este caso), haciéndose mas evidente cuando los niveles de ρ_{ij} son altos.

La gráfica también muestra que para el *Escenario 7* (en el que la correlación entre las variables del *Grupo 1* es media ($\rho_{ij} = 0.5$)) el punto de mínimo MSE no corresponde con el punto en el que se eliminan todas las variables irrelevantes del *Grupo 1*. Esto se debe a que, tal como se mencionó en la sección 4.1, las magnitudes de las medidas de importancia por permutación para las variables de dos grupos (*Grupo 1* y *Grupo 4* en este caso) tienen valores muy similares que hacen que el algoritmo no pueda diferenciarlas correctamente y cometa errores

al momento de eliminar las variables.

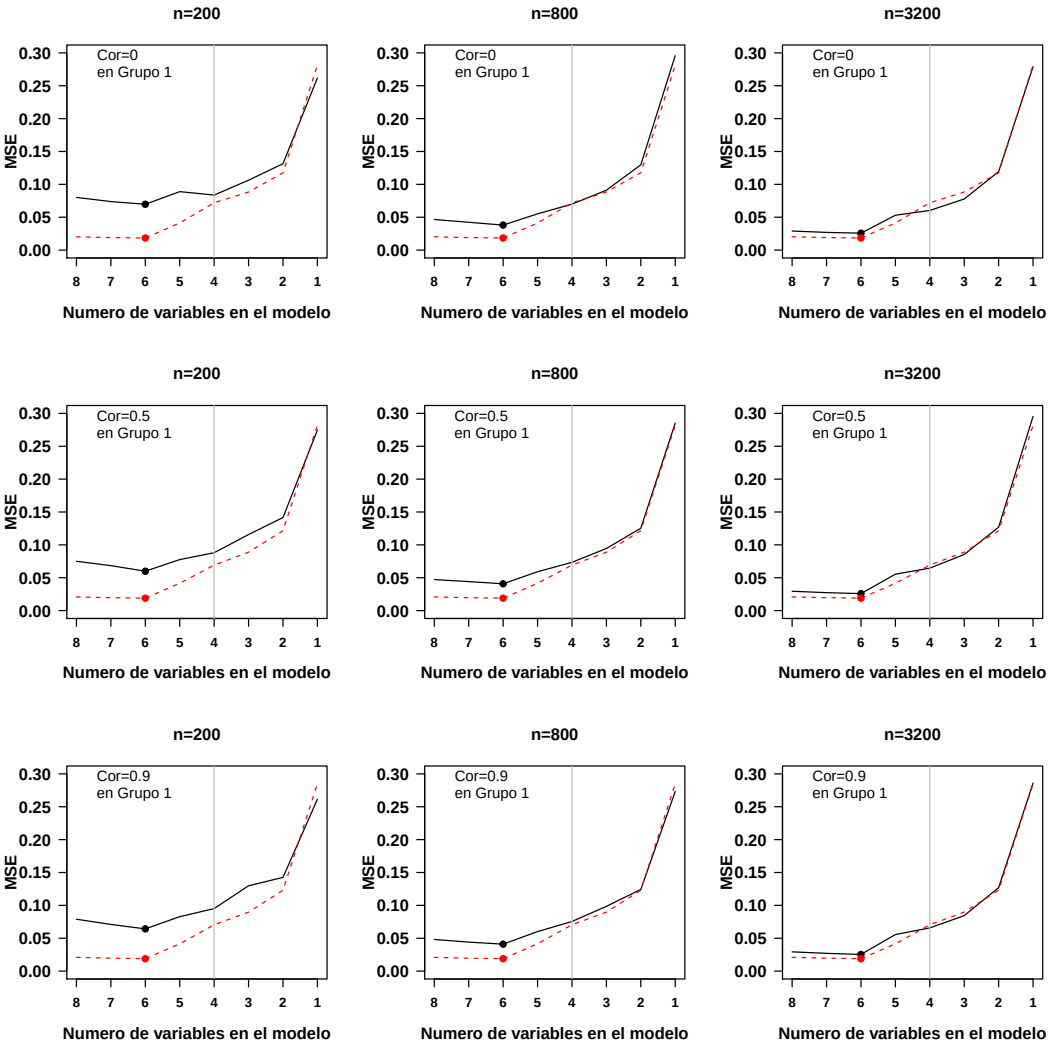

Figura 4.1: *MSE al aplicar el algoritmo de selección de variables con bosques aleatorios sobre conjuntos de datos que inicialmente contienen $p = 8$ variables predictoras. El conjunto de referencia (linea punteada) cuenta con $n = 12800$ observaciones y la linea vertical indica el punto en el que se eliminan la mitad de las variables.*

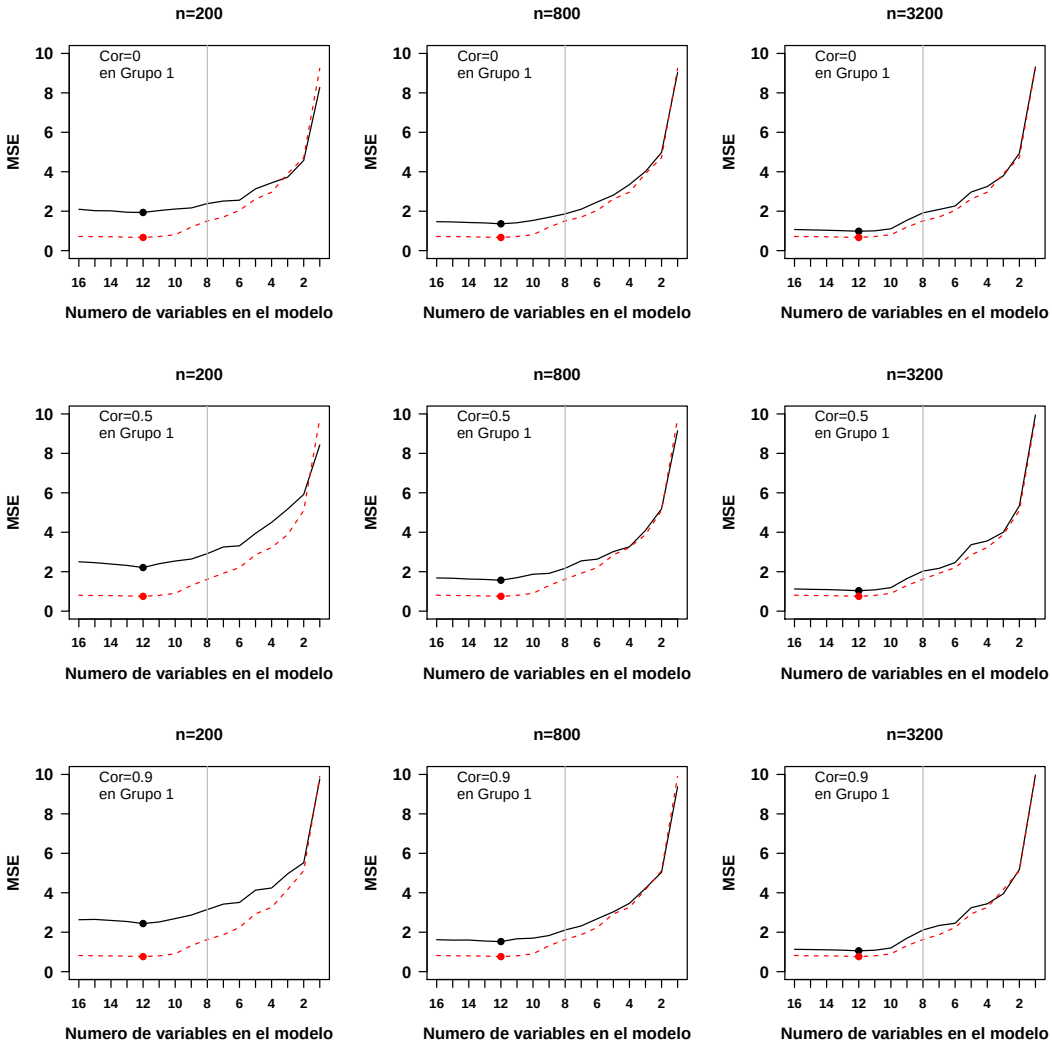

Figura 4.2: *MSE al aplicar el algoritmo de selección de variables con bosques aleatorios sobre conjuntos de datos que inicialmente contienen $p = 16$ variables predictoras. El conjunto de referencia (linea punteada) cuenta con $n = 12800$ observaciones y la linea vertical indica el punto en el que se eliminan la mitad de las variables.*

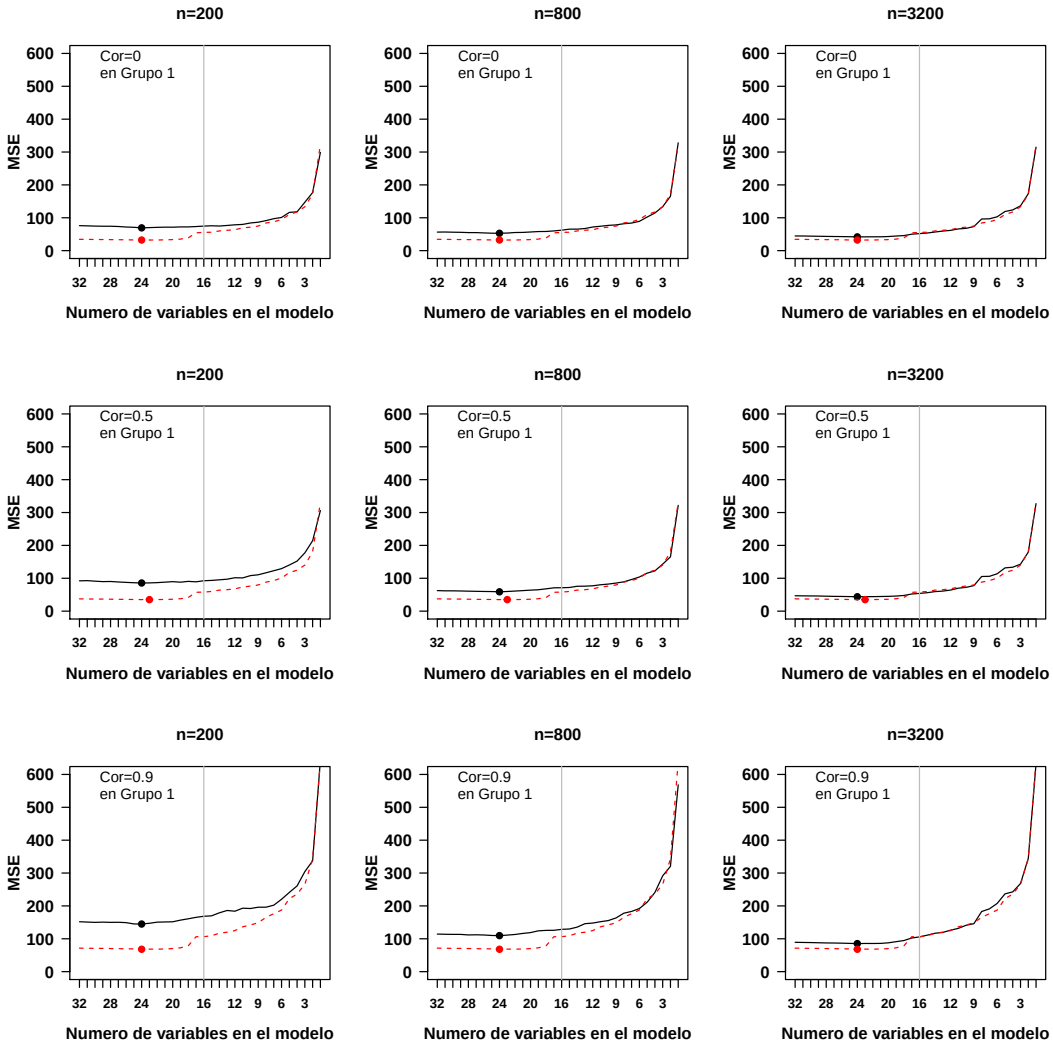

Figura 4.3: *MSE al aplicar el algoritmo de selección de variables con bosques aleatorios sobre conjuntos de datos que inicialmente contienen $p = 32$ variables predictoras. El conjunto de referencia (linea punteada) cuenta con $n = 12800$ observaciones y la linea vertical indica el punto en el que se eliminan la mitad de las variables.*

Capítulo 5

Conclusiones y trabajos futuros

Los resultados obtenidos en los diferentes escenarios implementados en este trabajo, evidencian que la metodología de selección de variables en bosques aleatorios propuesta por Gregorutti et al. (2017) es una alternativa eficiente para eliminar variables, ya que permite obtener modelos mas parsimoniosos (en concordancia con el principio de la *cuchilla de Occam*), con menor error de predicción y compuestos solo por las variables mas relevantes.
Sin embargo, en este trabajo dicha eficiencia comienza a perderse cuando los conjuntos de datos contienen grupos de variables predictoras cuyo tamaño es igual o superior a 8 y cuentan con niveles altos de correlación, ya que bajo esta condición es posible que el algoritmo no logre eliminar correctamente todas las variables irrelevantes y por ende el error de predicción no alcance el mínimo.

Los resultados también evidencian que el número de variables predictoras que contiene el conjunto de datos es un factor que influye de manera importante sobre el desempeño de la metodología de bosques aleatorios, ya que al incrementar la cantidad de variables se genera una degradación acelerada en los niveles de las curvas del error.

Otro factor que juega en contra del error de predicción, aunque no tan drásticamente como lo hace el número de variables predictoras, es la cantidad de observaciones que contiene el conjunto de datos, debido a que se requirieron tamaños muestrales grandes para obtener desempeños cercanos a los de las curvas de referencia.
Claro está, que para los resultados obtenidos acá, debe considerarse el hecho de que el método de bosques aleatorios esta siendo empleado bajo el escenario menos favorable, pues tal como se mencionó en la sección 2.1.1 haciendo referencia a James et al. (2013), el desempeño de los arboles y por ende el desempeño de bosques aleatorios tiende a ser superior "cuando la verdadera forma funcional

que relaciona las variables predictoras con la variable respuesta es altamente no lineal y compleja" y no bajo las condiciones de linealidad en las que fue empleado en estos estudios.

Lo anterior sugiere que una extensión plausible del presente trabajo sería considerar el caso en el que el modelo subyacente en los datos fuera de tipo **no-lineal**, tanto en escenarios de **regresión** como de **clasificación**. Lo cual favorecería al desempeño de bosques aleatorios y a la vez abordaría, al menos de forma exploratoria, el problema de selección de variables en bosques aleatorios en situaciones en las que no se sabe de antemano el comportamiento que puedan tener las medidas de importancia por permutación para las variables, ya que la ecuación (3.1) solo aplica para modelos lineales bajo escenarios de regresión.
Esto último indica que el caso de un modelo **lineal** bajo un escenario de **clasificación** es otra opción relevante que podría ser explorada en futuros trabajos.

Otra forma en la que podrían extenderse los estudios realizados en este trabajo es tratando de determinar valores óptimos para los parámetros del algoritmo de bosques aleatorios ($mtry$, $nodesize$ y $ntree$), esto con el fin de lograr minimizar el error de predicción en los conjuntos de datos considerados acá y para lo cual Probst, Wright, y Boulesteix (2019) presentan varias estrategias.

Apéndice A

Coeficientes del modelo lineal

Tabla A.1: *Coeficientes del modelo (4.1) cuando la correlación entre variables predictoras en el grupo 1 es* $\rho_{ij} = 0.0$.

		Grupo 1	*Grupo 2*	*Grupo 3*	*Grupo 4*
p	q	$\rho_{ij} = 0.0$	$\rho_{ij} = 0.1$	$\rho_{ij} = 0.5$	$\rho_{ij} = 0.9$
8	2	-0.001	0.14	0.22	0.27
16	4	0	0.60	0.66	0.78
32	8	0	2.00	2.10	2.32

Tabla A.2: *Coeficientes del modelo (4.1) cuando la correlación entre variables predictoras en el grupo 1 es* $\rho_{ij} = 0.5$.

		Grupo 1	*Grupo 2*	*Grupo 3*	*Grupo 4*
p	q	$\rho_{ij} = 0.5$	$\rho_{ij} = 0.1$	$\rho_{ij} = 0.5$	$\rho_{ij} = 0.9$
8	2	-0.001	0.15	0.215	0.27
16	4	0	0.64	0.67	0.80
32	8	0	2.11	2.10	2.35

Tabla A.3: *Coeficientes del modelo (4.1) cuando la correlación entre variables predictoras en el grupo 1 es* $\rho_{ij} = 0.9$.

		Grupo 1	*Grupo 2*	*Grupo 3*	*Grupo 4*
p	q	$\rho_{ij} = 0.9$	$\rho_{ij} = 0.1$	$\rho_{ij} = 0.5$	$\rho_{ij} = 0.9$
8	2	-0.001	0.150	0.215	0.270
16	4	-0.001	0.640	0.670	0.800
32	8	-0.070	3.000	2.900	3.100

Referencias

Altmann, A., Toloşi, L., Sander, O., y Lengauer, T. (2010, 04). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340-1347. Descargado de https://doi.org/10.1093/bioinformatics/btq134 doi: 10.1093/bioinformatics/btq134

Archer, K. J., y Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics Data Analysis*, *52*(4), 2249 - 2260. Descargado de http://www.sciencedirect.com/science/article/pii/S0167947307003076 doi: https://doi.org/10.1016/j.csda.2007.08.015

Biau, G., y Scornet, E. (2016, 01 de Jun). A random forest guided tour. *TEST*, *25*(2), 197–227. Descargado de https://doi.org/10.1007/s11749-016-0481-7 doi: 10.1007/s11749-016-0481-7

Blum, A. L., y Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1), 245 - 271. Descargado de http://www.sciencedirect.com/science/article/pii/S0004370297000635 doi: https://doi.org/10.1016/S0004-3702(97)00063-5

Boulesteix, A.-L., Janitza, S., Kruppa, J., y König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493–507. Descargado de http://dx.doi.org/10.1002/widm.1072 doi: 10.1002/widm.1072

Breiman, L. (1996, 01 de Aug). Bagging predictors. *Machine Learning*, *24*(2), 123–140. Descargado de https://doi.org/10.1023/A:1018054314350 doi: 10.1023/A:1018054314350

Breiman, L. (2001, 01 de Oct). Random forests. *Machine Learning*, *45*(1), 5–32. Descargado de https://doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Breiman, L. (2003). Manual–setting up, using, and understanding random forests v4.0. *Unpublished manuscript*, *available at: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf*.

Breiman, L., Friedman, J., Olshen, R., y Stone, C. (1984). *Classification and regression trees*. The Wadsworth.

Bühlmann, P., y Yu, B. (2002, 08). Analyzing bagging. *Ann. Statist.*, *30*(4), 927–961. Descargado de https://doi.org/10.1214/aos/1031689014 doi: 10.1214/aos/1031689014

Chandrashekar, G., y Sahin, F. (2014). A survey on feature selection methods. *Computers Electrical Engineering*, *40*(1), 16 - 28. Descargado de http://www.sciencedirect.com/science/article/pii/S0045790613003066 (40th-year commemorative issue) doi: https://doi.org/10.1016/j.compeleceng.2013.11.024

Degenhardt, F., Seifert, S., y Szymczak, S. (2017, 10). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, *20*(2), 492-503. Descargado de https://doi.org/10.1093/bib/bbx124 doi: 10.1093/bib/bbx124

Diaz-Uriarte, R. (2007). Genesrf and varselrf: a web-based tool and r package for gene selection and classification using random forest. *BMC Bioinformatics*, *8*. Descargado de https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-328 doi: 10.1186/1471-2105-8-328

Díaz-Uriarte, R., y Alvarez de Andrés, S. (2006, 06 de Jan). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. Descargado de https://doi.org/10.1186/1471-2105-7-3 doi: 10.1186/1471-2105-7-3

Efron, B. (1979a, 01). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, *7*(1), 1–26. Descargado de https://doi.org/10.1214/aos/1176344552 doi: 10.1214/aos/1176344552

Efron, B. (1979b). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, *21*(4), 460-480. Descargado de http://www.jstor.org/stable/2030104

Fawagreh, K., Gaber, M. M., y Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, *2*(1), 602-609. Descargado de https://doi.org/10.1080/21642583.2014.956265 doi: 10.1080/21642583.2014.956265

Genuer, R., Poggi, J.-M., y Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225 - 2236. Descargado de http://www.sciencedirect.com/science/article/pii/S0167865510000954 doi: https://doi.org/10.1016/j.patrec.2010.03.014

Genuer, R., Poggi, J.-M., y Tuleau-Malot, C. (2015). VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal*, *7*(2), 19–33. Descargado de https://doi.org/10.32614/RJ-2015-018 doi: 10.32614/RJ-2015-018

Gey, S., y Nedelec, E. (2005, Feb). Model selection for cart regression trees.

IEEE Transactions on Information Theory, *51*(2), 658-670. doi: 10.1109/TIT.2004.840903

Gregorutti, B., Michel, B., y Saint-Pierre, P. (2017, 01 de May). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678. Descargado de https://doi.org/10.1007/s11222-016-9646-1 doi: 10.1007/s11222-016-9646-1

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*(4), 308–319. Descargado de http://www.jstor.org/stable/25652309

Guyon, I., y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157-1182. Descargado de http://www.jmlr.org/papers/v3/guyon03a.html

Hapfelmeier, A., y Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics Data Analysis*, *60*, 50 - 69. Descargado de http://www.sciencedirect.com/science/article/pii/S0167947312003490 doi: https://doi.org/10.1016/j.csda.2012.09.020

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning* (2.ª ed.). Springer-Verlag New York. doi: 10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning*. Springer-Verlag New York. doi: 10.1007/978-1-4614-7138-7

Janitza, S., Celik, E., y Boulesteix, A.-L. (2018, 01 de Dec). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, *12*(4), 885–915. Descargado de https://doi.org/10.1007/s11634-016-0276-4 doi: 10.1007/s11634-016-0276-4

Kelly, K. T. (2007). Ockham's razor, empirical complexity, and truth-finding efficiency. *Theoretical Computer Science*, *383*(2), 270 - 289. Descargado de http://www.sciencedirect.com/science/article/pii/S0304397507003222 doi: https://doi.org/10.1016/j.tcs.2007.04.009

Kim, H., y Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*(454), 589-604. Descargado de https://doi.org/10.1198/016214501753168271 doi: 10.1198/016214501753168271

Kursa, M. B., y Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13. Descargado de http://www.jstatsoft.org/v36/i11/

Liaw, A., y Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18-22. Descargado de https://CRAN.R-project.org/doc/Rnews/

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. Descargado de http://dx.doi.org/10.1002/widm.8 doi: 10.1002/widm.8

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348. Descargado de http://dx.doi.org/10.1111/insr.12016 doi: 10.1111/insr.12016

Messenger, R., y Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, *67*(340), 768-772. Descargado de https://doi.org/10.1080/01621459.1972.10481290 doi: 10.1080/01621459.1972.10481290

Morgan, J. N., y Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*(302), 415-434. Descargado de http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1963.10500855 doi: 10.1080/01621459.1963.10500855

Probst, P., Wright, M. N., y Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301. Descargado de https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1301 doi: 10.1002/widm.1301

Quinlan, J. R. (1986, 01 de Mar). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. Descargado de https://doi.org/10.1007/BF00116251 doi: 10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

R Core Team. (2018). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de https://www.R-project.org/

Rokach, L., y Maimon, O. (2005, Nov). Top-down induction of decision trees classifiers - a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476-487. doi: 10.1109/TSMCC.2004.843247

Saeys, Y., Inza, I., y Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. Descargado de http://dx.doi.org/10.1093/bioinformatics/btm344 doi: 10.1093/bioinformatics/btm344

Sandri, M., y Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611-628. Descargado de https://doi.org/10.1198/106186008X344522 doi: 10.1198/106186008X344522

Strobl, C., Boulesteix, A.-L., Zeileis, A., y Hothorn, T. (2007, 25 de Jan). Bias

in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. Descargado de https://doi.org/10.1186/1471-2105-8-25 doi: 10.1186/1471-2105-8-25

Strobl, C., Malley, J., y Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*(4), 323–348. Descargado de https://psycnet.apa.org/doiLanding?doi=10.1037%2Fa0016973

Szymczak, S., Holzinger, E., Dasgupta, A., Malley, J. D., Molloy, A. M., Mills, J. L., ... Bailey-Wilson, J. E. (2016, 01 de Feb). r2vim: A new variable selection method for random forests in genome-wide association studies. *BioData Mining*, *9*(1), 7. Descargado de https://doi.org/10.1186/s13040-016-0087-3 doi: 10.1186/s13040-016-0087-3

Tolo̧şi, L., y Lengauer, T. (2011, 05). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986-1994. Descargado de https://doi.org/10.1093/bioinformatics/btr300 doi: 10.1093/bioinformatics/btr300

Venables, W. N., y Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Descargado de http://www.stats.ox.ac.uk/pub/MASS4 (ISBN 0-387-95457-0)

Wei, P., Lu, Z., y Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, *142*, 399-432. Descargado de https://doi.org/10.1016/j.ress.2015.05.018

Wickham, H. (2017). tidyverse: Easily install and load the 'tidyverse' [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=tidyverse (R package version 1.2.1)

Wright, M., y Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software, Articles*, *77*(1), 1–17. Descargado de https://www.jstatsoft.org/v077/i01 doi: 10.18637/jss.v077.i01

Xiang, C., Minghui, W., y Heping, Z. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 55-63. Descargado de https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.14 doi: 10.1002/widm.14

Zhou, Z.-H. (2012). *Ensemble methods* (1.ª ed.). Chapman and Hall/CRC.

Ziegler, A., y König, I. R. (2014). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), 55-63. Descargado de https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1114 doi: 10.1002/widm.1114