Análisis de Sensibilidad Mediante Random Forest

Presentado en la Escuela Técnica Superior de Ingenieros Industriales de la Universidad Politécnica de Madrid para la obtención del título de Grado de Ingeniería en Tecnologías Industriales

Alumno:	Marta García Ruiz de León
Especialidad:	Organización Industrial
Tutores:	José Manuel Mira McWilliams
	Ismael Ahrazem Dfuf

Julio 2018

Madrid

Dedicatoria y Agradecimientos

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem." -- John Tukey

Me gustaría dedicarles este trabajo a toda mi familia, mis padres Juan Antonio y María y mi hermana Alejandra, ya que sin su apoyo y ánimo durante estos años no hubiera llegado a donde estoy hoy en día.

Además me gustaría agradecer a mi tutor José Mira Mcwilliams por toda su ayuda durante el desarrollo de este proyecto. Especialmente, agradecerle a Ismael por todas las horas que ha dedicado en ayudarme a realizar este proyecto y su infinita paciencia a la hora de explicarme todo.

Agradezco y dedico también este trabajo a todos mis compañeros, por hacer estos últimos años únicos e inolvidables. Desde mis amigas con las que empecé esta larga y dura carrera, como Cris, Sofía, Ángela, Lucía, pero también María y Paula sin las que este año hubiera sido mucho más duro.

En especial quiero agradecer a Esther y Carlota por ser mi apoyo constante y ayuda en un año en el que la distancia nos ha separado pero que no ha evitado que hayamos seguido pasando momentos inolvidables.

Resumen Ejecutivo

El objetivo de este Trabajo de Fin de Grado es analizar la variabilidad en la respuesta de un modelo debido a sus efectos principales e interacciones. La importancia de variables se analizará a través de la creación de un modelo no paramétrico como es el *random forest* que se generará a partir de datos simulados o reales y sobre él se estimará la descomposición ANOVA para la variabilidad debida a los efectos principales e interacciones. Se compararán los resultados obtenidos al realizar la misma descomposición ANOVA a los datos iniciales, probando hasta qué punto la aproximación random forest es precisa.

Herramienta Random Forest

El algoritmo de random forest es un algoritmo de machine learning que surge como evolución de los árboles de decisión. Los árboles de decisión clasifican los datos de un modelo realizando una serie de particiones binarias, permitiendo realizar predicciones futuras en base a esta clasificación. Los random forests estarán formados por un gran número de árboles de decisión, creados a partir de la técnica del *bagging o bootstrap aggregating*. Este algoritmo mejora los árboles de decisión, creando un modelo más fiable y preciso evitando problemas como el *overfitting.*

Los random forests son un caso particular de la técnica de conjuntos (ensembles) de árboles, consiste en promediar los resultados de un conjunto de modelos individuales (árboles de decisión-CART en este caso). Cada modelo CART se estima partir de un remuestreo de la muestra original. Además, en cada divsión de los árboles no se consideran todas las variables de entrada sino un subconjunto elegido al azar.

La herramienta, es capaz de tratar modelos muy complejos con un gran número de variables y orbservaciones. Bien es cierto, que en ocasiones el hecho de que se comporte en cierta manera como una caja negra al que se le introducen unos valores y de la que obtenemos una respuesta, dificulta la comprensión de la misma. Es por esto, que como se va a analizar su sensiblidad como modelo de entrada-salida, se va a realizar una investigación a base de simulaciones y datos reales de cómo las propiedades de la herramienta afectan a su salida.

Para obtener la mejor salida del random forest, se tendrán que realizar experimentos de optimización de los principales parámetros del mismo. Estos serán:

- Mtry: Número de variables que se seleccionarán en cada partición de cada árbol del bosque. Es uno de los parámetros más importantes del bosque puesto que introduce una gran aleatoriedad al modelo, y dependiendo del número de variables seleccionadas se obtendrán mejores resultados en las predicciones del modelo. Cabe destacar que el valor recomendado para mtry en los problemas de regresión es de *p/3* y de $\sqrt{p}$ para un problema de clasificación.

- Ntree: Número de árboles que forman el random forest. Es un parámetro importante puesto que los experimentos indican que a mayor número de árboles menos problemas de *overfitting* deberían ocurrir.

Para encontrar los parámetros óptimos, se irán variando los parámetros, de manera que se dejará uno de los dos fijos con el valor predeterminado por la función de *R random forest* y se irá aumentando el otro. Para ecntrar el óptimo, se mostrará en una gráfica, la variabilidad total del modelo explicada por la variables del mismo.Se seleccionará aquel punto que explique mejor el modelo. Además, se comprobará que para este punto las medidas del error sean las mínimas

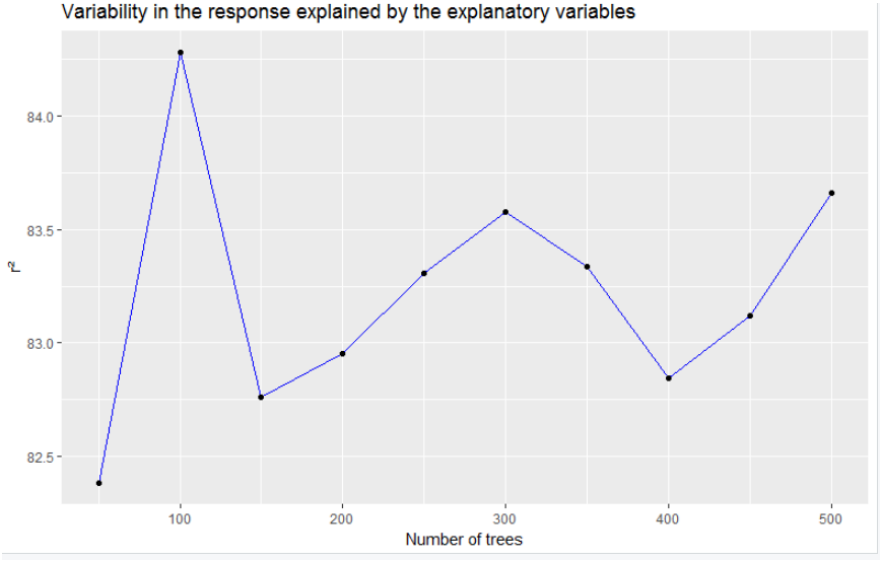


Ilustración 1 Ejemplo de gráfica de obtención óptimo RF

Para analizar la variabilidad de la salida de diferntes modelos, antes de realizar este experimentación con un modelo no paramétrico como es el modelo random forest, primero se va a estudiar un modelo paramétrico (regresión).

Análisis ANOVA

Un ANOVA (análisis de variancia) descompone la variabilidad de la variable de respuesta entre los diferentes efectos principales e interacciones de los factores.33

El procedimiento de descomsición de varianza, utilizando ANOVA (descompone la variabilidad de la variable de respuesta entre los diferentes factores), estima la contribución de cada efecto a la varianza de la variable dependiente, es decir, se respondererá a las siguientes preguntas principales:

- los factores que tienen un efecto significativo en la respuesta, y/o
- La parte de la variabilidad de la variable de respuesta que puede atribuirse a cada factor o a sus interacciones con los demás.

Un análisis ANOVA, realiza las siguiente hipótesis sobre el modelo:

- Distribución normal de los datos
- Muestras aleatorias simples
- Varianza constante

Por lo tanto, para poder expresar la variabilidad total del modelo como la suma de las variabilidades de cada factor e interacciónes es fundamental que el diseño sea otrogonal.

La primera parte de este trabajo será la generación de un modelo con variables ortogonales. Se utlizarán dos maneras de generar variables ortogonales:

- Sequencia Sobol: La función sobol, generará variables pseudoaleatorias con distribución deseada. Con este método no se obtendrán varibales 100% ortogonales pero aceptables para modelo sin interacciones entre variables.

- Codificación de variables: Con este método obtendrán variables 100% ortogonales.

Para lograr una compresión en profundidad, acerca de ambas herramientas (ANOVA y Random Forst) se van a estudiar varios modelos.

En este trabajo:

1) Se van a estimar modelos random forests a partir de un conjunto de datos originales obtenidos a partir de un experimento real o por simulación, con diseños ortogonales.

2) Se harán predicciones a partir de los modelos random forests para los valores del diseño ortogonal del apartado anterior.

3) Se estimarán, además de los random forests, modelos de regresión paramétricos com los que se presentan más abajo.

4) Se estimarán random forest a partir de las pedicciones de los modelos de regresión 3)

5) Se realizarán ANOVA sobre a) los datos originales b) los modelos random forest de apartado 2 c) los modelos de regresión del apartado 3) d) los random forests 4)

6) se compararán los resultados de los 4 ANOVA anteriores apara estudiar la bonda de la aproximación , a afectos de los ANOVA, que proporcionan los random forst y los modelos de regresión. La referencia para las comparaciones será el ANOVA de los datos originales.

Los modelos

Como es lógico, para poder sacar conclusiones generales a cerca de las herramientas analizadas es necesario analizar diferentes modelos. El primer modelo que se va a

analizar en la experimentación es el modelo de regresión lineal múltiple. Este modelo responde a la siguiente ecuación:

$$y = \sum_{i=0}^{n} \beta_i x_i$$

Este modelo, también se va a analizar con un término aleatorio como es el ruido, ya que introducirá una pequeña complicación al mismo.

$$y = \sum_{i=0}^{n} \beta_i x_i + \varepsilon_i$$

El modelo anterior no tiene interacciones, pero a contiuación se presenta otro más complejo que sí las inlcuye a través d los términos de productos cruzados, puesto que el modelo random forest a priori es capaz de encontrar las posibles interacciones entre las variables cuando se estima el modelo como un modelo lineal. Se va estimar un modelo con interacciones y se van a observar si estas son significativas. En caso de que lo sean, se cumplirá que efectivamente el modelo lineal no era el adecuado para explicar el modelo. La ecuación que podría seguir el modelo con interacciones es la siguiente:

$$y = \beta_1 X_1 + \beta_2 X_2 + \alpha_1 X_1 X_2 + \alpha_2 X_1^2 + \alpha_3 X_2^2$$

Por útlimo, se va a estudiar un modelo real para explicar el método de la obtención de las variables ortogonales mediante la codificación de las mismas por un diseño factorial 2^k.

Para realizar la comparación entre ambos modelos, se calcularán unos diagramas de barras en los que se mostrará en forma de porcentaje la varibilidad que cada factor e interacción introducen a la salida de cada modelo. A priori, se sabe que el modelo *random forest* es capaz de manejar modelos con interacciones, algo que el análsis ANOVA de un modelo lineal no es capaz de manejar

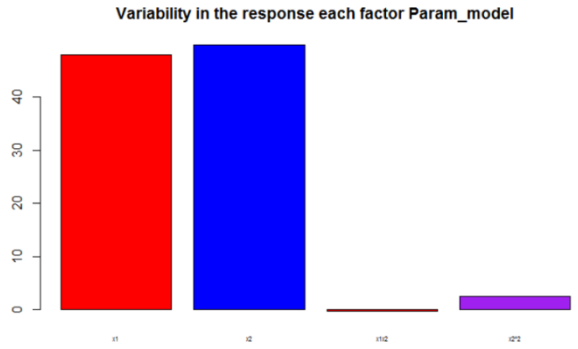


Ilustración 2 Ejemplo gráfica Importancia de Variables

Palabras Clave: Random forest, ANOVA, MAE, variabilidad total explicada, interacciones, modelo parametrizado, Ntree ,Mtry

Códigos UNESCO: 120304, 120326, 120903, 120904, 120909, 120914

Índice General

Capítulo 1: Introducción

1.1 Estado del Arte

Machine learning, o el aprendizaje automático, como es conocido en castellano, es una aplicación de inteligencia artificial (AI) que permite a los sistemas tener la habilidad de aprender automáticamente y mejorar con la experiencia, sin tener que ser programado específicamente para ello.

Con el paso del tiempo el aprendizaje automático, empezó a enfocarse en el razonamiento probabilístico y la investigación basada en la estadística.

El proceso de aprendizaje comienza con la recopilación de datos u observaciones. Con estas variables de entrada y sus respuestas se buscarán patrones, a través de los cuales se tomarán decisiones en el futuro. Estas decisiones estarán basadas en los datos que se han recopilado, de tal manera que se conseguirá que los ordenadores aprendan automáticamente sin necesidad de que tenga que haber intervención humana.

1.1.1 Algoritmos de Machine Learning

Los modelos de machine learning se basan básicamente en tres tipos de aprendizaje:

- **Aprendizaje supervisado**: Cuando se tenga un conjunto de variables y sus salidas perfectamente clasificadas, se creará una función, que se formará usando un conjunto (set) de los datos llamado set de entrenamiento (*training set)*. Después se aplicará esta función a otro set de datos, llamados el *test set*, o set de prueba, para realizar una predicción, y ver lo exacto que es modelo. La finalidad es crear una función que se ajuste bien a nuevos datos.
- **Aprendizaje no supervisado**: Este algoritmo, se utilizará cuando los datos no estén clasificados, por lo que no cuenta con ningún tipo de indicación previa. Por lo tanto, se tratarán los datos de entrada como variables aleatorias, y el sistema tendrá que ser capaz de reconocer patrones para clasificar las nuevas entradas.
- **Aprendizaje por refuerzo:** Este algoritmo, funciona a través del ensayo y error. Se obtienen los datos de entrada a través del feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones.

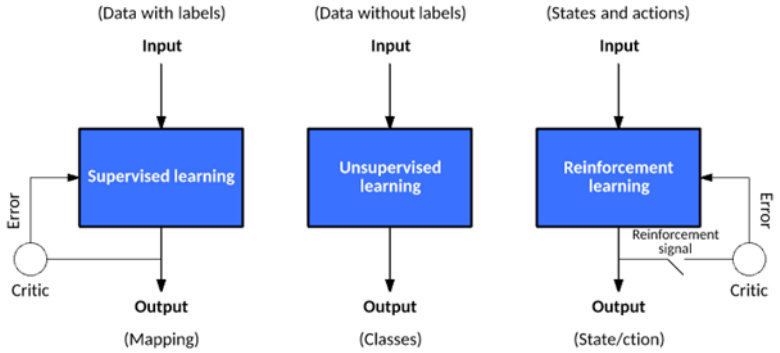


Ilustración 3 Algoritmos de Machine Learning

Cabe destacar también el aprendizaje semisupervisado, que mezcla datos clasificados con datos sin clasificación.

El aprendizaje supervisado, es el que más se utiliza en problemas estadísticos de regresión y predicción estadística, mientras que el no supervisado es más común en la clasificación de variables que se suministran al sistema.

Entre los muchos modelos de aprendizaje automático, vamos a destacar los siguientes:

1.1.2 Redes Neuronales

Las Redes neuronales, o *Artificial Neural Networks* (ANN), son modelos computacionales que están basados en la idea de que el aprendizaje del cerebro humano puede ser imitado. Las redes neuronales, están compuestas por una serie de nodos que imitan a las neuronas humanas, de manera que cada nodo recibe varios inputs y produce una salida. Por lo tanto, la red se va generando de forma autónoma.

La ventaja de este procedimiento es que es capaz de almacenar información redundante y es capaz de contrastarla en el caso de que la red maneje información equivocada. Se puede decir, que es un sistema robusto ante el error y es capaz de tolerar variables de entrada no relevantes o el ruido.

Las redes neuronales, están organizadas típicamente en capas paralelas, que están formadas por nodos interconectados. Comúnmente, las redes están formadas por tres capas, cada una contiendo una función, por lo que la salida será combinación de las mismas. La primera capa, contiene la Función de propagación, que dará como resultado una salida que es combinación de todas las entradas multiplicada por el peso. Teniendo en cuenta esta salida, se puede activar o no la Función de Activación, que dará lugar a la última capa, en la que se encuentra la Función de Transferencia, que se encarga de acotar la salida y comunicar los nodos correctamente.

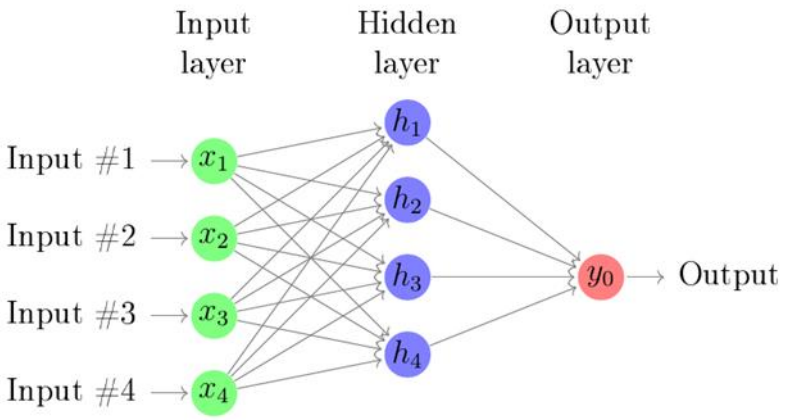


Ilustración 5 Red Neuronal

1.1.3 Máquinas de Vector Soporte

Las máquinas de Vector Soporte, son un algoritmo de machine learning de *aprendizaje supervisado* utilizados comúnmente para problemas de regresión y clasificación.

En las Support Vector Machines (SVM), dado un set de datos de entrenamiento (training set) ya clasificados, el algoritmo es capaz de dar como resultado un híperplano óptimo que categorizará nuevos datos. En el caso de un problema bidimensional, este hiperplano es una línea que separa los datos a ambos lados de la misma. Se llama Vector Soporte, al vector que está formado por el conjunto de puntos más cercano al hiperplano y son los datos más complicados de clasificar, teniendo una influencia muy elevada en la localización del hiperplano óptimo.

Se conocerán como *atributos* a las variables de predicción y como *característica* al factor principal de clasificación. La metodología para conseguir este hiperplano consistirá en la proyección de los atributos en subespacios de dimensión superior a la de los atributos, consiguiendo una separación de las variables muy eficiente. Los métodos más comunes son las funciones conocidas como Kernel, entre las que destacan la Polinomial-homogénea o la Radial Gaussiana.

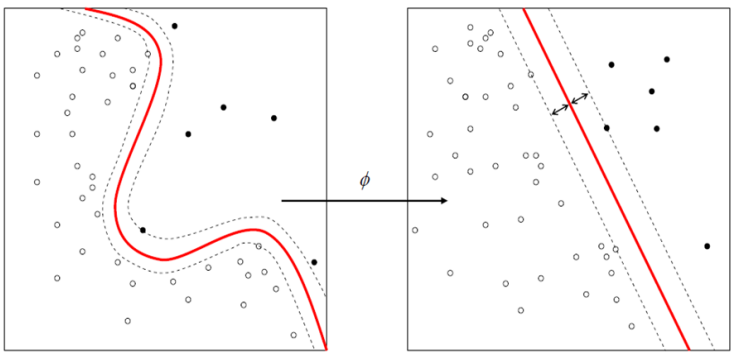


Ilustración 6 Máquina de Vector Soporte

1.1.4 Árboles de Decisión

Los árboles de decisión son un método de minería de datos comúnmente utilizados para establecer sistemas de clasificación basados en múltiples variables o para desarrollar algoritmos de predicción para una variable objetivo. Este método clasifica una población en segmentos en forma de rama que construyen un árbol invertido con un nodo raíz, nodos internos y nodos de hoja. El algoritmo no es paramétrico y puede manejar eficientemente grandes y complicados conjuntos de datos sin imponer una estructura paramétrica complicada. Cuando el tamaño de la muestra es lo suficientemente grande, los datos del estudio pueden dividirse en conjuntos de datos de entrenamiento y validación. Utilizando el conjunto de datos de formación para construir un modelo de árbol de decisión y un conjunto de datos de validación para decidir el tamaño de árbol apropiado necesario para lograr el modelo final óptimo.

Este algoritmo, se explicará de forma más detalla en el siguiente capítulo, ya que es el algoritmo en el que se basará este trabajo para realizar los experimentos y simulaciones.

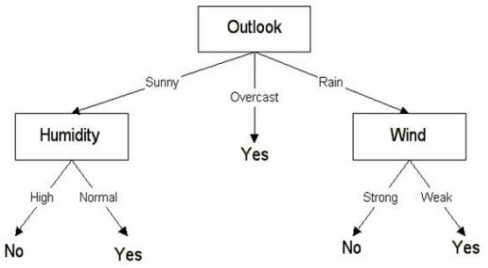


Ilustración 7 Ejemplo de formación árbol de decisión, mediante preguntas binarias

1.2 Objetivos

El objetivo principal de este Trabajo de Fin de Grado, es estudiar la variabilidad en la respuesta de un modelo atribuida a cada variable del modelo y a sus interacciones. Para ello, se hará uso de la herramienta *Random forest,* como un modelo de entrada-salida, haciendo una descomposición ANOVA, para ver las principales variaciones e interacciones.

Teniendo en cuenta la extensión temporal de este proyecto, no se podrá hacer un análisis extensivo de la herramienta, sino que se seleccionarán modelos sencillos que pero que den una respuesta completa al problema que se ha querido analizar. Dentro de estos modelos seleccionados, sí que se hará un análisis en profundidad, para poder sacar conclusiones válidas acerca del tratamiento de datos de la herramienta *random Forest.* Para ello, se variarán los parámetros fundamentales de la herramienta (número de árboles y el número de variables seleccionada en cada nodo) para observar cómo afecta su variación a los resultados.

Sin embargo, también es cierto que el análisis de la herramienta a través del análisis ANOVA, es un estudio novedoso del que todavía no se han realizado muchas investigaciones. Por lo tanto, los resultados van a resultar muy interesantes para explicar el funcionamiento de la herramienta.

Por otro lado, para poder realizar el tratamiento de datos a través del análisis ANOVA, el otro objetivo, era poder generar datos que cumplieran las condiciones específicas (como la ortogonalidad). Por lo tanto, este objetivo es fundamental para poder sacar las conclusiones necesarias a través de un análisis ANOVA, puesto que si no hubiera sido imposible de realizar el mismo.

1.3 Metodología

La metodología que se ha seguido a lo largo de este trabajo se puede resumir en los siguientes pasos:

1. Definición del modelo de simulación.
2. Programación del modelo en R.
3. Simulación del modelo y análisis de los resultados, comprobando que no se hayan cometido errores.
4. Conclusiones y contraste con los resultados esperados.

Los modelos utilizados van aumentando en complejidad, sin llegar a resultar excesivamente complejos de manera que resulte muy difícil interpretar los resultados.

Capítulo 2: Herramientas Empleadas

2.1 Introducción

En el siguiente apartado, se hará una explicación de las herramientas que se han utilizado a lo largo del trabajo. La comprensión de ambas será clave para poder analizar los resultados que se van a obtener.

2.2 CART, Classification and Regression Trees

Como se comentó brevemente anteriormente, los Árboles de Clasificación y Regresión (CART), son una herramienta estadística de tratamiento de datos, creada por Breiman, Friedman, Olshen y Stone en 1984, por lo que se trata de una técnica bastante moderna. Estos árboles, forman un modelo, que será capaz de predecir datos a partir de un algoritmo de decisión binario. Existen dos tipos de árboles como se indica en su nombre: Clasificación y Regresión. El tipo de árbol que se utilizará dependerá de la naturaleza de la variable de respuesta; siendo los de *regresión* utilizados para variables de salida continúas (numérica), y en el caso contrario, cuando se quieran categorizar datos, se utilizarán los árboles de *clasificación.*

Los CART, por lo tanto, serán una herramienta alternativa a los modelos tradicionales de regresión y de clasificación. Actualmente, es una herramienta utilizada en el data mining(minería de datos), machine learning y varios campos de la estadística.

2.2.1 Formación de los Árboles

Dado un conjunto inicial de datos, X_1, X_2, ⋯, X_n, el árbol se irá creando realizando una serie de particiones binarias. Estas particiones, serán lo que se van a denominar *nodos,* y van a ir dividiendo el espacio en dos regiones en función de una de las variables, con el fin de clasificar los datos. Al final del proceso, se obtendrá un árbol completo.

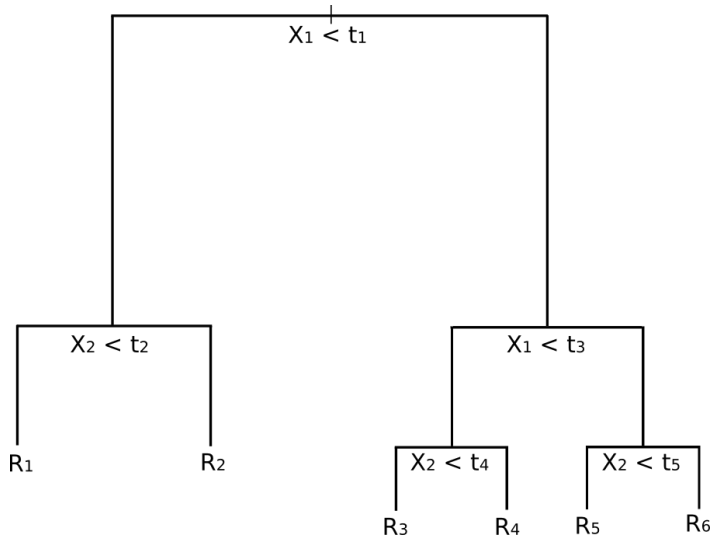


Ilustración 8 Ejemplo de Árbol de Decisión

Se puede establecer por tanto, que la creación del árbol de decisión se basa en tres principios:

- La selección de los *splits*, es decir, cúando y como partimos un nodo.
- El momento en el que decidimos parar de hacer *splits*, y establecer un nodo como terminal.
- La asignación de una categoría a cada uno de los nodos.

Por lo tanto, se deberán de establecer una serie de preguntas para realizar la partición binaria, un criterio para determinar la bondad de la partición y un criterio para determinar cuándo parar de realizar particiones. Para determinar la bondad, es decir, lo buena que es la partición a la hora de clasificar los datos, se tendrá una función llamada *Función de Impuridad (Impurity Function)* que se comentará en el siguiente apartado.

2.2.2 Función de Impuridad

A la hora de realizar las particiones en el árbol de decisión, en cada nodo se buscará realizar el *Split*, en la variable que reduzca al máximo la "impureza" del nodo. La suma de todas estas "purezas" en cada nodo, será lo que se denominará la *Función de Impureza (Φ)*.

Definiendo $\mathbf{p_{tk}}$ como la proporción de observaciones de clase K en el nodo m:

$$p_{tk=\frac{1}{N_t}\sum_{x_{i\in R_k}} I(y_i=k)}$$

Existen tres funciones de Impureza principales utilizadas en árboles de decisión:

- Tasa de error de clasificación: $1 - \max p_{tk}$
- Índice de Gini: $\sum_{k=1}^{k} p_{tk}(1 - p_{tk})$
- Entropía: $-\sum_{k=1}^{k} p_{tk}\, log\ \ p_{tk}$

La construcción del CART, se realizará de manera que se alcance la mínima impuridad nodal, para que se consiga un árbol lo más puro posible. También es importante de establecer un criterio de parada, puesto que sino, se seguiría construyendo el árbol hasta conseguir un 100% de pureza, algo que nos llevaría a tener un árbol de un tamaño inmanejable.

Estas tres funciones se pueden utilizar como medida de la impuridad nodal, tanto en la construcción del árbol como en la poda del mismo. No obstante, se suelen emplear el índice de Gini y la entropía para el crecimiento del árbol y la tasa de error de clasificación en la poda del mismo.

2.2.3 Ventajas de los CART

Las principales ventajas de los árboles de decisión son:

- Los árboles de decisión no son paramétricos, por lo que no están condicionados por el hecho de que los datos de entrada tengan un tipo de distribución específica.
- No requieren una especial preparación de los datos de entrada, ya que no se hacen distinciones en el tipo de distribución.
- Son muy fáciles de interpretar y de entender.
- Permiten que falten valores en alguna de las variables.
- Son capaces de tratar con grandes bases de datos formadas por un gran número de variables.

2.3 Random Forest

Los *Random forest*, son bosques de decisión aleatorios formados por un conjunto de árboles de decisión como los que se ha explicado anteriormente. Estos bosques, se forman mediante un algoritmo que introduce una aleatoriedad para reducir la correlación entre los árboles. Una vez construido el bosque, se utilizará para realizar la predicción.

2.3.1 Algoritmo de formación de Random Forest

Cada árbol de decisión se formará de la siguiente manera:

1. Teniendo un conjunto ***N*** observaciones diferentes, se elegirá una muestra ***N*** aleatoria con *reemplazamiento.* Esta técnica, recibe el nombre de ***bootstrapping*** y es una técnica utilizada en varios algoritmos de machine learning. Además, introduce aleatoriedad al algoritmo, ya que cada árbol se forma de manera ligeramente distinta.

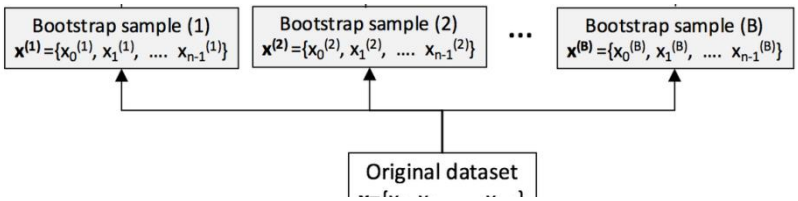


Ilustración 9 Técnica Boostrapping

2. Dadas las ***M*** variables de entrada, en cada nodo se seleccionarán de forma aleatoria ***p<<M*** variables. Este número ***p***, será constante en todo el proceso de formación del árbol e introducirá el segundo elemento de aleatoriedad en el algoritmo.

3. Se dejará crecer el árbol, sin podar hasta la máxima extensión posible.

Como se ha comentado, en la formación del forest se introduce una aleatoriedad para reducir la varianza en el modelo. La aleatoriedad, reducirá la correlación entre árboles, ya que en la formación de los mismos cada uno partirá de una muestra ligeramente distinta y en cada nodo la selección de las variables también será diferente.

La formación del forest, será igual si estamos tratando un problema de clasificación que uno de regresión. Una vez construido el bosque, se utilizará para realizar la predicción, siendo esta la media entre las predicciones de cada árbol en el caso de un problema de regresión. En el caso de un problema de clasificación la predicción será la clase más votada entre todos los árboles del bosque.

Por lo tanto, se puede establecer que los random forest dependen de dos parámetros fundamentales:

- Ntree: Número de árboles que forman el bosque.

- Mtree: Número de variables ***p*** que se seleccionan en cada nodo.

Se puede establecer que la tasa de error de los random forest está relacionada con estos parámetros. Al reducir el número ***p*** de variables, se reduce la correlación entre los árboles ya que en cada nodo se tienen menos posibilidades entre las que elegir. Sin

embargo, al reducir ***p***, también se reduce la precisión del árbol. Por lo tanto, en la práctica el valor de Mtree dependerá del problema. Al disminuir la correlación entre árboles, disminuirá la varianza, y por lo tanto, más preciso será el árbol.

Sus valores recomendados son:

- $\sqrt{p}$ para un problema de clasificación.
- $p/_3$ para un problema de regresión.

El número de árboles, Ntree, también tiene efecto en la precisión de la predicción. Como es lógico, a mayor número de árboles mejor será la predicción, puesto que el número de datos para hacer el promedio es mayor. Sin embargo, existe un valor para el cual, el error ya no disminuye y se estanca, aumentando solo el tiempo del algoritmo.

2.3.2 OOB Out of Bag Error

El *Out of bag error* (OOB) es una medida de error aplicada a modelos que utilizan la técina del ***bootstrapping.***

En la elección de los N datos con reemplazamiento, utilizando la técnica de bootstrapping, alrededor del 36% de los datos nunca son seleccionados. Por lo tanto, el OOB error, representa el error de predicción cometido por el bosque cuando se tienen en cuenta este conjunto de variables que han quedado "fuera de la bolsa".

Como se comentó anteriormente, el OOB disminuye al aumentar el número de árboles, llegándose a un valor asintótico para un elevado número de árboles. No obstante, al llegar a un número determinado, el OOB, se estanca y ya no disminuye más al aumentar el parámetro del número de árboles y puede producir problemas como el *overfitting*, que se explicará a continuación.

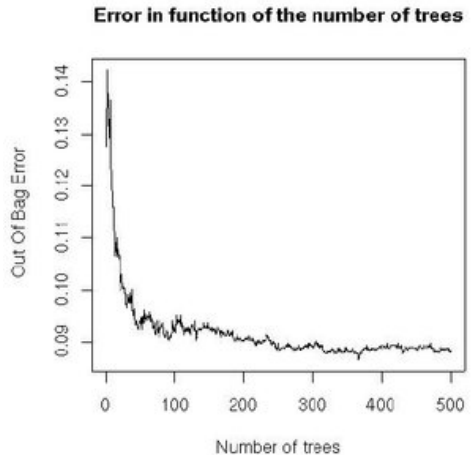


Ilustración 10 Ejemplo OOB error

2.3.3 Overfitting

El *overfitting*, o sobreajuste como es conocido en castellano, es un término muy usado en estadística y en machine learning. Este problema ocurre cuando un algoritmo es capaz de realizar una buena predicción con los datos de partida, pero pierde mucha precisión con datos diferentes a la muestra inicial.

En el caso de los CART, este es un problema que se puede solucionar realizando la poda de los árboles. En el caso de Random Forest, es más improbable que suceda este fenómeno ya que al introducir la aleatoriedad en la formación de los árboles hace que estos sean diferentes entre sí. Sin embargo, es posible que eligiendo un valor incorrecto de *Ntree* y *Mtree*, produzca un cierto sobreajuste.

2.3.4 Validación Cruzada

La validación cruzada, o ***Cross validation***, es un método de evaluación en modelos en los que se tratan un gran número de datos para establecer la independencia de los resultados (predicción) de la partición inicial de los datos en el *training* y *test set* (entrenamiento y prueba). Es un método muy utilizado en *Machine learning*, sobre todo el algoritmos de aprendizaje supervisado para garantizar la independencia de los resultados de la partición inicial en los datos de entrada.

El método ***holdout***, es el método de validación cruzada más simple. Este método se basa en dividir los datos en un set de entrenamiento (training) y en un set de prueba (test). Se utilizará sólo el set de entrenamiento para construir el modelo de predicción. Una vez construido el modelo, se hará una validación con el set de datos de prueba. Estableciendo su precisión realizando un comparación entre los datos predichos y los datos reales.

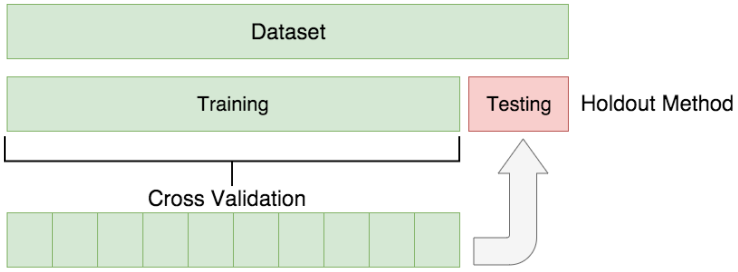


Ilustración 11 Hold Out Method

Sin embargo, el problema que surge con este método, es la elevada variancia que tiene, puesto que el resultado puede resultar muy diferente dependiendo de la división inicial de los datos. Para corregir el problema de la dependencia de la división inicial de los datos, existen otros métodos de *cross validation:*

- ***K-fold cross validation***: En este método se dividen los datos en ***K*** subconjuntos y se repite el *holdout method* ***k*** veces de manera que en cada iteración, se escoge uno de los ***K*** suconjuntos como el set de prueba y los ***K-1*** restantes, se utilizan como set de entrenamiento. La ventaja de este método, es que el resultado depende menos de la división que se hace de los datos en los sets de entrenamiento y prueba. Una vez realizados todos los modelos, se obtendrá el resultado final como la media aritmética de las ***K*** iteraciones.

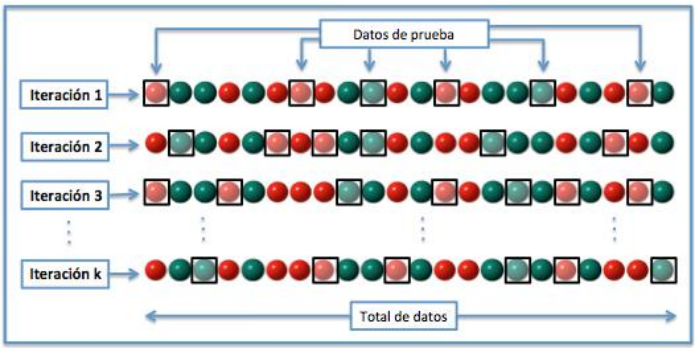


Ilustración 12 K-fold Cross Validation

- ***Aleatory Cross Validation:*** En este método, se divide en cada iteración de forma aleatoria los datos en test set y training set. De esta manera existe la posibilidad de que se solapen los subconjuntos en las diferentes iteraciones. Como en el método anterior, el resultado final será el promedio de todas las iteraciones. Como ventaja con respecto al modelo anterior se tiene además, que el número de iteraciones es menor que el número de subconjuntos.

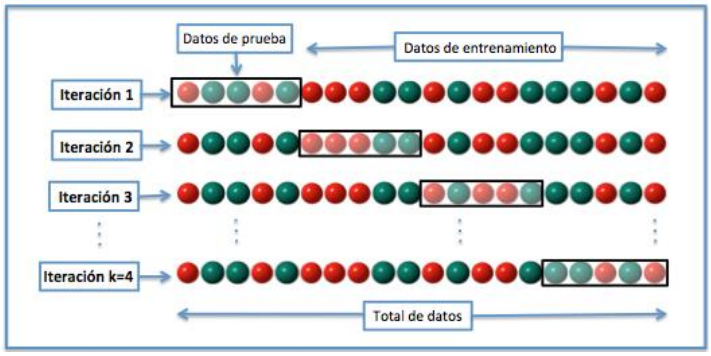


Ilustración 13 Validación Cruzada Aleatoria

Estos dos métodos, como se ha comentado anteriormente, tienen mejores resultados que el *holdout method*. Sin embargo, cabe destacar sus mayores costes debido a que el tiempo de computación es más elevado. Por ello, se elegirá un método u otro dependiendo del problema que se esté tratando.

2.3.5 Importancia de las variables

El concepto de la importancia de las variables, es muy utilizado en modelos tanto de regresión como de clasificación y otros modelos de machine learning. No obstante, es un concepto difícil de explicar debido su complejidad y difícil medición.

Se podría resumir este concepto en ver cómo afecta a la salida del modelo cuando se realizan cambios en las variables de entrada. Las variables de entrada que más variabilidad produzcan en la salida, serán aquellas que más influencia tengan y por tanto, aquellas que mejor explicarán el modelo y serán más importantes. Para realizar la medición de esta importancia, se van a tener dos maneras de realizarlo:

- **Reducción de la impuridad nodal media:** En cada árbol a la hora de realizar el *Split,* se mide la reducción en la impuridad que contribuye la variable elegida en la partición. Haciendo la media sobre todos los árboles, de todas las reducciones de todas las variables, se saca la media de este valor de impuridad, que como se mencionó anteriormente, se utiliza la medida del índice de Gini. La variable que más reduzca la impuridad del forest, será la más importante.

- **Incremento del Mean Squared Error (MSE):** El MSE, traducido al castellano como el error cuadrático medio (ECM), es una medida muy utilizada en estadística como medida de la calidad de un estimador. Para todas las respuestas, se calcula el error de la siguiente manera, denotando y como la respuesta real e $\hat{y}$ como la variable predicha.

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

En el caso de *random forest*, para medir la importancia de las variables, para cada árbol que forman el bosque, se calcula el MSE con las variables del OOB, que son las que han quedado fuera. A continuación, se realiza una permutación aleatoria en una de las variables de entrada. Esto produce un cambio en el en las y de salida que a su vez produce una cambio en el MSE de las variables que han quedado fuera.

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

Se realizará este proceso con todas las variables de entrada y en todos los árboles del bosque, permutando aleatoriamente. Para cada variable permutada, se compara el valor del MSE antes y después de permutar. Esta diferencia, para cada árbol se suma,

normaliza y se hace el promedio. Cuanto mayor sea este valor, más importante será la variable permutada.

Este proceso, es el propuesto por Breiman, en el paper sobre *random forest* publicado el 2002.

Entre los valores de medida de la importancia de las variables, el método utilizado más comúnmente, es la medida del incremento del MSE, ya que supone un tiempo computacional menor que el cálculo de la reducción de la impuridad nodal.

En este trabajo se va a tratar de explicar la importancia de las variables a través de un análisis ANOVA.

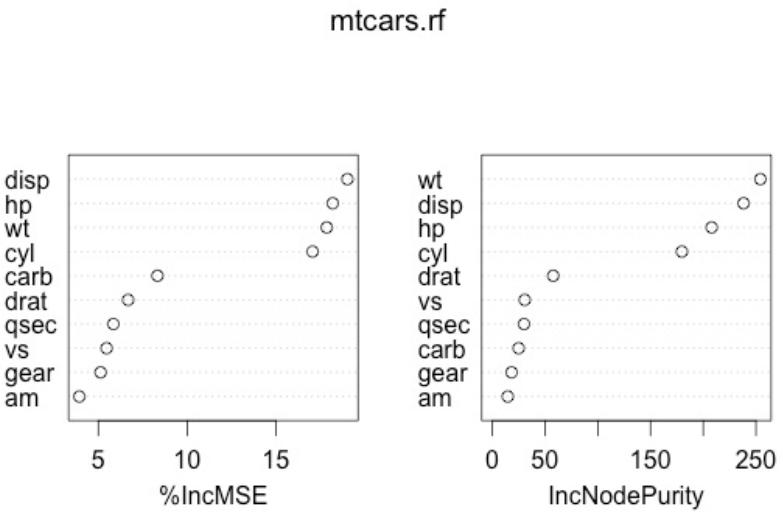


Ilustración 14 Ejemplo de análisis de la importancia de variables en R

2.4 Análisis ANOVA

El análisis ANOVA, *Analysis of Variance* en inglés (análisis ADEVA, en castellano) como su nombre indica, es un método estadístico para analizar datos experimentales. Es una técnica importante para analizar el efecto de los factores categóricos y sus interacciones en una respuesta. Un ANOVA descompone la variabilidad de la variable de respuesta entre los diferentes factores. Dependiendo del tipo de análisis, puede ser importante determinar:

- Qué factores tienen un efecto significativo en la respuesta.
- Qué parte de la variabilidad en la variable de respuesta es atribuible a cada factor.

El análisis de la varianza parte de los conceptos de regresión lineal. Encontramos diferentes tipos de análisis ANOVA:

- ANOVA unidireccional

Se utiliza un análisis unidireccional de la varianza cuando los datos se dividen en grupos según un solo factor. Se proporcionan pruebas estadísticas para comparar las medias de los grupos, las medianas de los grupos y las desviaciones estándar de los grupos. Cuando se comparan medias, se utilizan pruebas de rango múltiple, la más popular de las cuales es el procedimiento HSD de Tukey. Para muestras de igual tamaño, las diferencias significativas entre grupos pueden determinarse examinando el gráfico de medias e identificando los intervalos que no se superponen.

- ANOVA Multifactor

Cuando hay más de un factor presente y los factores están cruzados, un ANOVA multifactor es apropiado. Tanto los efectos principales como las interacciones entre los factores pueden estimarse como parte de esta prueba ANOVA. El resultado incluye un gráfico de interacción, que muestra la respuesta media estimada en cada combinación de 2 factores.

El análisis ANOVA se basa en estos tres conceptos:

- Distribución normal de los datos
- Muestras aleatorias simples independientes
- Variación constante

Para un análisis ANOVA unidireccional, la hipótesis será la siguiente:

- H0: todas las medias son iguales
- HA: no todos las medias son iguales

Cálculo de la relación F

ANOVA separa la variación en el conjunto de datos en 2 partes: entre el grupo y dentro del grupo. Estas variaciones se denominan sumas de cuadrados, que se pueden ver en las siguientes ecuaciones.

1) Variación entre grupos

La variación entre grupos (o sumas de cuadrados entre grupos, SS) se calcula comparando la media de cada grupo con la media global de los datos.

Específicamente, esto es:

$$SS = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$$

es decir, sumando el cuadrado de las diferencias entre la media de cada grupo y la media general de la población, multiplicada por el tamaño de la muestra, suponiendo que estamos comparando tres grupos (i = 1, 2 ó 3).

Luego dividimos el SRS por el número de grados de libertad [esto es como el tamaño de la muestra, excepto que es n-1, porque las desviaciones deben sumar a cero, y una vez que se conoce n-1, también se conoce el último] para obtener nuestra estimación de la variación media entre grupos.

2) Variación dentro de los grupos

La variación dentro del grupo (o las sumas de cuadrados dentro del grupo) es la variación de cada observación de su media de grupo.

$$SS_R = s^2_{group1}(n_{group1} - 1) + s^2_{group2}(n_{group2} - 1) + s^2_{group3}(n_{group3} - 1)$$

es decir, sumando la varianza de los tiempos de cada grupo por los grados de libertad de cada grupo.

Como antes, entonces dividimos por los grados totales de libertad para obtener la variación media dentro de los grupos.

3) La relación F

La relación F se calcula entonces como:

$$\frac{\text{Mean Between-group SS}}{\text{Mean Within-group SS}}$$

Si la diferencia promedio entre los grupos es similar a la que hay dentro de los grupos, la relación F es de aproximadamente 1. A medida que la diferencia promedio entre los grupos se vuelve mayor que la que hay dentro de los grupos, la relación F se vuelve mayor que 1.

Para obtener un valor P, se puede probar contra la distribución F de una variable aleatoria con los grados de libertad asociados con el numerador y el denominador de la relación. El valor P es la probabilidad de obtener esa relación F o una mayor. Las relaciones de F más grandes dan valores de P más pequeños.

Todos estos datos quedan recogidos en la siguiente tabla:

Source of Variation	**d.f.**	**SS**	**MS**	**F_0**
Factor A (between groups)	a-1	$SSA = \sum_{i=1}^{a} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$	$MSA = \frac{SSA}{(a-1)}$	$\frac{MSA}{MSE}$
Factor B (between groups)	b-1	$SSB = \sum_{j=1}^{b} n_j \left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$	$MSB = \frac{SSB}{(b-1)}$	$\frac{MSB}{MSE}$
Error (within groups)	(a-1)(b-1)	$SSE = SST - SSA - SSB$	$MSE = \frac{SSE}{(a-1)(b-1)}$	
Total	N-1	$SST = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{..}\right)^2$		

Ilustración 15 Ejemplo de tabla ANOVA

2.5 MAPE y MAE

El MAPE, *Mean Absolute Percentage Error*, es otra herramienta estadística para medir el error que se puede producir en un modelo al realizar una estimación o predicción.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

Donde y_i es la variable de salida real y $\hat{y}_i$ es la variable predicha por el modelo de *machine learning*. Sin embargo, el MAPE, tiene una serie de restricciones que pueden hacer que de un valor de error no real.

- Imposibilidad de dividir entre cero: No puede ser utilizado cuando trata datos con valores nulos o cercanos al cero, puesto que se estaría dividiendo entre cero. Esto, dispararía el valor del MAPE, aunque en realidad la predicción realizada sea buena.

- Inexistencia de límite superior: Para valores relativamente elevados, se pueden realizar errores superiores al 100%, lo que haría muy difícil la tarea de medir los errores.

- Cuando se utiliza el MAPE para seleccionar un modelo, sistemáticamente, se decantará por seleccionar aquel cuyas predicciones son más bajas, incluso para el mismo nivel de calidad.

Debido a estas restricciones, y teniendo en cuenta que en muchos de los modelos utilizados en este trabajo se utilizarán distribuciones N (0,1), se hará uso de la medida de error del MAE, *Mean Absolute Error*, para evitar usar valores irreales MAPE.

El MAE, se calcula de la siguiente manera:

$$MAE = \frac{100}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

De esta manera, evitamos dividir por valores cercanos al cero que nos darían valores erróneos de la medida del error. Se utilizará por tanto el MAE a lo largo de los experimentos de este trabajo como medida del error.

2.6 Paquete estadístico R

Para realizar los distintos análisis de este Trabajo de Fin de Grado, se ha hecho uso de la herramienta estadística de R.

El lenguaje de programación de R, es un lenguaje de programación utilizado en el campo de investigación estadística en temas de minería de datos, biomedicina, bioinformática y las matemáticas financieras.

R es parte del sistema *GNU Project*, por lo que forma parte de un proyecto colaborativo. Además, está formado por un amplio conjunto de *paquetes* que permiten utilizar una amplia gama de herramientas estadísticas como modelos lineales, no lineales, análisis de series temporales, algoritmos de clasificación…

Otra gran herramienta es la posibilidad de realizar gráficas muy completas. Todas las librerías se obtienen de la página web oficial de *CRAN projects*, sin embargo se ha utlizado ***Rstudio*** que es una *IDE (Integrated Development Environment)* que facilita el manejo del entorno de R.

2.6.1 Funciones más utilizadas

A la hora de desarrollar el código de R para simular el modelo, se van a utilizar un gran número de funciones del paquete R. A continuación, se van a explicar de manera simplificada las funciones más utilizadas en el código empleado en este trabajo para facilitar su explicación.

- *randomForest():* Dado un conjunto de datos formados por las variables de entrada y sus correspondientes salidas, esta función ajusta un modelo de random forest a estos datos. Se debe especificar cuál es la variable dependiente (variable de salida) y conviene que los datos de entrada sean aleatorios.

- *rnorm.Sobol():* Devuelve una serie de variables aleatorias con distribución normal dentro de una secuencia determinada. Se debe especificar el número de variables y observaciones que se desean.

- *Predict():* Función que predice un valor de un modelo dado. Se deberá de determinar el modelo sobre el que se realizará la predicción.

- *Chol():* Devuelve la matriz A de la descomposición de Cholesky.

- *aov():*Devuelve el análisis ANOVA (análisis de la varianza) de un modelo. Se debe especificar la variable dependiente para poder observar las interacciones y efectos en el análisis en variabilidad.

Capítulo 3

Experimentos con Modelo de Regresión Múltiple

3.1 Introducción

En este capítulo se van a exponer los experimentos realizados, se analizarán los resultados y se sacarán conclusiones sobre el modelo. Como el objetivo de este TFG es analizar la importancia de cada variable del modelo y la variabilidad que introduce cada una de estas a la respuesta del mismo, para realizar este análisis primero se ha tomado el modelo como un modelo lineal parametrizado realizando un análisis ANOVA. A continuación se ha querido realizar el mismo análisis pero en este caso con un modelo no paramétrico como es el modelo *random Forest*. Para sacar conclusiones a cerca de la herramienta *random Forest*, se han ido variando los parámetros fundamentales del mismo.

Por lo tanto, en este primer capítulo se tomará el modelo lineal de regresión múltiple que será un modelo sencillo de analizar pero, que sin embargo al ser un modelo sin interacciones entre sus variables, no nos permitirá observar el tratamiento de ambos análisis cuando si existan. Con lo cual, para observar la diferencia entre ambos resultados cuando sí existan interacciones, se va a estudiar en el próximo capítulo un modelo más complejo con interacciones.

3.2 El modelo

El primer modelo que se va a utilizar, va a ser un modelo muy utilizado en estadística como es modelo de regresión simple múltiple. Se comenzará con este modelo, ya que es un modelo del que se conocen muchas propiedades, y matemáticamente es un modelo bastante simple en cuanto a la relación de la variable de salida con las variables de entrada.

Dentro de este modelo, se harán experimentos para un modelo *no determinista*, es decir, un modelo sin ruido, y un *modelo determinista* en el que sí encontramos un término determinista, que introducirá una mayor complejidad a la hora de interpretar los resultados.

El modelo de regresión múltiple sin término determinista, sigue la siguiente ecuación:

$$y = \sum_{i=0}^{n} \beta_i x_i$$

En cambio, en el modelo determinista, se introduce un término aleatorio al que llamaremos ruido:

$$y = \sum_{i=0}^{n} \beta_i x_i + \varepsilon_i$$

En primer lugar, se debe de tener en cuenta que como se van a comparar distintos experimentos, es muy importante que todos los datos generados pertenezcan a la misma secuencia de datos. Esto es necesario para poder realizar comparaciones entre experimentos, algo que resultaría imposible si cada vez que se realiza una iteración del experimento, tenemos datos muy diferentes imposibles de ser comparados.

A la hora de crear las variables del modelo, se van a tener que tener en cuenta una serie de especificaciones para que el modelo sea válido para realizar un posterior análisis ANOVA.

En el capítulo dos, al explicar la teoría detrás del análisis ANOVA, se vieron que este análisis seguía una serie de asunciones sobre las variables del modelo. Estas eran las siguientes:

1) La salida del modelo, tiene una distribución normal
2) Las muestras aleatorias son independientes
3) Tienen varianza constante.

El punto dos es fundamental, puesto que para poder realizar un análisis ANOVA es necesario que las variables sean independientes entre ellas, por lo tanto, ortogonales. El porqué de esta ortogonalidad se explica en que el análisis ANOVA, establece que la variabilidad total del modelo es la suma de las variabilidades de cada factor. Esto sólo ocurre cuando las variables son independientes entre ellas, y por lo tanto ortogonales. De esta manera, conseguir la ortogonalidad de las variables del modelo va a ser fundamental para poder desarrollar el análisis ANOVA.

En este trabajo, se ha hecho uso de dos maneras diferentes de obtener variables ortogonales entre sí. Bien es cierto, que el primer método que es el que se va a utilizar en este primer capítulo no se obtienen variables 100% ortogonales, sin

embargo, en un modelo sin interacciones como es este primer modelo, el error cometido será tan pequeño que se podrá aceptar como válido y al no haber interacciones no se tendrán problemas. En el siguiente capítulo donde se va a tratar un modelo con interacciones, se hará uso del otro método (método de codificación) en el que si obtendremos variables completamente ortogonales.

El primer método que se utilizará para obtener el modelo con variables ortogonales, será el método utilizando una función sobol.

La función *sobol sequence* nos permite generar variables pseudoaleatorias con una distribución y secuencia determinada. En nuestro caso, se ha decidido para estos experimentos, utilizar una distribución $\boldsymbol{N}$ (0,1). Realizando un histograma de las variables generadas, se puede observar la distribución normal querida:

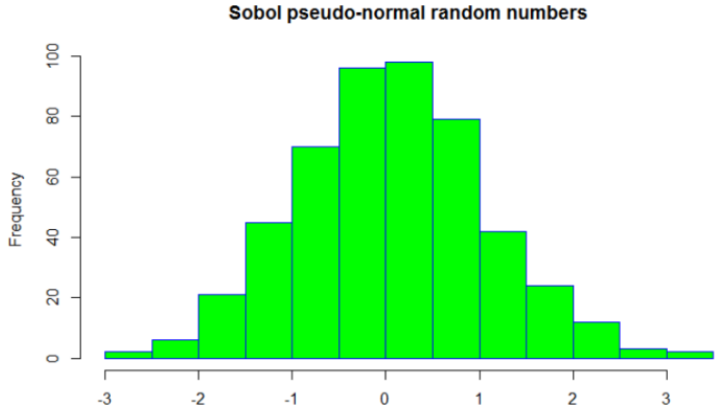


Ilustración 16 Distribución variables modelo

Sin embargo, estas variables no son independientes entre sí (ortogonales), y como es necesario para poder realizar el análisis ANOVA, se va realizar lo siguiente para ortogonalizar:

1. Se establece la matriz de covarianzas $\sum = \begin{bmatrix} \sigma_{11}^2 & \rho\sigma_{12}\sigma_{21} \\ \rho\sigma_{21}\sigma_{12} & \sigma_{22}^2 \end{bmatrix}$

   En esta matriz de covarianzas, se va a establecer como una matriz en la que las correlaciones entre variables son nulas, es decir, $\rho = 0$. Por lo tanto, vamos a obtener una matriz diagonal de unos.

2. Realizamos la descomposición de Cholesky en la matriz de covarianzas

$$\sum = AA^T$$

   Siendo A, la matriz triangular inferior $A = \begin{bmatrix} \sigma_{11}^2 & 0 \\ \rho\sigma_{21}\sigma_{12} & \sigma_{22}^2 \end{bmatrix}$

   Recordando que la matriz tiene A debe ser una matriz diagonal de unos.

3. Las variables X con distribución normal (μ,∑) quedarán de la siguiente manera:

$$X = \mu + Ax$$

Siendo x, una variable de distribución N (0,1)

Con este algoritmo, se va obtener una variable X de media μ=0 en este caso y con una matriz de covarianzas ∑, con valores de varianza de las variables ≈ 1.

Por lo tanto, para este modelo sin interacciones, esta aproximación va a ser aceptable.

Con este algoritmo por tanto, obtenemos las variables de entrada que utilizaremos para crear la salida del modelo. Este algoritmo, será el mismo para todos los modelos de este capítulo.

Por último, cabe señalar que para la variable determinista del ruido, se ha generado una variable aleatoria de distribución normal N (0,1).

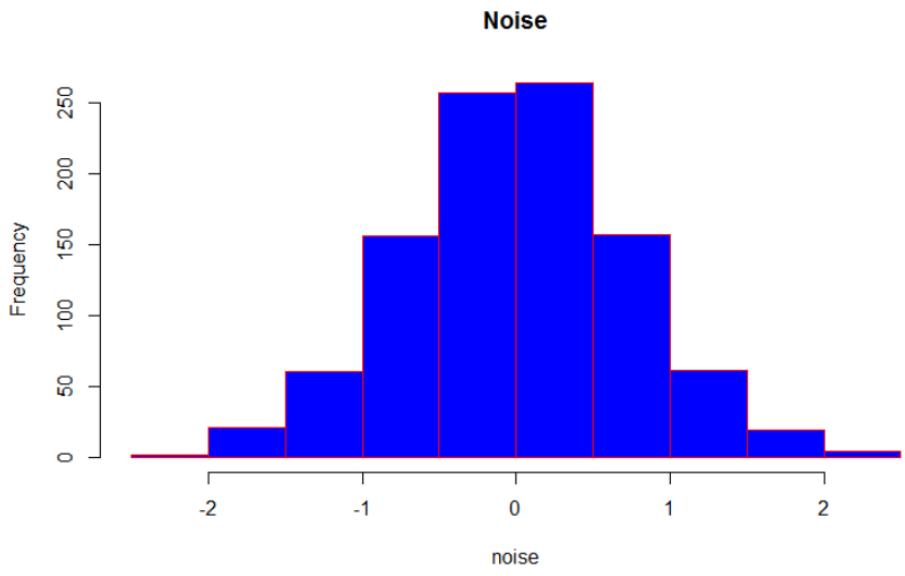


Ilustración 17 Distribución de variable ruido

3.3 Procedimiento

En este capítulo como se ha explicado se van a crear las variables ortogonales mediante el método sobol. Una vez creado el modelo, se va a analizar la variabilidad en la respuesta debida a los factores e interacciones. Para ello, se va a estudiar el modelo mediante un diseño paramétrico y a continuación realizaremos el diseño no paramétrico como es el *random forest*. A ambos modelos se les realizará un análisis ANOVA para obtener los resultados.

El procedimiento a seguir en cada experimento será el siguiente: En primer lugar, se generan N observaciones de p variables, como se ha indicado en el apartado anterior, generando una matriz de [N x p]. A continuación, se genera la variable de salida para las N observaciones, de acuerdo a la fórmula matemática correspondiente al modelo. En este caso la ecuación de un modelo de regresión lineal múltiple.

Con estos datos, se va a realizar un análisis ANOVA al modelo paramétrico creado. De la tabla ANOVA, se extraerá el valor de la suma de cuadrados que utilizaremos para calcular la variabilidad explicada de cada factor. Para ello, primero se va a calcular el error medio cuadrático de cada factor de la siguiente manera:

$$MS_i = \frac{SS_i}{DF_i}.$$

Con este valor se calculará la variabilidad de cada factor restando al error mínimo cuadrático el valor del error mínimo cuadrático de los residuos:

$$VE = \frac{MS_{factor} - MS_{error}}{nº\ Variables}$$

Con estos valores se hará una representación gráfica en un diagrama de barras representado los porcentajes de variabilidad de cada factor e interacción. A continuación se realizará lo mismo sobre el modelo *random forest* y se representará de la misma manera. Se procederá a comparar los resultados de ambos modelos.

Para crear el modelo *random forest*, primero se tendrá que calcular el valor óptimo de los parámetros que gobiernan la función. Estos parámetros como se ha mencionado a lo largo de trabajo son el número de árboles que forman el bosque (Ntree) y el número de variables que se seleccionan en cada split (Mtry). Para ello, se van a realizar dos experimentos en los que se van a ir variando cada uno de los parámetros. Estos experimentos se van a llevar a cabo en un bucle en el que se dejará constante uno de los parámetros, con el valor predeterminado por la función random Forest de R, y se irá variando el otro.

Para obtener el valor óptimo se va a representar de forma gráfica la variabilidad total del modelo. El mejor valor del parámetro será aquel que explique el modelo, por lo tanto se escogerá aquel valor con mayor variabilidad explicada. Para comprobar que en efectivo este es el mejor valor, se procederá a calcular también unas gráficas en las que se han calculado en cada una de ellas una medida de error. En la primera se ha representado el error mínimo cuadrático (MSE) y en la segunda se ha calculado el MAE (error mínimo absoluto). Se

deberá verificar que para el valor óptimo escogido se cumplirá que es el punto donde menor error se esté cometiendo.

Una vez que se ha calculado el valor óptimo para los parámetros de la función *random forest*, se construirá el modelo no paramétrico. Sobre este modelo realizaremos un análisis ANOVA. Se extraerán los mismos valores y se representará de la misma manera que se hizo con el modelo paramétrico. Por último, se realizará una comparación entre ambos modelos viendo que diseño explica mejor el modelo.

A continuación se expone un ejemplo del código en el que se calcula el valor óptimo de uno de los valores. El resto de código, se adjuntará en los anexos finales.

```
 for (mt in mtry)

 {  rf.Regression <- randomForest(y0 ~ ., data=trainSet,ntree=100L, mtry = mt)

  pred <- predict(rf.Regression, newdata = testSet)

 SSM <- sum((pred - mean(testSet$y0))^2)

  SST <- sum((testSet$y0 - mean(testSet$y0))^2)

  r2a <- SSM / SST

  r2nt2 <- c(r2nt2,r2a)

  mse2 <- c(mse2, MSE(y_pred =pred , y_true = testSet$y0))

  mae2 <- c(mae2, MAE(y_pred =pred , y_true = testSet$y0))}

#Cálculo varibilidad total del modelo

 r2nt2 <- 100 * r2nt2

 dfr22 <- data.frame(r2nt2,mtry)

 ggplot(data=dfr22, aes(x=mtry, y=r2nt2, group=1)) +

  geom_line(linetype = "solid", color="blue") +geom_point() +labs(x = "Number of predictors
sampled for spliting at each node", y = "r²", title = "Variability in the response explained by the
explanatory variables")

#Cálculo medida del error

  dfmae2 <- data.frame(mae2,mtry)

 ggplot(data=dfmae2, aes(x=mtry, y=mae2, group=1)) +    geom_line(linetype = "solid", color="purple") +
geom_point() +    labs(x = "Number of predictors sampled for spliting at each node", y = "MAE", title =
"MAE vs mtry")
```

3.4 Experimentos en el modelo no determinista

En este primer experimento, se tendrá el modelo de regresión múltiple formado por 10 variables de las que se obtendrán *1000* observaciones. Las variables de entrada y la variable dependiente se obtendrán como se ha mencionado en los apartados anteriores. Para este experimento se han elegido los siguientes valores para el vector beta = c (5, 5, 2,2, 1,-5,-5,-2,-2,-1). En este vector se muestran los valores para las variables X1, X2, X3, X4, X5, X6, X7, X8, X9 y X10 respectivamente.

Una vez construido el modelo, se procederá a realizar el análisis ANOVA al mismo. Como se ha explicado en el procedimiento, se calculará la variabilidad de cada factor obteniendo lo siguientes resultados:

Tabla 1 ANOVA Modelo No Determinista Paramétrico

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X1	1	19973	19973	1.468e+32	<2e-16	***
X2	1	21628	21628	1.590e+32	<2e-16	***
X3	1	6367	6367	4.680e+31	<2e-16	***
X4	1	9896	9896	7.273e+31	<2e-16	***
X5	1	8344	8344	6.132e+31	<2e-16	***
X6	1	6554	6554	4.817e+31	<2e-16	***
X7	1	17713	17713	1.302e+32	<2e-16	***
X8	1	1355	1355	9.958e+30	<2e-16	***
X9	1	2759	2759	2.028e+31	<2e-16	***
X10	1	937	937	6.886e+30	<2e-16	***
Residuals	989	0	0			

En la tabla, se puede observar en la primera columna con el nombre de Df, los grados de libertad de cada variable del modelo y de los residuos. A continuación se encuentra el término de la suma de cuadrados. Se recuerda que este término muestra la desviación del valor real con respecto al valor predicho o la media en este caso. Cuanto mayor sea este número indicará una peor predicción o ajuste del modelo con respecto de la realidad. La siguiente columna muestra el error cuadrático que resulta de dividir el valor de la suma de cuadrados entre el número de grados de libertad.
Por último, las dos últimas columnas muestra el valor de F que indica el ratio de la desviación de las medias de grupo con el ratio de la desviación agrupada dentro de la desviación de grupo. Cuanto mayor sea el valor F, mayor será la varianza relativa entre las medias del grupo. El valor de p indica la probabilidad de obtener un valor de F tan extremo o más extremo como el observado bajo el supuesto de que la hipótesis nula es verdadera.

Queda claro, que en este caso los residuos son nulos ya que del modelo que se está introduciendo en el análisis ANOVA, conocemos su entrada y su salida y no tiene interacciones.

Obteniendo los valores en porcentaje de las variabilidades de cada factor para el modelo se obtiene el diagrama de barras:

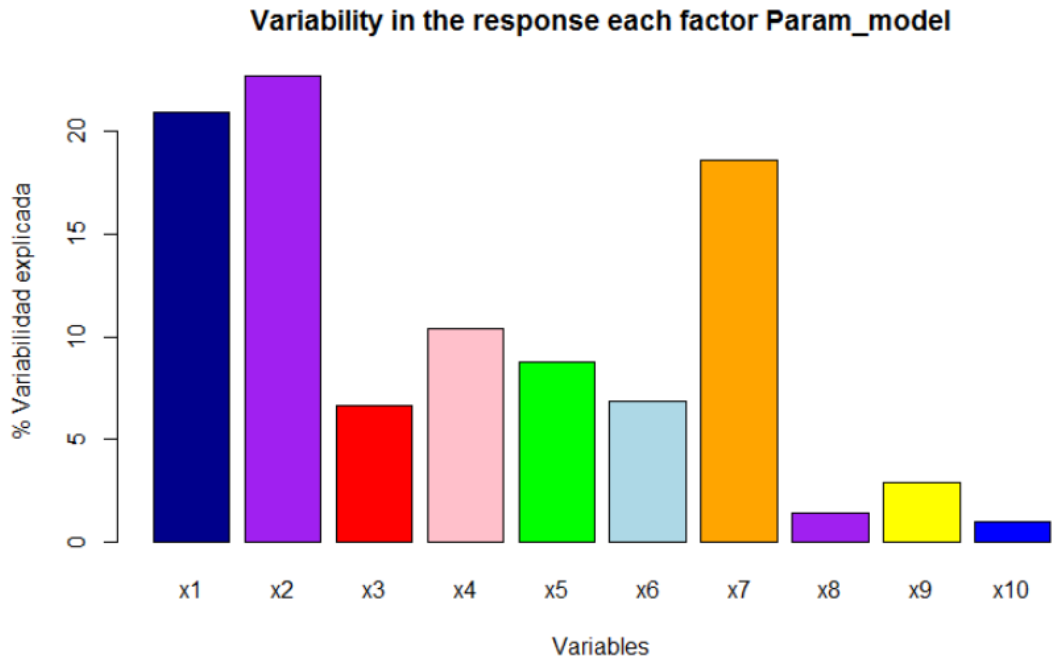


Ilustración 18 Diagrama % Variabilidad modelo Paramétrico

Variable	% Variabilidad Explicada sobre el modelo
X1	20.9%
X2	22.6%
X3	6.6%
X4	10.3%
X5	8.7%
X6	6.8%
X7	18.5%
X8	1.4%
X9	2.8%
X10	0.98%

A mayor porcentaje de variabilidad explicada de cada factor sobre el modelo, mayor importancia tiene esa variable puesto que explica mayor cantidad de la respuesta. Según los resultados obtenidos se concluye que la variable más importante es X2 y la que menos importancia tiene sobre el modelo es X10. Algo que resulta lógico viendo la ecuación que sigue el modelo.

Tras obtener estos resultados del análisis ANOVA, se va a proceder a crear el experimento sobre el *random Forest* para ver el resultado que se obtendrá al introducir un diseño no paramétrico. El objetivo de estos primeros experimentos es encontrar los valores de los parámetros del *random forest* explican mejor el modelo. Por lo tanto, en primer lugar, se irá variando únicamente un parámetro y se dejará el otro constante con el valor predeterminado por la función de R de *randomForest*, tal y como se ha explicado anteriormente.

3.4.1 Experimentos en Ntree

En estos experimentos se variará el parámetro del número de árboles que forman el bosque, desde un valor de 50 hasta el valor predeterminado en la función de R de 500 árboles. En cuanto a los demás parámetros, se mantendrán los valores predeterminados, siendo *mtree*, el número de variables que se seleccionan en cada nodo al hacer el split igual a *p/3*. En este experimento al tener 10 variables se tendrán 3 variables en cada split.

En primer lugar, se va a mostrar una gráfica que muestra el valor del número de árboles frente a la variabilidad total del modelo. De ella se obtendrá el valor del número de árboles que mejor explica el modelo en su conjunto.

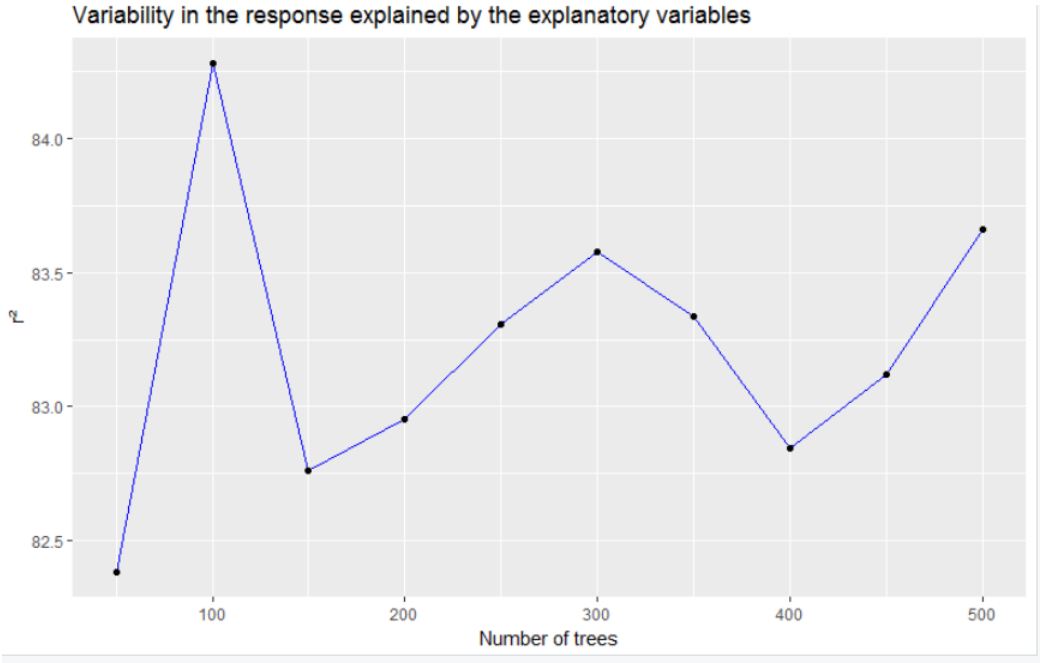


Ilustración 19 Gráfica % Variabilidad Total del modelo vs Número de Árboles

Se puede observar como con un valor de Ntree=100 árboles, se obtiene la mayor variabilidad explicada del modelo completo. Si ahora se muestra el valor del error mínimo cuadrático (MSE) y el MAE, se obtiene el mínimo valor de los mismos para el mismo número de árboles, algo que era de esperar, puesto que el punto que mejor explica el modelo será el punto en el que las predicciones son las mejores y por tanto se comete menor error.

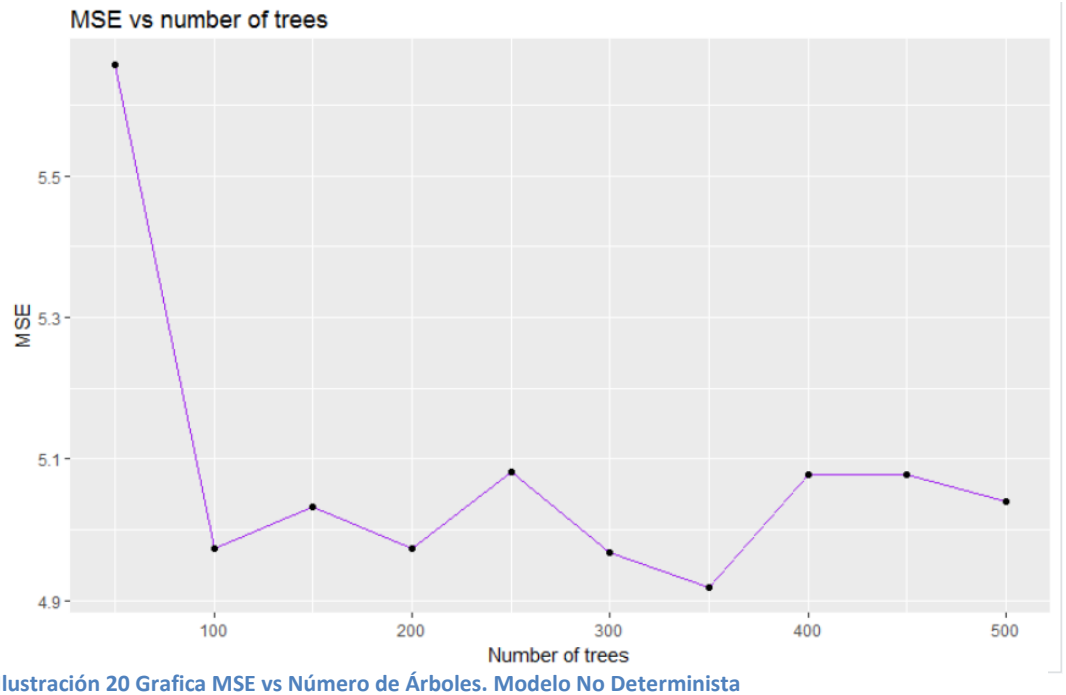


Ilustración 20 Grafica MSE vs Número de Árboles. Modelo No Determinista

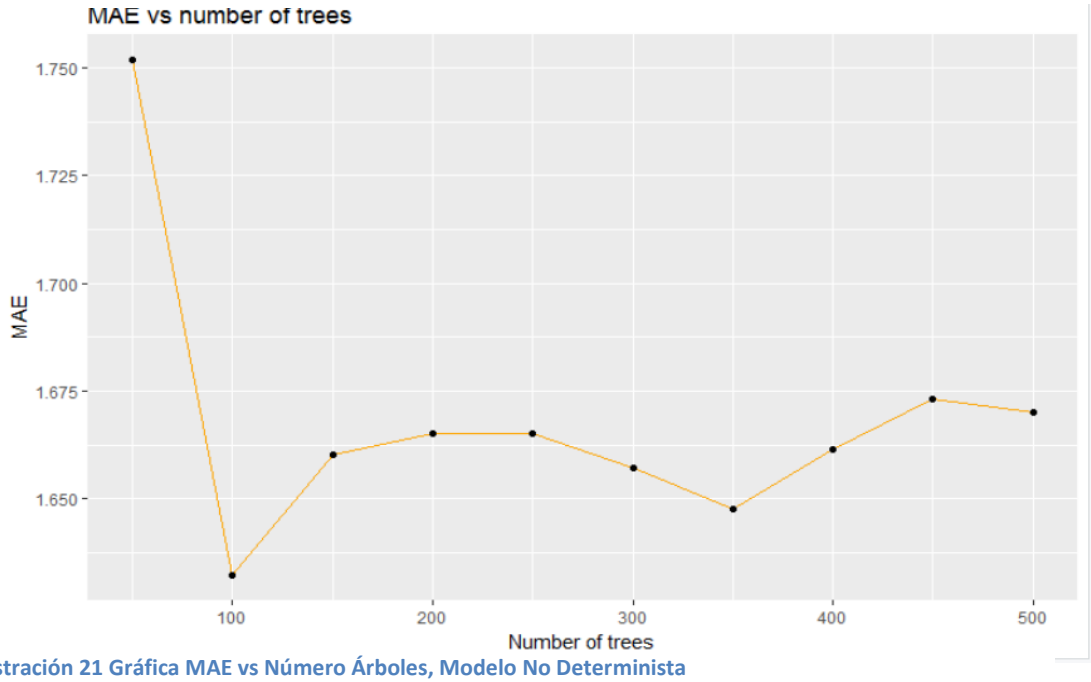


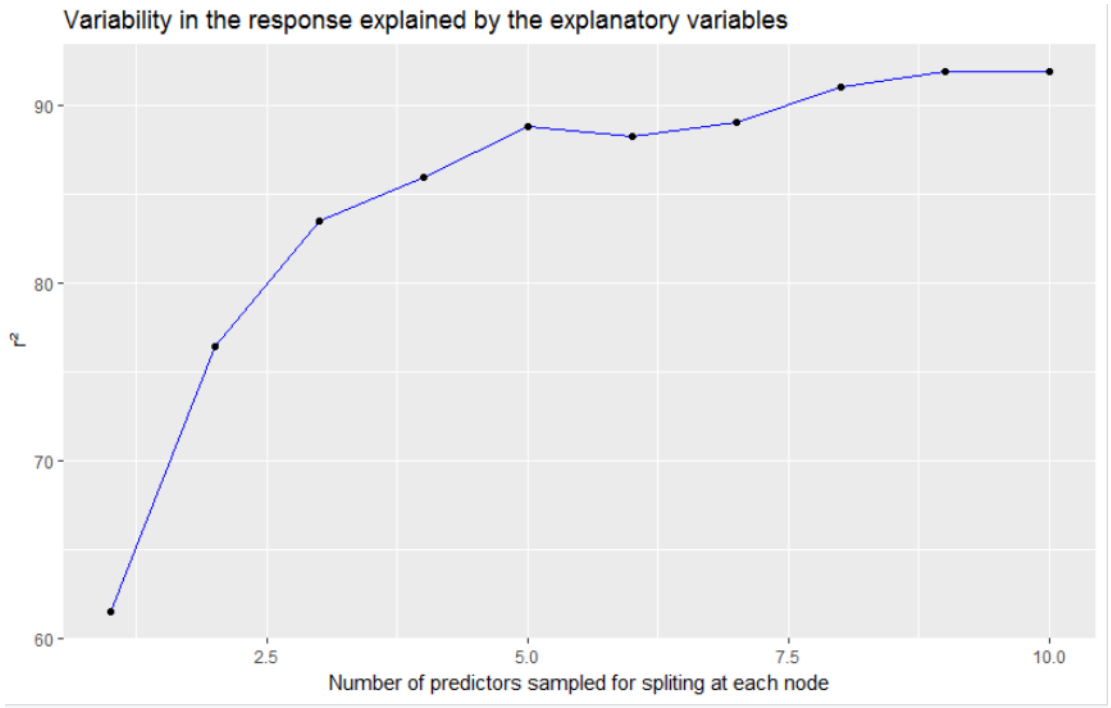
Ilustración 21 Gráfica MAE vs Número Árboles, Modelo No Determinista

3.4.2 Experimentos en Mtry

Del mismo modo que se han realizado los experimentos en el parámetro del número de árboles, se van a realizar experimentos en el parámetro del *random Forest* que elegirá el número de variables en cada partición. Para este experimento, se utilizarán las mismas variables que en el experimento anterior, es decir, se tendrán, N= 1000 observaciones de las p=10 variables anteriores.

Para estas variables, se seguirá el mismo procedimiento explicado anteriormente para generar el modelo *random Forest*. Con los mismos resultados gráficos, se explicará los resultados obtenidos. En este caso, el parámetro que se variará en el bucle el parámetro *mtry* de la función *random Forest*, desde un valor de 1, es decir, elegir sólo una variable en el split, hasta el número *p* variables del modelo.

Observando la gráfica, claramente se puede observar que cuantas más variables se seleccionen en cada nodo mejor se explicará en modelo. No obstante, al coger un número muy elevado de variables, se estaría restando al modelo su carácter aleatorio que va a ser fundamental para que se realicen predicciones correctas. Por lo tanto, el valor óptimo va a ser aquel para el cual el error no disminuye más, en este caso sería mtry= 4 variables.



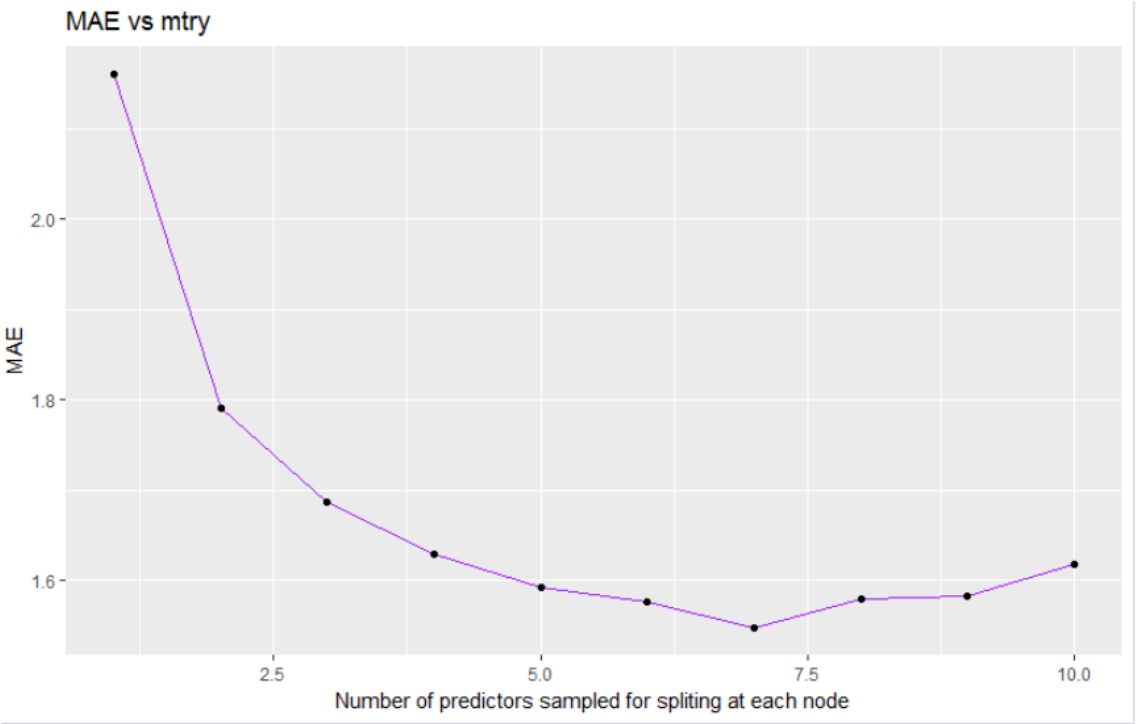


Ilustración 22 MAE vs Mtry, Modelo No Determinista

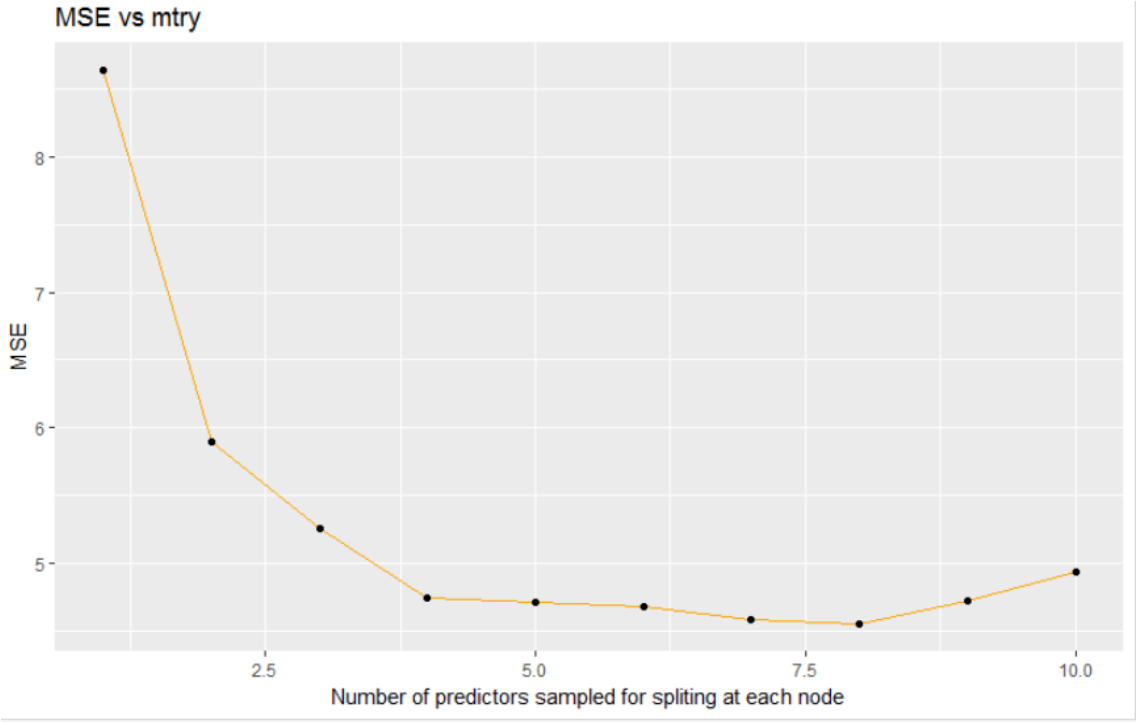


Ilustración 23MSE vs Mtry, Modelo No Determinista

Ahora, con esta pareja de valores obtenida, que es el óptimo para realizar el modelo random forest, construimos el modelo. Sobre este modelo, se realizará la predicción y con esta nueva salida del modelo y los datos de las variables de entrada generaremos un marco de datos sobre el que realizaremos el análisis ANOVA.

Los parámetros elegidos han sido:

- Ntree= 100
- Mtry=4 variables

La tabla obtenida es la siguiente:

Ilustración 24 Tabla ANOVA Modelo No Determinista Random Forest

X1	1	19486	19486	41896	<2e-16 ***
X2	1	20017	20017	43037	<2e-16 ***
X3	1	5819	5819	12511	<2e-16 ***
X4	1	9371	9371	20149	<2e-16 ***
X5	1	7863	7863	16906	<2e-16 ***
X6	1	6313	6313	13573	<2e-16 ***
X7	1	17928	17928	38547	<2e-16 ***
X8	1	786	786	1689	<2e-16 ***
X9	1	1624	1624	3491	<2e-16 ***
X10	1	543	543	1168	<2e-16 ***
Residuals	989	460	0		

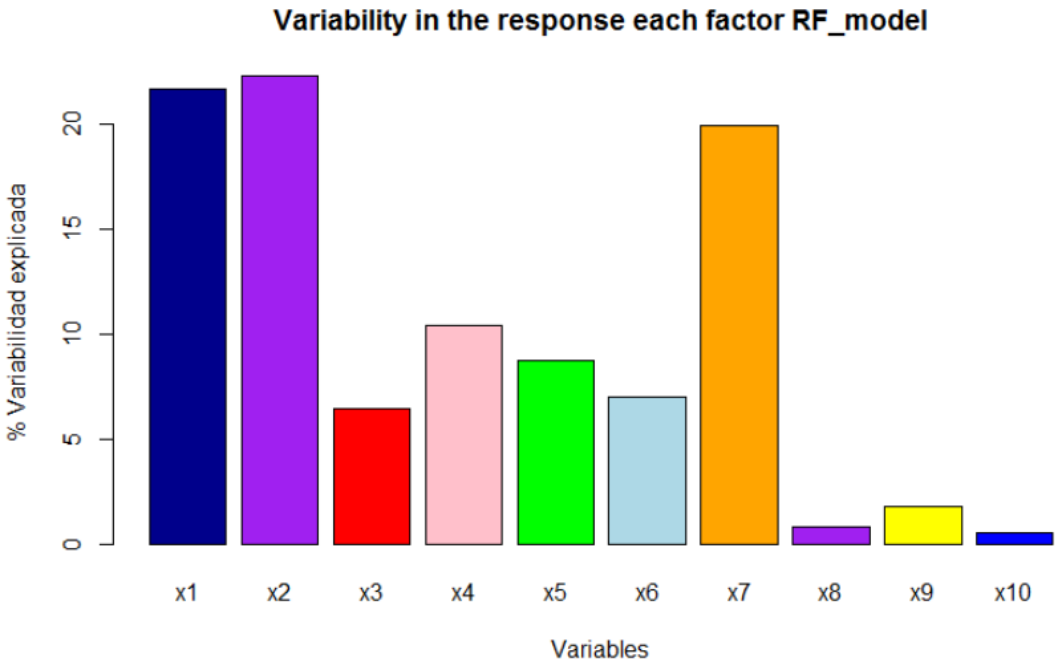


Ilustración 25 Gráfica % Variabilidad Factores Modelo No Dterminista, Random Forest

Variable	% Variabilidad Explicada sobre el modelo
X1	21.7%
X2	22.9%
X3	6.48%
X4	10.44%
X5	8.7%
X6	7.03%
X7	19.97%
X8	0.87%
X9	1.8%
X10	0.6%

Se observa según los datos obtenidos, que el orden de importancia es el mismo que cuando se ha realizado el modelo paramétrico, pero hay una diferencia en el porcentaje de variabilidad que introduce cada factor. Como conclusión se puede establecer por tanto, que el modelo no paramétrico *de random forest*, explica de manera más adecuada y más cerca de la realidad el modelo que el modelo paramétrico. Además en el modelo paramétrico las interacciones eran nulas, no obstante viendo que los residuos no son nulos ahora, se podría intuir que sí que existen interacciones reales entre las variables. Algo que el modelo paramétrico obviaba y que el random forest sí que es capaz de captar.

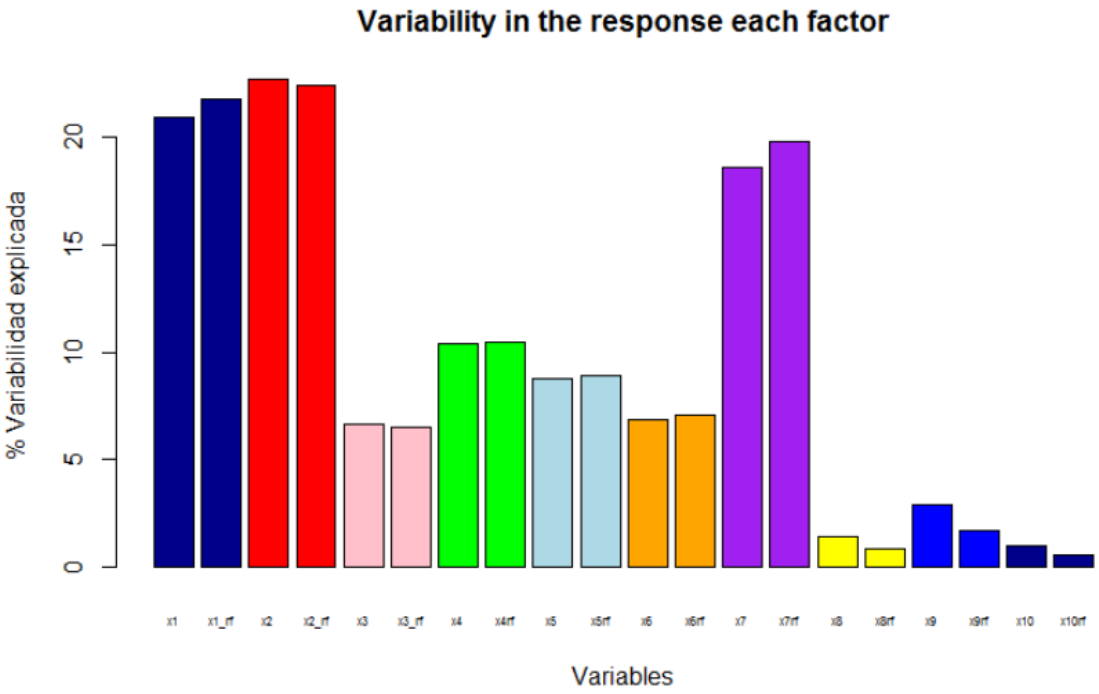


Ilustración 26 Gráfica Modelo Paramétrico vs Modelo Random Forest

3.5 Experimentos en el modelo Determinista

Una vez se han realizado los experimentos para el modelo no determinista, ahora se van a realizar para el mismo modelo al que se le añade el término aleatorio por lo que la fórmula final del modelo quedará:

$$y = \sum_{i=0}^{n} \beta_i x_i + \varepsilon_i$$

Siendo ε el término que va a introducir una mayor incertidumbre al modelo al ser un nuevo término aleatorio. Como se explicó anteriormente, este término, seguirá una variación normal, N (0,1).

Al igual que se hizo con el modelo no determinista, se van a llevar a cabo los mismos experimentos.

En primer lugar, se va a realizar el análisis ANOVA a los datos generados para el modelo: Al igual que en el modelo anterior, se va obtener la tabla del ANOVA, y después se extraerá el valor de la suma de cuadrados de cada variable para calcular la variabilidad explicada por cada factor.

Ilustración 27 Tabla ANOVA Modelo Determinista Paramétrico

X1	1	27012	27012	2.138e+32	<2e-16 ***
X2	1	21628	21628	1.712e+32	<2e-16 ***
X3	1	6367	6367	5.040e+31	<2e-16 ***
X4	1	9896	9896	7.833e+31	<2e-16 ***
X5	1	8344	8344	6.604e+31	<2e-16 ***
X6	1	6554	6554	5.188e+31	<2e-16 ***
X7	1	17713	17713	1.402e+32	<2e-16 ***
X8	1	1355	1355	1.072e+31	<2e-16 ***
X9	1	2759	2759	2.184e+31	<2e-16 ***
X10	1	937	937	7.416e+30	<2e-16 ***
Residuals	989	0	0		

Variable	%Variabilidad sobre el modelo
X1	26.3%
X2	21.08%
X3	6.21%
X4	9.64%
X5	8.13%
X6	6.39%

X7	17,26%
X8	1.32%
X9	2.69%
X10	0.91%

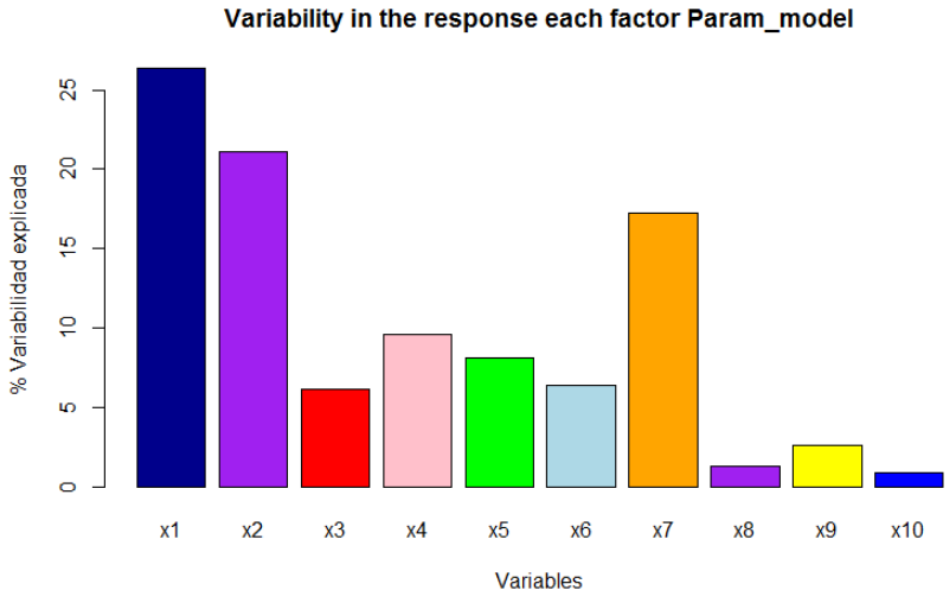


Ilustración 28 Gráfica Variabilidad Factores, Modelo Determinista Paramétrico

Se puede observar, que se obtienen resultados muy similares al modelo anterior pero con mayor variabilidad (mayor suma de cuadrados) debido a que el término aleatorio, introduce una mayor complicación a la hora de ajustar el modelo.

Una vez, explicado el análisis ANOVA, para los datos generados, se procederá una vez más a realizar los experimentos en el modelo *random forest* para ver cómo de bueno es el modelo, realizando este análisis con el análisis ANOVA.

3.5.1 Experimentos en *Ntree*

En estos experimentos se variará el parámetro del número de árboles que forman el bosque, desde un valor de 50 hasta el valor predeterminado en la función de R de 500 árboles. En cuanto a los demás parámetros, se mantendrán los valores predeterminados, siendo *mtree*, el número de variables que se seleccionan en cada nodo al hacer el split igual a *p/3*. En este experimento al tener 10 variables se tendrán 3 variables en cada split.

Haciendo, la variación en el número de árboles, vamos a obtener el número de árboles, que hace que el modelo se explique mejor, es decir, tenga una variabilidad explicada (r^2) mayor. Los resultados se presentan en la siguiente gráfica:

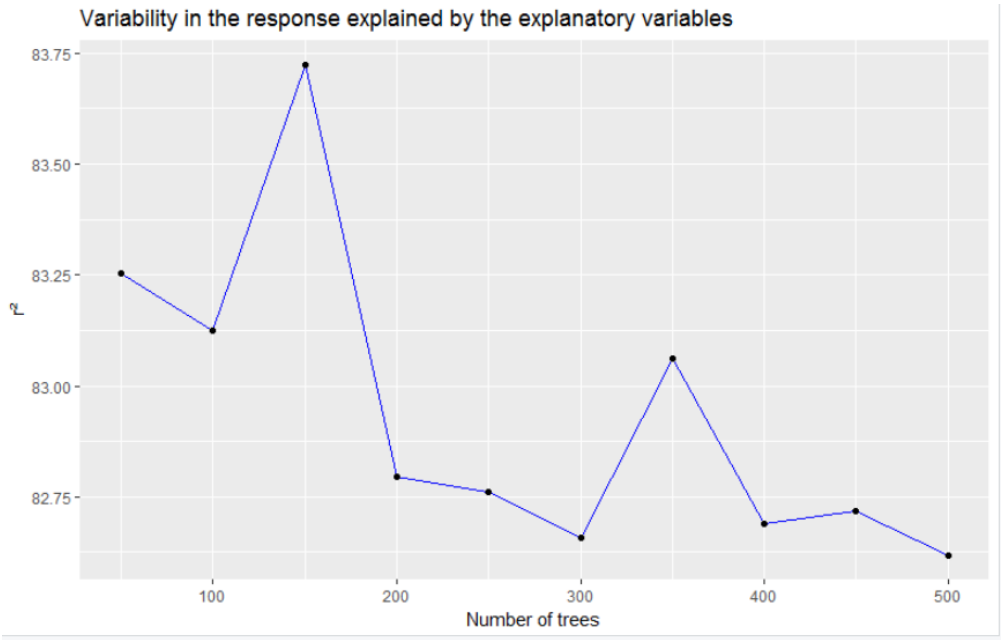


Ilustración 29 Gráfica % Variabilidad Total Modelo vs Ntree

Se observa, que es el punto con ntree=150 árboles, el punto que explica más el modelo, con un 83,75% aproximadamente. Esto quiere decir, que ese porcentaje del modelo es explicado por todas las variables cuando se tiene un modelo *random forest* formado por 150 árboles y 3 variables en cada split. Ahora se comprobará con las gráficas del error que efectivamente, este es el mejor punto:

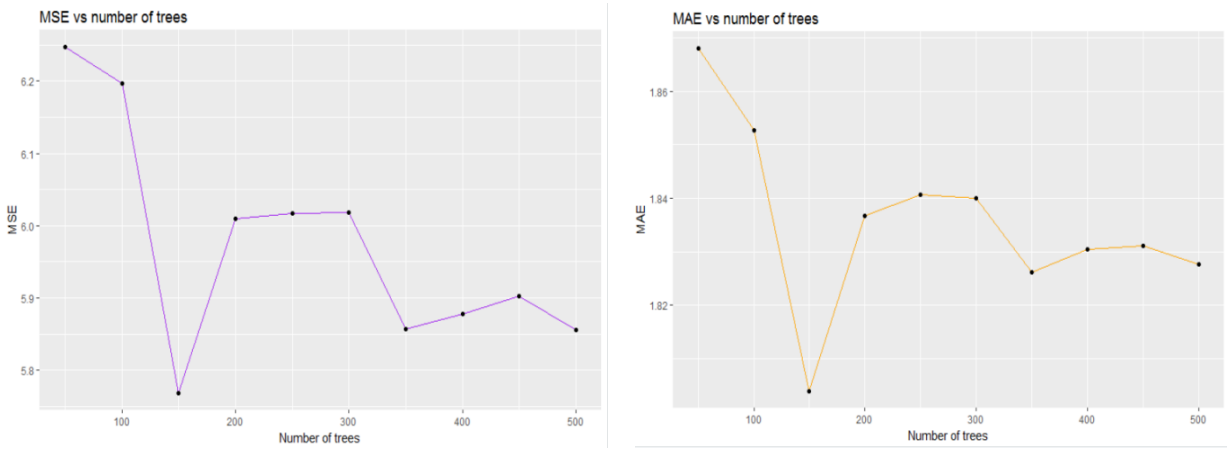


Ilustración 30 Gráficas MSE, MAE vs Ntee

Efectivamente, se observa, que para el punto mencionado anteriormente, se obtiene el menor error en el modelo. Por lo tanto, variando el parámetro del número de árboles, se obtiene que el mejor punto es 150 árboles.

3.5.2 Experimentos en *Mtry*

Del mismo modo que se han realizado los experimentos en el parámetro del número de árboles, se van a realizar experimentos en el parámetro del *random Forest* que elegirá el número de variables en cada split. Para este experimento, se utilizarán las mismas

variables que en el experimento anterior, es decir, se tendrán, N= 1000 observaciones de las p=6 variables anteriores.

Obteniendo de manera análoga, vamos a obtener el punto con mayor variabilidad explicada del modelo:

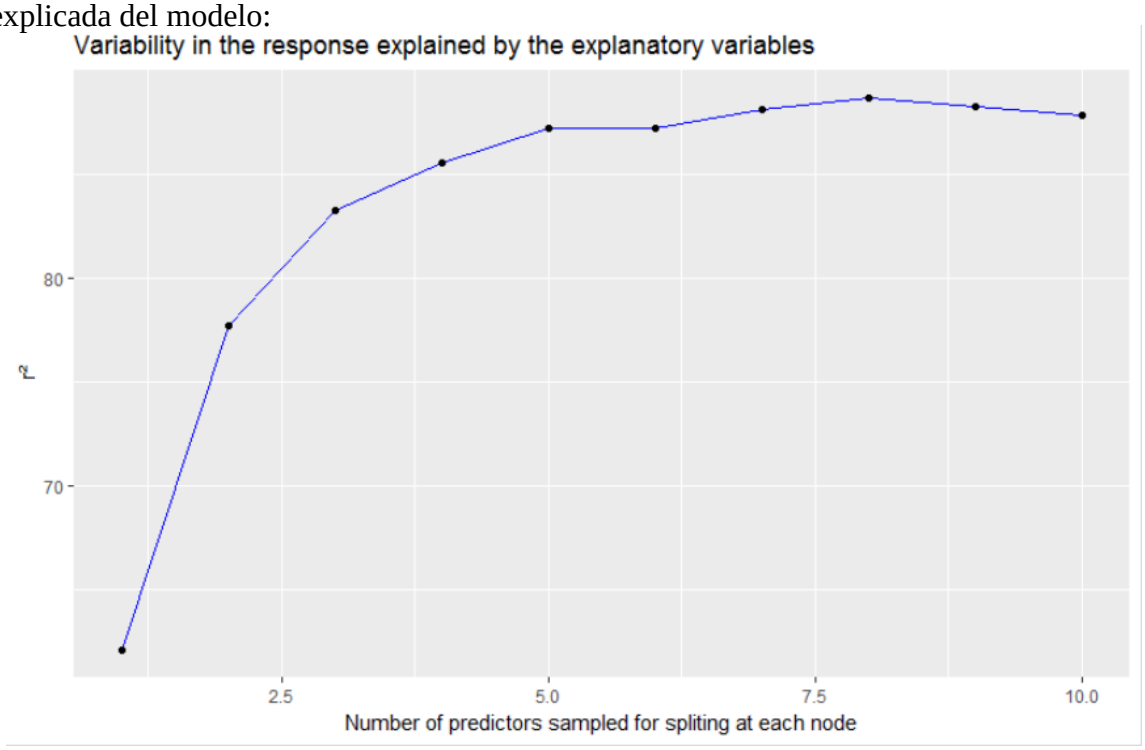


Ilustración 31 Gráfica % Varibilidad Total vs Mtry

Se puede ver, como a mayor número de variables escogidas en cada partición, mayor parte del modelo es explicado. No obstante, como se ha mencionado anteriormente, cuantas más variables se seleccionen en cada partición más aleatoriedad se perderá en el modelo.

Una vez más, calculando las gráficas del error, se obtiene que a mayor número de variables escogidas, que tiene como consecuencia una mayor variabilidad explicada del modelo, se tengan menores errores, hasta un punto donde aproximadamente se estabilizará y no se mejorará mucho más el modelo.

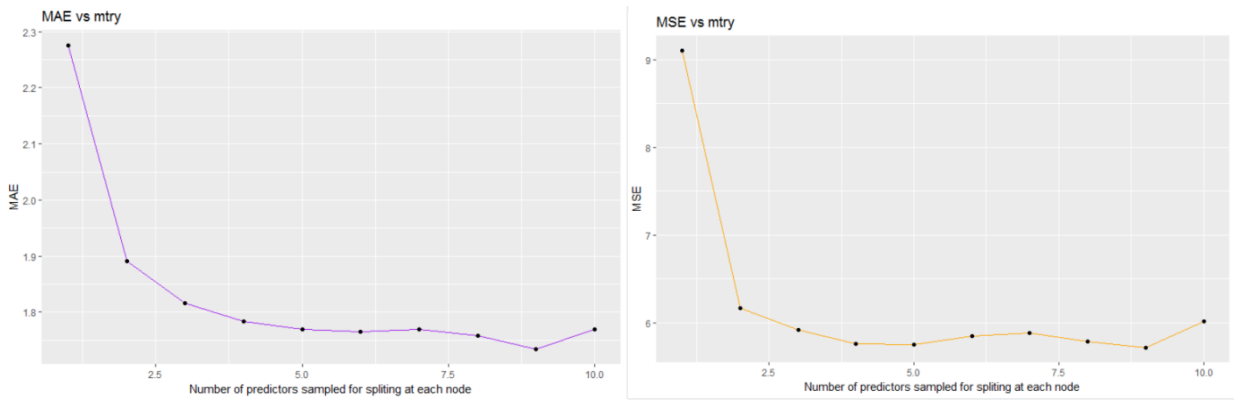


Ilustración 32 Gráficas MAE, MSE vs Mtry

Aproximadamente, para un valor de mtry=4 variables, se tiene el mejor punto del modelo con el menor error y una variabilidad explicada del modelo de un 90%.

Ahora, análogamente a como se hizo con el modelo no determinista, con este valor de los parámetros del *random Forest* construimos el modelo final sobre el que se realizará el análisis ANOVA. Se extraerán los mismos parámetros para poder calcular la variabilidad de cada factor de la misma forma que se realizó para el modelo anterior. Se obtienen los siguientes resultados:

Ilustración 33 Tabla ANOVA Modelo Determinista Random Forest

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X1	1	26207	26207	64024	<2e-16	***
X2	1	20122	20122	49157	<2e-16	***
X3	1	5977	5977	14602	<2e-16	***
X4	1	9452	9452	23092	<2e-16	***
X5	1	7919	7919	19347	<2e-16	***
X6	1	6304	6304	15402	<2e-16	***
X7	1	17536	17536	42842	<2e-16	***
X8	1	773	773	1889	<2e-16	***
X9	1	1536	1536	3752	<2e-16	***
X10	1	515	515	1258	<2e-16	***
Residuals	989	405	0			

De la tabla ANOVA, anterior, se calcula la siguiente gráfica de barras en la que se representa el porcentaje de variabilidad de cada factor del modelo.

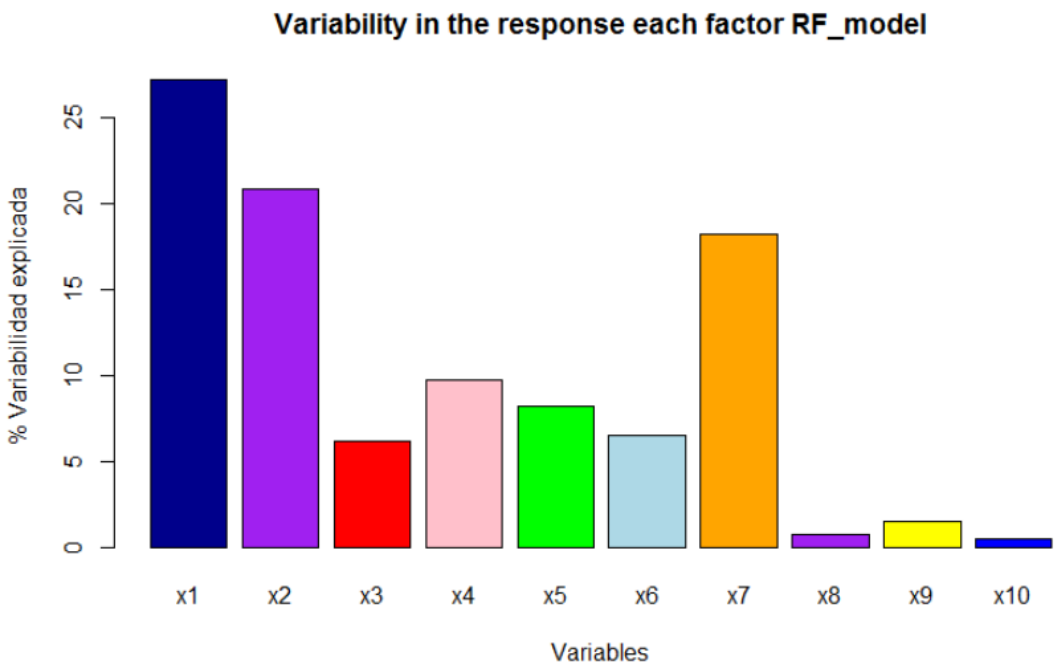


Ilustración 34 Gráfica % Variabilidad Factores Modelo Determinista Random Forest

Variable	%Variabilidad sobre el modelo
X1	27.2%
X2	20.88%
X3	6.21%
X4	9.8%
X5	8.22%
X6	6.54%
X7	18,22%
X8	0.8.%
X9	1.59%
X10	0.53%

Se observa, que en el orden de importancia de las variables es el mismo que en el diseño paramétrico. En cuanto a porcentajes de variabilidad se observa que aumenta para las variables más significativas del modelo. Comparando las dos tablas ANOVA, la del modelo original y la del modelo *random forest*, se observa nuevamente como en el modelo *random forest* la suma de cuadrados de este modelo es menor. Esto como se explicó anteriormente significa que la diferencia entre el valor real y el predicho es menor, por lo que la predicción realizada es mejor en este último modelo. Pero se observa que al no haber interacciones los resultados son muy similares.

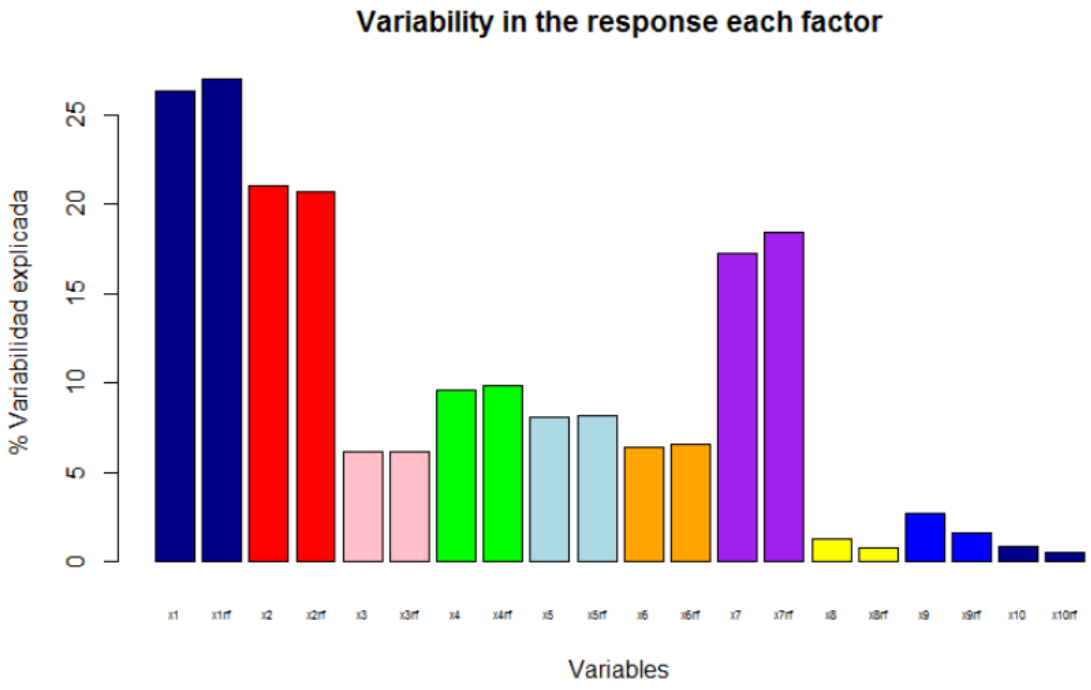


Ilustración 35 Variabilida explicad por los factores Modelo Paramétrico VS Random Forest

Capítulo 4

Experimentos en un sistema Lineal con interacciones

4.1 Introducción

En este Trabajo de Fin de Grado se quiere realizar un análisis de la importancia de las variables de un modelo. Para ello, se ha hecho uso de un análisis ANOVA a un modelo paramétrico, que explicará esta importancia en función de la variabilidad que los factores y sus interacciones introducen al modelo. Sin embargo, también se ha querido realizar este análisis de importancia de variables con un método no paramétrico como es el *random Forest*. En el capítulo anterior, se analizaba un modelo que se consideraba lineal sin interacciones, pudiéndose observar que ambos diseños explicaban de manera similar el modelo. No obstante, el modelo *random forest* explicaba mejor el modelo puesto que además de realizar muy buenas predicciones, es capaz de detectar las posibles interacciones entre las variables que si pudieran existir en realidad. Para ver el efecto que pueden tener estas interacciones, en este capítulo, se va a analizar un modelo con fuertes interacciones entre sus variables. Se querrá probar que a diferencia del modelo paramétrico, el random forest es capaz de manejar grandes y fuertes interacciones entre los factores del modelo.

En este capítulo por lo tanto, se irá un paso más allá a la hora de realizar el análisis de la sensibilidad del modelo a través de la herramienta *random Forest*.

4.2 El modelo

En este apartado, se va a explicar el modelo y los cambios en la construcción del mismo. En este capítulo, se ha hecho uso de un modelo ortogonal ya creado que se ha encontrado en el libro *Surfaces, Desings and Analyses.*[1] Este modelo, consta de cinco variables ortogonales entre sí. Esta ortogonalidad que ha sido calculada mediante la codificación de variables (segundo método que se

[1] "*Response Surfaces, Desing and Analyses,"* André I. Khuri, John A. Cornell. Statistics: textbooks and monographs. Volume 61.

proponía para conseguir variables independientes) ya ha sido calculado para este ejemplo. En el siguiente capítulo, se irá un paso más allá y se estudiarán unos datos no ortogonales que se ortogonalizarán. Tras este inciso, los datos a estudiar de este modelo, constan de cinco variables y una salida. De estos datos, no se conoce la ecuación que siguen los mismos. Sin embargo, si se conoce que las variables X_4 y X_5 no son significativas para el modelo. Sin embargo, para probar este hecho, en primer lugar se va a volver a tratar como un modelo lineal de la siguiente manera:

$$y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Al igual que se hizo en el apartado anterior, se calculará el ANOVA a este modelo, y posteriormente se generará el modelo *random forest*, realizando el ANOVA final a este segundo modelo. A través de este análisis, tal y como nos dice el ejemplo, las variables X_4 y X_5, no deberán de salir significativas. Además, a través del *random forest*, si existieran interacciones este diseño será capaz de encontrarlas.

Tras estos resultados, se va a decidir aproximar el modelo a la siguiente ecuación para ver si este modelo explica mejor el comportamiento real de los datos:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3$$

Este modelo, que es propuesto en el mismo ejemplo, se puede observar tiene mayores interacciones entre sus variables, lo que va a complicar el modelo de manera considerable. Si en este modelo, resultan significativas las interacciones (tras hacer los posteriores análisis ANOVA) entre las variables querrá decir que en efecto el modelo tiene interacciones entre sus variables.

4.3 Procedimiento

En el capítulo anterior, se explicó un procedimiento para obtener variables ortongonales. Se recuerda que la ortogonalidad de las variables es una condición estrictamente necesaria del análisis ANOVA para poder explicar la variabilidad total del modelo como la descomposición de las variabilidades de sus factores e interacciones. En el capítulo anterior al tratarse de un modelo sin interacciones, se utilizó el método de la función sobol. No obstante, con este método, obteníamos variables que no eran ortogonales al 100% pero el error era tan pequeño y como no existían interacciones entre las variables, se daban por validos los resultados.

Sin embargo, en este caso el modelo que se ha supuesto, sí que va a tener fuertes interacciones entre las variables. Es por esta razón que en este capítulo se va a hacer uso del segundo método que se mencionó para conseguir la ortogonalidad.

Como se ha explicado anteriormente, en este capítulo no se va a explicar el método de la codificación ya que las variables del modelo ya son ortogonales entre sí, ese estudio se realizará en el capítulo posterior.

Para la consideración lineal del modelo, se harán los mismos experimentos que en el capítulo anterior, un análisis ANOVA al modelo paramétrico y un análisis ANOVA al modelo *random forest.*

En cambio, cuando se considera que el modelo es un modelo con interacciones, como el que se ha supuesto en la ecuación dos, es necesario calcular el valor real de las β y la nueva salida $\hat{Y}$ estimada del modelo.

Para obtener estos parámetros se introducirá en R la siguiente matriz:

$$[Y] = [\beta][X]$$

La matriz X quedará de la siguiente manera:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	1	-1	-1	-1	1	1	1	1	1	1
[2,]	1	1	-1	-1	1	1	1	-1	-1	1
[3,]	1	-1	1	-1	1	1	1	-1	1	-1
[4,]	1	1	1	-1	1	1	1	1	-1	-1
[5,]	1	-1	-1	1	1	1	1	1	-1	-1
[6,]	1	1	-1	1	1	1	1	-1	1	-1
[7,]	1	-1	1	1	1	1	1	-1	-1	1
[8,]	1	1	1	1	1	1	1	1	1	1
[9,]	1	-1	-1	-1	1	1	1	1	1	1
[10,]	1	1	-1	-1	1	1	1	-1	-1	1
[11,]	1	-1	1	-1	1	1	1	-1	1	-1
[12,]	1	1	1	-1	1	1	1	1	-1	-1
[13,]	1	-1	-1	1	1	1	1	1	-1	-1
[14,]	1	1	-1	1	1	1	1	-1	1	-1
[15,]	1	-1	1	1	1	1	1	-1	-1	1
[16,]	1	1	1	1	1	1	1	1	1	1
[17,]	1	-2	0	0	4	0	0	0	0	0
[18,]	1	2	0	0	4	0	0	0	0	0
[19,]	1	0	-2	0	0	4	0	0	0	0
[20,]	1	0	2	0	0	4	0	0	0	0
[21,]	1	0	0	-2	0	0	4	0	0	0
[22,]	1	0	0	2	0	0	4	0	0	0
[23,]	1	0	0	0	0	0	0	0	0	0
[24,]	1	0	0	0	0	0	0	0	0	0
[25,]	1	0	0	0	0	0	0	0	0	0
[26,]	1	0	0	0	0	0	0	0	0	0
[27,]	1	0	0	0	0	0	0	0	0	0
[28,]	1	0	0	0	0	0	0	0	0	0
[29,]	1	0	0	0	0	0	0	0	0	0
[30,]	1	0	0	0	0	0	0	0	0	0
[31,]	1	0	0	0	0	0	0	0	0	0
[32,]	1	0	0	0	0	0	0	0	0	0

Siendo las variables de entrada al modelo, los términos de la ecuación de las interacciones entre las variables

Con el valor de las Y, se obtiene en R los siguientes valores de beta:

β= [77.17,-10.12,-8.68,-0.3,-4.10,-4.72,-2.30,-6.21, 2.77,-1.68]

Por lo tanto, el modelo final que paramétrico de segundo orden que se utilizará para el análisis ANOVA, será:

$$y = 77.17 - 10.12X_1 - 8.68X_2 - 0.10X_3 - 4.10X_1^2 - 4.72X_2^2 - 2.30X_3^2 - 6.20X_1X_2 + 2.72X_1X_3 - 1.68X_2X_3$$

Ahora, con este modelo, se calculará la nueva salida, y se procederá de la misma manera que se ha hecho con todos los modelos anteriores. Los resultados, se expondrán nuevamente, en un diagrama de barras en el que se muestra la variabilidad total que cada variable e interacciones introducen al modelo. Para extraer estos, se calcularán de la siguiente manera:

Para sacar la variabilidad explicada de cada factor e interacción del modelo, se va a calcular el mean square (media al cuadrado): Se dividirá la suma de cuadrados entre los grados de libertad:

$$MS_i = \frac{SS_i}{DF_i}$$

Donde SS_i es suma de cuadrados extraída de la tabla ANOVA, para cada factor e interacción y DF_i son los grados de libertad.

Ahora para calcular la varianza de cada factor:

$$VE = \frac{MS_{factor} - MS_{error}}{nº\,Variables}$$

Una vez calculada la variabilidad que introduce cada factor al modelo, se va expresar en forma de gráfico de barras. Este gráfico, se comparará con el gráfico que se obtendrá del modelo *random Forest* y así se observará la diferencia a la hora de explicar el modelo por ambas herramientas.

4.4 Consideración del Modelo Lineal

En primer lugar, como se ha explicado en el apartado anterior, primero se va a considerar el modelo como lineal. Realizando el ANOVA a los datos originales y calculando la gráfica de la variabilidad de cada factor obtenemos los siguientes resultados: Tabla 2 ANOVA Modelo Lineal

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	2458.4	2458.4	27.806	1.64e-05	***
X3	1	1807.9	1807.9	20.448	0.000119	***
X4	1	0.3	0.3	0.003	0.957133	
X5	1	12.2	12.2	0.138	0.713480	
X6	1	44.6	44.6	0.504	0.484096	
Residuals	26	2298.7	88.4			

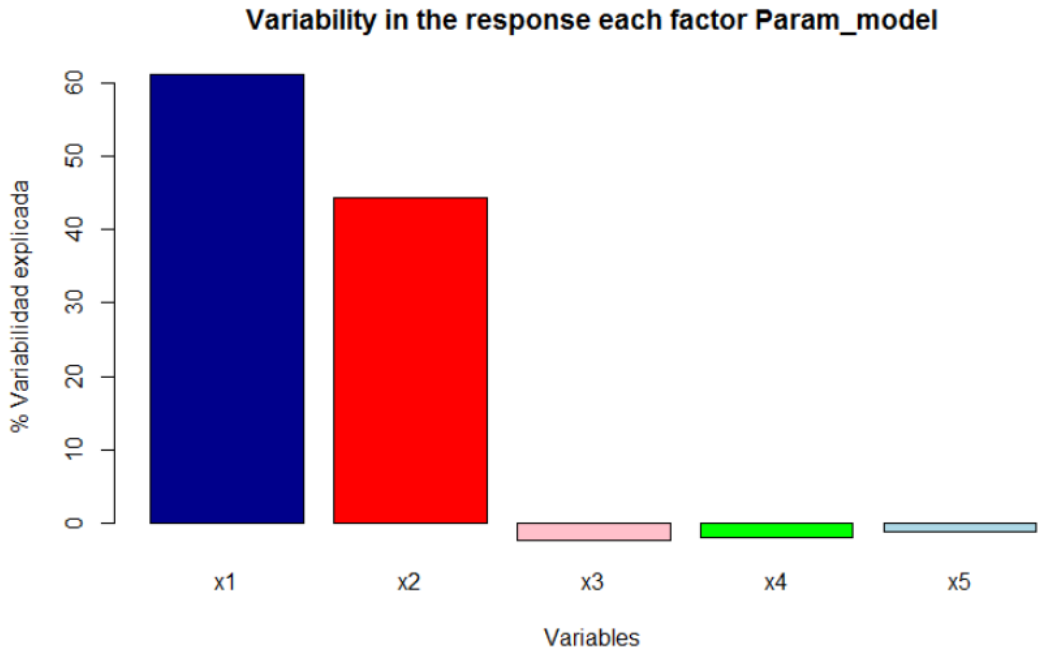


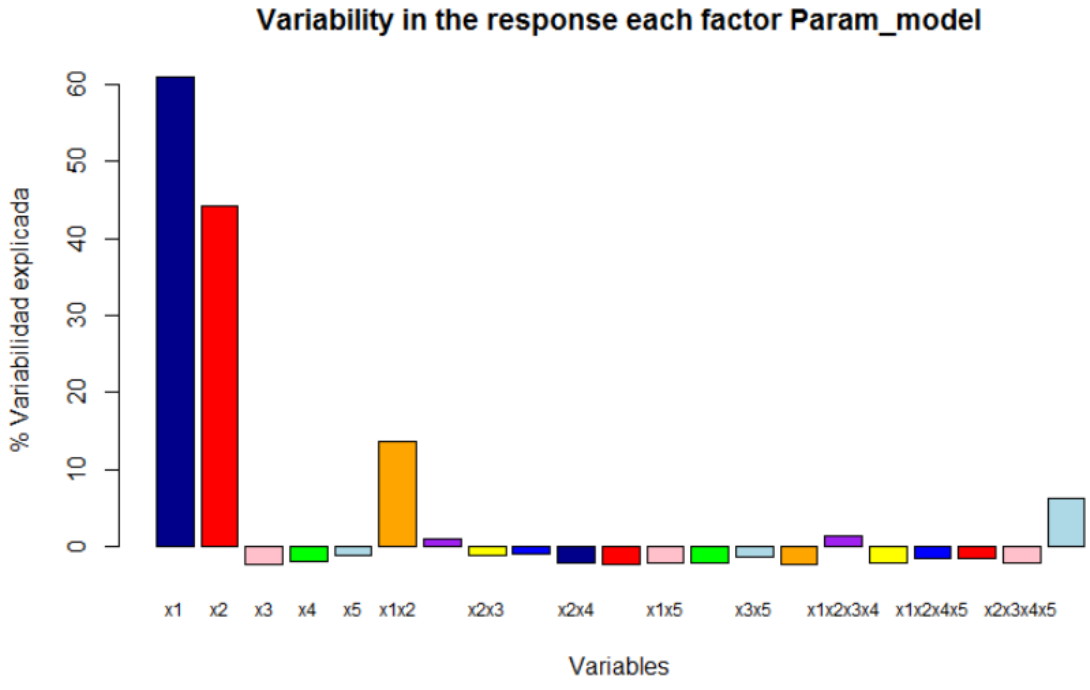
Ilustración 36 Variabilidad Cada Factor Modelo Lineal

Modelo	% Variabilidad Cada Factor
X_1	61.1%
X_2	44.3%
X_3	2.27%
X_4	-1.97%
X_5	-1.13%

Se observa, cómo se obtiene tal y como se dijo en el apartado de explicación del modelo que las variables X_4 y X_5 no son significativa ya que tienen porcentajes negativos en cuanto a la variabilidad en el modelo. En cambio se ve que las variables X_1 y X_2 son las variables que más explican el modelo. Si se realiza el ANOVA observando las posibles interacciones entre las variables del modelo obtenemos la siguiente tabla.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	870.7	870.7	25.950	0.000468
x2	1	767.1	767.1	22.862	0.000744
x3	1	1.1	1.1	0.031	0.862712
x4	1	1.7	1.7	0.051	0.826613
x5	1	17.2	17.2	0.513	0.490096
x1:x2	1	47.6	47.6	1.420	0.260937
x1:x3	1	5.1	5.1	0.151	0.705668
x2:x3	1	3.5	3.5	0.105	0.752438
x1:x4	1	3.1	3.1	0.092	0.767937
x2:x4	1	0.7	0.7	0.020	0.890848
x3:x4	1	0.4	0.4	0.011	0.919011
x1:x5	1	0.0	0.0	0.001	0.975414
x2:x5	1	0.7	0.7	0.021	0.887845
x3:x5	1	0.9	0.9	0.026	0.874880
x4:x5	1	0.1	0.1	0.003	0.960520
x1:x2:x3:x4	1	25.3	25.3	0.755	0.405131
x1:x2:x3:x5	1	0.1	0.1	0.002	0.967472
x1:x2:x4:x5	1	4.6	4.6	0.138	0.718115
x1:x3:x4:x5	1	6.5	6.5	0.194	0.669004
x2:x3:x4:x5	1	0.4	0.4	0.011	0.920263
x1:x2:x3:x4:x5	1	536.1	536.1	15.978	0.002530
Residuals	10	335.5	33.6		

Se observa como claramente existen interacciones hasta de quinto orden entre las variables del modelo de datos reales. Se observa como la interacción más importante es entre X_1 y X_2, como es lógico puesto que se ha visto que son las variables más importantes del modelo. Las demás interacciones no serían significativas.



Si ahora procedemos a realizar el mismo análisis pero al modelo *random forest*:

4.4.1 Experimentos en Random Forest

En primer lugar, al igual que se ha hecho en el capítulo anterior, se van a calcular los parámetros óptimos del árbol. Este método es exactamente el mismo que se realizó en el capítulo anterior por lo que para no ser redundantes, no se va a volver a explicar.

4.4.1.1 Experimentos en Ntree

En la gráfica de la variabilidad total del modelo, se puede observar como el punto óptimo es Ntree= 200. Se va a comprobar este punto con las gráficas del error mínimo cuadrático (MSE) y el error mínimo absoluto (MAE).

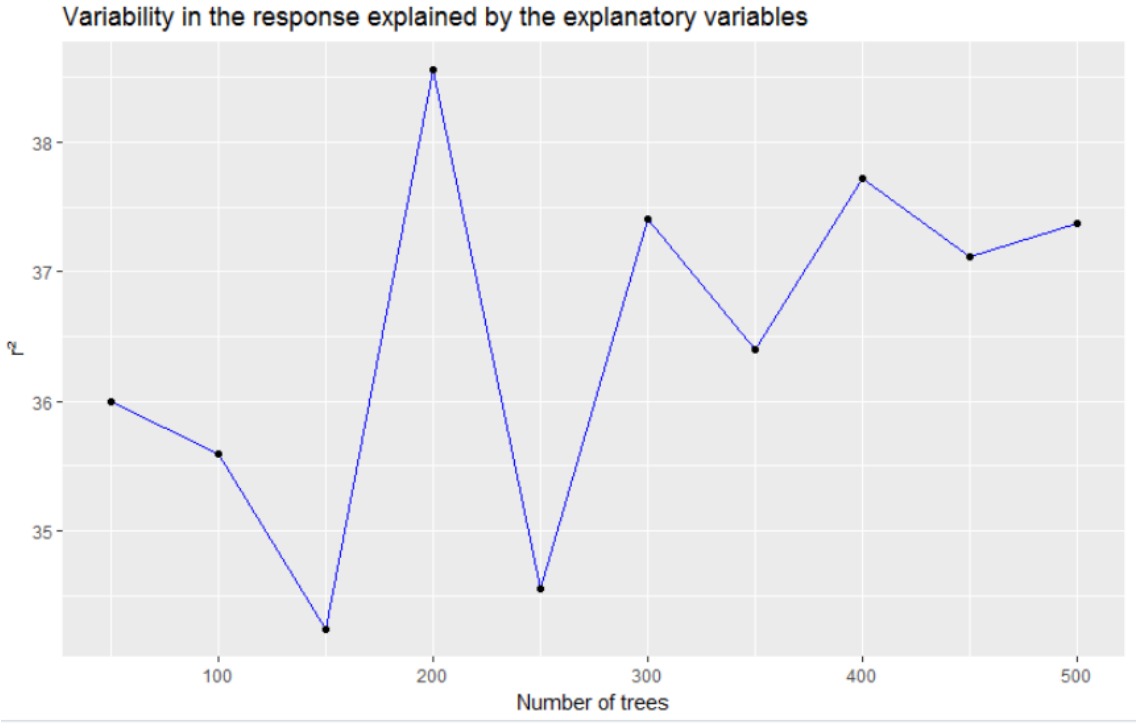


Ilustración 38 Gráfica Variabilidad Total del Modelo VS Ntree

Efectivamente, se obtiene el mismo punto que es el punto que va a explicar mejor el modelo.

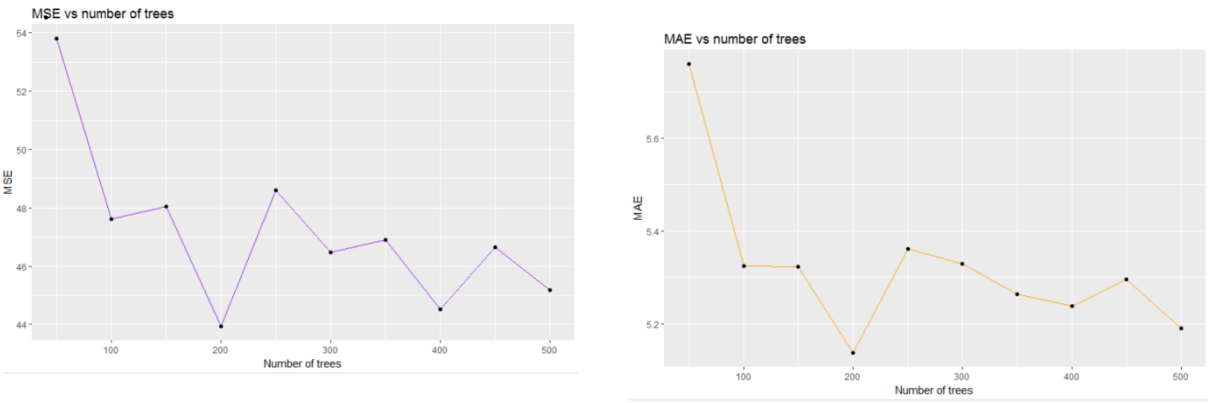


Ilustración 39 Gráficas MSE, MAE vs Ntree

4.4.1.2 Experimentos en Mtry

Análogamente, se van a realizar los mismos experimentos para el parámetro del número de variables seleccionadas en cada partición. Con los mismos resultados gráficos se va a obtener que a partir de Mtry=2 va a empezar a ocurrir overfitting por lo tanto, se va a seleccionar ese punto como punto óptimo.

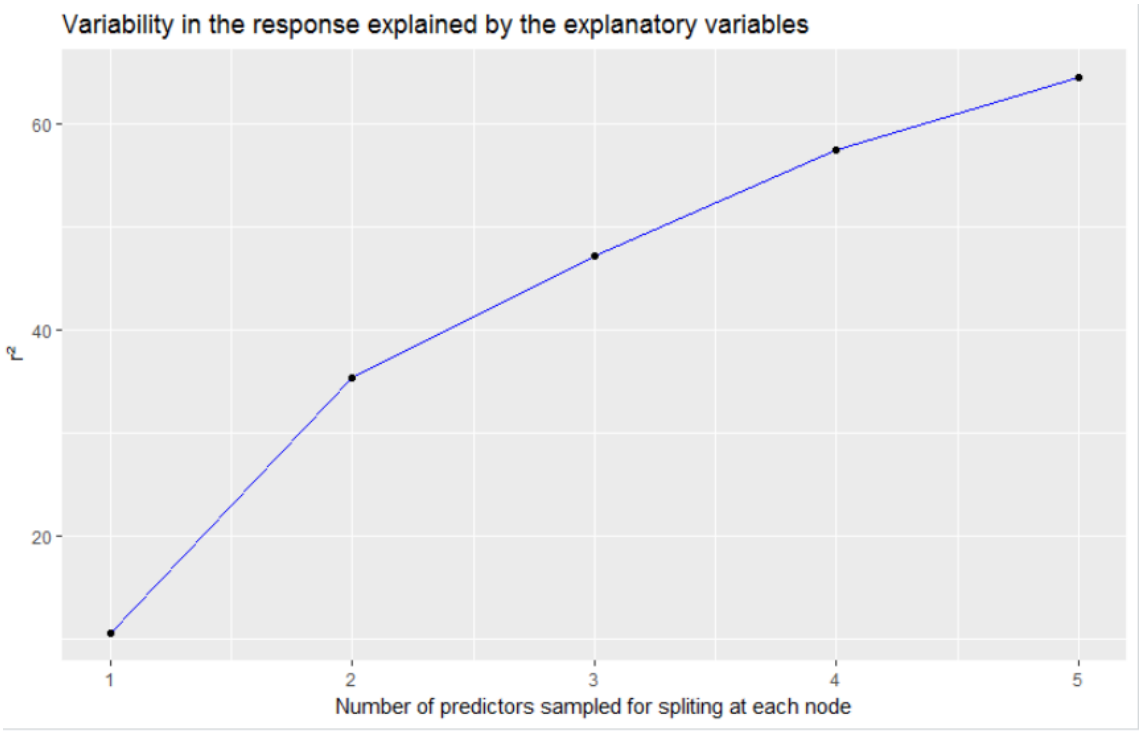
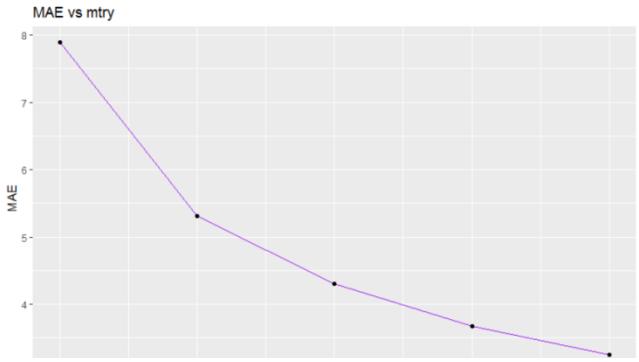


Ilustración 40 Gráfica % Variabilidad Total vs Mtry

Con las gráficas del error, se puede observar que cuantas más variables se seleccionen menor error existirá pero al igual que se ha explicado en capítulos anteriores no es aconsejable coger todas las variables del modelo puesto que se restaría aleatoriedad al modelo.



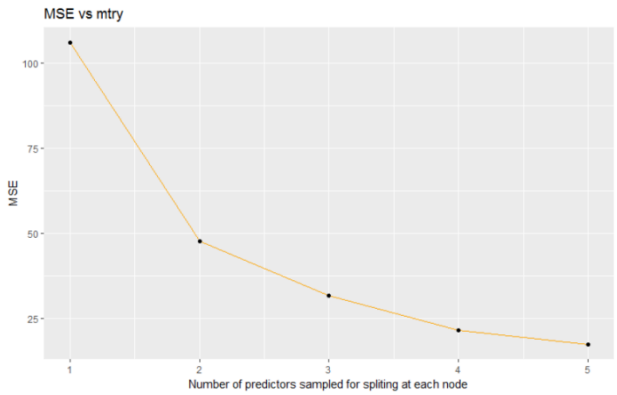


Ilustración 41 Gráficas MSE, MAE vs Mtry

Una vez que se tiene el punto óptimo del *random forest*, se va a proceder a crear el modelo. Con el modelo creado, se realizará una predicción y sobre ese nuevo modelo vamos a generar el ANOVA. Se calculará la variabilidad de cada factor de la misma manera que en los modelos anteriores obteniendo los siguientes resultados:

El punto óptimo seleccionado por tanto ha sido:

- Ntree= 200 árboles
- Mtry= 2 variables

Tabla 4 ANOVA Modelo RF

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	923.7	923.7	27.726	1.67e-05	***
X3	1	622.9	622.9	18.695	0.0002	***
X4	1	2.4	2.4	0.072	0.7908	
X5	1	5.6	5.6	0.167	0.6859	
X6	1	19.5	19.5	0.587	0.4506	
Residuals	26	866.2	33.3			

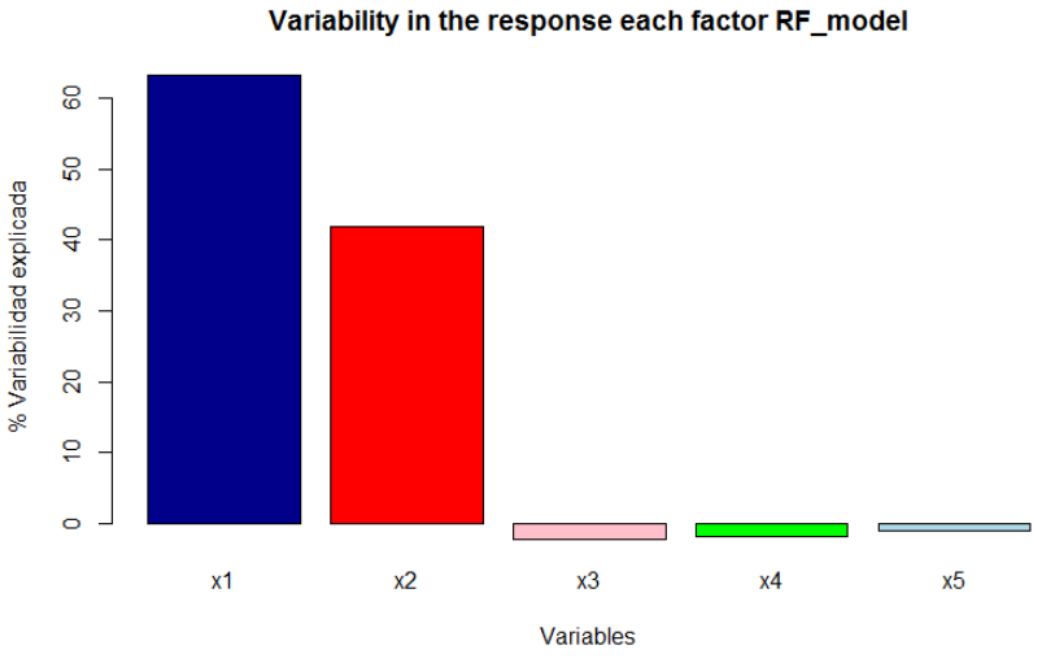


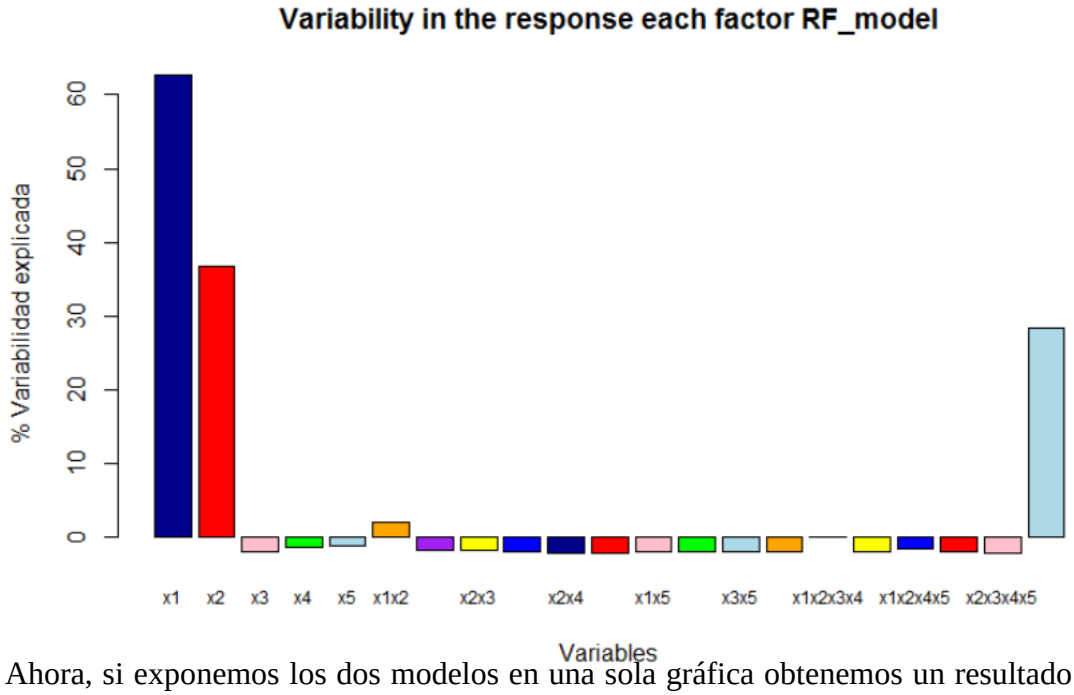
Ilustración 43 % Variabilidad Cada Factor Modelo RF

Modelo	% Variabilidad Cada Factor
X_1	63.26%
X_2	41.88%
X_3	2.19%
X_4	-1.97%
X_5	-0.97%

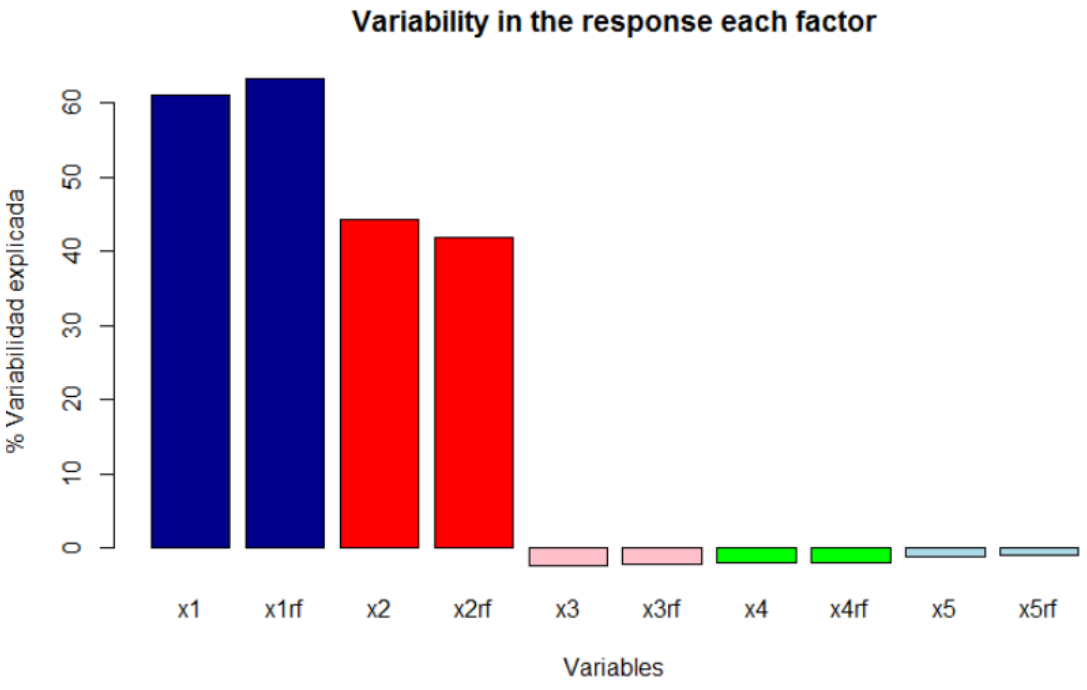
Se observa, como las variables que no eran significativas antes tampoco lo son ahora. Y aunque se obtienen resultados similares, como los porcentajes de variabilidad de los factores han cambiado, esto incita a creer que efectivamente el modelo tiene interacciones entre sus variables. Además, tal y como sucedía en el capítulo anterior, se puede ver cómo el modelo del random forest realiza mejores predicciones del modelo obteniendo una menor suma de cuadrados en las variables.

Calculando el ANOVA de las interacciones con el modelo random forest obtenemos:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	1065.3	1065.3	31.180	0.000233
x2	1	639.3	639.3	18.712	0.001500
x3	1	1.0	1.0	0.030	0.864893
x4	1	13.0	13.0	0.382	0.550607
x5	1	14.6	14.6	0.429	0.527422
x1:x2	1	68.0	68.0	1.991	0.188589
x1:x3	1	5.4	5.4	0.157	0.700180
x2:x3	1	4.2	4.2	0.122	0.734090
x1:x4	1	0.4	0.4	0.010	0.920720
x2:x4	1	0.1	0.1	0.002	0.962177
x3:x4	1	0.2	0.2	0.007	0.935644
x1:x5	1	1.3	1.3	0.038	0.849321
x2:x5	1	0.5	0.5	0.016	0.902691
x3:x5	1	2.5	2.5	0.072	0.793746
x4:x5	1	1.1	1.1	0.031	0.863698
x1:x2:x3:x4	1	36.0	36.0	1.054	0.328843
x1:x2:x3:x5	1	0.8	0.8	0.022	0.883800
x1:x2:x4:x5	1	7.3	7.3	0.215	0.653023
x1:x3:x4:x5	1	0.7	0.7	0.019	0.892170
x2:x3:x4:x5	1	0.3	0.3	0.008	0.929181
x1:x2:x3:x4:x5	1	501.7	501.7	14.685	0.003307
Residuals	10	341.7	34.2		



Ahora, si exponemos los dos modelos en una sola gráfica obtenemos un resultado muy visual de que ambos modelos explican de manera similar los datos. No obstante, la ligera diferencia puede radicar en la posibilidad de que existan interacciones entre las variables algo que random forest es capaz de detectar. Mediante el análisis ANOVA en las interacciones se ha probado que estas existen, no obstante no todas son significativas. Las interacciones más importante serán entre X_1 y X_2, como se ha podido observar en los resultados gráficos y entre las cinco variables. Por lo tanto, se concluye que el modelo de regresión múltiple lineal no es el modelo adecuado para explicar estos datos reales.



4.5 Consideración del modelo no lineal

Ahora, como se ha explicado en el apartado 4.3 en relación a la explicación del modelo, y al obtener los resultados esperados en el apartado anterior, se va a suponer que el modelo sigue una relación no lineal como la que se planteó. Si se obtienen que las interacciones entre las variables significativas son importantes, entonces se podrá establecer que el modelo efectivamente tenía interacciones. Además haciendo el resultado con el modelo no paramétrico de random forest, probaremos una vez más que este modelo es capaz de captar estas interacciones de mejor manera que el modelo paramétrico.

En primer lugar, vamos a crear el modelo, como se ha explicado y se va a realizar el análisis ANOVA a estos datos generados, obteniendo los siguientes resultados:

Tabla 6 ANOVA Modelo Interacciones

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	2458.4	2458.4	1.930e+31	<2e-16	***
X3	1	1807.9	1807.9	1.419e+31	<2e-16	***
X4	1	0.3	0.3	2.044e+27	<2e-16	***
X5	1	396.2	396.2	3.110e+30	<2e-16	***
X6	1	618.3	618.3	4.854e+30	<2e-16	***
X7	1	157.5	157.5	1.236e+30	<2e-16	***
X8	1	616.3	616.3	4.838e+30	<2e-16	***
X9	1	122.7	122.7	9.628e+29	<2e-16	***
X10	1	45.2	45.2	3.550e+29	<2e-16	***
Residuals	22	0.0	0.0			

En esta tabla hay que señalar la siguiente correspondencia entre el nombre de las variables:

Modelo	Tabla ANOVA
$\mathbf{X_1}$	X2
$\mathbf{X_2}$	X3
$\mathbf{X_3}$	X4
$\mathbf{X_1^2}$	X5
$\mathbf{X_2^2}$	X6
$\mathbf{X_3^2}$	X7
$\mathbf{X_1X_2}$	X8
$\mathbf{X_1X_3}$	X9
$\mathbf{X_2X_3}$	X10

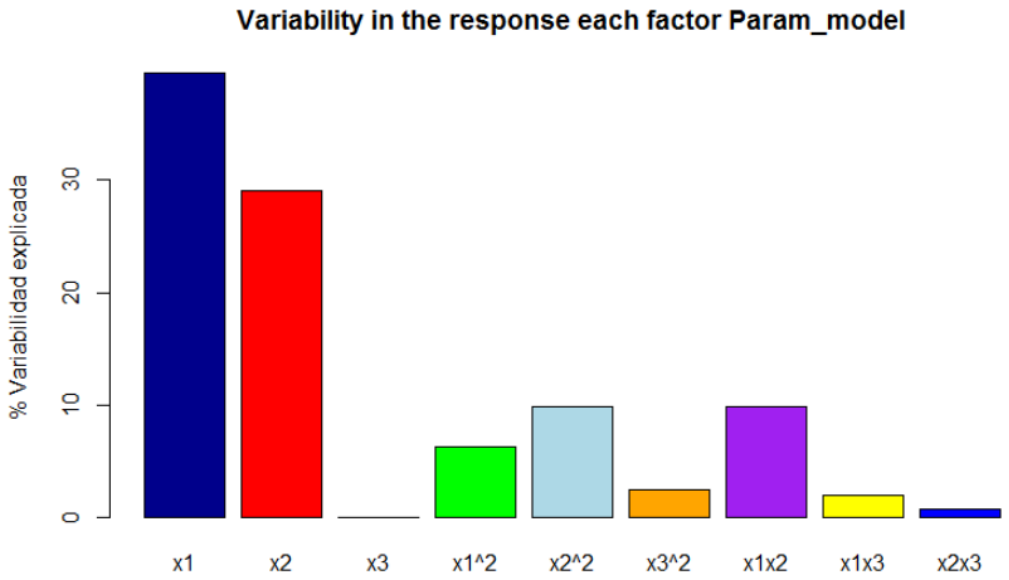


Ilustración 45 Gráfica % Variabilidad Factores e Interacciones Modelo Paramétrico

Variable	%Variabilidad sobre el modelo
X_1	39.5%
X_2	29.53%
X_3	0.04%
X_1^2	6.23%
X_2^2	9.99%
X_3^2	2.29%
X_1X_2	9.87%
X_1X_3	1.72%
X_2X_3	0.77%

Efectivamente, se puede observar cómo las interacciones entre las variables son significativas, por lo que este modelo va a explicar mejor los datos que el modelo lineal propuesto en un principio. Además, si calculamos la variabilidad total del modelo, se observa que se explica en un 93.95% un factor mayor que el 57.6% de considerar el modelo como un modelo lineal.

En el gráfico, se puede observar como el factor X_1 representa el 40% de la variabilidad total, por lo que será la variable que tenga el mayor efecto en la respuesta. En cambio, se observa como las interacciones entre las variables no tienen apenas contribución a la variabilidad de la respuesta, algo que se observaba en los datos originales.

A continuación, se va a realizar el mismo análisis con el modelo *random forest*. Este modelo como se ha mencionado anteriormente, a priori maneja mejor las interacciones entre las variables.

4.5.1 Experimentos en Random Forest

Para el modelo *random Forest*, se escogerán las variables de entrada codificadas y la salida del libro. De esta manera obtendremos los resultados para un modelo no parametrizado. Sin embargo, al igual que se hizo para el modelo lineal, se buscará la pareja de valores de Ntree y Mtry con los que obtenemos la mejor respuesta. Para ello, de la misma manera que se hizo anteriormente se hará una variación en un bucle de uno de los parámetros dejando con valor constante el otro. El valor que se dejará constante, será el valor en la función predeterminado por R.

4.5.1.1 Experimentos en Ntree

En este primer experimento, se hará igual que se hizo en el capítulo anterior se tomará el valor mtry como *p/3*. Se irá variando, el valor del número de árboles desde un valor de 50 árboles hasta 500 árboles. Para cada valor se calculará la variabilidad total del modelo que es explicada por las variables. Además, se calculará el valor del error mínimo cuadrático y el MAE para comprobar que en efectivo ese punto es el óptimo.

Expresando los resultados en forma gráfica:

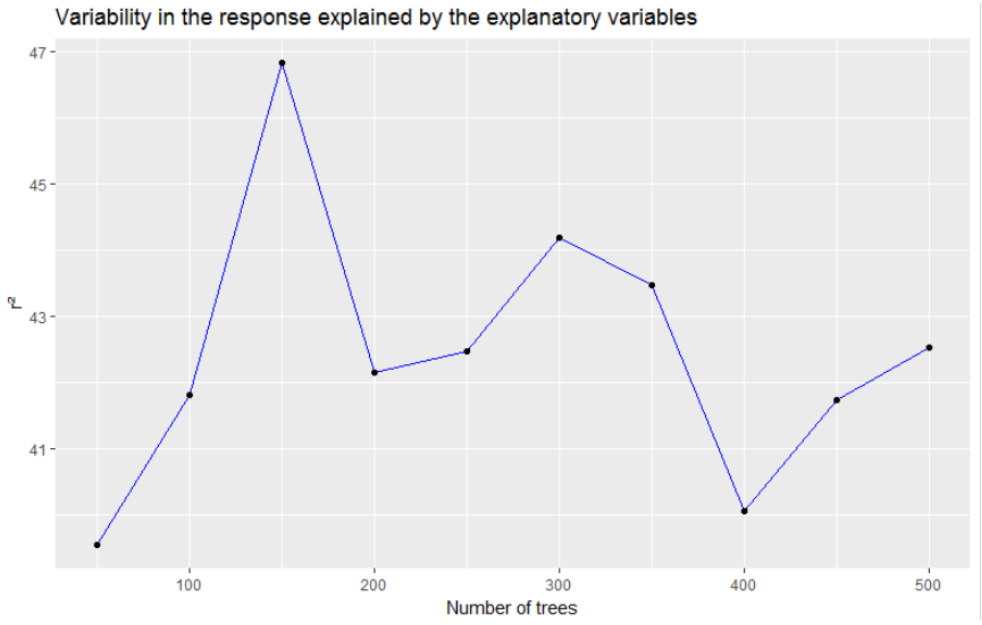


Ilustración 47 Gráfica % Variabilidad Total vs Ntree

En la gráfica, se puede observar que para un valor de árboles de 150, se obtiene la mayor variabilidad explicada total del modelo. Ahora, si calculamos las gráficas del error, tendríamos que obtener el menor error en el mismo punto:

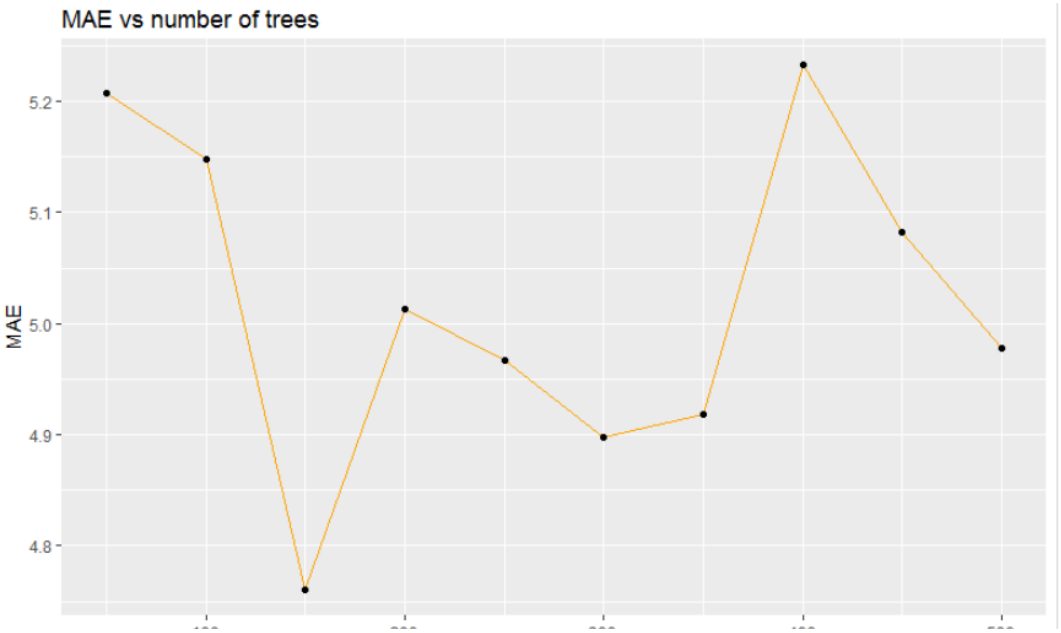


Ilustración 48 Gráfica MSE vs Ntree

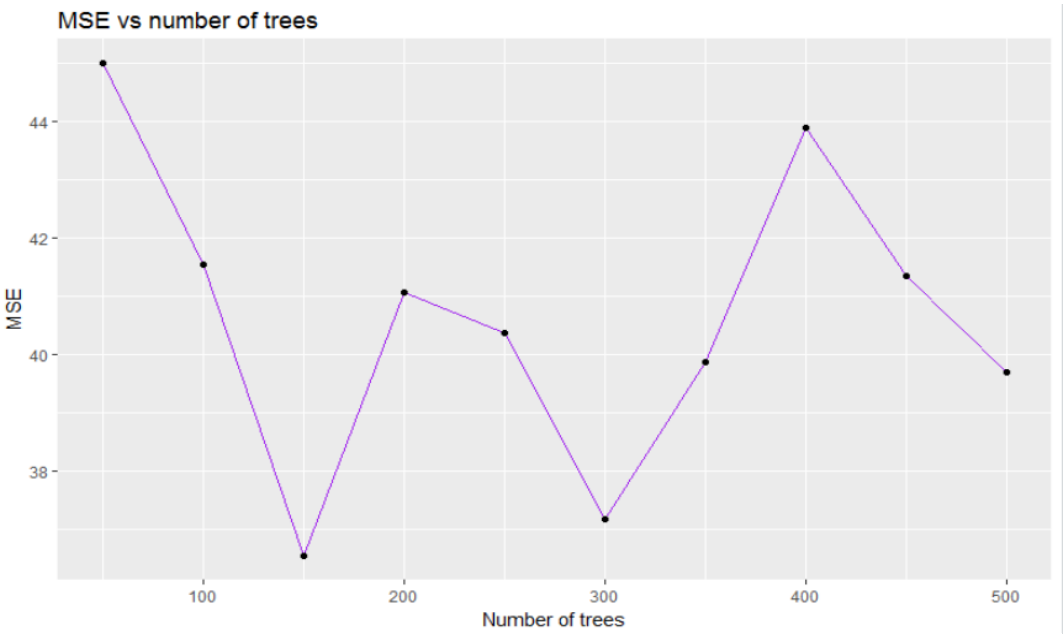
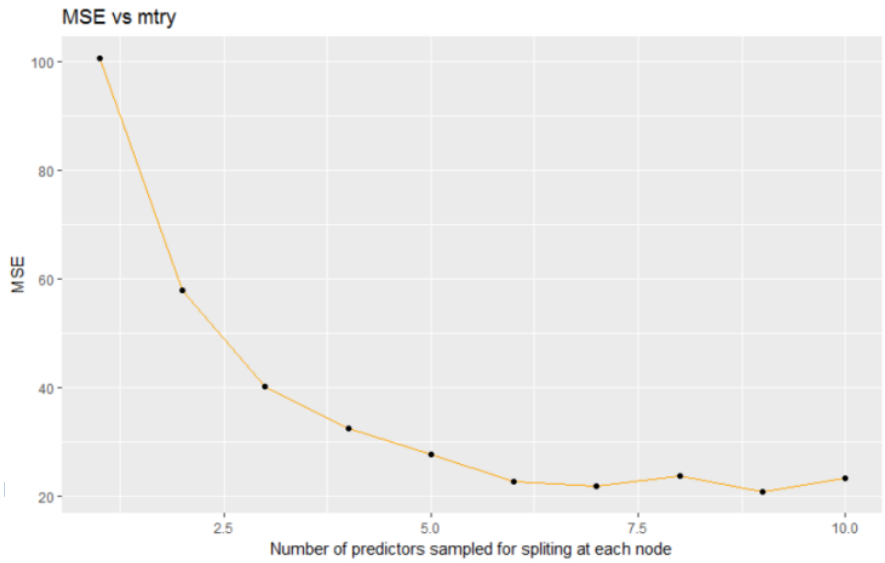


Ilustración 49 Gráfica MAE vs Ntree

Efectivamente, obtenemos el mismo punto, número de árboles igual 150 como mejor punto.

4.5.1.2 Experimentos en Mtry

Ahora si realizamos los mismos experimentos para el número de variables seleccionadas en cada partición.



Se observa claramente, como cuantas más variables se seleccionen, mejor se explicará el modelo. Sin embargo, no es aconsejable coger todas las variables en cada partición. Debido a que con valores altos, se le resta aleatoriedad al bosque pudiendo tener problemas de overfitting. Si volvemos a calcular las gráficas para las medidas del error se obtiene:

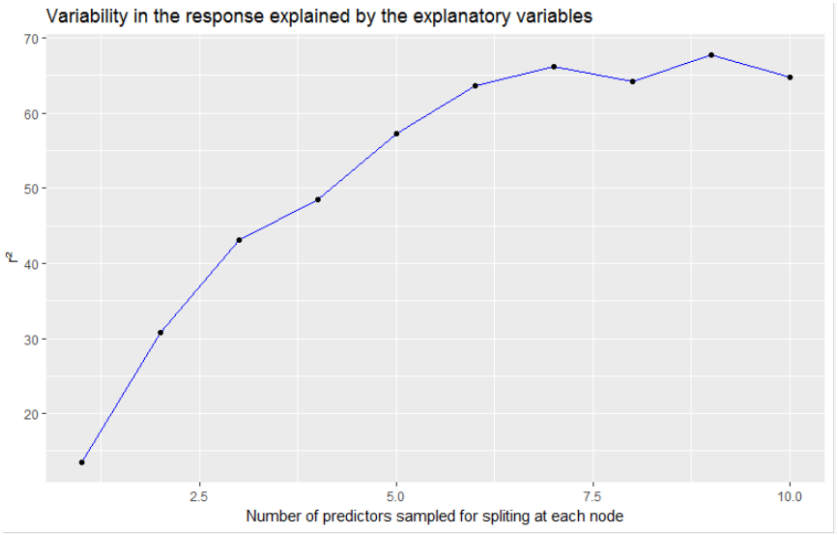


Ilustración 51 Gráfica MSE vs Mtry

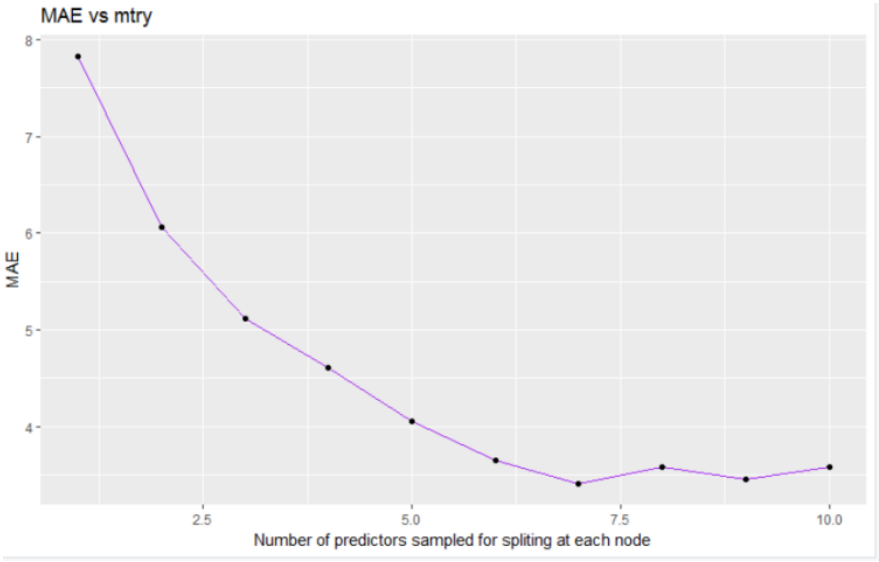


Ilustración 52 Gráfica MAE vs Mtry

Una vez realizados estos experimentos, obtenemos que los valores óptimos para realizar el modelo *random forest* son:

- Número de árboles: 150
- Número de variables elegidas en cada partición: 5

Con estos valores óptimos, se realizará el modelo *random forest*, y sobre este modelo, obtendremos el análisis ANOVA. Se calcularán, las variabilidades de cada factor y de sus interacciones al igual que se realizó para el modelo paramétrico. En el modelo *random forest* (modelo no paramétrico), se han obtenido los siguientes resultados de las variabilidades explicadas:

Tabla 8 ANOVA Modelo RF

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	1252.2	1252.2	195.673	1.98e-12	***
X3	1	884.9	884.9	138.277	5.86e-11	***
X4	1	0.0	0.0	0.005	0.9431	
X5	1	330.9	330.9	51.705	3.31e-07	***
X6	1	429.4	429.4	67.099	3.98e-08	***
X7	1	146.7	146.7	22.924	8.82e-05	***
X8	1	373.5	373.5	58.364	1.26e-07	***
X9	1	30.0	30.0	4.696	0.0413	*
X10	1	11.3	11.3	1.766	0.1975	
Residuals	22	140.8	6.4			

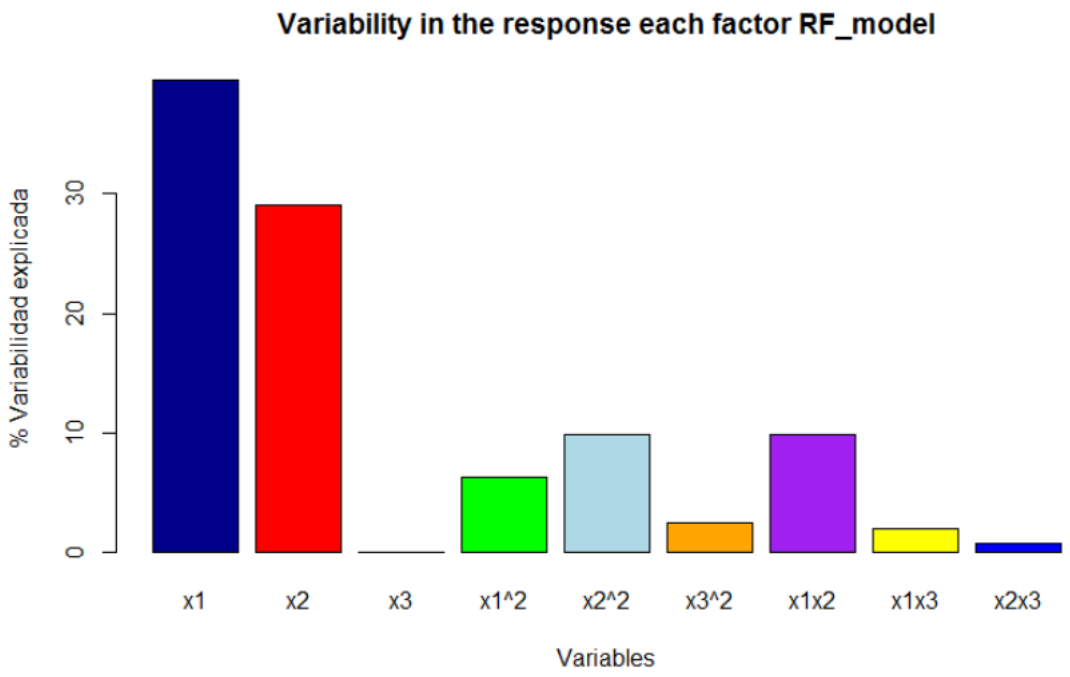


Ilustración 53 % Variabilidad Factores Modelo RF

Variable	%Variabilidad sobre el modelo
X1	36.62%
X2	25.82%
X3	-0.18%
X_1^2	9.53%
X_2^2	12.43%
X_3^2	4.12%
X_1X_2	10.79%
X_1X_3	0.69%
X_2X_3	0.14%

Se puede observar como en este modelo la variabilidad de la respuesta debida a los factores es menor, esto es debido a que las contribuciones en variabilidad por las interacciones son mayores.

Por lo tanto, se puede concluir que el modelo *random forest*, explicará mejor el modelo, ya que explica de mejor manera las posibles interacciones del modelo.

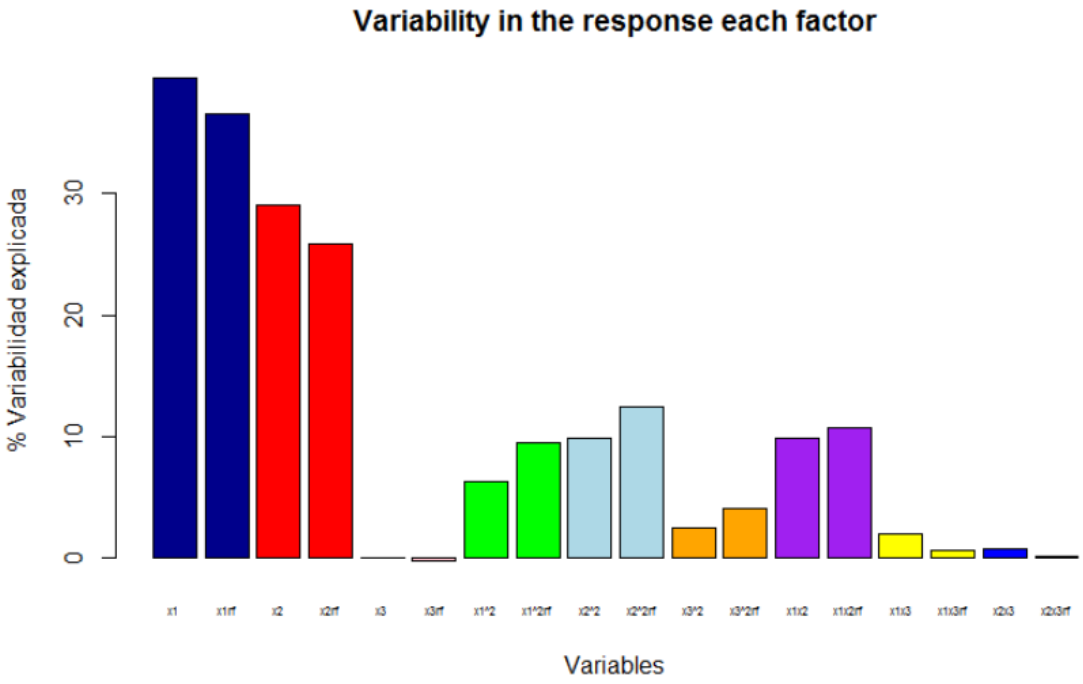


Ilustración 55 Gráfica Comparación Modelo Paramétrico vs RF

En este gráfico, se pueden observar ambos modelos juntos. Se ha señalado la misma variable en el mismo color para los dos modelo, siendo la variable de la izquierda la del modelo paramétrico y la de la derecha la del modelo *random forest*. Se puede ver claramente como en el modelo *random forest*, la variabilidad de las interacciones es mayor que en el modelo paramétrico, mientras que la variabilidad de los factores es menor.

Si se comparan los cuatro ANOVA (datos originales, random Forest modelo original, modelo paramétrico y random Forest del modelo paramétrico) se observa como los primeros son muy similares, sin embargo, al existir interacciones entre las variables, el modelo original no es capaz de mostrar las mismas al tomar el modelo como un modelo lineal. El modelo random forest de los datos originales tiene mayor variabilidad en X_1, esto se puede deber a que este ha encontrado las interacciones existentes, pero no es capaz de mostrarlas por separado sino que suma las interacciones de X_1 a la variabilidad total de la misma.

Por último, observamos como los modelos ANOVA del random forest del modelo paramétrico y el ANOVA original difieren mucho. Esto se debe a que el modelo original una vez más era incapaz de reconocer las interacciones, pero una vez que se han puesto como condición en el modelo, se ha visto que en efecto eran importantes y por lo tanto este diseño explicaría mejor las variables e interacciones del modelo.

Capítulo 5

Experimentos con un modelo Real

5.1 Introducción

En el capítulo anterior, se ha hecho uso de un modelo ortogonal, obtenido mediante el método de la codificación de variables. En este capítulo, se va a calcular la ortogonalidad a partir de datos reales no ortogonales. A continuación, se explicará la metodología que se va a seguir y se van a realizar los mismos experimentos que en los capítulos anteriores.

5.2 Método de Codificación de Variables

Existen varias ventajas al utilizar variables codificadas en vez de las variables originales, cuando se está ajustando un modelo polinómico. Las principales ventajas son:

- Aumento en la facilidad computacional y mejora de la precisión en la estimación del modelo.
- Una mejor interpretabilidad de las estimaciones de los coeficientes en el modelo.

Los posibles métodos de obtención de variables codificadas son los siguientes:

- Diseño 2^k factorial
- Réplicas fraccionarias de los diseños factoriales 2^k
- Diseños Simplex
- Diseños Plackett.Burman

En este Trabajo de Fin de Grado, se va hacer uso del primer método, el diseño 2^k factorial.

En este diseño, cada factor o variable de entrada son medidos en dos niveles que son codificados para tener el valor de -1 cuando se encuentra en el nivel bajo y valen 1 cuando se encuentran en el nivel alto. Teniendo en cuenta todas las posibles combinaciones de los niveles de los K factores, se obtiene una matriz **D** con las variables codificadas formada por 2^k filas. En la fila, n los elementos son iguales a 1 o -1 y representan las coordinadas del punto de diseño de las iteración n del experimento.

Cada fila, por lo tanto, representa una combinación de los niveles de los factores. La codificación por tanto se puede obtener mediante la siguiente fórmula:

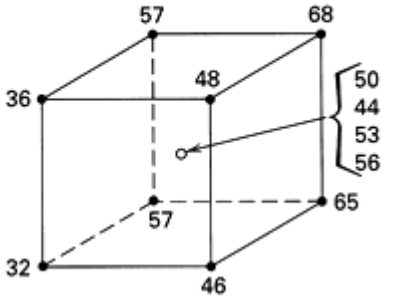
$$X_{ni} = \frac{2(X_{ni} - \bar{X}_i)}{R_i}$$

Donde X_{ni} es el valor original de la X en el factor i en la n iteración, mientas que $\bar{X}_i$ es la media de los niveles alto y bajo del factor i. Por último, R_i es el rango entre los niveles alto y bajo. De esta manera obtendremos las variables codificadas y por tanto ortogonales de cualquier set de datos reales.

5.3 Experimentos Modelo Real

Una vez explicado el método de obtención de las variables codificadas se va a realizar para un modelo real. Este modelo, representa el rendimiento de un proceso, que será la salida del modelo en función de tres variables como son la temperatura (T), Presión (psig) y la concentración (g/l).

Process Variables			
Run	Temp (°C)	Pressure (psig)	Conc. (g/l)
1	120	40	15
2	160	40	15
3	120	80	15
4	160	80	15
5	120	40	30
6	160	40	30
7	120	80	30
8	160	80	30
9	140	60	22.5
10	140	60	22.5
11	140	60	22.5
12	140	60	22.5



En primer lugar, se van a codificar las variables de entrada siguiendo el método de diseño factorial, en este caso un diseño 2^3, por tener tres variables de entrada.

$$x_1 = \frac{\text{Temp} - 140}{20}, \quad x_2 = \frac{\text{Pressure} - 60}{20}, \quad x_3 = \frac{\text{Conc} - 22.5}{7.5}$$

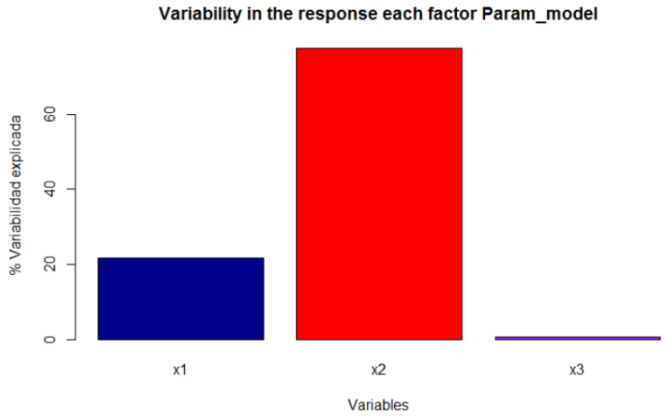
Con estas ecuaciones, obtendremos las siguientes X codificadas:

Coded Variables		
x_1	x_2	x_3
−1	−1	−1
1	−1	−1
−1	1	−1
1	1	−1
−1	−1	1
1	−1	1
−1	1	1
1	1	1
0	0	0
0	0	0
0	0	0
0	0	0

En primer lugar, se realiza el ANOVA a los datos originales, obteniendo:

Tabla 10 ANOVA modelo Original

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X1	1	253.1	253.1	1.566e+31	<2e-16	***
X2	1	903.1	903.1	5.586e+31	<2e-16	***
X3	1	10.1	10.1	6.263e+29	<2e-16	***
Residuals	8	0.0	0.0			



Si se analizan las posibles interacciones que pueda haber en el modelo obtenemos:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	253.1	253.1	12.796	0.0232	*
x2	1	903.1	903.1	45.656	0.0025	**
x3	1	10.1	10.1	0.512	0.5139	
x1:x2	1	6.1	6.1	0.310	0.6076	
x1:x3	1	0.1	0.1	0.006	0.9405	
x2:x3	1	1.1	1.1	0.057	0.8232	
x1:x2:x3	1	3.1	3.1	0.158	0.7113	
Residuals	4	79.1	19.8			

Tabla 11 ANOVA Modelo Original

Se ve como las variables si se toman do los datos originales tienen interacciones entre las variables. Por lo tanto, se realizará la estimación del modelo de regresión con interacciones tal y como se realizó en el capítulo anterior.

Ahora si estimamos un modelo random forest sobre estos datos originales:

Tabla 12 ANOVA RF Modelo Original

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X1	1	45.03	45.03	132.527	2.94e-06	***
X2	1	213.50	213.50	628.277	6.87e-09	***
X3	1	0.42	0.42	1.231	0.299	
Residuals	8	2.72	0.34			

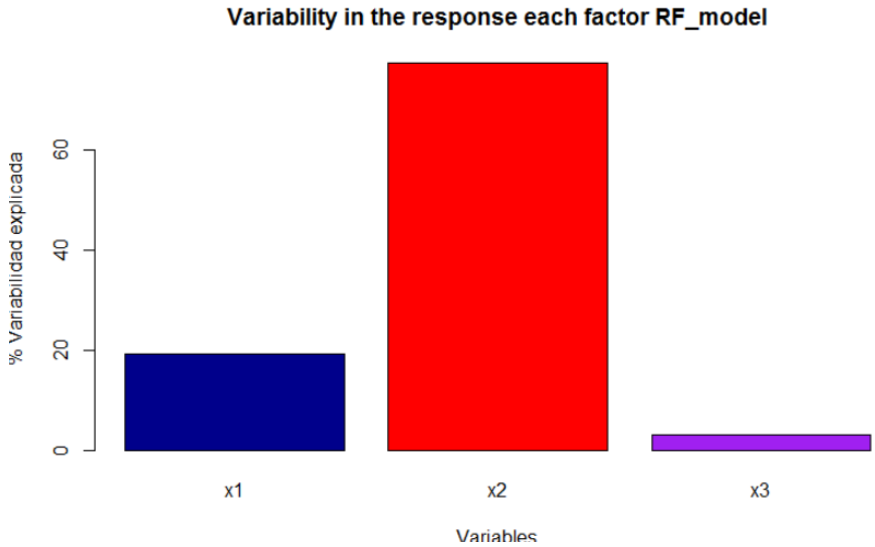


Ilustración 58 %Variabilidad factores RF

Ahora, se va a considerar el modelo como un modelo de regresión simple y se va a observar si existen diferencias significativas. Una vez que se ha realizado la regresión se calculará la nueva salida. A este modelo, se le realizarán los mismos experimentos que se han ido practicando a lo largo de todo este Trabajo de Fin de Grado. Con estos experimentos, probaremos para un sistema de datos reales, el hecho de que el modelo *random forest* debería en un principio explicar mejor el modelo que un modelo paramétrico.

El modelo que se supone va a seguir la siguiente ecuación para ver las interacciones significativas entre X1 y X_2

Con las X codificadas y las salidas del modelo, se estima el parámetro beta haciendo la regresión:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 68 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Por lo tanto, ahora con este parámetro, obtenemos que la ecuación del modelo sea:

$$\hat{Y} = 50.75 + 5.625X_1 + 10.625X_2 + 0.375X_1^2 - 0.875X_1X_2$$

Tabla 13 ANOVA Modelo Lineal

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	253.1	253.1	3.651e+30	<2e-16	***
X3	1	903.1	903.1	1.303e+31	<2e-16	***
X4	1	0.4	0.4	5.409e+27	<2e-16	***
X5	1	6.1	6.1	8.835e+28	<2e-16	***
Residuals	7	0.0	0.0			

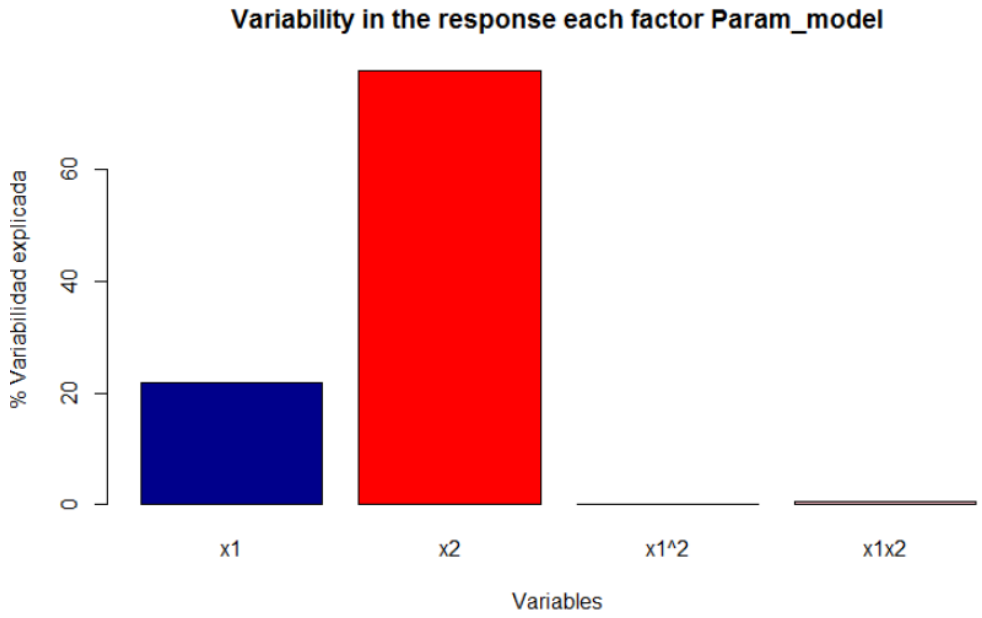


Ilustración 59 Gráfica % Variabilidad Factores Modelo Paramétrico

Modelo	% Variabilidad Cada Factor
$\mathbf{X_1}$	21%
$\mathbf{X_2}$	77.67%
$\mathbf{X_1^2}$	0.03%
$\mathbf{X_1X_2}$	0.56%

Se puede observar, cómo según el análisis ANOVA al modelo paramétrico se establece que la presión y la temperatura van a tener un efecto mucho más importante en el rendimiento que la concentración. Ahora ajustando el modelo *random forest* y realizando el análisis de la variancia sobre estos datos:

Tabla 15 ANOVA Modelo RF

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	25.80	25.80	3.385e+29	<2e-16	***
X3	1	78.98	78.98	1.036e+30	<2e-16	***
X4	1	1.77	1.77	2.325e+28	<2e-16	***
X5	1	0.11	0.11	1.459e+27	<2e-16	***
Residuals	7	0.00	0.00			

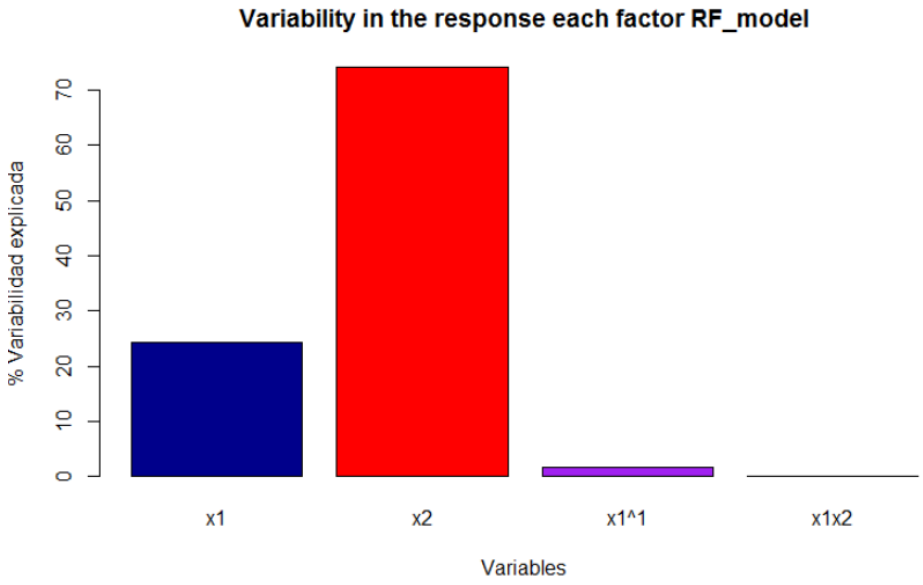


Ilustración 61 Gráfica % Variabilidad Factores Modelo RF

Modelo	% Variabilidad Cada Factor
X_1	24%
X_2	74%
X_1^2	1.76%
X_1X_2	0.1%

Se observa cómo una vez más, la presión es la variable con más importancia, sin embargo, en esta ocasión la temperatura explica en un porcentaje menor el modelo. En cambio la interacción entre X_1 y X_2 no resulta tan significativa como en el modelo original.

Ahora, si se muestran ambos valores en una sola gráfica para poder comparar ambos modelos, se pueden observar que una vez más al considerar el modelo como un modelo paramétrico con interacciones obtienen resultados parecidos, pero la variabilidad diferente del modelo random forest Se observa claramente, como se obtienen resultados muy similares al caso estudiado anteriormente, en el que el random forest es capaz de captar las interacciones reales del modelo.

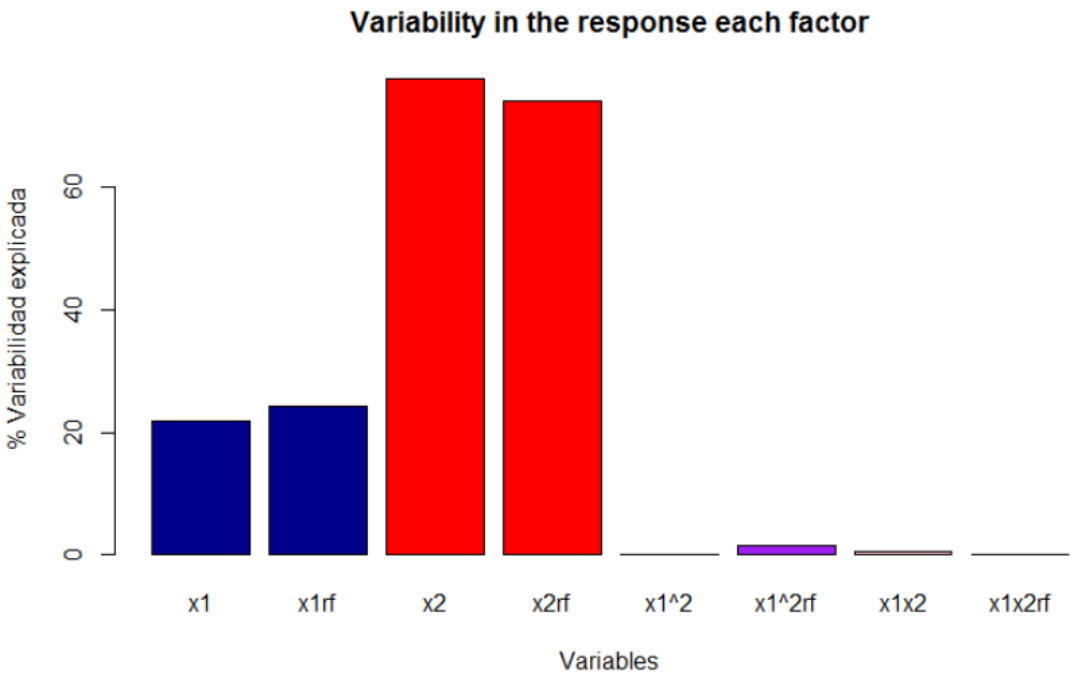


Ilustración 63 % Variabilidad Modelo Paramétrico vs Random Forest

Capítulo 6

Conclusiones y Líneas Futuras

6.1 Conclusión

En el presente Trabajo de Fin de Grado, se ha tratado de analizar la importancia de las variables de un modelo y sus interacciones, analizando la variabilidad que introducen en la respuesta del modelo. Para ello, se analizaron dos modelos: un modelo lineal sin interacciones entre las variables y otro modelo con interacciones entre las mismas. El objetivo, era analizar estos modelos con la herramienta *random forest* y probar que se obtienen mejores resultados que cuando se hace un análisis lineal del mismo. Para probar estos resultados, se realizaron mediante un análisis ANOVA de los datos.

Por lo tanto, por un lado se han sacado conclusiones de la comparación entre los diferentes análisis así como de la propia herramienta *random Forest*.

- En primer lugar, se obtiene la conclusión de que el modelo *random forest* es un modelo mucho más fiable y con mejores resultados en cuanto a la variabilidad explicada de cada factor sobre la respuesta del modelo, que el modelo paramétrico cuando se tienen interacciones entre sus variables. Esto se debe a que el *random forest* es capaz de captar las posibles interacciones entre las variables cuando se ha estimado un modelo lineal.

- En modelos en los que se ha estimado un diseño con interacciones entre las variables, el modelo random forest es capaz de explicar estas interacciones de mejor manera que el modelo paramétrico.

- En modelos con interacciones, se ha probado que si no se tienen variables 100% ortogonales, es decir, independientes entre sí, no se podrá hacer de manera correcta el análisis ANOVA.

También se han podido sacar conclusiones sobre la herramienta *random forest*:

- En la experimentación de la obtención de los parámetros óptimos de la función *random forest*, se puede concluir que en cuanto al número de variables que se seleccionan en cada split, Mtry, el valor para el cual se obtiene la mayor variabilidad explicada total del modelo sin restar el carácter aleatorio a la herramienta e introducir overfitting, va a ser de p/3 o un valor un poco mayor.

- En modelos con pocas variables, con árboles no muy profundos se va a obtener la mayor explicación total del modelo.

- Es muy difícil que aparezcan problemas de overfitting.

6.2 Líneas Futuras

Este proyecto de investigación debido a que trata temas muy novedosos y poco estudiados en profundidad, tiene una gran variedad de posibles líneas futuras de investigación:

- Realización de experimentos en modelos más complejos y con mayores interacciones entre las variables.

- Experimentación con datos ortogonales reales.

- Experimentación en modelos con varias respuestas de salida en vez de una sola respuesta como se ha hecho en este trabajo.

Capítulo 7

Planificación Temporal y Presupuesto

7.1 Estructura de Descomposición del Proyecto

La EDP, Estructura de Descomposición del proyecto, es una herramienta que permite representar de forma esquemática las actividades que se han realizado a lo largo del proyecto. La EDP, tiene una estructura jerárquica dividiendo el proyecto en actividades, hasta que se llega a la unidad mínima, que formarán los entregables del proyecto.

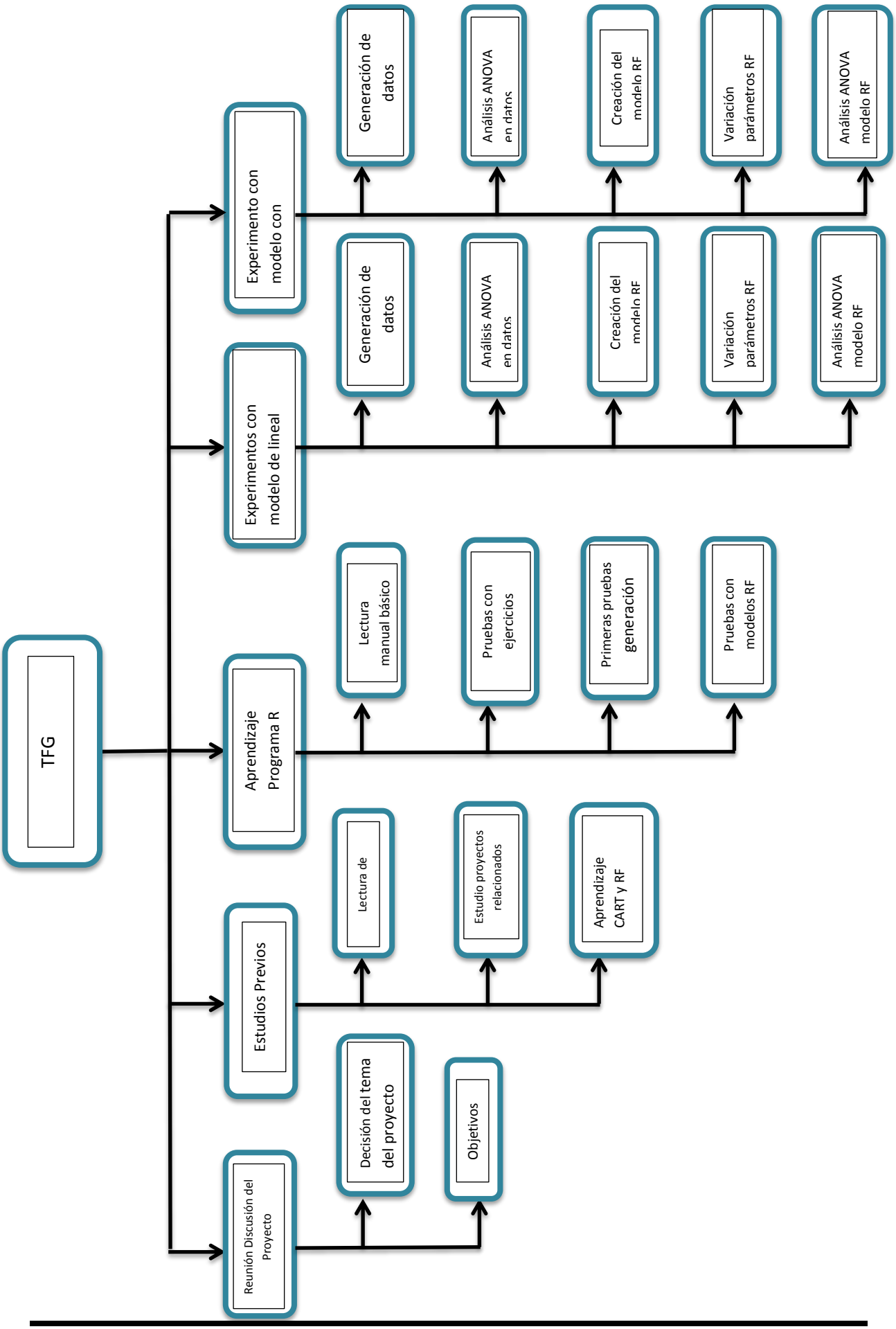
Cabe destacar, que no informa de la cronología de las actividades, pudiéndose solapar a lo largo del proyecto como se observará en el posterior diagrama de GANTT. Por lo tanto, su tamaño debe ser contenido para su interpretación rápida.

La EDP, de este trabajo de fin de grado se adjuntará a continuación.

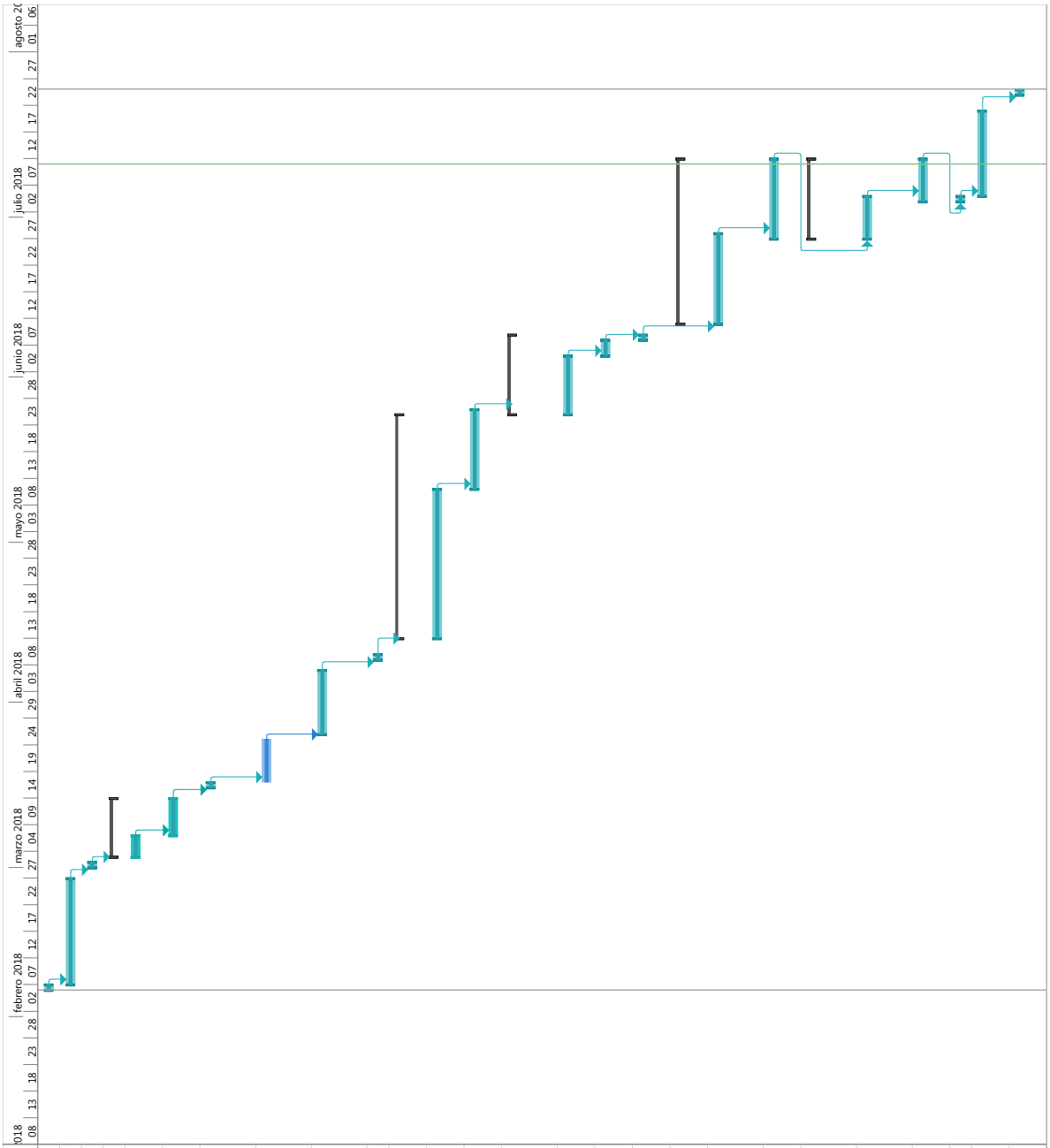
7.2 Diagrama de GANTT

El diagrama de GANTT, es una herramienta que permite a través de un diagrama de barras de exponer la cronología de todas las actividades de un proyecto. Sin embargo, en el diagrama de Gantt no se puede observar la relación entre actividades, por lo que la EDP será un complemento de este diagrama para ver estas observaciones.

El diagrama de Gantt al igual que la EDP, se adjuntarán a continuación.


TFG
Reunión Discusión del Proyecto
Decisión del tema del proyecto
Objetivos
Estudios Previos
Lectura de
Estudio proyectos relacionados
Aprendizaje CART y RF
Aprendizaje Programa R
Lectura manual básico
Pruebas con ejercicios
Primeras pruebas generación
Pruebas con modelos RF
Experimentos con modelo de lineal
Generación de datos
Análisis ANOVA en datos
Creación del modelo RF
Variación parámetros RF
Análisis ANOVA modelo RF
Experimento con modelo con
Generación de datos
Análisis ANOVA en datos
Creación del modelo RF
Variación parámetros RF
Análisis ANOVA modelo RF

Id	Nombre de tarea	Duración	Comienzo	Fin
1	Tutoría Inicial	1 día	mar 06/02/18	mar 06/02/18
2	Lectura de material	14 días	mié 07/02/18	lun 26/02/18
3	Definición proyecto	1 día	jue 01/03/18	jue 01/03/18
4	**Aprendizaje R**	**7 días**	**sáb 03/03/18**	**mar 13/03/18**
5	4.1 Lectura manual R	3 días	sáb 03/03/18	mar 06/03/18
6	4.2 Practica de codigos sencillos	5 días	mié 07/03/18	mar 13/03/18
7	Reunión resolución de dudas con Ismael	1 día	vie 16/03/18	vie 16/03/18
8	Primeros experimentos con R	5 días	sáb 17/03/18	sáb 24/03/18
9	Redacción Introducción Trabajo	10 días	lun 26/03/18	vie 06/04/18
10	Reunión Intermedia	1 día	lun 09/04/18	lun 09/04/18
11	**Experimentos con el modelo lineal**	**30 días**	**vie 13/04/18**	**jue 24/05/18**
12	12.1 Generación de datos	20 días	vie 13/04/18	jue 10/05/18
13	12.2 Experimentos en RF	11 días	vie 11/05/18	vie 25/05/18
14	**Experimentos con el modelo con interacciones**	**11 días**	**vie 25/05/18**	**vie 08/06/18**
15	15.1 Generación datos del modelo	7 días	vie 25/05/18	lun 04/06/18
16	15.2 Experimentos en RF	3 días	mar 05/06/18	jue 07/06/18
17	Reuión Ismael resolución dudas	1 día	vie 08/06/18	vie 08/06/18
18	**Experimento modelo lineal**	**23 días**	**lun 11/06/18**	**mié 11/07/18**
19	19.1 Experimentos en Ntree	13 días	lun 11/06/18	mié 27/06/18
20	19.2 Experimentos en Mtry	11 días	mié 27/06/18	mié 11/07/18
21	**Experimentos en modelo con interacciones**	**11 días**	**mié 27/06/18**	**mié 11/07/18**
22	22.1 Experimentos en Ntee	6 días	mié 27/06/18	mié 04/07/18
23	22.2 Experimentos en Mtry	6 días	mié 04/07/18	mié 11/07/18
24	Reunión final	1 día	mié 04/07/18	mié 04/07/18
25	Redacción final de trabajo	12 días	jue 05/07/18	vie 20/07/18
26	Impresión del trabajo	1 día	mar 24/07/18	mar 24/07/18



Id	Nombre de tarea	Duración	Comienzo	Fin
1	Tutoría Inicial	1 día	mar 06/02/18	mar 06/02/18
2	Lectura de material	14 días	mié 07/02/18	lun 26/02/18
3	Definición proyecto	1 día	jue 01/03/18	jue 01/03/18
4	**Aprendizaje R**	**7 días**	**sáb 03/03/18**	**mar 13/03/18**
5	4.1 Lectura manual R	3 días	sáb 03/03/18	mar 06/03/18
6	4.2 Practica de codigos sencillos	5 días	mié 07/03/18	mar 13/03/18
7	Reunión resolución de dudas con Ismael	1 día	vie 16/03/18	vie 16/03/18
8	Primeros experimentos con R	5 días	sáb 17/03/18	sáb 24/03/18
9	Redacción Introducción Trabajo	10 días	lun 26/03/18	vie 06/04/18
10	Reunión Intermedia	1 día	lun 09/04/18	lun 09/04/18
11	**Experimentos con el modelo lineal**	**30 días**	**vie 13/04/18**	**jue 24/05/18**
12	12.1 Generación de datos	20 días	vie 13/04/18	jue 10/05/18
13	12.2 Experimentos en RF	11 días	vie 11/05/18	vie 25/05/18
14	**Experimentos con el modelo con interacciones**	**11 días**	**vie 25/05/18**	**vie 08/06/18**
15	15.1 Generación datos del modelo	7 días	vie 25/05/18	lun 04/06/18
16	15.2 Experimentos en RF	3 días	mar 05/06/18	jue 07/06/18
17	Reuión Ismael resolución dudas	1 día	vie 08/06/18	vie 08/06/18
18	**Experimento modelo lineal**	**23 días**	**lun 11/06/18**	**mié 11/07/18**
19	19.1 Experimentos en Ntree	13 días	lun 11/06/18	mié 27/06/18
20	19.2 Experimentos en Mtry	11 días	mié 27/06/18	mié 11/07/18
21	**Experimentos en modelo con interacciones**	**11 días**	**mié 27/06/18**	**mié 11/07/18**
22	22.1 Experimentos en Ntee	6 días	mié 27/06/18	mié 04/07/18
23	22.2 Experimentos en Mtry	6 días	mié 04/07/18	mié 11/07/18
24	Reunión final	1 día	mié 04/07/18	mié 04/07/18
25	Redacción final de trabajo	12 días	jue 05/07/18	vie 20/07/18
26	Impresión del trabajo	1 día	mar 24/07/18	mar 24/07/18

7.3 Presupuesto

En este apartado, al igual que se hace para cualquier proyecto, se va a exponer el coste que ha conllevado este Trabajo de Fin de Grado.

Para realizar esta estimación se ha tenido en cuenta que este trabajo es un trabajo de simulación en el que software empleado no tiene ningún coste y no se ha tenido que adquirir ningún componente físico. Por lo tanto, el coste total de este proyecto viene dado por las horas trabajadas en el mismo. No obstante, para la redacción del trabajo se ha usado el paquete de Microsoft Office, por lo que el coste de sus licencia si se tendrá en cuenta.

Para obtener el coste por las horas trabajadas en el proyecto, hay que tener en cuenta en primer lugar, que el salario de un ingeniero sin graduar en prácticas ronda los 10€/hora. Estimando que se han trabajado un 360 horas, obtenemos un coste de 3600 €. En cuanto al salario de los dos tutores que han llevado este proyecto, se estima un salario de unos 40€/h con unas 15 y 30 horas trabajas respectivamente, que conllevan un coste total de 1800€.

En la siguiente tabla, se expone de forma clara el coste total de proyecto:

Concepto	Unidades	Coste unitario	Total
Salario Alumno	360 horas	10€/hora	3600€
Salario Tutor 1	15 horas	40€/hora	600€
Salario Tutor 2	30 horas	40€/hora	1200€
Licencia Paquete Office	1	59€	59€
Total sin IVA			5459€
IVA (21%)			1146.39€
Total			6605.39€

Capítulo 8

Bibliografía

1. Trevor Hastie, Robert Tibisharami, Jerome Friedman, "*The Elements of Statistical Learning, Data Mining, Inference and Prediction."* Standford. California.

2. André I. Khuri, John A. Cornell, "*Response Surfaces, Desing and Analyses,"* Statistics: textbooks and monographs. Volume 61.

3. Leo Breiman, *Random Forests*, The University of Berkley, California, https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (2001).

4. Penn State Eberly College of Sciene. *"Random Forest Online Course" https://onlinecourses.science.psu.edu/stat857/node/220/*.

5. Gilles Louppe, Louis Wehenkel, Antonio Sutera and Pierre Geurts "*Understanding Variable Importance in Random Forest"* Dept. of EE & CS, University of Liege, Belgium.

6. ANDREW GELMAN *"ANALYSIS OF VARIANCE—WHY IT IS MORE IMPORTANT THAN EVER"* Columbia University, New York.

7. Wei-Yin Loh *"Classification and regression trees"*.

8. Gerard Biau, "*Analysis of a Random Forests Model",* Universite Pierre et Marie Curie – Paris VI.

Índice de Figuras

Índice de Tablas

Códigos de R

1. Modelo Lineal

```
set.seed(1234)

N <- 1000 # Number of observations

p <- 10

normal <- rnorm.sobol(n = p, dimension = N, scrambling = 2) # Creates a matrix n(=p) by dimension(=N)

# for testing purpose

r1normal <- normal[1,] # First row of the matrix

hist(r1normal, main="Sobol pseudo-normal random numbers", border="blue", col="green")

# noise vector

noise_var <- rep(sqrt(0.5),N)

noise_mean <- rep(0,N)

noise_z<-rnorm.sobol(n = 1, dimension = N, scrambling = 2)

noise<- noise_mean + noise_var*noise_z

hist(noise, main="Noise", border="red", col="blue")


# Design your covariance ma

cov_mat <- matrix(0,p,p,byrow = TRUE)

diag(cov_mat)=1

c_matrix <- cov2cor(cov_mat)

# Cholesky decomposition

# If the measures of correlation used are product-moment coefficients, the correlation matric

# is the same as the covariance matrix of the standardized random variables

CH<-(chol(cov_mat))

CHT <- t(CH)

# Create the matrix whose rows are the covariates that we are looking for (correlated covariates)

X <- CHT%*%normal
```

```
XT <- t(X) # Matrix N by p
# Compute the correlation matrix to check
corr_mat <- cor(XT)
round(corr_mat, 2)
# vector of betas
beta <- c(5,5,2,2,1,-5,-5,-2,-2,-1)
# construct the model
beta_XT <- matrix(0,nrow = N,ncol = p,byrow = FALSE)
for(i in 1:N){for (j in 1:p){
  beta_XT[i,j] <- (beta[j]*XT[i,j])
}}
# Matrix beta*XT (each column is beta by the corresponding covariate)
y0 <- rowSums (beta_XT, na.rm = FALSE, dims = 1)
# The simulated y
y_sim <- t(y0 + noise)
mod.data <- data.frame(beta_XT, y0)
colnames(mod.data) <- c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10","y0")

nt <- c(50,100,150,200,250,300,350,400,450,500)
mtry <- 1:10
# Run Random forest algorithm
split<-sample.split(mod.data$y0,SplitRatio = 2/3)
trainSet<-subset(mod.data,split==TRUE)
testSet<-subset(mod.data,split==FALSE)
AOV=aov(formula= y0~.,data = testSet)
 r2nt <- c()
 mse1 <- c()
  mae1 <- c()
  mape1 <- c()
```

```
 for (mtree in nt)
 {
   rf.Regression <- randomForest(y0 ~ ., data=trainSet,ntree=mtree)
     pred <- predict(rf.Regression, newdata = testSet)
   SSM <- sum((pred - mean(testSet$y0))^2)
   SST <- sum((testSet$y0 - mean(testSet$y0))^2)
   r2a <- SSM / SST # Variability in the response explained by the explanatory variables
 r2nt <- c(r2nt,r2a)
 mse1 <- c(mse1, MSE(y_pred =pred , y_true = testSet$y0))
   mae1 <- c(mae1, MAE(y_pred =pred , y_true = testSet$y0))}
 r2nt <- 100 * r2nt
 dfr2 <- data.frame(r2nt,nt)
ggplot(data=dfr2, aes(x=nt, y=r2nt, group=1)) + geom_line(linetype = "solid", color="blue")
geom_point() + labs(x = "Number of trees", y = "r²", title = "Variability in the response
explained by the explanatory variables")
dfmse1 <- data.frame(mse1,nt)ggplot(data=dfmse1, aes(x=nt, y=mse1, group=1)) +
geom_line(linetype = "solid", color="purple") +geom_point() + labs(x = "Number of trees", y =
"MSE", title = "MSE vs number of trees")
dfmae1 <- data.frame(mae1,nt)
ggplot(data=dfmae1, aes(x=nt, y=mae1, group=1)) + geom_line(linetype = "solid",
color="orange") + geom_point() +  labs(x = "Number of trees", y = "MAE", title = "MAE vs
number of trees")
 r2nt2 <- c()
 mse2 <- c()
 mae2 <- c()
 mape2 <- c()
 sq1= c()
 for (mt in mtry)
 {
   rf.Regression <- randomForest(y0 ~ ., data=trainSet,ntree=100L, mtry = mt)
   pred <- predict(rf.Regression, newdata = testSet)
```

```
  SSM <- sum((pred - mean(testSet$y0))^2)

  SST <- sum((testSet$y0 - mean(testSet$y0))^2)

  r2a <- SSM / SST

  r2nt2 <- c(r2nt2,r2a)

  mse2 <- c(mse2, MSE(y_pred =pred , y_true = testSet$y0))

  mae2 <- c(mae2, MAE(y_pred =pred , y_true = testSet$y0))  }

 r2nt2 <- 100 * r2nt2

 dfr22 <- data.frame(r2nt2,mtry)

 ggplot(data=dfr22, aes(x=mtry, y=r2nt2, group=1)) +geom_line(linetype = "solid",
color="blue") +  geom_point() +labs(x = "Number of predictors sampled for spliting at each
node", y = "r²", title = "Variability in the response explained by the explanatory variables")

 dfmae2 <- data.frame(mae2,mtry)

 ggplot(data=dfmae2, aes(x=mtry, y=mae2, group=1)) +geom_line(linetype = "solid",
color="purple") + geom_point() +labs(x = "Number of predictors sampled for spliting at each
node", y = "MAE", title = "MAE vs mtry")
```

2. Modelo con Interacciones

```
Y <- c(32,46,57,65,36,48,57,68,50,44,53,56)

x0 <- c(1,1,1,1,1,1,1,1,1,1,1,1)

x1 <- c(-1,1,-1,1,-1,1,-1,1,0,0,0,0)

x2 <- c(-1,-1,1,1,-1,-1,1,1,0,0,0,0)

x3<-c(-1,-1,-1,-1,1,1,1,1,0,0,0,0)

X <- matrix(c(x0,x1,x2,x3),nrow=length(x0))


# Regression coefficients b (Parametric model)

b <- solve(t(X)%*%X)%*%t(X)%*%Y

Y1=t(b)%*%t(X)

Y1t=t(Y1)

mod.data1=data.frame(X,Y1t)

npk.aov <- aov(Y1t ~ . , mod.data1)
```

```
s_of_s <- as.matrix(anova(npk.aov)["Sum Sq"]) # Sum of Squares

d_of_f <- as.matrix(anova(npk.aov)["Df"])

m_of_s <- as.matrix((s_of_s / d_of_f)) # Mean of squares

var_comp <- c()

for (i in 1:3){

 aux <- (m_of_s[i] - m_of_s[4])/3

 var_comp <- c(var_comp,aux)}

Perc_var <- (var_comp / sum(var_comp))*100

barplot(Perc_var, main="Variability in the response each factor Param_model", horiz=FALSE,
names.arg=c("x1","x2","x3"),xlab = "Variables", ylab = "% Variabilidad explicada", col =
c("darkblue","red", "purple"))

mtree <- 150 # Optimal

mtry <- 1 # Optimal

rf.Regression <- randomForest(Y1t ~ ., data=mod.data1,ntree=mtree, mtry=mtry)

pred_Y <- predict(rf.Regression, newdata = mod.data1)

mod.data <- data.frame(X, pred_Y)

npk.aov1 <- aov(pred_Y ~ . , mod.data)

s_of_s1 <- as.matrix(anova(npk.aov1)["Sum Sq"]) # Sum of Squares

d_of_f1 <- as.matrix(anova(npk.aov1)["Df"])

m_of_s1 <- as.matrix((s_of_s1 / d_of_f1)) # Mean of squares

var_comp1 <- c()

for (i in 1:3){

 aux <- (m_of_s1[i] - m_of_s1[4])/3

 var_comp1 <- c(var_comp1,aux)}

Perc_var1 <- (var_comp1 / sum(var_comp1))*100

barplot(Perc_var, main="Variability in the response each factor RF_model", horiz=FALSE,
names.arg=c("x1","x2","x3","x1^2","x2^2","x3^2","x1x2","x1x3","x2x3"),xlab = "Variables",
ylab = "% Variabilidad explicada", col = c("darkblue","red","pink","green","lightblue","orange",
"purple","yellow","blue"))
```

```
barplot(Perc_var1, main="Variability in the response each factor RF_model", horiz=FALSE,
names.arg=c("x1","x2","x3"),xlab = "Variables", ylab = "% Variabilidad explicada", col =
c("darkblue","red","purple"))

barplot(Perc_var1, main="Variability in the response each factor RF_model", horiz=FALSE,
names.arg=c("x1","x1rf","x2","x2rf","x3","x3rf","x4","x4rf","x5","x5rf"),xlab = "Variables", ylab
= "% Variabilidad explicada", col =
c("darkblue","red","pink","green","lightblue","orange","purple","yellow","blue"))

var=c(Perc_var[0], Perc_var1[0],
Perc_var[1],Perc_var1[1],Perc_var[2],Perc_var1[2],Perc_var[3],Perc_var1[3])

barplot(var, main="Variability in the response each factor", horiz=FALSE,
names.arg=c("x1","x1rf","x2","x2rf","x3","x3rf"),xlab = "Variables", ylab = "% Variabilidad
explicada", col = c("darkblue","darkblue","red","red","purple","purple"))
```