



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

**Máster Universitario en Ingeniería de Análisis de Datos, Mejora  
de Procesos y Toma de Decisiones**

**Análisis de accidentes LBLOCA en centrales nucleares mediante  
técnicas de regresión y clasificación**

**Trabajo Final de Máster**

**Presentado por**

Antonio Barranca Mexía

**Tutores**

Ana Isabel Sánchez Galdón

Sebastián Martorell Alsina

Septiembre 2019

## Resumen

En el presente trabajo se presenta la aplicación de técnicas de análisis estadístico, paramétricas y no paramétricas, a la modelización de variables de seguridad en el marco de un accidente con pérdida de refrigerante en un reactor nuclear de agua a presión. Dichas técnicas han sido aplicadas a la predicción de la máxima temperatura de vaina (MPCT) y a la identificación de la rama donde se ha producido el accidente en función tanto de la MPCT como de la evolución temporal de la temperatura de vaina. La base de datos utilizada se ha obtenido a partir de la simulación del escenario accidental utilizando un código computacional termohidráulico. La predicción de la MPCT, en función de las condiciones de la planta en el momento de producirse el accidente, se ha realizado mediante la aplicación de diferentes técnicas de regresión avanzadas; por ejemplo, regresión con penalización, el modelo de regresión lineal generalizado y random forest. Respecto a la identificación de la rama donde se ha producido la rotura, se ha llevado a cabo utilizando máquinas de soporte vectorial y árboles de clasificación. Los resultados obtenidos muestran la posibilidad de sustituir el código termohidráulico, de elevado coste computacional, por un modelo de regresión.

## Abstract

This paper presents the application of statistical analysis techniques, parametric and non-parametric, to the modeling of safety variables in the context of an accident with loss of refrigerant in a nuclear reactor of water under pressure. These techniques have been applied to the prediction of the maximum sheath temperature (MPCT) and to the identification of the branch where the accident occurred based on both the MPCT and the temporal evolution of the sheath temperature. The database used has been obtained from the simulation of the accidental scenario using a thermohydraulic computational code. The prediction of the MPCT, depending on the conditions of the plant at the time of the accident, has been made by applying different advanced regression techniques; for example, regression with penalty, the generalized linear regression model and random forest. Regarding the identification of the branch where the break has occurred, it has been carried out using vector support machines and classification trees. The results obtained show the possibility of replacing the thermohydraulic code, with a high computational cost, with a regression model.

## Resum

En el present treball es presenta l'aplicació de tècniques d'anàlisi estadística, paramètriques i no paramètriques, a la modelització de variables de seguretat en el marc d'un accident amb pèrdua de refrigerant en un reactor nuclear d'aigua a pressió. Aquestes tècniques han sigut aplicades a la predicció de la màxima temperatura de baina (MPCT) i a la identificació de la branca on s'ha produït l'accident en funció tant de la MPCT com de l'evolució temporal de la temperatura de baina. La base de dades utilitzada s'ha obtingut a partir de la simulació de l'escenari accidental utilitzant un codi computacional termohidràulic. La predicció de la MPCT, en funció de les condicions de la planta en el moment de produir-se l'accident, s'ha realitzat mitjançant l'aplicació de diferents tècniques de regressió avançades; per exemple, regressió amb penalització, el model de regressió lineal generalitzat i random forest. Respecte a la identificació de la branca on s'ha produït el trencament, s'ha dut a terme utilitzant màquines de suport vectorial i arbres de classificació. Els resultats obtinguts mostren la possibilitat de substituir el codi termohidràulic, d'elevat cost computacional, per un model de regressió.

## Contenido

1. Introducción .....	1
2. Objetivos. ....	2
2.1 Objetivo general.....	3
2.2 Objetivo específico.....	4
3. Clustering .....	4
3.1 K-means.....	5
4. Regresión.....	6
4.1 Técnicas de regresión con penalización.....	6
4.1.1 Regresión Ridge.....	6
4.1.2 Regresión Lasso.....	7
4.1.3 Elastic Net.....	8
4.2 Modelo Lineal Generalizado (GLM).....	8
4.3 Tree based model.....	9
4.3.1 Modelos ensamblados/Combinados.....	10
5. Clasificación.....	12
5.1 Máquinas de soporte vectorial. ....	13
5.2 Árboles de clasificación. ....	14
6. Caso de aplicación.....	15
6.1 Clasificación del transitorio PTC.....	18
6.2 Clustering. ....	21
6.3 Estimación de la MPCT.....	22
6.3.1 Regresión Lineal múltiple con técnicas de penalización. ....	22
6.3.2 Modelo Lineal Generalizado (GLM).....	39
6.3.3 Random Forest. ....	43
6.3.4 Boosting.....	49
6.4 Comparación de los modelos de regresión.....	54
6.4.1 Comparación de los modelos de regresión.....	54
6.4.2 Selección del modelo de regresión. ....	55
6.5 Clasificación de la localización de la rotura en función de la MPCT. ....	55
7. Conclusiones.....	57
Bibliografía .....	58

## 1. Introducción

Una de las principales preocupaciones en el diseño de centrales nucleares es asegurar que las mismas se encuentran diseñadas para garantizar la seguridad de los trabajadores de la central, el entorno donde se ubica y las poblaciones próximas. Así, el análisis de accidentes de las plantas nucleares constituye un instrumento esencial para confirmar la idoneidad de las medidas de seguridad.

En los análisis de accidentes en una central nuclear podemos encontrar dos tipos:

- Accidentes base de diseño (DBA).
- Accidente más allá de la base de diseño (BDBA).

Los DBA son definidos como accidentes que engloban todas las condiciones accidentales previstas; condiciones previamente analizadas y a las que la central nuclear está diseñada para soportar, sin fallo en los sistemas, estructura y componentes necesario para garantizar la seguridad del personal y el correcto funcionamiento de la planta. En contraposición, los BDBA son los accidentes que superan los DBA o que no han sido analizados y, por lo tanto, no se han incluido en el diseño de seguridad de la central nuclear.

El accidente base de diseño (DBA) más común es el accidente con pérdida de refrigerante o LOCA (*Loss Of Coolant Accident*) grande (LBLOCA) y, por eso, este estudio centra en este tipo de accidente en un reactor tipo Westinghouse de agua a presión (PWR). En concreto, el marco de sucesos a estudiar es el de la rotura grande con pérdida de refrigerante o LBLOCA en el circuito primario del sistema de refrigeración del reactor.

En 1974, la United States Nuclear Regulatory Commission (USNRC) estableció una serie de parámetros de obligado cumplimiento en el caso de un accidente tipo LOCA para los sistemas de refrigeración de emergencia. En concreto, para la vasija del combustible las variables, dependientes entre ellas, que se deben controlar son 3:

- Maximum Peaking Clad Temperature (MPCT) o temperatura máxima de la vaina: Esta temperatura no debe exceder los 1204°C, ya que a más temperatura la vaina de combustible se fundiría. En la evolución de esta temperatura en un accidente tipo LBLOCA, suelen producirse 2 máximos relativos en las primeras fases de actuación de los sistemas de emergencia, por lo que es recomendable la monitorización de ambas, si estas se llegan a dar.
- Maximum Local Oxidation (MLO) o espesor local del óxido de la vaina: No puede superar 0.17 veces el espesor de la vaina de combustible previa a la oxidación de esta.
- Core Wide Oxidation (CWO) o proporción total oxidada de la vaina en el núcleo: El hidrógeno generado en esta reacción no debe superar 0.01 veces la cantidad de hidrógeno generada si se oxidara todo el metal de las vainas de combustible.

De igual manera y debido a su importancia en la evolución de la MPCT y en el correcto funcionamiento de los sistemas de refrigeración de emergencia (ECCS), se recomienda la monitorización de otras variables de interés, que son:

- Tiempo hasta la MPCT: Se registran el tiempo hasta los dos máximos relativos, si estos se dieran.

- Tiempo hasta la inyección de los acumuladores.
- Tiempo hasta el enfriamiento rápido después de las inyecciones.

Este trabajo, al igual que la mayoría de los estudios presentados en la literatura LBLOCA, se centra en el análisis de la MPCT, ya que, como se ha dicho previamente, el resto de las variables de interés están muy correlacionadas. Así, si la MPCT no supera los 1204°C, se cumplirá tanto el MLO como el CWO.

Debido a la importancia de estos factores, tanto en el diseño como en control de la seguridad de la central, es necesario conocer la evolución temporal de las diferentes variables termohidráulicas del reactor y cuando se alcanza la MCTP para poder conocer e identificar condiciones operativas de la central en el caso de un LBLOCA.

Como es de esperar, no es viable la obtención de estos datos a través de la observación directa, debido a las posibles consecuencias que ocasionaría una simulación intencionada de un LBLOCA de un reactor nuclear. Ya que el estudio debería darse en dicho marco circunstancial, es necesaria la simulación de los diferentes escenarios accidentales mediante códigos computacionales termohidráulicos de tipo Best Estimate (BE). Un problema derivado del uso de códigos BE es el elevado coste computacional, por lo que resulta de interés analizar la sustitución de dichos códigos por modelos de regresión que permitan una estimación de las variables de seguridad con interés.

La estructura del presente trabajo se dividirá en los siguientes puntos:

- Objetivos.
- Clustering, donde se explicará la metodología del análisis de los clústeres creados.
- Regresión, donde se detallará las distintas técnicas de regresión que se van a utilizar.
- Clasificación, donde se desarrollarán la metodología de clasificación usada.
- Caso de aplicación, resultados obtenidos de aplicar las técnicas, anteriormente citadas, en la consecución de los objetivos.
- Conclusiones, donde se analizarán la consecución de los objetivos a partir de las técnicas aplicadas.

## 2. Objetivos.

Debido a la importancia de los factores previamente mencionados, como la temperatura MPCT en la identificación de fallos en el sistema de refrigeración del reactor, es necesario conocer y saber identificar el comportamiento de estas variables con objeto de mejorar la seguridad de la central y poder identificar fallos en el reactor, como un LBLOCA, antes de que se descubriera el núcleo del reactor u otro accidente similar.

La intención del presente estudio se centra en el tratamiento estadístico de los datos de un reactor Westinghouse-PWR de 4 circuitos de refrigeración obtenidos a través de simulación. Este análisis consistirá en un estudio de la aplicabilidad de distintas herramientas estadísticas en

la predicción y clasificación del transitorio y en una comparación de los resultados obtenidos para así elegir la mejor o mejores herramientas de análisis.

A continuación, se detallan los objetivos que se pretenden alcanzar en el presente trabajo.

### 2.1 Objetivo general.

El objetivo principal consiste en la identificación de la rama de refrigeración de un reactor de diseño Westinghouse-PWR en la que se ha producido un LBLOCA a partir de la MPCT. Para la consecución de dicho objetivo se seguirá el siguiente flujo de trabajo:

- 1) Agrupación de las observaciones en función de la MPCT: Ya que el sistema de refrigeración está constituido por 4 ramas, en las cuales puede suceder un LBLOCA, con ayuda de técnicas de clustering se clasificarán las observaciones en función de la temperatura MPCT, la cual alcanza valores distintos en función de en qué rama de refrigeración se produzca la rotura. En la figura 1 se puede ver un boceto del reactor anteriormente mencionado
- 2) Predicción de la temperatura MPCT: A partir de las variables termohidráulicas del reactor, se construirá un modelo para la predicción de la temperatura MPCT.
- 3) Predicción de la rama en la que se ha producido el LBLOCA: A partir de la temperatura MPCT, anteriormente predicha, se construirá un clasificador que nos indique la rama del sistema de refrigeración en el que se ha producido la fuga de refrigerante.

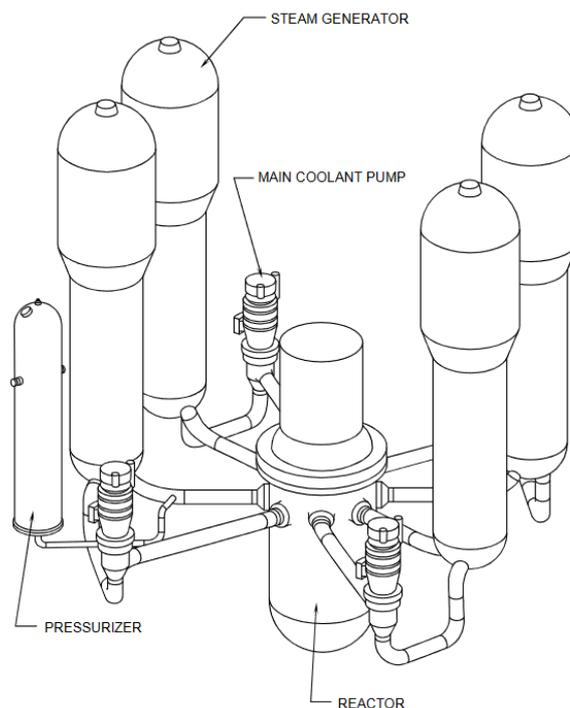


Figura 1: Reactor Westinghouse-PWR 4 loops.

## 2.2 Objetivo específico.

Como se ha dicho previamente, igual de importante es conocer el valor de la MPCT como su evolución temporal. El objetivo que se plantea es la de la construcción de un clasificador, como en el punto 3) del apartado anterior, que prediga la rama en la que se ha producido el LBLOCA en función de la evolución temporal de la PCT. La eficiencia de este clasificador residirá tanto en que acierte la rama en la que se ha producido la rotura, como que lo haga en el menor tiempo posible y antes de que se llegue a la MPCT. Un ejemplo de la evolución temporal de la PCT se puede ver en la figura 2.

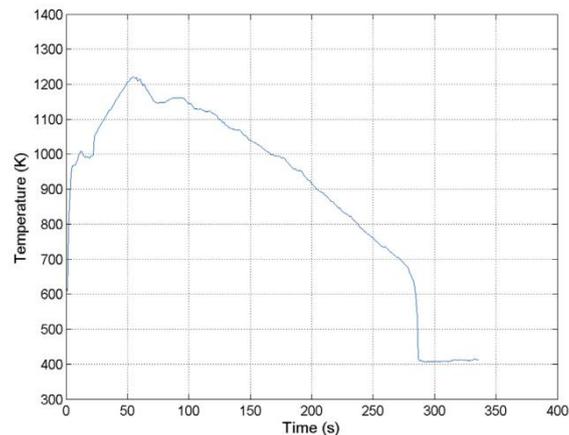


Figura 2: Evolución temporal de la PCT en un BLOCA

## 3. Clustering

El análisis clúster o clustering comprende una amplia variedad de técnicas utilizadas para clasificar un conjunto heterogéneo de individuos, observaciones o variables en un conjunto de grupos que reflejan las relaciones existentes entre los mismos (Macqueen, 1967). Estas técnicas generan grupos o clústeres formados por elementos similares entre sí, además, las técnicas propuestas en el análisis clúster son descriptivas, ya que tratan los datos de partida como la totalidad de una población, por lo que no son útiles para hacer un razonamiento de naturaleza inferencial.

De forma general, las técnicas de clustering se pueden agrupar en dos tipos:

- Clúster Jerárquico: El objetivo es el de estructurar los individuos de un conjunto de forma jerárquica por su similitud. Este tipo de clasificación implica que los datos se ordenan por niveles, por lo que los niveles inferiores están contenidos en los superiores.
- Clúster No Jerárquico: El objetivo de este análisis es el de conseguir agrupar un conjunto de datos heterogéneo en un número prefijado de  $k$  grupos. Las principales características de este tipo de clustering es que cada individuo pertenece a uno solo de los grupos, internamente homogéneo, y todos los individuos quedan clasificados.

Aunque los tipos de clústeres, anteriormente mencionados, tienen distintos objetivos, la operativa general para realizar ambos análisis es similar.

- 1) A partir de una matriz de datos  $I \times J$  se construye una matriz de distancias  $I \times I$ .
- 2) A partir de la matriz de distancias se construyen los grupos mediante algún algoritmo, jerárquico o no, de clustering.

En el presente trabajo el tipo de clustering que se ha usado es el clúster no jerárquico, en particular se ha usado el algoritmo *k-means*, debido a la naturaleza de los datos. A continuación, se detallará este tipo de clustering no jerárquico con más detalle.

### 3.1 K-means.

El término de *k-means* se usa para denotar el proceso de asignar cada individuo al clúster, uno de los  $k$  prefijados, con el centroide más próximo (Garbade, 2018). La clave de este procedimiento radica en que el centroide se calcula a partir de los miembros del clúster tras su asignación.

La secuencia del algoritmo es la siguiente:

- 1) Tomar los  $k$  primeros casos como clústeres iniciales.
- 2) Asignar cada uno de los  $(m - k)$  individuos restantes al clúster con el centroide más próximo. Tras cada asignación se calcula el centroide del nuevo clúster.
- 3) Tras la asignación de todos los individuos en el paso 2), se toman los centroides de los clústeres como puntos "semilla" y se hace una pasada más sobre los datos, asignando cada observación al punto de semilla más cercano.

A partir de la secuencia anterior se implementa un procedimiento convergente que se llama método de *k-means*:

- 1) Se comienza con una partición inicial de los individuos en clústeres. Se pueden usar los pasos 1) y 2), anteriormente mencionados, u otro método de agrupación.
- 2) Tomar cada individuo sucesivamente y calcular las distancias a todos los centroides de los clústeres; si el centroide más próximo no es el del clúster en el que se le ha agrupado inicialmente, se reasigna al clúster con el centroide más próximo y se vuelven a calcular los centroides de los clústeres involucrados.
- 3) Repetir el paso 2) hasta que se obtenga la convergencia, es decir, se continúa hasta que en un ciclo completo a través de los individuos no se produzca ningún cambio entre los miembros de los clústeres.

En la figura 3 se puede ver un esquema del algoritmo estándar *k-means*.

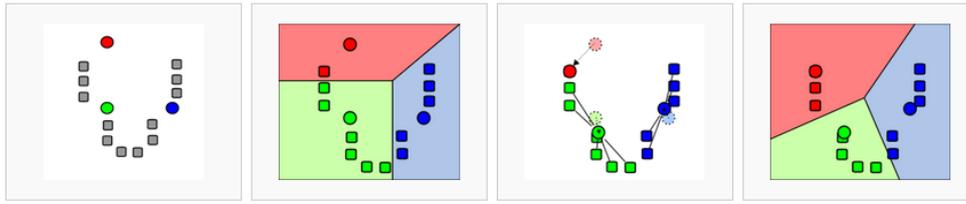


Figura 3: Secuencia algoritmo K-means (CC, 2015).

## 4. Regresión.

El análisis estadístico mediante regresión comprende un conjunto de técnicas que permiten estimar las relaciones entre variables dependientes e independientes. Este tipo de análisis ayudan a comprender como se comporta la o las variables dependientes al cambiar los valores de una o más variables independientes. Otra de las utilidades de la regresión es que permite estimar el valor promedio de las variables dependientes cuando se fijan las variables independientes.

Los dos principales usos que se le da al análisis de regresión son la predicción y el análisis de la relación entre las variables independientes y dependientes y, así poder explotar esa relación. En algunos casos este análisis se puede utilizar para inferir relaciones causales entre las variables, sin embargo, esto no siempre se da ya que la correlación de las variables no implica causalidad.

Los métodos de regresión estadística se pueden englobar en dos grandes familias. La primera de ellas sería los métodos paramétricos, donde la función de regresión se define entorno a un conjunto de parámetros desconocidos que se estiman a partir de los datos, como la regresión lineal y la regresión por mínimos cuadrados. La segunda sería la de las regresiones no paramétricas, donde la función de regresión consiste en un conjunto de funciones de dimensión variable. Un ejemplo de regresión no paramétrica sería la técnica conocida como Random Forest.

### 4.1 Técnicas de regresión con penalización.

Las técnicas de regresión con penalización comprenden una serie de técnicas de regresión donde existe una función creciente de penalización sobre el “tamaño” de los estimadores de los coeficientes de regresión  $\beta$  (Carrasco, 2016).

Las técnicas de regresión con penalización que se van a usar en este trabajo son la regresión Ridge, que tiene por función de penalización la norma L2; la regresión Lasso, que tiene como penalización la norma L1 y la regresión Elastic Net, la cual es una combinación de las dos anteriores por lo que permite un mayor rango de aplicación.

#### 4.1.1 Regresión Ridge.

La regresión Ridge (Kennard, 1970) se desarrolló para eludir los efectos adversos del problema de colinealidad en un modelo lineal estimado por mínimos cuadrados. Los coeficientes de regresión Ridge son los valores que minimizan la ecuación (1):

$$\hat{\beta}^{Ridge} = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2 \quad (1)$$

Donde los parámetros de la ecuación son:

- $\hat{\beta}^{Ridge}$ : Coeficiente estimado por la regresión Ridge.
- $\lambda$  : Parametro de contracción
- $\beta_j$  : Coeficiente de regresión por mínimos cuadrados
- $RSS$  : Suma del cuadrado de los residuos (*Residual Sum of Squares*)

El método Ridge tiende a contraer los coeficientes de regresión al incluir un término de penalización en la función objetivo: cuanto mayor sea este término de penalización, mayor será la penalización y, por consiguiente, mayor será la contracción de los coeficientes.  $\lambda$  es el parámetro de contracción. En función del valor de este parámetro estaremos ante distintos casos.

- $\lambda=0$ : mínimos cuadrados ordinarios.
- $\lambda= \infty$ : Entonces  $\beta_i=0$  y estaremos ante un estimador sesgado de  $\beta$ . Introducimos sesgo, pero reducimos varianza.

Una vez obtenidos los coeficientes hay que obtener el valor de  $\lambda$  que minimiza el error de estimación de la predicción. Esto se realizará a través de validación cruzada hasta obtener el mejor valor de  $\lambda$ .

Uno de los inconvenientes de este método es que puede llegar a contraer todos los coeficientes hacia cero, pero sin conseguir la nulidad de ninguno lo que provoca que no haya selección de variables, permaneciendo en el modelo todas las variables explicativas. Esto es un inconveniente para los modelos donde hay una gran cantidad de variables explicativas.

#### 4.1.2 Regresión Lasso.

Para eludir el inconveniente de que la regresión Ridge no consiga la contracción de los coeficientes a cero, se desarrolló la regresión Lasso (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1996). Esta técnica, a partir de cierto valor del parámetro de complejidad produce estimaciones nulas para algunos coeficientes, con lo cual se realiza una selección de variables debido a la norma L1. Debido a esto Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes y, al mismo tiempo, crea modelos interpretables por la reducción de alguno de los coeficientes.

Lasso resuelve el problema de mínimos cuadrados sobre la norma L1 del vector de coeficientes, la estimación de los coeficientes se obtiene minimizando la ecuación (2):

$$\hat{\beta}^{Lasso} = \min_{\beta} \left\{ \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 \right\} \text{ s. a } \sum_{j=1}^p |\beta_j| \leq s \quad (2)$$

Donde los parámetros de la ecuación son:

- $\hat{\beta}^{Lasso}$ : Coeficiente estimado por la regresión Lasso
- $\beta_j$  : Coeficiente de regresión por mínimos cuadrados
- $s$  : Parámetro de complejidad del modelo

La ecuación (2) es equivalente a la ecuación (3), donde se vuelve a tener el parámetro de complejidad  $\lambda$ .

$$\hat{\beta}^{Lasso} = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = RSS + \lambda \sum_{j=1}^p |\beta_j| \quad (3)$$

En general estos modelos son muchos más fáciles de interpretar que los obtenidos mediante Ridge. Al igual que en el caso de la regresión Ridge, el valor de  $\lambda$  se obtiene por validación cruzada.

#### 4.1.3 Elastic Net.

Esta técnica de regularización y de selección de variables (Hastie H. Z., 2005) retiene las ventajas de Lasso, selección automática de las variables y contracción continua, y de la regresión ridge superando, al mismo tiempo, alguna de sus limitaciones.

El criterio de penalización Elastic Net se define como:

Para cualesquiera  $\lambda_1$  y  $\lambda_2$  constantes fijas no negativas se define

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1 \quad (4)$$

donde:

$$|\beta|^2 = \sum_{j=1}^p \beta_j^2, \quad |\beta|_1 = \sum_{j=1}^p |\beta_j|, \quad |y - X\beta|^2 = RSS$$

Los coeficientes estimados por esta técnica es el vector que minimiza la ecuación (5).

$$\hat{\beta}^{enet} = \arg \min_{\beta} |y - X\beta|^2 \quad (5)$$

Si consideramos  $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ , el problema de optimización se convierte en el siguiente:

$$\hat{\beta}^{enet} = \arg \min_{\beta} |y - X\beta|^2 \quad (6)$$

$$s.a : \alpha |\beta|_1 + (1 - \alpha) |\beta|^2 \leq t \text{ para algún } t \quad (7)$$

A la función (7) se la conoce como penalización *Elastic Net*, la cual es una combinación de la penalización Ridge y Lasso.

#### 4.2 Modelo Lineal Generalizado (GLM).

Los modelos lineales clásicos (Regresión, ANOVA, MANOVA, etc...) se basan en una serie de supuestos (Diazaaraque), que son:

- Distribución normal de los residuos.
- Homocedasticidad.
- La variable dependiente presenta una relación lineal con las variables explicativas.

No obstante, en muchas ocasiones uno o varios de estos supuestos no se cumple por la naturaleza de los datos. Estos inconvenientes pueden solucionarse aplicando una transformación sobre la variable respuesta, como, por ejemplo, el logaritmo. Sin embargo, estas

transformaciones no siempre consiguen corregir la falta de normalidad, heterocedasticidad o no linealidad de los datos.

Una alternativa a las transformaciones y a la falta de normalidad de los datos son los modelos lineales generalizados, GLM (*Generalized Linear Models*) que son, por tanto, una extensión de los modelos lineales que permiten utilizar distribuciones no normales y varianzas no constantes.

La especificación de un GLM tiene tres componentes (Diazaraque):

- Componente aleatoria: Identifica la variable respuesta y su distribución de probabilidad
- Componente sistemática: Especifica las variables explicativas (Independientes o predictoras) utilizadas en la función predictora lineal.
- Función Vínculo: Es una función del valor esperado de  $Y$ ,  $E(Y)$ , como una combinación lineal de las variables predictoras.

En la construcción de GLM es necesario tener en cuenta que no existe un único modelo que sea válido. En la mayoría de los casos, habrá un número de modelos plausibles que puedan ajustarse a un conjunto determinado de datos. Parte del trabajo de construcción consiste en estimar los diferentes modelos y entre los adecuados seleccionar los de mayor varianza explicada. En el caso de que haya más de un modelo que describa los datos igual de bien, queda al criterio del usuario la elección de un modelo u otro (Turner, 2008).

En la construcción y evaluación de un GLM debemos tener en cuenta una serie de aspectos:

1. Elección de la estructura de errores y función vínculo: Para asegurarnos la correcta elección de la función vínculo es recomendable analizar los residuos del modelo para ver la distribución de los errores y, comparar modelos con distintas funciones vínculo para ver cual se adapta a nuestros datos mejor.
2. Ajuste del modelo a los datos: La cantidad de varianza explicada por los modelos GLM se conoce como devianza  $D^2$ . La devianza nos da una idea de la cuanta variabilidad, de los datos, consigue explicar el modelo. Para ello se debe obtener una medida de la variabilidad explicada por el modelo y la comparamos con la devianza residual del modelo nulo.
3. Simplificación del modelo: El mejor modelo es aquel que sea tan simple como sea posible. Esto es que no debe contener parámetros o niveles de un factor que sean redundantes o que no sean significativas.

#### 4.3 Tree based model.

Los algoritmos de aprendizaje basados en árboles están considerados como uno de los mejores y más utilizados métodos de aprendizaje supervisado. Los métodos basados en árboles potencian los modelos predictivos con alta precisión, estabilidad y facilidad de interpretación. A diferencia de los modelos lineales, mapean bastante bien las relaciones no lineales. Son adaptables para resolver tanto problemas de clasificación como de regresión (Torgo, 2019).

El árbol de decisión es un tipo de algoritmo de aprendizaje supervisado, que tiene una variable objetivo y predefinida, que se usa principalmente en problemas de clasificación. Funciona para variables de entrada y salida categóricas y continuas. En esta técnica, dividimos la población o muestra en dos o más conjuntos homogéneos o subpoblaciones basados en el divisor más

significativo en las variables de entrada. En la figura 4 se muestra un concepto del árbol de decisión.

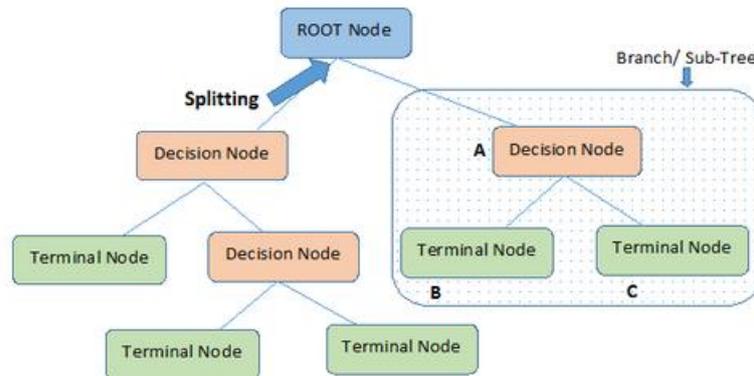


Figura 4: Esquema CART (Alvear, 2018).

Como se ha dicho previamente, los modelos basados en árboles de decisión pueden usarse tanto para clasificar como para tareas de regresión. Los cuales se denominan como:

- Árbol de decisión de variable categórica (Clasificación)
- Árbol de decisión de variable continua (Regresión)

Como se puede ver en la figura 4, cada nodo corresponde a una variable de entrada. Las hojas o *terminal node*, representa un valor de la variable objetivo dado los valores de las variables de entrada.

El árbol se construye dividiendo el conjunto de datos inicial, que constituye el *root node*, en subconjuntos. La división se basa en un conjunto de reglas de división basadas en las características de clasificación o regresión. Este proceso se repite en cada subconjunto, derivando de esta manera en una partición recursiva. El proceso recursivo finaliza cuando el subconjunto de un nodo tiene los mismos valores de la variable objetivo, o cuando la división ya no agrega valor a las predicciones. Este proceso de inducción desde la raíz hasta las hojas es un ejemplo de algoritmo egoísta, y es, la estrategia más común de aprendizaje de los árboles de decisión (Torgo, 2019).

Una de las ventajas de los árboles de decisión comparado con otros métodos de regresión es su interpretabilidad y la representación gráfica de este.

#### 4.3.1 Modelos ensamblados/Combinados.

Los métodos *bagging* y *boosting* nos permiten mejorar sustancialmente el rendimiento predictivo de los modelos basados en árboles, aunque con una considerable reducción en la facilidad de interpretación del modelo final. Estos métodos, conocidos como métodos combinados, utilizan múltiples algoritmos de aprendizaje para obtener predicciones que mejoren las que se obtendrían por medio de cualquiera de los algoritmos de forma individual. Estos métodos son aplicables tanto a regresión como clasificación (Martínez, 2018).

Los modelos combinados surgen como un método para solucionar el problema de sesgo y varianza de los árboles de decisión, es decir, cuanto en promedio son los valores predichos diferentes a los valores reales (sesgo) y, cuán diferentes son las predicciones de un modelo en un mismo punto si muestras diferentes se tomarán de la misma población (varianza). Esto se

debe a que al construir un árbol “pequeño” se obtendrá un modelo de baja varianza y alto sesgo, pero, a partir de un punto el modelo será muy complejo y se producirá un sobreajuste, momento en el cual la varianza empieza a aumentar. Por lo tanto, el modelo óptimo es el que mantiene un balance o *trade-off* entre el error de sesgo y de varianza. En la figura 5 se presenta un esquema conceptual de la evolución del *trade-off*.

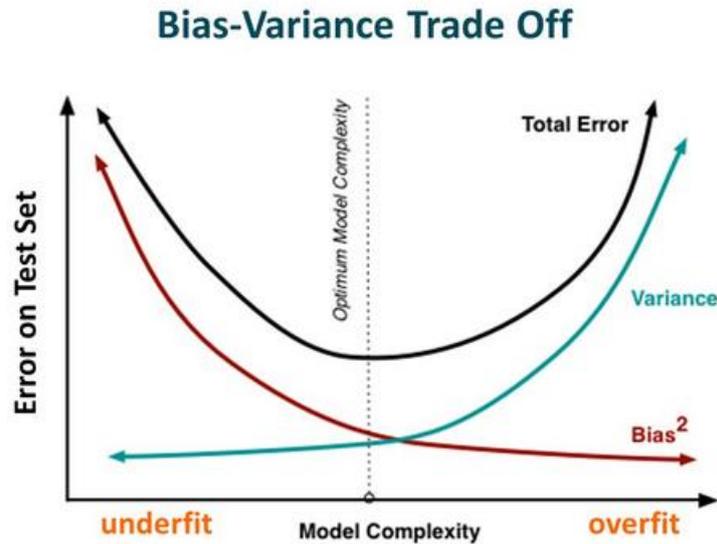


Figura 5: Esquema evolución *trade-off* (Alvear, 2018).

#### 4.3.1.1 Random Forest.

Los modelos predictivos, tanto de regresión como clasificación, Random Forest son modelos combinado a partir de *bagging*.

El *bagging* es una técnica usada para reducir la varianza de las predicciones a través de las combinaciones de los resultados de varios clasificadores, cada uno de ellos construido con diferentes subconjuntos de la misma población. Este proceso puede ser repetido el número de veces que sea necesario con el objetivo de obtener predicciones más precisas (Martínez, 2018). En definitiva, a partir de un conjunto de predictores débiles se construye un predictor robusto. Un esquema de proceso de *bagging* se puede ver en la figura 6.

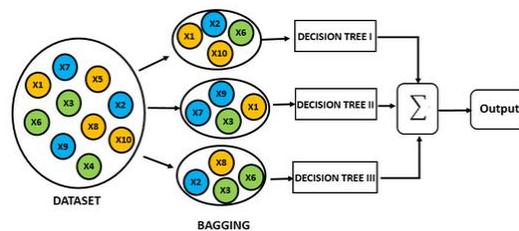


Figura 6: Esquema conceptual *bagging* (Alvear, 2018).

Para la construcción de un modelo *Random Forest* se siguen los siguientes pasos:

1. Dado un número  $N$ , casos en un conjunto de entrenamiento. Se selecciona de forma aleatoria, pero con reemplazo, una muestra de entre esas  $N$  observaciones. Esta muestra será el conjunto de entrenamiento para el árbol  $i$ .
2. Si el conjunto de datos tiene  $M$  variables de entrada, se seleccionan  $m \leq M$  variables aleatoriamente. La mejor división de estos  $m$  atributos se usan para el crecimiento del árbol. El valor de  $m$  se mantiene constante durante la generación del bosque.
3. Cada árbol crece hasta su máxima extensión posible, sin proceso de poda.
4. Las nuevas observaciones se predicen a partir de la agregación de las predicciones de los  $x$  árboles. En el caso de la regresión se toma el promedio de las salidas  $y$ , en de clasificación el resultado es la clase con mayor número de “votos” en todo el “bosque”.

Las principales ventajas de los modelos *Random Forest* sobre otros modelos es que, en *Random Forest*, existen muy pocos supuestos y, por lo tanto, el preprocesamiento de los datos es mínimo. Además, nos permite trabajar con gran cantidad de predictores e identificar los más importantes. Por lo que puede llegar a ser usado como un método de reducción de variables.

#### 4.3.1.2 Boosting.

El *boosting* funciona de manera similar al *bagging* en tanto en cuanto combina una gran cantidad de árboles, a excepción de que los árboles se construyen de manera secuencial. Cada árbol se genera usando la información retenida en los residuos de los árboles previamente generados. Otra diferencia es que el *boosting* no utiliza remuestreo a través de Bootstrap, sino que cada árbol se genera usando una versión modificada del conjunto de datos de entrenamiento (Martínez, 2018).

A diferencia del *random forest*, el método *boosting* puede sobreajustarse al conjunto de datos si el número de árboles es muy alto, por lo tanto, hay que buscar el número óptimo de árboles a partir de validación cruzada. También se puede controlar el número de interacciones, *interaction depth* de cada árbol con los árboles sucesivos y el parámetro de penalización  $\lambda$ , que es una medida del ritmo de aprendizaje del algoritmo.

## 5. Clasificación.

En el contexto de *machine learning* la clasificación se considera un problema de aprendizaje supervisado, en el cual, las clases o subpoblaciones son conocidas inicialmente y cuyo objetivo se centra en estimar a que clase pertenece una o múltiples observaciones (Torgo, 2019).

En general, la clasificación y el clustering son problemas de reconocimiento de patrones, que es la asignación de una etiqueta de salida a un valor de entrada.

Los tipos de clasificadores más comunes son los denominados *Clasificadores Lineales*. Los clasificadores lineales realizan la clasificación basándose en una combinación lineal de las características/variables de las observaciones. El funcionamiento de estos clasificadores consiste en la creación de una función lineal que asigna un valor o puntuación a cada categoría posible. Esto lo hace combinando el vector de características de cada observación con un vector de pesos, utilizando el producto escalar de ambos vectores. De esta forma, la categoría prevista es aquella que obtiene la puntuación más alta. En la figura 7 se muestra un ejemplo de clasificación binaria.

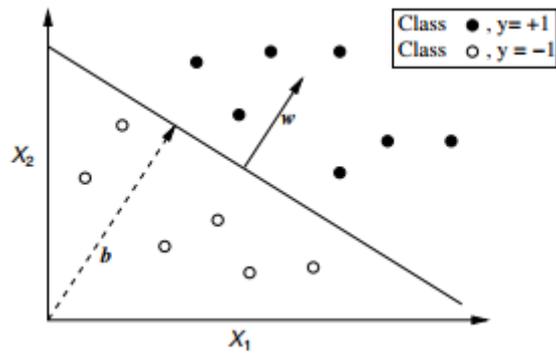


Figura 7: Esquema clasificación binaria (Torgo, 2019).

Existe un gran número de clasificadores, que se diferencian en función del tipo de algoritmo con el que hace la clasificación.

En este trabajo se ha realizado la clasificación utilizando dos algoritmos:

- SVM para la clasificación del transitorio de la PCT.
- Árbol de clasificación para la clasificación de la MPCT.

### 5.1 Máquinas de soporte vectorial.

Las máquinas de soporte vectorial o Support Vector Machines (SVM) son un tipo de algoritmo de machine learning supervisado aplicable a problemas de regresión y clasificación. Aunque inicialmente fueron pensadas para resolver problemas de clasificación binaria, actualmente ya se pueden utilizar tanto para resolver problemas de clasificación multinivel como de regresión de una variable continua.

En el caso de la regresión por SVM, la idea es seleccionar el hiperplano regresor que mejor se ajuste a nuestro conjunto de datos de entrenamiento. La idea se basa en considerar una distancia de margen  $\varepsilon$ , de modo que esperamos que todos los ejemplos se encuentren en una banda entorno a nuestro hiperplano, es decir, que diste una separación menor a  $\varepsilon$ . En la figura 8 podemos ver una representación de la idea de regresión por SVM.

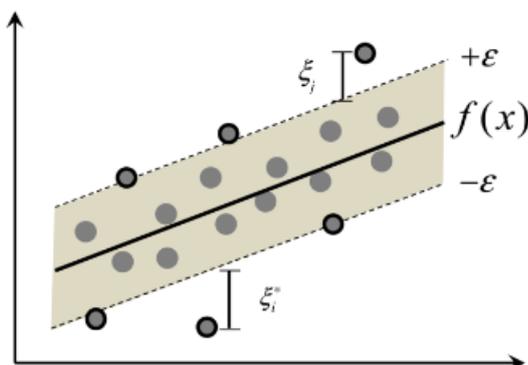


Figura 8: Concepto SVR (Torgo, 2019).

El funcionamiento de la clasificación a través de las máquinas de soporte vectorial es el contrario que la regresión por SVM. En vez de buscar el hiperplano que englobe el máximo número de

observaciones, el objetivo es encontrar el hiperplano que mejor separe las observaciones en función de la clase o subgrupo al que pertenezca. Más formalmente, el objetivo de la SVM es la selección de un hiperplano de separación que equidiste de los ejemplos más cercanos de clase para, de esta manera, conseguir lo que se denomina una distancia máxima a cada lado del hiperplano. A la hora de definir este hiperplano, sólo se considerarán las observaciones de entrenamiento que distan del hiperplano la distancia margen. Estas observaciones reciben el nombre de vectores de soporte. En la figura 9 se puede ver un esquema conceptual de clasificación binaria a partir de SVM.

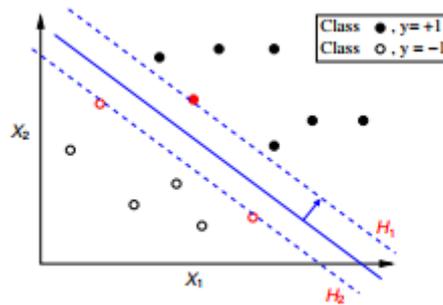


Figura 9: Concepto clasificación binaria a partir de SVM (CC, 2015).

## 5.2 Árboles de clasificación.

Los árboles de decisión donde la variable objetivo toma como valores un conjunto de valores discretos o etiquetas de clase se denominan árboles de clasificación. En este tipo de árbol de decisión las hojas representan las etiquetas de clase y las ramas representan una conjunción de características que conducen a esa etiqueta de clase (Torgo, 2019).

El funcionamiento de los árboles de clasificación sigue un modelo basado en pruebas lógicas, como se puede ver en la figura 10, a partir de las variables de entrada. Este modelo de pruebas lógicas puede verse como una partición del espacio de entrada definido por las variables de entrada. Estas particiones se definen en base a pruebas lógicas de una selección de las variables. Dentro de cada partición se asigna a todas las observaciones la misma predicción. Una de las principales ventajas de este tipo de clasificador frente a otros modelos es su interpretabilidad, ya que nos indican que variables son las más decisivas a la hora de clasificar las observaciones.

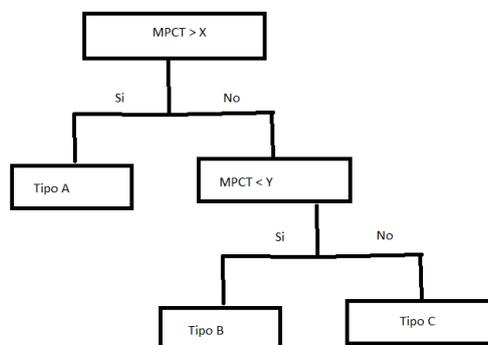


Figura 10: Concepto Árbol clasificación

## 6. Caso de aplicación

En esta sección se presenta una descripción del problema abordado y la base de datos que se ha analizado, además, se presentarán los resultados obtenidos del tratamiento estadístico con las técnicas descritas en las secciones anteriores.

Como se mencionó en la introducción, el presente trabajo se centra en el análisis de uno de los principales accidentes base de diseño, en concreto, un LBLOCA. Como consecuencia de un LBLOCA se produce una despresurización en el circuito primario de refrigeración, que debido a que el tamaño de la rotura es grande, es muy brusca. Esto activa la señal de alarma del reactor por baja presión y, simultáneamente, se descargan los acumuladores y se activa la inyección de seguridad.

Este disparo de la alarma del reactor, la descarga de los acumuladores, inyección de seguridad, en este caso de agua borada, y la misma despresurización del circuito primario crean huecos que ayudan y permiten a la reducción de la potencia del reactor hasta el nivel de potencia residual, potencia cuyo origen se encuentra en los productos de fisión y que es necesaria para mantener el reactor en una situación estable mientras los inyectores están actuando, introduciendo agua de los tanques de almacenamiento o mientras que se recircula el agua de las ramas frías.

En el análisis probabilístico de la seguridad, el árbol de sucesos iniciantes para este tipo de accidentes se muestra en la figura 11.



Figura 11: Árbol sucesos (nivel 1), LBLOCA

Con la intervención de los sistemas de seguridad, se puede diferenciar tres etapas en la secuencia accidental en función de la actuación de dichos sistemas en las diferentes partes del reactor (vasija, núcleo del reactor y plenum inferior). Estas fases son:

- 1) **Blowdown** o fase de vaciado: Vaciado o exposición del núcleo debido a la pérdida de refrigerante. Comprende desde la rotura del circuito de refrigerante hasta que la presión del sistema baja hasta el nivel de contención y el núcleo queda al descubierto.
- 2) **Refill** o fase de rellenado: Comienza tras la finalización del *blowdown* hasta que el suministro de agua de refrigeración de emergencia llena el plenum inferior y alcanza la cota inferior del combustible del reactor.
- 3) **Reflood** o fase de reinundación: Se extiende desde que el agua alcanza la cota inferior de los elementos de combustible hasta que el volumen de agua que se ha introducido en el reactor es suficiente como para detener el aumento de temperatura del núcleo y esta temperatura es estabilizada al nivel asociado a la disipación de calor de desintegración a largo plazo.

La base de datos que se ha utilizado para estimar la máxima temperatura de vaina (MPCT) en el caso de un LBOCA contiene un total 800 observaciones y las siguientes variables:

- **59 variables termohidráulicas:** Variables termohidráulicas que parametrizan el estado del circuito de refrigeración. Estas variables serán las variables explicativas o regresores que se usarán para predecir la temperatura MPCT. En la tabla 2 se muestran las diferentes variables.
- **Localización de la rotura:** Variable cualitativa que toma 4 valores distintos, los cuales corresponden a la rama del circuito refrigeración sobre la cual se ha simulado el LBLOCA y todos con el mismo número de observaciones asignadas,  $N= 200$ , y que se muestran en la tabla 1.

Tabla 1: Posibles localizaciones del LBLOCA

Localización de la rotura			
CI Rama fría	HI Rama caliente	II Rama intermedia	SI Línea de equilibrio del presionador

- **Temperatura MPCT:** Correspondiente a la máxima temperatura de vaina ( $K$ ) alcanzada durante el transitorio. Sera la variable dependiente/objetivo a predecir y, partir de esta, se clasificará los tipos de fallo.

En la tabla 2 se muestran los valores de  $\bar{x}$  y  $s$  para las cuatro localizaciones de la rotura consideradas.

Tabla 2: Distribución MPCT en función del lugar del LBLOCA

Localización de la rotura	Media ( $K$ )	Desviación estándar ( $K$ )
CI	1206.5	30.005
HI	716.003	19.43
II	784.79	19.54
SI	623.53	5.21

La función de densidad estimada, a partir del método kernel, de la temperatura MPCT para cada localización de la rotura, se muestra en la figura 12 y la caracterización de las 59 variables termohidráulicas en la tabla 3.

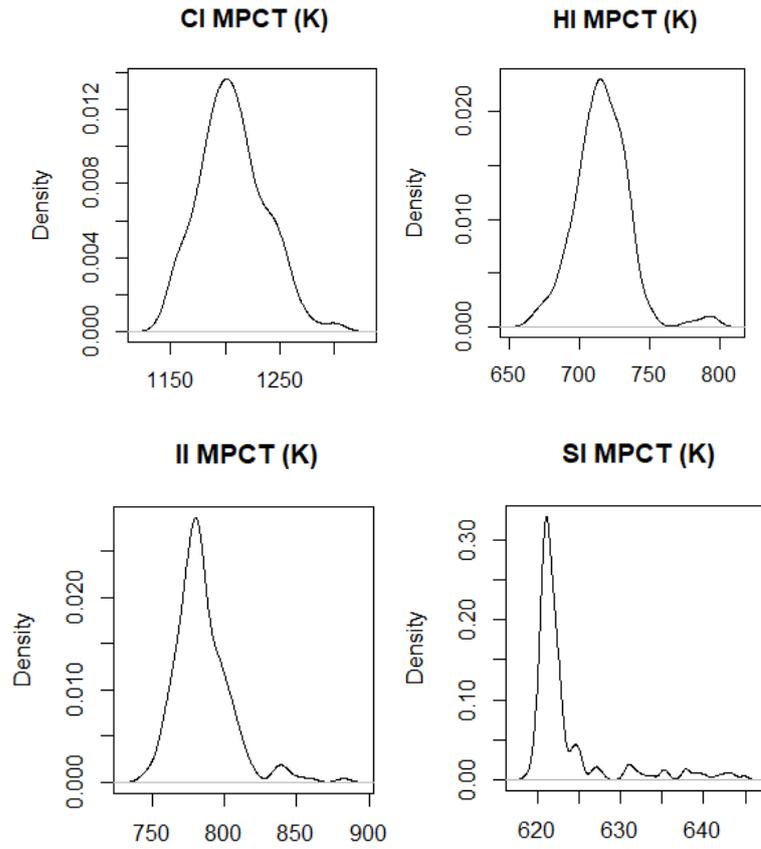


Figura 12: Grafico función de densidad de la MPCT

Tabla 3: Variables termohidráulicas.

Caracterización de las variables termohidráulicas					
Var.	Media	Des. Estándar	Var.	Media	Des. Estándar
INPOW	0,9994	0,0089	FILMB	1,0028	0,0577
PEAKF	0,9986	0,0201	TMIFS	0,9982	0,0560
HGSAV	1,0047	0,0855	TRANB	0,9998	0,0542
HGSHR	1,0094	0,0845	CRIHF	1,0109	0,1106
RPOWM	0,9993	0,0371	FLOSS	0,9958	0,0579
UO2TC	0,9971	0,0460	DRAG1	0,9968	0,0570
UO2SH	0,9999	0,0090	DRAG2	1,0052	0,0598

<b>IMCPS</b>	1,0000	0,0091	<b>DRAG3</b>	0,9975	0,0541
<b>BMCPS</b>	0,9937	0,0455	<b>DRAG4</b>	0,9996	0,0584
<b>ACCPR</b>	-0,0054	0,0910	<b>DRAG5</b>	0,9941	0,0586
<b>ACCLT</b>	-0,3418	4,3278	<b>DRAG7</b>	0,9998	0,0550
<b>LPISQ</b>	1,0012	0,0229	<b>CCFLC</b>	1,0032	0,0584
<b>PRZLE</b>	-0,0018	0,0422	<b>ACCLV</b>	-0,0270	0,2803
<b>MASFR</b>	1,0004	0,0176	<b>VAPIA</b>	1,0050	0,0545
<b>CLTEM</b>	0,0346	0,9029	<b>LIQIA</b>	1,0072	0,0622
<b>ACCKL</b>	1,0238	0,3020	<b>UHTEM</b>	5,3093	2,8531
<b>KSURG</b>	1,0096	0,3208	<b>LINT2</b>	1,0039	0,0572
<b>KDWNC</b>	1,0017	0,0299	<b>LINT3</b>	1,0084	0,0604
<b>KLCPL</b>	1,0038	0,0295	<b>VINT1</b>	1,0063	0,0601
<b>KBRCL</b>	1,0023	0,0283	<b>VINT2</b>	1,0010	0,0616
<b>LWFRI</b>	1,0009	0,0293	<b>VINT3</b>	1,0031	0,0610
<b>CHFC1</b>	0,9982	0,0302	<b>VINT4</b>	0,9914	0,0583
<b>CHFC2</b>	0,9991	0,0299	<b>SBHTC</b>	1,0031	0,0575
<b>GTHCO</b>	0,9982	0,0577	<b>DNBHC</b>	0,9987	0,0581
<b>CLCON</b>	0,9973	0,0568	<b>WH001</b>	0,9950	0,0569
<b>CLDEN</b>	0,9962	0,0278	<b>WH002</b>	1,0033	0,0594
<b>CONTP</b>	1,0063	0,0858	<b>WH003</b>	0,9958	0,0584
<b>CONVL</b>	0,9970	0,0565	<b>WH004</b>	1,0052	0,0562
<b>NUCBO</b>	0,9960	0,0580	<b>WH005</b>	0,9877	0,0563
<b>CONVV</b>	1,0037	0,0602			

### 6.1 Clasificación del transitorio PTC.

Como se ha comentado previamente en los objetivos específicos, apartado 2.2 del presente trabajo, no sólo es importante la estimación de la MPCT y localización de la rotura, también es importante la identificación de la rama en la que se ha producido el LBLOCA antes de que la PCT alcance su valor máximo, sea cual sea.

En la figura 13 se puede ver como la PCT evoluciona de una forma concreta en función de donde se haya producido la rotura y, alcanza su valor máximo en distintos instantes de tiempo.

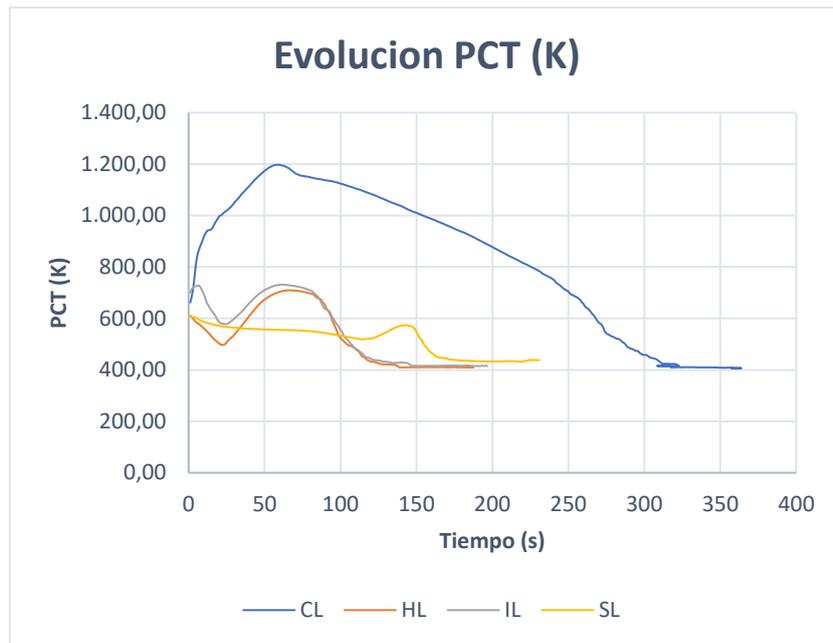


Figura 13: Evolución PCT media (K)

Para la construcción de un clasificador que identifique la localización del accidente en función de la evolución de la PTC, se van a tomar los valores de la PTC correspondiente a los 20 primeros instantes de tiempo. Esto se debe a que se quiere identificar la rama de la rotura antes de que se alcance la MPCT. Además, como se puede ver en la figura 13, el comportamiento de las series temporales PCT es muy distinto en los primeros instantes de tiempo, lo que indica que debería ser suficiente para una correcta clasificación. En la figura 14 se puede ver la evolución temporal de todas las observaciones con las que se va a construir el clasificador.

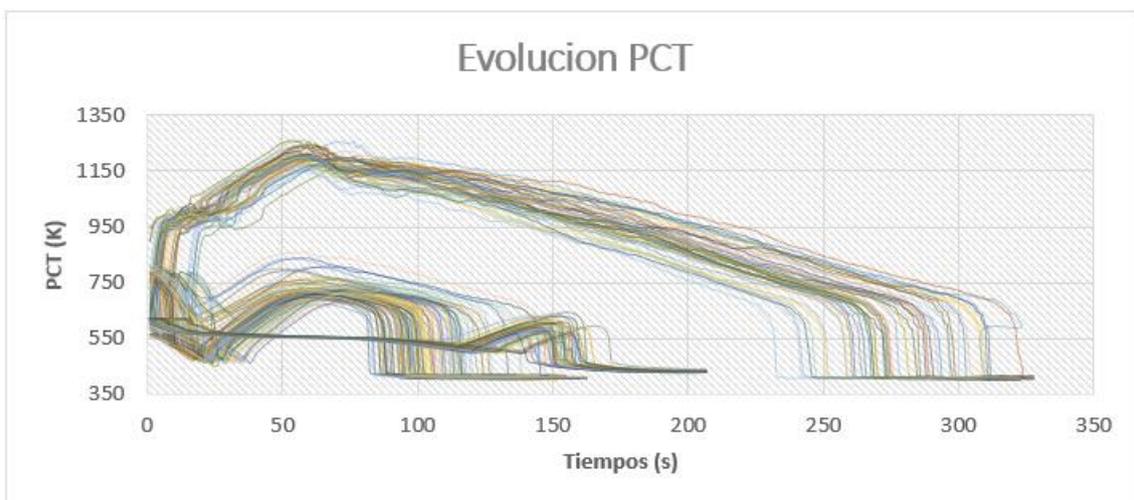


Figura 14: Evolución temporal de la PCT.

Como método de clasificación de las series temporales se ha hecho uso del algoritmo SVM. Como función kernel se ha utilizado la gaussiana o normal y se han estandarizado los datos. Para validar el modelo se ha utilizado una matriz de confusión para calcular la precisión del modelo y

se probará el modelo a través de validación cruzada con 5 particiones. Además, se mostrará las curvas ROC para cada clase, figura 15, y el diagrama de dispersión para los dos primeros instantes temporales, figura 16.

En la tabla 4 se muestra la matriz de confusión y el valor de la precisión del modelo es del 98%. Como se puede ver en la figura 15, se ha conseguido una muy buena clasificación, por lo que a partir de los instantes temporales iniciales ya se podría predecir donde se ha producido la rotura con una eficiencia muy alta. En la figura 16 se muestran las curvas ROC para las diferentes localizaciones de la rotura.

Tabla 4: Matriz confusión clasificador temporal PCT.

Loc. Real	Loc. Predicha			
	Cl pred.	HI pred	II pred.	SI pred.
Cl	48	0	1	0
HI	0	47	2	1
II	0	0	50	0
SI	0	0	0	50

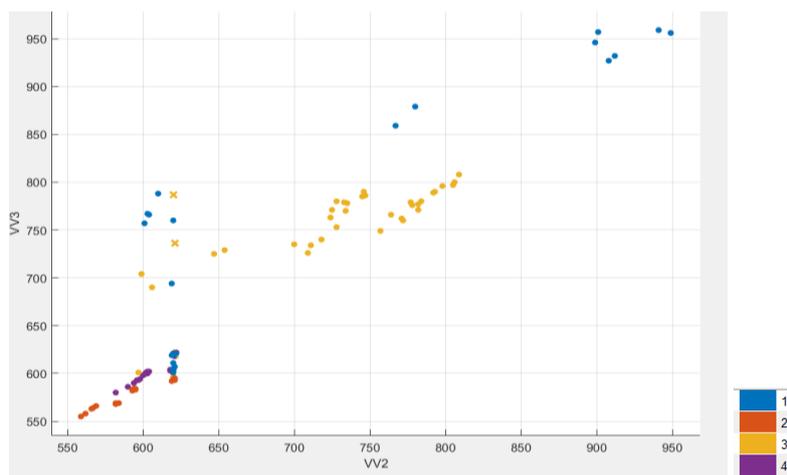


Figura 15: Grafico de dispersión de la PCT en t=1 y t=2.

Cl = 1    HI = 2    II = 3    SI = 4

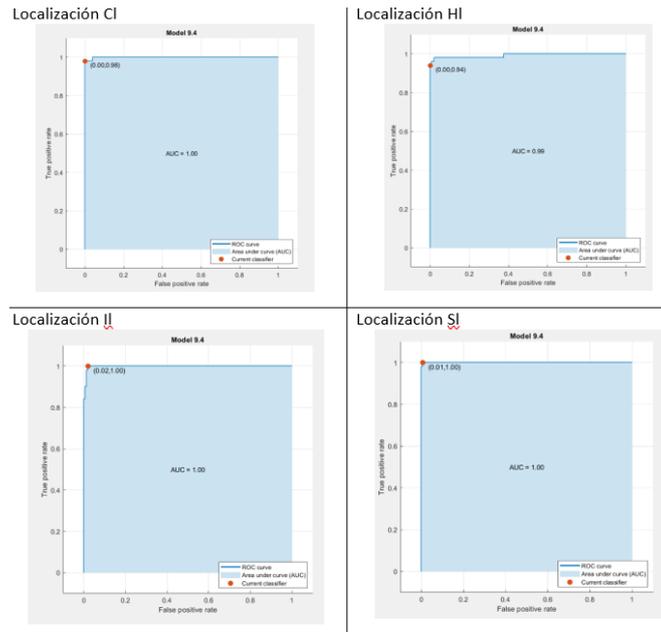


Figura 16: Cuvas ROC localización de la rotura.

## 6.2 Clustering.

Como se puede ver al inicio de este apartado, según la localización de la rotura la distribución de la temperatura MPCT cambia. Con la técnica de clustering k-medias, con  $k = 4$ , se ha agrupado las observaciones según la temperatura MPCT, con el objetivo de poder aplicar las técnicas de regresión y de clasificación. En la figura 17 y la tabla 5 se pueden ver los gráficos de las distribuciones de temperatura resultantes, su media y su desviación estándar.

Tabla 5: Caracterización de la MPCT de los clústeres formados.

Localización del fallo	Media (K)	Desviación estándar (K)
CI	1206.364	29.98
HI	716.003	19.433
II	783.676	17.294
SI	622.201	4.534

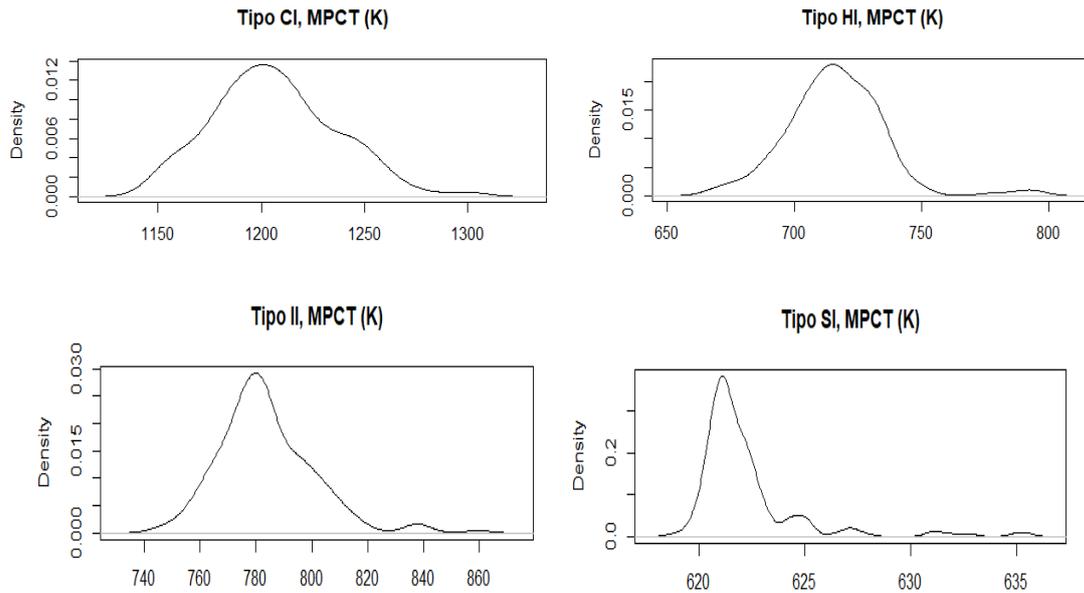


Figura 17: Distribución temperatura de los clústeres.

Como se puede ver en la tabla 5 y en la figura 17, las distribuciones de la MPCT de los 4 clústeres creados es muy similar a la de la clasificación inicial, que se puede ver en la tabla 2.

### 6.3 Estimación de la MPCT

Una vez obtenidos los 4 clústeres en el apartado anterior, se van a aplicar a cada uno de ellos las técnicas de regresión, descritas anteriormente, con objeto de saber cuál de ellas es la más adecuada para la predicción de la MPCT.

#### 6.3.1 Regresión Lineal múltiple con técnicas de penalización.

A continuación, se van a mostrar los resultados obtenidos con las distintas técnicas de penalización. Al final del apartado se mostrará una tabla comparativa.

Para la construcción de los modelos se ha usado una muestra de la población correspondiente al 70% de las observaciones de cada población, elegidas de forma semialeatoria y, para su validación, el 30% restante de las observaciones.

##### 6.3.1.1 Regresión Ridge.

El resultado de un ajuste por regresión ridge depende del parámetro  $\lambda$ , que determina el grado de penalización. Como no se sabe que valor es el que minimiza el error, se ha procedido a representar el valor de los 59 coeficientes de regresión en función de  $\log(\lambda)$ . En la figura 18, se puede ver la evolución de los coeficientes en función del parámetro de complejidad  $\lambda$ .

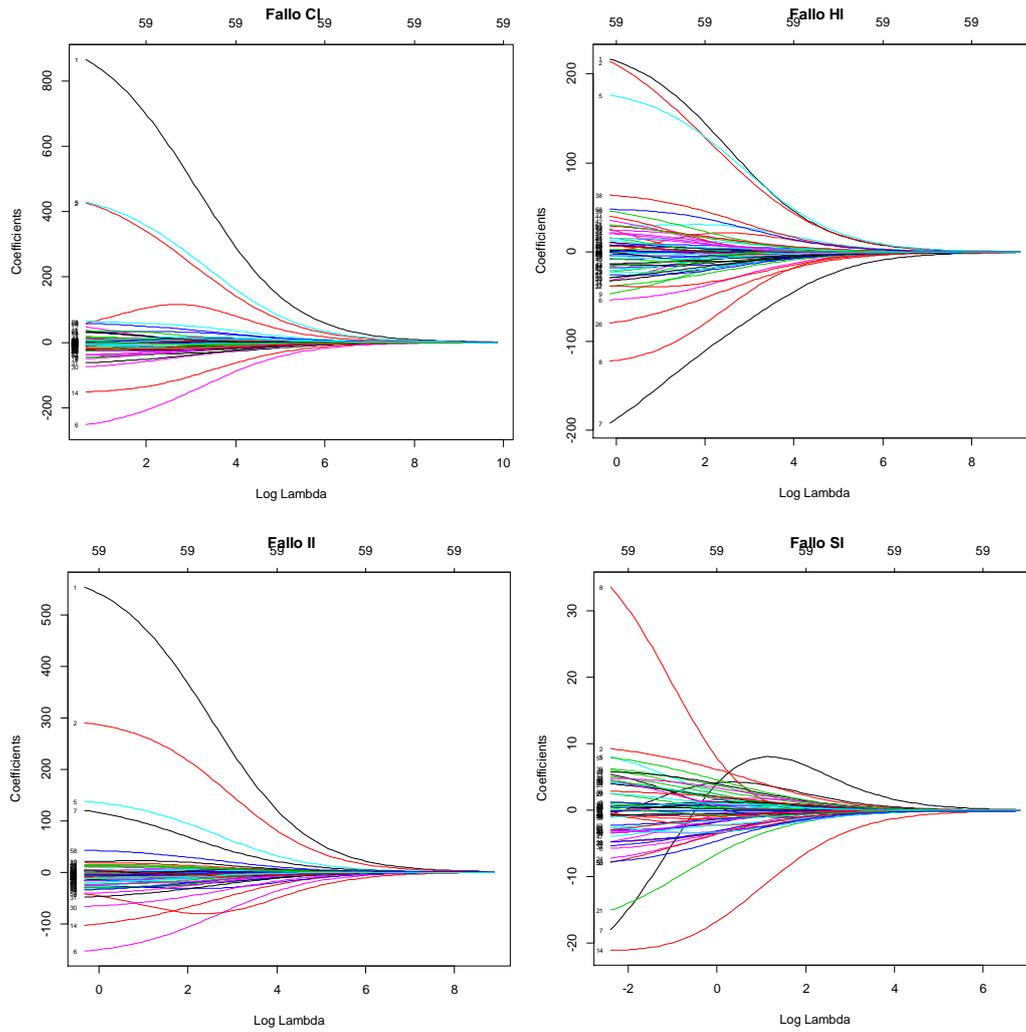


Figura 18: Evolución de los coef. de regresión en función de  $\lambda$ .

Como era de esperar, los coeficientes se van contrayendo a medida que aumenta el valor de  $\lambda$ . En la figura 19, se puede ver un gráfico complementario al anterior, donde vemos como aumenta la devianza en función del valor de los coeficientes de regresión.

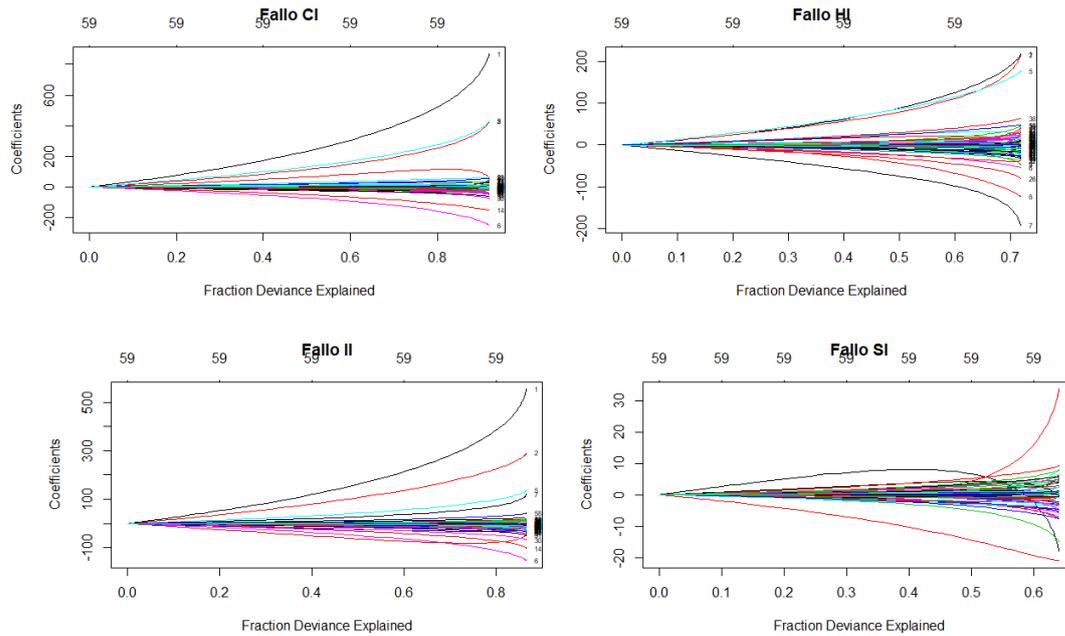


Figura 19: Fracción de la devianza explicada en función de los coef. de regresión.

Con el fin de identificar qué valor de  $\lambda$  da lugar a mejor modelo, se ha recurrido a su cálculo a través de una validación cruzada. En concreto se hará una validación cruzada con  $k = 20$  y para elegir el mejor modelo se evaluará el MSE del modelo según su valor de  $\lambda$ .

Los gráficos que se muestran a continuación, en las figuras 20 y 21, nos indican el MSE para cada valor de  $\lambda$ . A partir de estos gráficos podemos extraer el  $\lambda$  con el que se consigue el menor MSE y el  $\lambda$  con el que se consigue el modelo más sencillo que se aleja menos de una desviación estándar del mínimo MSE posible. Esos valores se muestran en la tabla 6.

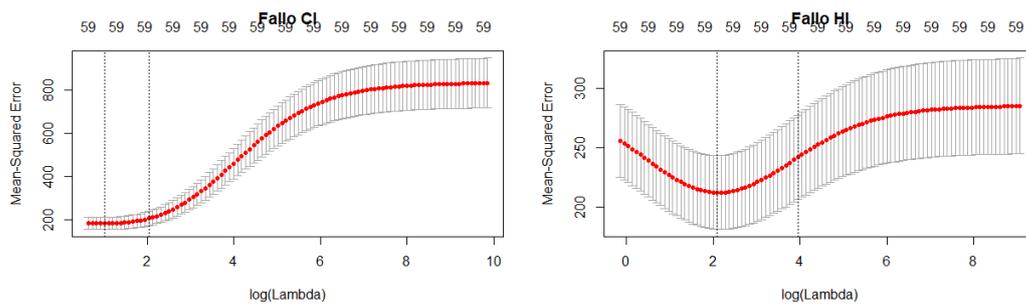


Figura 20: MSE en función de  $\lambda$  (1).

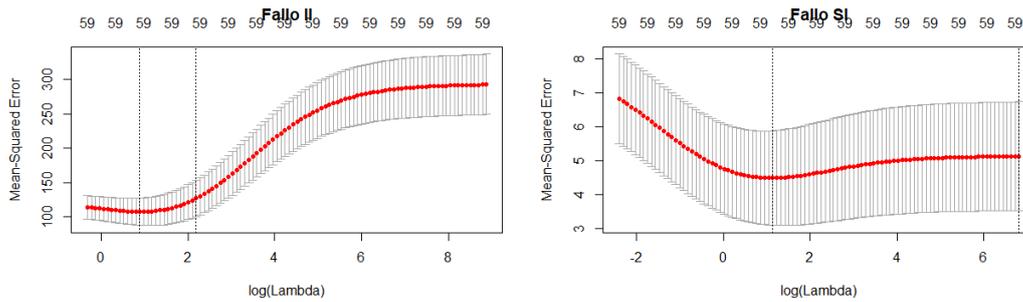


Figura 21: MSE en función de  $\lambda$  (2).

Tabla 6: Valores de  $\lambda$

Localización de la rotura	$\lambda_{min[MSE]}$	$\lambda_{1se}$
CI	2.754	7.664
HI	8.090	52.007
II	2.421	8.906
SI	3.116	908.499

Acorde a la norma de *one standard error rule*, el mejor modelo es el que se obtiene con los valores de  $\lambda_{1se}$ .

Una vez elegida el valor de  $\lambda$ , los valores de los coeficientes de regresión para los 4 modelos se muestran a continuación, en la tabla 7.

Tabla 7: Coef. regresión Ridge

$\beta$	Localización del Fallo			
	CI	HI	II	SI
(Intercept)	275,244	661,526	605,168	622,054
INPOW	691,628	47,569	342,291	0,034
PEAKF	337,037	44,760	205,172	0,026
HGSAV	9,331	-1,143	-4,520	0,00
HGSHR	32,541	3,746	-9,125	0,001
RPOWM	355,004	49,949	88,076	-0,003
UO2TC	-205,825	-12,686	-100,328	-0,003
UO2SH	-36,955	-45,563	62,719	0,112
IMCPS	107,163	-18,459	-80,602	-0,009
BMCPS	-37,496	-4,721	-18,029	0,014
ACCPR	-13,971	-1,408	-11,862	-0,002
ACCLT	-0,140	-0,039	-0,281	0,00
LPISQ	17,995	4,798	-5,322	0,013
PRZLE	17,694	-5,435	-3,810	-0,016
MASFR	-134,556	0,819	-63,759	-0,078
CLTEM	1,676	0,694	1,836	0,002
ACCKL	3,803	1,399	-1,894	-0,002
KSURG	2,460	-0,348	2,156	0,00
KDWNC	-35,982	-4,767	6,640	0,019
KLCPL	-12,410	0,059	18,052	0,012
KBRCL	3,772	13,784	9,187	-0,014

LWFRI	30,315	-4,707	-5,018	-0,017
CHFC1	0,661	-7,661	-31,817	0,007
CHFC2	57,927	17,716	-5,034	0,002
GTHCO	-3,008	-2,840	2,664	-0,011
CLCON	-18,501	-0,098	-23,722	0,008
CLDEN	-16,391	-18,592	1,247	-0,013
CONTP	-18,544	-9,052	6,120	0,00
CONVL	-26,770	3,172	-19,857	-0,019
NUCBO	-15,282	-0,851	-12,119	-0,001
CONVV	-57,501	0,671	-44,317	-0,019
FILMB	-51,162	-6,992	-30,405	0,011
TMIFS	2,250	-14,906	12,276	-0,002
TRANB	11,139	1,074	-21,957	0,007
CRIFH	-3,558	0,153	-19,583	-0,010
FLOSS	-0,836	-1,753	2,624	0,007
DRAG1	-4,975	5,849	3,248	-0,001
DRAG2	0,927	-4,558	-4,120	0,011
DRAG3	13,751	17,004	-3,568	0,010
DRAG4	7,936	4,887	-14,198	0,016
DRAG5	4,775	-4,372	-5,851	0,009
DRAG7	-6,229	4,936	-10,699	-0,007
CCFLC	-34,889	0,795	-17,483	-0,006
ACCLV	1,698	-3,198	0,660	0,001
VAPIA	2,292	-0,752	-4,181	0,00
LIQIA	-3,165	-6,075	-6,982	0,005
UHTEM	0,122	0,147	0,172	0,00
LINT2	-5,762	-6,909	5,275	-0,017
LINT3	-25,364	0,755	-10,114	0,016
VINT1	-24,005	-6,459	2,911	0,019
VINT2	1,277	6,310	-10,262	-0,010
VINT3	-20,049	5,101	10,192	0,018
VINT4	6,693	-7,235	-2,015	-0,007
SBHTC	3,840	3,212	-7,367	-0,020
DNBHC	-30,370	1,599	-24,572	-0,003
WH001	18,571	-5,287	-5,405	-0,008
WH002	-20,017	-2,419	-4,858	-0,001
WH003	-4,648	1,180	8,207	-0,007
WH004	48,747	14,571	27,297	-0,014
WH005	13,503	-2,486	-9,602	0,003

Como era de esperar, al tratarse de regresión Ridge, ha seleccionado todas las variables explicativas como regresores.

Los valores de MSE, BIC, AIC y de la devianza para los modelos se muestran en la tabla 8.

Tabla 8: Medición de parametros del ajuste.

Localización	MSE	Devianza	BIC	AIC
CI	168.20	0.88	-101160.2	-101243.7
HI	518.36	0.33	-11704.58	-11768.74
II	141.67	0.76	-30112.43	-30032.1
SI	9.38	0.004	280.85	229.60

Si representamos los valores predichos en función de los observados del set de datos test obtenemos la figura 22, que se muestra a continuación.

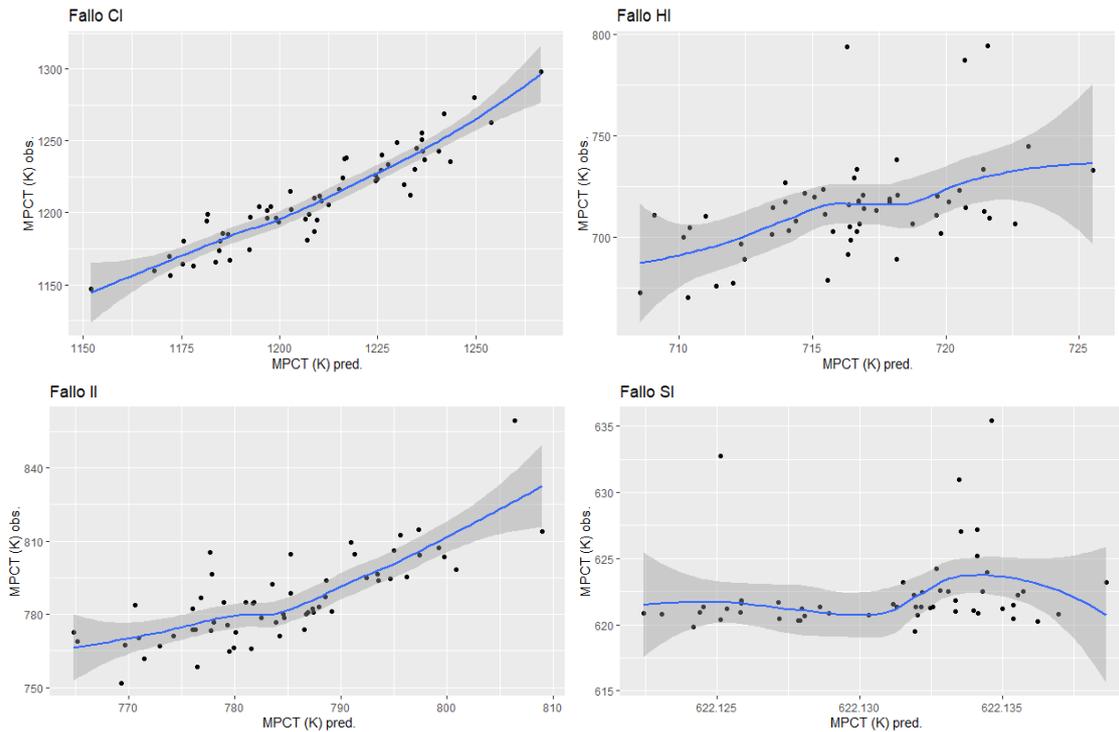


Figura 22: Obs. vs Pred. de los 4 modelo Ridge.

#### 6.3.1.1.1 Análisis residual.

Con objeto de validar los modelos se van a mostrar 2 gráficos de validación:

- Gráfico de los residuos en un gráfico probabilístico normal, para saber si siguen una distribución normal.
- Gráfico de observados vs residuos, para comprobar si los residuos se comportan como ruido blanco.

En la figura 23, se muestran los gráficos de normalidad y en la figura 24 los valores predichos frente a los residuos.

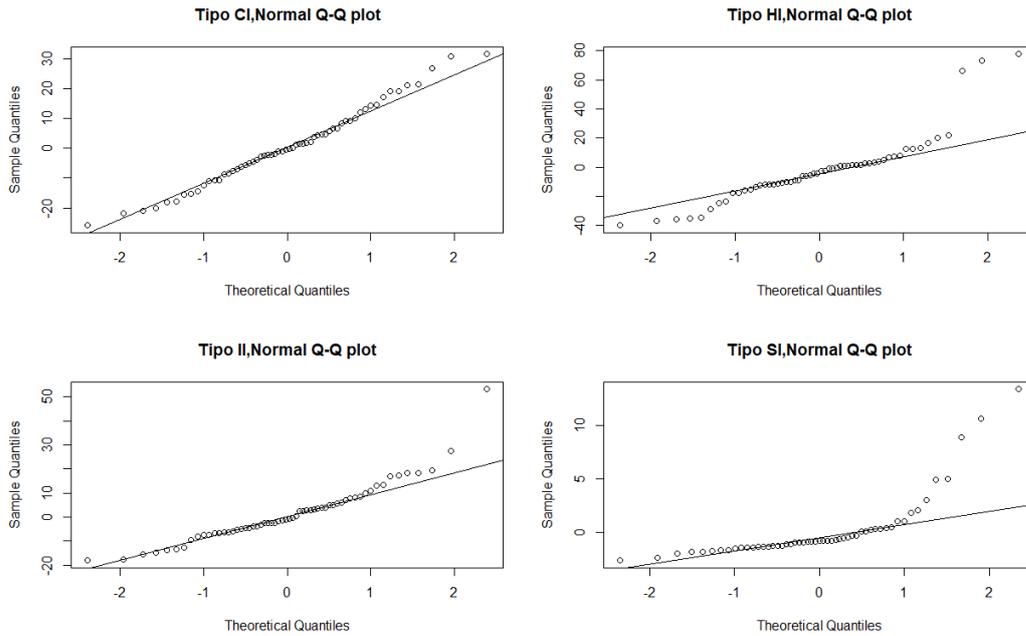


Figura 23: Grafico probabilístico normal de los residuos Ridge.

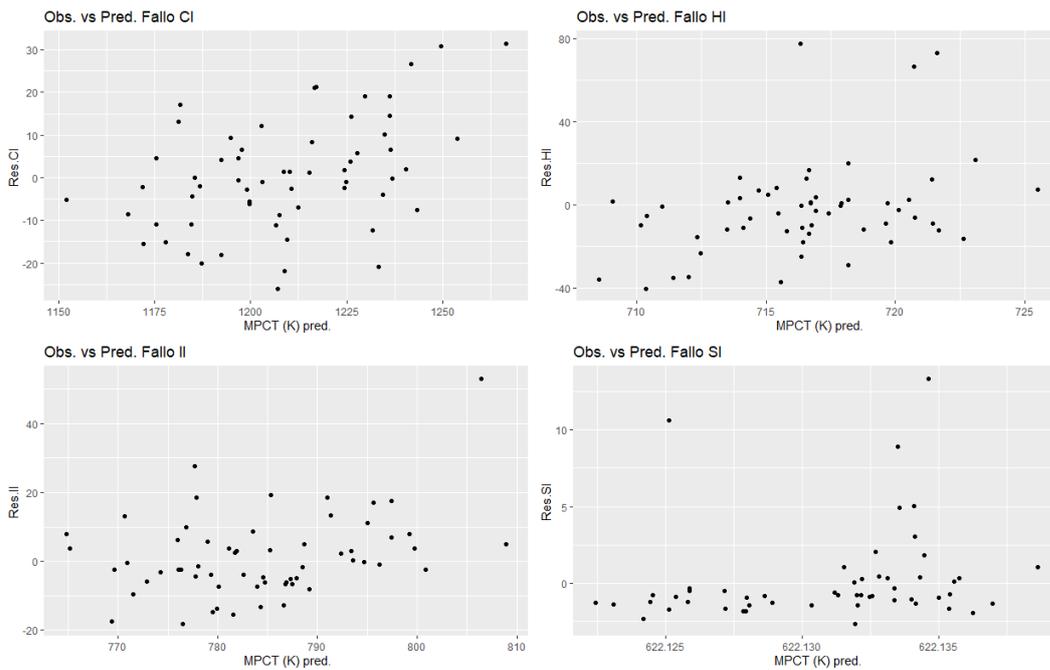


Figura 24: Residuos. vs Pred. de los modelos Ridge.

Como se puede ver en el grafico probabilístico normal, las predicciones de la MPCT en el caso de que la rotura se localice en la rama CI e II tiene un buen ajuste a una distribución normal. Sin embargo, en los casos en el que la rotura se localiza en las ramas HI y SI los valores estimados no se ajustan tan bien a una distribución normal. Esto puede ser debido a que la desv. Estándar de la MPCT en ambos casos es muy pequeña, en comparación con la magnitud de la MPCT, por lo que es normal encontrar valores que sea más de 3 veces superior a la desviación estándar.

En el gráfico de los residuos frente a los valores observados, podemos ver como hay homocedasticidad en el case de HI e II. No obstante, en el caso HI y SI no está tan claro debido a la presencia de valores muy extremos.

### 6.3.1.2 Regresión LASSO.

La principal diferencia entre la regresión Ridge y la Lasso, es que la Lasso tiene un mayor poder de contracción de los coeficientes. Esto significa que inicialmente ya sabremos que el número de coeficientes de regresión será menor y, por lo tanto, el modelo será más fácil de interpretar.

La forma de trabajar en este tipo de regresión será similar al de la regresión ridge, por lo que el primer paso será representar los valores de los coeficientes de regresión en función del parámetro de complejidad  $\lambda$ , el cual se muestra en la figura 25.

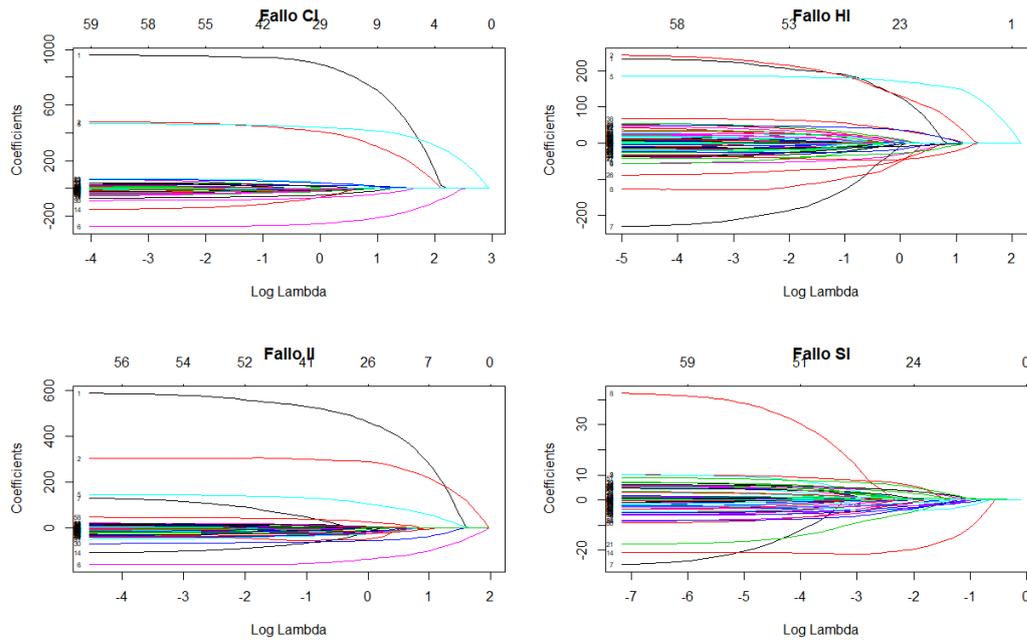


Figura 25: Evolución de los coef. de regresión en función de  $\lambda$ .

Como se puede ver, los valores de los coeficientes de regresión tienden mucho más rápido a contraerse que en la regresión Ridge.

A continuación, se muestran, en la figura 26 y 27, como aumenta el valor de la devianza en función de los valores de los coeficientes de regresión.

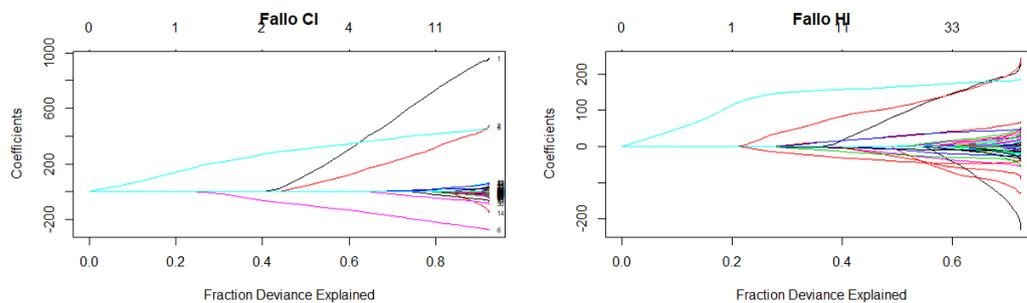


Figura 26: Fracción de la devianza explicada en función de los coef. de regresión (1).

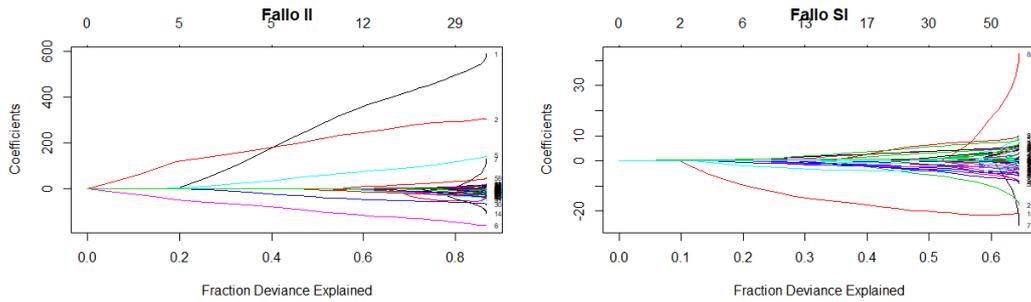


Figura 27: Fracción de la devianza explicada en función de los coef. de regresión (2).

Al igual que antes, para elegir el mejor valor de  $\lambda$  se hará una validación cruzada con  $k=20$  y se elegirá el valor de lambda que de un menor MSE. En la figura 28, que se muestra a continuación se ve la evolución del MSE en función de  $\lambda$ .

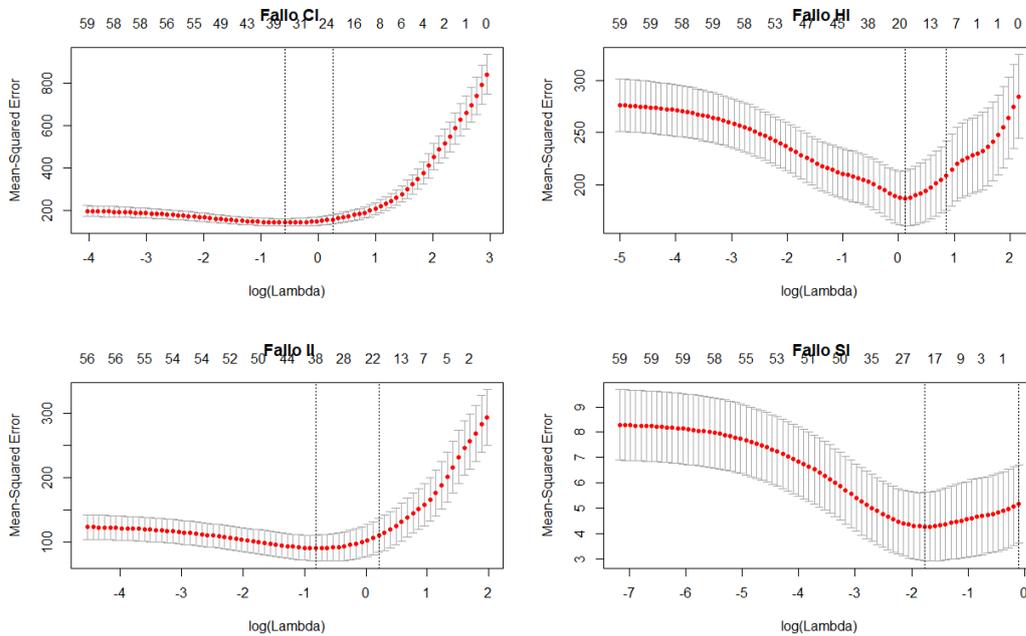


Figura 28: MSE en función de  $\lambda$ .

Como se puede ver en las figuras anteriores, el número de variables explicativas va disminuyendo a medida que aumenta el parámetro de complejidad  $\lambda$ . Esto no pasaba en la anterior regresión donde el número de variables explicativas era constante. Los valores de  $\lambda$  que minimizan el valor se pueden ver en la tabla 9.

Tabla 9: Valores de  $\lambda$ .

Localización	$\lambda_{min}[MSE]$	$\lambda_{lse}$
CI	0.553	1.278
HI	1.120	2.358
II	0.443	1.233
SI	0.170	0.908

Al igual que en el apartado anterior, y como norma para los siguientes apartados, se seleccionará el valor de  $\lambda_{lse}$  para el parámetro de complejidad.

Una vez elegido el parámetro de complejidad, los coeficientes de regresión obtenidos se muestran en la tabla 10.

Tabla 10: Coef. regresión Lasso.

Localización de la rotura				
B	CI	HI	II	SI
(Intercept)	-128,764	508,714	288,057	622,132
INPOW	861,650	0,00	439,835	0,00
PEAKF	393,496	67,601	281,140	0,00
HGSHR	23,265	0,00	-1,255	0,00
RPOWM	434,110	154,402	97,045	0,00
UO2TC	-244,407	0,00	-133,011	0,00
IMCPS	0,00	0,00	-41,552	0,00
BMCPS	-18,275	0,00	-6,259	0,00
ACCPR	0,00	0,00	-0,465	0,00
ACCLT	0,00	0,00	-0,070	0,00
MASFR	-43,193	0,00	-0,369	0,00
CLTEM	0,725	0,00	1,314	0,00
ACCKL	0,067	0,00	0,00	0,00
KSURG	1,715	0,00	0,00	0,00
CHFC2	36,054	0,00	0,00	0,00
GTHCO	0,00	0,00	0,00	0,00
CLCON	-11,288	0,00	-18,010	0,00
CLDEN	-4,594	0,00	0,00	0,00
CONTP	-7,190	-10,320	1,447	0,00
CONVL	0,00	0,00	-14,281	0,00
CONVV	-60,630	0,00	-55,284	0,00
FILMB	-41,130	0,00	-23,009	0,00
TMIFS	0,00	-26,901	0,00	0,00
TRANB	0,00	0,00	-12,634	0,00
CRIHF	0,00	0,00	-23,135	0,00
DRAG3	0,00	11,620	0,00	0,00
DRAG4	0,00	0,00	-1,361	0,00
DRAG7	-1,178	0,00	0,00	0,00
CCFLC	-7,284	0,00	0,00	0,00
ACCLV	0,00	-3,229	0,00	0,00
DNBHC	-9,341	0,00	-13,524	0,00
WH001	12,382	0,00	0,00	0,00
WH002	-3,371	0,00	0,00	0,00
WH004	35,820	11,268	20,265	0,00

Como era de esperar el número de coeficientes de regresión es mucho menor que el caso de la regresión ridge, además, este será el mínimo número de coeficientes obtenidos por los métodos de regresión con regularización.

Los valores de MSE, BIC, AIC y de la devianza para los diferentes modelos obtenidos se muestran en la tabla 11.

Tabla 11: Medicion de parametros del ajuste.

Localización	MSE	Devianza	BIC	AIC
CI	149.658	0.865	-99369.61	--99454.61
HI	467.983	0.361	-12936.4	-12955.43
II	120.189	0.721	-28483.12	-28536.41
SI	9.38	0	0	0

Si representamos los valores predichos en función de los observados del set de datos test, los resultados obtenidos se muestran en la figura 29.

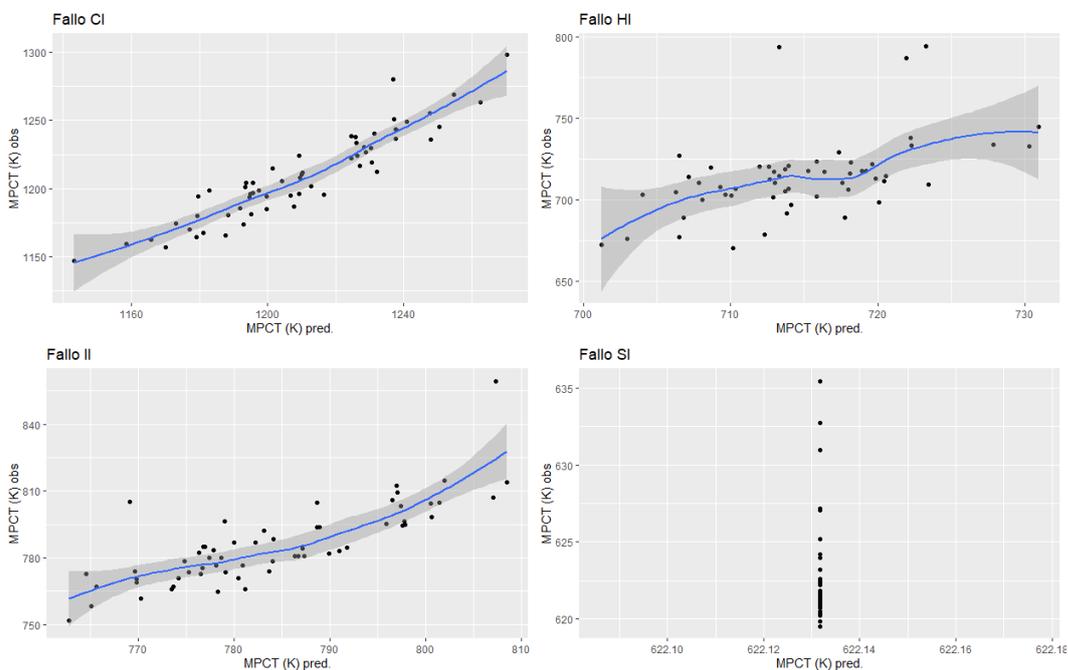


Figura 29: Obs. vs Pred. de los 4 modelo Lasso.

Como se puede ver en el caso de la rotura con localización SI, el método de regresión LASSO ha contraído todo lo posibles los coeficientes, dejando únicamente la media de los valores como regresor. Esto, repercutirá en la validación del modelo y en el análisis de los residuos.

### 6.3.1.2.1 Análisis de los residuos.

En la figura 30 se muestra el gráfico probabilístico normal de los residuos y el gráfico de valores predichos frente a residuos en la figura 31.

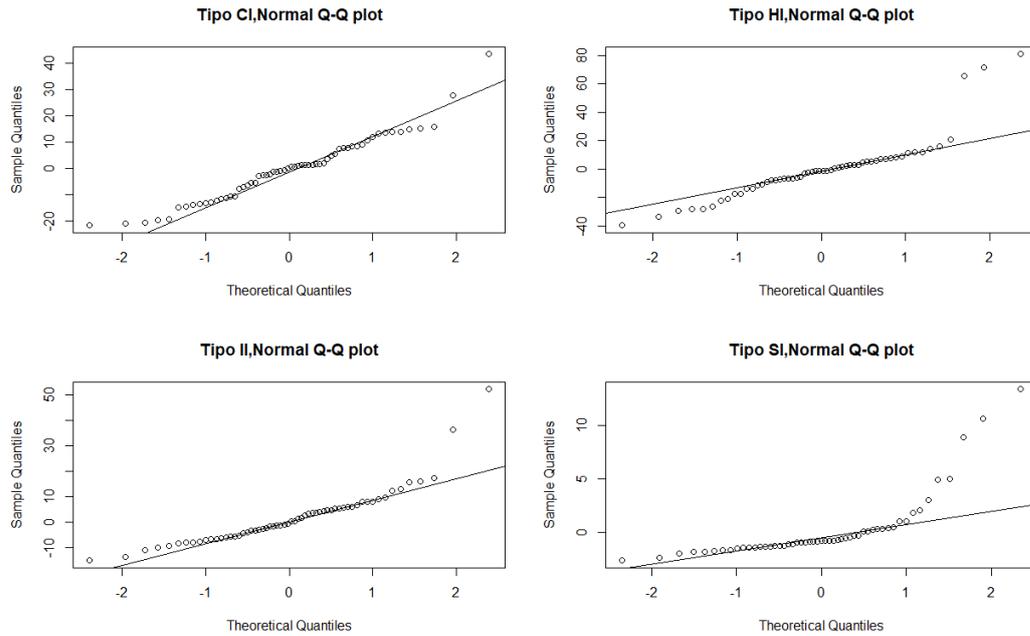


Figura 30: Grafico probabilístico normal de los residuos Lasso.

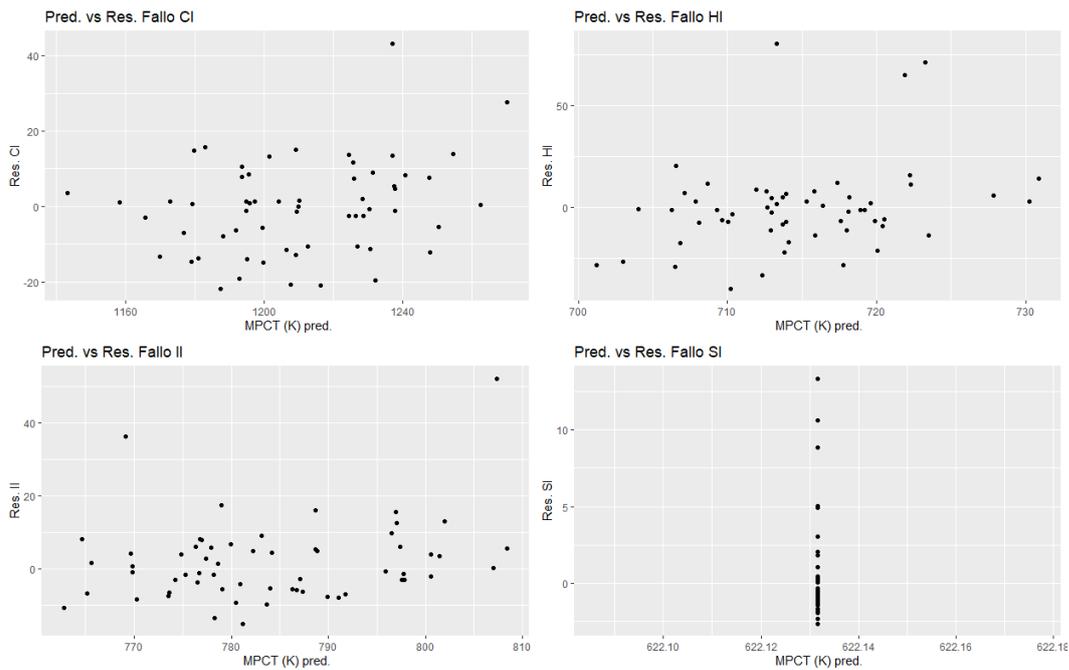


Figura 31: Residuos vs Pred. de los modelos Lasso.

Como se puede ver en la figura del gráfico probabilístico normal, los valores predichos para las localizaciones CI, II y HI se ajustan de forma aproximada a una distribución normal, aunque el ajuste es peor que en el caso de la regresión Ridge. En el caso de que la rotura se de en la localización SI tenemos una situación similar al de la regresión Ridge: Un ajuste aproximado a la normalidad, pero con valores muy extremos que rompen esa normalidad.

De la figura de predichos frente a los residuos, lo más destacable se da en la rotura SI donde, al haber contraído todos los coeficientes, solo tenemos como valores predichos la media.

### 6.3.1.3 Regresión Elastic Net.

Como se ha mencionado previamente, el método Elastic Net intenta aunar las bondades de la regresión Ridge y la Lasso. Es por esto por lo que tenemos que ajustar 2 parámetros.

- $\alpha$  : Comprendido entre 1 y 0. Si  $\alpha = 0$  estaremos antes el caso Ridge regresión y, si  $\alpha = 1$  ante la regresión Lasso.
- $\lambda$  : Parámetro de complejidad (El mismo parámetro de penalización por complejidad que en los casos anteriores).

A continuación, se va a repetir el mismo procedimiento que con las regresiones Ridge y Lasso, pero esta vez con la regresión Elastic Net, distintos valores de  $\alpha$ . Además, a través de validación cruzada, se ha obtenido el valor óptimo de  $\alpha$  que maximiza la devianza, este modelo se comparara con los anteriores.

Debido al gran número de imágenes y de coeficientes de regresión, se va a representar una serie de tablas, 12-16, con las métricas de error para los distintos valores de  $\alpha$ , y se van a presentar con más detalle los resultados obtenidos con el valor de  $\alpha$  seleccionado.

Tabla 12: Medición de parámetros del ajuste  $\alpha=0.1$ .

$\alpha = 0,1$						
Loc.	$\lambda_{\min}[\text{MSE}]$	$\lambda_{1\text{se}}$	MSE	Devianza	BIC	AIC
CI	2.395	6.073	159.336	0.868	-99663.43	-99741.66
HI	6.074	16.255	500.269	0.333	-11890.73	-11926.945
II	2.105	2.859	129.57	0.737	-29078.33	-29155.58
SI	1.287	9.084	9.389	0.0	4.804	2.03

Tabla 13: Medición de parámetros del ajuste  $\alpha=0.25$

$\alpha = 0,25$						
Loc.	$\lambda_{\min}[\text{MSE}]$	$\lambda_{1\text{se}}$	MSE	Devianza	BIC	AIC
CI	2.017	3.868	153.860	0.865	-99401.74	-99474.16
HI	3.089	9.433	495.165	0.304	-10878.05	-10899.66
II	1.341	7.732	122.791	0.724	-28566.51	-28629.67
SI	0.62	3.634	9.389	0.00	0	0

Tabla 14: Medición de parámetros del ajuste  $\alpha=0.5$

$\alpha = 0,5$						
Loc.	$\lambda_{\min}[\text{MSE}]$	$\lambda_{1\text{se}}$	MSE	Devianza	BIC	AIC
CI	1.008	2.557	157.38	0.857	-98483.89	-98541.8
HI	2.042	5.176	486.43	0.316	-11303.7	-11322.73
II	0.886	2.247	121.20	0.721	-28463.49	-28518.85
SI	0.310	1.817	9.389	0	0	0

Tabla 15: Medición de parámetros del ajuste  $\alpha = 0.75$

$\alpha = 0,75$						
Loc.	$\lambda_{\min}[\text{MSE}]$	$\lambda_{1se}$	MSE	Devianza	BIC	AIC
CI	0.738	1.553	145.69	0.87	-99920.12	-99981.9
HI	1.493	3.451	479.35	0.33	-11786.01	-11805.05
II	0.538	1.644	121.97	0.71	-28436.15	-28282.8
SI	0.206	1.211	9.389	0	0	0

Tabla 16: Medición de parámetros del ajuste  $\alpha = 0.99$

$\alpha = 0,99$						
Loc.	$\lambda_{\min}[\text{MSE}]$	$\lambda_{1se}$	MSE	Devianza	BIC	AIC
CI	0.559	1.417	155.164	0.85	-98551.7	-98605.5
HI	1.131	3.456	505.07	0.25	-8906.28	-8914.65
II	0.491	1.245	120.24	0.72	-28475.5	-28528.8
SI	0.156	0.917	9.389	0	0	0

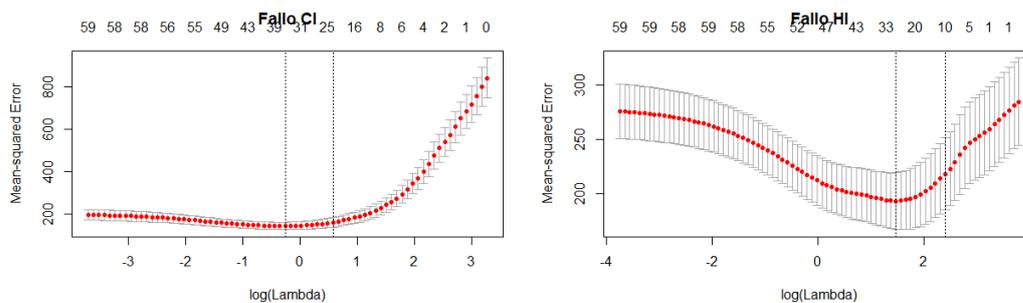
En la tabla 17 se muestra el valor de  $\alpha$ , que maximiza la devianza de las 4 localizaciones del fallo. Esta  $\alpha$ , se ha estimado a través de validación cruzada.

Tabla 17: Medición de parámetros del ajuste  $\alpha$  óptimo.

$\alpha_{cv50}$							
Loc.	$\lambda$	$\alpha$	Nº, Var	MSE	devianza	BIC	AIC
CI	0.466	1	39	114.91	0.90	-104399.4	-104482.7
HI	7.561	0.195	14	473.01	0.42	-15145.22	-15181.43
II	0.743	1	27	107.134	0.79	-31355.68	-31420.6
SI	0.696	0.953	1	9.141	0.07	-41.61	-44.38

Como era de esperar, el mejor modelo es el obtenido optimizando los valores de  $\alpha$  y  $\lambda$ .

La figura 32 muestran los gráficos para los valores del MSE en función del parámetro de complejidad  $\lambda$ .



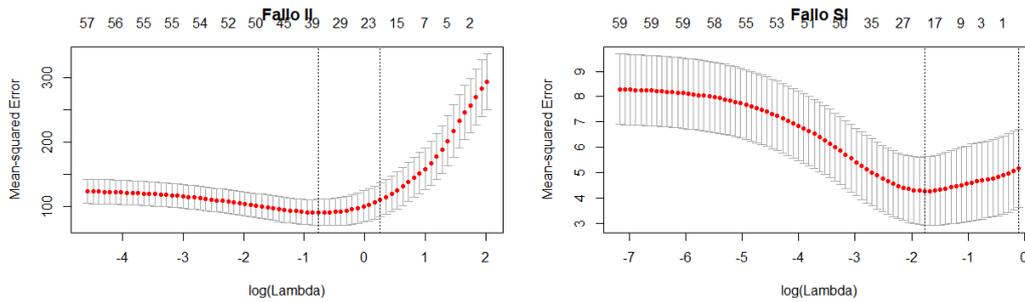


Figura 32: MSE en función de  $\lambda$ .

A continuación, en la tabla 18, se muestran los valores de los coeficientes de regresión para el modelo Elastic Net optimizado.

Tabla 18: Coef. regresión Elastic Net

Localización de la rotura				
$\beta$	CI	HI	II	SI
(Intercept)	-120,35	492,29	343,36	622,13
INPOW	939,34	49,31	491,43	0,00
PEAKF	439,09	83,62	291,77	0,00
HGSAV	6,88	0,00	0,00	0,00
HGSHR	23,20	0,00	-7,70	0,00
RPOWM	451,35	123,11	115,47	0,00
UO2TC	-264,67	-14,33	-145,57	0,00
UO2SH	0,00	0,00	0,00	0,00
IMCPS	0,33	0,00	-52,21	0,00
BMCP	-29,72	0,00	-15,99	0,00
ACCP	-3,24	0,00	-5,59	0,00
ACCL	0,00	0,00	-0,17	0,00
LPISQ	9,68	0,00	0,00	0,00
PRZLE	8,65	0,00	0,00	0,00
MASFR	-102,14	0,00	-41,08	0,00
CLTEM	1,63	0,62	1,92	0,26
ACCKL	4,94	0,00	0,00	0,00
KSURG	4,47	0,00	0,00	0,00
KDWNC	-19,42	0,00	0,00	0,00
KLCPL	0,00	0,00	0,00	0,00
KBRCL	0,00	0,00	0,00	0,00
LWFRI	28,96	0,00	0,00	0,00
CHFC1	0,00	0,00	0,00	0,00
CHFC2	58,49	7,83	0,00	0,00
GTHCO	0,00	0,00	0,00	0,00
CLCON	-19,49	0,00	-24,60	0,00
CLDEN	-19,24	-20,33	0,00	0,00
CONTP	-10,48	-16,43	7,51	0,00
CONVL	-11,70	0,00	-15,10	0,00

<b>NUCBO</b>	0,00	0,00	-3,11	0,00
<b>CONVV</b>	-76,56	0,00	-59,39	0,00
<b>FILMB</b>	-55,51	0,00	-32,43	0,00
<b>TMIFS</b>	3,52	-29,66	3,02	0,00
<b>TRANB</b>	0,00	0,00	-18,67	0,00
<b>CRIHF</b>	-0,13	0,00	-28,98	0,00
<b>FLOSS</b>	0,00	0,00	0,00	0,00
<b>DRAG1</b>	-6,50	0,00	0,00	0,00
<b>DRAG2</b>	0,51	0,00	0,00	0,00
<b>DRAG3</b>	0,00	23,48	0,00	0,00
<b>DRAG4</b>	0,00	0,00	-7,80	0,00
<b>DRAG5</b>	0,00	0,00	0,00	0,00
<b>DRAG7</b>	-10,42	0,00	-5,28	0,00
<b>CCFLC</b>	-23,29	0,00	-4,53	0,00
<b>ACCLV</b>	0,00	-5,40	0,00	0,00
<b>VAPIA</b>	4,96	0,00	0,00	0,00
<b>LIQIA</b>	0,00	0,00	0,00	0,00
<b>UHTEM</b>	0,00	0,11	0,00	0,00
<b>LINT2</b>	0,00	0,00	0,00	0,00
<b>LINT3</b>	-10,76	0,00	0,00	0,00
<b>VINT1</b>	-16,82	0,00	0,00	0,00
<b>VINT2</b>	0,00	0,00	-2,67	0,00
<b>VINT3</b>	-13,07	0,00	0,00	0,00
<b>VINT4</b>	0,00	-5,04	0,00	0,00
<b>SBHTC</b>	0,00	0,00	0,00	0,00
<b>DNBHC</b>	-26,84	0,00	-24,56	0,00
<b>WH001</b>	26,03	0,00	0,00	0,00
<b>WH002</b>	-9,25	0,00	0,00	0,00
<b>WH003</b>	0,00	0,00	0,00	0,00
<b>WH004</b>	51,54	21,97	29,67	0,00
<b>WH005</b>	0,04	0,00	-8,37	0,00

Como se puede ver, por el número de coeficientes de regresión con valor distinto de 0, los modelos no son tan simples como los obtenidos con la regresión Lasso, ni tan complejos como el de la regresión Ridge.

Por último, se muestran los gráficos de los valores predichos frente a los observados en la figura 33.

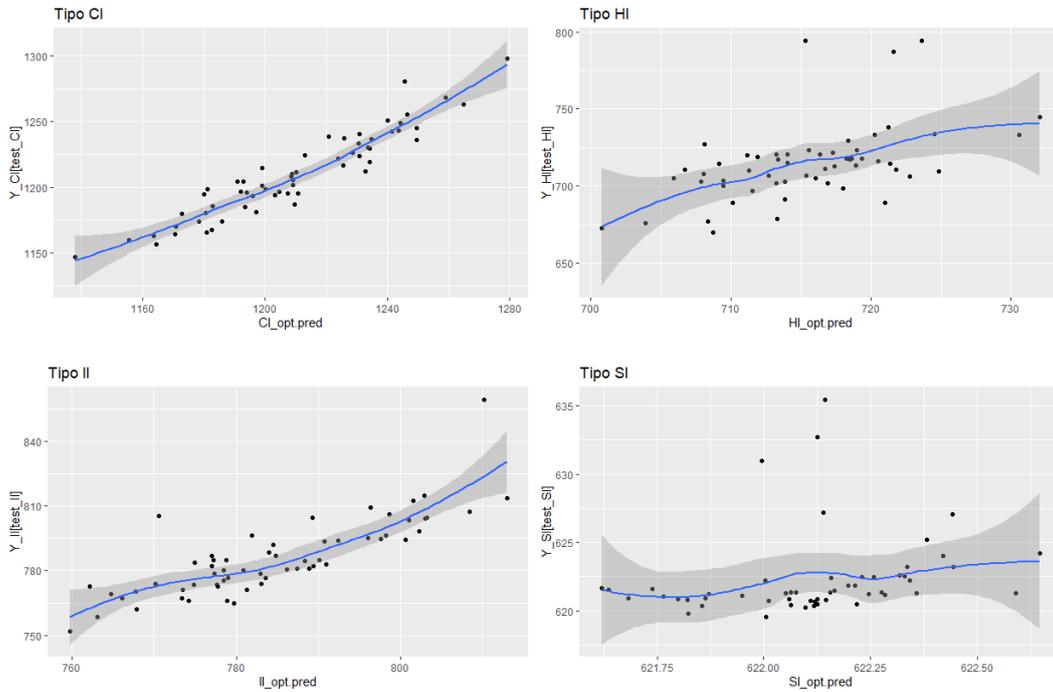


Figura 33: Obs. vs Pred. de los 4 modelo Elastic Net.

### 6.3.1.3.1 Análisis residual del modelo optimizado.

En primer lugar, en la figura 34, se puede ver el gráfico de los residuos en un gráfico probabilístico normal y en la figura 35, los gráficos de valores predichos frente a los residuos.

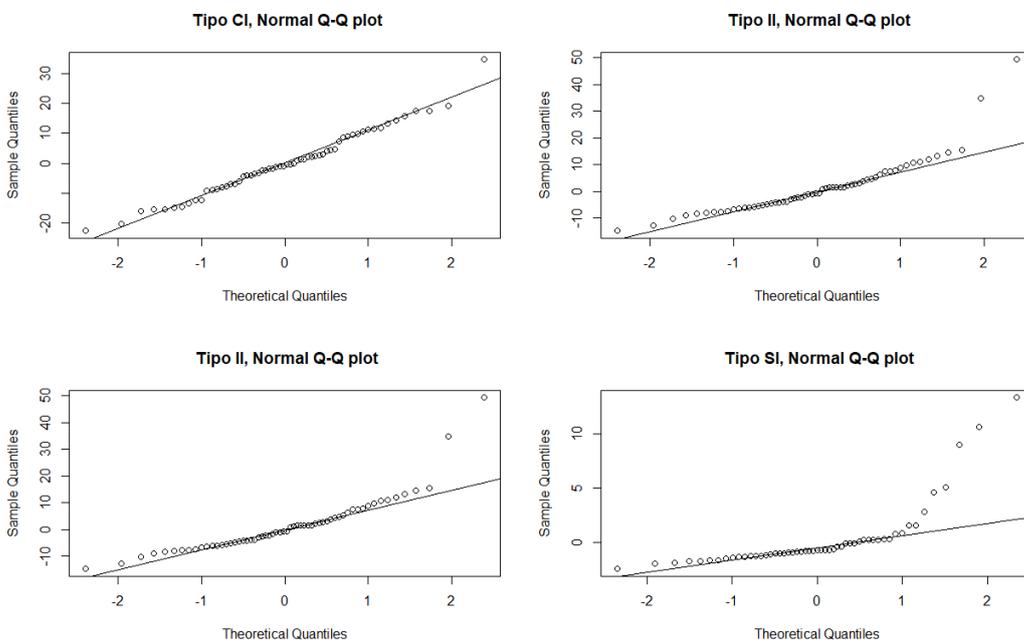


Figura 34: Grafico probabilístico normal de los residuos Elastic Net.

Como se puede ver, en la figura superior, en todos los casos se podría asumir que los residuos siguen una distribución normal. Aunque con los valores más extremos se pierde la normalidad cuando la localización del fallo se encuentra en SI.

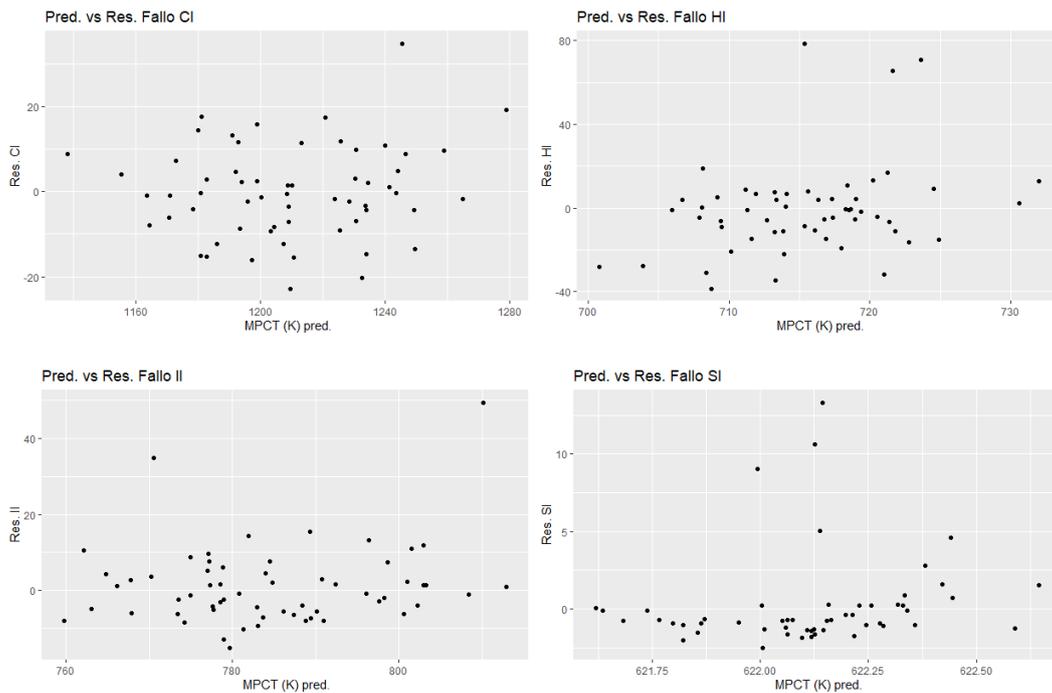


Figura 35: Residuos vs Pred. de los modelos Elastic Net.

De análisis de la figura 35, en el cual se analiza la homocedasticidad del conjunto de datos predichos, podemos ver que en principio los 4 gráficos se asemejan al gráfico que tendría una señal de ruido blanco, media 0 y homocedasticidad. Esto se cumple en los 4 casos, aunque cuando el fallo ocurre en la rama SI, se rompe un poco esta afirmación.

### 6.3.2 Modelo Lineal Generalizado (GLM).

La particularidad de este método de regresión es que permite ajustar el modelo a partir de una función enlace que ajuste la regresión a partir de la distribución de probabilidad de la variable dependiente. Se ha supuesto una función vínculo gaussiana.

En primer lugar, se han realizado 4 modelos, uno por cada localización de fallo. Una vez realizada el ajuste, se ha hecho una selección de variables por el método *stepwise*, tanto *backward* como *forward*.

Una vez hecha la selección de variables, se ha predicho los valores del conjunto de datos de prueba y se ha realizado un análisis de los residuos como en el apartado anterior.

En primero lugar, en la tabla 19, se muestran las variables seleccionadas por el algoritmo como significativas y los valores estimado para dichos coeficientes de regresión.

Tabla 19: Coef. regresión de los modelos GLM.

Localización de la rotura.				
$\beta$	CI	HI	II	SI
(Intercept)	1206,364	716,003	783,676	622,200
INPOW	8,721	-	4,863	-
PEAKF	9,477	3,581	6,475	0,285
HGSAV	2,414	-	-	-
HGSHR	2,630	-	-	-
RPOWM	16,588	8,031	5,606	0,310
UO2TC	-12,455	-2,347	-7,949	-
UO2SH	-	-	-	-
IMCPS	-	-	-0,940	-
BMCPS	-	-	-1,690	0,397
ACCPR	-	-	-1,266	-0,490
ACCLT	-1,127	-1,952	-1,247	0,265
LPISQ	-	-	-0,909	-
PRZLE	1,869	-	0,806	-
MASFR	-3,172	-	-2,435	-0,438
CLTEM	2,487	2,125	3,061	0,909
ACCKL	1,890	2,248	-	-0,257
KSURG	-	-	1,010	0,248
KDWNC	-0,977	-	-	-
KLCPL	-1,096	-	-	-0,317
KBRCL	-	-	0,814	-
LWFRI	-	-	-	-
CHFC1	1,280	-1,562	-	-
CHFC2	1,956	-	-1,122	-
GTHCO	-	1,616	-	-0,382
CLCON	-2,467	-	-2,415	-
CLDEN	-	-	-	-0,244
CONTP	-1,202	-2,143	0,827	-
CONVL	-1,250	-	-	-
NUCBO	-	-	-	-
CONVV	-5,204	-	-3,544	-
FILMB	-3,172	-2,589	-4,186	0,279
TMIFS	-	-4,727	-	-
TRANB	-	-	-	-
CRIHF	-1,493	-1,659	-4,557	-0,440
FLOSS	-	-	-	0,453
DRAG1	-	-	-	-
DRAG2	-	-1,539	-1,479	-
DRAG3	-	2,082	-	-
DRAG4	-	-	-1,328	0,268
DRAG5	-	-	-	-
DRAG7	-	-	-	-
CCFLC	-2,332	-	-	-

ACCLV	-	-2,265	-	-
VAPIA	-	1,532	-	-
LIQIA	-	-	-	0,264
UHTEM	-	3,015	-	-
LINT2	-	-	-	-
LINT3	-1,591	-	-	-
VINT1	-	-2,344	-	0,281
VINT2	-	-	-	-
VINT3	-1,988	-	-	0,439
VINT4	-	-2,432	-1,269	-
SBHTC	-	-	-1,821	-
DNBHC	-1,203	-	-2,615	-
WH001	2,145	-	-	-0,319
WH002	-1,075	-	-	-
WH003	-	-	-	-
WH004	4,439	2,730	2,369	-
WH005	-	-	-0,951	-

En la tabla 20, se muestran las mediciones de error de los 4 modelos construidos. Estas mediciones de error serán las mismas que en los modelos anteriores (MSE, Devianza, BIC y AIC)

Tabla 20: Medición de parámetros del ajuste.

Modelos GLM con selección de variables				
Localización	MSE	Devianza	BIC	AIC
CI	140.712	0.9176	1579.2	1480
HI	349.721	0.5301	1588.7	1518
II	109.867	0.8169	1499.8	1405
SI	7.222	0.409	842.4	772.8

Tras el cálculo de los parámetros de error se ha procedido a representar los valores predichos frente a los observados para los 4 modelos construidos. En las figuras 36 y 37 se pueden ver la representación para las 4 localizaciones de la rotura.

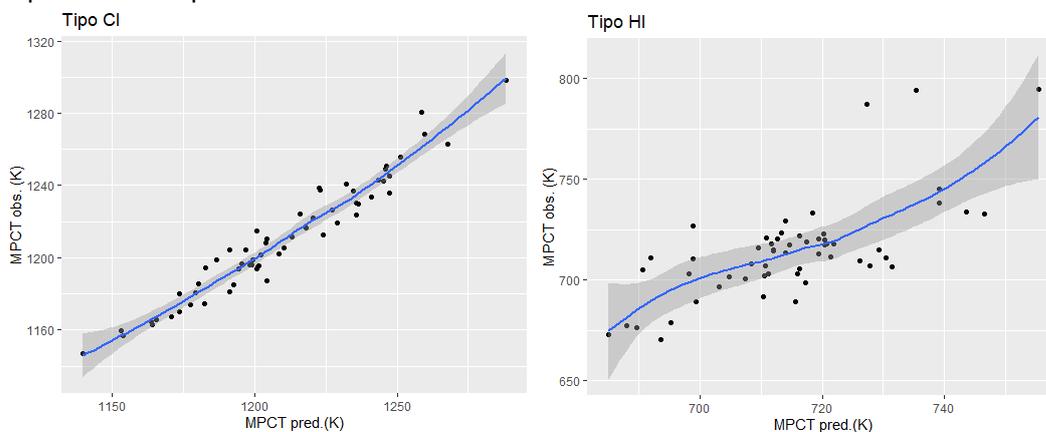


Figura 36: Obs. vs Pred. GLM (1).

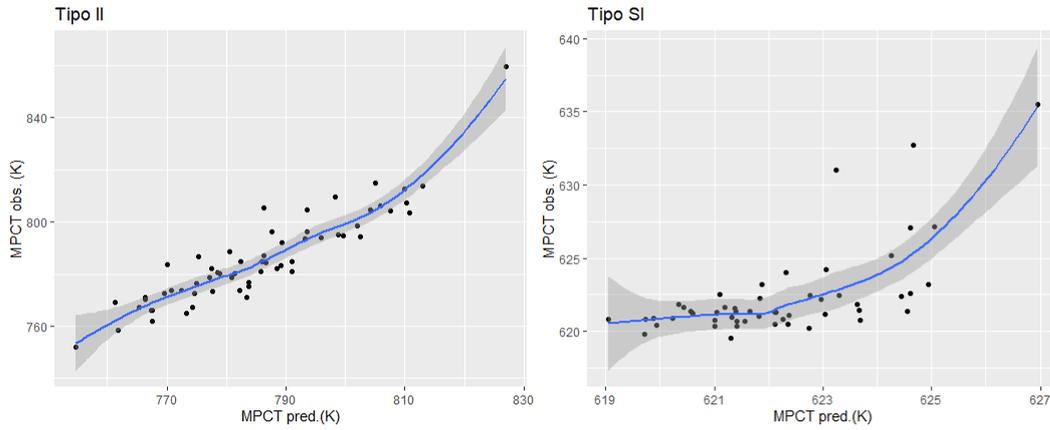


Figura 37: Obs. vs Pred. GLM (2).

### 6.3.2.1 Analisis residual.

En este apartado se va a comprobar la homocedasticidad de los residuos y su normalidad.

Para comprobar la normalidad, en la figura 38, se pueden ver los gráficos probabilísticos normales de la distribución de los residuos.

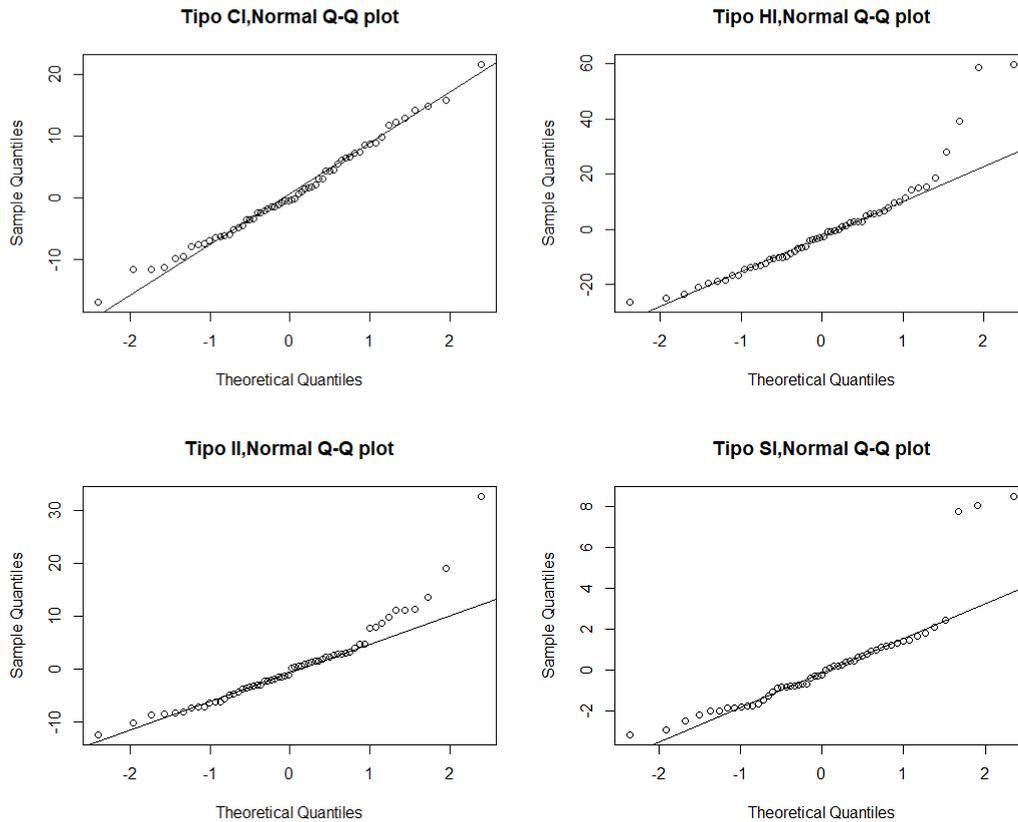


Figura 38: Grafico probabilístico normal de los residuos GLM.

Como se puede ver, en la figura superior, en los 4 modelos contruidos se mantiene la normalidad. Aunque hay valores extremos que rompen esa normalidad, son muy pocos en comparación con el volumen total de datos.

En la figura 39, que se muestra a continuación, están representados los valores predichos frente a los residuos.

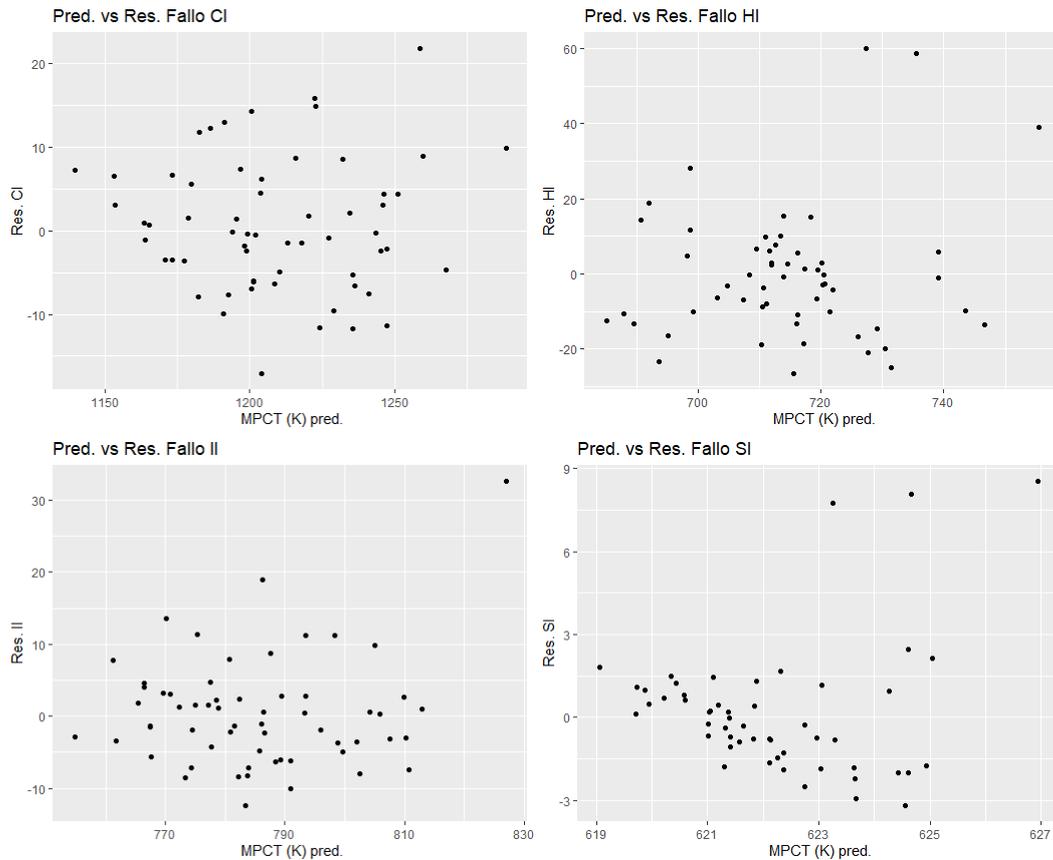


Figura 39: Residuos. vs Pred. de los modelos GLM.

Como se puede ver en la figura superior, en los casos CI, HI e II se podría decir que hay homocedasticidad de los residuos y que presentan un comportamiento de ruido blanco. Sobre el modelo SI decir que debido a la presencia de puntos influyentes a medida que aumenta el valor de la observación predicha el residuo se hace más negativo, lo que quiere decir que el modelo sobrestima la MPCT.

### 6.3.3 Random Forest.

Como se ha dicho anteriormente, la técnica random forest es una técnica no paramétrica que puede servir tanto para tareas de predicción como de clasificación.

En el caso de las 4 localizaciones se ha procedido de la misma manera. Se ha entrenado el modelo con el conjunto de entrenamiento y con un número muy alto de árboles,  $n = 5000$ , con el objetivo de saber cuándo se estabiliza el error.

En segundo lugar, se va a visualizar un gráfico que nos indicará el nivel de importancia de cada variable, tanto en el incremento del MSE con en la pureza de los nodos.

Por último, se usará el modelo para predecir lo valores del conjunto de datos test, se representará la solución gráfica y se calculará el MSE para la comparación de los modelos.

- Fallo CI:

Trans ajustar el modelo de random forest con  $n=5000$  a este tipo de fallo, nos encontramos que el error se estabiliza entorno a  $n=2000$ , tal como se puede ver en la figura 40.

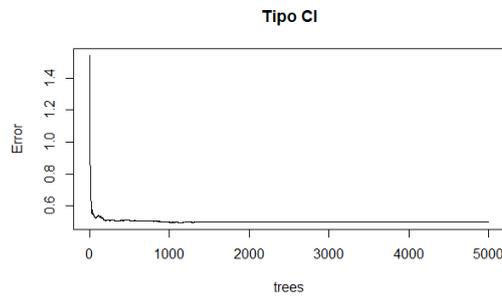


Figura 40: ntrees vs error, fallo CI

En la creación de nodos, las variables que más han influido se muestran en la figura 41, que se muestra a continuación.

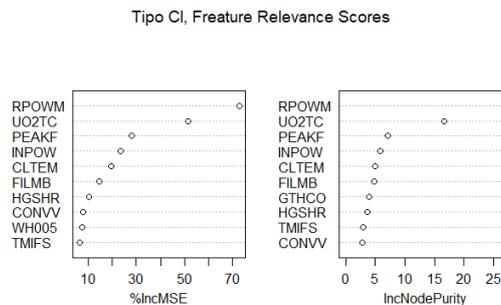


Figura 41: Var. Imp. CI

Como se puede ver la variable más importante es la **RPOWN**, seguido por **UO2TC** y **PEAKF**.

Los valores del MSE y de la devianza explicada se muestra en la tabla 21.

Tabla 21: MSE modelo RF para la loc. CI.

Fallo	MSE	Devianza
CI	444.52	0.49

En la figura 42 se muestra, el grafico de predichos frente a observados.

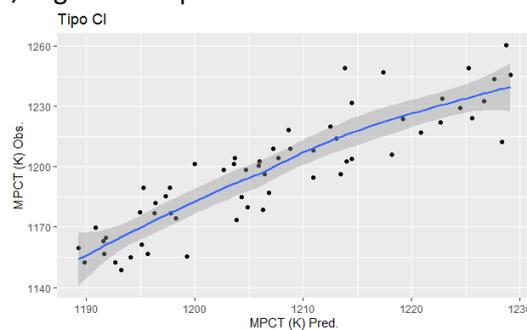


Figura 42: Pred vs Obs. CI RF.

- Fallo HI:

En primer lugar, se muestra, en la figura 43, como el error se estabiliza para  $n = 1500$ .

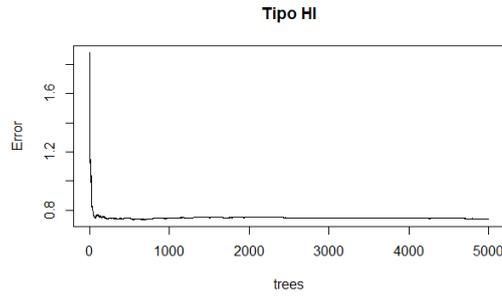


Figura 43 ntrees vs error, fallo HI

Las variables que más influyen en la pureza de los nodos y en la disminución del error se muestran en la figura 44.

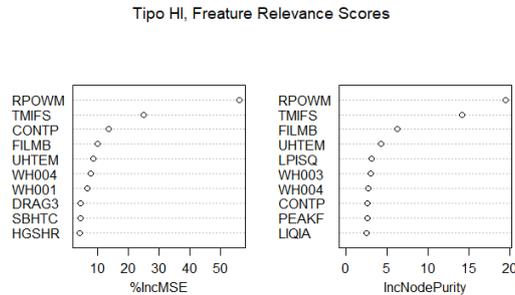


Figura 44 Var. Imp. HI

Como se puede ver en la figura 44, la variable más importante, en el caso de una rotura localizada en la HI, y con diferencia, es la **RPOWN** seguida de **TMIFS**.

La medición de errores y la gráfica de predichos frente a observados se muestran en la tabla 22 y la figura 45.

Tabla 22: MSE del modelo RF para la loc. HI.

Fallo	MSE	Devianza
HI	263.15	0.26

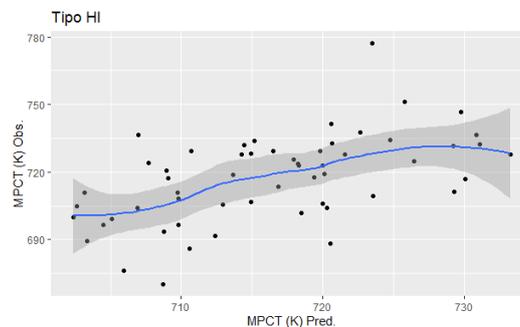


Figura 45: Pred vs Obs, fallo HI

- Fallo II.

En primer lugar, se muestra, en la figura 46, como el error se estabiliza para n = 1000.

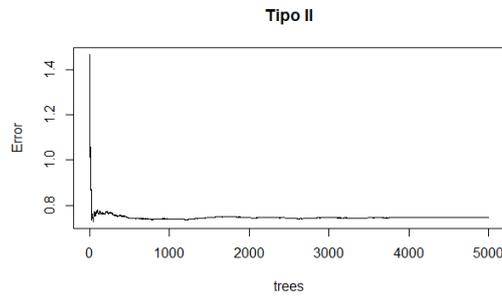


Figura 46 ntree vs error, fallo II

Las variables que más influyen en la pureza de los nodos y en la disminución del error se muestran en la figura 47.

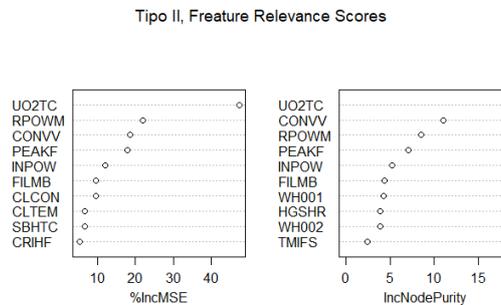


Figura 47 Var. Imp. II

Como se puede ver en la figura 11, la variable más importante, en este caso, y con diferencia, es la **UO2TC, RPOWM y CONV**.

La medición de errores y la gráfica de predichos frente a observado se muestran en la tabla 23 y la figura 48, respectivamente.

Tabla 23: MSE modelo RF para la loc. II.

Fallo	MSE	Devianza
II	161.82	0.27

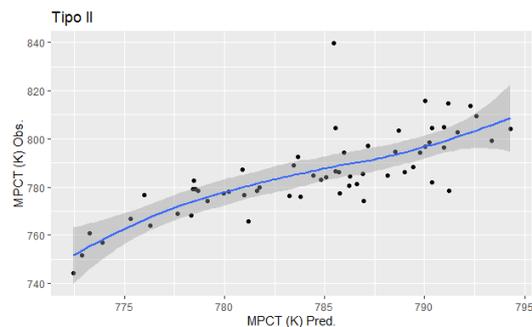


Figura 48: Pred vs Obs, fallo II

- Fallo SI.

En primer lugar, se muestra, en la figura 49, como el error se estabiliza para n = 1000.

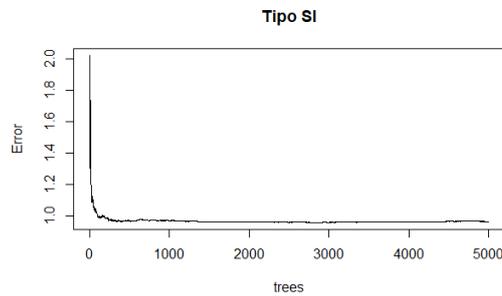


Figura 49: ntrees vs error, fallo SI

Las variables que más influyen en la pureza de los nodos y en la disminución del error se muestran en la figura 50.

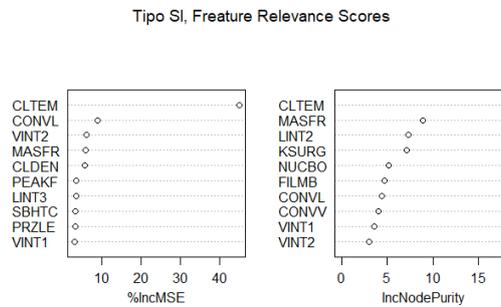


Figura 50: Var. Imp. SI

Como se puede ver en la figura 50, la variable más importante, y con diferencia, es la **CLTEM** seguido de **MASFR**.

La medición de errores y la gráfica de predichos frente a observado se muestran en la tabla 24 y la figura 51.

Tabla 24: MSE del modelo RF para la loc. SI.

Fallo	MSE	Devianza
SI	6.014	0.09

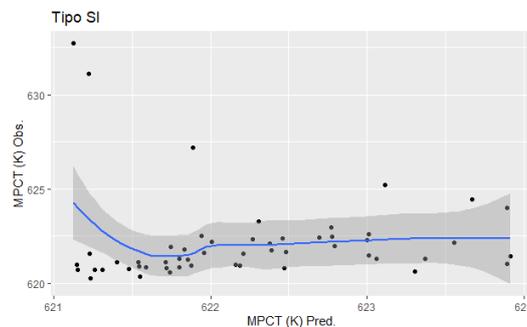


Figura 51: Pred vs Obs, fallo SI.

### 6.3.3.1 Análisis Residual.

En primer lugar, en la figura 52, se puede ver el gráfico probabilístico normal de los residuos de las observaciones predichas.

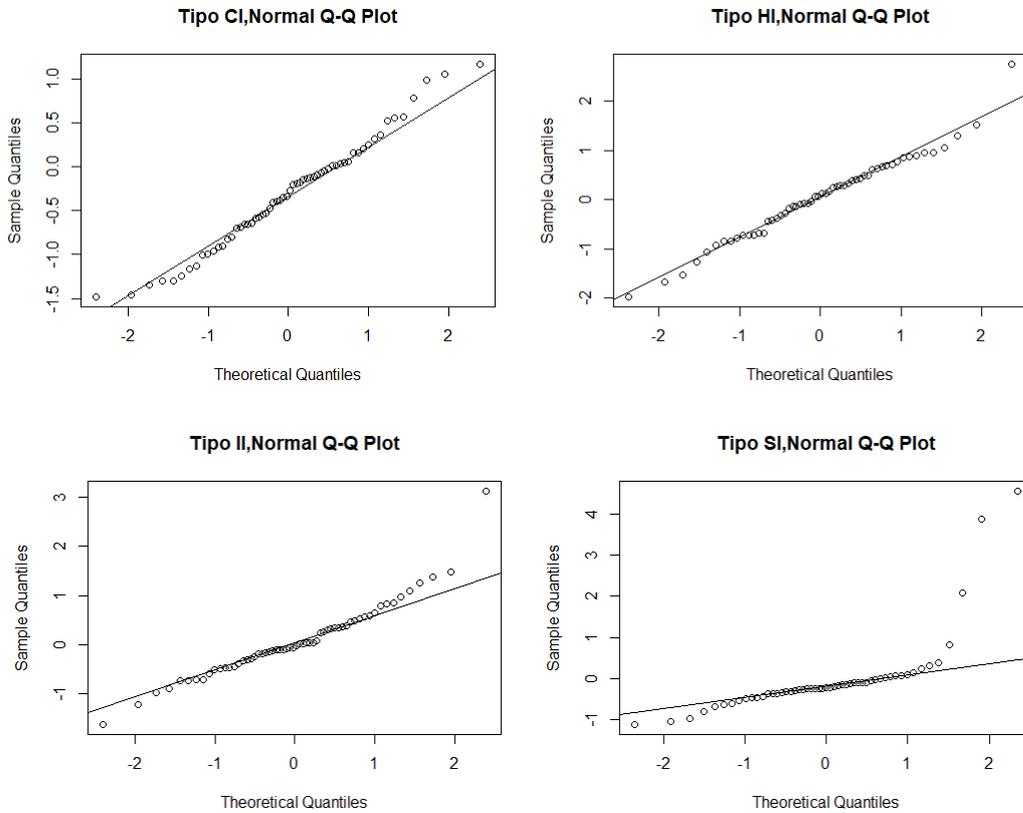


Figura 52: Gráfico probabilístico normal de los residuos RF.

Como se puede ver en la figura superior, en general se cumple la normalidad de los residuos excepto para los valores más altos, que se pierde la normalidad.

A continuación, en la figura 53 y 54, se puede ver el gráfico de predichos frente a los residuos.

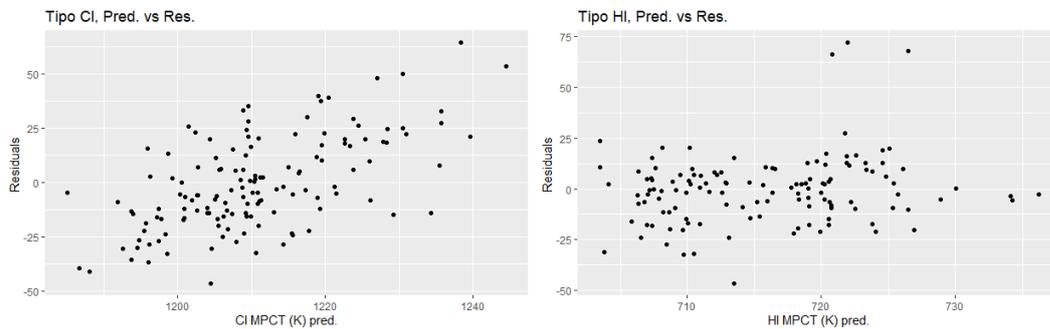


Figura 53: Obs. vs Pred. de los modelos RF (1).

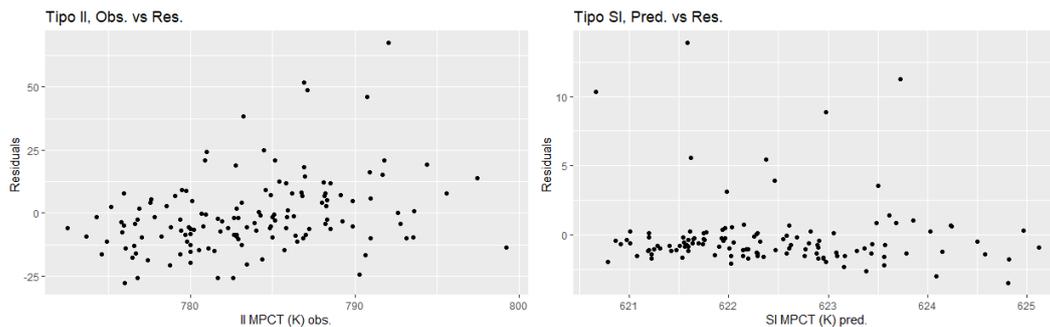


Figura 54: Obs. vs Pred. de los modelos RF (2).

Como se ve en las imágenes superiores se puede admitir homocedasticidad de los residuos salvo en el caso de la localización CI, donde se ve claramente que a medida que aumenta el valor de la observación predicha, aumenta los residuos de dicha predicción.

### 6.3.4 Boosting.

Como se ha dicho previamente, el *boosting* es un meta-algoritmo que reduce el sesgo y la varianza en el aprendizaje supervisado. Debido a que el *boosting* funciona como la combinación de dos o más predictores débiles, para crear uno fuerte. Los parámetros que se pueden modelar para obtener una mejor predicción se han optimizado a partir de validación cruzada, obteniéndose los resultados que se muestran en la tabla 25.

Tabla 25: Meta-parámetros de los modelos boosting.

Fallo	Shrinkage	Interaction. Depth	n.minobsinnode	n.trees
CI	0.01	5	15	3000
HI	0.001	9	15	4000
II	0.01	5	10	1000
SI	0.001	9	10	4000

El criterio de selección de estos parámetros es aquel que ha minimizado el MSE y, además, maximiza el  $R^2$ .

Una vez optimizado los parámetros para cada tipo de fallo, se ha ajustado el modelo y los resultados son los siguientes.

- Fallo CI.

Al igual que en el apartado anterior, podemos extraer que variables han sido más importante en la regresión de la temperatura MPCT. En la figura 55, se muestran las 10 variables más importantes a la hora de predecir el valor de la MPCT.

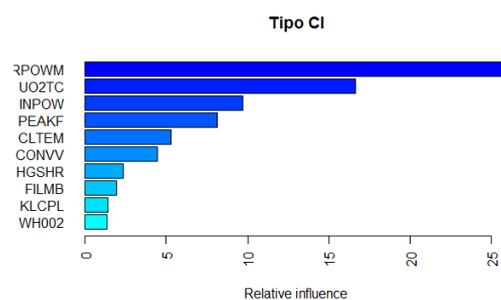


Figura 55: CI Var. Imp. Boosting.

Como se puede ver en la figura superior, la variable más importante es **RPOWM**, seguida de **UO2TC**, **INPOW** y **PEAKF**.

Una vez ajustado el modelo se ha calculado el MSE y la devianza para poder evaluar el modelo mostrándose los valores obtenidos en la tabla 26. El gráfico de valores predichos frente a observados se muestra en la figura 56.

Tabla 26: MSE modelo boosting CI.

Tipo de Fallo	MSE	Devianza
CI	223.02	0.65

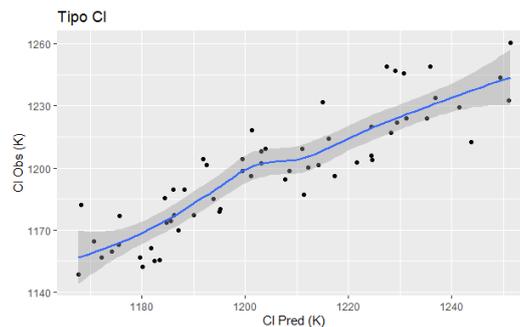


Figura 56: Obs. vs Pred. CI boost

- Fallo HI.

En la figura 57 se puede ver la importancia relativa de las variables.

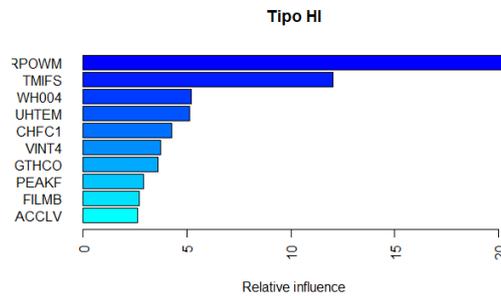


Figura 57: HI Var. Imp. Boosting.

En este caso, la variable más importante es la **RPOWN** seguido de **TMIFS**.

En la tabla 27 se muestran los valores de la medición de errores y, en la figura 58 el gráfico de predichos frente a observados.

Tabla 27: MSE modelo Boosting HI.

Tipo de Fallo	MSE	Devianza
HI	433.09	0.39

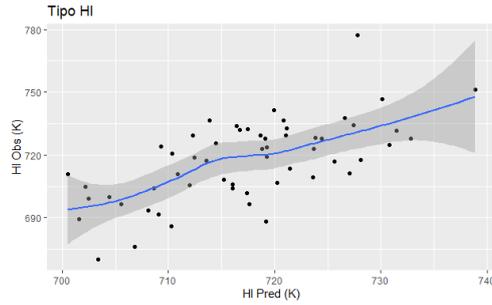


Figura 58: Obs vs Pred. HI.

- Fallo II.

En el fallo II las variables más importantes, representadas en la figura 59, son:

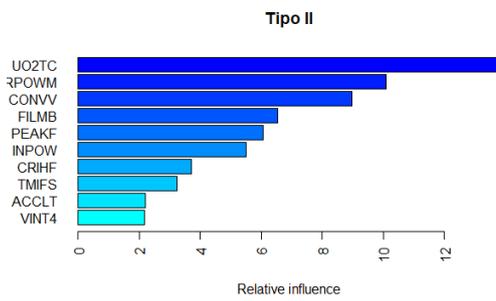


Figura 59: Il Var. Imp. Boosting.

En este caso hay 5 variables cuya importancia relativa es significativamente mayor a las demás: **UO2TC, RPOWN, CONVV y FILMB.**

A continuación, en la tabla 28 y en la figura 60 se muestran, respectivamente, los errores de medición y los valores predichos frente a los observados.

Tabla 28: MSE modelo boosting II.

Tipo de Fallo	MSE	Devianza
II	215.05	0.14

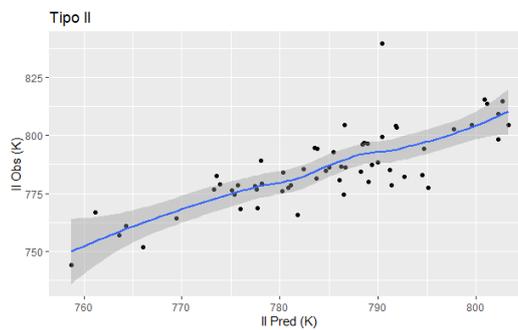


Figura 60: Obs vs Pred. II.

- Fallo SI.

Las variables más importantes en la regresión de la MPCT en la localización SI se muestran en la figura 61.

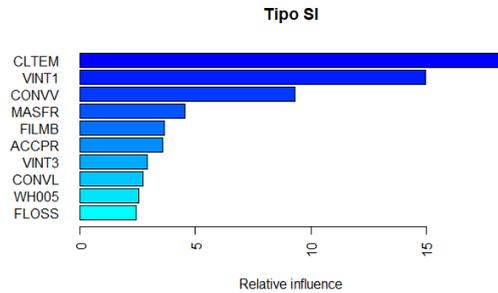


Figura 61: SI Var. Imp. Boosting.

En este caso las variables más importantes son: **CLTEM**, **VINT1** y, en menor medida **CONVV**.

El MSE y el grafico de predichos frente a observados se muestra en la tabla 29 y figura 62.

Tabla 29: MSE modelo boosting SI.

Tipo de Fallo	MSE	Devianza
SI	5.74	0.25

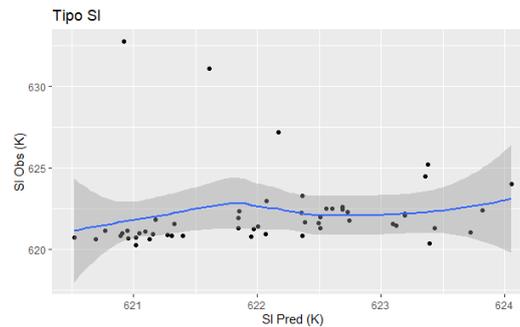


Figura 62: Obs vs Pred. SI.

#### 6.3.4.1 Análisis residual.

En primer lugar, en la figura 63, podemos ver el grafico probabilístico normal de los residuos, a continuación, en la figura 64, el gráfico de predichos frente a observados.

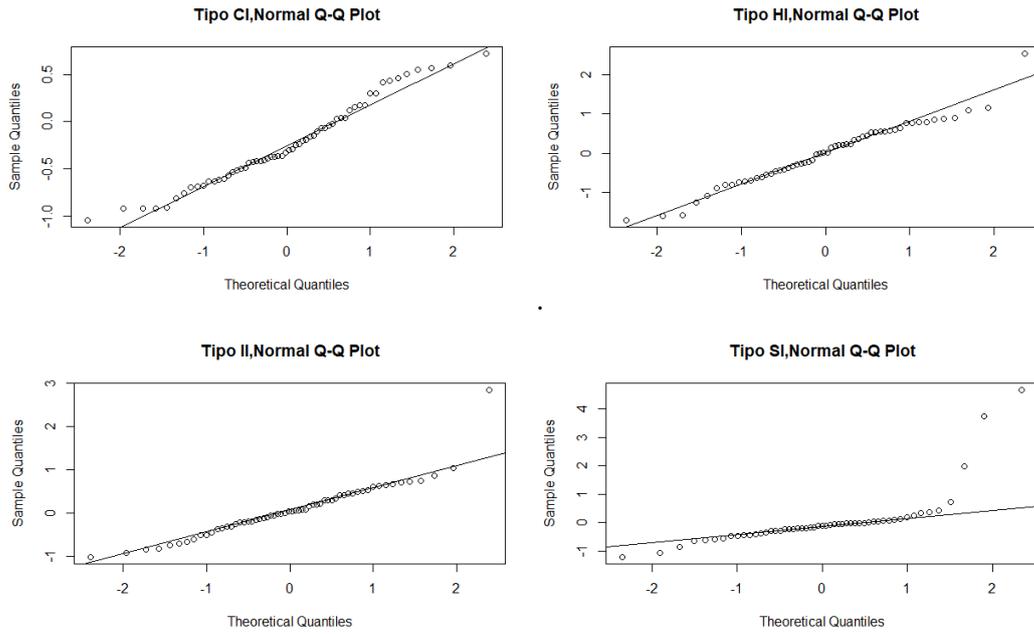


Figura 63: Grafico probabilístico normal de los residuos boosting.

Como se puede ver en la figura 63, los residuos de los 4 modelos cumplen con la hipótesis de normalidad. Aunque en el modelo para la localización SI, existen residuos cuyos valores rompen la esa hipótesis.

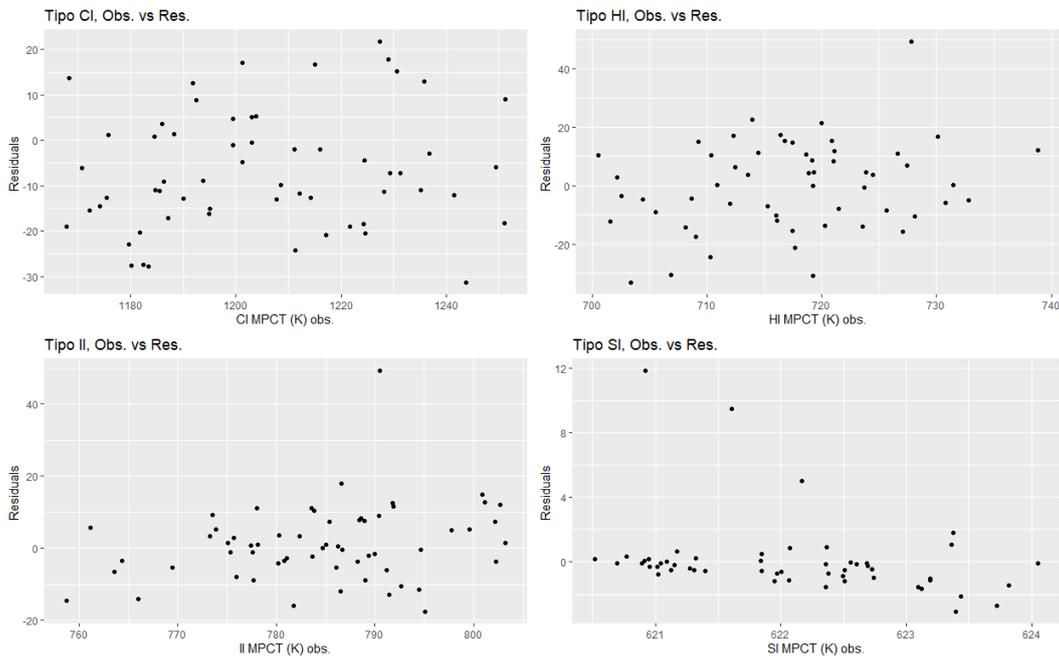


Figura 64: Obs. vs Pred. de los modelos boosting.

Sobre la distribución de los residuos decir que, según se puede ver en la figura 64, la nube de puntos de los 4 modelos parece ser homocedástica aunque, en el modelo SI y debido a la presencia de datos muy extremos, no se aprecia con claridad.

## 6.4 Comparación de los modelos de regresión.

En este apartado se van a comparar los modelos obtenidos a partir de las distintas técnicas de regresión lineal, con penalización y sin ella, y las técnicas de regresión basadas en árboles de decisión. Además, se seleccionará el mejor modelo para la predicción de la MPCT.

### 6.4.1 Comparación de los modelos de regresión.

En las tablas 30, 31, 32 y 33, que se muestran a continuación, se comparan los modelos de regresión lineal y los modelos basados en árboles de decisión de las 4 localizaciones de fallo posibles en función de los valores obtenidos de MSE, Devianza, BIC y AIC.

Tabla 30: Comparación modelo de la localización Cl.

Localización Cl				
	MSE	Devianza	BIC	AIC
Regresión Ridge	168.2	0.88	-101160.20	-101243.7
Regresión Lasso	149.66	0.86	-99369.61	-12955.43
Regresión Elastic Net	114.91	0.90	-104399.40	-104482.70
Modelo Lineal Generalizado	140.71	0.92	1579.20	1480.00
Random Forest	444.52	0.49	-	-
Boosting	223.02	0.65	-	-

Tabla 31: Comparación modelo de la localización Hl.

Localización Hl				
	MSE	Devianza	BIC	AIC
Regresión Ridge	518.36	0.33	-11704.6	-11768.74
Regresión Lasso	467.98	0.36	-12936.4	-12955.43
Regresión Elastic Net	473.01	0.42	-15145.2	-15181.4
Modelo Lineal Generalizado	349.72	0.53	1588.7	1518.2
Random Forest	263.15	0.26	-	-
Boosting	433.1	0.39	-	-

Tabla 32: Comparación modelo de la localización Il.

Localización Il				
	MSE	Devianza	BIC	AIC
Regresión Ridge	141.67	0.76	-30112.43	-30032.10
Regresión Lasso	120.189	0.72	-28483.12	-28536.41
Regresión Elasticnet	107.134	0.79	-31355.58	-31420.60
Modelo Lineal Generalizado	109.87	0.82	1499.80	1405.30
Random Forest	161.82	0.27	-	-
Boosting	215.05	0.14	-	-

Tabla 33: Comparación modelo de la localización SI.

Localización SI				
	MSE	Devianza	BIC	AIC
<b>Regresión Ridge</b>	9.38	0.004	280.85	229.60
<b>Regresión Lasso</b>	9.38	0	0	0
<b>Regresión Elastic Net</b>	9.14	0.07	-41.61	-44.38
<b>Modelo Lineal Generalizado</b>	7.22	0.41	842.40	772.8
<b>Random Forest</b>	6.01	0.09	-	-
<b>Boosting</b>	5.74	0.25	-	-

#### 6.4.2 Selección del modelo de regresión.

Tras el análisis de los distintos parámetros de ajuste y de las distintas figuras obtenidas, se seleccionó como el mejor modelo de regresión para la predicción de la MPCT el obtenido a partir del algoritmo de boosting.

Esto se debe a que dichos modelos no son los que tienen los valores más bajos y altos del MSE y de la devianza, respectivamente, pero son los que han obtenido unas mejores predicciones. Además, la distribución de los residuos de las MPCT predichas son los que se acercan más a la hipótesis de normalidad y de homocedasticidad.

Otra ventaja derivada de selección del modelo de boosting es que se puede usar como un método de reducción de variables, por lo que se podría llevar un control adicional de las variables que han sido seleccionadas como más importantes por el algoritmo, con el objetivo de identificar la localización del LBLOCA y la predicción de su MPCT.

#### 6.5 Clasificación de la localización de la rotura en función de la MPCT.

En el apartado anterior del presente trabajo se han aplicado distintas técnicas de regresión para predecir el valor de la temperatura MPCT en función de las variables termohidráulicas registradas. En este apartado se construye un clasificador, a partir del algoritmo de árbol de clasificación, para la predicción de la rama de refrigeración que ha sufrido un LBLOCA en función de la temperatura MPCT.

Las MPCT utilizadas son las que han sido predichas con los 4 modelos de boosting seleccionados en el apartado anterior. En la figura 65, se puede ver la secuencia de los diferentes pasos que se han realizado.

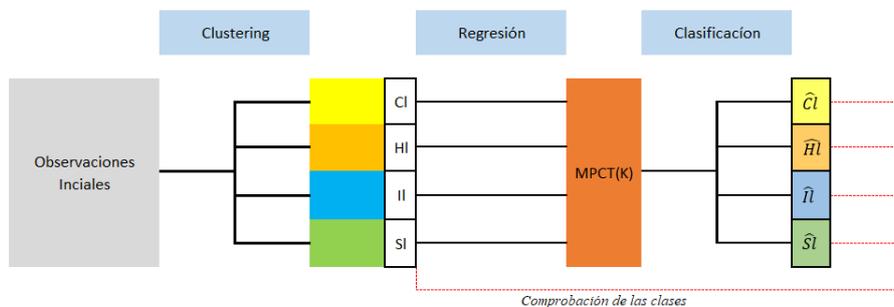


Figura 65: Secuencia de trabajos realizados.

En principio, debería ser una tarea sencilla y con un costo de computación bajo ya que sólo tenemos una variable para la construcción del clasificador, la temperatura MPCT, y 4 ramas en las que se ha podido producir el LBLOCA, CI, HI, II y SI.

A continuación, se mostrarán los resultados de la clasificación, obtenido a partir del algoritmo del árbol de clasificación.

En primero lugar, en la figura 66, podemos ver el árbol de clasificación resultante.

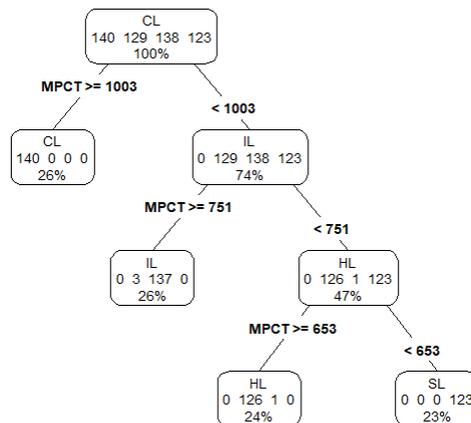


Figura 66: Árbol de clasificación resultante.

De la figura 66 podemos extraer que la temperatura MPCT, en el caso de que la rotura se de en la localización CI se caracteriza por ser superiores a los 1003 K. Por otro lado, la MPCT en la localización II y HI se caracterizan por encontrarse en el intervalo [751, 1003] K y [653, 751] K, respectivamente. Y, por último, la MPCT de la localización SI se caracteriza por ser inferior a los 653 K. Una observación pertinente sobre el árbol construido es que la mayoría de los nodos son puros o contienen muy pocos fallos de clasificación.

Tras la construcción del árbol de clasificación, se ha predicho la clase de la MPCT estimada en el apartado anterior y se ha calculado la tasa de acierto comparándola con la clase original. En la tabla 34, que se muestra a continuación, se presenta la matriz de confusión y la tasa de acierto.

Tabla 34: Matriz confusión de la clasificación de la MPCT.

	Localización predicha			
Loc. Observada	CI	HI	II	SI
CI	199	0	0	0
HI	0	181	1	0
II	0	1	196	0
SI	0	0	0	175

A partir de la tabla 34 se ha calculado la tasa de acierto, la cual es un 99.73% de acierto. Por lo que, pese a tener una regresión que, en los mejores casos explica el 65% de la variabilidad de los datos, la clasificación sí que lo es.

## 7. Conclusiones.

El presente trabajo se ha centrado en la aplicación de diferentes métodos de regresión y clasificación para el análisis de uno de los accidentes base de diseño (DBA) más importante en Centrales Nucleares, en concreto, el accidente con pérdida de refrigerante grande (LBLOCA).

El objetivo de aplicar técnicas de regresión ha sido analizar la posibilidad de sustituir el código de simulación termohidráulico, de elevado coste computacional, por un modelo de regresión que permitiese, a partir de las condiciones iniciales de la central, predecir el valor de variables de seguridad importantes como, por ejemplo, la máxima temperatura de la vaina (MPCT).

En relación a la clasificación, el principal objetivo es poder predecir de una forma rápida y antes de alcanzar la MPCT la localización de la rotura pudiendo utilizarse, en un futuro, como ayuda para el operador de la planta.

Respecto a los resultados específicos obtenidos mencionar los siguientes:

- Del clasificador de la evolución del transitorio de la PCT según la localización de LBLOCA se puede extraer que, debido a las diferencias entre las curvas de la temperatura PCT, con la medición de 20 instantes temporales se es capaz de predecir, a partir del algoritmo SVM, la rama donde se ha producido la rotura con una precisión del 98%.
- El método con el que se han obtenido mejores predicciones de la MPCT y que más robustez ha presentado es el construido a partir del algoritmo de boosting.
- Con los resultados obtenidos de la regresión de la MPCT y con ayuda de un árbol de clasificación simple se ha conseguido identificar en que rama se ha producido la rotura con precisión del 99.73%.

## Bibliografía

- Alexandros Karatzoglou, A. S. (2004). kernlab. *Journal of Statistical Software*. Obtenido de <http://www.jstatsoft.org/v11/i09/>
- Alvear, J. O. (2018). *Bokdown*. Obtenido de <https://bookdown.org/content/2031/>
- Brandon Greenwell, B. B. (2019). gbm: Generalized Boosted Regression Models. Obtenido de <https://CRAN.R-project.org/package=gbm>
- Carrasco, M. C. (2016). *Técnicas de regularización en regresión: Implementación y aplicaciones*. Universidad de Sevilla, Departamento de estadística e investigación operativa. Sevilla: Facultad de matemáticas. Obtenido de <https://idus.us.es/xmlui/bitstream/handle/11441/43746/Carrasco%20Carrasco%2C%20Mar%20C3%ADa%20TFG.pdf?sequence=1&isAllowed=y>
- CC. (2015). *Wikipedia*. Obtenido de [https://en.wikipedia.org/wiki/K-means\\_clustering#cite\\_note-hamerly4-9](https://en.wikipedia.org/wiki/K-means_clustering#cite_note-hamerly4-9)
- David Meyer, E. D. (2019). e1071: Misc Functions of the Department of Statistics. Obtenido de <https://CRAN.R-project.org/package=e1071>
- Diazaraque, J. M. (s.f.). *Tema 3: Modelos lineales generalizados*. Universidad Carlos III.
- Garbade, M. J. (12 de 9 de 2018). *towardsdatascience*. Obtenido de <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
- Hastie, H. Z. (2005). *Regularization and variable selection via the elasticnet*. Series B.
- Hastie, H. Z. (2018). elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. Obtenido de <https://CRAN.R-project.org/package=elasticnet>
- Kennard, A. H. (1970). *Ridge regression: biased estimation for non-orthogonal problems . . .* Technometrics.
- Kuhn, M. (2019). caret: Classification and Regression Training. Obtenido de <https://CRAN.R-project.org/package=caret>
- Macqueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. University of California Press.
- Maechler, M. R. (2018). *cluster: Cluster Analysis Basics and Extensions*.
- Martínez, C. G. (2018). *RPUBS*. Obtenido de [https://rpubs.com/Cristina\\_Gil/arboles\\_ensemble](https://rpubs.com/Cristina_Gil/arboles_ensemble)
- Noah Simon, J. F. (2011). *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent*. *Journal of Statistical Software*. Obtenido de <http://www.jstatsoft.org/v39/i05/>
- Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer. Obtenido de <http://lmdvr.r-forge.r-project.org>
- Tibshirani, R. (1996). *Regression shrinkage and selection via the Lasso . Journal of the Royal Statistical Society*. Series B.

- Torgo, L. (2019). *Predictive Analytics*. Institute for Big Data Analytics Dalhousie University.
- Turner, H. (2008). *Introduction to Generalized Linear Models*. ESRC National Centre for Research Methods, UK and Department of Statistics University of Warwick, UK.
- Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Obtenido de <http://www.jstatsoft.org/v40/i01/>
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Obtenido de <https://ggplot2.tidyverse.org>
- Wiener, A. L. (2002). Classification and Regression by randomForest. *R News*. Obtenido de <https://CRAN.R-project.org/doc/Rnews/>