
# Aplicación de Técnicas de Aprendizaje Automático en el Sector Ferroviario

## Memoria

Edición: 1
4 de julio de 2016

**Jesús Vázquez Galán**

UNIVERSIDAD PONTIFICIA ICAI ICADE COMILLAS MADRID

MASTER UNIVERSITARIO EN SISTEMAS FERROVIARIOS

European Accreditation of Engineering Programmes EUR-ACE®

# FICHA TÉCNICA
# TRABAJO FIN DE MASTER

**Datos del Proyecto**

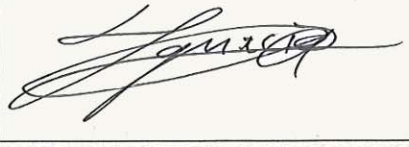
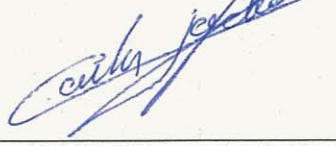
| | |
|---|---|
| Autor: | Jesús Vázquez Galán |
| Dirección de proyecto: | Ignacio Martínez González, Carlos Sancho de Mingo |
| Programa cursado: | Master Universitario en Sistemas Ferroviarios |
| Curso académico: | 2015-2016 |
| Título de proyecto: | Aplicación de Técnicas de Aprendizaje Automático en el Sector Ferroviario |

**Resumen del proyecto**


Los algoritmos de aprendizaje automático permiten buscar patrones en los datos, identificando las relaciones existentes entre todos los campos con el objetivo de obtener una función que sea capaz de realizar la predicción de variables. Esta predicción nos puede permitir, entre otros, identificar tipologías de fallos, determinar la vida útil de un elemento en función de su estado, optimizar recursos, etc.

Este tipo de técnicas se están utilizando en gran cantidad de ámbitos, en los cuales se están obteniendo grandes ventajas simplemente con el estudio de los datos de histórico. Dado que no se ha identificado un uso maduro de este tipo de técnicas en el sector ferroviario surge la motivación y oportunidad del presente proyecto con el objetivo final de realizar una descripción teórica acompañada de unos casos prácticos que sirven como demostración de su uso.

Con respecto a la parte teórica, el proyecto tiene una gran carga de recopilación y estudio de información referente a las técnicas de aprendizaje automático. Por otro lado, para la parte práctica se ha elegido el ámbito del mantenimiento predictivo, realizando dos pruebas de concepto con casos de uso diferentes. Por un lado, se ha llevado a cabo una prueba de algoritmos de clasificación y por otro se ha realizado una prueba con algoritmos de regresión.


| **Autor:** | **Dirección de Proyecto** | |
|---|---|---|
| Jesús Vázquez Galán | Ignacio Martínez González<br>*Ineco* | Dr. Carlos Sancho de Mingo<br>*Metro de Madrid* |
|  |  |  |

| Datos del documento | |
|---|---|
| **Nombre** | TFM_Memoria_JVG_Ed_Final |
| **Tipo** | Memoria |
| **Fecha** | 04/07/2016 |

**Control de versiones**

| Versión | Descripción de los cambios | Modificado por | Fecha | Revisado por | Fecha Revisión | Aprobado por | Fecha Aprobación |
|---|---|---|---|---|---|---|---|
| 1.0 | Documento inicial | Jesús Vázquez Galán | 01/07/2016 | Ignacio Martínez González<br><br>Carlos Sancho de Mingo | 04/07/2016 | Ignacio Martínez González<br><br>Carlos Sancho de Mingo | 04/07/2016 |

# ÍNDICE



# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# 1 INTRODUCCIÓN

La digitalización de los diferentes procesos ejecutados en el ámbito ferroviario está generando gran cantidad de información que, convenientemente tratada, podría desvelar comportamientos que actualmente están ocultos y que pueden aportar nuevos enfoques de actuación en diferentes ámbitos del sector.

El Aprendizaje Automático (Machine Learning) es la rama de la Inteligencia Artificial que se dedica al estudio de los modelos y algoritmos que aprenden o evolucionan basados en su experiencia o en un conjunto de datos de histórico, para realizar una tarea determinada cada vez mejor. El objetivo principal de todo proceso de aprendizaje es utilizar las evidencias conocidas para poder crear una hipótesis y poder dar una respuesta a nuevas situaciones no conocidas.

Por otro lado, cada vez aparecen nuevos paradigmas que demandan soluciones prácticas en numerosos aspectos del mundo real (por ejemplo, Big Data). Entre ellos el Aprendizaje Automático, en sus diferentes vertientes, ofrece métodos y estrategias para dar respuesta a estas demandas.

Hoy en día existen aplicaciones que utilizan agentes basados en aprendizaje automático en numerosas ramas de la industria y de la ciencia:

- *Procesamiento del Lenguaje Natural:* Para el análisis sintáctico y morfológico de los textos. Extracción de Información, Clasificación Automática de Documentos, Agrupamiento Semántico. Los modelos de Análisis de Sentimiento, por otro lado, permiten reconocer la orientación o polaridad subjetiva de un texto, es decir, si están hablando bien o mal de aquello que se está opinando e incluso ponderar el grado de aceptación o negación.
- *Sistemas de Recuperación de Información:* Los buscadores de Internet utilizan técnicas de aprendizaje automático para mejorar el rendimiento y precisión de sus búsquedas y confeccionar rankings personalizados según la experiencia de los usuarios.
- *Diagnóstico Médico:* Se utilizan este tipo de técnicas para asistir a médicos en el diagnóstico según la historia clínica y los síntomas que presenta el paciente.
- *Ciencias biológicas:* Para la clasificación de especies, reconocimiento de tumores, arritmias o patrones en cadenas de ADN.
- *Movilidad urbana:* Descubrimiento y predicción de patrones de movilidad.
- *Finanzas e Industria bancaria:* La industria bancaria utiliza modelos de riesgo crediticio para calificar a los solicitantes según su nivel de cumplimiento esperado en el pago de las cuotas del

crédito. También existen modelos de fraude de consumo de tarjetas de crédito y de predicción de comportamiento en el mercado de valores.

- *Análisis de imágenes:* El análisis de imágenes ha sido uno de los campos de aplicación más utilizados y que más ha avanzado, donde es posible reconocer escritura manuscrita, identificar direcciones y remitentes de un envío postal, facturas o recibos u otro tipo de documentos legales. También la identificación de objetos dentro de una imagen, como personas, rostros, monumentos arquitectónicos o accidentes geográficos, etc.

En el ámbito ferroviario[1], aunque se emplean este tipo de técnicas en casos puntuales, aun no se está realizando un uso maduro de esta tecnología. Por este motivo, surge la motivación y oportunidad de realización del presente proyecto.

## 1.1 RESUMEN

Los algoritmos de aprendizaje automático permiten buscar patrones en los datos, identificando las relaciones existentes entre todos los campos con el objetivo de obtener una función que sea capaz de realizar la predicción de variables. Esta predicción nos puede permitir, entre otros, identificar tipologías de fallos, determinar la vida útil de un elemento en función de su estado, optimizar recursos, etc.

Este tipo de técnicas se están utilizando en gran cantidad de ámbitos, en los cuales se están obteniendo grandes ventajas simplemente con el estudio de los datos de histórico. Dado que no se ha identificado un uso maduro de este tipo de técnicas en el sector ferroviario surge la motivación y oportunidad del presente proyecto con el objetivo final de realizar una descripción teórica acompañada de unos casos prácticos que sirven como demostración de su uso.

## 1.2 OBJETIVOS DEL TRABAJO

Los objetivos que se persiguen con el presente proyectos se pueden agrupar en una componente teórica y una componente práctica que se explican a continuación:

### *1.2.1 Objetivos teóricos*

El proyecto tiene un gran peso de recopilación y estudio de información referente a las técnicas de aprendizaje automático. Este trabajo tiene como misión cumplir con los siguientes objetivos:

1 En el apartado Aplicación en el Sector Ferroviario se tratarán ejemplos de posibles aplicaciones de estas técnicas dentro del sector ferroviario.

- Descripción general de las técnicas de aprendizaje automático.
- Clasificación y descripción de los principales algoritmos.
- Identificación de ámbitos de aplicación dentro del sector ferroviario.

### *1.2.2 Componente práctico*

El proyecto contempla la realización de casos de prácticos. El ámbito seleccionado para los casos prácticos está relacionado con el mantenimiento predictivo. Se han realizado dos pruebas de concepto con casos de usos diferentes. Por un lado, se ha llevado a cabo una prueba de algoritmos de clasificación y por otro se ha realizado una prueba con algoritmos de regresión. Para la realización de ambas pruebas han sido necesarios los siguientes pasos:

- Recopilar un conjunto de datos históricos de valor.
- Preparar convenientemente los datos de entrenamiento.
- Crear un modelo con los datos (entrenamiento del algoritmo).
- Ejecución de los modelos de predicción.

Los objetivos de estos ejemplos son poder contrastar tanto la viabilidad de este tipo de técnicas como identificar los problemas que puedan surgir en su aplicación.

## 1.3 TAREAS

**Estudio y descripción general de las técnicas de aprendizaje automático.**

Como primera actividad del proyecto se ha realizado una descripción general de las técnicas de aprendizaje automático.

**Identificación de los ámbitos de aplicación en el sector ferroviario.**

Se ha realizado un estudio de los principales ámbitos de aplicación donde las técnicas de aprendizaje automático pueden aportar un mayor valor añadido en el sector ferroviario.

**Realización de una prueba de concepto.**

*Identificación del ámbito de aplicación para la prueba de concepto*. Las capacidades de aprendizaje automático se pueden aplicar a gran cantidad de ámbitos del sector ferroviario, pero para la realización de las pruebas de concepto se determinará el ámbito de aplicación que permita evaluar el resultado de la aplicación de estas técnicas.

*Identificación de los parámetros a analizar*. Identificación de los parámetros que afectan al ámbito de aplicación seleccionado (estudio de auscultador dinámico y desgaste de ruedas) de manera que se pueda entrenar el modelo para posteriormente poder realizar los experimentos del predicción.

*Estudio y creación de los modelos de aprendizaje*. Identificación de los modelos de aprendizaje que mejor se adapten a las pruebas de concepto.

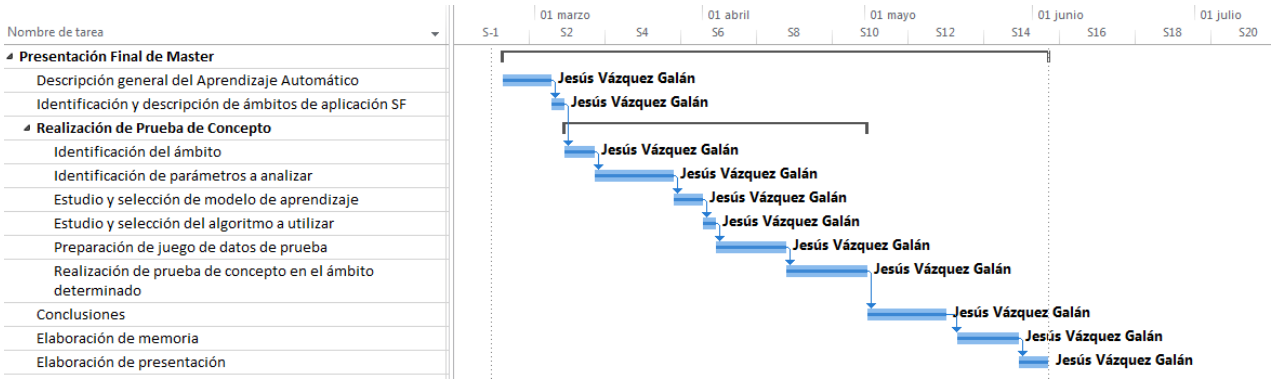
*Estudio y selección de algoritmos*. Realización de un estudio de los algoritmos disponibles que mejor se adapten a las necesidades del ámbito de aplicación identificado.

*Preparación de juego de datos de prueba*. Los modelos de aprendizaje automático requieren de una fase de aprendizaje basado en datos de histórico. En esta actividad se localizará la información necesaria para poder entrenar el modelo de la prueba de concepto.


*Figura 1-1: Esquema de realización de prueba de concepto*

## 1.4 PLANIFICACIÓN

La planificación establecida para el proyecto es la siguiente:



# 2 DESARROLLO TEÓRICO

## 2.1 APRENDIZAJE AUTOMÁTICO

Bajo el concepto de aprendizaje automático podemos englobar un conjunto de técnicas y algoritmos que nos pueden permitir:

- Convertir datos en bruto en información.
- Con base en un número limitado de observaciones, estimar una dependencia o estructura desconocida de un sistema.
- Realizar predicciones de valores o clasificaciones.

El aprendizaje es solo una parte del procedimiento seguido en las ciencias, la ingeniería, las ciencias sociales, la medicina, y en general en todos los campos que utilizan la estadística para extraer conclusiones.

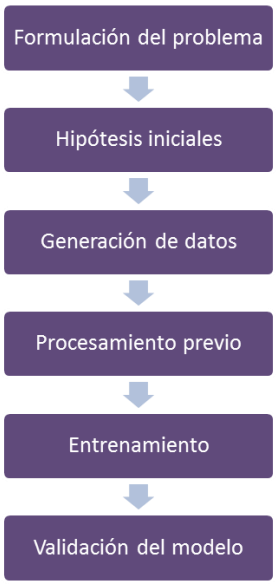



*Figura 2-1: Procedimiento para aprender un modelo de relación entre datos*

El proceso de aprendizaje completo se realimenta de las diferentes fases, ya que por ejemplo, un modelo que no resulta validado puede obligar a cambiar alguna o todas las hipótesis de partida o a cambiar el preprocesamiento de los datos.

### *2.1.1 Formulación del problema*

Todo conocimiento surge de la formulación, aunque sea implícita, de un problema. Se debe prestar atención en no focalizar la formulación en la forma de resolverlo, ya que puede provocar que no se haga

una exposición clara del problema que se desea resolver, lo que dificulta la comprensión de la aplicación del método.

### 2.1.2 *Hipótesis iniciales*

Se establece un conjunto de datos de entrada-salida y una dependencia desconocida, o de la que se tiene un conocimiento parcial, que se desea estimar a partir de unos datos experimentales.

Los datos pueden ser generados de forma controlada mediante la realización de experimentos, o bien recopilados del medio sin posibilidad de influir en su generación. Los datos de histórico entrarían dentro de la variedad de datos recopilados.

Dentro del conjunto de datos, podemos establecer la siguiente distinción:

- Datos para el entrenamiento del modelo
- Datos para probar el modelo

### 2.1.3 *Procesamiento previo*

Los datos recopilados, ya sea mediante sensores, observaciones o provenientes de una base de datos se deben tratar para ser codificados y poder así ser procesados.

En esta fase se deben realizar dos actividades principales:

- **Reducción de las dimensiones de los datos**, en caso de ser necesario. Es decir, se puede realizar una extracción de los datos que están directamente asociados a las características o propiedades de los datos relevantes para el problema. Este proceso se conoce como *Extracción de características*, y es un aspecto clave que va a influir en el éxito de las conclusiones. Los aspectos claves de estas características son:
  - Contener información relevante del problema
  - Deben ser sencillos de extraer.
  - Invariancia a las transformaciones irrelevantes del problema.

- **Detección de datos inusuales** que no son consistentes con la mayoría de los datos observados. Estos datos pueden estar motivados por errores de medida, mal funcionamiento o casos anormales que no son significativos, y que podrían provocar desviaciones en el modelo.

### 2.1.4 *Aprendizaje*

Este proceso también se denomina *estima* y consiste en seleccionar un modelo usando los datos como guía. Esta selección normalmente necesita, bien de una medida que permita comparar cuán bien se ajusta cada modelo a los datos, o bien una mediad que permita cuantificar la fiabilidad de la predicción del modelo a nuevos datos.

Junto con la medida es necesario un algoritmo que seleccione el modelo más adecuado. Este proceso de selección del mejor hace que casi siempre detrás de un algoritmo de estimación se encuentre un problema de optimización.

El aprendizaje se divide en dos áreas principales: aprendizaje supervisado y aprendizaje no supervisado.

#### 2.1.4.1 *Aprendizaje Supervisado*

El objetivo del aprendizaje supervisado es hacer predicciones a futuro basadas en comportamientos o características que se han observado en el histórico de datos. El aprendizaje supervisado nos permite buscar patrones en datos históricos.

Como ejemplo de aplicación del aprendizaje supervisado podemos pensar en un histórico de datos de funcionamiento de un motor. Este histórico debe incluir los valores de los principales atributos de funcionamiento tanto en circunstancias de funcionamiento normales como de fallo. Con estos datos de histórico podemos realizar el entrenamiento de nuestro algoritmo de manera que en base a los fallos producidos y a sus circunstancias (valores de los atributos de funcionamiento), el algoritmo nos pueda indicar cuándo se debería realizar la siguiente revisión del mismo para evitar fallos. Para ello, el algoritmo, con sus datos de entrenamiento basado en el histórico, podrá identificar las circunstancias previas al fallo que permita anticiparnos a la avería.

Dos conceptos íntimamente relacionados con el aprendizaje supervisado son la *Clasificación y la Regresión*. La diferencia radica en que un sistema de clasificación predice una categoría, mientras que una regresión predice un valor numérico.

Como ejemplo de clasificación podemos pensar en la determinación de la tipología de fallo de un sistema, mientras que los algoritmos de regresión determinarían por ejemplo el tiempo medio entre fallos de una manera predictiva.

#### 2.1.4.2 *Aprendizaje No Supervisado*

Por otro lado, el aprendizaje no supervisado usa datos históricos que no están etiquetados. El fin es explorarlos para encontrar alguna estructura o forma de organizarlos. Por ejemplo, el aprendizaje no supervisado podemos aplicarlo sobre datos en bruto sobre los cuales no tenemos un conocimiento de la relación de sus atributos y sobre la que queremos realizar una agrupación de características.

### 2.1.5 *Validación del modelo*

La validación del modelo trata de asegurar que el modelo es útil no solo para los datos usados en el aprendizaje, sino también para otro conjunto de datos que se suele denominar datos de validación.

Puede darse el caso de que el modelo se ajuste correctamente a los datos usados en el entrenamiento, pero puede tener malos resultados en los datos de validación. En ese caso se dice que el modelo está sobre ajustado.

## 2.2 ESQUEMA GENERAL DEL APRENDIZAJE

Los ejemplos clásicos de problemas de aprendizaje automático se pueden agrupar en las siguientes categorías:

- Clasificación
- Regresión
- Probabilidad

El planteamiento general para las categorías anteriores sería el siguiente:

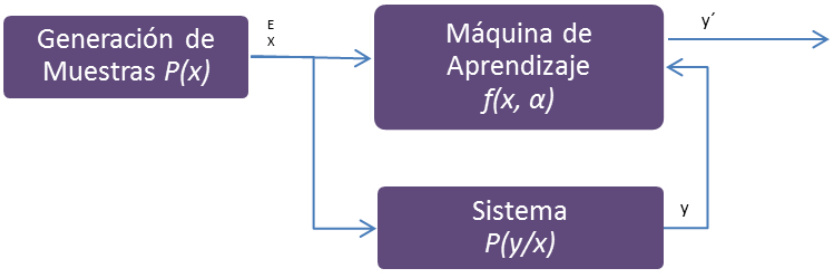



*Figura 2-2: Esquema general de aprendizaje*

Dado un conjunto de observación es X={x1, x2, … xn} que producen valores {y1, y2, …yn} de salida del sistema, las funciones de la máquina de aprendizaje proporcionan valores {y´1, y´2, … y´n}. Los valores de y´ dependerán de la función escogida, α, y el objetivo del aprendizaje es elegir aquel α que hace que el conjunto de observaciones aproxime de la mejor manera las salidas del sistema. El conjunto de

aproximaciones que soporta el aprendizaje refleja el conocimiento a priori que se tiene del sistema, o de la dependencia desconocida de la entrada y la salida.

De acuerdo con la figura 2.2 los datos de entrenamiento se generan de acuerdo con la siguiente función de probabilidad:

$$p(x, y) = p(x)\, p(y|x)$$

El número finito de datos generados por esta distribución es:

$$S = \{(xi, yi), i = 1, \dots N\}$$

La calidad de la aproximación realizada por el algoritmo de aprendizaje se mide mediante un función de pérdida o función de coste L(y, f(x, α)) que toma valores positivos o cero, y que representa el precio que pagamos, o pérdida que experimentamos cuando observamos (x, y), y predecimos f(x, α).

El valor medio de la función de pérdida se denomina riesgo funcional:

$$R(\alpha) = \int L\big(y, f(x, \alpha)\big) p(x, y) dx dy$$

Aprender es por tanto el proceso de estimar la función f(x, α) que minimiza R(α) sobre el conjunto de funciones que sostiene la máquina de aprendizaje utilizando sólo los datos de entrenamiento, es decir, sin conocer p(x,y).

## 2.3 ÁREAS RELACIONADAS CON EL APRENDIZAJE

El término aprendizaje automático (machine learning en inglés) apareció inicialmente ligado a la inteligencia artificial, campo en el que han desarrollado gran número de técnicas como las de Kohonen denominadas de autoorganización y autoaprendizaje, el método de aprendizaje en árboles binarios y los trabajos iniciales en el perceptrón, que condujeron posteriormente a las redes neuronales. Por otro lado, se han ido produciendo grandes avances en el aprendizaje dentro del campo de la estadística, dando lugar al aprendizaje estadístico.

En la actualidad existe la tendencia convergente en campos del aprendizaje automático, de la inteligencia artificial y el aprendizaje estadístico.

La minería de datos, o descubrimiento del conocimiento, se ha convertido en una palabra de moda para referirse a la búsqueda en grandes bases de datos para encontrar información relevante. Se aplica a gran variedad de problemas para obtener información de todo tipo, incluyendo la médica o genética. Este campo

está estrechamente relacionado con técnicas de aprendizaje automático como redes neuronales, árboles de decisión y técnicas estadísticas de clasificación.

## 2.4 ALGORITMOS Y SU CLASIFICACIÓN

Como veremos en el punto correspondiente a la aplicación en el sector ferroviario, en función de la naturaleza de nuestros datos y del objetivo que persigamos, será recomendable utilizar un tipo de algoritmo u otro.

Una clasificación funcional bastante intuitiva es la propuesta de Microsoft en su plataforma Azure:

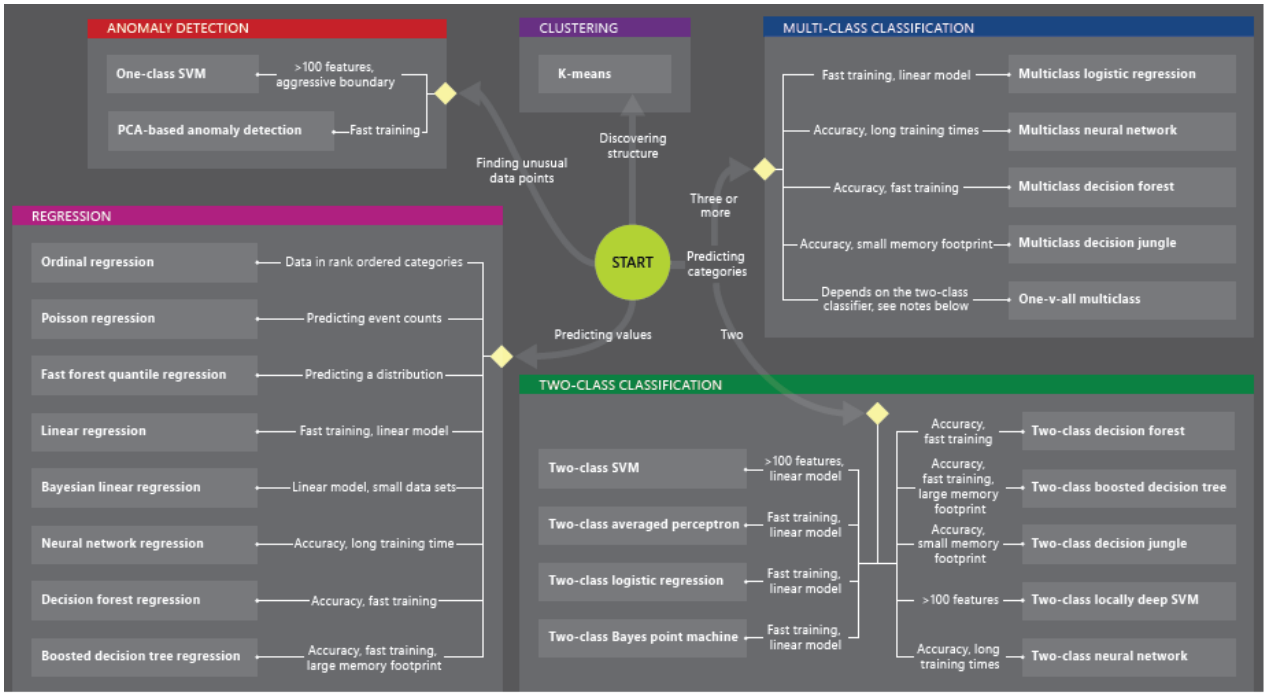



*Figura 2-3: Clasificación de algoritmos en función de su uso*

A continuación se describen y clasifican los tipos de algoritmos más utilizados en este tipo de técnicas:

### *2.4.1 Algoritmos de Regresión*

#### *2.4.1.1 Regresión línea*

La regresión lineal es una técnica estadística utilizada para investigar la relación entre dos o más variables. Se trata de un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε.

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

donde:

$Y_t$ : representa la variable dependiente, explicada o *regresando*.

$X_1, X_2 + \cdots + X_p$: representa las variables explicativas, independientes o *regresores*.

$\beta_1 + \beta_2 + \cdots + \beta_p$: representan los parámetros, miden la influencia que las variables explicativas tienen sobre la variable explicada..

$\beta_0$ : es la intersección o término "constante", las $\beta_i$ (i > 0) son los parámetros respectivos a cada variable independiente, y p es el número de parámetros independientes a tener en cuenta en la regresión.

El término lineal se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática. Los modelos lineales son una explicación simplificada de la realidad, mucho más ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadística.

Esta técnica es muy utilizada para realizar predicciones de una variable (objetivo) en términos de otras (regresivas). El término regresión fue acuñado por el inglés Francis Galton[2].

En función de las variables a investigar podemos tener regresión simple o regresión múltiple. Cuando la predicción de una variable se realizar únicamente en términos de una variables, hablamos de regresión simple. Por el contrario, cuando la predicción se hace teniendo en cuenta varias variables hablamos de regresión múltiple.

#### *2.4.1.2 Poisson*

La distribución de Poisson fue desarrollada por Siméon-Denis Poisson (1781-1840). Esta distribución de probabilidades es muy utilizada para situaciones donde los sucesos son impredecibles o de ocurrencia aleatoria.

La distribución Poisson es, junto con la distribución binomial, una de las más importantes distribución de probabilidad para variables discretas, es decir, sólo puede tomar los valores 0, 1, 2, 3, 4, ..., k.

Cada una de estas variables aleatorias representa el número total de ocurrencias de un fenómeno durante un periodo de tiempo fijo o en una región fija del espacio. Expresa la probabilidad de un número k de

2 Su trabajo con los guisantes y su posterior investigación en torno a la herencia de la altura lo condujeron a formular los conceptos de regresión y correlación y la ley de Galton de la herencia ancestral.

ocurrencias acaecidas en un tiempo fijo, si estos eventos ocurren con una frecuencia media conocida y son independientes del tiempo discurrido desde la última ocurrencia o suceso.

Se dice que X sigue una distribución de Poisson de parámetro λ y que se obtiene del producto n*p (que nombraremos como np, por simplicidad), que se representa con la siguiente notación:

$$X \sim Ps\ (\lambda)$$

La distribución de Poisson se caracteriza por las siguientes propiedades:

- Sea una población de tamaño ∞.
- Sea una muestra de tamaño n bastante elevado (se suele hablar de que tiende a ∞)
- Los sucesos son independientes entre sí.
- Sea A un suceso que tiene una probabilidad p de suceder durante un periodo de tiempo, siendo esta probabilidad de ocurrencia durante un periodo de tiempo concreto muy pequeña (se suele hablar de que tiende a 0).
- El producto n*p, tiende a aproximarse a un valor promedio o número medio, al que llamaremos λ. Por ejemplo, número medio de supresiones producidas en una línea AVE durante un mes.
- X: número de individuos de la muestra que cumplen A.
- El conjunto de posibles valores de A es, E = {0,1,2,3,4....}

Su función de probabilidad viene definida por:

$$F(X = x) = e^{-k} * \frac{\lambda^x}{x!}$$

donde x debe ser un entero positivo.

### *2.4.1.3 Redes Bayesianas*

Las redes bayesianas, también llamadas redes probabilísticas, son una herramienta que permite representar de forma estructurada la probabilidad conjunta de diferentes variables aleatorias, utilizando para ello un modelo gráfico de las relaciones existentes entre dichas variables. Su origen bayesiano[3] permite, entre otros, realizar cálculos sobre probabilidades subjetivas, tratar con la incertidumbre asociada a sistemas reales o incorporar en el razonamiento realizado el conocimiento a priori que se tiene sobre el sistema.

Aunque habitualmente son utilizadas para inferir probabilísticamente los valores de algunas de las variables que contienen, también es posible realizar otro tipo de operaciones como encontrar la configuración de

3 Utilización secuencial de Teorema de Bayes https://es.wikipedia.org/wiki/Teorema_de_Bayes

variables más probable, medir posibles conflictos entre los valores de las variables de la red o realizar análisis de sensibilidad sobre influencia de unas variables en otras.

Como se ha indicado anteriormente, una de las principales características de estas redes es la utilización de una representación gráfica para describir las relaciones de dependencia que se dan entre las variables que describen el sistema. Esta descripción se realiza sobre un grafo, cuyos nodos están asociados a las variables aleatorias que describen el sistema y cuyos arcos indican las relaciones existentes entre dichas variables.

Los grafos correspondientes a las redes bayesianas cumplen dos restricciones principales sobre las que se fundamentan todas las operaciones realizadas: tienen que ser grafos dirigidos y acíclicos (DAG, Directed Acyclic Graph). La utilización de un DAG como herramienta de modelado, no solo permite disponer de una descripción gráfica del sistema sirve también como soporte para desarrollar razonamientos eficaces. Por ejemplo, el concepto de separación direccional permite determinar si la evidencia de la que se dispone sobre algunas variables de la red es capaz de influir en el conocimiento que se tiene sobre otras variables.

Es importante destacar que los enlaces existentes entre las variables no tienen por qué representar impacto casual, ya que las propiedades asociadas al concepto de separación direccional forman parte de la semántica de la red. Por otro lado, la existencia de un enlace entre dos variables tampoco implica que haya una dependencia condicional entre las mismas, ya que este tipo de dependencias obedece también a la distribución condicional existente entre las variables. Por lo tanto, la estructura de la red bayesiana y el concepto de separación direccional permiten determinar la falta de dependencia entre variables de la res, pero no la existencia de la misma.

Existen distintos tipos de redes bayesianas:

- Clasificadas según el tipo de variables:
    - Redes discretas: Trabajan con variables discretas.
    - Redes continuas: Trabajan con variables continuas.
    - Redes mixtas: Compuestas por variables discretas y continuas.
- Organizadas según la estructura del DAG:
    - Árboles
    - Poli-árboles
    - Grafos múltiplemente conectados.
- Clasificadores bayesianos:
    - Gráfico dirigido y acíclico asociado a un clasificador bayesiano.
    - Red bayesiana dinámica

Es importante resaltar que en la definición del concepto de red no se establecen restricciones sobre el tipo de variables o de probabilidades que se pueden utilizar. Las restricciones de este tipo son impuestas posteriormente, a la hora de desarrollar los diferentes algoritmos de aprendizaje. Por lo tanto, entre las numerosas posibilidades existentes, habrá que elegir aquellas que permitan razonar sobre la red bayesiana utilizada para modelar el comportamiento de un determinado sistema.

#### *2.4.1.4 Redes de Neuronas*

Las redes neuronales artificiales son estructuras paralelas inspiradas en las neuronas biológicas. Una red neuronal artificial está compuesta por multitud de elementos simples, que denominamos neuronas, interconectadas de una forma más o menos densa y cuyo funcionamiento en conjunto puede dar lugar a un procesamiento no lineal complejo. Estas redes son capaces de ajustar su comportamiento a partir de datos experimentales de modo que son muy útiles en problemas donde el conocimiento del problema es incompleto o varía en el tiempo.

Una neurona biológica tiene tres partes principales: las dentrinas, el cuerpo de la neurona o soma y el axón. Las dentrinas son las fibras que reciben los impulsos al interior de la neurona. Cuando una neurona recibe un impulso, si el estímulo es lo suficientemente grande, producirá una excitación que provocará un cambio de estado en la neurona y que puede ser propagado a otras neuronas.

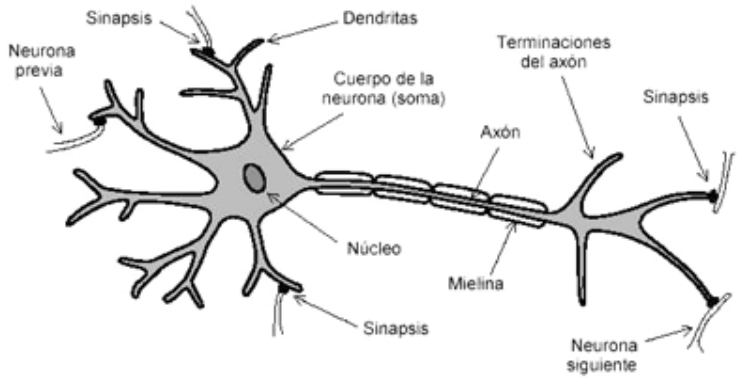



*Figura 2-4: Esquema red de neuronas biológicas*

De igual modo, una neurona artificial, se representa por nodos (cuerpo de la neurona) donde se realiza la suma de las señales que recibe de otros nodos o neuronas. El valor obtenido será utilizado para generar una señal que será transmitida a otras neuronas. Para el cálculo de la excitación de la neurona se tiene en cuenta los pesos asociados a las entradas.

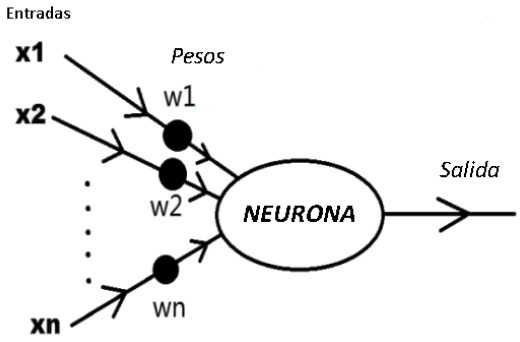



*Figura 2-5: Esquema red de neuronas artificial*

Existen varios elementos y parámetros que caracterizan una red neuronal como son el número de neuronas por capa, el grado y tipo de conectividad entre neuronas, el algoritmo de aprendizaje de las redes, la función de activación, la función de activación, la función de salida, etc.

- **Función de activación**: Se encarga de activar la neurona en función del estado de activación actual y de sus estradas.
- **Función de salida o transferencia**: En función de activación provoca una estado excitado, provocará una salida.
- **Niveles de neuronas y conectividad**: Las neuronas de una red se organizan en capas. Existen redes monocapa o multicapa. Las neuronas se pueden conectar entre neuronas de la misma capa o una capa y la siguiente. En general, una red de neuronas puede ser considerada como un grafo dirigido y ponderado. Cada uno de los nodos representa una neurona. Si el arco es de entrada llevará asociado un peso (grafo ponderado)

Atendiendo a las características del grafo al que se puede asemejar las redes de neuronas, así como si posee realimentación, podemos establecer la siguiente clasificación:

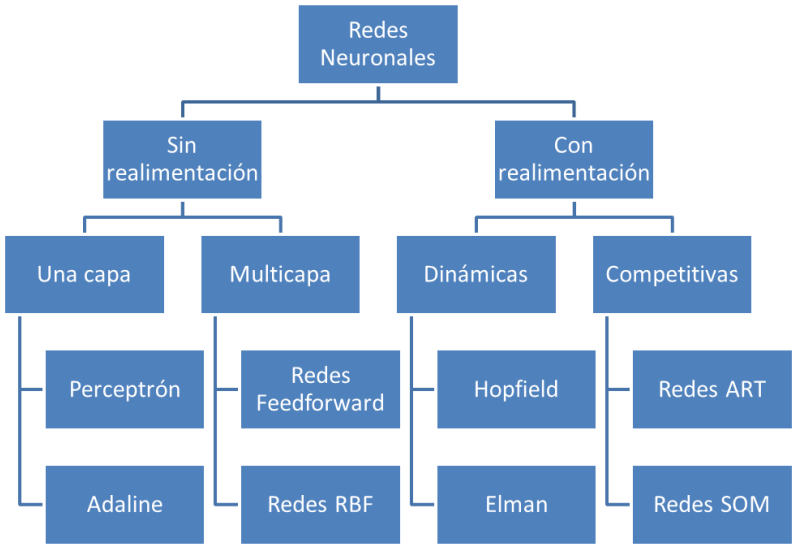



*Figura 2-6: Clasificación de redes de neuronas atendiendo al tipo de grafo*

En la aplicación de las redes neuronales podemos distinguir dos fases:

**Fase de aprendizaje o entrenamiento.**

La red es entrenada para realizar un tipo de procesamiento. Aprende la relación existente entre los datos presentados para devolver una determinada salida en cada situación.

Partiendo de un conjunto de pesos sinápticos aleatorio, se busca un conjunto de pesos que permitan a la red desarrollar correctamente una determinada tarea. Se trata por tanto del proceso de ajuste de los parámetros. Es un proceso iterativo, en el cual se va refinando la solución hasta alcanzar un nivel de operación suficientemente bueno. La mayoría de los métodos de entrenamiento consisten en proponer una función de error que mida el rendimiento actual de la red en función de los pesos sinápticos. El objetivo del método es encontrar el conjunto de pesos sinápticos que minimizan (o maximizan) la función.

**Fase de operación o ejecución:**

Consiste en evaluar el comportamiento de la red ante patrones nunca antes vistos. Esta fase es imprescindible para prevenir el sobreaprendizaje. En la mayoría de los casos se deja que el proceso de aprendizaje avance hasta alcanzar una cota de error razonable, guardando periódicamente las distintas configuraciones intermedias para luego seleccionar la de menor error de evaluación.

Las principales ventajas del uso de las redes neuronales artificiales como sistema de aprendizaje son:

- **Aprendizaje adaptativo**: Tienen la capacidad de aprender a realizar tareas basadas en la experiencia a partir de un entrenamiento con datos parciales.
- **Autoorganización**: La red puede crear su propia organización o su propia representación de la información después de la etapa de aprendizaje.
- **Tolerancia a fallos**: Una destrucción parcial de la red conduce a una degradación de los resultados, pero, en muchos casos, puede seguir funcionando.
- **Operación en tiempo real**: Las redes neuronales pueden ser realizadas para trabajar en paralelo, lo que permite su uso en tiempo real.
- **Fácil inserción dentro de la tecnología existente**: Una red puede ser rápidamente entrenada, comprobada, verificada y trasladada a hardware de bajo costo, lo que permite su inserción en tareas específicas.
- **No lineal**: En general la neurona realiza una transformación lineal de las entradas. La combinación de las distintas neuronas podrá representar funciones no lineales.

### *2.4.2 Algoritmos de Clasificación*

#### *2.4.2.1 Árboles de decisión*

Un árbol de decisión representa un conjunto de restricciones o condiciones que se organizan de forma jerárquica y que se aplican sucesivamente desde una raíz hasta llegar a un nodo terminal u hoja del árbol. Los árboles de decisión con conocimiento adquirido se pueden representar de manera alternativa como un conjunto de reglas "*Si-Entonces*" para mejorar su comprensión.

Este método de aprendizaje se encuentra entre los más populares de los algoritmos de inferencia por inducción[4] y ha sido utilizado con éxito en diferentes aplicaciones, desde sistemas de diagnóstico médico[5] hasta sistemas de determinación de crédito bancario. Pese a tratarse de sistemas de propósito general, las aplicaciones van dirigidas a procesos de clasificación.

La estrategia de aprendizaje es no incremental y está basada en ejemplos. Se presenta al sistema un conjunto de casos relevantes para clasificar y desarrollar un árbol de decisión basado en el conocimiento, desde la raíz hasta las hojas sin importar el orden de llegadas de los casos. Estos ejemplos pueden

4 http://www.ine.es/ss/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3B+filename%3D532%2F434%2F94_4.pdf&blobkey=urldata&blobtable=MungoBlobs&blobwhere=532%2F434%2F94_4.pdf&ssbinary=true

5 http://www.medigraphic.com/pdfs/veracruzana/muv-2009/muv092c.pdf

suministrarse a partir de una base de datos que contenga un histórico de observaciones, que será uno de los casos de uso que se desarrollarán en este trabajo.

Para realizar la formalización, un árbol de decisión se representa mediante un grafo con estructura arbórea inducido desde un conjunto de datos *IS* = <*U*, *A*, *De*>, donde *U* es un cierto conjunto de objetos denominado universo, *A* es el conjunto de condiciones sobre los atributos y *De* es el conjunto de decisiones sobre los mismos.

Un árbol de decisión puede considerarse como un grafo acíclico, donde podemos distinguir:

- Nodos internos: Se corresponden con una condición elegida en el conjunto de atributos, y donde cada ramificación de un nodo interno representa un valor del dominio de la condición del atributo.
- Nodo terminado: Contienen las clases o etiquetas de clasificación.

Los árboles de decisión tienen las siguientes ventajas:

- Pueden representar límites de decisión no lineal.
- Son eficientes en los cálculos y uso de memoria durante el entrenamiento y la predicción.
- Realizan la clasificación y selección de características integradas.
- Son resistentes en presencia de características ruidosas, es decir, fuera de los rangos habituales de la muestra

#### *2.4.2.1.1 Algoritmo ID3*

La mayoría de los algoritmos utilizados para el aprendizaje mediante árboles de decisión son variaciones de una algoritmo base denominado ID3. Este algoritmo plantea la búsqueda voraz[6] desde la raíz hasta las hojas en el espacio de los posibles árboles de decisión.

La estructura base del algoritmo ID3 es iterativa. Se elige de manera aleatoria un subconjunto del conjunto de datos de entrenamiento, llamado *ventana*, y con él se genera un árbol de decisión. Este árbol clasificará de manera correcta todos los objetos de la ventana ya que dispone de los resultados para el entrenamiento. Posteriormente el árbol intenta extender la clasificación al resto de elementos del conjunto de entrenamiento usando el mismo árbol de decisión. En el caso de resultar correctas las clasificaciones para todos los elementos, se da por terminado el proceso de aprendizaje. En caso contrario, una selección de aquellos objetos que no han sido bien clasificados se añade a la *ventana* y el proceso de clasificación vuelve a comenzar.

6 https://es.wikipedia.org/wiki/Algoritmo_voraz

El funcionamiento explicado anteriormente corresponde a la versión básica del algoritmo, que asume que el conjunto de datos de entrenamiento no tiene ninguna perturbación, es decir, que no tiene medidas imprecisas o juicios subjetivos, pero este funcionamiento perfecto no suele darse en la vida real. Esto provoca que algunos elementos del conjunto se clasifiquen mal.

Para resolver el problema de la perturbación inherente a los datos reales, se han desarrollado diferentes variaciones del algoritmo ID3, entre las que destaca la utilización de árboles de decisión basados en lógica borrosa[7] (fuzzy). Estudios detallados de los métodos basados en lógica borrosa, comparados con métodos numéricos sin tratamiento de la incertidumbre, han demostrado una gran mejora en la precisión de la predicción, debido principalmente a la reducción del modelo de variabilidad. Por otro lado, los parámetros presentan mayor estabilidad (casi un 50% menos de variabilidad que los árboles de decisión clásicos) y eso implica una mejora en la interpretación.

En los sistemas basados en la lógica borrosa, las reglas proporcionan una gran facilidad de comprensión y transfieren conocimiento de alto nivel. Por otro lado, la teoría de conjuntos, junto con la lógica borrosa y los métodos de razonamiento aproximado, proporcionan la habilidad de modelar tanto detalles del conocimiento como tendencias. Por este motivo, se están utilizando cada vez más para la resolución de problemas con incertidumbre, ruido o falta de precisión en los datos, abarcando tanto aplicaciones industriales, medioambientales, financieras, etc.

## 2.5 APLICACIÓN EN EL SECTOR FERROVIARIO

La inteligencia artificial y más concretamente el aprendizaje automático puede ser aplicado en gran cantidad de entornos ya sean ferroviarios, industriales o empresarial. Concretamente, dentro del sector ferroviario, como veremos en el presente apartado, puede ser aplicado a gran cantidad de ámbitos.

En la actualidad ya se están usando técnicas de Inteligencia Artificial en el sector, concretamente se usan algoritmos genéticos en la creación de mallas, resolución de conflictos o problemas de optimización. Por otro lado, se han encontrado muy pocos casos publicados del uso de técnicas de aprendizaje automático en el ámbito ferroviario, lo que ha motivado la realización del presente proyecto.

Como hemos visto en apartados anteriores, a grandes rasgos, el aprendizaje automático se basa en la extracción de conocimiento con base en la aplicación de diferentes técnicas sobre un conjunto de datos de histórico que nos permite aprender las dependencias existentes en la información aportada para el

7 http://arantxa.ii.uam.es/~dcamacho/logica/recursos/fuzzy-into-esp.pdf

aprendizaje. Estas conclusiones que obtenemos nos permiten aplicar lo aprendido sobre nuevos conjuntos de datos.

Debido a que el aprendizaje automático puede ser aplicado a la práctica totalidad de técnicas que forman el sistema ferroviario, en vez de enumerar cada una de las situaciones individuales parece más interesante realizar una agrupación de las mismas aportando algún ejemplo.

### *2.5.1 Mantenimiento predictivo*

El mantenimiento predictivo es uno de los campos de mayor potencial de las técnicas de aprendizaje automático. La tendencia en las actividades de mantenimiento, ya sea de infraestructuras como de sistemas, se está orientando hacia un mantenimiento según estado. Si bien este tipo de mantenimiento supone una gran mejora en cuanto a costes, depende de la monitorización o supervisión continua de los elementos a mantener para conocer el estado de sus características.

Este tipo de mantenimiento permite determinar cuándo un elemento va a requerir una actuación en función de su estado, pero no sabemos cuándo se van a dar esas condiciones ni cómo afectan en su vida útil.

El proceso de mantenimiento actual presenta las siguientes necesidades:

- La mayoría de las herramientas utilizadas para las actividades de mantenimiento se limitan al registro de la información y en la gestión de la actividad, pero esta información no se explota ni analiza convenientemente.
- Existen elementos de los que no se conoce como afectan los cambios de los parámetros del estado de un elemento en su vida útil.
- No se dispone de métodos que permitan determinar cómo afectarán cambios en el proceso de mantenimiento de un elemento a su ciclo de vida.
- El panorama económico actual requiere de mecanismos que permitan tanto la optimización de las tareas de mantenimiento como la diferenciación de la competencia por medio de nuevo métodos que aporten valor añadido al proceso.

Por otro lado, la digitalización de los procesos de mantenimiento actuales están generando gran cantidad de información, que convenientemente tratada podría desvelar patologías y comportamientos que actualmente están ocultos y que pueden aportar nuevos enfoques en las tareas de mantenimiento.

El uso de técnicas de aprendizaje automático puede ayudar en la determinación de los siguientes cuestiones:

- Predicción de vida útil.
- Detección de incidencias futuras.
- Catalogación de incidencias.
- Tasa de fallo y tasa de fallo media
- Tiempo medio entre fallos (MTBF)

#### 2.5.1.1 Infraestructura

En relación al mantenimiento predictivo de los elementos de la infraestructura, es necesario tener en cuenta dos factores:

- Normalmente son elementos que disponen de vidas útiles muy largas y por ese motivo es difícil disponer de un histórico de datos completo.
- Salvo elementos concretos (túneles, viaductos, etc.) la mayoría de los elementos no se encuentran instrumentados por lo que no se dispone de información para su estudio.
- Los ciclos de medición del estado están muy distanciados en el tiempo, por lo que a veces los elementos sufren actuaciones que modifican su vida útil, por lo que rompe la serie temporal.

Este tipo de técnicas pueden aportar un gran valor añadido en aquellos casos en los que se produzca situaciones con una componente de desgaste. Por otro lado, pueden utilizarse para la predicción de situaciones potencialmente peligrosas para la seguridad en la circulación como pueden ser crecidas de elementos hidráulicos (torrentes de agua).

Por otro lado, una vez calibrado un determinado modelo, puede utilizarse para la realización de simulaciones simplemente con cambiar las variables de entrada al modelo.

#### 2.5.1.2 Material rodante

El mantenimiento de los componentes del material rodante es otros de los casos de uso donde mayor partido se podría sacar a las técnicas de aprendizaje automático. El motivo es que este tipo de componentes no tiene los problemas que estaban presentes en la infraestructura. Por un lado, las vidas útiles son medias y por otro, son elementos suelen estar bien instrumentados y se realizan gran cantidad de mediciones. Esto facilita que se dispongan de datos suficientes para poder entrenar y calibrar los modelos.

#### 2.5.1.3 Instalaciones

Con respecto a los elementos de instalaciones, tiene la ventaja de que son elementos que disponen de un alto nivel de monitorización. Por otro lado, gran parte de los elementos disponen de un sistema de logs

(ficheros de auditoría de funcionamiento). Estos ficheros podrían ser procesados automáticamente en búsqueda de patologías que puedan ser identificadas como futuros fallos.

Por otro lado, en instalaciones es donde se pueden sacar partido a la capacidad de catalogación de incidencias, Tasa de fallo o Tiempo medio entre fallos (MTBF).

### 2.5.2 *Explotación de la infraestructura*

Con respecto al ámbito de la explotación de la infraestructura, las técnicas de aprendizaje automático deben estar integradas en el software de explotación ya que nos debe permitir:

- Detección temprana y propuesta de resolución de conflictos en la explotación.
- Propuestas de planes de acción en función de las diferentes situaciones.
- Optimización de recursos
- Previsión de puntualidad de las circulaciones.

### 2.5.3 *Planificación y Gestión*

Al igual que sucede con las aplicaciones para al explotación de la infraestructura, se pueden emplear este tipo de técnicas integradas dentro de las herramientas de gestión y planificación con el objetivo de ayudar en tareas como:

- Planificación y simulación de la disponibilidad de recurso (ya sean de mantenimiento o de explotación).
- Tratamiento de la información procedente del consumo de energía.

### 2.5.4 *Tipos de usos de estas técnicas*

Como hemos visto en los apartados anteriores, existen diferentes ámbitos en los tiene cabida el uso de este tipo de técnicas. Por ese motivo, es interesante identificar los tipos de usos que se puede realizar, basados principalmente en el modelo y volumen de datos analizados.

Se han identificados dos tipos de usos de los modelos predictivos:

- Predicción en tiempo de ejecución: Uso de los algoritmos para la estimación de un único registro, en el mismo momento en el que se producen los datos de entrada (por ejemplo, generación de alarmas en tiempo real). Podemos asemejarlo a la ejecución automática por parte de un sistema de tiempo real.

- Predicción por lotes: Uso en tiempo diferido de los algoritmos para un conjunto completo de datos (por ejemplo, evaluación de todos los datos del auscultador de una campaña). Podemos asemejarlo al trabajo en gabinete.

# 3 DESARROLLO PRÁCTICO

## 3.1 PLARAFORMA DE APRENDIZAJE AUTOMÁTICO

El uso del aprendizaje automático está experimentando una gran expansión en todos los sectores. Por este motivo, los fabricantes de software están realizando grandes inversiones para proporcionar este tipo de funcionalidades en sus productos. Gracias a ello, ya no es necesario (aunque es posible) que se tenga que realizar la implementación de los algoritmos de forma manual para poder utilizarlos.

En la actualidad existen diferentes posibilidades de aplicar algoritmos de aprendizaje automático en nuestros estudios o productos.

- Servicios en la nube: Plataformas que se encuentra accesibles vía internet y que ya tienen implementados diferentes algoritmos. Como ejemplos de software en la nube poder nombrar BigML o Azure.
- Software local: Utilizamos un software local para programar y ejecutar nuestros algoritmos. Como ejemplo podemos nombrar Matlab.

Las soluciones basadas en servicios en la nube se van a imponer a la contratación de la Infraestructura privada (on premise), ya que permite a las empresas liberarse de las tareas de instalación y mantenimiento de la Infraestructura, para centrarse en tareas que aporten valor al proyecto. Ya no hablamos de adquirir máquinas (virtuales o físicas) donde hemos instalado y configurado nuestra propia solución, sino de utilizar los servicios que necesitemos en cada momento, pagando solo por el tiempo de procesamiento y por el almacenamiento. Por ejemplo, si necesitamos un servicio de aprendizaje automático, donde poder definir un algoritmo de predicción que trabaje con nuestra propia información, basta con contratar el servicio en la nube y pagar solo por el tiempo de uso.

Por otro lado, las soluciones de software local podríamos clasificarlas en Software de Propósito General y Software Específico. El software de propósito general nos permite la implementación de nuestros algoritmos, como es el caso de Matlab, pero requiere de un conocimiento avanzado del software ya que

requiere de programación. Por el contrario, el Software Específico permite su uso mediante las funciones específicas que aporta su interfaz de usuario, lo que simplifica su uso.

Para la realización de las pruebas de concepto del presente proyecto se utilizará la plataforma de Machine Learning de Microsoft Azure debido a que dispone de los algoritmos de aprendizaje que se requieren para nuestros casos de uso, permitiéndonos centrarnos en las aplicaciones de la tecnología en vez de en la tecnología en sí.

## 3.2 CASO DE USO 1: ANÁLISIS DEL AUSCULTADOR DINÁMICO

Para este primer caso de uso de la aplicación de algoritmos de aprendizaje automático nos centraremos en los algoritmos de clasificación (Algoritmos de Clasificación 2.4.2). El objetivo es poder entrenar un modelo que automáticamente identifique el tipo de intervención que se debe realizar (no requiere intervención, requiere la programación de una intervención o requiere una intervención inmediata) en función de los datos aportados por el auscultador dinámico.

### *3.2.1 Datos de partida*

Partimos de los datos del auscultador dinámico de una línea de alta velocidad que denominaremos línea A, por motivos de confidencialidad. Se dispone de los datos de dos años completo correspondientes a los años 2010 y 2011. Los parámetros que aporta el auscultador y que serán la base de nuestro caso de uso son los siguientes:

| Campo | Descripción |
|---|---|
| Fecha | Fecha en la que se realiza la auscultación dinámica |
| Vía | Vía que se ha auscultador |
| Pk Inicio | Pk inicial del defectos encontrado, indicada en metros |
| Pk Final | Pk final del defecto encontrado, indicada en metros |
| Velocidad | Velocidad de paso del coche auscultador medida en Km/h. |
| ALB1 | Aceleración lateral 1, medido en $m/s^2$ |
| ALB2 | Aceleración lateral 2, medido en $m/s^2$ |
| AVCG1 | Aceleración vertical en caja de grasa 1, medido en $m/s^2$ |
| AVCG2 | Aceleración vertical en caja de grasa 1, medido en $m/s^2$ |
| ALC | Aceleración lateral, medido en $m/s^2$ |
| AVC | Aceleración vertical, medido en $m/s^2$ |

*Tabla 3-1: Datos de partida – Caso de uso 1*

Es importante destacar que para la realización del *Experimento*, como se denominan a los procesos de machine learning en Azure, no disponemos de la función matemática ni del conocimiento para determinar qué tipo de intervención se requiere en función de los parámetros. Solamente disponemos de datos con ejemplos de situaciones en las que se han tenido que realizar los diferentes tipos de intervención para facilitárselos al algoritmo de aprendizaje. Este dato es importante ya que el objetivo de esta prueba de concepto es evaluar si somos capaces de hacer que un modelo aprenda y extraiga el conocimiento necesario para la identificación del tipo de intervención necesaria solamente basándonos en un conjunto de datos.

### *3.2.2 Ejecución de la prueba*

Para la ejecución de la prueba el primer paso debe ser la preparación de los datos. Como se ha visto en el apartado 3.2.1 disponemos de los datos de partida con las variables que definen la clasificación de los elementos. Estos datos deben ser enriquecidos con la variable objetivo, es decir, la variable que debe aprender nuestro algoritmo para poder clasificar nuestras observaciones en: no requiere intervención, requiere la programación de una intervención o requiere una intervención inmediata.

Nuestros datos para el aprendizaje tienen la siguiente forma:

| Via | Fecha | Actuacion | PK Ini | PK Fin | Veloc | ALB1 | ALB2 | AVCG1 | AVCG2 | ALC | AVC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| V2 | 10/08/2011 | inmediata | 563243,903 | 563219,885 | 293,99 | 0,00 | | 0 | 0 | 0,00 | 2,21 |
| V2 | 10/08/2011 | Sin Intervencion | 563146,83 | 563133,82 | 295,173 | 0,00 | | 0 | 0 | 0,00 | 1,42 |
| V2 | 10/08/2011 | Sin Intervencion | 563050,757 | 563039,749 | 295,646 | 0,00 | | 0 | 0 | 0,00 | 1,11 |
| V2 | 10/08/2011 | programada | 562521,358 | 562515,353 | 297,382 | -3,00 | | 54,67 | 42,833 | 0,00 | 0,00 |
| V2 | 10/08/2011 | Sin Intervencion | 560916,147 | 0 | 300,381 | 0,00 | | 38,553 | 0 | 0,00 | 0,00 |
| V2 | 10/08/2011 | Sin Intervencion | 559697,227 | 559694,225 | 302,353 | 0,00 | | 35,041 | 0 | 0,00 | 0,00 |
| V2 | 10/08/2011 | Sin Intervencion | 555737,24 | 0 | 300,17 | 0,00 | | 30,089 | 0 | 0,00 | 0,00 |
| V2 | 10/08/2011 | Sin Intervencion | 554969,661 | 0 | 300,959 | 3,68 | | 0 | 0 | 0,00 | 0,00 |
| V2 | 10/08/2011 | programada | 553692,697 | 553687,694 | 300,091 | -4,41 | | 0 | 0 | 0,00 | 0,00 |

*Tabla 3-2: Muestra de datos de partida - Caso de uso 1*

Una vez que disponemos de los datos necesarios para que nuestro algoritmo aprenda como inferir la clasificación en función de variables de entrada (recordemos que disponemos de 19.328 filas), el siguiente paso será la transformación de formato para que sea entendido por la plataforma del aprendizaje automático que utilizaremos, en nuestro caso, Microsoft Azure[8]. Para poder cargar los datos de entrenamiento en Azure deberemos convertir nuestro Excel en formato CSV.

*"Via","Fecha","Actuacion","PK Ini","PK Fin","Veloc","ALB1","ALB2","AVCG1","AVCG2","ALC","AVC"*
*"V1",10/8/2011 0:00:00,"Sin Intervencion",341912.57,341916.57,280.57,-3.29,,32.97,0.00,0.00,0.00*
*"V1",10/8/2011 0:00:00,"Programada",341980.56,0.00,279.60,4.04,,0.00,0.00,0.00,0.00*
*"V1",10/8/2011 0:00:00,"Sin Intervencion",342604.42,0.00,271.71,3.37,,0.00,0.00,0.00,0.00*

8 https://azure.microsoft.com/es-es/services/machine-learning/

*"V1",10/8/2011 0:00:00,"Sin Intervencion",342829.37,0.00,270.66,0.00,,30.27,0.00,0.00,0.00*
*"V1",10/8/2011 0:00:00,"Sin Intervencion",343282.27,343273.27,266.84,0.00,,33.03,0.00,0.00,-1.16*
*"V1",10/8/2011 0:00:00,"Programada",346075.66,346079.66,276.84,0.00,,0.00,0.00,0.00,1.93*
*"V1",10/8/2011 0:00:00,"Programada",346546.55,346551.55,277.36,-4.56,,0.00,0.00,0.00,0.00*
*"V1",10/8/2011 0:00:00,"Programada",346599.54,0.00,277.05,5.69,,0.00,0.00,0.00,0.00*

Una vez que disponemos de nuestros datos en formato CSV, ya estamos en disposición de crear el Dataset que será el origen de las pruebas dentro de Azure.

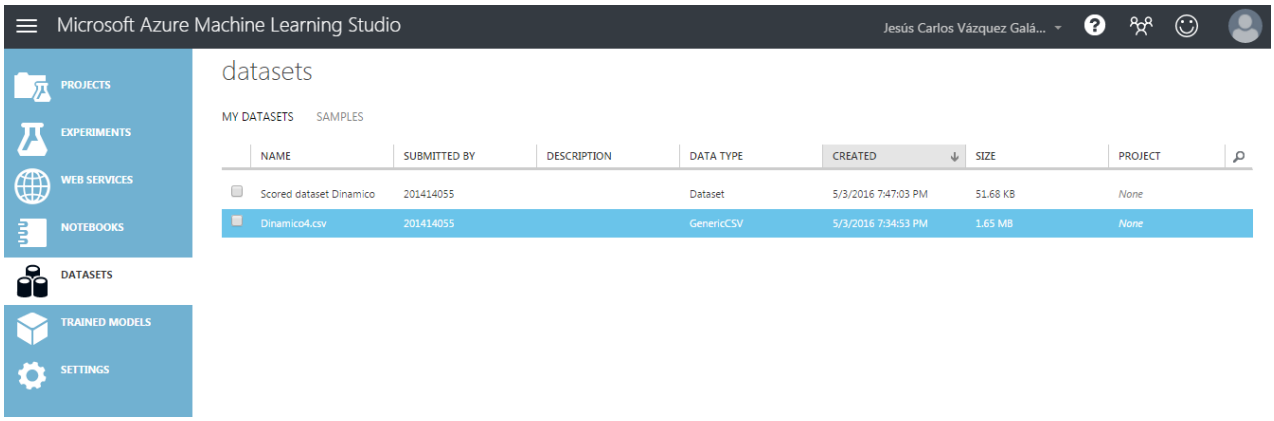



*Figura 3-1: Dataset para aprendizaje - Caso de uso 1*

Finalmente creamos le modelo. Los pasos que se deben seguir para la creación del modelo de aprendizaje será los siguientes:

- Importación de los datos de aprendizaje (dataset).
- Selección de las columnas correspondientes a los datos del aprendizaje. Es necesario indicar al modelo, del total de columnas que dispongamos en nuestro dataset, cuáles son las columnas que van a intervenir en el proceso de aprendizaje.
- Separación de los datos de aprendizaje en dos grupos:
    - Datos de aprendizaje: Filas que serán utilizadas para que el algoritmo aprenda como se deben clasificar los elementos en función de las variables de entrada.
    - Datos de verificación: Filas que se utilizarán para comparar la precisión del modelo de aprendizaje. El algoritmo realizará la predicción de la clasificación de estos elementos, gracias al aprendizaje realizado con el conjunto de datos anterior y comparará sus resultados con los resultados reales.
- Entrenamiento del modelo.
- Evaluación del modelo.

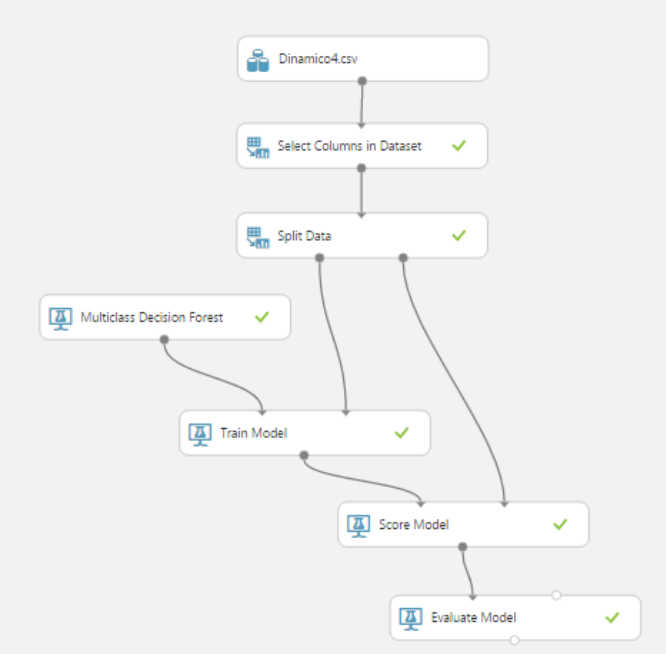



*Figura 3-2: Entrenamiento de modelo de clasificación - Caso de uso 1*

### *3.2.3 Resultados en función del algoritmo utilizado*

Partiendo del modelo anterior, simplemente con sustituir el algoritmo de aprendizaje (en la imagen anterior, MultiClass Decision Forest), podemos lanzar la ejecución con otro algoritmo diferente. En este caso de uso se ha preferido realizar el análisis caso a caso. En el siguiente caso de uso se realizará un análisis en paralelo.

Es importante tener en cuenta el número total de casos de ejemplo disponibles.

| Tipo Actuación | Ejemplos |
|---|---|
| Inmediata | 43 |
| Programada | 905 |
| Sin Intervencion | 18379 |

Por otro lado, también hay que tener en cuenta que del total de datos disponibles, el 70 % se ha utilizado para el entrenamiento y el 30% restante para la comprobación de la precisión.

A continuación se exponen los resultados obtenidos con los diferentes algoritmos utilizados.

### 3.2.3.1 *Multiclass Decision Forest*

El algoritmo Decision Forest es un método de aprendizaje de clasificación. El nombre de multiclass se debe al hecho de que permite la clasificación en más de dos clases (elementos de la clasificación). El algoritmo funciona creando bosques de árboles de decisión y, a continuación, vota la salida más popular.

El voto es una forma de agregación, en la que cada árbol de decisión genera un histograma de frecuencia no normalizado de etiquetas. El proceso de agregación suma estos histogramas y normaliza el resultado para obtener las "probabilidades" para cada una de las etiquetas. Los árboles que tienen una mayor fiabilidad en la predicción tienen un peso mayor en la decisión final.

En general, los árboles de decisión son modelos no paramétricos, lo que significa que admiten datos con distribuciones variadas. En cada árbol, ejecuta una secuencia de pruebas simples para cada clase, aumentando los niveles de una estructura de árbol hasta que se alcanza un nodo hoja (decisión).

Las parametrizaciones que se suelen hacer de este tipo de algoritmos se basan en definir el número de árbol de decisión y la profundidad que pueden tener los árboles. A continuación veremos pruebas con diferente número de árboles.

Los resultados de las predicciones realizadas con este algoritmo son las siguientes:

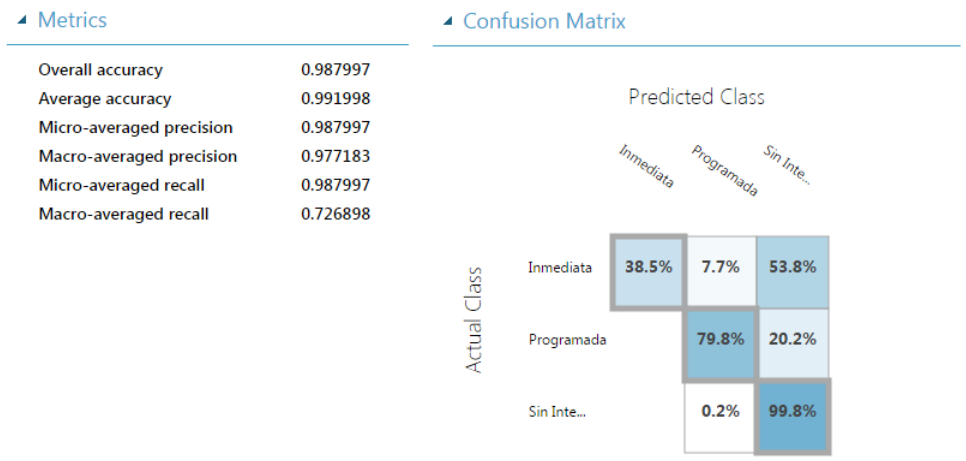



*Figura 3-3: Resultados algoritmo multiclass decision forest*

Se han realizado pruebas con diferentes configuraciones del algoritmo en función del número de árboles y de su profundidad, identificando los mejores resultados en verde en la siguiente tabla:

| Número de arboles de decisión | Profundidad | Overall accuracy | Average accuracy | Micro-averaged precision | Macro-averaged precision | No Requiere Intervención | Requiere Intervención Programada | Requiere Intervención Inmediata |
|---|---|---|---|---|---|---|---|---|
| 5 | 32 | 0,987997 | 0,991998 | 0,987997 | 0,977183 | 99,80% | 79,80% | 38,50% |
| **8** | **32** | **0,988618** | **0,992412** | **0,988618** | **0,974694** | **99,80%** | **82,80%** | **30,80%** |
| 15 | 32 | 0,987376 | 0,991584 | 0,987376 | 0,978423 | 99,80% | 78,30% | 30,80% |
| **15** | **64** | **0,988618** | **0,992412** | **0,986818** | **0,974694** | **99,80%** | **82,80%** | **30,80%** |

*Tabla 3-3: Comparativa de resultados algoritmo Decision Forest*

Para poder analizar los datos es necesario tener en cuenta el número de casos de ejemplo que se disponía para el entrenamiento. Por otro lado, también hay que tener en cuenta que del total de datos disponibles, el 70 % se ha utilizado para el entrenamiento y el 30% restante para la comprobación de la precisión.

Este modelo, con más casos de entrenamiento para el tipo de actuación *Inmediata* y *Programada*, puede realizar predicciones de clasificación con una presión superior al 99%.

### *3.2.3.2 Multiclass Decision Jungle*

Selvas de decisión son una extensión reciente de los bosques de decisión. Una selva decisión consiste en un conjunto de decisiones basadas en un grafo dirigido y acíclico (DAG, Directed Acyclic Graph).

En nuestro caso, este algoritmo se ha comportado bien con las categorías que disponían de más ejemplos, pero con los ejemplos del tipo de actuación "Inmediata" no ha sido capaz de aprender.

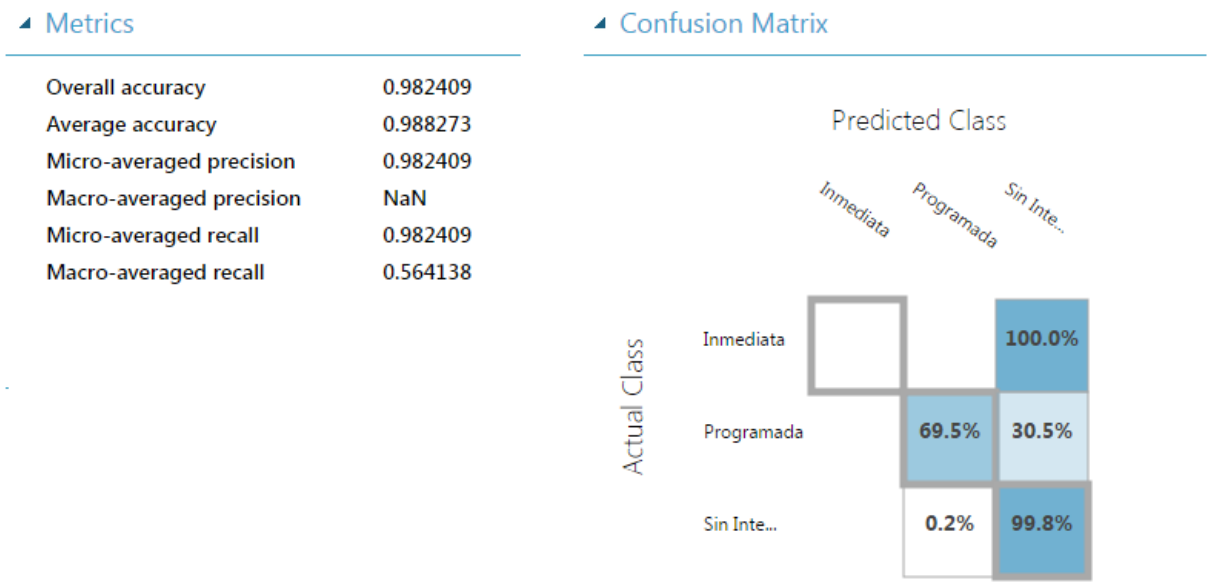



*Figura 3-4: Resultados algoritmo multiclass decision jungle*

### 3.2.3.3 *Multiclass Logistic Regression*

La regresión logística es un método muy utilizado en estadística con el objetivo de predecir la probabilidad de un resultado, normalmente utilizado para tareas de clasificación. Este algoritmo, basado en la regresión lógica, trata de predecir la probabilidad de ocurrencia de un evento mediante el ajuste de los datos a una función logística.

La ejecución de este algoritmo con nuestro conjunto de datos no ha sido capaz de realizar las predicciones esperadas.

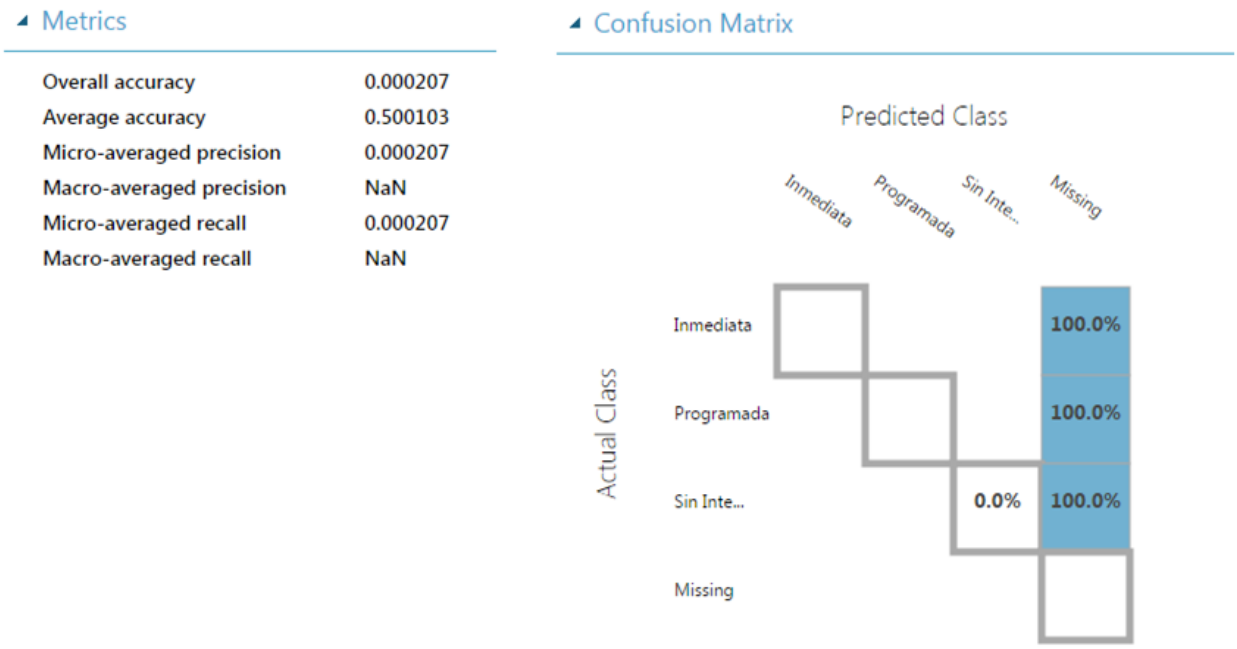
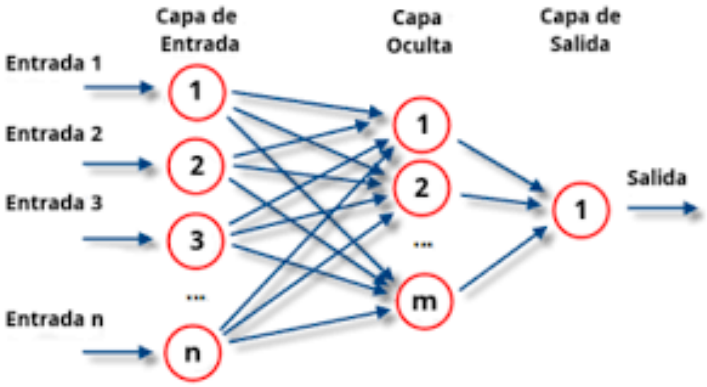



*Figura 3-5: Resultados algoritmo multiclass logistic regression*

### 3.2.3.4 *Multiclass Neural Network*

Tal como se ha explicado en el apartado referente a las redes de neuronas, una red neuronal es un conjunto de capas de neuronas interconectadas, en el que las entradas conducen a las salidas por medio de unos arcos ponderados y nodos. Los pesos en los arcos se aprenden cuando se entrena la red neuronal con los datos de entrada. La dirección de la gráfica va desde las entradas  hasta la salida a través de la capa oculta.




La mayoría de las tareas de predicción se puede lograr fácilmente con sólo una o unas pocas capas ocultas, aunque la investigación reciente ha demostrado que las redes neuronales profundos (DNN) pueden ser muy eficaces para tareas complejas (tales como imagen o de reconocimiento de voz), en el que un modelo de capas sucesivas aumenta los niveles de profundidad semántica.

La ejecución de este algoritmo con nuestro conjunto de datos no ha sido capaz de realizar las predicciones esperadas.

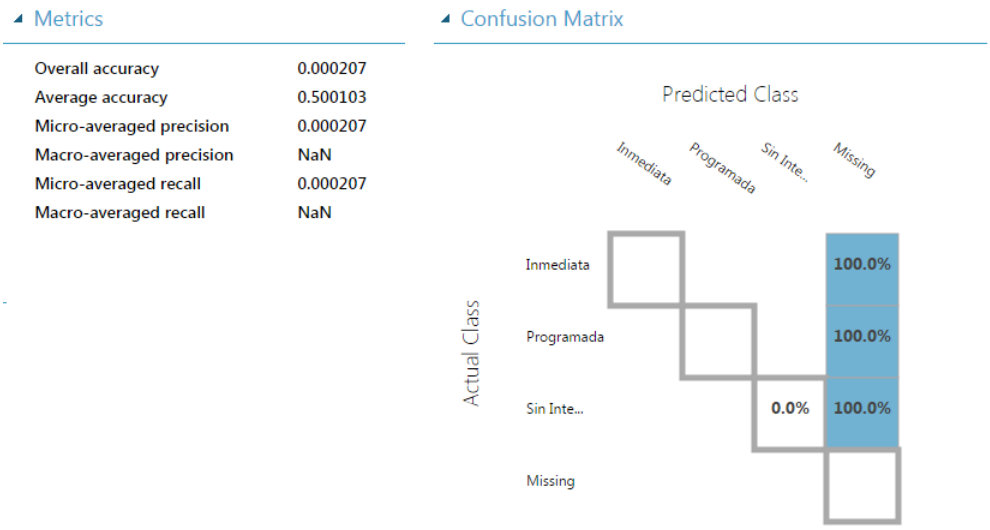



*Figura 3-6: Resultados algoritmo multiclass neural network*

### 3.2.4 Conclusiones del caso de uso

De la realización del caso de uso con los algoritmos anteriores podemos obtener las siguientes conclusiones.

- Los algoritmos que mejor se han comportado han sido los basados en árboles de decisión. El algoritmo *Multiclass Decision Forest*, ha obtenido un 99,80% de acierto en actuaciones *Sin*

*Intervención*, un 82,80% en *Programada* y un 30,80% en *Inmediata* (con solo 43 casos para aprender).

- El estado *Sin Intervención*, que obtiene un mayor número de aciertos, es debido a dos causas, por un lado es el estado más probable y por otro, es el que más casos tenía el algoritmo para aprender.
- El estado que menos aciertos presenta es el de *Inmediata*. Esto es debido a que al algoritmo no dispone de ejemplos suficientes (43 casos en la situación más favorable).

## 3.3 CASO DE USO 2: ANÁLISIS DE VIDA ÚTIL DE RUEDAS

El caso de uso anterior se centró en la realización de pruebas de algoritmos de clasificación. Como recordamos, los algoritmos de clasificación nos permiten, como en el caso anterior, clasificar datos, con el criterio que nosotros definamos, ya sea tipología de fallos, tipo de actuación requerida, etc.

En este caso de uso nos centraremos en los algoritmos de regresión. A diferencia de los algoritmos de clasificación, los de regresión nos van a permitir obtener valores discretos, como pueden ser la tasa de fallo, vida útil del elemento, tiempo medio entre fallos, etc.

Para la realización de esta prueba se ha elegido el estudio del diámetro de las ruedas ferroviarias. El motivo de su elección es el siguiente:

- Es un caso relativamente sencillo de realizar con métodos estadísticos, lo cual nos permite la comprobación de los resultados.
- Tiene una componente de desgaste que resulta interesante para ver su afección en el algoritmo de aprendizaje.
- El valor que debemos predecir es un valor discreto (días de vida útil)

El caso de uso, tiene como objetivo verificar si sería posible entrenar un algoritmo que nos prediga cuál es la vida útil de las ruedas de un eje (medido en días) en función de su diámetro actual.

### *3.3.1 Datos de partida*

Para la realización de la prueba se dispone de datos de mediciones de diámetros de ruedas correspondientes a una *Serie* de vehículos ferroviarios. Se dispone de la información desde el año 2002 hasta la actualidad. Los datos han sido anonimizados para mantener la confidencialidad de la información.

El origen de los datos disponible se agrupa en dos ficheros. El primer conjunto de datos dispone de mediciones de los diámetros de las ruedas en diferentes fechas. Este conjunto de datos dispone de los siguientes datos:

| Campo | Descripción |
|---|---|
| Fecha Medición | Fecha en la que se realiza la medición del diámetro a las ruedas |
| Unidad | Identificador de la unidad. La unidad tiene la siguiente forma: IdCoche1/IdCoche2/IdCoche3/… Por lo que gracias a la unidad se puede determinar la posición que ocupa un coche dentro de la composición. |
| Coche | Identificador del coche al que se realiza la medición. |
| Eje | Eje sobre el que se realiza la medición. |
| Diámetro | Diámetro mínimo en torno |

*Tabla 3-4: Datos de partida (diámetros) – Caso de uso 2*

Por otro lado, disponemos de un conjunto de mediciones de kilometración de las diferentes unidades. La estructura que dispone de la información es la siguiente:

| Campo | Descripción |
|---|---|
| Fecha | Fecha del registro de kilometración. |
| Unidad | Identificador de la unidad. La unidad tiene la siguiente forma: IdCoche1/IdCoche2/IdCoche3/… Por lo que gracias a la unidad se puede determinar la posición que ocupa un coche dentro de la composición. |
| Medida (Km) | Valor de los kilómetros medidos. |

*Tabla 3-5: Datos de partida (kilómetros) – Caso de uso 2*

### *3.3.2 Ejecución de la prueba*

Como en el caso anterior, el primer paso para poder realizar la prueba es la preparación de los datos de aprendizaje. Para ello, con base en el objetivo que queremos conseguir, que es la predicción de la vida útil de las ruedas de un determinado eje, se ha realizado un estudio de la información que se considera relevante para la determinación de dicha vida útil. El cálculo de la vida útil se ha realizado basándonos en días de funcionamiento, pero se podría realizar de igual manera en base a los kilómetros recorridos.

Las características identificadas han sido las siguientes:

| Característica | Descripción |
|---|---|
| Unidad | Identificador de la unidad. |
| Coche | Identificador del coche al que se realiza la medición. |
| Posición | Posición que ocupa el coche dentro de la composición. |
| Eje | Eje sobre el que se realiza la medición. |

| | |
|---|---|
| Diámetro | Diámetro mínimo en torno |
| Días | Días transcurridos desde la última medición |

*Tabla 3-6: Características principales – Caso de uso 2*

Por otro lado, además de las características que puede afectar al cálculo de la vida útil, será también necesario aportar la vida útil real de dichos ejemplos para que el algoritmo pueda estudiar la afección que tiene dichas características sobre la vida útil de esos elementos. Nuestro objetivo es comprobar si el algoritmo es capaz de identificar las relaciones que pueden existir entre los componentes que caracterizan la vida útil de las ruedas de un eje, para posteriormente poder predecir la vida útil de otras ruedas.

Para el cálculo de la vida útil de las ruedas que disponemos como ejemplo se han seguido los siguientes pasos:

- Ordenación de los datos por unidad, coche, eje, año y mes. De esta forma obtenemos la serie temporal de medidas de un mismo eje, ordenadas en el tiempo.
- Cálculo de los días transcurridos entre una medida y la anterior. De esta forma obtenemos los días transcurridos entre medidas.
- Eliminación de ruido de la muestra de los datos de entrenamiento. Eliminamos aquellos valores que pudieran resultar extraños por motivos de imprecisión en los cálculos intermedios.
- Cuando una rueda registra una medida de 876mm, ser considera que ha sido cambiado el eje.
- Cálculo de la vida útil teniendo en cuenta desde la primera medición hasta la última y la diferencia de diámetro derivada en ese tiempo.

Una vez realizadas las diferentes operaciones, disponemos de un conjunto de datos de 7.498 registros de información que tiene la siguiente forma:

| Unidad | Coche | Pos | Eje | Anio | Mes | Diametro | Diff. Días | da Util (día |
|---|---|---|---|---|---|---|---|---|
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2003 | 12 | 871 | 306 | 3834 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2004 | 5 | 869 | 152 | 3988 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2004 | 12 | 860 | 214 | 3926 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2005 | 10 | 848 | 304 | 3836 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2006 | 8 | 830 | 304 | 3836 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2007 | 11 | 830 | 426 | 3714 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2008 | 5 | 824 | 183 | 3957 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2010 | 1 | 810 | 610 | 3530 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2012 | 1 | 813 | 525 | 3615 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2013 | 2 | 807 | 151 | 3989 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2014 | 1 | 806 | 184 | 3956 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 1 | 2016 | 5 | 872 | 683 | 3457 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 2 | 2003 | 12 | 871 | 306 | 3834 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 2 | 2004 | 5 | 869 | 152 | 3988 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 2 | 2004 | 12 | 860 | 214 | 3926 |
| Coche01/Coche02/Coche03/Coche04/Coche04/Coche06 | Coche01 | 1 | 2 | 2005 | 10 | 850 | 304 | 3836 |

*Tabla 3-7: Muestra de datos de partida – Caso de uso 2*

Para la ejecución de la prueba, a diferencia del caso de uso anterior, se va a crear un escenario que lance en paralelo la ejecución de modelos de entrenamiento con los siguientes algoritmos:

- Decision Forest Regression.
- Boosted Decision Tree Regression.
- Poisson Regression.
- Neural Network Regression.

Una vez entrenados los cuadros modelos se evaluarán por separado y posteriormente se compararán los cuatros entre sí.

El esquema del escenario es el siguiente:

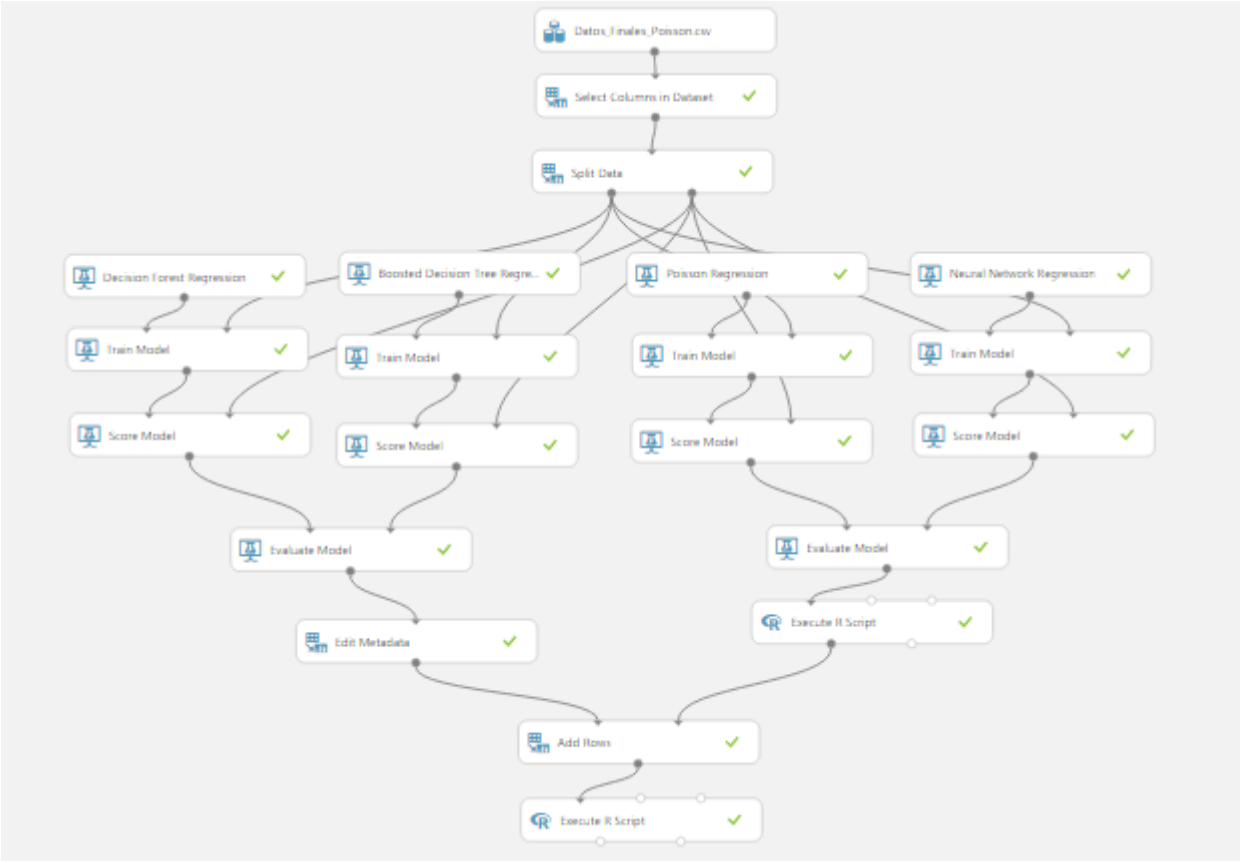



*Figura 3-7: Esquema comparativa algoritmos de regresión*

Tras la ejecución de la prueba, el último paso realiza la evaluación de los cuatro modelos, con el siguiente resultado:

| Algorithms | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| Decision Forest Regression | 144,600071 | 244,287397 | 0.18825 | 0.072215 | 0.927785 |
| **Boosted Decision Tree Regression** | **106,870073** | **152,775972** | **0.139131** | **0.028245** | **0.971755** |
| Poisson Regression | 447,99608 | 593,51874 | 0.583231 | 0.42628 | 0.57372 |
| Neural Network Regression | 785,953544 | 918,449966 | 1023207 | 1020791 | -0.020791 |

*Tabla 3-8: Resultado comparativa de algoritmos de regresión*

Como podemos ver en la tabla resumen, el algoritmo que obtiene mejores resultados ha sido un algoritmo basado en árboles de decisión denominado, Boosted Decission Tree Regression.

Los resultados, junto con la variación existente entre la realidad y los datos estimados son los siguientes:

| Posición del coche en la composición | Número de Eje | Diámetro | Días desde última medición | Vida útil restante REAL (días) | Vida útil restante PREDICCIÓN (días) | Diferencia (días) | Variación |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 814 | 525 | 3615 | 3606,95 | -8,05 | -0,22% |
| 1 | 3 | 813 | 525 | 3615 | 3606,95 | -8,05 | -0,22% |
| 5 | 2 | 866 | 243 | 2611 | 2671,51 | 60,51 | 2,32% |
| 2 | 1 | 867 | 61 | 3653 | 3501,66 | -151,34 | -4,14% |
| 4 | 2 | 823 | 426 | 3288 | 3424,29 | 136,29 | 4,14% |
| 5 | 2 | 849 | 608 | 3076 | 3055,75 | -20,25 | -0,66% |
| 6 | 3 | 871 | 153 | 4322 | 4252,16 | -69,84 | -1,62% |
| 1 | 3 | 815 | 486 | 3869 | 3722,77 | -146,23 | -3,78% |
| 3 | 4 | 857 | 365 | 3990 | 4087,58 | 97,58 | 2,45% |
| 5 | 2 | 820 | 335 | 4020 | 3941,01 | -78,99 | -1,96% |
| 6 | 3 | 863 | 579 | 1462 | 1590,11 | 128,11 | 8,76% |
| 1 | 4 | 845 | 366 | 2283 | 2467,12 | 184,12 | 8,06% |
| 2 | 1 | 829 | 486 | 4142 | 4118,83 | -23,17 | -0,56% |
| 5 | 2 | 849 | 883 | 1766 | 2010,01 | 244,01 | 13,82% |
| 3 | 1 | 852 | 306 | 3074 | 3217,56 | 143,56 | 4,67% |
| 4 | 3 | 807 | 395 | 3684 | 3736,93 | 52,93 | 1,44% |
| 5 | 4 | 867 | 365 | 4415 | 4504,71 | 89,71 | 2,03% |
| 3 | 1 | 859 | 457 | 2585 | 2521,71 | -63,29 | -2,45% |
| 5 | 4 | 818 | 123 | 2919 | 2967,42 | 48,42 | 1,66% |
| 2 | 3 | 806 | 335 | 3530 | 4087,78 | 557,78 | 15,80% |
| 1 | 4 | 859 | 184 | 3072 | 3134,48 | 62,48 | 2,03% |
| 2 | 2 | 805 | 30 | 3226 | 3273,85 | 47,85 | 1,48% |
| 3 | 3 | 854 | 184 | 3072 | 3033,66 | -38,34 | -1,25% |

*Tabla 3-9: Comparativa de predicción y valores reales*

Para poder analizar la tabla anterior, es necesario observar las columnas marcadas en gris claro, donde se muestran la vida útil restante tanto real como estimada. Si, por ejemplo, nos fijamos en la primera fila observamos que la vida útil real de ese eje fue de 3615 días y la estimación ha sido de 3606,95 días. Esto supone una diferencia real/estimación absoluta de 0,22%.

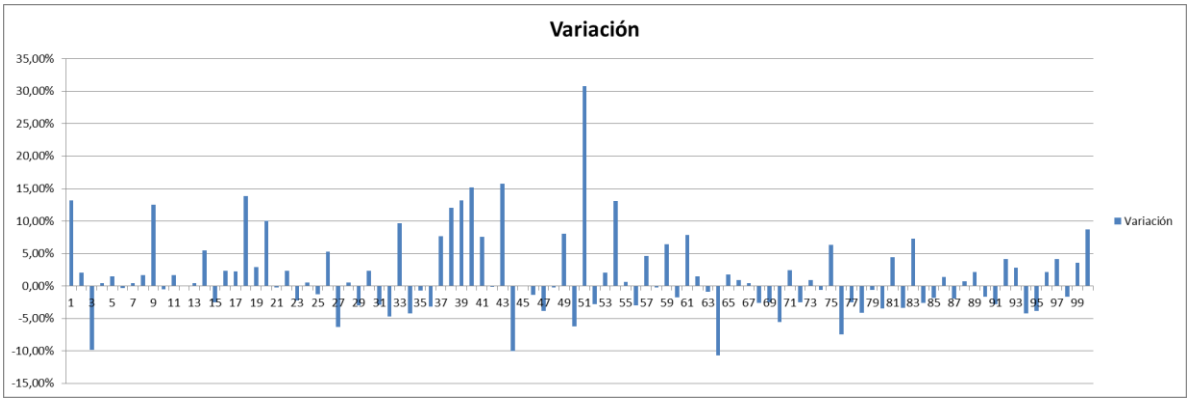



*Figura 3-8: Gráfica de porcentaje de variación con valor objetivo real*

Como podemos observar en la gráfica, salvo algún caso puntual, la mayoría de los resultados reflejan predicciones con una desviación mejor al 10%.

### *3.3.3 Conclusiones del caso de uso*

Con la experiencia recogida, tanto de este como de otros modelos intermedios, realizados con el objetivo de predecir valores discretos mediante algoritmos de regresión, se han extraído las siguientes conclusiones:

- Como podemos observar en los resultados del caso de uso, el algoritmo de regresión basado en árboles de decisión ha obtenido unos resultados bastante óptimos, donde salvo casos puntuales, se han obtenidos predicciones por debajo del 10% de variación con respecto al resultado esperado.
- La preparación de los datos es un paso fundamental en el proceso y la precisión final del modelo tiene una dependencia directa de esta fase.
- Los algoritmos de regresión tienen su máximo rendimiento en aquellos casos en los que disponemos de datos que representen múltiples características (variables), como podría ser el caso de diferentes sensores en un mismo motor. De esta forma, con el histórico adecuado, se podría entrenar un modelo para la predicción de Tasa de fallos, Tiempo medio entre fallos, etc.
- Es tedioso realizar estudios para intentar determinar qué algoritmo podrá funcionar bien para los datos que disponemos y nuestro objetivo, pero es sencillo lanzar la ejecución de 4 modelos de aprendizaje en paralelo para poder comparar sus resultados. Esto es debido, en parte, a la existencia de plataformas que permiten centrarse en la preparación de los datos y no en la programación de los modelos.
- Gracias a los modelos podemos ver cómo afecta el cambio de las características a la búsqueda del valor objetivo con un simple cambio de selección de columnas.

# 4 CONCLUSIONES DEL PROYECTO

Las técnicas de aprendizaje automático requieren de una gran cantidad de datos de histórico que nos permitan entrenar los modelos. Como se ha indicado en el apartado de aplicaciones en el sector ferroviario, la mayor parte de los elementos de infraestructura tienen ciclos de vida muy extensos. Por otro lado, la información de histórico disponible en la actualidad para este tipo de elementos, como puede ser la vía, no presenta la estructuración, periodicidad e interrelación con otros tipos de información necesaria. Esto impide aprovechar el potencial que este tipo de técnicas pueden aportar en campos como el mantenimiento predictivo.

Por el contrario, aquellos elementos con ciclos de vida más cortos, propensos a sufrir desgastes o disfunciones con el uso, son más adecuados para el uso de técnicas de aprendizaje automático. Esto es debido a varios aspectos:

- Al ser propensos al desgaste o fallo transcurridos una serie de ciclos, ha sido necesario realizar mediciones periódicas para conocer su estado en el mantenimiento según estado.
- En su mayoría son más fácilmente medibles. Pensemos por ejemplo en elementos del material rodante, donde ya se recopila el estado de gran cantidad de componentes.
- Al ser ciclos de vida más cortos, se dispone de más cantidad de información.

Con base en los puntos anteriormente indicados, una de las principales conclusión del presente trabajo es la necesidad de rediseñar las técnicas de recopilación y tratamiento de la información, especialmente en los elementos de infraestructura, y pensar en la información que necesitamos recoger en la actualidad para que pueda ser explotada en el futuro. Se debe construir un almacén de información que nos permita aprender de los comportamientos de aquellos elementos que debemos mantener.

Con respecto al uso de los diferentes algoritmos, en los casos de usos realizado, los algoritmos basados en árboles de decisión hay presentado un mayor grado de precisión, tanto para tareas de clasificación como de regresión.

En lo referente a los tipos de uso de estas técnicas, se han identificado dos formas de uso:

- Predicción en tiempo de ejecución: Similar a la ejecución automática de un sistema.
- Predicción por lotes: Similar al trabajo posterior de los datos que se realiza en gabinete.

Como conclusión final podemos indicar que el uso de las técnicas de aprendizaje automático es viable. Por otro lado, se debe poner mayor énfasis en la monitorización de infraestructuras y en la conservación de la información de histórico para su posterior aprovechamiento.

# 5 APORTACIONES

El presente proyecto pretende ser una primera aproximación a las técnicas de aprendizaje automático. Por este motivo, las aportaciones que persigue el proyecto son las siguientes:

- Presentar las técnicas de aprendizaje automático como un nuevo enfoque en las técnicas de análisis de la información en el ámbito de los sistemas ferroviarios, con el objetivo final de convertir la información en conocimiento. Este conocimiento puede ser empleado para la identificación de nuevas relaciones entre eventos, nuevas patologías, predicciones, etc.
- Recopilación de la información principal para conocer el fundamento teórico de estas técnicas.
- Disponer de ejemplos prácticos de utilización de algoritmos de clasificación y de regresión.
- Disponer de una comparativa de diferentes algoritmos trabajando con el mismo objetivo para poder comparar su resultado.
- Disponer de una orientación en cuanto a ámbitos de utilización de estas técnicas dentro del sector ferroviario.
- Disponer de unas conclusiones referentes al uso de este tipo de técnicas.

# 6 TERMINOLOGÍA

**Característica**: También llamado atributo o campo. Son las propiedades que describen cada una de las instancias del conjunto de datos. En el caso de uso de auscultador dinámico sería cada una de las medidas (aceleración vertical, horizontal, etc.) que proporciona la auscultación.

**Instancia**: También llamado registro, es cada uno de los datos de los que se disponen para hacer un análisis. Correspondería a cada fila de nuestro conjunto de datos.

**Conjunto de datos**: Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades.

**Confianza**: Es la probabilidad de acierto que calcula el sistema para cada una de las predicciones.

**Experimento**: Es como se identifica un procedimiento de aprendizaje automático dentro de la plataforma Azure de Microsoft.

# 7 BIBLIOGRAFÍA

**Recursos Web**

- Microsoft Azure (Abril 2016)

  https://msdn.microsoft.com/es-es/library/azure/dn905812.aspx

- Base de Conocimiento de IBM

  http://www.ibm.com/support/knowledgecenter/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/advanced/idh_plum.htm?lang=es

- ***Mónica Martínez Gómez, Manuel Marí Benlloch***. Universidad Politécnica de Valencia. Distribución de Poisson.

  https://riunet.upv.es/bitstream/handle/10251/7937/Distribucion%20Poisson.pdf

**Libros**

- ***Gonzalo Pajares Martinsanz, Jesús Manuel de la Cruz (Coordinadores) (Ra-Ma),*** Aprendizaje Automático, Un enfoque práctico.