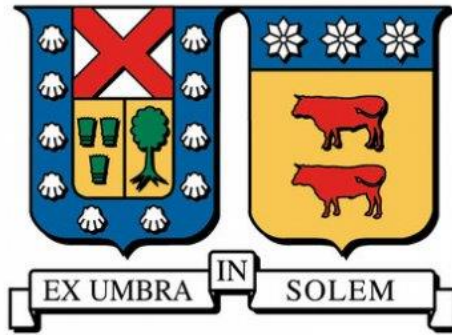




UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS

EVALUACIÓN DEL USO DE LA METODOLOGÍA CART EN EL
DIAGNÓSTICO DE FALLAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

AUTOR

RODRIGO RIFFO

PROFESOR GUÍA

MONICA LOPEZ C.

PROFESOR CORREFERENTE

FREDY KRISTJANPOLLER R.

VALPARAÍSO DE CHILE, 26 DE SEPTIEMBRE, 2018.

RESUMEN EJECUTIVO

En un contexto de mercados competitivos, surge la necesidad de usar de forma cada vez más eficiente los recursos, por lo tanto, el mantenimiento de las máquinas o equipos se debe hacer en el momento indicado. Es por esto por lo que en el presente documento se pretende evaluar el uso de los árboles de regresión y clasificación en el mantenimiento.

El mantenimiento busca mantener en correcto estado los equipos el mayor tiempo posible, pero también se pueden agregar otros beneficios como reducir costos, aumentar la vida útil o evitar accidentes. Existen varios tipos de mantenimiento como el correctivo, preventivo, predictivo y el mantenimiento productivo total, que se usan en distintas situaciones. Otro elemento importante que las compañías consideran es la demostración de su calidad, a través de normas internacionales, para las cuales es importante tener un apropiado plan de mantenimiento.

Con el contexto claro, se puede pasar a explicar la herramienta como tal, la cual fue planteada inicialmente en la publicación Classification and Regression Trees de Breiman, Freidman, Olshen y Stone en el año 1984. El árbol tiene una estructura de nodos, los cuales se particionan y generan 2 nodos hijos cada uno cada vez más homogéneos. Cada nodo en el árbol de regresión corresponde a una pregunta o un criterio de división de los datos, por lo tanto, si desea clasificar o determinar el comportamiento de una variable, se debe avanzar por el árbol desde el nodo raíz (primer nodo) hasta que se llegue a la clasificación final.

La metodología como tal consta de 3 pasos que son, la construcción del árbol, la selección del tamaño correcto y la clasificación de datos nuevos. Un dato que tomar en cuenta

es que las variables pueden ser categóricas o continuas, por lo cual la herramienta puede trabajar con distintos tipos de datos.

Para trabajar con los árboles de regresión se puede utilizar una programación de este en el programa Python que se encuentra en los tutoriales de Machine Learning de Josh Gordon publicados el 13 de septiembre del año 2017. Con esta información se puede pasar al desarrollo de esta memoria, donde se busca determinar algunos usos de la metodología en el área de mantenimiento con ayuda de la información recopilada, para posteriormente indicar como podría ser usada la herramienta en esas aplicaciones, mostrando las estructuras necesarias para esto. Finalmente se utilizará una base de datos para probar lo anteriormente planteado.

Con la información recopilada y los análisis respectivos se llegó a que existen 3 usos posibles para la herramienta que son la detección de fallas, determinar la causa de las fallas y la predicción de fallas. Con la base de datos se pretende trabajar con los usos sobre la determinación de la causa de la falla y la predicción de la falla, pero se deja de lado la detección de la falla, ya que se considera que la base de datos no aplica para este caso.

Para poder trabajar, se crearon 2 formas de estructurar la base de datos, siendo la primera una estructura que busca determinar el tiempo que permanecerá sin fallas una máquina y la segunda busca determinar si una máquina va a fallar en un tiempo determinado, pero complementario a esto se harán análisis de cada una de las variables que tiene la base de datos, para ver cómo se relacionan con lo que se busca obtener. De lo anterior, se puede destacar las formas de estructurar la base de datos para utilizar la herramienta, pero es importante saber si las estructuras planteadas rinden frutos. Un elemento importante que se concluyó es el saber que para el caso particular de la estructura que busca saber si una

máquina fallará en un tiempo determinado, los resultados no se deben juzgar por cualquier indicador, ya que se está trabajando con bases de datos desbalanceadas.

De lo ejecutado, también se destaca la importancia de hacer los análisis de las distintas variables clasificadoras de forma independiente, ya que no necesariamente más variables implica una mejor predicción, por lo que es importante discriminar correctamente que usar y que no.

Respecto al trabajo de las impurezas y su relación con los nodos, se puede decir que permitir más impureza es una opción para reducir los nodos finales, pero hay que ser cuidadosos con la reducción de los porcentajes de predicción, ya que es posible que al permitir más impureza dos nodos se mezclen y se pierda un tipo de respuesta. Esto lleva también a la necesidad de evaluar los porcentajes de predicción de cada una de las respuestas y en el caso de variables continuas se podrían agrupar las respuestas y evaluar el comportamiento de los grupos.

Respecto a los resultados de las estructuras, en el caso de si va a fallar una máquina en un periodo de tiempo fijo, se percibieron buenos porcentajes de predicción, superiores al 90%, pero solo cuando era un desfase de 1 día, ya que si se aumentaba el desfase los resultados empeoraban significativamente. De igual manera en el caso del tiempo sin fallas, los resultados no fueron buenos, pero no se puede culpar a la herramienta, sino más bien esto puede motivar a buscar otras formas de tomar y usar los datos.

Finalmente se puede decir que la herramienta para analizar las causas de una falla es bastante útil, ya que se presenta de una forma gráfica y fácil de entender, lo que permite

seguir el camino de preguntas y respuestas y analizar las causas para poder contemplar posibles soluciones.

TABLA DE CONTENIDO

1. PROBLEMA DE INVESTIGACIÓN	15
2. OBJETIVOS	18
1.1.1. Objetivo general	18
1.1.2. Objetivos específicos.....	18
3. MARCO TEÓRICO	19
3.1. Mantenimiento	19
3.2. Árboles de regresión o clasificación	23
3.2.1. Construcción del árbol máximo.....	25
3.2.2. Escoger el tamaño de árbol correcto.....	26
3.3. Árboles de regresión aplicados en el mantenimiento	28
3.4. Programación de los árboles de regresión	30
3.5. Formas de medir el desempeño del modelo.....	34
3.5.1. Likelihood ratios	35
3.5.2. Discriminant power (DP).....	36
3.5.3. Kappa (K)	36
3.5.4. Clasificador Kappa para más de 2 clasificaciones.....	37
4. METODOLOGÍA.....	40
5. DESARROLLO.....	42
5.1. Problema y forma de uso de la metodología CART	44
5.1.1. Detección de fallas.....	44
5.1.2. Causas de una falla	45
5.1.3. Predicción de fallas.....	46
5.2. Base de datos.....	48
5.2.2. Explicación de la base de datos	48
5.3. Uso de los datos	55
5.4. Estructuración de la base de datos	56
5.4.1. Tiempo sin fallas	56
5.4.2. Si fallará una máquina	59

5.5.	Análisis de la herramienta.....	61
5.5.1.	Análisis de los errores.....	61
5.5.2.	Análisis de las mediciones.....	64
5.5.3.	Análisis de los días desde una mantención o falla.....	66
5.5.4.	Análisis de estructura si fallará una máquina con 1 día de desfase	69
5.5.5.	Análisis de estructura si fallará una máquina con 2 días de desfase	85
5.5.6.	Análisis de la estructura de tiempo sin fallas	91
6.	CONCLUSION	96
7.	REFERENCIAS	99

ÍNDICE DE ILUSTRACIONES

Ilustración 1: Relación entre nodos en la metodología CART (Elaboración propia).....	23
Ilustración 2: Árbol de regresión referente al infarto al miocardio (Trujillano, y otros, 2008)	24
Ilustración 3: Fragmento de árbol construido 1 (Elaboración propia).....	32
Ilustración 4: Fragmento del árbol construido 2 (Elaboración propia)	32
Ilustración 5: Testing_data (Elaboración propia)	33
Ilustración 6: Predicción (Elaboración propia).....	33
Ilustración 7: Mediciones básicas basadas en la matriz de confusión (Bekkar, Djemaa, & Alitouche, 2013)	34
Ilustración 8: Lijelihood positivo (Bekkar, Djemaa, & Alitouche, 2013).....	35
Ilustración 9: Likelihood negativo (Bekkar, Djemaa, & Alitouche, 2013)	35
Ilustración 10 : Interpretación del Llikelihood positivo (Bekkar, Djemaa, & Alitouche, 2013)	36
Ilustración 11: Discriminant power (Bekkar, Djemaa, & Alitouche, 2013).....	36
Ilustración 12: Clasificación del Discriminant power (Bekkar, Djemaa, & Alitouche, 2013)	36
Ilustración 13: Indicador Kappa (Akosa, 2017)	36
Ilustración 14: Total accuracy (Akosa, 2017)	37
Ilustración 15: Random accuracy (Akosa, 2017)	37
Ilustración 16: Variable x_{ij} (Green, 1997)	38
Ilustración 17: Proporción de respuestas (Green, 1997).....	38
Ilustración 18: Valor de Kappa por categoría (Green, 1997)	38

Ilustración 19: Valor de Kappa (Green, 1997)	39
Ilustración 20: Diagrama de la metodología (Elaboración propia)	41
Ilustración 21: Diagrama del desarrollo (Elaboración propia)	43
Ilustración 22: Nodo resaltado (Elaboración propia)	69
Ilustración 23: Árbol de regresión y clasificación 10.000 datos parte 1 (Elaboración propia)	70
Ilustración 24: Árbol de regresión y clasificación 10.000 datos parte 2 (Elaboración propia)	71
Ilustración 25: Árbol de regresión y clasificación 10.000 datos parte 3 (Elaboración propia)	72
Ilustración 26: Árbol de regresión y clasificación 10.000 datos parte 4 (Elaboración propia)	73
Ilustración 27: Árbol de regresión y clasificación 10.000 datos parte 5 (Elaboración propia)	74
Ilustración 28: Árbol de regresión y clasificación 10.000 datos parte 6 (Elaboración propia)	75
Ilustración 29: Árbol de regresión y clasificación 10.000 datos parte 7 (Elaboración propia)	76
Ilustración 30: Nodo del componente 3 en la prueba de 10.000 datos con impureza 0.000 (Elaboración propia).....	81
Ilustración 31: Nodo del componente 3 en la prueba de 10.000 datos con impureza 0.001 (Elaboración propia).....	81
Ilustración 32: Nodo del componente 3 en la prueba de 10.000 datos con impureza 0.002 (Elaboración propia).....	82

Ilustración 33: Árbol de clasificación con 10.000 datos y 0,001 de impureza (Elaboración propia).....	90
Ilustración 34: Nodo con impureza (Elaboración propia)	93

ÍNDICE DE TABLAS

Tabla 1: Matriz de confusión (Elaboración propia).....	34
Tabla 2: Cuadro resumen del uso de la herramienta	47
Tabla 3: Archivo PdM_errors ((kaggle, 2018)).....	49
Tabla 4: Errores producidos (Elaboración propia)	49
Tabla 5: Errores por máquina (Elaboración propia).....	50
Tabla 6: PdM_failures ((kaggle, 2018))	50
Tabla 7: Número de fallas (Elaboración propia)	51
Tabla 8: PdM_machines ((kaggle, 2018))	51
Tabla 9: Fallas por modelo (Elaboración propia).....	52
Tabla 10: Fallas por edad de máquina (Elaboración propia).....	53
Tabla 11: Archivo PdM_maint ((kaggle, 2018)).....	54
Tabla 12: Relación entre fallas y cambios por componente (Elaboración propia).....	54
Tabla 13: Archivo PdM_telemetry ((kaggle, 2018)).....	55
Tabla 14: Tiempo sin fallas del mismo componente (Elaboración propia).....	57
Tabla 15: Estructura del tiempo de falla anterior en la estructura 1 (Elaboración propia)...	58
Tabla 16: Componente que se cambió (Elaboración propia)	58
Tabla 17: Estructura de los Modelos y edades de las máquinas en la estructura 1 (Elaboración propia).....	58
Tabla 18: Estructura de errores en la estructura 1 (Elaboración propia)	59
Tabla 19: Estructura de las mediciones en la estructura 1 (Elaboración propia)	59
Tabla 20: Variable a pronosticar en la estructura 2 (Elaboración propia).....	60
Tabla 21: Tiempos de mantención o falla por componente de la estructura 2 (Elaboración propia).....	61

Tabla 22: Pronostico sólo con errores con desfase de 1 día (Elaboración propia)	63
Tabla 23: Pronostico sólo con errores con desfase de 2 días (Elaboración propia)	63
Tabla 24: Pronostico con sólo las mediciones con desfase de 1 y 2 días (Elaboración propia):	65
Tabla 25: Pronostico con sólo las mediciones con desfase de 2 semanas (Elaboración propia)	65
Tabla 26: Pronostico sólo con los tiempos desde mantención o falla con desfase de 1 día (Elaboración propia)	67
Tabla 27: Pronostico sólo con los tiempos desde mantención o falla con desfase de 2 días (Elaboración propia)	67
Tabla 28: Pronostico sólo con los tiempos desde mantención o falla con desfase de 1 semana (Elaboración propia)	67
Tabla 29: Número de fallas en la base de construcción de 10.000 datos (Elaboración propia)	80
Tabla 30: Tabla de potencia de predicción (Elaboración propia).....	88
Tabla 31: Predicción de la estructura de tiempo sin fallas (Elaboración propia)	91
Tabla 32: Rangos de los días anticipados de mantenimiento (Elaboración propia)	92

ÍNDICE DE GRÁFICOS

Gráfico 1: Índice de Kappa en el análisis de errores (Elaboración propia)	64
Gráfico 2: Índice de Kappa en el análisis de mediciones (Elaboración propia).....	66
Gráfico 3: Grafico del Índice Sensitivity vs Impureza de los nodos con desfase de 1 día (Elaboración propia)	77
Gráfico 4: Grafico del N° de nodos finales vs Impureza de los nodos finales con desfase de 1 día (Elaboración propia)	78
Gráfico 5: Índice de Kappa vs Impureza en el análisis de estructura de si fallará una máquina con 1 día de desfase (Elaboración propia).....	79
Gráfico 6: Grafico del % de ac por componente vs Impureza de los nodos finales con desfase de 1 día (Elaboración propia)	80
Gráfico 7: N° de nodos finales vs Cantidad de datos usados en la construcción con desfase de 1 día (Elaboración propia)	83
Gráfico 8: Grafico del % de ac por componente vs Cantidad de datos usados en la construcción con desfase de 1 día (Elaboración propia)	84
Gráfico 9: Índice de Kappa vs N° de datos en el análisis de estructura de si fallará una máquina con 1 día de desfase (Elaboración propia)	84
Gráfico 10: Porcentaje de clasificación vs N° de datos con 2 días de desfase (Elaboración propia).....	86
Gráfico 11: N° de nodos vs N° de datos con 2 días de desfase (Elaboración propia)	87
Gráfico 12: Índice vs N° de datos en el análisis de estructura de si fallará una máquina con 2 días de desfase (Elaboración propia)	88
Gráfico 13: Tiempo de la predicción (Elaboración propia).....	93

Gráfico 14: N° de datos en el intervalo de los datos Antes vs Impureza de los nodos (Elaboración propia)	94
Gráfico 15: N° de nodos vs impureza (Elaboración propia)	94

1. PROBLEMA DE INVESTIGACIÓN

En la actualidad la competencia y las exigencias de los mercados, han llevado a que las empresas busquen ser cada día más eficientes en sus respectivas áreas, a lo cual se le suma el cumplimiento de las normas internacionales (ISO 9001, ISO14001, OHSAS 18001, entre otras). Es en este contexto, que el mantenimiento cobra una gran importancia, ya que parte relevante de la satisfacción de los requerimientos exigidos hoy en día, pasa por mantener apropiadamente los equipos, herramientas e instalaciones de la compañía (Olarte, Botero, & Cañon, 2010). Ciertamente aunque en estos tiempos este tema no es extraño para las organizaciones, no siempre fue así, ya que “El mantenimiento empieza a adquirir importancia a partir de los años 30 cuando Henry Ford implementó en su empresa un área destinada a las actividades de reparación de los equipos pertenecientes a su sistema de producción” (Cruz, 2017). Con este precedente, queda claro que los empresarios desde esta época poco a poco empezaban a valorar la relevancia del correcto funcionamiento de las maquinarias en las ganancias finales de la compañía.

Ahora bien, queda preguntarse de qué forma concreta el correcto funcionamiento de los equipos influye en las ganancias de una compañía. La lista es larga, ya que una falla en un momento inoportuno trae más de un problema a las empresas, los cuales varían en gravedad en función del rubro, tomando en cuenta que la detención de una o más maquinarias puede producir una detención parcial e incluso total del proceso productivo, lo que conlleva a atrasos en las entregas, multas e incluso la pérdida de contratos. Ahora bien, las consecuencias pueden ir más allá de lo monetario, ya sea por temas de prestigio o por efectos en industrias como las del transporte de pasajeros, donde un error inesperado, puede

desencadenar en un accidente que cobre vidas. Es por las razones ya nombradas, que las fallas inesperadas se convierten en un dolor de cabeza para los directivos, los cuales sienten la necesidad de anticiparse a estos sucesos y conocer la probabilidad de ocurrencia de los mismos. Claro está, los esfuerzos por lograr este objetivo dependerán de la naturaleza del proceso productivo, su importancia y los límites presupuestarios de la compañía, ya que existe la posibilidad de simplemente esperar a que la falla se presente y corregirla (mantenimiento correctivo) o revisar la máquina periódicamente, para así determinar su estado y arreglar lo que este mal (mantenimiento preventivo) (Molina, 2006).

Una de las opciones que surgen es la de pronosticar las fallas (mantenimiento predictivo) si es que el proceso productivo de la compañía justifica este esfuerzo, para así arreglar los equipos antes de que las fallas se presenten. La idea es trabajar en un momento donde la detención temporal no cause estragos. Este tipo de mantenimiento se utiliza, cuando se busca reducir costos en los métodos tradicionales antes nombrados, que son el correctivo y el preventivo (de León, 1998).

El mantenimiento predictivo, puede hacerse, a través de un seguimiento a las variables de interés de los equipos (Monitoreo), para así obtener signos de advertencia que indiquen qué podría fallar en el futuro. Los mantenimientos predictivos más comunes son por análisis de vibraciones, por termografía, análisis de ultrasonido y análisis de aceite (Olarte, Botero, & Cañon, 2010).

Además de hacer un seguimiento, también está la opción de utilizar datos históricos de las variables del equipo, para así poder determinar el tiempo que se tiene antes de que se presente una nueva falla. Esta predicción se logra modelando el comportamiento de las variables, por lo cual surge la pregunta de ¿Cómo es posible y conveniente modelar estas

variables? Una de las opciones a considerar es un modelo que fue propuesto en el año 1984 en la publicación “Classification and Regression Trees” hecha por Breiman, Freidman, Olshen y Stone (Serna Pineda, 2009), los cuales propusieron los árboles de clasificación y regresión o la metodología CART, la cual en términos básicos, busca crear grupos homogéneos de datos a través del particionamiento recursivo. La metodología CART se presenta como una opción interesante a tomar en cuenta ya sea para la clasificación de datos en una primera instancia o para la predicción misma, ya que “Los árboles de regresión CART son una alternativa de regresión que no requiere supuestos sobre los datos por analizar y es un método de fácil interpretación de los resultados” (Díaz & Correa, 2013).

Tomando en cuenta lo antes mencionado, es posible preguntarse, ¿de qué forma este método puede aportar al pronóstico de fallas?, ¿representa un aporte a los métodos ya existentes?, ¿podría servir como un complemento a las formas de mantenimiento actuales?

2. OBJETIVOS

1.1.1. Objetivo general

Evaluar el uso de la metodología de los árboles de regresión en el área de mantenimiento y fallas, mediante análisis y realización de un caso de estudio, para así ampliar el uso de la herramienta en este ámbito.

1.1.2. Objetivos específicos

- Determinar el funcionamiento de la metodología CART y conocer su utilización en distintos ámbitos, a través de la recopilación de información, para así tener una base conceptual.
- Proponer posibles usos de la metodología en el área de mantenimiento y predicción de fallas, para establecer un listado de opciones que se puedan analizar, a través del análisis de la información recopilada.
- Realizar el análisis de la metodología sobre una base de datos, para relacionarlo con alguna de las propuestas planteadas, identificando qué información se puede obtener a través con los árboles de regresión.
- Evaluar los resultados entregados por la metodología CART, con el fin de identificar que tan efectiva es para el caso planteado.

3. MARCO TEÓRICO

En lo que es el marco teórico o el estado del arte, se hablará en un principio sobre el mantenimiento en sí, ya que es el tema sobre el cual se piensa aplicar la herramienta.

Con el uso y la razón ya mostrados, se procederá a mostrar matemáticamente cómo funciona la metodología CART y algunos usos prácticos de la misma. Finalmente se hablará sobre la programación de la herramienta, para que pueda ser usada en la práctica.

3.1. Mantenimiento

En las industrias, se requiere que los instrumentos que se utilizan estén disponibles para su uso el mayor tiempo posible, para así poder ser eficientes en el proceso productivo y cumplir con los estándares de calidad requeridos (Olarte, Botero, & Cañon, 2010). Alineado con este objetivo, el mantenimiento busca primeramente mantener el correcto estado funcional de los equipos e instalaciones, aunque este objetivo inicial puede ser ampliamente sobrepasado por las consecuencias de este principio (León, 1998), esto por los distintos beneficios que puede traer esta práctica, ya sea porque se reducen los costos en la producción, se evitan accidentes o se amplía la vida útil de los equipos, entre otras cosas.

El mantenimiento, también está fuertemente relacionado con el concepto de las fallas, que es cuando el equipo o la maquinaria dejan de entregar o presenta problemas en la entrega del servicio para el cual está diseñada. Esta relación, se da porque el mantenimiento, busca reparar, evitar o reducir el número de fallas que la máquina o equipo sufre a lo largo de su vida útil (Molina, 2006).

A la hora de trabajar, no se puede considerar que todos los equipos se deben mantener de la misma forma, ya que cada equipo tiene características distintas, las cuales producen que una forma de mantener una máquina, sea ineficiente o incorrecta para otra, lo que nos lleva a concluir que en una misma empresa o planta de producción, se pueden utilizar varias formas de mantención. Es así como se identifican 4 tipos de mantenimientos.

3.1.1. Mantenimiento correctivo

En este tipo de mantenimiento, lo que se hace básicamente es mantener cuando ya se presentó la falla, es por esta razón que no se toman medidas previas. Es posible que para el lector esta práctica parezca irresponsable, pero no es así si se utiliza en la situación correcta, ya que se puede presentar que los costos de falla sean bajos, además de otras situaciones, como que sean fallas de equipos auxiliares o de equipos que no están directamente relacionados con la producción (León, 1998). Este tipo de mantenimiento se puede subdividir a su vez en 2 categorías (Molina, 2006):

- Mantenimiento paliativo o de campo (de arreglo): Lo que busca es hacer que la máquina vuelva a funcionar, pero no elimina la fuente que provocó la falla.
- Mantenimiento curativo (de reparación): Lo que busca es eliminar la causa de la falla.

Ciertamente hay que ser cuidadosos a la hora de utilizar este tipo de metodología, ya que al no tener un procedimiento previo, la falla se presenta de forma inesperada, lo cual en una situación crítica puede producir estragos, ya sea por la falta de personal en un momento determinado, o por una ausencia de repuestos, además esta detención inesperada, puede producir la detención de la producción o puede causar algún accidente que afecte al personal (León, 1998).

3.1.2. Mantenimiento preventivo

El mantenimiento preventivo, busca evitar las reparaciones, haciendo revisiones periódicas a los equipos, en donde se ve el estado de los componentes y se hacen los cambios que se estimen pertinentes, para así evitar que la máquina falle en el futuro. La dificultad en este método surge cuando hay que determinar cada cuanto tiempo se debe ejecutar la mantención, ya que si se hace con una alta frecuencia, los costos aumentan mucho y si se dejan pasar periodos muy largos, es posible que se produzcan fallas inesperadas. Es por esta razón que hay que tener cuidado en determinar el equilibrio de los costos (León, 1998).

3.1.3. Mantenimiento predictivo

El objetivo de este mantenimiento es reducir los costos producidos por los dos tipos de mantenimientos, antes mencionados. En términos generales, lo que se busca es conocer el estado de los equipos mediante variables representativas, las cuales son monitoreadas, para así determinar el instante correcto en el cual cambiar un determinado componente. Este tipo de mantención goza de varias ventajas, algunas de ellas son:

- Se puede detectar con antelación las posibles fallas que sufrirá el equipo sin necesidad de detenerlo.
- Se pueden programar los cambios o reparaciones en momentos donde la detención de la máquina no cause un problema en los procesos de la planta.
- Se reducen los tiempos de trabajo en la mantención, ya que se sabe cuál es el elemento que está pronto a fallar, por lo cual se saltan la revisión de las partes que no presentan problemas.
- Se puede programar la compra y recepción de piezas con una mayor precisión, para así tener un inventario más reducido.

- Permite identificar las causas de los fallos, para así poder tomar medidas.

A pesar de los beneficios de este método, existen 2 principales dificultades que causan problemas en su ejecución. La primera de ellas es la determinación de los parámetros de evaluación, ya que en ningún caso existe un parámetro o combinación de parámetros que indique de forma exacta las condiciones de la máquina, por lo que puede darse que, al hacer una mala elección, la predicción se vea afectada, ya sea haciendo un mal diagnóstico de la falla o simplemente no detectándola. La segunda dificultad, es que no es viable monitorear todos los parámetros de forma continua, por lo que se tiende a relajar este punto, seleccionando las variables más importantes, lo que podría llevar a los mismos problemas de la primera dificultad. Sumado a lo anterior, se puede dar que se decida hacer revisiones periódicas de los indicadores, lo que podría permitir que se produzcan fallas inesperadas entre revisiones (León, 1998).

3.1.4. Mantenimiento productivo total

Este tipo de mantenimiento surge y se desarrolla en Japón y busca recoger todos los aspectos que influyen en el uso de las maquinarias, por lo que este método implica a todos los estamentos y niveles de producción. Esto quiere decir que la responsabilidad del mantenimiento no recae únicamente en el departamento de mantenimiento, sino que en la estructura de la empresa (Molina, 2006).

Lo positivo de esta práctica, es que entrega un resultado final más completo, ya que se integra a toda la organización al mantenimiento, junto con que además este concepto está relacionado con la idea de calidad total y mejora continua. Por otro lado, también surgen problemáticas, ya que se requiere un fuerte trabajo con la cultura general de la empresa, la cual debe estar convencida y alineada con el concepto. Además, este mantenimiento, requiere

una inversión no solo en los cambios, sino también en tiempo, ya que se considera a la empresa en general (Molina, 2006).

Como se pudo apreciar, hay varias formas de mantener los equipos o maquinarias que una compañía posee, mostrándose que el uso de cada uno depende de las características del equipo y del proceso. Con lo anterior aclarado, no queda más que adentrarse en la explicación de la metodología CART, con el fin de entender que es y cómo funciona, para así poder posteriormente analizar cómo utilizarla en el contexto del mantenimiento.

3.2. Árboles de regresión o clasificación

La metodología CART fue inicialmente planteada por Breiman, Freidman, Olshen y Stone en su publicación ‘‘Classification and Regression Trees’’ del año 1984 (Breiman, Friedman, Stone, & Olshen, 1984) y causó interés en la comunidad, ya que es un método de fácil implementación e interpretación, que tiene la capacidad de trabajar con variables tanto numéricas como categóricas.

El árbol de regresión o de clasificación tiene una estructura de nodos, los cuales pueden identificarse como nodos padre y nodos hijo y se relacionan de la siguiente manera:

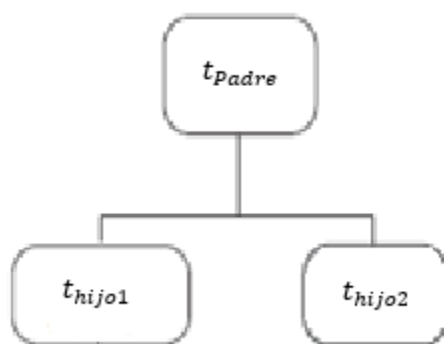


Ilustración 1: Relación entre nodos en la metodología CART (Elaboración propia)

Es de esta forma que se puede decir que producto del particionamiento de un nodo padre, surgen 2 nodos hijo. El método CART hace estas particiones repetidas veces con el fin de encontrar nodos hijos cada vez más homogéneos.

Cada nodo en el árbol de regresión corresponde a una pregunta o un criterio de división de los datos, por lo tanto, si desea clasificar o determinar el comportamiento de una variable, se debe avanzar por el árbol desde el nodo raíz (primer nodo) hasta que se llegue a la clasificación final. Un ejemplo de esto es el caso del siguiente árbol, que analiza la mortalidad hospitalaria del infarto agudo de miocardio (Trujillano, y otros, 2008), en donde para el caso de los hombres la primera pregunta es ¿está en shock o no? y luego para el caso de los que no están en shock, se hace otra pregunta, que es ¿son mayores o menores de 68,5 años?, para de esta forma seguir avanzando y determinar la clasificación final del dato.

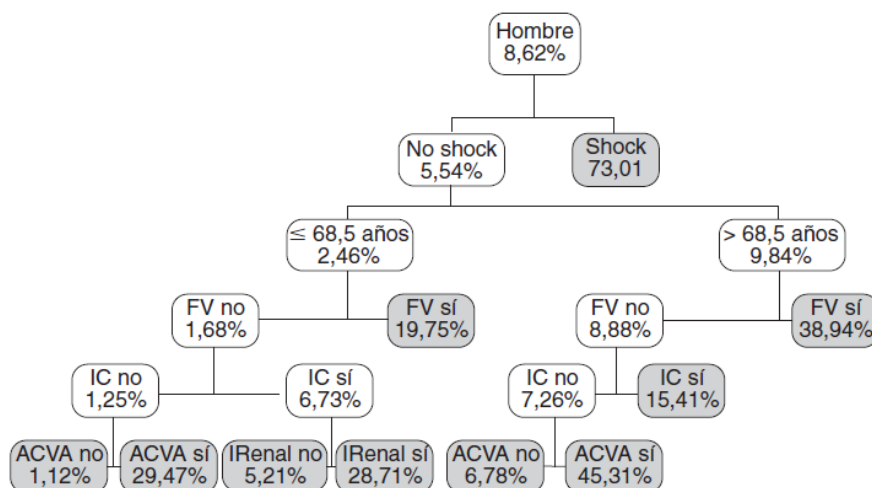


Ilustración 2: Árbol de regresión referente al infarto al miocardio (Trujillano, y otros, 2008)

Una vez entendida a grandes rasgos la estructura de los árboles de regresión es prudente adentrarse en lo que es la metodología de creación de los árboles, la cual consta de 3 partes (Timofeev, 2004):

- 1).- Construcción del árbol máximo.
- 2).- Escoger el tamaño de árbol correcto.
- 3).- Clasificar nuevos datos usando el árbol construido.

Hasta el momento se ha hablado de árboles de regresión y clasificación, como una sola cosa, pero a pesar de que ambos siguen la misma metodología nombrada anteriormente, estos trabajan con datos distintos y por lo tanto presentan diferencias a la hora del cálculo, ya que en el caso de los árboles de clasificación, la variable debe ser categórica, mientras que en los árboles de regresión la variable debe ser continua (Díaz & Correa, 2013). Es por esta razón que se justifica explicar por separado la metodología de construcción de cada tipo de árbol.

3.2.1. Construcción del árbol máximo

La construcción del árbol se produce a través de la partición de cada uno de los nodos en otros 2, pero esta partición debe tener un criterio, al cual llamaremos impureza (Serna Pineda, 2009). El objetivo es que cada vez que se ejecute la partición de un nodo padre, sus nodos hijos sean menos impuros (Trujillano, y otros, 2008), por lo que se va mejorando el pronóstico o la selección de los datos a medida que se desciende del árbol.

En el caso de los árboles de clasificación, existen varias formas de partición propuestas por los autores, en donde la más ampliamente utilizada es la regla de Gini (Timofeev, 2004):

La nomenclatura es la siguiente:

- $i(t)$: Impureza

- t : Nodo.
- $k, l = 1, \dots, K$: Categorías.
- p : Proporción.
- $p(k|t)$: Proporción de la categoría k en el nodo t .
- t_m, t_n : Nodos hijos.
- Y : Variable a clasificar.
- $\bar{Y}$: Promedio de la variable a clasificar.

$$i(t) = \sum_{k \neq l} p(k|t) \cdot p(l|t)$$

Para determinar que los nodos hijos son lo menos impuros posibles, hay que verificar que la variación de la impureza del nodo padre a los nodos hijos, sea lo más grande posible, lo que se logra maximizando la siguiente expresión (Díaz & Correa, 2013):

$$\Delta I(t) = i(t) - p[t_m] \cdot i(t_m) - p[t_n] \cdot i(t_n)$$

Como ya se tienen las funciones de impureza para los arboles de clasificación, falta determinarlas para los arboles de regresión, ya que estos trabajan con variables continuas. En este caso, se puede utilizar la varianza de la respuesta (Díaz & Correa, 2013):

$$i(t) = \sum_{i \in t} (Y_i - \bar{Y}(t))^2$$

Para este caso, el cálculo de la variación de la impureza sufre un ligero cambio, ya que ya no son necesarios los pesos de los nodos hijos (Díaz & Correa, 2013):

$$\Delta I(t) = i(t) - i(t_m) - i(t_n)$$

3.2.2. Escoger el tamaño de árbol correcto

Una vez construido el árbol, lo que se tiene es un árbol saturado, el cual tiene muchas ramas o divisiones, por lo que es necesario reducir su tamaño en un proceso que también es conocido como poda. El objetivo es obtener un árbol con una alta calidad y poco sensible al ruido de los datos (Díaz & Correa, 2013). Para cumplir este objetivo, es necesario crear un índice de calidad, el cual indicará el estado de los nodos terminales, ya que la calidad de un árbol está determinada por qué tan homogéneos son sus nodos terminales.

En este contexto se definen las siguientes medidas para determinar $R(t)$ que es el índice de calidad del árbol:

- Árbol de clasificación:

$$R(t) = \sum_{t \in \bar{t}} P[t] \cdot r(t)$$

- Árbol de regresión:

$$R(t) = \sum_{t \in \bar{t}} i(t)$$

Para lo indicado anteriormente, $\bar{t}$ es el conjunto de nodos terminales y $r(t)$ es una medida de calidad del nodo, que puede ser la función de impureza o también puede ser el costo de mala clasificación, el cual se define como:

$$r(t) = \sum_{i=1}^l c(j|i) \cdot P[Y = i|t]$$

En la ecuación $c(i|j)$ es el costo de mala clasificación de j en la clase i , con $i, j = 1, \dots, l$. Si se clasifica correctamente el costo es 0, por lo tanto $c(i|i)=0$ (Díaz & Correa, 2013).

El objetivo es encontrar el valor de $R(t)$ más pequeño.

3.2.2.1. Costo-complejidad

A la hora de seleccionar el árbol correcto, el tamaño del mismo cobra una gran importancia, tomando en cuenta que un árbol muy grande puede presentar problemas de sobreajuste (Díaz & Correa, 2013). En base a lo anterior se plantea el costo de complejidad, el cual toma en cuenta el costo de mala clasificación y el tamaño del árbol.

$$R_{\alpha}(t) = R(t) + \alpha \cdot |\bar{t}|$$

El parámetro de complejidad corresponde a α y $|\bar{t}|$ corresponde al número de nodos terminales. El parámetro α es un número real mayor o igual a cero, cuando $\alpha = 0$ se tiene el árbol más grande y a medida que α se incrementa, se reduce el tamaño del árbol (Serna Pineda, 2009).

Con la información de los árboles de regresión y clasificación, es posible adentrarse en lo que son algunos ejemplos que la literatura entrega respecto a la utilización de la metodología CART en el mantenimiento.

3.3. Árboles de regresión aplicados en el mantenimiento

Esta metodología ha sido utilizada en distintas áreas, donde se destaca principalmente el área de la salud, pero hace aproximadamente 2 décadas han surgido investigaciones, que buscan probar la efectividad de los árboles de regresión en el mantenimiento. El año 2002, se investigó la capacidad de varios modelos, para predecir la corrosión en los aviones. Para probar los modelos se utilizaron bases de datos de pruebas no destructivas y se llegó a la conclusión de que los árboles de regresión eran particularmente efectivos para predecir la corrosión (Brence & Brown, 2002). En el mismo año, se estableció la idea de hacer un diagnóstico y mantenimiento remoto, en donde uno de los pasos de esta investigación tuvo

como objetivo identificar la causa por la cual la máquina a mantener se descompuso. Es en este punto, donde la metodología CART fue utilizada, ya que las razones de la avería pueden ser varias, como factores ambientales, operadores, componentes u otros y se valoraron algunas características de los árboles de regresión, como el apoyo visual que entrega, que no se ven afectados por valores atípicos, su automatización y su velocidad (Luo, Hsieh, Su, & Tsai, 2002).

Más adelante en el año 2006, surge un estudio que busca desarrollar una herramienta que apoye la toma de decisiones respecto a la inspección de oleoductos de petróleo. Básicamente se busca determinar el tipo o clase de derrame utilizando algunas variables relevantes, para así finalmente decidir dónde y cómo monitorear. Para hacer este análisis se utilizan los árboles de regresión, sobre una base de datos de 30 años aproximadamente, la cual tiene registradas las condiciones de operación de los oleoductos de Europa occidental. El resultado de esta investigación, llevó a que los árboles de regresión funcionaron como una herramienta que orientaba a los gerentes de mantenimiento de petróleo, a planificar de una mejor forma las actividades de mantenimiento, para así mejorar la eficiencia y reducir los riesgos de derrames (Bertolini & Bevilacqua, 2006).

En el año 2009, se investigó la efectividad de los árboles de regresión en la predicción de fallas de un compresor de metano, utilizando una base de datos. La diferencia de esta investigación es que se utiliza una modificación a la metodología CART, conocida como “least square regression tree”, la cual busca en cada nodo establecer una función que relacione los valores a predecir, reemplazando al promedio.

$f(\Theta, x_i)$: Una función que utiliza a las variables clasificadoras.

$$i(t) = \sum_{Y_i, x_i \in t} (Y_i - f(\Theta, x_i))^2$$

Los resultados indicaron que la modificación del modelo fue más efectiva a la hora de predecir las fallas, además se establece que el nuevo modelo tiene la capacidad de rastrear las condiciones del cambio de funcionamiento de la máquina con una alta precisión (Yang, 2009).

En los sistemas de tratamiento de aguas, específicamente en el filtrado, el año 2017 se utilizó la metodología CART, para determinar las condiciones asociadas a un mayor riesgo de filtración ineficiente y se creó un método que apoya las decisiones. En este caso los árboles de regresión demostraron ser una metodología eficaz, entregando modelos precisos (Upton, Jefferson, Moore, & Jarvis, 2017).

Ahora que se conocen algunos ejemplos de usos de los árboles de regresión y clasificación, es importante conocer de qué forma o con que softwares se pueden generar estos árboles.

3.4. Programación de los árboles de regresión

Al buscar programas que generen árboles de regresión, las opciones que aparecen son escasas y por lo general permiten un uso limitado de la herramienta. Una de las opciones que surge es el uso del software estadístico SPSS, el cual tiene como una de sus opciones la construcción de árboles de regresión, pero a pesar de que existe esta herramienta, ahora se hará énfasis en otra forma de crear estos árboles, que es a través de los tutoriales de Machine Learning de Josh Gordon (Gordon, Youtube, 2018). Es aquí donde explica que Machine Learning son un conjunto de algoritmos que aprenden de los ejemplos y la experiencia, poniendo como ejemplo el hecho de diferenciar entre una manzana y una naranja, que es una

tarea difícil de programar si no se usan los métodos ya nombrados, ya que habría que escribir una gran cantidad de códigos para decidir que es una manzana o una naranja y si se presenta otra fruta, habría que escribir una nueva codificación para explicarlo. Lo que se busca en este caso entonces es un algoritmo que genere las reglas, para que no tengan que ser escritas manualmente y esto es precisamente lo que hace la metodología CART (Gordon, Youtube, 2018).

Para programar lo que hace, utiliza el software Python, el cual permite apreciar de mejor manera lo que se está trabajando y a su vez también facilita la modificación de los códigos.

En el capítulo 8 de su tutorial, Josh Gordon habla sobre los árboles de regresión y clasificación o la metodología CART y explica su funcionamiento. En este mismo capítulo, se muestra el código en Python de los árboles, para que pueda ser utilizado con otras bases de datos (Gordon, 2018). El código muestra los pasos que se siguen en la construcción de un árbol de regresión y clasificación, que utiliza el índice de Gini como clasificador.

Cuando se hace correr el programa, se entrega el árbol impreso y también entrega la opción de hacer la predicción de algunos ejemplos.

A pesar de que entrega el árbol impreso, este es difícil de leer si tiene un gran tamaño, ya que el formato de impresión es hacia abajo y no como un árbol normal.

```

--> True:
  Is Past >= 3?
--> True:
  Is Time >= 19?
--> True:
  Is Past >= 4?
--> True:
    Is Time >= 32?
--> True:
(' Predict', {212: 1, 219: 1, 227: 1, 223: 1})
--> False:
(' Predict', {286: 1})

```

Ilustración 3: Fragmento de árbol construido 1 (Elaboración propia)

En la ilustración 6 se muestra la pregunta y el desglose de su respuesta, mostrándose primero los verdaderos y después los falsos, por lo tanto, el falso de la primera pregunta que es “Is Past >= 3” no se muestra sino hasta que terminen todas las preguntas dentro de ese nodo. En el cuadro rojo se muestra un nodo final, que es la respuesta a la pregunta “Is Time>=32?”, mostrándose la respuesta si es verdadero o es falso. Es después del nodo ya mostrado que se muestra el falso de la pregunta “Is Past >=4”.

```

--> True:
  Is Past >= 4?
--> True:
  Is Time >= 32?
--> True:
    Predict', {212: 1, 219: 1, 227: 1, 223: 1})
--> False:
    Predict', {286: 1})
--> False:
  Is Time >= 50?

```

Ilustración 4: Fragmento del árbol construido 2 (Elaboración propia)

En la parte falsa de este nodo, se vuelve a hacer una nueva pregunta que es “Is Time >= 50?”.

Como se indicó anteriormente, también se pueden hacer predicciones, que entregan como resultado los posibles valores que pueden tomar la predicción y el porcentaje de peso que tiene dicho valor sobre el nodo. Mientras más pulcro este el nodo, menores serán las opciones que entrega.

Para hacer lo anterior, hay que ingresar una base de datos de prueba que tenga la misma estructura de la base utilizada para construir el árbol.

```
testing_data = [  
  ['D',0,'102',32,9],  
  ['D',0,'503',24,-3],  
  ['D',0,'3',-24,31],  
  ['D',1,'3',105,156],  
  ['D',2,'3',60,130],  
]
```

Ilustración 5: Testing_data (Elaboración propia)

Cuando corre el programa y se entrega el árbol de regresión, en la parte inferior se entregan las predicciones de la siguiente forma.

```
Actual: 156. Predicted: {73: '50%', -2: '50%'}  
Actual: 130. Predicted: {78: '33%', 118: '33%', 86: '33%'}
```

Ilustración 6: Predicción (Elaboración propia)

Lo que se muestra en la imagen es el valor que debería tener el nodo en “Actual” y luego la predicción. En el caso del primero que es 156, dice que un 50% de los valores del nodo son 73 y el otro 50% es -2. Estos resultados nos indican que para el ejemplo mostrado la predicción no es muy buena, ya que ninguno de los valores se acerca al valor real.

Una vez que se determinó la forma de programar los árboles de regresión y clasificación, es importante saber con qué criterios o indicadores evaluar o comparar los resultados que se obtengan.

3.5. Formas de medir el desempeño del modelo

En Machine Learning el clasificador es evaluado por una matriz de confusión (Bekkar, Djemaa, & Alitouche, 2013). Esta matriz es de 2x2, en donde se muestra la predicción del clasificador y la clase, además si se está trabajando con bases de datos desbalanceadas, la clase minoritaria se etiqueta como la positiva y la mayoritaria como la negativa, dejando la matriz de la siguiente forma.

Tabla 1: Matriz de confusión (Elaboración propia)

	Predicho como positivo	Predicho como negativo
Actual positivo	TP (Número de verdaderos positivos)	FN (Número de falsos negativos)
Actual negativo	FP (Número de falsos positivos)	TN (Número de verdaderos negativos)

En base a la información de la tabla x se pueden determinar algunas mediciones básicas para probar los resultados que se obtengan.

Measure	Formula	interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	
Error rate = 1-Accuracy	$\frac{FP + FN}{TP + TN + FP + FN}$	
Sensitivity (or Recall)	$\frac{TP}{TP + FN}$	Accuracy of positive examples
Specificity	$\frac{TN}{TN + FP}$	Accuracy of Negative examples
Precision	$\frac{TP}{TP + FP}$	

Ilustración 7: Mediciones básicas basadas en la matriz de confusión (Bekkar, Djemaa, & Alitouche, 2013)

El más común de todos es Accuracy, que indica la probabilidad de clasificar bien los valores, mientras que Error rate indica la probabilidad de una mala clasificación. Luego para determinar la exactitud de los valores positivos se utiliza el indicador Sensitivity y para medir la exactitud, pero de los valores negativos se usa el indicador Specificity. Finalmente, Precision indica que tan bien se clasificaron los valores positivos.

Para complementar el análisis, se pueden utilizar indicadores que se componen de combinaciones de los antes expuestos.

3.5.1. Likelihood ratios

Son dos, uno positivo y uno negativo. El positivo corresponde a la relación entre la probabilidad de clasificar un valor positivo como positivo y la probabilidad de clasificar un valor positivo como negativo.

$$L = \rho^+ = \frac{P(\text{Positive/Positive})}{P(\text{Positive/Negative})} = \frac{TP/(TP + FN)}{FP/(FP + TN)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

Ilustración 8: Lijelihood positivo (Bekkar, Djemaa, & Alitouche, 2013)

El negativo corresponde a la relación entre la probabilidad de clasificar a un valor negativo como positivo y la probabilidad de clasificar a un valor negativo como negativo.

$$\lambda = \rho^- = \frac{P(\text{Negative/Positive})}{P(\text{Negative/Negative})} = \frac{FN/(TP + FN)}{TN/(FP + TN)} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

Ilustración 9: Likelihood negativo (Bekkar, Djemaa, & Alitouche, 2013)

Un valor alto en el indicador positivo y un valor bajo en el indicador negativo indican un buen desempeño de los valores positivos y negativos respectivamente. La siguiente tabla muestra la interpretación del indicador positivo.

L. value	Model contribution
1	Negligible
1-5	Poor
5-10	Fair
> 10	Good

Ilustración 10 : Interpretación del Llikelihood positivo (Bekkar, Djemaa, & Alitouche, 2013)

3.5.2. Discriminant power (DP)

Esta medida busca incluir las medidas Sensitivity y Specificity de la siguiente forma:

$$DP = \frac{\sqrt{3}}{\pi} (\log X + \log Y)$$

Where : $X = \text{sensitivity} / (1 - \text{sensitivity})$, and $Y = \text{specificity} / (1 - \text{specificity})$

Ilustración 11: Discriminant power (Bekkar, Djemaa, & Alitouche, 2013)

La siguiente tabla muestra la clasificación de los resultados en función del valor del Discriminant power.

DP value	Model contribution
<1	Poor
>1 and <2	Limited
>2 and <3	Fair
> 3	Good

Ilustración 12: Clasificación del Discriminant power (Bekkar, Djemaa, & Alitouche, 2013)

3.5.3. Kappa (K)

Este indicador toma en cuenta la predicción que se produce solo por casualidad y se conforma de la siguiente forma.

$$kappa = \frac{\text{total accuracy} - \text{random accuracy}}{1 - \text{random accuracy}}$$

Ilustración 13: Indicador Kappa (Akosa, 2017)

Donde

$$total\ accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ilustración 14: Total accuracy (Akosa, 2017)

Y

$$random\ accuracy = \frac{(TN + FP)(TN + FN) + (FN + TP)(FP + TP)}{(TP + TN + FP + FN)^2}$$

Ilustración 15: Random accuracy (Akosa, 2017)

Este indicador tiene valores desde el -1 al 1, donde 0 implica que no hay acuerdo, 1 representa una perfecta concordancia del modelo con lo actual y un -1 representa un desacuerdo entre la predicción y lo actual.

Los indicadores ya expuestos trabajan solo con 2 posibles respuestas o clasificaciones, pero que sucede cuando ese no es el caso. Ahora se presentará una modificación del indicador Kappa para más posibles clasificaciones.

3.5.4. Clasificador Kappa para más de 2 clasificaciones

Lo primero es entender que el clasificador Kappa compara la relación entre 2 respuestas, siendo para este caso el valor real y la clasificación. Los valores que se toman en cuenta son la cantidad de respuestas (m), el número de datos (n) y la cantidad de categorías (k).

Se define una variable x_{ij} como el número de respuestas sobre el dato i ($i = 1, \dots, n$) en una categoría j ($1, \dots, k$) como

$$\sum_{j=1}^k x_{ij} = m$$

Ilustración 16: Variable x_{ij} (Green, 1997)

Para todo i.

También se define el número de respuestas promedio por dato ($\bar{m}$), pero en este caso como hay 2 respuestas para todos los datos, el valor siempre es 2.

Luego se define una proporción de respuestas por categoría ($\bar{p}_j$, j = 1, ... , k)

$$\bar{p}_j = \frac{\sum_{i=1}^n x_{ij}}{n\bar{m}}$$

Ilustración 17: Proporción de respuestas (Green, 1997)

Ahora se define un valor de Kappa por categoría de la siguiente forma

$$\hat{k}_j = 1 - \frac{\sum_{i=1}^n x_{ij}(m - x_{ij})}{nm(m-1)\bar{p}_j\bar{q}_j},$$

Ilustración 18: Valor de Kappa por categoría (Green, 1997)

Donde $\bar{q}_j = 1 - \bar{p}_j$

Finalmente, el valor de Kappa se define como una relación entre los valores de Kappa por categoría

$$\hat{k} = \frac{\sum_{j=1}^k \bar{p}_j \bar{q}_j \hat{k}_j}{\sum_{j=1}^k \bar{p}_j \bar{q}_j}$$

Ilustración 19: Valor de Kappa (Green, 1997)

4. METODOLOGÍA

Para hacer el análisis de la herramienta ya descrita anteriormente en el estado del arte, se procederá a realizar 2 etapas de trabajo, que son en un inicio un trabajo cualitativo en donde se busca determinar de forma general los principales problemas o necesidades con respecto al mantenimiento y un análisis cuantitativo de la herramienta. El análisis cualitativo se hará con ayuda de la información recopilada, utilizando las investigaciones pasadas y el marco teórico del mantenimiento para concluir al respecto.

Una vez que los problemas hayan sido identificados, para cada uno de ellos se buscará una relación con los árboles de regresión, mostrando como podrían ser usados y como deberían estructurarse las bases de datos, para que se pueda obtener la información que se requiere. Es por lo recién dicho, que se debe definir también de forma clara los resultados que se esperan obtener en cada uno de los casos.

Una vez terminado lo anterior, para cumplir con la segunda etapa (Análisis cuantitativo) se utilizará una base de datos que proviene de la página kaggle (kaggle, 2018), que corresponde a un sitio de bases de datos de libre acceso. La base de datos será explicada mostrando las distintas variables que se miden, luego se indicará lo que se espera analizar en la misma y se determinará de qué forma puede ser útil la metodología CART, mostrando que tan eficiente es, probando que tan bien clasifica o predice nuevos datos y como varía la clasificación de estos si la base de datos inicial es más grande o más pequeña. Complementario a lo anterior, se probará cómo funciona la predicción o clasificación si se modifican algunas variables de la misma, como la impureza de los nodos o los tiempos de predicción.

Una vez que se hayan hecho todas las pruebas sobre la base de datos ya nombrada, se hará un análisis de los resultados y se concluirá respecto a la eficiencia de la herramienta y sus posibles usos en el área de mantenimiento.

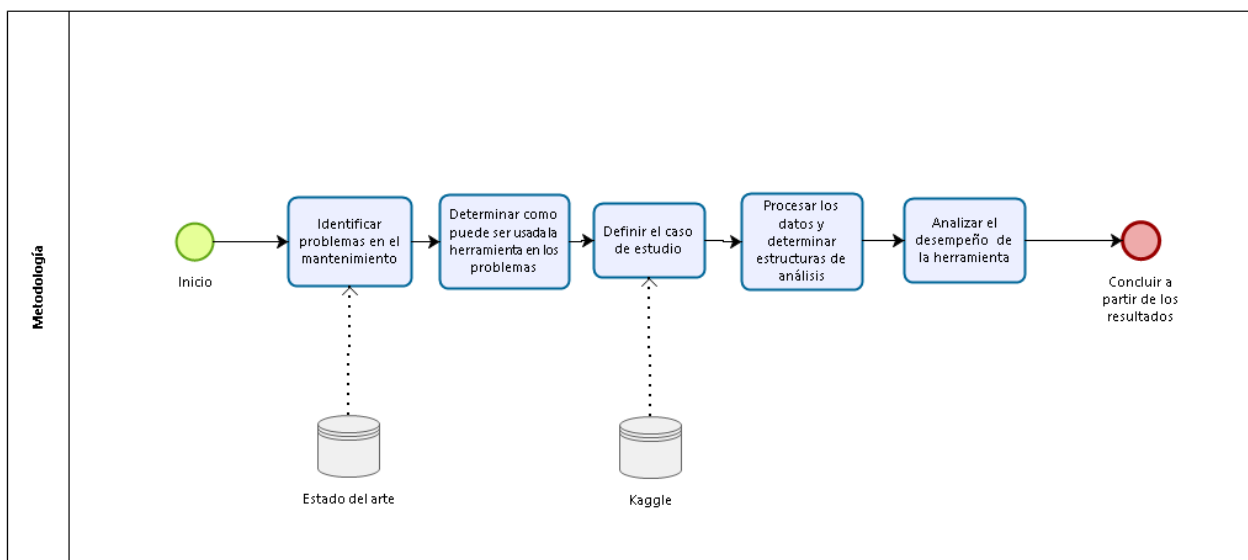


Ilustración 20: Diagrama de la metodología (Elaboración propia)

5. DESARROLLO

Antes de entrar directamente al desarrollo, es prudente para una mejor comprensión hablar en más detalle de lo que se pretende hacer en este capítulo. En un principio se pretende identificar los principales problemas en el mantenimiento y como la metodología CART puede ser usada para dar solución a los mismos. Luego se muestra la base de datos y se explica la información que contiene y como está estructurada.

Para poder trabajar con la información y obtener resultados, se presentan 2 tipos de estructuras que utilizan la información de la base de datos planteada (Si fallará una máquina y tiempo sin fallas). Posterior a eso, se comienza a hacer el análisis de la herramienta, primero haciendo análisis independientes de algunas variables clasificadoras. La evaluación se hará a través del indicador Sensitivity y el indicador Kappa, además de los porcentajes de los componentes de la máquina bien clasificados, para tener un análisis de los componentes por separado. Estos valores serán comparados en tres casos, con 1 día de desfase, 2 días de desfase y 2 semanas de desfase, mostrando los resultados generales en tablas, para cada desfase mostrando el indicador Sensitivity y los porcentajes de componentes bien clasificados además del indicador Accuracy como complemento. Junto a lo anterior se mostrará un gráfico mostrando el índice de Kappa en función de los distintos desfases.

Con el análisis de las variables por separado se puede pasar a el análisis de las estructuras. Para la estructura de si fallará una maquina (Que se explicará más adelante) se considerarán 2 desfases y se mostrará el resultado del árbol de regresión y clasificación de forma gráfica, para apreciar su desempeño como herramienta de análisis. Posteriormente se analizará la relación de los resultados frente a la variación de la impureza permitida,

contrastando este valor con los índices de predicción (Sensitivity, Kappa y porcentajes de los componentes) y el número de nodos finales, para analizar la reducción del tamaño del árbol final. En segundo lugar, se busca analizar, la relación del tamaño de la base de datos de prueba con los resultados ya sea de predicción (A través de los mismos índices anteriores) o de número de nodos finales.

Finalmente, para la estructura del tiempo sin fallas de una maquina el análisis será distinto, ya que se utilizarán clasificaciones para los datos, que mostrarán si la predicción indica que la maquina fallará después de lo que corresponde, antes de lo que corresponde o justo el día correspondiente. De esta información se desglosarán los valores de antes para determinar qué tan antes de la falla real indica la herramienta que se va a producir la falla. Con esta información se analizará la variación en función del grado de impureza y se agregará un análisis de la variación de los nodos finales en función de la impureza.

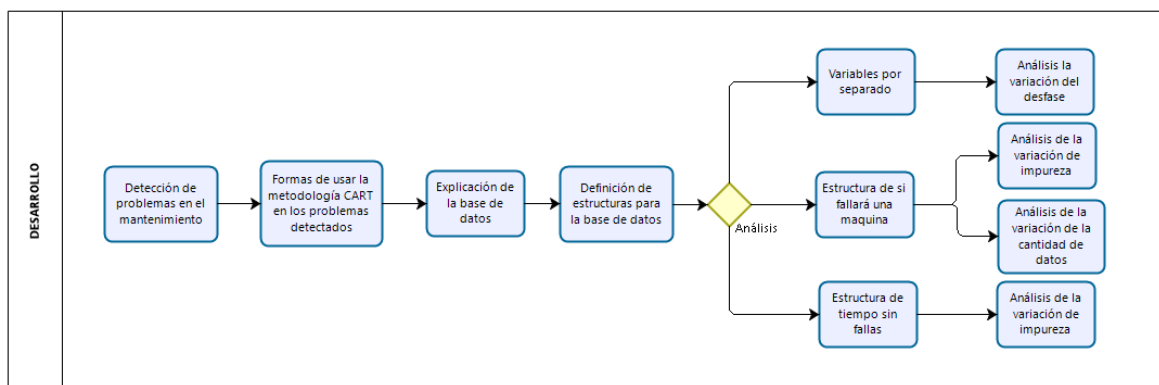


Ilustración 21; Diagrama del desarrollo (Elaboración propia)

5.1. Problema y forma de uso de la metodología CART

Si se analiza de qué forma la metodología CART puede ser usada en el área de mantenimiento y diagnóstico de fallas, es posible describir sus usos de forma general en tres grupos que se detallan a continuación.

5.1.1. Detección de fallas

Un problema que puede surgir en una empresa es la detección de fallas o más específicamente, saber dónde se encuentran las fallas cuando se presentan. Este problema se relaciona mucho con lo planteado respecto a los oleoductos de petróleo (Bertolini & Bevilacqua, 2006), donde la información entregada por la metodología CART les permitió saber dónde monitorear. En forma general, se puede decir que, con la información recopilada, se pueden identificar los lugares en donde el sistema es más propenso a fallar para poder monitorear esa zona evitando gastos excesivos, tomando en cuenta que, si no se tiene esta información, existe la posibilidad de monitorear lugares en donde no se justifique el gasto. Complementario a lo anterior, se puede agregar el ahorro de tiempo a la hora de encontrar el lugar de la falla, ya que, si es que no se está haciendo un monitoreo, también se puede reducir la lista de opciones de lugares.

La solución para el problema planteado suena beneficiosa, pero tiene un gran costo implicado, que es la recopilación de la información, ya que mientras más grande es el sistema o máquina que se desea monitorear, mayor debe ser la base de datos para que los resultados entregados sean precisos. Para el caso de los oleoductos, se requirió una base de datos de aproximadamente 30 años (Bertolini & Bevilacqua, 2006), por lo tanto, para usar este modelo, se requiere superar una brecha de recopilación de datos y temporal.

En el ejemplo ya mencionado se estaba abarcando un problema de gran envergadura, pero esto no necesariamente debe ser así, ya que, si se tienen sistemas o máquinas más pequeñas en grandes cantidades, la recopilación de información se vuelve más sencilla.

El enfoque de este problema es básicamente identificar donde es más probable que falle un sistema, por lo tanto, en ese lugar hay que poner los esfuerzos para prevenir posibles accidentes o pérdidas. Es por lo anterior, que la base de datos que se requiere debe tener al lugar de la falla como la variable a clasificar y a las condiciones de la falla como las variables clasificadoras.

5.1.2. Causas de una falla

Uno de los beneficios que entrega la metodología CART, es la posibilidad de cruzar información para poder encontrar o clasificar a una variable a través de muchas otras. Es esta cualidad la que permite encontrar el conjunto de razones que producen las fallas, para así poder tomar medidas al respecto, modificando las distintas variables que se consideren a la hora del análisis.

Complementario a lo anterior esta metodología, permite incluir variables que comúnmente no se toman en cuenta o son difíciles de agregar al análisis, como pueden ser la experiencia del operador o la temperatura que tiene la zona de trabajo, pero esta posibilidad a su vez produce la necesidad de tener una base de datos bien construida con todas las características que se desean utilizar.

En este caso las opciones son varias, ya que lo que se requiere es saber la magnitud del error o la falla de la máquina, que puede ser la disminución de la velocidad de trabajo, un retraso o un mal funcionamiento, como también podría ser el tiempo de uso de la máquina

sin fallas. Al mismo tiempo, las condiciones de uso pueden ser usadas como las variables clasificadoras, para que así el árbol indique qué combinación de condiciones nos entrega un mejor resultado. Si se desea complicar más el análisis, se podría considerar construir un árbol por tipo de falla, para así poder identificar que combinación de condiciones las produce o también que combinación de condiciones las evita. Otra forma de hacer lo antes descrito, es agregando al tipo de falla como una variable clasificadora.

En conclusión, en este caso el árbol mostraría las características que producen que una determinada máquina esté más tiempo en funcionamiento, por lo tanto, se puede saber qué factores modificar para tener mejores resultados.

5.1.3. Predicción de fallas

Hasta el momento no se ha hablado de predicción de fallas, por lo que ahora se van a tocar dos formas de predecir una falla con los árboles de regresión. La primera es teniendo una estructura similar a la anterior nombrada (Causa de la falla), identificando cuánto tiempo puede funcionar una máquina en determinadas condiciones, pero esta vez, el foco no está en ejecutar modificaciones, sino más bien en determinar cuánto tiempo le queda a la máquina para fallar, en otras palabras, si la máquina tiene ciertas variables clasificadoras, debería mantenerse operativa durante una cierta cantidad de tiempo. Esta información, nos permite hacer mantenciones en los momentos exactos y saber cuándo tener inventario de los repuestos, ya que como se indicó antes, con la base de datos apropiada se puede conocer el tipo de falla.

La segunda opción de predicción es utilizar la información que entregan los métodos predictivos hoy en día, como son por ejemplo el análisis de las vibraciones. Lo que se hace es analizar estos datos con árboles de regresión y determinar cuánto tiempo estará la máquina

operativa antes de fallar. En este caso es importante destacar, que esta información ya es recopilada en la actualidad, por lo que solo quedaría ver qué tan efectivos son los árboles de regresión como herramienta de predicción usando estos datos. Además de lo anterior se le pueden agregar otras variables al análisis, como el modelo de máquina, para así afinar aún más los resultados.

En conclusión, se definen a grandes rasgos 3 formas de utilizar los árboles de regresión en mantenimiento, que son la detección de fallas, la causa de las fallas y la predicción de las fallas. Es importante tomar en cuenta que estos usos no son necesariamente excluyentes entre sí, si se maneja la base de datos apropiada que lo permita. A continuación, se presenta un cuadro resumen que habla de los 3 usos.

Tabla 2: Cuadro resumen del uso de la herramienta

Uso	Descripción	Variable a clasificar	Variables clasificadoras	Observaciones
Detección de fallas	Se busca identificar donde se produce la falla de un sistema	Lugar de la falla	Condiciones de la falla	Ahorro en costo y en tiempo
Causa de la falla	Se busca determinar el factor o conjunto de factores que desembocan en una falla	Magnitud del error o el tiempo sin fallas	Condiciones de uso durante el periodo antes de la falla o durante la falla	Se puede agregar al tipo de falla como variable clasificadora
Predicción de fallas	Se busca pronosticar cuando se va a producir una falla en la máquina	Tiempo de funcionamiento sin fallas	Condiciones en este periodo o monitoreo características (Ej: Vibraciones)	Ayuda en la programación de mantenciones y en los tamaños de los inventarios

5.2. Base de datos

La base de datos se obtuvo de la página kaggle (kaggle, 2018), que contiene una gran cantidad de bases de datos de libre acceso. En este sitio los usuarios pueden usar y a la vez subir bases de datos para complementar más aun la información que ya se encuentra disponible en línea. De esta página, se utiliza la base de datos Xinjian (Kaggle, 2018), que según lo que se indica es para mantenimiento predictivo. Junto con la base de datos como tal, se adjunta un instructivo que la explica, indicando que es usada como una guía para hacer mantenimiento predictivo (Boylu, 2018).

En términos generales se puede decir que la información que se encuentra en la base de datos corresponde a distintas mediciones que se les hicieron a 100 máquinas diferentes durante el año 2015. Dentro de esta información, se encuentran las fallas, las mantenciones, los errores, la descripción de cada una de las máquinas e información obtenida por monitoreo, como la rotación o la vibración de las máquinas. Es importante destacar, que lo que se entrega no incluye información específica sobre el funcionamiento las máquinas, lo que hacen o en que industria trabajan, por lo que los análisis no podrán enfocarse en ese aspecto.

5.2.2. Explicación de la base de datos

La información se entrega en varios archivos csv, los cuales se dividen de la siguiente forma:

- **PdM_errors:** Este archivo contiene los errores de cada máquina a lo largo del año 2015. Un error, corresponde a un problema que presenta la máquina que no impide que siga funcionando. En el archivo se consideran 5 tipos de errores y se identifica a cada error con la fecha (Con la hora), la máquina que lo sufrió y el tipo de error.

Tabla 3: Archivo PdM_errors ((kaggle, 2018))

datetime	machineID	errorID
03-01-2015 7:00	1	error1
03-01-2015 20:00	1	error3
04-01-2015 6:00	1	error5
10-01-2015 15:00	1	error4
22-01-2015 10:00	1	error4
25-01-2015 15:00	1	error4
27-01-2015 4:00	1	error1
03-03-2015 22:00	1	error2
05-03-2015 6:00	1	error1
20-03-2015 18:00	1	error1
26-03-2015 1:00	1	error2
31-03-2015 23:00	1	error1

Si se revisan los datos más a fondo, se pueden apreciar los resultados indicados en la Tabla 3, referentes a la cantidad de cada uno de los errores durante el año 2015.

Tabla 4: Errores producidos (Elaboración propia)

Error	Cantidad
error1	1010
error2	987
error3	837
error4	727
error5	356

En el cuadro anterior, se puede apreciar que los errores más comunes son el 1 y el 2, en cambio el error 5 es menos recurrente, solo llegando a presentarse 356 veces durante el año. Complementario a eso, se puede hacer un conteo de la cantidad de errores en algunas máquinas para apreciar si se mantienen las mismas tendencias.

Tabla 5: Errores por máquina (Elaboración propia)

Maquina	error1	error2	error3	error4	error5
1	11	6	9	5	4
10	17	9	13	4	2
30	12	10	5	7	5
60	10	14	10	8	1
100	11	7	8	4	1

En el cuadro hay algunas máquinas seleccionadas aleatoriamente, en donde se puede apreciar que la estructura general no se cumple estrictamente en todos los casos, ya que en las máquinas 1, 10 y 100 por ejemplo, se produce que el error 3 es mayor al error 2.

- **PdM_failures:** Contiene las fallas que se produjeron a lo largo del año 2015, identificando cada una de ellas con la fecha, la máquina que sufrió la falla y el componente que falló. Una falla produce que la máquina no pueda seguir funcionando.

Tabla 6: PdM_failures ((kaggle, 2018))

datetime	machineID	failure
05-01-2015 6:00	1	comp4
06-03-2015 6:00	1	comp1
20-04-2015 6:00	1	comp2
19-06-2015 6:00	1	comp4
02-09-2015 6:00	1	comp4
17-10-2015 6:00	1	comp2
16-12-2015 6:00	1	comp4
19-03-2015 6:00	2	comp1
19-03-2015 6:00	2	comp2
18-04-2015 6:00	2	comp2
29-12-2015 6:00	2	comp2

Al igual que en el caso de los errores, se puede saber cuántos componentes de cada tipo fallaron durante el año 2015.

Tabla 7: Número de fallas (Elaboración propia)

Componente	Cantidad	
Comp1	192	25,2%
Comp2	259	34,0%
Comp3	131	17,2%
Comp4	179	23,5%

Se puede ver que el componente 2 falla más que el resto, llegando a un 34% de las fallas totales y también que el componente 3 es el que menos falló durante el año.

- **PdM_machines:** Contiene un listado con las máquinas que fueron monitoreadas durante el año 2015, agregándole como información adicional el modelo y la edad de cada una de las máquinas.

Tabla 8: PdM_machines ((kaggle, 2018))

machineID	model	age
1	model3	18
2	model4	7
3	model3	8
4	model3	7
5	model3	2
6	model3	7
7	model3	20
8	model3	16
9	model4	7
10	model3	10
11	model2	6

Para hacer un análisis previo de esta información, se puede hacer un cruce entre los modelos y las fallas, entregando el número de fallas por modelo y componente en el año 2015.

Tabla 9: Fallas por modelo (Elaboración propia)

	Comp 1	Comp 2	Comp 3	Comp 4
model1	30	45	65	41
model2	29	40	63	33
model3	64	88	0	64
model4	60	83	0	38

Del cuadro anterior, se puede apreciar que en el caso de los componentes 1 y 2, los modelos 3 y 4 fallan aproximadamente un 50% más que los modelos 1 y 2, además se puede destacar que en el caso del componente 3, los modelos 3 y 4 no presentan fallas, por lo cual se puede suponer que este componente tiene bajo índice de fallas en este modelo, por lo cual el componente no falla antes de que se le ejecute la mantención correspondiente.

Complementario al análisis anterior, también se puede hacer un cruce de información para conocer el número de fallas por edad de la máquina.

Tabla 10: Fallas por edad de máquina (Elaboración propia)

	Comp 1	Comp 2	Comp 3	Comp 4
1	7	10	0	0
2	13	15	9	0
3	7	11	5	0
4	6	10	3	0
5	8	8	4	0
6	6	12	7	0
7	11	17	4	0
8	1	4	0	0
9	13	11	0	0
10	14	21	17	0
11	5	2	4	0
12	2	1	5	0
13	0	0	0	0
14	24	46	20	50
15	12	14	7	23
16	6	15	11	19
17	15	13	6	23
18	11	15	10	22
19	4	14	3	16
20	17	14	13	23

Del cuadro anterior, lo que más se puede destacar es el comportamiento del componente 4, el cual falla únicamente cuando la máquina tiene más de 14 años.

- **PdM_maint:** En este documento se muestran los mantenimientos que se les han ejecutado a cada una de las máquinas a lo largo del año 2015, pero también se le agregan mantenimientos del año 2014 desde el mes de junio y se agregan los mantenimientos del año 2016 correspondientes al primero de enero. El archivo contiene la fecha, la máquina y el componente de la máquina que fue mantenido. Es importante tomar en cuenta que el archivo de mantención también incluye las fallas.

Tabla 11: Archivo PdM_maint ((kaggle, 2018))

datetime	machineID	comp
01-06-2014 6:00	1	comp2
16-07-2014 6:00	1	comp4
31-07-2014 6:00	1	comp3
13-12-2014 6:00	1	comp1
05-01-2015 6:00	1	comp4
05-01-2015 6:00	1	comp1
20-01-2015 6:00	1	comp3
20-01-2015 6:00	1	comp1
04-02-2015 6:00	1	comp4
04-02-2015 6:00	1	comp3
19-02-2015 6:00	1	comp3

Al analizar los datos, se puede apreciar que el 22,61% de los cambios de componentes son producto de fallas. Si se desglosa por componentes, se aprecian los siguientes resultados.

Tabla 12: Relación entre fallas y cambios por componente (Elaboración propia)

	Cambios	Fallas	
comp1	804	183	22,8%
comp2	863	256	29,7%
comp3	808	128	15,8%
comp4	811	176	21,7%

En la tabla anterior, se puede apreciar que el porcentaje de cambios producidos por fallas en el caso del componente 1 y 4 es muy similar, en cambio en el caso del componente 2, la tasa casi llega a un 30% siendo la más alta de los 4.

- **PdM_telemetry:** Este archivo contiene mediciones de 4 características que se le tomaron a cada una de las máquinas cada 1 hora. Cada medición, tiene una fecha (Con hora), el número de máquina y la medición.

Tabla 13: Archivo PdM_telemetry ((kaggle, 2018))

datetime	machineID	volt	rotate	pressure	vibration
01-01-2015 6:00	1	1,7622E+14	4,185E+14	1,1308E+14	4,5088E+14
01-01-2015 7:00	1	1,6288E+13	4,0275E+14	9,5461E+14	4,3414E+14
01-01-2015 8:00	1	1,7099E+14	5,2735E+14	7,5238E+14	3,4179E+14
01-01-2015 9:00	1	1,6246E+14	3,4615E+14	1,0925E+14	4,1122E+14
01-01-2015 10:00	1	1,5761E+13	4,3538E+14	1,1189E+14	2,5991E+14
01-01-2015 11:00	1	1,725E+14	4,3032E+14	9,5927E+14	3,5655E+14
01-01-2015 12:00	1	1,5656E+14	4,9907E+14	1,1176E+14	4,2754E+14
01-01-2015 13:00	1	1,7252E+14	4,0962E+14	1,01E+13	3,5482E+14
01-01-2015 14:00	1	1,7532E+14	3,9865E+14	1,1062E+14	4,5482E+14
01-01-2015 15:00	1	1,6922E+14	4,6085E+14	1,0485E+14	3,9902E+14
01-01-2015 16:00	1	1,6706E+14	3,8248E+14	1,0378E+14	4,2676E+14

5.3. Uso de los datos

Anteriormente se propusieron tres usos de la metodología sobre bases de datos y ahora se identificará cuál o cuáles de esos 3 usos pueden ser utilizados en la información que se tiene disponible.

El primero es la detección de fallas, en donde se pretende identificar la localización de las fallas en un sistema o máquina. Para este caso este uso no se aplica, dado que son solo 4 componentes los que pueden presentar problemas y no se indica que sean máquinas de gran tamaño, por lo que se supone que los síntomas son más fáciles de identificar o se puede saber que falló con una revisión rápida.

El segundo uso corresponde a la causa de la falla. Este sí es de utilidad, ya que, al tener 4 componentes, cobra importancia saber cuáles son las razones que producen que falle cada uno de ellos, para así tomar medidas al respecto y extender sus tiempos de vida.

Finalmente queda el método de predicción de fallas, el cual también se puede aplicar en este caso, ya que se tienen las fechas de las fallas y de las mantenciones, por lo cual se

puede estructurar el análisis de 2 formas. La primera forma es determinar los días restantes para que falle un determinado componente o el tiempo sin fallas y la segunda forma es simplificar lo anterior y decir si una máquina fallará en un determinado periodo de tiempo (Se deja el tiempo fijo, por ejemplo, se puede buscar saber si una máquina determinada fallará en 2 días más).

5.4. Estructuración de la base de datos

En función de lo que se busca determinar la estructura de los datos que se ingresarán al programa para crear el árbol será distinta, por lo que en esta sección se explicarán las estructuras por separado. Según la explicación anterior, se deben crear 2 estructuras distintas, siendo la primera para pronosticar los días restantes para que falle un componente y la segunda estructura tiene relación con la simplificación de la primera opción. Para analizar el posible uso de la herramienta para determinar las causas de las fallas, se pueden usar los árboles entregados por las estructuras usadas para los pronósticos.

5.4.1. Tiempo sin fallas

Lo primero que hay que hacer en este caso es encontrar el tiempo sin fallas de cada uno de los componentes. Para hacer esto, se trabaja con el archivo de las mantenciones, donde es importante recordar que se encuentran las mantenciones y las fallas de los componentes. De este archivo hay que identificar cuáles son mantenciones y cuáles son fallas, lo cual se hace comparando el archivo de mantenciones con el archivo de las fallas. Una vez que se establece esta diferencia, hay que indicar por cada falla cuanto tiempo pasó desde la mantención o falla más cercana del mismo componente.

Tabla 14: Tiempo sin fallas del mismo componente (Elaboración propia)

Tiempo sin fallas mismo comp (días) ▼
45
45
45
60
30
45

La variable recién determinada, corresponde a lo que se pretende pronosticar. Antes de obtener las variables clasificadoras, es importante hacer un análisis, que hace referencia al pronóstico de los datos. Si solo se está clasificando, no es relevante considerar este punto, pero en este caso hay que tomar en cuenta qué datos se pueden usar en el pronóstico, ya que si vamos a pronosticar el día de hoy para saber si una máquina determinada va a fallar en 2 semanas más, es necesario saber qué información se tiene disponible y cual no. En este caso, se tiene que los tiempos sin fallas son de un mínimo de 30 días, por lo que es posible recolectar información durante 29 días sin esperar percibir ninguna falla.

Tomando en cuenta la acotación anterior, todos los datos de las variables clasificadoras serán hasta 29 días después de la última mantención o falla de un componente. Este día será llamado “día de pronóstico”.

La primera variable clasificadora tiene relación con el tiempo que ha pasado desde la última mantención o falla de cada uno de los componentes en relación con el día de pronóstico. Esta variable se explica en el sentido de que lo que se espera es que, si un componente lleva más tiempo en uso, debería fallar antes. Es importante destacar que mucho o poco tiempo depende de cada uno de los componentes, por lo que no necesariamente debería fallar el componente con más tiempo de uso hasta la fecha.

Tabla 15: Estructura del tiempo de falla anterior en la estructura 1 (Elaboración propia)

comp1	comp2	comp3	comp4
53	203	53	218
110	185	80	140
170	215	185	170

Luego un elemento importante es qué componente se cambió, ya que como se quiere saber cuánto tiempo estará un componente sin percibir una falla, se debe saber qué componente es, por lo que la variable se estructura de la siguiente forma.

Tabla 16: Componente que se cambió (Elaboración propia)

Componente
comp2
comp1
comp1
comp4

Otras variables que considerar son el modelo y la edad de cada una de las máquinas, que se extraen por el número de cada una de ellas.

Tabla 17: Estructura de los Modelos y edades de las máquinas en la estructura 1 (Elaboración propia)

Modelo	Edad
model3	2
model3	20
model3	20
model4	7
model4	7
model4	7

Luego se encuentran los errores, los cuales se consideraron como la suma de ellos desde la última mantención o falla hasta 29 días después. Cada tipo de error se sumó por

separado, con el fin de determinar si algún tipo de error tiene alguna relación con algún componente o tiempo sin fallas.

Tabla 18: Estructura de errores en la estructura 1 (Elaboración propia)

E1	E2	E3	E4	E5
3	2	1	0	0
1	0	1	1	0
1	1	0	2	1

Finalmente quedan las mediciones, de las cuales se obtienen 2 datos, que son el promedio y la desviación en un determinado periodo de tiempo. Como 29 días es mucho tiempo para considerar un promedio, se puede tomar como periodo el día 29, siendo en este caso el día completo.

Tabla 19: Estructura de las mediciones en la estructura 1 (Elaboración propia)

Vibration	Volt des	Rotate des
383884975221409	48944626572292.3	133734530699373
301812100414699	48927593512893.4	132662031111090
307265685498778	48880664365846.7	129976643506536

5.4.2. Si fallará una máquina

En este caso se busca saber si una máquina va a fallar en un determinado periodo de tiempo, pero también es relevante determinar cuál es el componente que va a fallar. Para lograr esto, la variable a clasificar debe ser el componente que presentó la falla en el periodo de tiempo determinado o un “No” si es que no se presentó ninguna falla.

Tabla 20: Variable a pronosticar en la estructura 2 (Elaboración propia)

Pronosticar
No
No
No
No
No
comp1

Ahora que se determinó la variable a pronosticar, hay que determinar las variables clasificadoras, las cuales se calcularán de forma diaria, por lo tanto, se tomarán los datos por cada día del año 2015.

Algunas variables son iguales a las del modelo anterior, como son la edad y el modelo de las máquinas, algunas son similares, como son las mediciones que ahora se toman como un promedio del día y por otro lado están los errores que son la suma de los errores correspondientes al día.

En la misma línea anterior, la última variable es una similar a una variable de la estructura anterior y habla de los tiempos de mantención o falla de cada componente a la fecha. Lo que se busca es saber el tiempo que ha pasado desde la última mantención o falla de cada componente, para establecer una relación entre el tiempo de cambio del componente con una posible falla.

Tabla 21: Tiempos de mantención o falla por componente de la estructura 2 (Elaboración propia)

Comp1	Comp2	Comp3	Comp4
19	214	154	169
49	139	154	124
199	4	34	124
0	79	49	169

5.5. Análisis de la herramienta

A continuación, se va a utilizar la herramienta (Metodología CART) para verificar su poder predictivo en el caso expuesto. Antes de utilizar directamente las estructuras planteadas en la sección anterior, se pretende hacer un análisis usando solo algunos elementos de la base de datos, para así determinar que tanto aporta agregar más información. El análisis como tal, se hará basándose en la estructura de si fallará una máquina, ya que solo se tomarán en cuenta varios tipos de desfases de tiempo.

Es importante aclarar que los datos se ordenarán por fecha, por lo que la información con la que se construye el árbol es más antigua que la información con la que se prueba el mismo. Finalmente, la base para todas las pruebas tendrá un tamaño del 25% de los datos.

5.5.1. Análisis de los errores

Se utilizarán los 4 errores para intentar hacer una predicción de una falla con un desfase de 1 día, 2 días y 2 semanas. En el programa se ingresará la suma de los errores de un día antes del periodo correspondiente, dejando un total de 36400 para 1 día, 36300 datos para los 2 días y 35100 datos para las 2 semanas. La variación en la cantidad de datos es producto de que los últimos datos, que son las últimas fechas se quedan sin información, ya que no se sabe su comportamiento en el futuro. Para construir el árbol se utilizarán 20000 datos y la base de prueba tendrá un tamaño del 25% del desfase con menos datos y sólo se tomarán datos que sean válidos para los tres desfases, para que así sean comparables entre sí.

El primer caso es el del desfase de 1 día, donde del total de los datos de prueba, se pronosticó correctamente un 98,9% de los datos (Accuracy), pero solo un 44,9% de los componentes fueron pronosticados correctamente (Sensitivity). En este caso se puede apreciar que a pesar de que el pronóstico general es favorable, si se analizan los componentes independientemente no sucede lo mismo. A su vez, también es relevante indicar que los indicadores utilizados, usando la generalización de que los componentes son los valores positivos y los “No” son los negativos. A continuación, se agregará el porcentaje de predicciones correctas por componente (Número de predicciones de un componente X correctas / Número de fallas del componente X en la base de datos) para tener un análisis previo de cada uno de ellos.

Al revisar los componentes según su tipo, se puede apreciar que el mejor pronosticado es el componente 2, llegando a un 100% de acierto (porcentaje de predicciones correctas en el componente 2), lo que se contrapone a los resultados de los otros componentes, donde el que tiene el mejor desempeño es el componente 4 con un 8,8%. Al analizar al componente 2, se puede apreciar que los nodos finales en los cuales se reúnen los componentes 2 son muy limpios, por lo que se puede indicar o concluir que el componente 2 se explica claramente por los errores que se presenten en el día anterior a la falla.

Del cuadro que se encuentra a continuación, cada uno de los elementos significa lo siguiente:

- Nodos finales: La cantidad de nodos finales del árbol creado.
- comp1: Porcentaje de los componentes 1 bien clasificados.
- comp2: Porcentaje de los componentes 2 bien clasificados.
- comp3: Porcentaje de los componentes 3 bien clasificados.

- comp4: Porcentaje de los componentes 4 bien clasificados.
- No: Es el porcentaje de los valores “No” bien clasificados.

Tabla 22: Pronostico sólo con errores con desfase de 1 día (Elaboración propia)

Accuracy	98,9%	comp1	0,0%
Sensitivity	44,9%	comp2	100,0%
		comp3	7,4%
Nodos finales	28	comp4	8,8%
		No	99,9%

Lo explicado anteriormente cambia drásticamente cuando se habla de un desfase de 2 días y de 2 semanas, ya que los pronósticos de todos los componentes caen drásticamente a 0, tal como se aprecia en la imagen.

Tabla 23: Pronostico sólo con errores con desfase de 2 días (Elaboración propia)

Accuracy	98,2%	comp1	0,0%
Sensitivity	0,0%	comp2	0,0%
		comp3	0,0%
Nodos finales	22	comp4	0,0%
		No	100,0%

Esto indica que los errores no cobran importancia con un desfase, por lo que se puede concluir que solo cumple la función de ser un reflejo de la situación con un pequeño desfase de tiempo (1 día).

Hasta el momento se puede hablar de que la herramienta permite analizar los datos que se tienen y qué podemos lograr con ellos, determinando en este caso si los errores son de utilidad o no en qué momento.

Para hacer un análisis completo de la predicción, se va a utilizar el índice de Kappa generalizado.

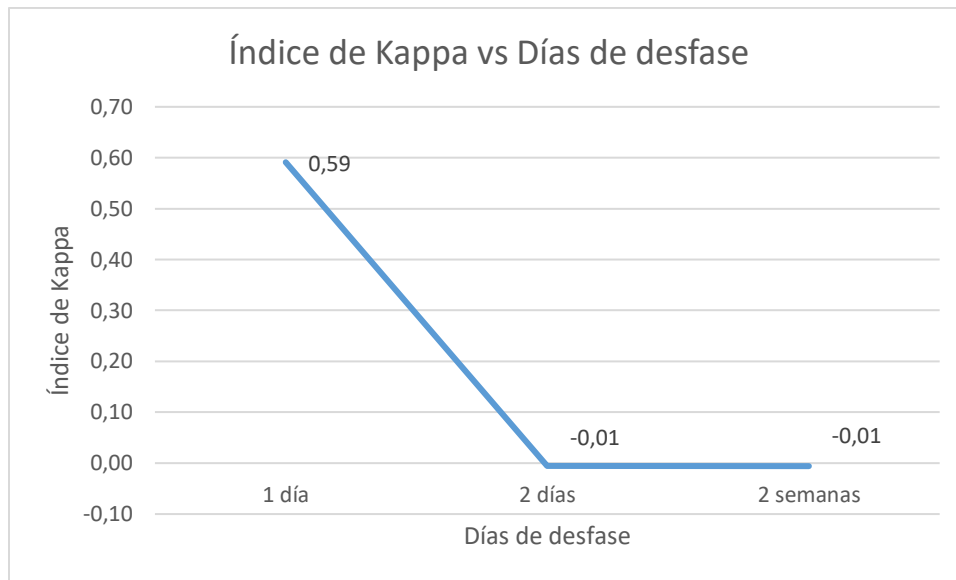


Gráfico 1: Índice de Kappa en el análisis de errores (Elaboración propia)

El índice en un principio parece tener una predicción relativamente decente, pero cuando se aumenta el desfase, cae a valores negativos, lo cual muestra un desacuerdo con la predicción y lo real. Según el análisis anterior, se puede decir que una parte importante del Kappa de 1 día de desfase está dada por la buena predicción del componente 2 que se mostró anteriormente.

5.5.2. Análisis de las mediciones

En la misma línea, se hará un análisis de las mediciones, para apreciar que tanto pueden pronosticar por si solas. Respecto a la cantidad de datos que se tienen y los utilizados, se cumple lo mismo que en el caso de los errores.

Para este caso, el primer y segundo desfase presentan valores similares, siendo el componente 3 el que más varía, aumentando de un 3,7% a un 16,0%.

Tabla 24: Pronostico con sólo las mediciones con desfase de 1 y 2 días (Elaboración propia):

Desfase 1				Desfase 2			
Accuracy	96,6%	comp1	16,1%	Accuracy	96,7%	comp1	22,2%
Sensitivity	17,1%	comp2	24,2%	Sensitivity	20,1%	comp2	21,5%
		comp3	3,7%			comp3	16,0%
Nodos finales	399	comp4	14,7%	Nodos finales	474	comp4	18,2%
		No	98,1%			No	98,1%

Finalmente, con el desfase de las 2 semanas se produce una caída en la predicción de los componentes y solo se llega a un 0,6%.

Tabla 25: Pronostico con sólo las mediciones con desfase de 2 semanas (Elaboración propia)

Accuracy	95,6%	comp1	0,0%
Sensitivity	0,6%	comp2	1,5%
	 (Ctrl) ▾	comp3	0,0%
Nodos finales	730	comp4	0,0%
		No	97,4%

También se puede apreciar que el único componente que presenta una predicción es el componente 2, llegando a un 1,5%.

Esto lleva a pensar que este elemento clasificador va a ser útil únicamente en los desfases de 1 y 2 días.

Al igual que en los errores se agrega el índice de Kappa.

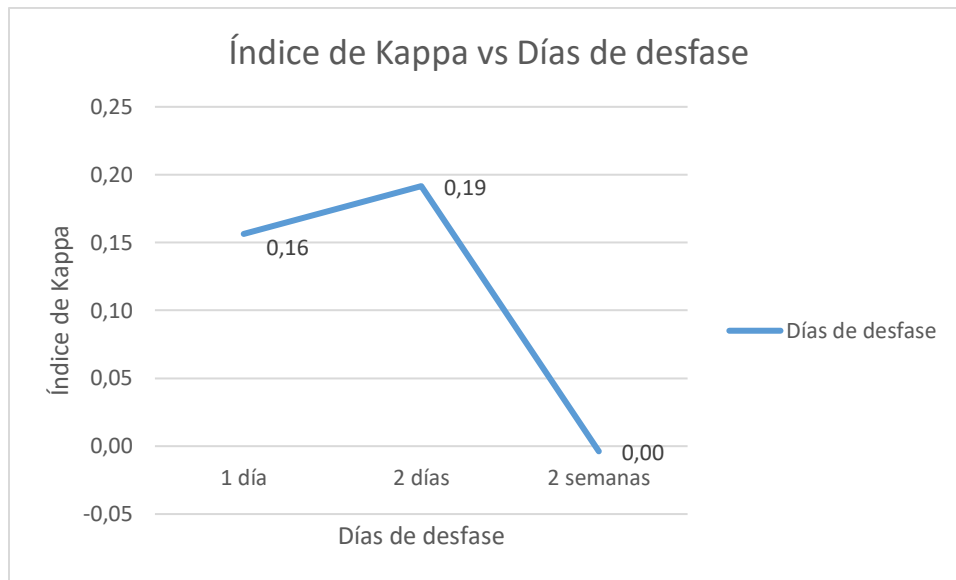


Gráfico 2: Índice de Kappa en el análisis de mediciones (Elaboración propia)

El índice muestra que, en los dos primeros días de desfase, se presenta una relación, pero no es muy lejana a 0, que representa el no acuerdo. Cuando se llega a las 2 semanas, la predicción cae al no acuerdo, por lo que se puede establecer al igual que con lo dicho anteriormente, que esta información (Desfase de 1 y 2 días) podría ser útil en conjunto con otros datos, pero por sí sola no es efectiva.

5.5.3. Análisis de los días desde una mantención o falla

Siguiendo la misma estructura de los análisis anteriores, se busca identificar cuándo se puede pronosticar en los tres desfases utilizando únicamente los días que han pasado desde la última mantención o falla.

En el caso del primer desfase, los porcentajes de acierto de los componentes son bajos, en particular el del componente 4, que solo llega al 2,9%.

Tabla 26: Pronostico sólo con los tiempos desde mantención o falla con desfase de 1 día (Elaboración propia)

Accuracy	96,6%	comp1	12,9%
Sensitivity	7,0%	comp2	6,1%
		comp3	7,4%
Nodos finales	635	comp4	2,9%
		No	98,2%

Cuando se pasa al desfase de 2 días, sólo varían algunos porcentajes, pero la situación es bastante similar.

Tabla 27: Pronostico sólo con los tiempos desde mantención o falla con desfase de 2 días (Elaboración propia)

Accuracy	96,7%	comp1	16,1%
Sensitivity	7,0%	comp2	6,0%
		comp3	3,6%
Nodos finales	625	comp4	3,1%
		No	98,4%

Finalmente queda el desfase de las 2 semanas, que al igual que en dos anteriores, las variaciones son muy pequeñas, siendo la única excepción el caso del componente 2, que aumenta de un 6,0% a un 12,3%, dando a entender que, para el caso de esta variable clasificadora, es mejor al aumentar el desfase.

Tabla 28: Pronostico sólo con los tiempos desde mantención o falla con desfase de 1 semana (Elaboración propia)

Accuracy	96,9%	comp1	13,9%
Sensitivity	10,1%	comp2	12,3%
		comp3	4,0%
Nodos finales	601	comp4	6,1%
		No	98,5%

De los resultados se puede decir que esta variable clasificadora en una primera instancia no parece variar significativamente su poder predictivo cuando se varía el desfase.

A continuación, se muestra el valor del índice de Kappa.

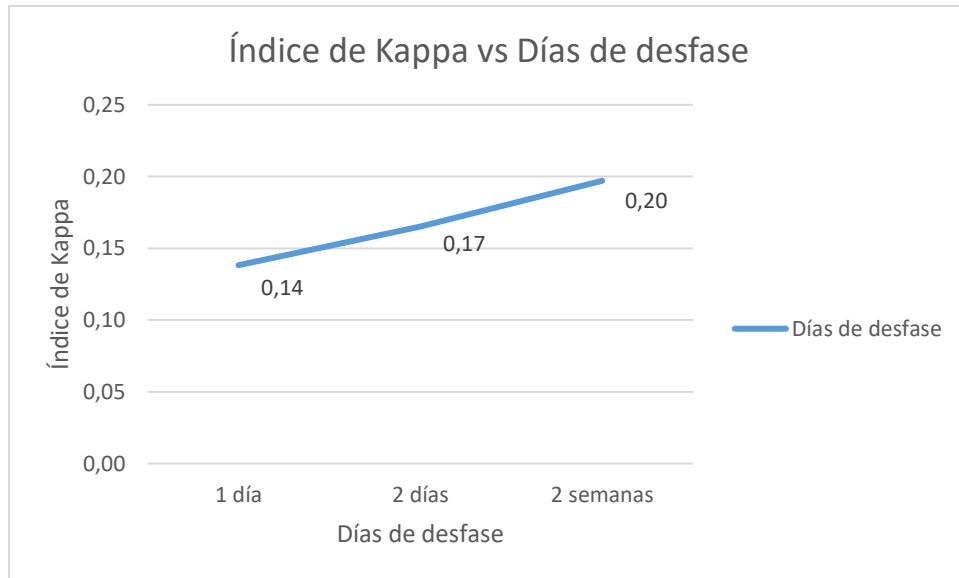


Gráfico 3: Índice de Kappa en el análisis de días desde una mantención o falla (Elaboración propia)

En este caso el índice de Kappa aumenta con los días de desfase, mostrando que el grado de relación entre la predicción y lo real siempre es mayor, pero son valores bajos, por lo tanto, se da una situación similar al análisis de mediciones, donde la capacidad de pronóstico por sí sola no es efectiva.

De los tres análisis anteriores, se puede apreciar que el indicador Accuracy es alto, pero esto no significa que se esté haciendo una buena predicción, ya que este alto valor es probable que se produzca porque aproximadamente el 98,1% de los datos son “No”, por lo tanto, si la variable clasificadora no es buena, sucede que en los nodos finales siempre hay valores “No” o que simplemente es lo más común. Tomando en cuenta la alta densidad de valores “No”, es que se llega a ese valor y como la mayoría de los datos tienen esa respuesta, se entiende que se presente un buen pronóstico (Base de datos desbalanceada). Producto de la explicación anterior, es prudente juzgar a la clasificación, en función de que tan bien clasifica un determinado componente y no según la clasificación Total. Complementario a lo

anterior el índice de Kappa muestra de una forma general el resultado de la predicción, evitando el problema antes mencionado.

5.5.4. Análisis de estructura si fallará una máquina con 1 día de desfase

Lo primero que se hará es crear el árbol de regresión o clasificación en esta estructura con 1 día de desfase, por lo tanto, se pueden utilizar 36400 datos, pero para poder hacer comparaciones con los análisis anteriores (Uso de los árboles de regresión usando solo algunas variables), se utilizará la misma base de prueba en cada uno de los árboles que se construyan. De esta estructura el enfoque se dará en 2 análisis, que son el pronóstico según el grado de impureza de los nodos finales y el pronóstico según la cantidad de datos que se utilicen.

Para mostrar los resultados, se usará el árbol que se construye con 10.000 datos y 0.001 de impureza. Este árbol consta de 39 nodos finales, pero se pueden apreciar nodos en donde se concentra la mayor cantidad de los resultados de un cierto tipo. Para resaltar estos nodos, se les colocó un color verde de la siguiente forma.

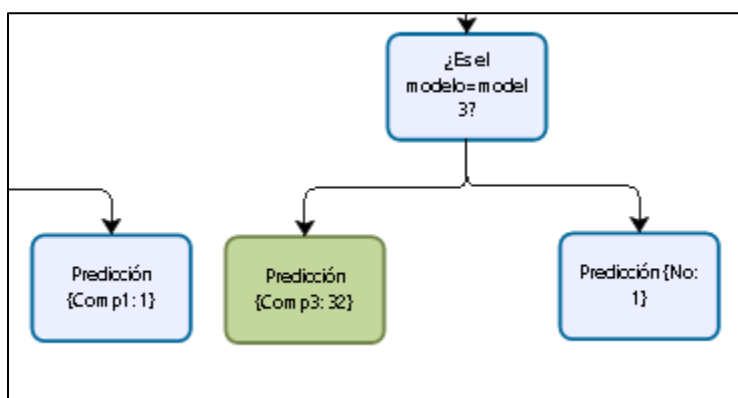


Ilustración 22: Nodo resaltado (Elaboración propia)

Este nodo contiene una cantidad importante de los componentes 3, por lo tanto, a la hora de hacer un análisis, el enfoque se dará en los nodos de color verde. También es

importante indicar que a la derecha se encuentra el resultado si la afirmación es verdadera y a la izquierda el resultado si la afirmación es falsa.

A continuación, se presenta el árbol de clasificación y regresión de izquierda a derecha, para que se aprecien bien los nodos y principalmente los nodos de color verde. A través de estas imágenes se buscará extraer información, para hacer un análisis respecto a las características que produce que un componente falle o no falle.

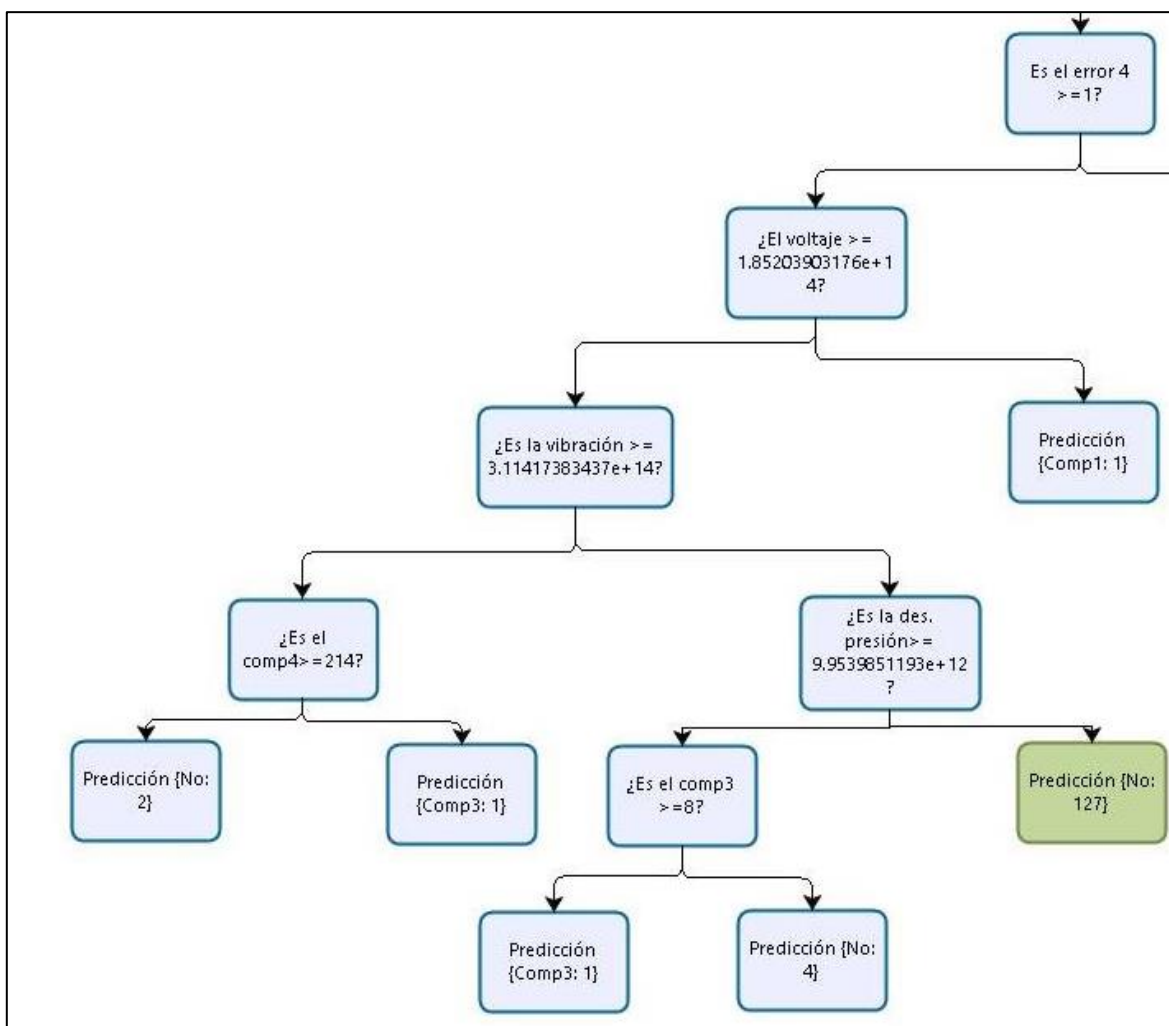


Ilustración 23: Árbol de regresión y clasificación 10.000 datos parte 1 (Elaboración propia)

En esta parte se encuentra un nodo que contiene una gran cantidad de valores No. Antes de llegar a la pregunta del error 4, hay una pregunta por el error 5, el error 3 y el error

1, siendo la respuesta de todas falso. Se podrían analizar las mediciones, pero los componentes que fallan bajo estas condiciones son tan pocos, que es posible quedarse con la conclusión de que si no se presentan errores lo más posible es que no se presente una falla.

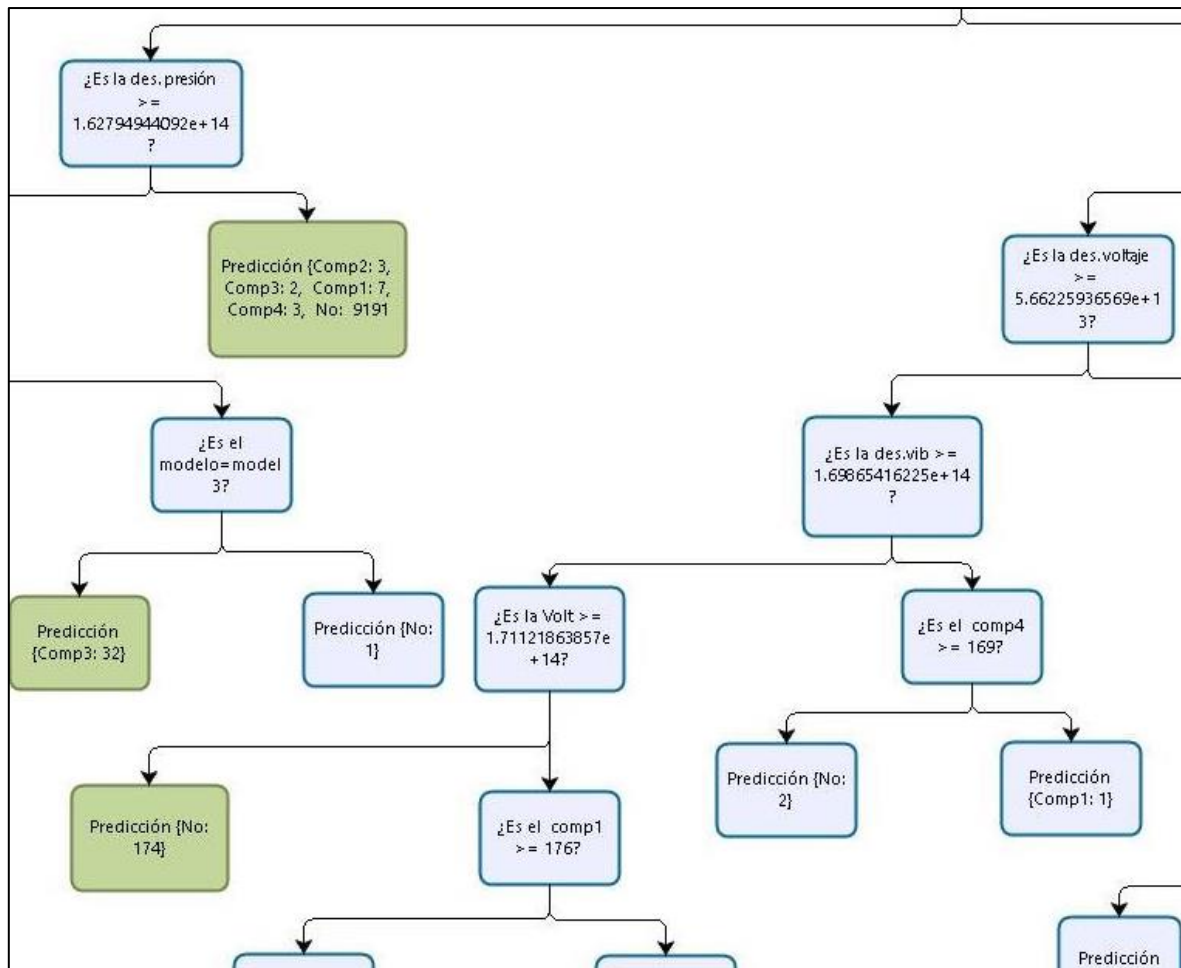


Ilustración 24: Árbol de regresión y clasificación 10.000 datos parte 2 (Elaboración propia)

En la imagen anterior se pueden apreciar 3 nodos importantes, siendo uno de ellos referente a los componentes 3 y los otros 2 son No. El nodo más grande de los No, indica que cuando no hay errores tipo 5, 3 y 1, sumado a que la desviación de la presión supera un cierto valor, no se producen errores. La mayoría de los valores No, se concentran en este nodo, por lo que sería interesante averiguar que puede producir que la desviación de la presión baje, para evitar estos elementos. El segundo nodo que contiene valores No es considerablemente

más pequeño que el ya nombrado y se produce cuando existe el error 1 y otras condiciones de voltaje y vibración, indicando que estos niveles se deben mantener bajo ciertos valores.

Ahora solo queda analizar el origen del nodo de los componentes 3, el cual se produce cuando no hay error tipo 5, 3 y 1, sumado a una desviación de la presión bajo un cierto valor. Luego se debe presentar el error 4 y el modelo de la máquina debe ser distinto a model 3. Lo anterior indica que la falla del componente 3 se produce por el error tipo 4 y no se presenta en el model 3, ya que si se analizan las fallas los modelos 3 no fallaron en el componente 3. Es importante para este caso analizar la razón de esta duración extendida, para identificar si es posible que esto pueda replicarse a otras máquinas.

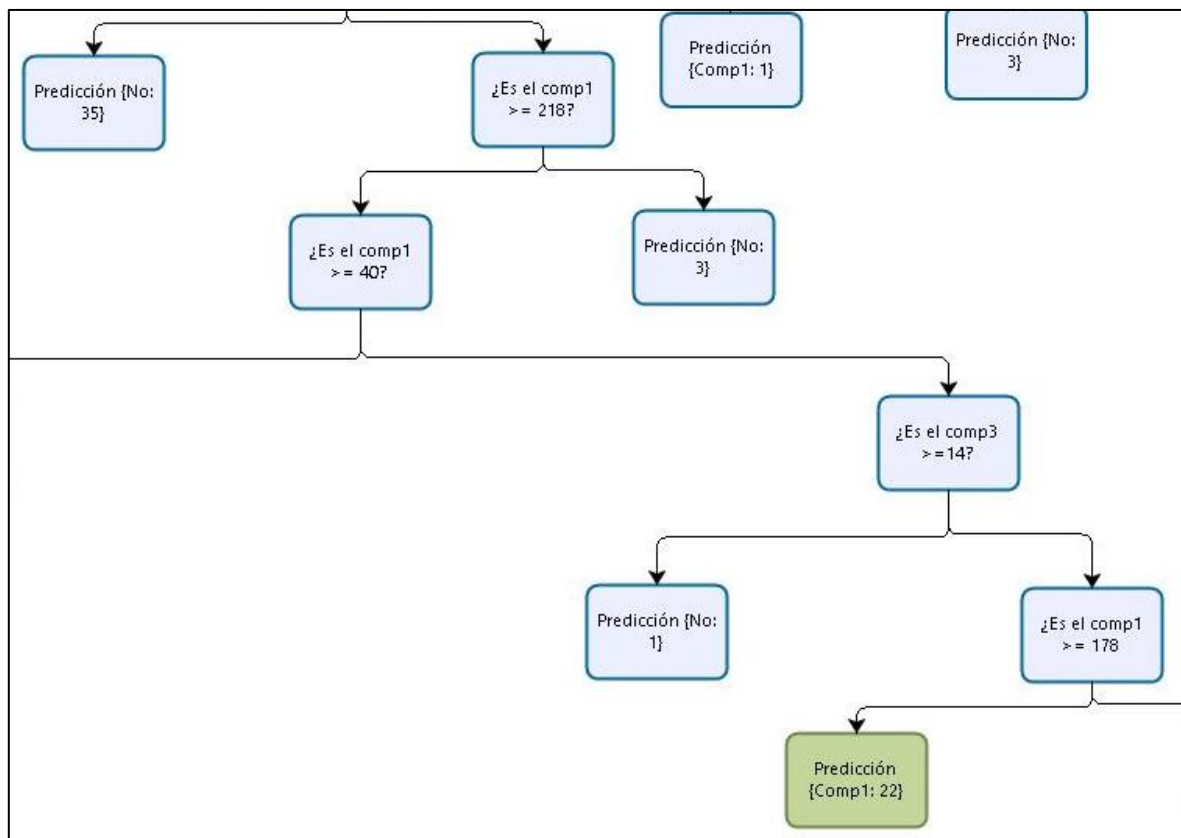


Ilustración 25: Árbol de regresión y clasificación 10.000 datos parte 3 (Elaboración propia)

En la imagen, se puede ver que el nodo más antiguo que se muestra es el de si el componente 1 se cambió hace más de 218 días, pero antes de ese nodo se contestó que no había error 3 y 5 y si había error 1, además de que se hicieron preguntas referentes al voltaje, como lo indica la imagen siguiente.

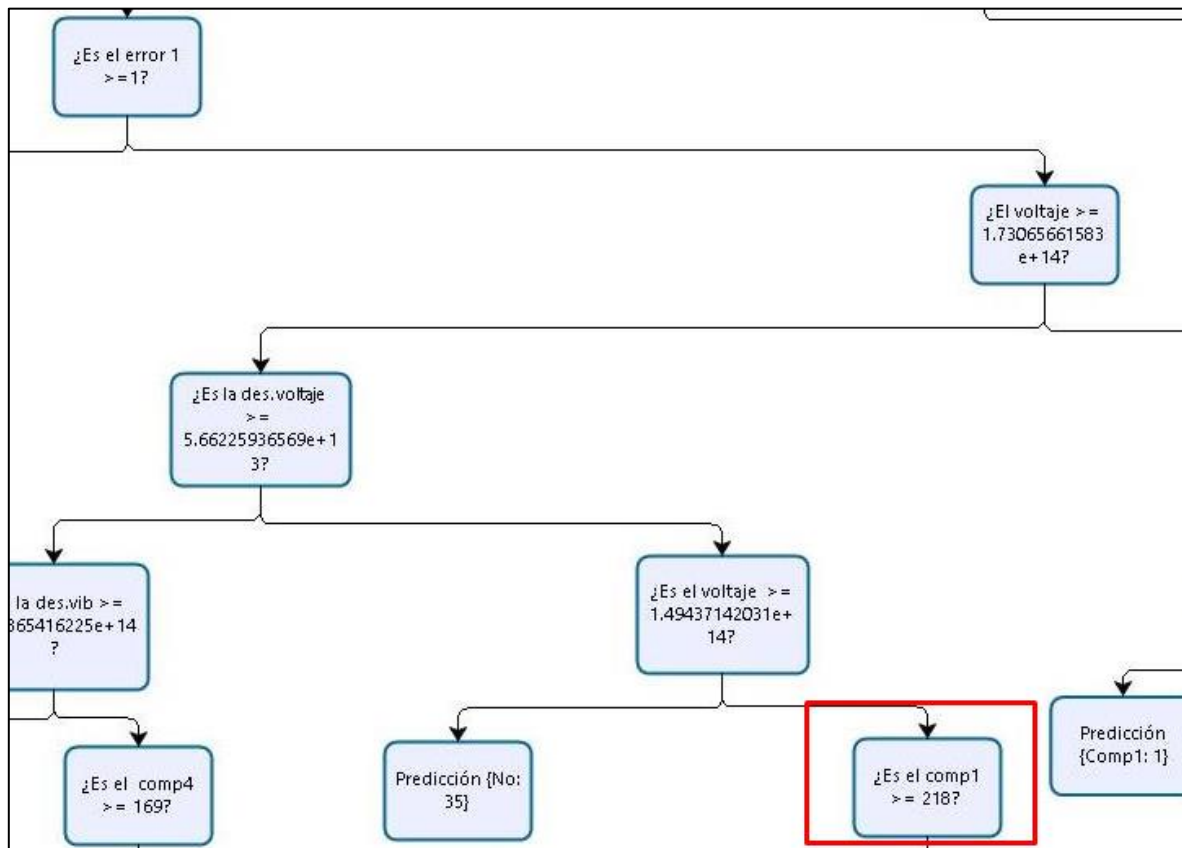


Ilustración 26: Árbol de regresión y clasificación 10.000 datos parte 4 (Elaboración propia)

Después, se puede ver que el último cambio del componente 1 no tiene más de 218 días, pero si tiene más de 40 días. Complementario a lo anterior se agrega un nodo que pregunta por el cambio del componente 3, lo que hace suponer que el componente 3 tiene alguna relación con el componente 1, pero también puede ser que solo sea que el programa considerara que se aumenta la homogeneidad de los datos, por lo que sin saber técnicamente de qué máquinas se está hablando no se pueden extraer mayores conclusiones. Finalmente,

la última pregunta es si el componente 1 se cambió hace más de 178 días, a lo cual la respuesta fue negativa.

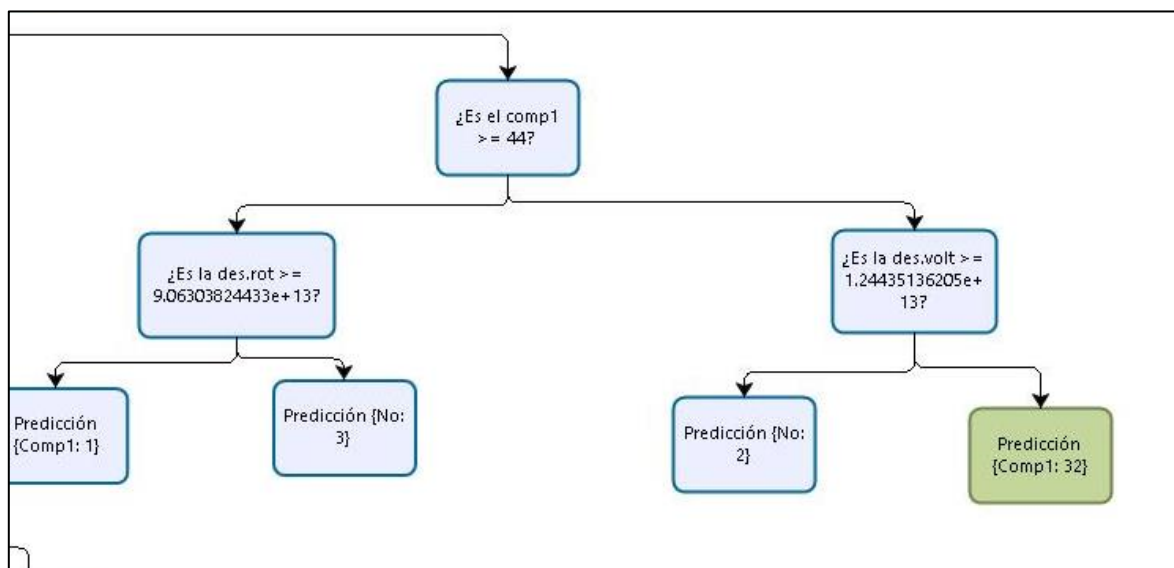


Ilustración 27: Árbol de regresión y clasificación 10.000 datos parte 5 (Elaboración propia)

Este nodo tiene condiciones similares a la anterior, pero tiene condiciones distintas de voltaje, ya que en el nodo que tiene la pregunta ¿El voltaje $\geq 1.73065661583 \times 10^{14}$?, en donde la respuesta para este nodo es verdadera y para el anterior es falsa. Luego aparece la pregunta si el componente 1 se cambió hace más de 44 días y finalmente se muestran condiciones de la desviación del voltaje. Se puede apreciar que los componentes 1 no duran muchos días, salvo algunas excepciones y además los componentes 1 están ligados a las modificaciones del voltaje.

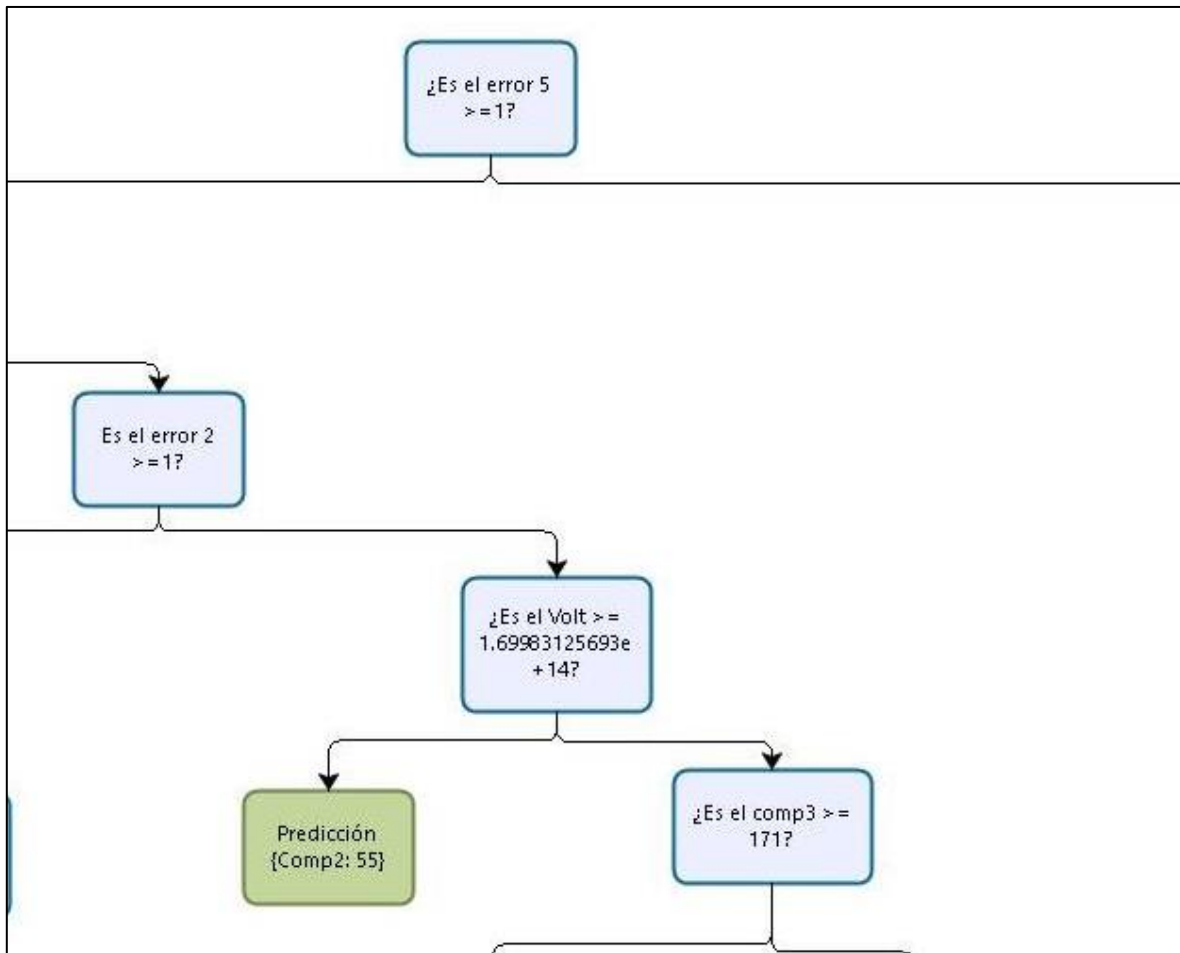


Ilustración 28: Árbol de regresión y clasificación 10.000 datos parte 6 (Elaboración propia)

En el caso de los componentes 2, se concentran en un nodo que responde que no hay error tipo 5, pero si hay error tipo 3 y tipo 2 al mismo tiempo y a esto se le suma que el voltaje debe ser menor a un cierto valor. En este caso es importante analizar las razones de estos errores y su relación con la falla del componente 2.

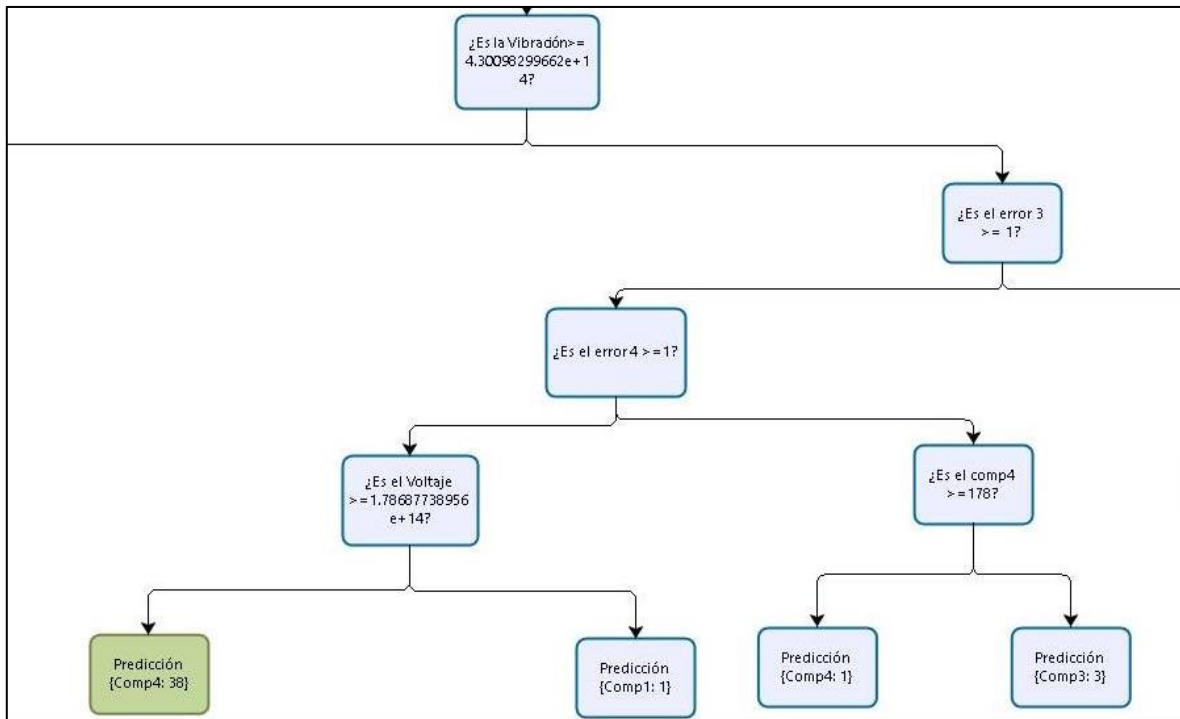


Ilustración 29: Árbol de regresión y clasificación 10.000 datos parte 7 (Elaboración propia)

Finalmente, el último nodo a destacar es el nodo de los componentes 4, el cual contesta primero que, si hay error 5, establece algunos límites a la vibración, indica que no hay error 4 ni 3 y establece que el voltaje debe ser inferior a un cierto valor.

Como se puede ver el árbol permite analizar las características que tiene la falla de cada componente de una forma visual y en función de esto se puede saber dónde enfocar los esfuerzos para mejorar el rendimiento de cada uno de los componentes, es por esto que mientras más características considera la base de datos, se pueden sacar conclusiones que incluyan más aspectos del uso y funcionamiento de las máquinas.

5.5.4.1. Variación del pronóstico según el grado de impureza

Para el caso de esta estructura, la impureza se calcula con el índice de Gini, ya que la variable a pronosticar no es cuantitativa. Para analizar la variación del pronóstico, se utilizará una cantidad de datos fija para construir el árbol y tendrá un tamaño de 10.000 datos. Se

probarán los resultados, exigiendo un grado de impureza de los nodos finales de 0.000, 0.001, 0.002, 0.003, 0.004 y 0.005, ya que para estos valores de exigencia se ven cambios en los resultados. Con este análisis se vera la variación de los indicadores y la variación de la cantidad de nodos finales, lo cual busca hacer un análisis similar al de la búsqueda del árbol óptimo.

Para comenzar se puede analizar el índice Sensitivity en función del grado de impureza. Los resultados se ven reflejados en el siguiente gráfico.

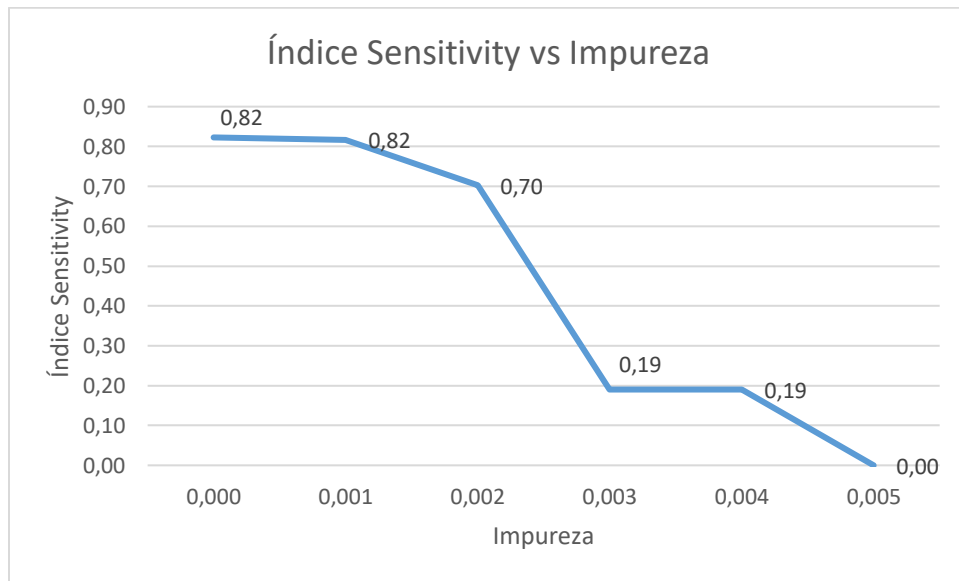


Gráfico 4: Grafico del Índice Sensitivity vs Impureza de los nodos con desfase de 1 día (Elaboración propia)

En el gráfico anterior se puede apreciar que a medida que aumenta la impureza permitida en los nodos finales, el índice Sensitivity cae. Esto cobra más sentido si se ve la variación de la cantidad de nodos finales en función de la impureza de los nodos finales.

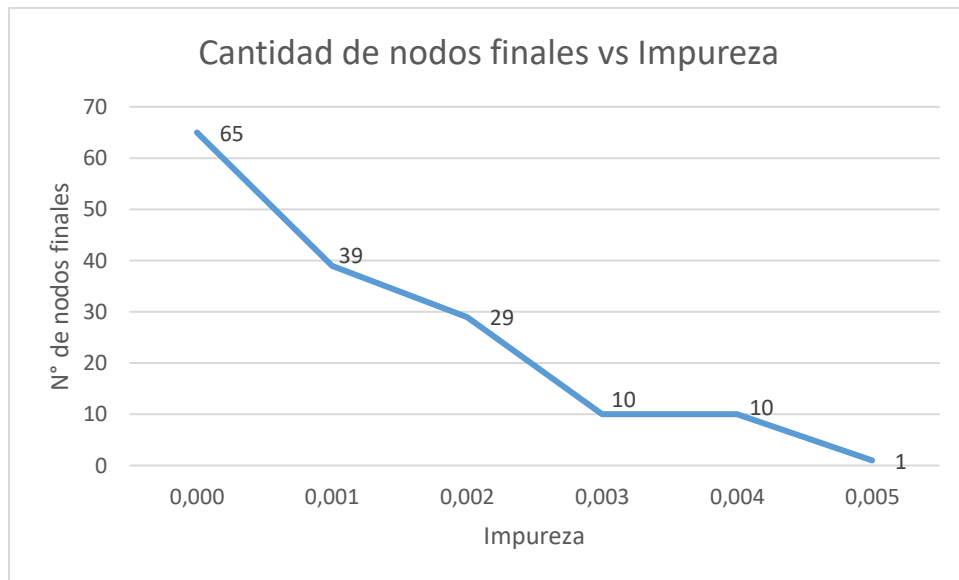


Gráfico 5: Grafico del N° de nodos finales vs Impureza de los nodos finales con desfase de 1 día (Elaboración propia)

Aquí se puede apreciar que la cantidad de nodos finales disminuye a medida que los nodos son más impuros. Esto se debe a que, si se les exige menos a los nodos, el programa no ve la necesidad de hacer más particiones, por lo tanto, si hay menos nodos y son más impuros, es natural pensar que el índice Sensitivity sea menor.

Otro factor para tomar en cuenta es el tamaño óptimo del árbol, ya que es posible tal como se explica en el estado del arte que el árbol tenga un gran tamaño por las distintas excepciones que se presentan en la base de datos, pero esto no necesariamente mejora o mejora significativamente la predicción. En el ejemplo expuesto anteriormente, se puede apreciar que el cambio del porcentaje de acierto desde la impureza 0.000 a la impureza 0.001, es el menor de todos, pero si se ve el gráfico de la cantidad de nodos finales se puede apreciar que la caída en el número de nodos es la más significativa. Esto lleva a pensar que utilizar el árbol a su máxima pureza no necesariamente es lo ideal, ya que queda un árbol más grande que es más difícil de analizar.

Otra forma de ver los resultados es con el índice de Kappa, en donde se muestra que el indicador para la impureza 0.000 es menor 0.001, por lo tanto, se reafirma la conclusión anterior en donde no necesariamente un nodo más impuro es mejor. Un dato relevante es que se produzca una mejora del indicador tomando en cuenta que se está reduciendo la impureza, pero esto se puede explicar porque, a diferencia del indicador Sensitivity, que utiliza solo los valores de los componentes, el indicador Kappa también toma en cuenta las respuestas “No” y al ser la diferencia tan pequeña entre una impureza y otra se produce el resultado que se puede apreciar.

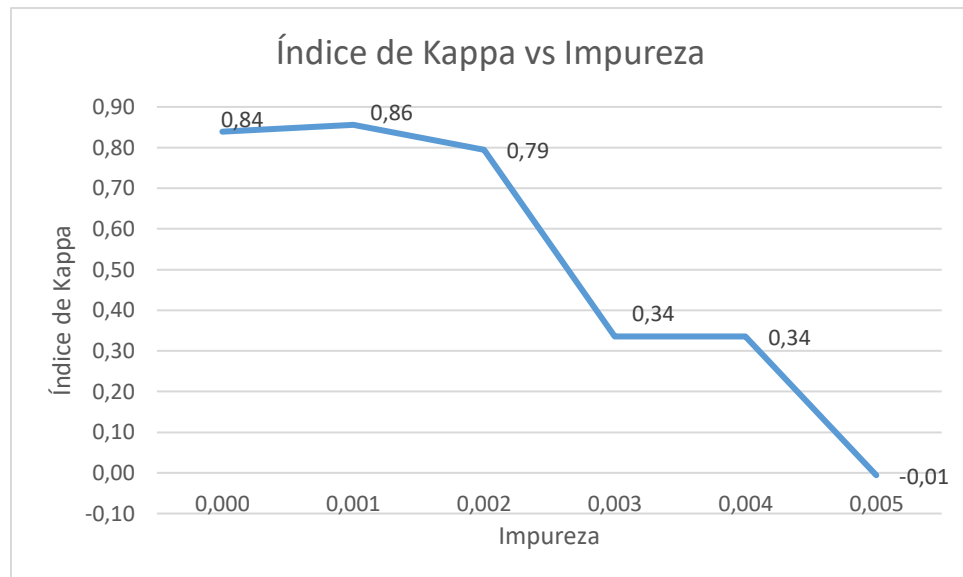


Gráfico 6: Índice de Kappa vs Impureza en el análisis de estructura de si fallará una máquina con 1 día de desfase
(Elaboración propia)

Una vez expuesta la variación de la predicción de los componentes en conjunto, suena interesante analizar cómo cambia el porcentaje de clasificación correcta de cada uno de los componentes por separado. El siguiente gráfico muestra cómo cambia el porcentaje de clasificación correcta de cada componente por separado en función de la impureza permitida para los nodos finales.

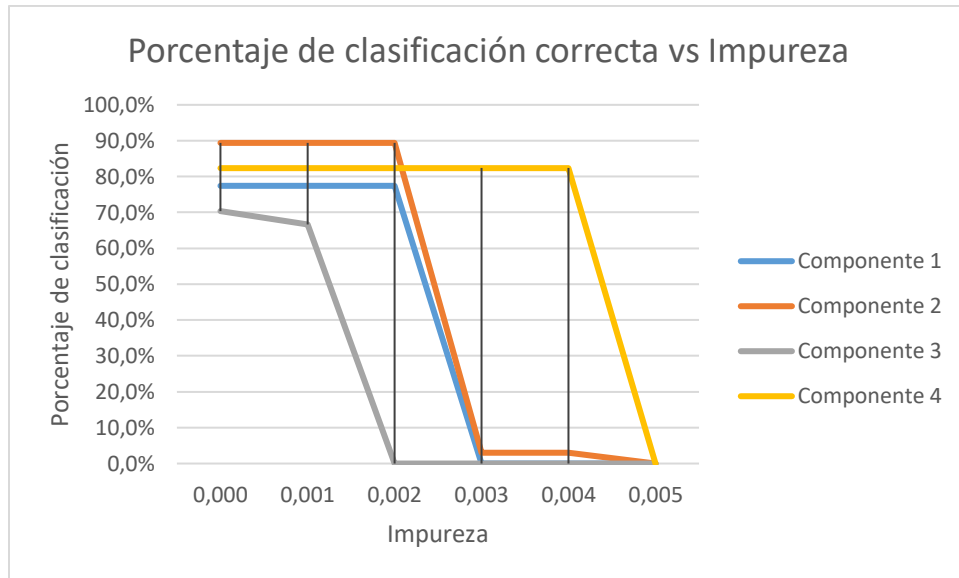


Gráfico 7: Grafico del % de ac por componente vs Impureza de los nodos finales con desfase de 1 día (Elaboración propia)

Se puede apreciar, que cuando se pasa a la impureza de 0.001, el único componente que reduce su porcentaje de acierto es el componente 3, el cual pasa de tener un 70.4% a tener un 66.7%. Este fenómeno muestra que el componente 3 es más sensible a la reducción de la pureza, ya que lo que sucede es que una parte de las respuestas del componente 3 se pierden en un nodo conjunto con otros valores.

Para dar respuesta al fenómeno anterior, hay que tomar algunos elementos en consideración, como es la cantidad de fallas de cada uno de los elementos en la base de datos.

Tabla 29: Número de fallas en la base de construcción de 10.000 datos (Elaboración propia)

Componente	N° de fallas
comp1	72
comp2	66
comp3	39
comp4	48

En la imagen anterior se muestra la cantidad de fallas por componente y se puede ver que el componente 3 es el que menos falla. Otro elemento en consideración es la estructura

del árbol, ya que para el árbol con impureza 0.000, existe un nodo que alberga 32 de los 39 componentes 3, mientras que los otros se encuentran separados en varios nodos.

```

--> False:
    Predict{'comp3': 2}
--> False:
    Is Err4 >= 1?
--> True:
    Is Modelo == model3?
--> True:
    Predict{'No': 1}
--> False:
    Predict{'comp3': 32}
--> False:

```

Ilustración 30: Nodo del componente 3 en la prueba de 10.000 datos con impureza 0.000 (Elaboración propia)

Cuando se pasa a la impureza 0.001, el nodo que contiene a la mayor cantidad de datos, se lleva algunos valores del componente 3, por lo tanto, es probable que debido a esto algunos valores sufran predicciones erróneas.

```

--> True:
    Predict{'comp2': 3, 'comp3': 2, 'comp1': 7, 'comp4': 3, 'No': 9191}
--> False:
    Is Err4 >= 1?
--> True:
    Is Modelo == model3?
--> True:
    Predict{'No': 1}
--> False:
    Predict{'comp3': 32}
--> False:
    Is Volt >= 1.85203903176e+14?

```

Ilustración 31: Nodo del componente 3 en la prueba de 10.000 datos con impureza 0.001 (Elaboración propia)

No es sino hasta que se pasa a la impureza de 0.002 que la mayor parte de las fallas del componente 3 caen en el nodo de los no, por lo tanto, la predicción es “No”.

```

--> False:
  Is des.vib >= 1.69865416225e+14?
--> True:
  Is comp4 >= 169?
--> True:
  Predict{'comp1': 1}
--> False:
  Predict{'No': 2}
--> False:
  Predict{'comp1': 1, 'No': 179}
-> False:
  Predict{'comp2': 3, 'comp3': 36, 'comp1': 8, 'comp4': 3, 'No': 9325}

```

Ilustración 32: Nodo del componente 3 en la prueba de 10.000 datos con impureza 0.002 (Elaboración propia)

Esto lleva a pensar que es probable que ciertas respuestas pasen por nodos similares, por lo que, si es que hay una clara superioridad numérica, es posible que se pierda la predicción de una de ellas si es que se aumenta la impureza para reducir el número de nodos finales en el árbol. El análisis anterior lleva a pensar que antes de aumentar la impureza permitida de los nodos, es prudente analizar cómo se comportan los grupos de fallas.

5.5.4.2. Variación del pronóstico según la cantidad de datos que se utilicen

Lo que se busca en esta sección, es determinar el comportamiento de la predicción total y la de cada uno de los componentes cuando se hace variar la cantidad de datos que se utilizan para crear el árbol. Las pruebas se ejecutarán con la misma base de prueba que en el caso anterior y para construir el árbol, se utilizarán bases de datos que contengan 10.000, 15.000, 20.000 y 25.000 datos, pero se utilizará una impureza de 0.001.

Al analizar el número de nodos, se puede apreciar que aumenta a medida que se aumenta la cantidad de datos, lo que tiene lógica tomando en cuenta que, al haber más datos, se pueden presentar más opciones, situación que se materializa en más nodos.

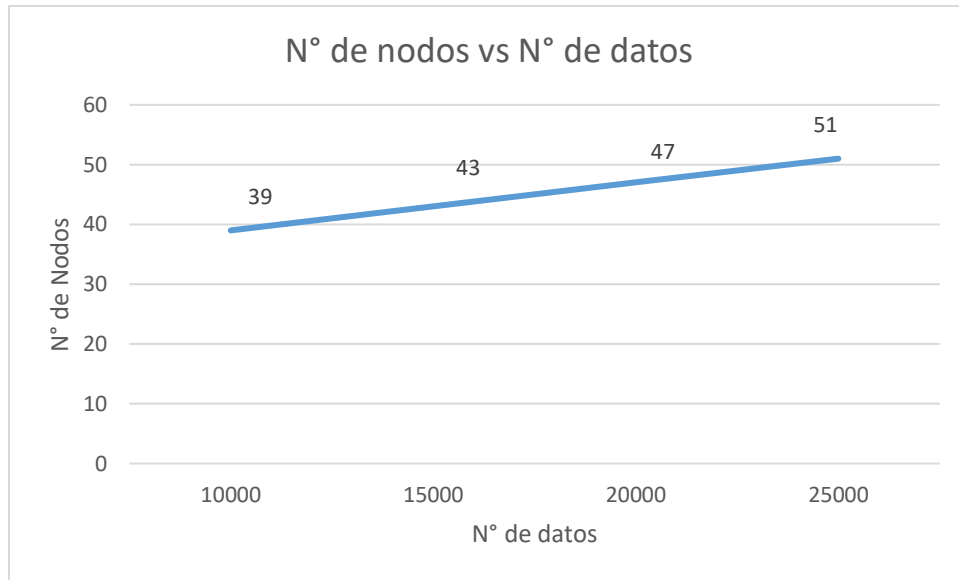


Gráfico 8: N° de nodos finales vs Cantidad de datos usados en la construcción con desfase de 1 día (Elaboración propia)

Lo relevante en este caso no es precisamente el aumento de los nodos, ya que no parece haber un cambio drástico, sino más bien si este aumento en los nodos se relaciona o se condice con un aumento en la predicción de cada uno de los componentes. En la siguiente gráfica se muestra el porcentaje de predicción de cada uno de los componentes en contraste con el número de datos que se usaron para crear el árbol de regresión.

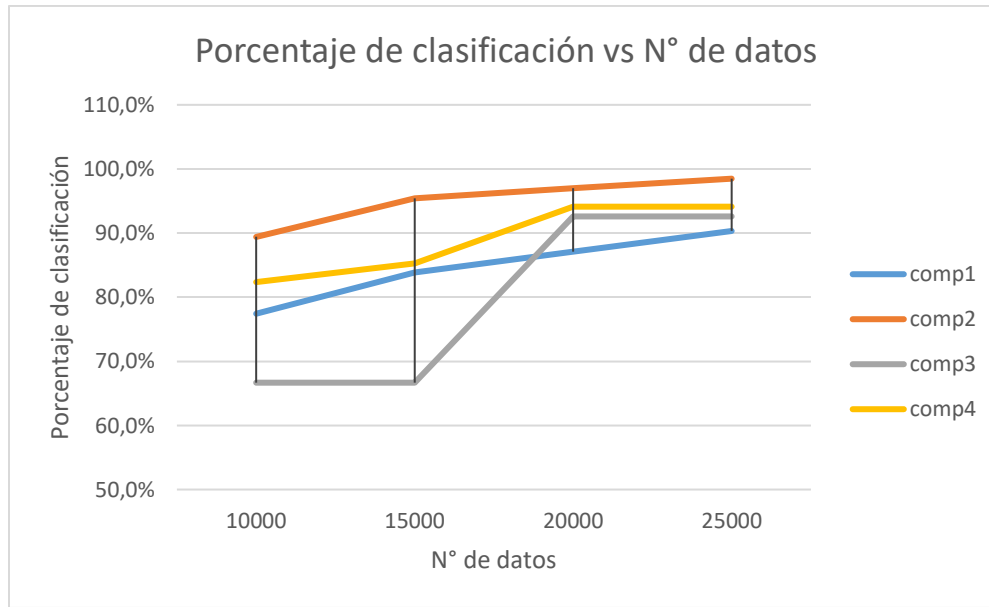


Gráfico 9: Grafico del % de ac por componente vs Cantidad de datos usados en la construcción con desfase de 1 día (Elaboración propia)

Se puede apreciar que en todos los casos los porcentajes de predicción se mantuvieron o aumentaron en todos los componentes, por lo tanto, hay que ver cómo se comporta el índice de Kappa con la variación del número de datos.

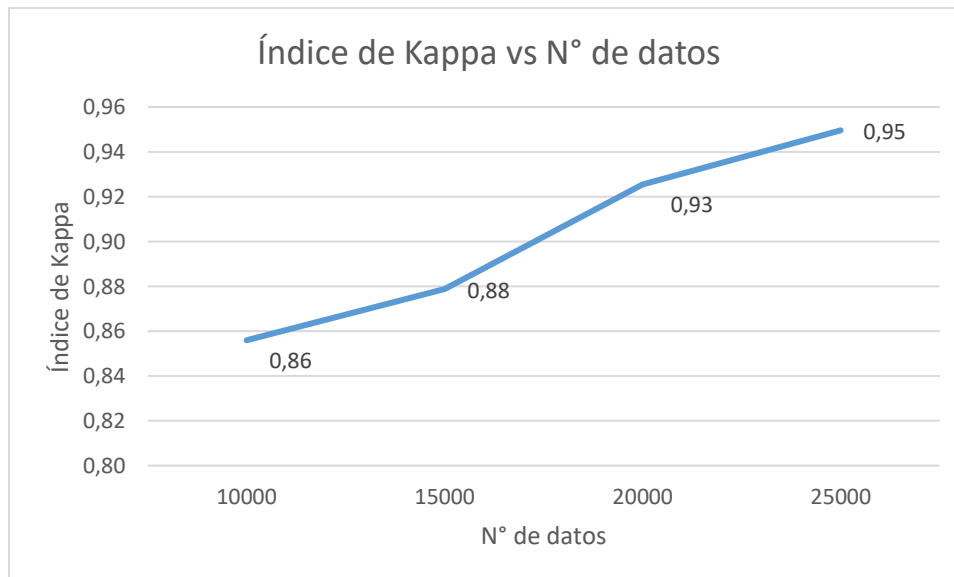


Gráfico 10: Índice de Kappa vs N° de datos en el análisis de estructura de si fallará una máquina con 1 día de desfase (Elaboración propia)

El índice de capa crece cada vez que se aumenta el número de datos, mostrando que para este caso más datos mejoran la predicción general. Respecto al índice de Kappa, está cercano a 1, por lo que la predicción parece estar relacionada con la falla.

Un elemento importante a destacar, es el caso del componente 2, el cual en el análisis de los errores, en donde se usó una base de construcción de 20.000 datos y la misma base de prueba, presentó una predicción del 100.0%, en cambio en este caso la predicción para el mismo número de datos fue de un 97,0%, incluso si es que se crea el árbol con los 20.000 datos e impureza 0, el resultado del componente 2 es de un 97,0%, por lo tanto se puede decir que para este caso agregar más variables al análisis no es mejor, por lo tanto es importante hacer análisis previos de las variables, para saber si efectivamente aportan a la predicción de la variable que se está buscando.

5.5.5. Análisis de estructura si fallará una máquina con 2 días de desfase

Una vez hecho el análisis de la estructura de si fallará una máquina con 1 día de desfase, hay que verificar cómo se comportan los resultados si se aumenta el desfase a 2 días. Para conseguir este objetivo, se compararán los resultados usando 10.000, 15.000, 20.000 y 25.000 datos para construir el árbol de regresión y clasificación con el desfase de 2 días.

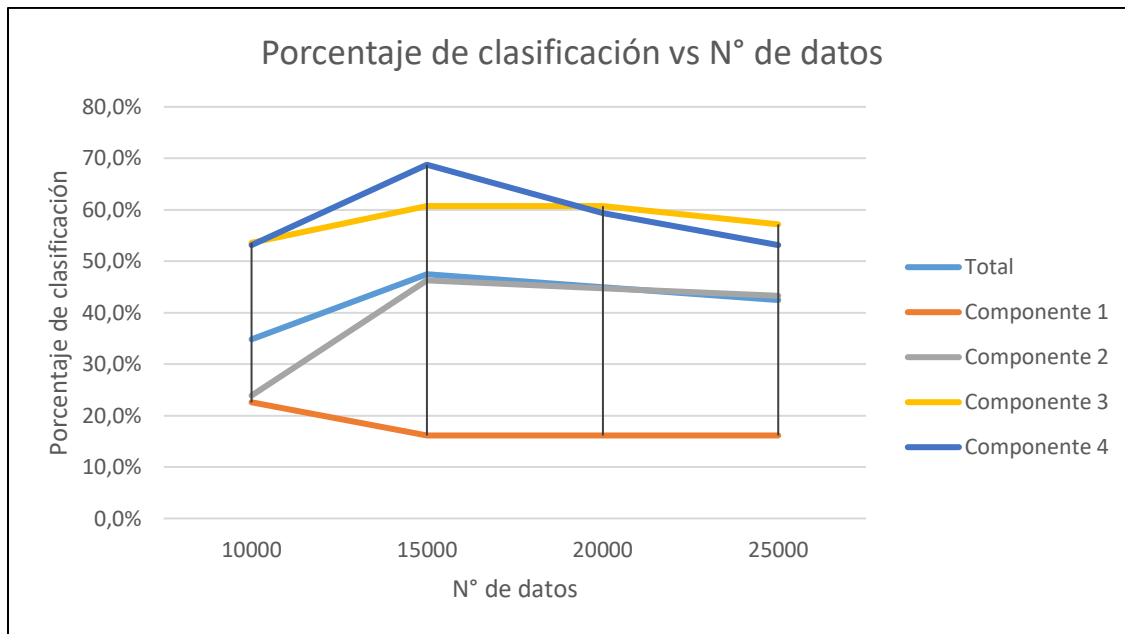


Gráfico 11: Porcentaje de clasificación vs N° de datos con 2 días de desfase (Elaboración propia)

En el porcentaje de clasificación correcta de los componentes, se puede ver que el mejor de los casos es el de 15.000 datos para construir el árbol, mostrando que, a diferencia del caso de desfase de 1 día, el mejor de los casos no es necesariamente el del árbol que se construye con más datos, lo cual lleva a pensar que no siempre más es mejor en esta metodología. Al hacer el análisis por componente la respuesta es la misma a excepción del componente 1, el cual tiene sus mejores resultados en el árbol con 10.000 datos. Finalmente, solo queda destacar que el componente con mejores resultados es el componente 4, mostrando que para este caso los datos que se entregan se acomodan mejor, pero aun así no se puede hablar de una buena predicción, ya que el valor más alto es de un 68,8%. Con esta información, se solidifica el hecho de que hay que hacer un análisis previo antes de usar la herramienta, ya que lo más probable es que en este caso como las variables clasificadoras no se acoplan bien, aumentar la cantidad de datos que se ingresan solo aumenta las posibles respuestas y perjudica a la predicción.

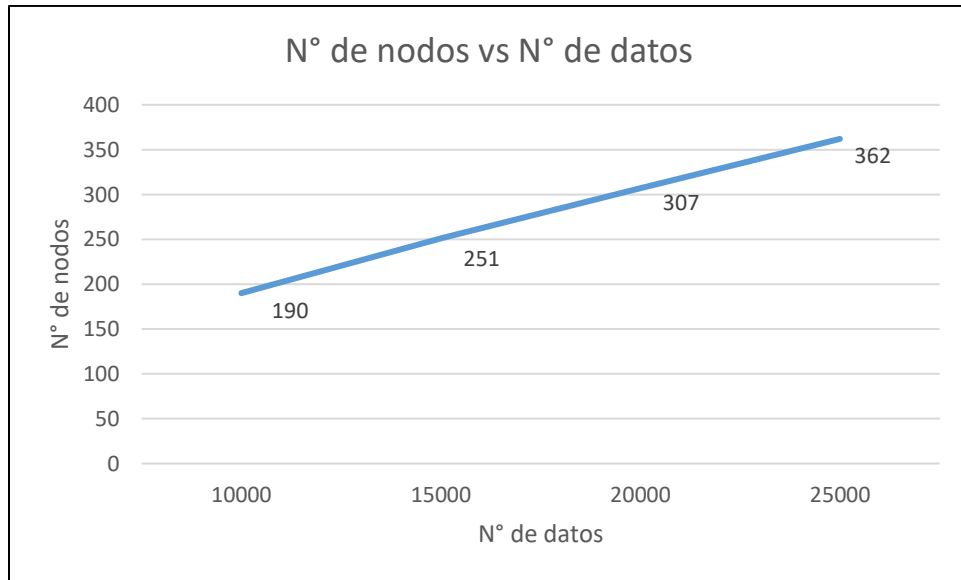


Gráfico 12: N° de nodos vs N° de datos con 2 días de desfase (Elaboración propia)

Como se puede apreciar en el gráfico de los nodos, hay un crecimiento de estos a medida que aumentan los datos de construcción al igual que en el caso anterior, pero los valores son mayores, siendo el de 10.000 datos con 1 día de desfase 65 nodos y el de 2 días de desfase 190 nodos. Esto se puede ver explicado por el hecho de que como no hay una buena relación entre las variables, cuesta llegar a nodos homogéneos, produciendo una división excesiva.

Por último, queda analizar el índice de Kappa y el índice Sensitivity para esta estructura con 2 días de desfase.

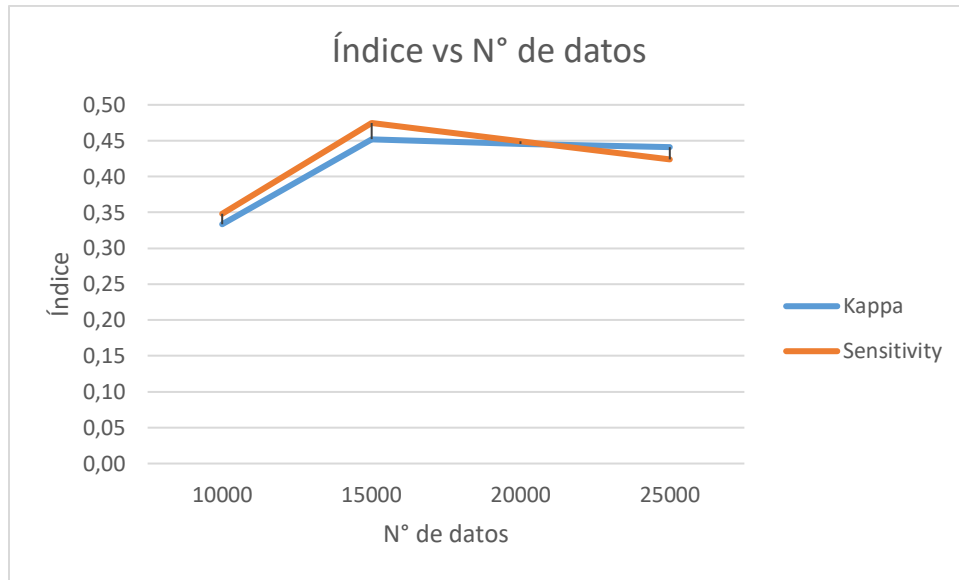


Gráfico 13: Índice vs N° de datos en el análisis de estructura de si fallará una máquina con 2 días de desfase (Elaboración propia)

Se comprueba lo dicho anteriormente, el tamaño con la mejor predicción es el de 15000 datos y por lo tanto no necesariamente más datos implican una mejor predicción. Respecto al valor del índice de Kappa, este es mayor a 1, mostrando que hay una concordancia, pero no es alta, lo cual se contrapone con lo obtenido en un día de desfase donde los resultados fueron mucho mejores.

A continuación, se muestra una tabla que contiene los valores del índice de Kappa las pruebas ejecutadas, para así resumir los resultados.

Tabla 30: Tabla de potencia de predicción (Elaboración propia)

Prueba	1 día de desfase	2 días de desfase
Análisis de errores (20.000 datos)	0,59	-0,01
Análisis de mediciones (20.000 datos)	0,16	0,19
Análisis de días desde una mantención o falla (20.000 datos)	0,14	0,17

Análisis de la estructura de si fallara una máquina (10.000 datos)	0,86	0,33
Análisis de la estructura de si fallara una máquina (15.000 datos)	0,88	0,45
Análisis de la estructura de si fallara una máquina (20.000 datos)	0,93	0,45
Análisis de la estructura de si fallara una máquina (25.000 datos)	0,95	0,44

Para finalizar esta sección, se agrega el árbol de 10.000 datos con 1 día de desfase completo.

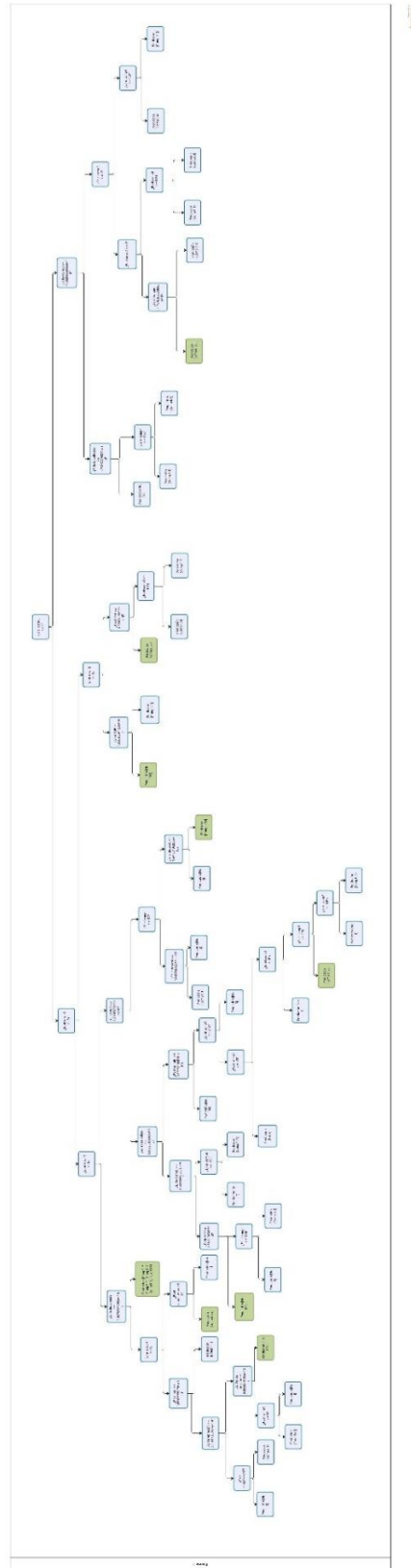


Ilustración 33: Árbol de clasificación con 10.000 datos y 0,001 de impureza (Elaboración propia)

5.5.6. Análisis de la estructura de tiempo sin fallas

En el análisis de esta estructura, hay un cambio importante, ya que la variable a clasificar es cuantitativa, por lo que ya no se utilizará el índice de Gini, sino que la varianza. Junto con lo anterior, la cantidad de datos disponibles se reduce considerablemente, ya que para esta estructura se utilizan solo las fallas, dejando un total de 399 datos para la construcción del árbol y 171 datos para la prueba del mismo, lo que es un 70% para la creación y un 30% para la prueba.

De los datos de prueba se puede obtener que hay pronósticos mayores, menores e iguales a el valor real, lo que en simples palabras se puede explicar, como el hecho de que la falla es después de lo que corresponde, antes de lo que corresponde y justo el día correspondiente.

Tabla 31: Predicción de la estructura de tiempo sin fallas (Elaboración propia)

N° de datos		
Después	85	49,7%
Antes	38	22,2%
Justo	48	28,1%
	171	

De la tabla anterior se puede ver que en el 50,3% de los casos la máquina recibirá la mantención, pero solamente en el 28,1% de los casos el pronóstico acertó al valor. De la información anterior, surge la duda en el caso de los pronósticos que indican que la falla se produce antes, ya que es importante saber que tan antes indica el pronóstico que falla la máquina.

Tabla 32: Rangos de los días anticipados de mantenimiento (Elaboración propia)

	N° de datos	
[-10,0[	0	0,0%
[-20,-10[	15	39,5%
[-30,-20[	8	21,1%
Menores	15	39,5%
	38	

En la imagen anterior, se puede apreciar que en el 100% de los casos se muestra que la falla es a lo menos 10 días antes de la falla real, además en cerca del 40% de los datos, se indica que la predicción es antes de 30 días de la falla real.

Los resultados anteriores, no parecen ser alentadores, pero no se puede indicar que es producto de la herramienta, ya que puede ser que falten datos o que dependiendo a qué tan largo plazo, el indicar cuando va a fallar la máquina es complicado, tomando en cuenta que los datos de las mediciones son independientes del componente, por lo que la falla de otros componentes dentro del periodo de medición puede ensuciar la predicción. Una forma de comprobar esto sería tomar mediciones que sean exclusivas de cada componente dentro de lo posible, para evitar que los efectos producidos por los otros componentes afecten a la predicción, si es que hay establecida una dependencia funcional entre estos.

Ahora, se analizará el cambio de los resultados producto de un aumento de la impureza.

Para hacer este análisis se utilizarán impurezas de 500, 1.000 y 1.500 (Las impurezas son mayores en este caso, porque no se está utilizando el índice de Gini) y se mostraran los resultados para cada una de ellas.

Se puede comenzar analizando los datos de antes, después y justo.

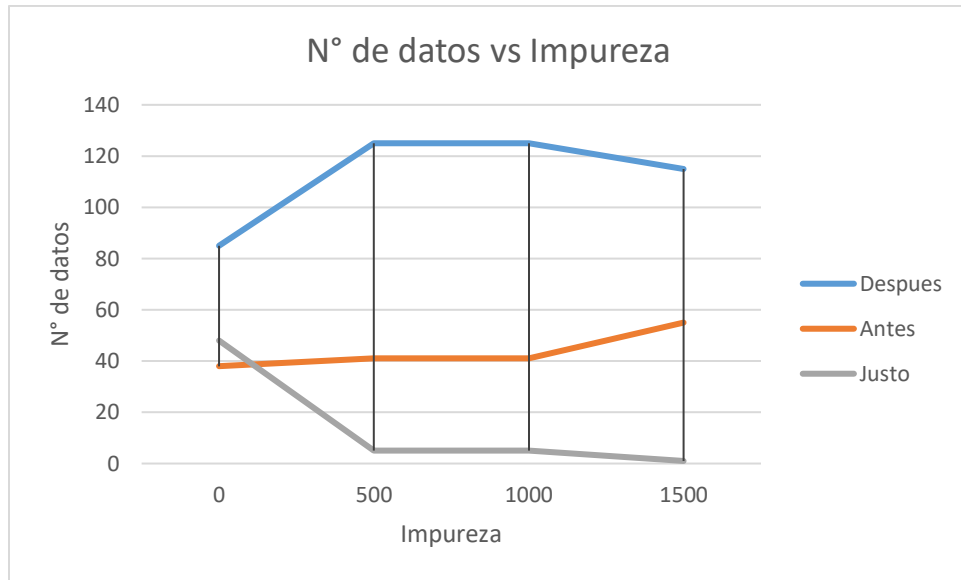


Gráfico 14: Tiempo de la predicción (Elaboración propia)

Se puede ver que los datos de antes y después son claramente mayores que los valores de justo (A excepción del caso de pureza máxima) tomando en cuenta que este último solo tiene un valor alto con impureza 0, en cambio en los casos impureza 500 y 1.000 su valor cae a 5 datos. Lo anterior se entiende tomando en cuenta que, si se permite impureza en los nodos, el valor del nodo se calcula como el promedio ponderado de los valores que contiene el nodo.

Actual: %s. Predicted: %s60{225: '0%', 195: '0%', 100: '0%', 165: '0%', 135: '4%', 105: '6%', 75: '16%', 45: '34%', 150: '1

Ilustración 34: Nodo con impureza (Elaboración propia)

En la imagen anterior se muestra un nodo y se puede apreciar que hay varios resultados con sus respectivos porcentajes, por lo cual lo que se hace para obtener un valor único es multiplicar cada valor con su porcentaje y sumarlos todos. Con esto aclarado, se explica que los valores justos sean menores, tomando en cuenta que el resultado es poco probable que sea exacto. Siguiendo en esta misma línea, se puede juzgar la calidad de la predicción, por la cantidad de días que tiene de desfase con el valor real.

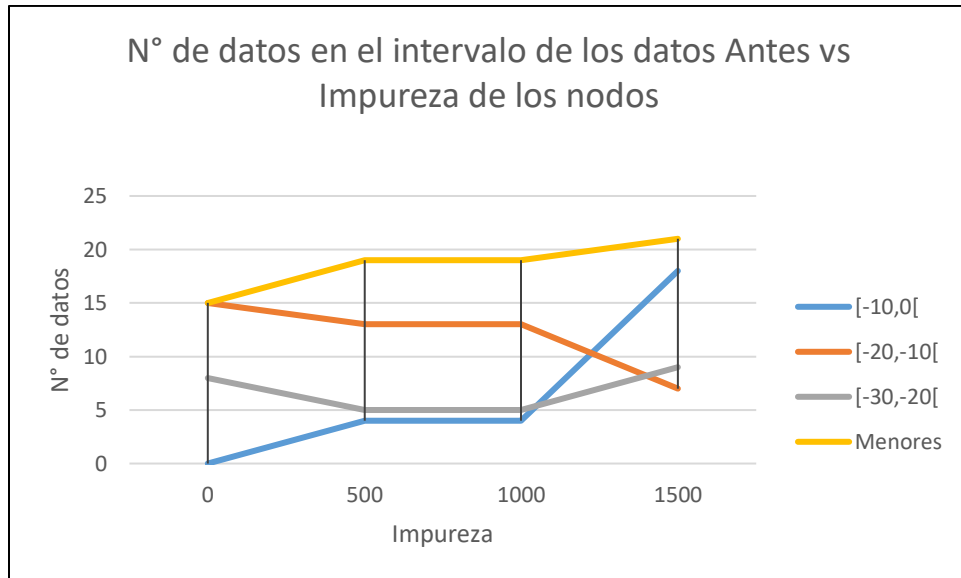


Gráfico 15: N° de datos en el intervalo de los datos Antes vs Impureza de los nodos (Elaboración propia)

Se puede ver que, en el intervalo de 0 a 10 días, sin impureza, no hay ningún dato, pero cuando se trabaja con más impureza, aumenta la cantidad de valores, pero no en la misma proporción que la caída de los valores justos. Finalmente queda indicar que la mayor cantidad de las predicciones indican que la falla es 30 días o más antes de la falla real.

Como análisis complementario se puede ver la caída de los nodos.

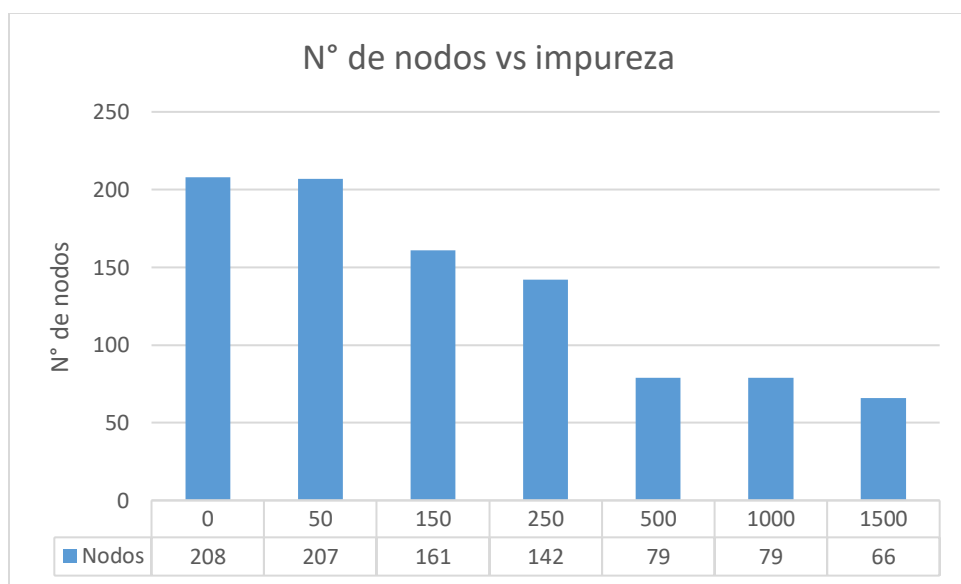


Gráfico 16: N° de nodos vs impureza (Elaboración propia)

Para poder ver mejor los cambios, se agregan algunos valores intermedios de impureza. Se puede apreciar que los valores caen, pero no tan fuertemente en un principio como en los casos antes expuestos.

6. CONCLUSION

Uno de los principales elementos que se pueden rescatar del presente documento es la forma de trabajar la base de datos, ya que lo que se muestra es la posibilidad de utilizar la herramienta para obtener la información o el análisis que se requiere, pero no solo basta con saber cómo estructurar la base, sino también es relevante saber si dicha estructura rinde frutos y en un caso más específico que la herramienta entregue una predicción correcta. Para dar respuesta a esta pregunta, se usó una base de datos de prueba y se descubrió que a pesar de que el índice Accuracy es alto, no se puede decir que la predicción es buena, ya que esto depende de la composición de la variable a clasificar, ya que para el caso expuesto como la densidad de valores “No” es muy alta, es fácil que una mala predicción solo entregue valores “No” y por ende el resultado de la predicción muestre valores correctos (Base de datos desbalanceada). Es por lo anterior que es relevante determinar qué elementos son los que nos indican si se cumple el objetivo o no y para este caso se puede usar el índice Sensitivity y el índice de Kappa.

Otro elemento a destacar fueron los análisis independientes de las variables clasificadoras, ya que de esto se puede concluir que más variables no necesariamente mejoran la predicción, ya que en el caso particular del componente 2 en el análisis de la estructura de si fallará una máquina con 1 día de desfase, el porcentaje de acierto fue de un 97%, en cambio cuando se analizaron los errores independientemente se logró un 100% de acierto, mostrando que a pesar de que con la herramienta se estaba llegando a bajos niveles de impureza, esto no se relacionaba con una mejor predicción. Es así como se reafirma la postura de hacer estos

análisis previos, para determinar las predicciones independientes y así evitar que se presenten estos casos.

En los análisis de las impurezas, se muestra que permitir más impureza, disminuye la cantidad de nodos, ya que el nodo requiere de menos particiones para llegar a su pureza ideal. Este análisis sumado al contraste entre la impureza y los porcentajes de predicción nos permite afirmar que es posible permitir niveles de impureza, para así reducir los nodos sin afectar significativamente a los niveles de predicción (Relación del tamaño de árbol optimo). Es en este punto donde hay que tener especial cuidado con la variable a clasificar, ya que puede darse la situación que ciertos valores se presenten en situaciones similares y al disminuir el grado de pureza, se pierdan en un solo nodo. Este fenómeno se presentó en la estructura de si fallará una máquina con 1 día de desfase en el componente 3, el cual después de ciertos niveles de impureza se unía a un nodo de valores “No”, llevando su predicción total a 0. Es en este contexto que hay que verificar que esto no suceda al disminuir la impureza, ya que si estamos hablando de un grupo de respuesta que se presenta con poca frecuencia se puede perder y no se vería en el análisis general, por lo tanto, se recomienda tomar estadísticas de todas o la mayor parte de las respuestas. Ligado a esto se puede pensar que, en el caso de las variables a predecir continuas, es difícil llevar un conteo de todos los valores, pero si se puede hacer una separación de los mismos en grupos identificables tal como se mostró de forma muy simplificada en el caso de la estructura del tiempo sin fallas.

Un análisis que entregó resultados deficientes fue el de la estructura del tiempo sin fallas, donde las respuestas que indicaban hacer el mantenimiento en el momento correcto, no superaron el 70%. Para este caso se plantea la tesis de que posiblemente es culpa parcialmente de las mediciones, ya que no son para cada componente, sino en general, lo que

puede desembocar que la información que entregan esté relacionada con la falla de otro componente. Es por esto por lo que como propuesta se podría analizar la factibilidad técnica de analizar los componentes por separado y ver que tanto aporta este cambio a la predicción final.

Pasando al uso de la herramienta para determinar la causa de las fallas, se llega a la conclusión de que la herramienta funciona como un buen apoyo visual que muestra todas las preguntas que se hacen y sus respuestas. Es así como se podría relacionar el árbol con la realidad, buscando porqué el grupo de condiciones que se presentan desembocan en una falla y como se podría dar solución a estas, para así aumentar el tiempo de vida de los componentes.

Con lo anterior, recordando las preguntas que se hicieron en el problema de investigación, se puede decir que la herramienta cumple con ser un buen elemento gráfico y en algunos casos cumple en lo referente al pronóstico, por lo que ciertamente puede ser visto como un complemento a las formas de mantenimiento actuales. Sumado a lo anterior, se destaca la forma de uso de la herramienta, la cual permite ingresar varias variables que comúnmente no son fáciles de considerar, por lo tanto, existe la posibilidad tomando en cuenta las nuevas tecnologías de tomar cada vez más y nuevos datos para ser usados.

Finalmente, solo queda hablar sobre algunos elementos que se podrían abarcar o podría ser objetivo de estudio en el futuro, como probar bases de datos que tengan una estructura distinta a la planteada o bases de datos más completas indicando que datos efectivamente vale la pena recopilar y cuáles no. También se podrían buscar nuevos criterios de particionamiento y probarlos para ver si funcionan mejor o cuales se acomodan más a determinados datos.

7. REFERENCIAS

- AENOR. (2013). ISO SURVEY 2013. 32-35.
- Akosa, J. (2017). Predictive Accuracy: A Misleading Performance Measure for Highly. *Oklahoma State University* .
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, vol 16; P412–424.
- Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Assessment over Imbalanced. *Journal of Information Engineering and Applications* , Vol.3, No.10.
- Bertolini, M., & Bevilacqua, M. (2006). Oil pipeline spill cause analysis: A classification tree approach. *Journal of Quality in Maintenance Engineering*, 12(2), 186-198.
- Boylu, F. (28 de 04 de 2018). *Microsoft Azure*. Obtenido de https://gallery.azure.ai/Notebook/Predictive-Maintenance-Modelling-Guide-R-Notebook-1#x_Problem-Description
- Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). Classification and regression trees. *CRC press*.
- Brence, J. R., & Brown, D. E. (2002). Data mining corrosion from eddy current non-destructive tests. *Computers & Industrial Engineering*, 43, 821.
- Cruz, R. d. (17 de 10 de 2017). *Duoc*. Obtenido de <http://www.duoc.cl/ver/noticia/la-importancia-del-mantenimiento-para-las-industrias-productivas?tags=alumnos>
- de León, F. C. (1998). *Tecnología del mantenimiento industrial*. Murcia: Universidad de Murcia.
- Díaz, J. F., & Correa, J. C. (2013). Comparación entre árboles de regreción CART y regresión lineal. *Comunicaciones en Estadística*, 6(2), 175-195.
- Gonzalez, H. (18 de 12 de 2015). *calidadgestion* . Obtenido de <https://calidadgestion.wordpress.com/2015/12/18/iso-550012014-gestion-de-activos/>
- Gordon, J. (2018 de 04 de 2018). *Youtube*. Obtenido de https://www.youtube.com/watch?v=cKxRvEZd3Mw&index=1&list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal
- Gordon, J. (13 de 04 de 2018). *Youtube*. Obtenido de https://www.youtube.com/watch?v=LDRbO9a6XPU&list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal&index=8

- Gorman, M. (19 de Marzo de 2018). *informs*. Obtenido de <http://connect.informs.org/railway-applications/awards/problem-solving-competition>
- Green, A. (1997). Kappa Statistics for Multiple Raters Using Categorical Classifications. *Proceedings of the 22nd annual SAS User Group International conference*, Vol. 2, p. 4.
- Heras-Saizarbitoria, I., & Boiral, O. (2013). ISO 9001 and ISO 14001: towards a research agenda on management system standards. *International Journal of Management Reviews*, 15(1), 47-65.
- ISO. (2016). *The ISO Survey of Management System Standard Certifications 2016*.
- ISO. (22 de 11 de 2017). ISO. Obtenido de <https://www.iso.org/the-iso-survey.html>
- kaggle. (22 de 05 de 2018). *kaggle*. Obtenido de <https://www.kaggle.com/datasets?sortBy=relevance&group=public&search=maintenance&page=1&pageSize=20&size=all&filetype=all&license=all>
- kaggle. (02 de 06 de 2018). *kaggle*. Obtenido de <https://www.kaggle.com/yuansaijie0604/xinjiang-pm/data>
- Kaggle. (28 de 04 de 2018). *Kaggle*. Obtenido de <https://www.kaggle.com/yuansaijie0604/xinjiang-pm>
- León, F. C. (1998). *Tecnología del mantenimiento industrial*. Murcia : Universidad de Murcia .
- Luo, R., Hsieh, T., Su, K., & Tsai, C. (2002). An intelligent remote maintenance and diagnostic system on mobile robot. In *IECON 02 [Industrial Electronics Society, IEEE 2002 28th Annual Conference of the]*, 2675-2680.
- Molina, J. (2006). Mantenimiento y seguridad industrial. *Monografías*.
- Olarte, W., Botero, M., & Cañon , B. (2010). Técnicas de mantenimiento predictivo utilizadas en la industria. *Scientia et technica*, 2(45), 224-226.
- Olarte, W., Botero, M., & Cañon, B. (2010). Importancia del mantenimiento industrial dentro de los procesos de producción. *Scientia et technica*, 1(44), 354.
- Serna Pineda, S. C. (2009). *Comparación de árboles de regresión y clasificación y regresión logística*. (Doctoral dissertation, Universidad Nacional de Colombia).
- Timofeev, R. (2004). Classification and Regression Trees (CART) Theory and Applications. *Humboldt University, Berlin* .
- Trujillano, J., Sarria-Santamera, A., Esquerda, A., Badia, M., Palma, M., & March, J. (2008). Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. *Gaceta Sanitaria*, 22(1), 69.

- Upton, A., Jefferson, B., Moore, G., & Jarvis, P. (2017). Rapid gravity filtration operational performance assessment and diagnosis for preventative maintenance from on-line data. *Chemical Engineering Journal*, 313, 250-260.
- Yang, B. (2009). Data-driven approach to machine condition prognosis using least square regression tree. *Journal of mechanical science and technology*, 23(5), 1468-1475.