Identificando Patrones de Predicción y Clasificación de Alarmas por Alto Spread en un Sistema de Combustión de Turbina a Gas

Roberto Prieto Morales
Ingeniero de Proyectos de Tecnología de la Información
Magíster en Ingeniería Informática UCN
Antofagasta, Chile
robertoprieto@vtr.net

Claudio Meneses Villegas
Departamento de Ingeniería de Sistemas y Computación
Universidad Católica del Norte
Antofagasta, Chile
cmeneses@ucn.cl

***Abstract*— En este artículo se analizan y modelan datos operacionales de un sistema de combustión de turbina a gas, para clasificar y predecir la condición de "alarma por alto spread". Esta condición de alarma indica que la combustión de la turbina no está siendo uniforme, lo cual puede llevar eventualmente a la deformación de la turbina. Con la generación de patrones de predicción y clasificación, se busca anticiparse a la activación de la alarma por alto spread en el sistema de combustión de la turbina, con lo cual se podría evitar o disminuir la indisponibilidad de la turbina. Mediante el entrenamiento de algoritmos de redes neuronales y árboles de decisión se obtuvieron dos modelos de clasificación y un modelo predictivo, los cuales fueron evaluados cuantitativamente y en base a la percepción de los usuarios, siendo los modelos de árbol mejor evaluados en este último aspecto.**

Keywords- Alarma por Alto Spread; Sistema Combustión de Turbina a Gas; Identificación de Patrones en Sistemas de Combustión

I. INTRODUCCIÓN

Actualmente las organizaciones están inmersas en un mercado muy competitivo, por lo que es importante para ellas, que sus ejecutivos posean información relevante y oportuna a la hora de tomar decisiones. Dentro de las herramientas que ocupan las organizaciones para apoyar la toma de decisiones, está la minería de datos.

Fayad (1996), define minería de datos como la búsqueda de patrones relevantes y de regularidades importantes en grades almacenes de datos [11]. Por otro lado, Michalski (1998) se refiere a minería de datos inteligente como la aplicación de métodos de aprendizaje automático u otros métodos similares, para descubrir y enumerar patrones presentes en los datos [12]. El aprendizaje automático es el área de la Ingeniería Informática, que estudia y desarrolla algoritmos que implementan distintos modelos de aprendizaje, y lo aplican en la resolución de problemas prácticos [16].

La minería de datos, se presenta como una etapa dentro de un proceso más amplio, que se refiere a la aplicación de algoritmos específicos para la extracción de patrones desde datos. Dicha etapa es parte del proceso de descubrimiento de conocimiento desde los datos, conocido como proceso KDD (Knowledge Discovery in Databases).

El término KDD fue acuñado por Piatetsky Shapiro (1989) [1] para enfatizar que el "conocimiento" es el producto final del descubrimiento accionado por los datos.

El conocimiento extraído, es muy valioso para las organizaciones a la hora de tomar decisiones. Para tomar decisiones correctas, confiables y acertadas se debe contar con la información adecuada [7].

Este artículo describe la aplicación de un proceso metodológico basado en la guía CRISP-DM (Cross Industry Standard Process for Data Mining), para la resolución de un problema operacional de alto impacto en una Central Generadora de Electricidad.

La aplicación de la metodología CRISP-DM en una Central Generadora de Electricidad busca entregar información a la empresa para apoyar la toma de decisiones. En términos concretos, se busca anticiparse al problema de la activación de la protección por alto spread del sistema de combustión de la TG (Turbina a Gas).

El resto del artículo está estructurado como se indica a continuación. La sección II describe el problema y su contexto. En la sección III se describe la metodología empleada en el desarrollo del trabajo. Las secciones IV y V describen los datos utilizados, su selección y transformación, respectivamente. La sección VI presenta y analiza los modelos de predicción y clasificación obtenidos. Finalmente, se presentan las conclusiones y trabajo futuro.

II. SISTEMA DE COMBUSTION DE TURBINA A GAS

El presente caso de estudio se llevó a cabo en una Central Generadora de Electricidad perteneciente al SING (Sistema Interconectado del Norte Grande), cuya matriz energética es gas y petróleo. Esta Central Generadora, posee dentro de sus objetivos tener la máxima disponibilidad posible para todas sus TG.

Por lo anterior, resulta imperioso para la organización trabajar en evitar fallas en sus TG, que puedan producir alguna indisponibilidad de las TG en la generación eléctrica.

En lo relacionado al sistema especifico objeto del estudio, éste corresponde a la turbina a gas, la cual es la principal máquina para generar electricidad que posee la organización. Esta turbina está compuesta por los sistemas de escape, enfriamiento y combustión.

En la Figura 1 [9], se puede apreciar las principales partes de la TG, separadas en dos partes, generación a gas y generación a energía.

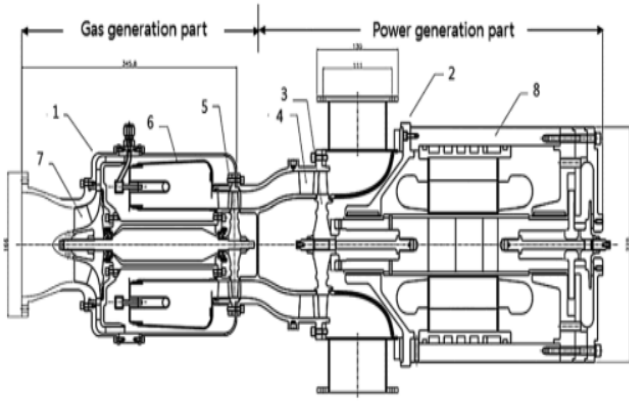

Figura 1. Corte longitudinal de la Turbina a gas. 1. generador de gas, 2. generador de energía, 3. álabe de turbina de potencia, 4. poder tobera turbina, 5. generador de turbina de gas cuchilla, 6. revestimiento de combustión, 7. impulsor del compresor, 8. el generador de estructuras.

El sistema de combustión de la turbina está compuesto por catorce cámaras de combustión, las cuales están dispuestas simétricamente alrededor del rotor de la turbina.

Periódicamente se registran mediciones de variables asociadas a la temperatura de las cámaras del sistema de combustión. Dichas temperaturas deben mantenerse uniformes, debido a que una alta diferencia de temperatura entre las cámaras de combustión, significaría que no se está produciendo una combustión eficiente y segura.

Por lo anterior, implicaría que se podría provocar una deformación en la estructura de la turbina, además de una pérdida de eficiencia en la capacidad de generación de electricidad.

El Sistema de Control Distribuido DCS (Distributed Control System) monitorea y controla el funcionamiento de la TG. En el DCS se encuentra configurada una protección por alto spread en el sistema de combustión, la cual opera al identificar una alta diferencia de temperatura entre las cámaras de combustión de la TG.

La protección actúa en primera instancia alarmando el alto spread. El sólo surgimiento de esta alarma, implica una disminución en la capacidad de generación en la TG, debido a que no son uniformes las fuerzas que hacen girar el eje del generador. En segunda instancia la alarma por alto spread, opera deteniendo el funcionamiento de la turbina, lo cual implica una indisponibilidad de la TG, conllevando a una pérdida de confianza ante sus clientes y la comunidad.

III. Aspectos Metodológicos del Desarrollo del Proyecto

El proyecto se abordó adoptando la guía CRISP-DM como marco de desarrollo del trabajo, el cual se instanció para este caso particular.

A. Guía Metodológica CRISP-DM

CRISP-DM en esencia corresponde a un modelo de proceso que proporciona un marco para el desarrollo de proyectos en el ámbito de Data Mining [8]. El cual, está siendo desarrollado por un consorcio de los principales usuarios y proveedores de minería de datos.

Este modelo de referencia, proporciona una visión general del ciclo de vida de un proyecto de minería de datos, el cual contiene las fases de un proyecto, sus tareas respectivas, y sus salidas.

El ciclo de vida de un proyecto de minería de datos se divide en seis fases que se muestran en la Figura 2.

La secuencia de las fases no es estricta, y en la práctica es un proceso iterativo. Las flechas indican sólo las secuencias y las dependencias más importantes entre las fases.

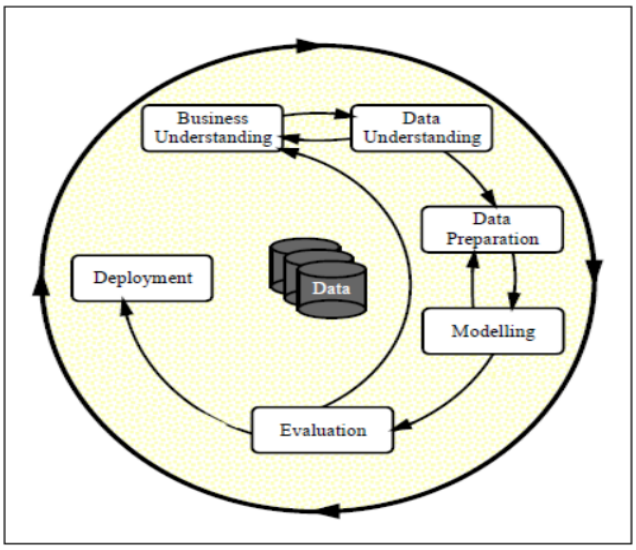

Figura 2. Fases del modelo de proceso CRISP-DM para el desarrollo de proyectos de minería de datos. (Fuente: http://www.crisp-dm.org/).

B. Instanciación de la Guía CRISP-DM

La guía CRISP-DM corresponde a un proceso genérico, el cual requiere ser instanciado para cada tipo de proyecto de data mining. A continuación se presenta la adaptación de CRISP-DM al problema particular descrito en la sección II y su aplicación.

La Figura 3 muestra las tareas desarrolladas para el problema analizado.

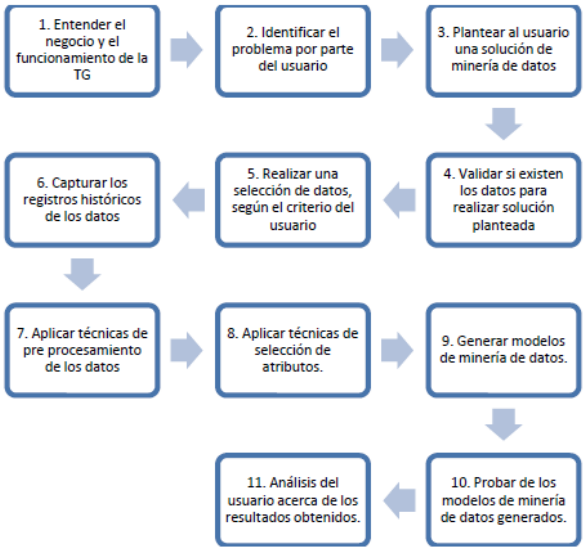

Figura 3. Secuencia de tareas desarrolladas

1) Entendimiento del negocio

El objetivo de negocio que se persigue con la realización del proyecto es mejorar el indicador de disponibilidad de la TG, mediante la aplicación de técnicas y herramientas de minería de datos.

Específicamente, se busca contribuir a maximizar la confiabilidad de arranque, minimizar salidas forzadas, lograr disponibilidad de acuerdo a programa de mantenimiento, disponibilidad media equivalente (92%), cumplir con la duración de los mantenimientos mayores.

Por lo anterior, se desea evitar las fallas o indisponibilidad de la TG por un tiempo prolongado, tal que, no se ponga en riesgo los contratos vigentes, y lograr menor índice de falla en el SING.

Desde el punto de vista técnico, los objetivos al aplicar las técnicas de minería de datos en esta situación particular, son generar patrones de predicción y clasificación, para apoyar la toma de decisiones, asociadas a evitar la activación de la protección por alto spread en el sistema de combustión de la TG.

La Turbina a Gas, es un motor térmico rotativo de flujo continuo que se caracteriza por presentar una baja relación peso-potencia y una velocidad de giro muy elevada. La TG está compuesta por los sub sistemas de combustión, enfriamiento y escape.

Se utiliza para la generación eléctrica, ya que la combustión generada al incinerar gas, hace que los gases calientes al escapar hagan girar el rotor del alternador de la TG, con lo cual se produce inducción, la cual se trasforma en energía eléctrica.

A continuación en la Figura 4 [10], se muestra el esquema de funcionamiento de la TG para generar electricidad.

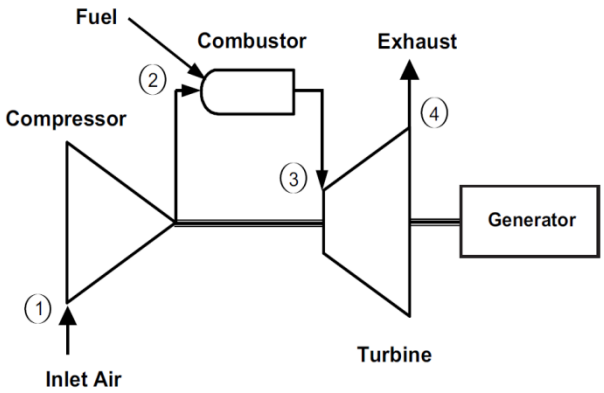

Figura 4. Esquema de funcionamiento de turbina a gas.

En la Figura 5 se describen en mayor detalle los pasos de la secuencia de funcionamiento de una TG.

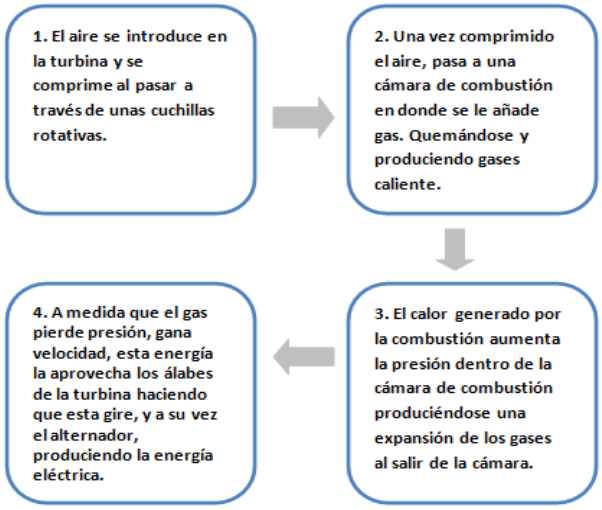

Figura 5. Secuencia de funcionamiento de turbina a gas.

2) Descripción del problema

El sub sistema de combustión de la TG se compone por catorce cámaras de combustión, en donde se incinera el gas natural o petróleo diesel, estas cámaras se encuentran distribuidas simétricamente alrededor del rotor de la TG.

Existe una protección configurada en el DCS, esta se produce por alto spread en el sistema de combustión de la TG. Esta protección se calcula empíricamente tomando como datos entre otros, las temperaturas de las cámaras de combustión de la TG. La activación de dicha protección implica que no se está produciendo una combustión uniforme entre las cámaras de combustión.

Lo anterior, conlleva a una disminución de la generación de electricidad de la TG, es decir que con la misma cantidad de combustible se genera menos carga eléctrica que en una situación óptima. Además, la generación de electricidad con una combustión no uniforme entre las cámaras provoca daños en la estructura de la TG, deformando sus piezas.

Actualmente, cuando opera la alarma por un alto spread el operador disminuye la potencia eléctrica de la TG, hasta encontrar la causa y corregir el problema.

Se propone analizar los datos patrones de predicción y clasificación, para anticipar la operación de la protección por alto spread en el sistema de combustión de la TG.

Existe antecedente de aplicación de técnica de minería de datos en Centrales de Generación Eléctrica para identificar otras fallas en equipos [14].

IV. ENTENDIMIENTO DE LOS DATOS

Para la realizar la solución de minería de datos propuesta, es necesario contar con registros históricos, que posibiliten la creación de patrones.

La Central, tiene implementada la plataforma industrial de gestión de información "PI SYSTEM". Esta plataforma está compuesta por software, que permiten mostrar datos de proceso en tiempo real y almacenarlos en una base de datos propietaria.

La TG está compuesta por los sub sistemas escape, enfriamiento y combustión, también existen señalen que influyen en la generación eléctrica como los equipos auxiliares, aparte de las señales propias de la generación eléctrica como la potencia eléctrica y la frecuencia.

Según el usuario del negocio como el alto spread se origina en el sub sistema de combustión de la TG, se seleccionaron todas las señales de ese sub sistema, aparte de señales que dependen directamente de la activación de la alarma por alto spread, como la potencia eléctrica y la frecuencia.

La inclusión de cualquier otra variable adicional no tendría relación con el alto spread de la TG. Por lo cual, sólo produciría ruido en la elaboración de los modelos.

Al conjunto de datos resultante, se agregó manualmente el atributo clase. Este atributo corresponde a la activación de la alarma por alto spread en el sistema de combustión de la TG. Dicho atributo es de tipo numérico, codificándose como el valor 0 para condición sin alarma y 1 para condición con alarma.

Para el presente caso de estudio, se seleccionaron registros históricos a partir de enero del 2008 y hasta diciembre del 2010. Por lo cual, el conjunto de datos final seleccionado por el usuario del negocio, incluye 54 variables, todas de tipo numérica. En la tabla 1 se muestra un resumen del conjunto de datos seleccionado.

Característica	**Valor**
Total atributos:	54
Total Instancias:	23430
Tipo de atributos:	Numérico
Total Instancias erróneas:	382

Tabla 1. Resumen del conjunto de datos

En la Figura 6 se describe la tabla de hechos del conjunto de variables seleccionadas.

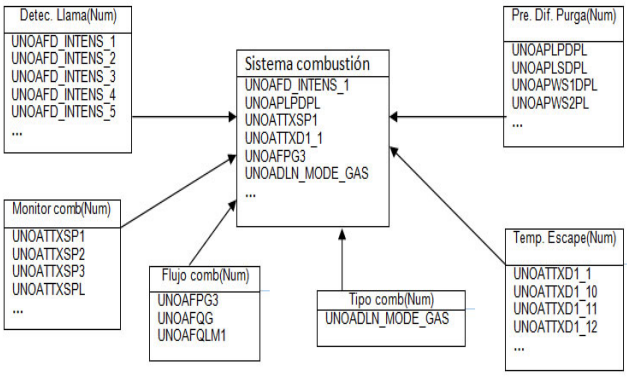

Figura 6. Tabla de hecho.

El conjunto obtenido de datos posee únicamente variables de tipo numérico, correspondiente a valores de variables de proceso como temperaturas de las cámaras de combustión, presión, intensidad de la llama, etc.

En la Tabla 2, se describe una caracterización de algunos atributos del conjunto de datos.

Nombre atributos	**Valores de variables**					
	Instancias distintas	Media	Desviación estándar	moda	mínimo	máximo
UNOADLN_MODE_GAS	16	0.04	0.486	0	0	9
UNOAFD_INTENS_2	3630	102.718	75.307	2012	0	231.245
UNOAFD_INTENS_3	1405	94.448	382.945	844	0	5708
UNOAFD_INTENS_4	1041	72.332	398.402	613	0	3822
UNOAFD_INTENS_5	15285	61.929	42.372	14253	0	162.668
UNOAFLUPSP	7005	0.315	0.899	5756	0	12.401
UNOAFPG3	5114	0.22	1.581	4578	0	39.306
UNOAPLPDPL	2310	0.037	0.281	1821	0	6.286
UNOAPLSDPL	1660	0.042	0.273	1232	0	8.082
UNOATTRF1	5698	27.619	83.892	4154	0	1154.33
UNOATTRXB	3358	7.31	19.405	2013	0	611.219
UNOATTXD1_1	4442	16.772	46.329	3065	0	614.838
UNOATTXD1_10	4360	16.364	45.691	2996	0	615.244
UNOATTXD1_11	4655	16.682	44.216	3277	0	614.838
class	2	-	-	-	-	-

Tabla 2. Caracterización del conjunto de datos.

La figura 7 muestra la relación de variables entre la temperatura de combustión de la cámara número 5 (eje X), y la temperatura de los gases de la cámara de combustión (eje Y) medidas en grados Celsius.

Este gráfico muestra que existe una relación directamente proporcional, es decir a mayor temperatura de combustión, mayor es la temperatura de gases de escape. Lo que implica, que se está realizando una combustión óptima en la cámara N° 5, ya que, que los inyectores de combustibles no se encuentran sucios y la cámara de combustión aún no necesita ser lavada para sacar los residuos de la combustión adheridos a ella.

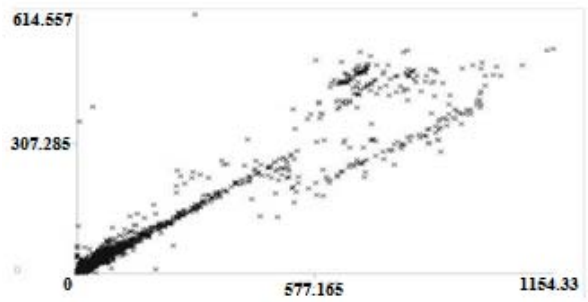

Figura 7. Gráfico temperatura de gases.

En la figura 8 se representa la relación de variables entre la diferencia de temperatura de las cámaras de combustión N° 5 y N° 10 (eje X), y la temperatura de la turbina (eje Y) medida en grados Celsius.

En este gráfico se aprecia que se encuentra delimitado el valor máximo para la diferencia de temperatura entre cámaras de combustión. Para el periodo de tiempo dado la diferencia de temperatura entre las cámaras de combustión N° 5 y N° 10 mayoritariamente fue baja, independiente de la temperatura de la turbina. Lo que implica, que para un funcionamiento normal de la TG, necesariamente debe existir una baja diferencia de temperatura entre las cámaras de combustión.

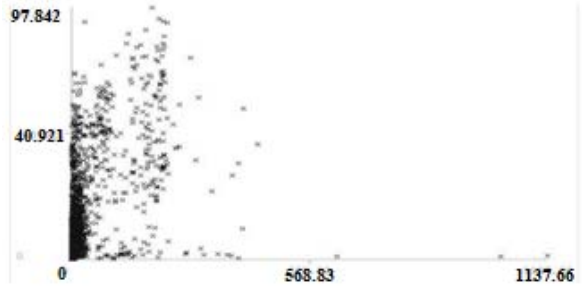

Figura 8. Gráfico diferencias de temperatura.

La tabla 3 describe la matriz de correlación para los atributos más relevantes relacionados con el atributo clase.

Atributos		Valor correlación
Primero	**Segundo**	
Unoafd_intens_5	Unoafsr_control	-0.002
Unoafd_intens_5	Unoal83llbm	-0.002
Unoafd_intens_5	Class	0.112
Unoafsr_control	Class	-0.018
Unoal83llbm	Class	-0.018

Tabla 3. Matriz de correlación de variables.

Las relaciones entre variables descritas en la Tabla 3, corresponden en su gran mayoría a correlación negativa débil. Excepto, la relación entre la temperatura de combustión de la cámara N° 5 y la variable clase. Esta relación es una correlación positiva débil. Por lo cual, la variable 'Unoafd_intens_5' está más correlacionada con el valor de la variable clase que el resto de las otras variables.

V. TRANSFORMACIÓN Y SELECCIÓN DE DATOS

La calidad de los datos es un factor importante en el éxito de la minería de datos en una tarea determinada. Por lo cual, es una etapa relevante dentro del modelo de proceso CRISP-DM. La selección de atributos es el proceso de identificar y eliminar la mayor cantidad de información irrelevante y redundante [2].

Para el presente caso de estudio, se realizó una limpieza de los registros erróneos, eliminando dichos registros. Estos eran producto de fallas en la captura de las señales. Además se verificó la integridad de los datos, no encontrando problemas de este tipo.

Dado que existe una gran cantidad de atributos (cincuenta y cuatro) se aplicó una técnica para evaluar a priori su importancia y disminuir el número de atributos, dejando sólo los más relevantes. A continuación se detalla la técnica utilizada de atributo evaluador y método de búsqueda, para reducir el número de atributo del conjunto de datos.

- Atributo evaluador CfsSubsetEval [3], este método evalúa un subconjunto de atributos considerando la habilidad predictiva individual de cada variable, así como el grado de redundancia entre ellas.
- Método de búsqueda BestFirst [3], este algoritmo de búsqueda, trata de expandir el nodo más próximo al objetivo, percibiendo que probablemente conduzca rápidamente a una solución. El algoritmo BestFirst puede comenzar su búsqueda por el conjunto vacío de atributos y de búsqueda hacía adelante, o empezar con todo el conjunto de atributos y búsqueda hacia tras, o empezar en cualquier momento y buscar en ambas direcciones (enfoque híbrido).

En la tabla 4, se detallan los resultados obtenidos de la aplicación de la técnica de selección de atributos BestFirst.

Resultados aplicación técnica selección de atributos BestFirst	
Inicio establecido:	Sin atributos.
Dirección de búsqueda:	Hacia adelante.
Búsqueda de rancio:	Después de 5 expansiones de nodos.
Número de subconjuntos de evaluación:	410
Mérito del mejor subconjunto:	0.113

Tabla 4. Resumen resultados BestFirst.

La aplicación de esta técnica obtuvo como resultado una disminución de atributos de 54 a 4. La Tabla 5 detalla los atributos seleccionados.

Variable	Descripción
UNOAFD_INTENS_5	Temperatura cámara de combustión N° 5
UNOAFSR_CONTROL	Temperatura de control de la turbina
UNOAL83LLBM	Intensidad de la llama cámara N° 8
Class	Protección

Tabla 5. Atributos seleccionados

VI. Modelos de Clasificación y Predicción

Existen antecedentes de estudios, en el cual se compara la performance de un conjunto de algoritmos de clasificación. Concluyendo que la elección de del algoritmo más adecuado, es altamente dependiente de la aplicación [15]. Además, se debe considerar que no existen antecedentes de aplicación de minería de datos para el problema particular de la alarma por alto spread en un sistema de combustión de la TG.

Para el presente caso de estudio, se desea que el usuario del negocio tenga modelos de varios tipos, para que los analice y escoja el modelo que mejor soluciona el problema planteado. Para lo cual, se aplicarán las técnicas de árbol de decisión y red neuronal artificial.

Los árboles de decisión son una técnica sencilla de aprendizaje de clasificación supervisada, pero exitosa. Los árboles están compuesto por segmentos más pequeños llamados nodos terminales u hojas. Estos nodos son homogéneos respecto a una variable de destino [17].

Las redes neuronales artificiales (ANN) han sido utilizadas por muchos investigadores para identificar ubicación y severidad de distintos tipos de variables de entrada y salida. Ya que, proporcionan una herramienta eficaz para el reconocimiento de patrones [18].

El algoritmo J48 es la implementación para Weka del algoritmo C 4.5. Este algoritmo J48, elige el atributo que posee el máximo de información relacionada con la ganancia, como criterio de la mejor división. Además, utiliza los atributos que mejor diferencia las salidas, generando una rama por cada salida.

El algoritmo RepTree presenta una poda rápida para corregir en el árbol de decisión los efectos de los ruidos en los datos de entrenamiento. El árbol podado reduce la complejidad en el proceso de clasificación.

El algoritmo Multilayer Perceptron posee funciones sigmoidales que se emplean como activación de funciones no lineales para todas las capas. Estas funciones sigmoidales minimizan el sobreajuste con un método de detección temprana.

El overfitting o sobreajuste se produce cuando un algoritmo busca las mejores variables para un modelo en particular, usando un conjunto de datos limitado, puede sobre ajustar los datos, resultando un rendimiento inferior del modelo sobre los datos de prueba [20].

Para la reducir la complejidad y evitar el exceso de sobreajuste, en el presente caso de estudio se ocuparán los clasificadores RepTree y J48, además del Algoritmo de ANN Multilayer Perceptron. Los tres algoritmos seleccionados se encuentran disponibles en Weka.

A. Algoritmo Multilayer Perceptron

Este algoritmo es una ANN (red neuronal artificial) multicapa. Según Rumenlhart (1986) las ANN multicapas de tipo feedforward con aprendizaje por algoritmo de retro propagación, son un tipo de estructura de computación paralela, en donde, varias pequeñas unidades de cálculo denominadas neuronas, están masivamente interconectadas con la capa anterior de donde reciben información, y con la capa posterior hacia donde la transmiten [5].

Las principales características del algoritmo Multilayer Perceptron son su capacidad para aprender las relaciones funcionales a partir de ejemplos, descubrir patrones y regularidades en los datos, a través, de la auto organización. Por lo cual, son muy adecuados para de problemas de mapeo no lineal [19].

En la Figura 9, se muestra la estructura de la red neuronal artificial generada. En donde se aprecia el ingreso de las tres variables a la capa de entrada, dos neuronas en la capa oculta y una neurona en la salida. Además se ilustra la distribución de los ocho pesos sinápticos generados por el modelo de predicción.

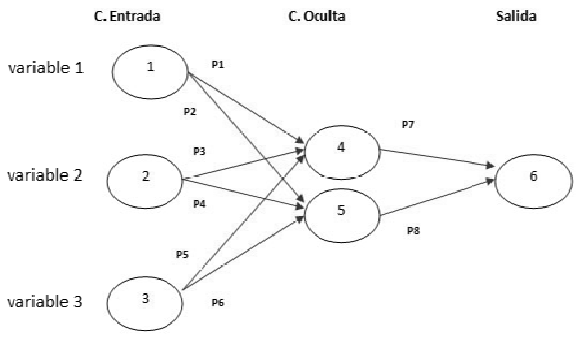

Figura 9. Estructura de la red neuronal articifial.

La ANN empieza con unos pesos aleatorios y aprende ajustando dichos valores iterativamente, hasta llegar a un estado en donde es capaz de determinar una relación funcional dentro de los objetivos preestablecidos entre los datos de entrada y el conjunto de datos de entrenamiento en su fase de aprendizaje.

El ajuste de los coeficientes, se realiza por medio de un mecanismo de retro propagación del error desde las capas de salidas hacia las capas ocultas, para posteriormente cotejar los resultados obtenidos de la salida de la red con el valor correcto entregado por el vector que contiene el conjunto de datos de entrenamiento.

El mecanismo de proceso formal para cada neurona es el siguiente [4]:

(1)

En donde, es la salida de la neurona, es el producto escalar entre el vector traspuesto de pesos sinápticos.

(2)

es el vector de pesos sinápticos.

(3)

es el vector de entradas a la neurona.

El subíndice m indica el número de entradas a la neurona, es un valor denominado umbral que permite ajustarse para disminuir el sesgo.

La función transforma el escalar resultante en la salida de la neurona, en la mayoría de los casos corresponde a una función sigmoidal.

(4)

El resultado de esta función produce salidas dentro del rango [0,1]. La salida de cada neurona, es una entrada para cada neurona de la capa siguiente, excepto en la primera capa, en donde la entrada es el vector con las variables independientes, tal como se aprecia en la Figura 9.

El modelo predictivo elaborado por algoritmo Multilayer Perceptron entregó como resultado la generación de ocho ponderaciones de pesos. Estos pesos ponderan las variables de entrada y empíricamente determinan la variable clase. Esta variable clase es la que indica si se produce el alto spread en el sistema de combustión de la TG, para unas variables de entrada en particular.

B. Algoritmo Rep Tree

El algoritmo RepTree se utiliza para la elaboración de un patrón de clasificación, obteniendo como resultado de la aplicación de dicho algoritmo una representación gráfica de un árbol de clasificación.

El funcionamiento del algoritmo RepTree se compone de dos fases, en la primera fase se crea un conjunto de reglas que se sobreajuste a los datos usados para el aprendizaje, en la segunda fase se poda el conjunto de reglas usando ejemplos que no participaron en el aprendizaje [3].

Para la aplicación de este algoritmo se utilizaron los parámetros por defectos en la construcción el modelo.

En la Figura 10 se muestra el árbol gráfico generado por el algoritmo Rep Tree. La rama izquierda del árbol muestra la parte no relacionada con la condición de alarma, en cambio en la rama derecha del árbol están los indicadores que corresponden a la alarma.

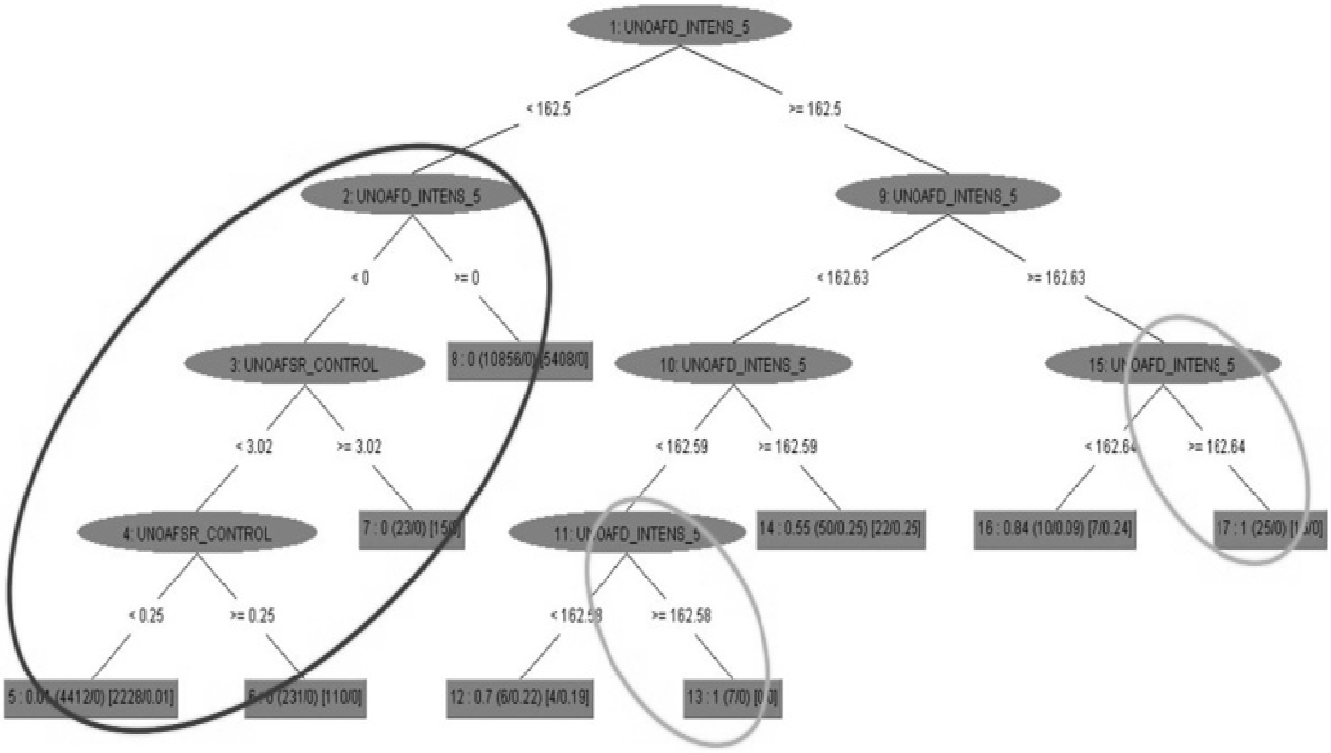

Figura 10: Árbol Gráfico generado por el algoritmo Rep Tree.

C. Algoritmo J48

El algoritmo J48 es una aplicación del algoritmo C4.5. Esta implementación genera árboles de decisión. El algoritmo J48 ocupa una técnica voraz para inducir la decisión para los árboles de clasificación [6]. Este algoritmo, utiliza en la elaboración del árbol de clasificación los atributos que mejor diferencia las salidas, creando una

rama por cada salida [13]. Además, termina la rama si todos los miembros poseen la misma clase, etiquetando la rama con dicha clase [3].

Para la aplicación de este algoritmo se utilizaron los parámetros por defectos en la construcción el modelo.

En la Figura 11, se muestra el patrón de clasificación generado. En la rama derecha del árbol, describe la clasificación de instancias relacionadas con el surgimiento de la condición de alarma.

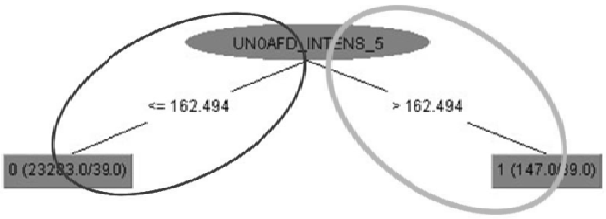

Figura 11: Árbol Gráfico generado por el algoritmo J48.

VII. RESULTADOS

Para la construcción de cada uno de los modelos de minería de datos generados, se ocupó un 95% de las instancias para entrenamiento y un 5% para prueba. Las métricas de rendimiento de cada modelo generado se resumen en la Tabla 6.

Variable	Algoritmos		
	MultiLayer Perceptron	**RepTree**	**J48**
Coeficiente de correlación	0.7395	0.8158	0.7992
Media de error absoluto	0.0045	0.0047	0.0043
Raiz de error cuadrado	0.0495	0.0483	0.0443
Error absoluto relativo	39.3034 %	41.5828 %	37.6059 %
Raiz de error relativo cuadrado	69.3076 %	67.6879 %	61.9838 %
Número de instancias	1171	1171	1171

Tabla 6: Resumen de resultados de modelos.

Una vez obtenidos los modelos de predicción y clasificación, se requirió la evaluación del usuario respecto del potencial grado de interés en cada uno de ellos. A continuación se resume esta evaluación.

A. Modelo de predicción ANN

El modelo de predicción generado por la red neuronal artificial, permite indicar si opera o no la alarma para valores específicos de las variables de entrada. Por lo que, operacionalmente no es una ayuda para anticiparse a la activación de la alarma.

B. Modelo de clasificación Rep Tree

El modelo de clasificación elaborado por el algoritmo Rep Tree clasifica la operación de la alarma, a partir de un rango de valores específicos para una variable en particular. Pero, además presenta un modelo confuso para el cliente, ya que el árbol de clasificación generado posee ramas que no aportan a la clasificación de la alarma.

C. Modelo de clasificación J48

El modelo de clasificación J48, presenta un modelo de clasificación similar al generado por el algoritmo Rep Tree. También clasifica la operación de la alarma, a partir de un rango de valores específicos para una variable en particular. La diferencia es que poda las ramas del árbol que no aportan a la clasificación de la alarma. Por lo anterior, se obtiene un modelo que es claro y visualmente aceptable por parte del cliente.

Criterio	Algoritmos		
	MultiLayer Perceptron	**RepTree**	**J48**
Es comprensible	Sí	No	Sí
Cumple con la función	Sí	Sí	Sí
Ayuda a los objetivos del negocio	No	Sí	Sí
Potencialmente útil	Sí	Sí	Sí
Aprobado	**No**	**No**	**Sí**

Tabla 7: Resumen de aceptación de los modelos.

El modelo de clasificación de J48, se basa sólo en un atributo que es la intensidad de la llama de la cámara de combustión número cinco. Según el usuario esto se explica porque en el sistema de combustión existen 14 cámaras de las cuales 3 poseen termocuplas, para la medición de temperatura, en el caso de la cámara número cinco es la que se encuentra más cercana a las otras dos cámaras que también poseen termocuplas. Esto explicaría el por qué cuando se produce una alta diferencia de temperatura por alto spread, se manifiesta dicha diferencia con mayor intensidad en la temperatura de la cámara de combustión número cinco.

VIII. CONCLUSIONES Y TRABAJO FUTURO

Con los patrones de predicción y clasificación generados, se cuenta con información para saber qué valores deben tener las señales claves, cuando se produce la activación de la protección por alto spread del sistema de combustión de la TG, para así evitar la activación de dicha protección. Con la utilización de esta información para tomar decisiones, se debería aumentar la disponibilidad de la TG para dar cumplimiento a los objetivos de la empresa, además de mejorar la confiabilidad ante sus clientes.

A partir del trabajo desarrollado se identifica como acción futura la generación de una simulación de las condiciones de operación de la planta. Esto permitiría validar el modelo de minería de datos generado. Una vez validado el modelo de clasificación, se procederá a su implantación. Esto es, identificar anticipadamente el surgimiento de la condición de una alarma que permita evitar el disparo de la turbina.

IX. AGRADECIMIENTOS

Este trabajo fue realizado en el marco del desarrollo de un caso de estudio en la asignatura de minería de datos 1 del programa de Magíster en Ingeniería Informática de la UCN.

X. GLOSARIO DE TÉRMINOS

Alternador: Máquina eléctrica generadora de corriente alterna.
Alto Spread TG: Es cuando se produce una alta diferencia de temperatura entre las cámaras de combustión de la turbina a gas.
Cámara de combustión: Consiste en un recipiente al cual ingresa aire comprimido, al que se le añade combustible y se quema en forma ininterrumpida.
Disponibilidad: Se produce cuando una turbina está declarada al controlador del SING, como utilizable para producir electricidad.
Matriz Energética: Son los posibles combustibles que ocupa una Central para generar electricidad.
PI System: El estándar de la industria en la infraestructura de la empresa para la gestión de datos en tiempo real y eventos.
Potencia eléctrica: Es la velocidad con que se agota la energía.
Rotor: Parte giratoria de una turbina.
SING: Agrupación de Centrales Generadoras de Electricidad que se encuentran ubicadas entre las ciudades de Arica y Antofagasta.
Salidas forzadas: Es la interrupción intempestiva de la turbina por falla o defecto de esta o cualquier otro motivo.
Sistema de Control Distribuido: Es un concepto en donde la filosofía del equipo se define como la distribución geográfica del control, enlazada por una red de comunicaciones, cuyo destino es centralizar un cuadro de control central.
Turbina a Gas: Es un motor térmico rotativo de flujo continuo que se caracteriza por presentar una baja relación peso-potencia y una velocidad de giro muy elevada

XI. REFERENCIAS

[1] U.M. Fayyad, G. Piatetsky-Sapiro, and P. Smyth. 1991. *From data mining to knowledge discovery in databases*. Editors, Advances in Knowledge Discovery and Data Mining, page 39. AAAI Press, (1997).

[2] Mark A. Hall, Geoffrey Holmes. 2003. *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*. IEEE transactions on knowledge and data engineering, vol. 15, NO. 3.

[3] I.H. Witten, E. Frank. (2005). *Data Mining: practical machine learning tools and techniques* 2nd. Edition. Morgan Kaufmann.

[4] D. Rumelhart, G. Hinton, and R. Williams. 1986. *Learning representation by error propagation*, In D.E. Rumelhart and J.L. Mclelland, editors, Paralled Distributed Procession (Cambridge, MA: MIT Press).

[5] Abrahart, R.J., See, L. y Kneale, P.E. 2001. *Investigating the role of saliency analysis with a neural network rainfall-runoff model*. Journal of Computers and Geosciences, 27: 921-928.

[6] Soman, T. and Bobbie, P.O. 2005. *Classification of Arrhythmia Using Machine Learning Techniques*. Southern Polytechnic State University (SPSU) 1100 S. Marietta Parkway, Marietta, GA 30060, USA.

[7] Elizabeth Vitt, Michael Luckevich, Stacia Misner. 2002. Making Better Business Intelligence Decisions Faster. editors Microsoft Press.

[8] R. Wirth, J. Hipp. 2000. *CRISP-DM: Towards a standard process model for data mining*, in: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 2000, pp. 29-39.

[9] Min Tae Kim, Si Woo Lee. 2012. Application of in situ oxidation-resistant coating technology to a home-made 100 kW class gas turbine an its performance analysis. Applied Thermal Engineering, Volume 40, Pages 304–310.

[10] Frank J. Brooks. 2001. *GE Gas Turbine Performance Characteristics*. GE Power Systems Schenectady, NY GER-3567H.

[11] M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. 1996. *Advances in Knowledge Discovery and Data Mining*. San Mateo, AAAI Press, EE.UU.

[12] Michalski, R.S., Bratko, I., Kubat M. 1998. *Machine Learning andData Mining. Methods and Applications*. Wiley & Sons Ltd., EE.UU.

[13] S.G. Jolandan, H Mobli, H Ahmadi, M Omid, S.S. Mohtasebi. 2012. *Fuzzy-Rule-Based Faults Classification of Gearbox Tractor*. Department of Agricultural Machinery Engineering, faculty of Agricultural Engineering and technology University of Tehran, Karaj, Iran.

[14] Christina Athanasopoulou, Vasilis Chatziathanasiou and Ioannis Petridis. 2007. Utilizing data mining algorithms for identification and reconstruction of sensor faults: a Thermal Power Plant case study. University of Thessaloniki.

[15] Vedrana Vidulin, Mitja Luštrek, Matjaž Gams. 2007. *Comparison of the performance of genre classifiers trained by different machine learning algorithms*. Department of Intelligent Systems. Jožef Stefan Institute. Jamova 39, 1000 Ljubljana, Slovenia.

[16] Michalski, R. S. 1983. *A Theory and Methodology of Inductive Learning*. En Michalski, R. S., Carbonell, J. G., Mitchell, T. M. (eds.). Machine Learning: An Artificial Intelligence Approach, Vol. I. Morgan-Kauffman, EE.UU.

[17] Jun Li, Shunyi Zhang, Yanqing Lu, Junrong Yan. 2008. *Real-time P2P Traffic Identification*. Nanjing University of Posts and Telecommunication, Nanjing, JiangSu, China. Zhejiang Wanli University, Ningbo, Zhejiang, China.

[18] Prechelt L. 1998. *Early stopping — but when?* In. Orr GB, Muller OR, editors. Neural networks: Tricks of the trade. Berlin: Springer-Verlag Telos.

[19] Ayman Ahmed Seleemah. 2012. *A multilayer perceptron for predicting the ultimate shear strength of reinforced concrete beams*. Journal of Civil Engineering and Construction Technology Vol. 3(2), pp. 64-79.

[20] U.M. Fayyad, G. P. Shapiro and P. Smyth. 1996. *The KDD process for extracting useful knowledge from volumes from data*. Communication of ACM, Vol. 39(11).