APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS Y CLASIFICACIÓN DE INFORMACIÓN PRODUCIDA DURANTE INSPECCIONES DE LÍNEAS DE TRANSPORTE DE HIDROCARBUROS

WILLIAM LEONARDO GARCÍA RUEDA

FACULTAD DE INGENIERÍA DE SISTEMAS
UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
Bucaramanga
2013

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS Y CLASIFICACIÓN DE INFORMACIÓN PRODUCIDA DURANTE INSPECCIONES DE LÍNEAS DE TRANSPORTE DE HIDROCARBUROS

WILLIAM LEONARDO GARCÍA RUEDA

Trabajo de grado para optar por el Título de
Master en Gestión, Aplicación y Desarrollo de Software

Directora
Ph.D. Cristina N. González Caro

FACULTAD DE INGENIERÍA DE SISTEMAS
UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
Bucaramanga
2013

DEDICATORIA

A:

Dios, por todas las bendiciones recibidas tanto espirituales como materiales gracias por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante todo el periodo de estudio.

Mis padres Olga y Leonardo, por darme la vida, quererme mucho, creer en mí y porque siempre están ahí apoyando mis metas.

Mi esposa Diana Patricia, por su incondicional ayuda y comprensión durante el desarrollo de este proyecto.

Amigos y familiares por motivarme a salir adelante cada día.

WILLIAM LEONARDO GARCÍA RUEDA

AGRADECIMIENTOS

A todas las personas que participaron e hicieron posible este proyecto
muchas gracias por su apoyo y enseñanza:
Phd Cristina Nayesca González Caro, directora
Ing Mario Alberto Quintero Carvajal, asesor
Phd Daniel Arenas Seleey, evaluador
Phd Maritza Liliana Calderón Benavides, evaluador

A la facultad de ingeniería de sistemas de la UNAB y
la Corporación para la Investigación de la Corrosión
por apoyar esta idea y permitir la realización de esta tesis.

Lista de acrónimos

CIC	Corporación para la Investigación de la Corrosión
ITION	Tecnología para inspección de tendencias inerciales y operacionales de líneas de transporte de hidrocarburos
MFL	Magnetic Flux Leakage
PIG	Pipeline Inspection Gauge
RNA	Red Neuronal Artificial
KDD	Knowledge Discovery in Databases

Glosario

Indicación: Es una señal que es registrada por un sistema de inspección en línea de transporte de hidrocarburos. Una indicación puede clasificarse o caracterizarse como una anomalía, imperfección, o componente.

Anomalía: Una desviación sin examinar en el material de la tubería, los recubrimientos y soldaduras.

Imperfección: Una anomalía cuyas características no exceden los límites aceptables.

Defecto: Una anomalía examinada físicamente con dimensiones o características que superan los límites aceptables.

Característica: Cualquier objeto físico detectado por un sistema de inspección en línea. Las características pueden ser anomalías, componentes, soldaduras, accesorios o algún otro elemento.

Técnica MFL: Consiste en crear un campo magnético temporal en la tubería usando imanes a medida que la herramienta se desplaza, con el fin de medir los cambios en el campo magnético en la pared de la tubería, donde se supone es uniforme si no existen defectos.

Tabla de Contenido

Índice de Tablas

Índice de figuras

Introducción

Hoy en día, las líneas de transporte de hidrocarburos son consideradas la forma más eficiente para llevar fluidos (petróleo y gas) por largas distancias [3]. A medida que pasan los años, la integridad de este tipo de infraestructura se vuelve un motivo de preocupación para las empresas, principalmente porque una gran parte de las tuberías va llegando al final de su vida útil o como en el caso de Colombia [4], son víctimas de atentados terroristas, operaciones incorrectas o vandalismo, lo cual afecta su funcionamiento.

En consecuencia, es necesario supervisar, evaluar y garantizar la confiabilidad de las tuberías, para prevenir la aparición de fugas, fallas y de esta forma evitar catástrofes que impacten el medio ambiente y la población. Debido a los antecedentes mencionados, la industria ha desarrollado una serie de técnicas y ensayos denominados no destructivos o invasivos, los cuales permiten revisar tanto interna como externamente el estado de la tubería sin necesidad de detener la operación.

Dentro de los ensayos más utilizados se encuentra la inspección de tubería con herramientas inteligentes. Una herramienta de inspección inteligente de tuberías (Ver figura 1) como se define en [5] es un dispositivo que insertado dentro de un ducto, viaja a través de su extensión, impulsado por el propio fluido y puede realizar varias funciones, tales como:

- Eliminar los residuos presentes.
- Medir las condiciones de operación.
- Detectar defectos y/o anomalías.

Por medio de una serie de sensores que se encuentran integrados en su interior. Los datos que se recopilan de una herramienta de inspección inteligente son procesados y analizados a partir de técnicas de inteligencia artificial y como resultado se obtiene, la identificación de posibles sitios dentro de la tubería con presencia de defectos, anomalías o la localización de elementos propios de las líneas como soldaduras, válvulas, cruces y puntos bajos, las cuales son decisivos para el funcionamiento del negocio.

Una de las técnicas de inteligencia artificial empleada es la minería de datos que se refiere a un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos o en un simple archivo.

En este contexto, el propósito de este trabajo de tesis consiste en aplicar una serie de técnicas de inteligencia artificial a los datos provenientes de una herramienta de inspección inteligente construida por la Corporación para la Investigación de la Corrosión (en adelante CIC) para reconocer patrones que permitan identificar dos fenómenos: soldaduras y válvulas. Finalmente, las técnicas utilizadas fueron implementadas en un prototipo software denominado ITIONdF.

El presente trabajo cuenta con los siguientes capítulos:

Figura 1: *Herramienta de inspección inteligente.* ***Fuente CIC.***

- **Capítulo I. Estado del arte**
  En este capítulo se describe los trabajos reportados más representativos en el área bajo tres aspectos, las diferentes técnicas computacionales empleadas para el reconocimiento de fenómenos en tuberías, las técnicas de inspección que utiliza la industria y los fenómenos que hasta el momento se pueden identificar.

- **Capítulo II. Marco teórico**
  Se describen los conceptos utilizados para el desarrollo del trabajo haciendo énfasis en el aprendizaje de máquina y las técnicas de inteligencia artificial. Este capítulo es muy importante para entender la manera como se abordó el problema. En los capítulos siguientes se hará referencia a algunos de estos conceptos para lo cual se sugiere una lectura detallada de cada tema.

- **Capítulo III. Reconocimiento del fenómeno: soldadura**
  En este capítulo se define la soldadura, la manera como el experto reconoce el fenómeno y las técnicas de inteligencia artificial empleadas para identificarla de forma automática, las pruebas y los resultados obtenidos.

- **Capítulo IV. Reconocimiento del fenómeno: válvula**
  En este capítulo se define la válvula, la forma como el experto reconoce el fenómeno y las técnicas de inteligencia artificial empleadas para identificarla de forma automática, las pruebas y los resultados obtenidos.

- **Capítulo V. Conclusiones y recomendaciones**
  Se mencionan las conclusiones del trabajo con base en los resultados obtenidos. Además, se hacen recomendaciones de acuerdo con las limitaciones encontradas.

- **Capítulo VII. Trabajos futuros**
  Se listan las diferentes acciones que pueden tomarse a futuro para mejorar los resultados encontrados y se mencionan otros tipos de fenómenos que pueden estudiarse y continuar así dando aportes en esta área de investigación.

- **Anexos**
  Anexo A. *Interfase de software*
  Se describe el prototipo software desarrollado que implementa las técnicas de inteligencia artificial utilizadas durante el desarrollo del proyecto.
  Anexo B. *Carta de soldaduras*
  Se muestra el documento que hace referencia a la ubicación de las soldaduras existentes en el gasoducto, el cual se utilizó para entrenar y validar el modelo propuesto.
  Anexo C. *Documento de localización de válvulas*
  Este anexo hace referencia a la ubicación de las válvulas empleadas en la fase de entrenamiento.

Capítulo 1

Estado del arte

1.1. Antecendentes

La inspección de tuberías es el proceso por el cual se inserta una herramienta mecánica en el interior de la tubería, que viaja a lo largo de su longitud con el fin de limpiar o inspeccionar la tubería. Según [6], pese a no conocerse con exactitud la fecha de creación, existen reportes de inspecciones a tuberías desde finales del siglo XIX, años después que iniciaron operación las primeras líneas de crudo en los Estados Unidos, las cuales presentaron una disminución en la velocidad y presión de transporte debido a la acumulación de parafinas y desechos inherentes generados por el crudo. Durante la primera mitad del siglo XX se construyeron herramientas de inspección equipados con accesorios como cepillos, raspadoras, hojas afiladas y otros dispositivos de limpieza. A partir de los años 60, producto de los avances en la electrónica se desarrollaron las primeros equipos de inspección de tuberías [5] con la capacidad de adquirir datos para medir deformaciones, detectar obstrucciones, corrosión e irregularidades, a este tipo de herramientas se les denominó marranos inteligentes. Estos equipos se fueron desarrollando para cumplir con las siguientes tareas:

- Conocer el estado de la geometría del tubo (abolladuras, deformidades entre otros).
- Mapear la tubería (Conocer el trazado de la tubería, altimetría, movimientos, deformaciones).
- Pérdida de Metal (detectar corrosión, medir la pérdida espesor de pared de la tubería etc).
- Agrietamiento (Medir la fatiga producida por el agrietamiento, fallas en soldaduras entre otros).

Los primeras herramientas de inspección inteligentes fueron construidos a partir de la técnica MFL Magnetic Flux Leakage y los datos eran almacenados en una grabadora de cinta analógica similar a las cajas negras de los aviones, años después se desarrollaron herramientas de inspección inteligente usando las *técnicas inercial* [7], *caliper* [8] , *rayos X* [9] y de *ultrasonido* [10]. Al principio, los datos se procesaban y analizaban para medir el diámetro interno, detectar posibles abolladuras, deformaciones y otros defectos causados en la construcción de la línea de transporte. Sin embargo, la tecnología de la época, sumado a las técnicas de análisis que empleaba la herramienta no permitía encontrar todas las irregularidades correctamente. En los años 90 nuevos adelantos en la electrónica y computación permiten mejorar la precisión de las técnicas y crear herramientas con capacidad de manejar una gran cantidad

de datos. Lo anterior, facilitó la detección de un mayor número de defectos (geométricos + mecánicos) y encontrar fenómenos de corrosión.

Hoy en día, los esfuerzos computacionales se enfocan no sólo en detectar la presencia de posibles anomalías en la tubería, también en conocer la forma, longitud, profundidad, ancho, orientación y ubicación de cada anomalía. La tendencia del mercado es construir equipos que incorporen varias de las tecnologías mencionadas. En Colombia este tipo de tecnología comienza a implementarse en el año 2003 cuando la compañía ECOPETROL decide crear un programa de inspección para revisar 6000 kilómetros de tuberías que llevaban más de 20 años en funcionamiento [11]. Desde entonces, algunos prototipos y herramientas han sido desarrollados en Universidades y centros de investigación del país [12], [13], [14].

Entre estos desarrollos se encuentra la herramienta ITION (sigla de tecnología para inspección de tendencias inerciales y operacionales de líneas de transporte de hidrocarburos) desarrollada por la corporación para la investigación de la corrosión (CIC) , la cual, implementando la técnica inercial permite medir e identificar los datos operacionales de la tubería, reconstruir altimetría, planimetría, la trayectoria sobre la topografía del suelo, localizar defectos, fugas, deformidades y ubicar componentes como válvulas, magnetos entre otros y cuenta a la fecha con aproximadamente 300 kilómetros de inspección. En el presente trabajo, se plantea tomar los datos recolectados por esta herramienta y emplearlos para identificar dos (2) tipos de elementos: soldaduras y válvulas.

1.2. Trabajos relacionados

En el campo de la inspección inteligente de tuberías, las técnicas computacionales más utilizadas para el procesamiento, análisis de datos y reconocimiento de patrones son: filtrado de señales, modelado por elementos finitos y aprendizaje de máquina (ver figura 1.1).

Filtrado de señales

El filtrado de señales consiste en tomar una señal analógica o digital y realizar un procesamiento matemático para obtener como salida una nueva señal, la cual contiene los cambios necesarios para el análisis requerido. En este contexto, los datos (señales digitales) registrados por la herramienta de inspección son descargados y almacenados, luego, si se requiere, se aplica una serie de algoritmos de filtrado como en [3] [15] a las señales con el objeto de eliminar posibles ruidos e interferencias.

Los filtros digitales más empleados han sido Fourier [3], Kalman [3], Wavelet [15], Savitzky-Golay [3], mínimos cuadrados promedio [15] y respuesta finita al impulso [15]. El tipo de filtro a emplearse depende de la naturaleza de los datos y el problema que se esté atacando, por ejemplo en [3] se realizaron pruebas con tres clase de filtros para eliminar el ruido a las señales y mejorar así, los resultados generados por el algoritmo de reconocimiento de soldaduras, para ese caso, el filtro que generó el mejor rendimiento fue Savitzky-Golay. En [3] además se llega a la conclusión de que pese a que la eficiencia del modelo mejoran sólo un 5 %, si se reduce considerablemente el tiempo de entrenamiento, lo cual es de gran importancia si se tiene en cuenta el volumen de datos a manejar.

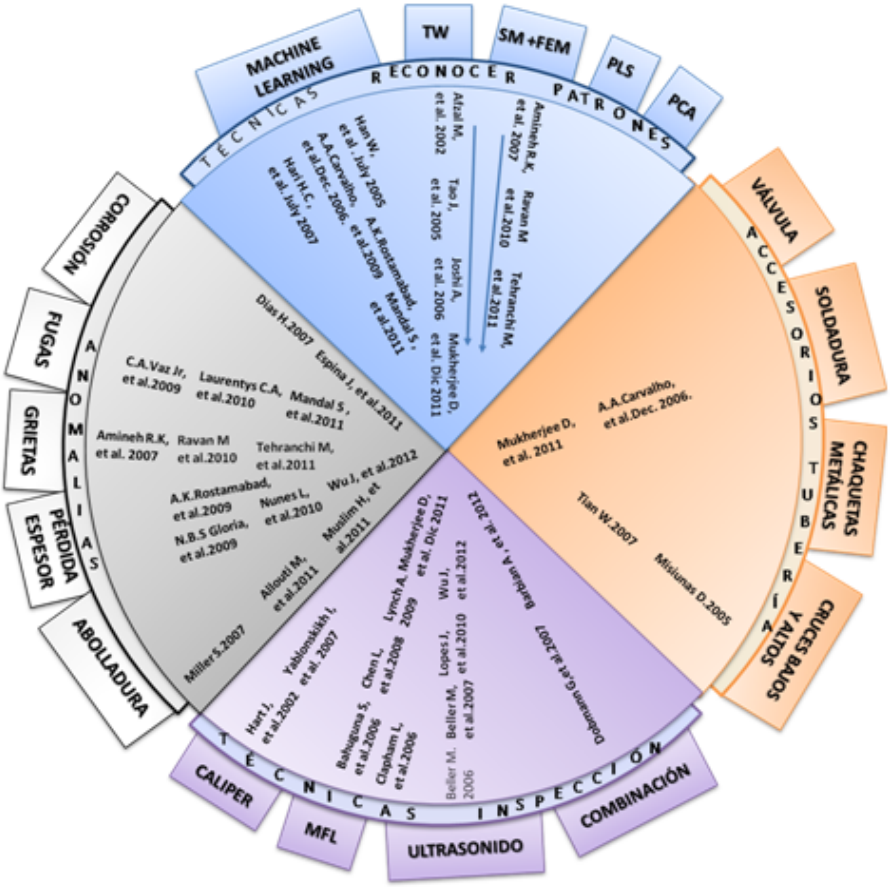


Figura 1.1: *Revisión de la literatura.*

Figura 1.2: *Trabajos representativos sobre aprendizaje de máquina.*

Análisis por elementos finitos

El análisis por elementos finitos es empleado para estudiar el fenómeno de las abolladuras y el agrietamiento en las paredes de las tuberías, como se menciona en [16], donde por medio de simulaciones se logra caracterizar la superficie de la tubería y hallar la orientación, longitud y profundidad de las grietas a partir de señales registradas por una herramienta de inspección.

Aprendizaje de máquina

El aprendizaje de máquina es una parte importante del proceso de descubrimiento de conocimiento (o KDD por sus siglas en inglés) que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de datos contenidos por un repositorio de información [1].

Con base en la revisión de la literatura, diferentes técnicas de aprendizaje se han utilizado para el reconocimiento de patrones en tuberías como en [15] donde se aplicaron técnicas de clasificación y regresión para reconocer soldaduras. Dichas técnicas de clasificación y regresión emplean algoritmos de análisis y procesamiento de datos como Kernelization [15], regresión por mínimos cuadrados [15], máquinas de soporte vectorial [15] y redes neuronales artificiales [15] con el objeto de reconocer patrones que conlleven a encontrar el fenómeno estudiado. A continuación (ver figura 1.2) se presenta un mapa que ilustra a nivel mundial los países donde se están trabajando con este tipo de técnicas.

Hacia donde se orienta la investigación

Los trabajos recientes buscan hallar la ubicación, tamaño, profundidad y como lo describe [15] la geometría del defecto. En resumen, en el presente proyecto se construirán modelos de aprendizaje automático con capacidad de identificar las siguientes anomalías: soldaduras y válvulas, tomando como referencia las técnicas encontradas en la literatura haciendo las adaptaciones necesarias. Además, se aplicará una metodología de minería de datos para garantizar un mejor análisis de la información. Finalmente, se desarrollará un prototipo que implemente los modelos desarrollados.

Otros tipos de anomalías diferentes a las 2 mencionadas se encuentran fuera del alcance del

proyecto debido principalmente a limitaciones de tiempo como son las abolladuras, localización de puntos bajos y altos y demás accesorios que se encuentran sobre la tubería. Además, para el caso de anomalías como pérdida de espesor, defectos por corrosión entre otros , los datos suministrados por la herramienta ITION en la actualidad no son suficientes para hacer estudios como los mencionados en la literatura. Una vez la CIC desarrolle una herramienta con tecnología MFL o de ultrasonido se podrá identificar y clasificar estos fenómenos lo cual abrirá la puerta para trabajos futuros.

Capítulo 2

Marco Teórico

2.1. Inspección de tuberías

Una herramienta de inspección de tuberías, (en inglés PIG [1] [17], también conocida como raspatubos [18], diablo [19] o marrano [13]), es un dispositivo mecánico que se inserta y viaja por el interior a lo largo de la longitud de una línea de transporte de hidrocarburos, agua u otro fluido con el objeto de limpiar o inspeccionar la tubería por medio de una serie de accesorios como cepillos, espumas, raspadores entre otros [2]. Adicionalmente, a las herramientas que se les adapta una serie de sensores electrónicos para registrar el estado de los tubos sin alterar sus propiedades físicas se les llama industrialmente como herramientas de inspección inteligente de tuberías [2].

Una herramienta inteligente, como se define en [5] *"es un dispositivo que insertado dentro de un ducto, viaja a través de su extensión, impulsado por el propio fluido y puede realizar varias funciones"*, tales como eliminar los residuos presentes, medir las condiciones de operación y detectar defectos y/o anomalías gracias a una serie de sensores que se encuentran integrados en su interior (Ver figura 2.2).

En Colombia, la CIC Corporación para la Investigación de la Corrosión ha desarrollado una herramienta de inspección de tuberías denominada ITION "herramienta inteligente para inspección y diagnóstico de líneas de transporte de hidrocarburos" , la cual está formada por dos elementos, un raspador convencional de limpieza de tuberías y el dispositivo electrónico (Ver figura 2.1). El raspador está constituido por un cuerpo metálico en el centro y un cuerpo blando que contiene 2 grupos de discos elaborados a partir de poliuretano, dentro del centro se ubica el dispositivo electrónico, el cual, se encuentra protegido por un encapsulado metálico.

2.2. Estudio de irregularidades en tuberías

Los datos que se recogen de una herramienta de inspección inteligente son procesados y analizados con el objeto de identificar posibles sitios dentro de la tubería con defectos, anomalías y a su vez, ubicar los componentes o accesorios como soldaduras, válvulas y cruces, las cuales son cruciales para el funcionamiento de la operación. Una anomalía en el contexto de la inspección de líneas de transporte de hidrocarburos (según la norma API Standard 1163) se define como una desviación no comprobada de los estándares de operación de un material,

[1] Pipeline Inspection Gauge

[2] En inglés se conoce con los nombres de *Smart Pigs* [20], *Intelligent Pigs*, *Internal Inspection Devices* ó *In-Line Inspection Tools* [21] y en español con el nombre de marranos inteligentes [13]

Figura 2.1: *Herramienta de Inspección Inteligente ITION.* **Fuente CIC.**

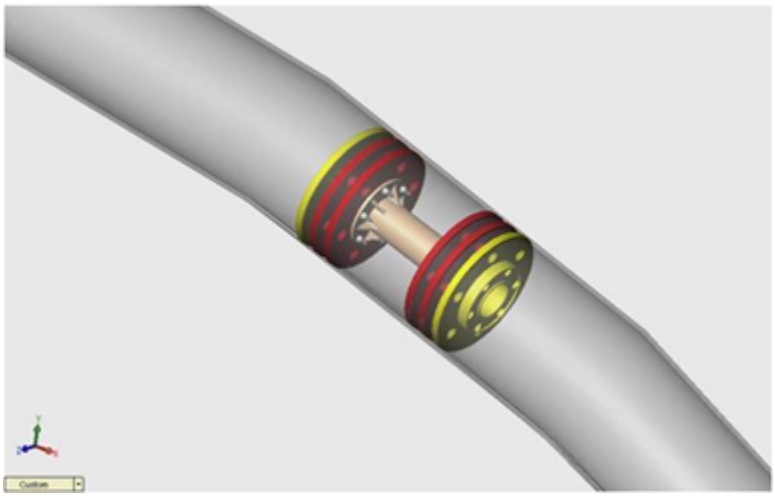


Figura 2.2: *Herramienta de Inspección Inteligente ITION.* **Fuente CIC.**

Figura 2.3: *Ejemplo de Válvula.* **Fuente CIC.**

Figura 2.4: *Ejemplo de Soldaduras.* **Fuente CIC.**

recubrimiento o soldadura aplicada a una tubería. A su vez, un defecto (lo define la norma API Standard 1163) es una anomalía examinada físicamente con unas dimensiones o características que exceden los límites aceptables. Existen diferentes tipos de posibles anomalías que pueden ser reconocidas a partir de los datos suministrados por una herramienta de inspección, algunos de los más importantes son:

Válvula: (ver figura 2.3) es un dispositivo mecánico que sirve para iniciar, regular o detener el paso de fluidos gracias a una pieza movible que permite abrir, detener o cerrar el orificio donde se transporta el material.

Soldadura: Se le denomina soldadura metálica al metal que se utiliza para unir otros dos metales base, en el negocio del transporte de hidrocarburos es fundamental para unir tubería o accesorios alrededor de una línea (ver figura 2.4).

Abolladura: es un daño que causa un cambio en la curvatura por deformación plástica permanente en la pared de la tubería sin reducir el espesor, es decir, no contiene otros defectos ni imperfecciones y es provocada por interferencia externa, por ejemplo golpes durante la construcción de un ducto, vandalismo, derrumbe de tierra entre otros, ver figura 2.5 [22]. De acuerdo con un estudio realizado en Colombia, por el instituto colombiano de petróleos al oleoducto Caño Limón Coveñas [4] mostró que el 65 % de los atentados terroristas realizados a la infraestructura producen abolladuras y fisuras en la unión de las soldaduras en las tuberías cercanas al sitio del atentado que de acuerdo con el grado de severidad pueden llegar a producir fallas y en consecuencia pérdidas económicas.

Actualmente, como lo afirma [15] los avances en esta tecnología han logrado no sólo reconocer el defecto, también es posible saber sus dimensiones (largo, ancho, profundidad) y su

Figura 2.5: *Ejemplo de abolladura.* **Fuente CIC.**

forma.

2.3. Reducción de la dimensionalidad

Es una técnica que se emplea para reducir el número de variables de entrada de un modelo con el objetivo de eliminar posibles correlaciones que afecten el funcionamiento de los algoritmos de aprendizaje. Para el caso de los algoritmos de clasificación, una ventaja de utilizar está técnica es la disminución en el sobre ajuste, que es un fenómeno en el cual el clasificador aprende las características contingentes y no sólo las constitutivas de los datos.

2.3.1. Análisis por componentes principales

Análisis de componentes principales (PCA) es una herramienta estadística estándar en el análisis de datos moderno, no paramétrico cuyo propósito es extraer la información relevante de los conjuntos de datos confusos. Consiste en reducir el número de variables perdiendo la menor cantidad de información posible. Los nuevos componentes serán una combinación lineal de las variables originales e independientes entre sí y se obtienen en orden decreciente de importancia. También puede entenderse que el análisis por componentes principales es la búsqueda del subespacio de mejor ajuste para los datos.

2.4. Selección de atributos

La selección de atributos es una de las técnicas de minería de datos para seleccionar un número de variables significativas de un set de datos. La función de la selección de atributos es elegir un subconjunto de variables de entrada mediante la eliminación de características con poca o ninguna información predictiva.

2.4.1. Algoritmo CfsSubsetEval

Este algoritmo evalúa un subconjunto de atributos considerando que estos datos poseen una habilidad predictiva individual en cada variable. Para ellos se necesitan que los subconjuntos de atributos estén muy correlacionados con la clase.

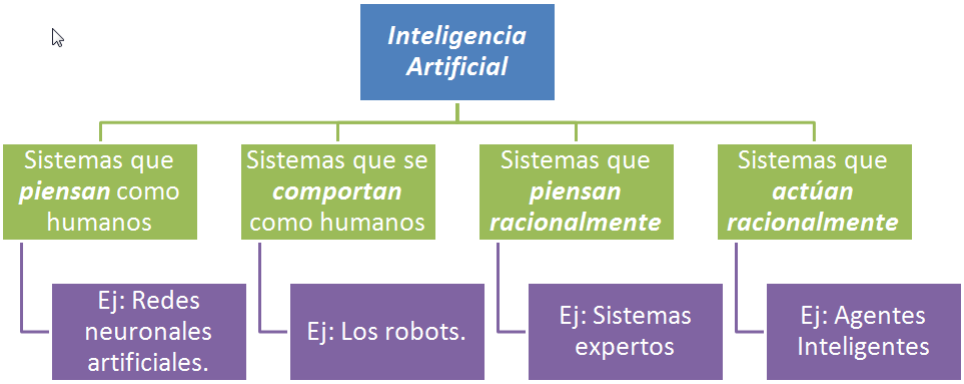


Figura 2.6: *Clasificación de sistemas en Inteligencia Artificial*

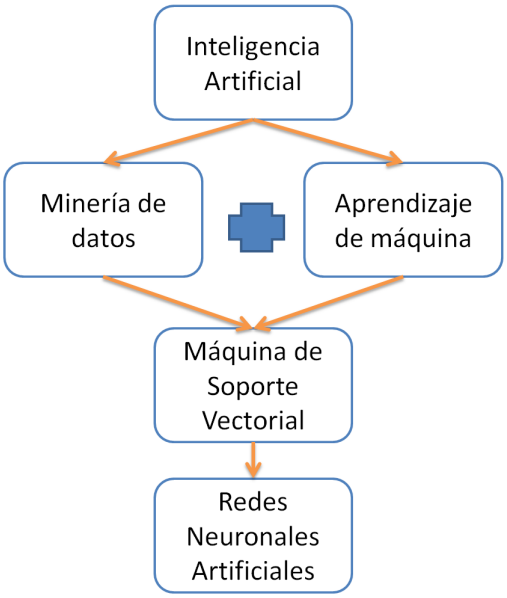


Figura 2.7: *Áreas de investigación de Inteligencia artificial.*

2.5. Técnicas de inteligencia artificial

Actualmente no se tiene un concepto sólido sobre la inteligencia, ya que involucra muchos procedimientos que aún no se comprenden en su totalidad. Dando una definición al término inteligencia, se describe como la parte del cálculo que tiene la capacidad de cumplir metas, esta habilidad se puede encontrar en las personas, algunos animales y máquinas. Según [23], la inteligencia artificial es la ciencia y la ingeniería de hacer máquinas inteligentes o sistemas especialmente inteligentes.
Según [24], la inteligencia artificial se puede clasificar de acuerdo a la técnica utilizada para resolver un problema. Ver figura 2.6

Entre las técnicas de inteligencia artificial que se usan en este trabajo se encuentran la minería de datos y el aprendizaje de máquina. Ver figura 2.7.

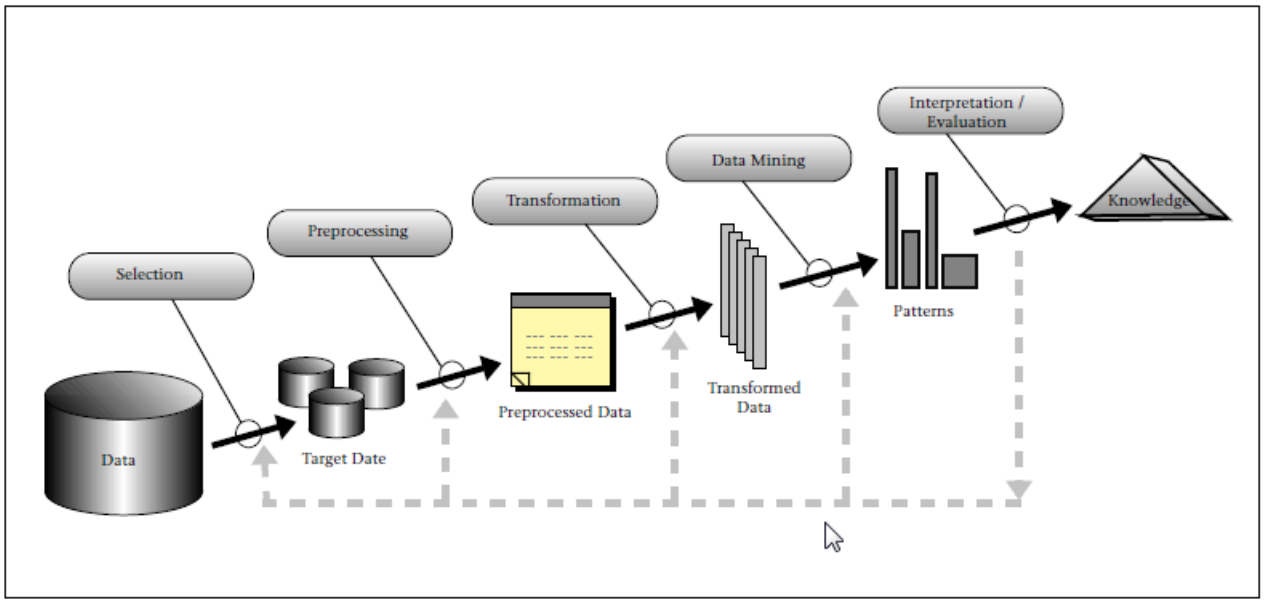


Figura 2.8: *Modelo de KDD.* **Fuente** [1]

2.5.1. Minería de datos

La minería de datos es el proceso de encontrar patrones en los datos [25], ya sea de forma automática o semiautomática. En el contexto del análisis de datos para inspecciones de tuberías se ha aplicado minería de datos para detectar fugas de crudo [26] o gas [27].

En [26] se utiliza el término Data Mining por primera vez como una solución al análisis de los grandes volúmenes de datos que son entregados por una herramienta de inspección que almacenan un gran número de variables que a su vez contribuyen a la ubicación de esos defectos. En el presente proyecto, para el estudio de cada uno de los patrones se utilizará la técnica de aprendizaje supervisado porque se cuenta con un conjunto de datos organizados para entrenamiento, cuya entrada y salida han sido validadas por un experto. Y se empleará la minería de datos como apoyo al proceso de reconocimiento de patrones.

La minería de datos es una parte importante del proceso de descubrimiento de conocimiento o KDD, el cual comienza (ver figura 2.8) con obtener los datos originales, a quienes, se les realiza selección, procesamiento y transformación para obtener unos datos que puedan ser utilizados por modelos o algoritmos de reconocimiento de patrones aplicando minería de datos con el fin de encontrar conocimiento.

2.5.2. Aprendizaje de máquina

Existen diferentes maneras para obtener conocimiento a partir datos, una de las formas más empleadas es por medio de aprendizaje de máquina, la cual según [28] "puede considerarse como una forma determinada de hacer que un agente que interactúa con su entorno cambie sus parámetros internos para que pueda adaptarse a nuevas situaciones y solucionar, a la larga, la tarea que le ha sido encomendada". También puede entenderse como "aquellos mecanismos, reglas, enfoques y tecnologías mediante el cual un agente puede aprender a desarrollar tareas que los seres humanos realizamos de forma natural y rápida".

Las técnicas de aprendizaje de máquina disponibles pueden clasificarse en al menos, cuatro grupos [29]:

Aprendizaje supervisado el cual consiste en crear una función capaz de predecir el valor correspondiente a cualquier objeto con datos de entrada válidos a partir de datos de entrenamiento (ejemplos etiquetados) y responder también a situaciones no vistas en el entrenamiento, los datos de entrenamiento se organizan por pares de objetos representados por un conjunto de atributos de entrada y un conjunto de respuesta o salida deseado.

En el **aprendizaje no supervisado** se crea un modelo a partir de unos datos de entrada, tratando a los objetos de entrada como un conjunto de variables aleatorias y sin información validada por un experto.

El **aprendizaje por optimización** estocástica se refiere a todas aquellas técnicas en las cuales el desempeño del modelo es calificado con algún índice y es necesario optimizar el modelo con alguna heurística particular de exploración aleatoria en el espacio de soluciones.

Finalmente, el **aprendizaje por refuerzo** consiste en aprender a decidir, ante una situación determinada, que acción es la más adecuada para lograr un objetivo.

Redes Neuronales

Las redes neuronales artificiales RNA son modelos matemáticos que se asemejan a la organización y funcionamiento del cerebro humano. Las ventajas que presenta este tipo de modelo para el análisis propuesto en este trabajo son:

- **Aprendizaje Adaptativo:** Las neuronas tienen la capacidad de adaptarse a las condiciones del entorno, por medio de un entrenamiento.
- **No linealidad:** Pueden ser lineales y no lineales. Esto les permite deducir relaciones complejas entre datos de entrada y salida.
- **Tolerancia a fallos:** Puede soportar daños parciales en su estructura, sin afectar el desempeño general.
- **Paralelismo:** Pueden procesar información de forma paralela y en tiempo real, lo cual las hace sistemas robustas, rápidas y confiables.
- **Capacidad de generalización:** tiene la capacidad de producir respuestas acertadas a entradas que no fueron suministradas durante el entrenamiento.

La estructura de una red neuronal artificial se compone de tres tipos de capas (Ver figura 2.9)

- **Capa de entrada:**Reciben la información proveniente de fuentes externas a la neurona.
- **Capa ocultas:** Son las capas que reciben de entrada los datos provenientes de capas anteriores y cuyas salidas pueden servir de entrada a capas posteriores.
- **Capa de salida:**Los valores de salida corresponden a las salidas de toda la red.

Para la resolución de un problema con ayuda de redes neuronales se parte de un conjunto de datos de entrenamiento significativo, con el cual se hace que la red aprenda automáticamente las propiedades deseadas. Para que el entrenamiento tenga éxito, se debe tener en

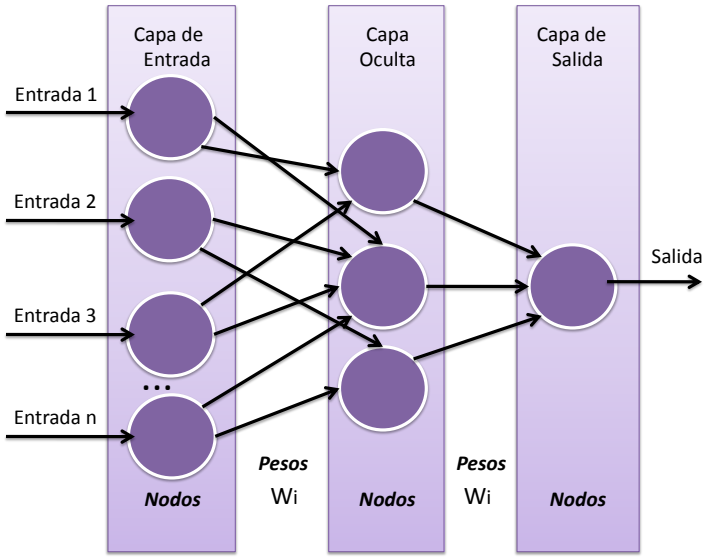


Figura 2.9: *Gráfica de la estructura de una Red Neuronal.*

cuenta la selección del modelo adecuado para la resolución del problema, las variables que se tendrán en cuenta y el preprocesamiento de la información con la cual se busca formar el conjunto de entrenamiento.

Perceptrón Multicapa

En este trabajo se utiliza una red neuronal del tipo Perceptrón multicapa. Este tipo de red tiene la ventaja de resolver problemas en donde las entradas no son linealmente separables. En la etapa de entrenamiento, realiza una actualización de los pesos de forma proporcional a la diferencia entre la salida esperada y la salida obtenida. Este tipo de algoritmo también es conocido como *Backpropagation* o retropropagación del error.

La retropropagación funciona de la siguiente manera: se aplica el conjunto de entrenamiento [3] a la entrada de la red, la cual se propaga por todas las capas hasta llegar a la capa de salida, allí se realiza la comparación de la salida esperada contra la salida obtenida y se calcula una señal de error para cada una de las salidas. Esta señal se propaga hacia atrás para todas las neuronas de la capa oculta que contribuyen con la salida. Este proceso continua hasta que se cumpla una de las siguientes condiciones: El error global es inferior al valor esperado, ó el número de iteraciones es inferior a un número determinado por el programador. Estos valores son configurados en el software WEKA junto a otros valores como: Número de neuronas en la capa oculta, tiempo de entrenamiento, normalización de los datos, etc.

Máquinas de Vectores de Soporte

Las máquinas de soporte vectorial son otro de los algoritmos de aprendizaje supervisado que se utiliza para clasificar patrones al enfrentar problemas de regresión y clasificación. La máquina de soporte vectorial mapea los datos a un espacio de características de n dimensiones,

[3] Este conjunto de entrenamiento esta formado por datos de entrada de los cuales se conoce su respectiva salida.

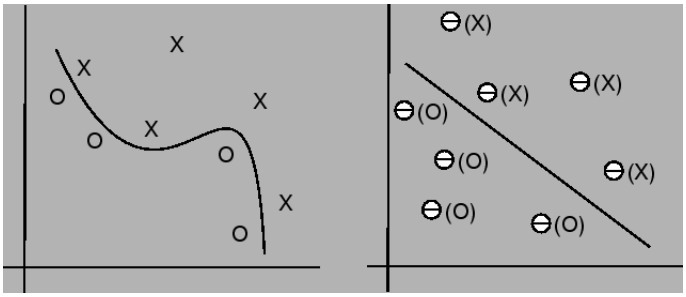


Figura 2.10: *Transformación del espacio de entrada con SVM's*

donde se halla un hiperplano de separación de manera más sencilla. Se lleva a cabo por un kernel, el cual hace una transformación de espacio de entrada en un espacio de características de mayor dimensión. Este hiperplano para el caso del uso de kernel puede ser no lineal.

El hiperplano de separación se calcula maximizando el espacio entre los patrones, este espacio es llamado margen. Esto ayuda a que la MSV encuentre el hiperplano más óptimo y así pueda tener mayor capacidad de generalización, de tal forma que al ingresar datos diferentes a los usados durante el entrenamiento, la MSV pueda clasificarlos correctamente. Ver figura 2.10 [4].

2.6. Técnicas de validación de modelos

Para la validación de los modelos y con base en los trabajos previos [3] y [15], se emplearon las técnicas de validación cruzada y análisis roc. Estas técnicas se apoyan en análisis estadísticos y se usan para determinar el algorirmo más apropiado para el aprendizaje de acuerdo con los datos suministrados.

2.6.1. Validación cruzada

También conocido como *cross validation*, es un método estadístico que tiene por objetivo estimar el grado de generalización de un modelo o comparar el rendimiento entre dos o más algoritmos para encontrar el mejor con base en los datos disponibles. Consiste en subdividir el conjunto de datos en varias partes, generalmente equivalentes en tamaño, una parte se usa para pruebas y el restante para entrenamiento. Luego, se definen n validaciones al modelo (ver figura 2.11) donde a medida que se hace cada validación se intercambia el conjunto de pruebas con una parte de los de entrenamiento a fin de evaluar todos los datos y se almacena el error producido. Finalmente, se calcula la media de todos los n errores generados.
La validación cruzada puede ser aplicada para estimar del rendimiento de un modelo, seleccionar un modelo o afinar los parámetros de un modelo de aprendizaje. En este proyecto se utiliza para comparar el mejor entre dos algoritmos de clasificación implementados durante el reconocimiento de soldaduras y válvulas.

2.6.2. Análisis ROC

Es una medida estadística que permite comparar el grado de sensibilidad contra la especificidad a las pruebas realizadas a un algoritmo de clasificación que tiene una salida binaria o

[4] Tomado de http://ccc.inaoep.mx/ emorales/Cursos/Aprendizaje2/node29.html

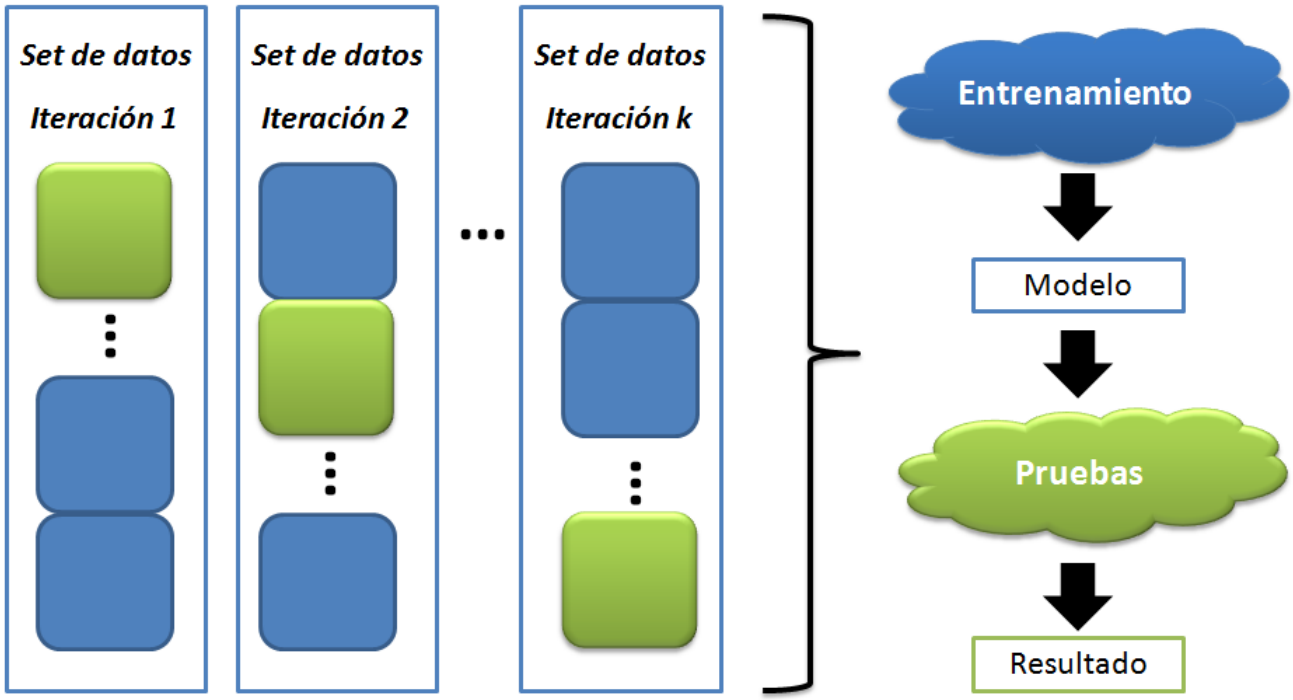


Figura 2.11: *Ejemplo Validación Cruzada.*

de dos estados. La sensibilidad es la proporción entre el número de pruebas clasificadas como positivas respecto al total de pruebas reales positivas. En este contexto, la sensibilidad mide la capacidad de identificar correctamente un fenómeno cuando está presente. La especificidad indica la probabilidad entre el número de pruebas clasificadas con resultado negativo sobre el total de pruebas negativas que realmente se presentaron. Para el presente trabajo, la especificidad está relacionada con la capacidad de descartar correctamente la presencia de un fenómeno cuando realmente no existe.

Para realizar el cálculo de la sensibilidad y la especificidad generalmente se construyen seis grupos de variables, los verdaderos positivos (VP) o casos clasificados como éxito por el algoritmo. Los verdaderos negativos (VN) o rechazos correctos. Falsos negativos (FN) es decir casos en los que el algoritmo de clasificación definió como positivo y en realidad era negativo. Falsos positivos (FP) aquellas pruebas donde la salida del algoritmo fue negativa pero en la realidad era positiva. Los verdaderos (V) son los casos cuya salida representa el fenómeno a clasificar y los negativos (N) son los casos cuya salida no representa el fenómeno.

Entonces, la sensibilidad se define como los verdaderos positivos sobre el total de casos verdaderos y la especificidad es la razón entre los verdaderos negativos sobre el total de casos negativos.

$$Sensibilidad = \frac{VP}{V} \tag{2.1}$$

$$Especificidad = \frac{VN}{N} \tag{2.2}$$

Para dibujar una curva ROC sólo es necesario las razones de Verdaderos Positivos (VPR) y de falsos positivos (FPR o también su equivalente 1-Especificidad). La VPR mide hasta

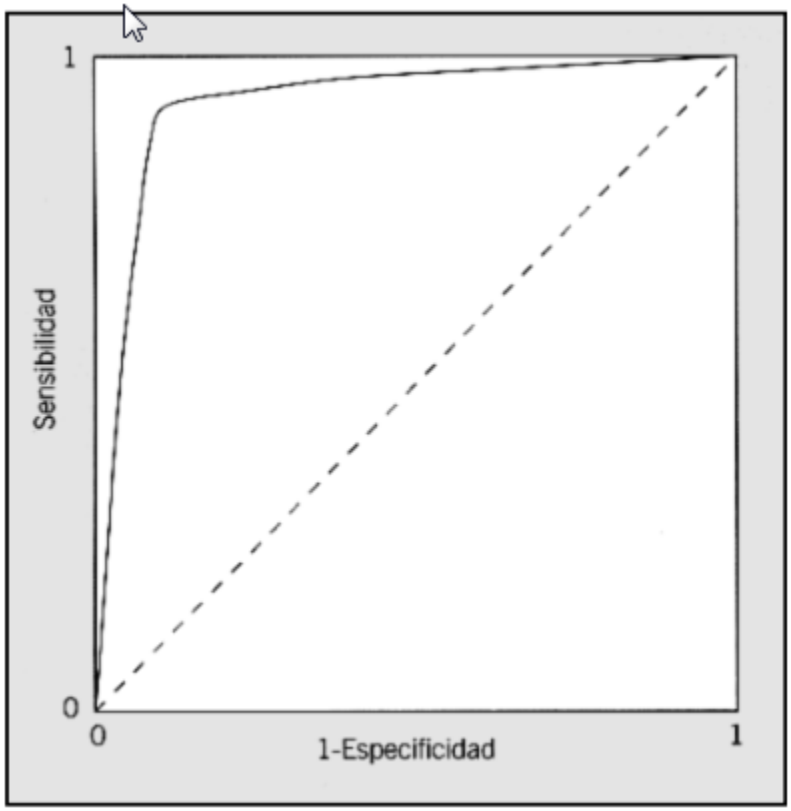


Figura 2.12: *Ejemplo de una curva ROC.* Fuente [30]

qué punto el clasificador es capaz de detectar o clasificar los casos positivos correctamente, de entre todos los casos positivos disponibles durante la prueba. La FPR define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles durante la prueba. En la figura 2.12 se muestra un ejemplo de una gráfica de este tipo.

Capítulo 3

Reconocimiento del Fenómeno: Soldaduras.

3.1. ¿Qué es una soldadura?

Descripción de la indicación:

Se le denomina soldadura al proceso mediante el cual se unen dos materiales generalmente metales por medio de un proceso físico denominado fusión. Los materiales se unen con la ayuda de otro tercer componente (también metálico) y como resultado se obtiene una unión fija y resistente. Al realizar este proceso alrededor de la tubería se logra una corona o cordón de soldadura (Figura 2.4) y esto es lo que finalmente busca detectar la herramienta de inspección. En el negocio del transporte de hidrocarburos es fundamental para unir tubería o accesorios alrededor de una línea.

Comportamiento físico:

Cuando la herramienta pasa por la soldadura, ella actúa como un obstáculo e intenta detener su paso, gracias al fluido con el que viaja, la herramienta toma impulso y continúa su recorrido. El fenómeno se repite dos veces a medida que cada grupo de discos de la herramienta pasan por el cordón de soldaduras. La distancia entre un grupo de discos y el otro es de 1 metro. Además en promedio la distancia entre un cordón de soldadura y otro es de doce (12) metros.

Comportamiento de las señales:

De acuerdo con los expertos de la CIC, existen tres tipos de sensores que sirven para detectar una soldadura, el acelerómetro, el giroscopio y el micrófono.

El acelerómetro es un dispositivo que mide los cambios de velocidad o aceleración que experimenta el cuerpo donde esté instalado. Cuando la herramienta pasa por un cordón de soldadura, inicialmente el sensor de aceleración registra una caída en su valor (desaceleración) producida por el efecto de frenado que ocasiona el choque con la soldadura , después, cuando el fluido impulsa la herramienta, el sensor registra una ganancia en su valor (aceleración). Los acelerómetros por su naturaleza son susceptibles a las vibraciones y los impactos, en la herramienta ITION el acelerómetro en el eje Y siempre se encuentra paralelo a la tubería, es por

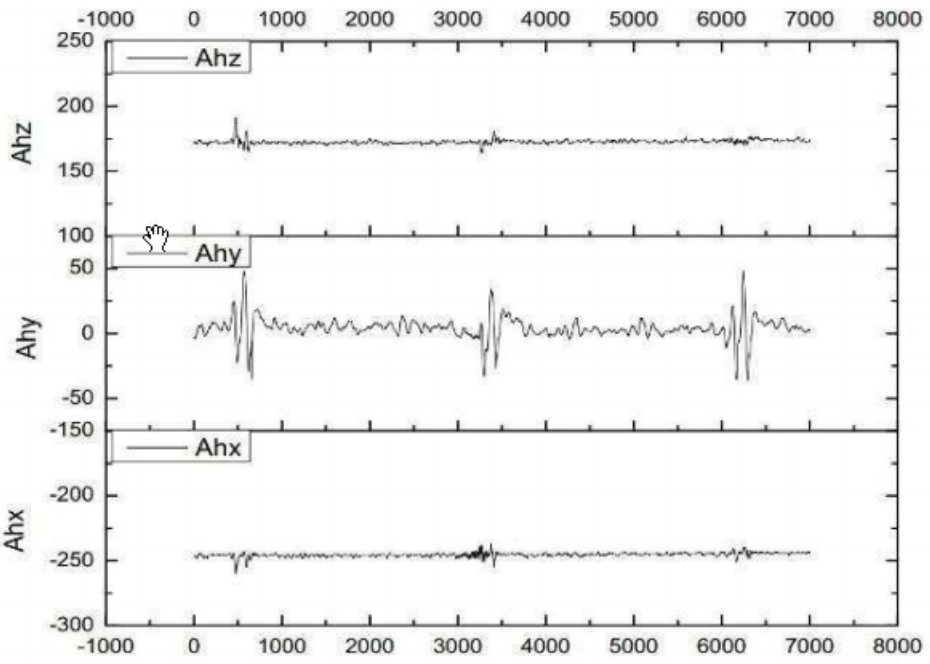


Figura 3.1: *Efecto del paso de la herramienta a través de tres cordones de soldadura registrado por el sensor acelerómetro.*

esta razón, que al momento de revisar los datos, la magnitud de su valor es más alto que en los otros dos ejes (x,z). En la figura 3.1, se muestra el comportamiento del sensor acelerómetro en los tres ejes al pasar por el cordón de soldaduras de 3 tubos. El eje X de la gráfica representa el número de muestras reportados por el sensor y en el eje Y el valor de la aceleración en lsb.

El giroscopio es un dispositivo que mide los cambios en la velocidad angular que experimenta el cuerpo donde se instale. En la figura 4.5 se muestra el comportamiento de los giroscopios en los tres componentes (X,Y,Z), el eje X de la gráfica representa el número de muestras reportados por el sensor y en el eje Y el valor de la aceleración en lsb.

El micrófono registra los cambios en la intensidad del ruido generado cuando el cuerpo al que se adhiere toca otras superficies. La herramienta de inspección cuenta con dos micrófonos, los cuales registran el sonido producido cuando los discos de la herramienta atraviesan el cordón de soldadura. En la figura 3.3 se muestra el comportamiento de la señal al pasar por las soldaduras de tres tubos. En este caso, AUD1 y AUD2 representa la intensidad del sonido generado cuando la herramienta choca con el cordón de soldadura, el eje X corresponde al número de registros entregados por el dispositivo. Por la amplitud de la señal registrada, el fenómeno se aprecia mejor en el micrófono 1. Lo anterior no significa que esto sea siempre correcto. Además es importante aclarar que los dos micrófonos están en capacidad de detectar los cambios en la intensidad del ruido al mismo tiempo. Para los expertos de la CIC, este sensor es el que ofrece la mayor confianza para reconocer la soldadura.

Finalmente, en la figura 3.4 se muestra la combinación de los tres sensores con los mejores resultados detectados por los sensores de la herramienta.

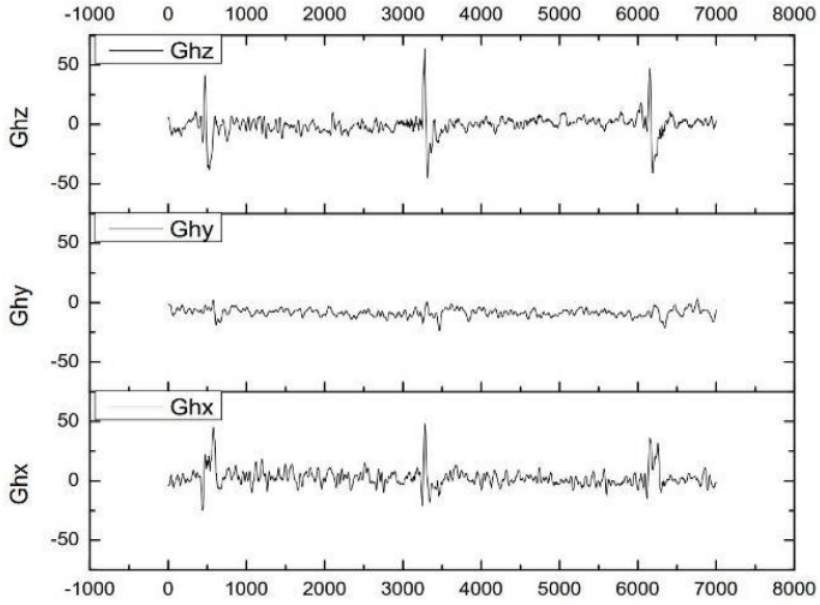


Figura 3.2: *Efecto del paso de la herramienta por tres soldaduras registrado por el sensor giroscopio.*

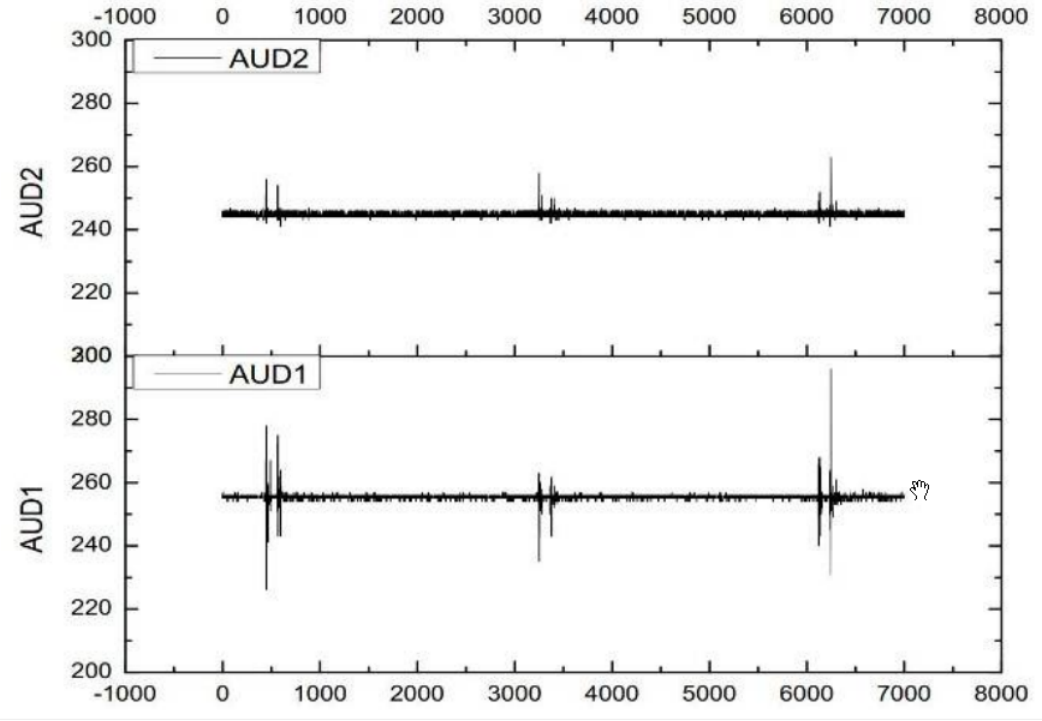


Figura 3.3: *Efecto del paso de la herramienta por tres soldaduras registrado por el sensor micrófono.*

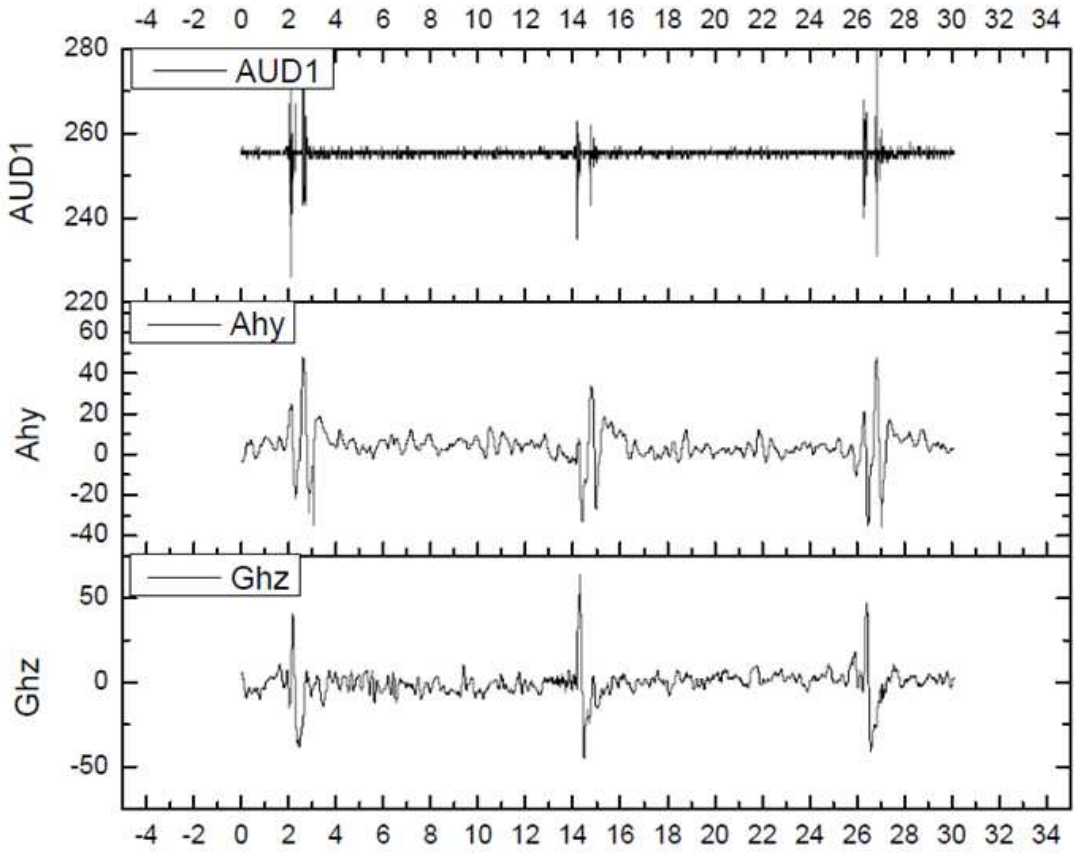


Figura 3.4: *Efecto del paso de la herramienta por tres soldaduras registrado por tres sensores.*

3.2. Selección y preparación de los datos

Los datos de entrada (atributos) corresponden a una inspección inteligente realizada en 15 de septiembre de 2012 a un gasoducto Colombiano que cuenta con las siguientes características técnicas:

- Longitud: 36 kilómetros [km].
- Diámetro nominal interno: 12 pulgadas [in].
- Material de diseño de la tubería: API 5LX-65.
- 3000 soldaduras reconocidas por manufactura según el documento denominado "Carta de soldaduras" (Ver Anexo 2) entregado por el propietario de la tubería.

Los datos suministrados por la Corporación para la investigación de la corrosión (CIC) presentan las siguientes características técnicas:

- Extensión del archivo: .ition
- Tamaño: 685 Megabyte [Mb].
- Número de Registros: 2'150.874.
- Tasa de muestreo: 250 registros por segundo.
- Variables de entrada: cincuenta y cinco (55) asociados con la información recolectada por los sensores del dispositivo.

Una tasa de muestreo de 250 registros por segundo significa que a medida que se desplaza la herramienta, esta es capaz de almacenar entre 4 y 6 registros de datos por cada punto que inspecciona.

Los atributos utilizados en la etapa de experimentación fueron seleccionados partiendo del criterio de los expertos CIC y se muestran en la tabla 3.1.

Atributo	Notación	Tipo dato	Unidad	Valor máximo	Valor mínimo
Giroscopio eje x	Ghx	Numérico	lsb	6000	-6000
Giroscopio eje y	Ghy	Numérico	lsb	6000	-6000
Giroscopio eje z	Ghz	Numérico	lsb	6000	-6000
Aceleración eje x	Ahx	Numérico	lsb	5454	-5454
Aceleración eje y	Ahy	Numérico	lsb	5454	-5454
Aceleración eje z	Ahz	Numérico	lsb	5454	-5454
Magnetómetro eje x	Mhx	Numérico	lsb	2500	-2500
Magnetómetro eje y	Mhy	Numérico	lsb	2500	-2500
Magnetómetro eje z	Mhz	Numérico	lsb	2500	-2500
Audio	AUD 1	Numérico	lsb	99999999	-99999999
Audio	AUD 2	Numérico	lsb	99999999	-99999999
Distancia	DIST	Numérico	lsb	99999999	-99999999

Tabla 3.1: Atributos seleccionados según criterio expertos CIC.

Los sensores Ahy, Ahx y Ahz corresponden al valor de la aceleración experimentado por la herramienta en los tres ejes espaciales a medida que se desplaza por la tubería. Los sensores Ghx, Ghy, Ghz miden los cambios en la inclinación y las vibraciones en los tres ejes espaciales.

GHX	GHY	GHZ	AHX	AHY	AHZ	MAGNITUD ACELERACION	MHX	MHY	MHZ	AUD1	AUD2	DISTANCIA	Soldadura
142	-36	1406	190	48	-305	362,53	-71	96	320	527	525	36,31	No
129	-33	1242	184	43	-302	356,24	-70	102	321	540	522	36,38	No
135	-28	1033	170	33	-298	344,66	-71	106	321	529	512	36,44	No
142	-22	908	160	20	-296	337,07	-71	110	322	527	506	36,50	No
149	-16	765	154	9	-296	333,79	-71	116	323	529	522	36,56	No
152	-5	650	146	-1	-297	330,95	-70	120	323	481	683	36,63	No
149	10	513	140	-8	-298	329,34	-71	129	324	1030	983	36,69	No
131	19	355	130	-15	-296	323,64	-71	133	325	1191	1439	36,75	Si
116	31	264	126	-11	-297	322,81	-70	138	325	1131	1375	36,81	Si
99	42	184	120	-11	-298	321,44	-70	144	326	543	1190	36,88	Si
74	44	111	115	-13	-303	324,35	-72	148	326	690	720	36,94	Si
47	45	78	112	-17	-307	327,23	-72	152	328	575	603	37,00	No
-17	44	26	108	-18	-318	336,32	-72	155	328	558	531	37,06	No
-65	40	-16	108	-17	-327	344,79	-72	157	329	565	536	37,13	No
-100	34	-51	110	-18	-332	350,21	-72	158	330	553	535	37,19	No
-154	-12	-81	112	-46	-346	366,57	-73	163	332	683	879	37,38	No
-112	-33	-70	111	-67	-350	373,24	-72	166	332	509	754	37,44	No
-81	-63	-84	109	-105	-352	383,16	-72	168	333	489	590	37,50	No
-56	-79	-110	105	-117	-351	384,60	-71	171	333	538	535	37,56	No
-31	-95	-143	103	-119	-349	382,85	-70	174	333	517	522	37,63	No

Figura 3.5: *Ejemplo del comportamiento de los datos en presencia de una soldadura.*

Los sensores Mhy, Mhx y Mhz miden el cambio en el campo magnético de la tierra en los tres ejes espaciales. Los sensores AUD1 y AUD2 obtienen el valor del nivel de ruido que se produce cuando la herramienta choca con la tubería. El sensor distancia mide el recorrido acumulado por el dispositivo.

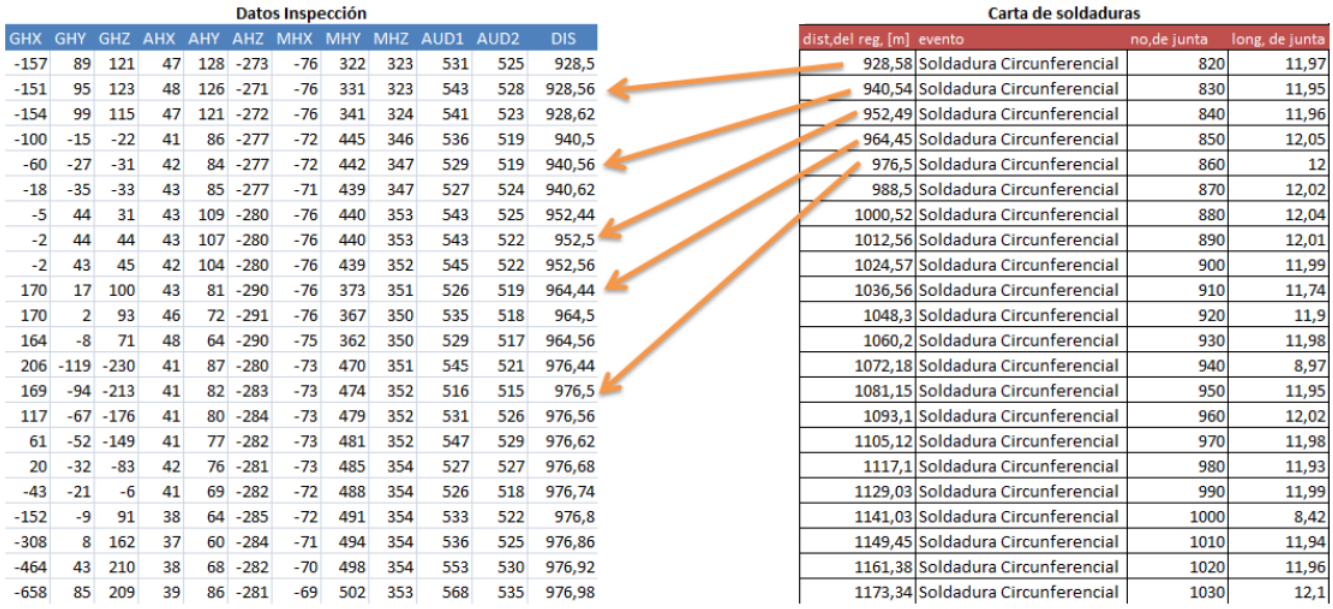
De manera experimental, se analizaron 20 soldaduras aleatoriamente (ver una muestra de los datos de una soldaduras en la figura 3.5) y los siguientes fueron los comportamientos comunes que se hallaron:

- Cuando la herramienta choca con una soldadura, los canales de audio registran una variación del nivel de ruido más alto que en los demás puntos de cada tubo.

- La magnitud de la aceleración es mayor o igual antes de llegar a la soldadura. Esto comprueba que la herramienta se "desacelera"justo antes de chocar con la soldadura.

- La magnitud de la aceleración es mayor o igual después de pasar por la soldadura. Lo anterior significa que la herramienta acelera por un instante de tiempo después de chocar con la soldadura.

Finalmente, los datos fueron separados en 67 hojas de cálculo, cada hoja con 32000 registros, las cuales se guardaron en formato xls.



Datos Inspección

GHX	GHY	GHZ	AHX	AHY	AHZ	MHX	MHY	MHZ	AUD1	AUD2	DIS
-157	89	121	47	128	-273	-76	322	323	531	525	928,5
-151	95	123	48	126	-271	-76	331	323	543	528	928,56
-154	99	115	47	121	-272	-76	341	324	541	523	928,62
-100	-15	-22	41	86	-277	-72	445	346	536	519	940,5
-60	-27	-31	42	84	-277	-72	442	347	529	519	940,56
-18	-35	-33	43	85	-277	-71	439	347	527	524	940,62
-5	44	31	43	109	-280	-76	440	353	543	525	952,44
-2	44	44	43	107	-280	-76	440	353	543	522	952,5
-2	43	45	42	104	-280	-76	439	352	545	522	952,56
170	17	100	43	81	-290	-76	373	351	526	519	964,44
170	2	93	46	72	-291	-76	367	350	535	518	964,5
164	-8	71	48	64	-290	-75	362	350	529	517	964,56
206	-119	-230	41	87	-280	-73	470	351	545	521	976,44
169	-94	-213	41	82	-283	-73	474	352	516	515	976,5
117	-67	-176	41	80	-284	-73	479	352	531	526	976,56
61	-52	-149	41	77	-282	-73	481	352	547	529	976,62
20	-32	-83	42	76	-281	-73	485	354	527	527	976,68
-43	-21	-6	41	69	-282	-72	488	354	526	518	976,74
-152	-9	91	38	64	-285	-72	491	354	533	522	976,8
-308	8	162	37	60	-284	-71	494	354	536	525	976,86
-464	43	210	38	68	-282	-70	498	354	553	530	976,92
-658	85	209	39	86	-281	-69	502	353	568	535	976,98

Carta de soldaduras

dist,del reg. [m]	evento	no,de junta	long. de junta
928,58	Soldadura Circunferencial	820	11,97
940,54	Soldadura Circunferencial	830	11,95
952,49	Soldadura Circunferencial	840	11,96
964,45	Soldadura Circunferencial	850	12,05
976,5	Soldadura Circunferencial	860	12
988,5	Soldadura Circunferencial	870	12,02
1000,52	Soldadura Circunferencial	880	12,04
1012,56	Soldadura Circunferencial	890	12,01
1024,57	Soldadura Circunferencial	900	11,99
1036,56	Soldadura Circunferencial	910	11,74
1048,3	Soldadura Circunferencial	920	11,9
1060,2	Soldadura Circunferencial	930	11,98
1072,18	Soldadura Circunferencial	940	8,97
1081,15	Soldadura Circunferencial	950	11,95
1093,1	Soldadura Circunferencial	960	12,02
1105,12	Soldadura Circunferencial	970	11,98
1117,1	Soldadura Circunferencial	980	11,93
1129,03	Soldadura Circunferencial	990	11,99
1141,03	Soldadura Circunferencial	1000	8,42
1149,45	Soldadura Circunferencial	1010	11,94
1161,38	Soldadura Circunferencial	1020	11,96
1173,34	Soldadura Circunferencial	1030	12,1

Figura 3.6: *Relación entre los datos de la carta de soldadura y la información suministrada por la herramienta de inspección.*

3.3. Procesamiento

Se realizó un proceso de verificación a cada hoja de cálculo con el fin de eliminar aquellos registros que tuviesen algún valor fuera del rango de aceptación, no operación de la herramienta o fueran registros duplicados. Luego, con el objeto de crear la etiqueta de clase, se desarrolló un algoritmo que permitiera relacionar los datos de entrada con la información suministrada en el documento carta de soldadura comparando el valor de la distancia con la abscisa donde se localiza la soldadura (ver figura 3.6). La carta de soldadura es un archivo que suministra el propietario de la línea y contiene los puntos geográficos (en metros) donde se encuentran las soldaduras. De esta forma, se genera una nueva columna denominada soldadura, la cual, posee dos valores 'S' si el registro corresponde a una soldadura y 'N' si no lo es. El algoritmo encuentra las soldaduras con un margen de error de más o menos 6 cm que es la mínima distancia medida por la herramienta entre un punto y otro. De esta forma, un registro o instancia lo conforman los 12 atributos definidos por los expertos y la clase soldadura.

Para la fase de modelado y con el objeto de realizar varias pruebas para medir la estabilidad del modelo, se construyen cuatro (4) conjuntos (sets) de datos para el entrenamiento en formato csv (comma-separated values) con el siguiente número de instancias: 148, 642, 1464 y 1838. Corresponden al 2 %, 11 %, 24 % y 31 % del total de las soldaduras. Cada archivo, contienen igual número de registros con la etiqueta soldadura y con no soldadura. Las soldaduras se eligen de forma consecutiva partiendo de la primera soldadura ubicada en la tubería mientras las no soldaduras se escogen al azar. También, se crearon 3 conjuntos (sets) en formato csv para la fase de pruebas usando los mismos criterios y con el siguiente número de instancias: 138, 434 y 600. Corresponden al 2 %, 7 % y 10 % de las soldaduras disponibles. Los datos empleados para el entrenamiento son diferentes a los seleccionados para las pruebas, sin embargo, es importante aclarar que dentro de los datos de entrenamiento el set superior si contiene los datos de los otros, en ese orden de ideas, el archivo con 1838 instancias contiene las instancias de los otros archivos (1464, 642 y 148). Ese mismo criterio se aplicó para el caso de los datos de validación.

3.4. Modelado

Se eligieron dos técnicas de aprendizaje de máquina: redes neuronales y máquinas de vectores de soporte (para una mayor comprensión de estos modelos favor dirigirse a la sección 2.5.2) con base en lo reportado en la literatura en el campo de la detección automática de soldaduras. El software seleccionado para realizar el entrenamiento y la validación fue Weka en su versión 3.7.9 porque es una de las herramientas software más completas para el análisis de datos gracias a las diferentes técnicas para el pre procesamiento y modelado de datos que implementa, adicionalmente, es muy portable porque se encuentra desarrollado con tecnología Java lo cual le permite correr en casi cualquier plataforma , está disponible libremente bajo licencia GNU y sus librerías se pueden integrar fácilmente en otras aplicaciones software.

Los datos para el entrenamiento del modelo se ingresaron al software Weka . Luego, como lo sugieren los trabajos previos ([3], [15]), se aplicó una técnica de selección de atributos específicamente el algoritmo *cfsSubsetEval* (para una mayor comprensión de este algoritmo dirigirse a la sección 2.4) y una técnica de reducción de la dimensionalidad: análisis por componentes principales (para una mayor comprensión de este algoritmo dirigirse a la sección 2.3) creándose así dos nuevos conjuntos de datos por cada archivo de entrenamiento y de pruebas. En total, dado que por cada uno de los 4 conjuntos de datos originales se aplicaron los algoritmos *cfsSubsetEval* y análisis por componentes principales, se construyeron doce (12) archivos y se guardaron en formato arff (Attribute-Relation File Format).

En la figura 3.7 se muestran los diferentes archivos utilizados durante el entrenamiento, la técnica de pre-procesamiento y los atributos más relevantes sugeridos por el algoritmo de pre-procesamiento.

3.5. Pruebas

3.5.1. Pruebas con redes neuronales

Cada archivo se entrenó con una red neuronal del tipo perceptrón multicapa (para mayor referencia acerca de redes neuronales diríjase a la sección 2.5.2). Después de realizar 12 pruebas, la configuración que se ajustó mejor a los datos fue la siguiente:

- N neuronas en la capa de entrada (con N entre 3 y 12 neuronas).
- M neuronas en la capa oculta (con M entre 6 y 8 neuronas)
- 2 neuronas en la capa de salida.

En la figura 3.8 se muestra en forma de tabla las pruebas realizadas a cada archivo de entrenamiento, la técnica de pre-procesamiento y el porcentaje de error alcanzado por el modelo, usando validación cruzada y conjuntos de pruebas. Para el caso de la validación cruzada se utilizaron 3 tipos de iteraciones (10, 20 y 40), cada tipo de iteración representa un subconjunto de los datos de entrenamiento que serán utilizados para validar el modelo mientras que los conjuntos de pruebas son datos independientes, los cuales no han sido utilizados durante el entrenamiento ni en validación cruzada, se crearon para medir el comportamiento del modelo con datos no conocidos. Existen tres conjuntos de datos independientes con 138, 434 y 600

Prueba Instancias	Técnica Pre procesamiento	Atributos más relevantes sugeridos por el algoritmo de pre-procesamiento
1838	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ, MHX, MHY, MHZ, AUD1, AUD2
1838	CfsSubsetEval	AHY, MHY, AUD2
1838	PCA	-0.466MHY +0.452AHZ +0.392AUD1+0.389AUD2 -0.292AHY -0.236MHX -0.226MHZ +0.217GHY +0.145AHX +0.101GHX -0.053GHZ 0.574AUD2 +0.571AUD1+0.31AHY -0.288AHZ -0.229GHX +0.201MHY +0.157MHX -0.125GHY -0.124AHX +0.101MHZ +0.077GHZ 0.596GHY -0.507GHZ +0.385AHY -0.311GHX -0.311AHX -0.186MHX -0.062AUD1 -0.059AUD2 -0.043MHY +0.04 AHZ +0.023MHZ -0.564AHX +0.537MHZ +0.466GHX +0.233AHZ +0.223MHX +0.183GHY +0.165GHZ -0.081AHY +0.049AUD1+0.047AUD2 -0.025MHY 0.634MHZ +0.456AHX +0.403MHY +0.273AHZ -0.214GHX -0.212MHX -0.189AHY -0.132GHZ +0.066GHY +0.052AUD1+0.038AUD2 -0.716GHZ +0.423MHX +0.413GHX +0.266AHX -0.197GHY +0.123AHY -0.073MHY +0.051AUD1+0.051AUD2 +0.045MHZ -0.018AHZ 0.763MHX -0.433GHX -0.317AHY +0.298GHY +0.141AHZ -0.105MHZ +0.072GHZ +0.06 AHX -0.03AUD1-0.025AUD2 +0.006MHY -0.457AHY -0.449GHY -0.385AHX -0.384GHX -0.337GHZ -0.308MHY +0.218MHZ -0.16AHZ -0.11MHX -0.018AUD2 -0.005AUD1 -0.624MHY +0.435AHY +0.358MHZ +0.232GHZ +0.229AHX -0.22GHX -0.166GHY +0.156AHZ +0.125MHX -0.098AUD1-0.076AUD2 -0.715AHZ +0.432GHY -0.295MHY +0.255AHX -0.243AHY +0.243MHZ +0.152GHX +0.064GHZ +0.062AUD2 -0.04MHX -0.018AUD1
1464	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ, MHX, MHY, MHZ, AUD1, AUD2
1464	CfsSubsetEval	AHY, AHX, AUD2
1464	PCA	0.551AHX +0.528AHZ -0.363MHY -0.328MHZ -0.319AHY -0.159MHX +0.133AUD1+0.128AUD2 +0.108GHY -0.057GHZ +0.052GHX 0.686AUD2 +0.683AUD1+0.167AHY -0.121AHX +0.067MHX +0.065MHZ +0.063GHZ -0.063AHZ -0.05GHX -0.038MHY -0.012GHY -0.528GHY +0.453GHZ +0.446GHX -0.379AHY +0.344MHX +0.2 MHZ +0.082AHX +0.064MHY +0.059AHZ +0.024AUD1+0.02 AUD2 -0.458MHZ -0.447MHY -0.43MHX -0.346AHZ +0.343GHX -0.303AHX -0.255GHY +0.094GHZ +0.043AHY -0.019AUD1-0.004AUD2 -0.65GHZ +0.521GHX +0.441MHX -0.275MHY +0.113MHZ +0.11AHY +0.093GHY -0.047AHZ -0.011AHX -0.006AUD1+0.003AUD2 0.763GHY +0.435GHZ +0.367GHX +0.224MHZ -0.114MHY -0.109AHZ -0.103AHY -0.068MHX -0.051AHX -0.028AUD1+0.012AUD2 -0.667MHX +0.425MHZ +0.373GHX +0.326MHY -0.242GHZ +0.2 AHZ -0.134GHY +0.068AUD1+0.053AHY +0.053AHX +0.036AUD2 -0.697AHY +0.381MHZ -0.317GHX -0.303MHY -0.249GHZ -0.236AHZ -0.198AHX -0.133MHX +0.038AUD1+0.034AUD2 -0.014GHY -0.613MHY +0.493MHZ +0.455AHY +0.212GHZ +0.193AHZ -0.182GHY -0.181GHX -0.099AUD1-0.093AUD2 +0.078AHX -0.037MHX
642	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ, MHX, MHY, MHZ, AUD1, AUD2
642	CfsSubsetEval	GHY, AHX, AHY, AHZ, MHX, MHZ, AUD1, AUD2
642	PCA	-0.469MHY -0.455MHZ +0.339AHX -0.334MHX +0.32 AHZ +0.297AUD2 +0.282AUD1-0.236GHZ +0.102AHY +0.101GHY -0.013GHX 0.552AUD1+0.53 AUD2 +0.361MHZ +0.335MHY +0.239MHX +0.239AHY +0.236AHZ +0.036GHZ -0.023GHY -0.004AHX -0.003GHX 0.542GHX -0.47GHY -0.356AHY -0.324AHZ +0.312GHZ +0.242AUD2 +0.223AUD1-0.156MHY -0.149MHZ -0.041AHX -0.019MHX 0.602AHX +0.462AHZ -0.441GHY +0.342MHX -0.225AUD2 -0.199AUD1+0.094GHZ +0.091GHX +0.05 MHY +0.045MHZ -0.024AHY -0.728GHX -0.533AHY -0.279GHY +0.184GHZ +0.153AHZ -0.138AHX +0.107AUD2 +0.106AUD1-0.077MHX -0.044MHZ -0.016MHY -0.718GHZ -0.551AHY +0.357MHX +0.139GHY +0.121GHX +0.079AUD1-0.067AHZ +0.06 MHY +0.057AHX +0.038AUD2 +0.02 MHZ 0.665GHY +0.483GHZ -0.404AHY +0.358AHX +0.113GHX +0.111MHY +0.056AHZ +0.052MHZ +0.037AUD2 -0.024MHX +0.024AUD1 0.76 MHX -0.381MHZ -0.335MHY +0.223GHZ +0.167GHY -0.164AHX +0.163AHY -0.152GHX -0.086AHZ +0.043AUD1+0.039AUD2 -0.696AHZ +0.586AHX -0.343GHX +0.162AHY +0.088AUD1-0.078GHY +0.076AUD2 -0.058GHZ +0.051MHZ +0.046MHY +0.016MHX
148	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ, MHX, MHY, MHZ, AUD1, AUD2
148	CfsSubsetEval	GHX, GHY, GHZ, AHX, AHY, AHZ, MHX, MHY, MHZ, AUD1, AUD2
148	PCA	0.451AUD2 +0.447AUD1-0.417AHX -0.403AHY +0.289MHY -0.237AHZ -0.229GHY +0.197MHX -0.144GHX +0.095GHZ -0.471MHY -0.468MHX +0.46 AUD2 +0.432AUD1+0.309AHZ +0.238AHX -0.063GHY -0.044GHX +0.044AHY -0.033GHZ 0.625GHX +0.525GHZ -0.316AHZ -0.309MHY -0.226MHX -0.183AHY -0.165AHX -0.113GHY -0.098AUD1-0.057AUD2 0.509AHZ +0.46 GHZ +0.43 MHX +0.417AHX -0.237AHY +0.207GHX +0.189MHY +0.141GHY +0.092AUD2 +0.083AUD1 -0.91GHY +0.259AHZ -0.201AUD1-0.165GHX -0.163AUD2 +0.092GHZ +0.043AHX -0.028MHY -0.021AHY +0.004MHX -0.631GHZ +0.51GHX -0.467AHY +0.128AHZ -0.102GHY +0.092MHX +0.061AHX -0.054AUD2 -0.05MHY +0.009AUD1 -0.717AHY -0.458GHX -0.294AUD1-0.292MHX +0.222GHY -0.146AUD2 -0.139MHY +0.104GHZ +0.06 AHZ +0.011AHX 0.651AHX -0.617AHZ +0.283MHX -0.232MHY -0.159GHX -0.131GHY -0.098AHY +0.096AUD1-0.035GHZ +0.011AUD2

Figura 3.7: *Archivos empleados durante el entrenamiento.*

PRUEBA instancias	Técnica pre-procesamiento	% Error en cada prueba					
		Validación Cruzada			Conjunto Entrenamiento		
		10 Folds	20 Folds	40 Folds	138	434	600
1838 (918 Soldaduras, 918 NO Soldaduras)	Ninguna	2,67	2,10	2,36	9,85	13,82	12,19
	CfsSubsetEval	3,67	3,74	3,54	5,09	5,15	4,98
	PCA	2,14	2,34	2,21			
1464 (732 Soldaduras, 732 NO Soldaduras)	Ninguna	1,95	2,18	2,03	49,63	31,92	32,59
	CfsSubsetEval	4,32	4,97	4,23	8,76	4,01	3,21
	PCA	2,30	2,58	2,46			
642 (321 Soldaduras, 321 NO Soldaduras)	Ninguna	2,89	2,68	2,53	27,20	31,30	28,11
	CfsSubsetEval	3,48	3,48	3,28	49,99	42,20	42,26
	PCA	2,64	2,58	2,80			
148 (74 Soldaduras, 74 NO Soldaduras)	Ninguna	3,29	3,22	3,12	49,64	46,15	45,78
	CfsSubsetEval	3,18	3,22	3,09	49,33	46,22	45,87
	PCA	2,19	2,31	2,32			

Figura 3.8: *Pruebas realizadas a los datos de entrenamiento y su porcentaje de error.*

instancias.

Se realizaron en total 60 pruebas. Al comparar los resultados obtenidos con la validación cruzada (*cross validation*) y los conjuntos de pruebas independientes se puede asegurar que el modelo reconoce soldaduras de forma más eficiente cuando utiliza como variables de entrada la aceleración en el eje Y, el campo magnético de la tierra en el eje Y y uno de los audios con 1838 instancias de entrenamiento, porque el porcentaje de error medido durante las pruebas estuvo en el rango entre 3.54 y 5.15 mientras que con otras instancias como 1434 y la misma configuración el porcentaje de error para las pruebas estuvo entre 3.21 y 8.76 y utilizando 642 instancias el porcentaje de error fue entre 2.53 y 62.6. Los resultados de estas pruebas demuestran que el modelo clasifica mejor después de 1434 instancias entrenadas utilizando el 31 % del total de las soldaduras.

En la figura 3.9 se muestran el comportamiento del modelo durante las 60 pruebas validado a partir de la técnica de análisis ROC. De acuerdo con estos resultados, la mayor efectividad se obtiene con 1838 instancias usando como entrada las señales de los sensores AHY, MHY y AUD2 porque la sensibilidad y especificidad los cuales determinan que también se reconoce una soldadura cuando realmente existe la soldadura o lo contrario está por encima del 95 %. No ocurre la mismo por ejemplo cuando se evaluó el modelo usando 642 instancias de entrenamiento donde la especificidad fue inferior al 20 % o con 148 instancias de entrenamiento donde la sensibilidad fue muy baja, inferior al 10 %. Lo anterior demuestra que es necesario un set de datos de entrenamiento por encima de 1464 instancias para que el algoritmo de clasificación reconozca eficientemente los datos. También es posible afirmar con los resultados que el modelo es más preciso para reconocer soldadura en comparación con la no soldadura porque en el 87 % de las pruebas realizadas con 1838 instancias entrenadas el valor de la sensibilidad es mayor o igual al de la especificidad.

En la figura 3.10 se muestra la evolución en el comportamiento del modelo para todas las instancias entrenadas al evaluarlas con el conjunto de 600 instancias independientes no empleadas en el entrenamiento ni en validación cruzada. Al principio, cuando se usa 148 instancias de entrenamiento la red neuronal no identifica correctamente la soldadura (sensibilidad del 8.5 %) , al aumentar el número de instancias la red reconoce mejor la soldadura pero comienza a no reconocer bien las no soldaduras (especificidad del 49 %), finalmente cuando se llega a las 1838 instancias de entrenamiento y se prueba con el archivo de 600 instancias, la red identifica sin error las soldaduras y en 6 de los casos la NO Soldadura la clasificó como soldadura (especifidad del 98 %). Esto comprueba que el modelo es sensible a la cantidad de datos con la cual se entrene.

3.5.2. Pruebas con máquinas de vectores de soporte

Cada archivo se entrenó usando dos tipos de kernel: Polinomio normalizado de grado 13 y el PUK ó *Pearson Universal Kernel* (para mayor información sobre estos tipos de kernel diríjase a la sección 2.5.2). Estos kernel se seleccionaron de acuerdo a trabajos encontrados en la literatura como [15] y [31] donde reportan un buen desempeño de estos algoritmos para resolver problemas de reconocimiento de defectos en soldaduras a partir de señales suministradas por herramientas de inspección inteligente.

Se realizaron pruebas con otros tipos de kernel como PolyKernel y RBFKernel que también se mencionan en esos trabajos, pero al final fueron descartados porque durante el entrena-

PRUEBA	Técnica pre-procesamiento empleada	Validación cruzada						Pruebas con datos independientes					
		10 Folds		20 Folds		40 Folds		138 instancias		434 instancias		600 instancias	
		Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)
1838 (918 Soldaduras, 918 NO Soldaduras)	Ninguna	99,2%	99,0%	99,2%	99,2%	99,0%	99,1%	100,0%	92,8%	99,5%	85,3%	99,7%	89,3%
1838 (918 Soldaduras, 918 NO Soldaduras)	*CfsSubsetEval*	98,6%	98,9%	99,1%	98,6%	99,0%	98,7%	100,0%	95,7%	100,0%	98,6%	100,0%	98,0%
1838 (918 Soldaduras, 918 NO Soldaduras)	*PCA*	99,2%	99,2%	99,1%	99,0%	99,2%	99,2%						
1464 (732 Soldaduras, 732 NO Soldaduras)	Ninguna	99,5%	99,3%	99,0%	99,2%	99,2%	99,3%	100,0%	0,0%	98,6%	34,1%	99,0%	32,3%
1464 (732 Soldaduras, 732 NO Soldaduras)	*CfsSubsetEval*	98,5%	98,8%	98,8%	98,4%	98,6%	98,8%	100,0%	94,2%	100,0%	98,2%	100,0%	98,7%
1464 (732 Soldaduras, 732 NO Soldaduras)	*PCA*	99,0%	98,9%	98,8%	99,3%	98,6%	99,0%						
642 (321 Soldaduras, 321 NO Soldaduras)	Ninguna	98,1%	99,1%	98,4%	99,1%	99,1%	99,1%	92,8%	53,6%	94,0%	41,9%	93,7%	49,7%
642 (321 Soldaduras, 321 NO Soldaduras)	*CfsSubsetEval*	98,8%	98,8%	98,8%	98,1%	98,4%	98,8%	100,0%	0,0%	100,0%	16,1%	100,0%	15,0%
642 (321 Soldaduras, 321 NO Soldaduras)	*PCA*	98,8%	99,1%	99,1%	99,1%	98,8%	98,8%						
148 (74 Soldaduras, 74 NO Soldaduras)	Ninguna	98,6%	100,0%	97,3%	100,0%	97,3%	100,0%	0,0%	100,0%	7,4%	100,0%	8,3%	100,0%
148 (74 Soldaduras, 74 NO Soldaduras)	*CfsSubsetEval*	98,6%	100,0%	97,3%	100,0%	97,3%	100,0%	1,4%	100,0%	7,8%	100,0%	8,3%	100,0%
148 (74 Soldaduras, 74 NO Soldaduras)	*PCA*	98,6%	100,0%	98,6%	100,0%	98,6%	100,0%						

Figura 3.9: *Pruebas realizadas a los datos de entrenamiento usando análisis ROC.*

PRUEBA # instancias	Técnica pre-procesamiento empleada	Prueba					
		600 instancias (300 Soldaduras y 300 NO Soldaduras)					
		Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Negativos (FN)	Falsos Positivos (FP)	Sensibilidad (VPR)	Especificidad (SPC)
1838 (918 Soldaduras, 918 NO Soldaduras)	Ninguna	299	268	1	32	99,7%	89,3%
1838 (918 Soldaduras, 918 NO Soldaduras)	*CfsSubsetEval*	300	294	0	6	100,0%	98,0%
1838 (918 Soldaduras, 918 NO Soldaduras)	*PCA*						
1464 (732 Soldaduras, 732 NO Soldaduras)	Ninguna	297	97	3	203	99,0%	32,3%
1464 (732 Soldaduras, 732 NO Soldaduras)	*CfsSubsetEval*	300	296	0	4	100,0%	98,7%
1464 (732 Soldaduras, 732 NO Soldaduras)	*PCA*						
642 (321 Soldaduras, 321 NO Soldaduras)	Ninguna	281	149	19	151	93,7%	49,7%
642 (321 Soldaduras, 321 NO Soldaduras)	*CfsSubsetEval*	300	45	0	255	100,0%	15,0%
642 (321 Soldaduras, 321 NO Soldaduras)	*PCA*						
148 (74 Soldaduras, 74 NO Soldaduras)	Ninguna	25	300	275	0	8,3%	100,0%
148 (74 Soldaduras, 74 NO Soldaduras)	*CfsSubsetEval*	25	300	275	0	8,3%	100,0%
148 (74 Soldaduras, 74 NO Soldaduras)	*PCA*						

Figura 3.10: *Evolución en el rendimiento del modelo a medida que se entrena con más instancias.*

miento , el porcentaje de error de la salida predicha por el modelo usando estos kernels era superior al 70 %.

Como se mencionó en la sección 3.4 se formaron inicialmente 4 archivos de datos que contienen 148, 642, 1466 y 1838 conjuntos de instancias. Posteriormente, fue necesario crear 8 nuevos archivos durante la fase de pre procesamiento, 4 con la técnica de selección de atributos y otros 4 para análisis por componentes principales, cabe aclarar que los datos de los nuevos archivos son transformaciones de los cuatro archivos originales. Esos 12 archivos son la entrada para los dos algoritmos de clasificación, los cuales se deferencian por el kernel que utilizan (PUK y Polinomio normalizado de grado 13). De esta forma se construyen y entrenan 24 casos (en este contexto, un caso corresponde a un algoritmo de clasificación que implementa uno de los dos kernels y emplea uno de los 12 archivos de entrada). Para validar estos 24 casos se utiliza validación cruzada (*cross validation*) con 10, 20 y 40 iteraciones y adicionalmente, se ejecutan pruebas con conjuntos de datos independientes los cuales no fueron usados en el entrenamiento ni forman parte de la validación cruzada. Estos datos independientes son archivos que contienen conjuntos de 138, 434 y 600 instancias respectivamente.

En la figura 3.11 se muestran las diferentes pruebas realizadas al modelo y el porcentaje de error obtenido. De acuerdo con los resultados, el modelo de máquinas de soporte vectorial reconoce con mayor eficiencia las soldaduras cuando utiliza un conjunto de 1838 instancias con los atributos AHY, MHY y AUD2 y los kernel PUK y polinomio de grado 13 porque el porcentaje de error durante las pruebas con validación cruzada se mantuvo entre 2.39 a 2.61 comparado con el 2.6 y 3.14 de las 1464 instancias ó el 2.8 a 3.43 de las 642 instancias. Esta diferencia es más notoria al validar el modelo con los conjuntos de datos independientes donde para el conjunto de entrenamiento de 1838 instancias el porcentaje de error estuvo entre el 1.67 al 4.65 comparado con el 49.8 al 50 del conjunto de 1464 instancias ó el 47 al 50 cuando se evaluó el conjunto con 642 instancias. Lo anterior demuestra que los datos utilizados durante las primeras pruebas no eran suficientes para que el clasificador realizara bien su trabajo.

En la figura 3.12 los mismos casos son analizados con la técnica de análisis ROC. De acuerdo con estos resultados, la mayor efectividad se obtiene con el conjunto de 1838 instancias entrenadas usando como entrada las señales de los sensores AHY, MHY y AUD2. El modelo es más preciso para reconocer soldadura en comparación con la no soldadura y siempre se presenta un margen de error en la sensibilidad y la especificidad inferior al cinco (5) por ciento tanto en las pruebas con validación cruzada como con los conjunto de datos independientes. No ocurre lo mismo por ejemplo, cuando se utilizaron 1464 instancias o 642 y se validaron con datos independientes. En dichas pruebas, el porcentaje de error de la especifidad fue muy alto cercano al 100, en consecuencia, el modelo tiene problemas para identificar correctamente las no soldaduras y se debe probablemente al tamaño de los datos empleados hasta ese momento.

En la figura 3.13 se muestra la evolución en el comportamiento del modelo para todas las instancias entrenadas frente a las 600 instancias de pruebas. Al usar el conjunto de 148 instancias para el entrenamiento la máquina de soporte vectorial no identifica correctamente la NO soldadura con una tasa de error del 100 %, al aumentar el número de instancias la máquina de soporte reconoce mejor la NO soldadura (pasa de 0 a 85 %), finalmente cuando se llega al conjunto con 1838 instancias de entrenamiento y se prueba con el archivo de 600 instancias, la máquina de soporte identifica las soldaduras correctamente exceptuando un caso si se utiliza el kernel PUK y dos casos con el kernel polinomio de grado 13. Respecto a las no soldaduras, se presentan 3 casos donde el clasificador no reconoce correctamente la instancia para los dos kernels.

PRUEBA	Técnica pre procesamiento	Algoritmo clasificación	% Error en cada prueba					
			Validación Cruzada			Conjunto Pruebas Independiente		
			10	20	40	138	434	600
1838 (919 Soldaduras, 919 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado 13	1,96	1,85	1,85	5,80	11,06	8,00
	Ninguna	SMO + PUKkernel omega 1 sigma 1	1,41	1,41	1,41	1,45	10,14	7,33
	CfsSubsetEval	SMO + NormalizedPolykernel grado 13	2,39	2,39	2,39	4,35	2,30	1,67
	CfsSubsetEval	SMO + PUKkernel omega 1 sigma 1	2,61	2,61	2,61	4,35	1,84	1,33
	PCA	SMO + NormalizedPolykernel grado 13	3,70	3,70	3,70			
	PCA	SMO + PUKkernel omega 1 sigma 1	1,41	1,41	1,41			
1464 (732 Soldaduras, 732 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado 13	1,78	1,78	1,78	50,00	35,48	25,67
	Ninguna	SMO + PUKkernel omega 1 sigma 1	1,64	1,64	1,64	34,78	17,51	15,00
	CfsSubsetEval	SMO + NormalizedPolykernel grado 13	2,60	2,60	2,60	50,00	50,00	49,80
	CfsSubsetEval	SMO + PUKkernel omega 1 sigma 1	3,14	3,14	3,14	50,00	62,67	45,33
	PCA	SMO + NormalizedPolykernel grado 13	1,50	1,50	1,50			
	PCA	SMO + PUKkernel omega 1 sigma 1	1,91	1,91	1,91			
642 (321 Soldaduras, 321 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado 13	2,80	2,80	2,80	50,00	50,00	48,67
	Ninguna	SMO + PUKkernel omega 1 sigma 1	1,87	1,87	1,87	50,00	50,00	50,00
	CfsSubsetEval	SMO + NormalizedPolykernel grado 13	3,43	3,43	3,43	50,00	50,00	47,00
	CfsSubsetEval	SMO + PUKkernel omega 1 sigma 1	2,80	2,80	2,80	50,00	50,00	50,00
	PCA	SMO + NormalizedPolykernel grado 13	5,29	5,29	5,29			
	PCA	SMO + PUKkernel omega 1 sigma 1	2,80	2,80	2,80			
148 (74 Soldaduras, 74 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado 13	0,00	0,00	0,00	50,00	50,00	50,00
	Ninguna	SMO + PUKkernel omega 1 sigma 1	0,00	0,00	0,00	50,00	50,00	50,00
	CfsSubsetEval	SMO + NormalizedPolykernel grado 13	0,00	0,00	0,00	50,00	50,00	50,00
	CfsSubsetEval	SMO + PUKkernel omega 1 sigma 1	0,00	0,00	0,00	50,00	50,00	50,00
	PCA	SMO + NormalizedPolykernel grado 13	0,00	0,00	0,00			
	PCA	SMO + PUKkernel omega 1 sigma 1	0,00	0,00	0,00			

Figura 3.11: *Resultados obtenidos para los modelos de máquinas de vectores de soporte.*

PRUEBA # instancias	Técnica pre-procesamiento empleada	Algoritmo clasificación empleado	Validación cruzada						Set independiente					
			10 Folds		20 Folds		40 Folds		136 instancias		400 instancias		600 instancias	
			Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)
1838	Ninguna	SMO + NormalizedPolykernel grado 13	98,8%	99,2%	98,9%	99,2%	98,9%	99,2%	100,0%	94,2%	100,0%	88,9%	100,0%	92,0%
1838	Ninguna	SMO + PUKkernel omega 1 sigma 1	99,2%	99,3%	99,2%	99,3%	99,2%	99,3%	100,0%	98,6%	99,5%	90,3%	99,7%	93,0%
1838	*CfsSubsetEval*	SMO + NormalizedPolykernel grado 13	98,7%	98,9%	98,7%	98,9%	98,7%	98,9%	100,0%	95,7%	99,1%	98,6%	99,3%	99,0%
1838	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	98,3%	99,1%	98,3%	99,1%	98,3%	99,1%	100,0%	95,7%	99,5%	98,6%	99,7%	99,0%
1464	Ninguna	SMO + NormalizedPolykernel grado 13	98,8%	99,5%	98,8%	99,5%	98,8%	99,5%	100,0%	0,0%	100,0%	29,0%	100,0%	48,7%
1464	Ninguna	SMO + PUKkernel omega 1 sigma 1	98,9%	99,5%	98,9%	99,5%	98,9%	99,5%	100,0%	65,2%	99,5%	82,9%	99,7%	85,3%
1464	*CfsSubsetEval*	SMO + NormalizedPolykernel grado 13	98,4%	99,0%	98,4%	99,0%	98,4%	99,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,3%
1464	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	97,8%	99,0%	97,8%	99,0%	97,8%	99,0%	100,0%	0,0%	100,0%	37,3%	100,0%	54,7%
642	Ninguna	SMO + NormalizedPolykernel grado 13	98,1%	99,1%	98,1%	99,1%	98,1%	99,1%	100,0%	0,0%	100,0%	0,0%	100,0%	2,7%
642	Ninguna	SMO + PUKkernel omega 1 sigma 1	98,8%	99,4%	98,8%	99,4%	98,8%	99,4%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%
642	*CfsSubsetEval*	SMO + NormalizedPolykernel grado 13	97,5%	99,1%	97,5%	99,1%	97,5%	99,1%	100,0%	0,0%	100,0%	0,0%	100,0%	6,0%
642	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	98,1%	99,1%	98,1%	99,1%	98,1%	99,1%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%
148	Ninguna	SMO + NormalizedPolykernel grado 13	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%
148	Ninguna	SMO + PUKkernel omega 1 sigma 1	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%
148	*CfsSubsetEval*	SMO + NormalizedPolykernel grado 13	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%
148	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%

Figura 3.12: *Pruebas realizadas a los datos de entrenamiento usando análisis ROC.*

Finalmente, si se unen los mejores resultados de cada técnica (ver tabla de la figura 3.14), se puede concluir que es posible reconocer de forma automática las soldaduras utilizando redes neuronales y máquinas de vectores de soporte y el modelo es más eficiente si se reduce el número de atributos a AHY, MHY y AUD2 comparado con el total de los atributos de entrada. Lo anterior contradice la opinión de los expertos quienes también utilizan los giroscopios como medio para reconocer soldaduras.

3.6. Discusión

- Como se mencionó en la sección anterior, luego de validar los modelos desarrollados con varios conjuntos de datos, se encontró que para algunas instancias los modelos no identificaron correctamente el fenómeno de la soldadura. En la figura 3.15 se muestra los registros que fallaron para el conjunto de 600 instancias, la salida real y las salidas predichas por los modelos. Durante la evaluación de las instancias número 77, 78 y 86 todos los modelos fallaron. La razón más probable para que se presente este error, es porque en esos registros el valor del audio está más cercano a valores donde se halla ubicada la soldadura que a valores cuando no hay soldadura (ver tabla figura 3.16). Lo anterior demuestra que los modelos de identificación de soldaduras son sensibles a la información que la herramienta de inspección recopila en el sensor audio. Para las instancias 595, 596 y 597 el modelo que utiliza redes neuronales puede estar clasificando esas instancias como soldaduras porque los valores de aceleración donde no hay presencia de soldadura se encuentran entre -41 y 90, sin embargo para estas instancias los valores de la aceleración se encuetran por fuera de este rango donde, según los datos debería existir soldadura. Para las instancias 147 y 322 no se encontró por inspección visual un argumento para explicar porque el modelo de máquinas de vectores de soporte generó ese resultado generado, lo cual implica que depende de la naturaleza de cada kernel para separar los datos de tal forma que reconozca una soldadura o una no soldadura.

- Con base en los resultados obtenidos, se puede afirmar que los dos modelos desarrollados permiten reconocer una soldadura de manera eficiente utilizando sólo 3 de los 12 variables de entrada. Lo anterior implica una reducción en el tiempo de procesamiento y análisis que será notoria en la medida que se procese una gran cantidad de datos.

- Los atributos que ofrecieron un mejor rendimiento para los modelos desarrollados fueron la aceleración y la variación del campo magnético en el eje Y así como uno de los sensores de audio. La herramienta de inspección inteligente se encuentra configurada de tal forma que el eje paralelo a la dirección del fluido es el eje Y. Esto coincide con lo que *apriori* los expertos de la CIC habían identificado. No obstante, a medida que se aumentaba el número de datos, la información del giroscopio en los tres ejes no fue clasificado por el algoritmo de selección de atributos como representativa y fue en estos conjuntos donde se obtienen los mejores resultados del modelo. Dicho hallazgo contradice la opinión de los expertos, quienes tienen en cuenta esta información durante su proceso de análisis. La razón por la cual el modelo sólo utiliza uno de los sensores de audio se debe a que los dos miden lo mismo y en consecuencia los valores que almacenan son similares.

- Con base en los resultados, se demuestra que es posible reconocer de forma automática los sitios donde se localizan las soldaduras en una tubería, lo anterior permitirá a los experto de la CIC contar con una herramienta de apoyo que facilite la toma de decisiones. Sin embargo, como los modelos desarrollados poseen un margen de error, queda a criterio del experto

PRUEBA # instancias	Técnica pre-procesamiento empleada	Algoritmo clasificación empleado	Conjunto de prueba					
			600 instancias : **300 Soldaduras** 300 NO Soldaduras					
			Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Negativos (FN)	Falsos Positivos (FP)	Sensibilidad (VPR)	Especificidad (SPC)
1838 instancias (918 Soldaduras, 918 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado	300	276	0	24	100,00%	92,00%
1838 instancias (918 Soldaduras, 918 NO Soldaduras)	Ninguna	SMO + PUKkernel omega 1 sigma 1	299	279	1	21	99,67%	93,00%
1838 instancias (918 Soldaduras, 918 NO Soldaduras)	*CfsSubsetEval*	SMO + NormalizedPolykernel grado	298	297	2	3	99,33%	99,00%
1838 instancias (918 Soldaduras, 918 NO Soldaduras)	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	299	297	1	3	99,67%	99,00%
1464 instancias (732 Soldaduras, 732 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado	300	146	0	154	100,00%	48,67%
1464 instancias (732 Soldaduras, 732 NO Soldaduras)	Ninguna	SMO + PUKkernel omega 1 sigma 1	299	256	1	44	99,67%	85,33%
1464 instancias (732 Soldaduras, 732 NO Soldaduras)	*CfsSubsetEval*	SMO + NormalizedPolykernel grado	300	1	0	299	100,00%	0,33%
1464 instancias (732 Soldaduras, 732 NO Soldaduras)	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	300	164	0	136	100,00%	54,67%
642 instancias (321 Soldaduras, 321 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado	300	8	0	292	100,00%	2,67%
642 instancias (321 Soldaduras, 321 NO Soldaduras)	Ninguna	SMO + PUKkernel omega 1 sigma 1	300	0	0	300	100,00%	0,00%
642 instancias (321 Soldaduras, 321 NO Soldaduras)	*CfsSubsetEval*	SMO + NormalizedPolykernel grado	300	18	0	282	1	0,06
642 instancias (321 Soldaduras, 321 NO Soldaduras)	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	300	0	0	300	100,00%	0,00%
148 instancias (74 Soldaduras, 74 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado	300	0	0	300	100%	0%
148 instancias (74 Soldaduras, 74 NO Soldaduras)	Ninguna	SMO + PUKkernel omega 1 sigma 1	300	0	0	300	100%	0%
148 instancias (74 Soldaduras, 74 NO Soldaduras)	*CfsSubsetEval*	SMO + NormalizedPolykernel grado	300	0	0	300	100%	0%
148 instancias (74 Soldaduras, 74 NO Soldaduras)	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	300	0	0	300	100%	0%

Figura 3.13: *Resultados análisis ROC para 600 instancias.*

PRUEBA # instancias	Técnica pre-procesamiento empleada	Algoritmo clasificación empleado	Conjunto de prueba					
			600 instancias : **300 Soldaduras** 300 NO Soldaduras					
			Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Negativos (FN)	Falsos Positivos (FP)	Sensibilidad (VPR)	Especificidad (SPC)
1838 (919 Soldaduras, 919 NO Soldaduras)	Ninguna	SMO + NormalizedPolykernel grado 13	300	276	0	24	100,00%	92,00%
1838 (919 Soldaduras, 919 NO Soldaduras)	Ninguna	SMO + PUKkernel omega 1 sigma 1	299	279	1	21	99,67%	93,00%
1838 (919 Soldaduras, 919 NO Soldaduras)	*CfsSubsetEval*	SMO + NormalizedPolykernel grado 13	298	297	2	3	99,33%	99,00%
1838 (919 Soldaduras, 919 NO Soldaduras)	*CfsSubsetEval*	SMO + PUKkernel omega 1 sigma 1	299	297	1	3	99,67%	99,00%
1838 (919 Soldaduras, 919 NO Soldaduras)	Ninguna	MultilayerPerceptron	299	268	1	32	99,7%	89,3%
1838 (919 Soldaduras, 919 NO Soldaduras)	*CfsSubsetEval*	MultilayerPerceptron	300	294	0	6	100,0%	98,0%

Figura 3.14: *Mejores resultados del modelo aplicando las dos algoritmos de clasificación.*

		RNA Perceptrón multicapa	**SVM** Kernel: Normalized Polykernel	**SVM** Kernel: PUK
# Instancia	Salida Real	Salidad Predicha	Salidad Predicha	Salidad Predicha
77	N	S	S	S
78	N	S	S	S
86	N	S	S	S
147	S	S	N	S
322	S	S	N	N
595	N	S	N	N
596	N	S	N	N
597	N	S	N	N

Figura 3.15: *Instancias donde los modelos no reconocieron correctamente la soldadura.*

GHX	GHY	GHZ	AHX	AHY	AHZ	MHX	MHY	MHZ	AUD1	AUD2	Soldadura
44	-14	34	-104	-109	-242	-129	76	177	661	630	S
35	-22	-15	-105	-145	-234	-58	-209	311	599	671	S
39	3	7	-102	-151	-235	-93	-13	322	600	689	S
44	-3	-90	-101	-159	-234	-91	-300	293	563	639	S
30	**57**	**13**	**-223**	**-42**	**-178**	**-54**	**162**	**294**	**587**	**606**	**N**
-16	**67**	**53**	**-224**	**-28**	**-176**	**-61**	**148**	**281**	**556**	**579**	**N**
-283	**87**	**-95**	**-227**	**-44**	**-182**	**-63**	**182**	**253**	**554**	**585**	**N**
-185	57	-46	-234	-36	-178	-72	161	253	544	540	N
-98	47	-19	-232	-55	-176	-77	149	252	549	541	N
40	51	-15	-216	-93	-186	-117	162	356	536	525	N
44	51	-5	-216	-92	-186	-117	160	356	542	525	N
44	49	6	-216	-91	-185	-117	156	356	540	526	N
39	47	8	-216	-91	-185	-119	151	355	544	527	N
36	45	7	-216	-90	-184	-120	146	354	537	526	N
32	42	4	-216	-90	-184	-122	140	353	542	526	N

Figura 3.16: *Ejemplo de instancias donde se presenta soldadura y no soldadura.*

definir en esos casos de acuerdo con su experiencia y la información de diseño de la línea, si el punto posee o no soldadura.

- Respecto al estado del arte se puede afirmar que si bien los trabajos de [3] y [15] resuelven un problema con un contexto diferente y también con otro tipo de tecnología, es posible usando la tecnología inercial que implementa la herramienta de inspección inteligente desarrollada por la corporación para la investigación de la corrosión identificar las soldaduras con una tasa de éxito comparable con la reporta por esos trabajos.

- La industria podrá verse beneficiada de los resultados en la medida que la CIC pueda entregar los informes de inspección de forma más oportuna. Esta fuera del alcance del presente trabajo medir el tiempo que se ahorra en el análisis de los datos de la inspección, lo que si se puede garantizar dado los resultados obtenidos es que el tiempo empleado durante los análisis de soldaduras será menor al que realiza en la actualidad.

Capítulo 4

Reconocimiento del Fenómeno: Válvulas

4.1. ¿Qué es una válvula?

Descripción de la indicación

Una válvula es un dispositivo mecánico que se utiliza para iniciar, regular o detener el paso de fluidos gracias a una pieza movible la cual facilita la manipulación del orificio donde se transporta el material. En la industria del transporte de hidrocarburos funcionan diferentes tipos de válvulas, las más utilizadas en Colombia son las de tipo bola y las de tipo cheque o anti retorno.

Las válvulas tipo bola poseen un mecanismo de regulación en forma de esfera perforada, de tal forma que al girar permite el paso del fluido cuando la perforación se encuentra alienada con la entrada y la salida de la válvula. Las válvulas tipo anti retorno o cheque poseen una compuerta metálica a la entrada (ver figura 4.1) que abre en el sentido por el que viaja el fluido que transporta la tubería y tienen por objeto cerrar por completo el paso del fluido en circulación en un solo sentido, son útiles para cortar el servicio en caso que se requiera.

Comportamiento físico

Las válvulas de bola por lo general, se encuentran presentes al inicio y final del recorrido de la inspección, por regulación, si la distancia del recorrido es mayor a 20 km o se presenta una conexión entre líneas, es posible encontrar al menos otra válvula de este tipo justo en el punto donde se conectan o en un punto cercano a la mitad de la longitud de la tubería.

Cuando se instala una válvula de este tipo en la tubería, debido a que todos sus componentes son metálicos y ocupan un volumen considerable dependiendo del diámetro de la tubería, se genera un aumento en la cantidad de metal que altera las propiedades del campo magnético de la tierra. Por lo tanto, los sensores magnéticos de la herramienta reconocen esos cambios y lo registran. Además, para acoplar la válvula a la tubería se requiere aplicar soldadura, cuando la herramienta pasa, choca contra las soldaduras y los sensores de audio registran la variación generada.

El comportamiento de la herramienta es diferente cuando se encuentra con válvulas de bola ubicadas al inicio y final del recorrido comparado con una válvula de bola ubicada en otra

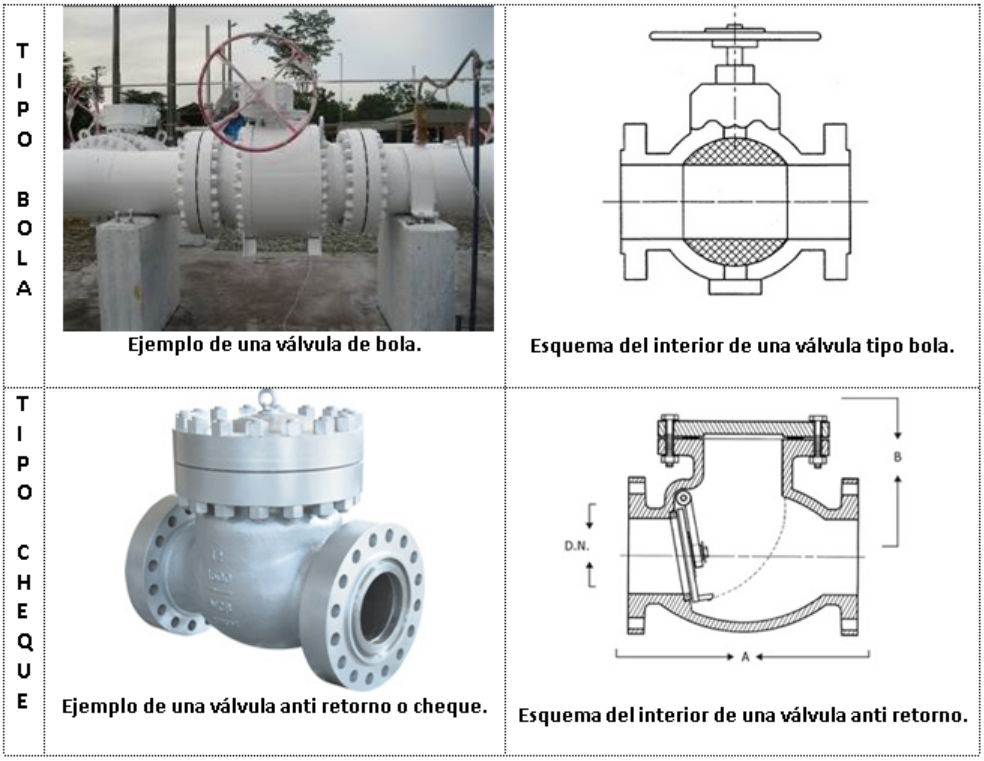


Figura 4.1: *Tipos de válvulas empleadas en la industria del transporte de hidrocarburos.* **Fuente fotos CIC y esquemas** [2]

región de la tubería debido a las diferencias de presión la cual es más alta en los extremos que en los demás puntos de la línea. Esto produce un desplazamiento más fuerte en los extremos y mayor número de vibraciones producto del choque del dispositivo con los componentes que conforman la válvula.

Cuando la herramienta pasa por una válvula anti retorno, ella actúa como un obstáculo e intenta detener su paso, gracias al fluido con el que viaja, la herramienta gana impulso y continúa su recorrido. Si la compuerta se encuentra cerrada, la herramienta necesariamente debe chocar produciendo un ruido que es captado por el sensor de micrófono, en una magnitud muy superior al de una soldadura.

Comportamiento de las señales:

De acuerdo con los expertos de la CIC, existen cuatro tipos de sensores que sirven para detectar una válvula, el acelerómetro, el giroscopio, la presión y el micrófono. Sin embargo, el comportamiento de la señal es diferente si la válvula se encuentra ubicada en los extremos o en otro punto de la línea.

El acelerómetro es un dispositivo que mide los cambios de velocidad o aceleración que experimenta el cuerpo donde esté instalado. Al comparar el comportamiento de la aceleración cuando la herramienta pasa por válvula que se encuentra en los extremos (ver figura 4.2) con una válvula ubicada en otro sitio (ver figura 4.2), se observa que existen dos tipos de comportamiento, en el primero se presenta un cambio brusco seguido por pequeñas variaciones por una distancia que sobrepasa la longitud de la válvula, en el segundo en cambio

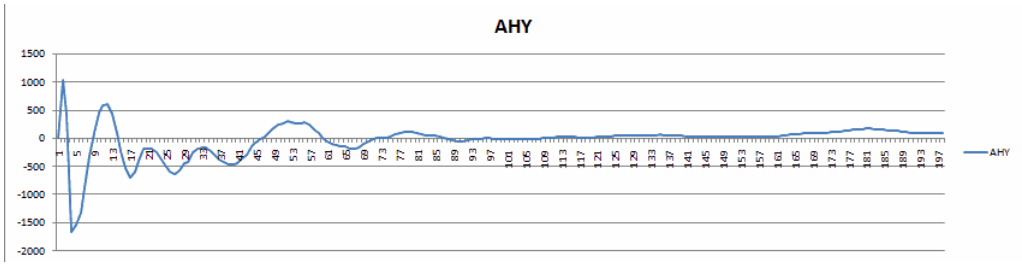


Figura 4.2: *Efecto del paso de la herramienta a través de una válvula ubicada al inicio o fin de una línea de transporte registrado por el sensor de aceleración.*

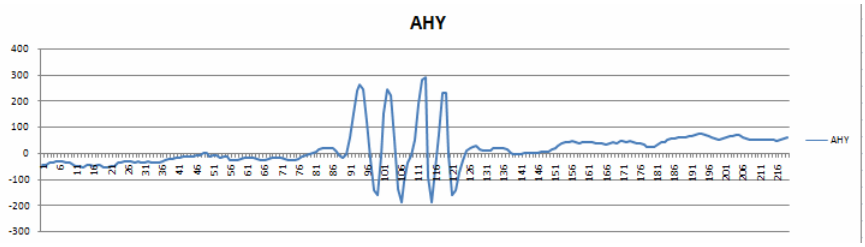


Figura 4.3: *Efecto del paso de la herramienta a través de una válvula ubicada al inicio o fin de una línea de transporte registrado por el sensor de aceleración.*

se nota claramente cuatro puntos con cambios muy similares sólo sobre la región de la válvula.

El giroscopio es un dispositivo que mide los cambios en la velocidad angular que experimenta el cuerpo a medida que se desplaza por la tubería. Al comparar las figuras (4.4 y 4.5) se observa que para el caso de la válvula ubicada en un extremo de la línea el comportamiento es en forma de montaña, el problema es que unos metros después ese mismo comportamiento se repite de forma más pronunciada, mientras que para el caso de la válvula ubicada a lo largo de la línea la señal que registra la herramienta es en forma de dos montañas, con la primera más alta que la segunda.

El micrófono registra los cambios en la intensidad del ruido generado cuando el cuerpo al que se adhiere toca la superficie de la tubería. La herramienta de inspección cuenta con dos micrófonos, los cuales registran el sonido producido cuando los discos de la herramienta atraviesan la válvula. Al comparar las figuras (4.6 y 4.7) se observa que en el primer caso existe una variación prolongada de cambios bruscos en la señal, la cual no se debe totalmente a la presencia de la válvula, mientras en el segundo caso la variación se presenta sólo en el

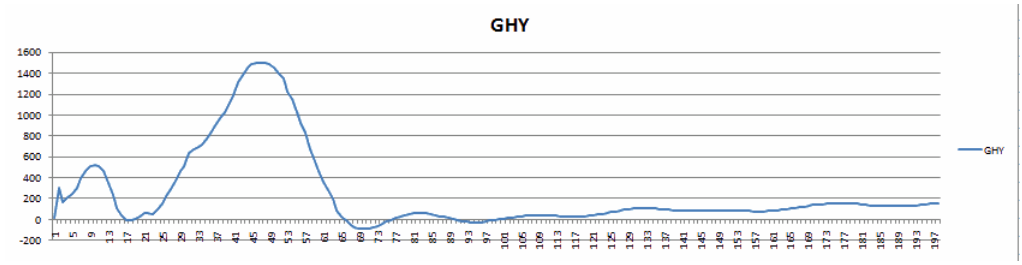


Figura 4.4: *Efecto del paso de la herramienta por una válvula ubicada en un extremo de la línea de transporte registrado por el sensor giroscopio.*

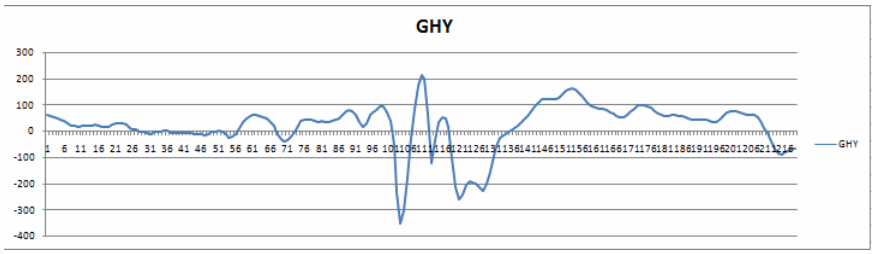


Figura 4.5: *Efecto del paso de la herramienta por una válvula ubicada a lo largo de la línea de transporte registrado por el sensor giroscopio.*

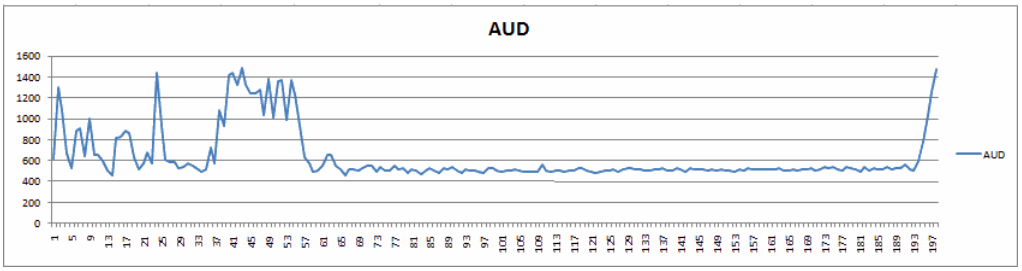


Figura 4.6: *Efecto del paso de la herramienta por una válvula ubicada en los extremos de la línea registrado por el sensor micrófono.*

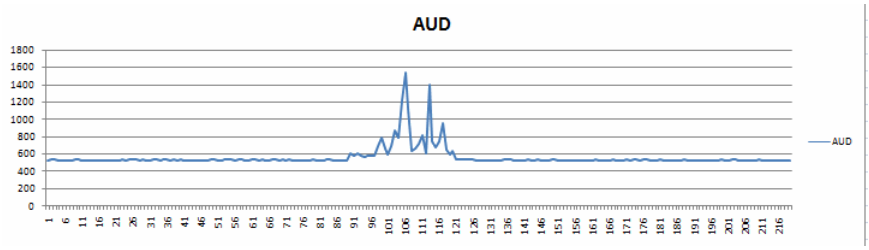


Figura 4.7: *Efecto del paso de la herramienta por una válvula ubicada a lo largo de una línea registrado por el sensor micrófono.*

punto donde está ubicada la válvula.

Partiendo del análisis de las señales con las cuales el experto reconoce una válvula se puede afirmar que existen dos tipos de comportamientos uno si la válvula se localiza en los extremos de la línea y otro si la válvula se encuentra a lo largo.

4.2. Selección y preparación de los datos

Los datos de entrada (atributos) corresponden a seis inspecciones inteligentes realizada durante los años 2009 a 2012 a tres gasoductos colombianos que cuenta con las siguientes características técnicas (ver tabla 4.1):

Característica	Gasoducto 1	Gasoducto 2	Gasoducto 3
Longitud [km]	35.8	23.8	110.4
Diámetro nominal interno [in]	24	20	20
# Válvulas	3	3	4
Total válvulas registradas	6	6	8
Año inicio operación	1990	2000	2000

Tabla 4.1: Características de cada gasoducto

Los datos suministrados por la Corporación para la investigación de la corrosión (CIC) presentan las siguientes características técnicas (ver tabla 4.2):

Característica	Gasoducto 1	Gasoducto 2	Gasoducto 3
Extensión del archivo	ition	ition	ition
Tamaño [Mb]	131	129	685
# Registros	418.363	2'150.874	1'279.886
Variables de entrada	55	55	55

Tabla 4.2: Características del archivo fuente

Los atributos utilizados en la etapa de experimentación fueron seleccionados partiendo del criterio de los expertos CIC y se muestran en la tabla 4.3.

Los sensores Ahy, Ahx y Ahz corresponden al valor de la aceleración experimentado por la herramienta en los tres ejes espaciales a medida que se desplaza por la tubería. Los sensores Ghx, Ghy, Ghz miden los cambios en la inclinación y las vibraciones en los tres ejes espaciales. Los sensores Mhy, Mhx y Mhz miden el cambio en el campo magnético de la tierra en los tres ejes espaciales. Los sensores PRES1 y PRES2 registran la presión a la cual viaja la herramienta. Los sensores AUD1 y AUD2 obtienen el valor del nivel de ruido que se produce cuando la herramienta choca con la tubería. El sensor DIS registra el desplazamiento de la herramienta a lo largo de la tubería.

Como se mencionó en la sección 4.1 existen dos tipos de comportamientos en los datos. Sin embargo, después de realizar una revisión a los históricos de inspección suministrada por la CIC no se encontró suficiente información para generar dos tipos de archivos de datos que permieran crear dos modelos uno para válvulas en los extremos y otro para válvulas a lo largo de la línea, por está razón se construyó sólo un archivo de datos.

Variable	Notación	Tipo dato	Unidad	Valor máximo	Valor mínimo
Aceleración eje X	Ahx	Entero	Lsb	5454	-5454
Aceleración eje Y	Ahy	Entero	Lsb	5454	-5454
Aceleración eje Z	Ahz	Entero	Lsb	5454	-5454
Giroscopio eje X	Ghx	Entero	Lsb	6000	-6000
Giroscopio eje Y	Ghy	Entero	Lsb	6000	-6000
Giroscopio eje Z	Ghz	Entero	Lsb	6000	-6000
Magnetómetro eje X	Mhx	Entero	Lsb	2500	-2500
Magnetómetro eje Y	Mhy	Entero	Lsb	2500	-2500
Magnetómetro eje Z	Mhz	Entero	Lsb	2500	-2500
Presión	PRES1	Entero	Lsb	2000	0
Presión	PRES2	Entero	Lsb	2000	0
Audio	AUD1	Entero	Lsb	99999999	-99999999
Audio	AUD2	Entero	Lsb	99999999	-99999999
Distancia	DIS	Flotante	m	99999999	0

Tabla 4.3: Atributos seleccionados para reconocer una válvula según criterio de expertos

La información se recolectó de forma secuencial comenzando por los registros ubicados antes de la válvula, luego los datos correspondientes a la válvula y finalizando con los registros después de la válvula. Los datos fueron separados en 20 hojas de cálculo (una por cada válvula) y cada hoja en promedio con 159 registros, las cuales se guardaron en formato xls.

4.3. Procesamiento

Se realizó un proceso de verificación a cada hoja de cálculo con el fin de eliminar aquellos registros que tuviesen algún valor fuera del rango de aceptación , registros duplicados o registros cuando la herramienta no se encontrara funcionando. Luego se continuó con la creación de la etiqueta de clase denominada válvula, la cual se le asignó dos posibles valores 'S' si el registro corresponde a la válvula y 'N' si no lo era. Para asignar el valor a la etiqueta fue necesario tomar los documentos entregados por el propietario del gasoducto,localizar el punto donde se encuentra instalada la válvula y marcar con una 'S' ese registro en el archivo de datos, a los demás registros se marcó con 'N'. Adicionalmente, se asumió un margen de un metro como la región donde se encuentra la válvula. Lo anterior porque la ubicación suministrada por el propietario del gasoducto en algunos casos no es exacta, por ejemplo para la empresa la válvula esta en el posición Km 100+ 50 m. Como la herramienta captura datos cada 6 cm, el experto no sabía exactamente el punto dentro del Km 100 + 50 m donde está la válvula, pero de acuerdo con los datos, la region del Km 100 + 50 m al Km 100 + 51m registra cambios

que son reconocidos de forma visual y permiten reconocer la válvula. También, fue necesario asumir esta hipótesis para aumentar el número de instancias positivas. Un registro o instancia para la válvula lo conforman los 14 atributos definidos por los expertos y la clase válvula.

Para la fase de modelado y con el objeto de realizar varias pruebas para medir la estabilidad del modelo, se construyen dos (2) conjuntos (sets) para el entrenamiento en formato csv (comma-separated values) con el siguiente número de instancias: 140, 244. Corresponden al 60 % y 80 % del total de las válvulas. El primer set de datos contiene 46 instancias clasificadas como válvulas y 94 instancias que son NO válvulas. En el segundo set de datos 87 instancias se clasifican como válvulas y 157 son NO válvulas. También, se crearon 2 conjuntos (sets) en formato csv para la fase de pruebas que contienen 60 y 94 instancias distribuido de la siguiente forma: El conjunto de 60 instancias contiene 21 registros clasificados como válvulas, 39 clasificados como NO válvulas. El conjunto de 90 instancias contiene 34 registros clasificados como válvulas y 56 clasificados como NO válvulas.

Los datos empleados para el entrenamiento son diferentes a los seleccionados para las pruebas, sin embargo, es importante aclarar que dentro de los datos de entrenamiento el set de 244 si contiene los datos del set de 140. Ese mismo criterio se aplicó para el caso de los datos de validación.

4.4. Modelado

Se construyen dos modelos para la detección automática de válvulas a partir de la información suministrada, aplicando dos técnicas de aprendizaje de máquina: redes neuronales y máquinas de vectores de soporte (para una mayor comprensión de estos modelos favor dirigirse a la sección 2.5.2). La razón por la cual se utilizaron estas dos técnicas se debe a que la información de entrada es casi la misma utilizada para el reconocimiento de soldaduras donde se obtuvieron buenos resultados, desafortunadamente, en la revisión de la literatura no se encontraron trabajos relacionados que ofrecieran una guía al respecto.

El software seleccionado para realizar el entrenamiento y la validación fue Weka en su versión 3.7.9 porque es una de las herramientas software más completas para el análisis de datos gracias a las diferentes técnicas para el pre procesamiento y modelado de datos que implementa, adicionalmente, es muy portable porque se encuentra desarrollado con tecnología Java lo cual le permite correr en casi cualquier plataforma , esta disponible libremente bajo licencia GNU y sus librerías se pueden integrar fácilmente en otras aplicaciones software. Los datos para el entrenamiento y las pruebas se ingresaron al software Weka . Luego, siguiendo el mismo procedimiento utilizado para la detección las soldaduras, se aplicó la técnica de selección de atributos (aplicando el algoritmo CfsSubsetEval) y reducción de la dimensionalidad (PCA) creándose así dos nuevos conjuntos de datos por cada archivo de entrenamiento y de pruebas. En total se crearon seis (6) archivos y se guardaron en formato arff (Attribute-Relation File Format).

La tabla 4.4 se muestran los diferentes archivos utilizados durante el entrenamiento, la técnica de pre-procesamiento y los atributos de entrada.

PRUEBA	**Técnica pre-procesamiento**	**Atributos relevantes de acuerdo con el algoritmo de pre-procesamiento**
140 instancias	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 , AUD1, AUD2.
140 instancias	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, AUD1, AUD2.
140 instancias	*PCA*	
244 instancias	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 , AUD1, AUD2.
244 instancias	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, AUD1, AUD2.
244 instancias	*PCA*	

Tabla 4.4: Atributos más relevantes para el reconocimiento de válvulas.

4.5. Pruebas y discusión

4.5.1. Pruebas con redes neuronales

Cada archivo se entrenó con una red neuronal del tipo perceptrón multicapa (para conocer más acerca de este modelo favor dirigirse a la sección 2.5.2). Después de realizar 6 pruebas, la configuración que se ajustó mejor a los datos fue la siguiente:

- N neuronas en la capa de entrada (con N entre 7 y 14 neuronas).
- M neuronas en la capa oculta (con M entre 4 y 8 neuronas)
- 2 neuronas en la capa de salida.

En la figura 4.8, se muestra las pruebas realizadas a cada archivo de entrenamiento, la técnica de pre procesamiento y el porcentaje de error alcanzado por el modelo, usando validación cruzada y conjuntos de pruebas. Para el caso de la validación cruzada se utilizaron 3 tipos de iteraciones (10, 20 y 40) mientras que los conjuntos de pruebas constan de 60, 94 instancias. El primero contiene 21 instancias clasificadas como válvulas y 39 como NO Válvulas. En el segundo conjunto independiente posee 34 instancias clasificada como válvula y 60 instancias como NO válvula.

Partiendo de los resultados obtenidos en la validación cruzada y los conjuntos de pruebas se puede concluir que el modelo reconoce válvulas de forma más eficiente cuando se aplica pre procesamiento con las variables de entrada GHY, AHY, MHY, AUD1 y AUD2. Además, al comparar los resultados por set de datos se aprecia una mejora en la tasa de éxito del clasificador entre el 5 y 20 % cuando se utiliza técnicas de selección de atributos. Esto significa que existen datos poco representativos los cuales impiden al algoritmo de entrenamiento que utiliza redes neuronales clasificar correctamente todas las instancias.

PRUEBA	Técnica pre procesamiento	Atributos entrada	% Error en cada prueba					
			Validación Cruzada			Conjunto Entrenamiento		
			10 Folds	20 Folds	40 Folds	60 (21 S)	94 (34 S)	
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	9,02	7,32	6,90	35,36	37,63	
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	18,25	18,06	17,11	18,32	42,11	
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*PCA*		18,09	15,69	14,27			
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	12,26	12,63	10,75	48,44	35,44	
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	29,52	19,87	17,85	27,34	37,69	
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*PCA*		31,99	24,87	24,57			

Figura 4.8: *Pruebas realizadas a los datos de entrenamiento y su porcentaje de error.*

PRUEBA	Técnica pre-procesamiento	Atributos entrada	Comportamiento métrica ROC									
			Validación cruzada						Set Independiente			
			10 Folds		20 Folds		40 Folds		60 (21 S)		94 (34 S)	
			Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	97,83%	96,81%	97,83%	97,87%	97,83%	98,94%	86%	58,97%	52,94%	73,33%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	82,61%	97,87%	84,78%	97,87%	84,78%	97,87%	81%	97,44%	50,00%	98,33%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*PCA*		89,13%	98,94%	89,13%	100,00%	89,13%	98,94%				
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	89,66%	100,00%	87,36%	99,36%	91,95%	100,00%	90%	33,33%	82,35%	56,67%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	78,16%	96,82%	85,06%	97,45%	87,36%	96,82%	71%	94,87%	61,76%	96,67%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*PCA*		68,97%	97,44%	74,71%	98,73%	73,56%	98,73%				

Figura 4.9: *Pruebas realizadas a los datos de entrenamiento usando análisis ROC.*

En la figura 4.9, se muestra el comportamiento del modelo validado a partir de la técnica de análisis ROC. De acuerdo con estos resultados, la mayor efectividad se obtiene con 244 instancias usando como entrada las señales de los sensores GHY, GHZ, AHY, MHY, MHZ, AUD1 y AUD2. El modelo es más preciso para reconocer las no válvulas comparado con las válvulas. Al analizar el comportamiento del modelo con cada conjunto de entrenamiento usando técnicas de pre procesamiento contra los datos originales, se puede afirmar que el algoritmo mejora la especificidad, es decir la probabilidad de encontrar eficientemente los casos clasificados como negativos, pero disminuye la sensibilidad por tanto la probabilidad de encontrar exitosamente los casos catalogados como positivos es menor cuando el modelo pierde información usando técnicas de selección de atributos o reducción de la dimensionalidad.

En la figura 4.10 se muestra el comportamiento del modelo para todas las instancias entrenadas frente a las 94 instancias de pruebas. Al usar las 140 instancias y la técnica de selección de atributos, la red neuronal sólo identifica en un cincuenta por ciento de forma correcta la válvula , al aumentar el número de instancias a 244 instancias y la técnica de selección de atributos se presenta un crecimiento al 61.76 % en la precisión para el reconocimiento de la válvula, sin embargo este valor sigue estando lejos del 100 %. Para el caso de las no válvulas, el modelo reconoce las no válvulas con un margen de error entre el 4 y 2 %.

4.5.2. Pruebas con máquinas de vectores de soporte

Cada archivo se entrenó usando dos tipos de kernel: Polinomio normalizado de grado 13 y el PUK ó *Pearson Universal Kernel* (para mayor información sobre estos tipos de kernel diríjase a la sección 2.5.2)se seleccionaron estos dos kernels con base en los resultados obtenidos durante el trabajo realizado para soldaduras tomando como referencia que la información de entrada era practicamente la misma. También se realizaron pruebas con otros tipos de kernel(como PolyKernel o RBFKernel)pero los resultados obtenidos no fueron significativos por lo tanto se descartaron.
Luego se estimó el error del modelo por medio de validación cruzada (*cross validation*) con 10, 20 y 40 iteraciones y conjuntos de pruebas de 60 y 94 instancias.

En la figura 4.11 se muestran las diferentes pruebas realizadas al modelo y el porcentaje de error obtenido. De acuerdo con los resultados, el modelo de máquinas de vectores de soporte clasifica con mayor eficiencia las válvulas cuando utiliza 244 instancias con todos los atributos utilizando el kernel PUK. Al comparar las pruebas por cada conjunto de datos, se puede afirmar que en principio si se utilizan técnicas de pre procesamiento y se evalua el algoritmo con validación cruzada el modelo posee un margen de error más bajo con menos variables de entrada. Sin embargo, al emplear conjuntos no entrenados los resultados de las pruebas realizadas al algoritmo que ha sido pre procesado son menos eficientes que si se utilizan todas las variables de entrada. Lo anterior, podría suponer que haría falta más información de entrenamiento para garantizar que el algoritmo de clasificación funcione mejor usando técnicas de pre procesamiento. Desafortunadamente, en este momento no se cuenta con los datos necesarios para validar esta hipótesis.

En la figura 4.12 los mismos resultados son analizados con la técnica de análisis ROC. De acuerdo con estos resultados, la mayor efectividad se obtiene con 244 instancias usando todas las variables de entrada. El modelo es más preciso para reconocer una no válvula en comparación con la válvula porque en el 95 % de las pruebas el valor de la especificidad fue mayor comparado con la sensibilidad.

Al comparar los resultados de las pruebas par cada set de entrenamiento, es posible hacer la siguiente afirmación: usando evaluación cruzada, la sensibilidad y especificidad es mayor cuando se utilizan técnicas de pre procesamiento. No obstante cuando el modelo se evalua con datos independientes, el algoritmo de reconocimiento presenta dificultad para identificar correctamente una válvula y este problema es más notorio cuando se utilizan técnicas de pre procesamiento. Lo anterior confirma que con la información que se cuenta, no es posible disminuir el número de variables sin que se pierde eficiencia a la hora de reconocer una válvula.

En la figura 4.13 se muestra la evolución en el comportamiento del modelo para todas las instancias entrenadas frente a las 94 instancias de pruebas. Al usar las 140 instancias la máquina de soporte vectorial no identifica correctamente las NO válvulas (especificidad de 65 %), al aumentar el número de instancias a 244 la máquina de soporte reconoce mejor la NO válvula, aunque se pierde precisión al reconocer la válvula (sensibilidad entre 58 y 85 % comparado con una sensibilidad entre el 70 y 85 % para el caso de las 144 instancias). Existe usando todos los atributos y el kernel PUK, 5 casos de falsos negativos, es decir instancias que están etiquetadas como no válvulas y el modelo las reconoce como válvulas y 3 casos de falsos positivos, registros etiquetados como NO válvulas y el modelo las reconoce como válvulas.

Finalmente, si al unir los mejores resultados de cada técnica (ver tabla de la figura 4.14), se puede afirmar que es posible reconocer de forma automática las válvulas utilizando máquinas de vectores de soporte y el modelo es más eficiente si se emplean todos los atributos de entrada con un kernel PUK con una sensibilidad del 85.29 % y especificdad del 95 %.

4.5.3. Discusión

- Con base en los resultados obtenidos, se puede afirmar que el modelo que implementa máquinas de vectores de soporte con kernel PUK posee un mejor desempeño para reconocer válvulas con una sensibilidad del 85.29 % y especificidad del 95 % comparado con el modelo de redes neuronales, el cual posee una sensibilidad del 61.78 % y especificidad del 96.6 % . No obstante, es importante revisar con detalle las pruebas para analizar las instancias que no fueron clasificadas correctamente. En la figura 4.15 se muestra un ejemplo de una instancia que fue clasificada como falso negativo, como se aprecia, todos los valores ubicados desde la posición 17729.2 hasta 17729.9 corresponden a una misma válvula, así que si la instancia ubicada en la posición 17729.77 no fue clasificada correctamente el experto puede asumir sin llegar a equivocarse que la válvula está en ese punto. Lo anterior, garantiza una mayor confianza en el modelo. No ocurre lo mismo con las instancias catalogadas como falsos positivos. Es necesario realizar un estudio más profundo sobre este tema una vez se cuenta con más información para entrenar.

- Con base en los resultados, se demuestra que es posible reconocer de forma automática los sitios donde se localizan las válvulas en una tubería, lo anterior permitirá a los experto de la CIC contar con una herramienta de apoyo que facilite la toma de decisiones. Sin embargo, como los modelos desarrollados poseen un margen de error, queda a criterio del experto definir para los puntos donde el modelo clasifica como válvula y no existe registro de la empresa. En esos casos, el experto podría apoyarse en la forma como se delimitó la válvula en el modelo, la cual considera la agrupación de varios registros consecutivos.

- La industria podrá verse beneficiada de los resultados en la medida que la CIC mejore el modelo agregando información de nuevas inspecciones. De acuerdo con estudios realizados por Ecopetrol el hurto de combustible representa pérdidas anuales de alrededor de 11 millones

de dólares.

PRUEBA	Técnica pre-procesamiento empleada	Técnica pre-procesamiento	Set de pruebas					
			94 (34 S)					
			Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Negativos (FN)	Falsos Positivos (FP)	Sensibilidad (VPR)	Especificidad (SPC)
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	18	44	16	16	52,94%	73,33%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	17	59	17	1	50,00%	98,33%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	28	34	6	26	82,35%	56,67%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	21	58	13	2	61,76%	96,67%

Figura 4.10: *Resultados análisis ROC para 94 instancias.*

PRUEBA	Técnica pre procesamiento	Atributos de entrada	Algoritmo clasificación	% Error en cada prueba					
				Validación Cruzada			Conjunto Pruebas Independiente		
				10	20	40	60 (21 S)	94 (34 S)	
140 instancias 46 instancias son Valvulas 94 instancias son NO	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO NormalizedPolykernel	14,53	14,53	14,53	37,10	36,38	
140 instancias 46 instancias son Valvulas 94 instancias son NO	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO PUKkernel	14,53	14,53	14,53	18,55	30,51	
140 instancias 46 instancias son Valvulas 94 instancias son NO	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	SMO NormalizedPolykernel	19,37	19,37	19,37	18,55	30,51	
140 instancias 46 instancias son Valvulas 94 instancias son NO	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	SMO PUKkernel	9,69	11,30	9,68	25,97	32,86	
140 instancias 46 instancias son Valvulas 94 instancias son NO	*PCA*		SMO NormalizedPolykernel	41,98	41,97	40,35			
140 instancias 46 instancias son Valvulas 94 instancias son NO	*PCA*		SMO PUKkernel	16,15	16,14	16,14			
244 instancias 87 instancias son Valvulas 157 instancias son NO	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO NormalizedPolykernel	24,09	24,09	23,20	29,16	41,57	
244 instancias 87 instancias son Valvulas 157 instancias son NO	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO PUKkernel	15,17	14,28	13,38	10,93	18,48	
244 instancias 87 instancias son Valvulas 157 instancias son NO	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	SMO NormalizedPolykernel	18,74	19,63	19,63	10,93	27,71	
244 instancias 87 instancias son Valvulas 157 instancias son NO	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	SMO PUKkernel	13,39	12,49	11,60	18,22	32,33	
244 instancias 87 instancias son Valvulas 157 instancias son NO	*PCA*		SMO NormalizedPolykernel	19,63	16,95	18,74			
244 instancias 87 instancias son Valvulas 157 instancias son NO	*PCA*		SMO PUKkernel	21,42	18,74	17,85			

Figura 4.11: *Resultados obtenidos para los modelos de máquinas de vectores de soporte.*

PRUEBA	Técnica pre procesamiento	Atributos entrada	Algoritmo clasificación	Validación cruzada						Set pruebas			
				10 Folds		20 Folds		40 Folds		60 (21 S)		94 (34 S)	
				Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)	Sensibilidad (VPR)	Especificidad (SPC)
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO NormalizedPolykernel	84,78%	97,87%	84,78%	97,87%	84,78%	97,87%	52,38%	100,00%	70,59%	65,00%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO PUKkernel	84,78%	97,87%	84,78%	97,87%	84,78%	97,87%	76,19%	100,00%	85,29%	65,00%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	SMO NormalizedPolykernel	76,09%	98,94%	76,09%	98,94%	76,09%	98,94%	76,19%	100,00%	79,41%	65,00%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	SMO PUKkernel	86,96%	100,00%	84,78%	100,00%	86,96%	100,00%	66,67%	100,00%	85,29%	65,00%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO NormalizedPolykernel	82,76%	98,73%	70,11%	99,36%	71,26%	99,36%	95,24%	82,05%	67,65%	88,33%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguna	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SMO PUKkernel	82,76%	98,73%	82,76%	99,36%	83,91%	99,36%	95,24%	94,87%	85,29%	95,00%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	SMO NormalizedPolykernel	82,76%	96,18%	82,76%	95,54%	82,76%	95,54%	95,24%	94,87%	70,59%	96,67%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	SMO PUKkernel	86,21%	98,09%	85,06%	99,36%	86,21%	99,36%	76,19%	100,00%	58,82%	100,00%

Figura 4.12: *Pruebas realizadas a los datos de entrenamiento usando análisis ROC.*

PRUEBA	Técnica pre procesamiento	Atributos entrada	Algoritmo clasificación	Conjunto validación 94 (34 S)					
				Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Negativos (FN)	Falsos Positivos (FP)	Sensibilidad (VPR)	Especificidad (SPC)
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SVM NormalizedPolykernel	24	39	10	21	70,59%	65,00%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SVM PUKkernel	29	39	5	21	85,29%	65,00%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	SVM NormalizedPolykernel	27	39	7	21	79,41%	65,00%
140 instancias 46 instancias son Valvulas 94 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, ,AUD1, AUD2.	SVM PUKkernel	29	39	5	21	85,29%	65,00%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SVM NormalizedPolykernel	23	53	11	7	67,65%	88,33%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, PRES1, PRES2 ,AUD1, AUD2.	SVM PUKkernel	29	57	5	3	85,29%	95,00%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	SVM NormalizedPolykernel	24	58	10	2	70,59%	96,67%
244 instancias 87 instancias son Valvulas 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY,AUD1, AUD2.	SVM PUKkernel	20	60	14	0	58,82%	100,00%

Figura 4.13: *Resultados análisis ROC para 94 instancias.*

PRUEBA	Técnica pre procesamiento	Atributos entrada	Algoritmo clasificación	Conjunto Entrenamiento					
				34 instancias : **34 Soldaduras** 56 NO Soldaduras					
				Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Negativos (FN)	Falsos Positivos (FP)	Sensibilidad (VPR)	Especificidad (SPC)
244 instancias **87 instancias son Valvulas** 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, GHZ, AHY, MHY, MHZ, AUD1, AUD2.	MultilayerPerceptron	21	58	13	2	61,76%	96,67%
244 instancias **87 instancias son Valvulas** 157 instancias son NO Valvulas	Ninguno	GHX, GHY, GHZ, AHX, AHY, AHZ MHX, MHY, MHZ, AUD1, AUD2	SMO PUKkernel	29	57	5	3	85,29%	95,00%
244 instancias **87 instancias son Valvulas** 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY, AUD1, AUD2	SMO NormalizedPolykernel	24	58	10	2	70,59%	96,67%
244 instancias **87 instancias son Valvulas** 157 instancias son NO Valvulas	*CfsSubsetEval*	GHY, AHY, MHY, AUD1, AUD2	SMO PUKkernel	20	60	14	0	58,82%	100,00%

Figura 4.14: *Mejores resultados RNA y SMO.*

		RNA Perceptrón multicapa	**SVM** Kernel: Normalized Polykernel	**SVM** Kernel: PUK
Distancia	Salida Real	Salidad Predicha	Salidad Predicha	Salidad Predicha
17727,82	N	N	N	S
17727,89	N	N	N	S
17727,94	N	N	N	S
17728,01	N	N	N	S
17729,2	S	S	N	N
17729,27	S	S	N	N
17729,35	S	S	N	N
17729,41	S	S	N	N
17729,47	S	S	S	S
17729,55	S	S	S	S
17729,62	S	S	S	S
17729,67	S	S	S	S
17729,77	S	N	N	N
17729,93	S	S	S	S
17730,8	N	N	N	N
17730,86	N	N	N	N
17730,92	N	N	N	N
17730,98	N	N	N	N

Figura 4.15: *Análisis de un caso de falso negativo en la identificación de válvulas.*

Capítulo 5

Conclusiones

A continuación se presentan las conclusiones del trabajo realizado:

- Las pruebas realizadas confirman que es posible identificar de forma automática soldaduras y válvulas usando técnicas de aprendizaje de máquina y de minería de datos logrando una eficiencia entre el 98 al 99 porciento para el caso de las soldaduras y entre el 82 al 90 porciento en las válvulas. Gracias a estos resultados es posible diseñar herramientas software que implementen dichas técnicas y le sirva como apoyo para el análisis de datos en menor tiempo eliminando así tareas manuales y repetitivas como se desarrollan en la actualidad.

- Con base en los resultados obtenidos se puede concluir que para reconocer una soldadura en tuberías que transportan gas con la herramienta de inspección inteligente desarrollada por la CIC sólo es necesario las señales obtenidas de los siguientes sensores: acelerómetro, magnetómetro en el eje Y y uno de los sensores de audio. Para el caso de la válvula se requiere la información suministrada por el giroscopio, el acelerómetro y el magnetómetro en el eje Y junto con los dos sensores de audio y los sensores de presión.

- Para el caso de la identificación de soldaduras, el uso de algoritmos de selección de atributos sirvió para reconocer de forma más eficiente el fenómeno. Además,el uso de este tipo de tecnicas representa un ahorro de operaciones computacionales que puede ser representativo en la medida que se analicen un gran volumen de datos.

- Se desarrolló un modelo para reconocer de forma automática soldaduras y válvulas en tuberías que transportan hidrocarburos utilizando datos de una herramienta de inspección inteligente con tecnología inercial, de la cual, no se tienen trabajos reportados en la literatura.

Capítulo 6

Trabajos Futuros

- Como se mencionó en las secciones 3.6 y 4.5.3 se requiere validar los modelos de reconocimiento de soldadura y válvula con datos de nuevas inspecciones donde por ejemplo exista variación en el diámetro, el material de la tubería, la edad, el tipo fluido que transporta entre otros factores para medir el grado de sensibilidad del modelo.

- Continuar trabajando en la identificación de otros fenómenos como las abolladuras, fugas, puntos bajos y altos a lo largo de la línea de transporte, magnetos entre otros. Los cuales son importantes para el correcto funcionamiento de las líneas de transporte de hidrocarburos y pueden ser reconocidos con la tecnología inercial utilizada por la herramienta de inspección inteligente de la CIC. Adicionalmente se encuentra en fase de desarrollo una nueva herramienta de inspección que combina las tecnologías inercial y MFL, lo anterior abre las puertas para que se haga investigación en áreas como la detección de defectos, pérdida de espesor de pared, problemas de corrosión entre otros.

- Para mejorar la eficiencia en el reconocimiento de válvulas se sugiere clasificar las válvulas entre las de inicio y fin de la línea de transporte y las del medio y construir modelos por separado, porque como se menciona en 4.1 y en 4.5.3, el hecho que la herramienta se comporte diferente y los datos no cumplan con un único patrón representa una dificultad para el modelo actual, si a este problema se añade la falta de datos hace que en estos momentos no sea posible alcanzar mejores resultados. Entonces es necesario alimentar el modelo con más información que se obtiene a medida que la CIC realice nuevas inspecciones. Además, sería interesante que los modelos cuenten con información de inspecciones realizadas a líneas de crudo o sus derivados, lo anterior con el fin de revisar la importancia que tienen variables como la presión, la cual para el caso de las líneas que transportan gas no fue representativa.

Bibliografía

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[2] J. Cordell and H. Vanzant, *All about pigging: the design of pipelines and facilities for conventional and intelligent pigging and a guide to pig selection, operation and maintenance and to pipeline pigging services.* On-Stream Systems. Firm, 1996.

[3] A. Carvalho, J. Rebello, L. Sagrilo, C. Camerini, and I. Miranda, "Mfl signals and artificial neural networks applied to detection and classification of pipe weld defects," *Ndt & E International*, vol. 39, no. 8, pp. 661–667, 2006.

[4] G. LATORRE, R. MORA, F. MEJÍA U, A. MARTÍNEZ, and R. SUÁREZ, "Análisis estructural de tuberias de oleoductos abolladas por carga explosiva," *CT&F-Ciencia, Tecnología y Futuro*, vol. 1, no. 4, pp. 101–110, 1998.

[5] R. D. Souza, "Avaliação estrutural de dutos com defeitos de corrosão reais," *Pós-Graduação em Engenharia Mecânica, PUC-Rio, Rio de Janeiro, Dissertação de Mestrado, 112p*, 2003.

[6] J. Tiratsoo, *Pipeline pigging and integrity technology.* Scientific Surveys Limited, 2003.

[7] D. Santana, N. Maruyama, and C. Furukawa, "Estimation of trajectories of pipeline pigs using inertial measurements and non linear sensor fusion," in *Industry Applications (INDUSCON), 2010 9th IEEE/IAS International Conference on.* IEEE, 2010, pp. 1–6.

[8] J. D. Hart, G. H. Powell, D. Hackney, and N. Zulfiqar, "Geometry monitoring of the trans-alaska pipeline," in *11-th International Conference on Cold Region Engineering, Anchorage*, 2002.

[9] R. Carneval, M. G. MARINHO, and J. Santos, "Flexible line inspection," in *European Conference on Nondestructive Testing (ECNDT)*, 2006.

[10] M. Beller, "Applying ultrasound for in-line inspection: Facts and issues," in *PPSA Aberdeen Seminar, UK*, 2006.

[11] E. C. de Petróleos, *Carta petrolera.* Oficina de Divulgación y Prensa, 2007, no. 72-82.

[12] J. D. Aldana Carvajal, A. A. Yepes Maldonado, D. Padilla, and N. Nabonazar, "Análisis de la hidrodinámica exterior de una herramienta multitareas mediante cfd." Ph.D. dissertation, 2010.

[13] G. A. Londoño Vélez *et al.*, "Prototipo pig intelligent," Ph.D. dissertation, Universidad Nacional de Colombia-Sede Manizales, 2003.

[14] J. Y and U. R, "Diseno y construcción de un prototipo "smart pig"que permita el monitoreo de tuberías en oleoductos, basado en la estrategia magnetic flux leakage (mfl)." 2012.

[15] A. Khodayari-Rostamabad, J. P. Reilly, N. K. Nikolova, J. R. Hare, and S. Pasha, "Machine learning techniques for the analysis of magnetic flux leakage images in pipeline inspection," *Magnetics, IEEE Transactions on*, vol. 45, no. 8, pp. 3073–3084, 2009.

[16] R. K. Amineh, N. K. Nikolova, J. P. Reilly, and J. R. Hare, "Characterization of surface-breaking cracks using one tangential component of magnetic leakage field measurements," *Magnetics, IEEE Transactions on*, vol. 44, no. 4, pp. 516–524, 2008.

[17] W. K. Muhlbauer, *Pipeline risk management manual: ideas, techniques, and resources.* Gulf Professional Pub, 2004.

[18] W. Villarreal Tapia, "Determinación del riesgo de falla por abolladuras en oleoducto usando método de elementos finitos," 2012.

[19] M. A. C. Ruiz, "Análisis comparativo de evaluación de defectos en ductos entre estudios realizados con equipos instrumentados inteligentes de segunda y tercera generació."

[20] H. Cordell, Jim; Vanzant, *Pipeline Pigging Handbook.*, 3rd ed. Clarion Technical Publishers., 2003. [Online]. Available: Onlineversionavailableat:http://www.knovel.com/web/portal/browse/display?_EXT_KNOVEL_DISPLAY_bookid=2934&VerticalID=0

[21] J. Pitchford, "Specification and requirements for the intelligent pig inspection of pipelines," *Pipes & pipelines international*, vol. 44, no. 1, pp. 17–27, 1999.

[22] A. G. Islas Garrido, "Análisis experimental de esfuerzos en tubos con abolladuras sujetos a presión interna," Ph.D. dissertation, 2010.

[23] J. McCarthy, "What is artificial intelligence," *URL: http://www-formal. stanford. edu/jmc/whatisai. html*, 2007.

[24] S. J. Russell and P. Norvig, *Inteligencia Artificial: un enfoque moderno*, 1996.

[25] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[26] C. A. Vaz Jr, O. de QF Araújo, and J. L. de Medeiros, "Failure diagnostics using data mining tools," *Computer Aided Chemical Engineering*, vol. 27, pp. 1539–1544, 2009.

[27] X.-f. Wang, Y. Wang, C.-l. Jiang, and H.-w. Liang, "Natural gas pipeline leak detection based on data mining," in *Consumer Electronics, Communications and Networks (CECNet), 2011 International Conference on.* IEEE, 2011, pp. 492–494.

[28] J. Moreno and D. Ovalle, "Modelo de apoyo a la comercialización de electricidad usando lógica difusa y aprendizaje de máquina." *Dyna-Medellin*, vol. 76, no. 159, p. 67, 2009.

[29] S. M. Weiss and N. Indurkhya, *Predictive data mining: a practical guide.* Morgan Kaufmann, 1998.

[30] M. Burgueño, J. García-Bastos, and J. González-Buitrago, "Las curvas roc en la evaluación de las pruebas diagnósticas," *Med Clin (Barc)*, vol. 104, no. 17, pp. 661–70, 1995.

[31] X. Yun, D. Bo, T. Xiaoping, and S. Sha, "Ultrasonic in-line inspection of pipeline corrosion based on support vector machine multi-classifier," in *Control Conference (CCC), 2010 29th Chinese*. IEEE, 2010, pp. 2894–2899.

[32] M. Allouti, C. Schmitt, G. Pluvinage, J. Gilgert, and S. Hariri, "Study of the influence of dent depth on the critical pressure of pipeline," *Engineering Failure Analysis*, vol. 21, pp. 40–51, 2012.

[33] P. M. Kurowski, "Finite element analysis for design engineers," *Warrendale, PA: Society of Automotive Engineers, 2004. 212*, 2004.

[34] G. Latorre, R. Mora, F. Mejía U, A. Martínez, and R. Suárez, "Análisis estructural de tuberías de oleoductos abolladas por carga explosiva," *CT&F-Ciencia, Tecnología y Futuro*, vol. 1, no. 4, pp. 101–110, 1998.

[35] E. B. Marigorta, S. V. Suárez, and J. F. Francos, *Sistemas de bombeo*. Universidad de Oviedo, Departamento de Energía, 1994.

[36] S. Miller, "Prediction of dent size using tri-axial magnetic flux leakage intelligent pigs," *CORROSION 2007*, 2007.