Universidad de Morón.

Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias.

Tesis de Grado

"Aplicación de Minería de Datos Para el Análisis del Proceso Industrial de Brazing"

Alumno: Panzitta Laureano.

Matrícula: 48011595

Mayo 2022.

Escuela Superior Ingeniería, Informática y Ciencias Agroalimentarias

ACTA DEL TRIBUNAL

DE EXAMEN CORRESPONDIENTE AL TESIS DE GRADO EN LA CARRERA DE

Licenciatura en Sistemas

Aplicación de Minería de Datos Para el Análisis del Proceso Industrial de Brazing

PRESENTADO POR

Panzitta Laureano Ezequiel

REALIZADO BAJO LA **TUTORÍA** DE LA PROFESORA

Mg. Lic. Marisa Daniela Panizzi

INTEGRANTES DEL TRIBUNAL

Completar nombres y apellidos

_____	_____	_____
Integrante Nro. 1	**Integrante Nro. 2**	**Integrante Nro. 3**

En el día de hoy, ___ de _____ del año 20__ , en el aula _____ de la Universidad de Morón se reúne el tribunal que ha de juzgar la tesis de grado, constituido por los señores que arriba se citan.

Se procedió a escuchar la exposición con la que el/los alumno/s defendiera/n su Trabajo de Diploma y posteriormente en turno abierto por el Presidente del Tribunal, se escucharon las contestaciones a las aclaraciones solicitadas por los miembros del Tribunal sobre el mismo.

Realizada la deliberación sobre las circunstancias antes referidas, el Tribunal decidió conceder a la Tesis presentada la calificación de "_____" (Libro _____ , Folio _____ , Año 20__ , Acta Nro. ____ , Fecha ___ / ___ / 20__)

Y para que así conste, se extiende la presente en E.S.I.I.C.A de la Universidad de Morón, en la fecha anteriormente mencionada y que firman los Miembros del Tribunal.

_____	_____	_____
Integrante Nro. 1	**Integrante Nro. 2**	**Integrante Nro. 3**

Agradecimientos

Este largo camino emprendido hace varios años concluye al fin y en estas líneas quiero agradecer a todos los que me acompañaron en este largo proceso, mi familia, amigos y docentes, agradezco todo el apoyo brindado y la confianza depositada. Agradezco por sobre todo haber confiado en mí mismo y a pesar de las adversidades, con una voluntad inclaudicable haber cumplido uno de mis mayores anhelos, el formarme en esta disciplina tan apasionante como lo es la informática.

Resumen

El proceso industrial de Brazing, como otros procesos industriales originan una gran cantidad de datos que son factibles de ser capturados y analizados. En la actualidad el auge de la minería de datos como herramienta de análisis, proporciona una importante posibilidad de obtener información valiosa para la toma de decisiones que permitan optimizar y mejorar estos procesos industriales.

Esta tesis tiene como objetivo identificar aquellas variables que son determinantes para la calidad del proceso y para este propósito se aplicaron diversas técnicas de minería de datos como: regresión logística, arboles de decisión, máquinas de vector soporte y redes neuronales. Con estos análisis se construyeron varios modelos para predecir en diferentes condiciones si las piezas que se hornean en el proceso de Brazing pueden tener o no problemas de calidad.

Todos estos modelos predictivos fueron entrenados con datos recabados en la empresa estudiada y luego se pusieron a prueba para verificar su eficacia en las predicciones. Si bien no se logró establecer una correlación entre las variables de entrada del proceso y el resultado del mismo, quedan abiertas futuras líneas de investigación para automatizar la recolección de datos, aplicar estos mismos modelos a otras industrias que realicen el proceso de Brazing, y ampliar el alcance de los modelos involucrando otros procesos de la cadena de elaboración de un panel de aluminio.

Índice de Contenido

Índice de Figuras

Índice de Tablas

1) Capítulo 1: Introducción

En este capítulo se realiza una introducción del objeto de estudio de esta tesis, detallando el problema que se aborda en este estudio, los objetivos y alcances del trabajo. Por último, se detalla un diagrama conceptual referido a la metodología llevada a cabo para la solución propuesta.

1.1) Motivación

El avance exponencial en la tecnología provoca que los individuos y organizaciones generen cada vez más datos. El énfasis actual en el análisis de estos datos para la toma de decisiones ha provocado que se desarrolle un vasto campo de investigación en herramientas y modelos para acelerar y mejorar las técnicas de análisis.

La industria en general, y específicamente la metalúrgica es un ambiente donde día a día se producen grandes volúmenes de datos derivados de procesos que son susceptibles a mejoras. Es en este campo donde la minería de datos aporta su mayor contribución al sector permitiendo a través de sus diversas técnicas de análisis obtener un mayor conocimiento del comportamiento de los procesos, ayudando a la mejora en los resultados de éstos.

Este trabajo vincula directamente el análisis de datos con un proceso industrial específico para mostrar como la tecnología en el análisis de datos puede contribuir al desarrollo de la industria en general, y especialmente a las pequeñas y medianas empresas, utilizando herramientas y conocimientos que están ampliamente divulgados.

1.2) Problema

Las industrias en general, y para este caso en particular la industria metalúrgica, son fuentes de diversos procesos industriales tales como; laminación, extrusión, fundición, galvanoplastia, forja, soldadura, etc. En cualquier industria los procesos industriales conforman la base o medio para la

manufactura de productos. Estos procesos que son desarrollados en parte sobre una base teórica y científica, y en parte sobre una base empírica y pragmática no están exentos de errores, tal es el caso del proceso industrial de soldadura fuerte, que de ahora en más llamaremos por sus siglas en inglés Brazing y el cual es objeto de estudio en este trabajo.

El proceso de Brazing es complejo por la cantidad de variables que intervienen, por un lado, una atmosfera controlada de nitrógeno, quemadores que proporcionan la energía (calor) necesaria para realizar la soladura, la velocidad de recirculación del aire en la cámara de horneado, y el tiempo de permanencia de las piezas en la cámara de horneado. El control de todas estas variables y su relación unas con otras hacen que resulte complejo realizar un análisis de las causas ante eventuales defectos en la calidad de las piezas.

La minería de datos como herramientas de análisis de datos, brinda la posibilidad de obtener información valiosa acerca de la correlación de todas las variables intervinientes en el proceso que indiquen desviaciones, para luego poder tomar decisiones con el objetivo de mejorar la calidad del proceso.

1.3) Alcance del Trabajo

Los modelos de minería de datos desarrollados en este trabajo deben ser capaces de identificar la correlación entre las diversas variables que componen al proceso de Brazing y como estas impactan en el resultado de una pieza conforme o no conforme. Los datos analizados serán datos históricos relevados por el personal que ejecuta el proceso.

De acuerdo a este alcance los modelos que se logren definir serán particulares para el proceso de Brazing puntualmente en la industria donde se están relevando los datos, es decir, no aplicará a procesos de Brazing en general.

1.4) Objetivos de la Tesis

Esta tesis aborda un proceso industrial específico como lo es la soldadura fuerte (Brazing) y como la minería de datos puede aportar información valiosa para identificar tendencias en el proceso que permitan mejorar problemas de calidad. Con este trabajo se busca:

Objetivo principal:

- Construir modelos de minería de datos que permitan resolver los problemas de calidad que presenta el proceso industrial de Brazing.

Objetivos específicos:

- Construir el estado del arte de la minería de datos con una metodología adecuada.
- Analizar los diferentes modelos de minería de datos aplicados en la industria.
- Aplicar diferentes modelos de minería de datos con una metodología adecuada.
- Analizar y seleccionar las herramientas y lenguajes adecuadas para la aplicación de la minería de datos.

Objetivos secundarios:

- Comprender a que se deben los fallos de calidad de soldadura en las piezas.
- Entender el proceso a partir del análisis de datos históricos.
- Comprender en base a muchos algoritmos de minería de datos existentes en la actualidad, cuáles son los más adecuados para trabajar con este tipo de proceso.
- Al identificar la combinación de variables que producen problemas de calidad, poder realizar correcciones en el proceso que permitan evitar estos incidentes.

Esto permitirá mejorar considerablemente los siguientes puntos:

- Mejorar la calidad en la soldadura de las piezas.
- Evitar el descarte de piezas por fallas de soldadura.
- Mejorar el entendimiento que se tiene sobre el proceso.
- Poder implementar a futuro sistemas de monitoreo en tiempo real para corregir las variables durante el proceso.

1.5) Estructura del Trabajo

El presente trabajo de tesis se estructura en seis capítulos los cuales componen la totalidad del trabajo desarrollado. A continuación se resume el contenido de cada uno de estos capítulos:

1.5.1) Capítulo 1 – Introducción

Este capítulo explica la problemática y la motivación por la cual se origina este trabajo de tesis, define el alcance de la misma y plantea cuales son los objetivos que se pretenden alcanzar una vez finalizada.

1.5.2) Capítulo 2 – Estado del Arte

Este capítulo se centra en el desarrollo de un mapeo sistemático de la literatura (SMS) abordando sus distintas etapas, que permite brindar información sobre el estado actual de las investigaciones acerca de la minería de datos aplicada a procesos industriales. También se extraen conclusiones a partir de un análisis cuantitativo de los artículos seleccionados.

1.5.3) Capítulo 3 – Descripción del Proceso Industrial

Este capítulo detalla como está constituido el proceso industrial de Brazing, se realiza un desarrollo técnico de los paneles de aluminio relevando sus características constructivas. Se analizan las diferentes técnicas de soldadura existentes en la actualidad de la industria y particularmente se detallan aquellas características asociadas al proceso de Brazing. Se realiza una descripción del proceso de horneado en sus

distintas etapas y se analizan las características de cada módulo del horno. Finalmente se explica cómo se realizan las pruebas de estanqueidad que permiten determinar si un panel presenta o no fallas de soldadura a causa del proceso de horneado.

1.5.4) Capítulo 4 – Propuesta de Solución

En este capítulo se realiza el desarrollo de la metodología KDD como solución para la problemática planteada previamente. Se abordan cada una de las fases de la metodología desde la selección de las herramientas y lenguajes de programación, pasando por la recopilación e integración de los datos, la selección, limpieza y transformación, la minería de datos y por último la evaluación e interpretación que permite arribar a las conclusiones.

1.5.5) Capítulo 5 – Conclusiones y Futuras Líneas de Investigación

En este capítulo se describen las conclusiones surgidas a partir del trabajo realizado y se detallan posibles líneas de investigación a desarrollar en el futuro.

1.5.6) Capítulo 6 – Bibliografía

En este capítulo se detallan las distintas fuentes de información que fueron consultadas para la realización de esta tesis.

1.5.7) Capítulo 7 – Anexos

En este capítulo se aportan una serie de anexos donde se realiza una comparativa del trabajo realizado en el capítulo 4 en la fase de minería de datos con Python comparado a otra herramienta mucho más automatizada denominada Rapidminer que permite aportar más evidencia a las conclusiones obtenidas en la propuesta de solución.

2) Capítulo 2: Estado del Arte

2.1) Introducción

La minería y el análisis de datos han desempeñado un papel importante en el descubrimiento del conocimiento y el soporte a la toma de decisiones en la industria de procesos durante las últimas décadas. Como motor computacional para minería y análisis de datos, el aprendizaje automático sirve como herramientas básicas para la extracción de información, patrón de datos reconocimiento y predicciones [1].

En [Witten & Frank 2000] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [2].

La Explotación de Información consiste en la aplicación de herramientas de análisis y síntesis con el objetivo de extraer conocimiento no trivial que se encuentra distribuido en forma implícita en los datos disponibles de diferentes fuentes de información dentro de una organización [3]. Este conocimiento es previamente desconocido y puede resultar útil para la toma de decisiones dentro de una organización [4].

Las enormes cantidades de datos provenientes de las bases de datos de producción las cuales contienen grandes cantidades de filas y varios atributos requieren un procesamiento simultáneo lo cual hace prácticamente imposible el análisis manual. Estos factores son indicadores de la necesidad de metodologías automatizadas para el análisis de datos las cuales permitan extraer conocimiento útil.

El proceso de “descubrimiento de conocimiento en bases de datos” (en inglés, Knowledge Discovery in Databases, KDD) surge como una reconocida metodología para la automatización y el análisis de datos que permite lograr el objetivo de la inteligencia y el análisis de datos automatizado. La minería de datos es una etapa en el proceso de KDD, que involucra la aplicación de algoritmos específicos para extraer patrones de datos. Los pasos adicionales en el proceso de KDD, como la preparación de los datos, su

limpieza, selección, incorporación de conocimiento previo apropiado e interpretación adecuada de los resultados garantizan que se deriven conocimientos útiles de los datos [5].

En los entornos industriales modernos, se recolectan inmensas cantidades de datos en sistemas de administración de base de datos y data warehouses[1] involucrando diferentes áreas tales como; diseño de productos y procesos, montaje, planificación de materiales, control de calidad, mantenimiento, programación, detección de fallas, etc. La minería de datos se ha convertido en una herramienta importante para la adquisición de conocimiento de las bases de datos de estos entornos [6].

La soldadura fuerte (Brazing) es una tecnología para unir materiales de sustrato similares diferentes mediante el uso de temperaturas superiores a 450° centígrados para efectuar el cambio de fase en un relleno o revestimiento añadido, preferiblemente sin afectar notablemente a la integridad del sustrato [7].

Dada la necesidad de las industrias de contar con conocimiento respecto a sus procesos industriales con el propósito de incrementar la calidad de estos para evitar inconvenientes como, la oxidación de los metales generando en el mejor de los casos retrabajo con las piezas u ocasionando el descarte de éstas. Por estos inconvenientes mencionados anteriormente, se propone llevar a cabo un proyecto de minería de datos respecto al proceso de Brazing.

Antes de comenzar con el diseño del proyecto de minería de datos, se decide realizar un mapeo sistemático de la literatura (en inglés, Systematic Mapping Studies o SMS) con el propósito de analizar el estado del arte y descubrir las contribuciones de la aplicación de minería de datos en los procesos industriales en la industria metalúrgica.

[1] Almacén de datos

La metodología de investigación para esta tarea será el SMS (Systematic Mapping Studies) ya que, permite recopilar información sobre el tema abordado de manera ordenada y otorga una visión global identificando la cantidad y tipo de investigación y resultados disponibles sobre el mismo.

Una vez recopilada la información acerca de los casos de estudio de la minería de datos aplicada a procesos en la industria se seleccionarán las principales herramientas y metodologías utilizadas. Con esta información se presentará el estado del arte, para luego seleccionar las técnicas y herramientas más convenientes y aplicar estas al proceso concreto de Brazing (soldadura fuerte).

2.2) Mapeo Sistemático de la Literatura (SMS)

Es una metodología que permite proporcionar una visión global sobre un tema de interés, identificando la cantidad, tipo de investigación y resultados sobre el mismo. Para realizar el SMS se siguieron los lineamientos propuestos por Kitchenham et al. [8] y por Petersen et al. [9].

Consta del desarrollo de una serie de pasos enumerados a continuación:

1. Planificación del SMS.
2. Ejecución del SMS.
3. Resultados del SMS

2.2.1) Planificación del SMS

En esta sección se presenta la definición del protocolo de revisión del SMS: preguntas de investigación (PI), estrategia de búsqueda, selección de los estudios, criterios y proceso de selección, formulario de extracción y el proceso de síntesis de los datos.

El objetivo de este SMS es responder la siguiente pregunta de investigación (PI): *¿Qué contribuciones existen de la aplicación de minería de datos en los procesos industriales en la industria metalúrgica?* Esta pregunta principal se descompone en un conjunto de sub-preguntas (PI1-5), las cuales se presentan en la Tabla 2-1 junto con su motivación.

Referencia	Pregunta de interés	Motivación
PI1	¿En qué tipos de industrias se aplica la minería de datos?	Encontrar los procesos industriales en los cuales se aplica minería de datos en la actualidad.
PI2	¿Qué técnicas de minería de datos se utilizan en procesos industriales?	Encontrar los algoritmos más usados frecuentemente en los procesos industriales.
PI3	¿Con qué herramientas y lenguajes de programación se trabaja en la minería de datos aplicada a la industria?	Encontrar las herramientas y lenguajes de programación más utilizadas en la industria para la minería de datos.
PI 4	¿Qué modelos de aprendizaje de minería de datos son los más utilizados para análisis de procesos industriales?	Encontrar que tipos de modelos de aprendizaje para minería de datos (supervisados, no supervisados y semi supervisados) son los más utilizados en las industrias.
PI 5	¿Qué tipos de investigación se presentan en los artículos?	Determinar el tipo de investigación propuesta en los artículos de acuerdo con la clasificación propuesta por Wieringa [10].

Tabla 2-1 Preguntas de investigación.

Se decidió realizar una búsqueda automática en las librerías y plataformas digitales que se presentan en la Tabla 2-2 considerando artículos de congresos y artículos de revistas.

Librerías y plataformas digitales
Google Scholar
IEEE Xplore
Digibuo
Idus
Sciencedirect

Tabla 2-2 Listado de librerías y plataformas digitales.

La búsqueda se realizó en el período comprendido entre enero del año 2000 hasta mayo del año 2021. Para el armado de la cadena de búsqueda se consideraron términos principales y alternativos definidos en la Tabla 2-3.

Principal	**Alternativo**
Minería de datos	Data minning
Procesos industriales	Industrial Proccess
Metalúrgica	Metallurgical

Tabla 2-3 Términos de búsqueda.

Las cadenas de búsqueda resultante son:

(Minería de datos aplicada a procesos industriales) OR (Minería de datos aplicada a la industria) OR (Minería de datos en la industria metalúrgica) OR (Data mining applied to industrial processes) OR (Data mining applied to industry) OR (Data mining in the metallurgical industry)

Los criterios de inclusión y exclusión utilizados para el proceso de selección de artículos se presentan en la Tabla 2-4**.**

Criterios de inclusión
Artículos duplicados: si hay varios artículos de un mismo autor que contemple la misma investigación, se considerará el más completo y el más reciente.
Artículos en idioma inglés y español.
Artículos publicados entre enero del año 2000 y junio del año 2021.
Artículos que contengan cadenas candidatas en el título, palabras clave y/o en el resumen.
Criterios de exclusión
Artículos a los cuales no se tenga acceso.
Artículos que no estén orientados a procesos industriales.
Literatura gris [11], artículos disponibles solo en forma de resúmenes, presentaciones en PowerPoint, tesis doctorales o libros.

Tabla 2-4 Listado de criterios de inclusión y exclusión.

El proceso de selección de los estudios consistió en los siguientes pasos:

1. Realizar la búsqueda en las fuentes definidas aplicando la cadena en el título y/o en el resumen.
2. Eliminar los artículos duplicados.
3. Aplicar los criterios de inclusión y exclusión en el título, resumen y palabras clave
4. Aplicar los criterios de inclusión y exclusión al texto completo. Este proceso permitió la selección de los estudios primarios que se analizaron para dar respuesta a las preguntas de investigación (PI) formuladas.

Para dar respuesta a cada una de las preguntas de investigación (PI) se definió un esquema de clasificación que se presenta en la Tabla 2-5.

Dimensión	Categorías
PI1/Tipos de industrias.	Otras industrias, no aplica, industria energética, industria farmacéutica, industria automotriz, industria metalúrgica.
PI2/Técnicas utilizadas.	Modelización estadística paramétrica, modelización estadística no paramétrica, reglas de asociación y dependencia, métodos bayesianos, árboles de decisión y sistemas de reglas, métodos relacionales y estructurales, redes neuronales artificiales, máquinas de vector soporte,

	extracción de conocimientos con algoritmos evolutivos y reglas difusas, métodos basados en casos y en vecindad, no Aplica,
PI3/Modelo de aprendizaje del algoritmo.	Supervisado, no supervisado, semi supervisado.
PI4/Herramientas/lenguajes de programación.	RapidMiner, WEKA, Orange, KNIME, SAS, IBM, SPSS, Clementine, R, Python, Spark, ninguno, otros.
PI5/Tipos de investigación.	Evaluación, propuesta de solución, experiencia personal, opinión, validación, filosófico [10].

Tabla 2-5 Esquema de clasificación.

En la Tabla 2-6 se presenta el formulario de extracción de datos de los estudios primarios. Se compone de dos partes; la primera se refiere a los metadatos de cada uno de los estudios primarios y la segunda se refiere a cada una de las preguntas de investigación (PI).

Metadatos	Año, título, autor(es), fuente de búsqueda, país, palabras clave, citas.
PI/Dimensión	**Categorías**
PI1/Tipos de industria.	Otras industrias, no aplica, industria energética, industria farmacéutica, industria automotriz, industria metalúrgica.
PI2/Técnicas utilizadas.	Modelización estadística paramétrica, modelización estadística no paramétrica, reglas de asociación y dependencia, métodos bayesianos, árboles de decisión y sistemas de reglas, métodos relacionales y estructurales, redes neuronales artificiales, máquinas de vector soporte, extracción de conocimientos con algoritmos evolutivos y reglas difusas, métodos basados en casos y en vecindad, no Aplica,
PI3/Modelo de aprendizaje del algoritmo.	Supervisado, no supervisado, semi supervisado.
PI4/Herramientas/lenguajes de programación.	RapidMiner, WEKA, Orange, KNIME, SAS, IBM, SPSS, Clementine, R, Python, Spark, ninguno, otros.
PI5/Tipos de investigación.	Evaluación, propuesta de solución, experiencia personal, opinión, validación, filosófico [10].

Tabla 2-6 Formulario de extracción de datos.

Se utiliza una síntesis temática basada en el esquema de clasificación que se representará a través de tablas y gráficos.

Se realizará una síntesis cuantitativa de cada estudio considerando la aplicación del mismo a cada dimensión del formulario, para luego, de esta manera, poder dar respuesta a cada pregunta de investigación planteada.

2.2.2) Ejecución del SMS

En esta sección, se presenta la búsqueda realizada en las librerías y plataformas digitales, la selección de estudios primarios de acuerdo con lo definido en el protocolo de revisión del SMS.

Se aplicaron las cadenas de búsqueda en las librerías y repositorios con algunas adecuaciones necesarias en función de las particularidades de cada una. En la Tabla 2-7 se presenta la cantidad de artículos encontrados en cada una de las librerías y repositorios.

Librería o repositorio digital	**Artículos Encontrados**
https://scholar.google.com/	53
https://ieeexplore.ieee.org	7
https://digibuo.uniovi.es	3
https://dialnet.unirioja.es	9
https://idus.us.es	3
https://www.sciencedirect.com	5
Total de artículos encontrados:	80

Tabla 2-7 Listado de artículos encontrados.

De un total de 80 artículos encontrados, se analizaron 14 estudios primarios. El listado de los estudios analizados se presenta en la Tabla 2-8.

Id	**Estudio primario**
[EP1]	Aldana Fransheska Dongo Pozo, Xiomara Pamela Silva Cama, Análisis de la minería de datos aplicada en empresas del sector retail, Perú, 2020.
[EP2]	Data Mining and Analytics in the Process Industry: The Role of Machine Learning, Zhiqiang Ge, Zhihuan Song, Steven X. Ding, Biao Huang, China, 2017.
[EP3]	Dra. Claudia Barreto Cabrera, Minería de datos aplicada a la mejora de procesos de extrusión de elastómeros, España, 2009.

[EP4]	D. Daniel González Ordóñez, Modelado de Procesos Industriales Complejos a través de Minería de Datos. Aplicación a un Tren de Laminación en Frío, España, 2012.
[EP5]	Ana González Marcos, Desarrollo de técnicas de minería de datos en procesos industriales modelización en líneas de producción de acero, España, 2006.
[EP6]	Ortiz Silva, José Luis, Optimización energética mediante técnicas de minería de datos en sistemas de refrigeración industrial. Aplicación a la industria agroalimentaria, España, 2017.
[EP7]	Francisco Javier Martínez de Pisón Ascacíbar, Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado, España, 2003.
[EP8]	Thembinkosi Nkonyana, Yanxia Sun, Bhekisipho Twala, Eustace Dogo, Performance Evaluation of Data Mining Techniques in Steel Manufacturing Industry, Sudáfrica, 2019.
[EP9]	Seyyed Soroush Rohanizadeh, Mohammad Bameni Moghadam, A Proposed Data Mining Methodology and its Application to Industrial Procedures, Iran, 2009.
[EP10]	Hamza Saad, The Application of Data Mining in the Production Processes, New York, 2018.
[EP11]	A. K. Choudhary · J. A. Harding · M. K. Tiwari, Data mining in manufacturing: a review based on the kind of knowledge, UK, 2008.
[EP12]	Jinlin Zhua, Zhiqiang Gea, Zhihuan Songa, Furong Gao, Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, China, 2018.
[EP13]	Michał Rogalewicz, Robert Sika, Methodologies of knowledge discovery from data and data mining methods in mechanical engineering, Polonia 2016.
[EP14]	Christoph Gröger, Florian Niedermann, and Bernhard Mitschang, Data Mining-driven Manufacturing Process Optimization, UK, 2012.

Tabla 2-8 Listado de estudios primarios.

2.2.3) Resultados del SMS

En esta sección se presentan los resultados del SMS. En la Tabla 2-9 se presenta una síntesis cualitativa de los resultados del análisis de los estudios primarios en base a lo establecido en el esquema de clasificación definido en el protocolo de revisión para dar respuesta a cada una de las preguntas de investigación (Ver Tabla 2-1).

Id	Resultados por cada PI (Pregunta de investigación)				
	Dimensión-Tipos de Industria/ (PI1)	Dimensión- Técnicas utilizadas/(PI2)	Dimensión-Modelo de aprendizaje del algoritmo/(PI3)	Dimensión-Herramienta y	Dimensión-Tipo de investigación/(PI5)

				lenguajes de programación/(PI4)	
[EP1]	No Aplica.	Modelización estadística paramétrica, reglas de asociación y dependencia, árboles de decisión y sistemas de reglas, métodos relacionales y estructurales, redes neuronales artificiales.	Supervisado, semi supervisado.	Otros.	Experiencia personal.
[EP2]	Otras industrias, industria automotriz.	Modelización estadística paramétrica, métodos bayesianos, árboles de decisión y sistemas de reglas, redes neuronales artificiales, máquinas de vector soporte, métodos basados en casos y en vecindad.	Supervisado, no supervisado, semi supervisado.	Ninguno.	Opinión.
[EP3]	Industria metalúrgica.	Modelización estadística paramétrica, máquinas de vector soporte.	Supervisado.	R.	Propuesta de solución.
[EP4]	Industria metalúrgica.	Modelización estadística paramétrica.	Supervisado.	Otros.	Propuesta de solución.
[EP5]	Industria metalúrgica.	Redes neuronales artificiales.	Supervisado.	Ninguno.	Validación.
[EP6]	Industria energética.	Redes neuronales artificiales.	Supervisado.	IBM SPSS Clementine.	Propuesta de solución.
[EP7]	Industria metalúrgica.	Modelización estadística paramétrica, redes neuronales artificiales.	Supervisado.	WEKA, IBM SPSS Clementine, R, Otros.	Propuesta de solución.
[EP8]	Industria metalúrgica.	Árboles de decisión y sistemas de reglas, redes neuronales artificiales, máquinas de vector soporte.	Supervisado, no supervisado, semi supervisado.	Python.	Evaluación.

[EP9]	Otras industrias, industria farmacéutica, industria automotriz, industria metalúrgica.	Árboles de decisión y sistemas de reglas, redes neuronales artificiales.	Supervisado.	Ninguno.	Propuesta de solución.
[EP10]	Otras industrias, industria energética, industria automotriz.	Métodos bayesianos, árboles de decisión y sistemas de reglas, redes neuronales artificiales, máquinas de vector soporte, métodos basados en casos y en vecindad.	Supervisado.	Orange.	Opinión.
[EP11]	Otras industrias, industria automotriz, industria metalúrgica.	Métodos bayesianos, árboles de decisión y sistemas de reglas, redes neuronales artificiales, máquinas de vector soporte, extracción de conocimientos con algoritmos evolutivos y reglas difusas.	Supervisado, no supervisado.	Ninguno.	Opinión.
[EP12]	Otras industrias, industria farmacéutica, industria automotriz, industria metalúrgica.	Modelización estadística no paramétrica, métodos bayesianos, métodos relacionales y estructurales, métodos basados en casos y en vecindad.	No supervisado.	Ninguno.	Opinión.
[EP13]	Otras industrias, industria farmacéutica, industria energética.	Modelización estadística paramétrica, modelización estadística no paramétrica, reglas de asociación y dependencia, métodos bayesianos, árboles de decisión y sistemas de reglas, métodos relacionales y estructurales, redes neuronales artificiales, máquinas de vector soporte. extracción de conocimientos con algoritmos evolutivos y reglas difusas, métodos	Supervisado, no supervisado.	Ninguno.	Opinión.

		basados en casos y en vecindad.			
[EP14]	Otras industrias, industria metalúrgica.	Métodos bayesianos, árboles de decisión y sistemas de reglas, redes neuronales artificiales, máquinas de vector soporte.	Supervisado, no supervisado.	WEKA.	Propuesta de solución.

Tabla 2-9 Listado de estudios primarios.

A continuación, se pretende dar respuesta a las preguntas de investigación en base a la literatura analizada.

PI1: ¿En qué tipos de industria se aplica la minería de datos?

Según los datos analizados son diversas las industrias donde se hace uso de la minería de datos. Se encuentran industrias específicas relacionados con la energía, metalúrgica, automotriz, control de calidad de productos, producción de acero, entre otras. En general, según los estudios analizados, industrias de diversa índole utilizan la minería de datos y lo hacen para mejorar sus procesos en cuanto a la optimización, relacionada con una reducción de costos, y en cuanto a la mejora de calidad, relacionada a prevenir fallos en los productos.

En cuanto a las industrias en general, la minería de datos aplicada a diferentes áreas en ingeniería de producción y fabricación para extraer conocimiento para aplicar en la predicción de mantenimiento, diseño, fallas detección, control de calidad, producción y apoyo a la toma de decisiones sistemas (Ver Figura 2-1).

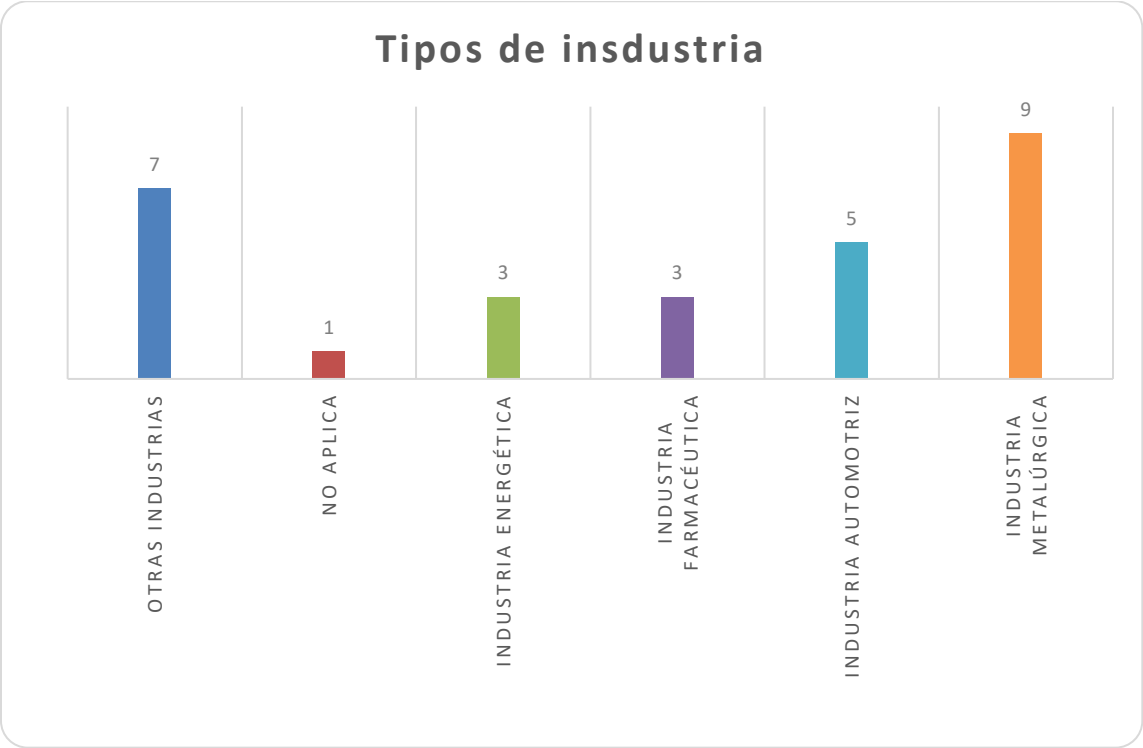


Figura 2-1 Cantidad de artículos primarios según tipo de industria.

La mayoría de los artículos hacen referencia a procesos industriales diversos como la industria alimenticia, la industria energética, automotriz.

PI2: ¿Qué técnicas de minería de datos se utilizan en procesos industriales?

La Figura 2-2 presenta que los algoritmos más utilizados en minería de datos para procesos industriales son las máquinas de vector soporte (SVM), los árboles de decisión y las redes neuronales artificiales.

En los diversos estudios analizados se determina respecto a la precisión para la clasificación de elementos que la técnica de árboles de decisión con sus algoritmos (Random Forests y AdaBoost) es una de las mejores.

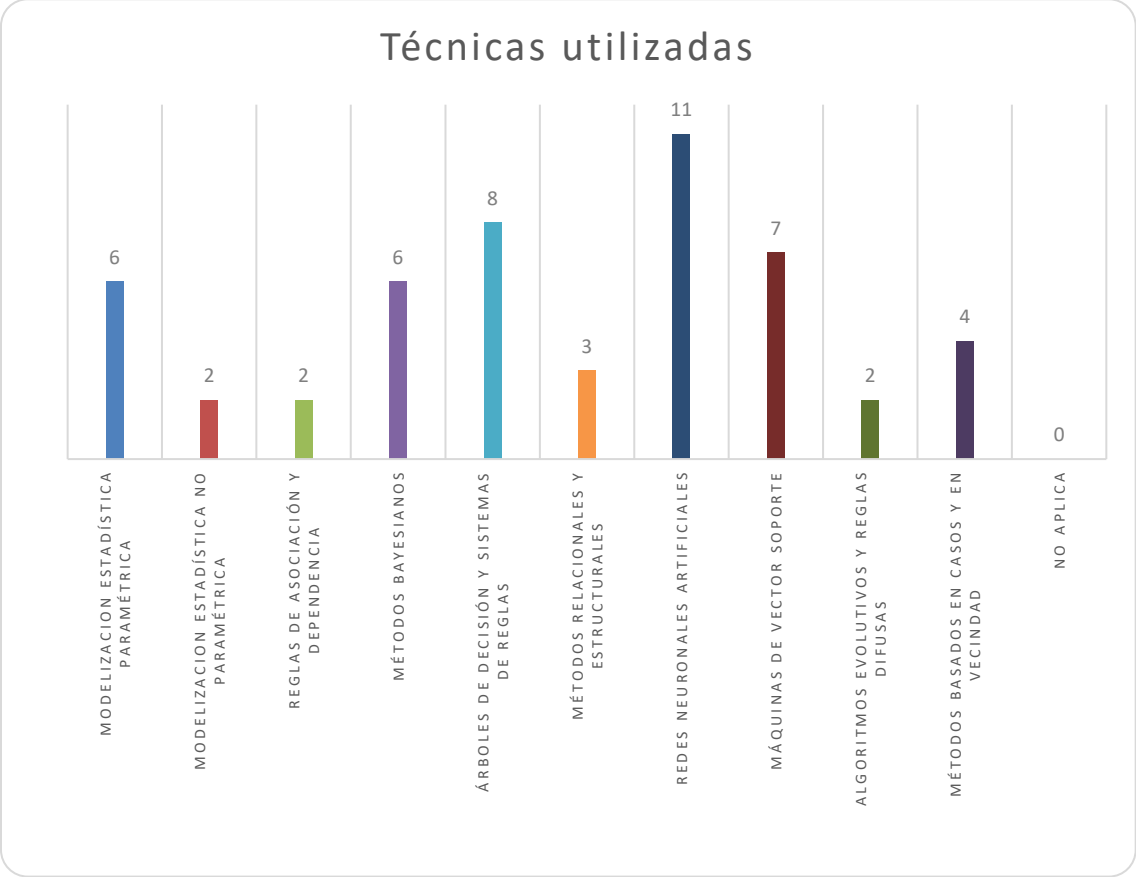


Figura 2-2 Técnicas de minería de datos.

PI3: ¿Qué modelos de aprendizaje de minería de datos son los más utilizados para análisis de procesos industriales?

La Figura 2-3 sintetiza que el modelo de aprendizaje más utilizado en los diversos algoritmos de aprendizaje automático que se emplean en las industrias es el **supervisado**.

El tipo de modelo que prevalece para las aplicaciones industriales es el **supervisado**. Los procesos industriales suelen contrastarse con datos previos de observaciones empíricas registradas y es por esto por lo que este tipo de aprendizaje es el que mejor se ajusta para el entrenamiento de los modelos.

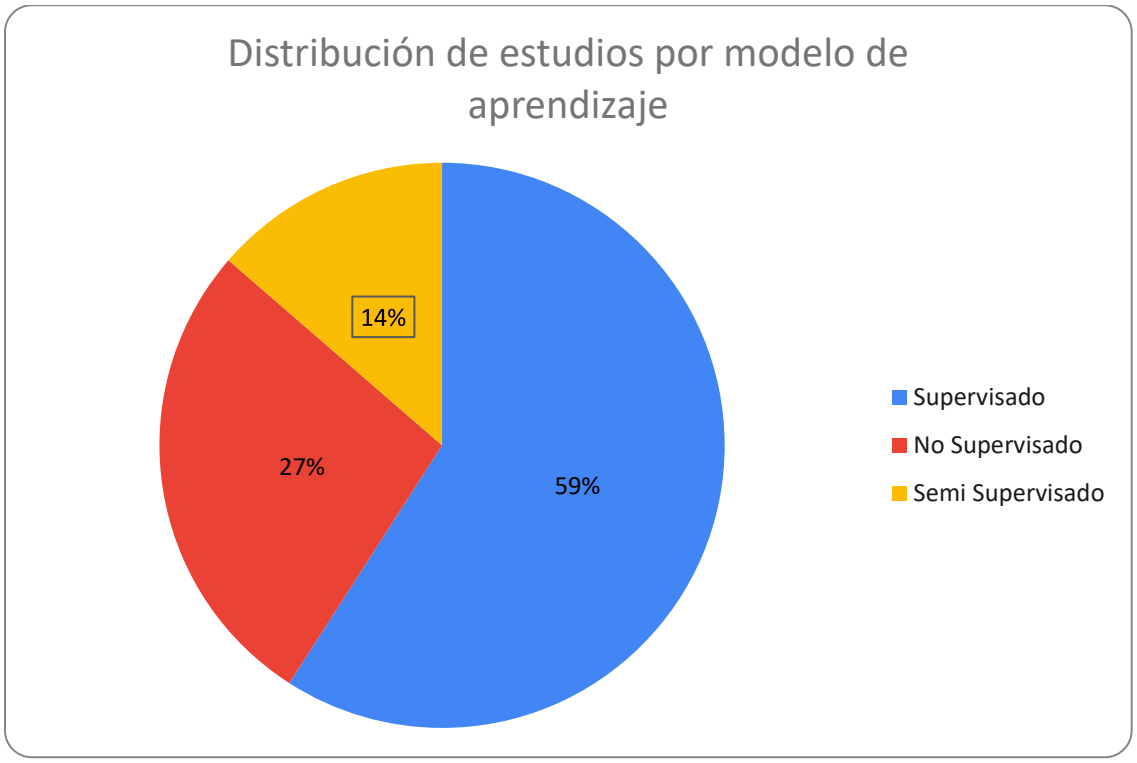


Figura 2-3 Distribución de estudios por modelos de aprendizaje.

La mayoría de industrias tienen información histórica sobre sus procesos que les permiten entrenar a los algoritmos en base a resultados anteriores y de esta manera ir generando una precisión cada vez mayor para predecir eventos futuros o clasificar de manera más óptima. Es por esto que los modelos de aprendizaje supervisado son los que mejor se adaptan para estos casos.

Los algoritmos basados en modelos no supervisados no disponen de información previa, para su entrenamiento, sino que intentan buscar una relación únicamente con los datos de entrada, es por esto que se dice que los algoritmos tienen un carácter exploratorio.

PI4: ¿Con qué herramientas y lenguajes de programación de minería de datos se trabaja actualmente en la industria?

En el siguiente grafico se presentan las herramientas más utilizadas para el proceso de minería de datos.

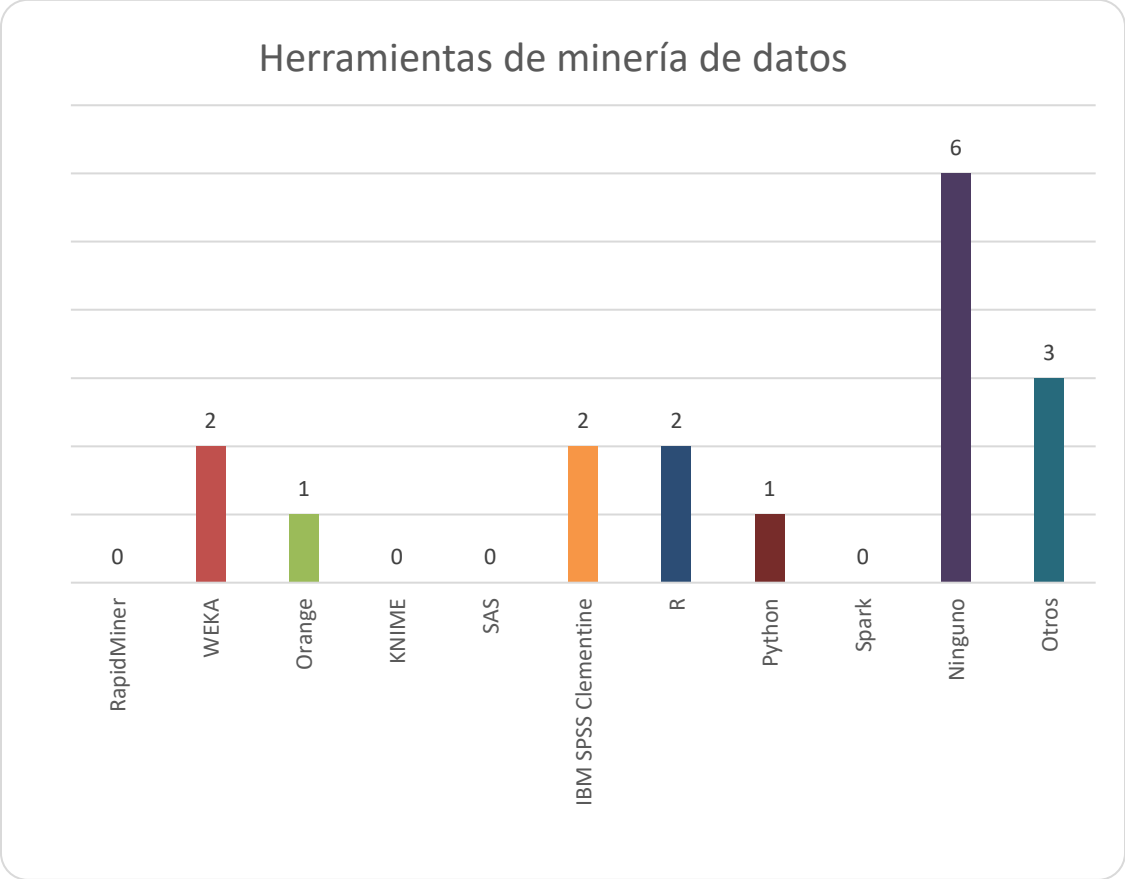


Figura 2-4 Herramientas de minería de datos.

La Figura 2-4 muestra las herramientas mencionadas en los diversos estudios primarios analizados. Al hacer referencia a herramientas también se están incluyendo lenguajes de programación que en algunos casos como el de Python son de propósito general, pero tienen un fuerte desarrollo orientado al análisis de datos y, en otros casos, lenguajes como R que están diseñados específicamente para analizar datos.

Se puede inferir a partir de la Figura 2-4 que existe un uso variado de herramientas y, en última instancia este aspecto dependerá mucho del usuario y la calidad que este perciba respecto a variables como la usabilidad, potencia y versatilidad que pueda ofrecer dicha herramienta. No es un tema menor el aspecto

de las licencias ya que herramientas como R y Python son de código abierto y no requieren licencias para su uso y tienen comunidades muy activas de desarrolladores y usuarios que trabajan permanentemente en su mantenimiento.

Si bien muchos artículos son de análisis teórico y no hacen mención a herramientas particulares resulta interesante los datos proporcionados por la herramienta Google trends [12] donde se realiza una comparativa entre las principales herramientas analizadas en la Figura 5 adicionando una popular librería llamada Sklearn del lenguaje Python.

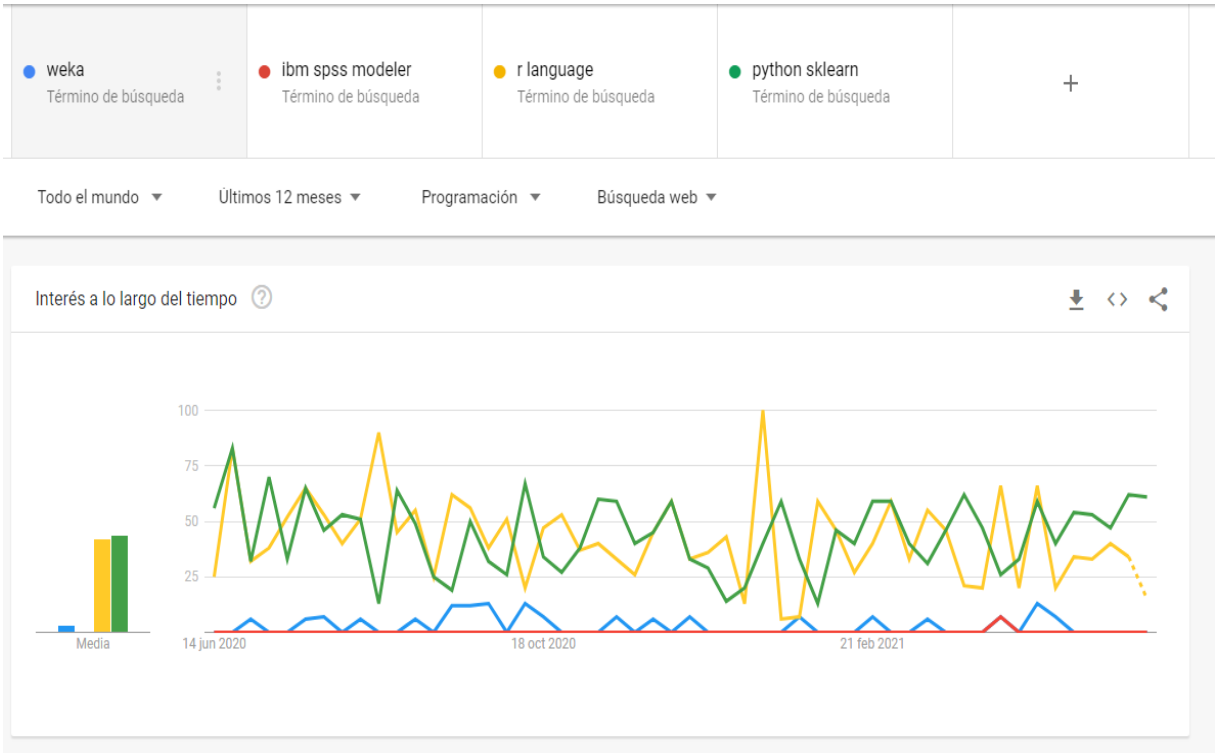


Figura 2-5 Tendencias en herramientas para minería de datos.

PI5: ¿Qué tipos de investigación se presentan en los artículos?

Se encontró que, del total de los estudios primarios, 6 estudios (43%) tienen como propósito de investigación realizar una propuesta de solución. Existe un solo artículo (7%) correspondiente a la clasificación, evaluación de la investigación. Con respecto a la clasificación, experiencia personal y a la

validación se encontró un artículo (7%) para cada categoría respectivamente. En cuanto a los artículos de opinión se encontraron cinco (36%) en total.

2.3) Amenazas a la validez

Se analizaron las potenciales amenazas a la validez que podrían afectar al SMS, respecto a las cuatro categorías sugeridas por Wohlin et al.

2.3.1) Validez del constructo

En este SMS, con el fin de mitigar estas amenazas, describimos el significado que le hemos dado a minería de datos y proceso Brazing basados en literatura reconocida [1], [2], [3], [4], [7].

2.3.2) Validez interna

Para mitigar las preocupaciones sobre la validez interna, el primer autor creó un protocolo de revisión como parte de la investigación de un trabajo de tesis de grado de la carrera Licenciatura en Sistemas de la Universidad de Morón y éste fue revisado por los otros dos autores (docentes de la asignatura).

2.3.3) Validez externa

Se tomó la decisión de utilizar cinco motores de búsqueda a los cuales se tuvo acceso. No se consideró la literatura gris, como los artículos disponibles solo en forma de resúmenes, presentaciones en PowerPoint, tesis doctorales o libros, porque incluirlos podría haber afectado la validez de nuestros resultados.

2.3.4) Fiabilidad

Se intentó mitigar el sesgo de las publicaciones definiendo cuidadosamente:

1. Los criterios de inclusión y exclusión para poder seleccionar estudios primarios.
2. Los criterios de exclusión específicamente, con el fin de seleccionar reglas basadas en las preguntas de investigación predefinidas en el trabajo.

Para aumentar la confiabilidad, los tres autores de manera paralela aplicaron los criterios, realizaron la catalogación de los estudios; se discutieron las discrepancias entre ellos, con el propósito de determinar si era apropiado incluir un artículo en particular o no, y de ese modo se obtuvo el listado final de estudios primarios. Además, se diseñó un formulario para la registración de los datos con Excel y se mapearon las preguntas de investigación de acuerdo con el esquema de clasificación definido para cumplir con los objetivos de este estudio. Se considera que el efecto potencial de este sesgo tiene menos importancia en estudios de mapeos sistemáticos que en las revisiones sistemáticas de literatura.

Para fortalecer la confiabilidad, luego de aplicar los criterios de inclusión y exclusión, se creó una matriz con las propiedades de los datos extraídos de los artículos y se los catalogó con las preguntas de investigación con el motivo de cumplir con el objetivo de este estudio.

2.4) Conclusiones y trabajos futuros

En esta tesis se presentó un mapeo sistemático de la literatura para analizar el estado del arte respecto a la aplicación de la minería de datos en procesos industriales. Se seleccionaron 14 estudios primarios de un conjunto inicial de 80 artículos resultantes de las búsquedas realizadas en Google Scholar, Dialnet, Sciencedirect, IEEE Xplore, Digibuo, en el período comprendido entre enero del año 2000 y junio del año 2021. Una vez analizados los estudios primarios, se concluye que:

- El 60% de los estudios utiliza modelos de entrenamiento supervisados.
- Los algoritmos más utilizados en procesos industriales son las máquinas de vector soporte (SVM), los arboles de decisión y las redes neuronales artificiales.
- Las herramientas/lenguajes de programación más utilizados para el machine learning son Python y R.
- En la mayoría de los estudios analizados se presentan propuestas de solución como tipo de investigación, 43% en total.

- Intrínsecamente, los métodos de minería de datos basados en modelos estadísticos multivariados y teorías de inferencia actúan como ejes para la comprensión y el seguimiento en una amplia gama de procesos y sistemas industriales complejos, como las plantas químicas, fabricación de semiconductores, sistemas mecánicos, producción de alimentos y procesos farmacéuticos.
- El Brazing, es un proceso industrial complejo y que posee diversas variables físicas cuantificables como temperatura, caudal de nitrógeno, caudal de aire y tiempo. Este proceso es susceptible de aprovechar los beneficios que aporta la minería de datos para el análisis de estas variables contribuyendo de esta manera a la mejora en la calidad del producto terminado.

3) Capítulo 3: Descripción del Proceso Industrial

3.1) Introducción

En este capítulo se realiza una descripción completa del proceso de soldadura fuerte el cual de ahora en adelante lo llamaremos “Brazing” como actualmente se lo conoce en la industria.

La soldadura fuerte (Brazing) es una tecnología para unir materiales de sustrato similares o diferentes mediante el uso de temperaturas superiores a 450 °C para efectuar el cambio de fase en un relleno o revestimiento añadido, preferiblemente sin afectar notablemente a la integridad del sustrato [7].

En este caso puntual el Brazing es sobre paneles de aluminio para refrigeración de fluidos como el agua, aceite y aire. En la industria estos paneles, en conjunto con tanques y ventiladores es lo que se conoce como radiadores y son los principales elementos de la actualidad utilizados para la refrigeración de fluidos en máquinas.

3.2) Tipos Constructivos de Paneles

3.2.1) Paneles Placa Barra

Son aquellos construidos con los siguientes componentes:

- Perfiles
- Chapas
- Aletas
- Turbulador (aleta de aceite)

A continuación en la Figura 3-1 se pueden ver los componentes de un panel placa barra en una vista de sección.

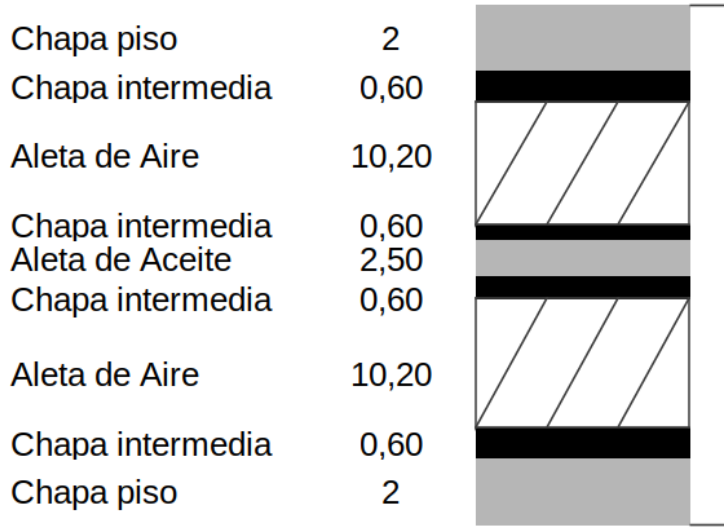


Figura 3-1 Sección de panel placa barra.

Su uso está indicado en general para presiones elevadas y para entornos hostiles donde se deben tolerar vibraciones que fatigan a los materiales. Por lo general son utilizados para la refrigeración de aceite.

3.2.2) Paneles de Tubo

A continuación, en la Figura 3-2 se pueden ver los componentes que conforman un panel de tubos.

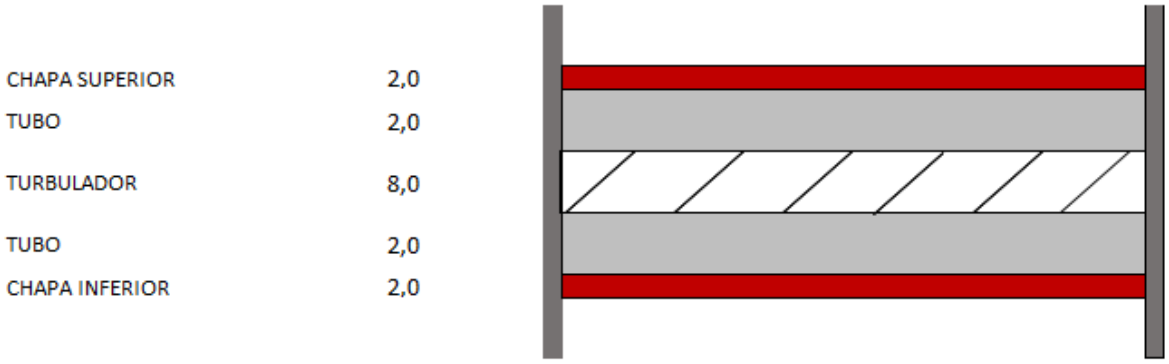


Figura 3-2 Panel de tubos y sus componentes.

Su uso está indicado para aplicaciones que no requieren una presión de trabajo elevada (menores a 10 bar) y son utilizados por lo general para refrigerar agua.

3.2.3) Paneles de Intercooler

A continuación, vemos en la Figura 3-3 los componentes correspondientes a un panel intercooler.

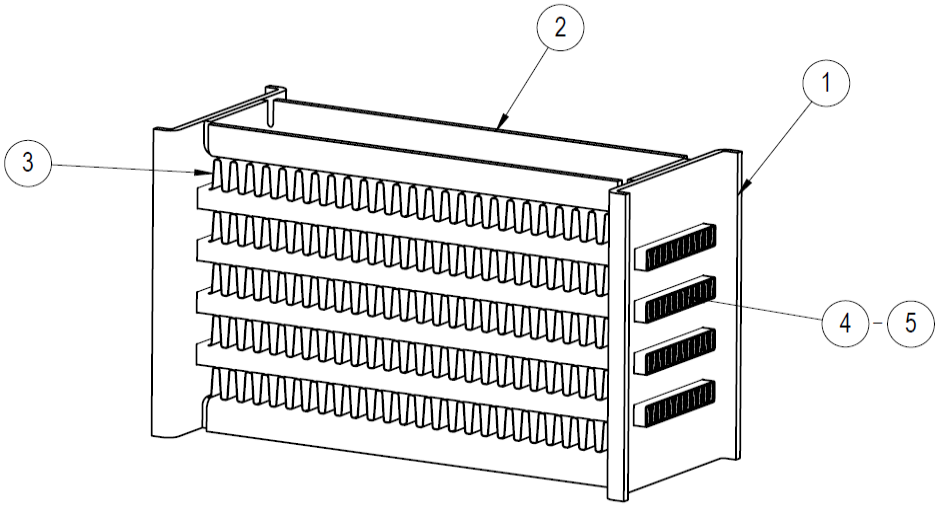


Figura 3-3 Panel intercooler y sus componentes.

1. Colectores
2. Chapas
3. Aletas
4. Chapas filtro

Su uso está indicado para la refrigeración de aire y al igual que los radiadores de tubo se indica no utilizarlo para aplicaciones donde circulen presiones mayores a 10 bar.

3.3) Tecnologías para procesos de soldadura

La industria automotriz siempre estuvo a la vanguardia y fue la pionera en el desarrollo de radiadores para refrigeración de fluidos. Hasta el año 1970 los desarrollos se realizaban con un proceso de Brazing en cobre. Entre los años 1970 y 1990 comienza a utilizarse el aluminio como principal material ya que su peso es aproximadamente 1/3 más bajo respecto al cobre, y tiene otras ventajas que mencionaremos en la siguiente sección. Respecto a los procesos para conseguir soldar los componentes, existen principalmente dos:

1. Brazing en batch: requiere una atmosfera controlada de gas inerte como el nitrógeno y la carga a soldar debe permanecer un tiempo determinado hasta llegar a temperatura y producir la fusión del metal.
2. Brazing en continuo: las condiciones son similares al proceso de batch pero la carga a soldar circula continuamente sin detenerse.
3. Horno de vacío: tienen gran calidad en la soldadura ya que no hay gases en la atmosfera que produzcan oxidación. El calentamiento es por inducción y de esta manera se logra una temperatura mucho más uniforme en todas las partes de la pieza respecto a la tecnología de Brazing.

Para este trabajo realizamos el análisis sobre un horno de Brazing en batch.

3.4) Características del Proceso de Brazing y Materiales

Los paneles que componen a los radiadores pueden ser de diversos metales según las características que tenga la aplicación donde estarán funcionando, por ejemplo, en industrias donde existe un gran contacto con agentes corrosivos como la sal se utilizan materiales como el comúnmente denominado acero inoxidable, el cual es una aleación de acero con cromo el cual brinda una mayor resistencia a los procesos de oxidación. También se utiliza el latón, que es una aleación de cobre y zinc, pero que en la actualidad tiene cada vez menos aplicación por su bajo coeficiente de disipación y un mayor peso específico respecto al aluminio lo cual lo hace más costoso y menos eficiente. Por último, tenemos el aluminio, el cual es el material sobre el que trabajaremos en este estudio. El aluminio reúne ciertas características que lo hacen en la actualidad el material más conveniente para la fabricación de radiadores, podemos nombrar algunas:

1. Buen coeficiente de disipación térmica respecto de otros metales (80 a 230 W/(m·K)).
2. Menor peso específico respecto al acero y el cobre (2700 kg/m^3).
3. Abunda en la corteza terrestre y puede ser reciclado sin perder propiedades

4. Gran resistencia al peso

5. Al no contener hierro no es magnético

Los paneles de aluminio de acuerdo a su aplicación están compuestos por distintos elementos según el fluido que se requiera refrigerar, las presiones de trabajo y la aplicación del radiador. Para este estudio analizaremos el tipo constructivo de radiador Paneles Placa Barra que es el más común en la industria.

Los perfiles idealmente son de aluminio puro, las chapas tienen en su centro una capa de aluminio puro y a los lados están revestidas por capas más delgadas de una aleación de aluminio y silicio, las aletas son también de aluminio puro al igual que el turbulador.

El proceso de Brazing consiste en juntar todos estos materiales armando un paquete donde todas las piezas estén en contacto mediante técnicas como el prensado. Luego de esto se realiza un proceso de aplicación de un fundente, esto permite que durante el proceso de soldadura el metal no se oxide, ya que las altas temperaturas aceleran el proceso de oxidación en los metales. La aplicación del fundente en este caso es por inmersión en una primera etapa, y luego, antes de entrar al horno se realiza una segunda aplicación por rociamiento.

3.5) Temperaturas de Brazing

Según la Tabla 3-1 podemos ver un detalle de las temperaturas de Brazing según los tipos de panel que se introducen al horno.

Tipo de panel	Precámara (ºC)	Cámara (ºC)	Precámara (min)	Cámara (min)
Placa barra	375	595 - 615	12	12
Panel de tubos	375	585	12	12
Panel intercooler	375	585	12	12

Tabla 3-1 Temperaturas y tiempos de horneado.

El tiempo que cada panel permanece en las cámaras está definido en la programación del tablero PLC. Por ejemplo en los paneles placa barra se programa un tiempo de 12 minutos, este tiempo no corresponde al tiempo total que el panel se encuentra en la cámara de horneado sino que representa el tiempo que el panel está a temperatura de Brazing (para este caso 595 a 615° C). Es posible entonces definir al tiempo en la cámara de horneado como un tiempo compuesto por una parte fija programada y una parte variable que depende de la velocidad con la que los quemadores lleguen a la temperatura de Brazing, esto dependerá de la temperatura ambiente y del peso total de la carga y del tipo de paneles introducidos.

$$Tiempo\ en\ cámara = Tiempo\ en\ alcanzar\ t°\ de\ Brazing + Tiempo\ programado$$

3.6) Proceso de Horneado

Una vez que los componentes están armados y tienen el fundente aplicado comienza la etapa de horneado. El horno está compuesto por tres módulos:

1. Módulo de precalentamiento: es el primer módulo donde ingresa la carga, aquí se lleva la carga a una temperatura de 375 grados centígrados. Esto permite que luego en la cámara posterior la carga llegue más rápido a la temperatura objetivo.
2. Módulo de horneado: es donde la carga llega a la temperatura objetivo la cual varía según el tipo de panel que se está horneando según la Tabla 3-1. Una vez alcanzada la temperatura y transcurrido el tiempo fijado es donde se produce el proceso de Brazing lo cual une todas las piezas formando un solo elemento. Tanto la precámara como la cámara principal del horno están ambas inundadas con gas nitrógeno para desplazar todo el oxígeno que pudiera haber y así evitar la oxidación, esto se lleva a cabo gracias a un recirculador que cumple la función de hacer recircular el nitrógeno y homogenizando la atmosfera para conseguir una temperatura

estable en todos los puntos del horno. La temperatura se alcanza a través de los cuatro quemadores que proporcionan la energía necesaria.

1. Módulo de enfriamiento: la carga llega a este módulo y se detiene unos minutos, para bajar su temperatura, esta cámara tiene una recirculación interna de un líquido refrigerante compuesto por agua desmineralizada y un aditivo orgánico con propiedades anticorrosivas, un sistema de refrigeración con radiadores y un recirculador que permiten disipar velozmente las altas temperaturas. El tiempo en esta cámara está directamente relacionado con el tipo de panel y el tiempo determinado en la cámara principal ya que el horno es continuo y la cadena que transporta los paneles es solidaria a todas las cámaras del horno A continuación se presenta un plano esquemático para identificar las distintas partes del horno.

3.7) Problemas del proceso de Brazing en batch

Si bien la tecnología ha avanzado mucho en cuanto a la mejora de procesos para la construcción de radiadores, aún existen deficiencias en técnicas como el Brazing en batch las cuales enumeramos a continuación:

1. Oxidación por atmosfera inestable.
2. Falta de homogeneidad de la temperatura en la cámara de horneado.
3. Variabilidad de tiempos de horneado en relación a los pesos de las cargas introducidas.
4. Soldadura con poca penetración e inestable.

3.8) Detalle de los Módulos del Horno

3.8.1) Módulo de Precalentamiento

Es la primera etapa del proceso de Brazing, aquí las piezas ingresan ya prensadas y con la inmersión del fundente. Este módulo cuenta con 2 unidades de tubos radiantes donde cada una tiene un quemador que suministra 100.000 Kcal por hora y permiten que las cargas lleguen a una temperatura de 375 grados centígrados antes de transcurridos 10 minutos de su ingreso. El programa de horneado es comandado por un tablero PLC, cuando el programa comieza el PLC da la orden a dos actuadores neumáticos que levantan la compuerta de acceso a la precámara y otra orden a un motor que tracciona la cadena para que avance la carga hasta la posición adecuada (dentro de la cámara de precalentamiento), también apaga el recirculador para evitar que se fugue la atmosfera de nitrogneno mientras la compuerta está abierta, para esto también existen unas cortinas de goma que ayudan a evitar la fuga de la atmosfera. Luego, el PLC ordena detener la cadena, pone a los actuadores a cerrar la compuerta y prende nuevamente el recirculador. El precalentamiento de la carga es necesario ya que, la carga en el módulo de horneado debe estar a temperatura de Brazing según la Tabla 3-1 y para alcanzar esta temperatura en el tiempo estipulado es necesario un precalentamiento estable a 375°C, un precalentamiento deficiente puede provocar que la carga deba permanecer más tiempo en el módulo de horneado para alcanzar la temperatura de Brazing, esto puede traer problemas como la oxidación o que los materiales directamente se fundan.

En la siguiente figura se visualiza una vista en corte del módulo:

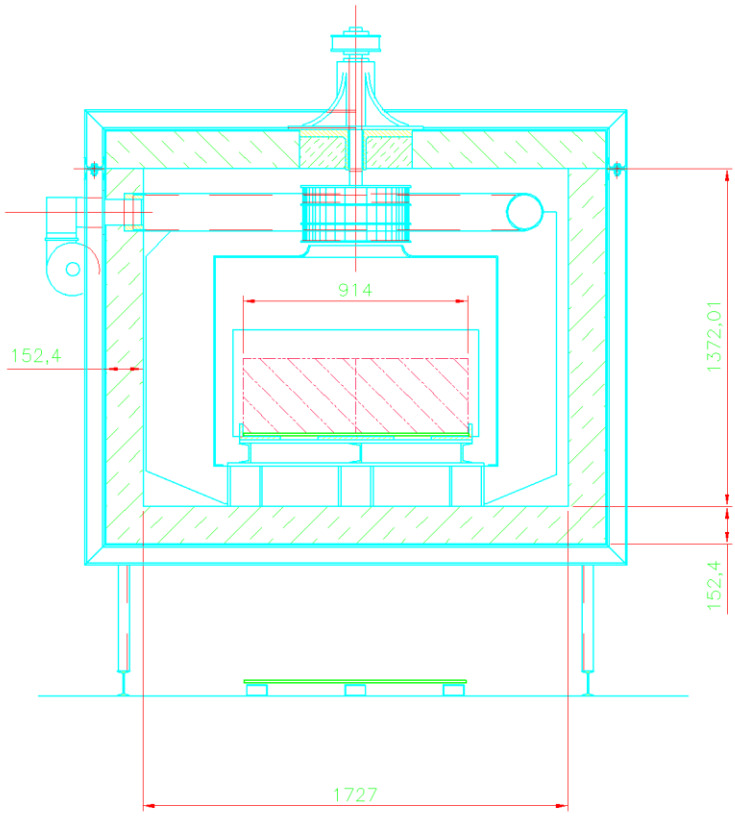


Figura 3-4 Módulo de precalentamiento.

3.8.2) Módulo de Horneado

Una vez finalizado el tiempo en la precámara el PLC ordena apagar el recirculador, levantar la compuerta que une la cámara de precalentamiento con la del horno y el avance de la cadena hasta que la carga se sitúa en la posición correspondiente a la cámara de horneado. Una vez situada la carga el PLC ordena frenar el avance de la cadena, el cierre de la compuerta y el encendido del recirculador. Es a partir de este momento donde por orden del PLC los cuatro quemadores de esta cámara comienzan a elevar su temperatura hasta alcanzar la temperatura correspondiente según el tipo de panel ver Tabla 3-1. El control de la temperatura lo realiza un sensor llamado termocupla que consiste en un termopar el cual funciona midiendo la diferencia de potencial que se genera al aplicar temperatura sobre los alambres (por lo general uno de hierro y otro de aleación cobre y níquel). Este pequeño voltaje se puede medir y mediante un simple

cálculo obtener su equivalencia en grados centígrados, este trabajo lo realiza el PLC y en función de la temperatura medida y la programada va ordenando a los quemadores otorgar más o menos potencia, este trabajo lo realiza accionando una serie de electroválvulas que permiten regular el pasaje de gas a los quemadores.

Al estabilizarse la temperatura real con la programada (con una tolerancia de ± 2 grados centígrados) es cuando comienza a ocurrir el proceso de Brazing, la capa de silicio de las chapas de aluminio comienza a fundirse y a soldarse con todas las piezas que están en contacto con esta. Una vez transcurrido el tiempo programado en el PLC, éste detiene el recirculador, abre la compuerta que une la cámara de horneado con la de enfriamiento y ordena el avance de la cadena hasta que la carga llega a la posición indicada en la cámara de enfriamiento, ahí el PLC detiene el avance de la cadena.

En la siguiente figura se visualiza una vista en corte del módulo:

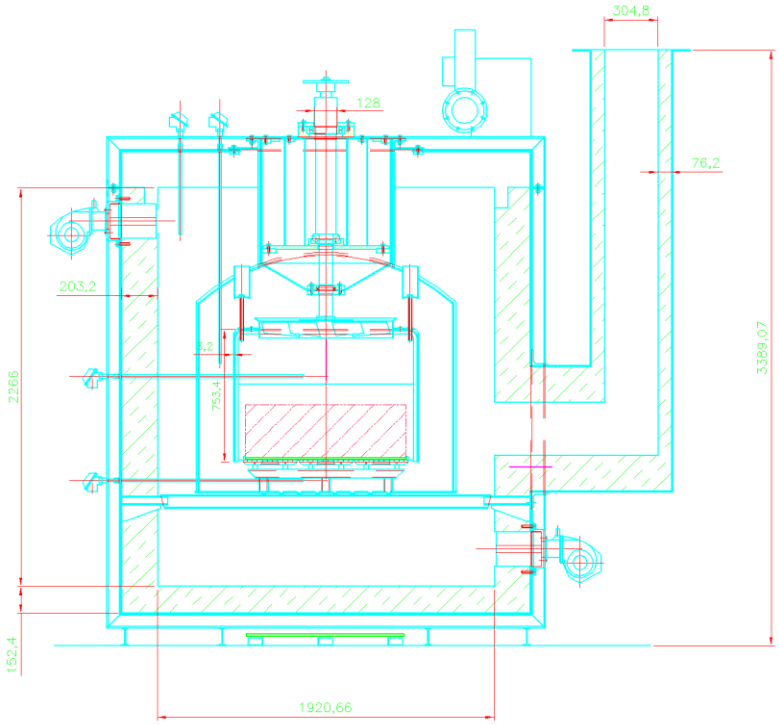


Figura 3-5 Módulo de horneado.

3.9) Módulo de Enfriamiento

Es el último módulo es el módulo de enfriamiento, una vez finalizado el proceso de horneado el PLC ordena detener el recirculador, abrir las compuertas y accionar la cadena para llevar la carga al módulo de enfriamiento. Una vez que la carga llega al módulo de enfriamiento el PLC ordena cerrar las compuertas, encender nuevamente el recirculador y encender las bombas de agua que realizaran la refrigeración de la cámara de enfriamiento. Las piezas permanecen en la zona de enfriamiento durante el mismo tiempo que la carga siguiente permanezca en horneado, luego el PLC ordena detener el recirculador, abrir las compuertas y hace avanzar la cadena hasta que las piezas salen de dicha cámara y quedan libres sobre la cadena para luego ser recogidas por un operador. En la siguiente figura se visualiza una vista en corte del módulo:

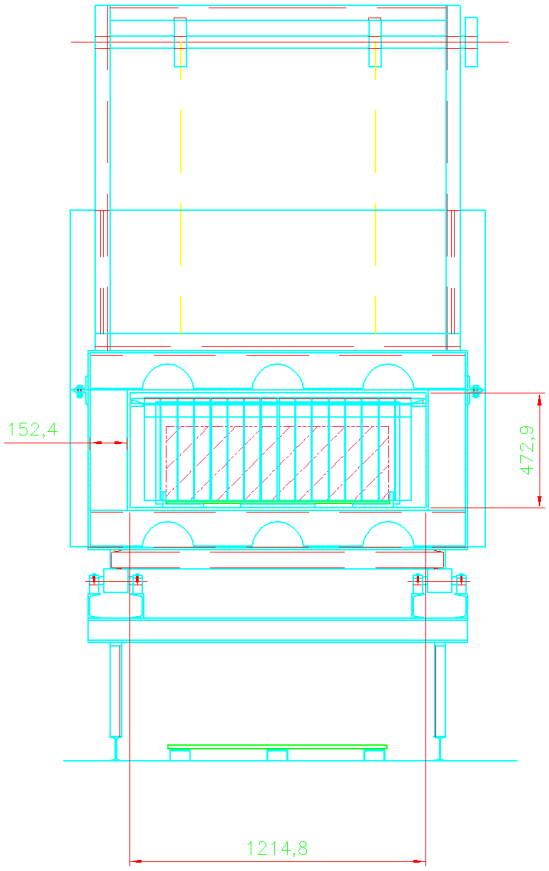


Figura 3-6 Cámara de enfriamiento.

3.10) Prueba de Estanqueidad

Del proceso de horneado se obtiene como producto el panel del radiador, este componente luego pasa al proceso de soldadura donde se le colocan las tapas laterales que permitirán la conexión de las mangueras por las cuales circulará el fluido a refrigerar. El proceso de soldadura de las tapas si bien es necesario para la construcción del radiador, no está bajo el análisis de este estudio por lo cual no se lo describirá en detalle.

Una vez que las tapas están soldadas al panel se obtiene como producto el radiador, llegado a este punto se da inicio al proceso de prueba de estanqueidad. La prueba de estanqueidad consiste en inyectar aire a presión por una de las bocas de entrada del radiador mientras la boca de salida se encuentra sellada. Esto provoca que el aire inunde todas las cavidades del panel y al no poder escapar comience a generar presión en todos los puntos de contacto. Todo este proceso debe realizarse con el radiador sumergido completamente en una pileta con agua, esto permitirá que ante una eventual fuga de aire del radiador ésta pueda ser observada por las burbujas generadas en el agua y así detectar donde se origina la pérdida para poder evaluar si es solucionable o no.

La presión de prueba es un requerimiento establecido por el cliente y la misma suele variar según sea la aplicación del radiador. Por ejemplo, para radiadores que son utilizados en los retornos de los circuitos hidráulicos no es necesario probar el radiador a una gran presión ya que en estos circuitos la presión no suele superar los 2 bares. Pero existen aplicaciones donde el radiador no puede ir conectado al retorno del circuito y debe colocarse dentro del circuito donde muchas veces pueden existir altas presiones (mayores a 20 bares).

Una vez alcanzada la presión solicitada por el cliente se verifica que no existan pérdidas (ausencia de burbujas de aire en la pileta), de ser así se procede a marcar el radiador como apto para el siguiente proceso que según sea el caso puede ser pintado o embalaje. En caso de presentar pérdidas el operador realiza un análisis visual del lugar donde viene la pérdida, si viene del panel puede clasificarse en los siguientes tipos de pérdidas:

3.10.1) Fallas en el panel

- Perfiles: son pérdidas originadas en las uniones de los perfiles de aluminio que a la vista parecen bien soldados pero en realidad no lo están.
- Chapa: son pérdidas en la superficie de las chapas del panel, estas pueden estar ocasionadas por corrosión o por perforaciones de objetos punzantes.

- Pisos despegados: son pérdidas que surgen de la unión de los perfiles con las chapas que conforman los pisos del panel y esto puede ser debido a varios factores como por ejemplo; suciedad en la superficie de los componentes, oxidación de las superficies, temperaturas de horneado insuficientes, etc.
- Chapa colectora: son pérdidas en los radiadores de tipo tubular y se originan en la soldadura entre la chapa colectora y el tubo de aluminio.

3.10.2) Fallas en las tapas

- Poros: son pérdidas en la superficie de las tapas que están soldadas al panel, ocurren en tapas que son de aluminio fundido y por las características del material fundido se producen microporos que provocan pequeñas pérdidas.
- Fijaciones: son pérdidas que se encuentran en fijaciones soldadas al panel o a las tapas.
- Roscas: son pérdidas que se encuentran en roscas que se realizan generalmente sobre las tapas de los paneles.
- Conectores: son pérdidas que se encuentran en los conectores las cuales suelen aparecer en las roscas.

3.10.3) Fallas en soldadura

- Tapas a panel: son pérdidas que se originan en la soldadura de las tapas con el panel.
- Conectores: son pérdidas que se originan por la soldadura de los conectores a las tapas.
- Planchuelas: son pérdidas que se originan por la soldadura de planchuelas de aluminio a las tapas. Dichas planchuelas son utilizadas como refuerzos y/o elementos de sujeción.
- Punteras: son pérdidas que se originan en la soldadura de aquellas tapas que requieren punteras en sus extremos.

4) Capítulo 4 Propuesta de Solución

4.1) Introducción

Es este capítulo se adopta la metodología de trabajo KDD (Knowledge Discovery in Databases) la cual se desarrollará en detalle a lo largo de todo el capítulo junto con herramientas de trabajo que darán soporte a la implementación de la solución. Posteriormente se desarrollará el estudio de minería de datos siguiendo cada uno de los pasos definidos en la metodología y aplicando los modelos resultantes a la problemática planteada.

4.2) Metodología KDD

4.2.1) Introducción

Se define el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" [13].

En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído:

- Válido: hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.
- Novedoso: que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- Potencialmente útil: la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.

- Comprensible: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

Debido a la definición expuesta se entiende que la minería de datos solo es una de las diversas etapas que conforman la metodología KDD, esta situación se puede observar en la siguiente figura.

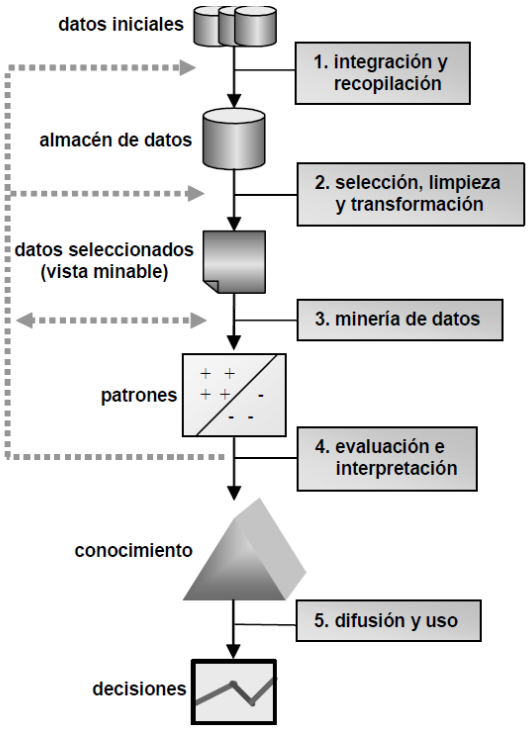


Figura 4-1 Fases de KDD [5].

4.3) Herramientas y lenguajes de programación para minería de datos

Los resultados del SMS desarrollado en el capítulo 2.2.3 no permitieron evidenciar las herramientas o lenguajes de programación más utilizados en proyectos de minería de datos. Para complementar el SMS se realiza una búsqueda de tendencias que permitan determinar cuáles son las herramientas con mayor

popularidad para el análisis de datos. Para este análisis se enumeran las herramientas más mencionadas en los papers y tesis analizados en el SMS. Si se comparan las herramientas estudiadas en el aplicativo de Google Trends [12] podemos ver que las tendencias de búsqueda posicionan al lenguaje de programación Python (puntualmente su librería más famosa para el análisis de datos "Sklearn") como el más buscado en internet.

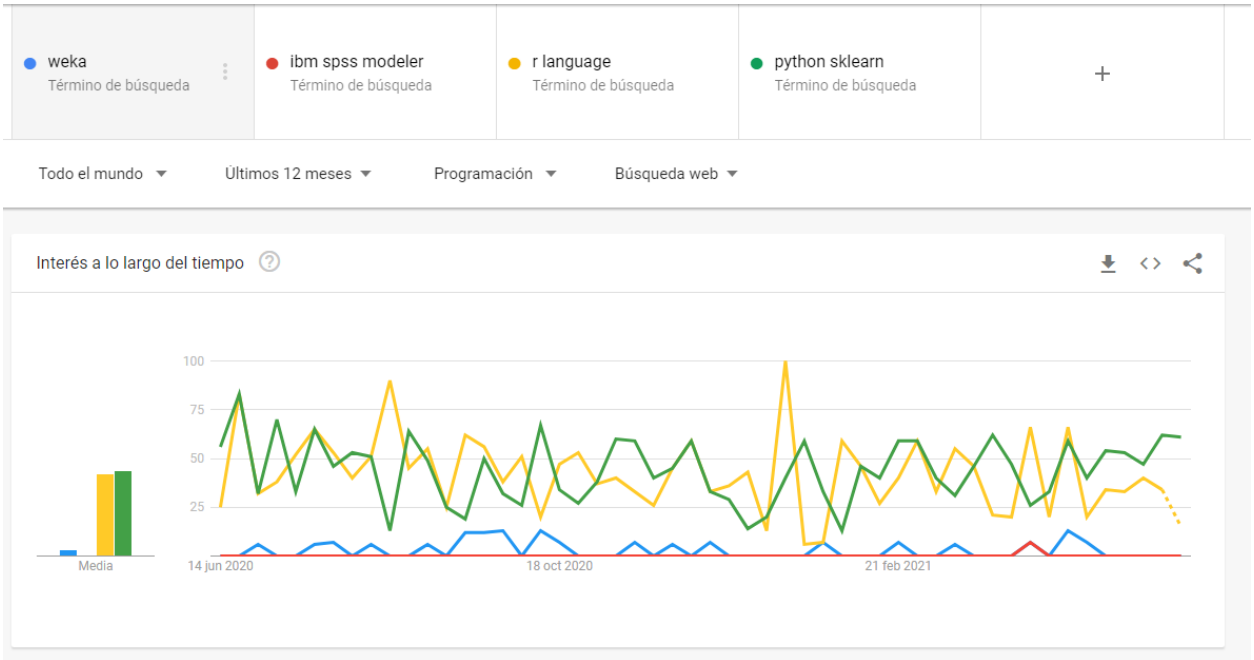


Figura 4-2 Tendencias en herramientas y lenguajes de programación para minería de datos.

Python es un lenguaje de alto nivel interpretado, con una curva de aprendizaje leve y una sintaxis limpia. En la actualidad Python tiene una gran comunidad desarrollando bibliotecas para la minería de datos, entre las principales se pueden destacar:

- Scikit-learn: Es una biblioteca para aprendizaje automático que incluye algoritmos de clasificación, regresión y análisis de grupos.
- Pandas: Es una biblioteca que incluye estructura de datos, y operaciones para manipular tablas numéricas y series temporales.

- Matplotlib: Es una biblioteca para la creación de gráficos a partir de listas o arrays.

Es por estas características que Python se selecciona como herramienta de desarrollo para este trabajo.

4.4) KDD Fase de Integración y Recopilación

La información que se registra en el proceso de Brazing es almacenada de manera manual en planillas donde se vuelcan los siguientes datos:

- Cantidad de paneles por carga
- Tipo de carga
- Ancho del panel
- Hora de ingreso a precámara
- Hora de ingreso en horno
- Hora de salida
- Ordenes de trabajo
- Caudal de nitrógeno
- Temperatura de quemador 1
- Temperatura de quemador 2
- Temperatura de quemador 3
- Temperatura de quemador 4
- Temperatura de precalentamiento
- Control visual
- Fecha de horneado

En la siguiente figura se muestra la plantilla de registro para el proceso de Brazing que utiliza la empresa.

MBP

CONTROL DIARIO DE BRAZADO DE PANELES

Nª	CANT.DEPANELES	DESCRIPCION DE CARGA		HS. INGRESO A PRECAMARA	HS. DEINGRESOEN HORNO	HORA DESALIDA	OTs
		TIPO DECARGA	ANCHO				
1							
2							
3							
4							
5							

	ESTIPULADO	Primer carga	C/2 Horas	C/2 Horas	C/2 Horas	C/2 Horas
CAUDAL NITROGENO EN HORNO	78					

TEMPERATURAS PROGRAMAD	Cada 2 Horas y/o Cambio de Temperatura						
ENFRIADOR SUPERIOR							
PURGA INFERIOR							
ENFRIADOR INFERIOR							
ENFRIADOR SUPERIIOR							
PRECALENTAMIENTO							

CONTROL FINAL DE RADIADORES

Nº OT	OBSERVACIONES	CANTIDAD	OK	NO OK	DESCARTE

		FIRMA			
OPERARIO :					
SUPERVISOR :				FECHA :	

Figura 4-3 Control diario de brazado de paneles RG-88.

Este registro al estar en soporte papel fue necesario digitalizarlo en principio a una planilla formato "xls" compatible con Microsoft Excel y otros editores de código libre para planillas de cálculo. Una vez

digitalizados los registros a una hoja de cálculo fue posible exportar estos datos en formato "csv[2]" el cual es muy utilizado por las bibliotecas de Python para importar y exportar datos de diversas fuentes.

El registro RG-88 contiene los datos referidos al proceso de Brazing, pero para este análisis, son necesarios los datos que indican cómo fue el resultado del proceso de Brazing, es decir, si la soldadura del panel fue correcta o tuvo deficiencias de calidad. La prueba de calidad que verifica la condición de soldadura del panel y define si esta correcta se efectúa en el proceso de control de estanqueidad. La descripción de este proceso ya fue realizada en el capítulo 3 **Prueba de Estanqueidad**, los datos que surgen en esta etapa son en principio registrados en formato papel en el registro de la empresa RG-10 el cual podemos ver en la siguiente imagen.

[2] "Comma-separated values." En informática, un archivo de valores separados por comas almacena datos tabulares en texto plano. Cada línea del archivo es un registro de datos. Cada registro consta de uno o más campos, separados por comas.

MBP

CONTROL FINAL DE ESTANQUEIDAD DE ENFRIADORES

Fecha	O.T. N°	Prueba N°	Presion (Kg)	Panel							Tapas				Soldadura					Reparado	OK	NO OK	Probador	Observaciones
				Perfil	Chapa	Pisos	Tubo	Chapa Colectora	Ciba	Anulado	Poros	Rosca	Fijacion	Conector	Tapa / panel	Conecto	Planchu	Puntera	Operari					
		69103																						
		69104																						
		69105																						
		69106																						
		69107																						
		69108																						
		69109																						
		69110																						
		69111																						
		69112																						
		69113																						
		69114																						
		69115																						
		69116																						
		69117																						
		69118																						
		69119																						
		69120																						

RG-010 Rev.06 24-11-2016 Pág. 1 de 1

Figura 4-4 Control de estanqueidad RG-10.

Luego del registro en soporte papel los datos son digitalizados en el sistema de la empresa quedando almacenados en una base de datos MySQL[3].

Para poder vincular los datos de ambos registros es necesario realizar algunas operaciones sobre los datos que permitan vincular los números de orden de trabajo que son el dato que tienen en común ambos registros y mediante el cual se relacionan para conseguir una trazabilidad de las piezas a los largos de los procesos.

3 MySQL es un sistema de gestión de bases de datos relacional desarrollado bajo licencia dual: Licencia pública general/Licencia comercial por Oracle Corporation y está considerada como la base de datos de código abierto más popular del mundo, y una de las más populares en general junto a Oracle y Microsoft SQL Server, todo para entornos de desarrollo web.

Para comprender el mecanismo de vinculación entre ambos registros primero es necesario comprender como se vinculan las ordenes de trabajo para la fabricación de una pieza terminada que está constituida por varios componentes. En la siguiente figura del sistema de la empresa se puede ver que la estructura de un radiador terminado está compuesta por la orden de trabajo de un panel y la orden de trabajo de un soporte metálico que forma parte del conjunto.

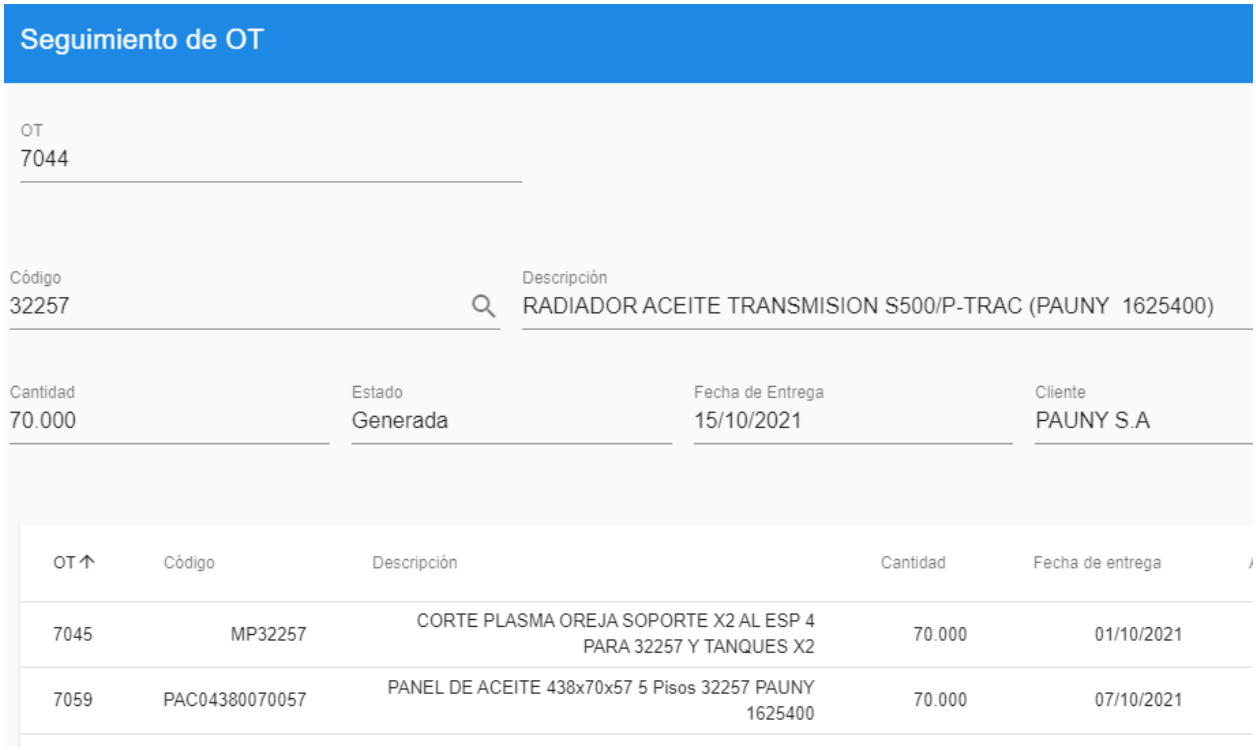


Figura 4-5 Relación entre órdenes de trabajo.

A partir de la Figura 4-5, tomando como ejemplo esa estructura de producto, en el proceso de Brazing (RG-088) se registran todas las cargas de paneles horneados bajo el número 7059, luego, en el proceso de control de estanqueidad (RG-013) se registran las pruebas de estanqueidad de los radiadores, en este caso el registro se realiza sobre el número de orden del radiador (7044).

Para lograr vincular ambos registros se parte de los datos almacenados de las pruebas de control de estanqueidad y mediante una consulta SQL se extraen los paneles asociados a los radiadores probados cada uno con su correspondiente orden de trabajo. Al conseguir el dato del número de orden de trabajo del panel

asociado al radiador éste mismo se puede vincular a los datos registrados en el proceso de Brazing, para este caso, utilizando las herramientas provistas por la librería Pandas del lenguaje de programación Python.

La unión que se realiza en estos dos tipos de registros consiste en un producto cartesiano que arroja como resultado un conjunto de filas que luego son agrupadas de acuerdo a la orden de trabajo del panel. El agrupamiento es necesario ya que, como están determinados los registros y procesos actualmente en la empresa, es imposible establecer una trazabilidad directa entre el radiador que se prueba en el proceso de estanqueidad y el panel horneado con sus correspondientes datos referidos a las características de la carga de horneado (tiempos, temperaturas, etc.).

Previo a esta unión es necesario agrupar los datos por la columna "orden de trabajo", como se mencionó anteriormente el producto cartesiano realizado en la consulta join de SQL arrojo el resultado de la asociación de los registros de estanqueidad cruzados con los de Brazing a partir del dato en común de las ordenes de trabajo de paneles. Pero como se analizó anteriormente, no es posible realizar una trazabilidad directa entre la prueba de los paneles y la carga en la que estos fueron horneados, por lo cual se realiza una agrupación de los datos de Brazing para evaluar el promedio de datos que representan a la carga de horneado del panel. Una vez realizada esta agrupación de los datos de Brazing estos ya se encuentran en condiciones de unirse a los datos del control de estanqueidad.

4.5) Fase de Selección, Limpieza y Transformación

En esta etapa se cuenta con los datos de ambos procesos (Brazing y control de estanqueidad) unificados en una misma tabla representada por una estructura de datos de Python denominada DataFrame[4]. Esta estructura de datos es la que permite a través de funciones de la biblioteca Pandas manipular los datos

[4] Estructura de la biblioteca Pandas que permite representar los datos en dos dimensiones (filas y columnas), donde cada columna puede adoptar un tipo específico de datos (entero, cadena de texto, fecha y hora, etc.).

para realizar la limpieza y transformaciones necesarias. El DataFrame inicial con el cual se comienza a trabajar tiene un total de 33 columnas. De todas estas columnas se realizan una serie de filtros, manipulaciones y transformaciones las cuales se enumeran en la siguiente lista:

1. Se formatean las columnas relacionadas a los tiempos de horneado, el formato actual es de tipo texto y se hace una transformación al objeto DateTime para poder realizar operación entre fechas y horas. También se agrega una columna adicional que representa el tiempo transcurrido de la carga dentro del módulo de horneado, esta columna se calcula como: $tiempo\ en\ horno\ =\ hs\ de\ salida\ -\ hs\ de\ ingreso\ a\ horno$

   En el proceso se afectan las columnas (“hs de ingreso a precámara”, “hs de ingreso a horno”, “hs de salida”, “tiempo en horno”).
2. Se transforma la columna de estado, el proceso define 3 estados posibles para la prueba de estanqueidad de los paneles, estos son:
   a. OK
   b. No OK
   c. Reparado

   Para el análisis que se requiere realizar es necesario transformar los valores del estado reparado en “No OK” en funciona de determinar si la falla en la prueba estanqueidad está asociada a un defecto en el panel. Para realizar esta transformación se analizan las columnas relacionadas a fallas en el panel (“mChapa”, “mBagueta”, “mPerfil”, “mPisoDesp”) y la columna de estado. Para comprender el significado de las columnas que representan las fallas en un panel se deja una breve explicación:

   a. mChapa: Son fallas que ocurren en los paneles de aceite con las chapas que funcionan como separadoras entre los pisos.

b. mBagueta: Son fallas que ocurren en el Brazing de las baguetas para los paneles de aceite.
c. mPerfil: Son fallas que ocurren entre los perfiles que separan los pisos del panel longitudinalmente.
d. mPisosDesp: Son fallas que ocurren en el los paneles de aceite y que representan pisos del panel no soldados.

3.

En este paso se quitan las columnas que no aportan relevancia para el estudio que se está realizando, este procesamiento nos deja como resultado las siguientes columnas útiles en el DataFrame:

a. cantidad de paneles,
b. ancho de panel,
c. enfriador superior,
d. purga inferior,
e. enfriador inferior,
f. enfriador superior,
g. precalentamiento,
h. nitrógeno,
i. tiempo en horno,
j. ot,
k. estado,
l. mChapa,
m. mBagueta,
n. mPerfil,
o. mPisoDesp,

p. tRosca,

q. tPoros,

r. sConector,

s. sTapaPanel,

t. sPlanchuelas,

u. presión,

v. mAnulado,

w. mCiba,

x. tFijacion,

y. mChapaColectora,

z. sPuntera,

aa. tConector,

bb. otPanel

4. Por último se realiza una normalización de datos para ajustar los rangos de las diferentes columnas y de esta manera evitar que columnas con mayor valor absoluto tengan más peso en el modelo. Para esta tarea utilizamos una de las técnicas más sencillas que permiten escalar los rangos de datos en [0,1], mediante la siguiente ecuación:

$$X_{(scaled)} = \frac{x}{max(|x|)}$$

En la siguiente imagen se puede ver el DataFrame escalado:

	cantidad de paneles	ancho de panel	enfriador superior	purga inferior	enfriador inferior	purga superior	precalentamiento	NITROGENO	tiempo en horno	estado
0	0.535714	0.48629	1.0	1.0	1.0	1.0	1.0	1.0	0.785714	1.0
1	0.535714	0.48629	1.0	1.0	1.0	1.0	1.0	1.0	0.785714	1.0
2	0.535714	0.48629	1.0	1.0	1.0	1.0	1.0	1.0	0.785714	1.0
3	0.535714	0.48629	1.0	1.0	1.0	1.0	1.0	1.0	0.785714	0.0
4	0.535714	0.48629	1.0	1.0	1.0	1.0	1.0	1.0	0.785714	1.0

Figura 4-6 DataFrame escalado.

4.5.1) Tratamiento de los Outliers

El outlier consiste en un dato anómalo, como una observación que se desvía marcadamente de otros miembros de la muestra en la cual se encuentra [14].

Un dato anómalo con estas características es muy probable que produzca ruido en los modelos de minería de datos y es por este motivo que resulta conveniente analizar estos desvíos y eliminar dichos datos anómalos.

Para este análisis descartamos todos los datos que se encuentran por debajo de 1.5 veces el primer cuartil y lo que se encuentra por encima de 1.5 veces el tercer cuartil.

```
Bigote inferior: -2.0
Bigote superior: 6.0
```

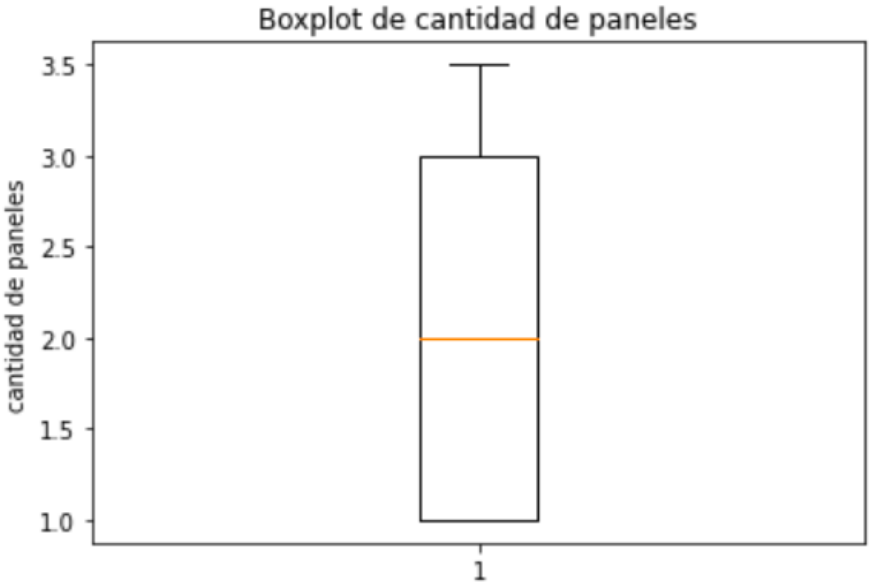


Figura 4-7 Outlier cantidad de paneles.

En la Figura 4-7 no se detectan outliers para el campo “cantidad de paneles”.

```
Bigote inferior: -45.0
Bigote superior: 227.0
```

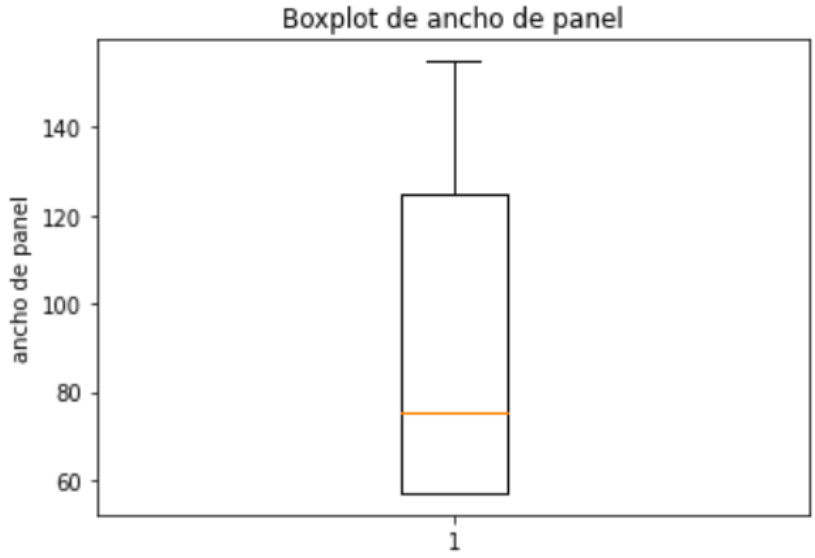


Figura 4-8 Outlier ancho de panel.

En la Figura 4-8 no se detectan outliers para el campo “ancho de panel”.

```
Bigote inferior: 554.6896551724137
Bigote superior: 657.5862068965519
```

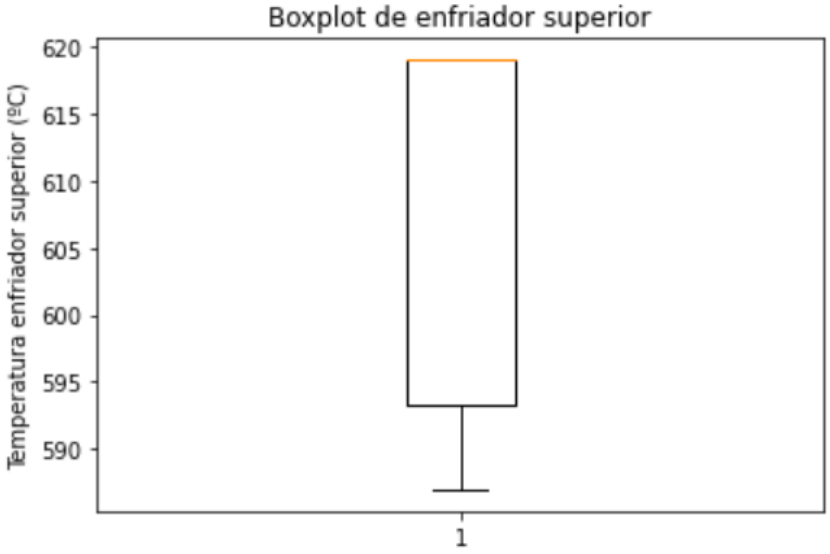


Figura 4-9 Outlier enfriador superior.

En la Figura 4-9 no se detectan outliers para el campo “enfriador superior”.

```
Bigote inferior: 550.5
Bigote superior: 650.5
```

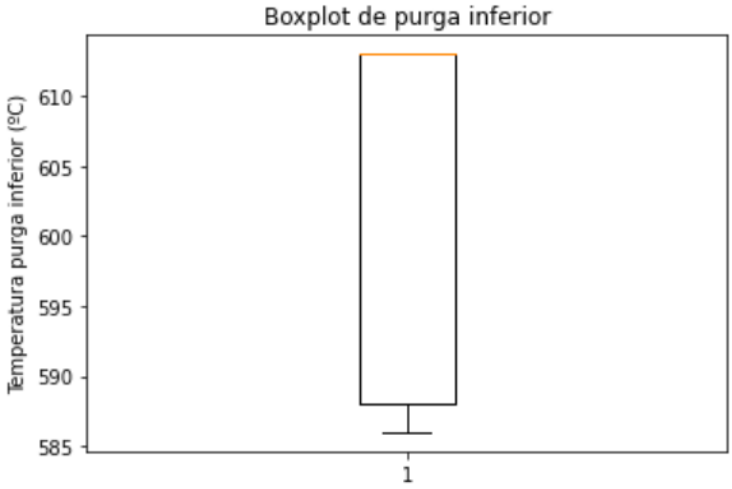


Figura 4-10 Outlier purga inferior.

En la Figura 4-10 no se detectan outliers para el campo “purga superior”.

```
Bigote inferior: 549.2931034482758
Bigote superior: 651.2241379310344
```

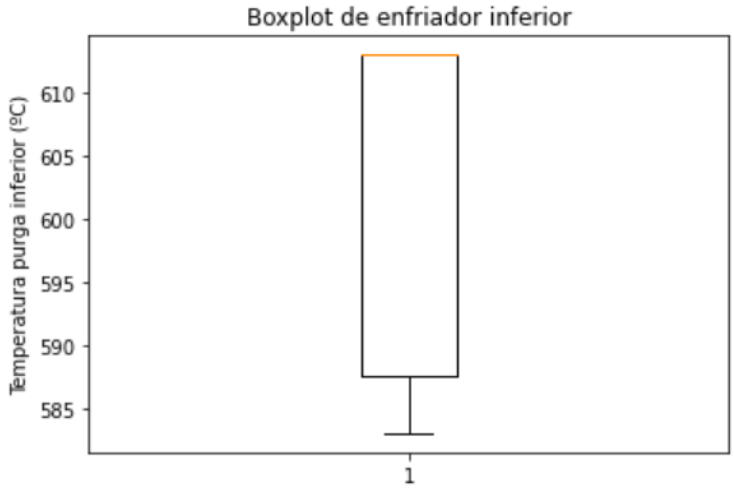
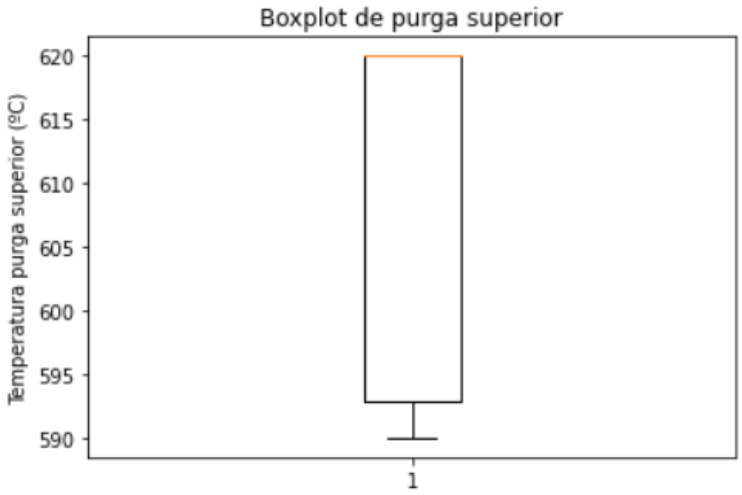


Figura 4-11 Outlier enfriador inferior.

En la Figura 4-11 no se detectan outliers para el campo “enfriador inferior”.

Figura 4-12 Outlier purga superior.

```
Bigote inferior: 552.155172413793
Bigote superior: 660.7068965517242
```



En la Figura 4-12 no se detectan outliers para el campo “purga superior”.

```
Bigote inferior: 375.0
Bigote superior: 375.0
```

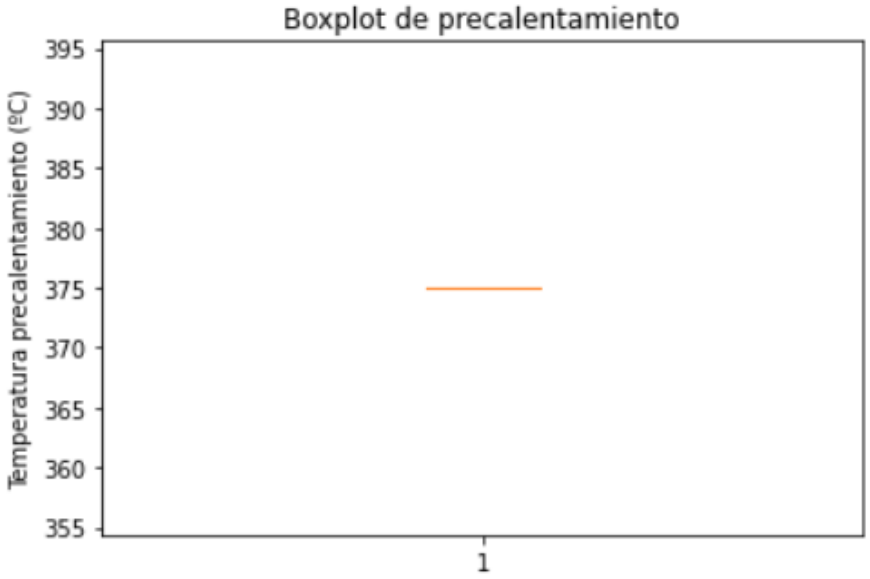


Figura 4-13 Outlier precalentamiento.

En la

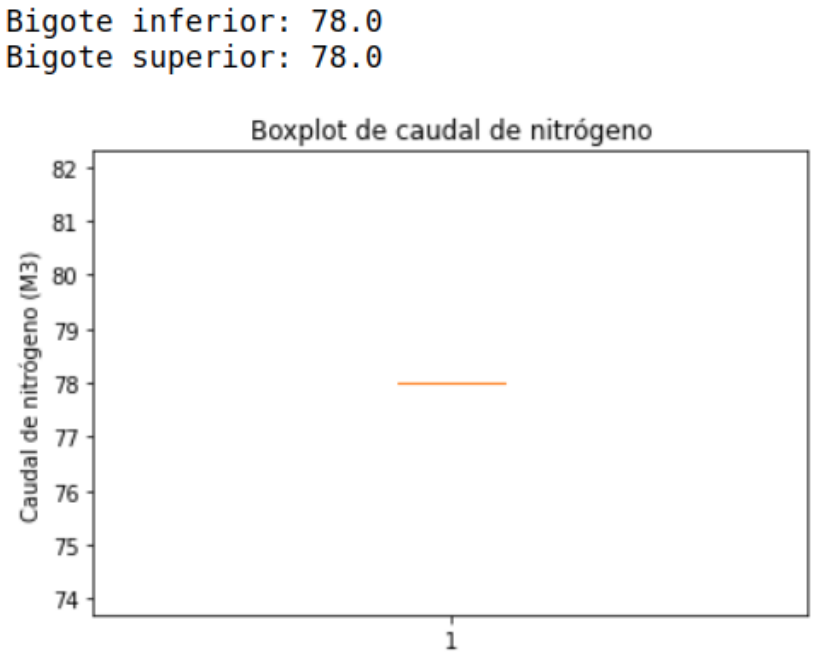


Figura 4-13 Outlier precalentamiento.

no se detectan outliers para el campo “precalentamiento”.

Figura 4-14 Outlier caudal de nitrógeno.

En la Figura 4-14 no se detectan outliers para el campo “nitrógeno”.

```
Bigote inferior: 2.857142857142854
Bigote superior: 40.952380952380956
```

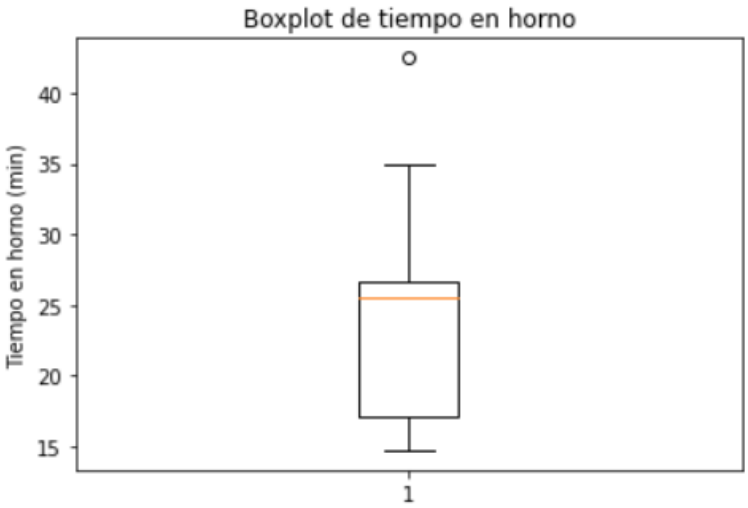


Figura 4-15 Outlier tiempo en horno.

En la

Figura 4-15 se detecta un valor outlier que indica un tiempo en el horno mayor a 40 minutos, por lo tanto, se procede a eliminar la fila que contiene este dato “anómalo”.

Las columnas restantes del DataFrame representan valores discretos (categóricos) por lo cual no es necesario realizar el análisis de descubrimiento de outliers.

Llegado a este punto se da por concluida la fase de selección, limpieza y transformación de datos. El conjunto de datos resultante se exporta a un archivo “csv” para luego poder ser utilizado como fuente de análisis en los modelos predictivos a desarrollar.

4.6) Fase de Minería de Datos

Para la construcción de un modelo predictivo que permita determinar en funciona de ciertas variables de entrada si el panel que se obtendrá del proceso de Brazing estará en buenas condiciones (OK) o en malas condiciones (No OK), se analizaran los siguientes modelos:

a. Regresión lineal múltiple
b. Regresión logística
c. Árbol de decisión
d. Máquinas de vector soporte (SVM)
e. Redes neuronales artificiales

4.6.1) Regresión Lineal Múltiple

El análisis de este modelo es mayormente teórico, para el caso de estudio la variable a predecir es el "estado" de los paneles luego de horneado, al ser una variable discreta, es decir que solo puede aceptar un numero finito de valores (para este caso "ok" y "no ok") el modelo de regresión lineal no es el más apropiado en este caso.

Como primer paso es conveniente determinar cuáles son las variables más importantes, es decir, las que tienen más peso en su relación con la variable a predecir. Para esta tarea se utilizan dos herramientas de selección de rasgos de la biblioteca Sklearn (feature_selection y svm).

```
from sklearn.svm import SVR
from sklearn.feature_selection import RFE
```

```
x = data_brazing_estanqueidad_filtered[colsIn]
y = data_brazing_estanqueidad_filtered["estado"]
```

```
estimator = SVR(kernel="linear")
selector = RFE(estimator, 5, step=1)
selector = selector.fit(x, y)
```

Figura 4-16 Código para selección de rasgos.

Para este caso indicamos que la cantidad de variables a utilizar serán 5 de un total de 11 variables iniciales. El procesamiento de este código arroja como resultado las 5 variables con más peso del conjunto las cuales se muestran a continuación.

Variable	¿Se incluye?
cantidad de paneles	True
ancho de panel	True
enfriador superior	False
purga inferior	True
enfriador inferior	True
purga superior	True
precalentamiento	False
Nitrógeno	False
tiempo en horno	False
Ot	False
otPanel	False

Tabla 4-1 Selección de rasgos.

Para este análisis utilizamos la función LinearRegression de la biblioteca Sklearn. Una vez separadas las variables predictoras y la variable a predecir ejecutamos el modelo con estos datos obteniendo los siguientes resultados.

```
x = data_brazing_estanqueidad_filtered[colsIn]
y = data_brazing_estanqueidad_filtered["estado"]
```

```
lm = LinearRegression()
lm.fit(x,y)
```

```
LinearRegression()
```

El score que arroja el modelo es igual a 0.07311022982421678 lo cual indica que no hay mucha relación entre las variables predictores y la variable a predecir.

Los coeficientes de cada una de las variables predictoras quedan representados en la siguiente tabla:

Variable predictora	Coeficiente
cantidad de paneles	0.0659825047545223
ancho de panel	-0.003372191793639701
enfriador superior	-0.24964386512105013
purga inferior	-0.2852372424915916
enfriador inferior	0.195102876086697
purga superior	0.30903898465757057
precalentamiento	-1.0269562977782698e-15
nitrógeno	0.0
tiempo en horno	-0.006024421086449287
ot	0.0005131314391071972
otPanel	-0.00048666100310844015

Tabla 4-2 Coeficientes de variables predictoras.

Los coeficientes de la Tabla 4-2 indican la poca relación entre cada una de las variables predictoras respecto a la variable a predecir, mientras más bajo es el coeficiente, más baja es la relación entre las variables.

La ecuación del modelo de regresión lineal múltiple queda formada de la siguiente manera:

$$estado = \text{cantidad de paneles} \times 0.0659825047545223 + \text{ancho de panel} \times -0.003372191793639701 + \text{enfriador superior} \times -0.24964386512105013 + \text{purga inferior} \times -0.2852372424915916 + \text{enfriador inferior} \times 0.195102876086697 + \text{purga superior} \times 0.30903898465757057 + \text{precalentamiento} \times -1.0269562977782698e-15 + \text{nitrogeno} \times 0.0 + \text{tiempo en horno} \times -0.006024421086449287 + \text{ot} \times 0.0005131314391071972 + \text{otPanel} \times -0.00048666100310844015$$

Este análisis permite determinar que el modelo de regresión logística múltiple no es un modelo adecuado para los datos en estudio.

4.6.2) Regresión Logística

Para el análisis sobre este modelo se parte sobre la misma base de datos del modelo de regresión lineal múltiple. Se importa el archivo de datos ya filtrado y se arma un DataFrame.

El modelo de regresión logística, a diferencia del lineal, es más adecuado para predecir variables discretas y debido a esta característica se ajusta mejor a los datos de estudio de este trabajo ya que el objetivo es predecir si un panel saldrá del proceso de horneado “OK” o “No OK”.

Para la implementación del modelo se utiliza la biblioteca “linear_model” del paquete Sklearn. En principio en el modelo se consideran todas las columnas de datos como variables predictoras, excepto la variable que queremos predecir.

Listado de variables predictoras:

a. cantidad de paneles,
b. ancho de panel,
c. enfriador superior,
d. purga inferior,
e. enfriador inferior,
f. purga superior,
g. precalentamiento,
h. nitrógeno,
i. tiempo en horno

Variable a predecir: “Estado”.

```
logistic_model = linear_model.LogisticRegression()
logistic_model.fit(x_columns, y_column)
```

Figura 4-17 Modelo de regresión logística en Python.

El modelo arroja como resultado un $R^2 = 0.76$, esto se interpreta como una eficacia de prediccion del 76%. Por otro lado podemos enumerar en la siguiente tabla los coeficientes de la funcion logistica:

Dato	Coeficiente
cantidad de paneles	- 0.5934048961290318

ancho de panel	- 0.04340900788593574
purga inferior	0.42141086034827474
enfriador inferior	0.5343856419217285
purga superior	0.5179631763813778

Tabla 4-3 Coeficientes del modelo logístico.

Validación del Modelo Logístico

Para asegurar la eficacia del modelo es necesario probarlo con un conjunto de datos distintos a los datos de entrenamiento y de esta manera verificar su precisión.

Con la biblioteca train_test_split del paquete Sklearn se puede separar del conjunto de datos disponibles una parte para entrenamiento y una parte para pruebas. En la práctica es habitual separar un 70% de los datos para el entrenamiento del modelo y un 30% para las pruebas. Una vez separados los datos se entrena al modelo con los datos de entrenamiento (70%) y luego se prueba el modelo entrenado con el conjunto de datos de prueba (30%) para analizar su efectividad en las predicciones.

El modelo calcula las probabilidades, no las clases, es decir no calcula si el panel saldrá OK o no OK, calcula la probabilidad de que salga OK y no OK. Después depende del analista definir un límite a partir del cual considere la probabilidad como OK o no OK.

Probabilidad No OK	**Probabilidad OK**
0.23127669	0.76872331
0.2664654	0.7335346
0.2664654	0.7335346
0.2664654	0.7335346
0.19952471	0.80047529
0.20071212	0.79928788
0.24538057	0.75461943

Tabla 4-4 Probabilidades según datos de prueba.

Según la Tabla 4-4 se puede ver en la segunda columna algunas de las probabilidades de que un panel salga OK de proceso de Brazing según los datos de prueba utilizados en el modelo.

Para poder probar la predicción categórica sobre el estado de los paneles solo es necesario ejecutar la función “predict” del modelo con los datos de prueba, esto arroja los resultados de la siguiente tabla:

Estado del panel
1
1
1
1
1
1
1

Tabla 4-5 Resultados de la predicción del conjunto de pruebas.

La Tabla 4-5 indica para estos datos de prueba que los paneles saldrán en estado “OK”, esto se basa en que el umbral de decisión por defecto en el algoritmo es de 0.5, y como se vio en la Tabla 4-4 todas las probabilidades de que el resultado sea “OK” son mayores a 0.5.

Este umbral definido por defecto en el algoritmo se puede cambiar en función de la necesidad, es posible que, por ejemplo, el analista desee considerar que un panel está OK cuando su probabilidad supera el 80%. Para este caso puntual el resultado de la predicción arroja los siguientes valores:

Probabilidad OK	Predicción
0.76872331	0
0.7335346	0
0.7335346	0
0.7335346	0
0.80047529	1
0.79928788	0
0.75461943	0

Tabla 4-6 Predicción con umbral del 80%.

Curvas ROC

Las curvas ROC[5] son una herramienta grafica que permiten comprender la eficacia de clasificación de un modelo. Para el caso de la regresión logística la predicción brindada por el modelo puede ser correcta o incorrecta.

Con esta técnica es posible determinar si existe overfitting[6] en el modelo, de esta manera se puede descubrir si el modelo se ajusta mucho a los puntos suministrados por el conjunto de datos y cuando recibe datos diferentes no tiene una buena precisión en sus predicciones.

Con la utilización de la biblioteca "metrics" del paquete Sklearn y el paquete "ggplot" podemos hacer uso de la función "roc_curve" la cual brinda un gráfico de la curva ROC en base a los parámetros de sensibilidad y especificidad. La función "rocu_auc_score" permite calcular la AUC ("área under the curve").

```
from sklearn.metrics import roc_auc_score

roc_auc_score(y_test, prob)

0.5091750841750842
```

Figura 4-18 Calculo del AUC.

[5] En la teoría de detección de señales, una curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

[6] En aprendizaje automático, el sobreajuste (también es frecuente emplear el término en inglés overfitting) es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.

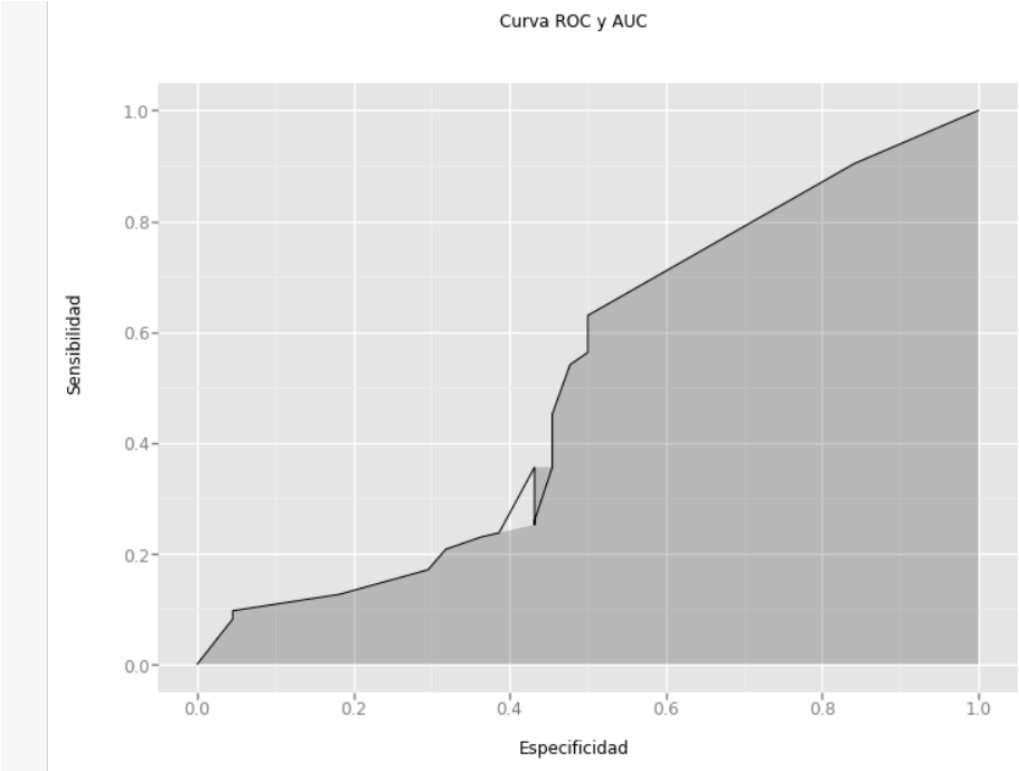


Figura 4-19 Curva ROC y AUC.

Según la Figura 4-19 se puede ver que la curva ROC está lejos de tener su forma ideal, un modelo válido es aquel en el cual su curva tiende a aproximarse a su vértice superior izquierdo. Lo mismo sucede con el área definida debajo de la curva (AUC) la cual dio como resultado 0.509, un modelo valido es aquel que tiende a un AUC=1. Por estos parámetros observados se puede concluir que el modelo de regresión logística programado no es válido para realizar predicciones.

4.6.3) Árbol de decisión

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas [5].

Al igual que con la regresión logística y lineal se parte del mismo conjunto de datos ya filtrados y transformados y se arma el correspondiente DataFrame.

Para la implementación del modelo se utiliza la biblioteca “DecisionTreeClassifier” del paquete Sklearn.

Se consideran las siguientes variables predictores para la construcción del modelo:

a. Cantidad de paneles
b. Ancho de panel
c. Purga inferior
d. Enfriador inferior
e. Purga superior

La variable a predecir es el estado del panel el cual puede ser “OK” o “NO OK”.

```
tree = DecisionTreeClassifier(criterion="entropy", min_samples_split=20, random_state=99)
tree.fit(predictors_train, target_train)
```

Figura 4-20 Modelo de árbol de decisión en Python.

En la Figura 4-20 se puede ver la ejecución del árbol de decisión con una serie de parámetros que serán determinantes para la definición de la estructura del árbol y la precisión que logre brindar el modelo en sus predicciones. El primer parámetro define como criterio de decisión la entropía la cual determina la aleatoriedad de los datos en cada nodo del árbol y el parámetro “random_state” define la aleatoriedad de permutaciones, al ingresar un valor entero garantizamos un comportamiento determinístico con lo cual podemos reproducir siempre los mismos resultados aunque ejecutemos varias veces el mismo código.

El resultado del árbol puede visualizarse con la biblioteca “export_graphviz” del paquete Sklearn como muestra la siguiente imagen.

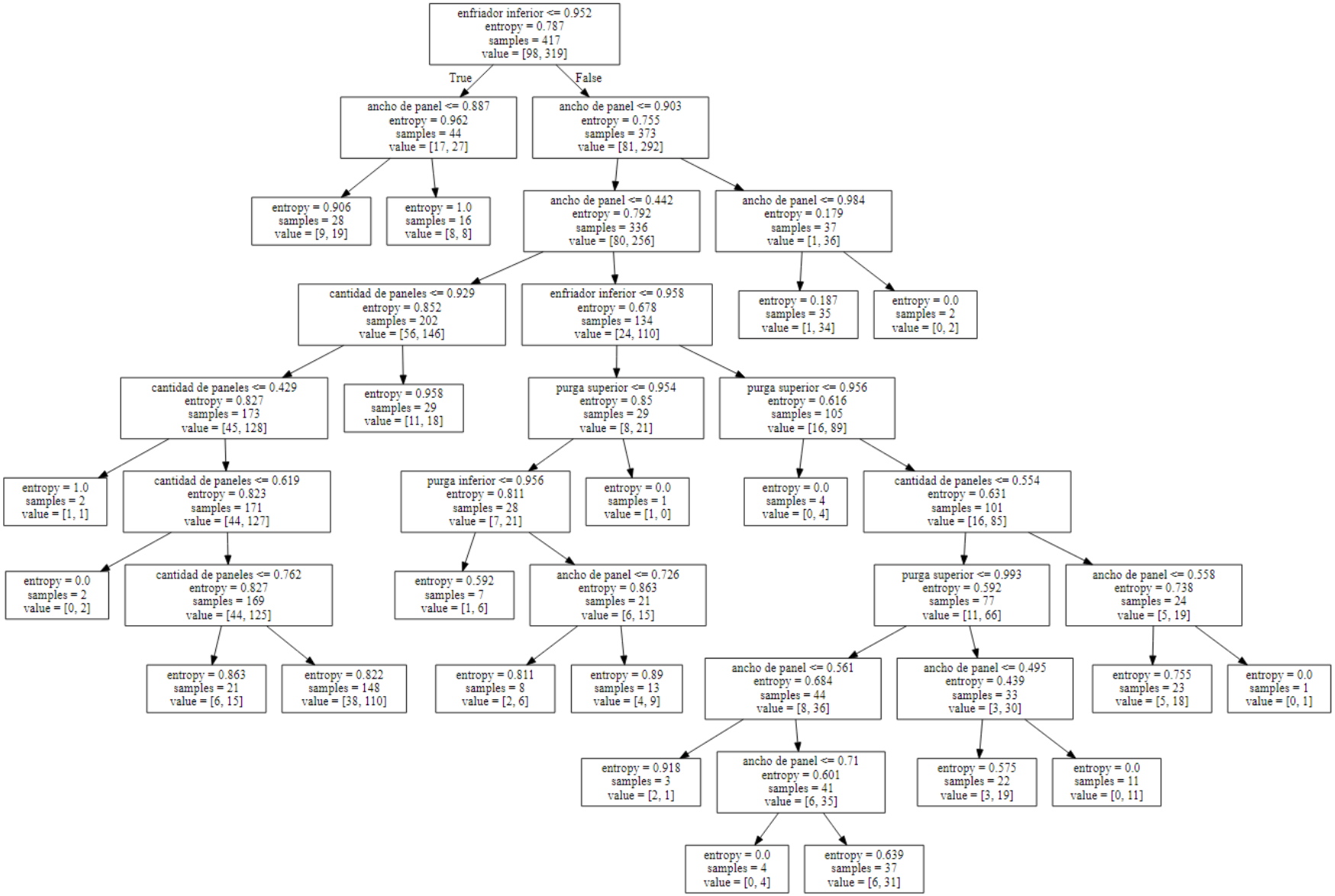

enfriador inferior <= 0.952
entropy = 0.787
samples = 417
value = [98, 319]
True
False
ancho de panel <= 0.887
entropy = 0.962
samples = 44
value = [17, 27]
ancho de panel <= 0.903
entropy = 0.755
samples = 373
value = [81, 292]
entropy = 0.906
samples = 28
value = [9, 19]
entropy = 1.0
samples = 16
value = [8, 8]
ancho de panel <= 0.442
entropy = 0.792
samples = 336
value = [80, 256]
ancho de panel <= 0.984
entropy = 0.179
samples = 37
value = [1, 36]
cantidad de paneles <= 0.929
entropy = 0.852
samples = 202
value = [56, 146]
enfriador inferior <= 0.958
entropy = 0.678
samples = 134
value = [24, 110]
entropy = 0.187
samples = 35
value = [1, 34]
entropy = 0.0
samples = 2
value = [0, 2]
cantidad de paneles <= 0.429
entropy = 0.827
samples = 173
value = [45, 128]
entropy = 0.958
samples = 29
value = [11, 18]
purga superior <= 0.954
entropy = 0.85
samples = 29
value = [8, 21]
purga superior <= 0.956
entropy = 0.616
samples = 105
value = [16, 89]
entropy = 1.0
samples = 2
value = [1, 1]
cantidad de paneles <= 0.619
entropy = 0.823
samples = 171
value = [44, 127]
purga inferior <= 0.956
entropy = 0.811
samples = 28
value = [7, 21]
entropy = 0.0
samples = 1
value = [1, 0]
entropy = 0.0
samples = 4
value = [0, 4]
cantidad de paneles <= 0.554
entropy = 0.631
samples = 101
value = [16, 85]
entropy = 0.0
samples = 2
value = [0, 2]
cantidad de paneles <= 0.762
entropy = 0.827
samples = 169
value = [44, 125]
entropy = 0.592
samples = 7
value = [1, 6]
ancho de panel <= 0.726
entropy = 0.863
samples = 21
value = [6, 15]
purga superior <= 0.993
entropy = 0.592
samples = 77
value = [11, 66]
ancho de panel <= 0.558
entropy = 0.738
samples = 24
value = [5, 19]
entropy = 0.863
samples = 21
value = [6, 15]
entropy = 0.822
samples = 148
value = [38, 110]
entropy = 0.811
samples = 8
value = [2, 6]
entropy = 0.89
samples = 13
value = [4, 9]
ancho de panel <= 0.561
entropy = 0.684
samples = 44
value = [8, 36]
ancho de panel <= 0.495
entropy = 0.439
samples = 33
value = [3, 30]
entropy = 0.755
samples = 23
value = [5, 18]
entropy = 0.0
samples = 1
value = [0, 1]
entropy = 0.918
samples = 3
value = [2, 1]
ancho de panel <= 0.71
entropy = 0.601
samples = 41
value = [6, 35]
entropy = 0.575
samples = 22
value = [3, 19]
entropy = 0.0
samples = 11
value = [0, 11]
entropy = 0.0
samples = 4
value = [0, 4]
entropy = 0.639
samples = 37
value = [6, 31]

Figura 4-21 Grafico de árbol de decisión

Una vez entrenado el modelo se utiliza una matriz de confusión para probar la precisión de sus predicciones contra los datos de prueba separados oportunamente.

Predicciones / Actual	0	1
0	7	37
1	4	131

Tabla 4-7 Matriz de confusión.

La Tabla 4-7 muestra los resultados de utilizar el modelo para predecir el conjunto de datos de pruebas.

Para un valor real 0 el modelo realizo 7 predicciones correctas y 37 incorrectas, y para un valor real 1 el modelo realizo 4 predicciones incorrectas y 131 correctas.

Se puede concluir que para el conjunto de datos suministrados el modelo tiene una buena tasa de acierto (97%) cuando el estado del panel es 1 (OK), pero por el contrario, cuando el estado del panel es 0, la tasa de acierto (15,9%) es muy baja y esto implica un problema para detectar paneles en mal estado.

Poda del árbol de decisión

La subdivisión del árbol de decisión en muchos nodos no siempre garantiza una mayor presión del modelo. Es necesario analizar hasta qué punto es conveniente expandir el árbol con más nodos y determinar en lo posible el punto óptimo donde se obtiene la mayor precisión en la predicción.

Para este propósito se utiliza la validación cruzada que nos permite a través del parámetro “score” medir la precisión del modelo con valores en el rango [0,1], donde 1 representa una precisión absoluta.

Desde Python se utiliza la función “cross_val_score” del paquete Sklearn y se obtienen los scores.

```
scores = cross_val_score(tree, predictors, target, scoring="accuracy", cv=cv, n_jobs=1)
scores
```

```
array([0.75      , 0.85      , 0.66666667, 0.85      , 0.73333333,
       0.75      , 0.79661017, 0.81355932, 0.69491525, 0.76271186])
```

Figura 4-22 Validación cruzada en Python.

Con el parámetro que arroja la validación cruzada ahora es posible construir "n" modelos ajustando la profundidad del árbol y analizando la variación de su score para determinar cuál modelo tiene mayor precisión.

Luego de realizar las iteraciones con una profundidad de 7 niveles el árbol arroja su mayor nivel de score (76.84%). También es posible obtener las variables predictoras de mayor peso para la predicción correspondiente al árbol de decisión modelado. Para este conjunto de datos las dos variables que resultaron más influyentes en la predicción son:

- ancho de panel
- enfriador inferior

4.6.4) Máquinas de vector soporte (SVM)

Las máquinas de vectores soporte pertenecen a la familia de los clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy dimensionalidad (introducidos por funciones núcleo o kernel) con un sesgo inductivo muy particular (maximización del margen) [5].

En la siguiente imagen se visualiza un conjunto de datos separado por un hiperplano.

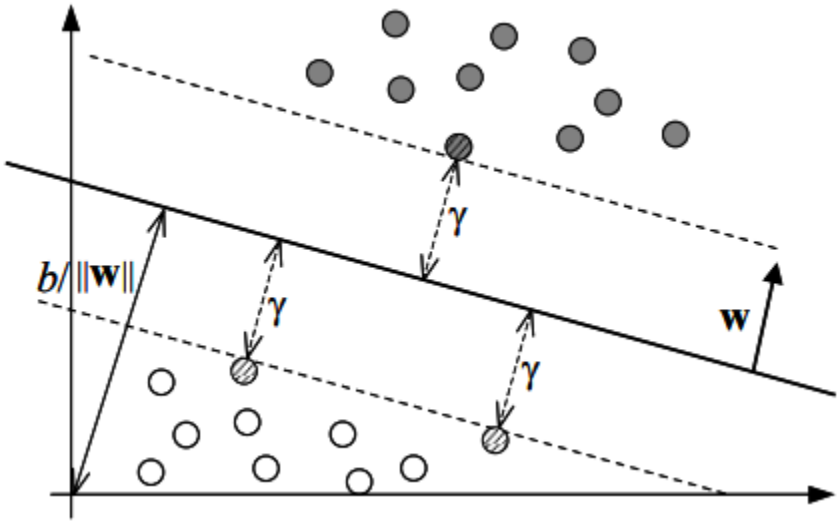


Figura 4-23 Hiperplano (w,b) equidistante a dos clases, margen geométrico (γ) y vectores soporte (puntos rayados) [5].

El objetivo de las SVM es seleccionar el mejor hiperplano separador entra las diferentes categorías de datos, y el mejor será aquel que maximice la distancia mínima entre las categorías de datos representadas. En la Figura 4-23 se representa un caso bidimensional, para el caso de estudio actual se etiquetaron 5 variables predictoras por lo cual no es posible representar gráficamente en un plano las 5 dimensiones.

El conjunto de variables predictoras seleccionado es:

a. Cantidad de paneles
b. Ancho de panel
c. Enfriador superior
d. Purga inferior
e. Enfriador inferior
f. Purga superior
g. Tiempo en horno

La variable a predecir es: “estado”.

Para la creación del modelo se utiliza la biblioteca SVC de paquete Sklearn, en la

La función SVC admite diversos parámetros:

a. Kernel: Puede tomar los valores “linear”, “poly”, “rbf”, “sigmoid”, “precomputed”. El kernel toma los datos que tienen una dimensión “N” y los traslada a un espacio de dimensión mucho mayor “M” donde resulte más fácil clasificarlos. Dicho kernel no es más que una función que recibe dos muestras y devuelve un valor. De acuerdo al tipo de kernel que se selecciona se aplicará un tipo diferente de función que será más o menos apta para clasificar de acuerdo a la dimensionalidad y distribución de los datos.

b. C: Es un parámetro que contempla una penalización por la clasificación incorrecta de los datos, también denominado como holgura. Cuando el parámetro se acerca a 0 la holgura es mayor y esto implica que el hiperplano de separación admite más puntos por fuera de las restricciones.

```
svc_lin = SVC(kernel="linear", C=1)
```

Figura 4-24 Modelo de maquina vector soporte en Python.

En la Figura 4-24 se intenta una primera aproximación con un kernel lineal y un parámetro de holgura bajo.

Al igual que con los otros modelos se realiza un entrenamiento con un 70% del conjunto de datos y pruebas con el 30% restante. Luego de entrenar el modelo se realizan las predicciones para el conjunto de datos de prueba y sobre estas predicciones se calcula el nivel de precisión que obtuvo el modelo, que para este caso fue de un 76.49%, lo cual en un principio indica que la capacidad predictiva del modelo es aceptable.

Si embargo para validar esta precisión se realiza una matriz de confusión que permita visualizar como respondió el modelo con sus predicciones a los datos reales.

Predicciones / Actual	1
0	44
1	135

Tabla 4-8 Matriz de confusión kernel lineal.

Del análisis de la Tabla 4-8 se puede ver que el modelo predice 135/179 datos de manera correcta (75.42%) y 44/179 datos de manera incorrecta (24.58%). El problema que se puede visualizar en este resultado es que el modelo no arroja predicciones para el valor 0, con lo cual no puede detectar paneles que están mal horneados.

Se realiza otra prueba con un parámetro de holgura mayor (C=1000) para ajustar las restricciones y comparar los nuevos resultados pero se vuelven a obtener los mismos resultados.

La siguiente prueba que se realiza recrea el modelo pero esta vez con un kernel del tipo “rbf”, se realiza el entrenamiento y las pruebas con el mismo conjunto de datos que el kernel lineal.

El score obtenido con este kernel es de 54.17%, es un valor muy bajo para una predicción lo cual lo hace un modelo inútil.

Si se realizan la matriz de confusión para este kernel se pueden visualizar el resultado de las predicciones.

Predicciones /	0	1

Actual		
0	1	43
1	0	135

Tabla 4-9 Matriz de confusión para kernel "rbf"

Según la Tabla 4-9 el modelo predice bien 1/44 de resultados negativos (2.27%) y de resultados positivos predice 135/135 (100%). La conclusión vuelve a ser la misma con respecto a que el modelo no puede predecir los resultados negativos, es decir, cuando los paneles salen en mal estado del proceso de horneado.

4.6.5) Redes Neuronales Artificiales

Las redes neuronales artificiales son un método de aprendizaje cuya finalidad inicial era la de emular los procesadores biológicos de información. Las RNA parten de la presunción de que la capacidad humana de procesar información se debe a la naturaleza biológica de nuestro cerebro [5].

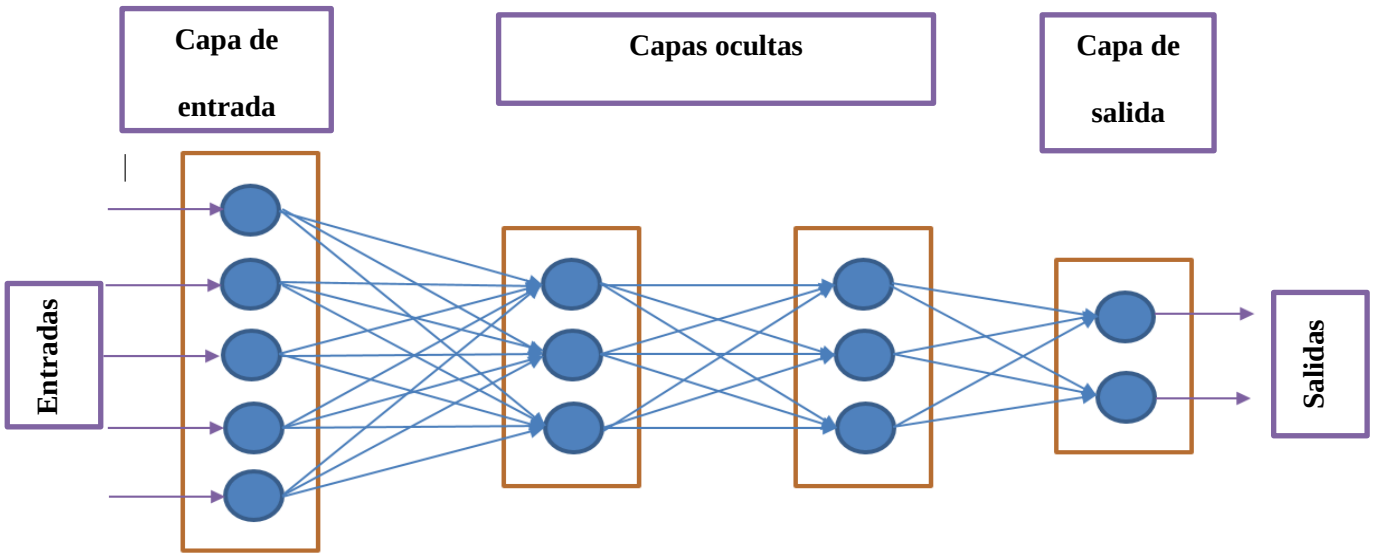


Figura 4-25 Ejemplo de red neuronal artificial de cuatro capas

En la Figura 4-25 se observa una red neuronal construida por cuatro capas, la primera capa (capa de entrada) recibe los valores que corresponden a los datos de entrada, en este caso estos valores están representados por las siguientes variables predictoras:

a. Cantidad de paneles
b. Ancho de panel
c. Enfriador superior
d. Purga inferior
e. Enfriador inferior
f. Purga superior
g. Tiempo en horno

Las capas siguientes realizan una suma ponderada de todos los valores de entrada, para realizar esta operación se construye una matriz de pesos comúnmente denominada W. Esta matriz tiene tantas filas como neuronas la capa anterior y tantas columnas como neuronas la capa actual. El siguiente paso consiste en agregar a la suma ponderada otro parámetro que tiene cada una de las neuronas denominado “bias”. Por último a la suma ponderada más la suma del “bias” se le aplica una función de activación y el resultado de esta función será el resultado de la neurona.

```
model = Sequential()
model.add(Dense(16, input_dim=7, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

Figura 4-26 Modelo de red neuronal en Python

En la Figura 4-26 se puede observar un modelo de red neuronal artificial implementado en Python, el modelo de tipo “Sequential” indica que serán creadas las capas una delante de la otra tal como se vio en la Figura 4-25. Se agregan dos capas “Dense” mediante la aplicación de la función “model.add()”. Mediante el parámetro “input_dim” se definen la cantidad de variables de entrada de la red, que para este caso son 7 utilización la función de activación “relu”. La última capa que se agrega es la de salida y posee una sola neurona con la función de activación “sigmoid”.

```
model.compile(loss='mean_squared_error',
              optimizer='adam',
              metrics=['binary_accuracy'])

model.fit(x_train, y_train, epochs=1000)
```

Figura 4-27 Entrenamiento de red neuronal en Python

En la Figura 4-27 se observa la configuración de la red neuronal, donde se indica en principio el tipo de función de perdida para que en las recursivas iteraciones se reduzca la misma, luego se indica el parámetro de optimizador, para este caso fue seleccionado “Adam” (Adaptative moment estimation), aunque también existe otros como SGD, Momentum, NAG, Adagrad, etc. Cada optimizador tendrá resultados diferentes y es conveniente ir probando para detectar cual arroja mejores resultados. Por último se define las métricas que se desean obtener, para este caso “binary_accuracy”.

Una vez configurada la red con sus parámetros se inicia el entrenamiento con el conjunto de datos previamente separados, indicando la cantidad de iteraciones, para este caso 1000.

```
Epoch 1/1000
14/14 [==============================] - 0s 719us/step - loss: 0.2358 - binary_accuracy: 0.6764
Epoch 2/1000
14/14 [==============================] - 0s 575us/step - loss: 0.2083 - binary_accuracy: 0.7745
Epoch 3/1000
14/14 [==============================] - 0s 560us/step - loss: 0.1987 - binary_accuracy: 0.7559
Epoch 4/1000
14/14 [==============================] - 0s 526us/step - loss: 0.1865 - binary_accuracy: 0.7740
```

Figura 4-28 Resultados para la evaluación de una red neuronal en Python.

Como se puede observar en la Figura 4-28 en las primeras 4 iteraciones los pesos se van ajustando, en principio comenzando con una efectividad del 67,64% y en la cuarta iteración una efectividad del 77,40%, este proceso se repite hasta llegar a la iteración número mil, arrojando finalmente una efectividad promedio de 75,42%.

Esta efectividad si bien resulta razonable y similar a las otras técnicas analizadas previamente no aporta suficiente información sobre la eficacia del modelo para predecir paneles con problemas de horneado que es lo que se pretende con el modelo.

```
# Matriz de confusión de las predicciones de test
# ==============================================
confusion_matrix = pd.crosstab(
    y_test.ravel(),
    predicciones[0],
    rownames=['Real'],
    colnames=['Predicción']
)
confusion_matrix
```

Predicción	1.0
Real	
0.0	44
1.0	135

Figura 4-29 Matriz de confusión de red neuronal en Python.

En la Figura 4-29 se realiza una matriz de confusión que permite visibilizar el desempeño del modelo para sus predicciones. Se observa que el modelo no realiza ninguna predicción de valor 0, esto implica que no reconoce ningún panel como mal horneado del conjunto de datos.

4.7) Fase de Evaluación e Interpretación

La siguiente tabla muestra una comparativa de los modelos analizados en la sección anterior.

Modelo	Score (%)	Validación cruzada (%)			
		VP	VN	FP	FN
Regresión lineal	8,89	n/a	n/a	n/a	n/a
Regresión logística	76,17	62,96	50	37,04	50
Árboles de decisión	76,84	97	16,6	3	83,4
Máquinas de vector soporte (kernel lineal)	76,49	75,4	0	24,6	100
Máquinas de vector soporte (kernel rbf)	54,17	100	2,27	0	97,73
Redes neuronales artificiales	75,42	75,41	0	24,58	0

Tabla 4-10 Análisis comparativo entre modelos.

De todos los modelos analizados el que arroja un mejor score es el árbol de decisión con un 76.84%. Sin embargo al realizar la validación cruzada ningún modelo puede predecir con un buen margen (>80%) los verdaderos negativos (VN), siendo en este caso también el árbol de decisión el modelo que mejores resultados arroja (16.6%). Para el caso de los verdaderos positivos (VP) que para este caso de estudio representan a los paneles que salen OK, el modelo de maquina vector soporte es el que mejor predice (100%).

A partir de las matrices de confusión generadas es posible evaluar el desempeño de los modelos con distintas métricas:

- Exactitud: Representa la cantidad de predicciones positivas que fueron correctas y está representada por la ecuación.

$$Ex = (VP + VN) / (VP + VN + FP + FN)$$

- Precisión: Representa el porcentaje de los casos positivos detectados, y se representa por la ecuación.

$$Pr = VP / (VP + FP)$$

- Sensibilidad: Representa la tasa de verdaderos positivos y se representa por la ecuación.

$$Se = VP / (VP + FN)$$

- Especificidad: Representa la tasa de verdaderos negativos y se representa por la ecuación.

$$Esp = VN / (VN + FP)$$

Modelo	Métricas			
	Exactitud	Precisión	Sensibilidad	Especificidad
Regresión logística	0,60	0,79	0,63	0,5
Árboles de decisión	0,77	0,78	0,97	0,16
Máquinas de vector soporte (kernel lineal)	0,75	0,75	1,00	0,00
Máquinas de vector soporte (kernel rbf)	0,76	0,76	1,00	0,02
Redes neuronales artificiales	0,75	0,75	1	0

Tabla 4-11 Métricas de desempeño

Los diversos análisis con modelos supervisados arrojan similares resultados respecto a la incapacidad de poder predecir piezas que salen en mal estado (la especificidad es baja), esto da lugar a las siguientes conclusiones:

1. No existe una correlación entre las variables de entrada a los modelos, es decir, no hay una relación que permita modelar que bajo determinado estado de las variables de entrada un panel sale NO OK.

2. Faltan datos de entrada para mejorar la precisión del modelo. Las variables relevadas como por ejemplo la temperatura se toman a intervalos relacionados con el cambio de carga de tipo de panel en el horno, esto sucede entre dos y 3 veces por día. Estas temperaturas deberían medirse en intervalos de tiempo constante para cada carga para de esta manera poder tener una mayor precisión en las fluctuaciones. También deberían relevarse para cada carga los caudales de nitrógeno y la velocidad de recirculación del recirculador en la cámara que se encarga de homogeneizar la atmosfera. Otro dato sumamente importante y que no es relevado en el proceso tiene que ver con las partes por millón de oxígeno en la atmosfera del horno, es un dato clave para determinar si ocurrirá oxidación en el aluminio lo que luego supondrá una falla casi segura en el proceso de Brazing.
3. El proceso de Brazing es uno más de una cadena de procesos para la fabricación de un radiador. Este análisis no tiene en cuenta que pueden ocurrir problemas de calidad en procesos anteriores que no sean detectados oportunamente y que éstos originen fallas en el proceso de Brazing. Esta situación obliga a tener una estructura que asegure la calidad de todos los procesos para poder analizar cada uno de estos de manera independiente.

5) Conclusiones y Futuras Líneas de Investigación

5.1) Conclusiones

Para la elaboración del estado del arte se utilizó el mapeo sistemático de datos como metodología para la recopilación y análisis de la información. Esta metodología resultó adecuada para la definición de las preguntas de investigación y luego el análisis e interpretación de los resultados.

De las metodologías más utilizadas para la aplicación de minería de datos se optó por KDD, ya que el proceso abarca diversas fases que integran el tratamiento de datos, su recolección, limpieza, transformación e interpretaciones de los resultados. La ventaja de aplicar esta metodología es que permitió detectar errores de formato en los datos de entrada (que se recolectaban manualmente) y corregirlos y transformarlos para su posterior análisis.

Del análisis de las herramientas y lenguajes de programación se optó por trabajar con el lenguaje Python, ya que es el lenguaje más popular para trabajar con datos, esto implica que existen diversas bibliotecas con algoritmos de minería de datos ya desarrollados para poder implementar fácilmente. Otro beneficio de optar por este lenguaje es la gran comunidad de desarrolladores que hay detrás lo cual brinda un buen soporte y garantiza el desarrollo continuo y la corrección inmediata de eventuales problemas.

En cuanto a los objetivos secundarios planteados por este trabajo de investigación, se concluye que lo siguiente:

- Los dos algoritmos con mejor desempeño para el análisis del proceso de Brazing resultaron ser los árboles de decisión y, las maquinas vector soporte (con kernel RBF).
- No se logró establecer una correlación valida entre las variables de entrada del proceso de Brazing y los resultados acerca de si un panel presentaba o no problemas de calidad.

5.2) Futuras Líneas de Investigación

Como parte de futuras líneas de investigación se podrían definir las siguientes:

5.2.1) Redefinir el proceso de recolección de datos

Como se expresó en el capitulo Fase de Selección, Limpieza y Transformación de la metodología KDD los datos relevados para este trabajo surgieron a partir de planillas que los operarios completaban manualmente, posteriormente esta información fue digitalizada en planillas de cálculo para luego ser limpiada y transformadas. El proceso de registro manual implica una gran posibilidad de cometer errores, sería conveniente automatizar el registro de datos tales como caudales, temperaturas y tiempos para reducir al mínimo la posibilidad de error y tener una frecuencia de registro más alta lo que implicaría un mayor volumen de datos para luego ser analizados (más datos y de mejor calidad). Dispositivos de bajo costo como las placas Arduino podrían contribuir en esta tarea comunicándose con los dispositivos físicos del horno y registrando los datos en un soporte físico de la empresa o una nube.

5.2.2) Replicar Estos Modelos en Otras Industrias de Brazing

Actualmente en la Argentina, Metalúrgica BP es la única industria que aplica el proceso de Brazing pero sería posible replicar estos modelos en industrias brasileras donde el proceso de Brazing está desarrollado y gestionado por empresas de similares características en cuanto a infraestructura y cantidad de personal, esto permitiría probar los modelos desarrollados con datos de otra industria y poder comparar los resultados con los obtenidos en este trabajo para obtener conclusiones más específicas en cuanto a la utilidad de los modelos.

5.2.3) Ampliar el Alcance de los Modelos

Es posible ampliar la cantidad de datos que se analizan involucrando otros procesos que integran la cadena de producción de un radiador. Recabar más cantidad de datos puede aportar información valiosa que permita establecer una correlación entre el resultado del proceso de horneado y variables tales como

las dimensiones de los componentes, su aleación y el resultado de los procesos previos al Brazing (corte de componentes, lavado, aplicación de fundente, conformado de aletas, armado y prensado, etc.).

6) Bibliografía

[1] G. ZHIQIANG, S. ZHIHUAN, D. STEVEN y H. BIAO, «Data Mining and Analytics in the Process Industry: The Role of Machine Learning,» 2017.

[2] I. Witten, E. Frank y M. Hall, Data Mining Practical Machine Learning Tools and Techniques, USA: Elsevier, 2011.

[3] J. Schiefer, J. Jeng, S. Kapoor y P. Chowdary, «Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence,» *IEEE International Conference on ECommerce Technology*, 2004.

[4] E. Thomsen, «BI's Promised Land,» de *Intelligent Enterprise*, 2003.

[5] J. H. Orallo, J. Quintana y C. Ramírez, Introducción a la Minería de Datos, Madrid: Pearson, 2004.

[6] A. K. Choudhary, J. A. Harding y M. K. Tiwari, «Data mining in manufacturing: a review based on the kind of knowledge,» *Springer Science+Business Media*, 2008.

[7] D. Sekulić, Advances in Brazing: Science, Technology and Applications, United Kingdom: Woodhead Publishing Limited, 2013.

[8] B. Kitchenham, D. Budgen y P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, USA: CRC Press, 2015.

[9] K. Petersen, R. Feldt, S. Mujtaba y M. Mattsson, «Systematic mapping studies in software engineering,» de *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, 2008.

[10] R. Wieringa, N. Maiden, N. Mead y C. Rolland, «Requirements engineering paper classification and evaluation criteria: A proposal and a discussion,» *Requirements Engineering,* 2005.

[11] M. Genero, J. Cruz-Lemus y M. Piattini, Métodos de investigación en ingeniería del software, Grupo Editorial RA-MA, 2014.

[12] «www.trends.google.es,» 2021. [En línea]. Available: https://trends.google.es/trends/explore?cat=31&q=weka,ibm%20spss%20modeler,r%20language,python%20sklearn. [Último acceso: 13 6 2021].

[13] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, From Data Mining to Knowledge Discovery: An Overview, 1996.

[14] V. Barnett y T. Lewis, Outliers in Statistical Data, 1994.

[15] R. Wieringa, N. Maiden, N. Mead y C. Rolland, «Requirements engineering paper classification and evaluation criteria: A proposal and a discussion,» *Requirements Engineering,* 2005.

7) Anexos

Como comparativa a los modelos realizados en Python se utilizó la herramienta RapidMiner[7], con la cual se realizó un análisis automático con diversos modelos.

Una vez instalado el programa se comienza a trabajar con un modo interactivo.

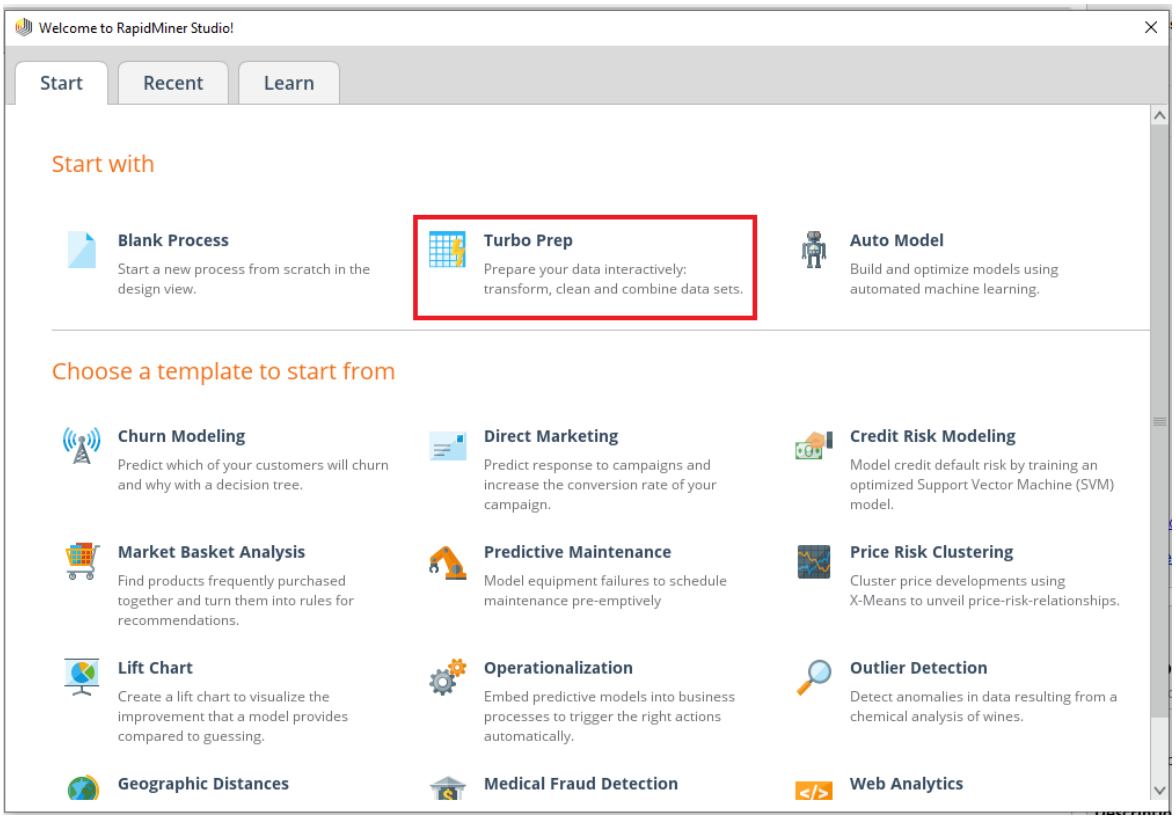


Figura 7-1 Modo interactivo Turbo Prep de Rapidminer.

Se comienza cargando el archivo csv con los datos del proceso de Brazing.

[7] Es un programa informático para el análisis y minería de datos mediante un entorno gráfico.

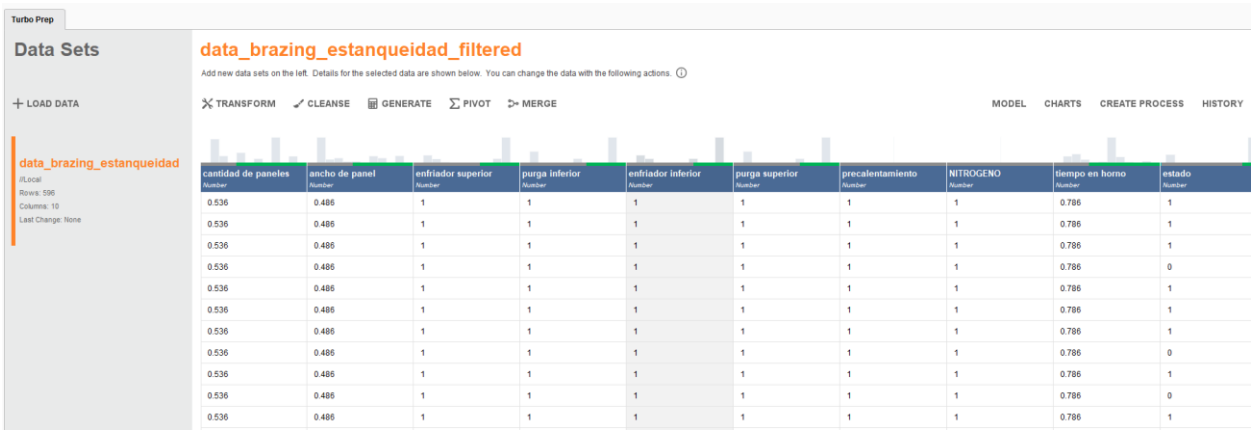


Figura 7-2 Carga de datos en Rapidminer.

El programa permite realizar varias operaciones de manipulación y limpieza de datos, se opta en principio por realizar una limpieza automática.

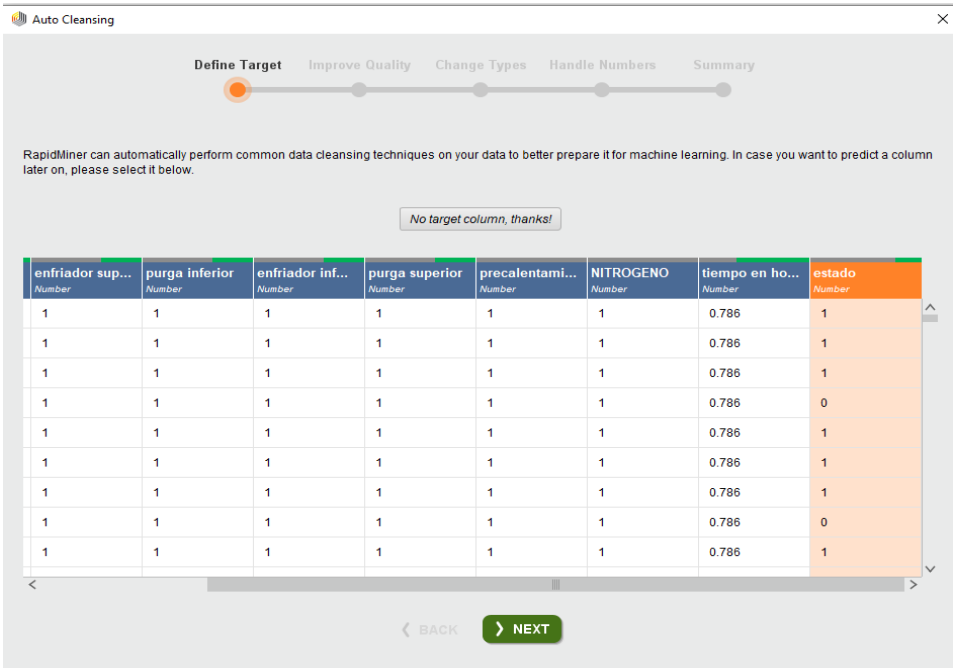


Figura 7-3 Operaciones de limpieza de datos Rapidminer.

En la Figura 7-3 se selecciona la variable que se desea predecir, luego el programa determina las columnas que tienen una alta estabilidad (para este caso “precalentamiento” y “nitrógeno”) y las elimina por no tener relevancia. Posteriormente realiza un reemplazo de valores ausentes y de esta manera finaliza el proceso.

El siguiente paso consiste en seleccionar un modelo para intentar predecir la variable objetivo (estado). Se seleccionan las entradas al modelo en función de la correlación que indica el programa, el mismo programa advierte en función de las variables de entrada que dispone cuales son convenientes descartar indicando esto a través de la columna status como se puede ver en la Figura 7-4.

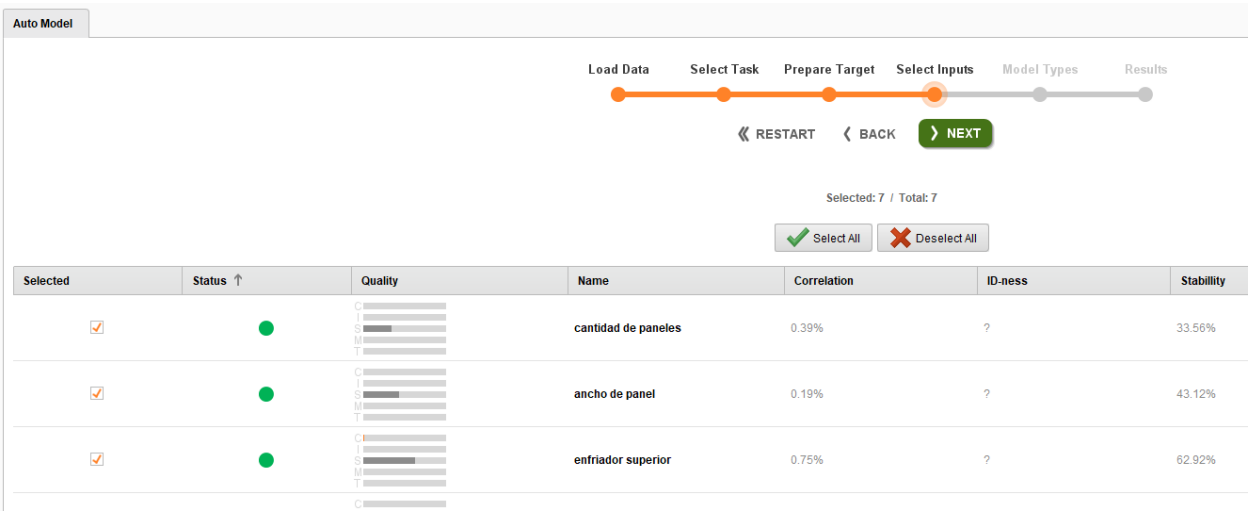


Figura 7-4 Variables de entrada al modelo.

El paso siguiente consiste en seleccionar el tipo de modelo que se desea aplicar para el conjunto de datos ya manipulado.

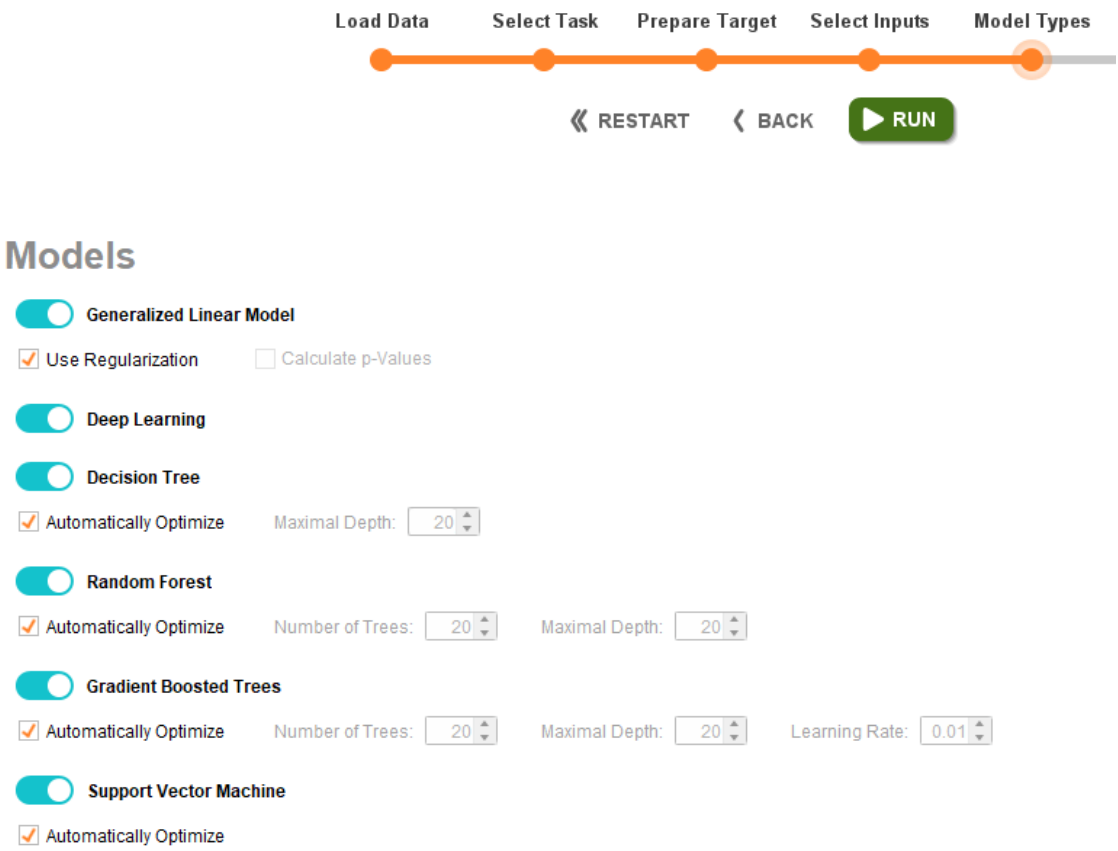


Figura 7-5 Modelos para minería de datos en Rapidminer

Como se muestra en la Figura 7-5, se seleccionan todos los modelos disponibles para realizar el procesamiento de los datos y poder concluir cual resulta más efectivo.

El paso siguiente consiste en procesar cada uno de los modelos seleccionados obteniendo luego varias graficas para su análisis.

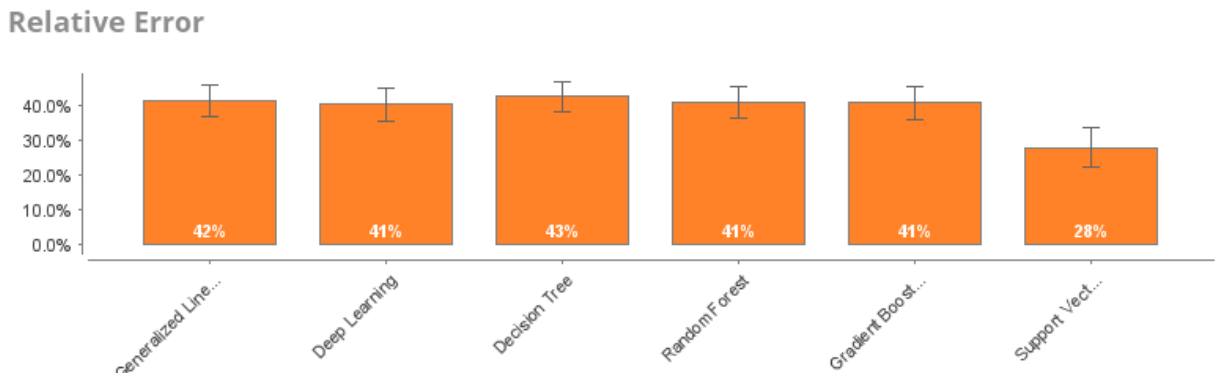


Figura 7-6 Gráfica de error relativo de Rapidminer.

De acuerdo a los resultados obtenidos el modelo que mejor se comportó en cuanto al error relativo es el de máquinas vector soporte (28%).

Model	Relative Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Ro...
Generalized Linear Model	41.6%	± 4.7%	?	2 s	185 ms
Deep Learning	40.6%	± 4.8%	?	2 s	723 ms
Decision Tree	42.7%	± 4.4%	?	362 ms	7 ms
Random Forest	41.3%	± 4.6%	?	2 s	284 ms
Gradient Boosted Trees	41.0%	± 4.7%	?	19 s	91 ms
Support Vector Machine	28.0%	± 5.7%	?	19 s	4 s

Tabla 7-1 Tabla de modelos general Rapidminer.

Según la Tabla 7-1 el modelo más eficaz para la predicción es el de máquinas vector soporte, aunque juntos con el modelo Gradient Boosted Tree son los que más tiempo tardaron en ejecutarse.

El análisis del software Rapidminer permite confirmar lo desarrollado en el Capítulo 4 respecto a que es evidente que no existe correlación entre las variables de entrada y el dato a predecir, por lo tanto ningún modelo logra ajustarse a una predicción aceptable de piezas que salen en mal estado.