**PROPUESTA DE DESARROLLO DE MODELOS DE ANALÍTICA DESCRIPTIVA Y PREDICTIVA COMO HERRAMIENTA PARA GOBIERNO DE DATOS CLAVES EN LA PREVENCIÓN Y ANÁLISIS DE FALLA: ENTREGABLE 1 - BASE DE DATOS**

**INFORME DE AVANCES**

**DEPARTAMENTO DE INGENIERÍA INDUSTRIAL**

**UNIVERSIDAD DE ANTIOQUIA**

**JULIO 2024**

**INTRODUCCIÓN**

En el marco de la creciente necesidad de las empresas por optimizar sus procesos y mejorar la calidad de sus productos, Forjas Bolívar, una destacada empresa metalmecánica, se encuentra en una posición crucial para implementar herramientas avanzadas de análisis de datos. La propuesta de desarrollo de modelos de analítica descriptiva y predictiva surge como una solución integral para mejorar el gobierno de datos clave, especialmente en la prevención y análisis de fallas. Este informe presenta un diagnóstico detallado de la situación actual, así como un planteamiento estructurado para la creación de una base de datos robusta que soporte dichos modelos.

**DESARROLLO**

# **DIAGNÓSTICO**

Forjas Bolívar es una empresa metalmecánica especializada en el proceso de forjado en caliente, dedicados a brindar soluciones de transporte de materiales a los clientes, complementados con servicios de ingeniería y montaje (Forjar, 2024), según la información suministrada el proceso de trabajo de un cliente se divide en las siguientes fases.

**Cotización**: Es el primer contacto con el cliente donde se relaciona el tipo de producto con la necesidad especifica, por parte de forjas bolívar se entregaron los siguientes formatos: Datos de entrada para cotización de equipos, Especificación cliente y Requerimiento cliente.

Estos formatos sirven para contextualizar los requerimientos del cliente y las características del equipo para solucionar dicha necesidad, incluyendo parámetros importantes como condiciones de operación, condiciones de servicio, etc. Sin embargo, todos estos datos pertenecen a los formatos como información no estructurada, por lo que, extraer de forma automatizada no es posible a menos que se haga un preprocesamiento de la información y se estandarice el tipo de formato, es decir que al menos la información tenga una forma de relacionarse en el algoritmo, ya sea con variables claves o posiciones.

Por ejemplo, el formulario “DATOS DE ENTRADA PARA LA COTIZACIÓN DE EQUIPOS” con código DD-FO-01, presenta una fuente de entrada de datos relacionados con las condiciones de operación del equipo, tales como: Nombre del equipo en planta. (Ejemplo: 491TP1), Tipo de necesidad (arrastre, elevador, válvula, silos, tornillos sin fin, fabricaciones especiales, etc.), Material a transportar (Densidad, Granulometría, Humedad, temperatura), documentos técnicos anexos soporte (planos, distancia entre centros, voltaje, frecuencia, caracterización de material, capacidad, datos de control, esquemas. Lo cual representa información clave para hacer un modelo de predicción preciso, sin embargo, al ser un formato de Excel (ver figura 1) no hay una forma estandarizada de diligenciarlo por lo que el usuario podría llenarlo como transcripción verbal de la situación, por lo que dificulta el proceso de extraer información para la generación de base de datos de forma automática; En consecuencia se sugiere utilizar un formato tipo formulario, que permita generar una base de datos mas precisa facilitando el diligenciado por el área correspondiente pero que a la vez relacione los datos de forma más concisa en la base de datos.

|  |  |
| --- | --- |
| **Especificaciones Técnicas** | |
| Material para transportar (Densidad, Granulometría, Humedad, temperatura)\* Si es equipo de transporte. |  |

**Fig. 1**. Ejemplo de sección del formato DD-FO-01, donde se observa la celda que llena el usuario de forma textual.

En la figura 2, se observa un ejemplo de equipo con la información que se requiere para conformar la base de datos para el modelo y que se puede extraer de forma manual o automática de los formatos suministrados.

Aplicación

Descripción generada automáticamente con confianza media

**Fig. 2.** Información representativa del modelo comparado con los archivos suministrados.

De la figura 2 se puede observar que del ideal de la información que se requiere versus lo que se puede extraer de los formatos, en su mayoría son datos no estructurados debido a que se diligencian de forma manual y no siguen siempre el mismo esquema; de las condiciones de operaciones que son las 5 primeras filas se puede extraer de distintos formatos entre cotización y fabricación; del seguimiento hasta una falla se puede extraer una parte de la información del informe de servicio técnico y de los formatos de reclamaciones, sin embargo persiste el mismo inconveniente de los datos no estructurados y la limitación en el numero de datos, debido a que los informes de servicio técnico son pocos y la trazabilidad hasta el equipo que falla se debe hacer de forma manual. El color rojo destaca información importante que no se encuentra en ningún archivo y el color amarillo destaca información que no se encuentra, pero se supone que debería estar consignado en ese formato.

A partir de la información suministrada del informe de servicio técnico y de las reclamaciones (Acciones CA-FO-20 2019-2022) se implementaron técnicas de imputación de datos y conteo de palabras para estudiar la información no estructurada y así poder analizar el texto escrito de forma textual que no sigue un orden especifico, de tal forma que se encontró de la información suministrada aquellos componentes que más fallaban y poder enfocar la base de datos del modelo a esos fallos recurrentes, tal como se observa en la figura 3 que representa el Pareto de componente con fallas mas recurrentes y la figura 4 nube de palabras para las fallas más recurrentes.

Icono

Descripción generada automáticamente con confianza media

**Fig. 3**. Pareto de fallas más recurrentes.

Texto

Descripción generada automáticamente

**Fig. 4**. Nube de palabras de fallas mas recurrentes.

Concluyendo entonces que a partir del Pareto y la nube de palabras, se observa que las fallas más recurrentes están relacionadas con tornillería y con el sistema de cadenas, lo cual permite generar una idea de las variables principales que se deberían incluir en la base de datos del modelo predictivo.

**Fabricación:** En esta fase del proceso se relacionan los elementos y actividades que se requieren para la fabricación de los equipos, la información se encuentra caracterizada en archivos de OP ordenes de proceso el cual relacionan el detalle de la fabricación de cada pieza y componente del equipo, y van desde planos hasta ordenes de producción que incluyen lista de materiales, actividades y recursos para la manufactura de la pieza o equipo. Sin embargo, esta información, poco aporta al proceso de modelación de fallas y no es posible extraer información por métodos exploratorios. Sin embargo cabe resaltar que las ordenes de producción u OP permiten a su vez establecer y entender los parámetros de estándares de fabricación, lo cual facilita el entendimiento para hacerle la trazabilidad a los equipos debido a que en algunos casos la nomenclatura de los equipos no es muy clara, por ejemplo a los equipos o componentes se relacionan con sus especificaciones de forma simplificada, entonces si se quiere decir Cadena (Referencia)-(Tipo aditamento)-(Cadencia)-(CS)-(Paso), se utiliza CDNA FB 864-K443-C2-P7"; adicionalmente BUJE (ØExterior)X(ØInterior)X(Longitud) se escribe con la nomenclatura Buje ØE16XØI11X43, lo que permite relacionar la nomenclatura que usa internamente Forjas Bolívar con el equipo instalado y así poderle hacer la trazabilidad de formas mas sencilla en la base de datos.

Por otro lado, está la información de planta virtual que en primer lugar es una base de datos estructurada la cual permita extraer y manipular información de forma más sencilla, además de que incluye datos para la trazabilidad de los equipos y el seguimiento en áreas posteriores, por ejemplo, el “ConsecutivoOP” y “ConsecPedido”.

**Servicio Técnico:** De la información suministrada por Forjas Bolívar, servicio técnico puede ser aquellos datos de mayor relevancia para el diseño del modelo de predicción; de forma general fueron suministrado algunos ejemplos de informes de servicio técnico con modos de fallas y métodos de medición de elongación y rangos permitidos, sin embargo, la dificultad principal con estos formatos es la limitación en la cantidad de datos suministrados debido a que en los ejemplos solo fueron entregados 3 , además como el formato es llenado por quien realiza el servicio técnico estos pueden tener una estructura diferente en cada caso, lo que la convierte en información no estructurada que complica el proceso de extracción de datos, debido a que no siempre están en la misma posición o relacionados con la misma palabra clave. Sin embargo, también se encuentra un archivo de Excel llamado “informe de servicio técnico Diagnóstico equipos” el cual es una recopilación de datos de modos de falla extraídos de informes de servicio técnico con un total de 28 entradas, que representa una muestra pequeña pero relevante debido a que es información estructurada lo que permite hacerle trazabilidad de forma mas sencilla a los datos, además de que son datos relacionados directamente al modo de falla lo que permite alimentar información idónea al modelo de predicción.

**Reclamación:** En esta fase se encuentra ejemplos del proceso de reclamación que llevan los clientes, reportando incidentes presentados con los equipos, de forma general son informes con datos no estructurados del análisis de reclamación, sin embargo, existe un archivo llamado “ACCIONES CA-FO 2019-2022” que es un resumen de la información de reclamos de forma estructurada que permite conocer las variables de reclamación, pero dentro de la información estructurada hay variables cuya respuestas se encuentra de forma no estructurada, es decir no todos tienen la misma forma de presentación, por ejemplo en la descripción de la situación se escribe textualmente lo que se reporta y en algunos casos reportan información incompleta lo que dificulta hacerle trazabilidad a los equipos o componentes que reclaman.

# **ARQUITECTURA**

Dentro de la información suministrada por Forjas Bolívar existen datos que pueden ser muy relevantes para el modelo de predicción, sin embargo, se han presentado inconvenientes en la entrega de estos debido a que en algunos casos suele ser incompleta o limitada debido a la dificultad interna de forjas bolívar para extraerlos o hacerles la trazabilidad debido a que en la mayoría de los casos comprenden informes que se redactan en formato prosa que cambian según el usuario que lo realice. Por lo tanto, se plantea que a partir de la información de forjas bolívar se podría incluir variables para un modelo con suficiente información que permita hacer predicciones precisas acordes a la realidad, como se muestra en la figura 5.

Gráfico, Tabla

Descripción generada automáticamente

**Fig.5**. Base de datos con variable relevantes para un modelo preciso.

Para relacionar las bases de datos se plantea usar 3 Claves primarias siendo ConsecutivoOP, ConsecPedido y Producto; En los 3 casos las claves son datos numéricos que permiten relacionar la información entre las bases de datos, de tal forma que se generé una unión y trazabilidad desde la fabricación hasta el reporte de falla o reclamación. Se decide Analizar las bases de datos organizadas por forjas bolívar y que permiten extraer información relevante para la base de datos del modelo, así:

Para la base de datos Planta Virtual, se tienen los siguientes registros:

**Tabla 1.** Numero de datos planta virtual

|  |  |
| --- | --- |
| **Descripción** | **Valor** |
| Número de registros | 244637 |
| Número de columnas | 54 |

**Tabla 2**. Tipo de Variables Planta Virtual

|  |  |
| --- | --- |
| **Columna** | **Tipo de Dato** |
| ConsecutivoOP | int64 |
| ConsecPedido | float64 |
| Cliente | object |
| NombCliente | object |
| Producto | object |
| NombProducto | object |
| FechElab | object |
| Proceso | object |
| Version | int64 |
| FechEntregaPdcc | object |
| FechEntregaClte | object |
| OCCliente | object |
| QtyRequerida | float64 |
| Actividad | object |
| QtyControl | int64 |
| RecMAQ | object |
| RecHOM | object |
| SecAct | int64 |
| FechEstInicil | object |
| estado | object |
| TipoConsecutivo | object |

**Tabla 3**. Numero de Nulos Planta virtual.

|  |  |
| --- | --- |
| **Columna** | **Datos Faltantes** |
| ConsecutivoOP | 0 |
| ConsecPedido | 3170 |
| Cliente | 735 |
| NombCliente | 0 |
| Producto | 0 |
| NombProducto | 0 |
| FechElab | 0 |
| Proceso | 0 |
| Version | 0 |
| FechEntregaPdcc | 106930 |
| FechEntregaClte | 253 |
| OCCliente | 0 |
| QtyRequerida | 0 |
| Actividad | 0 |
| QtyControl | 0 |
| RecMAQ | 0 |
| RecHOM | 0 |
| SecAct | 0 |
| FechEstInicil | 66 |
| estado | 0 |
| TipoConsecutivo | 0 |

Para la base de datos Informe servicio técnico, se tienen los siguientes registros:

**Tabla 4.** Numero de datos Informe servicio técnico

|  |  |
| --- | --- |
| **Descripción** | **Valor** |
| Número de registros | 29 |
| Número de columnas | 33 |

**Tabla 5**. Tipo de Variables Informe servicio técnico

|  |  |
| --- | --- |
| **Columna** | **Tipo de Dato** |
| Fecha de visita | object |
| Empresa | object |
| Planta | object |
| País | object |
| Ciudad | object |
| Tipo de equipo | object |
| Orden de compra | object |
| Pedido N° | float64 |
| Fecha de instalación de la cadena | object |
| Material | object |
| Referencia equipo | object |
| Causa de desgaste | object |
| Si escogió otros, cuál? | object |
| Escoja el desgaste donde se presenta | object |
| Tipo de cadena | object |
| Referencia cadena | object |
| Medida referencia de control desgaste nueva en mm | object |
| Medida referencia control de desgaste actual en mm | object |
| Porcentaje de elongación actual % | object |
| Porcentaje de elongación máx % | object |
| Proyección en horas de operación restantes (H) | object |
| Medida sin desgaste en mm | object |
| Medida actual en mm1 | object |
| Porcentaje de desgaste actual % | object |
| Medida sin desgaste en mm1 | object |
| Medida sin desgaste en mm2 | float64 |
| Medida actual en mm2 | float64 |
| Medida actual en mm | float64 |
| Altura de la platina en mm | float64 |
| Medida actual de la platina en mm | float64 |
| Paso en mm | object |
| Cantidad de pasos (Muestra) | float64 |
| Recomendaciones | object |

**Tabla 6**. Numero de Nulos Informe servicio técnico

|  |  |
| --- | --- |
| **Columna** | **Datos Faltantes** |
| Fecha de visita | 1 |
| Empresa | 1 |
| Planta | 2 |
| País | 1 |
| Ciudad | 1 |
| Tipo de equipo | 1 |
| Orden de compra | 6 |
| Pedido N° | 3 |
| Fecha de instalación de la cadena | 4 |
| Material | 4 |
| Referencia equipo | 2 |
| Causa de desgaste | 3 |
| Si escogió otros, cuál? | 13 |
| Escoja el desgaste donde se presenta | 1 |
| Tipo de cadena | 1 |
| Referencia cadena | 1 |
| Medida referencia de control desgaste nueva en mm | 14 |
| Medida referencia control de desgaste actual en mm | 11 |
| Porcentaje de elongación actual % | 4 |
| Porcentaje de elongación máx % | 3 |
| Proyección en horas de operación restantes (H) | 10 |
| Medida sin desgaste en mm | 9 |
| Medida actual en mm1 | 13 |
| Porcentaje de desgaste actual % | 22 |
| Medida sin desgaste en mm1 | 20 |
| Medida sin desgaste en mm2 | 25 |
| Medida actual en mm2 | 25 |
| Medida actual en mm | 23 |
| Altura de la platina en mm | 15 |
| Medida actual de la platina en mm | 20 |
| Paso en mm | 4 |
| Cantidad de pasos (Muestra) | 9 |
| Recomendaciones | 3 |

Es importante destacar que algunas variables presentan una alta tasa de valores nulos (más del 90%). Esto se debe al reducido número total de registros (28) y a que no en todos los casos se realizaron todas las mediciones. Por ejemplo, la medición de "Medida sin desgaste en mm²" solo se realizó en 3 componentes. Por lo tanto, no se emplearon técnicas para imputar valores nulos en esta base de datos, ya que ciertos registros deben mantenerse así debido a la naturaleza del servicio técnico, que no requería realizar mediciones adicionales.

Para la base de datos reclamaciones CA FO 2019-2024, se tienen los siguientes registros:

**Tabla 7.** Numero de datos CA FO 2019-2024

|  |  |
| --- | --- |
| **Descripción** | **Valor** |
| Número de registros | 68 |
| Número de columnas | 27 |

**Tabla 8**. Tipo de Variables CA FO 2019-2024

|  |  |
| --- | --- |
| **Columna** | **Tipo de Dato** |
| ID | int64 |
| Hora de inicio | object |
| Hora de finalización | object |
| Correo electrónico | object |
| Nombre | object |
| Nombre de la persona | object |
| Fecha | object |
| Tipo de acción | object |
| Fuente de la acción | object |
| Proceso | object |
| Pedido inicial 1 | float64 |
| Pedido inicial 2 | float64 |
| Cliente | object |
| Orden de compra (OC) | object |
| Items (Separar los ítems por ; (punto y coma)) | object |
| Causa de Reclamación | object |
| Situación | object |
| 1. ¿Por qué? | object |
| 2. ¿Por qué ? | object |
| 3. ¿Por qué? | object |
| 4. ¿Por qué? | object |
| 5. ¿Por qué? | object |
| Escoja la M (De la causa principal) | object |
| Acción | object |
| Actividades - Plan de acción (Para eliminar la causa raíz del problema) | object |
| Fecha programada para el cierre del plan de acción | object |
| Responsable de verificación de la implementación | object |

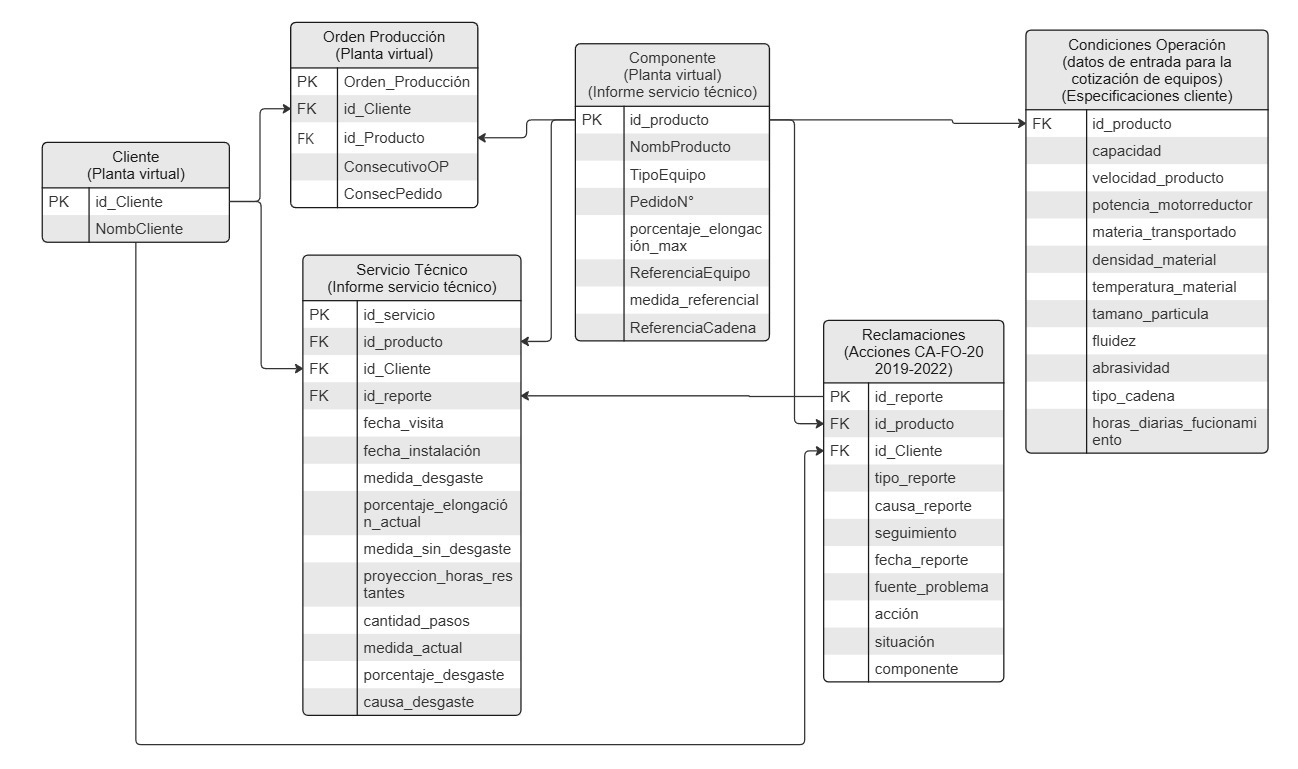
**Tabla 9**. Numero de nulos CA FO 2019-2024

|  |  |
| --- | --- |
| **Columna** | **Datos Faltantes** |
| ID | 0 |
| Hora de inicio | 0 |
| Hora de finalización | 0 |
| Correo electrónico | 0 |
| Nombre | 0 |
| Nombre de la persona | 0 |
| Fecha | 0 |
| Tipo de acción | 1 |
| Fuente de la acción | 0 |
| Proceso | 47 |
| Pedido inicial 1 | 20 |
| Pedido inicial 2 | 57 |
| Cliente | 20 |
| Orden de compra (OC) | 20 |
| Items (Separar los ítems por ; (punto y coma)) | 20 |
| Causa de Reclamación | 21 |
| Situación | 1 |
| 1. ¿Por qué? | 0 |
| 2. ¿Por qué ? | 8 |
| 3. ¿Por qué? | 18 |
| 4. ¿Por qué? | 34 |
| 5. ¿Por qué? | 38 |
| Escoja la M ( De la causa principal) | 0 |
| Acción | 0 |
| Actividades - Plan de acción ( Para eliminar la causa raíz del problema) | 0 |
| Fecha programada para el cierre del plan de acción | 0 |
| Responsable de verificación de la implementación | 0 |

En este caso los nulos corresponden a aquellos datos que no se les pudo hacer trazabilidad directa con la información suministrada.

Finalmente se resalta que condiciones de operación no tiene una base de datos, sin embargo, como es información crucial para hacer un modelo de predicción se debe incluir en el modelo de predicción, así se tenga que hacer una extracción manual de los informes de cotización, que es de donde debería salir la información respectiva para las condiciones.

Ahora, si se comparan los datos se tiene que Planta virtual tiene un número considerable de registros, sin embargo, el informe de servicio técnico y reclamaciones CAFO 2019-2024 no tanto, esto atribuido a la poca información que se puede hacer trazabilidad; por lo tanto, para generar una base de datos con un numero de resultados suficiente para hacer un entrenamiento y una validación del modelo con las variables mencionadas se plantea un diagrama relacional que se observa en la figura 6 , que permite cómo analizar como interactúan las bases individuales para generar la base de datos general como se muestra en la figura 5.



**Fig. 6**. Diagrama relacional para las bases de datos.

En el diagrama relacional se observa como se relacionan las claves primerias ConsecutivoOP, ConsecPedido y Producto, para conformar la base de datos. Por lo tanto, para crear un modelo lo suficientemente robusto se requiere tener la posibilidad de hacer una relación clara por ejemplo con el ID producto, ID Servicio o ID pedido, de las bases de datos de condiciones de operación ( que no hay bases de datos solo los informes de cotización) , Reclamaciones ( que tiene pocos datos y es información no estructurada), Servicio técnico ( que tiene pocos datos y es información no estructurada) y planta virtual ( siendo una base de datos bastante completa).

Por lo tanto, las variables con la información suministrada que si se puede hacer trazabilidad y generar una base de datos con la que se puede modelar es la mostrada en la figura 7.

|  |  |  |
| --- | --- | --- |
| **Base de datos** | **Variables utilizables** | **Tipo de Dato** |
| **Info del equipo: Planta virtual** | ConsecPedido | int |
| ConsecutivoOP | float |
| Producto | object |
| NombProducto | object |
| Cliente | object |
| NombCliente | object |
| OCCliente | object |
| TipoConsecutivo | object |
| Version | int |
| **Condiciones de operación** | Planta | object |
| Fecha de instalación | object |
| Horas diarias de operación | object |
| Tipo de transportador | object |
| Capacidad(t/H), | object |
| Velocidad del equipo (m/s) | object |
| Potencia Motoreductor (KW/h) | object |
| Material a transportar | object |
| Densidad (Kg/m3) | object |
| Temperatura (°C) | object |
| % Humedad | object |
| Tamaño de particula | object |
| Fluidez | object |
| Abrasividad | object |
| **Informe de visita diagnostico técnico de equipos** | Fecha de instalación de la Cadena | object |
| Material | object |
| Referencia equipo | object |
| Causa de desgaste | object |
| Lugar de desgaste | object |
| Medida de referencia cadena | object |
| Medida Actual cadena (mm) | object |
| Metodo de medición | object |
| Porcentaje de elongación Actual | object |
| Porcentaje de elongación Max | object |
| Medida referencia del lugar de desgaste | object |
| Medida actual del lugar de desgaste | object |
| Porcentaje de elongación actual % | object |
| Porcentaje de elongación máx % | object |
| Proyección en horas de operación restantes (H) | object |
| Medida sin desgaste en mm | object |
| Medida actual en mm1 | object |
| Porcentaje de desgaste actual % | object |
| Cantidad de pasos (Muestra) | float |

**Fig. 7.** Variables utilizables suministradas por Forjas Bolívar.

Cabe resaltar que los datos de condiciones de operación aun no se encuentran extraídos debido a que se debe realizar de forma manual una vez se tenga conformada la base de datos para relacionar los informes de cotización con los equipos, adicionalmente el tipo que se le atribuye a esta variables es general y esta sujeto a cambios.

# **INTEGRACIÓN**

En la generación de la base de datos para el modelo con la información suministrada se tiene que: Planta virtual es una base de datos con todos los equipos y componentes cargados por forjas bolívar con su identificador de servicio, sin embargo, la dificultad radica en relacionarlo con las otras variables que no son bases de datos estructuradas. Las condiciones de operación es información crucial, debido a que relaciona factores de servicio que son directamente incidentes a los modos de falla, sin embargo, debido a que es información que se debe extraer de informes de cotización que no siguen una misma estructura dificulta el proceso de extracción automático, además de que se suministraron muy pocos informes de estos. El Informe de visita diagnostico técnico de equipos permita conocer el componente de falla y el rango de elongación máxima relacionada, es una base de datos completa sin embargo no posee un numero muy elevado de datos, las ACCIONES CA FO presenta las reclamaciones presentada por los clientes lo cual también podría ser un insumo importante, sin embargo cuando se intenta hacerle trazabilidad a los equipos para poder relacionar el reclamo con el equipo y las condiciones de operación, en muchos casos no es posible porque no cuenta con el ID producto, ID Servicio o ID pedido, lo que dificulta la relación por ejemplo de las condiciones de operación o información del equipo.

Es de resaltar que las variables mostradas en la figura 3 parten de la información suministrada que, si permite construir una base de datos utilizable en el modelo de predicción, sin embargo, Planta virtual tiene 240.000 resultados entre equipos y componentes que hacen parte del mismo montaje por lo tanto tienen el mismo identificador. Informe de visita diagnóstico técnico de equipos tiene 28 resultados, es decir que tiene información de 28 informes a los cuales se les hizo el diagnostico de falla y se relaciono con valores fuera de los limites permitidos en la elongación. En condiciones de operación actualmente solo hay 3 resultados lo que es una información bastante limitada. Por lo tanto, se concluye que para el ejemplo el modelo de predicción se usaría la información de planta virtual cruzada con el ConsecPedido (debido a que es el más completo y permite realizar trazabilidad entre bases de datos) y los resultados del Informe de visita diagnóstico técnico de equipos, dejando como resultados 28 entradas utilizables para el modelo de predicción. Resaltando que se llegó al acuerdo que una vez se tengan los 28 equipos relacionados con planta virtual y las fallas de elongación se buscaría en los informes de cotización para poder extraer de forma manual las condiciones de operación. El formato en el que se entregaría la base de datos es un archivo \*.Csv que es un compatible con Excel y permite trabajar con las librerías de Python.

**Código de Python para unir las diferentes bases de datos**

# Instalación de Complementos

!pip install pandas openpyxl

!pip install wordcloud

!pip install sweetviz

!pip install openpyxl

#Cargamos librerias

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

from plotly.subplots import make\_subplots

from collections import Counter

import re

from wordcloud import WordCloud

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import sweetviz as sv

# Se carga la base de datos de servicio tecnico

file\_path = 'Informe Servicio Técnico Diagnostico de equipos (2)(Sheet1).csv'

# Lee el archivo CSV especificando el códec

S\_tecnico = pd.read\_csv(file\_path, encoding='latin1')

# Muestra las primeras filas del DataFrame

S\_tecnico.head()

# Se carga la base de datos de planta virtual

file\_path = 'Planta Virtual(Hoja1).csv'

# Lee el archivo CSV especificando el códec

Planta = pd.read\_csv(file\_path, encoding='latin1')

# Muestra las primeras filas del DataFrame

Planta.head()

# Se verificar valores unicos de pedidos en servicio tecnico

S\_tecnico['Pedido N°'].unique()

# Reemplazar el valor '28508;28509' por '28508' en la columna 'Pedido N°' para quedarse con una unico valor debido a

# que es el unico con valor diferente

S\_tecnico['Pedido N°'] = S\_tecnico['Pedido N°'].replace('28508;28509', '28508')

S\_tecnico['Pedido N°'] = S\_tecnico['Pedido N°'].replace('27380, 26425', '27380')

# Verificar el cambio

print(S\_tecnico['Pedido N°'].unique())

# se verifica tamaño

S\_tecnico.shape

# Función para limpiar los valores de la columna

def limpiar\_valor(valor):

    if isinstance(valor, str):  # Solo aplica las transformaciones si el valor es de tipo str

        valor = valor.strip()

        valor = valor.replace(',', '')

    return valor

# Aplica la función de limpieza a la columna 'Pedido N°'

S\_tecnico['Pedido N°'] = S\_tecnico['Pedido N°'].apply(limpiar\_valor)

# Ahora intenta convertir a float64

S\_tecnico['Pedido N°'] = pd.to\_numeric(S\_tecnico['Pedido N°'], errors='coerce')

# Verifica el nuevo tipo de la columna

print(S\_tecnico['Pedido N°'].dtype)

# Muestra los primeros datos para verificar el cambio

print(S\_tecnico.head())

# Se veririca tipo de variable

S\_tecnico.info()

print("\nNúmero de datos faltantes por columna:")

print(S\_tecnico.isnull().sum())

# Se verifica tipo de variable

Planta.info()

# Reemplaza 'ConsecPedido' con el nombre exacto de la columna en el dataframe Planta

consec\_pedido\_col = 'ConsecPedido'

# Realiza la comparación y actualización

for index, row in S\_tecnico.iterrows():

    pedido\_num = row['Pedido N°']

    # Verifica si el valor de "Pedido N°" está en "ConsecPedido" de Planta

    if pedido\_num in Planta[consec\_pedido\_col].values:

        # Encuentra las filas correspondientes en Planta

        planta\_indices = Planta[Planta[consec\_pedido\_col] == pedido\_num].index

        for i in planta\_indices:

            # Copia todas las columnas de S\_tecnico a Planta

            for col in S\_tecnico.columns:

                Planta.at[i, col] = row[col]

    else:

        print(f"No hay coincidencias para Pedido N°: {pedido\_num}")

# Se verifica el tamañao del nuevo dataframe

Planta.head(10)

# Se verifica las columnas del nuevo dataframe

Planta.columns

#Se verifica el tamaño del DF

Planta.shape

# Mostrar las primeras filas del DataFrame para asegurarse de que se ha cargado correctamente

# Número de registros y columnas

print("\nNúmero de registros y columnas:")

print(Planta.shape)

# Tipos de variables para cada columna

print("\nTipos de variables para cada columna:")

print(Planta.dtypes)

# Número de datos faltantes por columna

print("\nNúmero de datos faltantes por columna:")

print(Planta.isnull().sum())

# Generación de información de datos por dataframe

# Mostrar las primeras filas del DataFrame para asegurarse de que se ha cargado correctamente

primeras\_filas = Planta.head()

# Número de registros y columnas

num\_registros\_columnas = pd.DataFrame({

    'Descripción': ['Número de registros', 'Número de columnas'],

    'Valor': [Planta.shape[0], Planta.shape[1]]

})

# Tipos de variables para cada columna

tipos\_de\_variables = Planta.dtypes.reset\_index()

tipos\_de\_variables.columns = ['Columna', 'Tipo de Dato']

# Número de datos faltantes por columna

datos\_faltantes = Planta.isnull().sum().reset\_index()

datos\_faltantes.columns = ['Columna', 'Datos Faltantes']

# Crear un ExcelWriter para guardar múltiples DataFrames en un solo archivo Excel

with pd.ExcelWriter('informe\_analisis\_planta.xlsx', engine='openpyxl') as writer:

    # Guardar las primeras filas del DataFrame

    primeras\_filas.to\_excel(writer, sheet\_name='Primeras Filas', index=False)

    # Guardar el número de registros y columnas

    num\_registros\_columnas.to\_excel(writer, sheet\_name='Registros y Columnas', index=False)

    # Guardar los tipos de variables en una hoja separada

    tipos\_de\_variables.to\_excel(writer, sheet\_name='Tipos de Variables', index=False)

    # Guardar los datos faltantes en una hoja separada

    datos\_faltantes.to\_excel(writer, sheet\_name='Datos Faltantes', index=False)

print("Informe generado y guardado como 'informe\_analisis\_planta.xlsx'")

# Observamos cuantos valores unicos existen en planta

Valores\_unicos=Planta['Pedido N°'].nunique()

print(f"El número de valores únicos en la columna 'Pedido N°' es: {Valores\_unicos}")

# Obtenemos los valores únicos en la columna "Pedido N°"

unique\_values = Planta['Pedido N°'].unique()

print("Los valores únicos en la columna 'Pedido N°' son:")

print(unique\_values)

# Eliminar filas con valores nulos en la columna "Pedido N°"

Planta = Planta.dropna(subset=['Pedido N°'])

# Se vrrifica nuevamente el tamaño

Planta.shape

# dataframe actualizado de planta en un archivo CSV

Planta.to\_csv('Planta\_act.csv', index=False)

# **PRE-PROCESAMIENTO**

En el proceso de Extracción y cargue de archivos se utilizó el software visual studio code para generar un repositorio en línea privado para cargar y analizar la información utilizando lenguaje de programación pyhton y las librerías de análisis de datos pandas, numpy, matplotlib, sweetviz, Entre otras.

En el cargue de información se montaron los archivos en el entorno de programación generando dataframes separados, para luego realizar un preprocesamiento de información para limpiar, organizar y tratar las bases de datos para formar dataframes más amigables de trabajar y unirlos a partir del ConsecPedido. Construyendo así un dataframe con las variables relacionadas en la figura 7, exceptuando las condiciones de operación que aún no están relacionadas porque hay que cargarlas manualmente después de revisar los informes de cotizaciones.

Una vez cruzado los datos de planta virtual con los del informe de servicio técnico se tiene que la base de datos final tiene las siguientes coincidencias:

**Tabla 10**. Numero de registros base de datos cruzados

|  |  |
| --- | --- |
| **Descripción** | **Valor** |
| Número de registros | 1903 |
| Número de columnas | 54 |

**Tabla 11**. Numero de datos faltantes de base de datos cruzados

|  |  |
| --- | --- |
| **Columna** | **Datos Faltantes** |
| ConsecutivoOP | 0 |
| ConsecPedido | 0 |
| Cliente | 0 |
| NombCliente | 0 |
| Producto | 0 |
| NombProducto | 0 |
| FechElab | 0 |
| Proceso | 0 |
| Version | 0 |
| FechEntregaPdcc | 833 |
| FechEntregaClte | 0 |
| OCCliente | 0 |
| QtyRequerida | 0 |
| Actividad | 0 |
| QtyControl | 0 |
| RecMAQ | 0 |
| RecHOM | 0 |
| SecAct | 0 |
| FechEstInicil | 0 |
| estado | 0 |
| TipoConsecutivo | 0 |
| Fecha de visita | 0 |
| Empresa | 0 |
| Planta | 91 |
| País | 0 |
| Ciudad | 0 |
| Tipo de equipo | 0 |
| Orden de compra | 242 |
| Pedido N° | 0 |
| Fecha de instalación de la cadena | 331 |
| Material | 170 |
| Referencia equipo | 206 |
| Causa de desgaste | 211 |
| Si escogió otros, cuál? | 748 |
| Escoja el desgaste donde se presenta | 0 |
| Tipo de cadena | 0 |
| Referencia cadena | 0 |
| Medida referencia de control desgaste nueva en mm | 747 |
| Medida referencia control de desgaste actual en mm | 620 |
| Porcentaje de elongación actual % | 285 |
| Porcentaje de elongación máx % | 206 |
| Proyección en horas de operación restantes (H) | 738 |
| Medida sin desgaste en mm | 603 |
| Medida actual en mm1 | 778 |
| Porcentaje de desgaste actual % | 1519 |
| Medida sin desgaste en mm1 | 1205 |
| Medida sin desgaste en mm2 | 1720 |
| Medida actual en mm2 | 1720 |
| Medida actual en mm | 1388 |
| Altura de la platina en mm | 1071 |
| Medida actual de la platina en mm | 1212 |
| Paso en mm | 168 |
| Cantidad de pasos (Muestra) | 528 |
| Recomendaciones | 43 |

De la generación de base de datos final con el cruce de variables se tiene que en teoría se deben cruzar 28 datos de fallas pertenecientes a informes de servicio técnico con los 244637 de planta virtual, pero al momento de realizar esta actividad se generaron 1903 registros, esto debido a que para un ConsecPedido que es el identificador o clave primaria de servicio técnico pueden existir varios registros en plata virtual con el mismo consecutivo, por ejemplo el ConsecPedido *“23996*” tiene una sola coincidencia en el informe de servicio técnico, pero tiene 80 coincidencias en planta virtual, esto se debe a que en planta virtual un ConsecPedido se puede atribuir al equipo, pero también a los componentes del equipo; en el caso del *“23996*” los 80 registros resultan de todos los componentes de ese elemento, resultando que se les otorga el mismo consecutivo para hacerle trazabilidad y saber que todos pertenecen al mismo elemento. Sin embargo, los valores de informe de servicio técnico, especialmente, las elongaciones registradas no aplican para todos los componentes de un equipos, debido a que depende del servicio técnico que se solicite, se relacionan las mediciones a hacer, por ejemplo, en el informe de servicio técnico puede aparecer medidas de elongaciones para cadenas, bujes, platinas, todos o solo uno, por lo tanto para hacer una base de datos efectiva se debe relacionar cada elongación solamente con los componentes que le corresponden. Adicionalmente la situación anteriormente mencionada sustenta el aumento de nulos al momento de hacer el cruce con la base de datos, en especial en la zona de las variables del informe de servicio técnico.

**Trabajos Futuros**

Una vez generada la base de datos con los valores cruzados se debe realizar un posterior filtrado para asignar las elongaciones registradas en el informe de servicio técnico única y exclusivamente para los componentes que lo requieren, es decir, en algunos casos el servicio técnico incluyó mediciones de varios elementos como bujes, cadenas y platinas, entonces se deben asignar esas elongaciones solo a esos componentes pertenecientes dentro de la planta virtual, por lo que se espera que los registros de la base de datos cruzados pasen de 1903 registros a mínimo 28 y máximo 84 registros, como consecuencia de la asignación de la elongación correspondiente solo a los elementos que lo representan. Posteriormente se hará un pre tratamiento de datos, un análisis exploratorio y finalmente empezar con las pruebas del modelo.

**BIBLIOGRAFÍA**

* Quienes somos – Forjas. (n.d.). Retrieved July 30, 2024, from <https://www.forjasbolivar.com/quienes-somos/>
* Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. Revista Ibérica de Sistemas e Tecnologias de Informação, (E28), 586-599.
* Bobadilla, J. (2021). Machine learning y deep learning: usando Python, Scikit y Keras. Ediciones de la U.
* Moreno González-Páramo, L. (2017). Herramientas avanzadas de análisis de datos de aplicación en ingeniería: Machine Learning using Python.