



Instituto Tecnológico y de Estudios Superiores
de Monterrey

Campus Monterrey

Unidad de Formación - TC2004B Análisis de
ciencia de datos (Gpo 102)

NLC Ternium

Rubí Isela Gutierrez López
Ma. Angelina Alarcón Romero

DATA MANIACS (EQUIPO 1):

Fernando Varela Roman	A01425020 Desarrollador Web
Yuu Ricardo Akachi Tanaka	A01351969 CDO
Juan Carlos Sala Pulido	A00832952 Data Engineering
Valeria Edith Lugo Gutiérrez	A00830523 Data Scientist
Pablo Monzón Terrazas	A01562619 PM
Sabrina Carsellé Aldaco	A01368666 Data Scientist
Donnet Emmanuel Hernández Franco	A01352049 Data Engineering
Itzel Yacquelín Beltrán Reyes	A00832700 Diseñador UX

17 de marzo de 2023

Índice

Índice	2
1. Descripción del problema específico	4
1.1. Objetivos del Negocio	4
1.2. Evaluación de situación actual	4
1.3. Solución: Idea central del proyecto	5
1.4. Hipótesis	5
1.5. Objetivos del Proyecto	5
a) Resultados clave	5
1.6. Justificación	6
1.7. Mercado potencial (descripción general)	6
1.8. Identificación de clientes/consumidor y usuarios	6
1.9. Plan de actividades del proyecto	7
2. Comprensión de los datos	8
2.1. Descripción del set de datos	8
3. Preparación de los datos	16
3.1. Limpieza de datos	16
3.2. Transformación de Datos	19
Volviendo a la limpieza de datos:	20
4. Aplicación de técnicas de Modelación	22
4.1. Extracción de características	22
4.2. Explicación breve de los diferentes modelos de aprendizaje e hiperparámetros	22
K Neighbors((KNN)):	22
Random forest	23
SVM (Support Vector Machine)	24
NAIVE BAYES	24
LOGISTIC REGRESSION	25
K-MEANS	26
4.3. Explicación de la metodología para el entrenamiento y prueba.	26
4.4. Generación de Modelos	26
5. Evaluación	30
5.1 Evaluación de resultados: Entender e interpretar los resultados obtenidos, su impacto y utilidad, considerando los criterios de éxito del negocio.	30
5.2 Revisión del proceso: Sumarizar todo el proceso, principales problemas, posibles mejoras.	31
5.3 Impacto social principal	32
5.4 Impacto hacia los Objetivos de Desarrollo Sostenible.	32
6. Despliegue	33

6.1 Descripción del prototipo funcional	33
7. Recomendaciones	34
7.1 Recomendaciones al negocio	34
7.2 Recomendaciones técnicas	35
Referencias	37

1. Descripción del problema específico

Los datos proporcionados por la empresa Ternium no se encuentran ordenados ni estructurados de la manera en la que la empresa socio formadora encuentra de su agrado, son datos en crudo, es decir, son plasmados como datos antes del procesamiento, contando con faltas de ortografía, falta de información, errores de acentuación, valores nulos y renglones sin llenar, lo cual dificulta mucho el entendimiento, el procesamiento y el análisis de estos mismos datos para poder sacar conclusiones o hacer mejoras dentro de la industria.

1.1. Objetivos del Negocio

El objetivo de la empresa con este proyecto de Ciencia de Datos en sus propias palabras es: “*Interpretar y clasificar automáticamente textos compuestos en categorías definidas de un proceso sin tener un set de datos inicialmente clasificados*” En relación con los beneficios esperados se tiene:

- Disminuir la cantidad de trabajo humano que deban de realizar los responsables del manejo de datos dentro de Ternium.
- Mejorar la organización, y por consiguiente la facilidad para analizar los procesos llevados a cabo.
- Preparar la base de datos de Ternium para hacer posible la implementación de tecnologías y métodos como el machine learning, que contribuyan en una mejor toma de decisiones.
- Fortalecer los procesos de segmentación con el fin de la creación de nuevas oportunidades para optimizar procesos.

La meta que se tiene es clasificar de manera idónea los avisos en el proceso de producción. Los criterios de éxito se basan en las métricas de validación de los modelos de Machine Learning, con el fin que se puedan detectar patrones de comportamiento en el proceso de producción con la finalidad de corrección en sus pasos.

1.2. Evaluación de situación actual

Actualmente Ternium no cuenta con un adecuado sistema de recolección de datos, es por esto que la base de datos de fallas de equipo contienen numerosas complicaciones para el entendimiento, procesamiento y análisis de la data, lo que genera que no se puedan tomar

decisiones en función a la erradicación de las fallas en sistema, identificación de zonas, razones, síntomas, entre otros factores posibles. Hoy en día los factores de éxito para la empresa se basan en los competidores, la demanda del mercado y la influencia que se tiene dentro de esta misma, es por esto, que si la empresa no logra adaptarse a los requerimientos para poder tomar decisiones, no se podrá tener un avance en cuestión a los competidores del mercado, debido a que hay mayor probabilidad de fracaso para estos mismos y de éxito para los mencionados anteriormente.

1.3. Solución: Idea central del proyecto

Realizar una propuesta que conlleve un tipo de clasificación de datos, el cual tome principalmente en cuenta las columnas de las variables de datos nombradas **SÍNTOMA** y **CAUSA**, para lo cual será necesario la implementación de métodos de limpieza y transformación de datos, consiguiendo de esta manera, que el modelo de clasificación funcione de la manera más óptima posible.

1.4. Hipótesis

Si realizamos una clasificación eficiente sobre las variables **SÍNTOMA** y **CAUSA** utilizando técnicas de clustering, agrupación de datos, árboles de decisión y algoritmos NLPs, nos ayudarán a reducir el ruido, reducir dimensionalidad y clasificar datos que serán de ayuda para el Socio Formador.

1.5. Objetivos del Proyecto

a) Resultados clave

- Disminuir la cantidad de trabajo humano que deban de realizar los responsables del manejo de datos dentro de Ternium.
- Mejorar la organización, y por consiguiente la facilidad para analizar los procesos llevados a cabo.
- Preparar la base de datos de Ternium para hacer posible la implementación de tecnologías y métodos como el machine learning, que contribuyan en una mejor toma de decisiones.
- Fortalecer los procesos de segmentación con el fin de la creación de nuevas oportunidades para optimizar procesos.

1.6. Justificación

A partir de la realización de diversas investigaciones, las cuales se presentan anteriormente, se identificó que dichos papers contenían información sumamente importante para el principal objetivo del proyecto, es decir, poder encontrar el modelo de Machine Learning óptimo que pudiera clasificar los comentarios descritos en la data set, gracias a esta investigación nos dimos cuenta que aunque hay numerosos métodos, como lo son los árboles de decisión o las redes neuronales, en este caso el que funcionaría de la mejor manera es el NLP, procesamiento de lenguaje natural, el cual pertenece al aprendizaje no supervisado. Con este método, una vez clasificados los comentarios, generará eficiencia en la empresa, así como también, una manera más sencilla de manejar y prevenir las fallas en los equipos.

1.7. Mercado potencial (descripción general)

Ternium es un Centro Industrial productor de aceros, los cuales ofrecen sus productos a millones de mexicanos para el uso en sus hogares. Es por esto, que sus productos y maquinarias pueden llegar a tener algunos errores o defectos. Para esto es que registran miles de datos con los avisos de falla, así como también, llevan el registro de lo que ha salido mal y cómo se podría resolver. Por esa razón es que es de suma importancia tener un modelo capaz de clasificar los datos de la mejor manera, así como también tener los datos preparados y listos para lograr obtener el resultado deseado para los clientes.

Una vez dicho lo anterior, el mercado potencial para este proyecto son las empresas vendedoras de cualquier tipo de producto doméstico que tengan contacto directo con los clientes a los que les ha fallado o han tenido problemas con lo vendido; y que tengan una base de datos sin limpieza y con datos en crudo. De igual manera podría funcionar para empresas pequeñas que tengan la necesidad de mantener una mejor organización y entendimiento de sus datos y mejoras o errores, así como también, para empresas que cuenten con máquinas muy complicadas y tiendan a fallar sus productos.

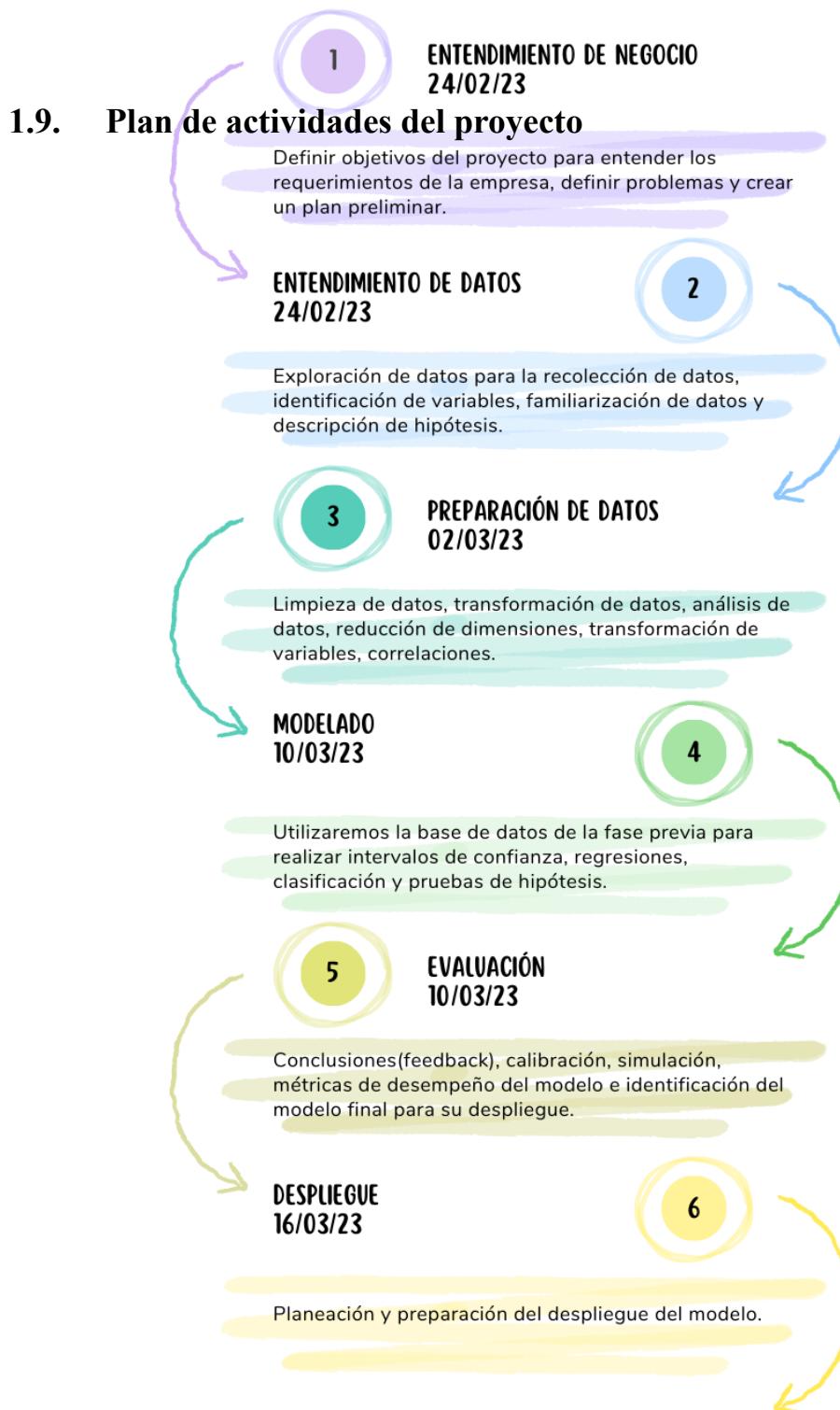
1.8. Identificación de clientes/consumidor y usuarios

- Instituciones que buscan recabar información mediante cuestionarios como son las que realizan censos demográficos.
- Empresas que tengan un trato directo con clientes y tengan encuestas de evaluación.

- Departamentos que se encarguen de dar mantenimiento y solución a casos reportados.

PLANEACIÓN DE ACTIVIDADES

DATA MANIACS



2. Comprensión de los datos

2.1. Descripción del set de datos

Para el desarrollo de esta etapa, en donde se aplicó la limpieza de los datos, se comenzó por comprender la estructura del dataset, así como también el significado de cada registro, para de esta manera poder obtener una idea amplia de lo que se debía hacer y cómo.

Para empezar, con ayuda de funciones de la librería **pandas** como lo fue `.shape`, se identificó que el data frame creado contenía 64,719 registros clasificados en 16 variables diferentes, para lo cual, con `.info()` se verificó que únicamente se contaba con una variable numérica, una booleana y todas las demás eran categóricas. A partir de esto se utilizó otra función llamada `.isnull()` para obtener los valores nulos en cada columna del data frame, lo cual más adelante nos ayudó para la limpieza de los datos y la selección de variables a utilizar.

Para la descripción de los datos se utilizó el método de visualizar los primeros 5 registros y de igual manera se logró identificar con `.unique()` todos los posibles valores que podía tomar cierta variable. A continuación se muestra un resumen de la descripción de las variables, su tipo, valores nulos y posibles valores a tomar:

- **Aviso:** Es una variable categórica que corresponde al número de identificación del aviso, toma valores enteros entre 56103947 y 57143776, además, no cuenta con valores nulos.
- **Descripción:** Es una variable categórica que contiene la descripción del problema presentado en la fábrica, el input de esta variable es ingresado por los trabajadores. Tiene un valor nulo, y al ser un string puede contener cualquier valor del mismo tipo.
- **Qué pasó?:** Variable categórica que describe lo sucedido (lo que pasó) con la problemática, tiene 30402 posibles valores y tiene 3 valores nulos.
- **Por qué pasó?:** Variable categórica que explica la razón de lo sucedido con la problemática, cuenta con 41345 posibles valores, tiene 3 valores nulos.
- **Qué se hizo?:** Variable categórica que responde a lo que se hizo para responder el problema, y cuenta con 43735 posibles valores, tiene 3 valores nulos.
- **Parte Objeto:** Variable categórica que menciona la parte del objeto que falló. Esta variable tiene 632 posibles valores, no cuenta con valores nulos.
- **Síntoma:** Variable categórica que explica la manera en que se dieron cuenta que la máquina estaba fallando, no tiene valores nulos y tiene 51 valores posibles.
- **Texto Síntoma:** Variable categórica con la descripción detallada del síntoma, tiene 40124 valores nulos y tiene 15520 posibles valores.
- **Causa:** Variable categórica que explica la razón que provocó el problema, tiene 0 valores nulos y 60 valores posibles.
- **Equipo:** Variable categórica que muestra el número de identificación del equipo que tuvo problema, tiene 4317 valores nulos y 7967 posibles valores.
- **Denominación:** Se refiere a la variable categórica que define el nombre del equipo, tiene 4317 valores nulos y 6764 posibles valores.
- **Parada:** Variable categórica booleana (Dummy) que identifica si el equipo se paró (X) o no. Tiene 26233 valores nulos que se refiere a que el equipo no se paró y dos valores posibles.
- **Duración parada:** Variable numérica que indica el tiempo en que el equipo se paró, no tiene valores nulos y tiene 565 posibles valores.
- **Inicio avería:** Variable categórica que contiene la fecha del inicio de la avería. No tiene valores nulos y cuenta con 750 posibles valores.

Imagen 1: Descripción de los datos

Para la exploración de los datos se comenzó con la única variable cuantitativa, la cual corresponde a “Duración parada”, sin embargo, al no ser una variable de interés para el proyecto a realizar, únicamente se obtuvo un resumen de los estadísticos más importantes utilizando la función `.describe()`.

```
count      64719.000000
mean        3.169491
std         142.704152
min         0.000000
25%        0.000000
50%        0.060000
75%        0.150000
max        6991.500000
Name: Duración parada, dtype: float64
```

Imagen 2: Análisis descriptivo de la variable numérica Duración parada

Por otro lado, como el resto de las variables eran cualitativas, se realizaron tablas de frecuencia para cada una de las variables, así como también, se encontraron las medias de cada una puesto que eso ayudó a obtener información valiosa del comportamiento de estas mismas. Las tablas de frecuencia fueron realizadas con la función `crosstab()`.

Descripción	¿Qué pasó?																																																																								
<p>col_0 count</p> <table border="1"> <thead> <tr> <th>Descripción</th> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr> <td># DE REPORTE 47892,INHIBIRSE PANTALLAS D</td> <td>2</td> <td></td> </tr> <tr> <td>(falla bomba rectificadores)</td> <td></td> <td>1</td> </tr> <tr> <td>(se apaga pantalla de i2 y de coronas)</td> <td></td> <td>8</td> </tr> <tr> <td>*(NO PODER CERRAR GUIAS PARA CENTRAR LAM</td> <td>1</td> <td></td> </tr> <tr> <td>*PONER PERMISIVO PARA NO SACAR GUIAS X A</td> <td>2</td> <td></td> </tr> <tr> <td>...</td> <td>...</td> <td></td> </tr> <tr> <td>zonal alarma hid.de baja</td> <td></td> <td>1</td> </tr> <tr> <td>zonal alarma hid.de media</td> <td></td> <td>1</td> </tr> <tr> <td>zonal larma del oil mist</td> <td></td> <td>1</td> </tr> <tr> <td> </td> <td></td> <td>1</td> </tr> <tr> <td>'pruebas RTM</td> <td></td> <td>1</td> </tr> </tbody> </table> <p>39716 rows × 1 columns</p>	Descripción	col_0	count	# DE REPORTE 47892,INHIBIRSE PANTALLAS D	2		(falla bomba rectificadores)		1	(se apaga pantalla de i2 y de coronas)		8	*(NO PODER CERRAR GUIAS PARA CENTRAR LAM	1		*PONER PERMISIVO PARA NO SACAR GUIAS X A	2			zonal alarma hid.de baja		1	zonal alarma hid.de media		1	zonal larma del oil mist		1			1	'pruebas RTM		1	<p>col_0 count</p> <table border="1"> <thead> <tr> <th>Qué pasó?</th> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr> <td>-NO ENTRAR TENSIÓN PROCESO</td> <td></td> <td>1</td> </tr> <tr> <td>-SE CONTINUA CON EL SEPARADOR MAGNETICO.</td> <td></td> <td>1</td> </tr> <tr> <td>.</td> <td></td> <td>1</td> </tr> <tr> <td>.REPORTAN CAMA DE SALIDA BOTADO</td> <td></td> <td>1</td> </tr> <tr> <td>.SE DAÑÓ BASE DE MOTRO DE GIRO DE CHAROLA</td> <td></td> <td>1</td> </tr> <tr> <td>...</td> <td></td> <td>...</td> </tr> <tr> <td>'PARO LINEA SECCION SALIDA</td> <td></td> <td>1</td> </tr> <tr> <td>'PIDEN CAMBIO DE VIAS.</td> <td></td> <td>1</td> </tr> <tr> <td>'REPORTAN NO SALIR TERCER APOYO</td> <td></td> <td>1</td> </tr> <tr> <td>?DESENROLLADOR NO ALIMENTA MATERIAL A LINEA</td> <td></td> <td>1</td> </tr> <tr> <td>ÁRA MLOINO.</td> <td></td> <td>1</td> </tr> </tbody> </table> <p>39401 rows × 1 columns</p>	Qué pasó?	col_0	count	-NO ENTRAR TENSIÓN PROCESO		1	-SE CONTINUA CON EL SEPARADOR MAGNETICO.		1	.		1	.REPORTAN CAMA DE SALIDA BOTADO		1	.SE DAÑÓ BASE DE MOTRO DE GIRO DE CHAROLA		1	'PARO LINEA SECCION SALIDA		1	'PIDEN CAMBIO DE VIAS.		1	'REPORTAN NO SALIR TERCER APOYO		1	?DESENROLLADOR NO ALIMENTA MATERIAL A LINEA		1	ÁRA MLOINO.		1
Descripción	col_0	count																																																																							
# DE REPORTE 47892,INHIBIRSE PANTALLAS D	2																																																																								
(falla bomba rectificadores)		1																																																																							
(se apaga pantalla de i2 y de coronas)		8																																																																							
*(NO PODER CERRAR GUIAS PARA CENTRAR LAM	1																																																																								
*PONER PERMISIVO PARA NO SACAR GUIAS X A	2																																																																								
...	...																																																																								
zonal alarma hid.de baja		1																																																																							
zonal alarma hid.de media		1																																																																							
zonal larma del oil mist		1																																																																							
		1																																																																							
'pruebas RTM		1																																																																							
Qué pasó?	col_0	count																																																																							
-NO ENTRAR TENSIÓN PROCESO		1																																																																							
-SE CONTINUA CON EL SEPARADOR MAGNETICO.		1																																																																							
.		1																																																																							
.REPORTAN CAMA DE SALIDA BOTADO		1																																																																							
.SE DAÑÓ BASE DE MOTRO DE GIRO DE CHAROLA		1																																																																							
...		...																																																																							
'PARO LINEA SECCION SALIDA		1																																																																							
'PIDEN CAMBIO DE VIAS.		1																																																																							
'REPORTAN NO SALIR TERCER APOYO		1																																																																							
?DESENROLLADOR NO ALIMENTA MATERIAL A LINEA		1																																																																							
ÁRA MLOINO.		1																																																																							

¿Por qué pasó?		¿Qué se hizo?	
	col_0 count		col_0 count
Por qué pasó?		Qué se hizo?	
" DESAJUSTE EN V DE PRIMER BRAZO "	1	" SE RESTABLECE" NO ME HABLARON	1
"O" RING DAÑADO	1	"REPARTA OPERADOR DEL DANIELI TENER ABIERTOS LOS INTERRUPORE	2
"O"RING DAÑADO	1	#1-SE LE HIZO ORIFICIOS MAS HOLGUEADOS A TENEDOR GUIA DE TU-	1
"O"RING DAÑADO DEL REGULADOR DE PRESION, TAPA DE VALVULA HUNT	1	* SE AJUSTAN CUENTAS DEL HIDRAULICO SALIDA.	1
"V" DESALINEADA	1	* SE APoyo EN BLOQUEOS PARA LA CALIBRACION DE LA TIJERA.	1
...
'POR TENER BAJO NIVEL EN DEPOSITO DE ACEITE	1	ya libre el reel se expande y se aplica grasa entre mordazas para que	2
'por rechuparse rollo en rel ote	1	OPERADOR REPORTA QUE NO TIENE PERMISO DE CONEXIÓN , SE COME	1
ÁREA CONTAMINADA,FUGA DE GRASA EN BOMBAS	1	REPORTA OPERADOR ALARMA DE BAJO CAUDAL EN RETORNO DE CENTRO	1
ÁREA SE DETECTA QUE MOTOR DIÉSEL ESTÁ APAGADO Y ESTA BOTADO	1	REPORTE DE FALLA ALARMA BAJO FLUJO EN PANEL S3 , SE REVISA	1
ESTO SE DEBE A QUE OPTO DE SALIDA DIGITAL DEL EQUIPO ESTA CRU	1	'SE MOVIO MANUALMENTE	1
41344 rows x 1 columns			

Tabla 1: Tablas de frecuencia de 4 variables categóricas

Para las variables Descripción, Qué pasó?, Por qué pasó? y Qué se hizo?, se observa en la Tabla 1, que se tratan de tablas de frecuencias en las cuales sí existen casos en la que la entrada en dichas variables es la misma, lo cual se puede verificar al existir un menor número de renglones en las tablas de frecuencia que en el data frame original, así como también se pueden ver casos con frecuencias distintas a 1, sin embargo, son mínimas los textos repetidos. Esto ocurre gracias a que se trata de variables con una entrada libre de texto, posteriormente, se verificará si no se tratan de registros duplicados dentro de la base de datos, lo cual representaría una inconsistencia gracias a que concuerda la cantidad de valores únicos de la variable “Avisos” (que es una ID única para cada caso), con la cantidad de registros existentes en el Data Frame.

Seguimos analizando las tablas de frecuencia de las demás variables:

Parte Objeto		Síntoma	
	col_0 count		Síntoma
Parte Objeto			
ABANICO	174	A TIERRA	311
ABRAZADERA	872	ABIERTO	8008
ACEITE	1129	AFLOJAMIENTO	6443
ACOPLADOR	11	ALTA TEMPERATURA	281
ACOPLAMIENTO	1727	ALTA VIBRACION	132
...	...	AMARRADO	104
Valvula Reguladora (pres/caudal)	26	ANTIVIRUS	1
Valvula Retención/Alivio	12	ATORADO	4594
Valvula de amortiguación/caudal	21	Automático por defecto	5536
Valvula de cierre/Shut off	27	BAJA RESISTENCIA	215
Ventiladores de Drives	2	BLOQUEADO	104
631 rows x 1 columns			

<p>Se puede ver que se disminuyó la dimensión de un poco más de 60k registros a menos de mil, lo cual permite observar aquellas en las cuales se presentan mayores problemas, así como de igual manera revisar la existencia de fallas a la hora del registro de los datos.</p>	<p>Se disminuyó la dimensión del DataFrame original a una cantidad mucho menor, lo cual a su vez permite detectar los principales síntomas existentes, así como de igual forma revisar la existencia de fallas a la hora del registro de los datos. Esto nos dice que se repitieron varios síntomas.</p>																																																						
<p>Texto Síntoma</p> <table border="1" data-bbox="314 707 659 1111"> <thead> <tr> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr> <td>Texto Síntoma</td> <td>15519</td> </tr> <tr> <td>(OL) ABIERTO</td> <td>1</td> </tr> <tr> <td>.</td> <td>1</td> </tr> <tr> <td>. se ajusta sensor reset coche 2</td> <td>1</td> </tr> <tr> <td>. se ajusta sensor rodillo tijera</td> <td>1</td> </tr> <tr> <td>..P.T.A. molino w200 volteador</td> <td>1</td> </tr> <tr> <td>...</td> <td>...</td> </tr> <tr> <td>'Reportan no salir tercer apoyo</td> <td>1</td> </tr> <tr> <td>'paro linea sección salida</td> <td>1</td> </tr> <tr> <td>'piden apoyo para hacer cambio de vía</td> <td>1</td> </tr> <tr> <td>NO EN BOTONERA CARRO SUR CENTRADO</td> <td>1</td> </tr> <tr> <td>óperador pide sacar muestra</td> <td>1</td> </tr> </tbody> </table> <p>15519 rows × 1 columns</p> <p>Existe una menor cantidad de registros en la tabla de frecuencias al analizar la variable Texto Síntoma, pero esto se puede presentar de igual forma gracias a que cuenta con una gran cantidad de datos nulos, lo cual no se podría indicar que se trata de un espacio opcional de llenar a la hora de registrar un aviso, por lo cual no se considera demasiado relevante esta variable.</p>	col_0	count	Texto Síntoma	15519	(OL) ABIERTO	1	.	1	. se ajusta sensor reset coche 2	1	. se ajusta sensor rodillo tijera	1	..P.T.A. molino w200 volteador	1	'Reportan no salir tercer apoyo	1	'paro linea sección salida	1	'piden apoyo para hacer cambio de vía	1	NO EN BOTONERA CARRO SUR CENTRADO	1	óperador pide sacar muestra	1	<p>Causa</p> <table border="1" data-bbox="966 707 1232 1111"> <thead> <tr> <th>Causa</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>ACIDEZ</td> <td>42</td> </tr> <tr> <td>AFLOJAMIENTO</td> <td>2202</td> </tr> <tr> <td>ALTA TEMPERATURA</td> <td>610</td> </tr> <tr> <td>ATERRIZADO</td> <td>7</td> </tr> <tr> <td>Ajeno al proveedor</td> <td>3</td> </tr> <tr> <td>Automático por defecto</td> <td>5569</td> </tr> <tr> <td>BAJO NIVEL ACEITE</td> <td>1365</td> </tr> <tr> <td>BLOQUEO DE PUESRTOS</td> <td>3</td> </tr> <tr> <td>Beschaffung</td> <td>1</td> </tr> <tr> <td>CAVITACION</td> <td>89</td> </tr> <tr> <td>CONTAMINACION</td> <td>4552</td> </tr> <tr> <td>CORTE DE ENERGIA</td> <td>296</td> </tr> <tr> <td>CRUZADO</td> <td>10</td> </tr> </tbody> </table> <p>Se observa una disminución en la cantidad de renglones en la tabla de frecuencias y se trata de una variable que no cuenta con datos nulos. Lo cual indica que es necesario el llenado de este campo para dar de alta un registro. Observar la tabla de frecuencias será de mucha utilidad a la hora de revisar patrones de comportamientos y para conocer cuáles son las principales causas de las problemáticas.</p>	Causa	Count	ACIDEZ	42	AFLOJAMIENTO	2202	ALTA TEMPERATURA	610	ATERRIZADO	7	Ajeno al proveedor	3	Automático por defecto	5569	BAJO NIVEL ACEITE	1365	BLOQUEO DE PUESRTOS	3	Beschaffung	1	CAVITACION	89	CONTAMINACION	4552	CORTE DE ENERGIA	296	CRUZADO	10
col_0	count																																																						
Texto Síntoma	15519																																																						
(OL) ABIERTO	1																																																						
.	1																																																						
. se ajusta sensor reset coche 2	1																																																						
. se ajusta sensor rodillo tijera	1																																																						
..P.T.A. molino w200 volteador	1																																																						
...	...																																																						
'Reportan no salir tercer apoyo	1																																																						
'paro linea sección salida	1																																																						
'piden apoyo para hacer cambio de vía	1																																																						
NO EN BOTONERA CARRO SUR CENTRADO	1																																																						
óperador pide sacar muestra	1																																																						
Causa	Count																																																						
ACIDEZ	42																																																						
AFLOJAMIENTO	2202																																																						
ALTA TEMPERATURA	610																																																						
ATERRIZADO	7																																																						
Ajeno al proveedor	3																																																						
Automático por defecto	5569																																																						
BAJO NIVEL ACEITE	1365																																																						
BLOQUEO DE PUESRTOS	3																																																						
Beschaffung	1																																																						
CAVITACION	89																																																						
CONTAMINACION	4552																																																						
CORTE DE ENERGIA	296																																																						
CRUZADO	10																																																						

Tabla 2: Tablas de frecuencias de 4 variables categóricas

<p>Equipo</p> <table border="1"> <thead> <tr> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr><td>Equipo</td><td></td></tr> <tr><td>21000000.0</td><td>2</td></tr> <tr><td>21000001.0</td><td>2</td></tr> <tr><td>21000010.0</td><td>3</td></tr> <tr><td>21000012.0</td><td>3</td></tr> <tr><td>21000013.0</td><td>5</td></tr> <tr><td>...</td><td>...</td></tr> <tr><td>46000886.0</td><td>1</td></tr> <tr><td>46001011.0</td><td>2</td></tr> <tr><td>46001070.0</td><td>8</td></tr> <tr><td>46001071.0</td><td>3</td></tr> <tr><td>46001072.0</td><td>9</td></tr> </tbody> </table> <p>7966 rows × 1 columns</p>	col_0	count	Equipo		21000000.0	2	21000001.0	2	21000010.0	3	21000012.0	3	21000013.0	5	46000886.0	1	46001011.0	2	46001070.0	8	46001071.0	3	46001072.0	9	<p>Denominación</p> <table border="1"> <thead> <tr> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr><td>Denominación</td><td></td></tr> <tr><td>013-F TOLVA ALMACEN MINERAL</td><td>5</td></tr> <tr><td>014-F1 ALMEA DUCTO DE CARGA</td><td>10</td></tr> <tr><td>014-F2 ALMEA DUCTO DE CARGA</td><td>14</td></tr> <tr><td>085-V BANDA DE TRANSPORTADOR</td><td>36</td></tr> <tr><td>085-V M MOTOR OTE</td><td>8</td></tr> <tr><td>...</td><td>...</td></tr> <tr><td>ZSHL63711 - POSICION RE-637-L1</td><td>1</td></tr> <tr><td>ZSHL67421 - POSICION RE-674-L2</td><td>1</td></tr> <tr><td>ZSHL67521 - POSICION RE-675-L2</td><td>1</td></tr> <tr><td>ZYH61531 - SOLENOIDE APERTURA RE-615-L3</td><td>1</td></tr> <tr><td>tLOCKS DE VALVULAS BENDING MOLINO 5</td><td>1</td></tr> </tbody> </table> <p>6763 rows × 1 columns</p>	col_0	count	Denominación		013-F TOLVA ALMACEN MINERAL	5	014-F1 ALMEA DUCTO DE CARGA	10	014-F2 ALMEA DUCTO DE CARGA	14	085-V BANDA DE TRANSPORTADOR	36	085-V M MOTOR OTE	8	ZSHL63711 - POSICION RE-637-L1	1	ZSHL67421 - POSICION RE-674-L2	1	ZSHL67521 - POSICION RE-675-L2	1	ZYH61531 - SOLENOIDE APERTURA RE-615-L3	1	tLOCKS DE VALVULAS BENDING MOLINO 5	1
col_0	count																																																				
Equipo																																																					
21000000.0	2																																																				
21000001.0	2																																																				
21000010.0	3																																																				
21000012.0	3																																																				
21000013.0	5																																																				
...	...																																																				
46000886.0	1																																																				
46001011.0	2																																																				
46001070.0	8																																																				
46001071.0	3																																																				
46001072.0	9																																																				
col_0	count																																																				
Denominación																																																					
013-F TOLVA ALMACEN MINERAL	5																																																				
014-F1 ALMEA DUCTO DE CARGA	10																																																				
014-F2 ALMEA DUCTO DE CARGA	14																																																				
085-V BANDA DE TRANSPORTADOR	36																																																				
085-V M MOTOR OTE	8																																																				
...	...																																																				
ZSHL63711 - POSICION RE-637-L1	1																																																				
ZSHL67421 - POSICION RE-674-L2	1																																																				
ZSHL67521 - POSICION RE-675-L2	1																																																				
ZYH61531 - SOLENOIDE APERTURA RE-615-L3	1																																																				
tLOCKS DE VALVULAS BENDING MOLINO 5	1																																																				
<p>Denominación.1</p> <table border="1"> <thead> <tr> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr><td>Denominación.1</td><td></td></tr> <tr><td>601-V1 TUB DE 601-V A 735-L1 (TN-40102)</td><td>2</td></tr> <tr><td>601-V2 TUB DE 601-V A 735-L2 (TN-40105)</td><td>2</td></tr> <tr><td>633-J - COMPRESOR ENFRRIAMIENTO EXTERNO</td><td>1</td></tr> <tr><td>701-V1 TUB DE 735-L1 A 736-L1 (TN-40103)</td><td>1</td></tr> <tr><td>702-V1 TUB DE 736-L1 A 754-F (TN-40104)</td><td>1</td></tr> <tr><td>...</td><td>...</td></tr> <tr><td>WALKING BEAM</td><td>14</td></tr> <tr><td>ZONA DE ENFRRIAMIENTO</td><td>3</td></tr> <tr><td>ZONA DE ENTRADA</td><td>1</td></tr> <tr><td>ZONA ENFRRIAMIENTO</td><td>3</td></tr> <tr><td>ZONA MUERTA</td><td>3</td></tr> </tbody> </table> <p>2007 rows × 1 columns</p>	col_0	count	Denominación.1		601-V1 TUB DE 601-V A 735-L1 (TN-40102)	2	601-V2 TUB DE 601-V A 735-L2 (TN-40105)	2	633-J - COMPRESOR ENFRRIAMIENTO EXTERNO	1	701-V1 TUB DE 735-L1 A 736-L1 (TN-40103)	1	702-V1 TUB DE 736-L1 A 754-F (TN-40104)	1	WALKING BEAM	14	ZONA DE ENFRRIAMIENTO	3	ZONA DE ENTRADA	1	ZONA ENFRRIAMIENTO	3	ZONA MUERTA	3	<p>Ubicac.técnica</p> <table border="1"> <thead> <tr> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr><td>Ubicac.técnica</td><td></td></tr> <tr><td>HM-ACE-ACHF</td><td>2</td></tr> <tr><td>HM-ACE-ACHF-GRUA-GNC2</td><td>4</td></tr> <tr><td>HM-ACE-ACHF-GRUA-GNC2-CAR-EST</td><td>1</td></tr> <tr><td>HM-ACE-ACHF-GRUA-GNC2-CAR-TRL</td><td>2</td></tr> <tr><td>HM-ACE-ACHF-GRUA-GNC2-CAR-TTC</td><td>39</td></tr> <tr><td>...</td><td>...</td></tr> <tr><td>HM-TUB-SERP-ALUM-AIPT</td><td>1</td></tr> <tr><td>HM-TUB-SERP-ALUM-ALEX</td><td>1</td></tr> <tr><td>HM-TUB-SERP-TORR-T200</td><td>2</td></tr> <tr><td>HM-TUB-SERP-TORR-T3Y6</td><td>1</td></tr> <tr><td>HM-TUB-SERP-TORR-TW35</td><td>3</td></tr> </tbody> </table> <p>2545 rows × 1 columns</p>	col_0	count	Ubicac.técnica		HM-ACE-ACHF	2	HM-ACE-ACHF-GRUA-GNC2	4	HM-ACE-ACHF-GRUA-GNC2-CAR-EST	1	HM-ACE-ACHF-GRUA-GNC2-CAR-TRL	2	HM-ACE-ACHF-GRUA-GNC2-CAR-TTC	39	HM-TUB-SERP-ALUM-AIPT	1	HM-TUB-SERP-ALUM-ALEX	1	HM-TUB-SERP-TORR-T200	2	HM-TUB-SERP-TORR-T3Y6	1	HM-TUB-SERP-TORR-TW35	3
col_0	count																																																				
Denominación.1																																																					
601-V1 TUB DE 601-V A 735-L1 (TN-40102)	2																																																				
601-V2 TUB DE 601-V A 735-L2 (TN-40105)	2																																																				
633-J - COMPRESOR ENFRRIAMIENTO EXTERNO	1																																																				
701-V1 TUB DE 735-L1 A 736-L1 (TN-40103)	1																																																				
702-V1 TUB DE 736-L1 A 754-F (TN-40104)	1																																																				
...	...																																																				
WALKING BEAM	14																																																				
ZONA DE ENFRRIAMIENTO	3																																																				
ZONA DE ENTRADA	1																																																				
ZONA ENFRRIAMIENTO	3																																																				
ZONA MUERTA	3																																																				
col_0	count																																																				
Ubicac.técnica																																																					
HM-ACE-ACHF	2																																																				
HM-ACE-ACHF-GRUA-GNC2	4																																																				
HM-ACE-ACHF-GRUA-GNC2-CAR-EST	1																																																				
HM-ACE-ACHF-GRUA-GNC2-CAR-TRL	2																																																				
HM-ACE-ACHF-GRUA-GNC2-CAR-TTC	39																																																				
...	...																																																				
HM-TUB-SERP-ALUM-AIPT	1																																																				
HM-TUB-SERP-ALUM-ALEX	1																																																				
HM-TUB-SERP-TORR-T200	2																																																				
HM-TUB-SERP-TORR-T3Y6	1																																																				
HM-TUB-SERP-TORR-TW35	3																																																				
<p>Parada</p> <p>Parada</p> <table border="1"> <tbody> <tr><td>X</td><td>38486</td></tr> </tbody> </table>	X	38486	<p>Inicio avería</p> <table border="1"> <thead> <tr> <th>col_0</th> <th>count</th> </tr> </thead> <tbody> <tr><td>Inicio avería</td><td></td></tr> <tr><td>1/1/2021</td><td>44</td></tr> <tr><td>1/1/2022</td><td>35</td></tr> <tr><td>1/1/2023</td><td>55</td></tr> <tr><td>1/10/2021</td><td>46</td></tr> <tr><td>1/10/2022</td><td>70</td></tr> <tr><td>...</td><td>...</td></tr> <tr><td>9/7/2022</td><td>85</td></tr> <tr><td>9/8/2021</td><td>76</td></tr> <tr><td>9/8/2022</td><td>106</td></tr> <tr><td>9/9/2021</td><td>82</td></tr> <tr><td>9/9/2022</td><td>113</td></tr> </tbody> </table> <p>750 rows × 1 columns</p>	col_0	count	Inicio avería		1/1/2021	44	1/1/2022	35	1/1/2023	55	1/10/2021	46	1/10/2022	70	9/7/2022	85	9/8/2021	76	9/8/2022	106	9/9/2021	82	9/9/2022	113																								
X	38486																																																				
col_0	count																																																				
Inicio avería																																																					
1/1/2021	44																																																				
1/1/2022	35																																																				
1/1/2023	55																																																				
1/10/2021	46																																																				
1/10/2022	70																																																				
...	...																																																				
9/7/2022	85																																																				
9/8/2021	76																																																				
9/8/2022	106																																																				
9/9/2021	82																																																				
9/9/2022	113																																																				

Tabla 3: Tablas de frecuencias de 6 variables

Las anteriores variables a las cuales se sacó la frecuencia, permiten obtener información extra a la hora de revisar un aviso en específico de la empresa Ternium, pero de acuerdo con la empresa Ternium, no son relevantes en el proyecto de Ciencia de Datos que se está realizando actualmente.

Para obtener las modas de las variables se utilizó `.mode()`, para de esta manera ver el comportamiento de las variables y verificar cuál era la que más se repetía.

Variable	Moda
Descripción	ATORON OPERATIVO
Qué pasó?	VINCULOS
Por qué pasó?	VINCULOS
Qué se hizo?	VINCULOS
Parte Objeto	Automático por defecto
Síntoma	PROTECCION
Texto Síntoma	HORNO DEMORADO
Causa	FATIGA
Equipo	26009610.0
Denominación	MENTOR # 2 TRANSP ROLLOS - 15HP CA
Ubicac.Técnica	HM-SER-TACE-EMOV-LCOM
Denominación.1	LOCOMOTORAS
Parada	X
Duración Parada	0.0
Inicio Avería	2/7/2021

Tabla 4: Modas de las variables

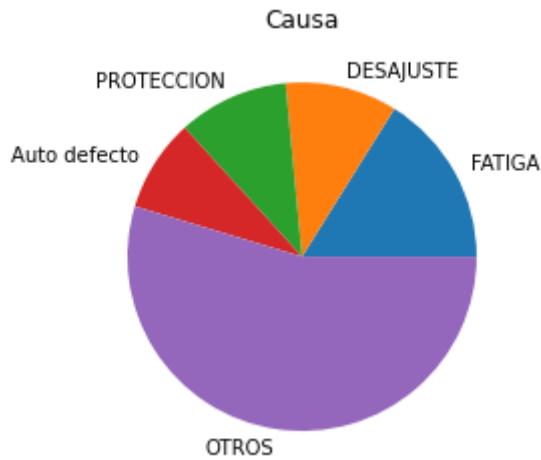
Como resumen de la Tabla 4, se observa que existe una mayor cantidad de avisos debido a "Atorones operativos", y al observar el dataframe, hay una relación con el valor de vínculos dentro de las variables de entrada libre de texto, como lo son las variables "Qué pasó?", "Porqué pasó?" y "Qué se hizo?". De igual manera se pudo detectar que la mayor cantidad de avisos se dio dentro de la fecha: "2/7/2021". Conocer la moda puede ser de gran utilidad para el Socio Formador ya que le da información de las problemáticas más comunes y empezar a corregirlas.

Para la visualización de los datos se seleccionaron dos variables que fueran lo suficientemente explicativas para el objetivo a alcanzar, en este caso fueron las variables

categóricas “Síntoma” y “Causa”, esto debido a que cuentan con la menor cantidad de valores posibles a tomar, ya que las otras variables categóricas son de entrada libre de texto lo cual no nos serviría visualizar sus frecuencias en gráficas. Esto permitirá una mejor visualización con las herramientas aplicadas. Además, en la plática con Ternium se mencionó que esas dos variables eran de suma importancia y fundamentales para el proyecto de Ciencias de Datos.

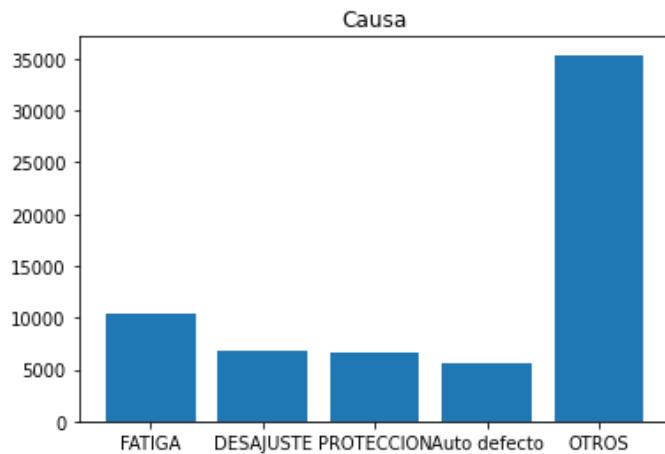
Para esto se realizaron gráficas de pastel para obtener las causas y los síntomas más repetidos, así como también, para poder obtener la duración de parada más común y los días en que se presentaron más avisos.

Para la variable “Causa” se tomaron las primeras 4 causas de una tabla de frecuencias, siendo “PROTECCION”, “DESAJUSTE”, “FATIGA” y “AUTO DEFECTO”, las demás se catalogaron como “Otras”, obteniendo la siguiente gráfica.



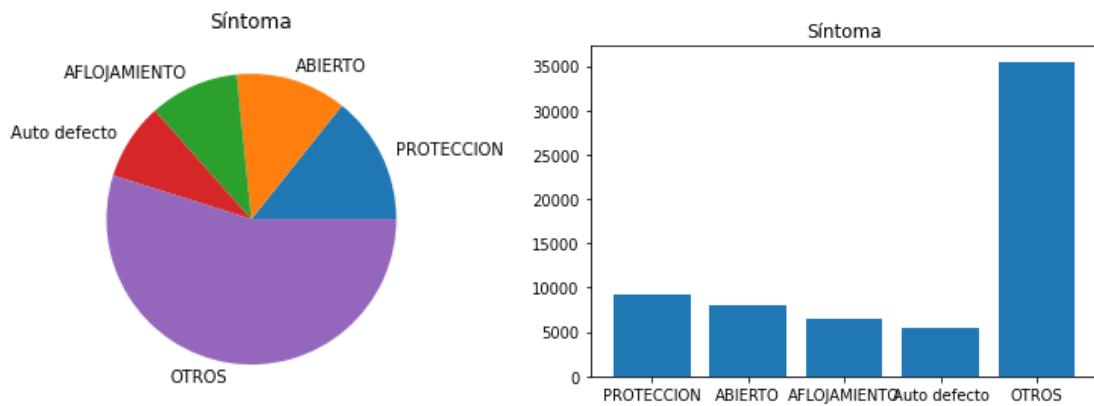
Gráfica 1: Gráfica de pastel “Causa”

También se creó una gráfica de barras con las mismas variables seleccionadas para la columna “Causa”.



Gráfica 2: Gráfica de barras “Causa”

Para la variable “Síntoma” se realizó el mismo procedimiento tanto para la gráfica de pastel como para la de barras, tomando únicamente los primeros 4 valores de la tabla de frecuencia: “AFLOJAMIENTO”, “ABIERTO”, “PROTECCION”, “Auto defecto”.



Gráficas 3 y 4: Gráficas de pastel y de barras de “Síntoma”

Para las variables **cuantitativas**, como se mencionó anteriormente, únicamente se cuenta con una, sin embargo, al no ser considerada como explicativa para el proyecto, se decidió no hacer exploración ni visualización de la misma.

A partir de la información presentada anteriormente, se puede decir que la calidad de los datos es generalmente buena, dado que las variables que se toman en consideración para el proyecto únicamente cuentan con 3 valores faltantes, además, en cuestión a la veracidad de los datos, como bien se nos mencionó en la plática con la OSF, existen casos en que los valores que se implementan a la hora del registro de los avisos son erróneos, debido a la

pereza de registrarlos correctamente. Por último, debido a que el propósito del proyecto de Ciencia de Datos, se enfoca en el uso de algoritmo de NLP, se utilizarán estas variables para que exista una congruencia entre las variables de Síntoma y Causa, junto con las variables libres de entrada de texto, por lo cual los valores de las variables libres de entrada de texto se consideran como útiles y correctos.

Sin embargo, como se ha repetido en varias ocasiones la base de datos proporcionada por la OSF cuenta con varios errores de ortografía, valores nulos, mayúsculas y minúsculas por lo que será necesaria una purificación de la misma asegurando así que se tomen en cuenta de la forma correcta todos los datos.

3. Preparación de los datos

Antes de pasar a la preparación de los datos fue necesario seleccionar las variables óptimas para el desarrollo del mismo proyecto, esto se realizó a partir del análisis descriptivo de cada variable, seleccionando lo mejor de acuerdo a las gráficas, a las tablas de frecuencia y medidas centrales como la moda, además, tomando cada una de las consideraciones de la OSF Ternium. A partir de esto, el conjunto de variables seleccionadas para el proyecto de Ciencia de Datos fueron: **Descripción**, **Qué pasó?**, **Porqué pasó?**, **Qué se hizo?**, **Síntoma** y **Causa**. Por esto mismo, no son necesarios los métodos de selección de variables. Añadiendo que estas columnas se enfocan en la **variable objetivo** *Descripción*, dado que cada una presenta información para lograr llegar a la clasificación de los avisos existentes en la base de datos.

3.1. Limpieza de datos

Para esta etapa se busca corregir la base de datos de Ternium, enfocándonos en reducir la base de datos únicamente al conjunto de las variables objetivo seleccionadas, esto con el fin de poder hacer más eficiente el futuro modelo de Machine Learning (ML) a crear. Además, considerando que uno de los principales problemas de la base de datos de Ternium es que cuentan con valores mal escritos o con diferentes modalidades de escrituras, ya sean minúsculas o mayúsculas, en este caso se cambiarán todos a minúsculas para una implementación más amena de los algoritmos del sistema de ML.

Para implementar lo anteriormente mencionado, se creó un nuevo dataframe (df2) la cual únicamente contenía las variables seleccionadas, así, la nueva dimensión del dataset es de 64,719 registros con 6 columnas o variables.

Posteriormente, para la transformación de la escritura de las variables de mayúsculas a minúsculas se utilizó `.lower()` para cada una de las columnas del nuevo data frame.

Descripción	Qué pasó?	Por qué pasó?	Qué se hizo?	Síntoma	Causa
fuga en cilindro rotativo	reporta operador fuga en cilindro rotativo de ...	por dañarse valvula rotativa de enrollador ot...	se cambia cilindro dañado por refaccion trabaj...	fugas	bajo nivel aceite
se-44927-1-hh servicio de emergencia equ	falla plc	programacion	se efectua servicio de cia control tech en ten...	señal/indicación/alarma errónea	error de programaci
cabledo de comunicacion de plc	se pide poner cable de comunicacion a tablero ...	por falta de comunicacion de plc a plc	se adaptaron gabinetes y conectores por donde ...	otros	otros
se-48312-1-hh servicio de emergencia equ	fallas en plc control logics	programacion	se configura logica	desprogramado	error de programaci
se-48189-1-hh servicio de emergencia equ	falla en plc control logix	desprogramado	se modifica programacion plc	falla de control	falla de comunicaci

Tabla 5: Visualización de los primeros registros del nuevo data frame

Como se puede observar en la Tabla 5, ahora los registros están escritos en minúsculas, por lo cual ahora se tiene mayor orden en la base de datos y mayor control en cada uno de los registros. Una vez teniendo esto se continuó con los valores nulos del nuevo data frame, para lo cual, de nuevo, con la función `.isnull()` se obtuvo la suma de los valores nulos de cada columna, esto para poder analizar si se deberían eliminar o sustituir por algún otro valor importante. Con esto se obtuvo lo siguiente:

Variable(valores nulos): Descripción(1), Qué pasó?(3), Por qué pasó?(3), Qué se hizo?(3), Síntoma(0), Causa(0).

Con esto se observa que los valores nulos son mínimos en las variables seleccionadas, con un total de 10 valores nulos. Por esto mismo, se decidió eliminar los registros con valores nulos, ya que como los datos se tratan de valores categóricos no se pueden sustituir por otro valor como la media. Para esto se utilizó la función `.dropna()`. Si bien es cierto que eliminar registros puede afectar negativamente nuestra base de datos, al ser máximo 10 registros a eliminar de 64,719, es decir un porcentaje muy bajo, no afecta casi nada la base de datos, dicho de una forma coloquial “es como redondear 1.01 a 1”.

Una vez eliminados los registros con valores nulos de la base de datos, la nueva dimensión del data frame es de 64,716, por lo cual se nota que únicamente se eliminaron 3 registros de la base de datos inicial, una cantidad muy pequeña.

De igual manera se realizó un análisis en la base de datos para confirmar la existencia de valores duplicados. Con la función de `.drop_duplicates()` y obteniendo la diferencia entre la longitud del data frame original y el segundo se obtuvo que hay un total de 10,777 registros

idénticos, los cuales fueron eliminados debido a que al momento de entrenar el modelo de ML para realizar la clasificación, el tener valores duplicados podrá resultar redundante clasificar los mismos registros, además de esta manera se puede agilizar dicho proceso. Una vez eliminados estos registros se tiene una dimensión de 53,939 registros.

También se quitaron los signos de puntuación y los caracteres no alfanuméricos para facilitar el entrenamiento del modelo en una etapa posterior. Para esto se importaron dos librerías importantes para el desarrollo, que fueron **string** y **re**, además, se creó una función llamada **clean**, la cual contiene diversas funciones como lo es **.lower()** para tener los registros en minúsculas, también **.translate()** que recibe como argumento a la función **.maketrans()** que remueve los signos de puntuación y por último con **.join()** se filtran las palabras que forma una cadena de caracteres, quitando los que no son alfanuméricos.

Una vez teniendo esto se eliminaron los espacios encontrados al principio y al final de los registros, para de igual manera agilizar el proceso de entrenamiento del modelo. Para esto se creó otra función llamada **remove_spaces** que recibe un string y te devuelve una cadena de caracteres sin los espacios iniciales y finales; esto se aplicó para cada una de las variables seleccionadas.

Por último, se quitaron los acentos con el mismo objetivo que los anteriores pasos; en este caso corresponde a la **normalización** de los datos. Para esto se reemplazaron los acentos de las vocales de cada una de las columnas por la vocal sin tilde, esto fue con la función **.replace()**.

	Descripción	Qué pasó?	Por qué pasó?	Qué se hizo?	Síntoma	Causa
0	fuga en cilindro rotativo	reporta operador fuga en cilindro rotativo de ...	por dañarse valvula rotativa de enrollador ot...	se cambia cilindro dañado por refaccion trabaj...	fugas	bajo nivel aceite
1	se449271hh servicio de emergencia equ	falla plc	programacion	se efectua servicio de cia control tech en ten...	señalindicacionalarmarretornea	error de programaci
2	cableido de comunicacion de plc	se pide poner cable de comunicacion a tablero ...	por falta de comunicacion de plc a plc	se adaptaron gabinetes y conectores por donde ...	otros	otros
3	se483121hh servicio de emergencia equ	fallas en plc control logics	programacion	se configura logica	desprogramado	error de programaci
4	se481891hh servicio de emergencia equ	falla en plc control logix	desprogramado	se modifica programacion plc	falla de control	falla de comunicaci
...
64714	dañarse engrane en transmision entrada	reportan fuerte ruido en transmision de entrada	se acude a destapar y se encuentra engrane dañado	se comienzan preparativos para cambio de engrane	fractura	fatiga
64715	fv2164 amarrada	reportan atoramiento	por falta de lubricacion	la lavan con petroleo y la lubrican	rayada	contaminacion
64716	255f2 cambio de solenoide de barrido	reportan falla en solenoide	atoramiento en solenoide	se pide solenoide para cambi	corrosion	fatiga
64717	671j revisar control de arrancador	el arrancador tiene falla	falla control de arrancador	se movieron tiempos de temporizadores	proteccion	proteccion
64718	instalar tapa de 223g	se daña empaque de tapa inferior en 223g	por fin de vida util	se cambia empaque	desgaste	fatiga

Tabla 6: Data frame sin acentuación, puntuación, espacios iniciales y finales y en minúscula

3.2. Transformación de Datos

Con el fin de poder obtener la mejor transformación de los datos, para de esta manera lograr cumplir con el objetivo planteado, se investigaron y seleccionaron diversas técnicas de transformación, la primera fue la de **binning**. Consideramos que **no** es necesario realizar esta técnica de discretización, dado que las variables necesarias para realizar este tipo de agrupamiento de datos se requieren conjuntos numéricos o continuos en un número más pequeño de intervalos o categorías discretos. Además, es importante mencionar que esta técnica también se usa en el análisis de datos para reducir la complejidad de los datos continuos y así, simplificar el proceso de análisis por medio de la agrupación de data en contenedores, con lo cual es posible identificar tendencias, patrones y distribuciones en los datos, que de otro modo, no habrían sido evidentes. Además, una de las ventajas es que se puede realizar manualmente o mediante herramientas de software automatizadas.

Por último, en relación a la **construcción de atributos** derivados, se sabe que estos son usados normalmente para generar variables nuevas que se crean a partir de variables existentes en un Dataframe, por lo cual **no** se considera conveniente, dado que lo que se busca es la limpieza y clasificación de los datos en las clases previamente establecidas, además normalmente se llega a utilizar para dar información adicional a partir de los datos originales, lo cual no se requiere en este caso.

La única transformación que se llevará a cabo será sustituir los valores posibles de las variables Causa y Síntoma de tipo de datos **string** a valores **numéricos discretos**, ya que se observó la posibilidad de elaborar modelos de categorización y agrupamiento como el método de K-means. La elección de estas variables fue debido a que no son textos libres, además que son aquellas columnas que tienen menos posibles valores del dataframe. Lo que se realizó para esto fue asignarle un valor a todos los posibles valores de las variables Causa y Síntoma y se le asignó ese valor a cada uno de los registros. Esto se hizo con la función *.replace* para sustituir los string por valores numéricos discretos. Se sustituyó el array que despliega la función *unique()*, que serán cambiados a un valor entero desde 1 hasta el número de valores posibles que se pueda seleccionar en cada variable, esto se realiza gracias a la función *arange()* de la biblioteca numpy que permite crear un array de números discretos. La forma en la que aparezcan los valores posibles en la función *unique()* se relaciona sucesivamente con el número desplegado por la función *arange()*. Dandonos como resultado el siguiente dataframe:

	Descripción	Qué pasó?	Por qué pasó?	Qué se hizo?	Síntoma	Causa
0	fuga en cilindro rotativo	reporta operador fuga en cilindro rotativo de ...	por dañarse valvula rotativa de enrollador ot...	se cambia cilindro dañado por refaccion trabaj...	1	1
1	se449271hh servicio de emergencia equ	falla plc	programacion	se efectua servicio de dia control tech en ten...	2	2
2	cableido de comunicacion de plc	se pide poner cable de comunicacion a tablero ...	por falta de comunicacion de plc a plc	se adaptaron gabinetes y conectores por donde ...	3	3
3	se483121hh servicio de emergencia equ	fallas en plc control logics	programacion	se configura logica	4	2
4	se481891hh servicio de emergencia equ	falla en plc control logix	desprogramado	se modifica programacion plc	5	4
...
64714	dañarse engrane en transmision entrada	reportan fuerte ruido en transmision de entrada	se acude a destapar y se encontro engrane dañado	se comienzan preparativos para cambio de engrane	8	6
64715	fv2164 amarrada	reportan atoramiento	por falta de lubricacion	la lavan con petroleo y la lubrican	40	9
64716	255f2 cambio de solenoide de barrido	reportan falla en solenoide	atoramiento en solenoide	se pide solenoide para cambi	24	6
64717	671j revisar control de arrancador	el arrancador tiene falla	falla control de arrancador	se movieron tiempos de temporizadores	14	5
64718	Instalar tapa de 223g	se daña empaque de tapa inferior en 223g	por fin de vida util	se cambia empaque	7	6

53939 rows x 6 columns

Tabla 7: Data frame con la transformación de Síntoma y Causa

Volviendo a la limpieza de datos:

Una vez planteada la transformación de datos, es necesario continuar con su limpieza, para poder realizar una clasificación y manejo de datos más eficiente al momento de realizar el procesamiento de los mismos. Se comenzó realizando el proceso de tokenización, el cual es indispensable para el procesamiento de lenguajes naturales, este proceso consiste en separar palabras del texto en entidades llamadas tokens, para que esta manera se pueda analizar mejor cada palabra y realizar de mejor manera los procesos de categorización. La tokenización se hizo por medio de la función `str.split()`; esta función permite separar las palabras por cada espacio; de tal forma que crea una lista de strings por cada espacio detectado, generando arrays. Finalmente se aplicó a cada columna del dataframe a excepción de síntoma y causa.

	Descripción	Qué pasó?	Por qué pasó?	Qué se hizo?	Síntoma	Causa
0	[fuga, en, cilindro, rotativo]	[reporta, operador, fuga, en, cilindro, rotati...]	[por, dañarse, valvula, rotativa, de, enrolla...]	[se, cambia, cilindro, dañado, por, refaccion,...]	1	1
1	[se449271hh, servicio, de, emergencia, equ]	[falla, plc]	[programacion]	[se, efectua, servicio, de, dia, control, tech...]	2	2
2	[cableido, de, comunicacion, de, plc]	[se, pide, poner, cable, de, comunicacion, a, ...]	[por, falta, de, comunicacion, de, plc, a, plc]	[se, adaptaron, gabinetes, y, conectores, por,...]	3	3
3	[se483121hh, servicio, de, emergencia, equ]	[fallas, en, plc, control, logics]	[programacion]	[se, configura, logica]	4	2
4	[se481891hh, servicio, de, emergencia, equ]	[falla, en, plc, control, logix]	[desprogramado]	[se, modifica, programacion, plc]	5	4
...
64714	[dañarse, engrane, en, transmision, entrada]	[reportan, fuerte, ruido, en, transmision, de,...]	[se, acude, a, destapar, y, se, encontro, engr...]	[se, comienzan, preparativos, para, cambio, de...]	8	6
64715	[fv2164, amarrada]	[reportan, atoramiento]	[por, falta, de, lubricacion]	[la, lavan, con, petroleo, y, la, lubrican]	40	9
64716	[255f2, cambio, de, solenoide, de, barrido]	[reportan, falla, en, solenoide]	[atoramiento, en, solenoide]	[se, pide, solenoide, para, cambi]	24	6
64717	[671j, revisar, control, de, arrancador]	[el, arrancador, tiene, falla]	[falla, control, de, arrancador]	[se, movieron, tiempos, de, temporizadores]	14	5
64718	[Instalar, tapa, de, 223g]	[se, daña, empaque, de, tapa, inferior, en, 223g]	[por, fin, de, vida, util]	[se, cambia, empaque]	7	6

53939 rows x 6 columns

Tabla 8: Data frame tokenizado

Una vez finalizado el proceso de tokenización, se dio a la tarea de la eliminación de palabras llamadas “stopwords”, dichas palabras son los artículos, preposiciones, conjunciones, pronombres, etc, palabras que permiten mantener cohesión en el lenguaje, pero carecen de un significado real por si solas. El proceso de eliminación de *stopwords*, se llevó a

cabo con la librería NLTK, que dentro de ella cuenta con una función que permite detectar palabras de dicho tipo y eliminarlas de nuestro Data Frame.

	Descripción	Qué pasó?	Por qué pasó?	Qué se hizo?	Síntoma	Causa
0	[fuga, cilindro, rotativo]	[reporta, operador, fuga, cilindro, rotativo, ...]	[dañarse, valvula, rotativa, enrollador, ote,...]	[cambia, cilindro, dañado, refaccion, trabajan...]	1	1
1	[se449271hh, servicio, emergencia, equ]	[falla, plc]	[programacion]	[efectua, servicio, clia, control, tech, tenson...]	2	2
2	[cabledo, comunicacion, plc]	[pide, poner, cable, comunicacion, tablero, plc]	[falta, comunicacion, plc, plc]	[adaptaron, gabinetes, conectores, pasa, cable]	3	3
3	[se483121hh, servicio, emergencia, equ]	[fallas, plc, control, logics]	[programacion]	[configura, logica]	4	2
4	[se481891hh, servicio, emergencia, equ]	[falla, plc, control, logix]	[desprogramado]	[modifica, programacion, plc]	5	4
...
64714	[dañarse, engrane, transimion, entrada]	[reportan, fuerte, ruido, transmision, entrada]	[acude, destapar, encontro, engrane, dañado]	[comienzan, preparativos, cambio, engrane]	8	6
64715	[fv2164, amarrada]	[reportan, atoramiento]	[falta, lubricacion]	[lavan, petroleo, lubrican]	40	9
64716	[255f2, cambio, solenoide, barrido]	[reportan, falla, solenoide]	[atoramiento, solenoide]	[pide, solenoide, cambi]	24	6
64717	[671j, revisar, control, arrancador]	[arrancador, falla]	[falla, control, arrancador]	[movieron, tiempos, temporizadores]	14	5
64718	[Instalar, tapa, 223g]	[daña, empaque, tapa, inferior, 223g]	[fin, vida, util]	[cambia, empaque]	7	6

53939 rows x 6 columns

Tabla 9: Data frame con la eliminación de stopwords

Para que el modelo funcione de la manera más óptima posible, se tomó en cuenta los posibles errores ortográficos que el dataframe pudiera tener, así que con ayuda de la librería *pyspellchecker* se desarrolló una función de corrección ortográfica de palabras que se observaba que se utilizaban más frecuentemente, esto para evitar que se crearan más clusters producto solamente de errores ortográficos. Sin embargo, al tratar de implementar esta función de corrección en el código se tomó la decisión de que era muy poco funcional, debido a que a pesar de que pueda funcionar correctamente, este conlleva demasiado tiempo en llevarse a cabo y además al tener algunos textos con números y letras para identificar modelos nos regresaba valores None el cual nos arruinaba la limpieza de datos hechos. Por esta razón llegamos a la conclusión de que no es óptimo incluirlo por la operatividad y funcionalidad del proyecto.

Con esto terminamos la Comprensión y la Preparación de Datos para poder continuar con la siguiente etapa que será el modelado del sistema a partir de Machine Learning con la metodología NLP.

	Descripción	Qué pasó?	Por qué pasó?	Qué se hizo?	Síntoma	Causa
0	[fuga, cilindro, rotativo]	[reporta, operador, fuga, cilindro, rotativo, ...]	[dañarse, valvula, rotativa, enrollador, ote,...]	[cambia, cilindro, dañado, refaccion, trabajan...]	1	1
1	[se449271hh, servicio, emergencia, equ]	[falla, plc]	[programacion]	[efectua, servicio, clia, control, tech, tenson...]	2	2
2	[cabledo, comunicacion, plc]	[pide, poner, cable, comunicacion, tablero, plc]	[falta, comunicacion, plc, plc]	[adaptaron, gabinetes, conectores, pasa, cable]	3	3
3	[se483121hh, servicio, emergencia, equ]	[fallas, plc, control, logics]	[programacion]	[configura, logica]	4	2
4	[se481891hh, servicio, emergencia, equ]	[falla, plc, control, logix]	[desprogramado]	[modifica, programacion, plc]	5	4
...
64714	[dañarse, engrane, transimion, entrada]	[reportan, fuerte, ruido, transmision, entrada]	[acude, destapar, encontro, engrane, dañado]	[comienzan, preparativos, cambio, engrane]	8	6
64715	[fv2164, amarrada]	[reportan, atoramiento]	[falta, lubricacion]	[lavan, petroleo, lubrican]	40	9
64716	[255f2, cambio, solenoide, barrido]	[reportan, falla, solenoide]	[atoramiento, solenoide]	[pide, solenoide, cambi]	24	6
64717	[671j, revisar, control, arrancador]	[arrancador, falla]	[falla, control, arrancador]	[movieron, tiempos, temporizadores]	14	5
64718	[Instalar, tapa, 223g]	[daña, empaque, tapa, inferior, 223g]	[fin, vida, util]	[cambia, empaque]	7	6

53939 rows x 6 columns

Tabla 10: Data frame FINAL

4. Aplicación de técnicas de Modelación

4.1. Extracción de características

La extracción de características es un paso fundamental en el proceso de aprendizaje automático que permite a los modelos analizar y procesar datos en nuestro caso de lenguaje natural de manera más efectiva.

Para la extracción de características en este proyecto de ciencia de datos, se utilizó el módulo incluido en la librería de sklearn llamada TF-IDF, el cual nos permite asignar un peso a cada palabra en nuestro dataset de acuerdo a la importancia de la misma, lo cual facilita el procesamiento de modelos de Machine Learning.

4.2. Explicación breve de los diferentes modelos de aprendizaje e hiperparámetros

K Neighbors(KNN):

Este tipo de modelo de aprendizaje supervisado generalmente es utilizado para clasificación como para regresión, este modelo calcula la distancia entre un punto de datos y los demás puntos del conjunto de entrenamiento. En lugar de aprender un modelo a partir de los datos de entrenamiento, el algoritmo KNN simplemente almacena los datos y clasifica las nuevas observaciones basándose en la similitud con los datos de entrenamiento y los k vecinos más cercanos son elegidos para determinar la clase del punto de datos.

Los hiperparametros de este modelo de clasificación son:

- n_neighbors: el número de puntos más cercanos al nuevo que queremos que voten. Los “N” más cercanos.
- weights: La ponderación de los votos de cada punto, puede ser:
 - uniforme: todos los puntos valen lo mismo.
 - distance: los votos de los puntos más cercanos tienen más valor
- algorithm: controla el algoritmo que computa las distancias, las opciones son:
- brute: es el algoritmo de búsqueda de fuerza bruta, lo que hace es calcular todas las distancias de todos los puntos entre sí.

- kd_tree: es el recomendado para dataset de tamaño medio. Está basado en árboles, la idea es sencilla, si sabemos que un punto “A” está muy distante de un punto “B” y “B” está muy cercano a un punto “C”, entonces podemos inferir que “A” está lejos de “C”.
- ball_tree: es muy similar al kd_tree, pero el algoritmo funciona más rápido para datasets grandes.
- auto: el propio algoritmo decide cual de los tres anteriores es el más óptimo.

Random forest

El modelo de aprendizaje supervisado Random Forest se basa en la combinación de múltiples salidas de árboles de decisión unificadas en un simple resultado, en este caso, para poder predecir a la variable clase, para los cuales, cada árbol analiza diversas partes aleatorias de los datos de entrenamiento, construyéndose de manera independiente. Es importante mencionar que el modelo de Random Forest, al estar compuesto de diversos árboles de decisión (dependiendo el número de árboles seleccionado para el modelo), cuentan con varias preguntas en los nodos de las hojas, para de esta manera poder verificar si se debería continuar o no dependiendo de qué se esté buscando y tomando en cuenta la métrica utilizada, repitiendo el proceso iterativamente hasta conseguir el bosque completo. Por último, ya que se construyen todos los árboles de decisión, este algoritmo combina las predicciones de cada árbol para generar una predicción final.

- n_estimators: El número de árboles en el bosque.
- criterion: La función de pérdida que se utiliza para medir la calidad de una división. Los valores posibles son «gini» (impureza de Gini) y «entropy» (entropía).
- max_depth: La profundidad máxima de cada árbol en el bosque. Si se establece en None, los árboles se expanden hasta que todas las hojas sean puros o hasta que se alcance el límite máximo de hojas.
- min_samples_split: El número mínimo de muestras necesarias para dividir un nodo interno.
- min_samples_leaf: El número mínimo de muestras necesarias para estar en una hoja.
- min_weight_fraction_leaf: El porcentaje mínimo de la suma total de pesos (en el conjunto de entrenamiento) requerido para estar en una hoja.

- `max_leaf_nodes`: El número máximo de hojas en un árbol. Si se establece en `None`, los árboles se expanden hasta que todas las hojas sean puros o hasta que se alcance el límite máximo de hojas.

SVM (Support Vector Machine)

Es un modelo de aprendizaje supervisado para problemas de clasificación y regresión. SVM funciona encontrando el mejor hiperplano que separa capas de datos en un espacio multidimensional. El objetivo de SVM es encontrar el hiperplano que maximiza la separación entre clases. Para ello, el SVM busca un hiperplano que esté lo más alejado posible de las observaciones de entrenamiento más recientes de ambas clases. Estas observaciones de entrenamiento más cercanas al hiperplano se denominan vectores de soporte. El hiperplano encontrado por SVM es la línea o región que separa las dos clases de observaciones. En problemas de clasificación binaria, este hiperplano divide el espacio en dos regiones, una para cada clase. Sin embargo, los datos no pueden estar separados por un hiperplano en el espacio original.

Los hiperparámetros de SVM son:

- `C`: Es el parámetro de regularización, que controla la penalización por clasificar mal un punto de datos.
- `kernel`: Es el tipo de función kernel utilizada para transformar los datos de entrada en un espacio de características de alta dimensión.
- `gamma`: Es el parámetro del kernel gaussiano (RBF), que controla la amplitud de la función gaussiana.
- `degree`: Es el grado del kernel polinómico.
- `coef0`: Es un término independiente en las funciones kernel polinómicas y sigmoidales.

NAIVE BAYES

Este modelo es un tipo de clasificación probabilística, lo que quiere decir que hace clasificaciones en base a las probabilidades de las variables y también clasifica en diversas categorías, además es considerado como uno de los métodos más simples, eficaces y rápidos para hacer predicciones. El método que usa este modelo para calcular la probabilidad de que una serie de datos específica pertenezca a una clase, se calcula la probabilidad de cada clase,

y con el uso de las suposiciones de independencia de características calcula la probabilidad de cada característica dada cada clase; por último se obtiene la probabilidad conjunta con el producto de ambas probabilidades. Con el teorema de Bayes se calcula la probabilidad posterior, en donde la probabilidad posterior más grande, corresponde a la clasificación predicha para los datos.

Los hiperparámetros de Naive Bayes son:

- Alfa (α): Es un parámetro de suavizado utilizado para prevenir la probabilidad cero.
Fit_prior: Es un booleano que indica si se deben calcular o no las probabilidades previas de clase.
- Class_prior: Es una lista o matriz que contiene las probabilidades previas de clase.
- Binarize: Es un valor umbral que se utiliza para binarizar las características.
- Multinomial_nb: Se utiliza para especificar si se va a utilizar el modelo Naive Bayes multinomial o no.

LOGISTIC REGRESSION

La regresión logística es otro método comúnmente utilizado en el procesamiento de lenguaje natural para modelar la relación entre variables predictoras lingüísticas (como la frecuencia de las palabras, las características sintácticas, la longitud de los textos, etc.) y la probabilidad de una cierta categoría o etiqueta para un texto dado. Se puede utilizar este método para modelar la probabilidad de que el texto pertenezca a cada categoría, dado un conjunto de características del texto.

Los hiperparámetros son:

- Regularización: La regularización es un proceso que reduce la complejidad del modelo al penalizar los coeficientes que son demasiado grandes.
- Tasa de aprendizaje: la tasa de aprendizaje determina qué tan rápido el modelo actualiza los pesos durante el proceso de entrenamiento.
- Número de iteraciones: el número de iteraciones o épocas se refiere a cuántas veces el modelo revisa todo el conjunto de datos de entrenamiento.
- Umbral de decisión: el umbral de decisión se utiliza para determinar a qué clase pertenece una observación.

- Métodos de solución: existen diferentes métodos de solución para encontrar los valores de los coeficientes que mejor ajustan el modelo a los datos.

K-MEANS

Al ser un modelo que está planeado para funcionar con variables numéricas cuando le ingresamos variables categóricas lo que hace es tomar en cuenta las modas que hay por lo que incluso pasarlo a números (como se hizo con **SÍNTOMA** y **CAUSA**) podría ser una opción viable, sin embargo el hacer esto no toma en cuenta una gran cantidad de datos, en especial porque se tiene que juntar palabra por palabra con las otras variables x entonces aunque sean similares el programa lo toma en cuenta como 2 distintas.

Por estas y más razones este modelo no funcionaba para esta situación y se decidió terminar al momento en el que el programa encontró la cantidad de cluster que había en nuestros datos, dando una gráfica de codo en la que después de cálculos demuestra que hay 35 clusters.

4.3. Explicación de la metodología para el entrenamiento y prueba.

La metodología utilizada en este proyecto de ciencia de datos para el entrenamiento y prueba, se basó en utilizar la librería de sklearn, la cual nos permite dividir nuestra base de datos en dos subconjuntos excluyentes: el de entrenamiento y el de prueba.

La fase del entrenamiento del modelo tiene la finalidad de utilizar algoritmos de aprendizaje para ajustar sus parámetros de tal manera que pueda realizar predicciones precisas sobre datos futuros. El objetivo es que el modelo sea capaz de generalizar bien y hacer predicciones precisas sobre datos que no ha visto antes.

Una vez realizada la fase de entrenamiento del modelo, se hace la prueba del modelo, en esta fase se utiliza un conjunto de datos que no se utilizó durante el entrenamiento (subconjunto test) para hacer predicciones. Esto permite evaluar el rendimiento del modelo en datos desconocidos y determinar si el modelo ha sido capaz de generalizar bien.

4.4. Generación de Modelos

Para la generación de modelos se utilizaron métodos de *feature extraction* para las variables que contenían texto de lenguaje natural, una vez realizado el proceso de feature extraction se utilizó la función de *pipeline* que permite anidar partes de preprocesamiento

(TF-IDF) junto con los modelos de clasificación que se seleccionaron. En este proyecto de ciencia de datos se optó por realizar modelos supervisados, gracias a que dentro de las bases de datos ya se encuentran “etiquetados” los avisos, por medio de las variables de **Síntoma** y **Causa**, lo que nos permite evaluar de manera óptima los resultados que arrojen la clasificación de los modelos.

De métodos de clasificación supervisados se realizaron distintos algoritmos para obtener un mejor resultado con base en los *accuracy score* de cada uno de los modelos. Los modelos utilizados en este proyecto de ciencia de datos fueron: K-Neighbors, Random Forest y Naive Bayes.

Una vez adentrados en el proceso de generación de modelos, se tomó la decisión de unificar las variables independientes en una sola por medio de la función de concatenación, esto gracias a que permite introducir una mayor información al modelo de clasificación en lugar de considerar una sola variable independiente de las cuatro existentes, lo que provoca una mejora en la métrica de accuracy score y genera una predicción más realista con el subconjunto de prueba de las variables.

Se hicieron pruebas de los modelos con las variables independientes de forma individual en cada uno de los modelos de clasificación, así como las posibles combinaciones de fusión de ellas. Observando la métrica de accuracy score se observa un mejor rendimiento de clasificación con el método de **Random Forest** cuando se utilizan las 4 variables independientes unificadas.

```
TF-IDF Transform
Empezamos la modelación con este método para que nos haga la tokenización y la indexación y nos será de utilidad para realizar los otros modelos.

[ ] from sklearn.feature_extraction.text import TfidfVectorizer

[ ] # vemos que hace TfidfVectorizer()
v = TfidfVectorizer()
transformed_output = v.fit_transform(df4["Descripción"])
print(v.vocabulary_)

{'fuga': 4946, 'cilindro': 2149, 'rotativo': 9178, 'sehh': 9429, 'servicio': 9518, 'emergencia': 3966, 'equ': 4299, 'cabledo': 1556, 'comunicacion': 2377, '}

EMPEZAMOS POR SÍNTOMA Y DESCRIPCIÓN NADA MÁS

➊ # TRAIN TEST
# x = df4["Descripción"].str.cat(df4["Qué pasó?"], sep = " ")
x = df4["Descripción"]
y = df4["Síntoma"]

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 1) # entrenamos con el 80% y probamos con el 20%

Ahora para entrenar el modelo de TF-IDF Transform necesitamos de clasificadores de sklearn como KNN, Random Forest, etc.

Usamos los clasificadores para ver cuál es el mejor modelo.
```

K Neighbors

```
[ ] from sklearn.neighbors import KNeighborsClassifier
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report

clf = Pipeline([
    ("vectorizer_tfidf", TfidfVectorizer()),
    #Kmeans,
    ("KNN", KNeighborsClassifier())
])

clf.fit(x_train, y_train)

y_pred = clf.predict(x_test)

print(classification_report(y_test, y_pred))

[ ] # TRAIN TEST
# x = df4["Descripción"].str.cat(df4["Qué pasó?"], sep = " ")
x = df4["Descripción"] + df4["Qué pasó?"]
y = df4["Síntoma"]

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 1) # entrenamos con el 80% y probamos con el 20% . . .

# TRAIN TEST
# x = df4["Descripción"].str.cat(df4["Qué pasó?"], sep = " ")
x = df4["Descripción"] + df4["Qué pasó?"] +df4["Por qué pasó?"] + df4["Qué se hizo?"]
y = df4["Síntoma"]

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 1) # entrenamos con el 80% y probamos con el 20% . . .

from sklearn.neighbors import KNeighborsClassifier
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report

clf = Pipeline([
    ("vectorizer_tfidf", TfidfVectorizer()),
    #Kmeans,
    ("KNN", KNeighborsClassifier())
])

clf.fit(x_train, y_train)

y_pred = clf.predict(x_test)

print(classification_report(y_test, y_pred))
```

Random Forest

```
[ ] from sklearn.ensemble import RandomForestClassifier

#Creamos el pipeline que permite primero cambiar el dataframe con TfidfVectorizer y luego usar el clasificador
clf = Pipeline([
    ("vectorizer_tfidf", TfidfVectorizer()),
    ("Random Forest", RandomForestClassifier())
])

clf.fit(x_train, y_train)

# hacemos las predicciones para x_test y lo guardamos en y_pred
y_pred = clf.predict(x_test)

# imprimimos el reporte de la clasificación
print(classification_report(y_test, y_pred))
```

Vamos a ver cómo funcionó

```
[ ] x_test[:5]
[ ] y_test[:5]
[ ] y_pred[:5]
```

Observamos que Random Forest tuvo un mejor desempeño que KNN ya que aumentó la precisión (51%) y además observamos que ahora solo se equivocó en la última línea en la muestra.

NAIVE BAYES

```
[ ] from sklearn.naive_bayes import MultinomialNB  
clf = Pipeline([  
    ("vectorizer_tfidf", TfidfVectorizer()),  
    ("Multi NB", MultinomialNB())  
])  
clf.fit(x_train, y_train)  
y_pred = clf.predict(x_test)  
print(classification_report(y_test, y_pred))
```

Vamos a ver cómo funcionó

```
[ ] x_test[:5]  
[ ] y_test[:5]  
[ ] y_pred[:5]
```

El clasificador Naive Bayes tiene 40% de precisión y vemos errores en la tercera y última fila en la muestra.

4.5 Breve discusión que justifique la validez del modelo

Previo al ajuste de parámetros, el modelo de **Random Forest** obtuvo un *accuracy score* del 61%, siendo el más alto en comparación a los demás modelos, sin considerar que Naive Bayes obtuvo el mismo valor de *accuracy* pero esto sucedió al usar la modificación de parámetros. Después de haber ajustado los hiperparámetros para el modelo seleccionado, se lograron encontrar los parámetros que mejor explicaran, generalizaran y mejoraran el rendimiento del modelo, a partir de esto se obtuvo la métrica de desempeño *accuracy* para validar la calidad del modelo posterior al ajuste de los hiperparámetros encontrados, con esto hecho, el porcentaje obtenido fue de 86%. Se sabe que este valor indica la proporción de predicciones correctas que realiza el modelo sobre el total de predicciones realizadas, siendo un valor sumamente alto en comparación al porcentaje de datos que se tomaron para el set de entrenamiento. Es por esto que se considera que la validez del modelo es correcta, puesto que lo que se buscaba era la clasificación y predicción de valores correctamente y se tuvo un porcentaje bastante alto.

4.6 Ajuste de modelos a través de hiperparámetros.

El ajuste de los hiperparámetros se basó en la validación cruzada proporcionada por el módulo de la biblioteca de sklearn llamado SearchGridCV, que permite obtener los hiperparametros más óptimos de acuerdo con el modelo seleccionado.

4.7 Evaluación y selección de modelo(s) de acuerdo a la metodología de validación. Hiperparámetros utilizados en los modelos.

Una vez evaluado el modelo y haber obtenido un *accuracy* de 61%, se decidió ajustar los hiperparámetros para lograr obtener una *accuracy* mayor como en el caso de Naive Bayes. La selección de estos nuevos hiperparámetros se realizó mediante la función *GridSearchCV* obteniendo los siguientes resultados `{'rf_max_depth': None, 'rfmin_samples_split': 5, 'rfn_estimators': 200, 'tfidf_max_df': 0.5}`, una vez corriendo el modelo implementando esta modificación se obtiene un *accuracy* del 86%.

5. Evaluación

5.1 Evaluación de resultados: Entender e interpretar los resultados obtenidos, su impacto y utilidad, considerando los criterios de éxito del negocio.

Después de haber aplicado diversos algoritmos para la clasificación de los datos de Ternium, en este caso, los avisos y los síntomas de las fallas de equipo, y posterior a la obtención de modelos como lo fue de k-means, regresión logística, Naive Bayes, KNN, SVM y Random Forest; al igual después de haber obtenido el valor de *accuracy* para cada uno de ellos, con el fin poder brindar una comparación y elección del modelo más óptimo para los objetivos de la empresa socio formadora, se considera que los resultados obtenidos para el modelo de Random Forest, el cual obtuvo un *accuracy* suficientemente bueno de 61% sin la necesidad de modificar hiperparámetros, se puede decir que el modelo tiene un gran impacto y utilidad considerando los criterios de éxito de Ternium, los cuales giran en torno a los objetivos principales del negocio: “*Interpretar y clasificar automáticamente textos compuestos en categorías definidas de un proceso sin tener un set de datos inicialmente clasificados*”, para lo cual se buscaba un modelo capaz disminuir la cantidad del trabajo humano, mejorar la organización, preparar la base de datos y fortalecer los procesos de

segmentación; para finalmente clasificar de manera idónea los avisos en el proceso de producción. Al obtener un *accuracy* relativamente alto, lo cual explica la calidad del modelo y después de analizar el comportamiento del modelo, se observó que en efecto se pudieron cumplir los objetivos de Ternium, dándoles una mejor forma de validar los modelos de Machine Learning y de implementarlo en su base de datos. Añadiendo que con lo dicho anteriormente se puede confirmar la hipótesis establecida al inicio del proyecto, que indica aún más la satisfactoria obtención del resultado.

5.2 Revisión del proceso: Sumarizar todo el proceso, principales problemas, posibles mejoras.

Para la selección del modelo con mayor calidad considerando los objetivos del proyecto y de la organización socio formadora, se tuvo que realizar un breve análisis de desempeño para cada uno. Esto se realizó a partir de la obtención del valor de *accuracy* para cada uno, así como también, la modificación de hiperparámetros para algunos de ellos. Después de realizar lo anteriormente dicho, se tomaron en consideración los mejores modelos, que fueron Random Forest y Naive Bayes, debido a su practicidad, y aunque los dos son modelos bastante diferentes tienen varias similitudes como su constante uso en la clasificación de datos, ambos modelos son modelos de aprendizaje supervisado y ambos modelos se usan para la clasificación de múltiples clases lo que nos permite tener un análisis de datos de más de dos clases de datos, de suma importancia en este caso debido a la selección de dos variables dependientes, sin embargo, al utilizarlos en el data frame proporcionado, se concluyó que el mejor modelo a utilizar fue **Random Forest**, por su enfoque en clasificación de datos de todo tipo y porque al utilizarlo a la base de datos se mantuvo un mayor nivel de predicción al modelo de Naive Bayes.

A pesar de que al inicio el modelo de Naive Bayes tuvo un *accuracy* del 47%, para verificar si este podía aumentar el desempeño se utilizó la modificación de hiperparámetros, para lo cual se logró empatar al modelo de Random Forest, sin embargo, se intentó modificar los hiperparámetros para este modelo y no se logró debido al tiempo de ejecución del programa, aunque se considera que en caso de haber sido posible, el valor de calidad pudo haber mejorado significativamente.

5.3 Impacto social principal

La implementación del lenguaje natural en las empresas, como lo es Ternium en este caso, tiene un impacto significativo en varios aspectos sociales, desde la forma en cómo las empresas interactúan con sus clientes y empleados, hasta la forma en la que operan internamente. De forma general podemos mencionar que la implementación del lenguaje aumenta la eficiencia operativa, ya que permite a las personas a ser más eficientes al automatizar las tareas y procesos, como lo es la atención al cliente y la gestión de datos. Esto contribuye a liberar recursos y tiempo para que los empleados se centren en tareas de valor agregado, lo cual aumenta su productividad y reduce costos.

Así mismo, la clasificación de datos automatizada a través de NLP mejora la precisión y eficiencia en los procesos, además que contribuye a hacer la clasificación de datos más transparente, de tal forma que sean auditados y explicados para asegurar que las decisiones de clasificación sean justas y precisas. Finalmente otra contribución social que tiene en procesamiento de lenguaje natural es que la clasificación automatizada ayuda a proteger los datos personales, ya que al automatizar los procesos de clasificación se reduce la posibilidad que los datos sean vistos por personas no autorizadas.

5.4 Impacto hacia los Objetivos de Desarrollo Sostenible.

Previo al desarrollo del proyecto se investigaron los ODS que más se relacionan con la problemática que Ternium tiene, y se vieron impactados en el desarrollo de este, los cuales se muestran a continuación.

Industria, innovación e infraestructura, ya que gracias al uso de modelos no supervisados se logró detectar patrones de problemas constantes y por consiguiente una mejor organización en la base de datos, para así hacer más eficiente el proceso de toma de acciones para resolver dichos problemas.

Trabajo decente y crecimiento económico, ya una vez teniendo un plan que permita atender las problemáticas más frecuentes, se tiene una mayor oportunidad de tener productos de buena calidad y así permitir asegurar el posicionamiento de Ternium como uno de los líderes de producción de acero plano en Latinoamérica, así como generar mayores ganancias.

Producción y consumo responsable, ya que se tienen vistas las problemáticas más frecuentes, es posible atenderlas de forma eficiente, reduciendo así el tiempo y consumo de recursos que se utilizarían al no conocer la forma correcta o más rápida de atender dicho problema, y con esto aportar al objetivo que tiene Ternium de reducir las emisiones de dióxido de carbono un 20% para 2023, mediante el reciclaje de acero y la captura de carbono, esto gracias a su ruta de descarbonización.

6. Despliegue

6.1 Descripción del prototipo funcional

A través de streamlit y anaconda, se pudo realizar un prototipo funcional de un modelo de clasificación, el cual permitirá a los trabajadores de Ternium, introducir las variables independientes de forma que se pueda hacer la clasificación correcta.

Para este caso, se crea un environment para el despliegue de la página web, así como el uso del modelo a través de su descarga del notebook trabajado, de forma que pudiera ser introducido al entorno del desarrollo web.

Aquí hay unas breves fotos del código del prototipo funcional, así como su funcionamiento.

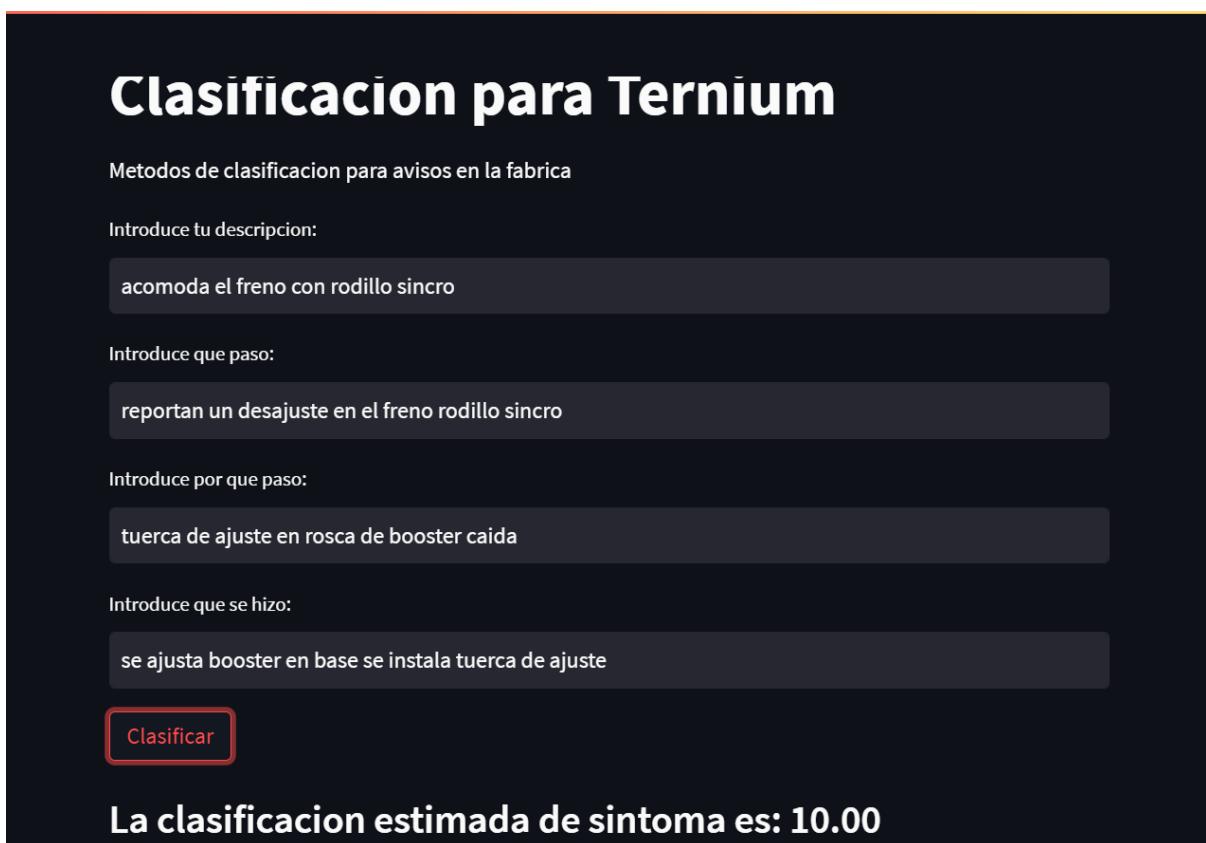
```
predict_page.py > show_predict_page
1  import streamlit as st
2  import pickle
3  import numpy as np
4
5  def load_model():
6      modelo=open('NBbueno.pkl', 'rb')
7      clasificacion=pickle.load(modelo)
8      return clasificacion
9
10 clasificacion=load_model()
11
12 def show_predict_page():
13     st.title("Clasificación para Ternium")
14
15     st.write("Métodos de clasificación para avisos en la fábrica")
16     cate= ("falla drive rda run procesomarca falla driverestablece",
17            "tornillo desgastado vielas lado motri mal corte material abre tapa viela cambio tornillo")
18
19     descripción = st.text_input("Introduce tu descripción: ")
20     qp=st.text_input("Introduce qué pasó: ")
21     pq=st.text_input("Introduce por qué pasó: ")
22     qh=st.text_input("Introduce qué se hizo: ")
23     opciones= descripción+qp+pq+qh
24     opciones=[str(opciones)]
25     ok=st.button("Clasificar")
26
27     if ok:
28
29         predicciones = clasificacion.predict(opciones)
30         st.subheader(f"La clasificación estimada de síntoma es: {predicciones[0]:.2f}")
```

```

26     if OK:
27
28         predicciones = clasificacion.predict(opciones)
29         st.subheader(f"La clasificación estimada de síntoma es: {predicciones[0]:.2f}")
30
31     reg=st.button("Registrar")
32     top5=st.button("Revisar 5 síntomas más frecuentes")
33     if top5:
34         st.subheader("Los top 5 síntomas son: ")
35         st.write("1.- Protección")
36         st.write("2.- Abierto")
37         st.write("3.- Aflojamiento")
38         st.write("4.- Otros")
39         st.write("5.- Automático por defecto")
40     show_predict_page()

```

Así como una prueba de clasificación realizada, donde se obtenía el resultado correcto:



7. Recomendaciones

7.1 Recomendaciones al negocio

Algunas observaciones que se le pueden dar al socio formador para evitar problemas futuro al utilizar el modelo, son el mejoramiento en algunos aspectos en el data frame como el mejoramiento en encabezados y columnas, asegurándose que las columnas sean más descriptivas y significativas, sin dejar pasar el buscar técnicas de recopilación de datos que

ayuden a disminuir la limpieza y preparación de los datos, de igual forma se recomienda el uso de nombres cortos en columnas pero informativos, para así mejorar el comprendimiento de los datos que se almacenan, y poder clasificar los datos más fácil.

También, mejorar el orden de las columnas para organizarlas de una forma que sean más fáciles de leer y comprender, que mejoren su criterio para la verificación de datos para asegurar que los datos sean correctos y precisos, mejorar el formato de los datos para que puedan verificar consistencia. Además es de suma importancia considerar el tamaño del data frame, puesto que si el tamaño es demasiado grande, se deben de aplicar técnicas de reducción de dimensionalidad para conservar las variables más significativas, puesto que esto puede llegar a afectar la veracidad del modelo al incluir variables que no logren explicar correctamente la variabilidad de los datos, de igual manera es importante verificar la ortografía, asegurándose que los datos iguales están escritos de la misma manera y evitar la redundancia para no afectar en gran medida o poder arruinar el modelo y por último, desarrollar un glosario para el mejoramiento en la identificación y entendimiento de variables, así como para facilitar la clasificación de características futuras.

7.2 Recomendaciones técnicas

Algunas técnicas recomendadas para el uso del modelo son; la selección de características, la cual consiste en el manejo de una gran cantidad de características y variables, sin embargo es importante considerar que en caso de utilizar variables irrelevantes del dataframe, se podría afectar de manera negativa el modelo. Otra técnica es el ajuste de hiperparámetros, el cual no pudo ser implementado correctamente en el modelo por el tiempo de carga, pero no se descarta puesto que al contar con más memoria y tiempo de ejecución se podría obtener un índice de desempeño más alto, este consiste en la optimización del modelo para el ajuste adecuado de los hiperparámetros como el número de árboles, la profundidad máxima de los árboles, el número mínimo de muestras por hoja, entre otros utilizados para el algoritmo.

Otra técnica es el uso de validación cruzada la cual consiste en evaluar el rendimiento del modelo utilizando datos de entrenamiento y prueba, esto ayuda a la generalización de nuevos datos y tiene una gran utilidad debido a que existen diferentes tipos dependiendo los datos utilizados. La gestión del desequilibrio de clases es una técnica que consiste en verificar si el data frame tiene clases desequilibradas, en otras palabras una clase puede llegar a tener

muchas más instancias que las demás, lo que afecta el rendimiento y ajusta el peso de las clases en el modelo.

La última técnica para verificar el rendimiento del modelo es la interpretación de características, la cual consiste en el proporcionamiento de una medida de importancia de características que puede utilizarse para interpretar el modelo, esto puede aportar en la comprensión de variables importantes para la clasificación y de esta manera observar cómo afecta en el rendimiento del modelo. Por último, es importante considerar la limpieza previa de los datos, identificación de tipos de datos y preparación de los mismos para de esta manera aplicar el modelo de la manera óptima.

Las técnicas que puede proporcionar un modelo como Random Forest son bastante poderosas en el aspecto que pueden proporcionar una clasificación precisa en una amplia variedad de datos. Sin embargo, es importante aplicar técnicas como la selección de características, la validación cruzada y la gestión del desequilibrio de clases para maximizar su rendimiento y comprender mejor el modelo y su desarrollo.

Referencias

Berlanga, V., Rubio Hurtado, M., Vilá Baños, R. (2013). *Cómo aplicar árboles de decisión en SPSS.* ICE Universitat de Barcelona, DOI: 10.1344/reire2013.6.1615. <http://deposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf>

Charu Aggarwal, Cecilia Procopiuc, Joel Wolf, Phillip Yu, and Jong Park. "Fast algorithms for projected clustering", In ACM SIGMOD Conference, (1999).

F. Charrua-Santos *et al.*, "An Overview of Lean Production and Industry 4.0 in Different Context," *2020 9th International Conference on Industrial Technology and Management (ICITM)*, Oxford, UK, 2020, pp. 69-72, doi: 10.1109/ICITM48982.2020.9080386.

González, L. (2020, 18 agosto). *Algoritmos de Agrupamiento.*  Aprende IA. <https://aprendeia.com/algoritmos-de-clustering-agrupamiento-aprendizaje-no-supervisado/>

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbSCAN: *Fast density-based clustering with R.* *Journal of Statistical Software*, 91, 1-30.

Hernández, M.; Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica*, 32 (1), 87 - 96. https://revistapolitecnica.epn.edu.ec/ojs2/index.php/revista_politecnica2/article/view/32/pdf

Origel, C; Rendón, E; Abundez, I; Alejo, R. (2020). "Redes neuronales artificiales y árboles de decisión para la clasificación con datos categóricos", *Tecnológico Nacional de México, Instituto Tecnológico de Toluca, México.* https://rcs.cic.ipn.mx/2020_149_8/Redes%20neuronales%20artificiales%20y%20árboles%20de%20decisión%20para%20la%20clasificación%20con%20datos%20categoricos.pdf

Patrick, P. (2005). Randomized Algorithms and NLP: Using Locality Sensitive Hash Function for High Speed Noun Clustering. *Proceedings of the 43rd Annual Meeting of the ACL*, 622–629. <https://aclanthology.org/P05-1077.pdf>

Rodríguez Suárez, Y., & Díaz Amador, A. (2009). Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas*, 3(3-4), 73-80.