Introducción

En el contexto profesional cada vez más competitivo, surge la necesidad de prepararse cada vez más y diversificar la hoja de vida con lo que el mercado demanda es imperante. La demanda a nivel general es cada vez más exigente y la competencia entre productores hace que la minimización de los costos sea un factor determinante; este contexto permite que las fuerzas productivas avancen y es aquí que la tecnología se posicione como un actor clave.

En ese orden de ideas, partiendo nuestro análisis propio del mercado y nuestros proyectos futuros, además de los conocimientos adquiridos en el curso de Talento Especializado decidimos realizar el Reto Tecnológico 6 con nombre Análisis predictivo de fallos mecánicos. Este reto consiste en analizar las características que podrían hacer fallar a un grupo de máquinas haciendo diferentes procedimientos como filtrado, agrupamiento y en general modificación de la información suministrada acerca de diferentes máquinas de distintas industrias. Este reto ayuda a nuestra formación profesional y nuestra capacidad de analizar datos, de manipularlos y hacer conclusiones cada vez más acertadas.

En el presente documento comentaremos cómo fue el proceso para solucionar el reto; las consideraciones previas y en el desarrollo del mismo, los procedimientos que tuvimos que seguir para poder sacar adelante el código y la descripción de los datos. Además, quisimos ir más allá de los parámetros del reto como tal y desarrollamos un modelo de regresión logística y un Random Forest que permite hacer predicciones de las fallas mecánicas a partir de los datos suministrados.

El Reto

El reto consiste en analizar un dataset que tiene diversas características que podrían hacer fallar a un grupo de máquinas. Este análisis incluye varios procedimientos como agrupación, filtrados, listados entre otros.
En un principio se decidió realizar un análisis sobre las características que presenta el dataset como condiciones de funcionamiento y proceso de las máquinas, nombradas a continuación: Air Temperature, Process Temperature, Rotational Speed, TorqueTool Wear. Sobre estas características se realizó un chequeo para verificar la presencia de datos nulos o atípicos; con el uso de histogramas.

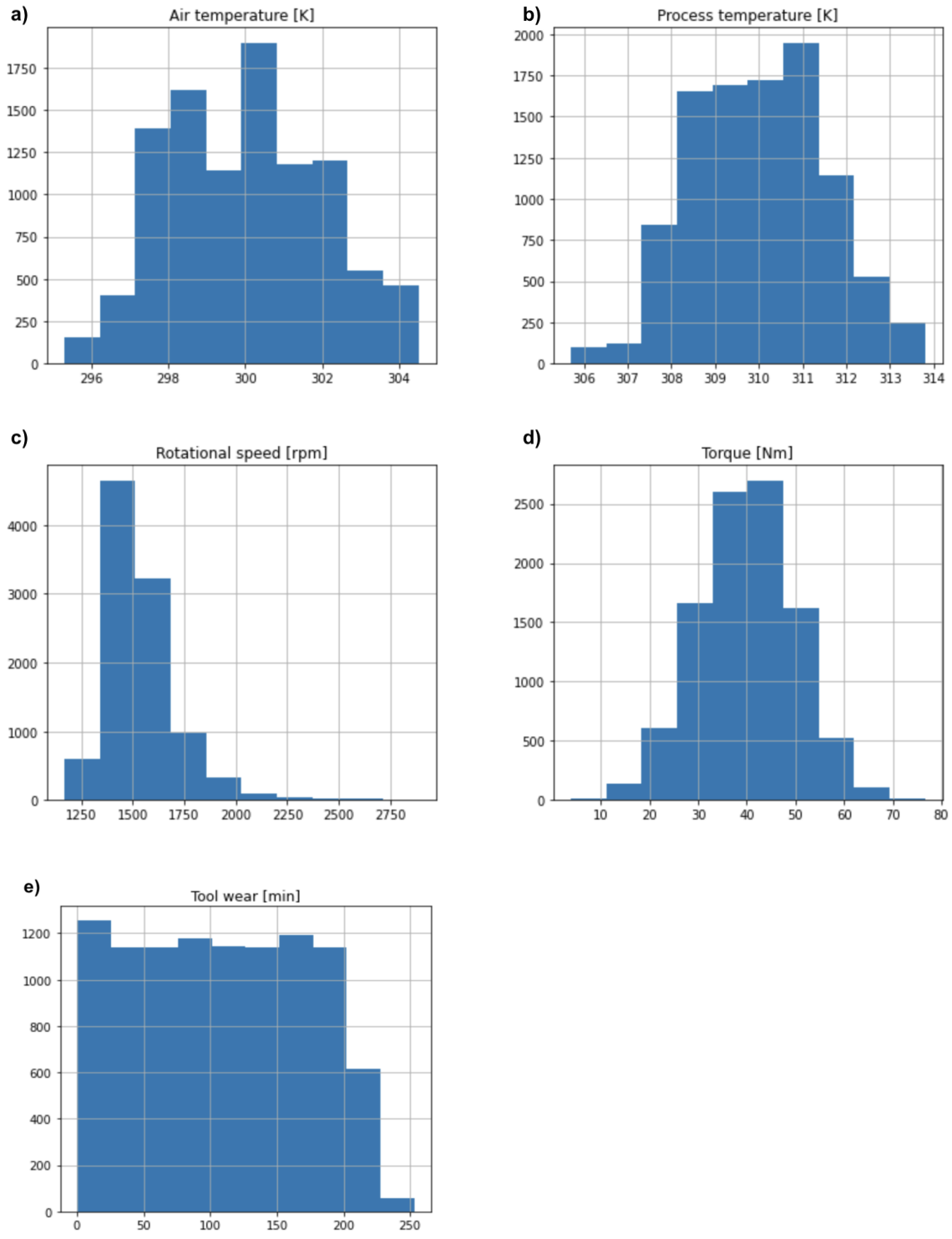

Fig 1. Histogramas de las características seleccionadas para el análisis de machine learning

En la figura 1 se ilustran los histogramas de las características identificadas en el dataset. Los histogramas a), b) y d) correspondientes a Air Temperature, Process Temperature y Torque respectivamente, presentan una distribución aproximadamente normal. En cuanto a los histogramas c) y e) correspondientes a Rotational Speed y Tool Wear,

presentan una distribución que no parece normal, por lo tanto puedan requerir un preproceso antes de ingresarlas en un modelo de machine learning. Por otro lado, el dataset no presenta datos nulos y es adecuado para los pasos siguientes en el reto.

Posteriormente, decidimos generar un DataFrame donde pudiésemos transformar las variables originales en variables categóricas o dicotómicas, con el objetivo de hacer predicciones en un posterior momento.

Entrando en materia del reto, la primera parte consistió en generar un reporte de información con ciertos tratamientos a la base de datos. Primero agrupamos las máquinas por tipo y por tipo de fallo, lo que permite tener una visualización más específica de cuáles máquinas, cuántos y qué tipo de fallo presentan. Una muestra del resultado es lo siguiente:

Type	Failure Type	UDI	Product ID	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Target
H	Heat Dissipation Failure	8	8	7	7	8	8	7	1
	No Failure	979	979	86	74	456	341	226	2
	Overstrain Failure	1	1	1	1	1	1	1	1
	Power Failure	5	5	5	5	5	5	5	1
	Random Failures	4	4	4	4	4	4	4	1
	Tool Wear Failure	6	6	6	5	6	6	5	1
L	Heat Dissipation Failure	74	74	22	21	48	65	60	1
	No Failure	5757	5757	93	82	803	488	241	2
	Overstrain Failure	73	73	44	38	61	60	37	1
	Power Failure	59	59	42	36	57	51	50	1
	Random Failures	12	12	11	11	12	12	12	1
	Tool Wear Failure	25	25	24	20	24	25	19	1
M	Heat Dissipation Failure	30	30	15	18	26	29	28	1
	No Failure	2916	2916	92	82	676	456	232	2
	Overstrain Failure	4	4	4	4	4	3	3	1
	Power Failure	31	31	30	24	31	29	30	1
	Random Failures	2	2	2	2	2	2	2	1
	Tool Wear Failure	14	14	12	13	14	14	13	1

Fig 2. Listado de máquinas organizadas por tipo y tipo de falla.

Se puede apreciar que es considerablemente mayor el número de máquinas que no presentan fallas; el tipo de máquina L tiene 6000 observaciones de las cuales 5757 no presenta fallo alguno.

Por otro lado, se generó un listado de las máquinas, ordenándose por la cantidad de rpm. Se puede apreciar que la menor cantidad de rpm es 1168 rpm y el máximo es 2886 rpm.

Product ID	Rotational speed [rpm]
M15854	1168
L55617	1181
M21264	1183
L50764	1192
L54943	1200
M22145	1202
L54452	1202
M16830	1202
L49796	1207
L47609	1208

Product ID	Rotational speed [rpm]
L48964	2886
L47643	2874
L47230	2861
L48027	2833
L51476	2825
L50549	2760
L48571	2737
L48275	2721
M22857	2710
M18727	2709

Fig 3. Listado de máquinas organizadas por velocidad de rotación

Como tercer paso, se solicita realizar un filtro que muestre la cantidad de máquinas con fallas que tienen una temperatura en proceso mayor a la indicada por algún usuario, en este caso ingresamos 280 K; miremos el código y el resultado:

```
1 temp_user=int(input('Ingrese temperatura en [K]:'))
2 df[['Product ID']][(df['Process temperature [K]'] > temp_user)&(df['Failure Type'] != 'No Failure')]
```

```
Ingrese temperatura en [K]:280
```

	Product ID
50	L47230
69	L47249
77	L47257
160	L47340
161	L47341
...	...
9758	L56938
9764	L56944
9822	L57002
9830	L57010
9974	L57154

348 rows × 1 columns

Fig 4. Listado de máquinas que presentan una falla a la temperatura de proceso indicada por el usuario.

En la figura 4 se puede observar la lista de máquinas que presentaron una falla a la temperatura de 280K. También se puede ver que un total de 348 máquinas fallan a esta temperatura de proceso.

Para el cuarto paso fue solicitada la cantidad de fallos por tipo de máquina, y el tipo de máquina que más fallas presentó. Sobre el dataset se realizó un filtrado usando la característica de tipo de fallo, posteriormente se agrupó por tipo de máquina y se encontraron los resultados presentados en la tabla 1.

Tabla 1. Fallas por tipo de máquina

Tipo de máquina	**Número de Fallas**
H	24
L	243
M	81

En la tabla 1 presenta la cantidad de fallas presentadas por cada tipo de máquina. Se puede apreciar que la máquina H, L y M presentaron 24, 243 y 81 fallas respectivamente, siendo la máquina tipo L la que mayor número de fallas presentó en el dataset. También fue requerido encontrar el número promedio de temperatura de aire en el cual las máquinas presentan fallas por tipo disipación de calor, se encontró que el promedio de temperatura de aire es de 302 K para las máquinas que presentan fallas por disipación de calor.

Posteriormente fue requerido encontrar la la temperatura de proceso promedio, mínima y máxima por tipo de máquina, para lo cual se aplicó el método groupby y agg para encontrar las temperaturas pedidas.

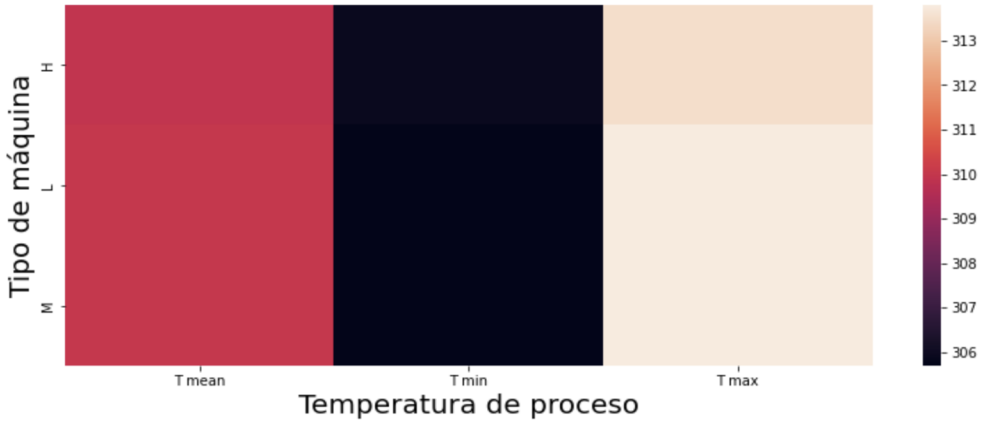

Fig 5. Temperatura de proceso media, mínima y máxima por tipo de máquina

La figura 5 ilustra un mapa de calor con las temperaturas de proceso media, mínima y máxima por tipo de máquina. Se puede observar que la temperatura de proceso media, mínima y máxima es similar para todos los tipos de máquina, solo es posible apreciar una pequeña diferencia en la temperatura de proceso máxima para la máquina tipo H, la cual corresponde a una diferencia de 0.3% levemente apreciable en el mapa de calor.

De igual forma que en el paso anterior, se realizó un agrupamiento usando el método groupby para obtener el porcentaje de tipos de máquinas presentes en el dataset.

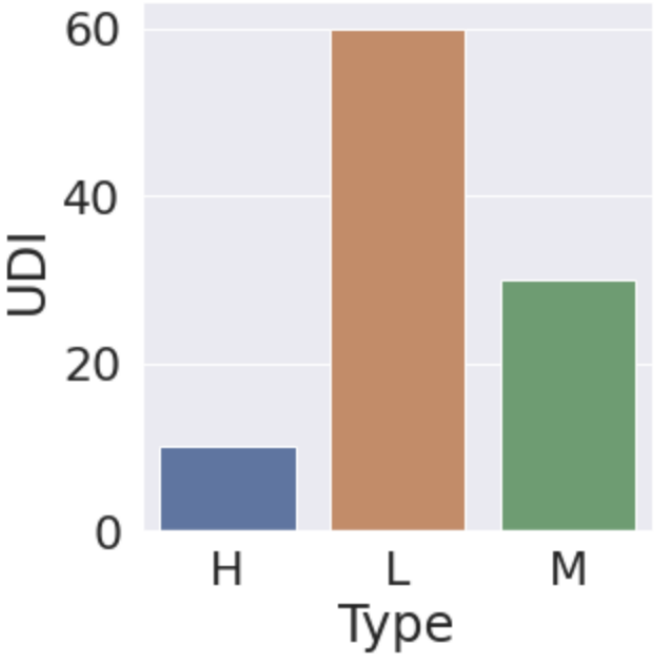

Fig 6. Porcentaje de máquinas por tipo.

La figura 6 presenta el porcentaje del tipo de máquinas presentes en el dataset. El tipo de máquina L es la de mayor porcentaje con un 60%, seguido por el tipo M con un porcentaje de 30% y finalmente el tipo H con un 10% de las máquinas.

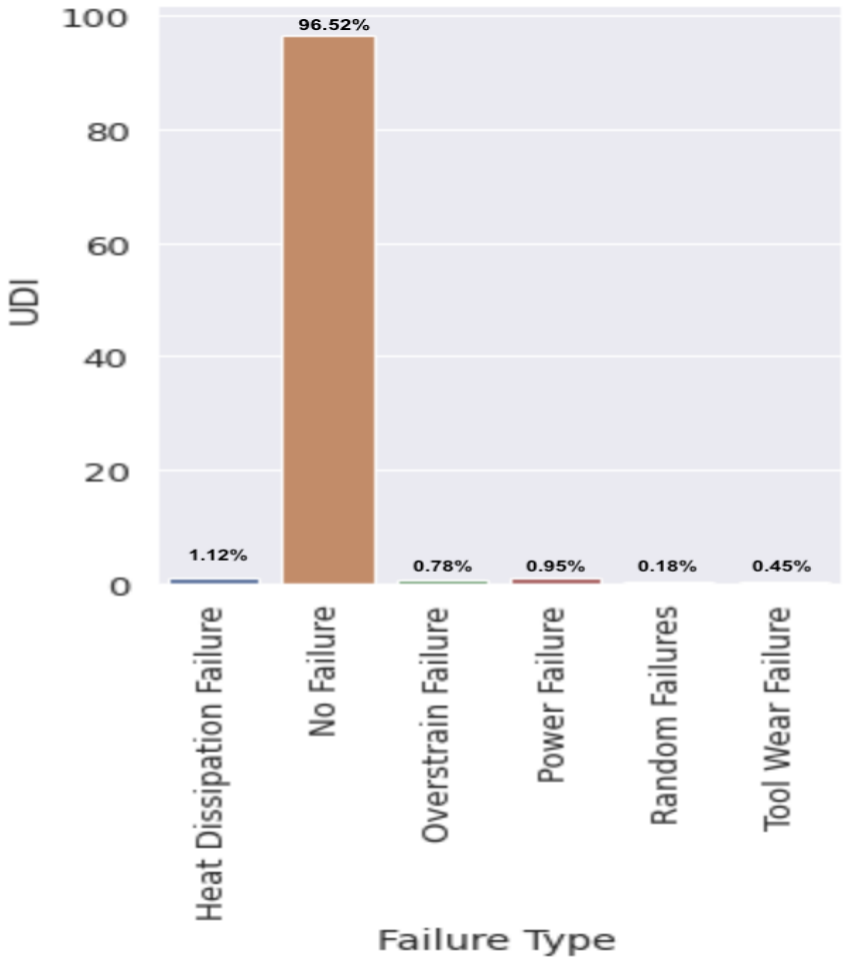

Fig 7. Porcentaje de máquinas por tipo de fallo

Así mismo, se determinó el porcentaje por tipo de falla en las máquinas usando el método groupby. La figura 7 ilustra el porcentaje de máquinas por tipo de fallas, se observa que la mayoría de máquinas no presenta fallas con un 96.52%, siendo menor a el 4% las fallas totales. El tipo de falla con mayor porcentaje es por disipación de calor con un 1.12%, seguido por los fallos por potencia, hipertensión y desgaste, con un porcentaje de 0.95%, 0.78% y 0.45% respectivamente. También se puede apreciar que los fallos aleatorios representan un 0.18% de las fallas.

Por otra parte se calculó la correlación entre las características consideradas importantes, usando el método *corr*. Luego las correlaciones encontradas fueron graficadas en un mapa de calor, presentado en la figura 8.

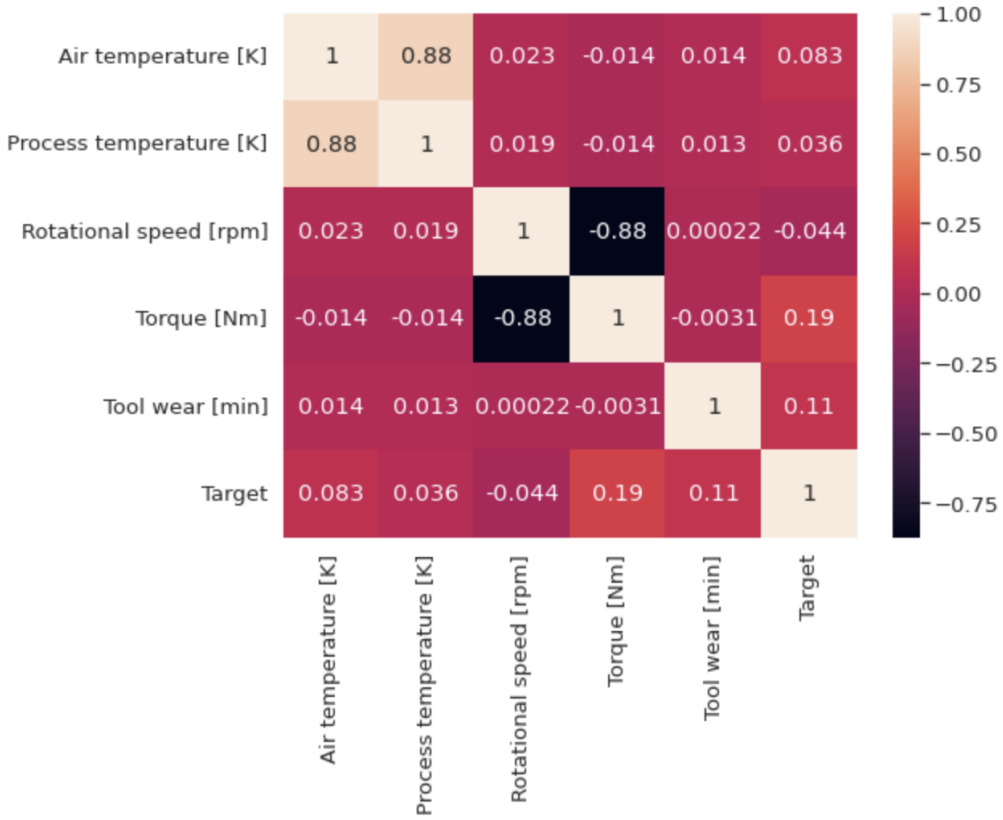

Fig 8. Mapa de calor con correlación entre las características del dataset.

La figura 8 ilustra el mapa de calor de las siguientes características; Air Temperature, Process Temperature, Rotational Speed, Torque y Tool Wear. La variable Target contiene la información sobre si falla o no una máquina, así que corresponde a la variable objetivo para un posterior proceso de machine learning. El mapa de calor presenta la intensidad de correlación, para una correlación negativa los colores son orientados hacia el negro, contrario para una correlación positiva, donde son orientados al colo piel. En la figura 8 se puede observar que las variable objetivo target es altamente correlacionada positivamente con las características Torque y Tool wear, con un valor de 0.19 y 0.11 respectivamente, superando en alrededor de un 75% las correlaciones

con las otras características. En ese orden de ideas también se puede apreciar que las características Air Temperature y Process Temperature presentan una correlación positiva de 0.88, lo cual era de esperarse ya que la temperatura del ambiente y proceso pueden ser similares por equilibrio térmico. También es posible ver una correlación negativa de -0.88 para las características Toque y Rotational speed.

Predicción de fallas (Machine learning)

En el desarrollo de este reto, se realizó un trabajo adicional sobre el dataset de fallas mecánicas, se planteó el uso de dos modelos de machine learning para un problema binario de predicción. La pregunta objetivo planteada fue: *Dada las condiciones de torque, velocidad rotacional, temperatura de aire y proceso de una máquina, ¿Podrá presentar fallas o no?*.
Inicialmente se realizó la selección de las características (x); Air Temperature, Process Temperature, Rotational Speed, Torque y Tool Wear en un nuevo dataframe. También se definió la característica de predicción (Y); Target. Posteriormente se utilizó la función *train_test_split* de sklearn para dividir los datos en dos secciones; la de entrenamiento y validación del modelo. La sección de entrenamiento y validación fueron definidas con un 70% y 30% de los datos respectivamente.

Luego se procede a usar los modelos de regresión logística y Random forest para realizar la predicción, a continuación se presentan los resultados obtenidos:

- Regresión logística:

Este modelo es comúnmente usado para la predicción de variables dicotómicas; es decir variables que pueden tomar dos valores. Para este caso la variable dependiente puede presentar uno de los siguientes valores; (1) para falla y (0) para no falla.

La regresión logística multinomial es una extensión de la regresión logística que agrega soporte nativo para problemas de clasificación de clases múltiples.

La regresión logística, por defecto, está limitada a problemas de clasificación de dos clases. Algunas extensiones como one-vs-rest pueden permitir que la regresión logística se use para problemas de clasificación de clases múltiples, aunque requieren que el problema de clasificación primero se transforme en problemas de clasificación binaria múltiple.

En cambio, el algoritmo de regresión logística multinomial es una extensión del modelo de regresión logística que implica cambiar la función de pérdida a pérdida de entropía cruzada y predecir la distribución de probabilidad a una distribución de probabilidad multinomial para admitir de forma nativa problemas de clasificación de clases múltiples.

```
Accuracy on training set: 0.971
Accuracy on test set: 0.967
precision score: 0.833
recall score: 0.212
f1 score: 0.338
```

Fig 9. Métricas de la regresión logística

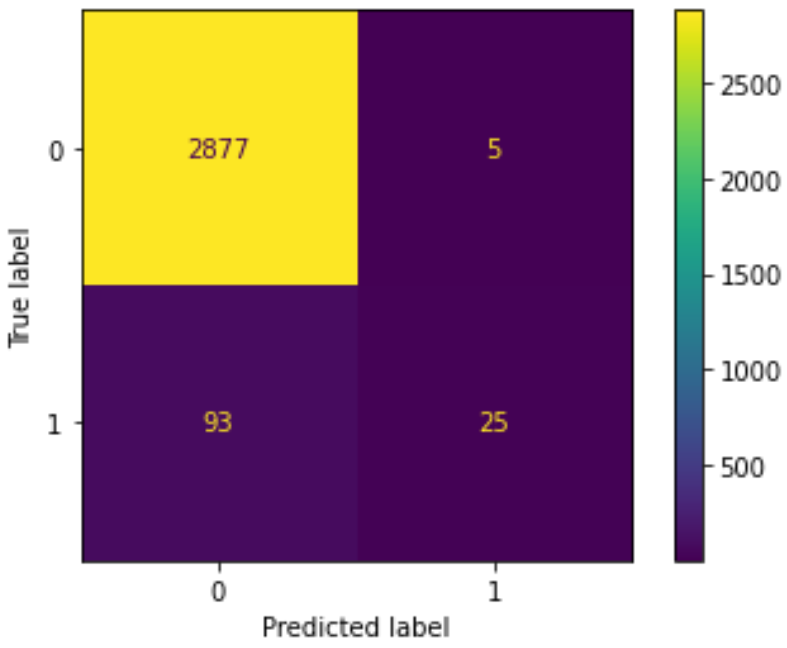

Fig 10. Matriz de confusión del modelo de regresión logística

La exactitud o “Accuracy” (AC) se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones. Lo que quiere decir que nuestro modelo acierta en el 96.7% de las veces.

Para contrastar la métrica de exactitud, ya que este depende en gran medida del balance de las clases, se aplicaron métricas como la precisión, la exhaustividad y el valor F1. Con la métrica de precisión podemos medir la calidad del modelo de machine learning en tareas de clasificación; nuestro modelo tiene una precisión del 83.3% lo que indica que clasifica de buena manera las máquinas que en realidad fallarán.

La métrica de exhaustividad (recall) nos va a informar sobre la cantidad que el modelo de machine learning es capaz de identificar ¿Qué porcentaje de casos positivos fueron capturados? La regresión sólo captura el 21.2% de los casos positivos. El valor F1 se utiliza para combinar las medidas de precisión y recall en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones. El modelo cuenta con un valor F1 del 33.8%

- Random forest:

RF es una versión de árboles de bagging decorrelacionados, esto se logra introduciendo variabilidad en la construcción de los árboles. El proceso de decorrelación de bosques aleatorios consiste en que cada vez que se hace un corte en un árbol de bagging, se escoge al azar un número de variables y se usan estas para buscar la mejor variable y el mejor punto de corte, como se realiza en la construcción de árboles.

El proceso de bagging se basa en el hecho de que, promediando un conjunto de modelos, se consigue reducir la varianza. Esto es cierto siempre y cuando los modelos agregados no estén correlacionados. Si la correlación es alta, la reducción de varianza que se puede lograr es pequeña.

Random Forest evita este problema haciendo una selección aleatoria de m predictores antes de evaluar cada división. De esta forma, un promedio de (p−m) /p divisiones no contemplarán el predictor influyente, permitiendo que otros predictores puedan ser seleccionados. Es por esto que con la decorrelación de árboles se consigue una mayor reducción de la varianza.

```
Accuracy on training set: 0.983
Accuracy on test set: 0.979
precision score: 0.854
recall score: 0.418
f1 score: 0.562
```

Fig 12. Métricas de Random Forest

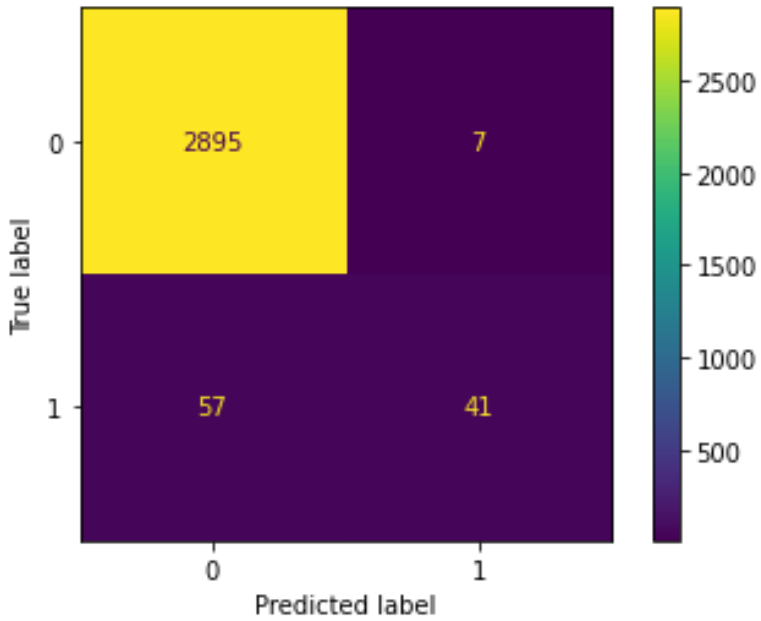

Fig 13. Matriz de confusión del modelo de Random Forest

Para el modelo RF se decidió incorporar una variable más, el tipo de máquina, se mapeó de 0 a 2 para poderla incluir. Se aprecia que las métricas aplicadas tuvieron mejora para este modelo en todos los indicadores, especialmente en en recall y el valor F1. También se puede

evidenciar que el error tipo I aumentó con respecto a la regresión mientras que el error tipo II disminuyó en una gran proporción.

Los valores de las métricas son: el modelo cuenta con una exactitud del 97.9%, una precisión del 85.4%, una exhaustividad del 41.8% y un valor F1 del 56.2%. Lo que se puede interpretar como que el modelo Random Forest es un mejor clasificador de las máquinas que van a fallar. En otras palabras, el modelo acierta en su predicción el 97.9% de las veces, el 85.4% de las veces que predice un fallo la máquina falla e identifica el 56.2% de las máquinas que van a fallar.