TP Final

Inferencia Estadística y Reconocimiento de Patrones

UNAB

Análisis de datos operativos y de mantenimiento de maquinas industriales

Raul Marusca

Diciembre 2021

Índice general

$\frac{1}{2}$

Índice de figuras

CAPÍTULO 1

Introducción

A pedido de nuestro cliente, en el presente informe se analizara un conjunto de datos de mantenimiento de equipos mecánicos en una industria.
El conjunto consta de 10000 mediciones de parámetros de funcionamiento de los equipos que incluyen 10 valores, a saber:

- UDI - Identificador de la muestra
- Product ID - Identificador de la maquina
- Type - Calidad de la maquina (Buena, regular y mala)
- Air temperature [K] - Temperatura promedio del entorno en kelvins
- Process temperature [K] - Temperatura del área de proceso en Kelvins
- Rotational speed [rpm] - Velocidad de rotación de la herramienta en revoluciones por minuto
- Torque [Nm] - Torque de la herramienta en Newton por metro
- Tool wear [min] - Tiempo en el que se desgasta la herramienta en minutos
- Target - Indica si el equipo fallo o no
- Failure Type - Tipo de falla encontrada

De estos valores solo tomaremos algunos descartando los identificadores de muestra y de maquina en particular. Tampoco tomaremos en cuenta para este análisis el tipo de falla encontrada. Solo nos interesa determinar que condiciones operativas pueden ser útiles para predecir que el equipo falle, no la falla especifica.

También vamos a convertir en una condición numérica los datos de la calidad de la maquina y el del target que en el dataset figuran como categóricos.
El objetivo de este análisis es el de evaluar algunos modelos predictivos (con sus posibles parametrizaciones) pera prever fallas en otros equipos.
A nuestro cliente le interesa conocer con antelación que equipos van a fallar, de manera de organizar anticipadamente los cambios necesarios para mantener la continuidad de la producción. Todos los modelos predictivos nos pueden sugerir que un equipo va a fallar cuando en realidad no lo hace (A esta condición la llamamos **F**also **P**ositivo). Y predicen que un equipo no va a fallar cuando si lo va a hacer (**F**also **N**egativo).
Este ultimo numero es fundamental para la continuidad de la producción, ya que una maquina que falla sin preaviso implica demoras, costos adicionales e incluso hasta cambios en la distribución de los turnos laborales.

En la segunda parte de este análisis intentaremos contestar la pregunta del cliente sobre si debería agrupar sus equipos de acuerdo con alguna o algunas características operativas que se pueda desprender de la información brindada en el conjunto de datos sobre el funcionamiento de esos equipos.
Es de interés del cliente conocer cuales son los agrupamientos de condiciones de operación de los equipos de manera de poder capacitar correctamente al personal de mantenimiento para así acelerar los tiempos de mantenimiento y facilitar el servicio preventivo.

Debemos hacer notar que el dataset adolece de ciertas informaciones tales como el nivel de vibración de la maquina, del consumo eléctrico y de su factor de potencia. Información que, en nuestra experiencia, indica ser una mejor medida de la probabilidad de falla de los equipos electromecánicos.

CAPÍTULO 2

Análisis exploratorio

Un detalle que surge a primera vista y que se puede evidenciar en la figura 2.1. de este conjunto de datos es el de que las fallas son de rara ocurrencia. Del total de 10000 muestras solo se presentan fallos en 339 de ellas. Esto representa tan solo un 3.4 %

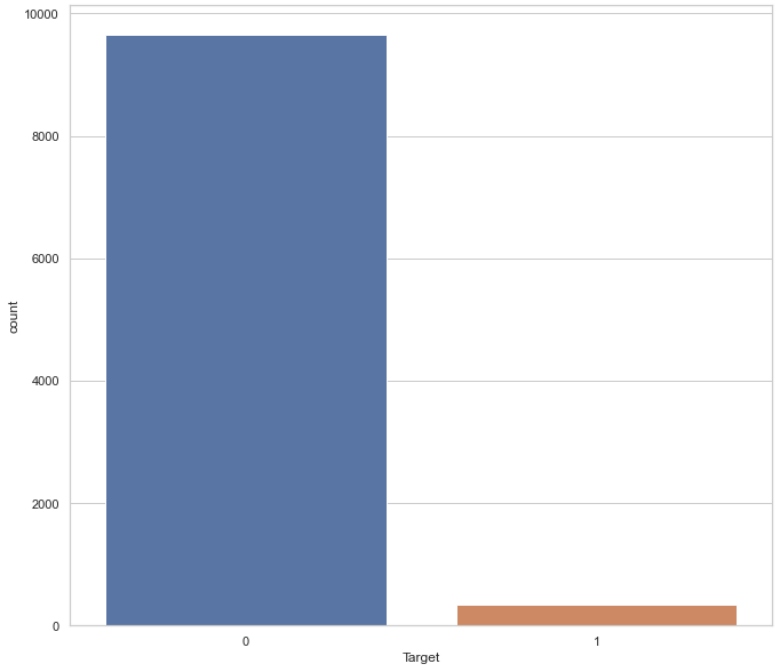

Figura 2.1: Relación entre equipos con fallas y sin fallas

Esto complica la predicciones ya que reportando que la maquina no va a fallar, acertaríamos el 96.6 % de las veces, sin realizar ningún análisis de ninguna variable operativa del equipo.

Un subsiguiente análisis que podemos hacer es el de interpretar como afectan todas las variables de entrada a el objetivo. Para comprender eso necesitaremos relacionar gráficamente 6 dimensiones. Como esto no podemos hacerlo en nuestra vida tridimensional, aplicaremos una técnica matemática de reducción de dimensiones llamada **PCA** (*Principal Components Analisys*)
Este método recorta la cantidad de información presentada, y en este caso nos permite representar el 62 % de la información aproximadamente con una reducción a dos dimensiones. Esto se puede ver en la figura 2.2

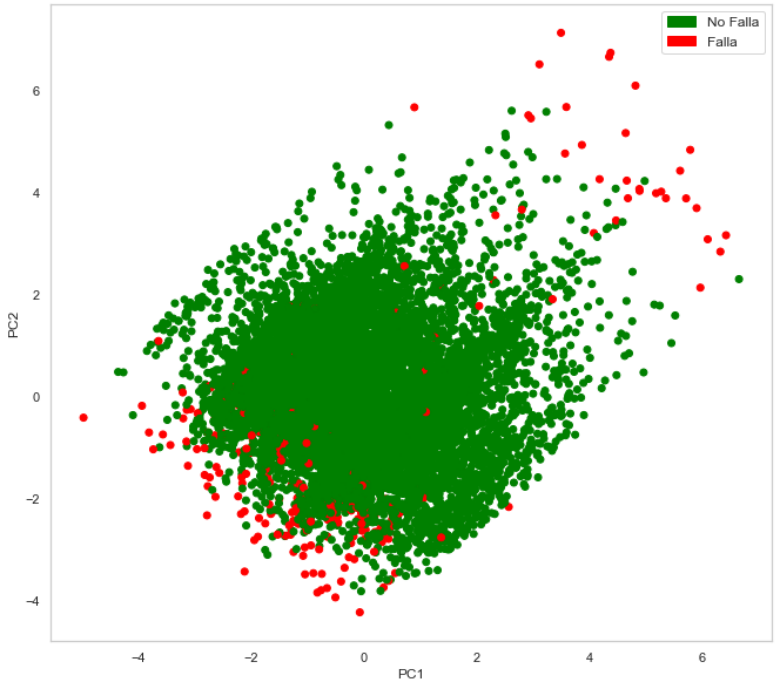

Figura 2.2: Análisis PCA

Se nota que hay un núcleo central de equipos funcionales y que los equipos que presentan fallas se concentran mayormente en los extremos diagonales. Esto es coincidente con la idea de que los equipos deberían fallar en condiciones extremas. Aun así se notan algunos casos donde eso no es así (los puntos rojos desperdigados dentro del grupo de puntos verdes)
Dado que nos interesa saber cuales serán las variables cuyos valore extremos provocaran las fallas, debemos **proyectar** las variables medidas sobre el plano del análisis PCA
Esta proyección se muestra en la figura 2.3

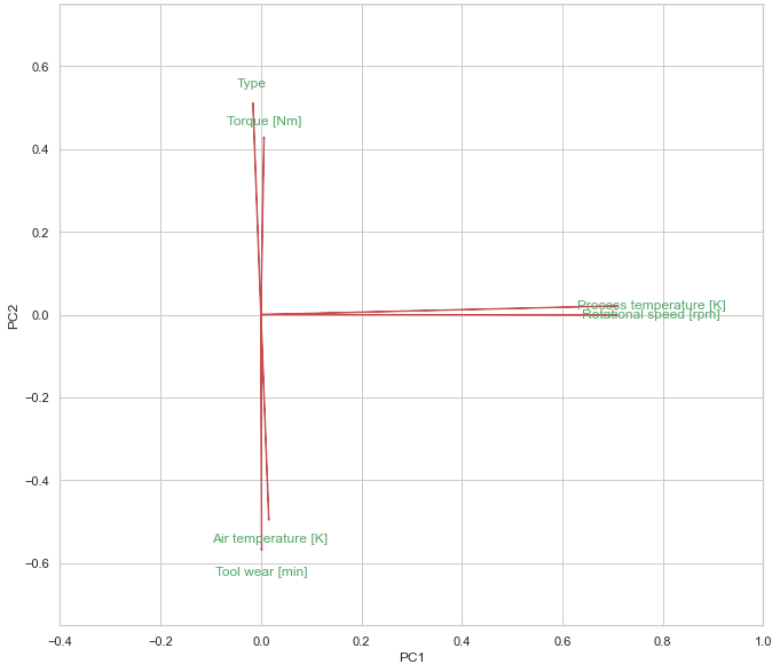

Figura 2.3: Proyección de variables originales sobre el plano PCA

Estas variables se alinean fuertemente con los ejes PCA. Pero si recordamos la figura 2.2 tendremos que los puntos rojos donde ese acumulaban las fallas estaban mas bien siguiendo un eje diagonal.

CAPÍTULO 3

Conclusiones del análisis exploratorio

Partiendo de una representación que incluye a mas del 60 % de la información que podemos extraer del dataset concluimos que:

- El porcentaje de fallas es bajo, por lo que debemos concentrarnos en encontrar un modelo que tenga una baja probabilidad de predecir como NO FALLA a un equipo que si vaya a fallar.
- No podemos determinar con claridad cual es la variable informada que mas incidencia tenga en la falla del equipo. Esto se ve en el gráfico donde todas las variable proyectadas tienen participación en la distribución de los puntos donde se encuentran fallas.
- Se evidencia también que los valores extremos de condiciones de funcionamiento son los que tienden a resultar en equipos con fallas.
- Del diagrama PCA se desprende que todos los valores medidos participan en el estado de falla o no falla. Esto dificulta la predicción de ese estado.

CAPÍTULO 4

Clasificación supervisada

Sobre este conjunto de datos de entrada vamos a evaluar algunos modelos predictivos básicos.
Cada uno de estos modelos evalúa distintas relaciones entre las variables de entrada. Y predice una probabilidad de que el equipo falle o no. El valor de corte por defecto de esa predicción es del 50 %.
Pero este valor se puede modificar y podemos entonces ajustar esa probabilidad e ir planteando distintos escenarios. Para mostrar como se adaptan esos escenarios a nuestras necesidades iremos mostrando una tabla donde enfrentaremos los valores predichos contra los que ya figuran en el *dataset*. Prestemos atención a los números de **F**alsos **N**egativos y **F**alsos **P**ositivos en cada figura.

4.1. Escenario 1

El primer escenario nos muestra en la figura 4.1 que la predicción de equipos con fallas es por debajo del promedio del total de la muestra (De un total de 2000 muestras predice correctamente 17 - lo que es un 0.85 %) El resto de los equipos fallados se concentran en los FN.

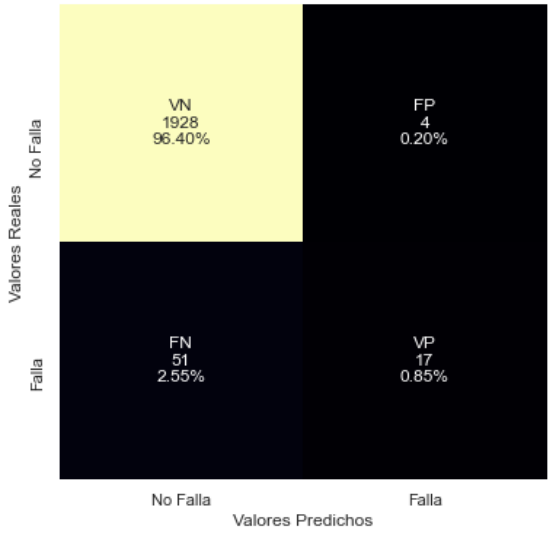

Figura 4.1: Matriz de confusión Escenario 1

4.2. Escenario 2

Ahora probamos cambiando la probabilidad de falla a un numero mas elevado y, en la figura 4.2, se observa que aunque mejore el numero de FN también se perjudican las otras condiciones.

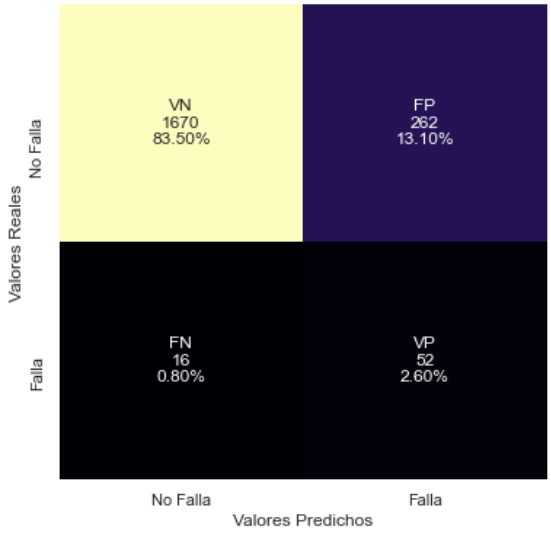

Figura 4.2: Matriz de confusión Escenario 2

4.3. Escenario 3

En este escenario empleamos otra técnica de predicción, sus resultados se pueden observar en la figura 4.3

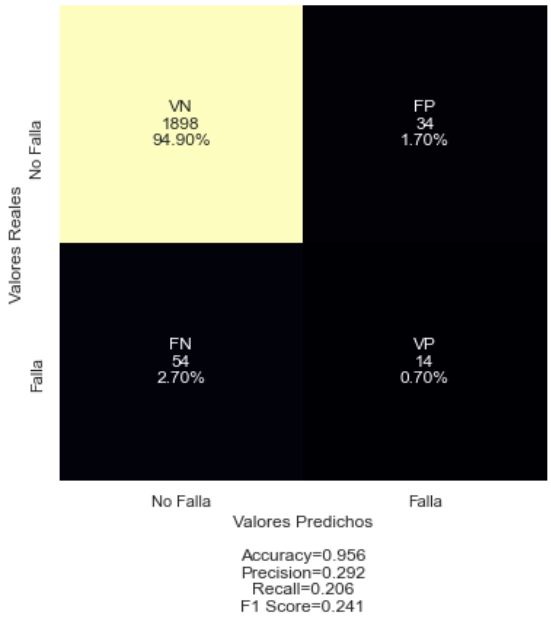

Figura 4.3: Matriz de confusión Escenario 3

4.4. Escenario 4

Usando esta misma técnica y modificando sus parámetros obtenemos los resultados en la figura 4.4

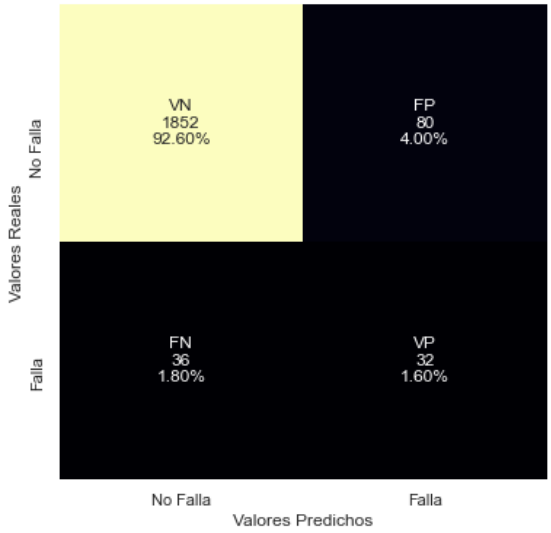

Figura 4.4: Matriz de confusión Escenario 4

4.5. Escenario 5

Por ultimo consideramos otro modelo predictivo como se vera en el Apartado Técnico (Modelo 3 Arbol de decisión), requiere de consideraciones especiales para su funcionamiento.
Los resultados que obtenemos eligiendo adecuadamente los parámetros del modelo se pueden observar en la figura 4.5

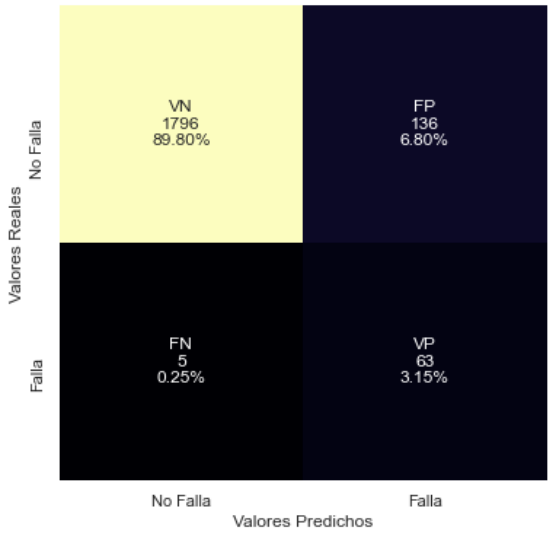

Figura 4.5: Matriz de confusión Escenario 5

CAPÍTULO 5

Conclusiones de la clasificación supervisada

De la aplicación de los métodos predictivos aquí consideramos podemos concluir que para atender el requisito del cliente de reducir al máximo el numero de FN (Esto seria el caso en el que el negocio tiene continuidad) el mejor de ellos se muestra en el Escenario 5 aunque esto hace incurrir al cliente en gastos de mantenimiento mas elevado que en los demás escenarios, ya que lo obliga a realizar mayores mantenimientos preventivos en los equipos sobre los que se predice falla pero que resulta siendo un Falso Positivo (que seria, sumando el de los Verdaderos Positivos en el 9.95 % del total de las maquinas)

Si, por otro lado, se prefiriera utilizar el escenario 4 el mantenimiento preventivo se debería realizar sobre solo el 5.6 % de las maquinas y se estaría teniendo un 1.8 % de falla de continuidad de la producción.
A continuación se presentan estos resultados en forma de tabla.

	Caída Continuidad	Mantenimiento
Escenario 1	2.55 %	1.05 %
Escenario 2	0.8 %	15.7 %
Escenario 3	2.7 %	2.4 %
Escenario 4	1.8 %	5.6 %
Escenario 5	0.25 %	9.95 %

Queda a cargo del cliente la decisión de el método definitivo de predicción de acuerdo a sus mejores interese económicos y productivos.

Si solo importara la continuidad de la producción el escenario 5 seria la mejor opción

CAPÍTULO 6

Clasificación no supervisada

En este capitulo del informe aplicaremos sobre el conjunto de datos una aproximación desde un punto de vista diferente. Como en este caso al cliente le interesa clasificar que diferentes tipos de fallos se puede encontrar y de que manera puede agruparlas por sus características, emplearemos técnicas de *clustering* que nos permitirán determinar cuantos y cuales son los valores que categorizan los distintos tipos de fallos.
Esto se hace con el objetivo de capacitar de forma adecuada al personal de mantenimiento de manera de prevenir fallos y de acelerar los tiempos de reparación.
Para ello vamos a prescindir de las columnas que nos reportan la condición de falla y el tipo de falla de cada equipo (además de las que identifican a la muestra, columnas que ya habíamos retirado del *dataset*.
Por lo tanto el conjunto de valores a considerar quedara limitado a:

- Type - Calidad de la maquina (Buena, regular y mala)
- Air temperature [K] - Temperatura promedio del entorno en kelvins
- Process temperature [K] - Temperatura del área de proceso en Kelvins
- Rotational speed [rpm] - Velocidad de rotación de la herramienta en revoluciones por minuto
- Torque [Nm] - Torque de la herramienta en Newton por metro
- Tool wear [min] - Tiempo en el que se desgasta la herramienta en minutos

Un paso importante es determinar el numero de *clusters* en el que podemos agrupar a las maquinas. Con el empleo de ciertas técnicas estadísticas determinamos que el numero mas correcto es el de 2.

6.1. Escenario 6

En este escenario la aplicación del método nos divide a las fallas de la forma que se puede ver en la figura 6.1

Figura 6.1: Resultado clusterizacion escenario 6

Vemos una separación bastante clara que corta al grupo perpendicularmente a su eje diagonal.
De los resultados del método matemático usado podemos ver que esa separación esta fuertemente indicada por los equipos que tienen altos rpm y bajo torque en color celeste, y los que tienen la situación contraria en marrón. Aunque se detecta también cierta participación del desgaste de la herramienta de corte en los criterios de separación

6.2. Escenario 7

Ahora usaremos otra técnica para definir los criterios de separación en dos cúmulos. Los resultados se trafican en la figura 6.2

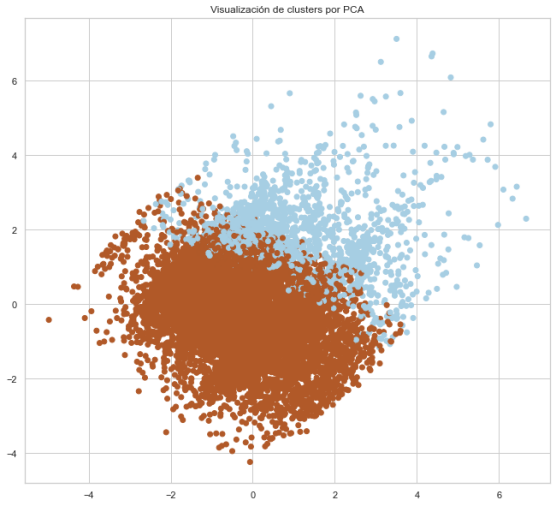

Figura 6.2: Resultado clusterizacion escenario 7

Esta figura muestra prácticamente la misma división que en el caso anterior (con la salvedad de la inversión de colores).
También se puede ver que los criterios para la separación también son los de la velocidad de rotación y el torque aplicado en la herramienta.
La separación entre *clusters* es un poco mas difusa que la que vemos en el escenario 6 pero es porque la participación del resto de las variables es también notoria.

CAPÍTULO 7

Conclusiones de la clasificación no supervisada

De la revisión de los resultados de las características de los cúmulos conformados en ambos escenarios podemos resumir la siguiente tabla:

	RPM herramienta	Torque Herramienta	Otros valores
Escenario 6 cumulo 1	↑	↓	Poca influencia
Escenario 6 cumulo 2	↓	↑	Poca influencia
Escenario 7 cumulo 1	↑	↓	Mayor influencia
Escenario 7 cumulo 2	↓	↑	Mayor influencia

Es claro entonces que el criterio que debemos emplear es el de la relación inversa entre Velocidad de rotación y Torque para diferenciar los diferentes tipos de funcionamiento de las maquinas.
Esto es coincidente con la categorización de los motores eléctricos, que según la forma de conformado de sus bobinados obtienen esta relación inversa.
Un detalle llamativo es el de que los equipos estaban categorizados en tres calidades, pero este parámetro prácticamente no tiene influencia en nuestro análisis de agrupamiento. Evidentemente el tipo de construcción del motor eléctrico es fundamental y sobrepasa cualquier otra característica que hayamos medido en este análisis.
Sugerimos emplear el sistema de clasificación propuesto en el escenario 7 ya que también involucra a las demás variables de funcionamiento

CAPÍTULO 8

Conclusión General

Hemos visto en ambos análisis que el funcionamiento de los equipos se puede categorizar en aquellos que tienen alto torque a bajas revoluciones o los que tienen bajo torque a altas revoluciones. Y que las fallas mayoritariamente se concentran cuando estos valores son extremos (Es decir que hay fallos a bajos RPM's/Altos torques y cuando tenemos bajos torques/altos RPM's.
Esto significa que el personal de mantenimiento debe centrar su atención en estos casos, intensificando la vigilancia y el mantenimiento preventivo en esos casos, recibiendo la capacitación adecuada.
De esta forma se asegurara la continuidad de la producción y el ordenamiento de los turnos laborales, reduciendo costos y aumentando la producción.

CAPÍTULO 9

Futuros trabajos

9.1. Clasificación supervisada

Como continuación a este análisis se debería evaluar los tipos de fallas puntuales que se presentan en los equipos. Y determinar cuales condiciones operativas (variables de entrada) son las que mas influyen en cada tipo de falla en particular.
También se deberían seguir evaluando otros modelos de clasificación supervisada y si alguno de ellos no provee de mejores predicciones respecto de los casos donde el equipo puede fallas, que son los que el cliente quiere reducir procediendo a mantenimientos preventivos mas exhaustivos. Pero fundamentalmente queda por determinar si podemos combinar los resultados de los dos modelos aquí tratados de manera tal que se reduzca en numero de Falso Negativos y que también se reduzca el numero de Falsos Positivos empleando una técnica llamada de *stacking*

9.2. Clasificación no supervisada

Debido a que se podría haber elegido una separación entre 3 *clusters* dado que la métrica Silhouette no era muy diferente que para solo 2, un futuro trabajo a realizar seria el análisis con ese numero de *clusters*. Pero debemos recordar que el objetivo de esta clasificación no supervisada era el de capacitar al personal técnico de manera de obtener mejoras en los tiempos de reparación y mantenimiento. Por lo tanto debería ser una decisión consensuada con los especialistas del área.
Aunque debemos hacer notar que la división en 2 cúmulos es bastante coincidente con el análisis previo de las condiciones de funcionamiento y coincidente con las causas probables de falla.

CAPÍTULO 10

Apéndice técnico

10.1. Métodos

Para los métodos empleados en la clasificación supervisada se separa del conjunto de datos una cantidad de muestras con las cuales se *entrena* al modelo. Y con el resto de las muestras se realiza la verificación del mismo. Las predicciones realizadas con este conjunto de *test* se contrastan contra los resultados reales y de esa forma obtenemos las matrices de confusión. Estas matrices sirven para realizar ciertas métricas del modelo (de las cuales la mas aceptada actualmente como indicadora de la fiabilidad del mismo es la de **F1 score**). Pero cada caso es distinto y cada análisis se puede fundamentar en solo alguna de ellas.
En nuestro caso la usada fue la de **Recall** que nos mide la cantidad de Falsos Negativos, valor que interesaba al cliente.
Otro punto a tener en cuenta aquí, es que el modelo da una probabilidad de Fallar y otra de No Fallar. La decisión final se toma eligiendo un punto de corte de esa probabilidad ajustándola a la situación real de la aplicación.

10.1.1. Modelo 1 Regresión Logística

Este modelo empleado en el escenario 1 con su punto de corte por defecto y en el escenario 2 con un punto de corte elegido a partir de la siguiente curva 10.1 donde se muestra como varían las métricas del modelo según varia ese punto. 10.1

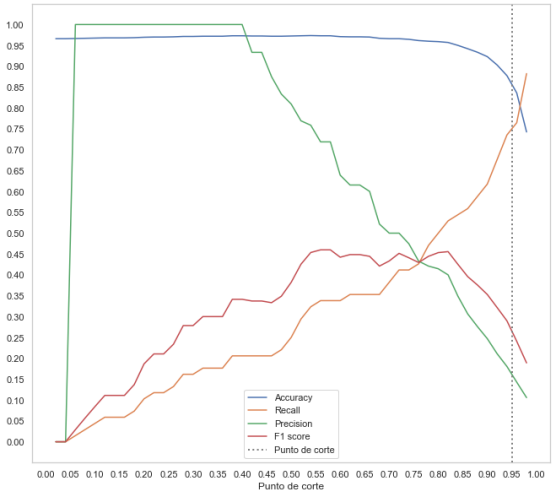

Figura 10.1: Métricas y punto de corte modelo Regresión Logística

Para mas información sobre sobre este modelo consultar en Modelo Regresión Logística

10.1.2. Modelo 2 Bayes Ingenuo

Este método que se basa en una técnica estadística muy conocida se empleo en los escenarios 3 y 4. En este ultimo caso fue con un punto de corte que se decidió en función de la siguiente figura 10.2

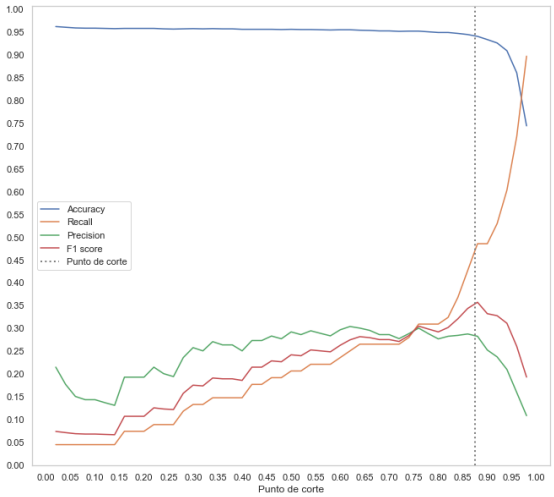

Figura 10.2: Métricas y punto de corte modelo Bayes Ingenuo

Para mas informacion del modelo referirse a modelo Bayes Ingenuo

10.1.3. Modelo 3 Arbol de decisión

Este modelo empleado en el escenario 5 se basa en realizar una serie de comparaciones sobre cada valor hasta lograr determinar que características identifican a cada muestra. Como tambien nos brinda una probabilidad para cada resultado podemos ver como varian las diferentes metricas en el siguiente grafico 10.3

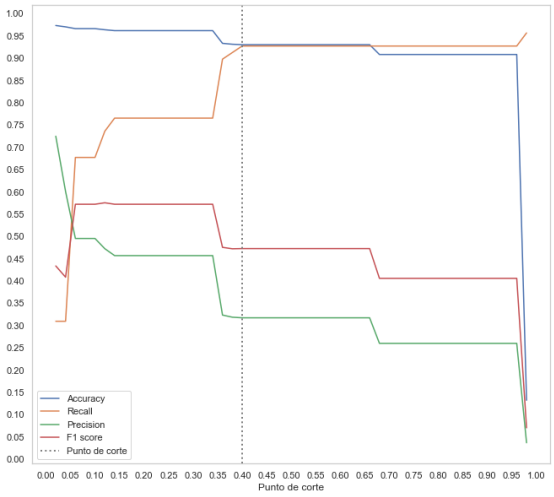

Figura 10.3: Metricas y punto de corte modelo Árbol de decisión

Como el punto de corte ideal provee las misma métricas que el por defecto, se uso este ultimo.
Para profundizar en este método se puede visitar Que es el Modelo de Árbol de Decisión y para que sirve

10.1.4. Modelo 4 K-means y Modelo 5 Aglomerativo

Estos métodos son utilizados en el capitulo de clasificación no supervisada son técnicas que tratan de agrupar muestras por similitud.
K-mean elige ciertos puntos al azar y agrupando los puntos cercanos en cúmulos cada vez mayores. Una vez que llego a todos los puntos ubica el centro de cada cumulo y toma esa posición como nuevo punto de inicio y repite el algoritmo hasta que se llegue a un numero de repeticiones o hasta que ya no haya cambios en la composición de los cúmulos.
El método aglomerativo inicialmente considera a cada punto como un núcleo y va agrupando los mismos uniendo estos grupos en otro mayor. Esto se sigue haciendo hasta llegar al numero de agrupamientos deseado.
Ambos métodos requieren de una forma de medir la distancia entre puntos (o entre puntos y cúmulos) y estos son los parámetros de configuración del modelo. En nuestro análisis hemos empleado la distancia euclidiana en K-means (es así por defecto) y la distancia çomplete"(que toma la mayor distancia posible entre cada muestra o entre

cúmulos) en el aglomerativo.
Otro paramentar que necesitan estos métodos es el numero de clusteres buscados. Para ello se emplea la técnica de la Silueta 10.2.3

10.2. Herramientas

10.2.1. PCA

En nuestro análisis podemos observar que tenemos 6 mediciones distintas en cada muestra. Se dice entonces que nuestro problema tiene 6 dimensiones. En teoría nosotros tendríamos que poder como interactúan esas 6 distintas variables, pero nuestro razonamiento esta limitado las 3 dimensiones de nuestro espacio. Para tratar de resolver esta situación se han desarrollado técnicas que *reducen las dimensiones*.
La que usamos para nuestros gráficos se denomina PCA (por las iniciales en ingles de *Análisis de Componentes Principales*. Esta técnica busca en las 6 dimensiones de nuestros datos unos ejes sobre los cuales al proyectar los datos, se proyecte la mayor variabilidad de los mismos.
Luego elegimos los dos primeros (que en nuestro caso acumulan aproximadamente un 62 % de esa variación) y los presentamos en un plano.
De esta manera tenemos una representación gráfica que podemos entender y que no nos oculta demasiada información.

10.2.2. Stacking

La herrramienta del *stacking* es una manera de combinar diversos modelos de clasificación de menor nivel en uno de mayor nivel que toma en cuenta los resultados de los de primer nivel como datos de entrada.
Esto permite una mejora en las métricas elegidas. Obteniendo un modelo final que posibilita mejores predicciones que los métodos que lo conforman.
Para mas información sobre este método podemos visitar:
Simple Model Stacking, Explained and Automated
Esta técnica no ha sido utilizada en este informe, pero se deja constancia de ella para futuros trabajos

10.2.3. Silhouette

El valor de la silueta es una medida de cuán similar es un objeto a su propio *cluster* en comparación con otros *clusters*.
La silueta va de -1 a +1, donde un valor alto indica que el objeto está bien emparejado con su propio cúmulo y mal emparejado con los cúmulos vecinos. Si la mayoría de los objetos tienen un valor alto, entonces la configuración del cúmulo es apropiada. Si muchos puntos tienen un valor bajo o negativo, entonces la configuración de cúmulos puede tener demasiados o muy pocos cúmulos.

En nuestro ejemplo al probar con un numero de *clusters* entre 2 y 6 nos resulto el siguiente grafico 10.4:

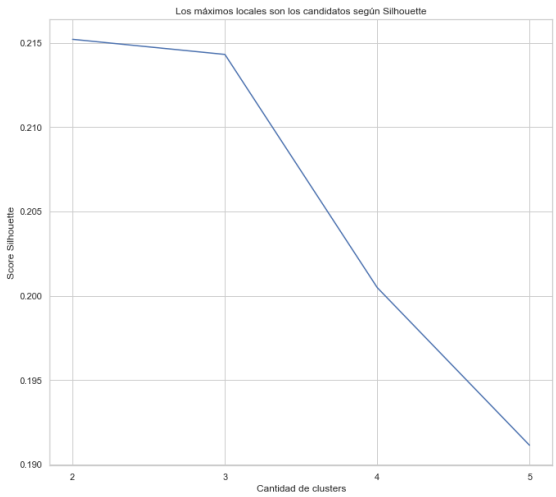

Figura 10.4: Calculo del numero óptimo de clusters por medio de Silhouette

Podemos ver que hay una mínima diferencia de score entre 2 y 3 clusters. El numero de 2 se decidió por el dominio del problema planteado