



1. Documento CRISP-DM

En esta actividad práctica se pretende aplicar de manera básica la metodología CRISP-DM a un caso de uso industrial.

CRISP-DM (Cross Industry Standard Process for Data Mining), es un **modelo de proceso de minería de datos** que describe una manera en la que los expertos en esta materia abordan el tema.

* 1. Objetivo de negocio

Esta es la primera tarea a desarrollar y tiene como objetivo determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar la minería de datos.

En nuestro caso el objetivo de negocio será el **mantenimiento predictivo de los fallos de operación en el funcionamiento de los motores de aviones**.

El mantenimiento predictivo abarca una variedad de temas, que incluyen entre otros, la predicción de fallos, el diagnóstico de fallos (análisis de causa raíz), la detección de fallos, la clasificación de tipos de fallos y la recomendación de acciones de mitigación o mantenimiento después de un fallo.

* 1. Objetivo de minería

Esta tarea tiene como finalidad representar los objetivos del negocio en términos de las metas del proyecto de minería de datos.

Puede ser que mi proyecto no tenga objeto de minería, por ejemplo, si quiero estudiar patrones o la eficiencia energética.

En el caso de nuestros datos no tenemos un campo de incidencias (fallos) como tal por lo tanto no tenemos un objetivo de minería.

Para poder realizar el análisis supervisado generaremos un **atributo nuevo** (campo objetivo) que será Incidencia. En esta actividad práctica el objeto de minería será el **campo de incidencias (fallos) de los motores de aviones.**

* 1. Tipo de proyecto funcional y tecnológico

El tipo de proyecto funcional será de **producto** (motores de avión) porque no buscamos analizar el proceso si no que analizamos el fallo del equipo. El proyecto tecnológico lo dividiremos en dos partes.

* + 1. Proyecto descriptivo

En este proceso se pretende comprobar la existencia de patrones entre sensores que se comportan igual, para poder agrupar y hacer el problema más sencillo.

Para ello, realizaremos en primer lugar una **PCA** (Principal Component Analysis), analizando los datos de entrada, sin objeto de minería, y reduciendo el espacio vectorial. Esto me permitirá ver:

1. La información en un plano 2D

2. La varianza de la información, lo dispersa que está mi información (si es muy dispersa el árbol de decisión será muy grande si no, será más pequeño)

A continuación, realizaremos un **clustering** para agrupar los sensores en grupos en función de su comportamiento. Eso, me permitirá saber dentro de los clusters cuáles se comportan de forma normal y cuáles de forma anómala. Además, me puede ayudar a ver la influencia de los outliers y si tengo que quitarlos para que no afecten a la generalidad de los datos.

Por otra parte, realizar una **correlación** que me dirá cuáles son las variables relevantes con respecto de mi objetivo y así poder reducir mis entradas a las más relevantes.

Una vez hecho el clustering y detectado los outliers haremos el **árbol de decisión**, que puede ser general o por cluster, depende de la varianza, si es muy grande divido el problema en cuatro modelos para que me dé reglas específicas para cada uno de los clusters y tengo más especificidad.

Por último, realizaremos una **validación cruzada**. Consiste en coger el árbol de decisión y aplicarlo al 80 % de mis datos, varias veces y después aplicarlo al 20 % de los datos restantes para que me determine la accuracy (la confianza) de mi modelo.

* + 1. Proyecto predictivo

El objetivo es estimar valores futuros o desconocidos de algunas variables de interés a partir de otras variables independientes (variables predictivas).

En nuestro caso, haremos los mismos pasos que en el caso anterior menos en el tercero que en vez de un árbol de decisión, utilizaremos una **red neuronal.**

* 1. Recursos

Los datos disponibles son los siguientes:

1. **Datos de entrenamiento:** Es el dato de fallo del motor del avión

2. **Datos de prueba:** Son los datos de operación del motor del avión sin eventos de fallo registrados

3. **Datos verdaderos básicos:** contiene la información de los ciclos restantes antes de que el motor de fallo

* 1. Granularidad

Este apartado hace referencia a la unidad mínima de información a analizar. En nuestro caso viene definida en la estructura de los propios datos, por lo que la granularidad será por **motor de avión** representado por el ID.

* 1. Cadencia

El número de veces que se repite la granularidad serán **los ciclos**.

El conjunto de la granularidad más cadencia es la clave de mi dataset. Tengo que tener una única fila por cada combinación de granularidad y cadencia y al resto de datos que tenga excepto el objeto de minería, tienen que estar agregados a la clave. 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ciclo | ID (Máquina) | S1 | S2 | S3 | S21 | Incidencia |
| 1 | 1 | X | X | X | X | Correcto (0) |
| 2 | 1 | X | X | X | X | Fallo (1) |
| 3 | 2 | X | X | X | X | Correcto (0) |

* 1. Presupuesto.

Hardware.

Software.

Infraestructura.

Roles. Jornadas. Esfuerzos.