**TITULO PROYECTO: YO**

**INTEGRANTES: SIMÓN CORREA MARÍN**

1. **ENTENDIMIENTO DEL NEGOCIO**
   1. DESCRIPCIÓN DEL NEGOCIO

No es histórico, misión, visión.

Describir de manera corta de que se trata el negocio (muy especifico\*).

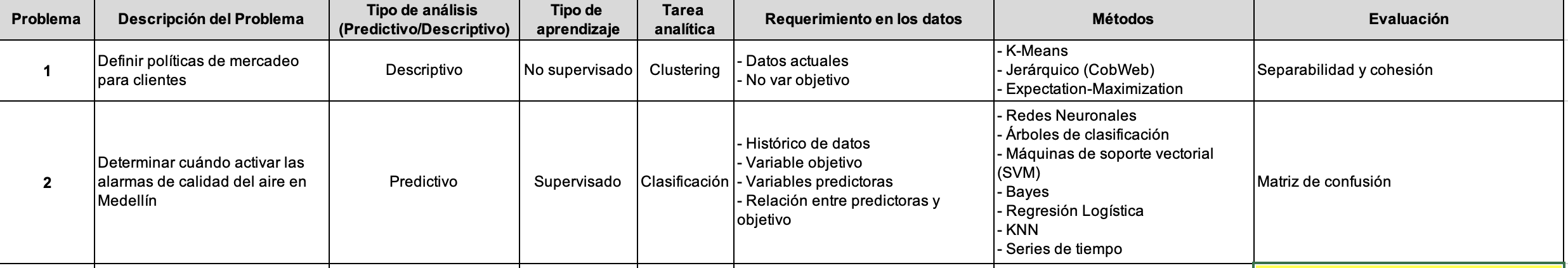
* 1. DESCRIPCIÓN DEL PROBLEMA

Cada proyecto es cerrado a cada problema que se debe resolver, un problema específico que se quiere solucionar, lo da el experto del negocio.

* 1. OBJETIVOS DE LA MINERÍA

Los objetivos empiezan por un verbo en infinitivo, dicen la tarea analítica específica que se va a realizar. Pueden ser varios objetivos específicos, uno por cada tarea analítica que se quiere observar. Ej. Un Clustering, un KNN.

* 1. DISEÑO DE SOLUCIÓN



+ Calidad esperada.

Se especifica con el cliente que es lo que voy a hacer, muy específicamente. Aun no se ha programado ni se ha creado el modelo (evitar los “entre otros”, “etc” no permitir ambigüedades, todo claro y especificado P.ej: cual es la calidad esperada, que métodos quiere).

* 1. RECURSOS PARA CREACIÓN DEL MODELO Y PARA DESPLIEGUE

Que necesito tanto como para la creación del modelo como para el desarrollo de sw del despliegue. Ej: será en la nube, será local, como conversa el modelo con la interfaz gráfica. ¿Será un servicio web en la nube? ¿Será una aplicación? ¿Qué lenguajes de programación? ¿Qué licencias? ¿Qué IDE’s?

1. **ENTENDIMIENTO DE LOS DATOS (Datos específicos del problema)**
   1. CICLO DE LOS DATOS

Se hace en compañía de alguien encargado de gestionar las BD’s o área de TI.

* ¿Cómo se generan/capturan los datos?
* ¿Dónde se almacenan los datos?
* ¿Quién puede modificar los datos?
* ¿Cada cuánto tengo datos nuevos? Periodicidad de los datos. **De allí se puede estimar el cronograma de entrenamiento**. Ej: UPB recibe datos nuevos cada 6 meses, se reentrena cada 3 años.
  1. DICCIONARIO DE DATOS

Tiene que existir (en teoría). El diccionario de datos de la base de datos describe la variable, el tipo de la variable y su descripción.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Tipo** | **Descripción** |
|  |  |  |

* 1. REGLAS DE CALIDAD

No se construyen a partir de los datos.

|  |  |
| --- | --- |
| **Variable** | **Regla de calidad** |
| Edad | 18 - 60 |

Cualquier valor fuera del rango de datos, es un error. Estas reglas las aporta un experto en TI del negocio. Reglas de calidad por cada variable. Servirá en un futuro para identificar atípicos.

**2 primeras etapas son las más complicadas.**

1. **PREPARACIÓN DE DATOS (Estadística)**
   1. INTEGRACIÓN

Se crea la sabana de datos, unir tablas (join).

* 1. SELECCIÓN DE VARIABLES

Borrar variables irrelevantes, las que no sirven para el modelo o que hagan parte de una regulación de privacidad de datos (nombres, direcciones, cédulas, teléfonos).

* 1. DESCRIPCIÓN ESTADÍSTICA

Medidas estadísticas, gráficos, etc.

* 1. LIMPIEZA DE ATÍPICOS

Cualquier valor por fuera del rango expuesto en las reglas de calidad es considerado atípico.

* 1. LIMPIEZA DE NULOS

No siempre se puede hacer imputación. Definir si borro o imputo (inventar datos).

* 1. CREACIÓN DE NUEVAS VARIABLES

Creación de nuevas variables a partir de datos libres.

* 1. ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA

Se usa para eliminar variables que no me sirven porque son redundantes.

* 1. ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA (PREDICCIONES)

Cuando una variable definitivamente no me sirve para nada.

* 1. REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)

Análisis de componentes principales (PCA), se asumen riesgos como perdida de información al reducir variables. Puedo perder el 0.25 de la varianza original de los datos.

* 1. BALANCEO (CLASIFICACIÓN)

Cuando la variable objetivo está muy desbalanceada se puede optar por hacer balanceo sobre muestreo. Muy riesgoso también porque evalúas mediante datos ficticios.

* 1. TRANSFORMACIONES

Tener en cuenta si debo crear dummies o no, es decir, tener en cuenta con que estoy trabajando.

1. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN**
   1. CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING

Se aplican los métodos que se especificaron en el diseño de solución, y se configuran.

* 1. ANALISIS DE MEDIDAS DE CALIDAD

Análisis de las medidas de calidad definidas previamente.

* 1. SELECCIÓN DEL MEJOR MODELO

Además de la medida de calidad se deben tener en cuenta otros aspectos como la complejidad computacional. Se usan métodos como la ANOVA o la prueba de Tuckey. Si no hay diferencia significativa en los modelos se parte de la comparación a partir de la complejidad computacional.

1. **DESPLIEGUE**
   1. PREDICCIÓN DE DATOS FUTUROS

Despliegue por interfaz gráfica, app, nube, web, etc.

* 1. CRONOGRAMA DE MANTENIMIENTO

Haya o no errores en el modelo, se hace un mantenimiento cada cierto periodo de tiempo definidos en las primeras etapas.