Entregable Módulo de Introducción: CRISP-DM

Laura Rodríguez Navas

December 2019

En este entregable se considera la aplicación de cada una de las fases de la metodología CRISP-DM al problema práctico que se nos plantea, que es la extracción y explotación de datos de un sistema de salud.

1 Comprensión del Negocio

1.1 Determinar los Objetivos del Negocio

El objetivo de la minería de datos que se va a aplicar en este problema práctico es el de hacer predicciones lo más fiables a partir de los atributos recogidos por un screening general que se ha realizado durante 20 años a todos los hombres cuando cumplían los 45 años. El objetivo de les predicciones es reducir el sobrediagnóstico (falsos positivos) de cáncer de próstata, manteniendo los falsos negativos en $< 1\%$, usando la información de muchos de estos pacientes acerca de si han sufrido o no cáncer de próstata en los años posteriores al screening.

1.1.1 Contexto

En referencia a la situación de negocio en el sistema de salud al principio de este problema práctico se puede decir que se cuenta con una base de datos de pacientes y existe un estudio en profundidad acerca de si han sufrido o no cáncer de próstata en los años posteriores al screening general del que se pueden sacar conclusiones o patrones para hacer predicciones sobre los futuros pacientes.

1.1.2 Objetivos del negocio

El objetivo del negocio, como ya se ha mencionado, es la reducción de los falsos positivos de pacientes de cáncer de próstata, de tal manera que se puedan hacer unas predicciones fiables partiendo de los datos que ya tenemos de dichos pacientes. Las predicciones pueden ser muy útiles para la detección i reducción de errores no deseados en las exploraciones físicas de esta enfermedad. Todo esto permitirá al sistema de salud disminuir la cantidad de ansiedad y angustia que un falso positivo causa tanto a los médicos como a los pacientes.

1.1.3 Criterios de éxito del negocio

Desde el punto de vista del negocio se establece como criterio de éxito la posibilidad de realizar predicciones a cinco años sobre nuevos pacientes con un elevado porcentaje de fiabilidad, de tal forma que se puedan reducir los futuros falsos positivos, manteniendo los falsos negativos en $< 1\%$.

1.2 Evaluación de la Situación

1.2.1 Inventario de recursos

En cuanto a recursos de software disponemos del programa de minería de datos Oracle BI Data Miner que proporciona herramientas para realizar tareas de minería de datos sobre una base de datos Oracle que es con la que contamos para el almacenamiento de los datos. Los recursos de hardware de los que disponemos son un ordenador portátil con las siguientes características:

- Marca: Apple ©
- Modelo: MacBook Pro del 2013
- Procesador: Intel © Core 7 de cuatro núcleos a 3,2 GHz
- Memoria RAM: 8 GB
- Capacidad de almacenamiento: 256 GB

La fuente de datos es una base de datos Oracle con la información de los pacientes que ha sido almacenada durante 20 años. Esta información incluye el identificador de cada paciente y si han sufrido o no cáncer de próstata a la edad de 45 años.

1.2.2 Requisitos, supuestos y restricciones

Al no poder utilizar los datos personales de los pacientes reales debido a cuestiones legales, se ha tenido que utilizar una base de datos ficticia con datos no reales de pacientes inventados.

1.2.3 Terminología

- Falso positivo: es un error por el cual al realizar una exploración física o una prueba complementaria médica su resultado indica una enfermedad determinada, cuando en realidad no la hay.
- Falso negativo: es un error por el cual al realizar una exploración física o una prueba complementaria médica su resultado es normal o no detecta la alteración, cuando en realidad hay una enfermedad en el paciente.
- Cáncer de próstata: es un tipo de cáncer que se desarrolla en uno de los órganos glandulares del sistema reproductor masculino llamado próstata.

- Screening: es una estrategia utilizada para buscar afecciones o marcadores de riesgo aún no reconocidos en pacientes sin signos o síntomas.

1.2.4 Costes y beneficios

Los datos de este proyecto no suponen ningún coste adicional al sistema de salud ya que estos datos pertenecen al propio sistema de salud.

En cuanto a beneficios, no se puede decir que este proyecto genere algún beneficio económico para el sistema de salud directamente, pero sí que se puede suponer indirectamente ya que el objetivo principal es reducir la cantidad de falsos positivos, y por tanto la satisfacción de los pacientes y los médicos augmenta, y esto se traduce en prestigio para el sistema de salud, lo cual hará que más pacientes consideren hacerse la exploración física y que más médicos trabajen en él.

1.3 Determinar los Objetivos de la Minería de Datos

1.4 Realizar el Plan del Proyecto

1.5 Comprensión de los Datos