Entregable Módulo de Introducción: CRISP-DM

Laura Rodríguez Navas

Enero 2020

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.

Índice

1 Comprensión del Negocio **3**

1.1 Determinar los Objetivos del Negocio . . . 3

1.1.1 Contexto . . . 3

1.1.2 Objetivos del negocio . . . 3

1.1.3 Criterios de éxito del negocio . . . 3

1.2 Evaluación de la Situación . . . 3

1.2.1 Inventario de recursos . . . 3

1.2.2 Requisitos, supuestos y restricciones . . . 4

1.2.3 Terminología . . . 4

1.2.4 Costes y beneficios . . . 4

1.3 Determinar los Objetivos de la Minería de Datos . . . 5

1.3.1 Criterios de éxito de minería de datos . . . 5

1.4 Realizar el Plan del Proyecto . . . 5

1.4.1 Evaluación inicial de herramientas y técnicas . . . 5

2 Comprensión de los Datos **6**

2.1 Recolectar los Datos Iniciales . . . 6

2.2 Descripción de los Datos . . . 6

2.3 Exploración de los Datos 6
2.4 Verificar la Calidad de los Datos 6

3 Preparación de los Datos **7**

3.1 Seleccionar los Datos 7
3.2 Limpiar los Datos 7
3.3 Construir los Datos 7
3.4 Integrar los Datos 7
3.5 Formateo de los Datos 7

4 Modelado **8**

4.1 Escoger la Técnica de Modelado 8
4.2 Generar el Plan de Prueba 8
4.3 Construir el Modelo 8
4.4 Evaluar el Modelo 8

5 Evaluación **9**

5.1 Evaluar los Resultados 9
5.2 Revisar el Proceso 9
5.3 Determinar los Próximos Pasos 9

6 Implantación **9**

6.1 Planear la Implantación 9
6.2 Planear la Monitorización y el Mantenimiento 10
6.3 Producir el Informe Final 10
6.4 Revisar el Proyecto 11

1 Comprensión del Negocio

1.1 Determinar los Objetivos del Negocio

El objetivo de la minería de datos que se aplica en este problema práctico es el de hacer predicciones lo más fiables a partir de los atributos recogidos por un screening general en un sistema de salud que se ha realizado durante 20 años a todos los hombres cuando cumplían los 45 años. El objetivo de les predicciones es reducir el sobre-diagnóstico (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, durante el screening general y en los años posteriores a este.

1.1.1 Contexto

En referencia a la situación de negocio del sistema de salud, inicialmente se puede decir que se cuenta con una base de datos de pacientes reales de entre 45 y 50 años, y existe un estudio acerca de si han sufrido o no cáncer de próstata durante los 20 años de la realización del screening general y en los 5 años posteriores a este, del que se pueden sacar conclusiones o patrones para hacer predicciones sobre futuros pacientes de cáncer de próstata.

1.1.2 Objetivos del negocio

El objetivo del negocio, como ya se ha mencionado anteriormente, es la reducción de los falsos positivos en pacientes de cáncer de próstata, manteniendo los falsos negativos en $< 1\%$, de tal manera que se puedan hacer unas predicciones fiables partiendo de los datos que ya tenemos de dichos pacientes.

En este caso, las predicciones pueden ser muy útiles para la detección i reducción de errores no deseados en las exploraciones físicas de esta enfermedad. Además permitirá al sistema de salud disminuir la cantidad de ansiedad y angustia que provoca un falso positivo, 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 sobre los pacientes diagnosticados con un elevado porcentaje de fiabilidad, de tal forma que se puedan reducir los 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 WEKA que proporciona herramientas para realizar tareas de minería de datos sobre una base de datos SQLite que es con la que he decidido contar para el almacenamiento de los datos.

Además, los recursos de hardware de los que disponemos son un ordenador portátil con las siguientes características:

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

La fuente de datos, como se acaba de comentar, es una base de datos SQLite , creada a partir de un fichero CSV que contiene la información de todos los pacientes diagnosticados. Cada línea del fichero CSV corresponde a un paciente diferente, que incluye su identificador único y si ha sufrido o no cáncer de próstata a la edad de 45 años y en los 5 años posteriores.

1.2.2 Requisitos, supuestos y restricciones

Al poder utilizar los datos personales de pacientes reales y que son propiedad del mismo sistema de salud, no existen requisitos, supuestos y restricciones.

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 como consecuencia aumentará la satisfacción de los pacientes y los médicos, y esto se traduce en mayor prestigio para el sistema de salud.

1.3 Determinar los Objetivos de la Minería de Datos

El objetivo en términos de minería de datos es:

- Predecir falsos positivos para reducir el sobre-diagnóstico de cáncer de próstata de futuros pacientes, manteniendo las predicciones de falsos negativos inferiores al 1%, acorde a la predicción de falsos positivos.

1.3.1 Criterios de éxito de minería de datos

Desde el punto de vista de la minería de datos se establece como criterio de éxito la posibilidad de realizar predicciones sobre los pacientes diagnosticados con un elevado porcentaje de fiabilidad, concretamente se podría definir este porcentaje en un 80%. Aunque el grado de fiabilidad lo determina el algoritmo específico que se emplea a la hora de conseguir el modelo de la minería de datos, por lo que este tema se aborda más adelante en la metodología (evaluación).

1.4 Realizar el Plan del Proyecto

El proyecto se dividirá en las siguientes etapas para facilitar su organización y estimar el tiempo de realización del mismo:

- Etapa 1: Análisis de la estructura de los datos y la información de la base de datos. Tiempo estimado: 2 semanas.
- Etapa 2: Preparación de los datos (selección) para facilitar la minería de datos sobre ellos. Tiempo estimado: 1 semana.
- Etapa 3: Elección de las técnicas de modelado y ejecución de las mismas sobre los datos. Tiempo estimado: 2 semanas.
- Etapa 4: Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa 3. Tiempo estimado: 3 semanas.
- Etapa 5: Producción de informes con los resultados obtenidos en función del objetivo de negocio y los criterios de éxito establecidos. Tiempo estimado: 1 semana.

1.4.1 Evaluación inicial de herramientas y técnicas

La herramienta que se va a utilizar para llevar a cabo este proyecto de minería de datos es WEKA, como ya se comentó en el apartado 1.2.1 .

La elección de WEKA es porqué, aparte de ser un programa de código abierto y compatible con las bases de datos SQLite , ofrece una colección de herramientas de visualización y algoritmos para el análisis de datos y modelado predictivo, considerado inicialmente para llevar a cabo este proyecto de minería de datos.

2 Comprensión de los Datos

2.1 Recolectar los Datos Iniciales

Los datos utilizados en este proyecto son datos referentes a pacientes que incluyen información personal sobre ellos. Estos son propiedad del sistema de salud y por tanto el coste de recolección será inferior a lo normal en estos casos, y como se utilizan datos reales de pacientes existentes en el sistema de salud, las predicciones y estudios en el proyecto serán muy realistas.

Aunque, debido a la gran cantidad de pacientes que son necesarios para poder realizar el proyecto de minería de datos con éxito, la opción de insertar los datos de los pacientes manualmente uno a uno en la base de datos no es viable, por lo que se opta por crear un programa de inserción de datos en el lenguaje de programación Python, cuya salida es una tabla para la base de datos.

2.2 Descripción de los Datos

Los datos se encontrarán almacenados en una tabla SQLite , llamada Pacientes. Cada fila de la tabla corresponde a un paciente, y cada paciente estará identificado por su número de paciente, que es un valor numérico y único. También se indicará con un valor booleano, si los pacientes han sufrido (T) o no (F) cáncer de próstata a la edad de 45 años y en los 5 años posteriores.

Por ejemplo, la tabla Pacientes se estructuraría:

IDPaciente	45	+45
...	...	...
5002	T	T
5003	F	T
5004	F	F
...	...	...

2.3 Exploración de los Datos

Una vez que se han descrito los datos, se procede a explorarlos, y esto implica aplicar pruebas estadísticas básicas que revelarán propiedades de los datos, y crear tablas de frecuencia y gráficos de distribución de los datos. Esto sirve principalmente para determinar la consistencia y completitud de los datos.

Pero en este caso, no he llevado a cabo esta sección ya que no puedo realizar la exploración inicial sin los datos del proyecto.

2.4 Verificar la Calidad de los Datos

Después de hacer la exploración inicial de los datos supondremos que estos son completos. Los datos cubren los casos requeridos para la obtención de los resultados necesarios para poder cumplir los objetivos del proyecto y no contienen errores, ya que son datos generados automáticamente por el script de inserción de datos. Tampoco se encontrarían valores fuera de rango, ya que los datos son controlados desde el mismo script, por lo que no habrá riesgo de ruido en el proceso de la

minería de datos, aunque es un fenómeno muy común en conjuntos de datos muy grandes, como el que se utiliza en este proyecto.

3 Preparación de los Datos

3.1 Seleccionar los Datos

En términos de registros, para el análisis se van a utilizar todos los campos de las columnas IDPaciente, 45 y +45, dentro de la tabla Pacientes que compone la base de datos y que podemos ver en el apartado 2.2 , ya que al ser ésta una base de datos específicamente creada para este proyecto, el número de pacientes que se han insertado ha sido elegido a propósito.

3.2 Limpiar los Datos

La base de datos con la que se cuenta para el proyecto contiene toda la información necesaria para poder cumplir el objetivo de la minería de datos, además estos datos al haber sido introducidos ex profeso para el caso práctico que se presenta, son datos limpios que no contienen valores nulos y por lo tanto no hay necesidad de hacer una limpieza sobre ellos.

3.3 Construir los Datos

No seria aplicable la construcción de datos en este proyecto porqué no es necesario que se realicen operaciones de transformación sobre atributos derivados de los datos. Tampoco será necesario generar nuevos atributos ni nuevos registros sobre la base de datos, ya que esta está completa y ha sido creada específicamente para su uso en este proyecto.

3.4 Integrar los Datos

Como se acaba de comentar en el apartado 3.3 , no seria necesaria la creación de nuevas estructuras (atributos, registros, campos, etc.), ni la fusión entre distintas tablas de la base de datos ya que solo existe una tabla.

3.5 Formateo de los Datos

No será necesario cambiar el orden de ningún campo dentro de los registros, ni tampoco la re-ordenación de los registros dentro de la tabla Pacientes. Tampoco será necesario cambiar el formato de ninguno de los campos que se van a utilizar para la minería de datos ya que el formato actual es admitido por la herramienta WEKA .

4 Modelado

4.1 Escoger la Técnica de Modelado

Debido a que se va a utilizar el programa WEKA para realizar la minería de datos, se utilizará alguna de las técnicas de modelado que nos ofrece esta herramienta de acuerdo con el objetivo del proyecto que está reflejado en el apartado 1.3 .

Concretamente, de los modelos que ofrece el programa WEKA , el que mejor se adapta al objetivo del proyecto es un modelo lineal de regresión, puesto que el problema que se quiere resolver es un problema de predicción, los campos que se quieren predecir contienen valores continuos y la variable de salida precedida es numérica: el % de falsos positivos. Así, se empleará el algoritmo GLM .

El modelo y el algoritmo de regresión lineal se utilizará para estudiar la relación entre los pacientes diagnosticados. Seria de interés conocer el efecto que uno o varios pacientes pueden causar sobre otro, e incluso predecir en mayor o menor grado relaciones de un paciente a partir de otro.

4.2 Generar el Plan de Prueba

El procedimiento que se emplea para probar la calidad y validez del modelo es el de utilizar las medidas del error cuadrático medio (root mean squared error) y el error absoluto medio (mean absolute error). Estas medidas de error las calcula automáticamente el programa WEKA al ejecutar los modelos de regresión.

WEKA también ofrece la opción de dividir los datos en dos grupos automáticamente antes de generar el modelo: por un lado tendríamos el conjunto de datos que se van a utilizar para generar el modelo, basados en el screening general, y un segundo conjunto de datos basados en los diagnósticos de los 5 años posteriores al screening general, que se empleará para realizar las pruebas y medir la calidad del modelo.

4.3 Construir el Modelo

Los campos sobre los cuales se va hacer la predicción son "cancer_con_45_años", "cancer_con_+45_años" y el identificador del paciente. En cuanto a los parámetros empleados para el algoritmo de GLM, se utilizan los parámetros que vienen por defecto en WEKA .

Y para construir el modelo de regresión, el primer paso seria observar si puede existir o no dependencia/relación entre los pacientes diagnosticados, por ejemplo, representando gráficamente los datos observados mediante una nube de puntos, lo que se conoce como diagrama de dispersión .

Una vez representados los datos y tras detectar que entre dos o más pacientes existe una relación, el siguiente paso sería intentar modelar dicha relación usando una expresión matemática que permita predecir, de forma aproximada.

4.4 Evaluar el Modelo

Una buena manera de evaluar la efectividad del modelo es utilizando los dos indicadores que se establecieron en el plan de pruebas de este proyecto, en el apartado 4.2 . Además de estos dos

indicadores, la herramienta WEKA nos da más información acerca de los modelos que es muy útil a la hora de evaluarlos, como son el indicador de confianza predictiva (predictive confidence) y el valor predicho medio junto al valor medio real.

5 Evaluación

5.1 Evaluar los Resultados

Desde el punto de vista del negocio, se ha establecido como criterio de éxito principal, el poder realizar predicciones con un porcentaje de fiabilidad "elevada" (ver apartado 1.3.1).

Para la evaluación de las predicciones se observaran los indicadores estadísticos que se obtienen al ejecutar el modelo descrito en el apartado 4.2 . En cualquier caso, basándome en el conjunto de datos que disponemos, que esta formado por datos reales de pacientes, podría afirmar que el modelo será factible. Ya que se podrán hacer predicciones muy fiables acerca de que porcentaje de falsos positivos i falsos negativos existirán con los pacientes, lo cual se considera de fiabilidad elevada desde el punto de vista del objetivo de negocio.

5.2 Revisar el Proceso

En un escenario real, es decir, utilizando la base de datos real de la que dispone el sistema de salud, posiblemente no habrá complicaciones a la hora de realizar el modelo y se ejecutará como estaba previsto. Como resultado del proceso se obtendrán predicciones muy fiables como se comenta en el apartado 5.1 .

5.3 Determinar los Próximos Pasos

El siguiente paso a realizar es el de ejecutar la etapa de implantación para el objetivo del proyecto.

6 Implantación

6.1 Planear la Implantación

Para poder implantar este proyecto sería necesario en primer lugar tener acceso a la base de datos, es decir la base de datos que contiene toda la información relativa a los pacientes del sistema de salud. A partir de ahí, los pasos a seguir serían los mismos que se han seguido en este documento desde la comprensión del negocio hasta la implantación. Si bien, cabe decir que habrá algunas fases, como la de comprensión y preparación de los datos, que probablemente sea más compleja y llevará más tiempo ya que la base de datos real contiene muchos registros.

En segundo lugar sería necesario que en el sistema de salud se use una base de datos SQLite , de no ser así se tendrían dos opciones, la primera sería exportar la base de datos actual a otro tipo de base de datos, compatible con el programa de de minería de datos utilizado en este proyecto (WEKA). Si no fuera posible se buscaría otro software para este propósito.

6.2 Planear la Monitorización y el Mantenimiento

Como he comentado anteriormente, el volumen de los datos es muy grande por el cual la extracción de los datos debe ser realizada cuidadosamente y realizando siempre backups de los datos. La minería de datos debería ser realizada en periodos anuales ya que esta es la medida de tiempo utilizada en el sistema de salud para realizar las exploraciones físicas a los pacientes.

Como plan de supervisión y mantenimiento se podría establecer los siguientes procesos:

- Extracción y almacenamiento anual de los datos guardando la información obtenida en formato CSV.
- Distribución de los datos del modelo de regresión a trabajar en función de la herramienta WEKA .
- Los archivos de la explotación de datos deberán ser guardados usando el servicio de almacenamiento en la nube del propio sistema de salud, almacenándolos por ejemplo en carpetas ordenadas por diagnósticos anuales.
- Los resultados obtenidos en cada explotación de datos deberán ser transformados a formato CSV y generar gráficas de distintos tipos para una mejor visualización e interpretación de los resultados obtenidos en cada año.

6.3 Producir el Informe Final

El público al que va dirigido este informe es a los médicos del sistema de salud de tal manera que se pueda estudiar la situación actual y tomar medidas correctivas para la mejora del servicio de atención médica.

El uso de la metodología CRISP-DM en este proyecto permite encontrar un comportamiento predictivo a la hora de reducir los falsos positivos manteniendo los falsos negativos por debajo del 1%, un plan de extracción, normalización, y codificación de datos para la realización de procesos de minería de datos anualmente, para alcanzar el objetivo de minería de datos que se ha fijado inicialmente.

Para poder hacer una simulación lo más real posible, se ha tenido que desarrollar un pequeño programa en Python que crea la base de datos de manera automática, debido a la gran cantidad de datos que necesitábamos manejar para hacer una estimación lo más precisa posible.

Cuando ya disponíamos de la base de datos sobre la que ejercer la minería de datos, se hizo un análisis de la estructura de los datos y la información contenida.

El lado positivo de haber creado la base de datos con datos reales es que se ha podido acceder a los datos des de un primer momento sin la necesidad de espera o otros problemas legales que pudiera impedir la adquisición de estos. Esto redujo significativamente la duración estimada de la etapa 2 definida en el apartado 1.4 (Realización del Plan de Proyecto).

A continuación se realizó la elección de la técnica de modelado y la ejecución de dicha técnica sobre los datos empleando la herramienta escogida para ello, WEKA .

Por último, una vez obtenido el modelo, se analizó para determinar la adecuación o no del mismo. En este caso se determinó que el modelo podría ser válido para nuestro objetivo y no se descartó por ser muy fiable.

Y realizados todos estos pasos se presentan los resultados alcanzados al público que es el objetivo de este apartado.

6.4 Revisar el Proyecto

El mayor error que se ha podido cometer a lo largo de este proyecto es el de no realizar una limpieza, conversiones o formateo de los datos que se disponen. Ya que al utilizar una base de datos tan grande es normal que aparezcan problemas durante el plan de proceso del proyecto definido en el apartado 1.4 .