# Aprendizaje automático en problemas reales

(y cómo resolver el dilema via-gra/Joiner)


## Los tres pilares del data mining

**Técnicas**

IA, estadística, reconocimiento de patrones, visualización…

Métodos *off-the-shelf* son a menudo suficientes.

**Datos**

Volumen elevado, diferentes fuentes, diferentes formatos, incompletos, erróneos: no se guardan para hacer data mining.

**Modelado**

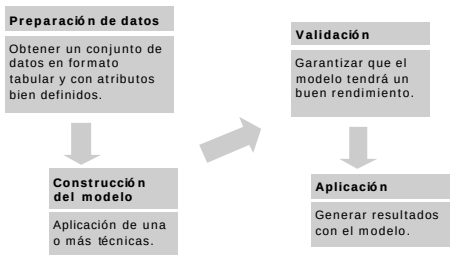
Comprender los datos.

Tener en cuenta conocimiento de negocio.

Saber ajustar los parámetros de las técnicas utilizadas.

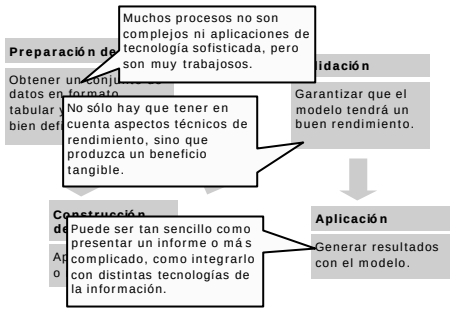
2

## Data mining como proceso




3

## Data mining como proceso




4

## Lo primero es identificar el problema

Existe una tendencia a tratar el análisis de datos como algo exclusivamente técnico. Sin embargo, es necesario hablar con la gente que conoce el negocio para conseguir lo siguiente:

- Seleccionar un problema bien definido y con el que se puedan obtener beneficios aplicando data mining.
- Especificar claramente la solución requerida.
- Definir cómo se va a utilizar la solución.

5

## El modelo perfecto

Vicepresidente: Quiero un modelo que identifique los compradores potenciales para una campaña.

Data miner: Ok.

(un mes y 50000 € después...)

Data miner: Este modelo permite contactar sólo un 20% de clientes y captar alrededor del 70% de compradores potenciales. Supone un gran ahorro.

Vicepresidente: No. Yo quiero **todos** los compradores potenciales.

Data miner: No existe el modelo perfecto. La única manera de lograrlo es contactar con todos.

6

## Comencemos... ¿pan comido?


7

## A ver, esos datos...

Lo que necesito...

| Id_cliente | Compra Vino | Compra Agua mineral | Compra Garbanzos | Compra Gel baño | Gasto total |
|---|---|---|---|---|---|
| 173423 | NO | SI | NO | SI | 7,05 |
| 186632 | SI | NO | SI | NO | 5,75 |

Lo que tengo...

| Id_cliente | Fecha | Producto | Cantidad | Precio |
|---|---|---|---|---|
| 173423 | 31/3/04 | 763 | 2 | 1,6 |
| 173423 | 31/3/04 | 87 | 1 | 0,60 |
| 186632 | 31/3/04 | 135 | 1 | 0,75 |
| 186632 | | | | |
| 173423 | | | | |
| 173423 | | | | |

| Id_Producto | Nombre | Familia |
|---|---|---|
| 35 | Vino | 8 |
| 87 | Agua mineral | 9 |
| 135 | | 23 |

| Id_cliente | E.Civil | Edad | CP |
|---|---|---|---|
| 173423 | Casado | 36 | 08025 |
| 173424 | Casado | 39 | 08032 |
| 173425 | Soltero | 23 | 08029 |
| 173426 | Soltero | 26 | 08005 |

| Id_Familia | Nombre |
|---|---|
| 8 | Bebidas alc. |
| 9 | Bebidas no alc. |
| 23 | Legumbres |

8

## Si comes vainilla no conduzcas

Queja al servicio de atención al cliente de una persona que acaba de adquirir un coche:

*"Cuando salgo a comprar helado y paro el coche en la puerta de la tienda, le cuesta volver a arrancar si compro helado de vainilla. Si compro otro sabor arranca a la primera".*

9

## GIGO: Garbage In, Garbage Out

El resultado de aplicar data mining en un problema determinado depende mucho de la calidad de los datos.

El data mining únicamente explicita conocimiento que está implícito en los datos. Si éstos no contienen información suficiente no obtendremos nada útil.

Es muy importante comprender la estructura de los datos, su significado y seleccionarlos y prepararlos de forma adecuada.

10

## Una lista de tareas de preparación

- Agrupar
- Pivotar
- Combinar
- Errores
- Datos redundantes
- Sinónimos con la columna a predecir
- Consistencia
- Valores únicos para cada instancia
- Columnas con un solo valor
- Columnas con casi un solo valor
- Valores extremos
- Valores nulos
- Columnas derivadas
- Extracción de nuevas columnas

11

## Agrupación

| Id_cliente | Fecha | Producto | Cantidad | Precio |
|---|---|---|---|---|
| 173423 | 31/3/04 | 763 | 2 | 1,6 |
| 173423 | 31/3/04 | 87 | 1 | 0,60 |
| 186632 | 31/3/04 | 135 | 1 | 0,75 |
| 186632 | 31/3/04 | 35 | 2 | 2,50 |
| 173423 | 1/4/04 | 87 | 1 | 0,60 |
| 173423 | 1/4/04 | 223 | 1 | 3,65 |

La unidad es la transacción.

Calcular el gasto a partir de precio y cantidad.

Agrupar por Id_cliente

La unidad de análisis es el cliente.

| Id_cliente | Compras | Gasto |
|---|---|---|
| 173423 | 4 | 7,05 |
| 186632 | 2 | 5,75 |

Es importante obtener el nivel de detalle necesario para el análisis (de transacción a cliente).

12

## Pivotaje / Combinación

Convertimos cada valor de la columna producto en una columna nueva.

| Id_Producto | Nombre | Familia |
|---|---|---|
| 35 | Vino | 8 |
| 87 | Agua mineral | 9 |
| 135 | Garbanzos | 23 |

| Id_cliente | Compra Vino | Compra Agua mineral | Compra Garbanzos | Compra Gel baño | Gasto total |
|---|---|---|---|---|---|
| 173423 | NO | SI | NO | SI | |
| 186632 | SI | NO | SI | NO | |

Combinamos con información de una tabla de productos.

Calcular el gasto a partir de precio y cantidad.

Agrupar por Id_cliente.

Pivotar por Producto.

*Join* con tabla Productos.

En lugar de guardar totales, simplemente especifica si se ha comprado.

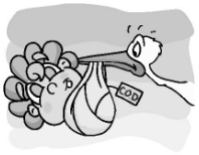
13

## Un 11 de noviembre productivo

Una compañía descubre que casi todos sus clientes han nacido el 11 de noviembre.

Y más… Casi el 5% nacieron en 1911, 11/11/11

¿Una coincidencia?



14

## Errores / integridad

Los datos críticos de negocio normalmente son correctos (p.e. facturas).

Existen muchos otros datos que se recogen sin prestarles demasiada atención.

Hay muchas fuentes de error:

- Resistencia de las personas a proporcionar o introducir datos.
- Formatos incorrectos (p.e. las comas en la columna *dirección* con importación de columnas separadas por comas).
- …

15

## Datos redundantes

Pueden provenir de la combinación de datos de fuentes distintas o de derivar nuevos columnas a partir de las existentes.

Ej: Fecha de nacimiento y edad.

Algunos métodos pueden tener problemas y siempre añaden complejidad inútil.

16

## Sinónimos con la columna a predecir

Si una columna está muy correlacionada con la clase puede indicar que es solamente un sinónimo.

Ejemplos:

- Número de cliente no nulo puede ser un sinónimo de cliente que ha respondido a una campaña.
- Prácticas de negocio: todos los clientes que contratan desvío de llamadas tienen llamada en espera.

17

## Consistencia

Si la entrada de datos es manual o provienen de fuentes distintas, puede haber diferencias de codificación.

Nombres distintos para lo mismo (valores inconsistentes):

Ej: Barcelona, BCN, Barc.

Mismos nombres para cosas distintas (nombres de columna inconsistentes):

Ej: qué constituye un cliente para distintos departamentos.

18

## Valores únicos para cada instancia

Hay columnas que tienen un valor distinto para cada instancia:

- DNI
- Identificador de cliente
- Dirección

Estas columnas no son útiles para construir el modelo.

(pero a veces podemos extraer información útil de ellas).

19

## Columnas con un solo valor

Pueden provenir de una selección previa de los datos (p.e. focalizar una campaña en clientes de una zona).

También pueden encontrarse cuando existen valores por defecto en formularios de entrada de datos.

Otra fuente son columnas en desuso.

No aportan ninguna información al análisis.

20

## Columnas con casi un solo valor

Hay que intentar entender por qué se dan los valores poco frecuentes en términos del negocio y de obtención de los datos.

Ej: Si los datos reflejan transacciones por producto, habrá muchos que no se compran a menudo.

Pueden provenir de una discretización previa de los datos.

21

## No tengo vacaciones pero soy longevo

Un portal en internet registra las fechas de inicio y final de vacaciones y la de nacimiento de los usuarios.

Un análisis detallado de los datos revela:

- Algunos usuarios acaban sus vacaciones antes de 1900.
- Alrededor de 1000 tienen más de 100 años.

22

## Valores extremos (outliers)

Son valores alejados de la tendencia de los demás. Pueden ser errores o reflejar casos reales.

Ej: cobro de seguros de cantidades elevadas.

Ciertos algoritmos son muy sensibles a estos valores.

Podemos eliminarlos, aunque también es posible sustituirlos o discretizar los datos.

En algunos problemas el objetivo es precisamente detectarlos. P.e. transacciones fraudulentas con tarjetas de crédito.

23

## Valores nulos

**Valores perdidos**: existen pero no se han introducido. P.e. Edad.

**Valores inexistentes**, p.e., valores históricos de clientes recientes.

Acciones:

- Dejarlos como están.
- Filtrar las instancias que los contienen.
- Ignorar la columna.
- Inferirlos.
- Construir varios modelos.

24

## Columnas derivadas

Combinar variables añade información que puede ayudar a encontrar mejores modelos. También pueden ayudar a la interpretación posterior.

Existen muchas combinaciones posibles y es conveniente intentar reflejar conocimiento y/o intuiciones. Ej: "Creo que puede haber una relación entre la edad del agente de seguros y la del cliente".

Hay que tener cuidado de no derivar muchas variables con la misma información. Ej: ratios y diferencias.

25

## Extracción de nuevas columnas

Algunas columnas contienen más de un tipo de información que puede extraerse como nuevas columnas.

- Los teléfonos contienen código de país y de área, si es móvil o fijo.
- Las direcciones contiene códigos postales.
- Las direcciones de internet contienen el tipo de dominio.
- Las fechas contienen día, mes, año, trimestre, festivo/laborable...
- ...

26

## Construcción del modelo

A priori, podemos emplear cualquiera de los algoritmos que hemos visto durante el curso (no vamos a repetirlos).

El criterio de elección depende de los factores que rodean al problema: requerimientos, rendimiento, coste computacional, coste económico...

Es un proceso de refinamiento continuo: si el modelo no acaba de funcionar como esperamos es posible que haya que volver a la preparación de datos y probar de nuevo.

27

## Evaluación de los métodos

Para cada método empleamos la metodología que hemos visto (hold-out, cross validation).

- Interna: ajustamos los parámetros del modelo para obtener el mejor resultado posible.
- Externa: comprobamos que el modelo realiza la tarea correctamente. Puede incluir seleccionar modelos de distintos métodos.

Es importante seleccionar criterios relacionados con los resultados que se esperan. Maximizar la precisión no tiene porqué ser el único objetivo.

28

## Matriz de confusión (del modelo)

A veces es necesario desglosar los resultados de la predicción para cada valor de la clase.

| | Valor real | |
|---|---|---|
| **Predicción** | **NO** | **SI** |
| **NO** | 2320 | 290 |
| **SI** | 260 | 170 |
| Precisión | 0.90 | 0.37 |

La precisión total es de 0.82 pero sólo acertamos un 37% de las respuestas positivas.

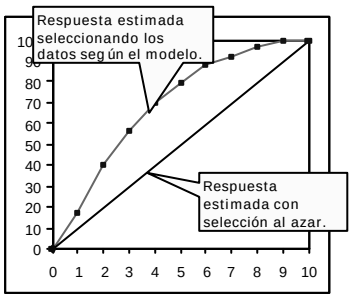
En ciertas aplicaciones es interesante distinguir entre falsos positivos y negativos.

Es muy importante analizarla si los valores de las clases tienen distribuciones muy distintas.

29

## Gráfico de ganancia

Mide en qué proporción una ordenación de los datos según el modelo acierta uno de los valores de la clase.




30

## Campaña de marketing de ACME

La compañía ACME especializada en equipamiento para capturar animales ficticios tiene un nuevo producto "el cazador de correcaminos". Está dirigido a su fiel audiencia de coyotes y quiere realizar una campaña por correo. Dispone de 60.000€.

Coste de un envío: 2€
Puede enviar 30.000 ofertas. Pero en su base de clientes hay 100.000 coyotes.

31

## Soluciones para ACME

- Seleccionar 30.000 clientes al azar.
- Realizar un análisis RFM y seleccionar los clientes que han realizado compras recientemente por una cantidad importante.
- Usar métodos no supervisados para encontrar grupos que reflejen perfiles interesantes para abordar con la campaña.
- Construir un modelo predictivo para determinar los clientes con más posibilidad de responder a la campaña.

32

## Selección del método

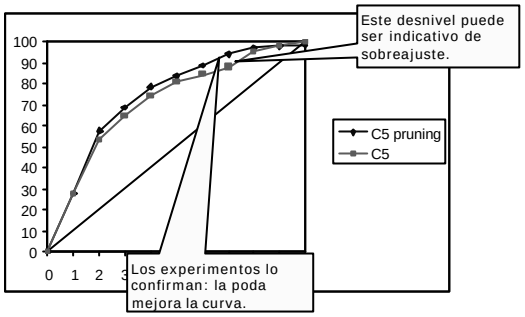
Hay que elegir un método que proporcione una probabilidad asociada a las predicciones (algunos que no lo hacen originalmente pueden adaptarse).

Seleccionamos dos candidatos que representan aproximaciones distintas y que se encuentran en productos comerciales:

- Árboles de decisión
- Redes neuronales

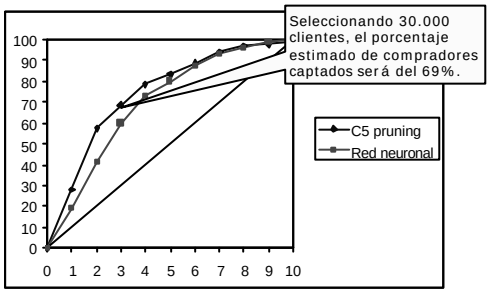
33

## Evaluación de parámetros




34

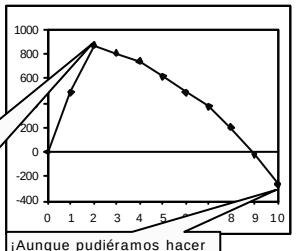
## Selección de modelo




35

## Hablar el lenguaje de los usuarios (I)

Utilizando el coste de envío y estimando un beneficio medio por respuesta, podemos cambiar el gráfico de ganancia por uno de **beneficio**.




36

## Aplicación

Formas de poner en funcionamiento el modelo:

- Informes: lista de clientes, gráficos de beneficio.
- Integración en las bases de datos ya existentes (*scoring*).
- Implementación de una aplicación vertical.

Independiente de lo anterior podemos obtener:

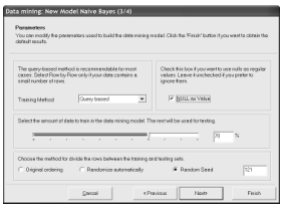
- Información sobre los perfiles de cliente.
- Posibles sugerencias para mejorar la recogida de datos.

37

## Hablar el lenguaje de los usuarios (II)

En aplicaciones verticales, interfaces en forma de asistente pueden simplificar el proceso de data mining y "educar" al usuario.

Incluso mejor si se adaptan los términos técnicos al lenguaje del usuario.



38

## El objetivo es básico en el proceso

ACME ha expandido el negocio y creado divisiones diferentes para correcaminos, conejos y patos. Cada división lleva a cabo sus propias campañas.

El método anterior ya no es adecuado porque puede ocurrir que un mismo cliente reciba varias ofertas y se canse.

El objetivo ha cambiado de maximizar el provecho de una campaña a encontrar la mejor campaña para cada cliente.

Esto no tiene nada que ver con la tecnología, es conocimiento de negocio.

39

## Aplicaciones no de negocio

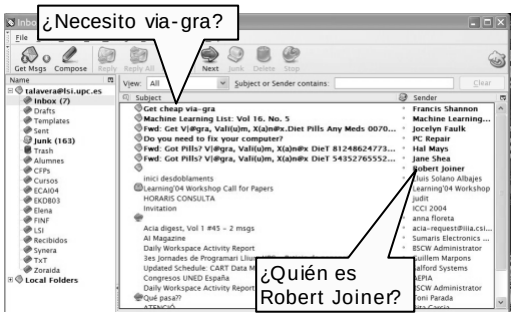
Aunque el proceso de data mining tiene relevancia en aplicaciones de negocio, puede generalizarse a cualquier problema de análisis complejo.

La preparación de datos puede ser igual de costosa. Las valoraciones económicas y el conocimiento de negocio se sustituyen por otros factores análogos.

Ejemplo: análisis de la interacción de los alumnos de un curso a través de internet para crear modelos de usuario.

40

## El dilema via-gra/Joiner




41

## Detección de spam

Un usuario de correo electrónico que gasta su tiempo dando clases en una universidad recibe enormes cantidades de correo no deseado (spam, junk) por haber hecho pública en internet su dirección a merced de robots sin piedad.

El usuario desearía poder identificar rápida y automáticamente los mensajes con más probabilidad de ser spam sin tener que definir reglas de filtrado.

42

## A ver, esos datos...

Lo que necesito...

| Id_mensaje | sex | Viagra | laboratori | UNED | Spam |
|---|---|---|---|---|---|
| 173423 | SI | SI | NO | NO | SI |
| 183555 | NO | NO | NO | SI | NO |
| 186632 | NO | NO | SI | NO | NO |

Lo que tengo...

Gone forever are the headaches, hassles and high costs of obtaining the pharmacy products you want and need. When you need them fast: ? V|@Gra ? XAN@x ' S.o.ma < Pnter/m/in > Vali.u.m $

Anexamos archivos con la información del I Congreso Internacional de Estilos de Aprendizaje y del IX Congreso Internacional de Informática > Educativa que realizaremos en julio de 2004 en la UNED Madrid España

Ad|pe:x, I0na:m|n, M3rid.ia, X'3nica|, Am:bi3n, S0naT.a

43

## Escriben raro ¿no?

Los nombres se escriben incluyendo caracteres aleatorios o especiales para confundir a los filtros: via-gra, V|@gra.

Gone forever are the headaches, hassles and high costs of obtaining the pharmacy products you want and need. When you need them fast: ? V|@Gra ? XAN@x ' S.o.ma < Pnter/m/in > Vali.u.m $ A.t!v@n

Necesito un método más flexible para detectar el spam.

44

## Elegir una representación

Hay que convertir la información textual en una forma adecuada para el análisis.

Lo más habitual es usar cada palabra como un atributo distinto (*bag of words*).

Se pueden representar de varias maneras: binaria (aparece o no), número de apariciones,

Hay alternativas menos triviales: n-grams (grupos de palabras), información lingüística, ontologías.

Se genera un número muy elevado de columnas.

45

## Seleccionar / ponderar atributos

Una práctica habitual es ponderar la frecuencia de aparición de cada palabra con su importancia.

Una forma muy popular es el método TFIDF (Term Frequency, Inverse Document Frequency).

Debido a la gran cantidad de atributos que se generan (miles) puede resultar conveniente realizar una selección.

Es normal aplicar métodos muy simples: palabras que se dan con poca frecuencia o alguna medida de correlación con la clase.

46

## Preparación de datos específica

En problemas de text mining, existen métodos específicos de preparación de datos.

Dos prácticas comunes son

**Stemming**: las palabras con la misma raíz se consideran el mismo atributo.

**Stop lists**: listas de palabras que no se incluyen en la representación por ser muy frecuentes y no proporconar información. Ej: artículos, preposiciones,...

47

## Selección del método

Los datos no son estáticos, sino que van llegando nuevos mensajes continuamente y las palabras también van cambiando.

El concepto de spam/no spam puede ir variando con el tiempo por lo que necesitamos un método que sea capaz de aprender de forma incremental.

En este caso elegimos el algoritmo Naive Bayes por ser muy eficiente y adaptarse de forma natural al aprendizaje incremental.

Para cada nuevo mensaje, calcularemos la probabilidad de que sea spam y lo marcaremos si supera un umbral.

48

## Ajustes específicos para el problema

Cada vez que llega un nuevo mensaje, se obtiene una representación y se seleccionan las 15 palabras más relevantes.

La relevancia se calcula midiendo la desviación respecto al valor 0.5 de la probabilidad de la palabra entre el spam.

Con estas 15 palabras se aplica Naive Bayes para obtener la probabilidad de que sea spam.

Las palabras que se encuentran en los mensajes normales se cuentan dos veces (ver evaluación).

49

## Evaluación

En este problema lo importante no es únicamente la precisión total, sino la distinción entre falsos positivos y negativos.

Un falso positivo puede ser mucho más crítico que un falso negativo porque podemos perder un mensaje importante.

Doblando las frecuencias de las palabras que aparecen en el no spam se sesgan las probabilidades para tender a no detectar falsos positivos.

50

## Otras cuestiones

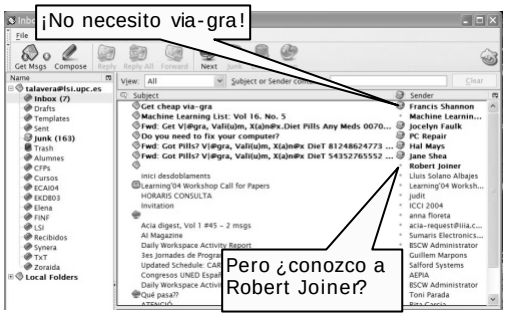
Adicionalmente, puede ser necesario resolver otras cuestiones específicas para el método concreto que usamos aplicado a este problema específico.

A menudo estas cuestiones deben resolverse encontrando la respuesta de forma experimental.

Ej: ¿Qué probabilidad de spam asignamos a palabras que no hemos visto para calcular su relevancia?

51

## ¿Dilema solucionado?




52

## El mensaje de Robert Joiner

Why not buy V-I-A-G-R-A - No Prescription Needed !!

Costs over 50% less than Viagra®

http://www.8Eq.34edmnr5.com/qp/default.asp?ID=bw

We also have these medications in highly discounted generic form:

Ambien, Xanax, Phentermine, Lipitor, Nexium, Paxil, and Vioxx.

Physician Consultation: FREE! Fast, FREE delivery

EZ online form

También ofrece viagra (bueno V-I-A-G-R-A).

¿Por qué le cuesta de detectar?

Algo impide que mire los términos adecuados.

53

## El truco de Robert Joiner

rhetorician dielectric terpsichore insight meridian chokeberry borealis whatever tx constantinople brutal kink harmon banjo scotsman substitute vat yang sound voluntary decomposition chalmers honoree amethystine bellini you've buff monotreme avoidance chugging debussy dragonfly moduli waxwork chimique draftsperson unanimity diamond mckesson corrosion annual alden augmentation timothy polytope headache boo conferred coupon bezel borden contemplate moreland

compensatory oberlin scarf infrequent liquidus lobule coriolanus newsstand farmhouse r cheater mathematician delicate ballroom celesta bergland pedestal aesthetic uplift ego coverall transposition chattanooga dynamic delhi rood genre

Al final del mensaje incluye una lista aleatorias de palabras que no se encuentran con frecuencia en mensajes spam.

54

## Lo siento Robert...

Es posible indicarle al programa de correo que un mensaje es o no spam para que Naive Bayes modifique las probabilidades.

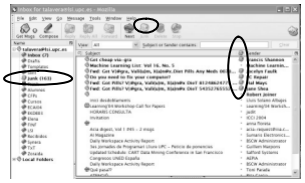
Dependiendo del grado de entrenamiento del sistema, puedo tener que indicárselo varias veces. Con el tiempo, el rendimiento mejora.

El sistema satisface las necesidades del usuario, se adapta a nuevas tácticas de los spammers y evita tener que hacer largas listas de filtros.

55

## Aplicación

No parece muy complejo de implementar.



La herramienta de correo electrónico de Mozilla ya lo incorpora, creo que me lo voy a ahorrar ☺

(no es publicidad, es gratis)

56

## Gracias

...por escuchar hasta el final.

(especialmente a los que estáis despiertos)

57