9. Conclusiones

Se ha partido de una base de datos no estructurada de información relacionada sobre diversos tipos de medicamentos y sus efectos según los pacientes que tomaron dicha medicina. Viendo la información en general, hemos observado que era imprescindible realizar un preprocesamiento de los datos, ya que había demasiados comentarios textuales y era necesario reducir la información prescindible para poder aplicar posteriormente una serie de técnicas.

Tras realizar dicho preprocesamiento, se ha realizado un **análisis exploratorio de la información**, de donde se ha concluido que en general los usuarios tienen buenas opiniones sobre los medicamentos, ya que en general son bastantes efectivos y no tienen efectos secundarios severos.

A continuación, se realizó un **análisis de sentimientos** teniendo en cuenta las opiniones escritas en los comentarios de dichos usuarios, y hemos observado que, aunque puntúen con una buena valoración al medicamento, los comentarios no son del todo positivos, ya que los usuarios suelen relatar más los posibles efectos negativos que los positivos, aunque los positivos estén en una mayor proporción. De aquí podemos decir que *"Tendemos a quejarnos más sobre los efectos negativos, que a mencionar los positivos"*, y esta sentencia se puede aplicar a todos los ámbitos de nuestra vida (sí, los seres humanos somos así por naturaleza).

Por otra parte, respecto a las **técnicas de tipo descriptivo** que hemos utilizado, debemos mencionar que en general, consideramos que funcionan bien a pesar de la dificultad adicional que traía nuestro conjunto de datos de forma implícita. En el caso de *clustering*, recordemos que, muchos de los datos se solapaban, dando lugar a resultados un poco confusos, pero, a medida que íbamos aplicando más versiones/extensiones de dicha técnica, comenzamos a entender más cosas acerca de nuestros datos y sus dependencias, y al final fuimos capaces, tanto de darle sentido a algunos de los agrupamientos, como de obtener buenos resultados con el coeficiente de silueta (como podemos ver en la tabla 1), además de ser capaces de obtener datos no solapados. El aplicar distintas técnicas de agrupamiento nos ha hecho ver, que en la práctica, aquellas técnicas que se esperan que mejoren las clásicas no tienen por qué dar buenos resultados; al final, **el papel que juega la naturaleza de los datos con los que trabajamos, es crucial en este tipo de problemas**.

Técnicas	**Coeficiente de Silueta**
K-medias	0.43
K-medioides	0.43
Clustering Difuso	0.4
Clustering Jerárquico	0.51

Table 1: Coefciente de Silueta para Clustering

Por otra parte, respecto a las **reglas de asociación**, pensamos que también hemos sido capaces de obtener buenos resultados y coherentes a partir de esta técnica, y además, cabe decir, que es una de las que más información adicional nos ha aportado, y que el hecho de realizarla previa a las técnicas predictivas, nos ha ayudado a entender el fallo o acierto de estas. Además, esta técnica nos ha parecido especialmente interesante a la hora de aplicarla en textos, ya que consideramos que resume muy bien la información textual, y que es importante a la hora de filtrar información y quedarnos con aquella que más nos interesa.

Con respecto a los **modelos predictivos**, debemos destacar la componente subjetiva que tienen las reglas de asociación a la hora de su elección. Se observó, cómo los consecuentes más frecuentes fueron **effect** y **side**, lo que es lógico ya que estamos midiendo los efectos que tienen los medicamentos.

En referencia a las técnicas de clasificación, se ha observado como los arboles decisión han generado unas aceptables predicciones, y posteriormente se han mejorado utilizando la técnica llamada *randomForest*. Aun así, y con todas las mejoras que se llevan a cabo esta técnica, **pensamos que las técnicas de clasificación basadas en árboles no resultan adecuadas en nuestro problema**, debido a dos cuestiones. La primera de ellas, se basa en el hecho de que tenemos una cantidad muy amplia de datos, y nuestra información se sesga hacia unas etiquetas concretas. Al utilizar todo el conjunto de train, no somos capaces de obtener un árbol con cinco nodos hoja (uno por etiqueta). Por otra parte, el que todas las palabras utilizadas en el

conjunto de test tengan que formar parte del conjunto con el que entrenamos (aunque al final no se haga uso de ellas), hace que la técnica sea muy limitada y de difícil generalización. Por tanto, no vemos adecuado, en nuestro caso, este tipo de técnica, porque ya no es sólo las palabras que no existan, sino las faltas ortográficas y topográficas que puedan tener los usuarios al redactar (y que serían tenidas en cuenta como palabras diferentes en algunos casos).

Por otro lado, SVM no funciona bien con texto a no ser que tengas las mismas palabras tanto en el test como en el train, cosa que no ocurre en nuestro caso. Esto hace que SVM no tenga todo lo necesario para generar una buena función, ya que nos impide predecir para futuros casos (puesto que tendrían una alta probabilidad de contener palabras no contempladas en el entrenamiento). Cosa que sí ocurre, por ejemplo, en este tutorial (https://www.svm-tutorial.com/2014/11/svm-classify-text-r/).

Por otro lado, se aplicó el **modelo de regresión** de regresión sobre las variables numéricas. Dicho método obtuvo unas predicciones de etiquetas muy buenas, obteniendo con la regresión logística multivariable el mínimo error fuera de la muestra, ya que dicho método permite estimar de manera muy aceptable la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa.

Técnicas	**Error Test (%)**	**Error Train (%)**
Naive Bayes (atributos textuales)	19.2331	22.4371
SVM	-	-
Random Forest (atributos numéricos)	35.14377	37.85
Random Forest (atributos textuales)	52.3	52.40
Regresión Lineal Simple (ratingLabel $\sim$ sideEffectsInverse)	12.1405	12.350
Regresión Lineal Simple (ratingLabel $\sim$ effectivenessNumber)	12.14058	10.1593
Regresión Lineal Múltiple (atributos numéricos)	9.265176	8.366534
Regresión Logística Simple (ratingLabel $\sim$ effectivenessNumber)	12.14058	10.15936
Regresión Logística Simple (ratingLabel $\sim$ sideEffectsInverse)	12.14058	12.3506
Regresión Logística Multivariable (atributos numéricos)	9.265176	8.366534
Ridge Regression (atributos numéricos)	10.46724	-
Regresión Polinomial (atributos numéricos)	8.945687	7.768924

Table 2: Errores Test y Train para técnicas de Clasificación y Regresión

Como se aprecia en la tabla, los errores obtenidos fuera del conjunto de la muestra no son muy altos, sin embargo, si que se aprecia como los errores del test obtenidos para atributos numéricos, tienen un error mucho menor que los obtenidos para atributos textuales. Esto es así, debido al tratamiento que tienen tanto los atributos textuales como lo numéricos. Ya que cuando estamos usando los comentarios de los pacientes, es posible que se introduzca mayor ruido en dichos datos. Por otro lado, cabe destacar el buen funcionamiento que tienen los modelos de regresión. Además, el error obtenido en Random Forest, es lógico, debido a que la predicción con datos derivados de un texto es mucho más difícil. Asimismo, se aprecia como no se han añadido los árboles de decisión a la tabla anterior, ya que para dicha técnicas hemos obtenido la precisión y no el error fuera de la muestra. En donde se obtiene una precisión del 35% para atributos numéricos y de un 52.3, para textuales, cosa que nos lleva a la misma conclusión de antes, en donde la predicción con datos derivados de un texto es mucho más difícil.