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).

Con respecto a las técnicas, debemos destacar la componente subjetiva que tienen las reglas de asociación a la hora de su elección. Se observó, como 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 este tipo de algoritmos de clasificación no resultan adecuados en nuestro problema**, debido a dos cuestiones:

1. 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).
2. El hecho de 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, se aplicó el método 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	**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

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.

Técnicas	**Error Test (%)**	**Error Train (%)**
Naive Bayes	19.2331	22.4371
SVM	-	-
Random Forest (atributos numéricos)	35.14377	
Random Forest (atributos textuales)	35.14377	
Árboles de Decisión (atributos numéricos)	35	-
Árboles de Decisión (atributos textuales)	52.3	-
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