MÁSTER EN BIG DATA. TECNOLOGÍA Y ANALÍTICA AVANZADA

TRABAJO FIN DE MÁSTER

MACHINE LEARNING APLICADO AL INCUMPLIMIENTO DE LA LEY ORGÁNICA DE PROTECCIÓN DE DATOS

Autor: María Patricia Medina de las Heras

Director: Carlos Morrás Ruiz-Falcó

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Machine Learning aplicado al incumplimiento de la Ley Orgánica de Protección de Datos en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2024/25 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: María Patricia Medina de las Heras Fecha: 16/05/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Firmado por MORRAS RUIZ-FALCO CARLOS - ***0907** el día 17/05/2025 con un certificado emitido por AC FNMT Usuarios

Fdo.: ……. Fecha: ……/ ……/ ……

MÁSTER EN BIG DATA. TECNOLOGÍA Y ANALÍTICA AVANZADA

TRABAJO FIN DE MÁSTER

MACHINE LEARNING APLICADO AL INCUMPLIMIENTO DE LA LEY ORGÁNICA DE PROTECCIÓN DE DATOS

Autor: María Patricia Medina de las Heras

Director: Carlos Morrás Ruiz-Falcó

Madrid

Abstract

The protection of personal data is an emerging issue due to the development of the most innovative technologies of Big Data. The “Agencia Estatal de protección de Datos” is the spanish public organization responsible for ensuring compliance with the Organic Personal Data Protection Law.

This project consists of a descriptive analysis of the resolutions published by the AEPD, the extraction of the most relevant characteristics of the resolutions and the analysis of the most suitable Machine Learning models in order to classify the resolutions.

The techniques used to carry out these analysis are Web Scraping techniques in order to obtain data, regular language techniques to extract the relevant information of this data and Natural Language Processing techniques to classify the resolutions.

Resumen

La protección de datos personales es un problema que va en aumento con el desarrollo de las tecnologías más innovadoras de tratamiento de grandes volúmenes de datos. La Agencia Española de Protección de Datos es el organismo público español que se encarga de velar por el cumplimiento de la Ley Orgánica de Protección de Datos Personales.

En este trabajo se recoge un análisis descriptivo de las resoluciones publicadas en la AEPD, la extracción de las características más relevantes de estas resoluciones y un estudio de los modelos más adecuados de Machine Learning para realizar una clasificación de las resoluciones.

Las técnicas empleadas para llevar a cabo estos análisis son técnicas de Web Scraping para la obtención de los datos, técnicas de lenguaje regular para la extracción de la información relevante y técnicas de Procesamiento del Lenguaje Natural para la clasificación de las resoluciones.

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

Con el desarrollo constante de las nuevas tecnologías, en particular del Big Data y la inteligencia artificial, se ha producido un cambio de paradigma en cuanto a términos de protección de datos, el desarrollo de estas tecnologías genera brechas en la privacidad de las personas, ya que el avance realizado muchas veces no se encuentra regulado y por tanto se produce un dilema entre lo éticamente correcto en términos de protección de datos y el desarrollo y crecimiento tecnológico.

El artículo 18.4 de la Constitución Española de 1978 establece que *la Ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos.* Sin embargo, a medida que la tecnología y la inteligencia artificial evoluciona, aumenta la capacidad de utilizar la información personal de formas que pueden entrometerse en los intereses de privacidad al elevar el análisis de la información personal a nuevos niveles de potencia y velocidad.

Para garantizar y proteger las libertades públicas y los derechos fundamentales de las personas físicas, en lo que afecta a datos personales, en 1992 se crea la Agencia Española de Protección de Datos (AEPD). Esta es la autoridad pública independiente encargada de velar por el cumplimiento de la Ley Orgánica de Protección de Datos de Carácter Personal (LOPD) en España.

En media, en la AEPD se publican 3.400 resoluciones anualmente, esto son 9’3 resoluciones diarias, todas ellas referentes al incumplimiento de la LOPD. Actualmente, no existe ningún proyecto en el ámbito tecnológico que recoja la información más relevante de estas resoluciones.

El objetivo principal del proyecto es estimar el importe de la multa asociada a una resolución, en función del análisis del texto de los hechos, el cuál recoge los artículos incumplidos, el tipo de infracciones relativas a datos de carácter personal cometidas, etc.

Como objetivos secundarios se quiere hacer un análisis de las palabras más importantes en las resoluciones y ver si hay diferencias entre las palabras más importantes del texto que recoge los hechos y el texto que recoge la conclusión de la resolución. También se quiere estimar el tipo de infracciones que ha cometido cada acusado, para poder realizar una clasificación de la gravedad del problema.

Además, se quiere comparar los resultados de este mismo proyecto realizado en 2022, con los resultados a fecha de mayo de 2025. Con la intención de entender si en tres años la evolución de la Inteligencia Artificial y los nuevos desarrollos tecnológicos han influido en el panorama general de protección de datos personales. Y entender si el desarrollo de los últimos tres años tiene suficiente peso como para suponer un cambio significativo en el análisis global de los datos.

Para abordar el proyecto se van a utilizar técnicas de Web Scraping para recoger los datos, técnicas de lenguaje regulares para extraer caracteres de interés del texto y para reemplazar o eliminar aquella información que no sea relevante en el análisis y técnicas de Procesamiento del Lenguaje Natural, del inglés Natural Language Procesing (NLP) para analizar el texto, que es información no estructurada y calcular los correspondientes modelos de clasificación.

Capítulo 2. ESTADO DEL ARTE

2.1 TF-IDF

El término TF-IDF proviene del inglés Term Frequency- Inverse Document Frequency, es decir, es una medida numérica que expresa la frecuencia de ocurrencia de cada término en la colección de documentos, es decir, cuán relevante es una palabra para un documento en una colección.

El valor TF-IDF aumenta proporcionalmente al número de ocurrencias de una palabra en un documento, pero es compensada por la frecuencia de esa palabra en la colección de documentos.

El número de veces que un término aparece en un documento se denomina frecuencia de término (TF). Y se denomina frecuencia inversa de un documento (IDF) a la medida de lo común que es un término en una colección de documentos.

De forma que la métrica TF-IDF se calcula como el producto de dos medidas, la frecuencia de término y la frecuencia inversa de documento.

Calculamos cada uno de los términos de la siguiente forma:

$$tf(t,d) = \frac{número\ de\ veces\ que\ aparece\ t\ en\ d}{número\ de\ palabras\ en\ d}$$

$$idf(t) = \log\left(\frac{N}{df+1}\right)$$

Siendo, t un término, d un documento y df el número de documentos en los que aparece el término t.

A la hora de calcular la frecuencia inversa de un documento se observa que el denominador se ajusta sumando 1 al número de documentos en los que aparece el término *t*. Esto se debe

a que puede darse el caso de que haya documentos que no contengan el término *t*, y entonces estaríamos dividiendo entre 0 y el resultado sería indefinido.

2.2 CLASIFICADOR DE NAÏVE BAYES

El clasificador de Naïve Bayes es un algoritmo simple de clasificación que se basa en el Teorema de Bayes y en un conjunto de hipótesis de independencia condicional de las variables predictoras dada la variable clase.

Teorema *(Bayes, 1764): Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por* $P(A)$ *y* $P(B)$ *respectivamente, verificándose que* $P(B) > 0$. *Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir,* $P(A)$ *y* $P(B)$*, así como la probabilidad condicionada del suceso B dado el suceso A, es decir* $P(B|A)$. *La probabilidad a posteriori del suceso A conocido que verifica el suceso B, es decir* $P(A|B)$*, puede calcularse a partir de la siguiente fórmula:*

$$P(A|\,B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{\sum_{A'} P(A')P(B|A')}$$

Se aplica el Teorema de Bayes a variables aleatorias multidimensionales y al uso de este como un clasificador. El objetivo es estimar el valor de una variable dependiente C, el cual está condicionado por un conjunto de variables independientes $F_1, F_2, \dots, F_n$, se calcula a partir de la siguiente fórmula:

$$P(C|\,F_1,\,F_2,\dots,F_n) = \frac{P(C)P(F_1,\,F_2,\dots,F_n|C)}{P(F_1,\,F_2,\dots,F_n)}$$

El denominador, en la práctica, se considera constante, ya que este no depende del valor de C, y los valores de F_i son los valores de los datos de partida. El numerador puede escribirse como una probabilidad compuesta, de la siguiente forma:

$$P(C,F_1,\dots,F_n) = P(C)\,P(F_1,\dots,F_n|C) = P(C)\,P(F_1|C)P(F_2,\dots,F_n|C,F_1)$$

Se asume independencia condicional, es decir se asume que cada F_i es independiente de cualquier F_j para todo $j \neq i$ cuando están condicionadas a C. Por tanto, la probabilidad compuesta se simplifica a:

$$P(C, F_1, \dots, F_n) = P(C)\, P(F_1|C)P(F_2|C)P(F_3|C) \cdots = P(C) \prod_{i=1}^{n} P(F_i|C)$$

Aplicando las hipótesis de la distribución condicional de C sobre las variables clasificatorias, el teorema de Bayes se puede expresar de la siguiente forma:

$$P(C|\, F_1, \dots, F_n) = \frac{1}{Z} P(C) \prod_{i=1}^{n} P(F_i|C)$$

Donde Z es un factor que depende únicamente de los valores de $F_1,\ F_2, \dots, F_n$, es decir, es constante si los valores de F_i son conocidos.

El objetivo del clasificador de Naïve Bayes (Good, 1965), es encontrar la clase más probable, una vez conocidas un conjunto de características, por tanto, el clasificador de Naïve Bayes se define como:

$$c^* = \arg\max_c P(C = c) \prod_{i=1}^{n} P(X_i = x_i|C = c)$$

Siendo c^*, la clase más probable a posteriori una vez conocido el valor de las variables independientes.

2.3 *SUPPORT VECTOR MACHINE*

Las máquinas de vectores de soporte, del inglés Support Vector Machine (SVM) son un conjunto de algoritmos (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995, Vapnik et al., 1997) que analizan los datos de entrada para resolver un problema de clasificación o regresión.

2.3.1 LINEAR SUPPORT VECTOR MACHINE

En concreto nos vamos a centrar en el modelo lineal. La idea básica del funcionamiento del algoritmo es crear una línea o hiperplano que separe los datos en las diferentes clases.

Como podemos observar en la Ilustración 1, la solución no es única, ya que hay más de una línea que separe los datos en las dos clases buscadas. Para encontrar la solución óptima, el algoritmo de SVM busca los dos puntos de cada clase más cercanos a la línea, a estos puntos se les llama vectores de soporte. Después, se calcula la distancia entre la línea y los vectores de soporte, a esta distancia se le denomina margen. La solución óptima es aquella para la cual el margen es máximo.

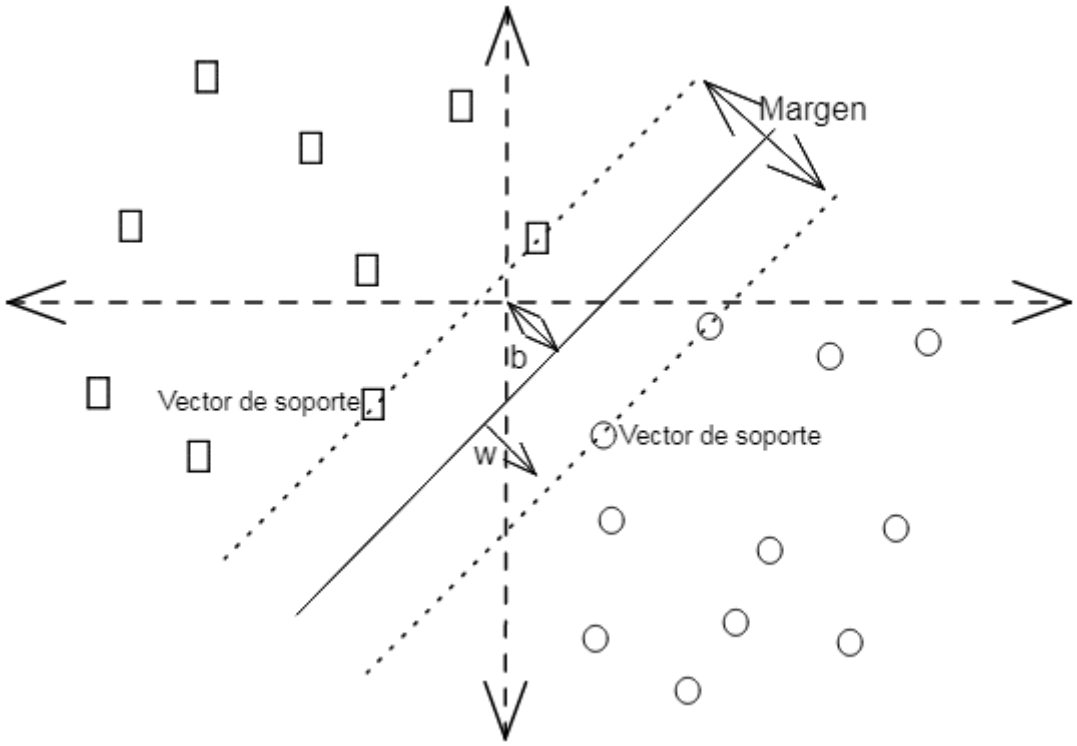


Ilustración 1. Ejemplo de clasificación en dos grupos mediante Linear SVM

Un hiperplano en un espacio euclídeo n-dimensional es un subconjunto plano de n-1 dimensiones de ese espacio, de forma que lo divide en dos partes disconexas. La ecuación del hiperplano es la siguiente:

$$w^T x + b = 0$$

Por tanto, la solución de la máquina de soporte vectorial será la siguiente:

$$u = \vec{w} \cdot \vec{x} - b$$

Donde, $\vec{w}$ es el vector normal del hiperplano y $\vec{x}$ es el vector de entrada

En el caso lineal, el margen está definido por la distancia al hiperplano de los puntos de cada clase más cercanos. Por tanto, el problema puede resolverse mediante una función de optimización, dado que maximizar el margen es equivalente a minimizar la norma del vector que define al hiperplano buscado, la función de optimización a resolver es la siguiente:

$$\min \frac{1}{2}||w||^2 \, , sujeto\ a, y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1\ \forall i$$

En la práctica, no siempre el espacio es linealmente separable, y para solucionar esto hay que aplicar distintas técnicas como variar la dimensionalidad, de forma que en el nuevo espacio el problema sí que sea linealmente separable. Utilizar variables de holgura, las cuáles penalizan el resultado de la optimización, pero permiten llegar a una solución. O aplicar otro tipo de kernel, como un kernel polinómico, gaussiano o sigmoidal.

2.4 MULTINOMIAL LOGISTIC REGRESSION

El modelo de regresión logística multinomial es una generalización del modelo de regresión logística binario clásico, de forma, que permite categorizar una variable dependiente en más de dos clases.

La fórmula para estimar la variable dependiente es similar a la de la regresión logística binaria, calculándose como el logaritmo de la razón de oportunidades, del inglés *Odds Ratio* (OR):

$$logit(P) = \log(OR) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

El valor logit puede tomar valores entre 0 e infinito y explica cuán más probable es que una observación pertenezca a la clase objetivo, en lugar de a otra clase.

La razón de oportunidades estima el cambio en las probabilidades de pertenecer al grupo objetivo cuando se produce un incremento de una unidad en el predictor.

Despejando de la fórmula anterior la probabilidad de que un caso esté en una categoría particular *p*, tenemos:

$$p = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n\right)}$$

Para estimar los parámetros $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$ del modelo de regresión logística multinomial se aplica el criterio de máxima verosimilitud. Es decir, se obtienen los parámetros que maximizan la función de verosimilitud, siendo la función de máxima verosimilitud asociada al modelo de regresión logística:

$$lnL\left(\left(x^{(1)}, c^{(1)}\right), \ldots, \left(x^{(N)}, c^{(N)}\right), \beta_0, \beta_1, \ldots, \beta_n\right) =$$

$$\sum_{j=1}^{N} c^{(j)} \left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right) - \sum_{j=1}^{N} \ln\left(1 + e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}\right)$$

2.5 ONE HOT ENCODING

La codificación One Hot es el proceso de transformar las variables categóricas para poder ser utilizadas en algoritmos de Machine Learning con el objetivo de mejorar las predicciones.

El nombre de One Hot proviene de la transformación que esta técnica realiza sobre la variable categórica, ya que crea un grupo de bits, para los cuales hay un único bit alto (1) y el resto son bits bajos (0).

Esta técnica se utiliza para distinguir inequívocamente a una palabra de un vocabulario del resto de palabras del vocabulario.

Además, permite asegurarse de que un algoritmo no asuma que valores mayores son más importantes que valores bajos. Aplicado al ámbito numérico, el 15 es mayor que el 3 y sin embargo esto no lo hace más importante. De igual manera esto se puede ver reflejado en un vocabulario, de forma que acidez es mayor que ácido y esto no lo hace más importante y puede darse la casuística de que queramos diferenciar entre estas dos palabras, sin necesidad de hacer creer al algoritmo que una es más importante que la otra.

2.6 REDES NEURONALES

Una red neuronal es un modelo computacional complejo compuesto por un conjunto de algoritmos modelados de forma que imitan vagamente el comportamiento del cerebro humano y que están diseñados para reconocer patrones en los datos.

Para el tratamiento de datos no estructurados, como imagen, audio o texto, la información se traduce a datos vectoriales, ya que una red neuronal es capaz de reconocer patrones numéricos, para posteriormente agruparlos o clasificarlos.

Una red neuronal está compuesta por un conjunto de capas interconectadas, que a su vez cada capa está compuesta por un conjunto de nodos o neuronas. En la Ilustración 2 puede verse el funcionamiento de una neurona.

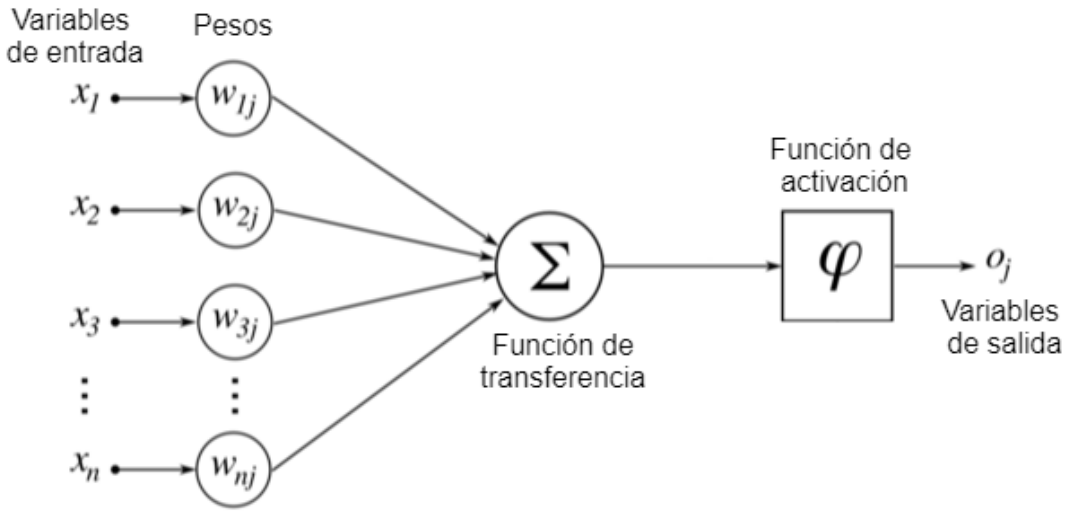


Ilustración 2. Funcionamiento de una neurona

Una neurona combina los datos de las variables de entrada con un conjunto de coeficientes o pesos que amplifican o amortiguan los valores de las variables de entrada, de forma que se asigna una importancia a las entradas con respecto a la tarea que el algoritmo está aprendiendo. Cada neurona genera una salida, esta salida se genera a partir de la suma de los productos de los pesos y las variables de entrada, también conocida como función de transferencia y de una función de activación.

Una capa, es un conjunto de neuronas puestas en paralelo, que hacen esta función de forma simultánea. En las redes neuronales clásicas, la salida de cada capa es la entrada de la siguiente. En la Ilustración 3 puede verse un esquema del funcionamiento de una red neuronal completa.

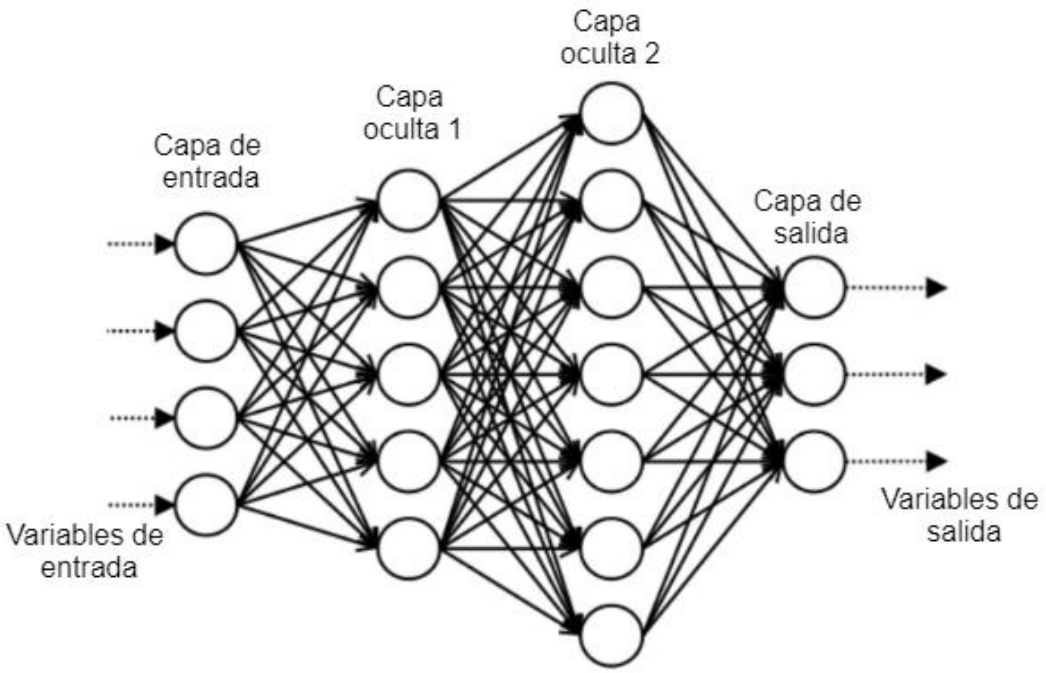


Ilustración 3. Esquema de una Red Neuronal

Capítulo 3. DESCRIPCIÓN DEL TRABAJO

3.1 OBTENCIÓN DE LOS DATOS

La obtención de los datos con los que se ha desarrollado este trabajo se ha realizado mediante técnicas de Web Scraping y un procesamiento posterior de estos. El flujo de datos está representado en la Ilustración 4. Para realizar el Web Scraping, fundamentalmente se han usado dos librerías de Python: Requests y BeautifulSoup.

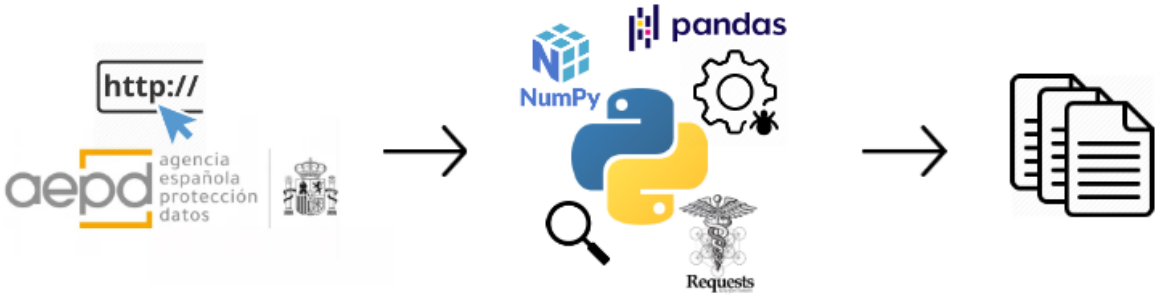


Ilustración 4. Diagrama de flujo de datos

La librería Request es una librería para construir y enviar peticiones HTTP a una Web, en nuestro caso se envían peticiones de tipo GET a la página de resoluciones de la AEPD, con el objetivo de obtener los datos de esta.

La estructura de la página Web de la AEPD es la que se muestra en la Ilustración 5, las resoluciones están publicadas en formato PDF y están distribuidas en varias páginas, por tanto, será necesario realizar varias peticiones GET, una por cada página. Una vez enviada la petición se obtiene una respuesta en formato binario que posteriormente se parsea.

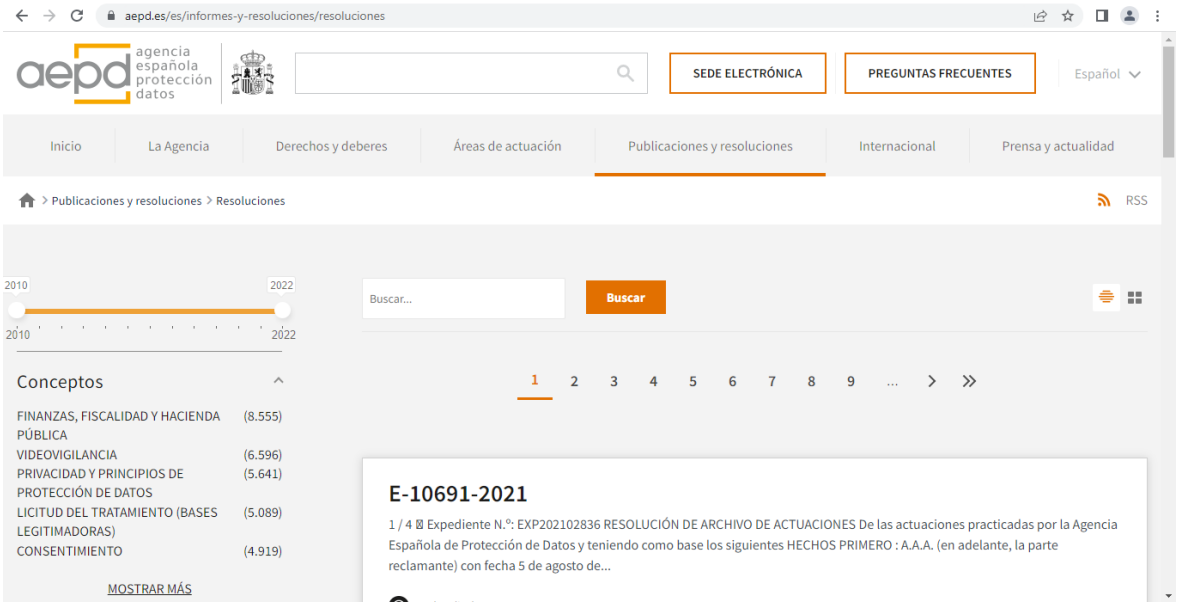


Ilustración 5. Página de resoluciones de la AEPD

La librería BeautifulSoup es una librería para extraer datos de archivos HTML y XML. Esta librería crea un árbol que representa la estructura sintáctica de una cadena de caracteres de acuerdo con una gramática concreta, en nuestro caso este árbol se crea a partir del código fuente de la página, de forma que permite extraer los datos de esta de una manera jerárquica y más legible.

Una vez que se ha parseado el HTML y se tienen organizadas todas las etiquetas de la página Web, extraer las resoluciones en formato PDF es relativamente sencillo. Se accede a las etiquetas de tipo “a” que, en HTML, son aquellas que contienen hipervínculos, dentro de estas se selecciona la información que contiene el atributo “href”, que es la referencia a la página de destino, en caso de que esta termine como .pdf, se guarda en una carpeta local.

3.2 PREPROCESAMIENTO DE LOS DATOS

Mediante el proceso de WebScraping se han obtenido un total de 42.417 resoluciones desde el año 2002 hasta junio del año 2022. Y un total de 45.210 resoluciones desde el año 2002 hasta el año 2025. Sin embargo, estas resoluciones son únicamente documentos de texto,

para poder realizar posteriormente modelos de Machine Learning, hay que preprocesar este texto.

En primer lugar, se importan todas las resoluciones en una estructura de datos de tipo DataFrame con una única columna que contiene el texto de la resolución. Además, se van a extraer algunas características que se han considerado importantes a la hora de analizar las resoluciones. Esta extracción se realiza mediante expresiones regulares y reglas duras, podemos realizar el preprocesado de este modo ya que partimos de un modelo de resolución estándar que explicaremos a continuación.

3.2.1 ESTRUCTURA DE LOS DOCUMENTOS

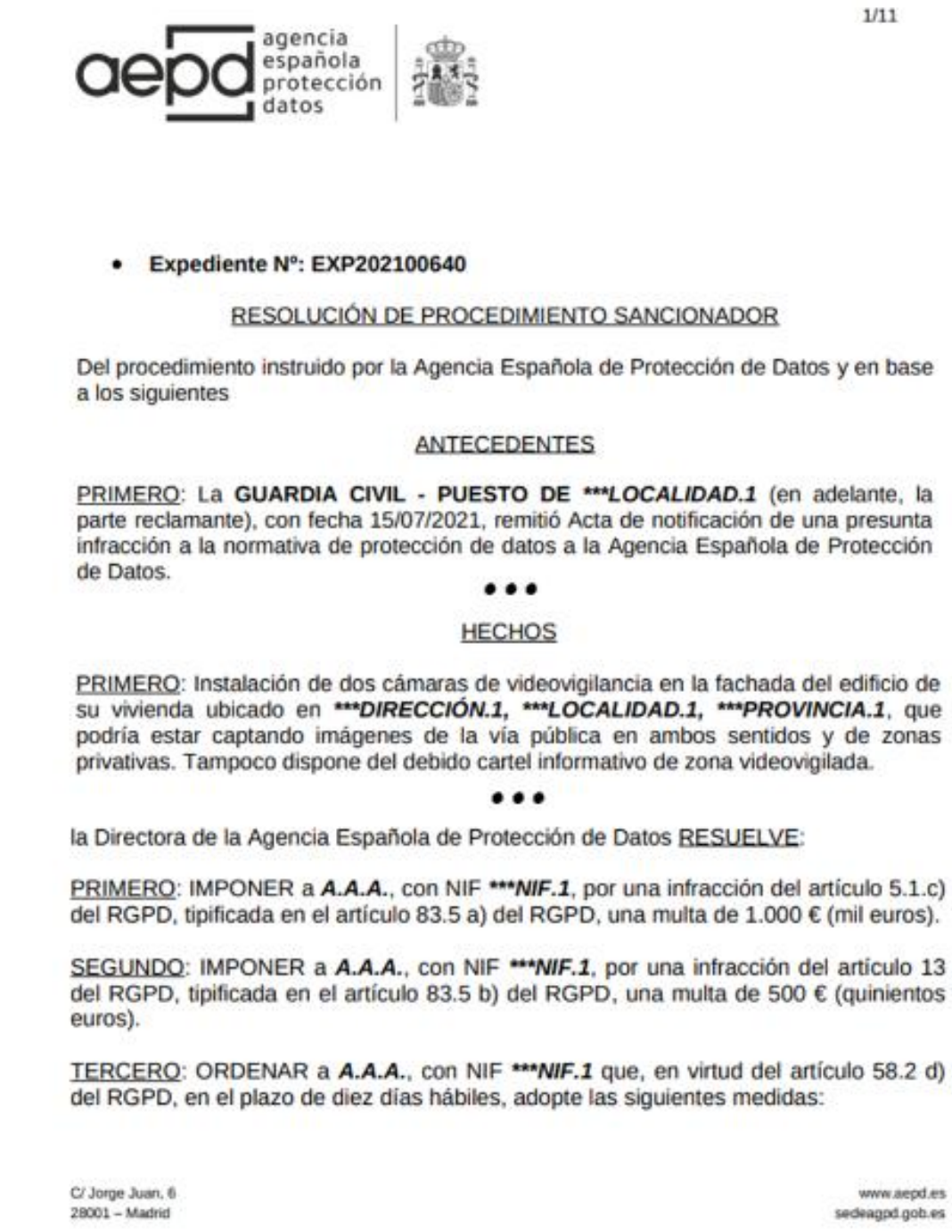
La mayoría de las resoluciones, siguen una estructura general que está establecida por la AEPD. En la Ilustración 6 se puede observar un ejemplo de una resolución.

Todas las resoluciones tienen una cabecera y un pie de página, en la cabecera está incluido el sello de la AEPD y el número de página y en el pie de página están incluidos los datos de contacto de la AEPD.

Lo primero que aparece en todas las resoluciones es el número de resolución, aunque no siempre viene precedido del texto “Expediente Nº”.

El apartado de antecedentes sólo aparece en las resoluciones que están vinculadas a alguna resolución anterior que ha sido recurrida, en este apartado se recoge la información principal de las resoluciones anteriores.

El apartado de los hechos aparece en todas las resoluciones y recoge los hechos que se denuncian. Por último, hay una conclusión en todas las resoluciones, en concreto en la del ejemplo anterior, viene expuesta detrás de la palabra clave “RESUELVE”, esta palabra clave puede variar, pero siempre hay una palabra clave y en mayúsculas con el resultado de la resolución.



1/11

agencia española protección datos

- **Expediente Nº: EXP202100640**

RESOLUCIÓN DE PROCEDIMIENTO SANCIONADOR

Del procedimiento instruido por la Agencia Española de Protección de Datos y en base a los siguientes

ANTECEDENTES

PRIMERO: La **GUARDIA CIVIL - PUESTO DE ****LOCALIDAD.1*** (en adelante, la parte reclamante), con fecha 15/07/2021, remitió Acta de notificación de una presunta infracción a la normativa de protección de datos a la Agencia Española de Protección de Datos.

• • •

HECHOS

PRIMERO: Instalación de dos cámaras de videovigilancia en la fachada del edificio de su vivienda ubicado en ******DIRECCIÓN.1, ***LOCALIDAD.1, ***PROVINCIA.1***, que podría estar captando imágenes de la vía pública en ambos sentidos y de zonas privativas. Tampoco dispone del debido cartel informativo de zona videovigilada.

• • •

la Directora de la Agencia Española de Protección de Datos RESUELVE:

PRIMERO: IMPONER a ***A.A.A.***, con NIF ******NIF.1***, por una infracción del artículo 5.1.c) del RGPD, tipificada en el artículo 83.5 a) del RGPD, una multa de 1.000 € (mil euros).

SEGUNDO: IMPONER a ***A.A.A.***, con NIF ******NIF.1***, por una infracción del artículo 13 del RGPD, tipificada en el artículo 83.5 b) del RGPD, una multa de 500 € (quinientos euros).

TERCERO: ORDENAR a ***A.A.A.***, con NIF ******NIF.1*** que, en virtud del artículo 58.2 d) del RGPD, en el plazo de diez días hábiles, adopte las siguientes medidas:

C/ Jorge Juan, 6
28001 – Madrid

www.aepd.es
sedeagpd.gob.es

Ilustración 6. Estructura de una Resolución

3.2.2 IMPUTACIÓN DE VARIABLES

3.2.2.1 Tipo de Infracción

Las resoluciones se clasifican como leves o graves, dependiendo de la naturaleza de los hechos y los artículos incumplidos. Por tanto, se ha creado una variable categórica que recoge si hay infracciones cometidas asociadas a una resolución y en tal caso si son leves, graves o de ambos tipos.

Para crear esta variable, se ha buscado con una expresión regular en cada texto las palabras: leve, leves, grave y graves. Ya que se ha observado que, en gran medida, estas palabras sólo aparecen en las resoluciones a la hora de calificar las infracciones y que en una misma resolución puede constar que se ha cometido más de una infracción.

3.2.2.2 Importe de la multa

Las resoluciones de las infracciones pueden ser varias, algunas de ellas son las siguientes:

- Resolución mediante el pago de una multa
- Resolución mediante exoneración de los cargos
- Resolución mediante apercibimiento de los hechos
- Resolución mediante el archivo de esta
- Resolución mediante una impugnación

Debido al gran abanico de casuísticas y que no se han podido investigar todas ellas, se ha querido estudiar con detalle la casuística de si la resolución es mediante multa o no y el importe de esta.

En este caso se ha buscado en el texto mediante la palabra euro y el símbolo €, como se puede ver en el ejemplo de la Ilustración 6, en una misma resolución puede haber más de una única multa, también hay casos para los que aparece una posible multa y esta se recoge en un rango de valores. Por tanto, se ha creado una variable que recoja todas las posibles multas que aparecen en el documento y se ha calculado la multa media.

3.2.3 PREPROCESAMIENTO DEL TEXTO

Para analizar la información mediante Procesamiento del Lenguaje Natural, del inglés Natural Language Processing (NLP), es necesario aplicar algunas técnicas de preprocesamiento del texto.

En primer lugar, se eliminan todos los pies de página, ya que a la hora de analizar el texto no aporta información y están repetidos. Además, eliminamos las URLs, y las direcciones de correo electrónico, ya que suelen crear problemas cuando se procesa el resto del texto.

A continuación, se convierte todo el texto a minúsculas, para que a la hora de analizar el texto no haya diferencias entre la misma palabra debidas a la grafía de esta. Se eliminan del texto los números y los caracteres especiales como los signos de puntuación.

Se eliminan las tildes, ya que esto ayuda a que si por error, falta alguna tilde, no se entienda la palabra como una palabra distinta. Hay que tener en cuenta que esto también puede crear ambigüedad, ya que, en castellano, hay palabras que se diferencian de otras únicamente por la tilde.

Se eliminan las stopwords, que son aquellas palabras que cuando no van acompañadas de otras palabras no aportan información relevante, como pueden ser las preposiciones, artículos, pronombres y adverbios.

Por último, se realiza un proceso de lematización que consiste en, dada una forma flexionada de una palabra, hallar el lema correspondiente. Así se consigue reducir el vocabulario de nuestro texto, ya que en castellano todos los verbos se conjugan y, por tanto, en caso de no realizar una lematización el mismo verbo en tiempo presente y en tiempo pasado actuaría como dos palabras completamente distintas. Se ha elegido el algoritmo de lematización sobre otros parecidos como puede ser el de stemming, porque este nos permite reducir las formas flexivas de cada palabra a una base común sin perder el sentido de las palabras que aparecen en el texto.

Capítulo 4. ANÁLISIS DE RESULTADOS

Como se ha comentado con anterioridad, partimos de 42.417 resoluciones hasta 2022 y 45.210 resoluciones hasta 2025. Vamos a utilizar técnicas de NLP para estimar la información que se ha extraído del texto en variables cualitativas y vamos a comparar si en los últimos 3 años hay una diferencia en los resultados.

4.1 ANÁLISIS DESCRIPTIVO DE LOS DATOS

Se realiza un análisis descriptivo de los datos con la finalidad de entender mejor de qué datos partimos. Al tratarse de texto, que son datos no estructurados, la técnica elegida para realizar este análisis descriptivo son las nubes de palabras.

Una nube de palabras, del inglés Word Cloud, es una representación visual de las palabras que pertenecen a un texto. En cada nube de palabras el tamaño de la palabra indica la frecuencia con la que esa palabra aparece en el texto. De forma que, a mayor tamaño, mayor frecuencia.

Como hemos visto en la estructura del documento, todas las resoluciones pueden separarse en los hechos y en la resolución, por tanto, se construirá una nube de palabras por cada fragmento. En el caso de las resoluciones que cuentan con antecedentes, esta parte del texto irá incluida junto con los hechos.

En primer lugar, se han construido las nubes de palabras teniendo en cuenta únicamente la frecuencia absoluta de las distintas palabras en los documentos.

En la Ilustración 7.1 se puede observar la nube de palabras para todos los hechos hasta 2022, y en la Ilustración 7.2 se observa la nube de palabras para todos los hechos hasta 2025. En particular, se ha decidido representar bigramas, que son el grupo de dos palabras, ya que

estos aportan más contexto. Como el texto ha sido lematizado se observa que todas las palabras están en género masculino y los bigramas carecen de preposiciones.

El resultado es el esperado, obteniéndose como bigramas más frecuente los que están en mayor tamaño, que, mirando algunas resoluciones, sabemos que vienen de: protección del dato, dato de carácter personal, artículo de la LOPD, ley orgánica, agencia española y responsable del tratamiento, entre otras. Además, vemos que aunque la frecuencia de los bigramas no es la misma con la actualización de las nuevas resoluciones, los bigramas más frecuentes sí son los mismos a pesar de las nuevas resoluciones incluidas en el análisis.

Ilustración 7.1. Nube de bigramas según la frecuencia para los Hechos hasta 2022

Ilustración 8.2. Nube de bigramas según la frecuencia para los Hechos hasta 2025

En las Ilustraciones 8.1 y 8.2 se puede observar la nube de bigramas para todas las resoluciones, hasta 2022 y hasta 2025, respectivamente. Al igual que para los hechos, el resultado es el esperado, obteniéndose como bigramas más frecuentes los mismos que para los hechos y añadiéndose: plazo de un mes, dispuesto en el artículo, contar con la notificación, conforme a lo establecido, etc. Y nuevamente vemos que la diferencia entre 2022 y 2025 solo se da en la frecuencia de los bigramas que es un poco diferente entre los dos periodos, lo que se ve en el tamaño de estos. Pero no hay nuevos bigramas con frecuencia alta en 2025 que no se hubiesen obtenido ya en 2022.

Ilustración 9.1. Nube de bigramas según la frecuencia para las Resoluciones hasta 2022

Ilustración 10.2. Nube de bigramas según la frecuencia para las Resoluciones hasta 2025

Estas nubes de palabras no aportan información significativa, porque ya partimos de la base de que todos los textos son de la AEPD y por tanto tratan temas relativos a protección de datos e incumplimientos de artículos. Para solucionar esto, se han construido las nubes de palabras a partir del TF-IDF, en este caso se representan palabras en lugar de bigramas.

En la Ilustración 9.1 y en la Ilustración 9.2 se puede ver que las palabras representadas en la nube de palabras construida mediante TF-IDF para los hechos son más representativas que las anteriores. Además, se ve que en este caso sí que hay diferencias entre las palabras obtenidas en 2022 y las de 2025, sin embargo, ahora hay ciertas palabras que a simple vista pueden no ser comprensibles, a continuación, se explican algunas de estas.

Se observa que, en el año 2022, la palabra "FT" es la más frecuente, se ha observado en más de una resolución que estas siglas provienen de PBC-FT que es la legislación sobre Prevención del Blanqueo de Capitales y Financiación del Terrorismo. Al estar separadas por un guión, después de eliminar los caracteres especiales, se reconoce PBC como una palabra y FC como otra palabra.

Otra palabra muy frecuente es "dubio" que proviene de la locución latina *in dubio pro reo*. Este es un principio de Derecho Penal que indica que, en caso de duda por parte del juez o el tribunal sobre la culpabilidad de un acusado, la sentencia o decisión judicial debe favorecer al acusado.

Las siglas MSP (Managed Services Providers) o proveedor de servicios gestionados, son empresas que se encargan de la administración remota de la infraestructura de tecnología de la información y los sistemas informáticos de un negocio bajo un modelo de subscripción por uso de servicio.

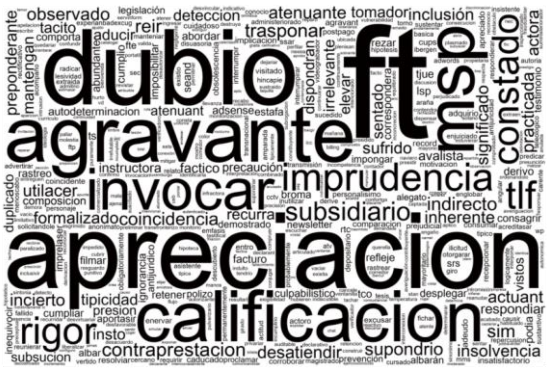


Ilustración 11.1. Nube de palabras según TF-IDF para los Hechos hasta 2022

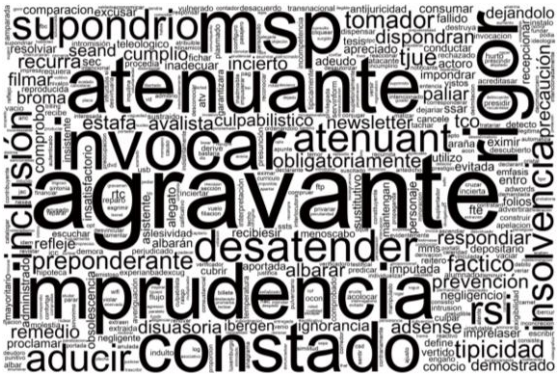


Ilustración 12.2. Nube de palabras según TF-IDF para los Hechos hasta 2025

De igual manera, en la Ilustración 10.1 y 10.2 está representada la nube de palabras a partir de TF-IDF para las resoluciones hasta 2022 y hasta 2025, respectivamente. Igual que para los hechos, vemos diferencias entre los dos periodos y encontramos siglas que son poco comprensibles.

Las siglas LAJG vienen de Ley de Asistencia Jurídica Gratuita. La Asistencia Jurídica Gratuita te proporciona los medios para que puedas defender tus derechos ante la Justicia si no dispones de recursos económicos para litigar, o independientemente de tu situación económica, si los procesos en los que necesitas defensa o tus circunstancias personales así lo determinan.

Tanto las palabras INZERTIA como SLUCONTEC son nombres propios. La primera es una empresa de consultoría financiera y la segunda es una empresa de consultoría de formación tecnológica.

Cotino es el apellido del presidente de la Agencia Estatal de Protección de Datos. Quien fue nombrado presidente en febrero de 2025. Esto explica que esta palabra se repita con mucha frecuencia en las Resoluciones para todo el periodo 2002-2025, sin embargo, no apareciese en el periodo 2002-2022.

Ilustración 13.1. Nube de palabras según TF-IDF para las Resoluciones hasta 2022

Ilustración 14.2. Nube de palabras según TF-IDF para las Resoluciones hasta 2025

En este caso, observamos que las palabras de la nube de puntos de los hechos y las palabras de la nube de puntos de las resoluciones son muy diferentes entre sí. Teniendo en la primera nube palabras más aplicadas al ámbito jurídico y en la segunda, palabras correspondientes a la resolución tomada por la AEPD y nombres propios de empresas y personas físicas, así como algunas palabras mal escritas, debido a que, al hacer la lematización, estas no se encuentran en el diccionario de vocabulario español.

4.2 MODELOS DE CLASIFICACIÓN DEL TIPO DE INFRACCIÓN

Se han aplicado varios modelos de clasificación para estimar el tipo de infracción. Como se ve en la Ilustración 11, partiendo del texto de las resoluciones de la AEPD, se usará un clasificador para estimar el tipo de infracción asociada a cada resolución. Como se ha comentado en el apartado 3.2.2.1. las clases posibles son: no infracción, grave, leve o ambas.

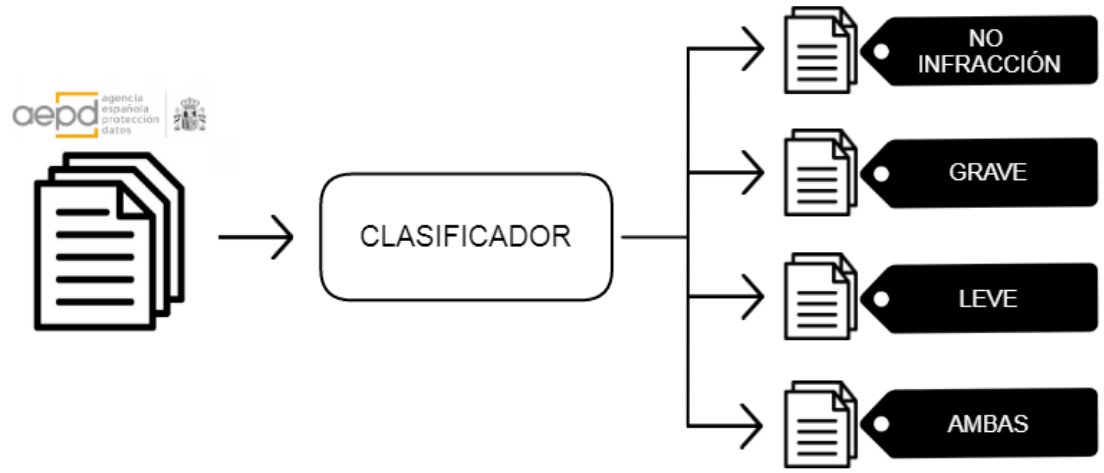


Ilustración 15. Diagrama del funcionamiento del clasificador

Antes de calcular los estimadores, se divide el conjunto de datos asignando un 80% de los datos al conjunto de datos de entrenamiento y un 20% al conjunto de datos de evaluación, manteniendo la distribución de las clases del conjunto de datos total, ya que estas están desbalanceadas. De forma que el 74% de las resoluciones están clasificadas como no infracción, el 2% están clasificadas como infracción grave, el 20% están clasificadas como infracción leve y el 4% están clasificadas como infracción doble.

4.2.1 CLASIFICADOR NAÏVE BAYES

Aun siendo un estimador sencillo, para el conjunto de datos 2002-2022, la precisión del modelo es bastante alta. Obteniendo una precisión del 85’7% en el conjunto de datos de entrenamiento y una precisión del 85% en el conjunto de datos de validación. Sin embargo, cabe destacar que este valor tampoco es un valor alto, ya que al estar tan desbalanceadas las clases, si el algoritmo clasificase el 100% de las resoluciones como “no infracción”, la precisión del modelo sería del 74%.

De igual forma, para el conjunto de datos 2002-2025, se obtiene una precisión del 83% para el conjunto de datos de entrenamiento y una precisión del 82.8% para el conjunto de datos de validación. Que no es elevada ya que, si el algoritmo clasificase el 100% de las resoluciones como “no infracción”, la precisión del modelo sería del 75,5% para este periodo.

En la Ilustración 12.1 y en la Ilustración 12.2 se puede ver la representación de la matriz de confusión para el conjunto de datos de validación. Se observa que el modelo ha optado por no clasificar ninguna de las observaciones en los grupos de las clases minoritarias, de forma que el modelo nunca predice que en una resolución haya una infracción grave o que en una resolución haya una infracción grave y una leve. Por tanto, el modelo se equivoca el 100% de las veces al clasificar una infracción grave o al clasificar una infracción doble. Esto pasa, tanto en el conjunto de datos del periodo 2002-2022, como en el del periodo 2002-2025.

También se observa que, entre los dos grupos restantes, el modelo se equivoca más al clasificar las resoluciones con una infracción leve, ya que para el periodo 2002-2022 el 44% de las resoluciones que son leves se están clasificando como que no hay infracción y para el periodo 2002-2025 el 61% de las resoluciones leves se clasifican como no infracción.

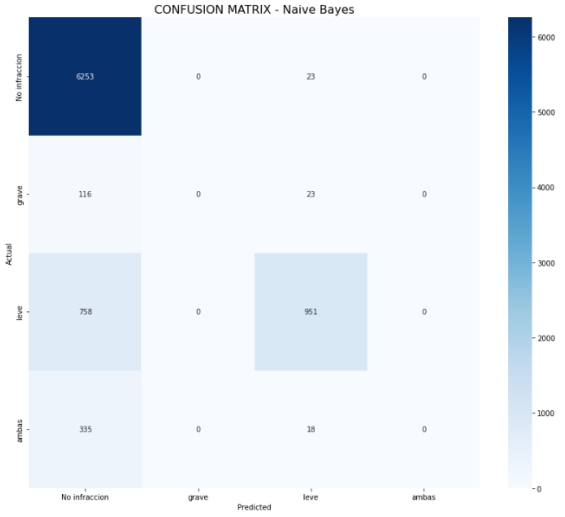


Ilustración 16.1. Matriz de Confusión para el modelo Naïve Bayes hasta 2022

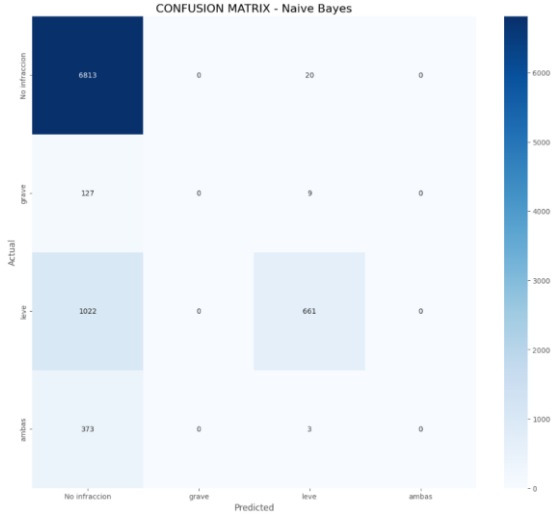


Ilustración 17.2. Matriz de Confusión para el modelo Naïve Bayes hasta 2025

4.2.2 CLASIFICADOR LINEAR SUPORT VECTOR MACHINE

Los resultados de este modelo son mejores que los del modelo anterior, teniendo una precisión del 92% en el conjunto de datos de entrenamiento y una precisión del 91’8% en el conjunto de datos de validación para el periodo 2002-2022. Y para el periodo 2002-2025, se

ha obtenido una precisión del 91.7% tanto para el conjunto de datos de entrenamiento como de validación.

En la Ilustración 13.1 y en la Ilustración 13.2 se puede ver la representación de la matriz de confusión para el conjunto de datos de validación. Se observa que la clasificación del grupo ambas, ha mejorado considerablemente, sin embargo, cuando la infracción es grave, el error del modelo sigue siendo del 100% de las observaciones, en los dos periodos temporales.

También se observa que, se ha mejorado la predicción de las resoluciones que están clasificadas con una infracción leve con respecto al modelo de clasifiación Naïve Bayes para los dos periodos temporales.

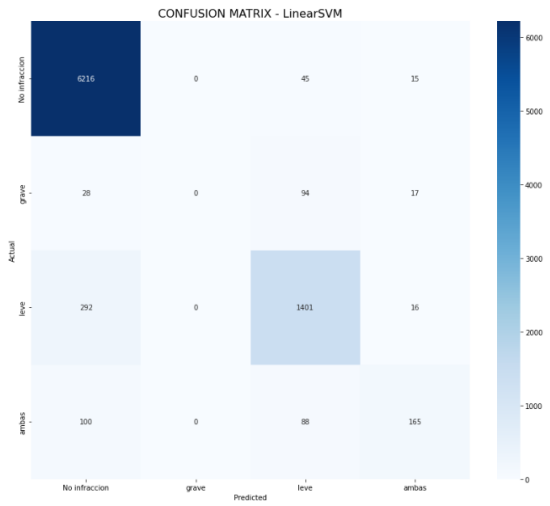


Ilustración 18.1. Matriz de Confusión para el modelo Linear Suport Vector Machine hasta 2022

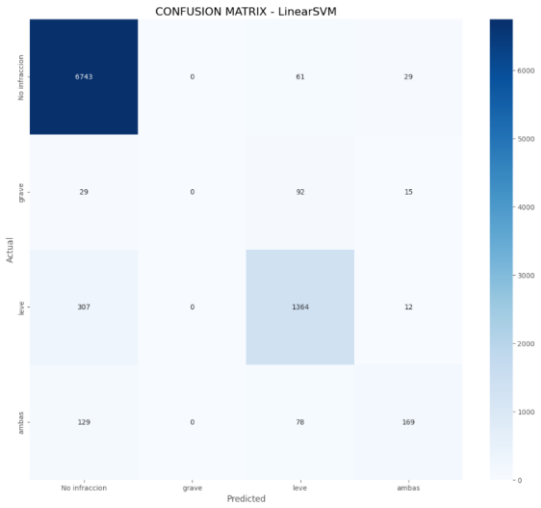


Ilustración 19.2. Matriz de Confusión para el modelo Linear Suport Vector Machine hasta 2025

4.2.3 REGRESIÓN LOGÍSTICA MULTINOMIAL

Los resultados obtenidos con este modelo son casi perfectos. Obteniéndose, para el periodo 2002-2022, una precisión del 100% en el conjunto de datos de entrenamiento y una precisión del 96’8% en el conjunto de datos de validación. Para el periodo 2002-2025, el resultado es prácticamente el mismo, se obtenido una precisión del 100% en el conjunto de datos de validación y una precisión del 96% en el conjunto de datos de validación.

Se puede observar en la matriz de confusión de la Ilustración 14.1 y 14.2 que las estimaciones son mucho mejores que para los modelos anteriores. La estimación de la clase grave sigue siendo problemática, pero recordamos que es la clase con menos observaciones, por tanto, este error mayor en la estimación tiene sentido. Vemos que para los dos periodos temporales obtenemos resultados muy similares, proporcionales al número de resoluciones que tienen alguna infracción.

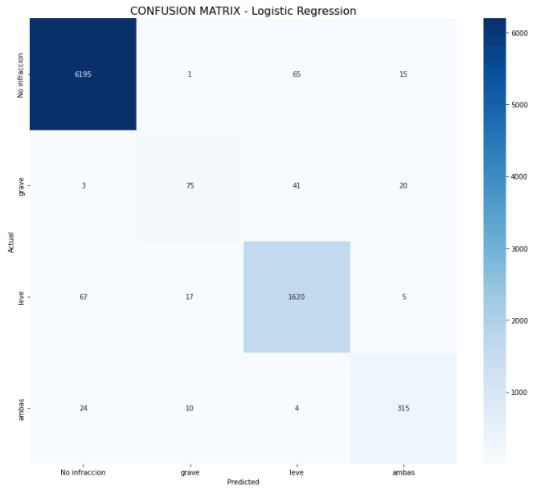


Ilustración 20.1. Matriz de Confusión para el modelo de Regresión Logística hasta 2022

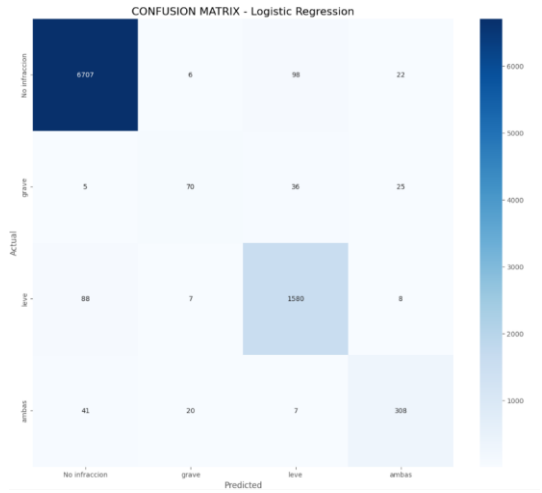


Ilustración 21.2. Matriz de Confusión para el modelo de Regresión Logística hasta 2025

Debido al nivel de precisión obtenido y habiendo comprobado que no hay sobre entrenamiento en el modelo, ya que la precisión del conjunto de validación es muy parecida a la precisión del conjunto de entrenamiento; se podría concluir la investigación con este modelo. Es un modelo relativamente sencillo, explicable, que requiere poco tiempo de procesamiento y recursos y con un resultado altamente positivo.

Sin embargo, se va a ajustar una red neuronal para estimar la salida, con la finalidad de valorar si un modelo más complejo podría mejorar los resultados obtenidos con el modelo de regresión logística multinomial.

4.2.4 REDES NEURONALES

Se ha optado por entrenar una red neuronal con dos capas ocultas, la primera de ellas con 10 neuronas de tipo "Dense", lo que significa que todas las neuronas de la capa oculta están conectadas con todas las neuronas de la capa anterior, en este caso la capa de entrada y con función de activación de tipo ReLu, cuya fórmula es:

$$f(x) = \max(0, x)$$

Donde x es el valor de entrada de la neurona, por tanto, es una función que devuelve el valor de entrada en caso de que este sea positivo o 0 en caso contrario.

La segunda capa oculta que se ha utilizado es una capa oculta de tipo "Dense" con 4 neuronas y función de activación softmax, la cuál es una generalización de la regresión logística y viene definida por la siguiente fórmula:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Es decir, la función softmax aplica la función exponencial estándar a cada elemento z_i del vector de entrada z y normaliza estos valores dividiendo por la suma de todos los exponenciales, esta normalización asegura que la suma de todas las componentes del vector de salida $\sigma(z)$ sume 1.

Se han realizado 30 epochs y el resultado obtenido, es muy similar al obtenido con la regresión logística. Obteniéndose, para el periodo 2002-2022 una precisión global del 99% para el conjunto de datos de entrenamiento y una precisión del 96% para el conjunto de datos de validación. Y para el periodo 2002-2025 una precisión del 99,6% para el conjunto de datos de entrenamiento y una precisión del 95,3% para el conjunto de datos de validación.

Se puede observar en la matriz de confusión de la Ilustración 15.1 y la Ilustración 15.2 que las estimaciones por clase son muy parecidas a las calculadas mediante el modelo de regresión logística. Sin embargo, la estimación de la clase grave, que en el apartado anterior se había mejorado, mediante el uso de este modelo ha vuelto a empeorar.

Igual que para el resto de modelos, se ve que, a la hora de clasificar las infracciones, el modelo se comporta de igual manera para los dos periodos temporales.

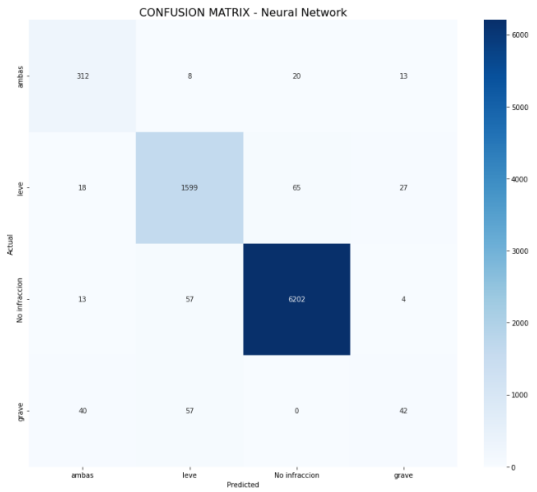


Ilustración 22.1. Matriz de Confusión para el modelo de Redes Neuronales hasta 2022

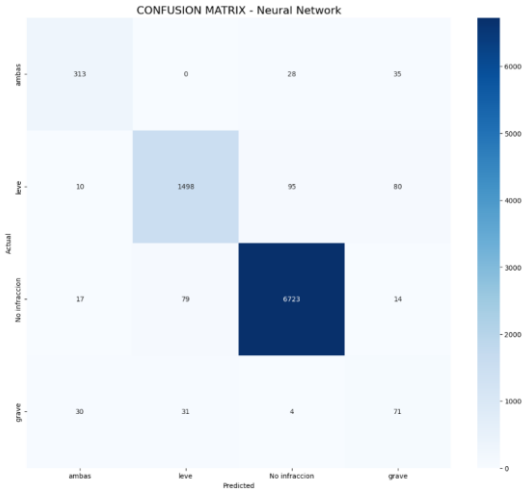


Ilustración 23.2. Matriz de Confusión para el modelo de Redes Neuronales hasta 2025

Por último, se ha representado la precisión del modelo y la función de pérdida para cada epoch, con la intención de saber si estos resultados podrían mejorarse añadiendo más iteraciones al modelo, los resultados obtenidos se encuentran en la Ilustración 16.1 y en la Ilustración 16.2.

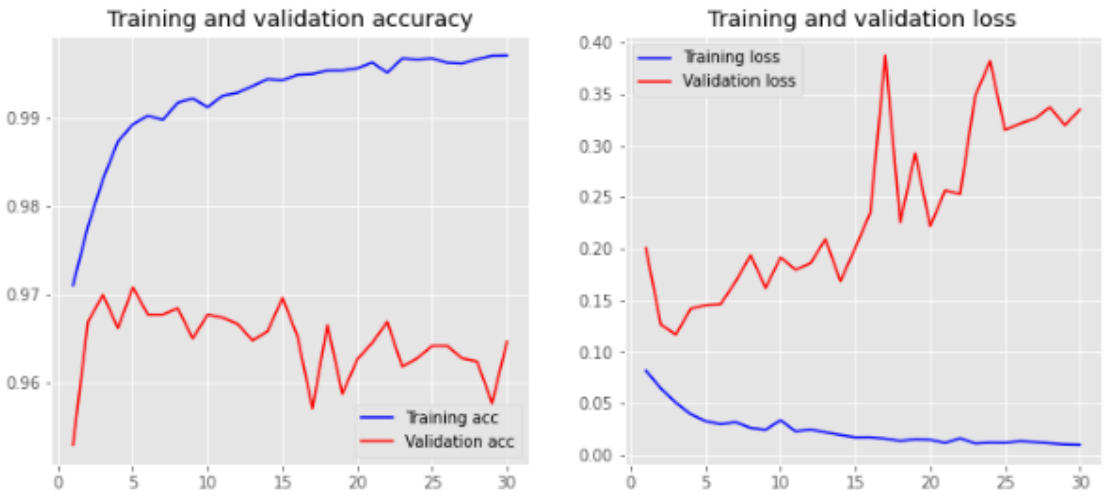


Ilustración 24.1. Precisión y función de pérdida del modelo para 30 epochs hasta 2022

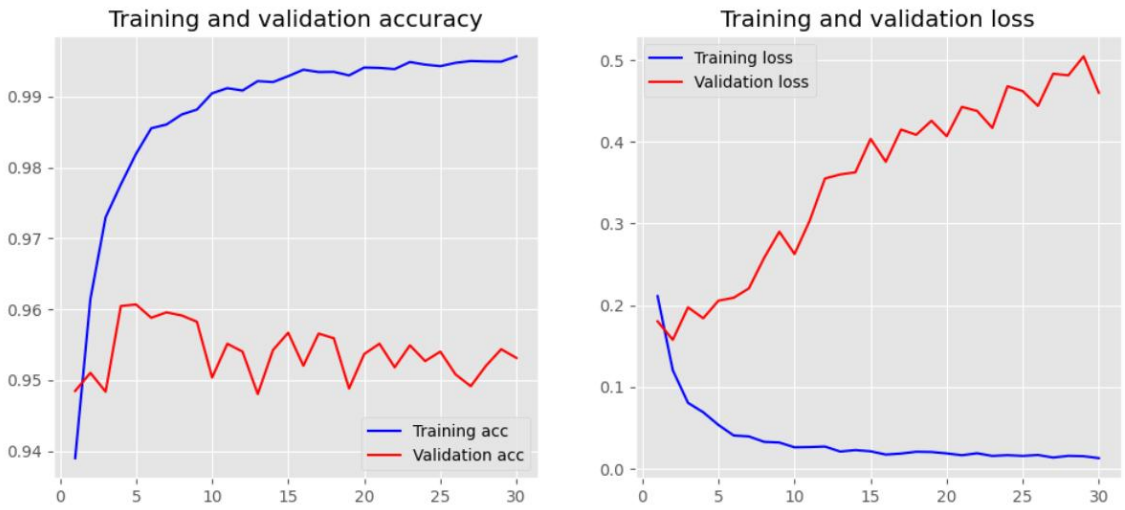


Ilustración 25.2. Precisión y función de pérdida del modelo para 30 epochs hasta 2025

En las gráficas se puede ver que ampliar el número de iteraciones no va a mejorar el modelo ya que se ve que la recta de entrenamiento y la recta de validación, tanto para la gráfica de precisión como para la de pérdida, se separan a medida que ampliamos el número de iteraciones. Esto quiere decir que el modelo está sobreentrenado, en las gráficas se ve que el número óptimo de epochs es entre 3 y 5, tanto para el periodo 2002-2022 como para el periodo 2002-2025.

Se vuelve a entrenar la red con 5 epochs y se obtiene un resultado muy similar. Para el periodo 2002-2022 se obtiene una precisión del 99% para el conjunto de entrenamiento y una precisión del 96% para el conjunto de validación. Para el periodo 2002-2025 se obtiene una precisión del 98,7% para el conjunto de datos de entrenamiento y una precisión del 96% para el conjunto de datos de validación.

4.2.5 COMPARACIÓN DE LOS MODELOS

En la tabla 1.1 y en la tabla 1.2 se ha recogido la precisión de cada modelo para cada clase y la precisión global, tanto para el conjunto de datos de entrenamiento como para el conjunto de datos de validación de los dos periodos temporales.

		Naïve Bayes	Linear SVM	LogReg	Red Neuronal 30 epochs	Red Neuronal 5 epochs
TRAIN	No Infracción	0.85	0.94	1	1	1
	Grave	0	0	1	0.88	0.89
	Leve	0.94	0.87	1	1	1
	Ambas	1	0.8	1	0.96	0.94
	Total	**0.857**	**0.92**	**1**	**0.99**	**0.99**
TEST	No Infracción	0.84	0.94	0.99	0.99	0.98
	Grave	0	0	0.73	0.49	0.42
	Leve	0.94	0.86	0.94	0.93	0.93
	Ambas	0	0.77	0.89	0.81	0.85
	Total	**0.85**	**0.918**	**0.968**	**0.96**	**0.96**

Tabla 1.1. Tabla resumen de la precisión de cada modelo utilizado para la clasificación del tipo de Infracción para el periodo 2002-2022

		Naïve Bayes	Linear SVM	LogReg	Red Neuronal 30 epochs	Red Neuronal 5 epochs
TRAIN	No Infracción	0.82	0.93	1	1	1
	Grave	0	0	1	0.81	0.73
	Leve	0.95	0.86	1	1	0.98
	Ambas	0.5	0.79	1	1	0.97
	Total	**0.831**	**0.917**	**1**	**0.996**	**0.99**
TEST	No Infracción	0.82	0.94	0.98	0.98	0.98
	Grave	0	0	0.68	0.35	0.52
	Leve	0.95	0.86	0.92	0.93	0.92
	Ambas	0	0.75	0.85	0.85	0.87
	Total	**0.828**	**0.917**	**0.96**	**0.953**	**0.96**

Tabla 2.2. Tabla resumen de la precisión de cada modelo utilizado para la clasificación del tipo de Infracción para el periodo 2002-2025

Teniendo en cuenta las precisiones recogidas en las tablas, el tiempo de entrenamiento y la complejidad de los modelos. El modelo óptimo para clasificar el tipo de infracción cometida es el modelo de regresión logística para los dos periodos temporales, ya que se observa que la red neuronal con 5 epochs tiene un resultado muy parecido a la regresión logística y el tiempo de entrenamiento es muy pequeño ya que son pocas iteraciones, sin embargo, la complejidad del modelo lo convierte en un modelo menos adecuado.

4.3 IMPORTE DE LA MULTA

Como se comentaba en el apartado 3.2.2.2. esta es una variable continua y en algunos casos hay más de una multa por resolución. Sin embargo, cuando se hace referencia a una posible multa, el rango de valores viene predefinido por el tipo de infracción cometida, siendo la más corriente una multa de 40.001€ a 300.000€. Además, hay muchas resoluciones que no tienen una multa asignada y por tanto el valor son 0€.

Debido a todo esto, se ha decidido hacer la media por resolución de todas las multas que aparecen en esta y agrupar estas multas en 4 clases, de forma que tenemos:

- Para el periodo 2002-2022:
  - No multa: Cuando la multa es de 0€ (73% de los datos)
  - Baja: Cuando la media de las multas está entre 1€ y 40.000€ (8.5% de los datos)
  - Media: Cuando la media de las multas está entre 40.001€ y 300.000€ (12.5% de los datos)
  - Alta: Cuando la media de las multas es superior a 300.000€ (6% de los datos)
- Para el periodo 2002-2025:
  - No multa: Cuando la multa es de 0€ (72% de los datos)
  - Baja: Cuando la media de las multas está entre 1€ y 40.000€ (10% de los datos)

- Media: Cuando la media de las multas está entre 40.001€ y 300.000€ (12% de los datos)
- Alta: Cuando la media de las multas es superior a 300.000€ (6% de los datos)

Al igual que para estimar el tipo de infracción, antes de ajustar los modelos de clasificación, se divide el conjunto de datos asignando un 80% de los datos al conjunto de datos de entrenamiento y un 20% al conjunto de datos de evaluación, manteniendo la distribución de las clases del conjunto de datos total, ya que igual que antes, estas están desbalanceadas.

Los clasificadores que se van a utilizar son los mismos que para clasificar el tipo de infracción cometida, pero en este apartado tenemos una complicación superior, ya que el texto del que partimos no tiene ningún número.

4.3.1 CLASIFICADOR NAÏVE BAYES

Mediante el clasificador de Naïve Bayes, para el periodo 2002-2022 se obtiene una precisión del 78% en el conjunto de datos de entrenamiento y una precisión del 77% en el conjunto de datos de validación. Para el periodo 2002-2025 se obtiene una precisión del 74% tanto para el conjunto de datos de entrenamiento como de validación.

Al igual que cuando se usaba el clasificador de Naïve Bayes para estimar el tipo de Infracción, con este modelo obtenemos que la precisión para la clase Alta, que es la minoritaria en este caso, es del 0% tanto para el conjunto de datos de entrenamiento como para el conjunto de datos de validación para los dos periodos.

4.3.2 CLASIFICADOR LINEAR SUPORT VECTOR MACHINE

Mediante la máquina de soporte vectorial se mejora la predicción anterior, obteniéndose, para el periodo 2002-2022 una precisión del 83% en el conjunto de datos de entrenamiento y una precisión del 82.5% en el conjunto de datos de validación. Para el periodo 2002-2025 se obtiene una precisión del 82% tanto en el conjunto de datos de entrenamiento como en el de validación.

La precisión del modelo por clase, en este apartado mejora considerablemente para las clases minoritarias, ya que todas las clases tienen una precisión superior al 61% tanto en el conjunto de datos de entrenamiento como en el conjunto de datos de validación para los dos periodos temporales.

4.3.3 REGRESIÓN LOGÍSTICA MULTINOMIAL

Al igual que para la estimación del tipo de infracción, la regresión logística multinomial es el modelo que mejores resultados da. Para el periodo 2002-2022 obtenemos una precisión del 97% en el conjunto de datos de entrenamiento y una precisión del 91% en el conjunto de datos de validación. Y para el periodo 2002-2025 obtenemos una precisión del 96.6% en el conjunto de datos de entrenamiento y una precisión del 89.7% en el conjunto de datos de validación.

Podríamos pensar que en este modelo hay un poco de sobre entrenamiento por la diferencia de la precisión para el conjunto de datos de entrenamiento y el conjunto de datos de validación. Si nos fijamos en las precisiones por clase, observamos que hay mucha diferencia entre la precisión de la clase Baja en el conjunto de datos de entrenamiento (91%) y en el conjunto de validación (68%) para el periodo 2002-2022. Existe la misma diferencia en el periodo 2002-2025, teniendo una precisión la clase multa Baja del 90% en el entrenamiento y del 67% en la validación.

4.3.4 REDES NEURONALES

Se ha ajustado la red neuronal con las mismas características que para la estimación del tipo de infracción, en este caso únicamente con 10 epochs. Pues habíamos visto anteriormente que el óptimo para el tipo de infracción estaba entre 3 y 5 epochs y tenemos unos datos parecidos para el importe de las multas, seguimos teniendo 4 clases, siendo una de ellas mayoritaria en cuanto a cantidad de observaciones.

En la Ilustración 17.1 y en la Ilustración 17.2 se puede ver la evolución de la precisión para cada epoch y cada periodo temporal.

Ilustración 26.1. Precisión y función de pérdida del modelo para 10 epochs hasta 2022

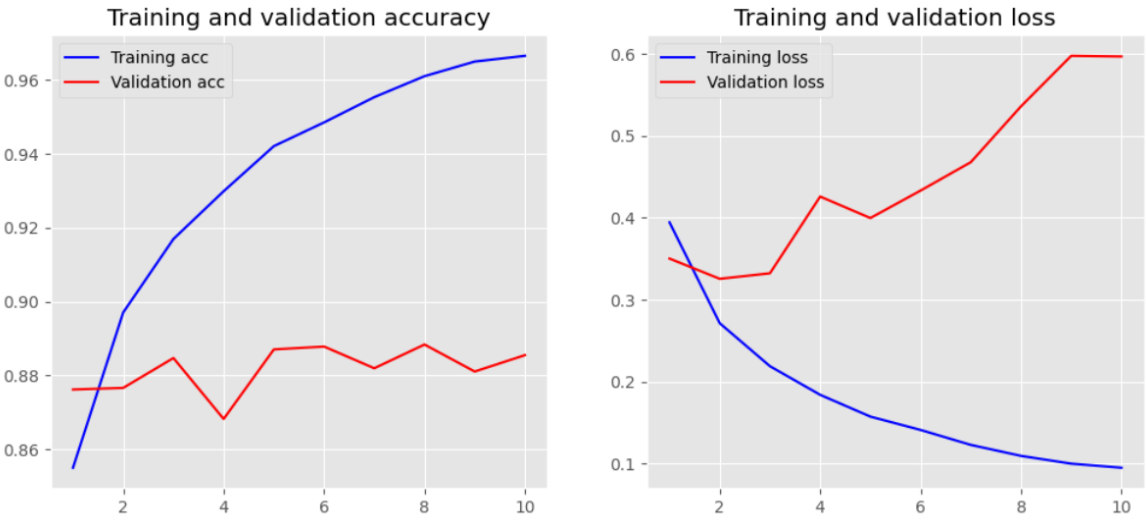


Ilustración 27.2. Precisión y función de pérdida del modelo para 10 epochs hasta 2022

Para el periodo 2002-2025, la precisión para el conjunto de datos de entrenamiento es del 99% y la precisión obtenida para el conjunto de datos de validación es del 90%. Si se analiza la precisión por clase, se obtiene que tanto la clase multa Baja como la clase multa Alta, pasa de tener un 97% de precisión en el conjunto de datos de entrenamiento a un 64% de precisión en el conjunto de datos de validación.

Para el periodo 2002-2025, la precisión para el conjunto de datos de entrenamiento es del 98% y la precisión obtenida para el conjunto de datos de validación es del 88%. Si se analiza la precisión por clase, se obtiene que la clase multa Baja, pasa de tener un 93% de precisión

en el conjunto de datos de entrenamiento a un 64% de precisión en el conjunto de datos de validación y la clase multa Alta pasa de un 88% de precisión en el conjunto de datos de entrenamiento a un 65% en el conjunto de datos de validación. Las otras dos clases también pierden precisión en el conjunto de datos de validación, pero la diferencia no es tan grande.

4.3.5 COMPARACIÓN DE LOS MODELOS

En la tabla 2.1 y en la tabla 2.2 se recoge la información de la precisión en los conjuntos de datos de entrenamiento y validación para cada clase de los modelos realizados en los apartados anteriores para los dos periodos.

		Naïve Bayes	**Linear SVM**	**LogReg**	**Red Neuronal 10 epochs**
TRAIN	No Multa	0.79	0.86	0.99	1
	Baja	1	0.75	0.91	0.97
	Media	0.61	0.64	0.92	0.95
	Alta	0	0.65	0.87	0.97
	Total	**0.78**	**0.83**	**0.97**	**0.99**
TEST	No Multa	0.79	0.86	0.96	1
	Baja	1	0.7	0.68	0.64
	Media	0.59	0.63	0.81	0.95
	Alta	0	0.61	0.67	0.64
	Total	**0.77**	**0.825**	**0.91**	**0.9**

Tabla 3.1. Tabla resumen de la precisión de cada modelo utilizado para la clasificación del importe de la Multa para el periodo 2002-2022

		Naïve Bayes	**Linear SVM**	**LogReg**	**Red Neuronal 10 epochs**
TRAIN	No Multa	0.75	0.85	0.99	1
	Baja	0.74	0.74	0.90	0.93
	Media	0.60	0.65	0.93	0.97
	Alta	0	0.65	0.87	0.88
	Total	**0.74**	**0.82**	**0.966**	**0.978**
TEST	No Multa	0.75	0.85	0.96	0.96
	Baja	0.71	0.66	0.67	0.64
	Media	0.61	0.63	0.79	0.76
	Alta	0	0.61	0.63	0.65
	Total	**0.74**	**0.818**	**0.897**	**0.886**

Tabla 4.2. Tabla resumen de la precisión de cada modelo utilizado para la clasificación del importe de la Multa para el periodo 2002-2025

Al igual que para la clasificación del tipo de infracción, los resultados obtenidos con la red neuronal y los resultados obtenidos con la regresión logística son muy similares. Sin embargo, teniendo en cuenta la complejidad de los modelos, el modelo óptimo para clasificar el importe de la multa es el modelo de regresión logística multinomial en los dos periodos.

Capítulo 5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1 CONCLUSIONES

Se han obtenido los datos de la página Web de la AEPD de forma satisfactoria y se han convertido a un formato de tabla de datos. A partir de estos se han extraído las principales características de interés del texto, tipo de infracción cometida e importe de la multa.

Se han ajustado varios modelos de clasificación, tanto para el tipo de infracción como para el importe de la multa y se ha observado que para ambas características el mejor modelo es el modelo de regresión logística multinomial.

El objetivo de clasificar el tipo de infracción se ha conseguido exitosamente obteniéndose una precisión de aproximadamente un 96%. El objetivo de estimar el importe de la multa se ha transformado de un problema de regresión a un problema de clasificación por la naturaleza de los datos y se ha clasificado el importe de la multa en 4 grupos obteniendo una precisión del 90%

Se ha observado que los modelos tienen más errores a la hora de predecir las clases que tienen menos observaciones con respecto al resto en ambos modelos, “grave” y “ambas” para el tipo de infracción y “alta” y “baja” para el importe de la multa.

Se ha comprobado que, tanto para el tipo de infracción como para el importe de la multa, más del 70% de las resoluciones publicadas en la AEPD recogen que no se ha cometido una infracción y que por tanto no se resuelve mediante una multa. Esto es un dato alentador, ya que uno de los motivos por los que se ha decidido hacer este trabajo era para evaluar la gravedad y el incremento del incumplimiento de la ley en términos de protección de datos.

Se ha comprobado que al añadir los últimos tres años al global de las resoluciones, no se ve diferencias en los resultados de forma general. Lo cual también es alentador, pues significa que a pesar del constante desarrollo de las teconologías, las resoluciones que se publican en

la AEPD no han sufrido un cambio drástico con respecto a las que se publicaban unos años atrás.

5.2 TRABAJOS FUTUROS

En cuanto a mejoras sobre lo ya realizado, se podrían realizar las clasificaciones en dos pasos, de forma que se haga una primera clasificación binaria cuya salida sea infracción y no infracción para el caso del tipo de infracción y multa o no multa para el caso del importe de la multa. Una vez con los datos separados, se podría volver a aplicar un modelo de clasificación para el resto de las clases. Esto duplica el número de modelos estimados, sin embargo, podría mejorar la precisión de los modelos para la estimación de las clases minoritarias, que a efectos de concienciación sobre la infracción de la LOPD son las clases más relevantes.

En cuanto a líneas futuras, se parte de una base de datos con mucha versatilidad, por tanto, se pueden plantear muchos objetivos y mejoras sobre el trabajo ya realizado. En caso de que se quiera poner en producción para estimar el importe de una posible multa en función de un texto que recoja los hechos, sería de gran utilidad crear un programa para descargar automáticamente las resoluciones que se publican nuevas en la AEPD cada cierto tiempo, el establecido oportuno por el desarrollador y añadir estas resoluciones a los datos ya disponibles para volver a ejecutar los modelos y ver si hay cambios. Ya que, con el tiempo, las leyes pueden cambiar y el modelo actual podría quedarse obsoleto.

Se podría utilizar el conjunto de datos para otra finalidad, como, por ejemplo, para hacer una investigación de las empresas que han realizado alguna infracción de datos personales, para saber si se recogen más resoluciones denunciando a una empresa o a un particular, o para saber qué tipo de resoluciones son las resoluciones que se recurren.

Capítulo 6. BIBLIOGRAFÍA

[1] AMAT RODRIGO, JOAQUÍN (2017) Máquinas de Vector Soporte (Support Vector Machines, SVMs)

[2] BANERJEE, JOYDEEP (2020) Neural Networks – the Rudiments and the Mathematics

[3] BROWNLEE, JASON (2022) Evaluate the Performance of Deep Learning Models in Keras

[4] CHOUBEY, VIJAY (2021) Multiclass Text Classification Using Deep Learning

[5] FLETCHER, TRISTAN (2008) Support Vector Machines Explained

[6] HARDESTY, LARRY (2017) Explained: Neural Networks

[7] JACOVI, ALON & SAR SHALOM, OREN & GOLDBERG, YOAV (2018) Understanding Convolutional Neural Networks for Text Classification

[8] JADRAQUE DE SORIA, DANIEL. Apuntes de la asignatura de Adquisición y Transformación de Datos.

[9] JANAKIEV, NIKOLAI (2018) Practical Text Classification with Python and Keras

[10] KONONENKO, IGOR. Semi Naïve Bayesian Classifier

[11] LARRAÑAGA, PEDRO & INZA, IÑAKI & MOUJAHID, ABDELMALIK. Clasificadores Bayesianos

[12] LARRAÑAGA, PEDRO & INZA, IÑAKI & MOUJAHID, ABDELMALIK. Regresión Logística

[13] LI, SUSAN (2018) Multi-Class Text Classification Model Comparison and Selection

[14] PANDO FERNÁNDEZ, V & SAN MARTÍN FERNÁNDEZ, R (2004) Regresión Logística Multinomial

[15] PERAMBAI, ABHISHEK (2020) Theory behind Word Embeddings in Word2vec

[16] PLATT, JOHN C. & DUMAIS, SUSAN & HECKERMAN, DAVID & SAHAMI, MEHRAN (2014) Inductive Learning Algorithms and Representations for Text Categorization

[17] PORTELA GONZÁLEZ, JOSÉ. Apuntes de la asignatura de Machine Learning I.

[18] SÁNCHEZ ÚBEDA, EUGENIO. Apuntes de la asignatura de Machine Learning II.

[19] SHARMA ABHINAV (2020) Easy Web Scraping using Python and BeautifulSoup4 and saving files as well HTML pages as PDF

[20] Requests User Guide Python. https://requests.readthedocs.io/en/latest/

[21] Web de la Agencia Epañola de Protección de Datos https://www.aepd.es/es