Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle

Trabajo Final de Máster

Máster Universitario en Ingeniería de Telecomunicaciones

Técnicas de Data Science para la caracterización del rendimiento de las noticias online

Alumno

Guillermo Brugarolas Sobejano

Profesor Ponente

Dr. Xavier Vilasís Cardona

ACTA DEL EXAMEN DEL TRABAJO FINAL DE MASTER

Reunit el Tribunal qualificador en el dia de la data, l'alumne

D. Guillermo Brugarolas Sobejano

va exposar el seu Treball de Fi de Màster, el qual va tractar sobre el tema següent:

Técnicas de Data Science para la caracterización del rendimiento de las noticias online

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL

Abstract

More and more, readers are turning to digital or online media for information. These have largely replaced the traditional press. Those responsible for managing and editing online newspapers, blogs and other content outlets are aware of the importance of maintaining their regular base readership. In recent years, different web data analytics and machine learning techniques have been used in these businesses to obtain useful information from the vast amount of available data on their article and news pages. This information can be in the form of pattern identification, characterization of the news articles according to different parameters, or the ability to predict their performance. This information is useful if it is used as a support tool the making of editorial and organizational decisions by the online media management staff.

This project is a broad and relatively high-level study on different predictive modelling and machine learning techniques that can be applied on a dataset that contains raw data about the pages of an online radio and news portal, with the aim of obtaining that previously mentioned useful information. The dataset has been obtained from a web analytics tool that periodically collects the value of different attributes of each article or news page of the web portal.

The first practical part of this project consists of the application of statistical modelling and clustering techniques on the data set. In the first place, different statistical models are applied, widely used in the theory of diffusion of innovations and of adoption of novelties by society, to characterize the curve of visits to news articles. Secondly, some clustering algorithms, which are unsupervised machine learning techniques, are employed to find different ways to categorize and classify the online news articles based on their performance parameters and initial characteristics.

The second practical part of the project consists of applying the different capabilities of the well-known XGBoost library to predict the performance parameters of online news articles based on their initial characteristics.

This project approximates the different possibilities that data science has applied to human behaviour, such as the visit of a user to a news page in a digital medium. The ability to extract useful information from it will depend on the techniques applied and on the data set.

Keywords

Data science, web analytics, metrics, statistical modelling, predictive analytics, online news articles, clustering algorithms, machine learning, regression.

Resumen

Cada vez más, los lectores acceden a medios digitales o en línea para informarse. Estos han ido sustituyendo en gran medida a la prensa tradicional. Los responsables de la dirección y edición de diarios, blogs y demás portales online son conscientes de la importancia de mantener su base de lectores más o menos regulares. En los últimos años, se vienen utilizando en estos negocios diferentes técnicas de analítica de datos web y de aprendizaje automático para obtener información útil a partir de la ingente cantidad de datos almacenados sobre sus páginas de artículos y noticias. Esta información puede ser en forma de identificación de patrones, caracterización de los artículos según diferentes parámetros, o de capacidad de predicción del rendimiento de estos. Esta información es útil si se utiliza como herramienta de soporte en la toma de decisiones editoriales y organizativas por parte de la dirección del medio o diario en línea.

Este proyecto es un estudio amplio y de relativo alto nivel sobre diferentes técnicas de modelado predictivo y de aprendizaje automático que se pueden aplicar sobre los datos de las páginas de un portal de radio y noticias en línea, con el objetivo de obtener esa información útil anteriormente mencionada. El conjunto de datos se ha obtenido de una herramienta de analítica web que recoge periódicamente el valor de diferentes atributos de cada página de artículo o noticia del portal web.

La primera parte práctica de este proyecto consiste en la aplicación de técnicas de modelado estadístico y de agrupamiento sobre el conjunto de datos. En primer lugar, se aplican diferentes modelos estadísticos, muy utilizados en teoría de difusión de innovaciones y de adopción de novedades por la sociedad, para caracterizar la curva de visitas a los artículos/noticias web. En segundo lugar, se emplean algoritmos de agrupamiento, que es una técnica de aprendizaje automático no supervisado, para encontrar diferentes maneras de categorizar y clasificar los artículos/noticias web, en función de sus parámetros de rendimiento y de sus características iniciales.

La segunda parte práctica del proyecto consiste en la aplicación de las diferentes capacidades de la reconocida librería *XGBoost* para predecir los parámetros de rendimiento de los artículos/noticias web en función de sus características iniciales.

Este proyecto es una aproximación de las diferentes posibilidades que tiene la ciencia de datos aplicada al comportamiento humano, como es la visita de un usuario a la página de una noticia de un medio digital. La capacidad de extraer información útil de ello dependerá de las técnicas aplicadas y también del conjunto de datos.

Palabras clave

Ciencia de datos, analítica web, métricas, modelado estadístico, análisis predictivo, noticias en línea, algoritmos de agrupamiento, aprendizaje automático, regresión.

Resum

Cada vegada més, els lectors accedeixen a mitjans digitals o en línia per a informar-se. Aquests han anat substituint en gran mesura a la premsa tradicional. Els responsables de la direcció i edició de diaris, blogs i altres portals online són conscients de la importància de mantenir la seva base de lectors més o menys regulars. En els últims anys, es venen utilitzant en aquests negocis diferents tècniques d'analítica de dades web i d'aprenentatge automàtic per obtenir informació útil a partir de la ingent quantitat de dades emmagatzemades sobre les seves pàgines d'articles i notícies. Aquesta informació pot ser en forma d'identificació de patrons, caracterització dels articles segons diferents paràmetres, o de capacitat de predicció de el rendiment d'aquests. Aquesta informació és útil si s'utilitza com a eina de suport en la presa de decisions editorials i organitzatives per part de la direcció del mitjà o diari en línia.

Aquest projecte és un estudi ampli i de relatiu alt nivell sobre diferents tècniques de modelatge predictiu i d'aprenentatge automàtic que es poden aplicar sobre les dades de les pàgines d'un portal de ràdio i notícies en línia, amb l'objectiu d'obtenir aquesta informació útil anteriorment esmentada. El conjunt de dades s'ha obtingut d'una eina d'analítica web que recull periòdicament el valor de diferents atributs de cada pàgina d'article o notícia del portal web.

La primera part pràctica d'aquest projecte consisteix en l'aplicació de tècniques de modelatge estadístic i d'agrupament sobre el conjunt de dades. En primer lloc, s'apliquen diferents models estadístics, molt utilitzats en teoria de difusió d'innovacions i d'adopció de novetats per la societat, per caracteritzar la corba de visites als articles/notícies web. En segon lloc, s'utilitzen algoritmes d'agrupament, que és una tècnica d'aprenentatge automàtic no supervisat, per a trobar diferents maneres de categoritzar i classificar els articles/notícies web, en funció dels seus paràmetres de rendiment i de les seves característiques inicials.

La segona part pràctica de el projecte consisteix en l'aplicació de les diferents capacitats de la reconeguda llibreria *XGBoost* per a predir els paràmetres de rendiment dels articles/notícies web en funció de les seves característiques inicials.

Aquest projecte és una aproximació de les diferents possibilitats que té la ciència de dades aplicada al comportament humà, com és la visita d'un usuari a la pàgina d'una notícia d'un mitjà digital. La capacitat d'extreure informació útil d'això dependrà de les tècniques aplicades i també del conjunt de dades.

Paraules clau

Ciència de dades, analítica web, mètriques, modelat estadístic, anàlisi predictiu, articles/notícies online, algorismes d'agrupament, aprenentatge automàtic, regressió.

Agradecimientos

Es un inmenso placer agradecer la asistencia y los comentarios de mi director de tesis, el Dr. Xavier Vilasís Cardona, quien estuvo allí para ayudarme en todo lo posible. Muchas gracias especiales para él por brindarme esta oportunidad a pesar de mi falta de experiencia en el campo, sus comentarios y orientación sobre la organización de la planificación de mi investigación y la supervisión de mi trabajo fueron de importancia crítica.

También me gustaría expresar mi gratitud a muchos grupos de personas que me han ayudado personal y emocionalmente durante la duración de mi proyecto. Quiero hacer una mención especial a la Srta. Jessie Martín Sujo, porque me asistió técnica y personalmente de la mejor manera en varios momentos durante el desarrollo de este proyecto.

Para mi familia, mi madre, mi padre, mi hermano y mi hermana. Sin ellos no podría haber terminado mis estudios.

A mis amigos, por escucharme, apoyarme y animarme a seguir adelante. Gracias por estar aquí.

A los profesores y compañeros de La Salle - Universitat Ramon Llull, que me han acompañado y enseñado durante mi estancia en la universidad.

Índice

Contenidos

1 Introducción

En este capítulo, se introduce y se describe el proyecto, se explican los antecedentes, y se detallan el alcance, los objetivos y la metodología del proyecto.

1.1 Descripción

En la era digital actual, una disciplina que ha ido tomando cada vez más relevancia es la inteligencia artificial (IA), y, más específicamente, el aprendizaje automático. La multitud de aplicaciones en la vida real que se le han encontrado, y que presumiblemente se le seguirán encontrando, la convierten en una disciplina muy atractiva para todo tipo de negocios. Entre ellos están los medios digitales, diarios en línea, etc., que pueden disponer de una ingente cantidad de datos acerca de sus páginas y artículos, pero pueden no saber cómo extraer información útil de ellos. Para extraer esta información, se pueden aplicar técnicas de aprendizaje automático y análisis predictivo. Esta información puede ser utilizada por la dirección del medio o diario en línea como herramienta de soporte en la toma de decisiones editoriales y organizativas, como, p.ej., la promoción de determinados artículos/noticias o la cantidad y el precio de la publicidad a contratar para las páginas de estos.

Los datos brutos de las páginas de artículos/noticias del portal en línea normalmente se obtienen utilizando herramientas de analítica web. La mayoría de ellas son productos software de pago, pero también las hay gratuitas. Estas herramientas de software son capaces de extraer y almacenar de manera periódica el valor de diferentes parámetros o atributos de un artículo/noticia en línea, tales como visitas a página (clicks), tiempo medio en página o tasa de salida. El conjunto de datos con el que se ha trabajado en este proyecto se ha obtenido de una de estas herramientas de analítica web, y contiene los datos brutos de artículos/noticias del portal en línea de una radio bastante popular. Más adelante, se explica más en detalle el contenido de este conjunto de datos.

Cuando ya se dispone del conjunto de datos que se quiere analizar, es cuando aparecen infinidad de opciones para realizar el análisis. Se puede hacer una caracterización del comportamiento de un determinado atributo, para comprender mejor su evolución. También se puede hacer una categorización o agrupamiento de los artículos/noticias en base a sus parámetros de rendimiento, o en base a sus características iniciales, para encontrar patrones de comportamiento no evidentes y relaciones entre los atributos (o características iniciales) y los parámetros de rendimiento no obvias. Como ya se ha comentado, existen muchas técnicas de IA y de aprendizaje automático diferentes, cada una con su coste y su ámbito de aplicación. En este proyecto de demostración, se explican y se aplican algunas de ellas, como las mencionadas en estas líneas.

1.2 Antecedentes

En los últimos diez a veinte años, coincidiendo con el uso masivo de Internet a nivel mundial, se han realizado muchas investigaciones sobre el modelado estadístico de diferentes conductas y decisiones humanas, como la visita a un sitio o página web, la adquisición de un producto, la adopción de una nueva tecnología, y muchas otras. Muchos de estos comportamientos se ven afectados por la estacionalidad, las sensaciones personales, el factor de conexión preferencial y otros factores aleatorios, y esto hace que analizarlos, predecirlos y modelarlos sea un reto para los científicos.

El campo de la analítica web para medios o diarios en línea lleva tiempo activo, con multitud de publicaciones e investigaciones a su respecto. La mayoría de los grandes diarios y portales de noticias en línea ya disponen de un equipo o departamento de analítica de datos en tiempo real. Es un campo que sigue creciendo cada año y es previsible que siga creciendo en un futuro próximo.

A nivel académico, este proyecto se concibió como una continuación o ampliación del **Trabajo Final de Grado** del autor, con título "*Data Science approach to the pageviews of online news articles*". Este proyecto se centraba en el modelado de la evolución del parámetro de rendimiento de visitas a página de los artículos/noticias en línea. Se presentaban y aplicaban una serie de modelos estadísticos muy populares en teoría de difusión de innovaciones y adopción de nuevas tecnologías. Se hacía una comparativa de los modelos en base a los resultados obtenidos con cada uno, y se seleccionaba uno de ellos para realizar pruebas de agrupamiento y de predicción.

En este proyecto también se han utilizado estos modelos estadísticos para hacer otro tipo de agrupamientos. Pero, además de ello, se ha ampliado el número de atributos analizados, y adicionalmente se han aplicado diferentes técnicas de IA y de aprendizaje automático.

1.3 Alcance

Ilustración 1. Proyecto SMARTDATA. Fuente: Grupo DS4DS de La Salle – URL.

Este proyecto se puede considerar una investigación/exploración independiente dentro del marco de un proyecto más amplio llamado ***SMARTDATA*** [1] , en el que colabora el grupo de investigación **DS4DS** de **La Salle – Universitat Ramon Llull**. Este equipo está centrado en proyectos de *Data Science* y *Analytics*. El proyecto *SMARTDATA* se realiza/ha realizado por un consorcio con el apoyo del Ministerio de Economía y Competitividad (ver Ilustración 2). Es un proyecto muy amplio y tiene varias ramificaciones, pero se centra en la aplicación de técnicas de análisis de datos y aprendizaje automático a los datos recopilados del sitio web de un medio/portal en línea.

Ilustración 2. Grupos, instituciones y organismos participantes en el proyecto SMARTDATA.

El alcance de este proyecto se resume en presentar y aplicar diferentes técnicas de análisis y aprendizaje automático sobre los datos de un portal de radio y noticias en línea, explicar los resultados y compararlos. Se incluye en el alcance de este proyecto la presentación y la aplicación de técnicas de preprocesado de datos, ampliación de parámetros analizables y limitación del conjunto de datos utilizable. Por último, se incluye la introducción de técnicas de predicción por regresión, y un ejemplo de la aplicación de una de estas técnicas sobre el conjunto de datos.

El alcance de este proyecto está limitado en primer lugar por las restricciones temporales, ya que la duración de este estaba indirectamente marcada por los plazos de entrega fijados por la universidad. En segundo lugar, el alcance de este proyecto está limitado por los datos disponibles, ya que el número de filas y columnas que contiene el conjunto de datos disponible es limitado.

Otros factores coyunturales que se considera que no hace falta especificar también impiden un mayor alcance de este proyecto de exploración.

1.4 Objetivos

Los objetivos de este proyecto son varios, y justifican el desarrollo de este. Se dividen en objetivos personales y objetivos académicos/técnicos.

1.4.1 Objetivos personales

A nivel personal, este proyecto me permite adquirir una madurez y experiencia en el sector de la ciencia de datos, especialmente en el campo del aprendizaje automático,

ya que esta es una verdadera investigación sobre los datos obtenidos de un portal de radio y noticias en línea, que se puede implementar en un futuro cercano.

Un objetivo personal más concreto es el de estudiar aprender sobre las técnicas de predicción y clasificación más populares y eficaces, para poder aplicarlas con solidez y coherencia en un caso real, aunque después los resultados no sean satisfactorios.

1.4.2 Objetivos académicos/técnicos

A nivel técnico, el primer objetivo es ampliar los conocimientos obtenidos en el anterior proyecto, el Trabajo Final de Grado, descrito anteriormente.

El objetivo principal a nivel académico es el de explorar diferentes técnicas de análisis de datos y de aprendizaje automático para obtener información útil de los datos brutos obtenidos.

Otro objetivo es caracterizar las curvas de visitas a página con modelos estadísticos, comparar los resultados y llegar a una conclusión sobre qué tipos de artículos/noticias siguen qué modelo en términos de visitas a página.

- ➔ Hallar nuevas maneras de clasificar y categorizar los objetos obtenidos, descubrir patrones ocultos y encontrar relaciones entre parámetros de rendimiento y atributos base.

Categorizar los objetos de datos recibidos de diferentes maneras, aplicando diferentes algoritmos y técnicas, comparar sus resultados, y llegar a una conclusión sobre su utilidad y aplicabilidad.

Explorar sobre las técnicas de aprendizaje automático supervisado para tareas de predicción y de clasificación, presentarlas, compararlas, seleccionar una y aplicarla sobre el conjunto de datos. Evaluar los resultados obtenidos y llegar a una conclusión sobre su utilidad y aplicabilidad.

- ➔ Ser capaz de predecir con un mínimo de precisión el rendimiento final de un objeto a partir de sus características o atributos base.

Identificar, a lo largo del desarrollo del proyecto, oportunidades de mejora en la aplicación de las diferentes técnicas.

A nivel general, el objetivo de fondo que tiene el desarrollo de este proyecto es:

- ➔ Demostrar cómo se podrían aplicar algunas técnicas de análisis de datos y de aprendizaje automático para extraer de un conjunto de datos brutos información útil que, en última instancia, aporte valor real tanto al negocio como al cliente o visitante.

1.5 Metodología

Para implementar el proyecto se siguieron varios procesos.

El proyecto se desarrolló íntegramente en el lenguaje de programación **Python** y el software utilizado para la programación fue el Jupyter Notebook, un notable entorno computacional basado en web, que permite disponer de *notebooks* donde se puede dividir el código por celdas y ejecutarlas por separado, así como añadir celdas de notas en cualquier punto.

Python es el lenguaje elegido para desarrollar proyectos sobre ciencia de datos, estadísticas y aprendizaje automático. En estos campos, existe una gran cantidad de paquetes y librerías para Python que contienen funciones y métodos muy útiles. Algunos paquetes y librerías que se han utilizado a lo largo del desarrollo de este proyecto son:

- NumPy
- SciPy
- pandas
- matplotlib
- Scikit-learn
- XGBoost

El trabajo de codificación se dividió en varios archivos de Python, cada uno de los cuales representaba un **cuaderno** en el ecosistema de **Jupyter**. Se ha trabajado con una estructura de varios directorios, cada uno dedicado a albergar los cuadernos relativos a cada una de las tres partes principales del proyecto en términos de programación:

A. Importación del *dataset*, pre-procesado y post-procesado (visualización)

B. Caracterización (modelado estadístico) y Categorización (algoritmos de agrupamiento)

C. Predicción de parámetros de rendimiento (*boosting* de gradiente para regresión)

2 Teoría y conceptos

A lo largo de este capítulo, se presenta y explica la teoría básica necesaria para comprender y seguir el desarrollo de este proyecto. Además, se explican varios conceptos más específicos que aparecen de forma relevante y recurrente en este proyecto.

2.1 Ciencia de Datos

El primer y más básico concepto es el de Ciencia de Datos, ya que este proyecto es fundamentalmente un proyecto de investigación de Ciencia de Datos.

En el sentido más amplio, la ciencia de datos es un campo multidisciplinar que engloba diferentes métodos, procesos, algoritmos y sistemas científicos para extraer conocimiento e información de datos estructurados y no estructurados [6] [7] . Uno de los principales objetivos de un científico de datos podría ser utilizar el hardware más potente, los sistemas de programación más potentes y los algoritmos más eficientes para resolver problemas de todo tipo: optimización, predicción, clasificación, etc.

La Ciencia de Datos es un concepto creado para unificar la estadística, el análisis de datos, el aprendizaje automático y otros campos/técnicas utilizados para comprender y analizar fenómenos reales con datos. En la Ciencia de Datos se utilizan diversas técnicas extraídas de muchos campos dentro de los contextos de las matemáticas, la estadística, la informática y la ciencia de la información.

Es un concepto tan amplio y difícil de definir que muchos creen que no existe per se, que es una palabra de moda o *buzzword* que se ha vuelto demasiado popular en la última década. Pero, para la mayoría de los científicos que trabajan con datos, es un campo muy real que abarca muchas ramas, campos, ciencias y métodos, y todos tienen algo en común: la necesidad de interpretar enormes cantidades de datos.

Ilustración 3. Ciencia de Datos. Fuente: https://online.stanford.edu/programs/foundations-data-science

Una de las principales razones de la existencia de la ciencia de datos es que permite a las empresas dar sentido a todos los datos que pueden recopilar de sus operaciones, actividades, usuarios y clientes. Continuará creciendo en importancia a medida que más y más empresas se conecten a Internet y adopten la transformación digital.

El ciclo de vida típico y clásico de un proceso de Ciencia de Datos se muestra en la Ilustración 4. Ciclo de vida típico de un proceso de Ciencia de Datos. Fuente: https://datascience.berkeley.edu/about/what-is-data-science/ a continuación. La imagen representa las cinco etapas del ciclo de vida de la ciencia de datos:

1. Capturar (adquisición de datos, entrada de datos, recepción de señales, extracción de datos);
2. Mantener (almacenamiento de datos, limpieza de datos, almacenamiento de datos, procesamiento de datos, arquitectura de datos);
3. Procesar (minería de datos, agrupación/clasificación, modelado de datos, resumen de datos);
4. Analizar (exploratorio/confirmatorio, análisis predictivo, regresión, minería de texto, análisis cualitativo);
5. Comunicar (informes de datos, visualización de datos, inteligencia empresarial, toma de decisiones).

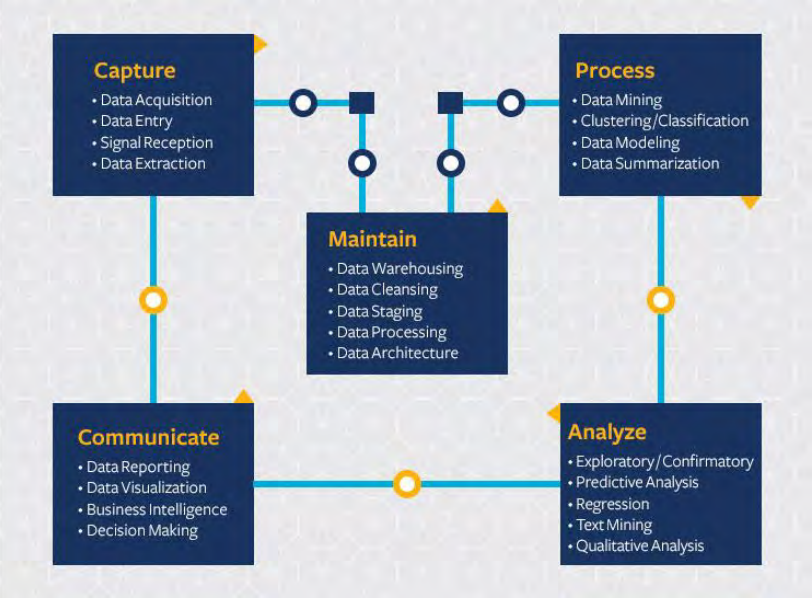

Ilustración 4. Ciclo de vida típico de un proceso de Ciencia de Datos. Fuente: https://datascience.berkeley.edu/about/what-is-data-science/

La ciencia de datos se usa principal y frecuentemente para:

- Análisis causal predictivo
- Análisis prescriptivo
- Aprendizaje automático para realizar predicciones

- Aprendizaje automático para el descubrimiento de patrones

2.2 Analítica web

El conjunto de datos utilizado en este proyecto contiene los datos por hora de una gran cantidad de artículos/noticias de un portal de radio y noticias en línea. El conjunto de datos se obtuvo de un servicio de analítica de sitios web, por lo que es razonable explicar qué es esto exactamente.

Ilustración 5. Analítica Web. Fuente: https://virtual.wemy.co/

Básicamente, la analítica web es la tecnología y el método para la recopilación, medición, análisis y generación de informes de sitios web y datos de uso de aplicaciones web. La analítica web ha ido creciendo desde el desarrollo de la WWW (World Wide Web). Ha pasado de ser una función simple de registro de tráfico HTTP (Protocolo de transferencia de hipertexto) a un conjunto más completo de seguimiento, análisis e informes de datos de uso. La industria y el mercado de la analítica web también están en auge con una amplia gama de herramientas, plataformas, trabajos y negocios [13] .

Una **visita a página, o *pageview* (PV)**, es una solicitud para cargar un solo archivo HTML (página web) de un sitio de Internet. En la World Wide Web, una solicitud de página sería el resultado de un internauta haciendo clic en un enlace en otra "página" que apunta a la página en cuestión.

Las visitas a página se cuentan como parte de la analítica web. Para el propietario de un sitio web, la información sobre las visitas a página puede ser útil para ver si algún cambio en la "página" (como la información en sí o la forma en que se presenta) genera más visitas. Si hay anuncios en la página, los editores también estarían interesados en el número de visitas a página para determinar sus ingresos esperados de los anuncios. Por esta razón, es un término que se utiliza ampliamente para el marketing y la publicidad en Internet [13] .

No existen definiciones acordadas a nivel mundial dentro de la analítica web, ya que los organismos de la industria han estado tratando de acordar definiciones que sean útiles y definitivas durante algún tiempo. Los principales organismos que han contribuido en esta área han sido el IAB (Interactive Advertising Bureau), JICWEBS (The Joint Industry Committee for Web Standards en el Reino Unido e Irlanda) y The DAA (Digital Analytics Association), anteriormente conocida como WAA (Asociación de análisis web, EE. UU.).

Muchas herramientas de analítica web recogen los mismos parámetros típicos y lo hacen de una forma periódica (por ejemplo, cada hora), registrando la marca temporal (*timestamp* en inglés) y los valores de los parámetros en aquel preciso instante. Algunos de estos **parámetros típicos** que se registran en analítica web son:

- *timestamp*: fecha y hora en milisegundos en formato Epoch.
- pageviews: número de visitas que llevaba la noticia en ese *timestamp*.
- avgTimeOnPage: tiempo total de todos los usuarios en la página dividido entre los usuarios que la han visitado. Es decir, una media del tiempo en página hasta el *timestamp*.
- uniquePageviews: número de visitas únicas (sin usuarios que repiten la visita).
- exitRate: porcentaje de usuarios que después de visitar esta noticia se van del portal de noticias.
- bounces: número de usuarios que entran al portal a través de esta noticia y salen del portal tras visitar esta noticia.
- bounceRate: porcentaje de usuarios que rebotan (definición en el campo anterior).
- versio: Campo que indica por qué versión de la noticia van. A veces, el diario cambia algo de la noticia y cambia la url aunque la id de la noticia es la misma; eso se considera un cambio de versión.

2.3 Inteligencia Artificial

En este proyecto se emplean conceptos y técnicas que, en última instancia, forman parte de la disciplina de la Inteligencia Artificial (IA), así que es interesante explicar qué es exactamente la IA y cómo aplica a este proyecto.

Ilustración 6. Inteligencia Artificial. Fuente: https://new.siemens.com/

La Inteligencia artificial (IA) es la capacidad de una computadora digital o un robot controlado por computadora para realizar tareas comúnmente asociadas con seres inteligentes. El término se aplica frecuentemente al proyecto de desarrollar sistemas dotados de los procesos intelectuales característicos de los humanos, como la capacidad de razonar, descubrir significados, generalizar o aprender de la experiencia pasada.

Desde el desarrollo de la computadora digital en la década de 1940, se ha demostrado que las computadoras pueden programarse para realizar tareas muy complejas, como, por ejemplo, descubrir pruebas de teoremas matemáticos o jugar al ajedrez, con gran destreza. Sin embargo, a pesar de los continuos avances en la velocidad de procesamiento de las computadoras y la capacidad de memoria, todavía no hay programas que puedan igualar la flexibilidad humana en dominios más amplios o en tareas que requieren mucho conocimiento diario.

Por otro lado, algunos programas han alcanzado los niveles de desempeño de humanos expertos y profesionales en la realización de determinadas tareas específicas, por lo que la inteligencia artificial en este sentido limitado se encuentra en aplicaciones tan diversas como diagnóstico médico, buscadores informáticos y reconocimiento de voz o escritura.

La IA es un área de la tecnología y la ciencia que engloba multitud de campos. En la Ilustración 7. Tecnologías parte de la IA. Fuente: https://bbvaopen4u.com/en/tags/artificial-intelligence se pueden observar los diferentes campos y las diferentes aplicaciones que típicamente se consideran parte de la IA:

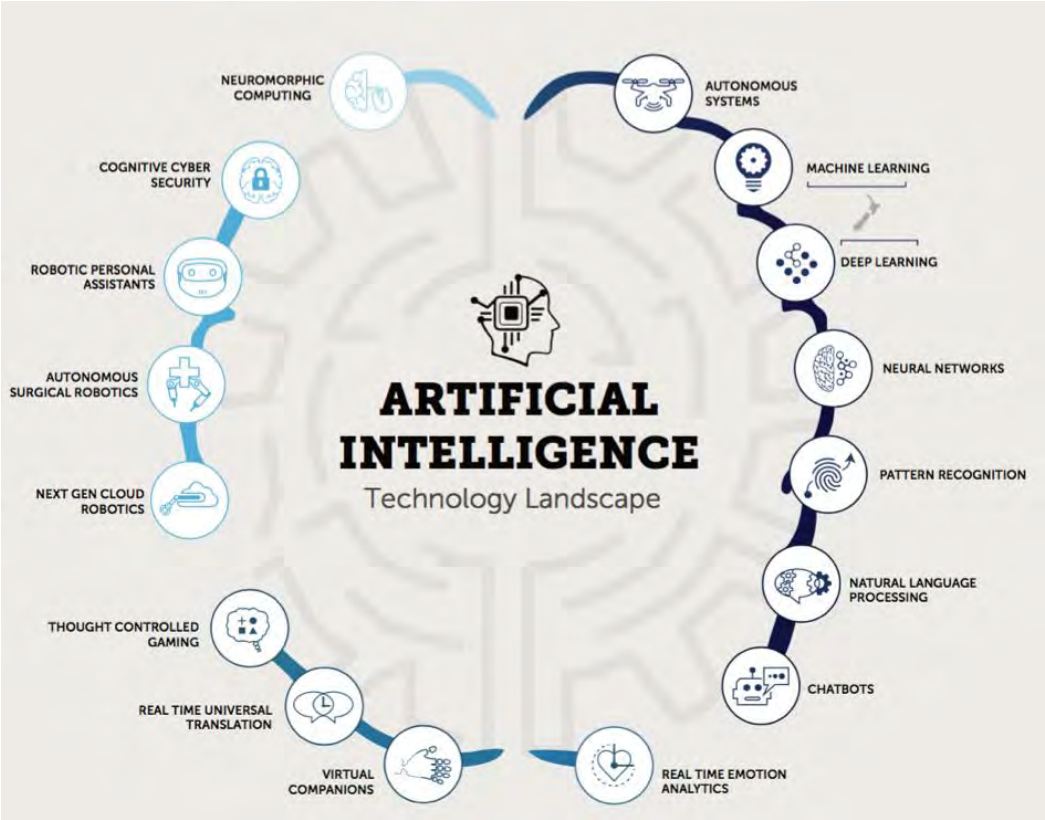

Ilustración 7. Tecnologías parte de la IA. Fuente: https://bbvaopen4u.com/en/tags/artificial-intelligence

2.3.1 Aprendizaje Automático

El aprendizaje automático, o *machine learning* (ML) en inglés, es uno de los campos más relevantes dentro del mundo de la inteligencia artificial.

En este proyecto, se aplican algunas técnicas de aprendizaje automático sobre el conjunto de datos. Por lo tanto, vale la pena hacer una revisión rápida sobre qué es y qué no es el aprendizaje automático y cómo se relaciona con este proyecto.

La definición más común de aprendizaje automático es: el estudio científico de algoritmos y modelos estadísticos que los sistemas informáticos utilizan para realizar de forma eficaz una tarea específica sin utilizar instrucciones explícitas, basándose en patrones e inferencias. Los algoritmos de ML construyen un modelo matemático basado en datos de muestra, conocidos como "datos de entrenamiento", para hacer predicciones sin estar programados para realizar la tarea.

“El aprendizaje automático también se puede definir como la ciencia de hacer que las computadoras actúen sin estar programadas explícitamente. En la última década, el aprendizaje automático nos ha brindado autos sin conductor, reconocimiento de voz práctico, búsqueda web efectiva y una comprensión mucho mejor del genoma humano. El aprendizaje automático está tan extendido hoy en día que probablemente lo usemos decenas de veces al día sin saberlo” [9] .

La ventaja de *machine learning* es que es posible utilizar los algoritmos y los modelos para prever los resultados. El truco consiste en asegurarse de que los científicos de datos que realizan el trabajo estén aplicando los algoritmos correctos, ingiriendo los datos más adecuados (precisos y limpios) y utilizando los modelos que ofrecen el mejor rendimiento. Si todos estos elementos se unen, es posible **entrenar el modelo** de forma continua y aprender de los resultados aprendiendo de los datos. La automatización de este proceso de modelar, entrenar el modelo y realizar pruebas permite generar previsiones precisas para impulsar el cambio de negocio.

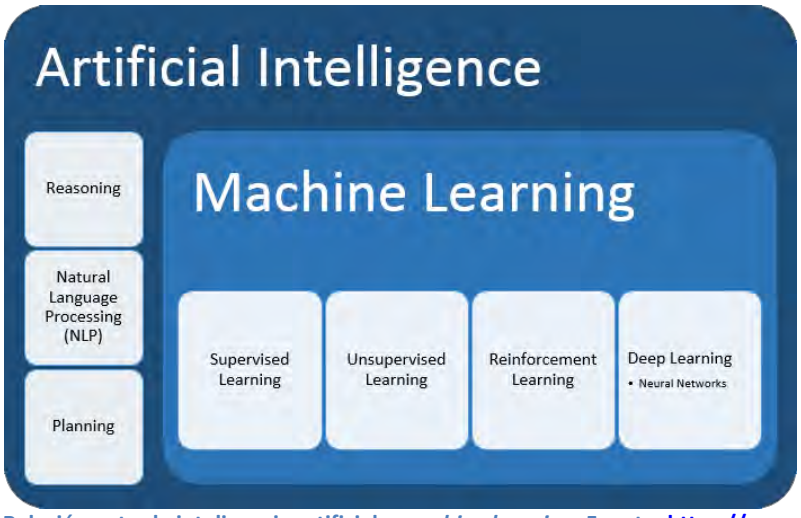

Ilustración 8. Relación entre la inteligencia artificial y *machine learning*. Fuente: https://www.ibm.com/es-es/analytics/machine-learning

2.3.1.1 Predicciones

El aprendizaje automático está muy relacionado con la estadística computacional, es decir, la ciencia de realizar predicciones con computadoras. En su aplicación a los problemas empresariales, el aprendizaje automático también se conoce como **análisis predictivo**.

El aprendizaje automático y la estadística son campos estrechamente relacionados. Algunos estadísticos han adoptado métodos de aprendizaje automático, dando paso a un campo combinado y compartido llamado **aprendizaje estadístico**.

El campo del aprendizaje automático se puede orientar hacia el análisis estadístico y el modelado, como es deseable para este proyecto. El objeto de esta orientación es **entrenar un sistema** de aprendizaje automático que pueda predecir resultados futuros. Esta capacidad predictiva es algo muy interesante para diferentes negocios. La Ilustración 9 a continuación muestra el diagrama de bloques clásico de este sistema de entrenamiento de un modelo de ML.

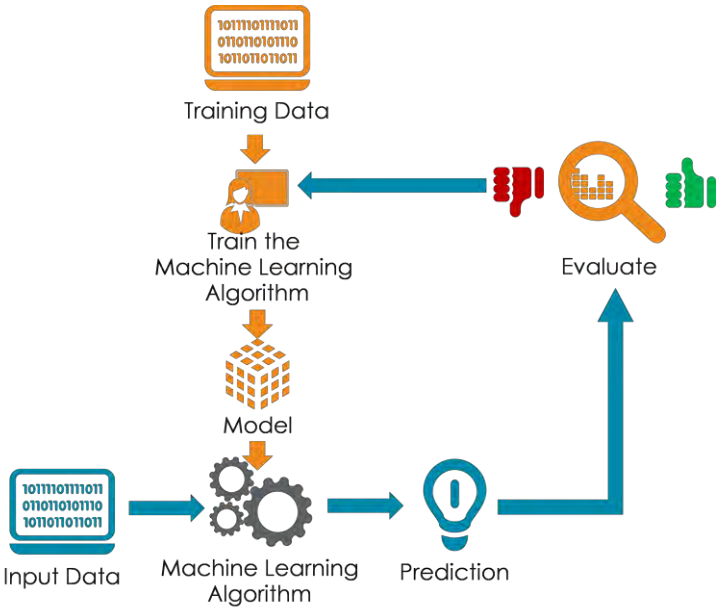

Ilustración 9. Bloques de un sistema predictivo de ML. Fuente: https://www.7wdata.be/big-data/

A continuación, se presentan y explican brevemente una serie de técnicas de aprendizaje automático que se han empleado en el desarrollo de este proyecto.

2.3.1.2 Técnicas de ML utilizadas (1): Procesado de Lenguaje Natural

Ilustración 10. Análisis de Sentimiento. Fuente: https://web.stanford.edu/class/

En este proyecto se ha aplicado, aunque de forma relativamente aproximativa y/o superficial, una técnica de Procesado de Lenguaje Natural, o *NLP*, por sus siglas en inglés, conocida como **Análisis de Sentimiento**.

El análisis de sentimiento (también conocido como minería de opiniones o IA de emociones) se refiere al uso del Procesamiento del Lenguaje Natural, análisis de texto, lingüística computacional y biometría para identificar, extraer, cuantificar y estudiar sistemáticamente **estados afectivos e información subjetiva**. El análisis de sentimiento se aplica ampliamente a los materiales de voz del cliente, como revisiones y respuestas a encuestas, redes sociales y en línea, etc., pero también se puede aplicar a textos como tweets en Twitter o como los **títulos y subtítulos de los**

artículos/noticias de un portal en línea, que es lo que se ha hecho en este proyecto.

En concreto, se ha utilizado una conocida herramienta software de análisis de texto para obtener dos métricas para cada artículo/noticia del portal de radio y noticias en línea:

a. Polaridad
b. Subjetividad

El detalle de este proceso, la explicación de la herramienta, y la descripción de estas dos métricas se encuentran más adelante, en el apartado correspondiente.

2.3.1.3 Técnicas de ML utilizadas (2): Clusterización

La clusterización o agrupamiento, en aprendizaje automático y análisis estadístico, es la agrupación de objetos de datos de un conjunto de datos en clústeres. Los clústeres son subgrupos de objetos de datos que comparten algunas funciones o características. Las técnicas de clusterización se utilizan en muchos campos, como la sociología y el análisis de mercado, para segmentar la población y descubrir patrones de comportamiento.

Aun así, no es fácil definir la noción de "clusterización" con precisión, y esta es una de las razones por las que existen tantos algoritmos de clusterización diferentes. Hay un denominador común: un conjunto de objetos de datos. Sin embargo, diferentes investigadores emplean diferentes modelos de clústeres, y para cada uno de estos modelos de clústeres se pueden dar nuevamente diferentes algoritmos. Así, el concepto de clúster, tal y como la encuentran diferentes algoritmos, varía significativamente en sus propiedades. Comprender estos "modelos de clúster" es clave para comprender las diferencias entre los distintos algoritmos. Los modelos de clúster típicos incluyen:

- Modelos de conectividad: por ejemplo, la clusterización jerárquica crea modelos basados en la conectividad de distancia.
- Modelos de centroides: por ejemplo, el algoritmo de *K-Means* representa cada clúster con un solo vector de media.
- Modelos de distribución: los clústeres se modelan mediante distribuciones estadísticas, como las distribuciones normales multivariadas.
- Modelos de densidad: por ejemplo, DBSCAN y OPTICS, que definen los clústeres como regiones densas conectadas en el espacio de datos.
- Modelos subespaciales: en *biclustering* (también conocido como co-clusterización), los clústeres se modelan con miembros del clúster y atributos relevantes.
- Modelos de grupo: algunos algoritmos no proporcionan un modelo refinado para sus resultados y solo proporcionan la información de clusterización.
- Modelos basados en gráficos: una camarilla, es decir, un subconjunto de nodos en un gráfico de manera tal que cada dos nodos del subconjunto están conectados por un borde, puede considerarse como un prototipo de clúster.

- Modelos de gráficos con signo: cada camino en un gráfico con signo tiene un signo del producto de los signos en los bordes.
- Modelos neuronales: la red neuronal no supervisada más conocida es el mapa autoorganizado y estos modelos generalmente se pueden caracterizar como similares a uno o más de los modelos anteriores, e incluyen modelos subespaciales cuando las redes neuronales implementan una forma de Análisis de Componentes Principales.

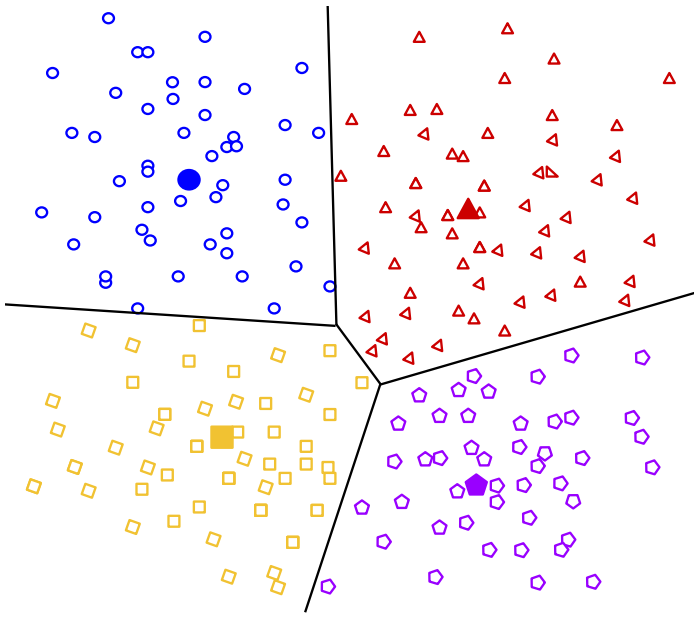

Ilustración 11. Ejemplo de clusterización por centroides. Fuente: https://developers.google.com/machine-learning/clustering/clustering-algorithms

En este proyecto, se han seleccionado un par de estos tipos de clusterización (atendiendo a su utilidad y disponibilidad) para aplicarlos sobre el conjunto de datos, con el objetivo de encontrar relaciones, descubrir patrones, detectar comunidades, etc.

2.3.1.4 Técnicas de ML utilizadas (3): *Boosting* del Gradiente

En el contexto de aprendizaje automático supervisado, el ***Boosting*** o potenciación se puede definir como una familia de algoritmos que convierten los modelos alumno débiles en modelos alumno fuertes. Un modelo alumno débil se define como un clasificador que está solo ligeramente correlacionado con la clasificación verdadera (puede etiquetar ejemplos mejor que adivinar al azar). Por el contrario, un alumno fuerte es un clasificador que está arbitrariamente bien correlacionado con la clasificación verdadera.

El árbol de decisión y ***boosting* de gradiente** (**GBDT**, por sus siglas en inglés) es una generalización del *boosting* a funciones de pérdida diferenciables arbitrarias. GBDT es un procedimiento de aprendizaje automático estándar, preciso y eficaz que se puede utilizar para problemas de **regresión** y **clasificación** en una variedad de áreas, incluida la clasificación de búsqueda web y la ecología.

En la Ilustración 12 se puede observar el esquema a alto nivel que sigue un proceso de *boosting* de gradiente. En este ejemplo, se trata de un problema de clasificación, pues se desea clasificar los objetos de un conjunto de datos en diferentes categorías o grupos según sus características. Como se puede observar, el árbol clasificador final es una combinación de los árboles clasificadores previos, que son débiles, como se puede comprobar observando sus funciones AUC. En cada iteración, se aumenta el peso relativo de las observaciones (puntos) que son difíciles de clasificar, y se reduce el peso relativo de aquellas cuya clasificación ha sido sencilla y directa.

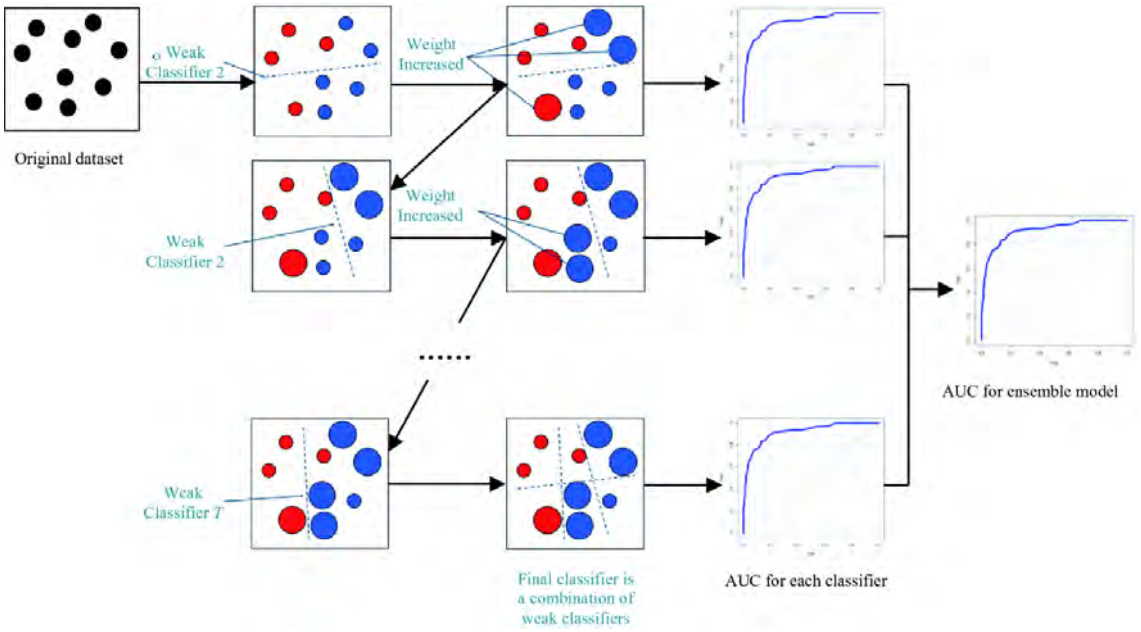

Ilustración 12. Proceso de gradient boosting. Fuente: https://datascience.eu/machine-learning/

En el caso de este proyecto, se ha utilizado la técnica del *gradient boosting* para el problema de regresión de predecir los parámetros de rendimiento de los artículos/noticias en línea en función de sus características iniciales (o atributos).

2.4 Análisis de regresión y modelos estadísticos

El concepto de **análisis de regresión** ya ha sido incluido implícitamente en los apartados anteriores, y es un concepto básico en este proyecto, pues en él se realizan varios tipos de análisis de regresión. Está interrelacionado con los conceptos de modelado estadístico y aprendizaje automático. Es, pues, menester detenerse y explicar este concepto general con detalle.

Se hace un análisis de regresión cuando se estudia la dependencia de un resultado (p.ej., el número total de visitas a un artículo/noticia en línea) de una o varias variables independientes, p.ej. su categoría, título, cantidad de imágenes/vídeos…

El análisis de regresión es una forma de clasificar matemáticamente cuál de esas variables tiene realmente un impacto. Responde a las preguntas: ¿Qué factores son

más importantes? ¿Qué podemos ignorar? ¿Cómo interactúan esos factores entre sí? Y, quizás lo más importante, ¿qué tan seguros estamos de todos estos factores?

En el análisis de regresión, esos factores se denominan variables. Está la variable dependiente, el factor principal que se está tratando de comprender o predecir. Y luego están las variables independientes: los factores que se sospecha que tienen un impacto en el valor de la variable dependiente.

Un **modelo estadístico** es la estructura o función matemática que se utiliza para realizar el análisis de regresión. Es decir, un modelo estadístico generalmente se especifica como una relación matemática entre una o más variables aleatorias y otras variables no aleatorias.

Dicho de otro modo, un modelo estadístico es un modelo matemático que incorpora un conjunto de supuestos estadísticos sobre la generación de datos de muestra. Un modelo estadístico representa, a menudo de forma idealizada, el proceso de generación de datos.

Los modelos estadísticos utilizan ecuaciones matemáticas para codificar la información extraída de los datos. En muchos casos, las técnicas de modelado estadístico pueden generar modelos adecuados muy rápidamente. Incluso para problemas en los que las técnicas de aprendizaje automático más flexibles (como las redes neuronales) pueden dar mejores resultados, se pueden utilizar algunos modelos estadísticos como modelos predictivos de referencia para juzgar el rendimiento de métodos más avanzados.

2.4.1 Análisis de regresión no lineal

En este proyecto, los modelos estadísticos se utilizan para realizar un análisis de regresión no lineal. En estadística, una regresión no lineal es una forma de análisis de regresión en la que los datos de observación son modelados por una función que es una combinación no lineal de los parámetros del modelo y depende de una o más variables independientes. Los puntos de datos se ajustan mediante el método de aproximaciones sucesivas.

En regresión no lineal, un modelo estadístico de la forma,

$$y = f(x; \beta)$$

relaciona un vector de variables independientes, x, y sus variables dependientes observadas asociadas, y. La función f no es lineal en las componentes del vector de parámetros β, pero por lo demás arbitraria. Para que la función f no sea lineal, no se puede expresar como una combinación lineal de los elementos del vector de parámetros β.

Ejemplos de funciones no lineales incluyen funciones exponenciales, funciones logarítmicas, funciones de potencia, la función gaussiana, curvas de Lorenz y las

funciones matemáticas que representan los modelos estadísticos utilizados en este proyecto.

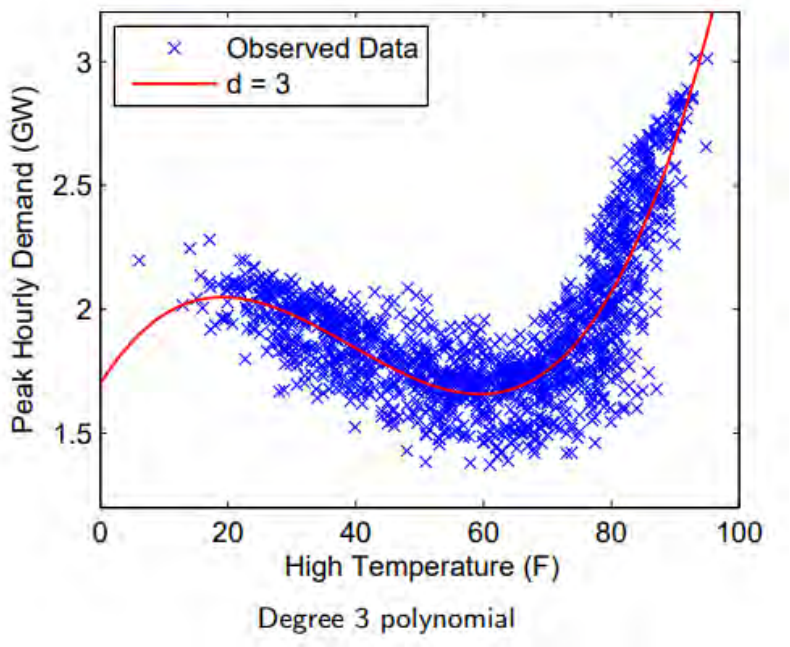

Ilustración 13. Ejemplo de regresión no lineal de orden 3. Fuente: https://cs.stanford.edu/

2.4.1.1 Mínimos cuadrados no lineales

El método de mínimos cuadrados no lineales es una variación o subtipo del método de mínimos cuadrados, por lo que este último se aborda en primer lugar.

El método de mínimos cuadrados (LS) es un enfoque estándar en el análisis de regresión para aproximar la solución de sistemas sobredeterminados, es decir, conjuntos de ecuaciones en las que hay más ecuaciones que incógnitas. "Mínimos cuadrados" significa que la solución general minimiza la suma de los cuadrados de los residuos obtenidos en los resultados de cada ecuación.

La aplicación más importante es el ajuste de datos. El mejor ajuste en el sentido de mínimos cuadrados minimiza la suma de los residuos al cuadrado (**residuo**: la diferencia entre un valor observado y el valor ajustado proporcionado por un modelo). Cuando el problema tiene incertidumbres sustanciales en la variable independiente (la variable x), los métodos de regresión simple y mínimos cuadrados tienen problemas; en tales casos, se puede considerar la metodología requerida para ajustar modelos de errores en variables en lugar de la de mínimos cuadrados.

Los problemas de mínimos cuadrados se dividen en dos categorías:

- mínimos cuadrados lineales u ordinarios
- mínimos cuadrados no lineales,

dependiendo de si los residuos son lineales en todas las incógnitas. El problema no lineal generalmente se resuelve mediante un refinamiento iterativo; en cada iteración el sistema se aproxima por uno lineal.

El método de mínimos cuadrados no lineales es la forma de análisis de mínimos cuadrados que se utiliza para ajustar un conjunto de *m* observaciones con un modelo que no es lineal en *n* parámetros desconocidos (*m* ≥ *n*). Se utiliza en algunas formas de regresión no lineal. La base del método es aproximar el modelo por uno lineal y refinar los parámetros por iteraciones sucesivas. Ejemplos de mínimos cuadrados no lineales son la regresión de umbral, la regresión suave, la regresión de enlace logístico, la regresión *probit* y muchos otros en teoría económica.

Considere un conjunto de ***m*** puntos de datos, $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ y una curva (función modelo) $y = f(x; \beta)$ que, además de la variable **x**, también depende de ***n*** parámetros, $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_n)$ con $m >= n$. Se desea encontrar el vector **β** de parámetros de manera que la curva se ajuste mejor a los datos dados en el sentido de mínimos cuadrados, es decir, la suma de cuadrados:

$$S = \sum_{i-1}^{m} r_i^2$$

sea minimizado, donde los residuos (error de las predicciones en la muestra) r_i son

$$r_i = y_i - f(x_i; \beta)$$

para $i = 1, 2, \ldots, m$.

En este proyecto, se ha utilizado el método de mínimos cuadrados no lineales para ajustar la curva de visitas a página de los artículos/noticias en línea con diferentes métodos estadísticos que son presentados y explicados más adelante.

2.5 Estado del arte

En la primera y única sección de este capítulo, se hace una breve introducción y revisión del estado del arte actual en los campos del modelado estadístico, la analítica web y el aprendizaje automático aplicado a los datos de los negocios online.

2.5.1 Perspectiva de negocio

En los últimos años, ha habido un gran crecimiento del uso por parte de muchos medios en línea, portales de radio, noticias en la web y demás, de herramientas de analítica web para aprovechar los datos de que disponen y tomar decisiones editoriales y organizativas más informadas. Estas **herramientas de analítica web** pueden ser gratuitas o de pago y pueden provenir de gigantes como Google, o ser un

software *open-source*. La demanda por estas herramientas ha crecido de manera exponencial y se prevé que siga creciendo fuertemente en el futuro cercano. Por eso, no es de extrañar que en los últimos años la oferta se haya ampliado y ensanchado de forma notable.

Ilustración 14. Herramientas de analítica web. Fuente: https://cihangir.co.uk/web-analytics-consulting/

En el contexto de los medios de comunicación digitales, portales de noticias en línea, diarios y/o radios, etc., apareció hace unos años el concepto de **analítica editorial**, que se considera una evolución de la analítica web clásica basada en los clics que venía siendo la regla la última década.

Una investigación del NiemanLab [34] , de Harvard, muestra que sólo unas pocas organizaciones de noticias en línea han desarrollado lo que se ha dado en llamar "analítica editorial": enfoques de análisis que van más allá del uso genérico de herramientas y técnicas estándar, y que desarrollan un **enfoque personalizado** de analítica alineado con las prioridades editoriales específicas y con los imperativos organizativos de un medio de noticias en concreto.

Desde el punto de vista **económico**, está claro que el interés de los grupos de medios de comunicación, periódicos y radios en línea está fijado en gran parte en los ingresos generados por la publicidad contratada en sus portales en la web. Un reciente análisis del Pew Research Center [35] muestra que, para los periódicos, diarios, etc. con versión de prensa y versión en línea, los ingresos por publicidad y anuncios en línea ya supusieron el 35% de los ingresos totales por publicidad y anuncios. Como se ve en la Ilustración 15, es una tendencia creciente.

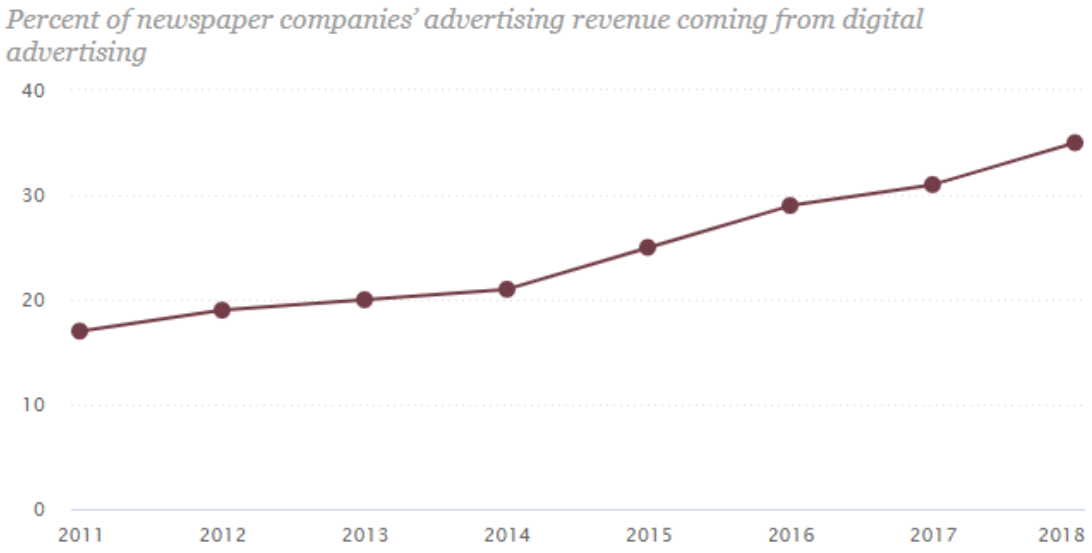

Ilustración 15. Porcentaje de ingresos por publicidad en línea para periódicos. Fuente: https://www.journalism.org/fact-sheet/newspapers/

Las herramientas de analítica web pueden ayudar a los periódicos y portales de noticias en línea a aprovecharse mejor de esta tendencia, siempre respetando tanto la cultura propia y la línea editorial como la independencia y criterio de los visitantes y lectores. En este sentido, una máxima que se ha extendido en este mundo de la analítica web es que se debe ser "**informado por los datos, no conducido por los datos**".

2.5.2 Perspectiva técnica

Desde el punto de vista científico-técnico, en los últimos tiempos muchas técnicas de aprendizaje automático han sido aplicadas con éxito sobre los datos de periódicos y portales de radio y noticias en línea.

Por otro lado, la comunidad de la ciencia de datos es muy activa y dinámica, y plataformas como **Kaggle** organizan competiciones en línea donde se plantean problemas de aprendizaje automático de clasificación y de regresión. Las competiciones han dado lugar a muchos proyectos exitosos, incluido el avance del estado del arte en la investigación del VIH, calificaciones de ajedrez y el pronosticado de tráfico en la red. Ello quiere decir que constantemente se generan soluciones de mucho valor añadido para la comunidad y para la sociedad, ya que pueden ser aplicadas para resolver problemas u optimizar procesos que tienen las organizaciones en el mundo real.

De estas competiciones en la red han salido en los últimos años varias soluciones tan eficaces y eficientes que han sido integradas en diferentes marcos de desarrollo y/o convertidas o incluidas en librerías/paquetes para R o Python.

Más concretamente, el subcampo de aprendizaje automático conocido como ***boosting*** de gradiente ha adquirido bastante relevancia recientemente, debido a que varias de las soluciones más rápidas y robustas a distintos problemas de clasificación y/o regresión se basan en las técnicas de *gradient boosting*. Una de ellas es ***XGBoost***, de la que se habla en detalle más adelante.

La compañía inglesa **Technavio**, líder mundial en investigación y asesoramiento en tecnología, publicó recientemente un informe de su investigación y análisis de las tendencias de los mercados emergentes, entre las que se incluyen las últimas tendencias y predicciones sobre la ciencia de datos y la analítica web [36] .

Según Technavio, la proliferación de IA en el análisis web tendrá un impacto positivo en el mercado y contribuirá a su crecimiento de manera significativa durante el período de pronóstico. Este informe de investigación también analiza otras tendencias importantes e impulsores del mercado que afectarán el crecimiento del mercado durante 2020-2024.

Los proveedores del mercado están integrando cada vez más la inteligencia artificial (IA) en el análisis web, que puede proporcionar información más procesada de los sitios web y hace que el análisis web sea más conveniente para los comerciantes en línea y los titulares de sitios web. Además, el aprendizaje automático ayuda a estudiar los patrones de comportamiento humano en los sitios web y puede correlacionarse y analizarse con herramientas de análisis web para encontrar el resultado deseado. Por lo tanto, los beneficios de integrar la inteligencia artificial en el análisis web impulsarán a las empresas a adoptar el análisis web durante el período de pronóstico.

Gartner Inc., compañía americana líder mundial en consultoría e investigación de las tecnologías de la información (TIC), publicó también recientemente un informe donde se identifican las principales tendencias mundiales en los campos de ciencia de datos y analítica web [37] . Dos de las principales tendencias que identifica Gartner en su investigación son:

1. El desarrollo de una IA más inteligente, rápida y responsable, que se integra eficientemente con la analítica web y de datos. En concreto, estima que, para finales de 2024, el 75% de las empresas pasarán de realizar pruebas piloto a poner en funcionamiento la IA, lo que generará un aumento de cinco veces más infraestructuras de análisis y datos de transmisión.

   En el contexto actual de la pandemia, las técnicas de IA como el aprendizaje automático (ML), la optimización y el **procesamiento del lenguaje natural** (NLP) están proporcionando conocimientos y predicciones vitales sobre la propagación del virus y la eficacia y el impacto de las contramedidas.

2. El advenimiento de una Inteligencia de Decisiones (***Decision Intelligence***): Para 2023, más del 33% de las grandes organizaciones tendrán analistas que practiquen la inteligencia de decisiones, incluido el modelado de decisiones. La inteligencia de decisiones reúne una serie de disciplinas, incluida la gestión de decisiones y el apoyo a las decisiones. Abarca aplicaciones en el campo de los

sistemas adaptativos complejos que reúnen múltiples disciplinas tradicionales y avanzadas.

Proporciona un marco para ayudar a los líderes de datos y análisis a diseñar, modelar, alinear, ejecutar, monitorear y ajustar los modelos y procesos de decisión en el contexto de los resultados y el comportamiento del negocio.

3 Dataset

A lo largo de este capítulo, se revisa y explica todo lo que tiene que ver con el conjunto de datos utilizado en este proyecto. Sin los datos brutos, este proyecto no podría haber comenzado. El conjunto de datos inicial es donde comienza todo y es lo que marca el tipo de estudio que se puede hacer, así como los resultados de este.

Primero, se hace una introducción general al conjunto de datos, desde cómo se obtuvo hasta una descripción completa del contenido de este.

En la siguiente sección, se hace una visualización del contenido del conjunto de datos, para tener una idea de la evolución de los parámetros de los artículos/noticias con el paso de las horas.

Las dos secciones que siguen están dedicadas a la presentación y explicación del limpiado y del preprocesado que se hace del conjunto de datos, respectivamente. Esto se hace para eliminar objetos de datos incompletos o incoherentes y para preparar el conjunto de datos para el estudio que se quiere hacer.

Por último, en la última sección de este capítulo se presentan y explican tres procedimientos que sirven para ampliar la base de características iniciales de los artículos/noticias en línea. Esto se hace para tener más atributos con los que trabajar a la hora de realizar predicciones o clusterizaciones.

3.1 Descripción

El conjunto de datos sobre el que se ha trabajado en este proyecto es una matriz que contiene datos sobre un número elevado de artículos/noticias del portal web de una emisora de radio bastante popular. Este conjunto de datos se ha considerado suficientemente completo, detallado y fiable para como para realizar una exploración estadística de este. Que el conjunto de datos contenga un número elevado de objetos es importante para la significancia y confiabilidad de los resultados del análisis estadístico.

Como se ha explicado en apartados anteriores, este conjunto de datos se ha obtenido de una conocida herramienta de analítica web, utilizada para registrar el valor de una serie de parámetros para cada artículo/noticia del portal web que se incluye en el alcance. En concreto, esta herramienta captura estos valores con una frecuencia de una hora, y, en general, se interrumpe la captura de estos valores cuando han transcurrido alrededor de 100 horas (4 días aprox.) desde la publicación del artículo/noticia en el portal web. No obstante, en el conjunto de datos obtenido existen tanto objetos que tienen menos de 100 líneas (horas) de datos como objetos que tienen más de 100. Esto se ha tenido en cuenta y se explica más adelante.

En concreto, este conjunto de datos no es más que una colección de marcas temporales (*timestamps* en inglés) junto con el identificador del artículo/noticia, sus características iniciales o atributos, que no cambian, y el valor de una serie de parámetros de rendimiento, que sí van cambiando o pueden ir cambiando cada hora.

Tiene las siguientes dimensiones: (1557933, 17). Si se fija el ID del artículo/noticia como clave primaria, el número total de objetos únicos es 17849.

La estructura es la siguiente:

ID de artículo/noticia 1; *marca temporal 1*; título; texto; subtítulo; longitud del texto; número de imágenes; número de otros*; categoría; visitas a página; salidas; tiempo medio en página; visitas a página únicas; tasa de salida; rebotes; tasa de rebotes;

ID de artículo/noticia 1; *marca temporal 2*; título; texto; subtítulo; longitud del texto; número de imágenes; número de otros*; categoría; visitas a página; salidas; tiempo medio en página; visitas a página únicas; tasa de salida; rebotes; tasa de rebotes;

.

.

.

ID de artículo/noticia 1; *marca temporal N*; título; texto; subtítulo; longitud del texto; número de imágenes; número de otros*; categoría; visitas a página; salidas; tiempo medio en página; visitas a página únicas; tasa de salida; rebotes; tasa de rebotes;

ID de artículo/noticia 2; *marca temporal 1*; título; texto; subtítulo; longitud del texto; número de imágenes; número de otros*; categoría; visitas a página; salidas; tiempo medio en página; visitas a página únicas; tasa de salida; rebotes; tasa de rebotes;

.

.

.

ID de artículo/noticia X; *marca temporal Y*; título; texto; subtítulo; longitud del texto; número de imágenes; número de otros*; categoría; visitas a página; salidas; tiempo medio en página; visitas a página únicas; tasa de salida; rebotes; tasa de rebotes;

*número de otros se refiere a número de vídeos y/u otros contenidos multimedia.

En la Tabla 1. Parámetros de los objetos del conjunto de datos. se puede observar la definición de cada uno de los parámetros de rendimiento, así como el tipo de dato.

Atributo/Parámetro	Definición	Tipo de dato
ID de artículo/noticia	El identificador individual de cada artículo de noticias en línea. Es un identificador numérico de seis dígitos. También conocido como índice.	Entero
Marca temporal	Tiempo UNIX. Va hora a hora. Cada ID de noticias tiene su propia matriz de marcas de tiempo cuando se capturan los demás parámetros.	Entero (convertido en ocasiones a tipo Fecha)
Visitas a página	La cantidad de visitas a página totales que tiene un artículo de noticias en línea en un momento dado.	Entero
salidas	La cantidad de visitas / sesiones que han finalizado en la página de un artículo de noticias en línea en un momento dado.	Entero
Tiempo medio en página	El tiempo promedio que se dedica realmente a interactuar con la página de un artículo de noticias en línea en un momento dado.	Entero
Visitas a página únicas	El número de visitas a página que no provienen del mismo visitante / máquina / IP que tiene un artículo de noticias en línea en un momento dado.	Entero
Tasa de salidas	El porcentaje de visitas / sesiones que han finalizado en la página de un artículo de noticias online en un momento dado.	Flotante

rebotes	La cantidad de visitas / sesiones que acceden a la página de un artículo de noticias online y luego salen inmediatamente, en cualquier momento.	Entero
Tasa de rebotes	El número de visitas / sesiones que acceden a la página de un artículo de noticias online dividido por el número total de visitas / sesiones, en un momento dado.	Flotante

Tabla 1. Parámetros de los objetos del conjunto de datos.

De todos estos parámetros de rendimiento, los que se han considerado los tres más importantes y relevantes son, por orden:

1. **Visitas a página**: Es el parámetro de rendimiento más importante en términos de analítica de clics, que sigue siendo la forma más funcional de hacer analítica web. De su valor puede depender la cantidad y calidad de publicidad que se contrate para uno o varios artículos/noticias en el portal web. Por lo tanto, para la dirección puede haber un interés económico claro en saber predecir aproximadamente cuántas visitas a página tendrá un artículo/noticia.
   En este proyecto, hay un capítulo entero dedicado al modelado de la curva de evolución del número de visitas a página de los artículos/noticias.

2. **Tiempo medio en página**: Es un parámetro de rendimiento bastante relevante en todo tipo de analíticas web. En principio, cuanto más tiempo pase un visitante en la página de un artículo/noticia más probable será que le haya gustado lo que ha leído/visto, y más probable será que haya visto los anuncios y la publicidad que hay en la página. Aun así, suele ser un parámetro difícil de interpretar correctamente.
   En este proyecto, se ha incluido este parámetro tanto en los análisis de clusterización como en los análisis de regresión para la predicción de su valor; ambos análisis se presentan y explican más adelante.

3. **Tasa de salidas**: Es un parámetro de rendimiento útil de conocer, ya que un valor alto es indicativo de que un determinado artículo o grupo de artículos está causando, por las razones que fuere, que los visitantes cierren su sesión de navegación en el portal web, es decir, salgan de él desde la página de ese artículo/noticia.
   En este proyecto, se ha incluido este parámetro en los análisis de clusterización.

Por la importancia relativa de estos tres parámetros de rendimiento respecto de resto, y por las limitaciones temporales y de otros tipos anteriormente explicadas, el alcance de este proyecto en términos de parámetros de rendimiento se ha limitado a estos tres.

3.2 Visualización

En esta sección, se realiza una descripción gráfica del conjunto de datos, con el objetivo de obtener una vista a alto nivel de la distribución de los valores de las características iniciales (o atributos) y de la evolución de los diferentes parámetros de rendimiento para todos los artículos de noticias en línea que están disponibles en el conjunto de datos.

Nota: Los puntos de tiempo en el conjunto de datos se formatearon como marcas de tiempo UNIX, que se convirtieron a fechas en el formato "DD/MM/AAAA hh:mm:ss". Se publicaron diferentes artículos en el sitio web en diferentes días y en diferentes momentos, por lo que los gráficos resultantes no serían inteligibles y agradables a la vista, debido a las diferencias con el eje x (tiempo). Entonces, en aras de la limpieza y la inteligibilidad, las fechas se convirtieron a horas y todas las curvas de tiempo de visitas a la página comienzan en x = 1, lo que indica la primera hora de datos recopilados para cada artículo de noticias en línea.

3.2.1 Visitas a página acumuladas

Las curvas de tiempo de visitas a página acumuladas muestran la cantidad de visitas a página acumuladas que tiene un artículo/noticia del portal web en un momento dado.

En la Ilustración 16 y la Ilustración 17 se puede observar la curva de visitas a página acumuladas por cada uno de los artículos/noticias que hay en el conjunto de datos con el que se ha trabajado finalmente, después de realizar las pertinentes tareas de limpieza y preprocesado del conjunto de datos original, que se presentan y explican en sus respectivas secciones, más adelante.

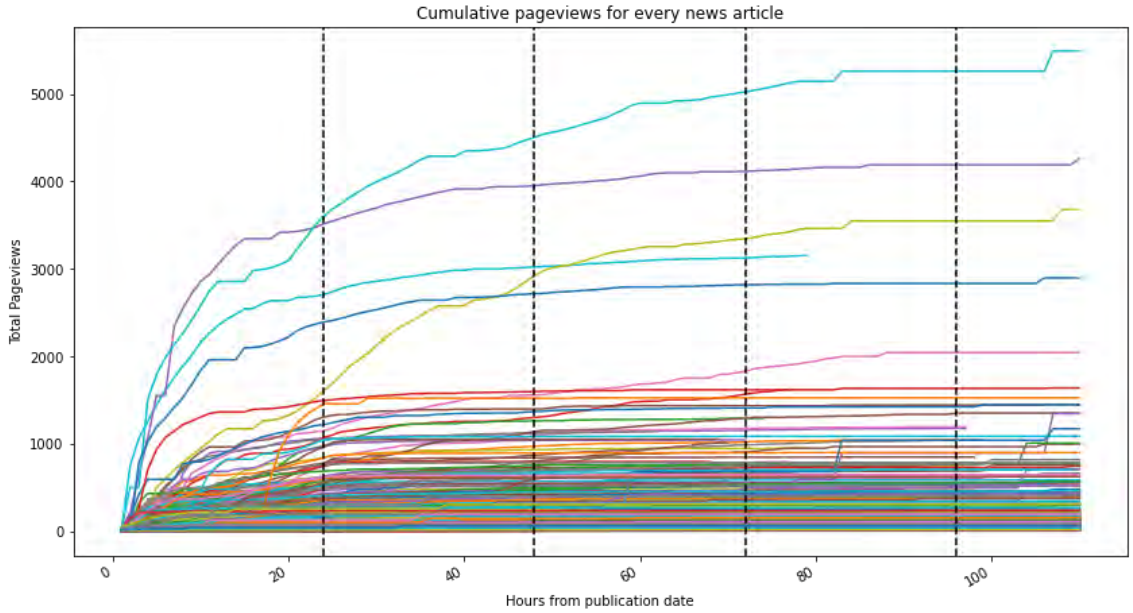

Ilustración 16. Curvas de visitas a página acumuladas por todos los artículos/noticias.

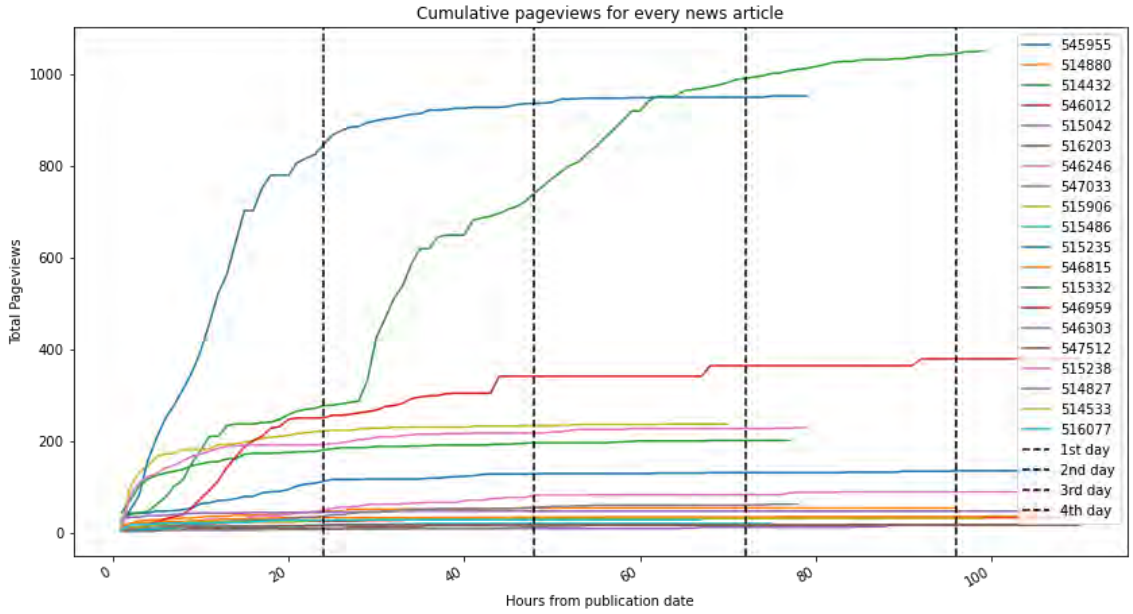

Ilustración 17. Curvas de visitas a página acumuladas por una muestra de 20 artículos/noticias.

Varias cosas se pueden observar:

- Las páginas de la inmensa mayoría de los artículos/noticias del portal web son visitadas menos de 1000 veces a lo largo de su vida. Esto puede ser indicativo del impacto/alcance social que tiene este portal web, y de la dificultad de generar contenidos que sean consultados un número importante de veces.

- Las primeras horas son críticas, y algunas horas antes o después de un día desde la fecha de publicación es cuando la mayoría de los artículos de noticias alcanzan su máximo impacto (si impacto = visitas a página totales). Estos

hechos demuestran la naturaleza acelerada de la industria editorial de noticias y la inmediatez que se requiere para los lectores y visitantes.

3.2.2 Visualización de la categoría

Los objetos (artículos/noticias) del conjunto de datos tienen un atributo de categoría, pero existen demasiadas categorías como para poder realizar cualquier tipo de análisis estadístico que incluya este atributo. Por esto, se ha realizado una recategorización de los artículos/noticias, que se presenta y se explica en otra sección más adelante.

Las 30 categorías más comunes en el conjunto de datos con el que finalmente se ha trabajado son las que se observan en la Ilustración 18.

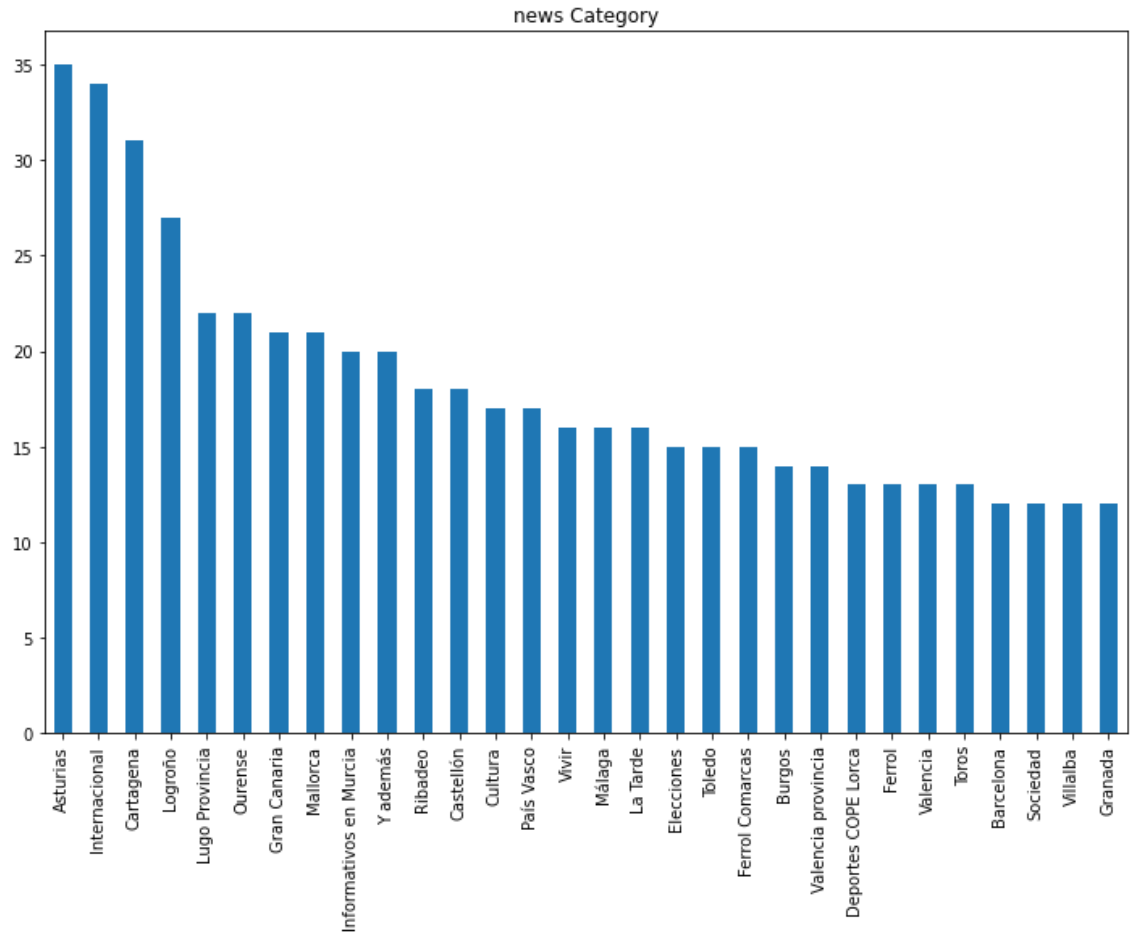

Ilustración 18. Las 30 categorías más comunes en el conjunto de artículos/noticias.

Como se puede observar, hay una sobrerrepresentación del ámbito local (comunidades, provincias, ciudades) en las categorías de los artículos/noticias, es decir, muchos, si no todos, estos artículos/noticias tratarán de algún tema (religión, economía, deportes…) que bien podría ser el atributo de categoría.

3.2.3 Visualización de otras características iniciales

A continuación, en la Ilustración 19, se observan las distribuciones de los valores de otras características iniciales (o atributos) de las que originalmente figuran en el conjunto de datos. Estos son, únicamente, el nº de imágenes y el nº de vídeos y otros en la página del artículo/noticia.

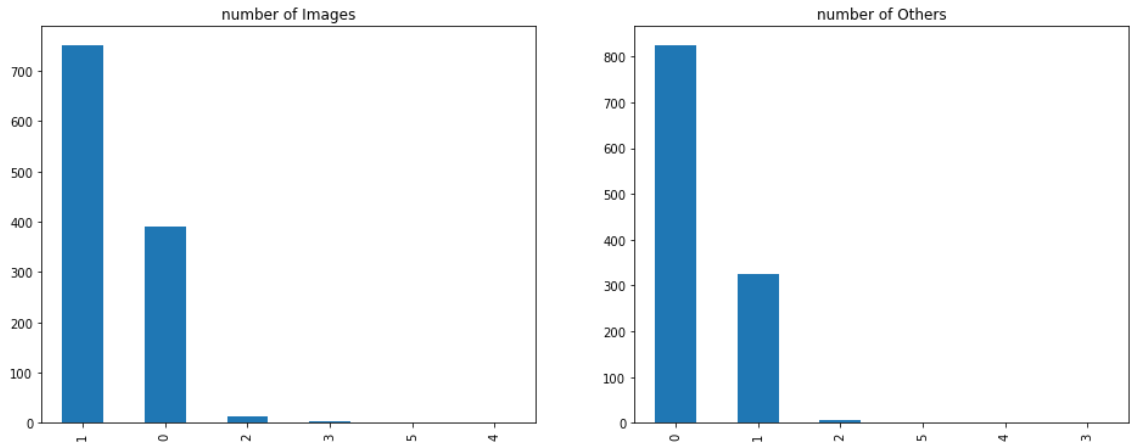

Ilustración 19. Histograma del nº de imágenes y de vídeos, etc. en las páginas de los artículos/noticias.

Se observa que la inmensa mayoría de artículos/noticias tienen una imagen, la de la cabecera, como es habitual. Muchos otros no tienen siquiera esta, y esto se puede deber a varios factores, como el tipo de contenido, etc.

Por otro lado, se puede observar, sin necesidad de cálculos, que la ratio entre el nº de páginas que no tienen ningún vídeo (u otro) y las que tienen uno es muy parecido a la ratio entre el nº de páginas que tienen una imagen y las que no tienen ninguna, por lo que es de suponer que habrá muchas "páginas-vídeo/audio", sin imágenes sueltas, como es relativamente lógico si se tiene en cuenta que se trata de un portal de radio.

3.3 Limpieza de datos

Ilustración 20. Limpieza de datos. Fuente: https://www.pngitem.com/

La limpieza de datos es el proceso de detectar, corregir o eliminar registros corruptos o inexactos de un conjunto de registros, de una tabla o de una base de datos. Se refiere a la identificación de partes incompletas, incorrectas, inexactas o irrelevantes de los datos y luego reemplazar, modificar, o borrar los datos sucios o malos. La limpieza de datos se puede realizar de forma interactiva con herramientas de gestión de datos o como procesamiento por lotes mediante secuencias de comandos.

Después de la limpieza, un conjunto de datos debe ser coherente con otros conjuntos de datos similares en el sistema. Las inconsistencias detectadas o eliminadas pueden haber sido causadas originalmente por errores de entrada del usuario, por corrupción en la transmisión o almacenamiento, o por diferentes definiciones de diccionario de datos de entidades similares en diferentes almacenes. La limpieza de datos difiere de la validación de datos en que la validación significa casi invariablemente que los datos se rechazan del sistema al ingresar y se realizan en el momento de la entrada, en lugar de en lotes de datos.

En este proyecto, **no** se ha realizado explícitamente ninguna tarea de limpieza de datos explícitamente, porque no se ha considerado necesario. Los datos del conjunto de datos se han considerado completos y correctos.

Sin embargo, sí se ha considerado reducir el gran tamaño del conjunto de datos original (17849 objetos únicos) a un tamaño mucho menor, cercano al 10%, para agilizar el procesado de datos y el análisis de estos, manteniendo un universo de objetos suficientemente grande.

3.4 Preprocesado de datos

El preprocesamiento de datos es un paso importante en el proceso de minería de datos. La frase "basura entra, basura sale" es particularmente aplicable a proyectos de minería de datos y de aprendizaje automático. Los métodos de recopilación de datos a menudo se controlan de manera poco estricta, lo que da como resultado valores fuera de rango (p. Ej., Ingresos: −100), combinaciones de datos imposibles (p. Ej., Sexo: Hombre, Embarazada: Sí), valores perdidos, etc. Si no se ha examinado cuidadosamente para detectar tales problemas, se pueden producir resultados engañosos. Por lo tanto, la representación y la calidad de los datos es ante todo antes de ejecutar un análisis. A menudo, el preprocesamiento de datos es la fase más importante de un proyecto de aprendizaje automático, especialmente en biología computacional.

En este proyecto, el preprocesamiento de los datos se ha realizado en base a dos restricciones:

a. La necesidad de reducir el tamaño del conjunto de datos con que se trabaja, para reducir el coste temporal y computacional de los análisis.

b. La necesidad de disponer en el conjunto de datos de valores correctos del parámetro de visitas a página, para poder realizar con seguridad el

estudio de caracterización de la evolución de este parámetro mediante modelos estadísticos. Por valores correctos se entiende valores que comiencen en 0 en el inicio de la serie temporal (timestamp = 0), y valores que se actualicen a cada hora después de la publicación.

Las tareas de preprocesamiento realizadas se pueden observar en la siguiente Ilustración 21, con las líneas de código en Python:

```
In [17]: #Número de objetos (IDs) únicos en el conjunto de datos original
         uniqueIds = data1.index.unique()
         uniqueIds

Out[17]: Index(['165189', '166883', '292957', '349240', '352770', '363368', '364454',
                '395449', '492973', '505567',
                ...
                '614163', '614172', '614176', '614180', '614182', '614192', '614193',
                '614194', '614198', '614202'],
               dtype='object', name='newsId', length=17849)
```

```
In [18]: #Inicialización del array de los IDs con los que se trabajará
         unIdsOK = []
         for unId in uniqueIds:
             #Método para recuperar la serie de fechas/horas y visitas
             #a la página de un determinado artículo/noticia
             datesC, hoursC, pvcumC, pvdiffC = getPageViews(unId, data1)
             totHoursC = len(hoursC)
             #Sólo se consideran los objetos con más de un día de datos
             if (totHoursC >= 24):
                 #Sólo se consideran los objetos cuyas visitas a página
                 #iniciales sean cercanas a cero (como debería ser)
                 if ((pvcumC[-1] > pvcumC[23]) and (pvcumC[0] <= 50)):
                     chng = 0
                     same = 0
                     pla = 0
                     #Bucle para considerar únicamente aquellos objetos
                     #cuyas visitas a página se actualicen cada hora
                     for hh in range(0, totHoursC - 1):
                         if ((pvcumC[hh + 1] > pvcumC[hh])):
                             chng += 1
                             if (same == 23):
                                 pla += 1
                             same = 0
                         else:
                             same += 1
                     if ((chng > 5) and (pla < 3)):
                         unIdsOK.append(unId)
```

```
In [19]: #Número final de objetos (IDs) con los que se trabaja
         print(len(unIdsOK))

         1161
```

Ilustración 21. Preprocesado de los datos originales en código Python.

3.5 Ampliación de atributos base

En esta sección, se presentan y explican tres tareas y/o procedimientos que se han realizado durante el desarrollo de este proyecto con el objetivo de ampliar y ensanchar el número de características iniciales o atributos base de los artículos/noticias. El disponer de más datos iniciales sobre estos permite incluir más variables y más posibilidades en los análisis de clasificación y de regresión.

3.5.1 Longitud de título y subtítulo

La primera tarea que se ha realizado en este sentido ha sido el cálculo de la longitud del título y del subtítulo de los artículos/noticias. A priori, es posible que estas dos características iniciales (suponiendo que no se cambian en ningún momento) tengan alguna influencia sobre los resultados de rendimiento, aunque sea mínima. Es decir, **podría ser** que los artículos/noticias con **títulos y subtítulos muy largos** sean menos atrayentes visualmente y, por ello, **reciban menos clics**. Por eso, se han incluido en los análisis realizados.

En la Ilustración 22 se observa la distribución del valor de estos dos atributos en el conjunto de datos con que se ha trabajado finalmente.

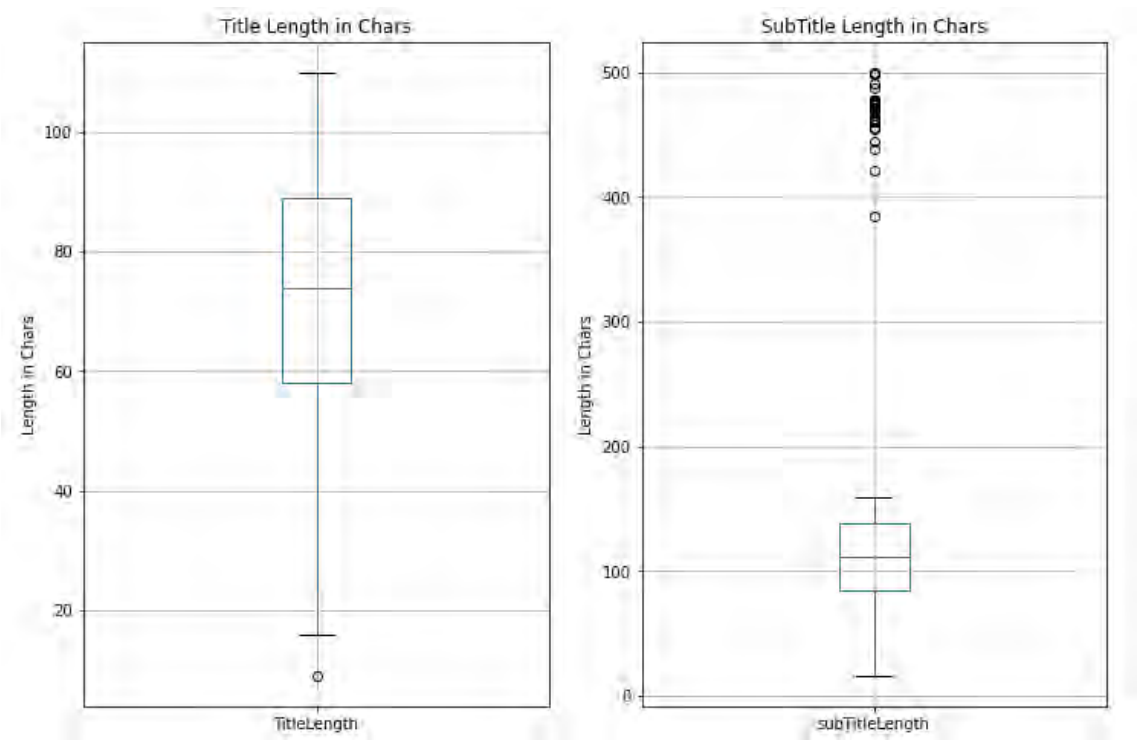

Ilustración 22. Distribución de la longitud del título (izq.) y del subtítulo (der.)

Se observa en estos diagramas de caja que la longitud de los títulos varía más.

3.5.2 Atributos de Análisis de Sentimiento

Como ya se ha comentado en el apartado introductorio, en este proyecto se han aplicado técnicas de Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) y en concreto técnicas de Análisis de Sentimiento (AS) de los títulos y subtítulos de los artículos/noticias del conjunto de datos.

Se ha utilizado la popular herramienta **Textblob**. TextBlob es una biblioteca de Python (2 y 3) para procesar datos textuales. Proporciona una API simple para sumergirse en tareas comunes de NLP, como etiquetado de parte del discurso, extracción de frases nominales, Análisis de Sentimiento (AS), clasificación, traducción y más.

Se han calculado, para cada par título-subtítulo, las métricas de AS:

1. **Polaridad**: Es un valor flotante entre [-1..1] que indica el tono del texto, es decir, si se trata de un texto, mensaje, opinión, información de carácter positivo o afirmativo (1), o, por el contrario, negativo o controvertido (-1).

2. **Subjetividad**: Indica el nivel de subjetividad u objetividad del texto. Es un valor flotante entre [0..1], donde 0 indica un texto totalmente objetivo, como podría ser el caso de una noticia internacional; y 1 indica un texto totalmente subjetivo, como sería el caso de una pieza de opinión.

Una vez obtenidas estas dos métricas para cada título y cada subtítulo, se ha obtenido la **media** entre el valor de ambos, y se ha guardado como el valor de polaridad y de subjetividad, respectivamente, del artículo/noticia.

En la Ilustración 23 se observa la distribución de estos valores aproximados de polaridad y subjetividad entre los artículos/noticias del conjunto de datos.

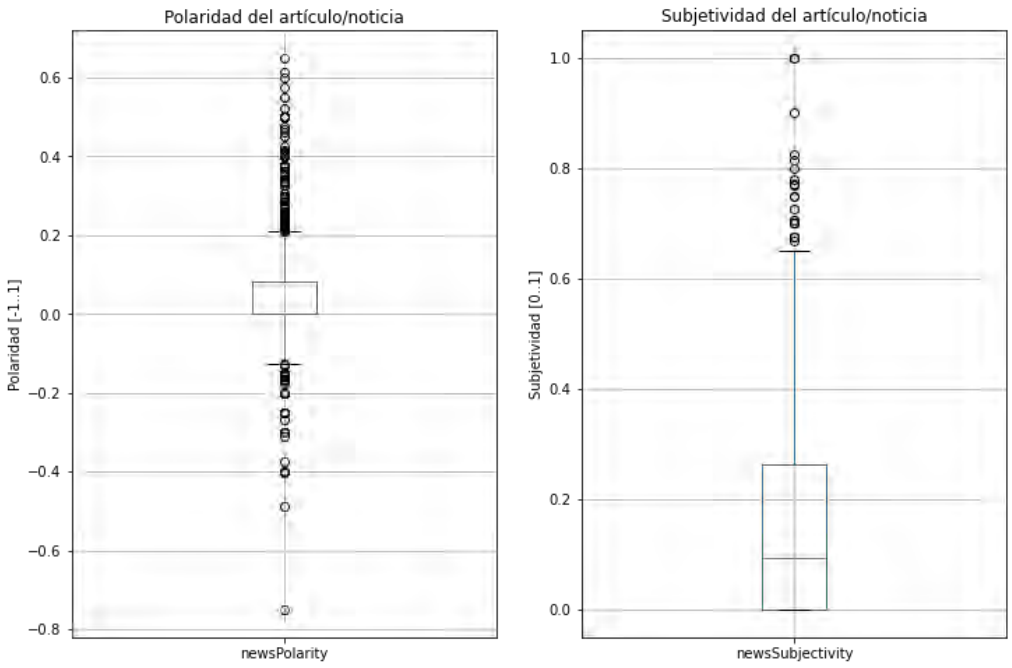

Ilustración 23. Distribución de los valores de polaridad y subjetividad de los artículos/noticias.

Se puede observar que la **polaridad** media de los artículos/noticias tiene un valor ligeramente superior a 0, con lo que, de media, tienen un tono neutro. Por otro lado, hay artículos/noticias muy polarizadoras, con valores que se salen de los percentiles de la distribución. Los valores de **subjetividad** media son muy bajos, cercanos a 0, indicando un tono general bastante objetivo, como debe ser la norma en el contexto de un portal de noticias en línea.

Estas dos métricas se han incluido en los análisis de regresión realizados para intentar predecir los valores finales de los parámetros de rendimiento de los artículos/noticias.

Nota: No se ha incluido el Análisis de Sentimiento del cuerpo del texto de los artículos/noticias, por dos razones:

a. Por el elevado coste temporal y computacional de realizar el AS de más de mil textos que pueden superar los 5000 caracteres de longitud.
b. Por la inutilidad de incluir los resultados de dicho AS en los análisis de clasificación y de regresión realizados en el proyecto, ya que el cuerpo del texto de un artículo/noticia no es visible al visitante hasta que no hace clic sobre el enlace en la página principal. Por lo tanto, no puede tener influencia.

3.5.3 Recategorización manual

Para poder trabajar con las categorías de los artículos/noticias, se ha realizado una recategorización manual de estas, ya que, como se ha explicado anteriormente, las categorías que aparecían en el conjunto de datos original eran demasiado específicas, el número era muy elevado, y ello no permitía un análisis rápido de este atributo.

En concreto, se ha pasado de un total de **341** categorías diferentes a un total de **22 categorías base**, cada una de ellas aglutinando un número de categorías iniciales.

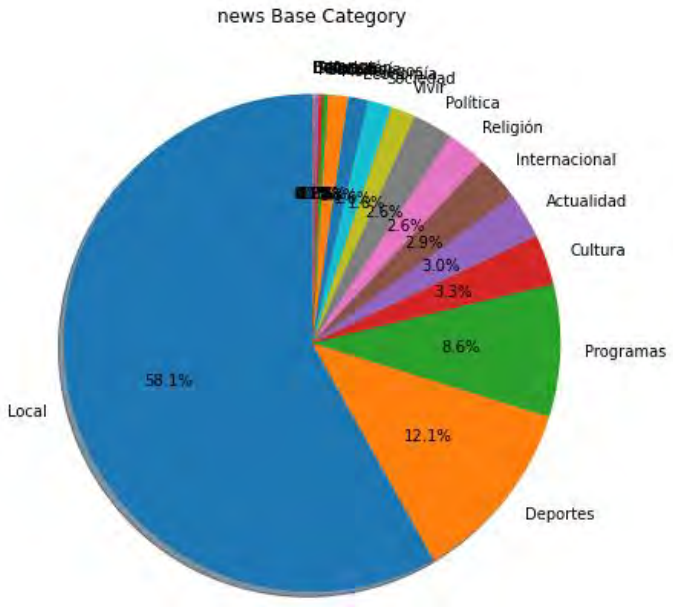

Ilustración 24. Categorías base en el conjunto de artículos/noticias.

En la Ilustración 24 se observa el resultado de la recategorización manual de los artículos/noticias. La mayoría de ellos han sido aglutinados en la nueva categoría "Local". El resto se reparten más o menos heterogéneamente entre el resto de nuevas categorías.

4 Caracterización por modelado estadístico

Este capítulo está completamente dedicado al estudio y revisión de los diferentes modelos estadísticos que han sido preseleccionados como candidatos para ser los que mejor se ajustan a las **curvas temporales de visitas a las páginas** de los artículos/noticias del conjunto de datos.

En la primera sección se hace una introducción sobre los modelos estadísticos en general, y luego se presentan y explican los utilizados en este proyecto.

La siguiente sección se centra en presentar los modelos de difusión de la innovación y por qué y cómo se pueden considerar para su aplicación en el caso de este proyecto.

Los últimos cuatro apartados tratan sobre cuatro modelos estadísticos que pertenecen al grupo mencionado y explicado en el apartado anterior. En cada uno de ellos, para cada uno de ellos, hay una explicación, una reinterpretación y una aplicación al caso de este proyecto.

4.1 Introducción

Como se mencionó en capítulos y secciones anteriores, hay una preselección de cinco modelos estadísticos candidatos. Hay otros modelos estadísticos que también podrían haber sido preseleccionados. El objetivo es encontrar el modelo que mejor ajusta las curvas de tiempo de visitas a las páginas de los artículos/noticias en línea que se encuentran en el conjunto de datos. Los mejores ajustes de los modelos a las curvas de visitas a página-tiempo se encontraron utilizando las aplicaciones y funciones proporcionadas por la biblioteca de Python **SciPy**. SciPy es una biblioteca que tiene paquetes utilizados para ciencia de datos, optimización, estadísticas y otros. Una de las funciones del paquete de estadísticas de SciPy realiza un ajuste de curva utilizando el método de mínimos cuadrados no lineales, que se explica brevemente en el primer capítulo. Esa función es la que se utiliza para ajustar las curvas a los modelos estadísticos. La biblioteca Scipy, junto con otras bibliotecas de Python que se muestran en la Ilustración 25, permitió un cálculo rápido y una visualización fácil.

Ilustración 25. Algunas librerías y paquetes de Python utilizados. Fuente: https://medium.com/

4.2 Modelos estadísticos

Como ya se ha mencionado, en esta sección, se presentan y explican los diferentes modelos estadísticos considerados para el ajuste de la curva temporal de visitas a las páginas de los artículos/noticias en línea.

4.2.1 Modelo de Barabási

El primer modelo estadístico que se presenta y se aplica es el Modelo de Barabási, llamado así por el apellido de uno de los autores de dos publicaciones de referencia [1] [3] , en las que se presenta este modelo y se demuestra su utilidad para el ajuste de la curva de la evolución temporal de las citas obtenidas por un artículo científico a lo largo de los años, y de las ventas semanales de los libros superventas, respectivamente.

En el artículo científico original [1] , los autores identificaron tres factores centrales que impulsan el historial de citas de artículos individuales:

1. El concepto de **conexión preferencial** captura el hecho bien documentado de que los artículos muy citados son más visibles y es más probable que se vuelvan a citar que las contribuciones menos citadas. En consecuencia, la probabilidad de que un artículo i sea citado nuevamente es proporcional al número total de citas c_i que el artículo recibió anteriormente.

2. El concepto de **envejecimiento o longevidad** captura el hecho de que las nuevas ideas se integran en el trabajo posterior; por tanto, la novedad de cada artículo se desvanece con el tiempo. La desintegración a largo plazo resultante se describe mejor mediante una probabilidad de supervivencia logarítmica normal, donde t es el tiempo; m indica inmediatez, que rige el tiempo que tarda un artículo en alcanzar su punto máximo de citas; y σ es la longevidad, capturando la tasa de desintegración.

$$P_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t}\exp\left[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}\right]$$

3. El concepto de **aptitud o adecuación**, η_i, captura las diferencias inherentes entre los artículos, lo que explica la novedad percibida y la importancia de un descubrimiento. La novedad y la importancia dependen de tantas dimensiones subjetivas e intangibles que es imposible cuantificarlas objetivamente todas. Entonces, la aptitud η_i es una medida colectiva que captura la respuesta de la comunidad a un trabajo.

Luego, los autores combinaron estos tres factores, realizaron un procesamiento matemático avanzado y crearon la ecuación modelo que define el número acumulado de citas adquiridas c por el artículo i en el momento t:

$$c_i^t = m\left[e^{\lambda_i \phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right]$$

dónde:

- $\phi(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$ es la distribución normal acumulada.
- m es el número medio de referencias o citas que contiene un artículo nuevo. Los autores descubrieron que este parámetro no afecta los resultados y decidieron establecerlo en m = 30.
- λ_i es la adecuación relativa que captura la importancia de un artículo en comparación con otros artículos.
- μ_i es el factor de inmediatez mencionado anteriormente de un artículo.
- σ_i es el factor de longevidad de un artículo mencionado anteriormente, que captura su tasa de deterioro en las citas.

Esta ecuación es la formulación matemática del modelo de Barabási, y es la que se utiliza en este proyecto, aunque se necesita una reinterpretación de los parámetros de este modelo para el caso de este proyecto.

Es necesario reinterpretar los tres conceptos sobre los que se basa el modelo Barabási, porque el caso de las visitas a página recibidas por los artículos de noticias en línea es diferente al caso de la evolución de las ventas de los libros más vendidos, o el caso del historial de citas. de artículos científicos.

1. El factor de aptitud o adecuación del artículo/noticia

   Cada artículo de noticias publicado en el periódico digital tiene un valor diferente para su audiencia. Este factor representa lo atractivo e interesante que es un artículo de noticias en línea para el público que navega por las páginas del periódico digital. Se aproxima a la capacidad de un artículo de noticias publicado para responder a los intereses de un amplio número de lectores. Esta capacidad está marcada principalmente por la propia noticia, el contenido y su momento. Otros factores importantes son el titular, la situación en la página de inicio, el/los autor/es, etc.

2. El factor de conexión preferencial de los artículos/noticias

   Dentro de un portal de noticias en línea, un artículo de noticias en el que se hace clic masivamente se coloca generalmente en un espacio destacado de la página de inicio, por ejemplo, colocándolo en la parte superior. También se puede incluir en una o más listas más leídas. Estas dos cosas harán que el artículo de noticias sea aún más visible y, por lo tanto, atraerá aún más visitas a página. Esto tiene que ver con el concepto de apego preferencial explicado anteriormente. El apego preferencial en este contexto también se basa probablemente en efectos colectivos, como recomendaciones de amigos, periodistas, celebridades y reseñas en línea. Matemáticamente, implica que la probabilidad de visitar la página de un artículo de noticias depende de sus visitas a página actualizadas.

3. El factor de envejecimiento o longevidad del artículo/noticia

   Este factor, en el caso de una noticia de reciente publicación, representa su atractivo, es decir, su lugar entre las noticias interesantes de actualidad. Inevitablemente, cualquier artículo de noticias perderá gradualmente su lugar en favor de artículos de noticias más recientes y / o interesantes. Esto significa que incluso los mejores artículos de noticias pierden su novedad y se desvanecen del ojo público algún tiempo después de su fecha de publicación.

La ecuación del modelo que define el total acumulado de visitas a página que tiene el artículo de noticias en línea i en un momento dado t es la misma que la de los dos artículos de referencia:

$$pv_i^t = m\left[e^{\lambda_i \phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right]$$

El significado de los parámetros de aptitud relativa (λ*i*), inmediatez (μ*i*) y longevidad (σ*i*) del artículo de noticias i es aproximadamente el mismo que para los casos de los dos artículos de referencia.

El parámetro *m* para el caso de artículos científicos representa el número promedio de referencias a otros artículos (citas) que contiene un artículo nuevo. En este caso, no está claro qué representa, pero dado que los autores de los artículos descubrieron que el valor del parámetro *m* no afecta directamente los resultados del ajuste del modelo, se establece en *m* = 30.

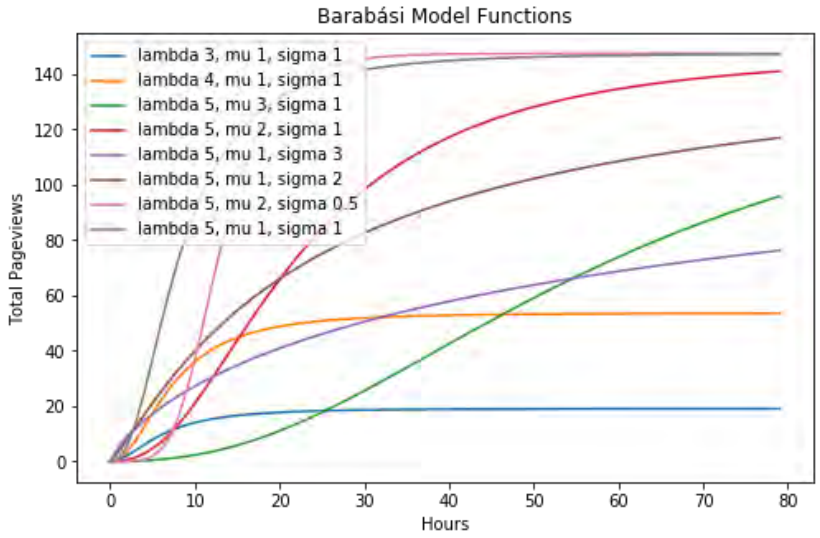

Ilustración 26. Ejemplos de curvas del modelo de Barabási.

4.2.1.1 Ajuste del modelo a la curva PVs-tiempo

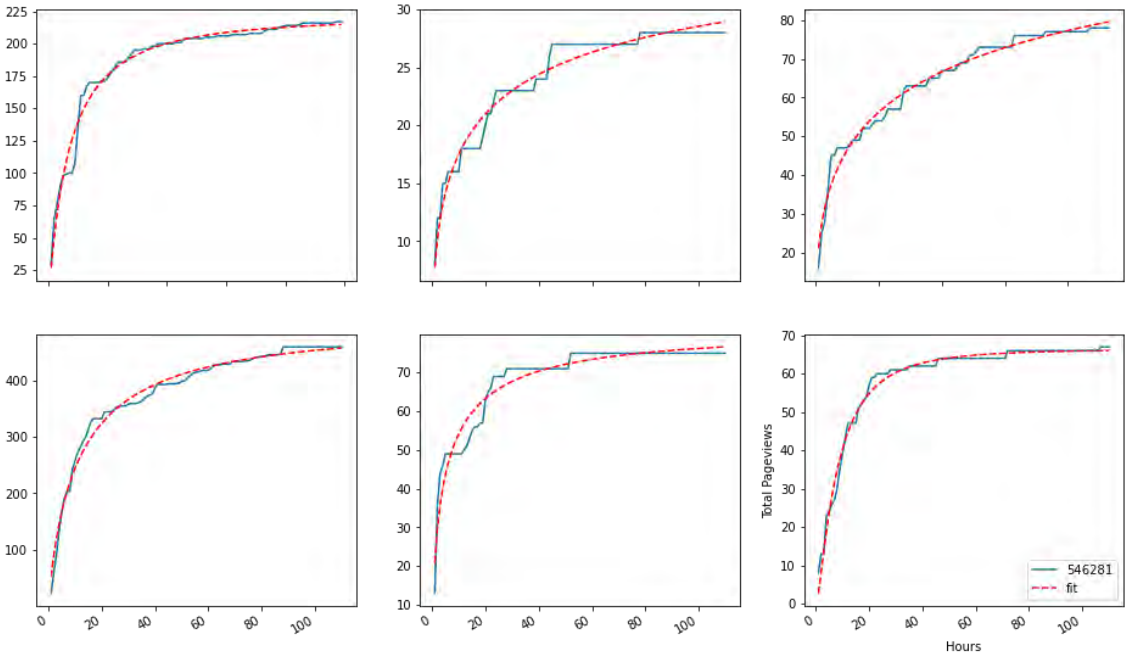

Ilustración 27. Ajuste del modelo de Barabási a la curva de visitas a página de 6 artículos/noticias seleccionados al azar. Azul espeso = real. Rojo punteado = ajuste.

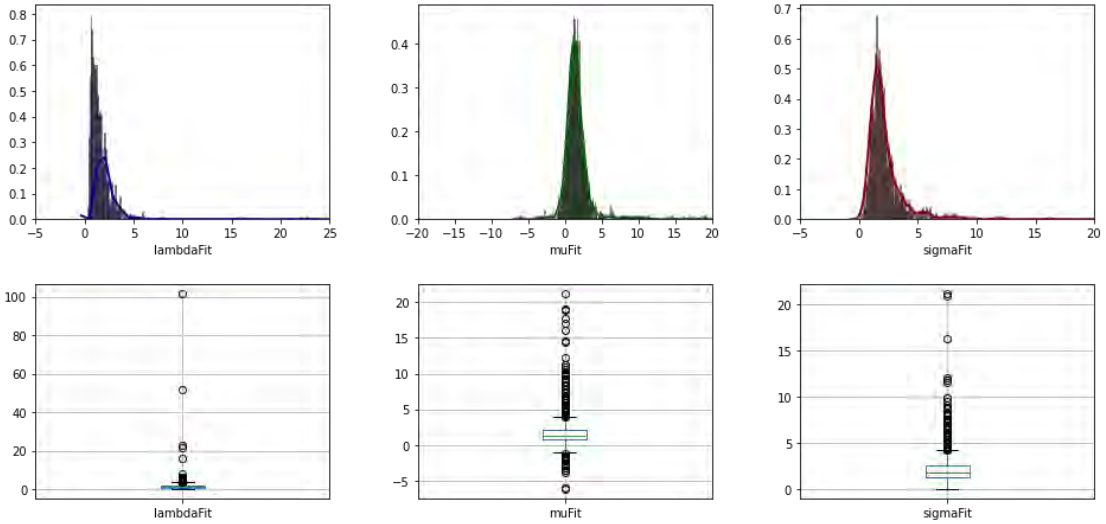

Ilustración 28. Distribución y diagramas de caja de los valores de los tres parámetros del Modelo de Barabási.

4.2.2 Modelo exponencial

Se propone un nuevo modelo de base exponencial para trabajar con las funciones de tiempo de visitas a página de los artículos de noticias en línea. Al igual que el modelo anterior, se basa en tres parámetros principales:

1. El factor de amplitud máxima

Este parámetro A está relacionado con la amplitud de las visitas a página del artículo de noticias.

2. El factor de crecimiento

Este parámetro α representa el factor de crecimiento de las visitas a página en el tiempo. Gobierna la primera parte ascendente de la curva de tiempo de visitas a la página.

3. El factor de descomposición

Este parámetro β representa la disminución gradual de las visitas a página en el tiempo. Gobierna la segunda parte descendente de la curva de visitas a página-tiempo.

La función general que define el total de visitas a página acumuladas de un artículo de noticias en un momento dado es:

$$\mathrm{pv}_i(\mathrm{t}) = \frac{100A_i}{\alpha_i - \beta_i}\left[\frac{1}{\alpha_i}(e^{-\alpha_i t} - 1) - \frac{1}{\beta_i}\left(e^{-\beta_i t} - 1\right)\right]$$

Si $t \rightarrow \infty$, el total de visitas a página finales que el artículo/noticia en línea que obtendría, de acuerdo con este modelo exponencial, se define por:

$$\mathrm{pvT}_i = \frac{100A_i}{\alpha_i - \beta_i}\left[\frac{1}{\beta_i} - \frac{1}{\alpha_i}\right]$$

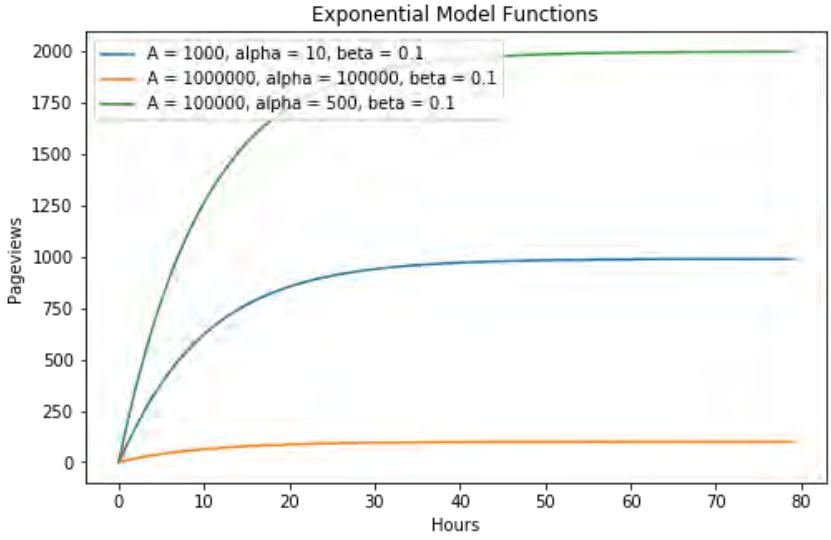

Ilustración 29. Ejemplos de curvas del modelo exponencial.

4.2.2.1 Ajuste del modelo a la curva PVs-tiempo

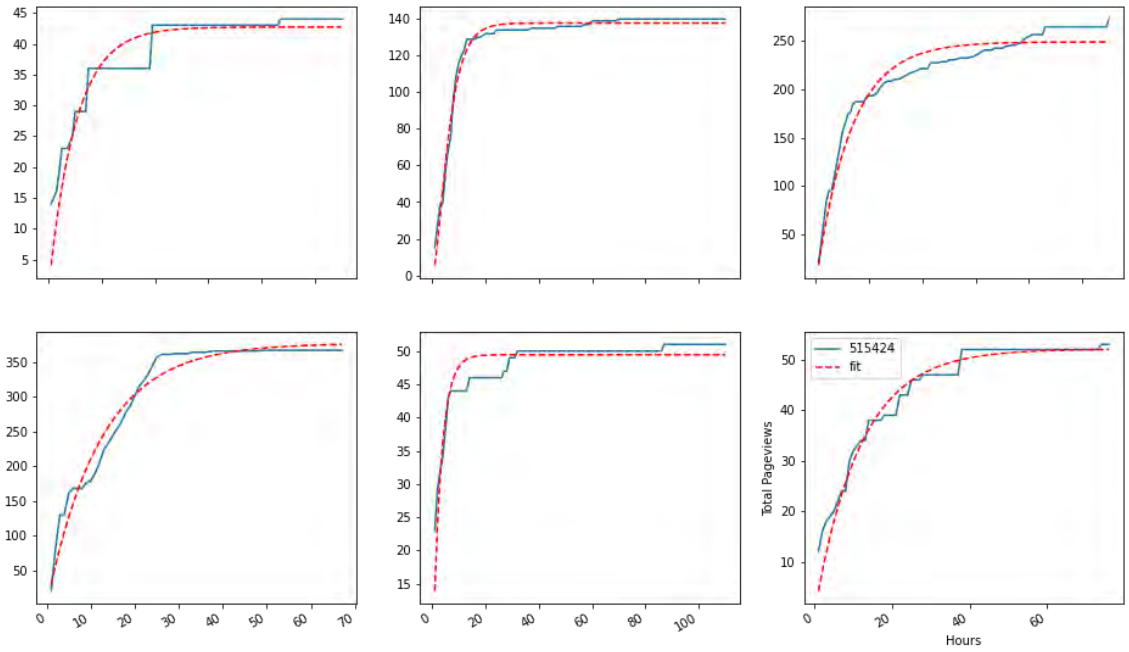

Ilustración 30. Ajuste del modelo exponencial a la curva de visitas a página de 6 artículos/noticias seleccionados al azar. Azul espeso = real. Rojo punteado = ajuste.

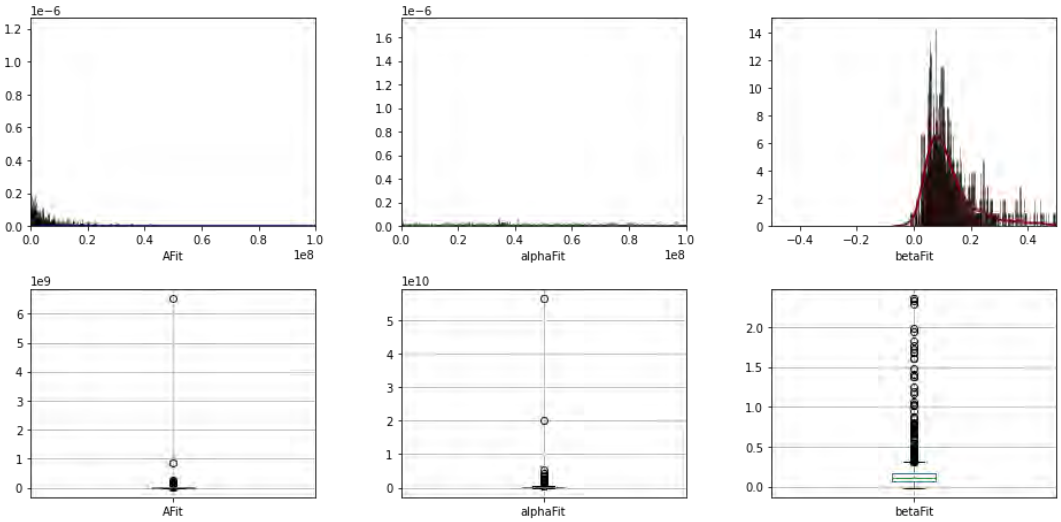

Ilustración 31. Distribución y diagramas de caja de los valores de los tres parámetros del Modelo exponencial.

4.2.3 Modelo de Bass

Este es el primer modelo que se utiliza en una rama de la estadística y la sociología llamada **Teoría de Difusión de Innovaciones**. Por ello, antes de presentar el modelo de Bass y los resultados de su aplicación, se realiza una breve introducción a la teoría.

La teoría de la difusión de innovaciones tiene como objetivo explicar la adopción de nuevas ideas, productos y tecnologías por parte del público.

El concepto de difusión de innovaciones y la teoría que lo sustenta fueron propuestos y popularizados por Everett Rogers con su libro "Diffusion of Innovations", publicado en 1962 [39] . Rogers sostiene que la difusión es el proceso por el cual una innovación se comunica a lo largo del tiempo entre los participantes de un sistema social.

Rogers propone que cuatro elementos principales influyen en la difusión de una nueva idea: la innovación en sí, los canales de comunicación, el tiempo y un sistema social. Este proceso depende en gran medida del capital humano. La innovación debe adoptarse ampliamente para ser autosuficiente. Dentro de la tasa de adopción, hay un punto en el que una innovación alcanza una masa crítica.

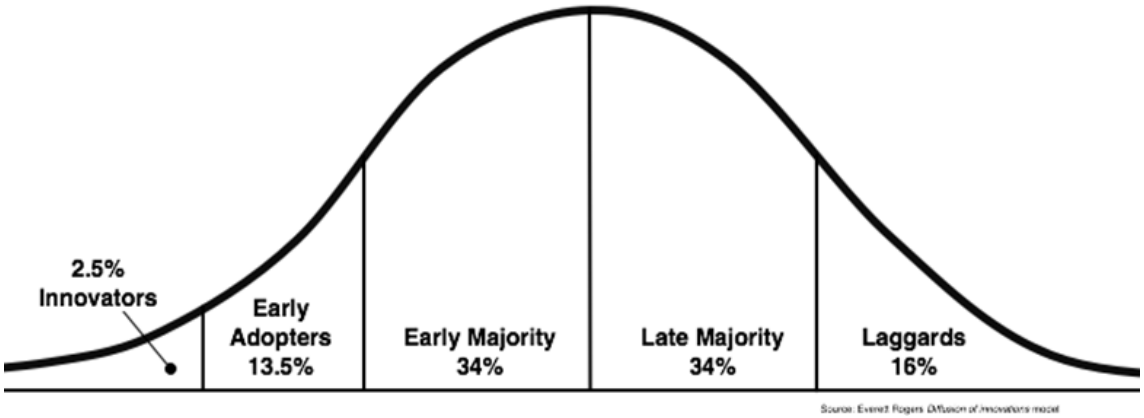

Ilustración 32. Difusión de innovaciones. Fuente: http://blog.leanmonitor.com/early-adopters-allies-launching-product/

Aunque el enfoque de esta teoría es determinar el éxito y el fracaso de un producto, los modelos estadísticos a menudo predicen curvas en S cercanas a la logística y que son como las curvas de tiempo de visitas a páginas de los artículos de noticias en línea. Por lo tanto, en el contexto de este proyecto, una noticia nueva publicada recientemente en el sitio web de un periódico digital es un nuevo producto que espera ser adoptado (cliqueado) por el público en general que visita el sitio web. Por lo tanto, en las siguientes secciones se explora la posibilidad de utilizar curvas S de difusión para describir el historial de visitas a página de los artículos de noticias en línea.

Uno de los modelos más famosos en las ciencias de la gestión y el marketing es el modelo de Bass, que describe el proceso de adopción de nuevos productos por parte de las poblaciones masivas. El modelo Bass asume que quienes adoptan un producto están influenciados por dos aspectos: los medios de comunicación y el boca a boca.

Fue desarrollado por Frank Bass en un artículo suyo [18] . Bass fue influenciado por la teoría de la difusión de la innovación de Rogers y le aportó algunas ideas matemáticas.

Por tanto, los compradores comprenden dos grupos. Un grupo, los innovadores como los acuñó Bass, está influenciado solo por los medios de comunicación, mientras que el otro grupo, los imitadores, está influenciado por otros (efecto boca a boca). Tales suposiciones son razonables en el contexto de los artículos de noticias en línea. Los

innovadores corresponden a personas que hacen clic en (y presumiblemente leen) la noticia de forma espontánea, poco influenciados por la cantidad de personas que ya visitaron su página, dejaron un comentario o lo compartieron. Al mismo tiempo, las visitas a página de un artículo de noticias en línea son impulsadas por la difusión de boca en boca (los imitadores).

La función general que define el total de visitas a página acumuladas, *pv*, de un artículo de noticias *i* en un momento dado *t* es:

$$pv_i^t = pv_i^\infty \frac{1 - e^{-(p_i+q_i)t}}{1 + \frac{q_i}{p_i} e^{-(p_i+q_i)t}}$$

donde p_i caracteriza a los "innovadores", reflejando una influencia que es independiente de las visitas a página acumuladas en el momento dado (pv_i), y q_i refleja la parte del modelo correspondiente a los "imitadores".

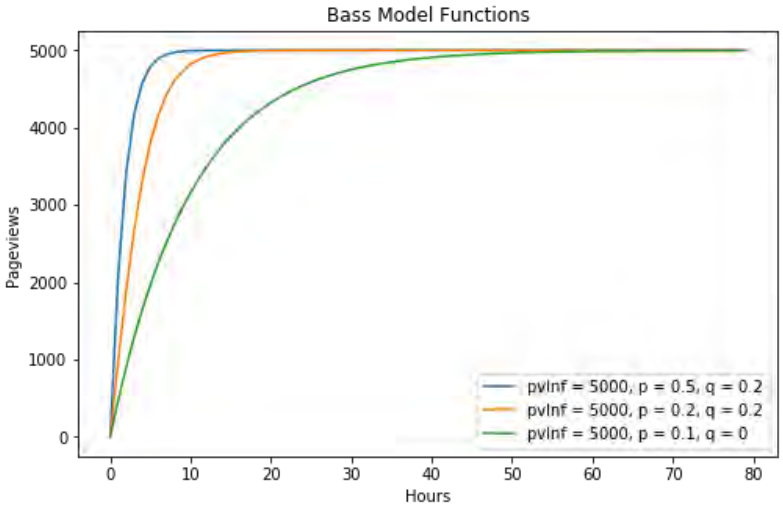

Ilustración 33. Ejemplos de curvas del modelo de Bass.

4.2.3.1 Ajuste del modelo a la curva PVs-tiempo

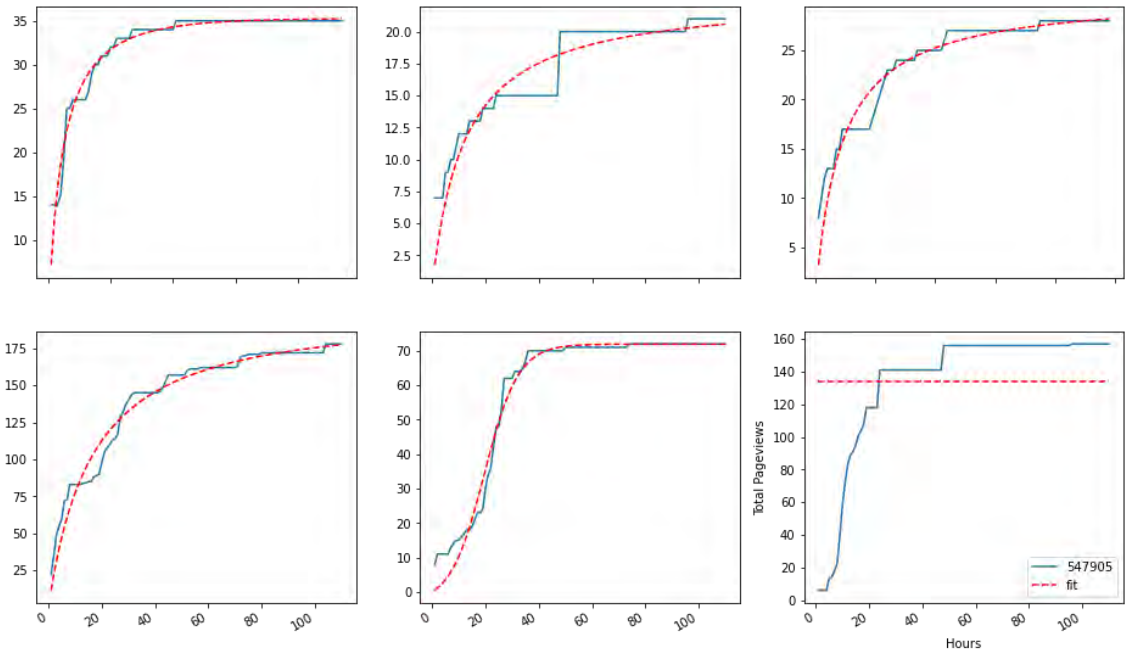

Ilustración 34. Ajuste del modelo de Bass a la curva de visitas a página de 6 artículos/noticias seleccionados al azar. Azul espeso = real. Rojo punteado = ajuste.

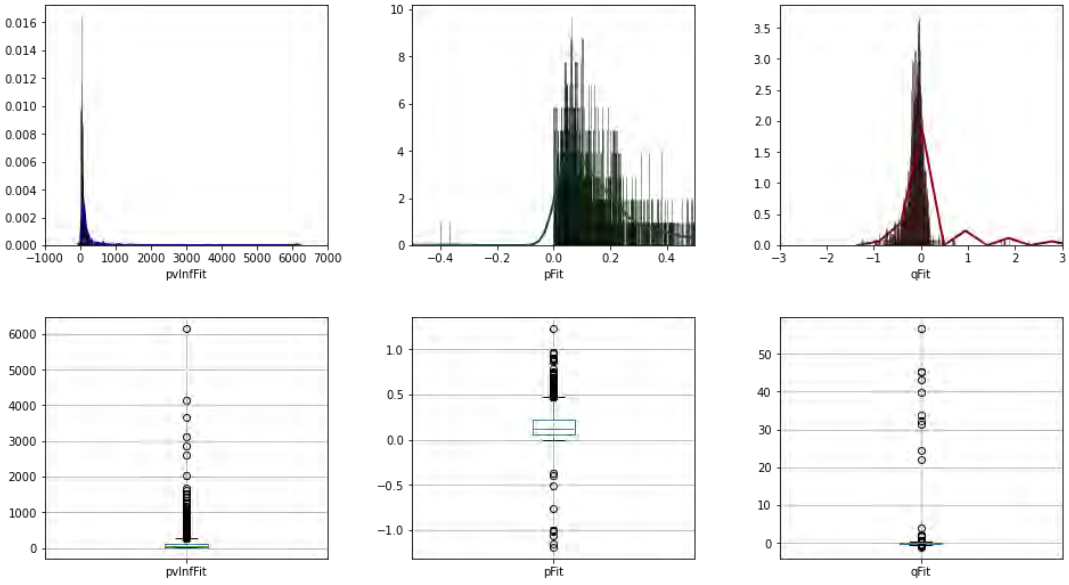

Ilustración 35. Distribución y diagramas de caja de los valores de los tres parámetros del Modelo de Bass.

4.2.4 Modelo logístico

Una curva o función logística es una curva sigmoide (S) muy común, con la siguiente ecuación matemática:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

donde:

- e = la base del logaritmo natural (también conocido como número de Euler).
- x_0 = el valor x del punto medio del sigmoide.
- L = valor máximo de la curva.
- k = tasa de crecimiento logístico o pendiente de la curva.

La función logística se usa ampliamente para modelar el crecimiento de la población y la adopción de productos, con aplicaciones en muchos campos como la estadística. Por ejemplo, la función logística es la función de distribución acumulativa de la familia logística de distribuciones, y se utiliza para modelar las posibilidades de que un jugador de ajedrez derrote a su oponente en el sistema Elo.

En el contexto de las visitas a página, uno podría ver un artículo de noticias en línea como un producto nuevo, cuya adopción conduce a un aumento en las visitas a página. Cada artículo de noticias se caracteriza por una tasa de aumento diferente r y un número total de visitas a página c_i^{∞} que captura las diferencias en el impacto, si las visitas a página se toman como el único o el factor más importante que determina el impacto final de un artículo de noticias. Con el tiempo, el atractivo y el atractivo de un artículo de noticias se desvanece, ya que la información contenida en él ha llegado a todos los posibles usuarios, de ahí que las visitas a página del artículo de noticias se acerquen a c_i^t.

Entonces, la ecuación del modelo logístico que define el total de visitas a página acumuladas (pv_i^t) por el artículo de noticias i en el momento t es:

$$pv_i^t = \frac{pv_i^{\infty}}{1 + e^{-r_i(t-\tau_i)}}$$

donde c_i^{∞}, r_i y tau_i corresponden, respectivamente, a las visitas a página finales, la longevidad y la inmediatez del artículo/noticia i.

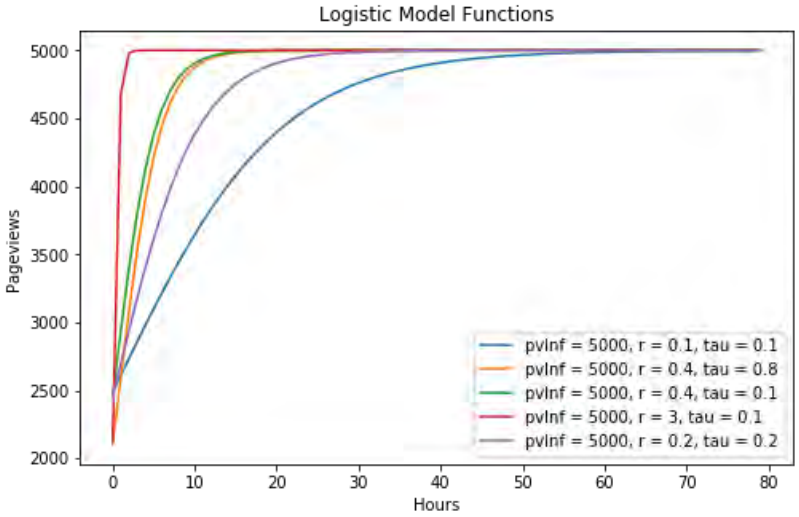

Ilustración 36. Ejemplos de curvas del modelo logístico.

4.2.4.1 Ajuste del modelo a la curva PVs-tiempo

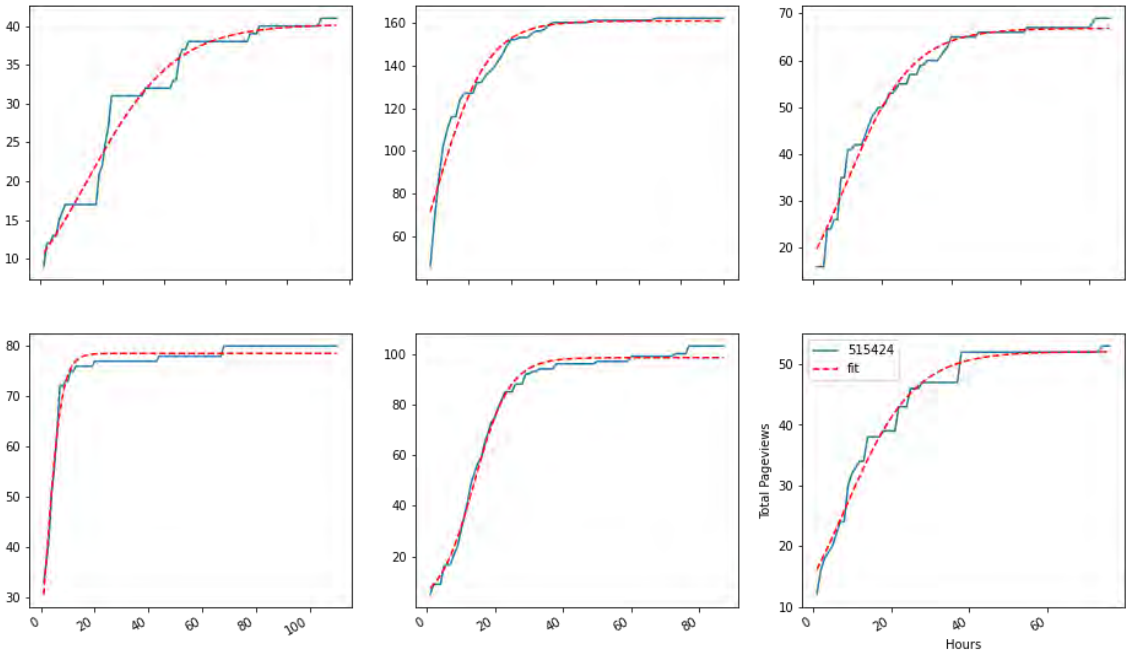

Ilustración 37. Ajuste del modelo logístico a la curva de visitas a página de 6 artículos/noticias seleccionados al azar. Azul espeso = real. Rojo punteado = ajuste.

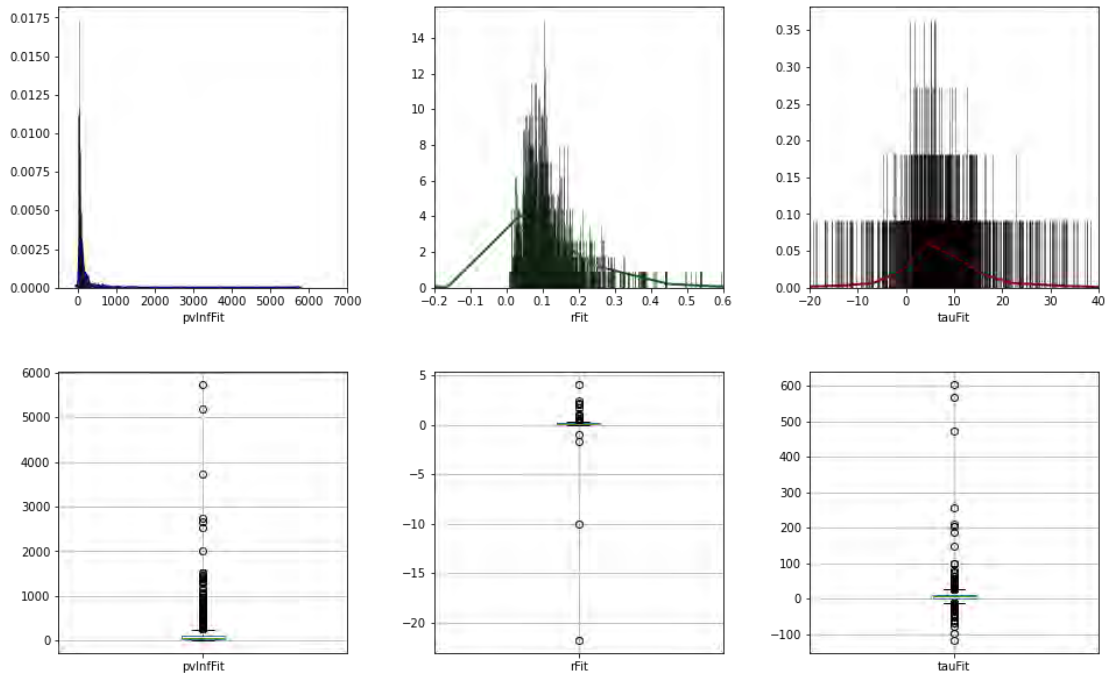

Ilustración 38. Distribución y diagramas de caja de los valores de los tres parámetros del Modelo logístico.

4.2.5 Modelo de Gompertz

El modelo de Gompertz, que lleva el nombre de Benjamin Gompertz, se propuso por primera vez en 1825 para modelar la mortalidad humana. Es una función sigmoidea que describe el crecimiento como más lento al comienzo y al final de un período de tiempo determinado. El modelo genera una curva de difusión sesgada con colas largas.

La función fue diseñada originalmente para describir la mortalidad humana, pero desde entonces ha sido modificada para ser aplicada en biología, para detallar poblaciones y muchos otros campos científicos.

En el contexto de los artículos de noticias en línea, las primeras páginas vistas allanan el camino para nuevas páginas vistas e impulsan la dinámica de páginas vistas, por lo que la tasa de publicaciones se desarrolla y aumenta a un ritmo exponencial. Esto se puede formular como:

$$pv_i^t = pv_i^\infty e^{-e^{-(a+qt)}}$$

donde a establece el desplazamiento en c_i^t mientras que q caracteriza la tasa de crecimiento de las visitas a página.

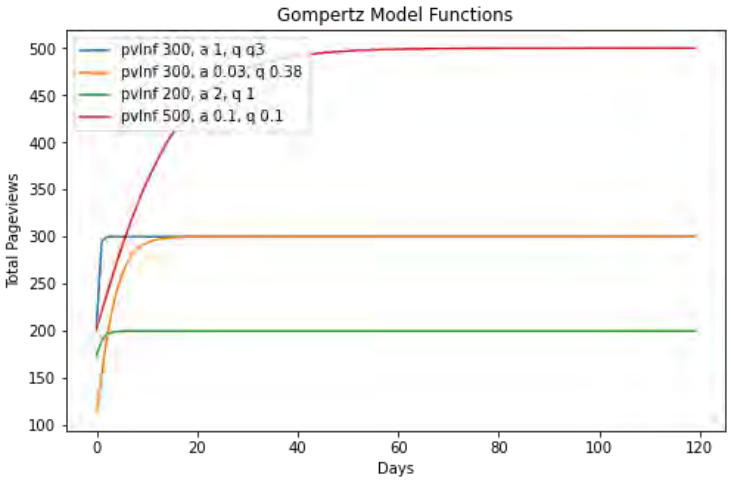

Ilustración 39. Ejemplos de curvas del modelo de Gompertz.

4.2.5.1 Ajuste del modelo a la curva PVs-tiempo

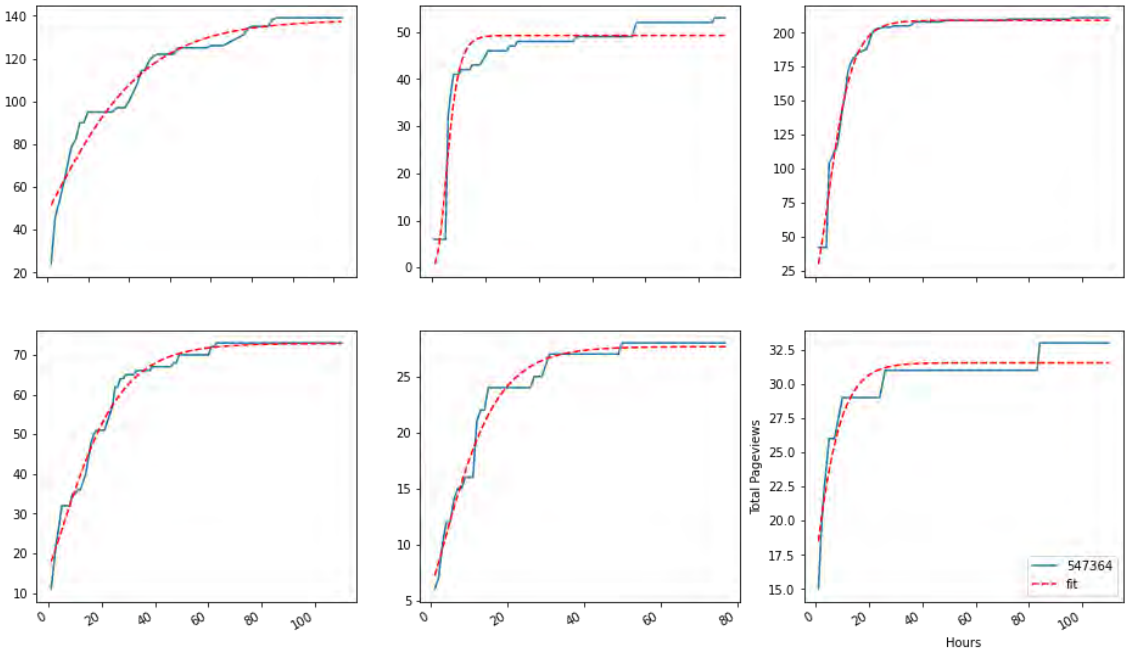

Ilustración 40. Ajuste del modelo de Gompertz a la curva de visitas a página de 6 artículos/noticias seleccionados al azar. Azul espeso = real. Rojo punteado = ajuste.

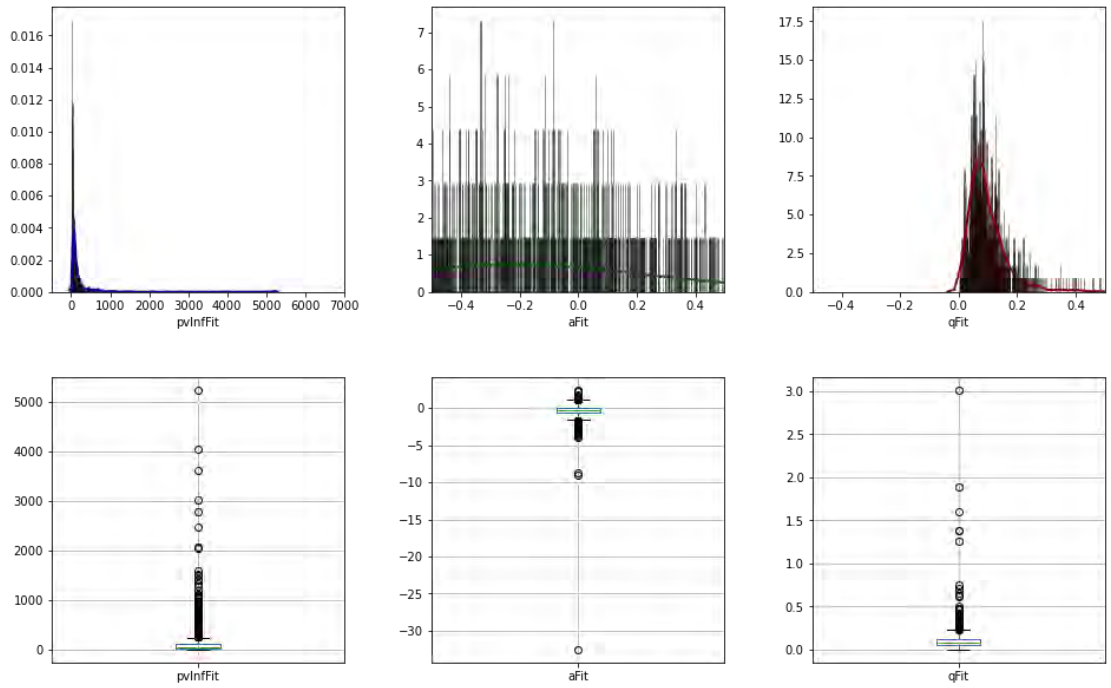

Ilustración 41. Distribución y diagramas de caja de los valores de los tres parámetros del Modelo de Gompertz.

4.3 Comparativa de modelos

Primero, debe tenerse en cuenta que los modelos estadísticos aplicados y comparados en este proyecto son solo cinco de un gran número de posibles candidatos de modelos estadísticos que podrían probarse para el ajuste de las curvas de visitas a página-tiempo de los artículos/noticias en línea.

Se quiere comparar los resultados de los ajustes de los cinco modelos probados. Para ello, se han seleccionado dos métricas que se pueden interpretar como medidas de **bondad del ajuste** a la curva. La bondad del ajuste es un criterio de selección de modelos estadísticos. El hecho de que los cinco modelos estadísticos probados tengan todos tres parámetros facilita la comparativa ente ellos.

4.3.1 Coeficiente de determinación (R^2)

El coeficiente de determinación (R^2) es una estadística ampliamente utilizada para evaluar la bondad de ajuste y/o el poder predictivo de un modelo estadístico. Es la proporción de la varianza en la variable dependiente que es predecible a partir de la(s) variable(s) independiente(s).

Matemáticamente, se define como:

$$R^2 = 1 - \frac{SSE}{SST}$$

donde:

- $SSE = \sum_i (y_i - f_i)^2$
- $SST = \sum_i (y_i - \hat{y})^2$, $\hat{y}$ siendo la media de los valores observados.

La salida suele ser un porcentaje de 0 a 1, que indica el porcentaje de desviación de los datos que podría explicarse o contabilizarse mediante el modelo estadístico. Sin embargo, cuando se utiliza un modelo de ajuste no lineal, como es el caso de este proyecto, R^2 puede producir valores negativos o valores superiores a 1. R^2 tiene muchos inconvenientes y deficiencias como criterio de selección del modelo.

Cuando se trabaja con modelos de ajuste no lineales como los de este proyecto, R^2 debe interpretarse con cuidado. Sin embargo, en general, R^2 es una estadística interesante de medir debido a su prominencia y simplicidad.

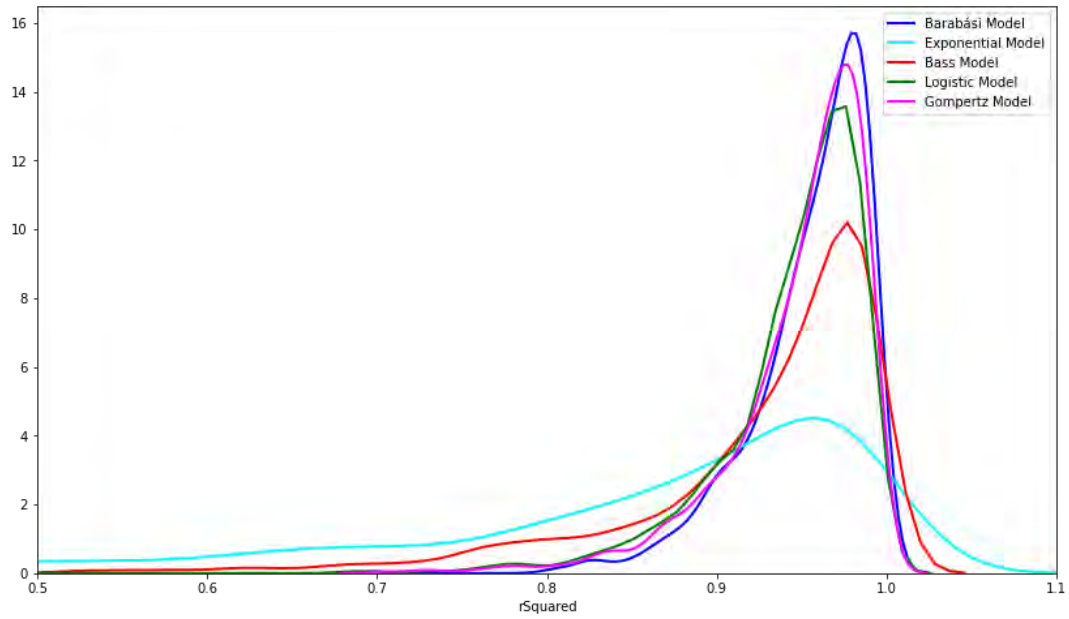

Ilustración 42. Distribución del coeficiente de determinación para los cinco modelos estadísticos probados.

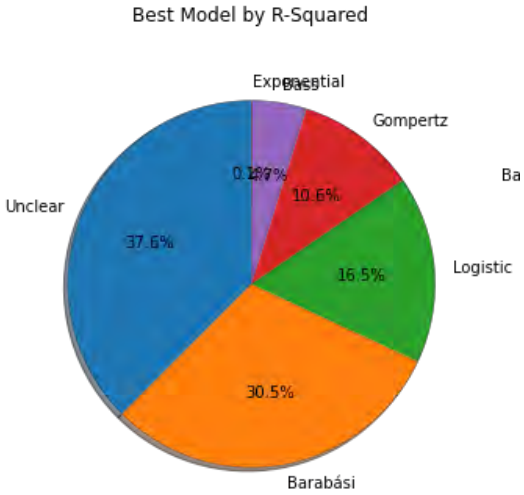

Ilustración 43. Repartición del Mejor modelo según coeficiente de determinación.

En la Ilustración 42, se observa que todos los modelos tienen unos valores medios de R^2 muy altos en sus ajustes a las curvas de *pv*-tiempo del conjunto de datos.

Además, se ha realizado un análisis adicional para seleccionar exactamente el mejor modelo estadístico para la curva de visitas a página de cada artículo/noticia. Concretamente, se ha seleccionado, para cada una de ellas, el modelo con cuyo ajuste se ha obtenido **un valor de R^2 un 1% superior a la media** de los valores de R^2 obtenidos con los ajustes de todos los modelos.

En la Ilustración 43 se observan los resultados. La gran mayoría de curvas no son ajustadas mucho mejor con un modelo estadístico de manera clara.

4.3.2 KS-test ponderado

El test de Kolmogorov-Smirnov (KS-test) es una estadística muy utilizada para comparar una función de distribución acumulativa empírica (ECDF) con una función de distribución acumulativa de referencia (CDF).

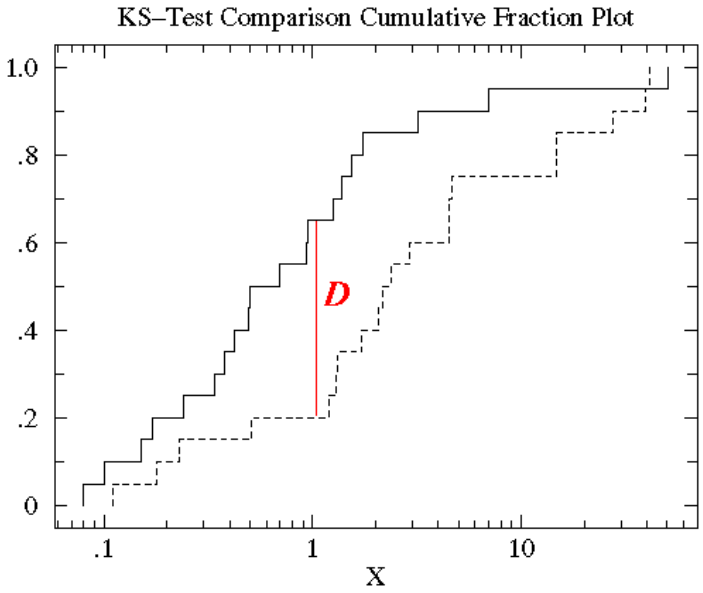

Ilustración 44. Ejemplo de un KS-test.

El KS-test se puede modificar para que sirva como una prueba de bondad de ajuste, y esto es lo que se hace en este proyecto.

Para cuantificar qué tan bien se ajusta cada uno de los cinco modelos a los datos reales, para cada artículo de noticias *i*, se obtuvo la siguiente medida KS ponderada:

$$D_i = \max_{t\in[0,T]} \frac{|pv_i^t - f_i^t|}{\sqrt{(1+pv_i^t)(pv_i^T - pv_i^t + 1)}}$$

donde:

- f_i^t representa las visitas a página computadas por el modelo.
- pv_i^T representa el total de visitas a página acumuladas en el periodo de tiempo $t \in [0, T]$.

Como en el KS-test original, una D_i más pequeña implica un mejor ajuste. En este caso, significa un mejor ajuste del modelo a la curva de visitas a página-tiempo del artículo de noticias *i*.

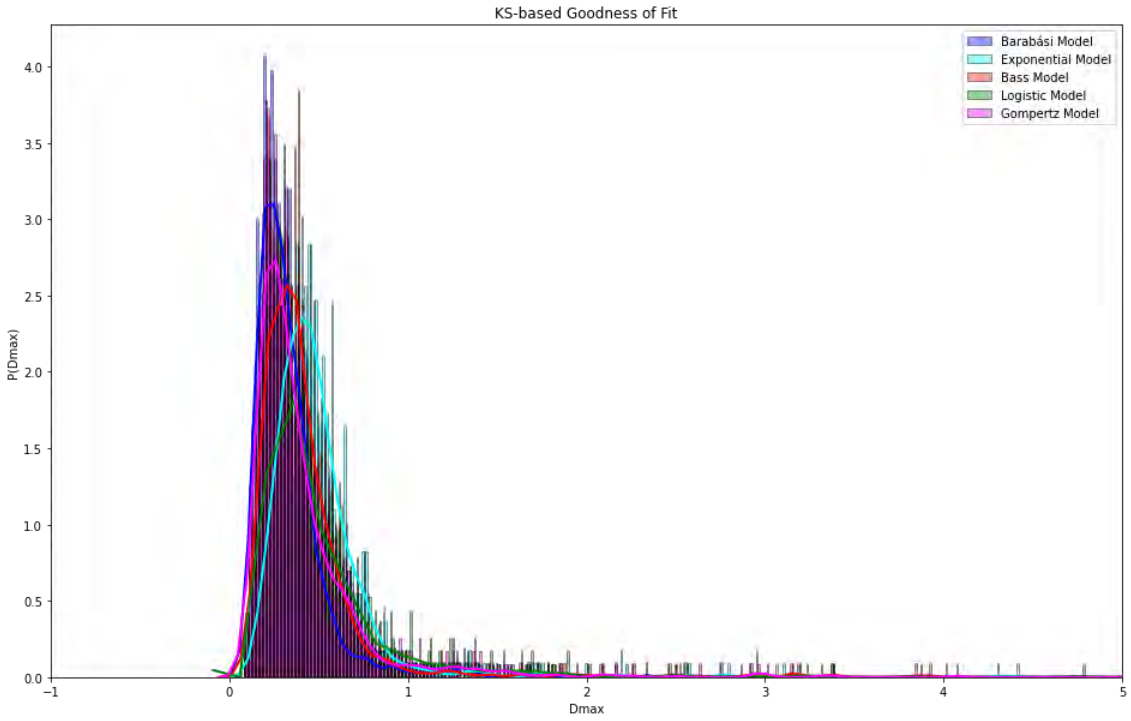

Ilustración 45. Medida de KS-Test ponderada para todos los modelos.

Los resultados obtenidos y mostrados en la Ilustración 45. Medida de KS-Test ponderada para todos los modelos. indican que todos los modelos funcionan, pero en diversos grados. Las diferencias en la *Dmax* media son pequeñas entre los modelos.

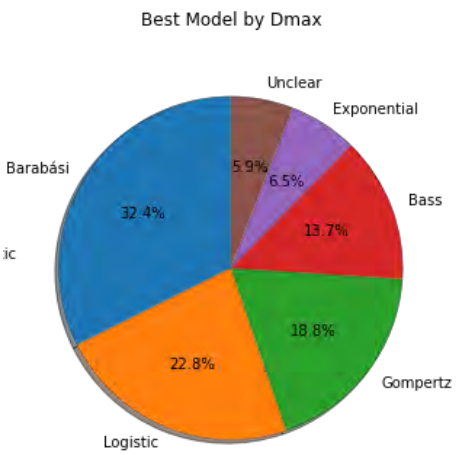

Ilustración 46. Repartición del Mejor modelo según medida de KS-test ponderado (Dmax).

Por otro lado, se ha realizado un análisis adicional para seleccionar exactamente el mejor modelo estadístico para la curva de visitas a página de cada artículo/noticia. Concretamente, se ha seleccionado, para cada una de ellas, el modelo con cuyo ajuste se ha obtenido **un valor de *Dmax* un 10% inferior a la media** de los valores de *Dmax* obtenidos con los ajustes de todos los modelos.

En la Ilustración 46. Repartición del Mejor modelo según medida de KS-test ponderado (Dmax). se observan los resultados. La gran mayoría de curvas no son ajustadas mucho mejor con un modelo estadístico de manera clara.

4.4 Resultados y observaciones

Antes que nada, se considera que todos los modelos estadísticos probados arrojan unos resultados suficientemente satisfactorios cuando se aplican al ajuste de la curva temporal de visitas a las páginas de los artículos/noticias en línea.

Sin embargo, el objetivo de este estudio era encontrar diferencias entre los artículos/noticias cuyas páginas reciben visitas según un modelo o según otro.

Si se hace una unión (∪) entre los resultados de mejor modelo según el coeficiente de determinación (R^2) y los de mejor modelo según la medida de KS-test ponderado (*Dmax*), presentados y explicados en las dos secciones previas, se obtienen los resultados que se observan en la Ilustración 47. Repartición del mejor modelo en conjunto..

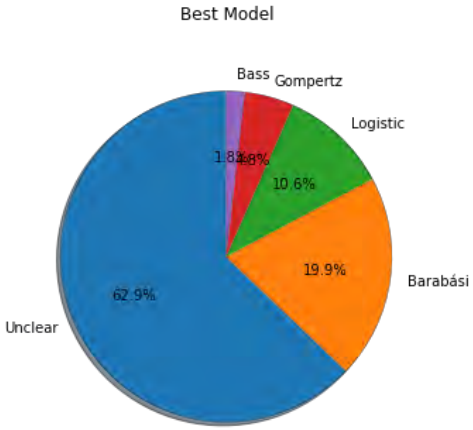

Ilustración 47. Repartición del mejor modelo en conjunto.

4.4.1 Según parámetros de rendimiento

Se ha estudiado la relación entre el tipo de modelo estadístico que mejor ajusta la evolución temporal de las visitas a las páginas de los artículos/noticias del portal web y los valores finales de los dos parámetros de rendimiento más relevantes, como son las visitas a página y el tiempo medio en página. Los resultados se observan en la Ilustración 48. Distribución de las visitas a página y del tiempo medio en página según el modelo estadístico que mejor ajusta la curva de pv-tiempo..

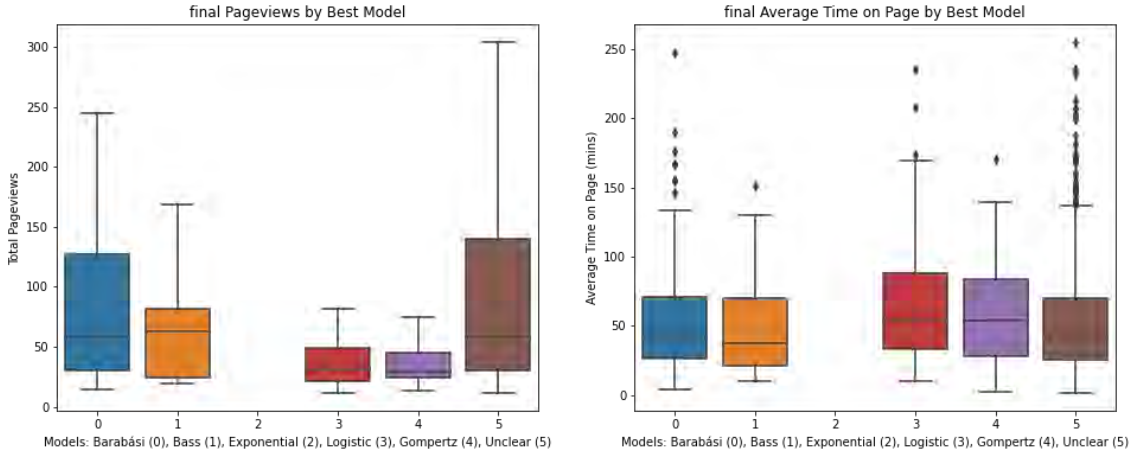

Ilustración 48. Distribución de las visitas a página y del tiempo medio en página según el modelo estadístico que mejor ajusta la curva de pv-tiempo.

4.4.2 Según características base

Se ha estudiado la relación entre el tipo de modelo estadístico que mejor ajusta la evolución temporal de las visitas a las páginas de los artículos/noticias del portal y sus características o atributos base. Los resultados se observan en la Ilustración 49.

Distribución de atributos base según el modelo estadístico que mejor ajusta la curva de pv-tiempo..

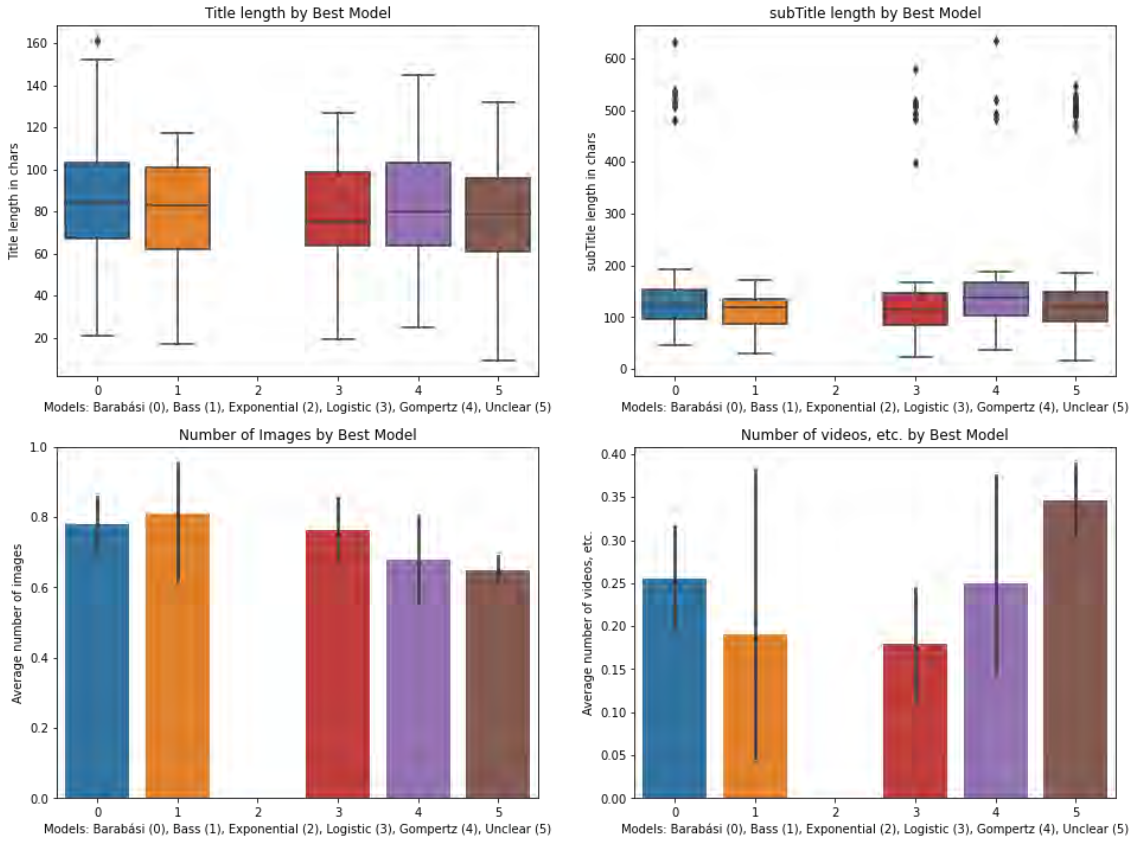

Ilustración 49. Distribución de atributos base según el modelo estadístico que mejor ajusta la curva de pv-tiempo.

Se ha querido observar también la posible relación entre el tipo de modelo estadístico que mejor ajusta la evolución temporal de las visitas a las páginas de los artículos/noticias del portal web y la categoría base de estos. Los resultados se observan en la Ilustración 50. Repartición de las categorías base según el modelo estadístico que mejor ajusta la curva de pv-t..

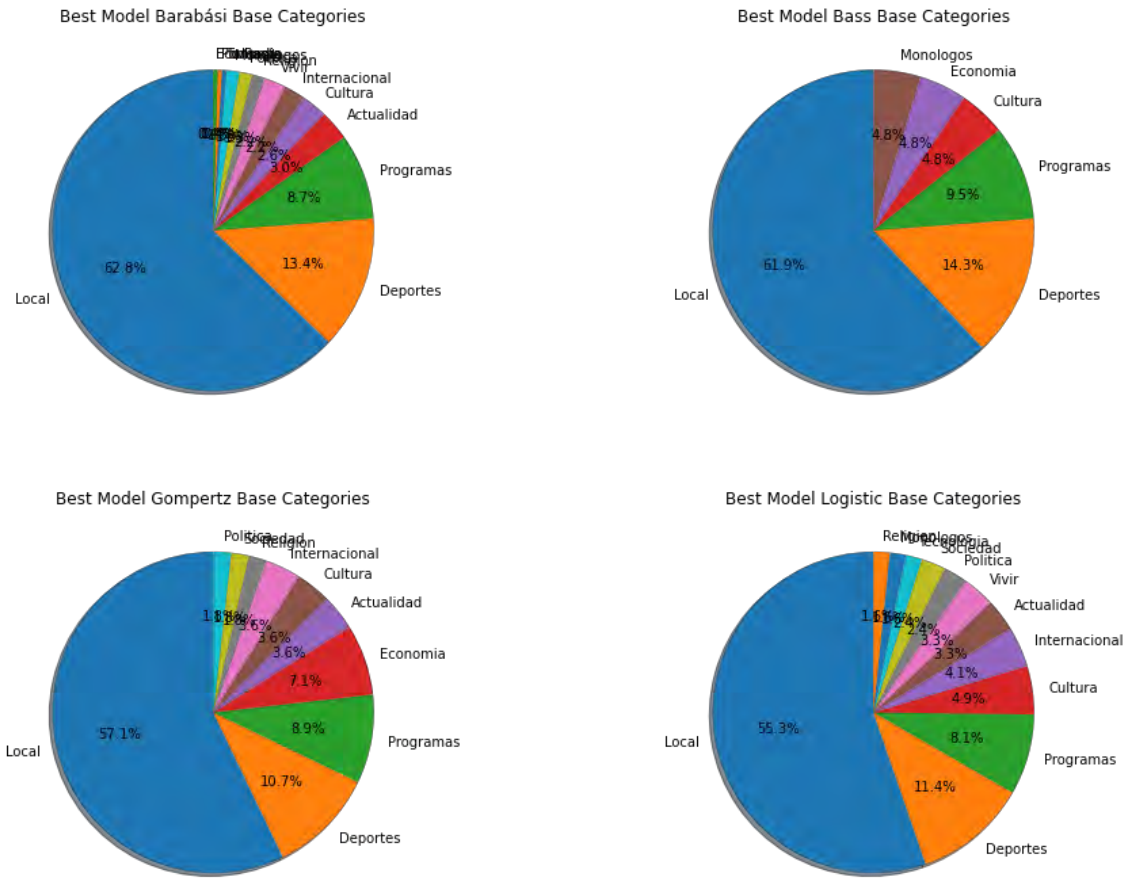

Ilustración 50. Repartición de las categorías base según el modelo estadístico que mejor ajusta la curva de pv-t.

Observaciones:

- Los artículos/noticias de **Economía** siguen preferencialmente el modelo de **Gompertz** en cuanto a la evolución temporal de las visitas a sus páginas.

- El modelo de **Bass**, comparado con el resto de los modelos, **no es apto** para el ajuste de las curvas de visitas a las páginas de los artículos/noticias de **Actualidad**. Esto se puede deber a que este tipo de piezas suelen ser informativas más que "innovativas", con lo que este modelo no sería el más adecuado.

- El modelo de **Barabási** es no sólo el que mejores resultados obtiene a nivel global, sino también el más **versátil** en el sentido de que es capaz de ser el mejor modelo para ajustar las curvas de visitas a las páginas de artículos/noticias de todas o casi todas las categorías.

5 Categorización por Clusterización

Este capítulo está dedicado a la presentación de diferentes técnicas de aprendizaje automático no supervisado conocidas como técnicas de clusterización, y de los resultados de su aplicación sobre el conjunto de datos.

En la primera sección, se realiza una introducción a esta rama del aprendizaje automático, se explican las diferentes técnicas existentes y sus aplicaciones en diferentes casos y áreas del mundo real, incluido el de este proyecto.

Después, se revisan en detalle dos de las técnicas introducidas en la sección anterior, sus ámbitos de aplicación, ventajas y desventajas, y demás aspectos. Se incluye en esta sección el análisis de clusterización realizado con cada una de estas dos técnicas.

Por último, se comparan los resultados obtenidos con ambas técnicas y se realizan una serie de observaciones sobre ello.

5.1 Introducción

En este proyecto se han aplicado técnicas de clusterización sobre el conjunto de datos con diferentes objetivos:

- Introducir y demostrar el proceso de la aplicación de estas técnicas en un caso real como el de este proyecto.
- Utilizar estas técnicas como herramientas independientes para obtener información sobre la distribución de datos del conjunto.
- Encontrar relaciones ocultas y descubrir patrones de comportamiento de los atributos/parámetros.
- Hallar nuevas formas de categorizar y clasificar los objetos de datos.

Toda vez que la clusterización ya ha sido introducida en el primer capítulo, a continuación, se resume en puntos brevemente su esencia, sus tipos y sus aplicaciones.

- Clúster: una colección de objetos de datos
  - Similares entre sí dentro del mismo grupo.
  - Diferente a los objetos de otros grupos.
- Análisis de clústeres:
  - Agrupar un conjunto de objetos de datos en grupos.
- La clusterización es una clasificación no supervisada: no hay clases predefinidas.
- Aplicaciones Típicas:
  - Como una herramienta independiente para obtener información sobre la distribución de datos.

    - Como paso de preprocesamiento para otros algoritmos.
- Un buen método de clusterización producirá clústeres de alta calidad con:
    - alta similitud intraclase
    - baja similitud entre clases

5.2 Tipos de clusterización

A nivel general, y como ya se ha explicado en el primer capítulo, existen diferentes algoritmos de clusterización según su función de distancia, su objetivo, o sus condiciones iniciales. Se pueden encontrar los siguientes tipos principales:

- Algoritmos de particionamiento: construyen varias particiones y luego las evalúa siguiendo algún criterio.
- Algoritmos de jerarquía: crean una descomposición jerárquica del conjunto de datos (u objetos) utilizando algún criterio.
- Basados en densidad: basado en funciones de conectividad y densidad.
- Basados en cuadrícula (grid): basado en una estructura de granularidad multinivel.
- Basado en modelos: se formula la hipótesis de un modelo para cada uno de los grupos y la idea es encontrar el mejor ajuste de este modelo entre ellos.

En este proyecto, se ha decidido aplicar dos tipos de clusterización:

1. un algoritmo de clusterización de particionamiento, o basado en centroides;
2. y un algoritmo de clusterización basado en densidad.

Esta decisión ha sido tomada en base a varias razones:

- Se ha analizado el conjunto de datos y se ha concluido que estos dos tipos de algoritmo, que son a su vez los más populares en análisis de clusterización, son adecuados y suficientes para obtener resultados variados y satisfactorios.
- Las restricciones temporales, computacionales y de conocimientos no han permitido explorar más de dos opciones de clusterización.

5.2.1 Clusterización basada en centroides: K-Means

La clusterización basada en centroides organiza los datos en clústeres no jerárquicos. K-Means es el algoritmo de clusterización basada en centroides más utilizado. Los algoritmos basados en centroides son eficientes, pero sensibles a las condiciones iniciales y a los valores atípicos. K-Means es un algoritmo de clusterización simple, eficaz y eficiente, aunque, como señalan muchos expertos en análisis de clusterización, el algoritmo K-Means no es muy flexible y el tamaño de las K agrupaciones tiende a ser el mismo, lo que no siempre es útil.

En este caso, se ha realizado una **clusterización 3D con K-Means**, donde las tres dimensiones son los **tres parámetros de rendimiento** que se han considerado más importantes: las visitas a página, el tiempo medio en página y la tasa de salidas.

El primer paso en el proceso de clusterización por centroides es seleccionar o adivinar el número de clústeres (K). Este es un problema frecuente en la clusterización de datos y no hay una respuesta clara sobre cuál es el mejor método. La selección de K a menudo se basa en interpretaciones subjetivas e instinto. El método elegido que se utiliza aquí se llama **Método del codo**. El Método del codo es un método de interpretación y validación de la coherencia dentro del análisis de clústeres diseñado para ayudar a encontrar el número correcto de clústeres en un conjunto de datos. Este método analiza el % de varianza explicado en función del número de clústeres. La regla es elegir un nº de clústeres (K) tal que agregar otro clúster no proporciona un mejor modelado de los datos.

La curva de puntuación para la agrupación de K-Means realizada en el conjunto de datos se muestra en la Ilustración 51. Método del codo para la clusterización 3D con K-Means.. Se han seleccionado 7 clústeres, aunque también se podrían haber seleccionado 8 o incluso 9.

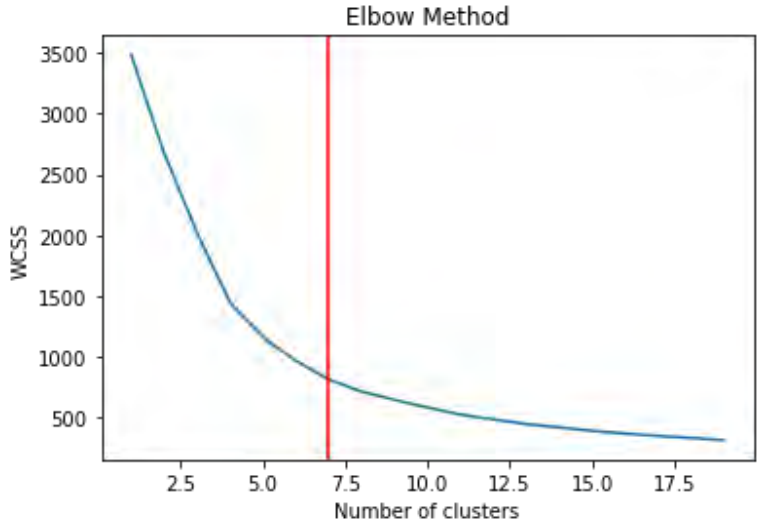

Ilustración 51. Método del codo para la clusterización 3D con K-Means.

La Ilustración 52, la Ilustración 53, la Ilustración 54 y la Ilustración 55 muestran los resultados gráficos de la clusterización 3D con K-Means.

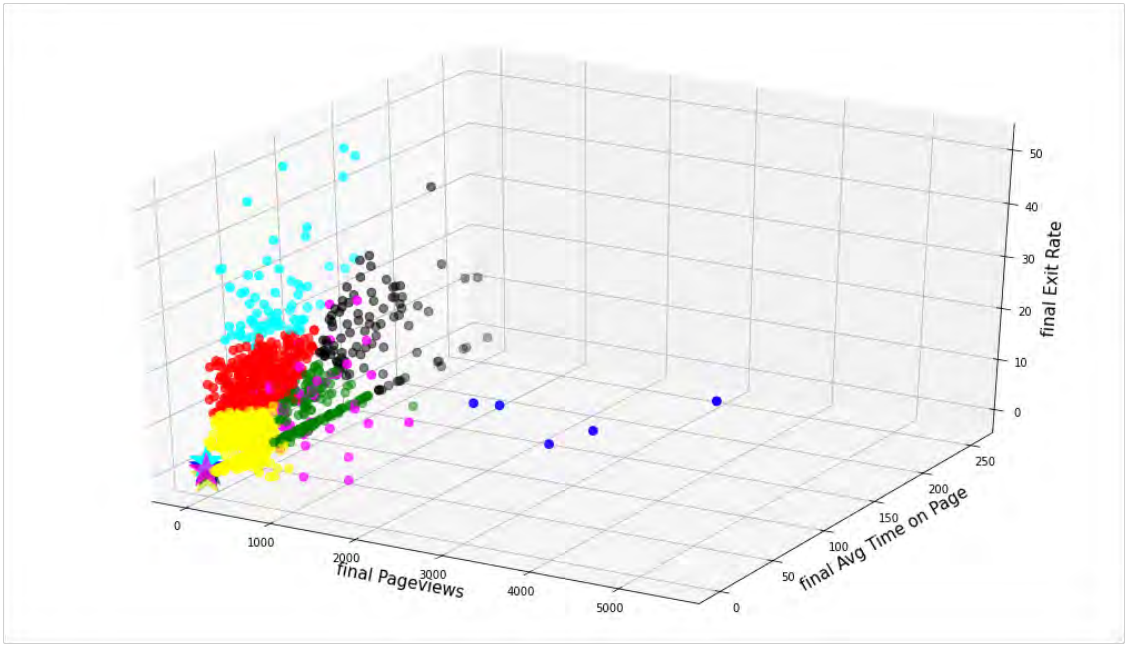

Ilustración 52. Clusterización 3D 7-Means.

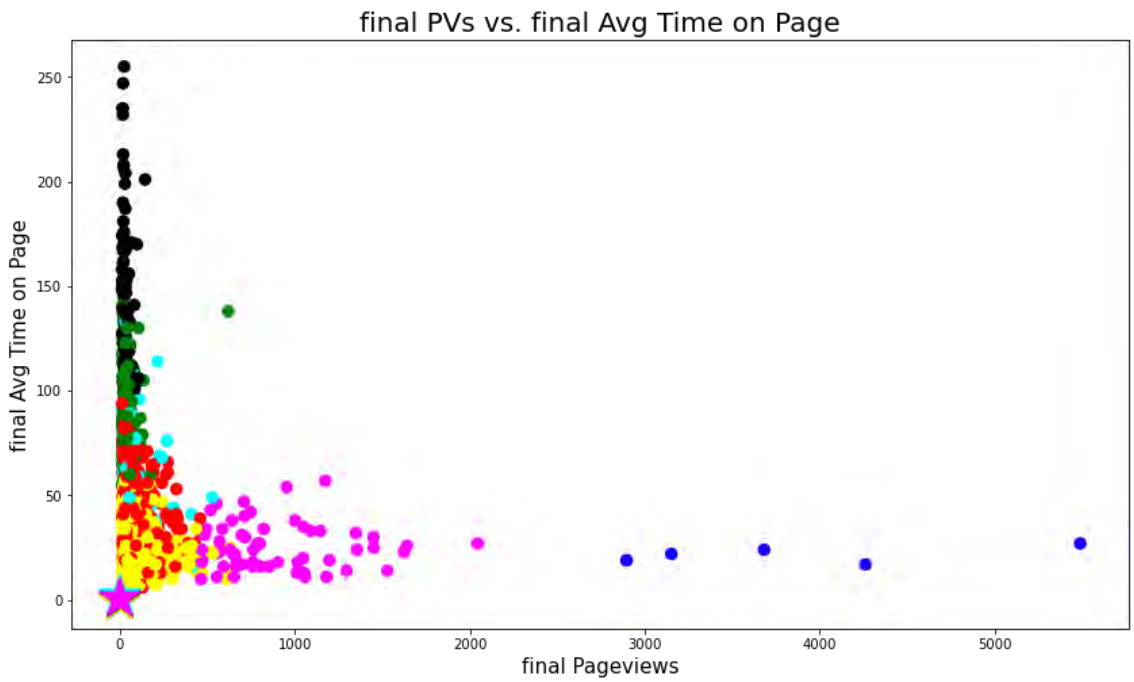

Ilustración 53. Dispersión 2D visitas a página-tiempo medio en página de la clusterización 7-Means.

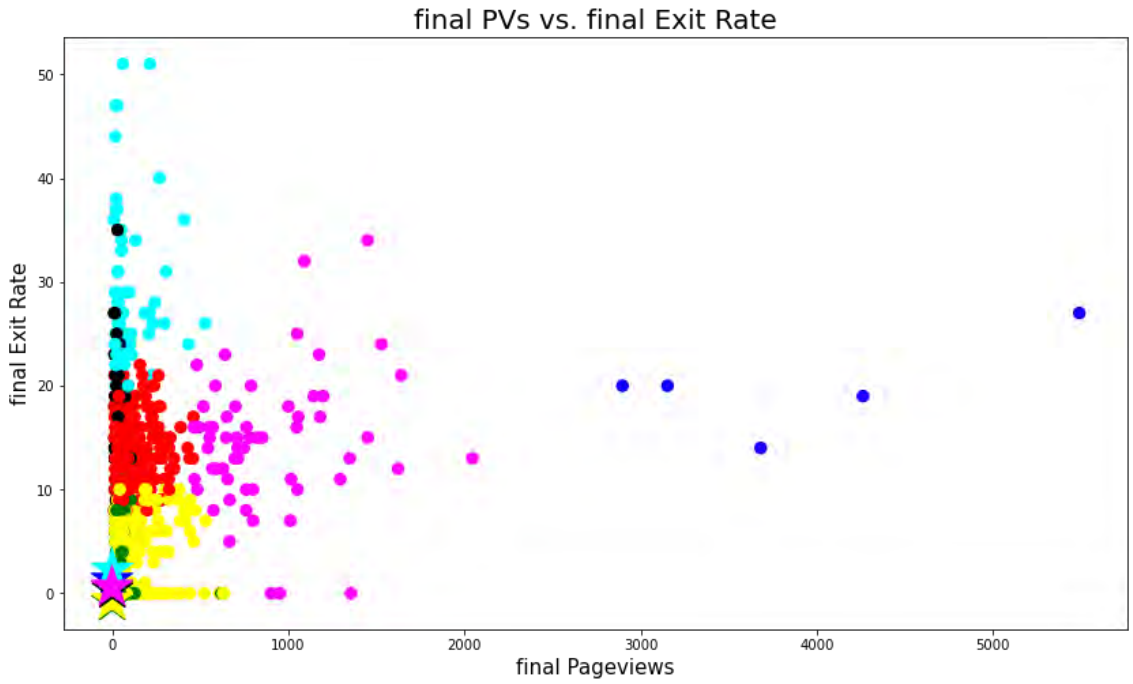

Ilustración 54. Dispersión 2D visitas a página-tasa de salidas de la clusterización 7-Means.

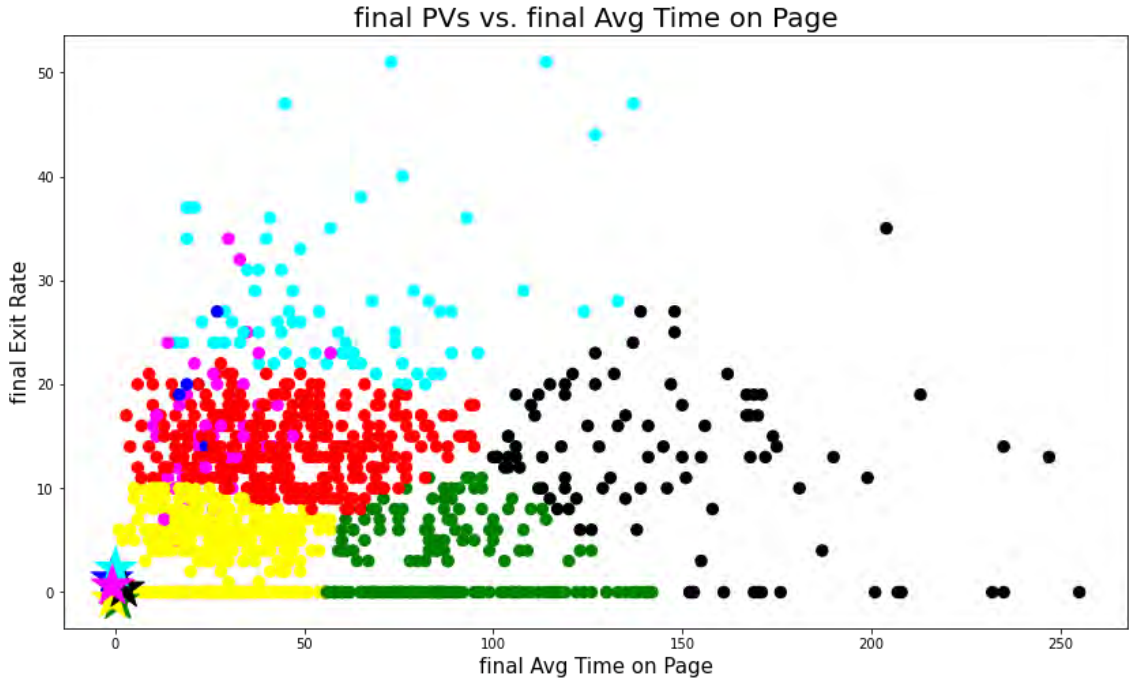

Ilustración 55. Dispersión 2D Tiempo medio en página-tasa de salidas de la clusterización 7-Means.

Se comprueba que el método puede dividir los artículos/noticias, según sus parámetros de rendimiento finales, en 7 clústeres de tamaños diferentes, pero del mismo "orden de magnitud", excepto uno que tiene una población de sólo cinco objetos.

La Ilustración 56, la Ilustración 57 y la Ilustración 58, a continuación, muestran los valores medianos y la distribución de los tres parámetros de rendimiento en los siete clústeres creados.

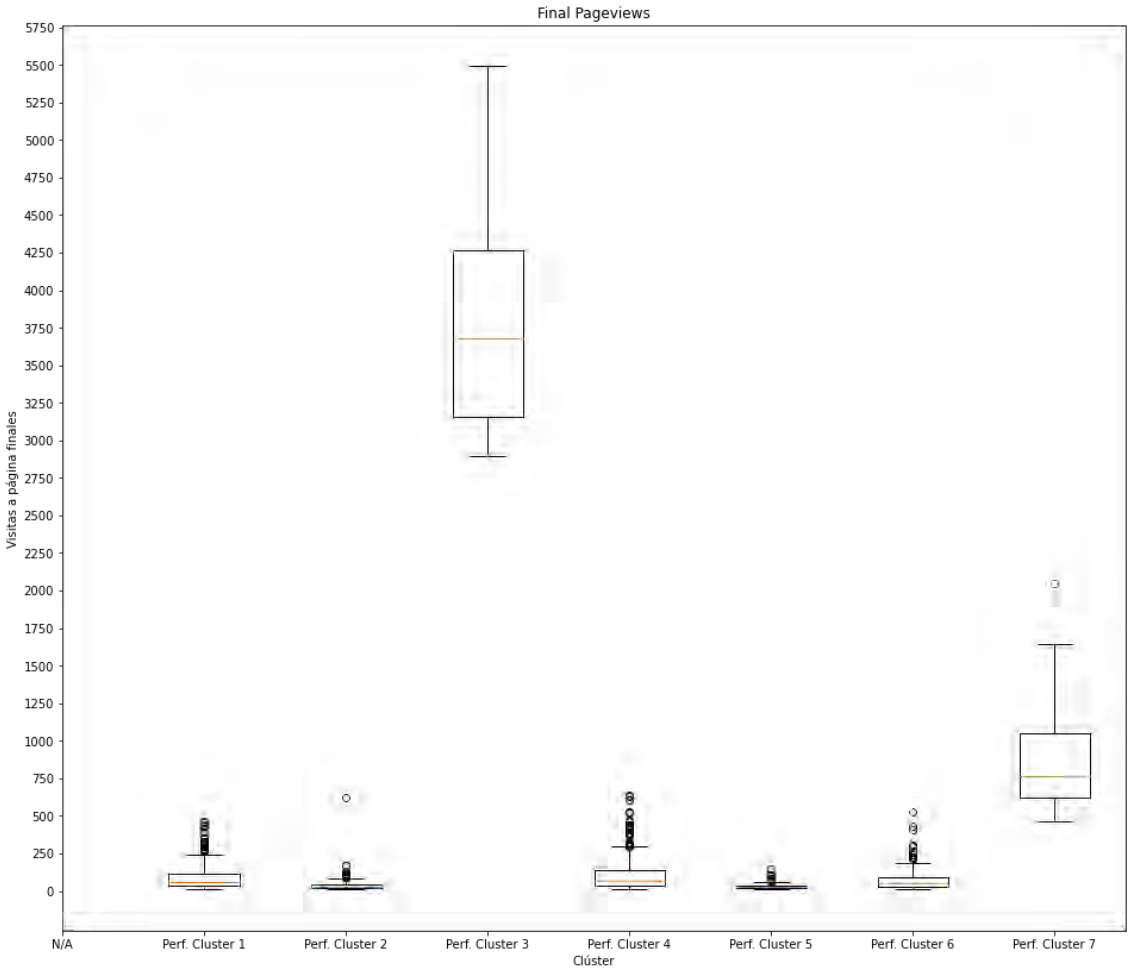

Ilustración 56. Distribución de las Visitas a página finales por clúster.

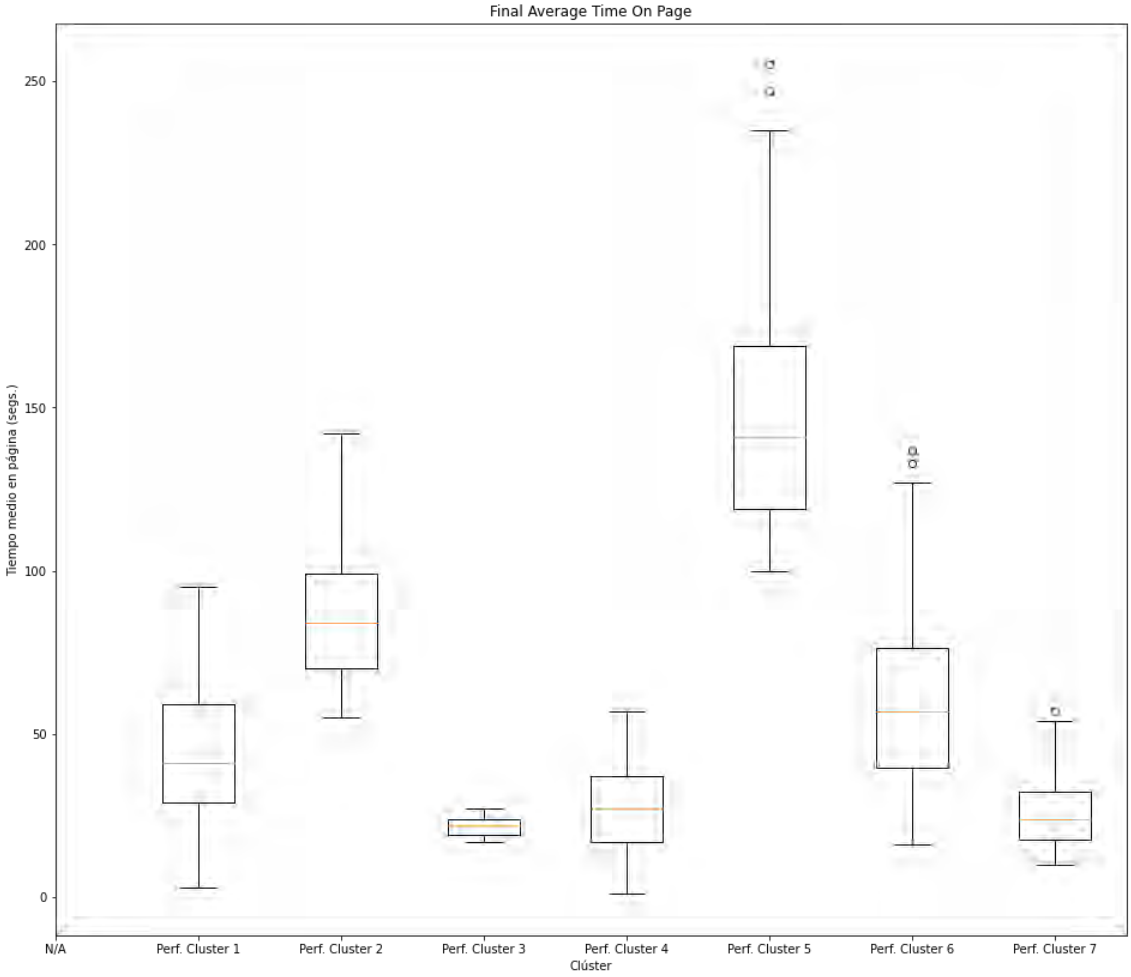

Ilustración 57. Distribución del tiempo medio en página final por clúster.

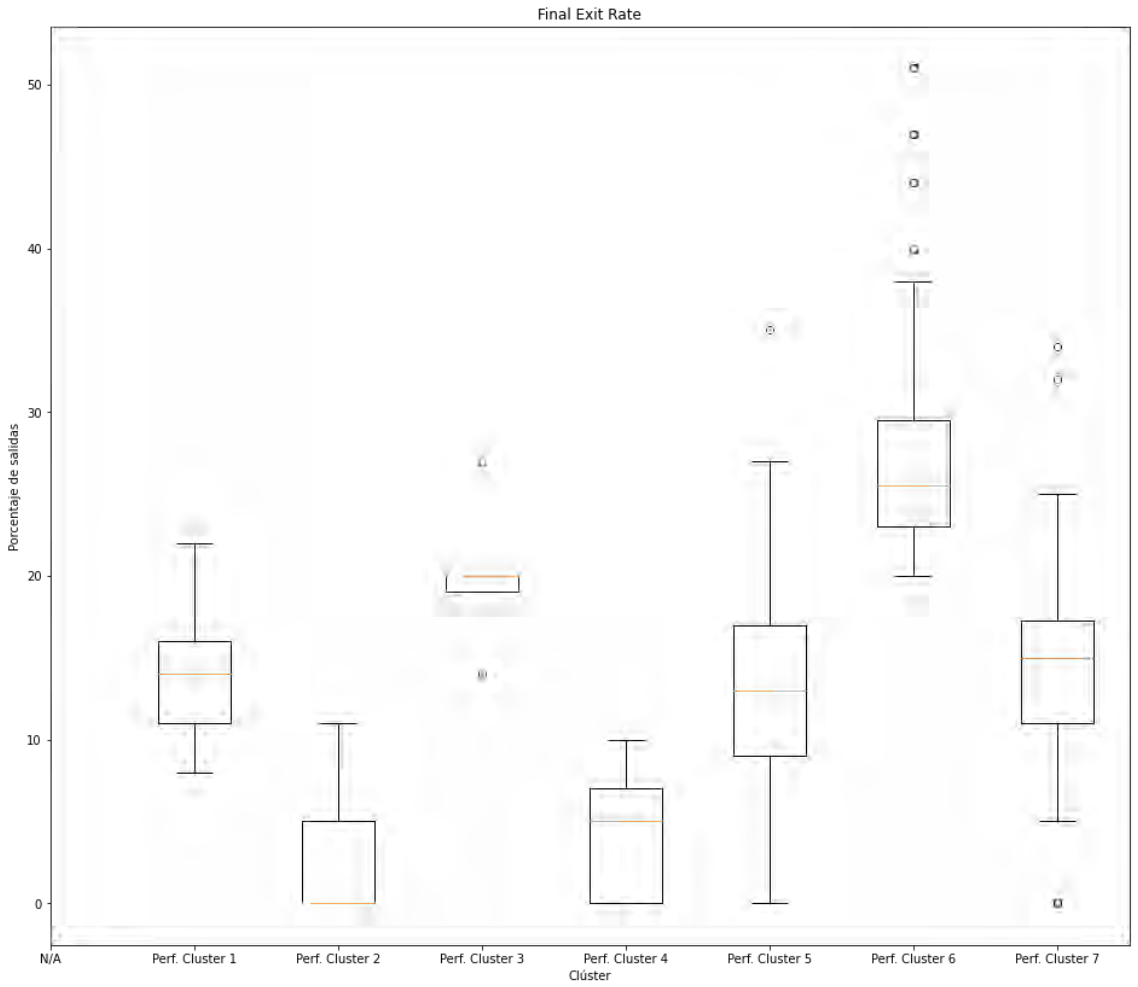

Ilustración 58. Distribución de la tasa de salidas final por clúster

5.2.1.1 Análisis de los clústeres

Ahora la idea es echar un vistazo a los artículos/noticias dentro de cada clúster para verificar si tienen características y/o atributos similares.

Para ello, se han creado una serie de gráficos y de diagramas de cajas donde se han juntado las distribuciones de los valores de los atributos base/características de los artículos/noticias de todos los clústeres creados. El objetivo ha sido poder encontrar rápidamente y de manera visual relaciones directas entre un clúster de rendimiento y los atributos o características base de los artículos/noticias del portal web, ya que se considera que puede ser información útil.

En las siguientes imágenes, que van de la Ilustración 59 a la Ilustración 65, se pueden observar los diferentes gráficos y diagramas obtenidos del análisis de los clústeres de rendimiento obtenidos con el algoritmo de K-Means que se ha aplicado sobre el espacio 3D conformado por los valores finales de los tres parámetros de rendimiento: visitas a página, tiempo medio en página y tasa de salidas.

Ilustración 59. K-Means 3D: Distribución de la longitud del título de los artículos/noticias en cada clúster de rendimiento.

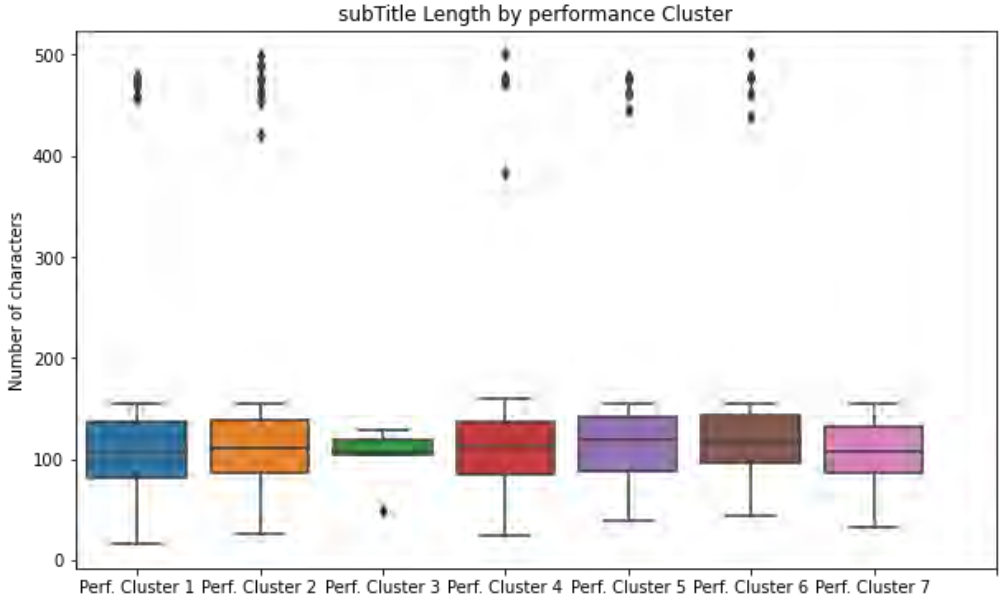

Ilustración 60. K-Means 3D: Distribución de la longitud del subtítulo de los artículos/noticias en cada clúster de rendimiento.

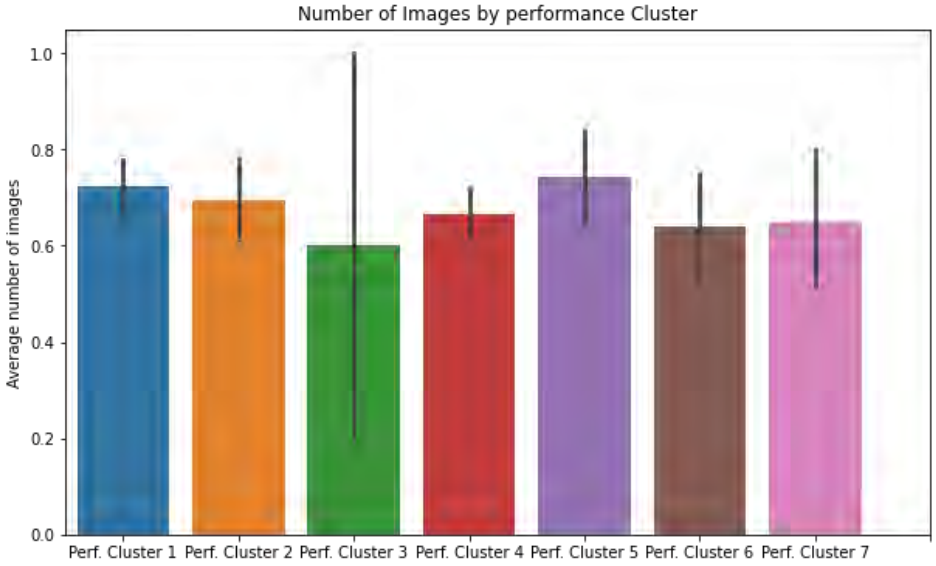

Ilustración 61. K-Means 3D: Número medio de imágenes en los artículos/noticias en cada clúster de rendimiento.

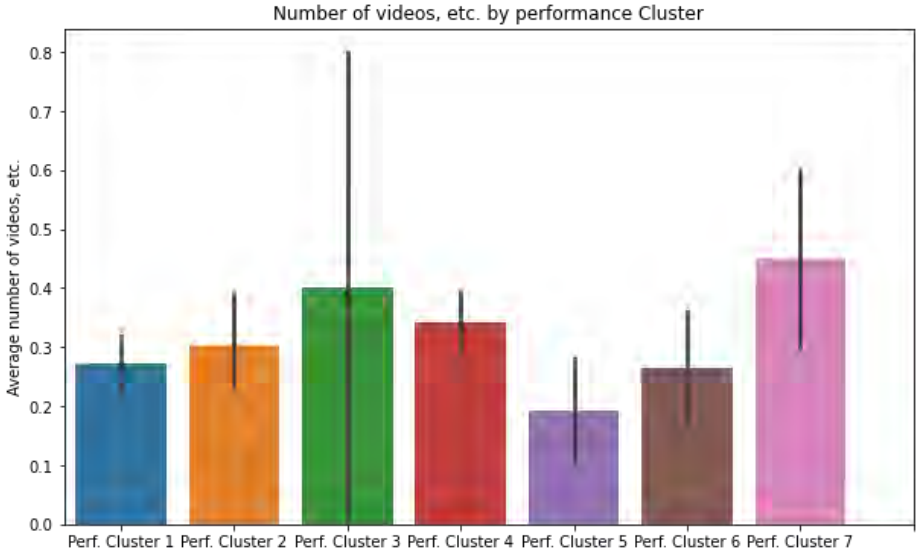

Ilustración 62. K-Means 3D: Número medio de vídeos, audios, etc. en los artículos/noticias en cada clúster de rendimiento.

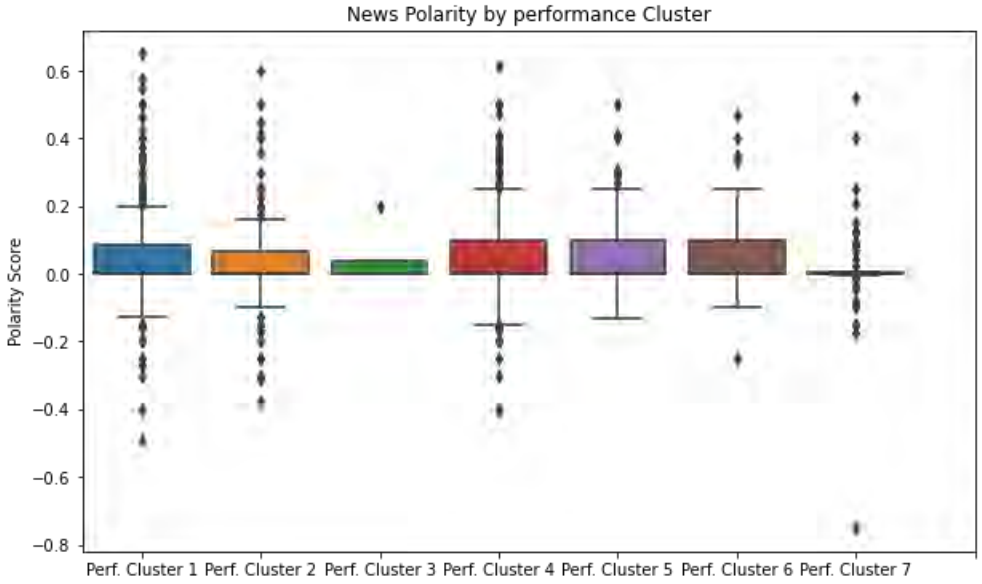

Ilustración 63. K-Means 3D: Valor medio de la polaridad de los artículos/noticias en cada clúster de rendimiento.

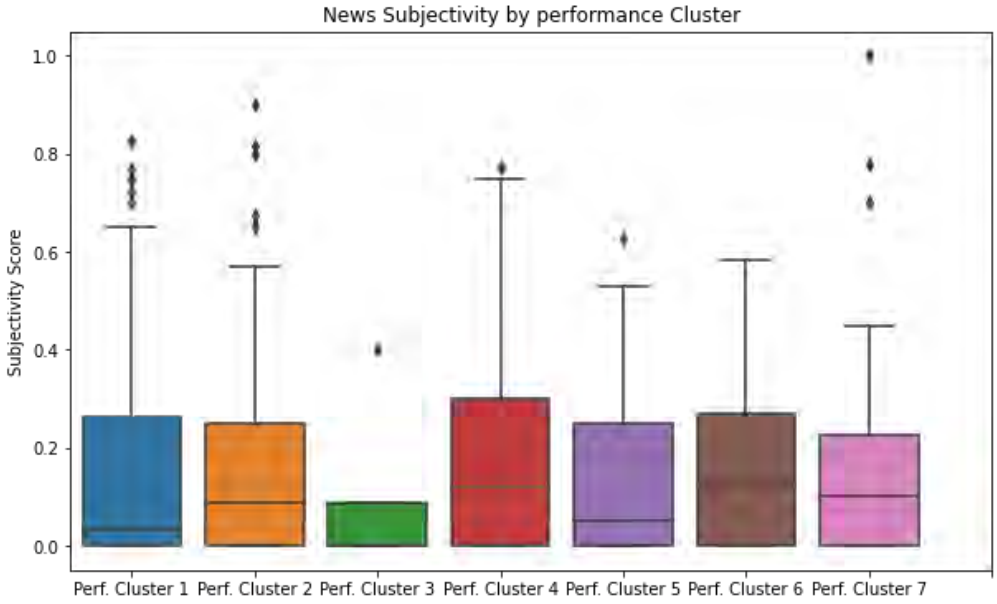

Ilustración 64. K-Means 3D: Valor medio de la subjetividad de los artículos/noticias en cada clúster de rendimiento.

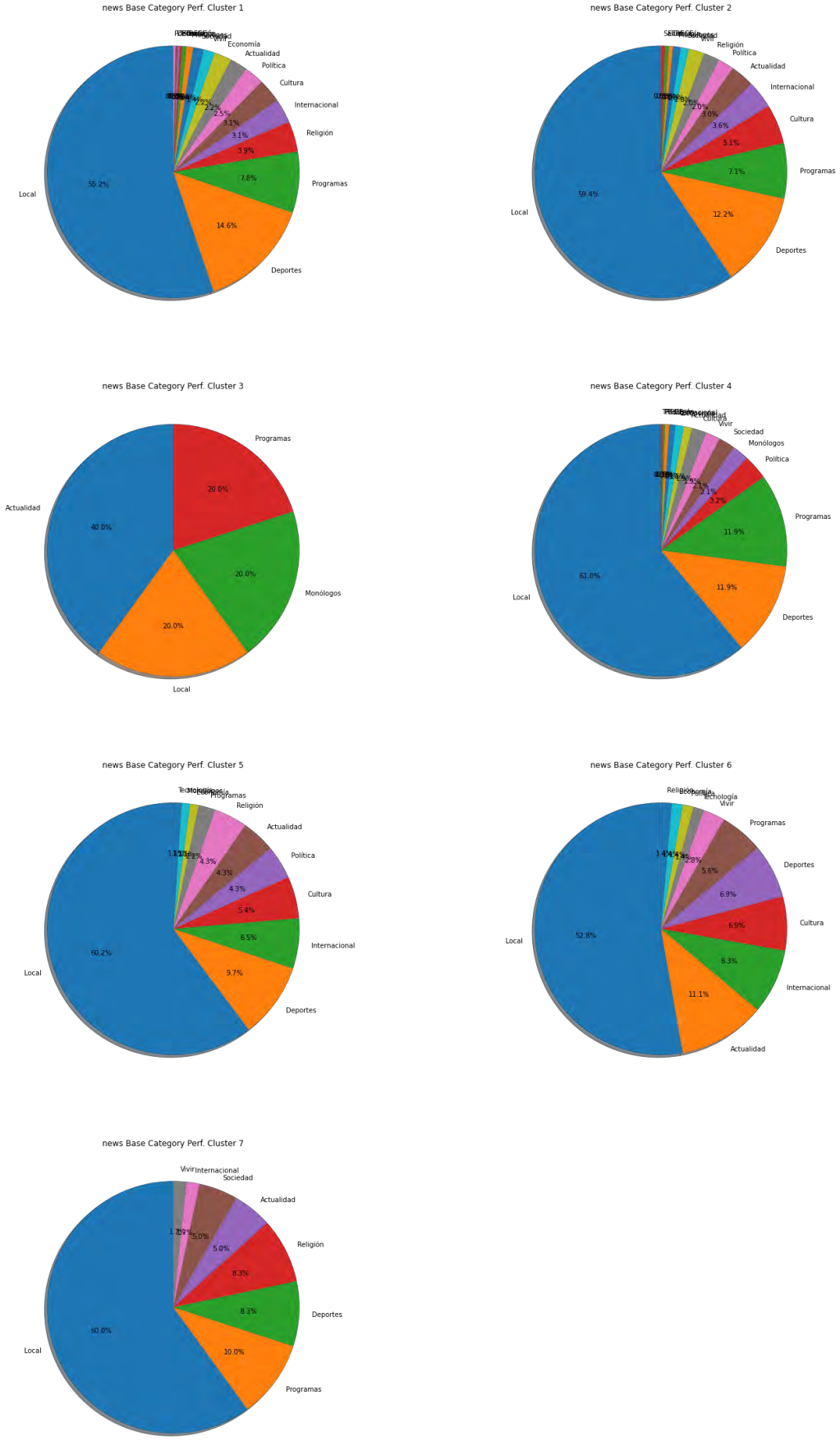

Ilustración 65. K-Means 3D: Categorías base de los artículos/noticias en cada clúster de rendimiento.

5.2.2 Clusterización basada en densidad: DBSCAN

Como ya se ha comentado anteriormente, la clusterización basada en densidad conecta áreas de alta densidad de ejemplo en clústeres. Esto permite distribuciones de formas arbitrarias, siempre que se puedan conectar áreas densas. Estos algoritmos tienen dificultades con datos de densidades variables y dimensiones elevadas. Además, por diseño, estos algoritmos no asignan valores atípicos a los clústeres.

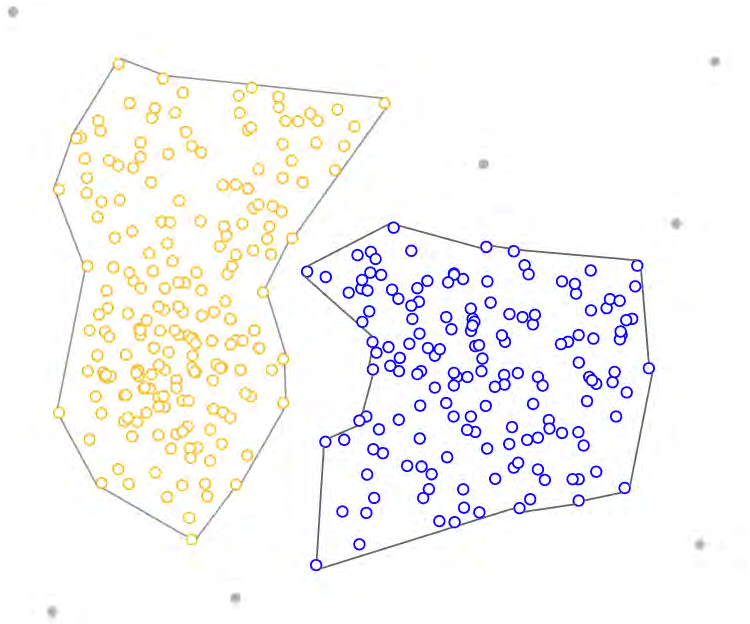

Ilustración 66. Ejemplo de clusterización basada en densidades.

En este proyecto, como se ha mencionado en secciones previas, se ha realizado la clusterización a partir de los pocos parámetros de rendimiento más importantes. En el caso de la clusterización basada en centroides, presentada y detallada en las secciones anteriores, se utilizaron tres: las visitas a página finales, el tiempo medio en página final y la tasa de salidas final. Para la clusterización basada en densidad no se ha encontrado una librería o herramienta que permitiera un análisis en tres dimensiones, así que se ha decidido eliminar el parámetro de rendimiento de tasa de salidas final, que se considera el menos relevante, y se ha realizado el análisis de clusterización basado en densidad en dos dimensiones, correspondientes a los **dos primeros parámetros de rendimiento**.

El algoritmo basado en densidad más eficaz y popular es el **DBSCAN** (*Density-based spatial clustering of applications with noise*). Es el que se ha utilizado en esta parte, gracias a que la popular librería **scikit-learn** de Python incluye una implementación del algoritmo de DBSCAN, para realizar la programación del análisis basado en densidad.

Algunos aspectos destacados sobre la clusterización de DBSCAN son:

- no requiere que el usuario establezca el número de clústeres a priori.
- puede capturar grupos de formas complejas.

- puede identificar puntos que no forman parte de ningún clúster (muy útil como detector de valores atípicos).
- es algo más lento que el algoritmo de K-Means, pero aún se puede escalar a conjuntos de datos relativamente grandes.
- funciona identificando puntos que se encuentran en regiones abarrotadas del espacio de características, donde muchos puntos de datos están muy juntos (regiones densas en el espacio de características).
- Los puntos que están dentro de una región densa se denominan muestras centrales (o puntos centrales).
- Hay dos parámetros en DBSCAN: ***min_samples*** y ***eps***.
- Si hay al menos *min_samples* puntos de datos a una distancia de *eps* de un punto de datos dado, ese punto de datos se clasifica como una muestra central.
- DBSCAN coloca en el mismo clúster muestras centrales que están más cerca entre sí que la distancia *eps*.

Una vez se ha realizado una primera ejecución del algoritmo DBSCAN con los valores por defecto de los parámetros *min_samples* y *eps* (definidos en los puntos anteriores), normalmente se realiza un ajuste del valor de estos, conocido como **optimización de hiperparámetros** del modelo. En este caso, se ha realizado una optimización manual de estos dos parámetros, para encontrar los valores iniciales con los que el algoritmo arroja los mejores resultados.

- Para el ajuste del hiperparámetro *min_samples*, se ha considerado el tamaño del conjunto de datos y el nº dimensiones utilizadas para la clusterización. Se han seguido las recomendaciones de los expertos de fijar *min_samples* >= 2 * D, donde D = 2 dimensiones.

- Para el ajuste del hiperparámetro *eps*, se ha utilizado el método del codo con el uso del método de distancia euclidiana para encontrar la distancia épsilon óptima (eps). Esto se hace utilizando la búsqueda por K-NN (nearest neighbours).
  OPTICS puede verse como una generalización de DBSCAN que reemplaza el parámetro ε (*eps*) con un valor máximo que afecta principalmente al rendimiento. *min_samples* entonces se convierte esencialmente en el tamaño mínimo de clúster para encontrar. Dicho de otro modo, OPTICS es una variación DBSCAN que elimina el parámetro épsilon; produce un resultado jerárquico que puede verse aproximadamente como "ejecutar DBSCAN con todas las épsilon posibles".

En la Ilustración 67, la Ilustración 68 y la Ilustración 69 se observa el resultado de la clusterización 2D con DBSCAN con diferentes valores de *eps*, para un *min_samples* fijo de 4.

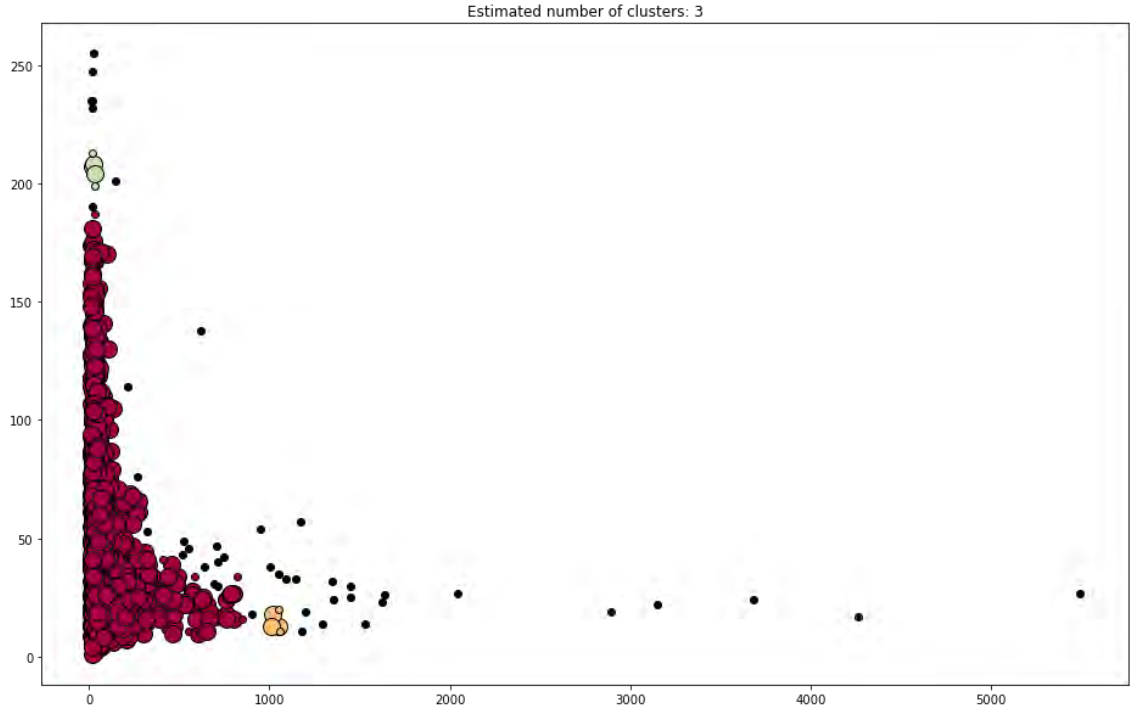

Ilustración 67. DBSCAN 2D: Clústeres con eps = 0.20, min_samples = 4.

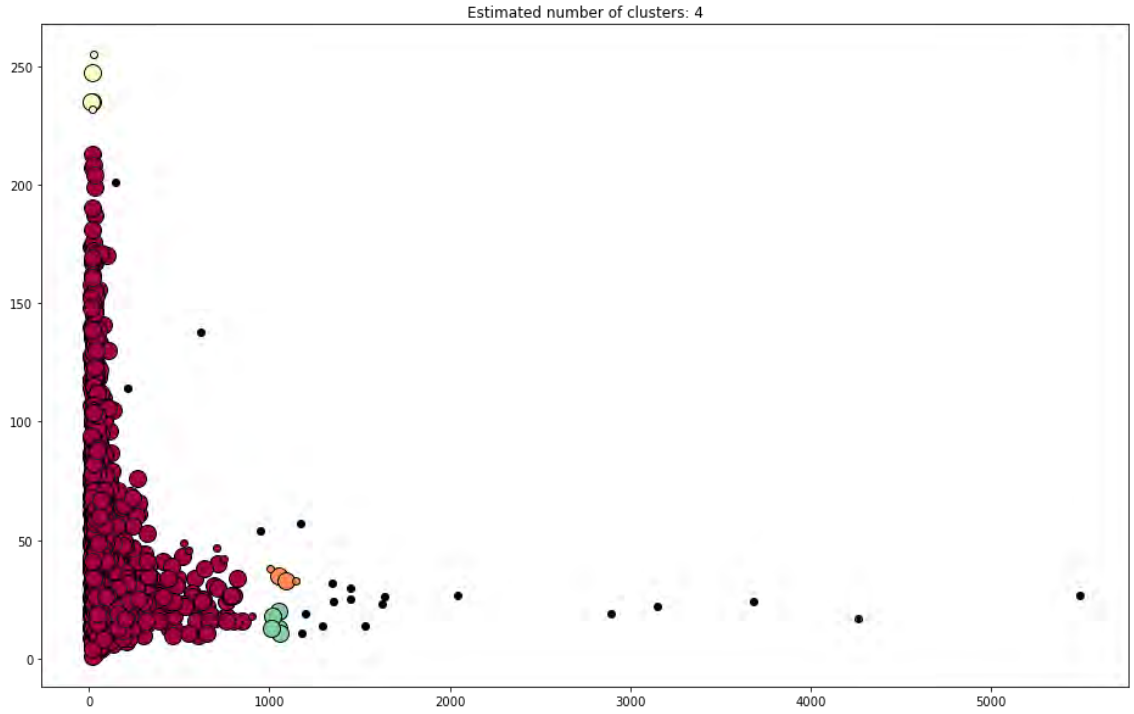

Ilustración 68. DBSCAN 2D: Clústeres con eps = 0.30, min_samples = 4.

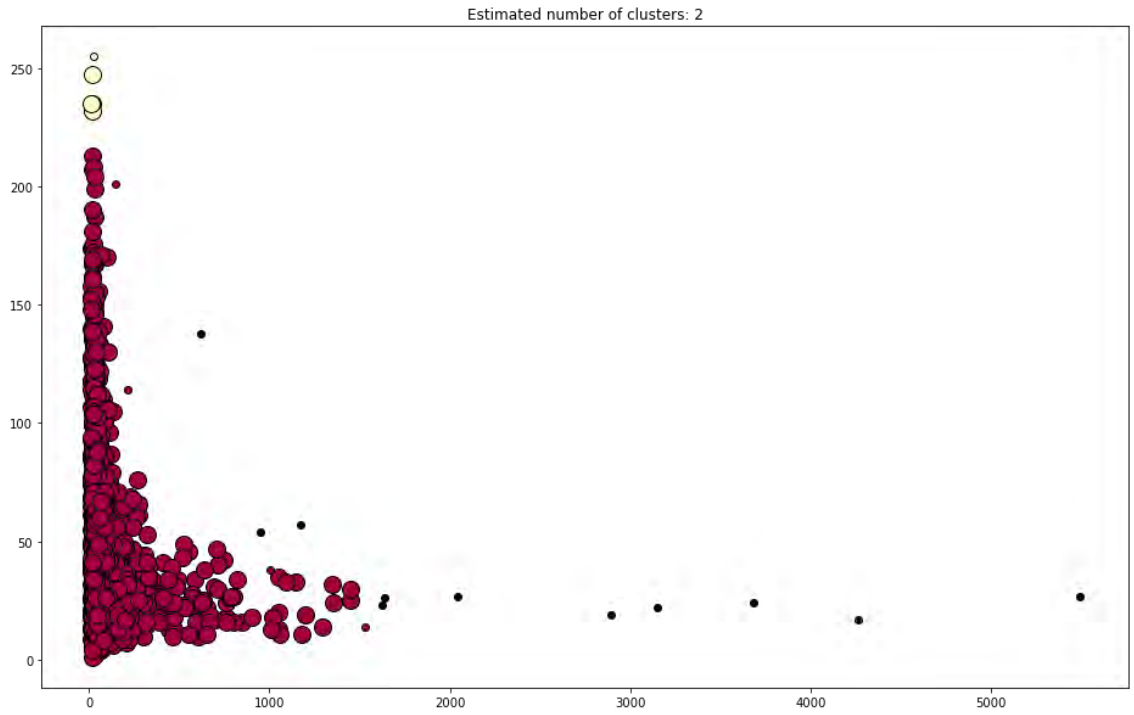

Ilustración 69. DBSCAN 2D: Clústeres con eps = 0.40, min_samples = 4.

5.2.2.1 Análisis de los clústeres

Como se ve en las imágenes anteriores, con diferentes valores ajustados de los hiperparámetros, el algoritmo DBSCAN no genera más de 4 clústeres. Uno de ellos es el que concentra la inmensa mayoría de la población del conjunto de datos y el resto no tienen más que 4 ó 5 elementos. Los puntos negros son objetos que no entran en ningún clúster de densidad, es decir, son los valores extremos, o *outliers*, en inglés. Además, tampoco se ha podido determinar con claridad el valor óptimo de *eps*.

Se ha considerado, pues, que no aporta valor el analizar las características y/o atributos base de los artículos/noticias de estos clústeres de densidad.

5.3 Comparativa de métodos

Teniendo en cuenta que, como ya se ha explicado, existen más tipos de clusterización que se podrían aplicar al caso de este proyecto, a continuación, se comparan los dos algoritmos utilizados (centroides y densidad) en base a los resultados.

En concreto, se ha realizado una clusterización basada en centroides con el algoritmo K-Means pero de dos dimensiones (2D), correspondientes a los valores finales de los dos parámetros de rendimiento más importantes: visitas a página finales y tiempo medio en página final. Esto se ha hecho para poder comparar directamente los

resultados de este análisis de clusterización con los que se presentan y explican en la sección anterior (clusterización 2D con DBSCAN).

Así pues, en la Ilustración 70 se observa el resultado de la clusterización en 2D con K-Means, que se puede comparar con el resultado de la clusterización 2D con DBSCAN para un valor razonable de eps como es 0.30, que se observa en la Ilustración 71.

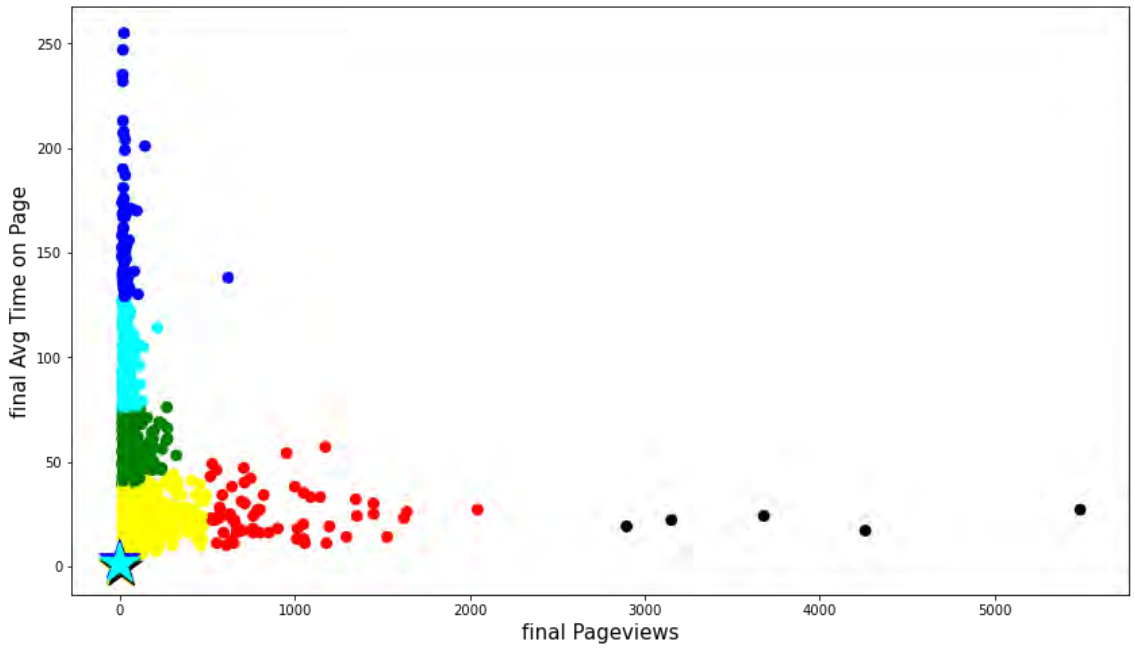

Ilustración 70. K-Means 2D: Resultado de la clusterización 6-Means.

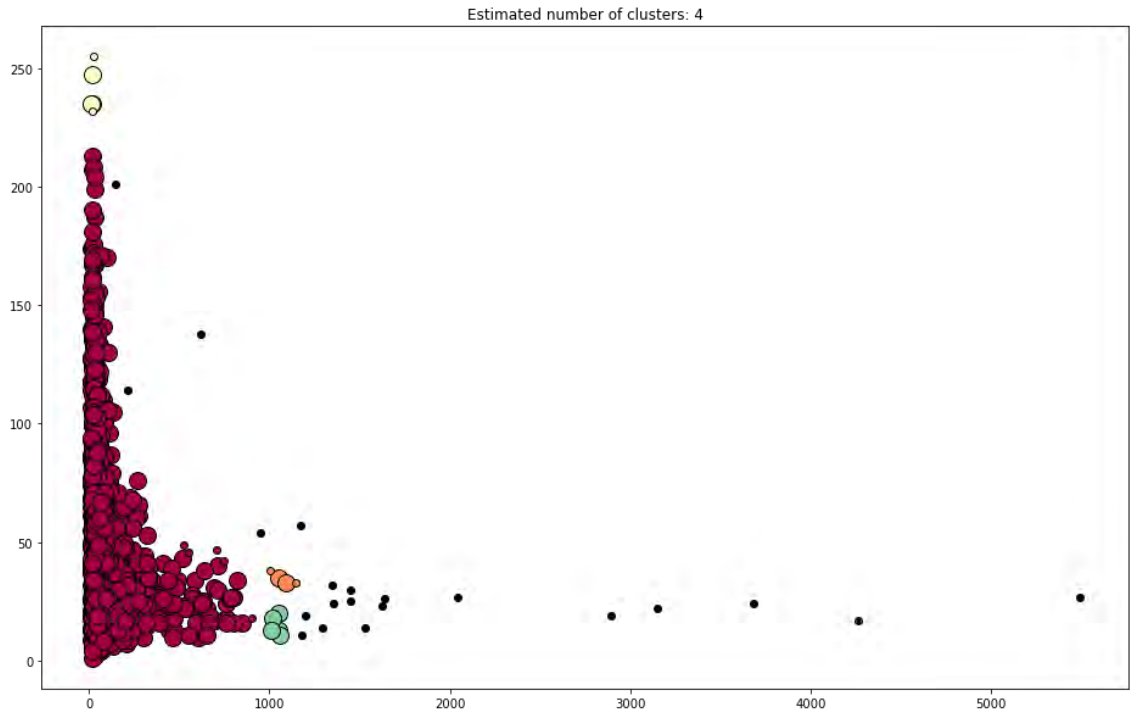

Ilustración 71. DBSCAN 2D: Clústeres con eps = 0.30, min_samples = 4 (bis).

5.4 Resultados y observaciones

A partir de los resultados de los dos tipos de clusterización, que se observan en las imágenes de la sección anterior, se pueden sacar una serie de conclusiones y realizar una serie de observaciones:

- Por el tipo de distribución en el espacio 2D de los valores finales de los dos parámetros de rendimiento de todos los artículos/noticias, resulta claro y visible que un algoritmo basado en densidad generará muy pocos clústeres si se compara con los generados por un algoritmo basado en centroides.
  Es decir, en cierto modo, el algoritmo basado en centroides genera, "demasiados" clústeres, mientras que el algoritmo basado en densidades genera "demasiado pocos".

- El algoritmo basado en centroides genera un clúster casi específico para los *outliers* en una de dos las dimensiones, mientras que el algoritmo basado en densidades descarta automáticamente todos estos *outliers*.

- Ambas soluciones tienen sus ventajas y sus desventajas; sin embargo, para este caso en concreto, se considera que un algoritmo basado en centroides con un número de clústeres muy ajustado puede arrojar resultados más informativos y variados.

6 Predicción

Este capítulo está dedicado enteramente al estudio y la aplicación de una técnica de análisis de regresión para la predicción de los valores finales de los parámetros de rendimiento de los artículos/noticias.

En las dos primeras secciones, se hace una introducción al análisis de regresión para predicciones y a las técnicas de aprendizaje automático que lo permiten, respectivamente.

En la tercera sección, se presenta la solución de aprendizaje automático que se ha seleccionado para realizar el análisis predictivo, y se detallan y visualizan los resultados de dicho análisis.

Por último, se presentan y explican las conclusiones a las que se ha llegado para esta parte, a partir de los resultados detallados en la sección previa.

6.1 Introducción

En cualquier ámbito de la vida personal y/o de los negocios, es muy atractiva la idea de poder predecir con más o menos precisión el valor o estado futuro de algo. Si se consigue, se puede obtener una ventaja o ganancia que puede tomar muchas formas.

Ilustración 72. Ramas y aplicaciones del análisis predictivo. Fuente: https://www.log-hub.com/predictive-analytics/

En este proyecto, se ha realizado un estudio a alto nivel sobre la posibilidad de predecir cuántas visitas habrán acumulado las páginas de las noticias/artículos del portal web al final de su vida útil, así como la duración media de dichas visitas, en función de la serie de características iniciales o atributos base que todos los artículos/noticias tienen, y que se presentan en el Capítulo 3.

Como ya se introduce y explica más en detalle en el capítulo 2, aquí únicamente se hace un breve repaso de lo que es el **análisis de regresión**.

En estadística, el análisis de la regresión es un proceso estadístico para estimar las relaciones entre variables. Incluye muchas técnicas para el modelado y análisis de diversas variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes (o predictoras). Más específicamente, el análisis de regresión ayuda a entender cómo el valor de la variable dependiente varía al cambiar el valor de una de las variables independientes, manteniendo el valor de las otras variables independientes fijas. Más comúnmente, el análisis de regresión estima la esperanza condicional de la variable dependiente dadas las variables independientes -es decir, el valor promedio de la variable dependiente cuando se fijan las variables independientes. En todos los casos, el objetivo de la estimación es una función de las variables independientes llamada la función de regresión. En el análisis de regresión, también es de interés caracterizar la variación de la variable dependiente en torno a la función de regresión, la cual puede ser descrita por una distribución de probabilidad.

El análisis de regresión es ampliamente utilizado para la predicción y previsión, donde su uso tiene superposición sustancial en el campo de aprendizaje automático. El análisis de regresión se utiliza también para comprender cuales de las variables independientes están relacionadas con la variable dependiente, y explorar las formas de estas relaciones. En circunstancias limitadas, el análisis de regresión puede utilizarse para inferir relaciones causales entre las variables independientes y dependientes. Sin embargo, esto puede llevar a ilusiones o relaciones falsas, por lo que se recomienda precaución. La correlación no implica causalidad.

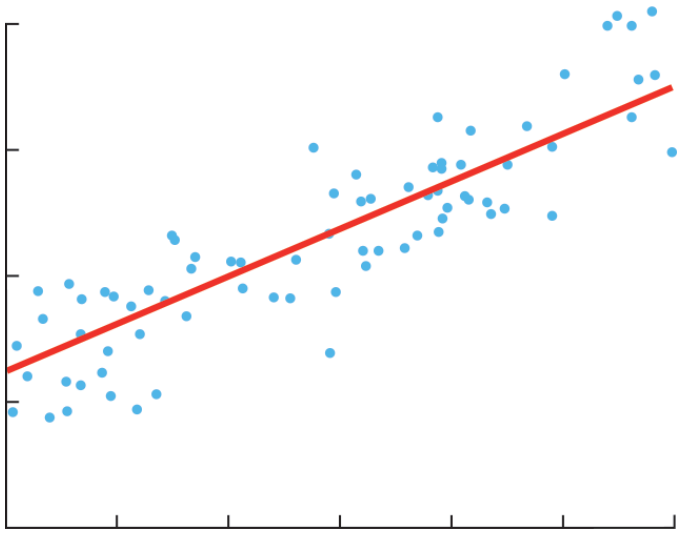

Ilustración 73. Ejemplo de regresión lineal.

6.2 Gradient Boosting

Al haber sido este concepto ya introducido en el capítulo 2, aquí únicamente se hace un breve repaso de su esencia y se añaden algunos detalles más acerca del mismo.

Gradient Boosting Machine (GBM) es un poderoso algoritmo de aprendizaje supervisado que combina múltiples alumnos débiles en conjunto con un excelente desempeño predictivo. Funciona muy bien en varias tareas de predicción en filtrado de spam, publicidad online, detección de fraudes y/o anomalías, física computacional (por ejemplo, el descubrimiento del bosón de Higgs), etc.; y ha aparecido habitualmente como el algoritmo superior en las competiciones de Kaggle y la KDDCup. GBM puede manejar naturalmente conjuntos de datos heterogéneos (datos altamente correlacionados, datos faltantes, datos no normalizados, etc.) y lleva a modelos interpretables mediante la construcción de un modelo aditivo. También es bastante fácil de usar con varias implementaciones disponibles públicamente como la de scikit-learn, gbm, Spark MLLib, LightGBM, XGBoost, etc. [43]

La idea principal del *gradient boosting* es agregar nuevos modelos al conjunto de forma secuencial. En cada iteración particular, se entrena un nuevo modelo débil de aprendiz básico con respecto al error de todo el conjunto aprendido hasta ahora.

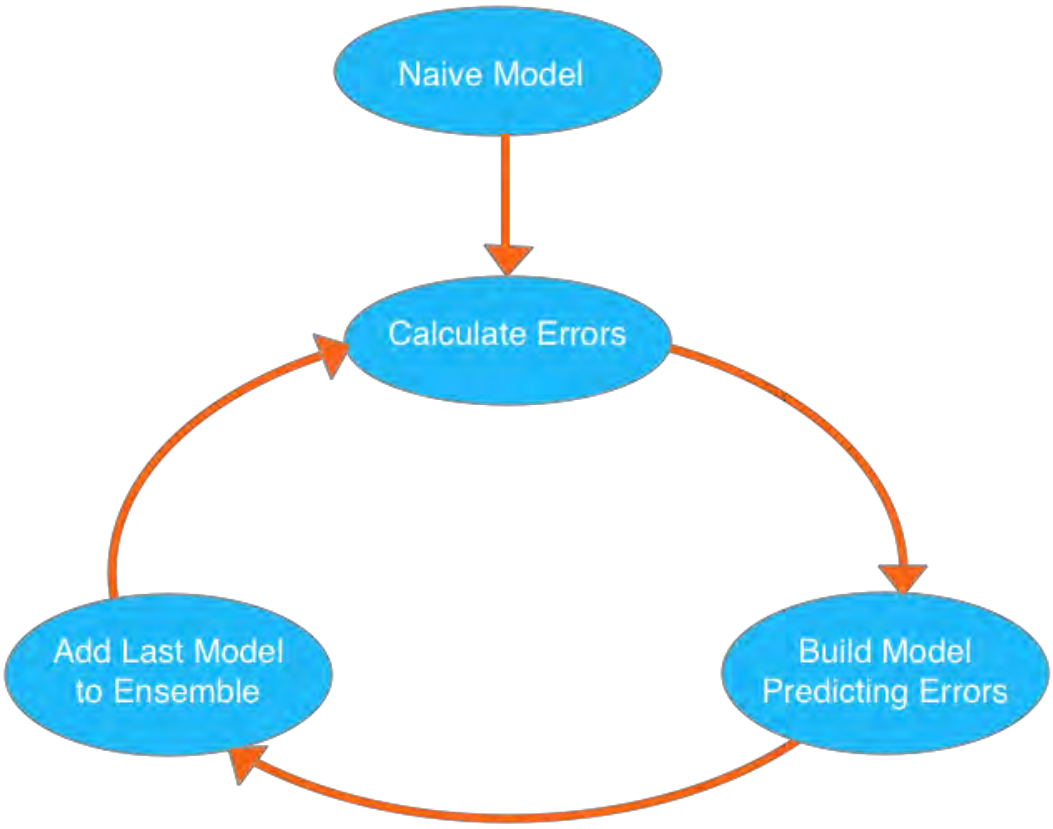

Ilustración 74. Diagrama del algoritmo Gradient Boost Decision Trees. Fuente: https://www.kaggle.com/dansbecker/xgboost

6.3 XGBoost

En este proyecto, se ha utilizado XGBoost para realizar el análisis predictivo mediante *gradient boosting*.

Ilustración 75. XGBoost. Fuente: https://www.kdnuggets.com/

XGBoost es una biblioteca distribuida optimizada de potenciación de gradiente (*gradient boosting*) diseñada para ser altamente eficiente, flexible y portátil. Implementa algoritmos de aprendizaje automático en el marco de *Gradient Boosting*. *XGBoost* proporciona un *boosting* de árbol paralelizado (también conocido como GBDT, GBM) que resuelve muchos problemas de ciencia de datos de una manera rápida y precisa. El mismo código se ejecuta en los principales entornos distribuidos (Hadoop, SGE, MPI) y puede resolver problemas de más de miles de millones de ejemplos. Recientemente, ha ganado mucha popularidad y atención como el algoritmo elegido por muchos equipos ganadores de competiciones de aprendizaje automático, actualmente es el modelo líder para trabajar con datos tabulares estándar, como los del conjunto de datos de que se dispone en este proyecto.

6.3.1 Codificación One-hot

Para poder incluir el importante atributo de la categoría base del artículo/noticia (local, deportes, sociedad, etc.) en el análisis predictivo que se realiza con un objeto regresor de *XGBoost*, se ha realizado la tarea previa de codificar estos atributos categóricos en formato cadena de caracteres con el método ***One-Hot***.

Para ello, y más concretamente, se ha aprovechado que los DataFrame de la librería pandas de Python tienen la función ***get_dummies()***, que retorna los valores de una columna codificados en varias columnas, que equivale a aplicar el método *One-Hot* sobre esta columna. Esto se puede observar en la Ilustración 76. Codificación One-Hot de las categorías base de los artículos/noticias del portal web..

newsId	newsBaseCategory_Actualidad	newsBaseCategory_Cultura	newsBaseCategory_Defensa	newsBaseCategory_Deportes
546079	0	0	0	0
545998	0	0	0	0
547538	0	0	0	1
515904	0	0	0	0
546058	0	1	0	0

Ilustración 76. Codificación One-Hot de las categorías base de los artículos/noticias del portal web.

6.3.2 Optimización de hiperparámetros

XGBoost tiene algunos parámetros que pueden afectar drásticamente a la precisión y a la velocidad de entrenamiento del modelo. Los primeros parámetros que se deben tener en cuenta son:

n_estimators y **early_stopping_rounds**

n_estimators especifica cuántas veces se pasa por el ciclo de modelado descrito anteriormente.

En el gráfico de sobreajuste vs infraajuste, *n_estimators* lo mueve más a la derecha. Un valor demasiado bajo provoca un ajuste insuficiente, que son predicciones inexactas tanto en los datos de entrenamiento como en los datos nuevos. Un valor demasiado grande provoca un sobreajuste, que son predicciones precisas sobre los datos de entrenamiento, pero predicciones inexactas sobre los datos nuevos (que es lo que importa). Los valores típicos oscilan entre 100 y 1000.

El argumento *early_stopping_rounds* ofrece una forma de encontrar automáticamente el valor ideal. La detención anticipada hace que el modelo deje de iterar cuando el resultado de validación deja de mejorar, incluso si no se está en la parada difícil para *n_estimators*. Es inteligente establecer un valor alto para *n_estimators* y luego usar *early_stopping_rounds* para encontrar el momento óptimo para dejar de iterar.

Dado que la probabilidad aleatoria a veces causa una sola ronda en la que los resultados de validación no mejoran, se debe especificar un número de cuántas rondas de empeoramiento directo se permiten antes de detenerse. *early_stopping_rounds* = 5 es un valor razonable que indica que el proceso se detendrá después de 5 rondas seguidas de empeoramiento de los resultados de validación.

learning_rate

En general, una tasa de aprendizaje (*learning_rate*) de valor pequeño (y una gran cantidad de estimadores) producirá modelos *XGBoost* más precisos, aunque también llevará más tiempo entrenar el modelo, ya que realiza más iteraciones a lo largo del ciclo.

n_jobs

En conjuntos de datos más grandes donde el tiempo de ejecución es una consideración, se puede usar la paralelización para construir los modelos más rápido. Es común establecer el parámetro *n_jobs* igual al número de núcleos en la máquina. En conjuntos de datos más pequeños, esto no ayudará.

XGBoost tiene una multitud de otros parámetros, pero estos son de los más relevantes a la hora de ajustar el modelo *XGBoost* para un rendimiento óptimo.

Existen numerosas técnicas para realizar la optimización o ajuste de estos y otros hiperparámetros de *XGBoost*. Para el caso de este proyecto, se ha empleado la técnica de ***grid search***, que no es más que una simple búsqueda exhaustiva a través de un subconjunto especificado manualmente del espacio de hiperparámetros. Para validar sus resultados de utiliza una **validación cruzada** con los subconjuntos de datos seleccionados. Estas funcionalidades vienen en el paquete/objeto de *XGBoost* de scikit-learn.

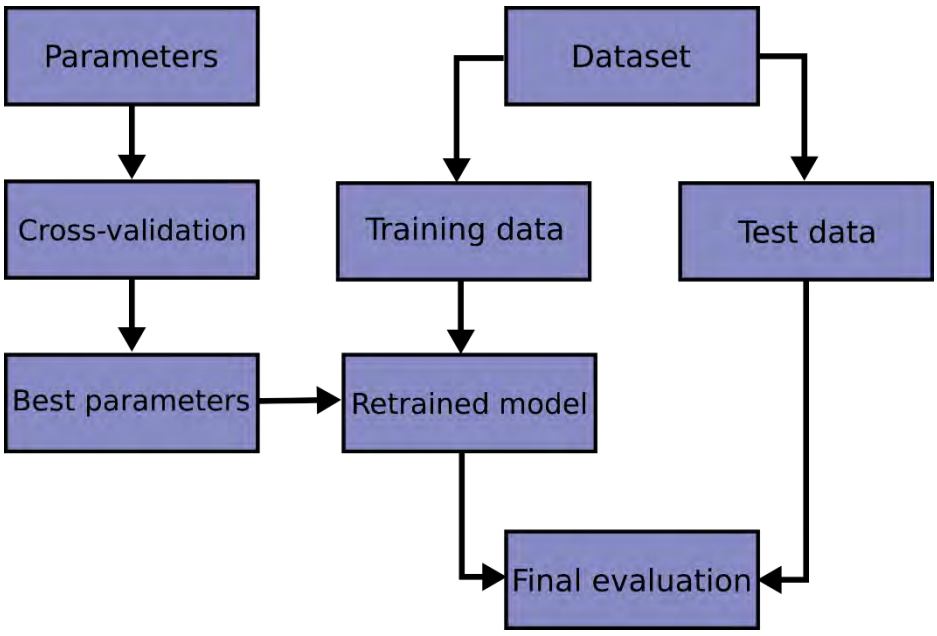

Ilustración 77. Esquema de la Validación Cruzada. Fuente: https://scikit-learn.org/stable/modules/

6.4 Predicción del nº final de visitas a página

En esta sección, se presenta la aplicación del objeto regresor de *XGBoost* para la predicción del nº final de visitas a las páginas de los artículos/noticias del portal web en función de sus características iniciales o atributos base.

En la Ilustración 78. Gráfico de dispersión de visitas a página finales reales (X) vs. predichas (Y). se observa el resultado de las predicciones en forma de gráfico de dispersión.

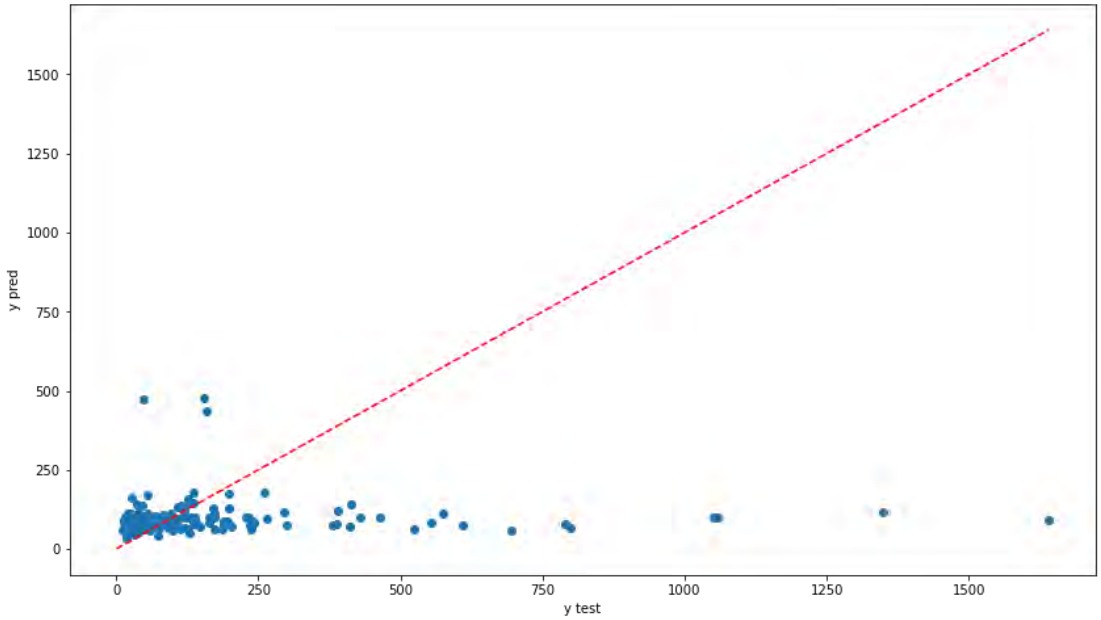

Ilustración 78. Gráfico de dispersión de visitas a página finales reales (X) vs. predichas (Y).

En la Ilustración 79 se observa la importancia relativa que ha tenido cada una de las características o atributos base en la predicción del número final de visitas a página.

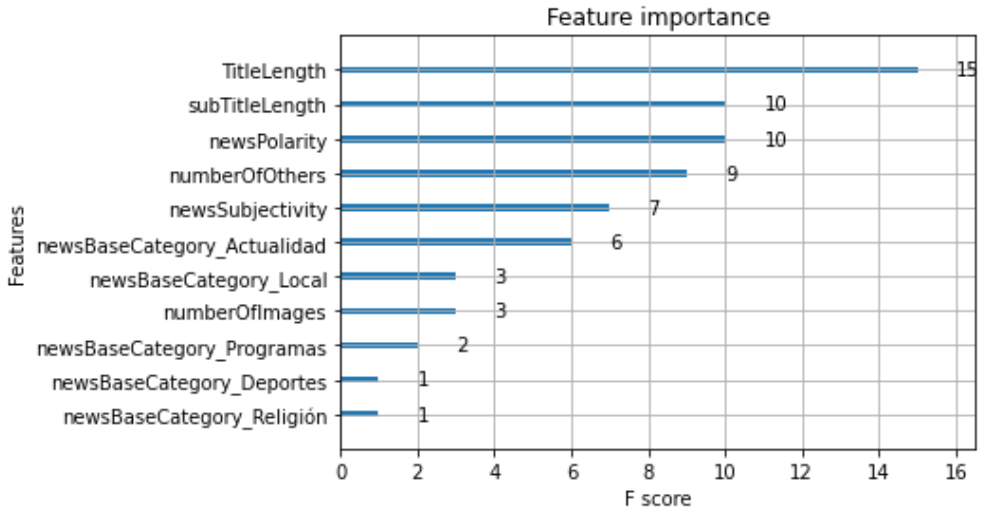

Ilustración 79. Importancia de las diferentes características en la predicción de las visitas a página finales.

Se puede concluir, en base a los resultados que se observan en las imágenes anteriores, que no es posible predecir con una mínima precisión el número final de visitas a las páginas de los artículos/noticias

a) en base a las características base o atributos disponibles,

b) utilizando la técnica de *boosting* del gradiente (*XGBoost*), y

c) con la optimización de hiperparámetros mediante *grid search* con validación cruzada.

Por otro lado, atendiendo a los resultados de importancia de las características base, se puede aventurar que:

- la longitud de título y subtítulo del artículo/noticia son factores importantes que influyen en el número final de visitas a su página.
- la presencia o no de vídeos y/o audios en la página es más determinante que la presencia de imágenes.
- la categoría base de "Actualidad", a falta de un atributo numérico que lo cuantifique, es un factor determinante en el número final de visitas a página.

6.5 Predicción del tiempo medio en página final

En esta sección, se presenta la aplicación del objeto regresor de *XGBoost* para la predicción del valor final tiempo medio que pasan los visitantes en la página de los artículos/noticias del portal web en función de sus características o atributos base.

En la Ilustración 80 se observa el resultado de la predicción en un gráfico de dispersión.

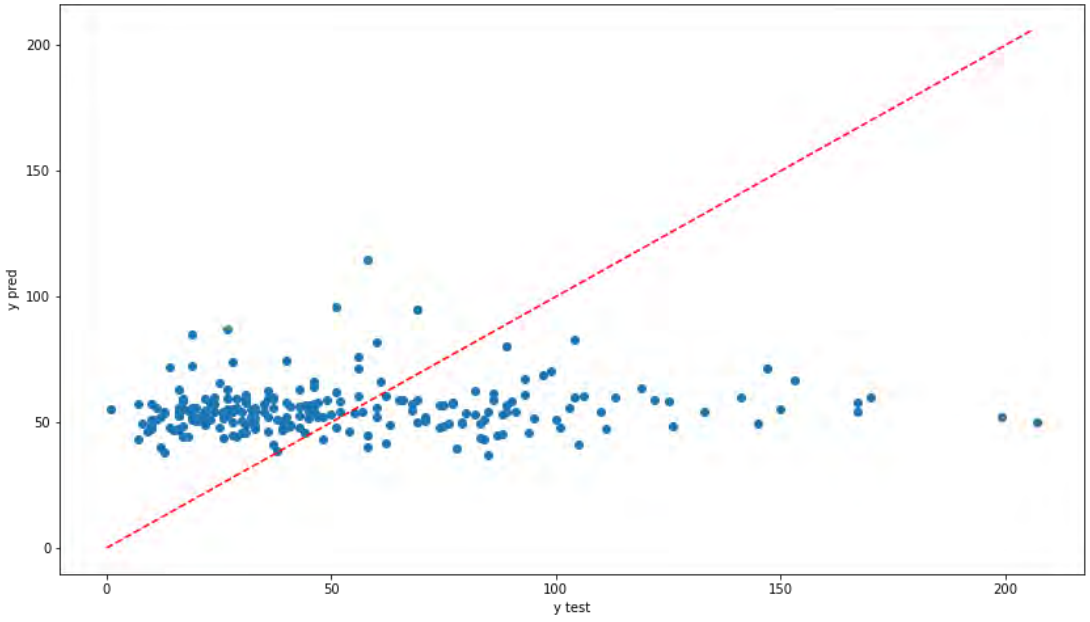

Ilustración 80. Gráfico de dispersión del valor final del tiempo medio en página real (X) vs. predicho (Y).

En la Ilustración 81 se observa la importancia relativa que ha tenido cada una de las características iniciales o atributos base en la predicción del valor final del tiempo medio en cada página.

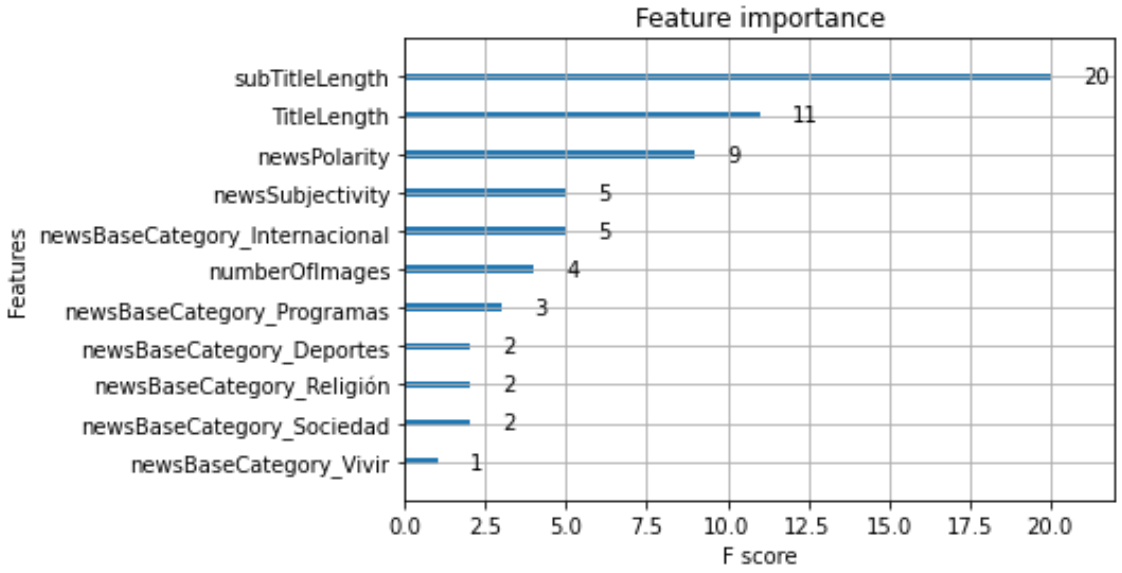

Ilustración 81. Importancia de las diferentes características en la predicción del valor final del tiempo medio en página.

Se puede concluir, en base a los resultados que se observan en las imágenes anteriores, que no es posible predecir con una mínima precisión el valor final del tiempo medio que pasan los visitantes en la página de los artículos/noticias.

6.6 Clasificación de clústeres de rendimiento

Adicionalmente, se ha realizado un estudio de la capacidad de *XGBoost* de clasificar los artículos/noticias del portal web en los **7 clústeres** de rendimiento obtenidos con el algoritmo de clusterización **K-Means**, presentados y explicados en el capítulo 5. Es decir, se ha probado el **objeto clasificador de *XGBoost*** en base a las características iniciales o atributos base de los artículos/noticias.

Primeramente, se ha aplicado el objeto clasificador de *XGBoost* con los hiperparámetros por defecto, y el resultado que se ha obtenido es que se han clasificado en su clúster de rendimiento correcto el **33%** de los artículos/noticias. Una serie de ajustes manuales de los hiperparámetros no han cambiado en demasía este resultado.

En vista de este pobre resultado, se ha decidido aplicar la técnica de *grid search* con validación cruzada para llegar a los valores óptimos de varios hiperparámetros importantes. En la Ilustración 82 se observan los fragmentos de código en Python que representan la aplicación de esta técnica.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 33)
```

```
xgb_clas = xgb.XGBClassifier()
```

```
parameters = {
     "eta"    : [0.05, 0.10, 0.15, 0.20, 0.25, 0.30 ] ,
     "max_depth"        : [ 3, 4, 5, 6, 8, 10, 12, 15],
     "min_child_weight" : [ 1, 3, 5, 7 ],
     "gamma"            : [ 0.0, 0.1, 0.2 , 0.3, 0.4 ],
     "colsample_bytree" : [ 0.3, 0.4, 0.5 , 0.7 ]
     }

grid = GridSearchCV(xgb_clas,
                    parameters, n_jobs=4,
                    scoring="neg_log_loss",
                    cv=3)

grid.fit(X_train, y_train)
```

Ilustración 82. Código de la técnica de grid search con validación cruzada.

El resultado de la clasificación con los hiperparámetros optimizados según esta técnica es prácticamente el mismo: se clasifican en su clúster de rendimiento correcto el **34,33%** de los artículos/noticias.

Así pues, se concluye que **no** es posible clasificar correctamente los artículos/noticias en los clústeres de rendimiento en base a los atributos disponibles. Se concluye que la técnica de optimización de los hiperparámetros del modelo clasificador no tiene efecto práctico en este caso.

6.7 Resultados y observaciones

Antes de entrar en el detalle del análisis de los resultados de las predicciones, se considera oportuno recalcar la dificultad intrínseca que tiene realizar predicciones como estas, con tanto potencial, de forma precisa y consistente. En la comunidad de la ciencia de datos y el aprendizaje automático se invierte mucho en encontrar modelos predictivos que funcionan en diferentes grados, sobre diferentes conjuntos datos y de diferentes maneras.

En este caso, a partir de los resultados que se muestran en las imágenes de las secciones anteriores, se llega a la conclusión de que las predicciones **no** son correctas en ningún caso. Esto se puede deber a varios factores:

- Falta de características iniciales o atributos base que pueden ser realmente importantes a la hora de influir en el rendimiento final de un artículo/noticia. Ejemplos: número de **comentarios**, número de comparticiones en **redes sociales**, autor/a, parámetros de análisis de sentimiento del cuerpo del texto (polaridad, subjetividad), atributo de **actualidad**, etc.

- Aplicación incorrecta del modelo y/o **optimización de hiperparámetros incompleta**. En términos de programación, es posible que la definición de los objetos regresor y clasificador, o el ajuste de los mejores hiperparámetros del modelo del objeto no hayan sido sujetos al nivel de revisión iterativa que seguramente requieren.

7 Limitaciones

Como ya se ha detallado en algún capítulo, antes del inicio y a lo largo del desarrollo de este proyecto se han manifestado una serie de limitaciones y restricciones de diversa índole, que a continuación se presentan y explican.

7.1 Limitaciones en el alcance

En primer lugar, han existido una serie de condicionantes que han restringido el alcance de este proyecto. Estos condicionantes son:

a. **Conjunto de datos**:

   Los parámetros que se capturan con la herramienta de analítica web son los que son, y, en este sentido, el conjunto de datos original a partir del cual se ha comenzado el desarrollo no contiene toda la información sobre los artículos/noticias del portal web que quizás se necesitaba para realizar algunos análisis como el de regresión.

   Además, en el conjunto de datos había cantidad de objetos con datos incompletos o irrelevantes.

b. **Capacidad predictiva**:

   La capacidad predictiva de los modelos algorítmicos de aprendizaje automático supervisado es una limitación clara al alcance del proyecto. A la hora de realizar el análisis predictivo del rendimiento de los artículos/noticias, se han encontrado dificultades conceptuales, técnicas y temporales que no fueron contempladas en un inicio. La capacidad de predecir el rendimiento futuro de un determinado objeto es algo tan interesante, y, a su vez, normalmente tan complejo de adquirir que, en muchas ocasiones, constituye todo el alcance de una determinada investigación o desarrollo, y no es el caso de este proyecto, que tiene un alcance más abierto, aunque no tan profundo.

7.2 Limitaciones técnicas

En segundo lugar, ha habido una serie de limitaciones técnicas que no se contemplaron en su totalidad a la hora de definir los objetivos y el alcance de este proyecto. Se presentan y explican a continuación:

a. **Capacidad de computación**:

Cabe destacar que un estudio como este tiene un alcance que no es comparable al de los estudios de audiencia/de lectores que realizan los grandes grupos de medios de comunicación en línea, de prensa y radio. Con un ordenador que no disponga de componentes con una gran capacidad computacional, existe una limitación técnica a la hora de operar con grandes volúmenes de datos.

b. **Optimización de hiperparámetros**:

Se han encontrado problemas técnicos y de computación a la hora de realizar la optimización de los hiperparámetros de los algoritmos de clusterización y de *boosting* de gradiente. Se han probado técnicas "automáticas" como *grid search* y también se han probado ajustes manuales, pero no se han obtenido resultados satisfactorios.

7.3 Limitaciones temporales y coyunturales

Por último, se repasan brevemente las limitaciones en tiempo y en contexto coyuntural que se han encontrado durante el desarrollo del proyecto:

a. **Plazos académicos**:

Es evidente que las restricciones temporales marcadas por los plazos de entrega de la universidad son una limitación en la duración y el desarrollo del proyecto.

b. **Crisis de COVID-19**:

La crisis sociosanitaria derivada de la pandemia de COVID-19 ha resultado ser una limitación del todo imprevista, que ha tenido un impacto moderado sobre el desarrollo de este proyecto.

8 Conclusiones y líneas de futuro

A medida que el desarrollo de este proyecto llega a su fin, es el momento de sacar algunas conclusiones y comprobar si se alcanzaron los objetivos iniciales.

El objetivo de realizar una pequeña contribución independiente al proyecto principal *SmartData* se cumple con el desarrollo de este proyecto, al tratarse de una ampliación del Trabajo Final de Grado, que ya se concibió y consideró de ese modo.

A nivel personal, los objetivos marcados de aprendizaje, gestión de datos y de gestión de proyectos se han alcanzado con creces, una vez finalizado el desarrollo del proyecto.

Un objetivo académico de este proyecto era comprobar que al menos uno de los cinco modelos estadísticos es adecuado para modelar la evolución de las páginas vistas de los artículos/noticias online. Los resultados obtenidos en el cuarto capítulo indican que sí, los cinco modelos estadísticos se pueden utilizar para modelar la evolución temporal de las visitas a página acumuladas. Todos los modelos estadísticos analizados obtienen muy buenos resultados cuando se observan los criterios de selección del modelo.

Se concluye que los resultados de la aplicación de técnicas de clusterización y agrupamiento según los parámetros de rendimiento son moderadamente satisfactorios y definitivamente utilizables, aunque no se hayan encontrado relaciones directas ni descubierto patrones ocultos entre las características base y los parámetros de rendimiento. Sin embargo, algunos de los resultados obtenidos sí pueden servir de indicio de estas relaciones y patrones.
De las dos técnicas de clusterización aplicadas, la que ofrece resultados más fácilmente interpretables y utilizables es la clusterización basada en centroides. La clusterización basada en densidades ha resultado mucho más compleja de parametrizar, y sus resultados más difíciles de manejar.

En marcado contraste con los resultados explicados en el anterior párrafo, los resultados de la aplicación de técnicas de predicción y de clasificación mediante *boosting* de gradiente son insatisfactorios: no se ha conseguido predecir con un mínimo de precisión el rendimiento de los artículos/noticias a partir de sus características o atributos base. No se puede descartar ninguna hipótesis sobre la razón de ello. Puede ser que únicamente se deba que no se incluyen (porque no los hay) atributos base que podrían tener una gran influencia. Puede ser que la técnica de boosting de gradiente no sea la más adecuada para el conjunto de datos con que se ha trabajado y el tipo de predicción deseado, aunque esto es relativamente improbable, pues el paquete XGBoost es conocido por su versatilidad, rapidez y eficacia. Puede deberse exclusivamente al método utilizado para la optimización de los hiperparámetros de los objetos regresores y clasificadores utilizados.

Los resultados obtenidos aquí para el portal web de radio y noticias que proporcionó el conjunto de datos podrían extenderse a los casos de otros portales web similares, con

secciones similares y una base de visitantes similar. Para ello habría que hacer un estudio previo de la demografía y de los modelos de negocio.

A nivel general, sí que se ha cumplido el objetivo de obtener de un conjunto de datos información que podría ser útil, como el indicio de la influencia de ciertas características base en el rendimiento final, o la validez de ciertos modelos estadísticos para ajustar ciertas curvas temporales, o la constatación de que hay atributos base que tienen más presencia en los grupos de objetos con mejor rendimiento final.

Nota del autor: Estoy muy contento con el desarrollo y el resultado de este proyecto de investigación que me sirve como tesis final. He aprendido mucho sobre diversos temas, desde modelos estadísticos hasta métodos de clusterización, a medida que avanzaba el proyecto. En estas últimas líneas, quisiera agradecer una vez más a mi mentor y tutor, el Dr. Xavier Vilasís Cardona, así como a todos los que han hecho posible el desarrollo de este proyecto.

8.1 Líneas de futuro

Antes de describir las posibles soluciones y mejoras que se pueden hacer, hay que destacar que la manera de encarar y definir la estrategia para resolver un estudio de este tipo puede hacer variar el resultado significativamente. Sin embargo, hay unas consideraciones generales y más en concreto sobre este que se hacen a continuación, para futuras líneas de investigación.

- Dedicar más recursos para poder obtener unos datos más amplios, tanto en filas como sobre todo en columnas (atributos). El hecho de disponer de más datos, como por ejemplo el número de comentarios o de comparticiones en redes sociales, permite, por un lado, diseñar modelos de ML que de otro modo no son posibles, por la poca cantidad de datos, como, por ejemplo, redes neuronales; y, por otro, mejorar los resultados de las predicciones de rendimiento mediante análisis de regresión y clasificación basados en los atributos base.

- Realizar un Análisis de Sentimiento (AS) completo y sólido tanto del par título-subtítulo como del cuerpo del texto del artículo/noticia. Así, se podría, por un lado, mejorar el estudio de la situación “pre-clic”, que incluye el AS del par título-subtítulo; y, por el otro, realizar un estudio de la situación “post-clic”, que puede incluir el AS del cuerpo del texto del artículo/noticia.

- Diseñar un algoritmo de recategorización automática. Para casos como el encontrado en este proyecto, donde los artículos/noticias tienen un atributo de categoría demasiado específico (subcategoría) que impide su uso fácil en cualquier análisis o algoritmo. Se debería realizar un análisis de todos los textos, crear modelos de texto basados en palabras/conceptos base, etc.

- Incluir un estudio de la robustez de los modelos estadísticos utilizados para ajustar la curva de evolución temporal de las visitas a las páginas de los artículos/noticias, para afinar la selección de los mejores modelos. Incluir otros posibles modelos candidatos para esa tarea.

- Estudiar maneras de añadir un atributo de "Actualidad" (no confundir con la categoría) a las características base de un artículo/noticia. Requeriría de consultas a otras fuentes de noticias y de información, como por ejemplo Twitter; más un exhaustivo análisis de los textos y de las palabras clave.

- Afinar el análisis de clusterización encontrando el algoritmo que mejor funciona y parametrizándolo de la manera óptima para obtener resultados de agrupamiento interesantes.

9 Referencias

[1] Retos Colaboración 2016 - RTC-2016-5503-7

[2] Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342(6154):127–132

[3] Yucesoy B, Wang X, Junming H, Barabási A-L (2018) Success in Books: a big data approach to best-sellers. EPJ Data Science; 7:7

[4] Dhar V (2013) Data science and prediction. Communications of the ACM. 56 (12): 64–73

[5] Leek J (12 December 2013) The key word in "Data Science" is not Data, it is Science. Simply Statistics

[6] Hayashi C (1 January 1998) What is Data Science? Fundamental Concepts and a Heuristic Example. In Hayashi C; Yajima K; Bock H-H; Ohsumi N; Tanaka Y; Baba Y (eds.) Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51

[7] https://datascience.berkeley.edu/about/what-is-data-science/ Accedido: Septiembre 2020

[8] Bishop C. M. (2006) Pattern Recognition and Machine Learning. Springer

[9] cita de Andrew Ng (2014) Machine Learning (online course: https://es.coursera.org/learn/machine-learning) Accedido: Septiembre 2020

[10] James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning. Springer. p. vii.

[11] https://www.datascience.com/blog/k-means-clustering Accedido: Septiembre 2020

[12] https://stanford.edu/~cpiech/cs221/handouts/kmeans.html Accedido: Septiembre 2020

[13] Zheng J, Peltsverger S (2015) Web Analytics Overview. Encyclopedia of Information Science and Technology

[14] Bennett, L (January 10, 2012) Metrics that Matter & the Death of the Page View. Chartbeat Blog

[15] https://www.ibm.com/support/knowledgecenter/en/SS3RA7_17.0.0/clementine/nodes_statisticalmodels.html Accedido: Septiembre 2020

[16] Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

[17] Rogers E. M. (1962) Diffusion of innovations (1st ed.). New York: Free Press of Glencoe

[18] Bass, F. M. (2004) Comments on "A New Product Growth for Model Consumer Durables": The Bass Model. Management Science. 50 (12): 1833–1840.

[19] https://en.wikipedia.org/wiki/Logistic_function Accedido: Septiembre 2020

[20] Maydeu-Olivares A, García-Forero C (2010) Goodness-of-fit Testing. Universitat de Barcelona

[21] Calvo M (2018) Regresiones. Máster Universitario en Data Science, La Salle – URL

[22] David T. Mage (1984) Pseudo-Lognormal Distributions, Journal of the Air Pollution Control Association, 34:4, 374-376

[23] van Gelder P.H.A.J.M., Beijer G, Berger M. Statistical analysis of pageviews on web sites. Delft University of Technology

[24] Sobre la clusterización K-means: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html Accedido: Septiembre 2020

[25] Cynthia Rudin. 15.097 Prediction: Machine Learning and Statistics. Spring 2012. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

[26] Cox, D. R. (2006) Principles of Statistical Inference, Cambridge University Press.

[27] Strutz, T. (2016) Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond). 2nd edition, Springer Vieweg

[28] Sobre el modelo de Bass: www.bassbasement.org/ Accedido: Septiembre 2020

[29] Sobre la aplicación de modelos estadísticos sobre conjuntos de datos: https://towardsdatascience.com Accedido: Septiembre 2020

[30] Sobre la construcción de modelos predictivos: https://medium.com Accedido: Septiembre 2020

[31] Kumar, Naresh. (2015). Review of Innovation Diffusion Models. 10.13140/RG.2.1.2413.0728. Accedido: Septiembre 2020

[32] Sobre inteligencia artificial: https://www.britannica.com/technology/artificial-intelligence Accedido: Septiembre 2020

[33] Sobre análisis de clusterización: https://en.wikipedia.org/wiki/Cluster_analysis Accedido: Septiembre 2020

[34] Sobre analítica editorial: https://www.niemanlab.org/2016/02/the-next-step-moving-from-generic-analytics-to-editorial-analytics/ Accedido: Septiembre 2020

[35] Sobre estudio de Pew Research Center: https://www.journalism.org/fact-sheet/newspapers/ Accedido: Septiembre 2020

[36] Sobre el informe de Technavio: https://www.businesswire.com/news/home/20200113005358/en/Global-Web-Analytics-Market-2020-2024-19-CAGR Accedido: Septiembre 2020

[37] Sobre el estudio de Gartner: https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020/ Accedido: Septiembre 2020

[38] Sobre limpieza de datos: Rahm, E., Hong, H. Data Cleaning: Problems and Current Approaches, University of Leipzig, Germany, 2000.

[39] Rogers, Everett (16 August 2003). Diffusion of Innovations, 5th Edition. Simon and Schuster. ISBN 978-0-7432-5823-4.

[40] Sobre técnicas de clusterización: http://www.stat.columbia.edu/ Accedido: Septiembre 2020

[41] Sobre DBSCAN: "Introduction to Machine Learning with Python", de Andreas C. Müller & Sarah Guido.

[42] Sobre análisis de regresión: https://es.wikipedia.org/wiki/An%C3%A1lisis_de_la_regresi%C3%B3n, https://hbr.org/2015/11/a-refresher-on-regression-analysis Accedidos: Septiembre 2020

[43] Sobre GBM: http://web.mit.edu/haihao/www/papers/RGBM.pdf Accedido: Septiembre 2020