Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

GRADO EN INGENIERÍA MATEMÁTICA APLICADA AL ANÁLISIS DE DATOS

PROYECTO FIN DE GRADO

Modelo de Datos Abiertos para un Análisis Comercial de la Ciudad de Madrid

María Fernández Morín

Dirigido por

Álvaro Sánchez Pérez

CURSO 2021-2022

TÍTULO: Modelo de Datos Abiertos para un Análisis Comercial de la Ciudad de Madrid

AUTOR: María Fernández Morín

TITULACIÓN: Grado en Ingeniería Matemática aplicada al Análisis de Datos

DIRECTOR/ES DEL PROYECTO: Álvaro Sánchez Pérez

FECHA: Junio de 2022

RESUMEN

La pequeña y mediana empresa (PYME) sustenta el mayor porcentaje del tejido empresarial español. Si bien somos un país de emprendedores, a medida que los años pasan el temor a emprender se acrecienta entre los ciudadanos. Por esta razón, fomentar el emprendimiento es una de las tareas pendientes de gobiernos y ayuntamientos para no perder el pilar básico de nuestra economía.

En el presente proyecto se plantea aportar a todos aquellos inversores interesados en iniciar un negocio una ayuda que dé seguridad a su sentimiento emprendedor, de tal forma que se posibilita a cualquier usuario reducir su miedo a invertir sus ahorros mediante la consulta de un pronóstico de éxito.

La propuesta desarrollada consiste en la creación de un modelo de datos relativos al censo de locales que cuente con un análisis y representación de la actividad comercial de la ciudad de Madrid. El objetivo será lograr un sistema que automatice la estandarización de los registros publicados por el Ayuntamiento y que, a su vez, nos facilite configurar un modelo de aprendizaje por medio de técnicas *machine learning*, de modo que su propósito será encontrar un patrón entre la duración de los establecimientos y sus respectivas localizaciones y actividades económicas. Esto permitirá ofrecer a un potencial cliente la probabilidad de éxito del tipo de negocio que está interesado en emprender en una ubicación concreta.

Palabras clave: *datos abiertos, machine learning, ETL, normalización, cuadro de mando, emprendimiento, actividad comercial, CNAE.*

ABSTRACT

Small and medium-sized enterprises (SMEs) make up the largest percentage of the Spanish business fabric. Although we are a country of entrepreneurs, as the years go by the fear of entrepreneurship increases among citizens. For this reason, encouraging business development is one of the pending tasks of governments and city councils in order not to lose the basic pillar of our economy.

This project proposes to provide all those investors interested in starting a business with a help that gives security to their entrepreneurial feeling, in such a way that any user can reduce their fear of investing their savings by consulting a success forecast.

The developed proposal consists of the creation of a data model related to the census of premises with an analysis and representation of the commercial activity in the city of Madrid. The objective will be to achieve a system that automates the standardization of the data published by the City Council and that, in turn, will allow us to configure a machine learning model, thus its purpose will be to find a pattern between the duration of the establishments and their respective locations and economic activities. This will enable us to offer a potential client the probability of success of the type of business he is interested in undertaking in a specific location.

Keywords: *open data, machine learning, ETL, normalization, dashboard, entrepreneurship, business activities, CNAE.*

AGRADECIMIENTOS

En primer lugar, agradecer a Álvaro su enseñanza, confianza y guía constante en el proyecto. Me ha transmitido la capacidad de ver el dato desde todas las perspectivas posibles y ha sido un profesor ejemplar.

En segundo lugar, transmitir mi más sincero agradecimiento a todas aquellas personas que han estado presentes en mis años de carrera. A mis compañeros y amigos por motivarme día a día, a Pablo, por animarme siempre a seguir adelante y a mi familia por llegar hasta aquí conmigo.

Mención especial a mi madre por enseñarme la importancia de la formación y ser el mejor ejemplo de dedicación y constancia al que dirigirse. Por estar en todo y por su paciencia y apoyo diarios.

A todos ellos, gracias.

La guerra es 90% información.

Napoleón Bonaparte

TABLA RESUMEN

	DATOS
Nombre y apellidos:	María Fernández Morín
Título del proyecto:	Modelo de Datos Abiertos para un Análisis Comercial de la Ciudad de Madrid
Directores del proyecto:	Álvaro Sánchez Pérez
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto: (esta entrada se puede marcar junto a la siguiente)	NO
El proyecto ha consistido en el desarrollo de una investigación o innovación: (esta entrada se puede marcar junto a la anterior)	SI
Objetivo general del proyecto:	Creación de un modelo de datos que muestre la probabilidad de éxito de un determinado negocio en una localización concreta de la ciudad de Madrid.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

Si todas las PYMES echaran el cierre, el PIB y la economía mundial caerían en picado. No solo son el pilar de la economía, sino que, según datos de la ONU, generan entre el 60% y 70% del empleo y producen el 50% del PIB mundial. Su pequeña estructura les otorga muchos beneficios frente al gran negocio, como puede ser la adaptación rápida a los cambios, la especialización en nichos de mercado, una menor complicación del negocio…, pero la lista no es precisamente larga.

En el territorio nacional, representan el 99% del tejido empresarial. A pesar de su gran importancia en la economía española, el temor a emprender avanza sin pausa cada año. El emprendedor medio considera que no cuentan con los apoyos suficientes para desarrollar su negocio. Por ello, la problemática a la que se enfrentan hoy en día las instituciones madrileñas es conseguir dar con nuevas técnicas para fomentar el emprendimiento.

El lanzamiento de un negocio es, sin duda, todo un reto, pero una modernización del sector a través de la tecnología y la transformación digital permitirá entrever un futuro más brillante.

1.2 Planteamiento del problema

El presente proyecto se enfoca en resolver un problema específico empresarial, lo cual se realiza por medio de la implementación de un innovador sistema de modelado de datos para obtener un análisis comercial automatizado en una localización concreta.

Se pretende ofrecer un servicio de recomendación a todos aquellos inversores interesados en entrar en el mundo del emprendimiento. Se plantea aportar a todos ellos una ayuda que favorezca su sentimiento emprendedor, de forma que un usuario podrá reducir su temor a inaugurar un comercio en cualquier emplazamiento de la ciudad mediante la consulta de un pronóstico de éxito, basado en el histórico de otros establecimientos del mismo tipo.

1.3 Objetivos del proyecto

Haciendo uso de un preprocesamiento exhaustivo de datos relativos al censo de locales de la ciudad de Madrid, una posterior creación de un modelo de aprendizaje a través de técnicas *machine learning*, la creación de una visualización de datos y una interfaz de usuario, el objetivo será conseguir obtener un sistema que automatice la limpieza de los datos publicados por el Ayuntamiento de Madrid y un modelo basado en datos abiertos capaz de ofrecer al cliente la probabilidad de éxito del tipo de negocio que está interesado en emprender en una localización concreta.

1.4 Resultados obtenidos

A través del presente proyecto se alcanzan tres soluciones diferentes, todas ellas relacionadas. En primer lugar, se obtiene un proceso automatizado de homogenización de los ficheros de datos relativos al censo de locales y actividades de la ciudad de Madrid, junto con un cuadro de mando encargado de mostrar el análisis comercial actual de la ciudad. Finalmente, se elabora un modelo de *machine learning* basado en el procesamiento de los ficheros mencionados, el cual se integra en una interfaz de usuario cuya finalidad es permitir a un potencial inversor interactuar con ella, de forma que sea de utilidad en la toma de decisiones sobre la localización y tipo de negocio a emprender en la capital.

1.5 Estructura de la memoria

- Capítulo 1: Resumen. Se realiza una breve introducción y resumen del presente proyecto. Se expone la razón principal de la idea propuesta y se sintetizan los medios para obtener el resultado deseado.
- Capítulo 2: Estado del Arte. Exposición exhaustiva del hilo conductor de la historia que ha llevado a la generación e ideación del trabajo.
- Capítulo 3: Objetivos. Se elabora un listado de los objetivos a alcanzar a lo largo del proyecto para llegar a lograr el resultado final.
- Capítulo 4: Desarrollo del proyecto. Se trata del capítulo más extenso. En él se detalla la planificación seguida y todos los recursos, herramientas y pasos necesarios para realizar el proyecto. Consiste en el manual que indica todo lo necesario para llegar al fin último del trabajo. En añadido, se presentan los diferentes resultados obtenidos en el proceso.
- Capítulo 5: Discusión. Se discuten los resultados y el camino seguido a lo largo del proyecto para llegar hasta los resultados obtenidos.
- Capítulo 6: Conclusiones. Es el punto final del proyecto realizado. En él se enumeran las conclusiones obtenidas a partir del trabajo realizado junto con las conclusiones personales tras la finalización del mismo.
- Capítulo 7: Futuras líneas de trabajo. En este caso se detallan los diferentes caminos a seguir para optimizar los resultados y convertir el proyecto en un producto real.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

2.1.1 Definición de Comercio Minorista

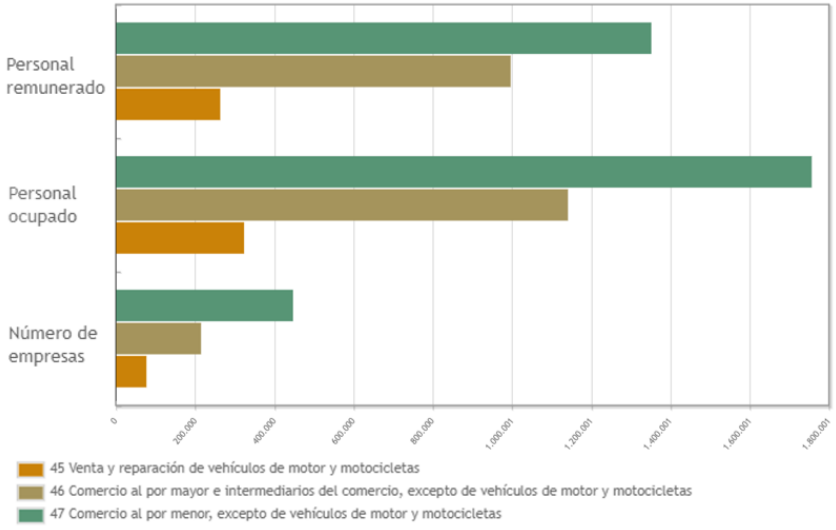
El comercio minorista forma una parte esencial del sistema productivo de cualquier economía desarrollada. Se encuentra dentro de las actividades incluidas en el sector servicios, más concretamente en la rama relativa al comercio. Su característica principal es que se trata del último eslabón de la cadena de distribución de un producto: es el sector que vende los productos al cliente final. La Asociación General de Consumidores define la actividad de las empresas de comercio minorista como aquella actividad profesional, desarrollada profesionalmente con ánimo de lucro, consistente en ofertar la venta de cualquier clase de artículos a los destinatarios finales de los mismos, utilizando o no un establecimiento [5].

2.1.2 La importancia del comercio minorista en la economía española

La creciente relevancia de las actividades terciarias supone uno de los elementos clave en el funcionamiento del aparato productivo nacional. El sector servicios en nuestra economía ocupa el mayor porcentaje de población activa ocupada, aproximadamente un 70% de la población empleada está en este sector. Es el pilar básico y el que sostiene la economía española [3].

El comercio minorista es el de mayor peso en el sector servicios español representando un 44% del valor añadido bruto (VAB) del total [3], con gran relevancia respecto a número de locales y empresas y personal ocupado (Ilustración 1). Su presencia se extiende por todo el territorio nacional siendo a su vez también el primer sector generador de empleo, el cual en un 77,5% es empleo fijo [4].

Ilustración 1. Principales magnitudes según actividad principal (CNAE-2009)
Fuente: elaboración a través de los datos de Estadística estructural de empresas: sector comercio (INE)



En términos generales, representa el 4,2% del VAB, lo que le consolida en una actividad muy importante para la economía española y que ha venido creciendo con fuerza en los últimos años (+3,2% anual promedio entre 2014 y 2018). Además, su importancia es algo mayor que en otras economías europeas con una estructura económica similar (3,9% en Alemania y en Francia, y 3,5% en Italia) y también frente a la media de la UE (3,9%) [1].

2.1.3 Ecosistema emprendedor en España

El emprendimiento se refiere al concepto de desarrollar y gestionar una empresa con el fin de obtener beneficios asumiendo diferentes riesgos en el mundo empresarial. En su nivel más básico, el emprendimiento se refiere a un individuo o a un pequeño grupo de socios que emprenden un camino original para crear un nuevo negocio. Un aspirante a empresario busca activamente una determinada aventura empresarial y es el empresario quien asume la mayor cantidad de riesgos asociados al proyecto. El emprendimiento ha desempeñado un papel vital en el desarrollo económico del mercado global en expansión [6].

Un ecosistema emprendedor es un entorno que facilita el surgimiento de nuevos negocios creando conexiones valiosas con otras empresas, instituciones o inversionistas para desarrollar ideas innovadoras que se puedan capitalizar [7].

Un ecosistema emprendedor combina elementos sociales, políticos, económicos y culturales para crear un entorno físico o digital en el que interactúan sus distintos actores para fortalecer distintos tipos de emprendimiento [7].

En la época de transformación digital que estamos viviendo, una gran cantidad de autónomos emprendedores han visto en la tecnología una oportunidad para crear su propia empresa. Crear una empresa no es una tarea fácil. Además de la idea innovadora, se requiere perseverancia, planificación y organización. La ventaja de montar un negocio propio actualmente es la ayuda que nos proporcionan las diferentes herramientas tecnológicas que, más allá de facilitar la administración, pueden garantizar el acercamiento al éxito.

2.1.4 Situación de las Pymes y autónomos en España

Pyme es el acrónimo para pequeña y mediana empresa, que incluye a todas aquellas empresas con menos de 250 trabajadores y una facturación anual inferior a 50 millones de euros [24].

Entre las pymes se distinguen tres tipos de empresa según los siguientes criterios: el número de trabajadores, el volumen de negocio anual y el balance general anual [24]:

- Microempresa. Tiene menos de 10 trabajadores, factura menos de 2 millones de euros al año y su balance general anual es inferior a 2 millones de euros.
- Pequeña empresa. Tiene entre 10 y 49 trabajadores, un volumen de negocio anual de entre 2 y 10 millones de euros y un balance general de entre 2 y 10 millones de euros anuales.
- Mediana empresa. Tiene entre 50 y 249 trabajadores, un volumen de negocio anual de entre 10 y 50 millones de euros y un balance general al año de entre 10 y 43 millones de euros.

A su vez un autónomo es la persona física que realiza por cuenta propia una actividad económica o profesional, fuera del ámbito de dirección de otra persona, con o sin trabajadores a su cargo [26].

En cuanto a la distribución del número de empresas por tamaño, es el siguiente:

- 53,9% autónomos.
- 39,9% microempresas (1 a 9 empleados).
- 5,2% pequeñas empresas de (de 10 a 49 empleados).
- 0,8% medianas empresas (50 a 249 empleados).
- 0,2 % grandes empresas (250 empleados o más).

Estos datos demuestran cómo las microempresas y los autónomos son parte esencial del tejido empresarial en España [25].

2.1.5 La tecnología para el emprendedor

Hoy en día la actividad empresarial y la tecnología están estrechamente vinculadas. Ya sea con fines industriales o sociales, no dependemos tanto de los métodos y medios tradicionales, en virtud de lo cual las empresas también han adoptado la tecnología en su beneficio. Esta está presente a lo largo de todo el aparato operativo de los negocios y cuando hablamos de expansión de una compañía, se presupone que la tecnología ha formado parte de ella directamente.

Sin embargo, en muchos casos, entre los que encontramos gran cantidad de pequeños comercios, encontramos un escaso aprovechamiento de todos los beneficios de la tecnología, de forma que el poco valor que muchos autónomos le prestan acaba pasando factura.

Las Tecnologías de la Información y Comunicación son un campo que los estudiantes universitarios utilizan como parte de la conformación de su carrera profesional, pero la vía de comunicación entre la universidad y el mundo de la empresa no es directa y a través de ella se ha creado una especie de laguna que no ayuda a conectar ambos extremos. Los estudiantes con enseñanzas superiores suponen grandes fuerzas tecnológicas en continuo aprendizaje que interactúan de diferentes formas con empresas locales.

Las empresas tienen que hacer un esfuerzo por valorar realmente la inmersión del mundo académico en su entorno, además de las universidades promocionar sus habilidades y su valor comercial para apoyar a las pequeñas y medianas empresas y conseguir ventajas competitivas.

¿En qué sentido podrán todos estos actores relacionarse en armonía para aprovechar nuevos caminos de desarrollo para el provecho mutuo y de la comunidad?, ¿es posible generar escenarios de participación que involucren a estudiantes y empresas?, ¿de qué manera se pueden fomentar estas actuaciones innovadoras desde el entorno empresarial? Para adquirir las respuestas a estas cuestiones hace falta esquematizar qué componentes tecnológicos serán el común denominador entre las empresas y la innovación [8].

En los últimos años, se han producido nuevos inventos y desarrollos en el campo de la tecnología. Herramientas modernas como la inteligencia artificial (IA), el internet de las cosas, el *big data* y el *cloud computing* son algunos ejemplos. Estos desarrollos han transformado las formas tradicionales de hacer las cosas. Hoy en día, todo se está llevando a las plataformas digitales. El *e-commerce*, el *e-learning*, el trabajo a distancia y la banca online, casi todo requiere ahora algún tipo de tecnología.

En términos sencillos, *big data* es una combinación de todos los procesos y herramientas relacionados con la utilización y gestión de grandes conjuntos de datos. El concepto de *big data* nació de la necesidad de comprender las tendencias, las preferencias y los patrones en la enorme base de datos que se genera cuando las personas interactúan con diferentes sistemas y entre sí. Con el *big data*, las organizaciones empresariales pueden utilizar la analítica y averiguar cuáles son los clientes más valiosos. También puede ayudar a las empresas a crear nuevas experiencias, servicios y productos.

Según los expertos, el *big data* puede crear un montón de nuevas oportunidades de crecimiento. Incluso puede dar lugar a una nueva categoría de empresas, como las que analizan y agregan datos del sector. La mayoría de estas empresas se encontrarán en medio de grandes flujos de información sobre servicios y productos, proveedores y compradores, intenciones y preferencias de los consumidores, etc. Las empresas de todos los sectores deberían empezar a desarrollar sus capacidades de Big Data de forma agresiva.

Además de la amplia escala de los *big data*, la alta frecuencia y la naturaleza en tiempo real de los datos son cruciales. Por ejemplo, la capacidad de estimar las métricas, incluida la lealtad del consumidor, se manejaba anteriormente de forma retrospectiva. Con el *big data*, este tipo de prácticas se están utilizando más ampliamente. Esto añade mucho al poder de predicción. Del mismo modo, la alta frecuencia permite a las empresas probar teorías en tiempo real.

2.1.6 Cómo fomentar la innovación y el emprendimiento utilizando datos abiertos

La Open Knowledge Foundation define los datos abiertos como "datos que pueden ser utilizados, compartidos y construidos libremente por cualquier persona, en cualquier lugar y con cualquier propósito". Otra definición es aquella que dice que los datos abiertos son datos en línea, gratuitos y accesibles que pueden ser utilizados, reutilizados y distribuidos siempre que se atribuya la fuente de datos [11].

En 2012, el gigante de los servicios profesionales Deloitte declaró que los datos eran "el nuevo capital de la economía mundial". Su visión era que "los datos abiertos, y no solo el *big data*, serán un motor vital para el crecimiento, el ingenio y la innovación en la economía del Reino Unido", y esa predicción se ha cumplido.

Los datos abiertos son un campo de estudio relativamente nuevo y emergente, pero la documentación relacionada con ellos es relativamente escasa. Con la apertura de los datos gubernamentales, una enorme cantidad de datos es accesible de forma gratuita. No hay duda de que las empresas pueden crear valor de estos datos, sin embargo, no hay directrices generales de cómo hacerlo y, sobre todo, lo más importante, cómo captar ese valor y convertirlo en beneficio. Los datos abiertos pueden considerarse como un puente entre el

sector público y el privado que promueve la innovación, y con la posibilidad de crear una causalidad acumulativa en la que ambos sectores se benefician mutuamente del desarrollo del otro.

Los mecanismos de valor de los Datos Abiertos operan en múltiples niveles. Se pueden esbozar múltiples beneficios de los Datos Abiertos, como la transparencia, el aumento de la eficacia y la eficiencia del gobierno, el aumento de la participación y el compromiso y la liberación de valor social y comercial a través de la innovación. En lo que respecta a esta última, los datos abiertos están cada vez más relacionados con la creación de nuevos productos y servicios, así como con la ayuda a las empresas para obtener nuevas ideas y conocimientos. Su aportación la economía es directa, ya que las empresas de esta forma pueden conseguir una mayor consciencia de su entorno, tomar decisiones y utilizarlos como materia prima para su funcionamiento, mejorando así los ingresos, creando empleo y ahorrando costes.

Aunque la innovación en materia de datos abiertos tiene lugar en organizaciones de todos los tamaños y niveles de madurez, los emprendedores y las empresas de nueva creación son actores cruciales para aprovechar todo el potencial de los datos abiertos, creando nuevos productos y modelos de negocio basados en ellos. Los empresarios son conocidos por su creatividad y su capacidad para convertir las ideas en negocios exitosos. Cada vez son más los emprendedores que tienen un modelo de negocio basado en los Datos Abiertos que dependen de la disponibilidad y la alta calidad de los datos y la información del sector público. Los Datos Abiertos tienen un efecto facilitador para los emprendedores, ya que crean oportunidades para nuevos modelos de negocio, productos y servicios, al proporcionar una gran cantidad de material gratuito para desarrollar aplicaciones de software para uso en ordenadores y móviles.

Estas empresas emergentes refuerzan la línea general de argumentación relacionada con el alto impacto económico que tienen los Datos Abiertos. Muestran -a través de sus actividades empresariales- los beneficios inmediatos que los Datos Abiertos pueden tener para la creación de empleo, nuevos modelos de negocio y la economía en su conjunto. Los empresarios que trabajan como analistas de datos, intermediarios de datos, desarrolladores de software y otros profesionales relacionados con los datos, conforman un importante grupo de reutilizadores. Al presentar productos y servicios tangibles de Datos Abiertos, las organizaciones públicas se convencen de invertir más en programas de Datos Abiertos. Además, las pequeñas empresas innovadoras que trabajan con datos abiertos pueden allanar el camino para que las grandes corporaciones y organizaciones públicas adopten nuevas soluciones de datos abiertos.

Según los beneficios económicos de los datos abiertos, el *European Data Portal* publicó en 2020 su estudio "The Economic Impact of Open Data: Opportunities for value creation in Europe", el cual investiga el valor creado por los datos abiertos en Europa en términos de empleo, ahorros y eficiencia. En él se estima su tamaño de mercado en 184.000 millones de euros y se prevé que alcance entre 199.510 y 334.210 millones de euros en 2025 [13]. El mismo informe destaca que se crearán más de 25.000 puestos de trabajo en el sector privado, en el mismo periodo.

Aunque el panorama es positivo, es cierto que las empresas que quieren emprender utilizando datos abiertos también tienen que enfrentarse a una serie de retos organizativos, técnicos y

financieros. La posible lentitud en una elaboración de un caso de negocio es un aspecto a tener en cuenta, especialmente si no se cuenta con las habilidades técnicas necesarios o todos los recursos económicos. Aquí, las administraciones públicas pueden desempeñar un papel importante [10].

La importancia del efecto facilitador de los datos abiertos en las nuevas empresas y viceversa se pone de manifiesto en el último informe analítico sobre datos abiertos y emprendimiento [14]. El informe analiza el papel de los emprendedores en la innovación de los Datos Abiertos, la relación entre los Datos Abiertos y la sostenibilidad, así como las principales barreras de los Datos Abiertos para el emprendimiento. El informe también examina las principales políticas aplicadas a nivel de la UE para fomentar el emprendimiento impulsado por los Datos Abiertos y ofrece una visión general de las mejores prácticas de Datos Abiertos. Ejemplos como FarmDog (una empresa impulsada por los datos que hace que los alimentos saludables y ambientalmente sostenibles sean accesibles para todos), OpenCorporates (la mayor base de datos abierta de empresas del mundo), PlumeLabs (un ejemplo de una empresa que fue capaz de lograr un producto mínimo viable y un prototipo con Datos Abiertos) o Synergise (compañía que actúa como intermediaria de bases de datos abiertos complejas y grupos interesados en ellas, permitiendo que estas últimas no se tengan que encargar de la captación, limpieza y gestión de los datos).

2.1.7 Machine Learning

Las máquinas pueden llegar a ser inteligentes por sí mismas, la inteligencia se imprime en ellas y aquí es donde entra el aprendizaje automático. El aprendizaje automático se centra en el desarrollo de programas informáticos que pueden acceder a los datos y utilizarlos para aprender de forma autónoma [15]. El objetivo principal es hacer que los ordenadores sean capaces de aprender automáticamente sin ninguna intervención humana.

En palabras de la compañía IBM Cloud Education, el aprendizaje automático es una rama de la inteligencia artificial y de la informática que se centra en el uso de datos y algoritmos para imitar el modo en que aprenden los humanos, mejorando gradualmente su precisión [16].

El aprendizaje automático es un componente importante del creciente campo de la ciencia de los datos. Mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones, descubriendo ideas clave dentro de los proyectos de minería de datos. Estos conocimientos impulsan posteriormente la toma de decisiones dentro de las aplicaciones y las empresas, lo que idealmente repercute en las métricas de crecimiento clave. A medida que los macrodatos sigan expandiéndose y creciendo, aumentará la demanda de del uso de la ciencia de datos en el mercado [16].

La mayoría de los sectores que trabajan con grandes cantidades de datos han reconocido el valor de la tecnología de aprendizaje automático. Al obtener información de estos datos, a menudo en tiempo real, las organizaciones son capaces de trabajar de forma más eficiente o de obtener una ventaja sobre sus competidores.

Dos de los métodos de aprendizaje automático más adoptados son el aprendizaje supervisado y el aprendizaje no supervisado, pero también hay otros métodos como el aprendizaje semisupervisado y el aprendizaje por refuerzo [17].

Los algoritmos de aprendizaje supervisado se entrenan utilizando ejemplos etiquetados, que son aquellos en los que se conoce la entrada con su resultado deseado. Por ejemplo, las piezas de un equipo estarían etiquetadas si se clasifican como "F" (falla) o "R" (funciona). El algoritmo de aprendizaje recibe un conjunto de entradas junto con las correspondientes salidas correctas, y el algoritmo aprende comparando su salida real con las salidas correctas para encontrar errores. A continuación, modifica el modelo en base a ello. Mediante métodos como la clasificación, la regresión, la predicción y el *gradient boosting* (potenciación del gradiente en español), el aprendizaje supervisado utiliza patrones para predecir los valores de la etiqueta en otros datos no etiquetados. El aprendizaje supervisado se utiliza habitualmente en aplicaciones en las que los datos históricos predicen probables acontecimientos futuros. Por ejemplo, puede anticipar cuándo es probable que las transacciones con tarjetas de crédito sean fraudulentas o qué cliente de seguros es probable que presente una reclamación [17].

El aprendizaje no supervisado se utiliza con datos que no tienen etiquetas registradas. No se le dice al sistema la "respuesta correcta". El algoritmo debe averiguar lo que se le muestra. El objetivo es explorar los datos y encontrar alguna estructura en ellos. El aprendizaje no supervisado funciona bien con datos transaccionales. Por ejemplo, puede identificar segmentos de clientes con atributos comunes que pueden ser tratados de forma similar en las campañas de marketing. O puede encontrar los principales atributos que separan los segmentos de clientes entre sí. Entre las técnicas más populares se encuentran los mapas autoorganizados *(self-organizing maps*), el mapeo de vecinos más cercanos (*nearest-neighbor mapping*), la agrupación de k-means (*clustering*) y la descomposición en valores singulares (*singular value decomposition*). Estos algoritmos también se utilizan para segmentar temas de texto, recomendar artículos e identificar datos atípicos [17].

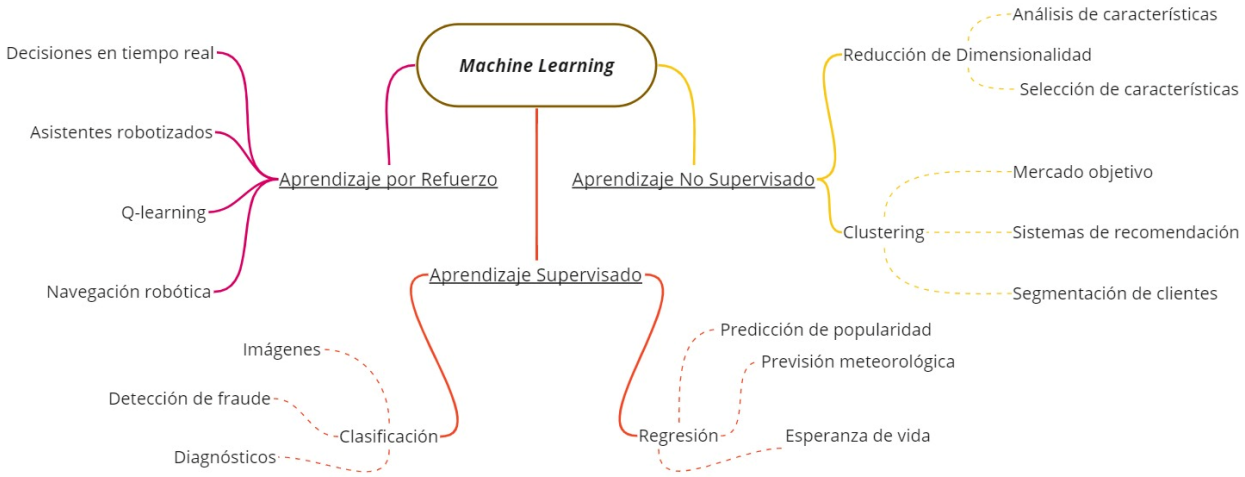
El aprendizaje semisupervisado se utiliza para las mismas aplicaciones que el aprendizaje supervisado, pero utiliza tanto datos etiquetados como no etiquetados para el entrenamiento: normalmente una pequeña cantidad de datos etiquetados con una gran cantidad de datos no etiquetados, debido a que estos últimos son menos costosos y cuesta menos esfuerzo adquirirlos. Este tipo de aprendizaje puede utilizarse con métodos como la clasificación, la regresión y la predicción. Es útil cuando el coste asociado al etiquetado es demasiado alto para permitir un proceso de entrenamiento totalmente etiquetado. Entre los primeros ejemplos de ello se encuentra la identificación de la cara de una persona en una cámara web [17].

Por último, el aprendizaje por refuerzo se utiliza a menudo en robótica, juegos y navegación. Con el aprendizaje por refuerzo, el algoritmo descubre, mediante el método de ensayo y error, qué acciones producen las mayores recompensas. Este tipo de aprendizaje tiene tres componentes principales: el agente (el que aprende o toma las decisiones), el entorno (todo aquello con lo que el agente interactúa) y las acciones (lo que el agente puede hacer). El objetivo es que el agente elija las acciones que maximicen la recompensa esperada en un tiempo determinado. El agente alcanzará el objetivo mucho más rápido si sigue una buena

estrategia. Por tanto, el objetivo del aprendizaje por refuerzo es aprender la mejor técnica [17].

Ilustración 2. Esquema del Aprendizaje Automático.

Fuente: elaboración propia



2.1.8 Modelos de Aprendizaje Supervisado

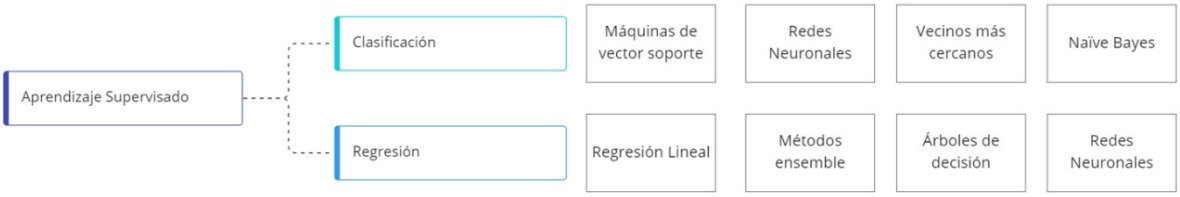
El aprendizaje supervisado se clasifica en dos grandes algoritmos de resolución de problemas, la clasificación y la regresión. A lo largo del presente proyecto utilizaremos ambos para conformar un modelo de datos abiertos completo y usable.

Clasificación: La clasificación utiliza un algoritmo para predecir salidas categóricas o de clase. Su objetivo es clasificar las entradas en un número determinado de clases o categorías en función de la etiqueta de datos con la que se ha entrenado. Determina si una determinada transacción es un fraude o no, si un correo electrónico es spam o no, etc. Algunos de los algoritmos de clasificación más conocidos son los árboles de decisión, los clasificadores lineales, los bosques aleatorios, las máquinas de vectores de apoyo y los vecinos más cercanos.

Regresión: La regresión es un proceso estadístico predictivo en el que los modelos de aprendizaje automático supervisado se esfuerzan por establecer una conexión entre las variables dependientes e independientes. El algoritmo de regresión tiene como objetivo encontrar un número continuo como los ingresos, las ventas o las puntuaciones de los exámenes. Los algoritmos de regresión más utilizados son la regresión logística, la regresión lineal, la regresión lineal bayesiana y la regresión polinómica.

Ilustración 3. Algoritmos más utilizados.

Fuente: elaboración propia



2.2 Contexto y justificación

2.2.1 Destrucción de empresas en Madrid

En el primer cuatrimestre del año 2019 en la Comunidad de Madrid batió récord de destrucción de empresas, ya que un total de 2825 compañías echaron el cierre [18].

Nunca antes en los primeros cuatro meses de un año se habían disuelto en la Comunidad tantas empresas, de hecho han dejado de operar el doble de las que se destruyeron en ese primer cuatrimestre de 2008, en el que fueron un total de 1.467 compañías las que cerraron su negocio. Sin embargo, a nivel nacional, el número de empresas destruidas no se encuentra en su máximo, lo que quiere decir que es una tendencia que se observa en Madrid en mayor medida, ya que el resto del país no está sufriendo un fenómeno parecido (al menos no hasta la época pre-covid19). En este mismo periodo también se registra que cada año se cierran en torno a 600 pequeños comercios en Madrid.

En el siguiente gráfico se puede observar claramente la preocupante tendencia:

Ilustración 4. Destrucción de empresas hasta abril de 2019 en la Comunidad

Fuente: Alejandra Olcese para Vozpópuli

Lorenzo Amor, presidente de la Asociación de Trabajadores Autónomos (ATA) explica al periódico Vozpópuli que los datos de la Comunidad de Madrid confirman que el crecimiento económico se está desacelerando y que "existe menor confianza, lo que impide que los autónomos monten empresas".

Esta es una realidad que muchos madrileños observamos día a día. En mi caso, siempre recurro al mismo ejemplo, y es que en la esquina de la calle en la que vivo hay un local que casi cada dos años pertenece a un dueño diferente. En primer lugar, fue una perfumería, posteriormente pasó a ser una cadena de patatas fritas, no sin antes albergar un local de venta de carcasas para móviles. Hace un año fue una churrería, actualmente es un bar y no me va a sorprender cuando este bar se convierta en una tintorería, o incluso en una tienda de segunda mano, por ejemplo. Es por esto mismo que he estado pensando en lo mucho que ahorrarían

los emprendedores si tuviesen un modelo de saber qué comercios faltan en según qué distritos, cuáles sobran, cuáles no son propensos a prosperar y cuáles pueden sorprender al vecindario. Esto se debe a que hay muchos factores que influyen en los distintos tipos de comercios que se necesitan en cada zona. Uno de esos factores es que el promedio de las rentas en Madrid varía desde los 17.786 euros del barrio más pobre hasta los 114.186 del más rico.

Dado que la preocupante tendencia se observa en el año 2019, se puede pensar que en el año 2022 no sucede de la misma manera. A fecha 22 de abril de 2022, el diario Cinco Días publicó un artículo en el que se afirma que, según el informe sobre la Declaración de la Renta y Política Fiscal Española realizado por TaxScouts, un 66% de los españoles no considera que en nuestro país se fomente el emprendimiento [19]. Además, declara que las cifras continúan siendo preocupantes, desde el Global Entrepreneurship Monitor (GEM) se avisa de que el emprendimiento ha ido en continúa caída desde el año 2007, momento el que se situaba en un 7,6%. En cambio, según los datos de 2020, únicamente un 5,2% de los españoles emprendió.

2.2.2 Temor por emprender

Según el último Informe GEM 20-21, España encabeza el ranking mundial en cuanto al miedo al fracaso como obstáculo para emprender, con un 64% de la población frente al 47% de media mundial [20].

A pesar de que la preocupante tendencia es más frecuente entre la franja de edad de los 18 a los 24, tanto entre jóvenes como seniors se percibe que este miedo les impide iniciar un nuevo negocio aun contando con el capital humano y social necesarios para emprender.

Asimismo, es preciso comprender las condiciones del entorno de cada emprendedor español y, más importante si cabe, cómo las perciben para lanzarse. En 2020, según los 36 expertos españoles entrevistados para el informe GEM, la infraestructura comercial (6,5), el acceso a infraestructura física (5,9), los programas gubernamentales (5,7) y a la educación en etapa post escolar (5,1) son las condiciones del entorno que han recibido las mejores valoraciones. En contraste, la educación y formación emprendedora en etapa escolar (2,2) y la burocracia e impuestos (3,9) han recibido las más bajas [20].

En virtud de todos los datos expuestos, podemos firmar con seguridad que el español medio teme a convertirse en emprendedor por miedo a todos aquellos obstáculos que puedan llevarle al fracaso. Perciben que existe una falta de empuje por parte de las instituciones para tomar esta decisión y que no existen herramientas o plataformas que ayuden a disipar sus dudas. Es cierto que tomar iniciativa implica asumir riesgos, salir de la comodidad y lanzarse al vacío con un plan de acción, sin embargo, el aspirante a emprendedor no solo no encuentra asistencia para conformar dicho plan de acción, sino que los datos que se recogen sobre los últimos años acrecentan las barreras.

2.2.3 Repercusión en el pequeño comercio

Respecto a dichos datos, existen numerosos artículos e informes que registran cómo el comercio minorista ha estado arrastrando pérdidas en los últimos años. El pequeño negocio sigue en crisis. De acuerdo a los datos publicados por la Confederación Española de Comercio, institución que ya advierte de la situación complicada en la que se sitúa el comercio minorista, sus ventas han bajado un 0,5% en el periodo de diciembre de 2021 a enero de 2022 [22]. A pesar de que pueda no parecer alarmante, cabe destacar que durante esta fase se supone que las cifras no deberían ser negativas gracias a la recuperación tras la pandemia COVID-19. ¿A qué se debe la inestabilidad y dificultades que sufre el pequeño comercio?

Su crisis empezó años atrás, se calcula que por el año 2015 comenzó el declive. A partir de ese momento, las malas noticias empezaron a brillar sobre los buenos resultados y las pérdidas se hacían cada vez más comunes. Sin embargo, la alarma no sonó hasta el año 2017, fecha en la cual la Asociación de Trabajadores Autónomos (ATA) alertó de que este tipo de negocios se encontraban acumulando una pérdida de 36 autónomos diarios de media en distintas comunidades autónomas [21].

Pese a que en ese mismo año las ventas aumentaron un 1,2% en esos negocios con respecto al año 2016, estas descendieron en un 1,6% cuando llegó la campaña de Navidad, época en la que acostumbran a darse buenos resultados. En aquel momento las pérdidas que ya sufrían debido al surgimiento y el éxito de las grandes superficies empezaron a acumularse.

Asimismo, teniendo en cuenta el auge de estas grandes superficies y afectado por el crecimiento de los E-*commerce* y los *Marketplace*, nos damos cuenta de que el comercio minorista necesita renovarse de alguna manera. No se trata únicamente de los problemas que tienen los emprendedores para iniciar un negocio y el riesgo al que se someten, sino que es el propio pequeño negocio quien tampoco está recibiendo buenos resultados. Es aquí donde debe tomar acción la tecnología y puede contribuir a su innovación. Esta será la razón de ser del presente trabajo. A partir de este punto es cuando decidimos tomar acción y proponer el modelo que posteriormente se expondrá.

En añadido, hay unos 14.000 locales comerciales vacíos en la ciudad de Madrid [23], según una estimación de la consultora barcelonesa Eixos Economic Observatory que se basa en su propio trabajo de campo. De acuerdo con los datos de Eixos, en 39 barrios de Madrid el índice de ocupación comercial es menor del 80%, el nivel que los expertos inmobiliarios consideran como indicador de mala salud comercial en una zona.

2.2.4 Propuestas ya existentes

En un intento de solucionar las problemáticas expuestas, existe una iniciativa lanzada en el año 2017 por la inmobiliaria barcelonense Forcadell que permite ahora a los vecinos votar qué negocio prefieren en su zona. Se trata de *Shopinion*, a través de la cual pretenden conocer la opinión de los ciudadanos respecto a los vacíos comerciales de cada barrio. De momento, está solo en Barcelona y en el sector comercios, aunque la intención es que salga de la capital catalana y que se pueda extender a otros ámbitos.

En el caso de esta plataforma, hemos podido observar que no se encuentran operativos desde 2018. En este caso no realizan un análisis y estudio de datos y estadísticas para conocer cuál es la actividad comercial que se puede necesitar en un barrio, sino que lo llevan a cabo a través de encuestas. Estas encuestas nunca recogen la opinión de una cantidad considerable de vecinos y por norma general tienden a estar sesgadas.

Ilustración 5. Página de inicio de Shopinion

Además, el banco BBVA lanzó en 2017 *Dinámicas del Turismo en la Ciudad de Madrid* [29], un estudio basado en la actividad comercial real del año 2012 que analiza el comportamiento de los turistas a partir de su actividad comercial, usando tecnologías de datos masivos o *big data*. El estudio es el primer resultado práctico de la colaboración que mantienen BBVA y el Ayuntamiento de Madrid en materia de promoción de la innovación y de las ciudades inteligentes. En añadido, en colaboración con CARTO, presentaron una *data story* que muestra cómo se puede aplicar la ciencia de datos, sobre dato anonimizado y agregado, para identificar áreas funcionales, describir dinámicas y establecer comparaciones entre tres grandes ciudades como Madrid, Barcelona y Ciudad de México. De esta manera, transformaron Madrid en datos dejando entrever su huella comercial y los niveles de gasto por zonas. En el caso de esta iniciativa, se llevó a cabo con el fin de llevar a cabo una transformación digital para conseguir datos tanto comerciales, como sociales y demográficos. Al fin y al cabo llevaron a cabo una herramienta que sirviese de sistema de recomendación para usuarios tanto en busca de casa, como de ocio, por ejemplo. Además de que al tratarse de un proyecto llevado a cabo por una gran entidad, su consumo de recursos es considerablemente alto, por lo que el margen de beneficio que necesitan obtener también lo es, lo que es sinónimo de precios elevados. Asimismo, este se trata de un informa puntual del año 2017 basado en información relativa al año 2012, por lo que actualmente no existe ninguna plataforma que haga uso de datos que se mantengan actualizados en base a lo que el ayuntamiento de Madrid permita.

2.3 Planteamiento del problema

La idea es crear una plataforma que cuente con un análisis y representación de los temas expuestos a lo largo de los apartados anteriores: qué tipo de negocios se mantienen mayor tiempo abiertos en una zona, qué negocios no tienen éxito, además de un estudio sobre las carencias comerciales de esas zonas, es decir, si no hay suficientes tintorerías en Usera, tal vez

un inversor o emprendedor determinado esté interesado en hacerse con un local libre para abordar el establecimiento de una. Hay muchos barrios cuyos índices de desempleo, e incluso de criminalidad, se deben a la falta de ciertos comercios en sus calles, por lo que mi proyecto ayudaría a la reactivación de ciertas zonas.

El objetivo del proyecto es predecir la probabilidad de éxito de un determinado negocio en una localización concreta abarcando únicamente la ciudad de Madrid. El proceso a seguir para llegar a ello será crear una especie de plataforma que cuente con una clasificación sobre qué tipo de negocios se conservan mayor tiempo abiertos en una zona, qué negocios no tienen éxito, además de un estudio sobre las carencias comerciales de esas zonas. En caso de no ser posible el análisis de algunas zonas de Madrid, estará enfocado a un único distrito con el objetivo de obtener conclusiones robustas poco a poco.

La razón principal del proyecto no es crear una plataforma, este será el proceso, pero el punto final consiste en conseguir un modelo de datos basado en datos abiertos, por ello el nombre del trabajo. Existirá un punto de contacto con el usuario final que servirá a modo de plataforma y los resultados del análisis se muestran sobre un cuadro de mando, por lo que el corazón de todo el trabajo es el modelado de datos que se llevará a cabo.

En añadido a lo anterior, en caso de que las circunstancias y el número de horas marcadas en la planificación lo permita, a modo de objetivo adicional surge la idea de “optimización de instalaciones”, además de haber visto fuentes de datos suficientes para todo lo anterior, también cuento con el censo de locales y actividades comerciales de Madrid que ofrece información sobre la situación de ciertos locales, si están abiertos, cerrados, en reforma o inutilizados. En el caso de los inutilizados, mi idea, junto con lo anterior, es promover una optimización de espacios ya construidos. Hay miles de infraestructuras abandonadas a lo largo de todo Madrid que suponen más gasto que beneficio.

Es debido a todo lo mencionado por lo que este proyecto está ideado para crear un modelo que indique qué actividad comercial es más adecuada para las distintas zonas de la ciudad de Madrid.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo general del proyecto es la creación de un modelo de datos abiertos que muestre la probabilidad de éxito de un determinado negocio en una localización concreta de la ciudad de Madrid. Este servirá como sistema de recomendación para inversores y se elaborará a través del entrenamiento de un algoritmo conformado a partir de datos abiertos del portal del ayuntamiento de la ciudad. Conjuntamente, el análisis de dichos datos se mostrará sobre un cuadro de mando interactivo a modo de contacto con el usuario.

3.2 Objetivos específicos

En primer lugar, para llegar a alcanzar el objetivo general del proyecto y obtener los resultados, es necesaria una primera fase de obtención y procesamiento de los datos. Esta etapa se denomina ETL, siglas en inglés de Extraer, Transformar y Cargar. Se trata de un proceso genérico en el que primero se adquieren los datos, luego se modifican o procesan y finalmente se cargan en un *datawarehouse*[1], bases de datos u otros archivos como Excel o CSV. Dentro de este proceso encontramos los siguientes objetivos:

1. Extraer los datos del portal de datos abiertos.
2. Limpiar y homogeneizar los campos de los conjuntos de datos.
3. Obtener dos datasets. Un primero que indique el estado actual de la actividad comercial de la ciudad y un segundo que recoja los registros a introducir en el modelo.

Tras esto, contaremos con los datos necesarios para realizar todas las tareas. En cuanto a análisis de los datos y visualización, se marcan los siguientes objetivos:

4. Elaborar un análisis de la actividad comercial a través de Power BI.
5. Crear un cuadro de mando que muestre la situación comercial actual de Madrid.
6. Configurar una serie de conclusiones iniciales.

Por último, llega la fase de creación del modelo de aprendizaje supervisado que tomará como entrada los datos que han sido procesados previamente.

7. Adaptar el conjunto de datos al modelo.
8. Entrenar un algoritmo y testearlo para configurar el modelo.
9. Conseguir un modelo capaz de realizar recomendaciones al usuario pudiendo probar sus resultados con diferentes combinaciones de negocio y localización.
10. Diseñar una ventana de muestra del modelo.

[1] Repositorio central de datos una vez cargados y procesados.

11. Mostrar el modelo al usuario a través de una interfaz a fin de establecer una vía de contacto directa con él.
12. Contrastar los resultados obtenidos con las conclusiones recibidas a través de la visualización y análisis de los datos.

3.3 Beneficios del proyecto

Hoy por hoy en España, más concretamente en la ciudad de Madrid, es preocupante el nivel de destrucción de empresas que estamos sufriendo. Según los datos aportados anteriormente, el pequeño comerciante no confía en sacar adelante su negocio ni emprender ideas nuevas. Gracias al modelo elaborado en este proyecto, se ofrecerá una ayuda directa al futuro emprendedor, lo cual permitirá que se fomente una mayor inversión en el sector empleador con mayor número de cotizantes en el territorio nacional. La necesidad de lanzar al mercado una plataforma dedicada al apoyo al emprendimiento aumenta tras el periodo vivido por todos los españoles debido a la pandemia del COVID-19. Los modelos de negocio están en un constante cambio, pero más aún a causa de dicha época, en la cual se ha intensificado el uso de la tecnología en los negocios como un requisito fundamental para salir adelante. Si gracias a su uso hemos conseguido mantener las conexiones, los negocios, el trabajo diario y la vida estudiantil, entre otros, también debería apoyar en el incentivo del pequeño comercio, gran olvidado como aliado de las nuevas tecnologías.

En línea con la transformación digital, es necesaria una automatización del procesamiento de los datos abiertos. Gracias al código elaborado en la fase de ETL del proyecto, se puede introducir cualquier fichero del censo de locales y actividades de Madrid y este se devolverá sin errores de escritura, en los formatos correctos y con el texto debidamente normalizado.

Adicionalmente, el modelo de datos abiertos se dirige directamente al usuario final que requiera de un servicio de recomendación en su futura inversión. De esta forma, no surgen intermediarios ni obstáculos que impidan su uso directo por el emprendedor y resulta fácilmente accesible para cualquier usuario interesado. Los costes de elaboración disminuyen considerablemente frente a los que tendría una gran empresa sobre el mismo proyecto, lo que conforma a la plataforma como una idea de negocio notablemente asequible. Por último, el nivel de adaptación del modelo a nuevas corrientes de negocio es alto, ya que es posible su constante actualización a través de los datos que proporciona el ayuntamiento de Madrid mensualmente desde hace ocho años.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

Fase 1. Establecimiento de objetivos

- Planteamiento detallado del contexto: definir y redactar de forma concreta la problemática a resolver, los motivos e hipotéticos resultados a obtener gracias al presente trabajo.
- Planteamiento de objetivos. Determinar todas las metas a alcanzar en el curso del proyecto y elaborar un plan de actuación. Definir cuáles son las tareas que se van a llevar a cabo para conseguir los resultados deseados y qué tipo y forma de datos son esenciales para elaborar el modelo.
- Alcance de las tareas u objetivos. Plantear a dónde se pretende llegar, qué abarcará el proyecto y la cobertura de cada una de las tareas. Detallar cuidadosamente todos aquellos frentes que no tenemos que cubrir con nuestros resultados y diferenciar lo que sí haremos frente a lo que no, o lo que puede que no se cumpla.

Fase 2. Planificación

- Elaboración del plan de trabajo. Organizar y esquematizar los objetivos detallados en la fase anterior. De esta manera realizaremos una agenda con los diferentes entregables a tener hechos en cada momento, teniendo en cuenta la dependencia entre tareas, ya que es importante distinguir cuáles se necesitan en un primer lugar para avanzar.
- Esquema y exposición del estado del arte y contexto. Redactar cuáles son los pasos que nos han llevado a realizar este proyecto, de forma que consigamos en este punto un hilo conductor de la justificación del proyecto con su modo de actuación y los posibles beneficios, así obtendremos una visión de lo más importante a abarcar.

Fase 3. Diseño

- Definición de herramientas a utilizar. Es una de las tareas más importantes. En ella se incluyen, en primer lugar, el lenguaje de programación con el que se llevarán a cabo las diferentes tareas, empezando desde la obtención de datos, su preprocesamiento, análisis y terminando con el modelo de aprendizaje supervisado. En segundo lugar, definición de cómo se mostrarán al usuario las visualizaciones del análisis y con qué plataforma. Por último, concretar cuál será el punto de contacto del modelo con el usuario y si esta será por medio de una plataforma o de un cuadro de mando.
- Arquitectura. Una vez tomadas las decisiones anteriores, se elaborará un esquema de la arquitectura de etapa con las respectivas plantillas de los *dashboards*.
- Adecuar las tareas a las herramientas. Esto quiere decir que una vez se supo que se utilizaría Python como lenguaje, se decidió que se podrían hacer cuadros de mandos con el mismo, a través de *Plotly* y *Dash*, y la tarea de obtención de datos además tendría *ParseHub* como apoyo. Además, dado que la plataforma de visualización sería

Power BI, los datos deberán cargarse en formato Excel también, por rapidez de importación en la herramienta.

Fase 4. Implementación

- Obtención de datos. Empezar con *ParseHub* para obtener links de descarga y posteriormente hacer web scraping de su contenido con Python. Cargar todos los registros en archivos de cara a su procesamiento. No solo incluye los datos utilizados para el modelo, sino también los actuales de Madrid.
- Transformación. Limpieza y preprocesamiento de todos los registros. Se trata de elaborar un sistema con funciones de normalización de texto, numerizar los campos adecuados, obtener fechas que nos permitan calcular las duraciones de los negocios y eliminar todos los registros sin información útil.
- Carga de los datos. Obtener varios ficheros en diferentes formatos de los conjuntos de datos a utilizar en el modelo, el que servirá para la visualización y el actual de la ciudad de Madrid.
- Análisis descriptivo de los datos. Esto nos permitirá tener diferentes consideraciones de cara al modelo. No solo se realiza a través de Python, sino que el propio cuadro de mando obtenido con Power BI nos permitirá tomar decisiones.
- Visualización de datos. A través de Power BI y la creación de diversos *dashboards* obtendremos las primeras conclusiones.
- Creación y entrenamiento del modelo en Python. Elaboración y entrenamiento de un modelo de regresión y otro de clasificación. Probaremos el que mejor funcione a través de la fase de test y será el que se utilice como resultado.
- Conexión del modelo con el usuario. Elaborar una interfaz sencilla con *Plotly* y *Dash* en Python que nos permita ser sirva como punto de recomendación en el mundo del emprendimiento.

Fase 5. Resultados

- Redacción de conclusiones. Explicación de todos los puntos calientes que han surgido a lo largo del proyecto y las decisiones tomadas para resolverlos que han llevado al resultado obtenido. A través de ellos, redactar las conclusiones del análisis y modelo realizado además de trabajos a futuro.

Fase 6. Subsanación de errores

- Pruebas del modelo. Simulación de un usuario real.
- Revisión de la memoria y bibliografía.
- Presentación del TFG.

Tabla 1. Planificación del proyecto

	ENE	FEB	MAR	ABR	MAY	JUN	Horas
Planteamiento del contexto							
Planteamiento de objetivos							*10*
Definir alcance de tareas							
Elaboración del plan de trabajo							*25*
Esquema: estado del arte							
Decisión de la arquitectura							*5*
Definición de herramientas							
Obtención de los datos							
Transformación							
Carga de los datos							
Análisis descriptivo							*150*
Visualización de datos							
Creación y entrenamiento modelos							
Creación interfaz de usuario							
Pruebas y revisión del modelo							*20*
Redacción memoria							
Presentación proyecto							*160*
Total Horas							*370*

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1 Herramientas empleadas

4.2.1.1 Tecnologías

- Lenguaje de programación: Python

En todas las fases del proyecto se hace uso de Python como lenguaje. Es la herramienta principal en la extracción, selección, procesamiento, análisis y modelado de los datos. La visualización del modelo también se realiza con Python

- Web Scraping: ParseHub

Se trata de una herramienta visual de extracción de datos para obtener datos de páginas webs que en este caso se ha utilizado para la obtención de los hipervínculos de descarga a los ficheros del portal de datos abiertos.

- Creación de mind maps: Miro

Se trata de una pizarra digital utilizada para la creación de los diagramas y mapas mentales elaborados en el proyecto y presentados a lo largo de la memoria.

- Visualización de Datos: Power BI y Plotly

La primera se utiliza como herramienta principal para el diseño de los cuadros de mando y la segunda es la base de la plataforma de exposición del modelo al cliente.

- Control de versiones: Git, OneDrive
- Interfaz gráfica de usuario (GUI): Anaconda

Se trata de una distribución de Python, que incluye el propio lenguaje y muchas bibliotecas de software comunes y un administrador de paquetes. Permite lanzar *Jupyter Notebooks*, que es la herramienta específica utilizada para toda la programación del análisis de datos completo.

- IDE: PyCharm

PyCharm es un IDE, lo que significa que los desarrolladores que escriben código Python interactúan directamente con él. Se ha utilizado para manejar archivos en código Python para la creación de la visualización del modelo en *Dash*. Además, PyCharm ofrece soporte para múltiples plataformas, así como los lenguajes complementarios HTML y JavaScript, también utilizados para manejar *Plotly* y *Dash*.

Ilustración 6. Logos de las herramientas empleadas

4.2.1.2 Librerías

Tabla 2. Listado de librerías utilizadas

NOMBRE	DESCRIPCIÓN
OS	Librería destinada para interactuar con el sistema operativo base.
PANDAS	Librería diseñada para la manipulación de datos y análisis de los mismos.
SELENIUM	Librería utilizada para automatizar la interacción con un navegador web.
REGEX	Librería que permite utilizar expresiones regulares en Python. Se utiliza para comprobar si una cadena de caracteres contiene un patrón de búsqueda específico.
NUMPY	Librería que contiene estructuras de datos y arrays multidimensionales. Proporciona objetos y métodos para operar sobre ellos.
SCIKIT-KLEARN	Librería que contiene herramientas para el aprendizaje automático y modelado estadístico.
STATSMODELS	Librería que proporciona clases y funciones para la estimación de diferentes modelos estadísticos, así como para la realización de pruebas estadísticas, y la exploración estadística de los datos.
MATPLOTLIB	Librería para crear visualizaciones estadísticas, animadas e interactivas.
SEABORN	Librería de visualización basada en matplotlib que proporciona una interfaz de alto nivel para dibujar gráficos estadísticos.
PLOTLY	Librería interactiva de código abierto que permite más de 40 tipos de gráficos únicos que cubren una amplia gama de casos de uso estadísticos, financieros, geográficos, científicos y tridimensionales. Está construida sobre la biblioteca Plotly JavaScript.

DASH	Librería en Python de código abierto construida sobre React. js, Flask y Plotly. js. Permite unificar las aplicaciones que normalmente requerirían un equipo de front-end, back-end y DevOps para conformarse.
GEOPY	Librería de python para acceder a servicios de geocodificación
UTM	Librería utilizada para convertir coordenadas UTM-WGS84 bidireccionalmente
PICKLE	Librería que nos permite almacenar casi cualquier objeto Python directamente en un archivo o cadena sin necesidad de realizar ninguna conversión.

Una vez expuestos los recursos necesarios, a continuación se detalla ordenadamente el procedimiento seguido para elaborar el modelo de datos al completo.

4.2.2 Obtención de los datos

El proyecto realizado se basa en el uso de la información que publica el portal de datos abiertos del Ayuntamiento de Madrid sobre el censo de locales, sus actividades y terrazas de hostelería y restauración. A fin de obtener la información sobre los negocios que se mantienen abiertos y los que no consiguen salir adelante, recurrimos al histórico, el cual ofrece la información existente desde el 31 de marzo de 2014 hasta el 31 de diciembre de 2014 en actualizaciones trimestrales. A partir de esa fecha, la información se ofrece en actualizaciones mensuales [27].

Cada fichero de microdatos del censo se sube al portal independientemente del anterior, de forma que cada mes se sube un registro actual del censo de locales. Como consecuencia, debemos descargarnos 89 conjuntos de datos con información repetida entre ellos. El trabajo posterior consistirá en eliminar los duplicados que no aporten información relevante al estudio.

Tal y como he mencionado, dentro del apartado del censo de locales encontramos en cada actualización mensual la posibilidad de descargarnos el CSV referente a los locales, el que contiene estos mismos con información de licencia, el que informa sobre sus actividades y el relativo a las terrazas. Para la problemática que estamos intentando resolver, escogemos descargarnos los que se encuentran en el apartado de actividades, debido a que es el que contiene registros sobre la sección comercial y no tiene información irrelevante.

Para la descarga de los conjuntos elegidos, en primer lugar acudimos a ParseHub, herramienta de scraping web gratuita y fácil de usar. Una vez descargada como aplicación, se pueden extraer datos de los sitios web para su análisis. Gracias a ella conseguimos obtener de manera rápida todas las direcciones de enlace (URL) de descarga junto con el título. El título de cada

conjunto de datos será necesario debido a que dentro del propio archivo no se encuentra la fecha de actualización de los datos. Esto quiere decir que no tenemos información en fechas sobre cuando se abrió o cerró cada negocio, únicamente la fecha en la que el portal de datos publicó un nuevo censo de locales. De este modo, en caso de que en un determinado mes surja un nuevo comercio, contaremos con su fecha de apertura, que se corresponderá con el mes en el que se subió la actualización. En la ilustración 7 se puede observar el formato comentado.

Ilustración 7. Captura de la página del censo de locales del Ayuntamiento de Madrid

Una vez obtenemos el resultado del scraping a través de ParseHub, contamos con un CSV que contiene cada URL con su fecha. Con el propósito de bajarlos en el ordenador, pasamos a su descarga a través de Python.

Cargamos el set de datos en un notebook de Python y limpiamos la distribución en la que ParseHub lo ha obtenido para poder hacer uso de él. Gracias a las librerías pandas y re, cuadramos las filas y obtenemos el mes correspondiente a cada enlace, ya que había que recoger expresamente la expresión entre paréntesis dentro de cada título, tal y como se aprecia en la ilustración 7.

Tras esto, llega la parte importante: obtener en local todos los CSV. Para ello, se elaboró un bucle a modo de función que recorría cada enlace y accedía a él gracias a la librería Selenium. Al entrar en cada uno a través de un driver, que en este caso será Chrome, la descarga inicia automáticamente. Para añadirles el título, utilizo una función lambda que accede al último CSV descargado, lo abre, añade una columna con la fecha y lo guarda de nuevo, pero en una ubicación diferente. A continuación, se muestra un pequeño fragmento del código utilizado, en

el que se ve cómo el dataframe2 correspondiente al resultado en ParseHub es “Descargar” y se trata a cada CSV como “df”:

```
for i in descargar.index:
  if i > 49: break
  else:
    driver.get(descargar.iloc[i,0])

    time.sleep(10)

    fld = 'C:\\Users\\María\\Downloads'
    files = Path(fld).glob('*.csv')
    latest = max(files, key=lambda f: f.stat().st_mtime)
    latest = str(latest)
    df = pd.read_csv(latest, sep = ';', encoding = 'ISO-8859-1', header = 0)

    df['fecha']=descargar.iloc[i,1]

    path = ('csvs/') + str(i) + ('.csv')
    df.to_csv(path)
```

Una vez ejecutado, dada la magnitud de las descargas, se concatenan todos los CSV en dos partes diferentes. Por un lado, los 45 primeros y por otro, los 44 últimos. Así pues, contamos con todos los registros subidos por el Ayuntamiento de Madrid en un entorno local.

4.2.3 Preprocesamiento de todos los registros

Llegados al punto en el que todos los registros de cada comercio de la ciudad de Madrid están en dos conjuntos de datos en local, es necesario pasar un preprocesamiento que consiste en su limpieza, transformación y carga en un último fichero de datos.

Los dos archivos que contienen los registros ocupan 3.66 GB y 3.93 GB respectivamente, por lo que es clave eliminar todo aquello que no sea de utilidad para el análisis, visualizaciones y modelo. La razón de que los ficheros sean tan pesados se debe a que contamos con todas las veces que se ha registrado cada local desde el año 2014, por lo que habrá un registro por cada mes que un local se mantenía abierto, cerrado, de baja, reunificado o utilizado como vivienda. Es necesario eliminar todas las filas que estén duplicadas y no contengan información nueva. No obstante, realizaremos primero una limpieza y homogenización de los campos para eliminar debidamente todo aquello que se repite, de manera que no se queden registros repetidos que no han sido eliminados por estar mal escritos o similares. Para realizarlo, continuaremos utilizando el lenguaje de programación Python a través de notebooks en Jupyter.

[2] Estructura de datos con dos dimensiones en que se puede guardar datos de distintos tipos en columnas. A diferencia de un dataset, este únicamente contiene una estructura en 2D.

4.2.3.1 Eliminar campos innecesarios

En este punto cada set de datos contiene 133 campos, de los cuales muchos están repetidos y la mayoría vacíos, debido a que se han generado por una mala separación de los registros. En la Ilustración 8 se puede observar lo comentado.

Ilustración 8. Campos innecesarios

```
('Unnamed: 66', '74'),
('Unnamed: 67', '75'),
('Unnamed: 68', '76'),
('Unnamed: 69', '77'),
('Unnamed: 70', '78'),
('Unnamed: 71', '79'),
('Unnamed: 72', '80'),
('Unnamed: 73', '81'),
('Unnamed: 74', '82'),
('Unnamed: 75', '83'),
('Unnamed: 76', '84'),
('Unnamed: 77', '85'),
('Unnamed: 78', '86'),
('ï»¿"id_local"', '87'),
('coordenada_y_agrupacion', '88'),
('id_vial_local', '89'),
('280015659', '90'),
('1', '91'),
('CENTRO', '92'),
('14,000', '93'),
('Unnamed: 4', '94'),
('Justicia', '95'),
('Unnamed: 6', '96'),
('88', '97'),
('0,000', '98'),
('0,000.1', '99'),
('Unnamed: 10', '100'),
('Agrupado', '101'),
('563004', '132'),
('Taberna', '133')]
```

```
('Agrupado', '101'),
('Unnamed: 12', '102'),
('Abierto', '103'),
('31035592', '104'),
('CALLE', '105'),
('Unnamed: 16', '106'),
('BARCELO', '107'),
('NUM', '108'),
('2', '109'),
('B', '110'),
('0', '111'),
('31035592.1', '112'),
('CALLE.1', '113'),
('Unnamed: 24', '114'),
('BARCELO.1', '115'),
('NUM.1', '116'),
('2.1', '117'),
('B.1', '118'),
('0,000.2', '119'),
('0,000.3', '120'),
('99000056', '121'),
('CENTRO COMERCIAL BARCELO-MERCADO TEMPORAL', '122'),
('MM', '123'),
('Mercado Municipal', '124'),
('Unnamed: 35', '125'),
('503', '126'),
('LA TABERNA DE BARCELO', '127'),
('I', '128'),
('Hostelería', '129'),
('56', '130'),
('Servicios de comidas y bebidas', '131'),
('563004', '132'),
('Taberna', '133')]
```

Tal y como se puede observar, se han creado una infinidad de columnas que no corresponden a ningún campo de valor. Se trata de registros que no fueron debidamente separados por el creador en alguno de los ficheros de microdatos y al concatenarlos con otros datasets han generado una nueva columna. Dado que es un problema irreversible en este punto y es primordial reducir dimensionalidad, eliminamos todos estos campos.

En añadido, nos deshacemos de las columnas que no aportan información al modelo, como pueden ser el número de acceso a la calle, el nombre del edificio en caso de existir, el tipo de acceso al local, el id de la agrupación en caso de tratarse de una agrupación de locales y similares…

Por consiguiente, reducimos a los siguientes 18 campos:

Tabla 3. Campos del conjunto de datos

Nombre	*Descripción*
id_local	Código numérico que identifica cada local
id_distrito_local	Código numérico con el distrito municipal
desc_distrito_local	Literal del distrito municipal
id_barrio_local	Código numérico con del barrio municipal (incluye el código de distrito)

desc_barrio_local	Literal del barrio municipal
desc_seccion_censal_local	Código de distrito más sección censal
coordenada_x_local	Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle
coordenada_y_local	Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle
id_situacion_local	Código numérico que identifica la situación de un local
desc_situacion_local	Tipos de situación de un local: Abierto: Local activo. Cerrado: Local cerrado (sin actividad). Obras: Local en obras. Baja: Local que ha desaparecido. Baja R: Local que ha desaparecido uniéndose a otro
rotulo	Nombre comercial del establecimiento
id_seccion	Código de “Sección” de actividad
desc_seccion	Literal de la “Sección” de actividad
id_division	Código de “División” de actividad
desc_division	Literal de “División” de actividad
id_epigrafe	Código de epígrafe según una tabla de clasificación de actividad de elaboración propia por parte del Ayuntamiento.
desc_epigrafe	Literal de epígrafe de actividad
fecha	Fecha en la que se registró la actividad y situación del establecimiento

Una vez ya contamos con el número de campos a utilizar en la generación del modelo, cargamos todos los registros de nuevo en dos ficheros. Estos serán de tipo CSV y, gracias a la reducción recientemente realizada, hemos conseguido que su tamaño quede en 1.92 y 1.74 GB cada uno.

4.2.4 Limpieza de cada campo

En este punto del proyecto ya tenemos una especie de *data lake*[3] del cual tendremos que obtener la información que nos sea útil. En este caso, si bien el símil es válido para entender el punto en el que nos encontramos, cabe destacar que no se trata de un lago de datos al uso, ya que, pese a tener los datos en crudo y sin procesar, están cargados en ficheros estructurados.

Estos conjuntos de datos en crudo están separados en dos ficheros debido a su elevada dimensión, por lo que realizaremos la limpieza en dos etapas diferenciadas, pero el código a desarrollar será aplicado de igual manera en ambas iteraciones.

El objetivo de realizar esta limpieza es que actualmente tenemos cada registro de cada local repetido 89 veces, en caso de que esté continue registrado desde el año 2014 hasta ahora. Por tanto, la finalidad de todo el proceso de limpieza será eliminar cada registro duplicado que no aporte nueva información, es decir, sí deberemos guardar cada vez que la situación del local cambie o cada vez que la actividad comercial del negocio pase a ser una diferente.

Este borrado de duplicados no podremos aplicarlo en bruto, ya que el conjunto de datos está mayormente basado en campos de texto, por lo que el proceso de limpieza, homogeneización y normalización de las columnas debe ser exhaustivo. De lo contrario, nos quedaríamos con dos registros que nos aportan la misma información con la única diferencia de que el primero se dedica al sector de la “Hostelería”, mientras el segundo está informado como “Hosteleria”. Este ejemplo sería un caso fácil de resolver, pero el problema viene cuando se dan palabras mal escritas, caracteres no reconocidos por la codificación del ordenador que se utilice o cambios de nombre en la consideración de una sección comercial, por ejemplo.

4.2.4.1 Normalización de los campos

La gran mayoría de campos, tal y como he mencionado anteriormente, son de texto, por lo que el proceso de limpieza resultó tedioso y repetitivo. Esto se debe a que no existe en programación ningún método estandarizado que nos permita corregir las palabras mal escritas por un error humano. No obstante, sí que hay funciones de normalización de texto ya creadas.

Normalización Unicode

El estándar Unicode define dos tipos formales de equivalencia entre caracteres: la equivalencia canónica y la equivalencia de compatibilidad. La equivalencia canónica es una equivalencia fundamental entre caracteres o secuencias de caracteres que representan el mismo carácter abstracto y que, cuando se muestran correctamente, deben tener siempre el mismo aspecto y comportamiento visual. La ilustración 9 muestra este tipo de equivalencia con ejemplos de varios subtipos. [28]

[3] Data lake: Un lago de datos es un depósito de almacenamiento que guarda una gran cantidad de datos en bruto en su formato nativo hasta que se necesite para las aplicaciones de análisis.

Ilustración 9. Ejemplos de equivalencia canónica
Fuente: Unicode Reports

Subtype	Examples
Combining sequence	Ç ⟷ C+◌̧
Ordering of combining marks	q+◌̇+◌̣ ⟷ q+◌̣+◌̇
Hangul & conjoining jamo	가 ⟷ ᄀ + ᅡ
Singleton equivalence	Ω ⟷ Ω

La equivalencia de compatibilidad es un tipo de equivalencia más débil entre caracteres o secuencias de caracteres que representan el mismo carácter abstracto (o secuencia de caracteres abstractos), pero que pueden tener apariencias visuales o comportamientos distintos. Las apariencias visuales de las formas equivalentes de compatibilidad suelen constituir un subconjunto de la gama esperada de apariencias visuales del carácter (o secuencia de caracteres) al que son equivalentes. Sin embargo, estas formas variantes pueden representar una distinción visual que sea significativa en algunos contextos textuales, pero no en otros. Por ello, es necesario tener más cuidado para determinar cuándo es apropiado utilizar un equivalente de compatibilidad. Si la distinción visual es estilística, podría utilizarse el marcado o el estilo para representar la información de formato. Sin embargo, algunos caracteres con descomposiciones de compatibilidad se utilizan en la notación matemática para representar una distinción de naturaleza semántica; sustituir el uso de códigos de caracteres distintos por un formato en tales contextos puede causar problemas. La ilustración 10 ofrece ejemplos de equivalencia de compatibilidad. [28]

Ilustración 10. Ejemplo de caracteres con equivalencia de compatibilidad.
Fuente: Unicode Reports

Subtype	Examples
Font variants	ℌ → H
	ℍ → H
Linebreaking differences	[NBSP] → [SPACE]
Positional variant forms	ﻉ → ع
	ﻊ → ع
	ﻋ → ع
	ﻌ → ع
Circled variants	① → 1

Esencialmente, el algoritmo de normalización de Unicode reglas de descomposición y composición para transformar cada cadena en una de las formas de normalización de Unicode.

Las cuatro formas de normalización se resumen en la ilustración 11.

Ilustración 11. Formas de normalización

Fuente: Unicode Reports

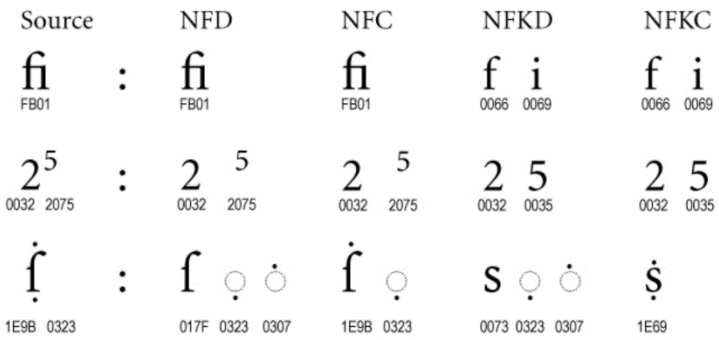
Form	Description
Normalization Form D (NFD)	Canonical Decomposition
Normalization Form C (NFC)	Canonical Decomposition, followed by Canonical Composition
Normalization Form KD (NFKD)	Compatibility Decomposition
Normalization Form KC (NFKC)	Compatibility Decomposition, followed by Canonical Composition

Al algoritmo tendremos que indicarle una serie de parámetros. Existen dos formas de normalización que convierten a los caracteres compuestos: La forma de normalización C y la forma de normalización KC. La diferencia entre ellas depende de si el texto resultante debe ser un equivalente canónico al texto original no normalizado o un equivalente de compatibilidad al texto original no normalizado. En NFKC y NFKD, se utiliza una K para compatibilidad y así evitar la confusión con la C, que significa composición.

En definitiva, el uso de una forma u otra se traducirá en lo siguiente sobre un texto:

Ilustración 12. Formas resultantes en la normalización

Fuente: Unicode Reports



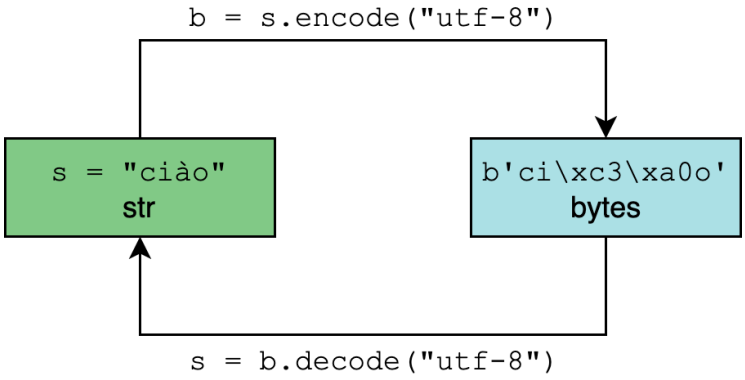
Este método tendrá que ser utilizado en todos los campos de texto de nuestros datos, ya que al tratarse de registros recogidos a lo largo de más de 8 años, no están homogéneamente documentados. No solo a la hora de anotar cada registro, sino que en cada tiempo distinto se pueden haber utilizado diferentes codificadores.

El proceso por seguir en cada campo es el siguiente: En primer lugar, ponemos el texto en minúscula. Posteriormente, utilizamos el método *normalize* introduciendo en cada caso los parámetros que sean necesarios. Por último, aplicamos la codificación 'ascii' para después decodificar en 'utf-8'. Un ejemplo de esta secuencia de código sería el siguiente:

```
df['desc_epigrafe'].str.lower().str.normalize('NFKD').str.encode('ascii','ignore').str.decode('utf-8')
```

La razón de este proceso es que primero debemos traducir cada carácter a su forma descompuesta con el método 'normalize'. En segundo lugar, el texto normalizado se convierte en una cadena de bytes ASCII y por último, decodificamos en UTF-8 y volvemos a la cadena decodificada a partir de los bytes dados. La razón de no decodificar de nuevo a ASCII reside en que UTF-8 es un superconjunto de esta. Cualquier cadena ASCII se puede decodificar en cualquiera de las codificaciones compatibles y obtener el mismo resultado.

Ilustración 13. Codificar vs Decodificar
Fuente: Realpython.com



Tras esto, hay que añadir nuevas funcionalidades a nuestra función de normalización. En este caso modificaremos en base al campo a tratar, ya que toca corregir errores dados por eliminar signos diacríticos[4] o debidos a un mal registro. En el caso de los barrios, los distritos y la descripción de la situación de local, reemplazamos uno a uno cada registro con fallos, como puede ser 'nio jess' en 'niño jesus'.

En el caso de las descripciones de la actividad comercial, lo cual incluye a los campos de sección, división y epígrafe, dado a la gran cantidad de valores únicos, debemos corregirlo haciendo uso de los datos del Código CNAE-2009 publicados por el INE. Los registros a corregir se basan en este código, por lo que al tener las columnas de identificador y el literal, podemos cruzar los identificadores con los reales publicados por el INE y sustituir la descripción por su forma correcta. Para ello, creamos una función que cruza nuestro conjunto de datos junto con

[4] En español: á, é, í, ó, ú, la diéresis (¨) o la virgulilla (˜).

el del CNAE y en aquellos registros en los que el identificador coincide, reemplaza el literal por la descripción utilizada en el código de Clasificación Nacional de Actividades Económicas.

Por último en lo relativo al texto, eliminamos espacios innecesarios en algunos registros. Esto sucede debido a que, con el objetivo de embellecer la apariencia del conjunto de datos en Excel, se han registrado algunos campos con espacios delante y detrás del texto. Para solucionarlo aplicamos una función que se basa en el uso del método *strip*, que elimina todos aquellos espacios que existan delante o detrás de la cadena de texto.

Con todo ello tendríamos todas las columnas de texto homogeneizadas en nuestro *dataset*, por lo que es el momento de enfocarse en corregir errores observados en el resto. Más allá, una conversión de los datos numéricos y de fecha al tipo de datos adecuado, encontramos un fallo en el registro de los números identificadores de los barrios. Resulta que en diferentes momentos del tiempo, el mismo barrio se identifica con un número diferente.

En el portal de datos abiertos del Gobierno de España, *Datos.gob*, existe un fichero de datos que contiene el nombre y código de cada uno de los 131 barrios de la ciudad de Madrid[5]. Dicho fichero lo utilizaremos para corregir nuestro campo 'id_barrio_local'. Puesto que previamente hemos limpiado la columna con el nombre, en este caso cruzaremos dicha columna con su homóloga ubicada en el conjunto de datos llamado Barrios, obtenido a través de *Datos.gob*. De esta forma, seguimos el mismo proceso seguido con los literales de la actividad comercial, pero de forma inversa.

Llegados a este punto, el proceso de normalización de los campos habría terminado. La razón de haberlo configurado a través de funciones es, una vez se haya aplicado sobre la primera mitad de los datos, hacer lo mismo sobre la segunda y última partición y que los resultados sean los deseados.

4.2.4.2 Tratamiento de los datos faltantes

En el proyecto que nos atañe, la presencia de registros nulos es notablemente destacable. Al igual que sucedía en el apartado de limpieza, puesto que se trata de un conjunto de datos formado a partir de información recogida en 8 años diferentes, los formatos a veces cambian y el error humano es bastante frecuente. Ello no solo resulta en un tedioso proceso de limpieza de los campos, sino que a su vez se generan muchos datos nulos.

Al obtener el conteo de datos faltantes en ambas particiones del conjunto de datos, vemos cantidades muy elevadas, sin embargo esto sucede a causa de que todavía no se ha llevado a cabo el borrado de registros duplicados y en muchos casos, cuando no se tiene cierta información de un establecimiento un año, sucede de la misma forma en el resto, por lo que tendríamos 89 registros nulos de dicho local.

[5] https://datos.gob.es/en/catalogo/l01280796-barrios-municipales-de-madrid

En el caso del campo 'id_situacion_local', la cantidad es mayor al resto. Dado que se trata de una simple variable categórica sobre la de descripción de la situación del local, decidimos deshacernos de ella. No obstante, no actuaremos de la misma manera sobre el resto de nulos.

Registros nulos encontramos sobre todos los campos. Esto se debe a que, en la mayoría de ocasiones, en aquellas filas en las que la sección del CNAE no está registrada, tampoco lo está la información de la división ni del epígrafe, y por tanto sus respectivos identificadores tampoco. Además, a su vez sucede que hay locales que son nulos en todos o casi todos sus diversos registros a lo largo del tiempo. Es aquí donde vamos a encontrar la manera de rellenarlos, ya que, como he mencionado, debido a que tenemos varios registros de la información de cada establecimiento, podremos rellenar registros de otros años que estén faltantes con sus mismos datos. Esto podremos hacerlo sobre todas las columnas excepto sobre la de situación del local, ya que esta cambia según la fecha, puede ser que un local abierto pase a estar de baja o utilizándose como vivienda de un momento a otro. Sin embargo, su actividad comercial, siempre que se mantenga un local con el mismo rótulo (nombre del establecimiento) e identificador, su localización o su actividad comercial será la misma. Estos serán nuestros objetivos por rellenar. Para ello, creamos un registro único de cada local sin nulos y lo guardamos en otro *dataset*. A través de él rellenaremos en nuestro conjunto de datos con todos aquellos valores que coincidan como el mismo negocio. Gracias a ello, hemos conseguido rellenar más de 500000 valores nulos sobre el total de registros, que de momento continúa siendo 14 millones.

Tras ello, seguimos teniendo valores nulos que no podemos rellenar, tal y como se puede ver en la figura 14. Todos aquellos locales que no contengan información sobre su actividad comercial o localización no podrán ser introducidos al modelo de aprendizaje supervisado, pero de cara a la visualización los vamos a mantener, ya que no es correcto eliminar valores nulos, dado que representan un establecimiento existente en la ciudad. Además, la dimensión de estos valores disminuirá cuando procedamos a eliminar los registros duplicados, ya que en muchos casos se trata de registros de un local que faltan durante varios años.

Ilustración 14. Valores faltantes resultantes

```
id_local                      152200
desc_situacion_local          152200
id_distrito_local             152200
desc_distrito_local           152200
desc_seccion_censal_local     152218
coordenada_x_local            152200
coordenada_y_local            152200
rotulo                        292282
fecha                              0
id_barrio_local              1866754
desc_barrio_local            1866754
id_seccion                   1866754
desc_seccion                 1866754
id_division                  1866754
desc_division                1866754
id_epigrafe                  1866754
desc_epigrafe                1866754
id_epigrafe_bueno            1866754
desc_epigrafe_bueno          1866754
```

4.2.4.3 Eliminar registros duplicados

El fichero de datos final se conforma tras este paso. En esta fase del preprocesamiento, ambas mitades de nuestro conjunto de datos están debidamente limpiadas y tratadas, por lo que lo único restante por aplicar es la eliminación de los famosos registros duplicados mencionados en diversas ocasiones a lo largo de todo el proceso.

El objetivo a lo largo de todo el ETL se trataba de configurar un conjunto de datos final cargado en un fichero. La razón por la que debemos eliminar duplicados es para conseguir un registro de cada local existente o que haya existido en la ciudad de Madrid en los últimos 8 años con su información comercial, localización y duración, a fin de utilizar esa información en el modelo que realizamos posteriormente. Por ello nos quedamos con el primer registro en el tiempo que haya de cada local y en cada situación (abierto, cerrado, uso vivienda, baja y baja reunificado). Esto es, mantenemos el primer registro de un establecimiento, pero a su vez también el primero que se tenga de cuando cerró, se dio de baja o se empezó a utilizar como vivienda. De esta forma calcularemos la longevidad de cada negocio en base a ello. Por esta razón antes de hacer nada es necesario ordenar toda la información según el dato de la fecha y el identificador de local para así tener un dataframe agrupado por los diferentes establecimientos y con cada registro ordenado por la fecha en la que se incorporó al censo de locales.

¿Por qué no se ha realizado antes? Es probable que sea una cuestión que haya surgido a lo largo de estos últimos pasos. Si se hubiera realizado antes no hubiésemos tenido que reducir a dos mitades todos los registros a fin de introducirlos por el procedimiento de limpieza. En añadido, habría resultado más rápido e incluso podrían haber disminuido los casos de limpieza.

No obstante, a pesar de todo lo apuntado, la razón de ser de la decisión tomada ha sido conseguir eliminar todo lo que realmente esté duplicado. Si aplicamos el método 'drop_duplicates', cuya función es precisamente realizar lo mencionado a lo largo de este apartado, sobre el conjunto obtenido en el punto 4.2.2.1, un negocio cuya localización es

'Pacífico' y otro ubicado en 'Pacfico' se habrían mantenido, o el local llamado "Vinoteca" se consideraría diferente al mismo bajo el nombre "Vinoteca ", cuando ambos aportan la misma información. La idea de realizar la limpieza antes surge de esto, debido a que se probó a realizarlo de esta forma, mas no se consiguió obtener un conjunto final de valor y acorde a la realidad de la ciudad de Madrid.

En virtud de todo ello, el primer paso a ejecutar será concatenar los dos dataframes en los que habíamos partido todos los datos, ya que en caso contrario no podríamos ordenarlo según locales y fechas adecuadamente. Así pues, resulta en un total de 14903355 registros preprocesados. Es en este momento cuando ordenamos como he comentado y aplicamos en este momento el método que nos permite deshacernos de registros duplicados. Si bien es fácil de emplear, debemos tener en cuenta que la supresión debe hacerse teniendo en cuenta que aquellos locales que se repitan, pero cuya situación o rótulo cambie, deben permanecer. Esto último se debe a que, en algunas ocasiones, un local cambia de denominación comercial y pasa a ser otro diferente.

A la hora de aplicarlo, se deben tener en cuenta un par más de consideraciones. Es importante destacar que eliminaremos duplicados, pero quedándonos siempre con el registro más antiguo de cada local. Con todo lo indicado, aplicamos el método para que se supriman en caso de que se repitan los siguientes campos: el identificador de local, distrito, barrio, situación del establecimiento, rótulo y sección, división y epígrafe del CNAE, y a su vez indicamos que debe quedarse con el primero de los registros repetidos, el cual se corresponderá con el momento en el que el negocio se agregó al censo.

Por último, al realizarlo encontramos que se han mantenido locales duplicados debido a lo que se puede observar en la ilustración 15:

Ilustración 15. Ejemplo de duplicado.

4473534,81	Abierto	rótulo no informado	2014-06-01	I	Hostelería	55.0	Servicios de alojamiento	551001.0	hoteles y moteles con restaurante
4473534,81	Abierto	hotel mediodia	2014-09-01	I	Hostelería	55.0	Servicios de alojamiento	551001.0	hoteles y moteles con restaurante

El mismo local se ha mantenido con su misma información dos veces a causa de que en su primera incorporación al censo no se añadió el rótulo y una vez se notificó, se replicó el registro. Aunque tengamos este problema, cargaremos el conjunto de datos resultante en un CSV. Este queda con una dimensión de 379230 registros y 19 variables.

Aun así, realizo de nuevo la eliminación de duplicados de forma que no dependa del rótulo aquellos casos en los que se repita un mismo local, con una misma actividad, id y localización, pero se cambie su rótulo por problemas como el de la imagen, no se tengan en cuenta. Esto lo hago porque de cara a la visualización de los datos y elaboración del modelo, no debo tener en cuenta locales que aporten la misma información. Este último conjunto de datos es también cargado en un CSV y esta vez resulta en 247086 registros y los mismos campos, deshaciéndonos así de un 98,12% de los registros extraídos del censo, que no aportaban información útil.

4.2.5 Foto actual de la actividad comercial madrileña

El sentido del presente proyecto gira en torno al análisis de los sectores de actividad económica en la ciudad de Madrid. Gracias al histórico del censo de locales y su actividad comercial que aporta el portal de datos abiertos del ayuntamiento de Madrid, hemos podido llevar a cabo este proyecto. No obstante, como añadido a las conclusiones y resultados que pretendemos obtener, es de gran interés a su vez realizar un análisis sobre la actualidad comercial en la ciudad. Esto es posible porque contamos ya con unas funciones de preprocesamiento aplicables a cualquier conjunto de datos con las mismas características y, únicamente con descargase la última actualización del censo aportada por el ayuntamiento, ya tendríamos los datos necesarios para realizar una foto actual del comercio en Madrid.

Así pues, una vez descargados los datos, los insertamos en el mismo código expuesto en el apartado anterior y en el orden idéntico. La forma de los datos, sus tipos y las variables son las mismas que se expusieron en la tabla 3. Es por ello que el tratamiento será idéntico, a excepción del registro de coordenadas, que esta vez se va a utilizar para realizar un mapa de los establecimientos actuales. En el caso del histórico no tenía sentido tratarlo, además de que el objetivo del conjunto anterior era dar con la duración de cada negocio. Sin embargo, en este punto realizaremos una conversión de coordenadas UTM a coordenadas geográficas.

En el sistema de coordenadas UTM se utiliza una cuadrícula para especificar ubicaciones en la superficie de la Tierra. El sistema UTM no es una única proyección cartográfica, sino una serie de zonas. Las unidades para las coordenadas empleadas en este sistema son metros, por lo que no servirán para graficar en la herramienta de visualización a utilizadas y el paso a seguir ahora será convertirlas a coordenadas geográficas, expresadas en grados decimales que miden los grados de latitud y de longitud.

Este proceso también se realizará con Python, esta vez a través de las librerías *geopy* y *utm*. Gracias a la segunda mencionada, que contiene un método específico para realizar nuestro objetivo, conseguimos convertir las coordenadas x e y del local en un vector de coordenadas geográficas. Dicho vector posteriormente lo separamos en latitud y longitud para conseguir dos campos nuevos.

Así pues, conseguimos un *set* de datos con 155975 registros que representa la realidad comercial de la ciudad de Madrid a fecha de marzo de 2022.

Los resultados de la visualización del conjunto resultante se expondrán en el correspondiente apartado.

4.2.6 Modelos Machine Learning

Alcanzado este punto del proyecto es el momento de generar un modelo de datos entrenado con el conjunto de datos que hemos preprocesado. El objetivo del modelo es funcionar como motor de recomendación de nuestra propuesta. Inferirá sobre la duración de un tipo de comercio en una localización determinada y así ofrecerá al cliente una propuesta con la mejor idea para su inversión, es decir, en caso de que un inversor esté interesado en abrir un negocio de peluquería canina, a través el modelo se le generará un orden con las localizaciones en las

que peor funciona ese tipo de comercio y aquellas en las que suelen durar más. En el caso contrario, si un cliente quiere realizar una inversión en un local en el barrio de Acacias, se le ofrecerá un listado con los comercios que mejor y que peor funcionan en dicha localización, todo ello basándonos en la duración de cada local.

Una vez sabemos cuál será el objetivo del modelo, podemos comprobar que la duración será la variable objetivo, es decir, el campo a inferir. Dicha variable no está disponible en el conjunto de datos por lo que tendremos que crearla como primer paso.

4.2.6.1 Creación de la variable objetivo

Este proceso de nuevo se realizará en el lenguaje Python a través de bucles que vayan calculando una a una cada duración.

Partimos del conjunto de datos resultado del apartado 4.2.4.3. Sobre él creamos un nuevo campo de tipo numérico y, filtrando sobre los locales abiertos, realizamos una resta entre la fecha actual y su fecha de apertura.

Para realizar lo mismo sobre los locales cuya situación no es “Abierto” primero los seleccionamos y guardamos en un *dataframe*. Posteriormente deberemos buscar esos mismos locales en la partición de abiertos y restar la fecha del registro del local no abierto con su fecha registrada como apertura. De esta forma, ya tenemos el número de días de duración de cada negocio.

4.2.6.2 Transformación adicional para el modelo

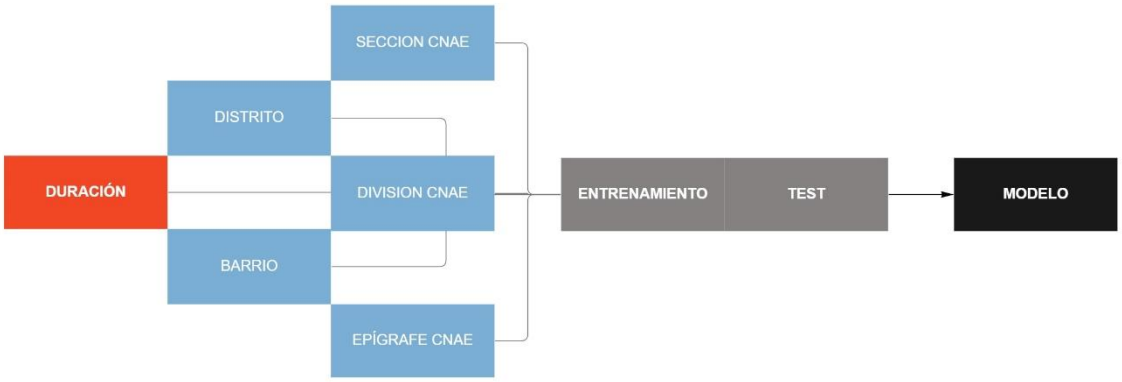
Contamos con conjunto de datos que tiene registrados todos los locales que han surgido desde 2014 con su duración en días. Tenemos registros de todas las situaciones diferentes de cada local, es decir, todo el ciclo de vida que ha seguido desde que abrió hasta que cerró, en caso de que lo hiciera. Esto quiere decir que tenemos el registro perteneciente al supermercado Ferpal acerca del momento en el que se registró por primera vez con una duración de 2891 días, esto es, la resta entre la fecha actual y su apertura, mas a su vez contamos con el registro de cuando cerró y una duración de 2771 días. Es por ello que debemos eliminar todos aquellos negocios que aparecen como abiertos, pero que también cuentan con un registro de cierre en una fecha posterior. Así pues contaremos introduciremos al modelo todos los locales que siguen abiertos en Madrid y los que alguna vez lo estuvieron y cerraron, o se dieron de baja, se convirtieron en vivienda…

4.2.7 Modelo de aprendizaje supervisado: regresión

Hemos comentado anteriormente que el objetivo es inferir la probabilidad de éxito de un negocio dependiendo de su actividad comercial y localización. Para ello hemos calculado cuánto ha durado cada negocio en la ciudad de Madrid desde el año 2014 y ahora pretendemos introducir este registro de históricos en un modelo que sea capaz de reconocer un patrón y aplicarlo sobre datos nuevos que introduzcamos.

Vamos a predecir la duración que tendrá un local que quiera inaugurar un inversor. El problema se deberá resolver mediante aprendizaje supervisado, ya que contamos con una serie de datos con todas sus variables y su duración final, es decir, sus entradas y resultados.

Ilustración 16. Estructura del modelo



En la ilustración 16 se muestra, *grosso modo*, cuál será la estructura del modelo que pretendemos crear. Con las entradas que se indican, entrenaremos un modelo que aplicaremos sobre un conjunto de pruebas para obtener la evaluación de su funcionamiento. La finalidad es conseguir un modelo al que seamos capaces de introducir los campos predictores y obtener la Duración esperada.

Este, tal y como lo hemos planteado hasta aquí es de regresión, debido a que se trata de un proceso estadístico predictivo cuyo objetivo es encontrar un número continuo. No obstante, en la búsqueda de encontrar el mejor modelo, probaremos más adelante con un modelo de clasificación, ya que en este caso podemos traducir la problemática a una categorización entre éxito o no éxito. Continuando con la regresión, aplicaremos el algoritmo *Linear Regression*, basado en una regresión lineal.

4.2.7.1 Base matemática de las regresiones lineales

La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos [30]. Estos modelos pueden emplearse para estimar la influencia que tienen los predictores sobre la variable objetivo, pero se debe analizar con cautela para no confundir causa y efecto.

Las técnicas de regresión lineal permiten crear un modelo lineal. Este modelo describe la relación entre una variable dependiente y (también conocida como la respuesta o variable objetivo) como una función de una o varias variables independientes Xi (denominadas predictores o variables regresoras). La ecuación general correspondiente a un modelo de regresión lineal es:

$$Y=\beta_0+\sum \beta_i X_i+\epsilon_i$$

donde β representa las estimaciones de parámetros lineales que se deben calcular y ϵ representa los términos de error.

En nuestro caso, dado que contamos con varias variables predictoras, emplearemos una regresión lineal múltiple. Esta regresión tiene múltiples Xi para predecir la respuesta, Y. Este es un ejemplo de la ecuación para 2 variables independientes:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\epsilon$$

Nuestra ecuación tendrá una forma similar, pero con tantas betas como variables regresoras introduzcamos. En ella, cada término de la ecuación significa lo siguiente:

β0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero. Nos ayudará ligeramente a observar si hay efectos que influyen sobre la variable y que no están contemplados por las predictoras [31].

βi: es el efecto promedio que tiene el incremento en una unidad de la variable predictora Xi sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión [31].

ei: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo [31].

Esta es la base para entender el porqué de las fórmulas y coeficientes de la regresión lineal múltiple, sin embargo a lo largo de la explicación del modelo introduciremos el resto de términos necesarios.

4.2.7.2 Variables categóricas

Una variable categórica (a veces llamada variable nominal) es aquella que tiene dos o más categorías, pero no hay un orden intrínseco para las categorías. Por ejemplo, una variable binaria (como una pregunta de sí/no) es una variable categórica que tiene dos categorías y no hay un orden intrínseco en las categorías. El color del pelo también es una variable categórica que tiene varias categorías (rubio, castaño, moreno, pelirrojo, etc.) y, de nuevo, no hay una forma acordada de ordenarlas de mayor a menor.

Cuando se introduce una variable categórica como predictor, una de las categorías, o nivel, se considera la de referencia (normalmente codificado como 0) y el resto se comparan con ella. En el caso de que el predictor categórico tenga más de dos niveles, se generan lo que se conoce como variables *dummy* o ficticias, que son variables que surgen para cada uno de los niveles del predictor categórico y que pueden tomar el valor de 0 o 1. Cada vez que se emplee el modelo para predecir un valor, solo una variable *dummy* por predictor adquiere el valor 1 (la que coincida con el valor que adquiere el predictor en ese caso) mientras que el resto se consideran 0. El valor del coeficiente parcial de regresión βi de cada variable *dummy* indica el porcentaje promedio en el que influye dicho nivel sobre la variable dependiente Y en comparación con el nivel de referencia de dicho predictor [31].

Hay un concepto a tener en cuenta cuando tratamos con variables de este estilo y se denomina la “trampa de la variable ficticia”. Esta se produce cuando dos o más variables

ficticias creadas por la codificación de una sola vez están correlacionadas (multicolineales). Es causado por introducir todas las variables *dummy* creadas a partir de un campo categórico en un modelo. Si introducimos todas, cada una de esas *dummy* será combinación lineal exacta del resto. Esto significa que una variable puede predecirse a partir de las otras, lo que dificulta la interpretación de los coeficientes predichos en los modelos de regresión. En otras palabras, el efecto individual de las variables ficticias en el modelo de predicción no puede interpretarse bien debido a la multicolinealidad [34]. Es fácil de solucionar, se consigue no generando una de las categorías en *dummy* y considerándola como el nivel de referencia que mencionamos al comienzo.

En el modelo que vamos a crear en este proyecto, todas las variables predictoras son categóricas, por lo que deberemos crear variables *dummy* para todos nuestros campos.

4.2.7.3 Nulos en el modelo

A pesar del tratamiento de nulos que hemos realizado en la fase de preprocesamiento, aún contamos con una serie de nulos que no hemos sido capaces de rellenar. Como se puede ver en la ilustración 17, la gran mayoría de datos faltantes se encuentran en registros de establecimientos no abiertos. Esto tiene sentido si consideramos que un local, una vez cerrado, pasado a ser vivienda, o dado de baja, se deja de registrar su actividad comercial, que es donde encontramos casi la totalidad de todos los nulos.

Ilustración 17. Nulos en el modelo

```
Cerrado             59779
Uso vivienda        15104
Abierto              4904
Baja                 2314
Baja R               1691
En obras             1219
Baja PC Asociado        5
Name: desc_situacion_local
```

Estos datos no pueden ser introducidos para entrenar en ningún modelo, sin embargo borrarlos no es una opción, ya que pueden tener un comportamiento propio que vayamos a ignorar, por tanto los sustituiremos por un valor concreto y diferente al resto.

4.2.7.4 Creación del modelo de regresión

El algoritmo a utilizar en este modelo será el expuesto anteriormente, una regresión lineal múltiple, la cual la implementaré en primer lugar a través de la librería *scikit-learn* y posteriormente realizaremos un modelo son *statsmodels.api* para obtener más métricas. Con *Statsmodels* seguiremos en gran medida el modelo tradicional en el que queremos saber lo bien que un modelo determinado se ajusta a los datos, y qué variables "explican" o afectan al resultado, o cuál es el tamaño del efecto. *Scikit-learn* sigue la tradición del aprendizaje automático, en la que la tarea principal es elegir el mejor modelo de predicción.

Para el primer modelo de regresión, generamos las variables dummy para todos los campos e introducimos las indicadas en azul en la ilustración 16 dentro del vector de variables

predictoras. Separamos el conjunto de X e Y en entrenamiento y test siguiendo la proporción 80-20 respectivamente. Entrenamos el modelo con el conjunto de *train* y predecimos sobre el *test*. Los resultados de la evaluación se mostrarán en el apartado correspondiente a ellos.

Para el segundo modelo el proceso es similar. En primer lugar, definimos por separado cada una de las 5 variables regresoras y la variable objetivo. Posteriormente definimos la ecuación de regresión y ajustamos el modelo de regresión lineal en los parámetros siguiendo el método de los mínimos cuadrados ordinarios (MCO). De nuevo, los resultados y la conclusión de los modelos se encuentran más adelante.

Tras la realización de ambos, visto en el apartado de resultados que los resultados del modelo no son lo suficientemente buenos y que, en añadido, la cantidad de variables para una regresión lineal es desorbitada, debemos probar con otro modelo. Dado que hemos tenido que realizar variables *dummy*, contamos con un modelo con muchos campos y, si son necesarios tantos campos para explicar el comportamiento de una variable, es que no se encuentra un comportamiento claro. Teniendo en cuenta que el error en las predicciones de duración obtenido y la cantidad de campos introducidos, esperamos un mejor resultado, por lo que traduciremos el problema a un algoritmo de clasificación.

4.2.8 Modelo de aprendizaje supervisado: clasificación

Según lo comentado al comienzo del trabajo, la clasificación utiliza un algoritmo para predecir salidas categóricas o de clase. Hasta este punto, el objetivo ha sido inferir sobre la duración de un negocio, por lo que esta no podrá ser la entrada del modelo. Seguiremos el mismo esquema que en la ilustración 16, mas con un cambio en la variable a predecir.

El algoritmo a utilizar en este caso será una Regresión Logística, también a través de la librería *Scikit-learn*. En este caso también realizaremos dos modelos, uno con las variables categóricas convertidas en *dummies* y un segundo realizado con una previa selección de variables.

4.2.8.1 Base matemática de la regresión logística

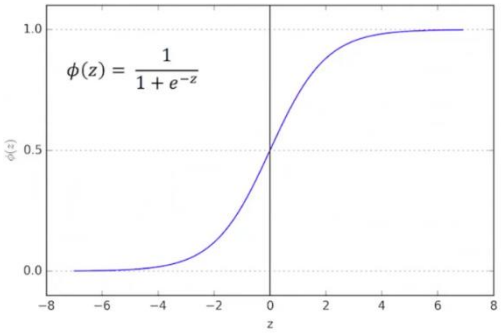
La regresión logística es un algoritmo que realiza una clasificación binaria modelando una variable dependiente (Y) en términos de una o más variables independientes (X). En otras palabras, es un modelo generalizado que predice la probabilidad de que ocurra un evento. En concreto, la regresión logística utiliza la regresión lineal regular para modelar la función logística, también llamada sigmoide [32]:

$$l = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_i x_i ...$$

Esta función devuelve una salida entre 0 y 1 para todos los valores posibles de X. Una representación visual de la función logística (sigmoide) es la siguiente:

Ilustración 18. Función sigmoide

Fuente: medium.com



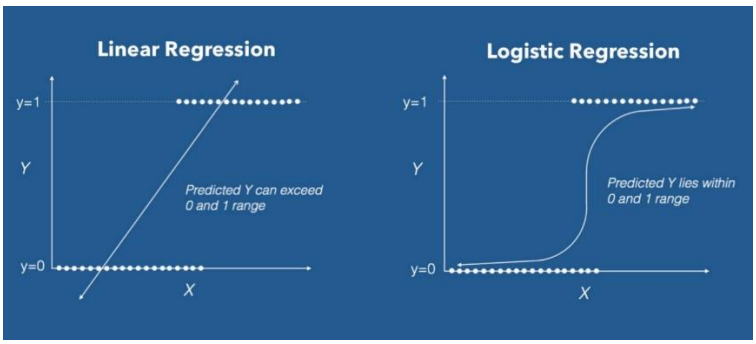
Esta función ayuda al modelo de regresión logística a ajustar los valores de (-k,k) a (0,1). La regresión logística se utiliza principalmente para tareas de clasificación binaria, sin embargo, puede utilizarse para la clasificación multiclase.

4.2.8.1.1 ¿Por qué se le llama "regresión"?

La razón es que, al igual que la regresión lineal, la regresión logística parte de una ecuación lineal. Sin embargo, esta ecuación consta de probabilidades logarítmicas que, además, pasan por una función sigmoide que reduce la salida de la ecuación lineal a una probabilidad entre 0 y 1. Y, podemos decidir un límite de decisión y utilizar esta probabilidad para realizar la tarea de clasificación. Por ejemplo, supongamos que estamos prediciendo si va a llover mañana o no basándonos en el conjunto de datos dado, y si después de aplicar el modelo logístico, la probabilidad resulta ser del 90%, entonces podemos decir con seguridad que es muy posible que llueva mañana. Por otro lado, si la probabilidad resulta ser del 10%, podemos decir que no va a llover mañana, y así es como podemos transformar las probabilidades en binarias [33].

Ilustración 19. Diferencia con la regresión lineal

Fuente: medium.com/@curryrowan



4.2.8.2 Nueva variable objetivo

La finalidad en este punto es traducir una variable numérica y ordinal a una categórica. Esto lo haremos incluyendo un conjunto de duración de días en una clase y el restante en otra clase.

Para ello definiremos un nuevo concepto en el modelo, el éxito. Esto quiere decir que definiremos que una longevidad del negocio conllevará la rentabilidad y duración larga del

establecimiento y todo lo que esté debajo de ese umbral, se tratará de una opción de negocio más inestable. Con el objetivo de determinar ese umbral, obtenemos los cuartiles de la variable:

Ilustración 20. Cuartiles de la variable objetivo

```
count    215566.000000
mean       1455.233316
std        1341.139158
min           0.000000
25%           0.000000
50%        1553.000000
75%        2891.000000
max        2891.000000
Name: Duracion, dtype: float64
```

Aquí podemos observar que todo el último cuartil de los datos ya alcanza la duración máxima de días, que se corresponde con todos los establecimientos que continúan abiertos desde la primera vez que se comenzó a registrar el censo, esto es, hace ahora 8 años. Debido a que existe una cantidad considerable de registros que alcanzan el máximo absoluto, definimos este como el corte entre el éxito y no éxito. Por tanto, todos aquellos que se encuentren por debajo recibirán el valor 0 en la nueva variable objetivo y los restantes, un 1.

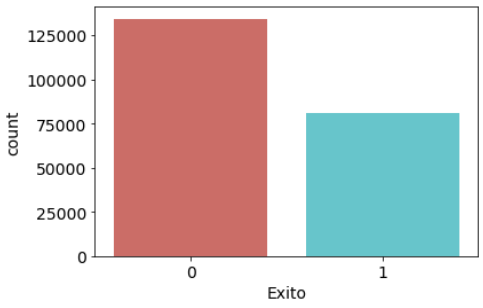
Esto no quiere decir que todo se reduzca a éxito o no definitivamente, sino que, regulando los parámetros del algoritmo, obtendremos la probabilidad de encontrarse en una de las dos clases.

4.2.8.3 Creación del modelo de clasificación

En primer lugar, realizamos una serie de gráficos para obtener las primeras hipótesis.

En la ilustración 21 observamos que no contamos con un set de datos balanceado, lo que quiere decir que hay más registros de una de las clases. Esto puede provocar un pequeño sesgo en el entrenamiento del modelo, pero dado que la diferencia no es muy grande,

Ilustración 21. Balance del set de datos



Seguidamente, podemos observar en qué lugares y actividades predominan los establecimientos con lo que nosotros hemos considerado como éxito, lo cual considera los establecimientos con una duración mayor a 2890.

Ilustración 22. Frecuencia de éxito por sección

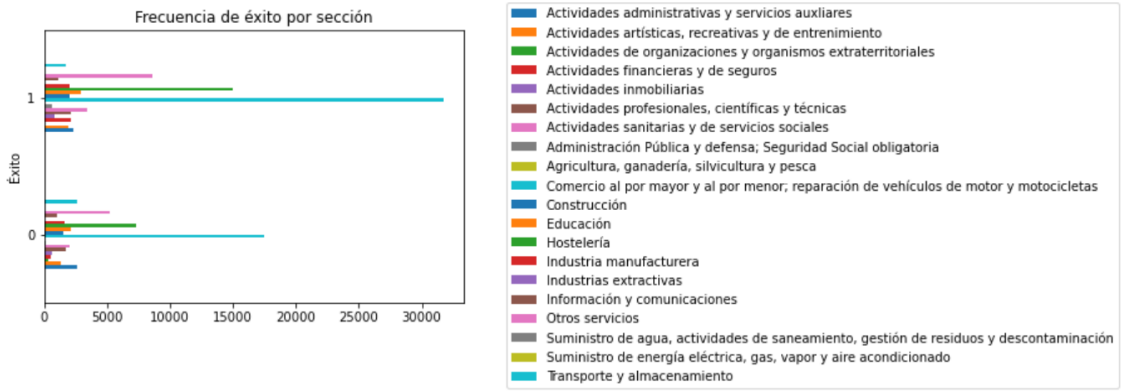


Ilustración 23. Frecuencia de éxito por distrito

En ellos podemos observar que en los distritos 10, 13 y 15, correspondientes a los barrios latina, puente de Vallecas y ciudad lineal respectivamente, es donde hay mayor número de ceros. Además, en el anterior vemos que el mayor número de unos pertenece al comercio al por menor y al por mayor, la hostelería y actividades sanitarias. Sucede de la misma manera en el caso de los no éxitos. Estas son las primeras conclusiones que podemos ir obteniendo, entre otras, pero no será hasta que lancemos el modelo que consigamos un resultado más claro, ya que la correlación entre las categorías mencionadas no tiene por qué significar que sean la causa de que así suceda.

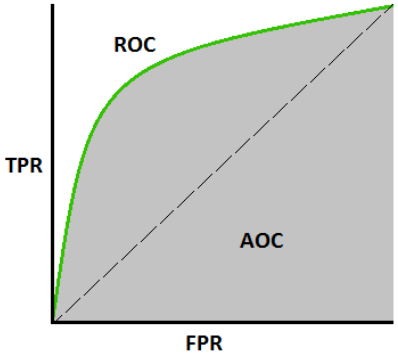
Tras este pequeño análisis de los datos, no realizamos ninguna modificación y continuamos con la creación de los modelos.

En el primer modelo, importamos la regresión logística de la librería y generamos las variables ficticias de todas las categorías. Esto lo haremos porque a la hora de introducir categorías, tal y como explicamos previamente en la memoria, si estas no son ordinales, podemos confundir al algoritmo si las introducimos sin modificar. Tomaremos un 75% del conjunto para entrenar y un 25% sobre el que realizaremos las pruebas.

Una vez entrenado y testeado, obtenemos las métricas de evaluación y la matriz de confusión, una técnica que nos ayuda a resumir el rendimiento de los algoritmos de clasificación. En ella,

observaremos los siguientes 4 valores: los verdaderos positivos y verdadero negativo, que son el número de casos en los que la clase real es 1 y la clase predicha también es 1, y dos tipos de errores, el falso positivo y falso negativo. En añadido, a su vez obtenemos la curva AUC – ROC, la cual indica en qué medida el modelo es capaz de distinguir entre clases. Cuanto más alto sea el AUC (por sus siglas en inglés, área bajo la curva), mejor será el modelo para predecir las clases 0 como 0 y las clases 1 como 1. Por analogía, cuanto más alto sea el AUC, mejor será el modelo para distinguir entre establecimientos con éxito y sin ellos. La curva ROC se representa con el ratio de verdaderos positivos (TPR) frente al ratio de falsos positivos (FPR), donde el TPR está en el eje de ordenadas y el FPR en el eje x [35].

Ilustración 24. Curva AUC - ROC
Fuente: My Photoshopped Collection



Por último, el segundo modelo se realizará de manera similar, pero con pasos añadidos. Primeramente, tras la generación de los *dummies* y la selección de las variables independientes y la dependiente, aplicamos al conjunto de datos una selección de variables llamada “Information value”. Una reducción de variables podría ser interesante para obtener un mejor resultado, ya que así podremos optimizar su uso y facilitar la comprensión del modelo.

Selección de variables

Information value es una técnica de exploración de datos que ayuda a determinar qué columnas de un conjunto de datos tienen poder predictivo o influencia en el valor de una variable dependiente específica. Esta define lo que es el IV, un valor numérico que cuantifica el poder predictivo de una variable independiente x en la variable dependiente binaria y. El IV es útil para reducir el número de variables como paso inicial en la preparación de la regresión logística, especialmente cuando hay una gran cantidad de variables potenciales, tal y como es nuestro caso. Se basa en un análisis de cada una de las variables independientes por separado, sin tener en cuenta otras variables predictoras. Según el IV que nosotros decidamos, devuelve una serie de campos u otros. Este rango de valores define la capacidad de predicción de las variables que devuelve, tal y como se puede ver en la ilustración 25.

Ilustración 25. Reglas de IV
Fuente: listendata.com

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

Las funciones para obtenerlo, según el lenguaje de programación, se encuentran ya en una librería o es necesario crearla. En nuestro caso, puesto que se trata de un método comúnmente conocido, la hemos creado a partir de un código ya existente y seleccionando un IV de 0.15, ya que encontraremos pocas variables con fuerte poder de predicción, por tanto debemos reducir el umbral.

Gracias al IV obtenemos la lista de variables que aportan mayor influencia sobre la variable dependiente. En añadido, realizamos otra selección de variables para comprobar su eficacia. En este caso aplicaremos el **test de hipótesis estadística Chi-Cuadrado de Pearson**, un ejemplo de prueba de independencia entre variables categóricas. Los resultados de esta prueba pueden utilizarse para la selección de campos donde todos aquellos que sean independientes de la variable objetivo pueden eliminarse del conjunto de datos. El estadístico Chi-Cuadrado se calcula mediante la siguiente fórmula, en la que "O" representa el valor observado o real y "E" el valor esperado si estas dos categorías fuesen independientes. Si son independientes, estos valores O y E serán similares, y si tienen alguna asociación, el valor de Chi-cuadrado será alto.

Ilustración 26. Fórmula del test Chi-Cuadrado
Fuente: towardsdatascience.com

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \cdots + \frac{(O_{mn} - E_{mn})^2}{E_{mn}}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Continuando con el modelo, puesto que en ambos procesos de selección obtenemos resultados prácticamente iguales, procedemos a entrenarlo modelo con los campos seleccionados. Dividiremos los conjuntos de *train* y *test* en la misma proporción que antes y, en este caso, lo que cambia es que traduciremos las predicciones a una probabilidad, esto es, obtenemos la probabilidad de que cada tipo de local a predecir esté en una clase u otra, en otras palabras, su probabilidad de éxito.

Así, se debe definir un valor de umbral de clasificación (***threshold***) si se quiere transferir un valor de regresión logística a una categoría binaria. Este **corte óptimo** para el algoritmo conseguirá que todas aquellas predicciones de probabilidad de éxito que se encuentren por

encima de ese valor, se considerarán éxito, mientras que el resto no. Lo iremos ajustando según lo que resulte de mayor interés para la calidad de nuestro modelo, ya que probablemente predecir correctamente los verdaderos negativos y disminuir los falsos positivos sea lo que importe más a un inversor. Con esto queremos decir que consideramos que calificar a los establecimientos sin éxito con un 0 y no equivocarnos es de mayor utilidad, ya que dar pronosticar éxitos a un negocio que luego no los alcance resultaría en un gran perjuicio económico para el inversor. Por ello seleccionamos un *threshold* de 0.64, de forma que los negocios con una predicción de éxito menor al 64% de probabilidad no se consideren éxito. Esto se ha escogido así para obtener el menor ratio de falsos positivos sin perder calidad.

Tras ello, pasamos a graficar las mismas métricas que comentamos en el modelo previo y que posteriormente valoraremos en el apartado de resultados.

4.2.9 Conexión del modelo con el usuario

Una vez contamos con un modelo de clasificación con buenos resultados, tal y como se puede comprobar en el apartado de resultados, procedemos a crear una interfaz en la que el cliente pueda interactuar con el modelo y obtener un resultado. Esta la realizaremos con Dash, un *framework* de código abierto para construir interfaces de visualización de datos que, a diferencia de otras plataformas tradicionales, nos permite crear aplicaciones web. En las aplicaciones Dash se puede desplegar aplicaciones en máquinas virtuales o clústeres y luego compartirlas a través de URLs. Dado que se ven en el navegador web, Dash es intrínsecamente multiplataforma y está adaptado a dispositivos móviles.

En primer lugar, nos descargaremos el modelo gracias a la librería *Pickle*, que lo almacena en un archivo *pickle* que posteriormente será abierto en PyCharm, el IDE que utilizaremos para codificar el *dashboard.*

Una vez lo tenemos descargado, lo abrimos en un archivo Python dentro del IDE y procedemos a crear entradas para que cada uno de los campos sea rellenado por el usuario. En la ilustración 27 se puede ver cómo se va añadiendo cada input una vez creada la distribución, con un formato muy semejante a una página web al uso. Esto se realizará tantas veces como sean necesarias y seguidamente definiremos una función que ejecutará el modelo con los valores que han sido introducidos por el cliente. No obstante, para más detalle, se puede consultar el código en el repositorio enlazado en el anexo.

Ilustración 27. Código Interfaz de Usuario

```
app = dash.Dash(__name__)
app.title = 'Machine Learning Model Deployment'
server = app.server
### cargar modelo ML ######################################
with open('modelo.pickle', 'rb') as f:
    clf = pickle.load(f)
### App Layout ############################################
app.layout = html.Div([
    dbc.Row([html.H3(children='Dónde tendrá éxito tu negocio')]),
    dbc.Row([
        dbc.Col(html.Label(children='Escoge una localización concreta'), width={"order": "first"}),
        dbc.Col([dcc.Dropdown(
                id='distrito', #asi se va a llamar lo escogido
                placeholder="Selecciona un distrito",
                options=[
                    {'label': 'Centro', 'value': 'desc_distrito_local_centro'},
                    {'label': 'Chamberí', 'value': 'desc_distrito_local_chamberi'},
                    {'label': 'Puente de Vallecas', 'value': 'desc_distrito_local_puente_de_vallecas'},
```

De esta forma, contamos con una pequeña página web desplegada en un puerto determinado, de momento únicamente disponible en local. Podremos consultar el resultado en el apartado respectivo.

4.3 Recursos requeridos

Los recursos necesarios para llevar a cabo este proyecto se componen de equipos hardware y software:

Recursos materiales y técnicos

- Portátil: más de 80 GB de almacenamiento libre y un mínimo de 8 GB de RAM.
- Monitor: más de 22”
- Licencia de Office365: Word, Power BI, Power Point, Excel y One Drive.
- Software open-source:
    - Anaconda: Jupyter Notebooks
    - Programas mencionados previamente: PyCharm, ParseHub, Git, Plotly
    - Lenguaje de programación: Python

Recursos Humanos

- Estudiante de Ingeniería matemática y/o análisis de datos
- Tutor de guía en el proyecto

4.4 Presupuesto

Tabla 4. Tabla de Presupuesto

Tipo de coste	Valor	Comentarios
Recursos Humanos		
Horas de trabajo en el proyecto	Estudiante: 13€/h Tutor: 60€/h	Duración estimada de 370h a lo largo de cinco meses. Añadido de 15 horas invertidas por el tutor.
Recursos para el desarrollo		
Ordenador portátil Dell LATITUDE E7450	600€	Valor aproximado en el mercado actualmente
Monitor AOC Asus – 23.9”	240€	Valor en el mercado actual
Software para el despliegue		

Licencia Office 365	69€/año	Valor en el mercado actual
Software Open-Source	0€	No requiere pagar precio por licencia

4.5 Viabilidad

El prototipo actual del proyecto puede ponerse en producción y ser probado por cualquier usuario, sin embargo todavía no podría monetizarse debido a que necesitaríamos una cantidad mayor de resultados, pruebas y/o alguna institución que nos avale. No obstante, la viabilidad financiera de un proyecto está muy condicionada a los costes en caso de fracaso y el capital necesario para completarlo, lo cual será positivo en el caso que nos atañe, debido que al tratarse de un trabajo realizado a través de software y datos open-source, los costes asociados no son elevados. Por ello, en caso de fracaso, la pérdida no sería notablemente alta. No obstante, se incluye un apartado en el anexo dedicado al *Business Model Canvas* del prototipo de negocio propuesto hasta aquí para una mejor toma de decisiones.

4.6 Resultados del proyecto

En el presente apartado expondremos los diferentes resultados obtenidos en las etapas del proyecto expuestas previamente. En primer lugar, se discutirán los resultados de los modelos *machine learning* utilizados a fin de escoger el definitivo que se dedicará al cliente. Seguidamente, el resultado de la visualización de los establecimientos gracias al código de limpieza automatizado con una comparación de lo obtenido en el modelo. Por último, finalmente se presentará la interfaz de usuario.

4.6.1 Resultados de los modelos de regresión

Recordamos que realizamos un modelo utilizando *scikit-learn* y *statsmodels*. El primero, más utilizado para *big data* y *machine learning*, tendrá mejores funcionalidades, pero utilizamos el segundo como método para analizar los datos. De este último obtenemos los siguientes resultados, de los cuales destacaremos los estadísticos más significativos.

Ilustración 28. Resultados Regresión en Statsmodels

OLS Regression Results

Dep. Variable:	y	R-squared:	0.665
Model:	OLS	Adj. R-squared:	0.664
Method:	Least Squares	F-statistic:	885.9
Date:	Thu, 23 Jun 2022	Prob (F-statistic):	0.00
Time:	01:27:30	Log-Likelihood:	-1.7405e+06
No. Observations:	215566	AIC:	3.482e+06
Df Residuals:	215084	BIC:	3.487e+06
Df Model:	481		
Covariance Type:	nonrobust		

En ellos podemos observar el F-Estadístico con su p-valor asociado (*Prob (F-statistic)*). Este es el estadístico general del modelo de regresión, que nos indica si el propio modelo en su conjunto es estadísticamente significativo. El p-valor nos permite evaluar el contraste de significatividad conjunta, que dicta como hipótesis nula que ninguna de las variables dependientes nos permite explicar la variable dependiente. Puesto que el p-valor es el menor nivel de significación para el cual esta hipótesis es rechazada y este es muy bajo, rechazamos la hipótesis nula a un nivel de confianza muy alto, casi al 100%.

Por otro lado, el R^2 ajustado, conocido como coeficiente de determinación, determina el grado de variación en la variable dependiente que puede explicarse por las variables independientes. Según el modelo, un 66,4% de la variación en la duración del local se explica por las variables introducidas, lo cual consideramos un buen resultado.

En la siguiente imagen podemos ver los coeficientes de cada variable y los p-valores individuales, los cuales indicarán si cada variable predictora es estadísticamente significativa o no. A través de los coeficientes obtenemos el cambio esperado en la variable de respuesta, asumiendo que el resto permanece constante. Por ejemplo, si el local se encuentra en el distrito Chamartín, esperamos que la duración media del establecimiento aumente en 1604 días. A su vez, el p-valor de este campo nos indica que "Chamartín" es estadísticamente significativo a un 90% de confianza estadística, por lo que deducimos que es una de las variables que más impacta sobre la duración de los negocios.

Ilustración 29. Coeficientes de Regresión

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3220	388.809	0.001	0.999	-761.735	762.379
x1[T.barajas]	729.5386	511.060	1.428	0.153	-272.126	1731.203
x1[T.carabanchel]	547.4855	494.474	1.107	0.268	-421.671	1516.642
x1[T.centro]	527.0800	461.233	1.143	0.253	-376.926	1431.086
x1[T.chamartin]	1604.7030	869.590	1.845	0.065	-99.671	3309.077

En el caso de los barrios, Casco Histórico de Vallecas es que tiene un menor p-valor, por tanto una mayor significatividad estadística. Sin embargo, el coeficiente es negativo, por lo que si ubicamos allí el local, la duración disminuye en 2026 días manteniendo el resto de campos constante. Lo mismo sucede en Ensanche de Vallecas y El Pilar. Para el resto de campos no encontramos influencia estadística relevante.

Esto último no ayuda a obtener un resultado claro, además de que el modelo nos avisa de los siguiente:

Ilustración 30. Multicolinealidad en el modelo

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.9e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Si existe multicolinealidad quiere decir que hay dependencia lineal entre variables explicativas, lo que puede ser causado por incorporar variables que no aportan información adicional. Esto afecta en la significatividad y el valor del R^2, por lo que el estimador pierde robustez. En añadido, el error cuadrático medio es demasiado alto:

Ilustración 31. Error cuadrático medio

```
mod.mse_model
```

```
535692991.2220841
```

Por último, realizamos el Contraste de White para detectar heterocedasticidad.

Ilustración 32. Contraste de White

```
white_test = het_white(mod.resid,  mod.model.exog)
white_test
```

```
(13835.511128362816, 0.0, 34.47426425356265, 0.0)
```

Nos fijamos en los dos primeros valores, que se corresponden con el T-Estadístico y su p-valor. Dado que este último es muy pequeño, rechazamos la hipótesis nula que dicta que el modelo es homocedástico. Hay heterocedasticidad y esta provoca que nuestro estimador no sea eficiente, por lo que realizamos una estimación robusta para solucionarlo. En esta, obtenemos diferentes resultados. En la ilustración 33 observamos que perdemos la significatividad conjunta. Asimismo, las individuales también cambian, pero el error se mantiene igual.

Ilustración 33.Estimaciones robustas

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.665
Model:                            OLS   Adj. R-squared:                  0.664
Method:                 Least Squares   F-statistic:                     0.6437
Date:                Fri, 24 Jun 2022   Prob (F-statistic):              0.695
Time:                        19:21:27   Log-Likelihood:            -1.7405e+06
No. Observations:              215566   AIC:                         3.482e+06
Df Residuals:                  215084   BIC:                         3.487e+06
Df Model:                         481
Covariance Type:                  HC0
```

Acto seguido, comenzamos con el siguiente modelo realizando un proceso similar y obtenemos lo siguiente:

Ilustración 34. Precisión de la regresión en scikit-learn

```
Precisión del modelo:
0.6637910960026533
```

Este valor de precisión indica que los resultados son similares a los anteriores. No es de extrañar, ya que utilizamos el mismo modelo y muestra. Además, el error es el mismo. A pesar de ello, en este caso logramos encontrar los valores que lo elevan a ese nivel, que son las que se observan en la ilustración 35. Eliminándolas, el error resultante es un valor más comedido.

Ilustración 35. Errores más altos de la predicción

	pred	real	error
16232	1.198986e+15	0	1.198986e+15
17391	-8.409012e+13	0	8.409012e+13
28255	6.804782e+13	0	6.804782e+13
28361	-2.393349e+14	2891	2.393349e+14
31156	7.875103e+13	1003	7.875103e+13
31519	1.250339e+13	2891	1.250339e+13
32801	5.203670e+14	183	5.203670e+14

```
predicciones['error'].mean()
```

```
521.8053704261416
```

Esto no soluciona el problema, no podemos eliminar una serie de predicciones del modelo, no afectan sobre la capacidad de estimación.

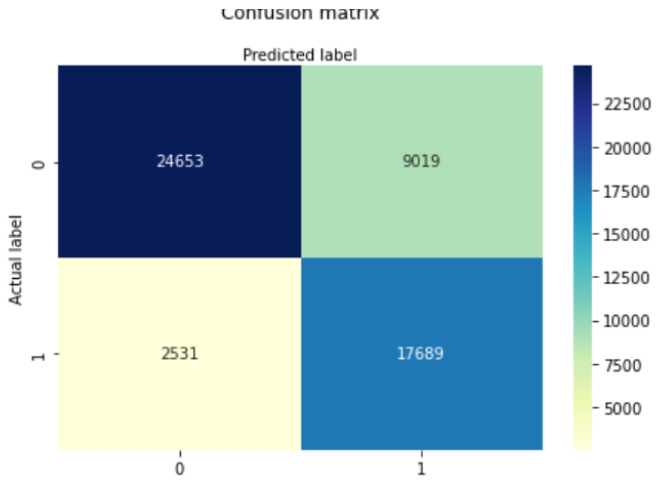
La conclusión en este punto es que debemos hacer una selección de variables que elimine la multicolinealidad y reducir el error.

4.6.2 Resultados del modelo de clasificación

Obtener una respuesta útil en una regresión es complicado cuando se trata de datos no muy exactos. El error en días cometido es notablemente grande teniendo en cuenta la cantidad de variables introducidas. Si aún con todas ellas, no encuentra un patrón claro, es necesario probar con otros algoritmos.

En la regresión logística conseguimos la matriz de confusión y métricas que se observan a continuación.

Ilustración 36. Matriz de Confusión



En ella se observa el número de éxitos (1) y no éxitos (0) predichos frente al valor real.

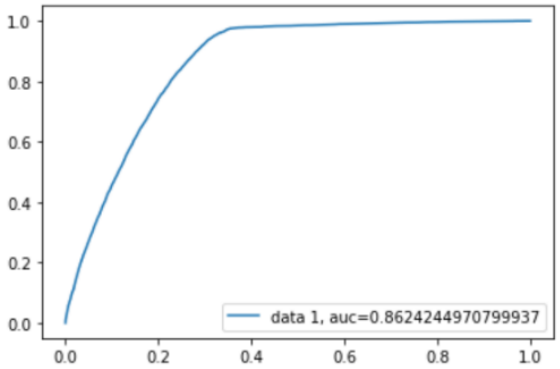
Ilustración 37. Accuracy, Precision y Recall

```
Accuracy: 0.7856824760632376
Precision: 0.6623109180769807
Recall: 0.8748269040553907
```

El *accuracy* nos muestra que hay un 78.56% de aciertos (verdaderos positivos y verdaderos negativos) sobre el total de datos de entrada. La precisión, que hay un 66.23% de verdaderos éxitos sobre la cantidad total de todo lo que se ha predicho como éxito, por lo que se enfoca en la importancia de los falsos positivos. En cambio, el *recall* mide la cantidad de locales clasificados como verdaderos éxitos sobre todo lo que realmente era éxito, por tanto será importante mantenerlo alto cuando no debamos dejar escapar los falsos negativos. En este caso, vemos que se trata del mejor resultado, un 87.48%, lo que significa que no hay muchos locales con éxito que hayan sido categorizados sin éxito.

A través de la siguiente ilustración podemos afirmar que el clasificador separa con muy buen resultado las dos diferentes clases, ya que se nos queda un área bajo la curva muy cercano a 1, el AUC es 0.86.

Ilustración 38. Curva ROC-AUC



Podemos concluir en que la clasificación funciona mejor con nuestros datos, por lo que pasamos a realizar una selección de variables a fin de conseguir reducir el modelo y alcanzar estas mismas métricas.

Para ello, elaboramos un segundo modelo precedido por la aplicación de la función *Information Value*, de la cual obtenemos el conjunto de variables que se muestra en la tabla.

Tabla 5. IV: Selección de variables

Distrito	Centro, Chamberí, Puente de Vallecas, Salamanca
Barrio	Justicia, Recoletos
Sección CNAE	Financiera, Sanitaria, Comercio al por menor y al por mayor, Educación,

	Hostelería, Otros Servicios
División CNAE	Sanitaria, Comercio al por menor, Educación, Servicios Personales, Comidas y Bebidas, Financiera, Venta y reparación de vehículos
Epígrafe CNAE	Ferretería, Prendas de vestir, Farmacéuticos, Bebidas, Reparación de vehículos, Artículos nuevos, Productos alimenticios, Peluquería y belleza, Restaurantes, Intermediación monetaria

No solo se devolvieron estos campos, sino que la categoría que creamos para rellenar los datos nulos también ha resultado significativa frente al resto, con una de las valoraciones más altas.

El modelo con estas variables nos da los siguientes resultados:

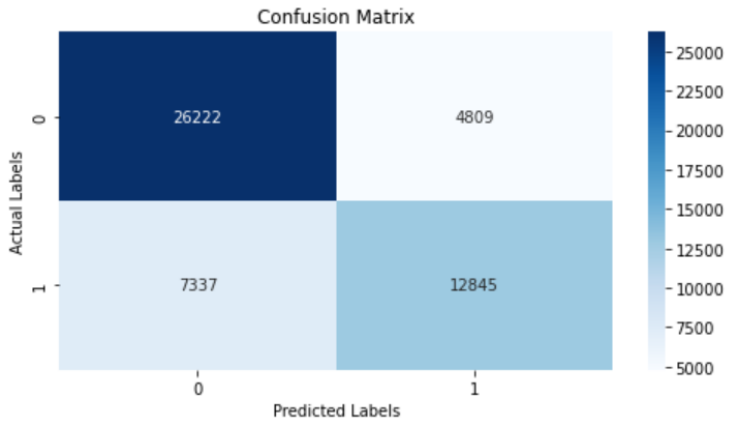
Ilustración 39. Accuracy, precision y recall del segundo modelo

```
Accuracy: 0.7628336555171539
Precision: 0.7275971451229183
Recall: 0.6364582301060351
```

Puesto que fijamos un corte óptimo de 0.64 bajo el cual no se considerará éxito a ningún negocio, aumentan los falsos negativos y el *recall* disminuye considerablemente. No obstante, este umbral se estableció para disminuir los falsos positivos, por lo que el dato de *precisión* ahora es notablemente mejor. En añadido, esto lo conseguimos sin disminuir de forma preocupante el *accuracy*, por lo que nos acercamos al resultado deseado.

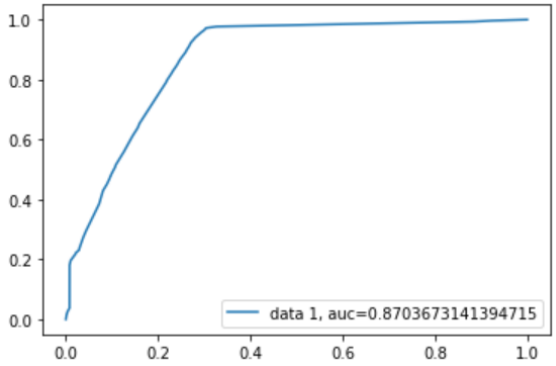
En la matriz de confusión se puede observar como la mayoría de negativos se han clasificado correctamente y hay pocos sin éxito que se hayan categorizado como un éxito, lo cual hemos priorizado a la hora de realizar el modelo, tal y como comentamos en el apartado de creación.

Ilustración 40. Matriz de confusión del segundo modelo



En añadido, el AUC score se mantiene similar al anterior, por lo que concluimos que este modelo ha optimizado el funcionamiento del algoritmo para obtener los resultados que buscábamos.

Ilustración 41. Curva ROC-AUC del segundo modelo



Este último será el que implementemos en la oferta al usuario.

4.6.3 Foto Actual de la ciudad de Madrid

En primer lugar, procedemos a mostrar cuál es el resultado obtenido tras la limpieza y procesamiento de los datos comerciales actuales. Este se muestra en una serie de *dashboards* que nos permiten ver de forma más visual cuáles son las tendencias comerciales actuales en la ciudad de Madrid.

En la ilustración 42 se pueden ver las localizaciones de todos los locales coloreados según su situación. En él se incluyen los filtros de sección y situación de forma que se pueda interactuar con el mapa según lo deseado. Además, la fecha se muestra a fin de ser conscientes del momento en el tiempo al que pertenece el análisis.

Ilustración 42. Dashboard Localización de Establecimientos

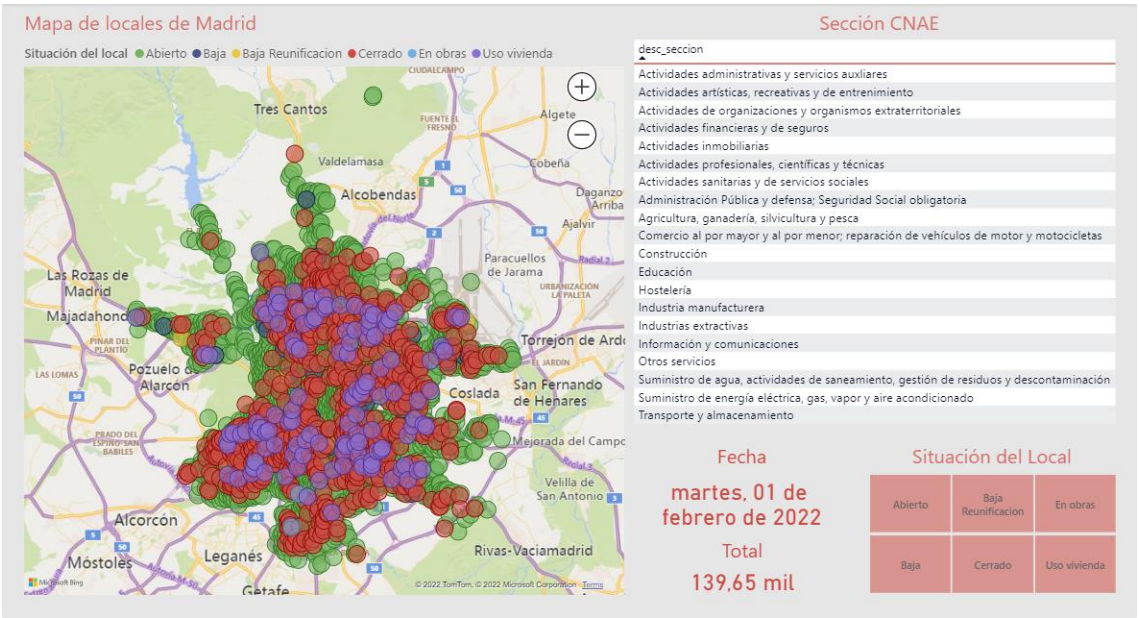
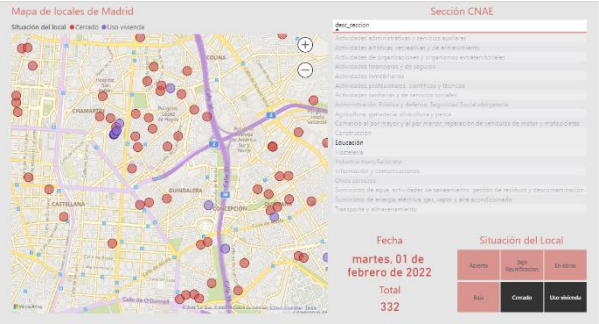


Ilustración 43. Ejemplo de filtro

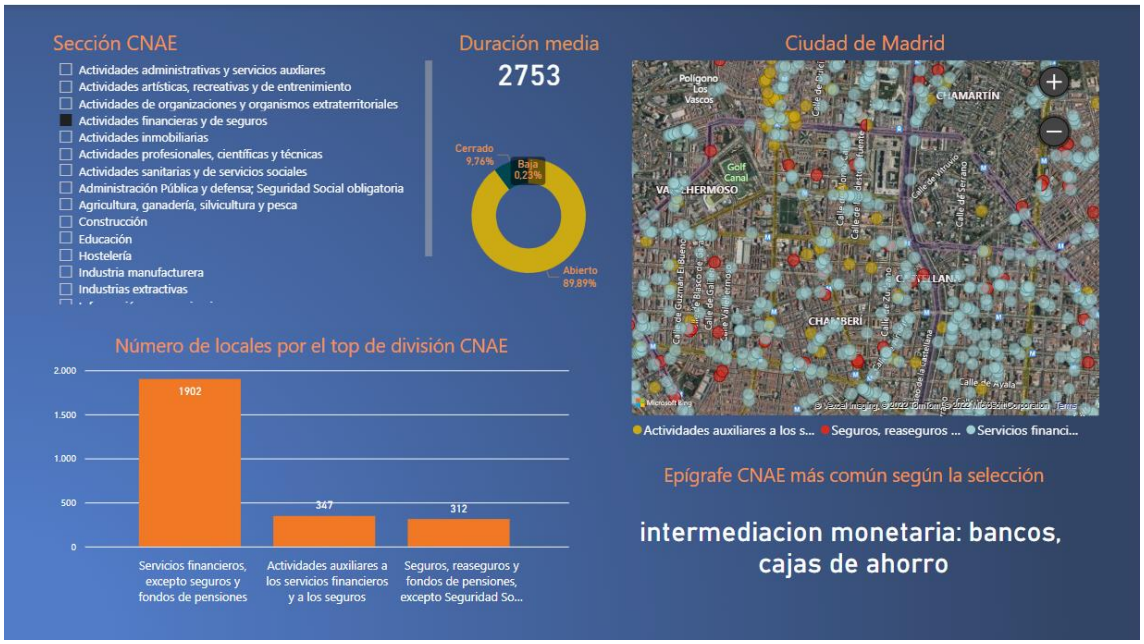


A través de él podemos ver que contamos aproximadamente con 90 mil locales abiertos en la ciudad frente a unos 36 mil cerrados. Los establecimientos en obras, de baja y utilizados como vivienda alcanzan la cantidad de 11 mil.

En la siguiente ilustración podemos observar un análisis más exhaustivo de las actividades comerciales más comunes dentro de la ciudad. Según la sección, se obtiene el ránking de sus divisiones CNAE con mayor número de locales para posteriormente obtener concretamente el epígrafe que más se repite y ubicar los establecimientos dentro de un mapa.

Analizando una de las secciones introducidas en el modelo podemos ver que los locales dedicados a actividades financieras tienen una longevidad muy elevada teniendo en cuenta que la frecuencia no es muy alta. Asimismo, hay muy pocos que estén cerrados y el epígrafe más común también ha sido insertado en el modelo

Ilustración 44. Ranking de Actividad Comercial por Sección CNAE



Por último, en la ilustración 45 contamos con un análisis de las diferentes zonas de Madrid. En él, se observan los 12 distritos madrileños cuya media de duración es más alta y cuyo tamaño depende del número de locales, gráfico que no se ve afectado por el filtro de distritos, y una distribución por situación, lo cual tampoco varía con el filtro de situación. En el resto de

objetos se pueden observar el número de locales en los barrios más longevos, junto con el tipo de negocio más duradero, el número total de locales en el distrito seleccionado y la duración media según los filtros. En la imagen podemos ver el caso de Moratalaz, pero este no ha sido escogido al azar. Gracias a la visualización observamos que el establecimiento más duradero es el de explotaciones agrícolas. Encontramos además que los entre los distritos con mayor promedio de duración están los que seleccionamos para el modelo. Fuencarral-El Pardo desaparece de la segunda posición como distrito duradero y con muchos locales cuando seleccionamos los comercios abiertos, por lo que tiene sentido que no se seleccionase en el modelo. A su vez, dos de los barrios del podio también se incluyeron en el modelo.

Ilustración 45. Análisis de localizaciones por duración

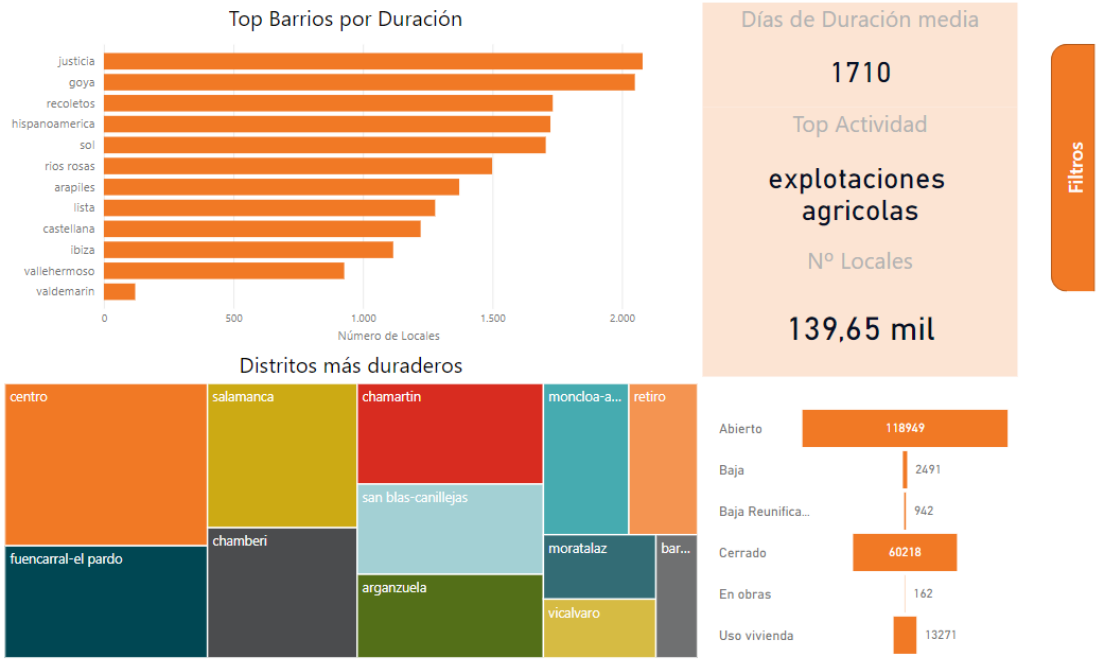
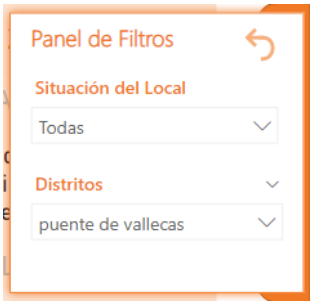


Ilustración 46. Puente de Vallecas

Si seleccionamos el distrito Puente de Vallecas, que también fue escogido por la selección de variables, vemos que cuenta con el promedio más bajo de todos los distritos y con un ratio de establecimientos no abiertos mucho mayor al resto.

Ilustración 47. Panel de Filtros deslizable

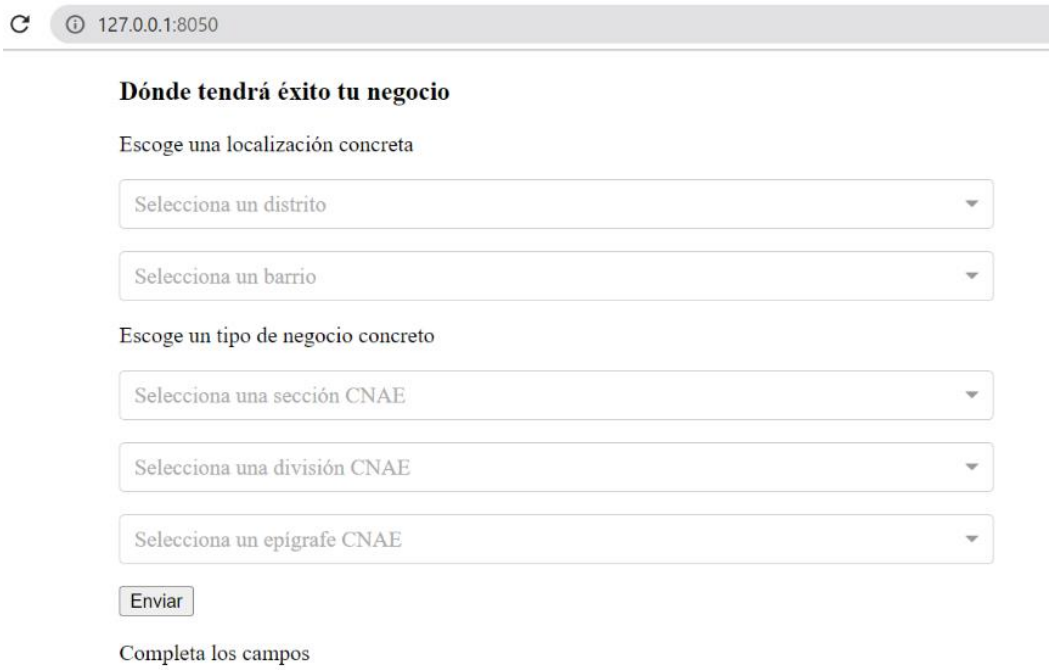


4.6.4 Conexión del modelo con el usuario

Finalmente, implementamos el modelo en una aplicación a través de la librería *Dash* que se ejecuta en el puerto 8050.

En ella hemos ubicado 5 selectores diferentes, uno para cada categoría del modelo. En ellas un cliente selecciona las diferentes opciones que desea consultar y al presionar el botón situado en el extremo inferior izquierdo de la página, se muestra una probabilidad de éxito para la combinación introducida.

Ilustración 48. Interfaz de Usuario



En la ilustración 48 se observa cómo se ve la página de cara al usuario, mientras que en las ilustraciones 49 y 50 podemos ver un ejemplo de selección con la probabilidad de éxito que tendría el hipotético negocio.

Ilustración 49. Negocio con éxito

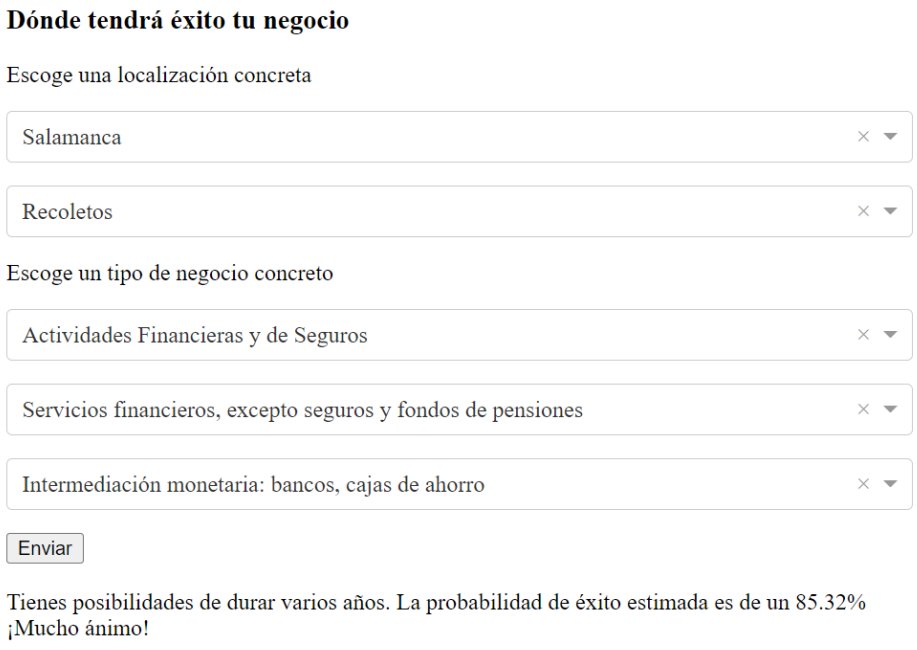
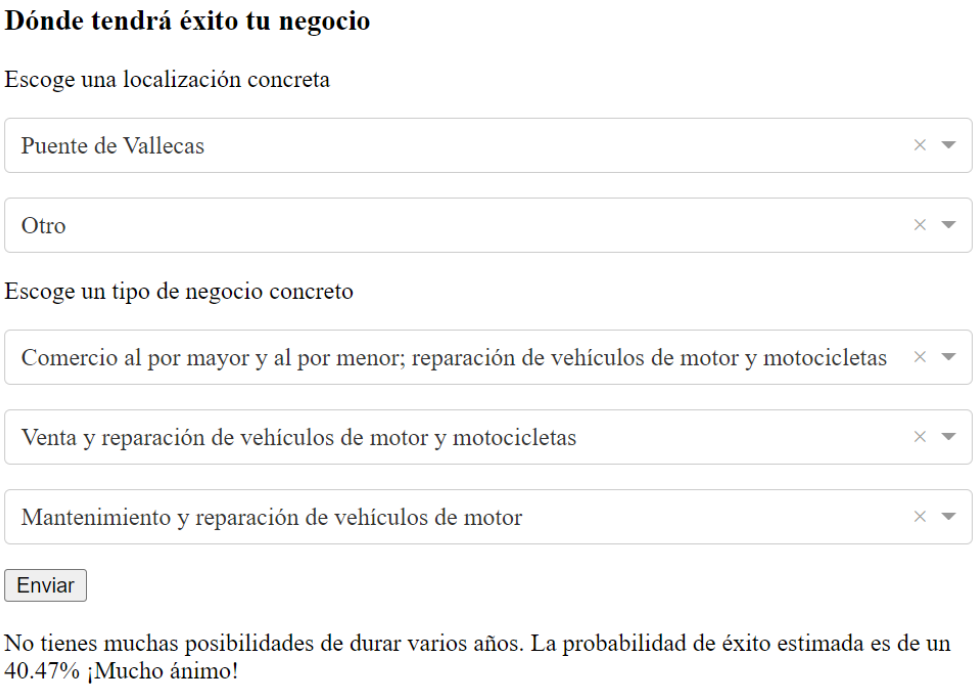


Ilustración 50. Negocio sin éxito



Capítulo 5. DISCUSIÓN

El presente proyecto se propuso como ayuda a futuros emprendedores para conocer la actividad comercial en la ciudad de Madrid, de forma que sea de utilidad en la toma de decisiones a la hora de iniciar un negocio. A lo largo del trabajo se ha definido el éxito de un establecimiento como una variable dependiente de su localización y actividad económica a fin de obtener un patrón, pero esta ha resultado ser algo más allá que una tarea trivial.

A partir del análisis y procesamiento de los datos del censo de actividades comerciales, publicado por el Ayuntamiento de Madrid, se han elaborado una serie de modelos de aprendizaje supervisado que infieren sobre la duración de los negocios. El modelo de clasificación ha resultado ser el que mejor trabaja con los datos, por lo que se ha implementado en una interfaz de usuario, a modo de página web, en la cual, según diferentes combinaciones de negocio, se devuelve una estimación de probabilidad de éxito basada en la posible duración en días del local.

Con objeto de introducir y estudiar estos datos, se ha tenido que realizar un extenso proceso de ETL. El mayor obstáculo dentro del proyecto ha sido homogeneizar toda la lista de campos, ya que se trata en su mayoría de texto, todos ellos relativos a registros recogidos en 8 años diferentes. Conlleva la concatenación de datos en diferente orden, formatos y bajo diferentes reglas, que han provocado un proyecto en constante cambio para corregir errores que iban apareciendo en el transcurso del trabajo.

En añadido, el modelo se ha basado en campos de datos categóricos, en virtud de lo cual encontramos una limitación en el estudio al no contar con variables continuas que influyan sobre el objetivo sin necesidad de crear una gran cantidad de variables ficticias.

Sin embargo, a pesar de las barreras acometidas, obtenemos la serie de campos que más impactan sobre la duración de un negocio y una normalización de ficheros capaz de actuar sobre cualquier archivo proveniente del censo. Asimismo, se ha generado un cuadro de mando con un análisis en función a la longevidad del establecimiento, calculada a través del procesamiento realizado en el proyecto, que avala los resultados obtenidos en el modelo de *machine learning.*

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

Como bien indica la frase que titula el proyecto, se ha llevado a cabo un modelado de datos abiertos que pretende cubrir una necesidad. Este modelo, valga la redundancia, está enteramente basado en datos, de modo que son el corazón y el motor de todo lo realizado. Dependemos de ellos para obtener uno u otro resultado, y cada vez que no se trata un registro debidamente, estamos perdiendo o modificando información. Nos encontramos ante la máxima representación de la importancia del proceso de extracción de conocimiento del dato.

El trabajo pretende servir de consulta para que todos aquellos interesados en emprender tengan una idea sobre las probabilidades de éxito de su negocio. Esto se consigue a través de la elaboración de un modelo de *machine learning* basado en registros históricos. Sin embargo, el proyecto ha consistido en su mayoría en el tratamiento del dato. Es por ello que se recalca su importancia al inicio del capítulo. Habiendo introducido los registros tal cual se obtuvieron al comienzo del trabajo, no se hubiera encontrado ningún patrón ni resultado a partir de ellos.

Aun así, todavía persiste una serie de datos faltantes que no se pudieron rellenar ni eliminar, ya que esto último supondría dejar de tener en cuenta el propio comportamiento de los datos nulos. Tal y como se ha mencionado en el apartado correspondiente, estos campos resultan finalmente significativos, en razón de lo cual se afirma que el nivel de ruido en los datos del modelo es alto y posiblemente sea difícil encontrar un patrón entre los registros de entrada y los resultados.

En muchas ocasiones, este es el inconveniente que acarrean los datos abiertos. En los ocho años de recopilación del censo, los formatos y las categorías han cambiado, en gran parte debido a errores de recolección manual. Se ha tratado al dato como si no tuviera el poder de dar información y generar dinero en el futuro. El modelo elaborado en el presente proyecto podría tener un valor incalculable si contase con todos los datos debidamente registrados y se diera con el algoritmo que mejor trabaja con ellos. Sin embargo, este último caso jamás se alcanza y ahí emerge la diferencia entre un buen ingeniero de datos sobre el resto.

Pese a todo ello, la información disponible y su tratamiento y análisis nos han permitido elaborar un cuadro de mando explicativo sobre la actividad comercial en la ciudad junto con un modelo basado en las variables que más influencia ejercen sobre la duración de los negocios. Entre las localizaciones destacan los distritos Centro, Salamanca, Chamberí y Puente de Vallecas, unido a los barrios Justicia y Recoletos. Gracias al *dashboard* se logra ver que todos repercuten de manera positiva, excepto Puente de Vallecas. Asimismo, respecto a las actividades económicas podemos comprobar que aquellas que derivan de las secciones del CNAE correspondientes a educación, comercio al por menor, actividades financieras y de seguros, actividades sanitarias y de servicios sociales, hostelería y otros servicios (que incluye servicios personales o relativos al uso doméstico), son las que más influyen sobre la duración, en estos casos, de forma positiva.

Finalmente, el modelo se ha podido integrar en una interfaz de usuario que permite a cualquier potencial cliente interactuar con el modelo y ajustar los parámetros a su gusto a fin obtener una probabilidad de éxito. Esta es una primera versión, ubicada en el equipo local, de lo que puede convertirse en un proyecto real y monetizable.

Retomando las palabras del principio, la calidad de los datos influye en gran medida en un trabajo, pero lo más importante es tener en cuenta que sin un buen análisis, el *big data* no es más que un *big basurero*.

6.2 Conclusiones personales

Comencé este proyecto motivada por el sentimiento de encontrar una solución real a lo que yo veía como un problema que se reproducía en las calles a diario. Observaba cómo determinados locales situados cerca de la zona donde vivo cambiaban de actividad comercial cada dos años, tres como mucho, sin ningún tipo de lógica. No solo veía un establecimiento en obras, nuevos dependientes, diferentes panfletos de publicidad, escaparates cambiantes… detrás de todo ello se encontraba una serie de inversores que decidieron emprender un negocio, pero no lograron mantener una clientela fiel que sustentase sus gastos. ¿Cómo era posible que tras tantos intentos nadie supiera lo que realmente funciona en el local de la esquina de la calle? De esta forma surgió la idea y mi entusiasmo fue tal, que decidí hacer un trabajo sobre ello.

Sin embargo, choqué contra el primer obstáculo nada más empezar: los datos. El censo no está homogéneamente organizado, no hay información sobre la apertura de los establecimientos y, por ende, tampoco de su duración, entre otras muchas carencias. Ha sido un camino constante de barreras, cambios, volver atrás, empezar desde el principio… Sin duda alguna mi capacidad de resolución de problemas tras este proyecto ha crecido enormemente.

A medida que avanza nuestra carrera profesional, se nos implanta la importancia del trabajo en grupo. Si bien es crucial en la vida laboral, gracias a este proyecto he aprendido a no dar por supuesto el peso del trabajo autónomo. Debido a ello ha resultado ser un trabajo mucho más retador de lo que parecía en un principio. El tratamiento de los datos ha supuesto gran parte del proceso y no pensaba que fuese a dar con la manera de estandarizar absolutamente todos los campos con la intención de conseguir encontrar un patrón a través de ellos.

Por todos los motivos expuestos puedo afirmar lo gratificante que es ver cómo ha cogido forma aquello en lo que he invertido tiempo durante muchos meses. Estoy muy satisfecha con el trabajo realizado y espero algún día retomar lo que empezó como un pequeño pensamiento de negocio.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

El trabajo realizado hasta este punto no conforma un producto palpable y accesible para usuarios externos, por lo que las futuras líneas de trabajo están encaminadas en la producción de este como un producto real.

En primer lugar, investigar la forma de corregir los errores en los datos y minimizar la pérdida de información. Puede ser que una extracción de los datos por medio de APIs y su procesamiento en clústeres o la nube permitan optimizar la calidad de los mismos.

En segundo lugar, utilizar otros algoritmos de *machine learning* capaces de procesar grandes cantidades de datos y que soporten la entrada de gran cantidad de campos, con el objetivo de conseguir un modelo apto para encontrar patrones entre muchas variables categóricas diferentes. Asimismo, sería de gran utilidad añadir otra serie de campos que aporten información sobre los negocios y estén en diferentes formatos a los que ya se incluyen, como pueden ser variables numéricas sobre ingresos, ayudas, empleados, entre otros.

Seguidamente se encuentra una de las futuras funcionalidades que ya mencionamos como añadido al comienzo del proyecto: la optimización de establecimientos cerrados. Esta se incluyó fuera del alcance en caso de que las horas de trabajo lo permitiesen. No se ha llevado a cabo del todo, pero constaría en situar los establecimientos cerrados o en baja, listarlos con sus respectivas localizaciones y definir qué actividades económicas funcionarían mejor en ellos.

En línea con lo anterior está el trabajo con mapas. Visto que la localización es un factor importante en el proyecto, resultaría más visual importar mapas interactivos que permitan interactuar con ellos y ayuden, tanto en la visualización de resultados del modelo, como en la optimización de locales vacíos.

Por último, mejorar la fase de conexión con el usuario. Sería interesante agregar estos mapas interactivos en ella y posibilitar a los futuros clientes una comunicación con el modelo mucho más fluida. Una vez alcanzado esto, se lanzaría de cara al público con el propósito de transformarlo en un producto.

Capítulo 8. REFERENCIAS

[1] ÁLVAREZ ONDINA, P., 2021. El papel del comercio minorista en la economía española. CaixaBank Research [en línea]. Disponible en: https://www.caixabankresearch.com/es/analisis-sectorial/comercio-minorista/papel-del-comercio-minorista-economia-espanola.

[2] DIAZGRANADOS CASTAÑO, R. El pequeño minorista en España: Situación actual y análisis de su competitividad en el mercado. Director: Juan Carlos Revilla Castro. Universidad Complutense de Madrid, Facultad de Comercio y Turismo, 2016/2017. Disponible en: https://eprints.ucm.es/id/eprint/45798/1/El%20Peque%C3%B1o%20Minorista%20en%20Espa%C3%B1a_Situaci%C3%B3n%20Actual%20y%20An%C3%A1lisis%20de%20su%20Competitividad%20en%20el%20Mercado.pdf.

[3] REDONDO, S., 2013. El sector terciario en España. Core.ac.uk [en línea]. Disponible en: https://core.ac.uk/display/235863574?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1.

[4] Relevancia económica del comercio minorista en España - Unión de comerciantes del Principado de Asturias. Unión de comerciantes del Principado de Asturias [en línea], 2019. Disponible en: https://comercioasturias.com/relevancia-economica-del-comercio-minorista-en-espana/.

[5] España. Ley 7/1996, de 15 de enero, de Ordenación del Comercio Minorista. Boletín Oficial del Estado, 17 de febrero de 1996, núm. 15, p. 1243 a 1254.

[6] VEENA, 2022. What is entrepreneurship? Stanford Online [en línea]. Disponible en: https://online.stanford.edu/what-entrepreneurship.

[7] ¿Cómo es el Ecosistema Emprendedor en España? - Gestron. Gestron [en línea], 2021. Disponible en: https://ayudatpymes.com/gestron/ecosistema-emprendedor/.

[8] VENTURA, R., MANUEL, J., ADÁN LÓPEZ MENDOZA y ANTONIO, J., 2017. La tecnología: una herramienta de apoyo para pymes y emprendedores desde el entorno universitario. Ciencia Ergo Sum [en línea], vol. 24, no. 1, pp. 75-82. Disponible en: https://www.redalyc.org/journal/104/10449880008/html/.

[9] SIMPLILEARN, 2012. How Big Data Can Help You Do Wonders In Your Business. Simplilearn.com [en línea]. Disponible en: https://www.simplilearn.com/how-big-data-can-help-do-wonders-in-business-rar398-article.

[10] DATOS.GOB.ES, 2018. How to encourage innovation and entrepreneurship using open data. Datos.gob.es [en línea]. Disponible en: https://datos.gob.es/en/noticia/how-encourage-innovation-and-entrepreneurship-using-open-data#:~:text=Open%20data%20has%20direct%20benefits,material%20for%20their%20business%20development.

[11] The Open Data Handbook. Opendatahandbook.org [en línea], 2022. Disponible en: https://opendatahandbook.org/.

[12] Open Data and Entrepreneurship | data.europa.eu. Europa.eu [en línea], 2018. Disponible en: https://data.europa.eu/en/datastories/open-data-and-entrepreneurship.

[13] The Economic Impact of Open Data Opportunities for value creation in Europe. [en línea], [sin fecha]. DOI 10.2830/63132. Disponible en: https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf.

[14] Analytical Report 10: Open Data and Entrepreneurship Analytical Report n10. [en línea], [sin fecha]. DOI 10.2830/357754. Disponible en: https://data.europa.eu/sites/default/files/analytical_report_10_open_data_and_entrepreneurship.pdf.

[15] What Is Machine Learning? A Definition. Expert.ai [en línea], 2022. Disponible en: https://www.expert.ai/blog/machine-learning-definition/.

[16] IBM CLOUD EDUCATION, 2020. What is Machine Learning? Ibm.com [en línea]. Disponible en: https://www.ibm.com/cloud/learn/machine-learning.

[17] Machine Learning: What it is and why it matters. Sas.com [en línea], 2021. Disponible en: https://www.sas.com/en_us/insights/analytics/machine-learning.html.

[18] OLCESE, A., 2019. La Comunidad de Madrid bate récord de destrucción de empresas hasta abril. Vozpópuli [en línea]. Disponible en: https://www.vozpopuli.com/economia_y_finanzas/comunidad-madrid-record-destruccion-empresas_0_1253275477.html.

[19] GONZÁLEZ, R., 2022. Los españoles no creen que se fomente el emprendimiento en nuestro país. Cinco Días [en línea]. Disponible en: https://cincodias.elpais.com/cincodias/2022/04/21/emprendedores/1650568082_618847.html.

[20] TOMÁS PALACÍN, J., 2021. Emprendimiento en España: ni tantos jóvenes, ni tantas «ganas». innovaspain [en línea]. Disponible en: https://www.innovaspain.com/emprendimiento-en-espana-informe-gem/.

[21] GONZÁLEZ, R., 2019. Comercio minorista: motivos por los que necesita modernizarse. Sage Advice España [en línea]. Disponible en: https://www.sage.com/es-es/blog/por-que-el-comercio-minorista-necesita-modernizarse/.

[22] LA INFORMACIÓN, 2022. Los pequeños comercios advierten que la recuperación pierde fuerza en 2022. La Información [en línea]. Disponible en: https://www.lainformacion.com/economia-negocios-y-finanzas/pequenos-comercios-advierten-recuperacion-pierde-fuerza-2022/2862211/.

[23] PEINADO, F., 2018. EL PAÍS: el periódico global. El País [en línea]. Disponible en: https://elpais.com/ccaa/2018/09/19/madrid/1537354582_144510.html#:~:text=Hay%20unos%2014.000%20locales%20comerciales,su%20propio%20trabajo%20de%20campo.

[24] PYME - ¿Qué es una PYME? | SumUp Facturas. SumUp - Una forma más inteligente de cobrar [en línea], 2022. Disponible en: https://sumup.es/facturas/glosario/pyme/.

[25] ANDÚJAR, J.A., 2018. ¿Autónomo o pyme? Conoce las diferencias - Sage Advice España. Sage Advice España [en línea]. Disponible en: https://www.sage.com/es-es/blog/diferencias-pymes-vs-autonomos/.

[26] REDAUTONOMOS, 2012. Empresario Individual o Autónomo. Redautonomos.es [en línea]. Disponible en: https://redautonomos.es/tipos-de-empresa/autonomo.

[27] Censo de locales, sus actividades y terrazas de hostelería y restauración. Histórico - Portal de datos abiertos del Ayuntamiento de Madrid. Datos.madrid [en línea], 2014. Disponible en: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=23160329ff639410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default.

[28] UAX #15: Unicode Normalization Forms. Unicode.org [en línea], 2021. Disponible en: https://unicode.org/reports/tr15/#Introduction.

[29] BBVA, 2013. BBVA y el Ayuntamiento de Madrid presentan un innovador estudio basado en «Big Data». BBVA NOTICIAS [en línea]. Disponible en: https://www.bbva.com/es/bbva-ayuntamiento-madrid-presentan-innovador-estudio-basado-big-data/.

[30] ¿Qué es la regresión lineal? Mathworks.com [en línea], 2022. Disponible en: https://la.mathworks.com/discovery/linear-regression.html#:~:text=La%20regresi%C3%B3n%20lineal%20es%20una,datos%20experimentales%2C%20financieros%20y%20biol%C3%B3gicos.

[31] AMAT RODRIGO, J., 2015. Introducción a la Regresión Lineal Múltiple. Cienciadedatos.net [en línea]. Disponible en: https://www.cienciadedatos.net/documentos/25_regresion_lineal_multiple#:~:text=La%20regresi%C3%B3n%20lineal%20m%C3%BAltiple%20permite,2%2C%20X3%E2%80%A6.

[32] CURRY, R., 2021. Simplified Logistic Regression: Classification With Categorical Variables in Python. Medium [en línea]. Disponible en: https://medium.com/@curryrowan/simplified-logistic-regression-classification-with-categorical-variables-in-python-1ce50c4b137.

[33] RAI, K., 2020. The math behind Logistic Regression - Analytics Vidhya - Medium. Medium [en línea]. Disponible en: https://medium.com/analytics-vidhya/the-math-behind-logistic-regression-c2f04ca27bca.

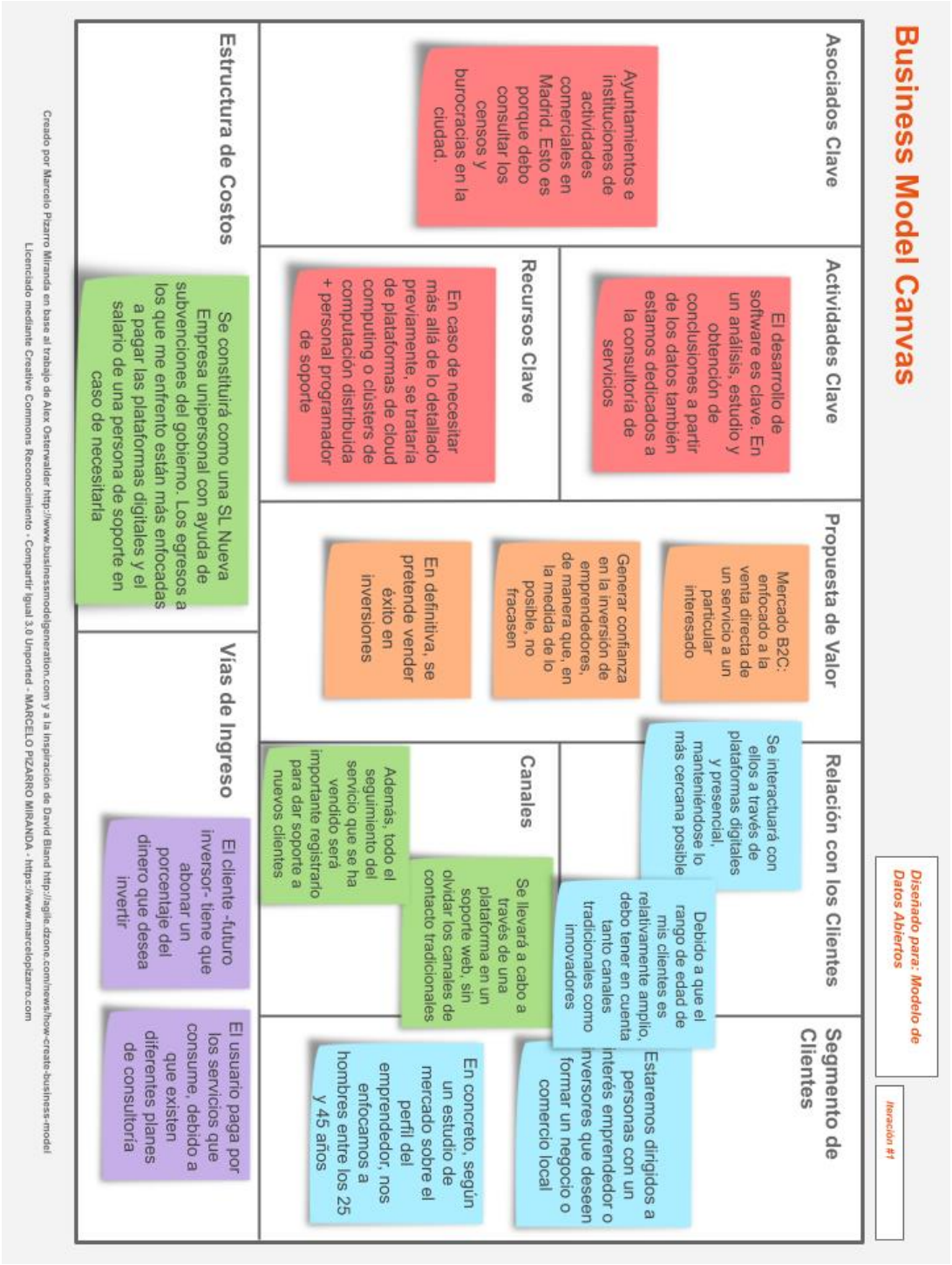
[34] KARABIBER, F., 2022. Dummy Variable Trap – LearnDataSci. Learndatasci.com [en línea]. Disponible en: https://www.learndatasci.com/glossary/dummy-variable-trap/#:~:text=machine%20learning%20courses.-,What%20is%20the%20Dummy%20Variable%20Trap%3F,coefficient%20variables%20in%20regression%20models.

[35] NARKHEDE, S., 2018. Understanding AUC - ROC Curve - Towards Data Science. Medium [en línea]. Disponible en: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=AUC%20%2D%20ROC%20curve%20is%20a,capable%20of%20distinguishing%20between%20classes.

Capítulo 9. ANEXOS

9.1 Business Model Canvas

Business Model Canvas: modelo de negocio



9.2 Código fuente

El código necesario para la realización del proyecto se encuentra disponible en el siguiente repositorio de GitHub: Repositorio TFG - María Fernández Morín

En él se encuentran los *notebooks* con el código realizado, el cuadro de mando en Power BI y el archivo python configurado para la interfaz de usuario.