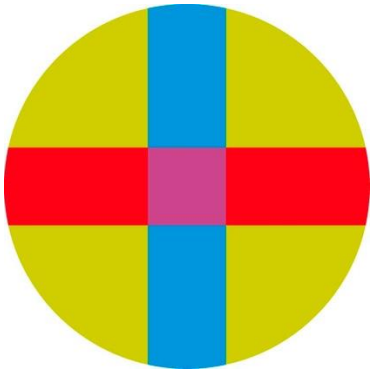
I

UNIVERSIDAD SAN PABLO - CEU

ESCUELA POLITÉCNICA SUPERIOR

GRADO EN INGENIERÍA DE SISTEMAS DE INFORMACIÓN



TRABAJO FIN DE GRADO

Análisis, predicción y representanción de las ventas de 1C Company

Autor: José Morán Nebot

Tutor: Pedro Garrido Gutiérrez

Junio 2023

UNIVERSIDAD SAN PABLO-CEU
ESCUELA POLITÉCNICA SUPERIOR
División de Ingeniería

Calificación del Trabajo Fin de Grado

Datos del alumno

NOMBRE: JOSÉ MORÁN NEBOT

Datos del Trabajo

TÍTULO DEL PROYECTO:

Tribunal calificador

PRESIDENTE:	FDO.:
SECRETARIO:	FDO.:
VOCAL:	FDO.:

Reunido este tribunal el ___ /___ /_____, acuerda otorgar al Trabajo Fin de Grado presentado por D./Dña. _____ la calificación de _____

Resumen

El trabajo realizado se centra en el análisis de una base de datos de una tienda de componentes electrónicos rusa llamada “1C Company” que contiene información sobre su facturación en base al tipo de producto, tipo de tienda y ciudades en las que están presentes en los años 2013,2014 y 2015.

El objetivo principal es representar esta información en la aplicación Power BI y predecir las ventas para un producto determinado en el año 2016 utilizando el algoritmo ARIMA, siendo este desarrollado en Python.

Para lograrlo se realizó un análisis de los datos, identificando patrones estacionales y tendencias en la facturación. Se aplicó el modelo ARIMA a los datos de 2013 y 2014 y se realizaron ajustes y evaluaciones para seleccionar el modelo que se aproximase más a los datos reales de 2015. Tras ello, se hizo el pronóstico para el año 2016.

Posteriormente, se utilizó Power BI para representar la información de manera gráfica e intuitiva, integrando a su vez el pronóstico realizado con Python en esta misma aplicación.

En resumen, se combinan estas dos herramientas para proporcionar una visión de la situación comercial de la empresa, facilitando la toma de decisiones estratégicas.

Palabras Clave

Power BI, aprendizaje automático, Inteligencia Artificial, Inteligencia de negocios, ARIMA, autoregresión, 1C Company.

Abstract

The present work focuses on the analysis of a database from a russian electronic components store called "1C Company," which contains information about its sales based on city, product and store type in the years 2013, 2014, and 2015.

The main objective is to represent this information in the Power BI application and predict the sales for a specific product in 2016 using the ARIMA algorithm, which is developed in Python.

To achieve this, a study and analysis of the data were conducted, identifying seasonal patterns and trends in the sales. The ARIMA model was applied to the data from 2013 and 2014, adjustments and evaluations were made to select the model that best approximated the actual data from 2015. After that, the forecast for the year 2016 was made.

Subsequently, Power BI was used to represent the information visually and intuitively, integrating the forecast made with Python into the same application.

In summary, these two tools are combined to provide an overview of the company's commercial situation, facilitating strategic decision-making.

Keywords

Power BI, Machine Learning, Artificial Intelligence, Business Intelligence, ARIMA, autoregression, 1C Company.

Índice de contenidos

Índice de ilustraciones

Índice de tablas

Capítulo 1

Introducción

1.1 Justificación técnica

Durante el grado, se ha hecho uso de Power BI y se ha podido ver el potencial que tiene esta herramienta para la representación de datos y la importancia que tiene la información que se obtiene para las empresas. Si se combina esta herramienta junto con los conocimientos que se han obtenido en otras asignaturas como estadística o Inteligencia Artificial, se puede programar un algoritmo que sea capaz de realizar predicciones según los datos proporcionados.

Power BI es una aplicación gratuita, al ser estudiante del CEU, en constante evolución, con mejoras continuas y valorada por el cuadrante mágico de Gartner como la mejor herramienta del mercado de Business Intelligence (BI) en el año 2022 [1]. Además, ante un conjunto grande de datos, permite facilitar la toma de decisiones gracias a la visualización de los datos de forma sencilla y eficiente. Por ello, cada vez hay más demanda de profesionales con los conocimientos técnicos necesarios para manejar esta aplicación de BI.

Si combinamos la utilización de esta poderosa herramienta de visualización de datos junto con un modelo de Machine Learning que pueda predecir una variable dada una serie temporal, agrupamos dos áreas que están muy demandadas por las empresas en el mercado actual y que tiene un crecimiento potencial altísimo de aquí a cinco años [2]. En España, se calcula que en estos dos últimos años la demanda de profesionales en Big Data y Machine Learning ha crecido un 92% [3].

1.2 Objetivos

Los objetivos del proyecto son los siguientes:

- Encontrar un dataset con una serie temporal de al menos tres años para poder realizar predicciones.
- Analizar los datos obtenidos de un dataset.
- Tratamiento del dataset.
- Filtrar la información que sea relevante.
- Seleccionar las librerías adecuadas para la arquitectura y diseño del sistema.
- Estudiar el algoritmo de Autoregressive Integrated Moving Average (ARIMA) y utilizar los parámetros que mejores resultados ofrezcan.
- Proporcionar datos estadísticos que respalden la utilización del modelo.
- Predecir las ventas de un producto para el año 2015 y hacer el cálculo para el año 2016.
- Representar la información en Power BI de forma clara mediante cuadros de mando para mejorar y facilitar la toma de decisiones.

1.3 Estructura

En el capítulo 1 se ha realizado una breve introducción al proyecto y se han establecido los objetivos que se persiguen en el mismo y que ya se recogieron en el anteproyecto.

En el siguiente capítulo, se pretende exponer cómo se ha organizado y planificado el proyecto de forma inicial y los cambios que han surgido a lo largo de la realización de este.

En el capítulo 3 se analizan los sistemas existentes en el mercado que tienen relación con el trabajo, así como los algoritmos que mejor pueden encajar con los objetivos que se han establecido. Además, se estudian las diferentes tecnologías

disponibles para realizar este tipo de estudio y se elige la que mejores resultados proporciona en base a unas métricas. Por otra parte, se establecen los requisitos que se deberán alcanzar en el proyecto.

A continuación, en el diseño e implementación, se detalla el diseño físico de los datos, el ajuste del modelo ARIMA y la evaluación de los resultados obtenidos. Asimismo, se muestran las pestañas del informe de Power BI realizado junto con una explicación de los gráficos utilizados y las librerías e IDE empleado.

A lo largo del apartado de construcción se especifica el repositorio en el que se encuentra el script de Python y enumera los pasos a seguir para desplegar el informe de Power BI y el código en la máquina deseada.

Por último, se recogen las conclusiones y posibles líneas de futuro extraídas durante la realización del proyecto.

Capítulo 2

Gestión del proyecto

2.1 Modelo de ciclo de vida

Tras examinar las diferentes metodologías para la construcción y organización de proyectos software se ha decidido utilizar un modelo en cascada. Como se estudió en la asignatura de Ingeniería del Software, este permite realizar un desarrollo secuencial con diferentes fases de forma lineal. Requiere una planificación precisa y estructurada, tarea que se ha realizado con el tutor desde el comienzo del proyecto. Además, el trabajo de fin de grado valora tanto el desarrollo de software como la documentación y este modelo se basa en la documentación del proyecto de forma detallada en cada una de sus secciones.

2.2 Papeles desempeñados en el proyecto

El tutor actúa como representante de la empresa que cuenta con una gran cantidad de datos que precisan ser analizados, procesados y representados de forma gráfica. El alumno actuará como científico de datos realizando esas tareas y proporcionará una predicción de las futuras ventas de un producto específico que oferta la empresa para que se pueda preparar el stock correspondiente.

2.3 Planificación

Para la organización del proyecto, se ha hecho uso del programa para gestión de proyectos de código abierto y gratuito llamado Project Libre. Dicha planificación se elaboró a fecha 28 de febrero de 2023 tras la aprobación del anteproyecto entregado el día 27 de febrero de 2023.

En primer lugar, se procedió a la búsqueda de un conjunto de datos que cumpliese con los requisitos que se habían impuesto desde un primer momento, es decir, que fuese una serie temporal de algún comercio con datos históricos de al menos tres años sobre sus ventas. Una vez encontrado, se analizaron los datos de manera preliminar para cerciorarse de que encajaban con nuestros planes de acción.

A continuación, se pasó a representar los datos en la aplicación Power BI en diferentes cuadros de mando para proporcionar unos informes más visuales sobre la empresa en cuestión. Después, hubo que realizar algunos cambios tras exponer los gráficos al tutor y recibir feedback por su parte para realizar ciertas modificaciones y mostrar información relevante para el negocio.

Terminada la parte de la representación, se escogió un producto de todo el catálogo disponible para realizar las predicciones de ventas futuras. El producto seleccionado debía ser de los más vendidos y contar con ventas durante todas las semanas de dos años seguidos. Cumplidos los requisitos propuestos, se investigaron algoritmos para la predicción de series temporales, concluyendo junto con la ayuda del tutor que el modelo ARIMA sería el más adecuado.

Seguidamente, se practicó con algunos ejemplos ya resueltos del algoritmo ARIMA en Python para familiarizarse con las diferentes librerías de las que hace uso y así poder utilizar los conocimientos adquiridos en la predicción de los datos escogidos. Gracias a ello, se realizó el estudio de parámetros como la estacionariedad, estacionalidad o correlación de los datos para ajustar los valores de entrenamiento y encontrar aquellos que obtuviesen una predicción más acertada.

Finalmente, era importante buscar una solución que integrase el código realizado en Python con los gráficos y medidas representadas junto con los informes ya realizados anteriormente en Power BI.

Todas estas tareas se compaginaron con la redacción de la memoria, la cual se fue completando siguiendo el orden establecido de gestión, análisis y diseño para realizar la primera entrega el día 10 de abril de 2023.

Esta planificación, como se ha indicado al comienzo de este punto, ha sido elaborada con anterioridad y ha habido modificaciones de tareas y plazos que han sido diferentes a lo que se esperaba. Por ello, en el siguiente apartado se indicará cómo ha transcurrido realmente el proyecto, pero la planificación inicial se puede ver a continuación:

	Nombre	Inicio	Terminado	Duracion	Predecesores
1	**Trabajo final de grado**	**27/02/23 8:00**	**18/04/23 17:00**	**34 days**	
2	Anteproyecto	27/02/23 8:00	27/02/23 17:00	1 day	
3	Búsqueda del dataset	28/02/23 8:00	3/03/23 17:00	4 days	2
4	Análisis del dataset	6/03/23 8:00	7/03/23 17:00	2 days	3
5	Representación en power BI	8/03/23 8:00	14/03/23 17:00	5 days	4
6	Cambios en representación	15/03/23 8:00	16/03/23 17:00	2 days	5
7	Selección de productos	17/03/23 8:00	17/03/23 17:00	1 day	6
8	Búsqueda de algoritmos	20/03/23 8:00	21/03/23 17:00	2 days	7
9	Prueba modelo ARIMA	22/03/23 8:00	23/03/23 17:00	2 days	8
10	Ajuste ARIMA	24/03/23 8:00	30/03/23 17:00	5 days	9
11	Python embebido power BI	31/03/23 8:00	4/04/23 17:00	3 days	10
12	**Redacción de la memoria**	**27/02/23 8:00**	**10/04/23 17:00**	**28 days**	
13	**1. Introducción**	**27/02/23 8:00**	**27/02/23 15:00**	**0,75 days**	
14	Justificación técnica	27/02/23 8:00	27/02/23 13:00	0,5 days	
15	Objetivos	27/02/23 13:00	27/02/23 15:00	0,25 days	14
16	**2. Gestión del proyecto**	**28/02/23 8:00**	**1/03/23 10:00**	**1,25 days**	**13**
17	Modelo de ciclo de vida	28/02/23 8:00	28/02/23 10:00	0,25 days	
18	Planificación	28/02/23 10:00	28/02/23 15:00	0,5 days	17
19	Diagrama de gantt	28/02/23 15:00	1/03/23 10:00	0,5 days	18
20	**3. Análisis**	**1/03/23 10:00**	**10/04/23 17:00**	**25,75 days**	**16**
21	Estado del arte	1/03/23 10:00	3/03/23 10:00	2 days	
22	Análisis de los datos	8/03/23 8:00	9/03/23 17:00	2 days	4;21
23	Tecnologías disponibles	10/03/23 8:00	13/03/23 13:00	1,5 days	22
24	Diagrama de entidad relación	13/03/23 13:00	13/03/23 17:00	0,5 days	23
25	Algoritmos de machine learni...	22/03/23 8:00	22/03/23 17:00	1 day	8;24
26	Representación power BI	23/03/23 8:00	24/03/23 17:00	2 days	6;25
27	**4. Diseño e implementación**	**31/03/23 8:00**	**10/04/23 17:00**	**4 days**	
28	Evaluación de resultados	31/03/23 8:00	3/04/23 17:00	2 days	10
29	Python embebido con powe...	10/04/23 8:00	10/04/23 17:00	1 day	11;28
30	5. Construcción	11/04/23 8:00	12/04/23 17:00	2 days	27
31	6. Conlusiones y lineas futuras	13/04/23 8:00	18/04/23 17:00	4 days	30
32	Documentación de la gestión de...	10/04/23 7:00	10/04/23 8:00	0 days	
33	Entrega de TFG	9/06/23 8:00	9/06/23 8:00	0 days	

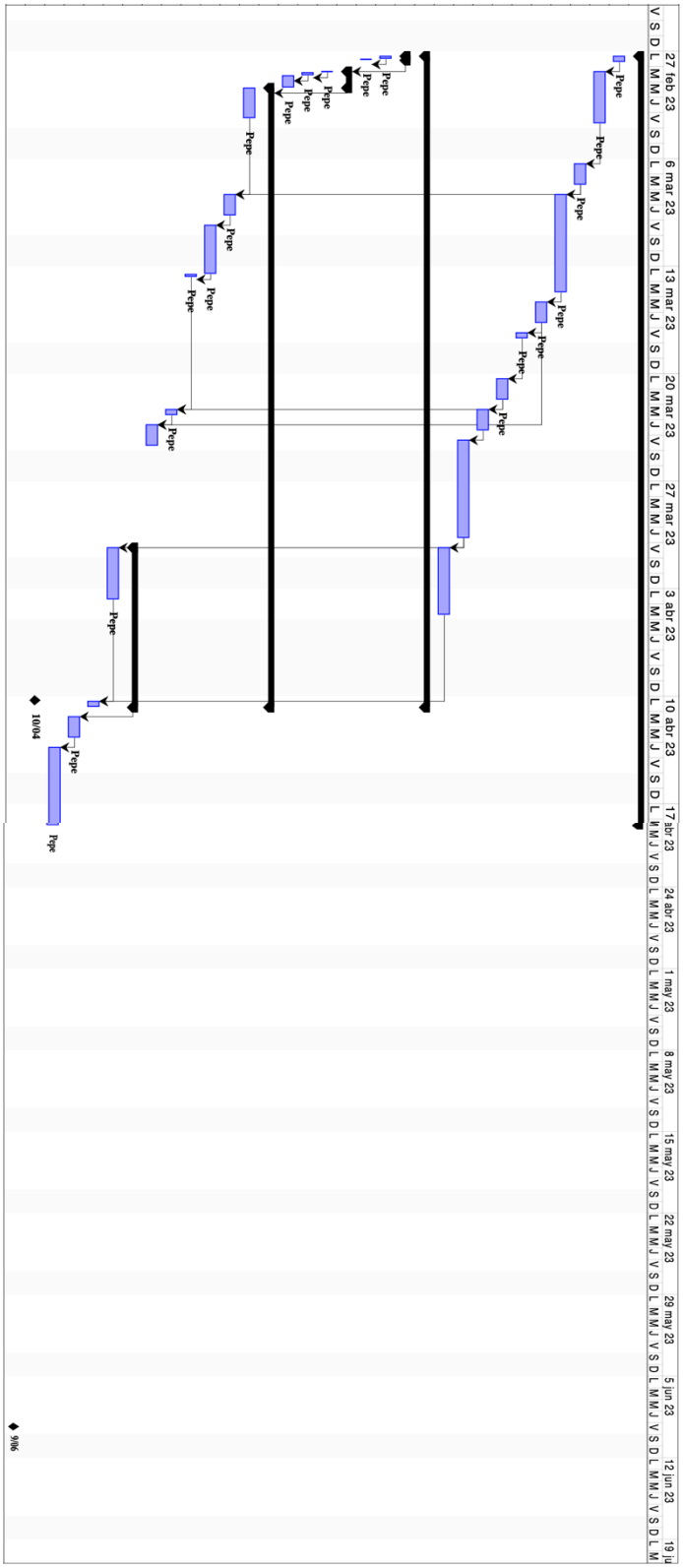


Ilustración 1: Planificación inicial.

2.4 Ejecución

Una vez finalizado el proyecto, al observar la planificación que se realizó al comienzo se puede concluir que hubo algunas tareas que se subestimaron y que llevaron más tiempo del que se pensaba. Esto es algo que puede ocurrir, sobre todo en un proyecto de larga duración y contando con poca experiencia previa, pero es algo que se previó desde el primer momento al dejar un margen de tiempo para posibles imprevistos que pudieran surgir durante los cuatro meses en los que se ha trabajado.

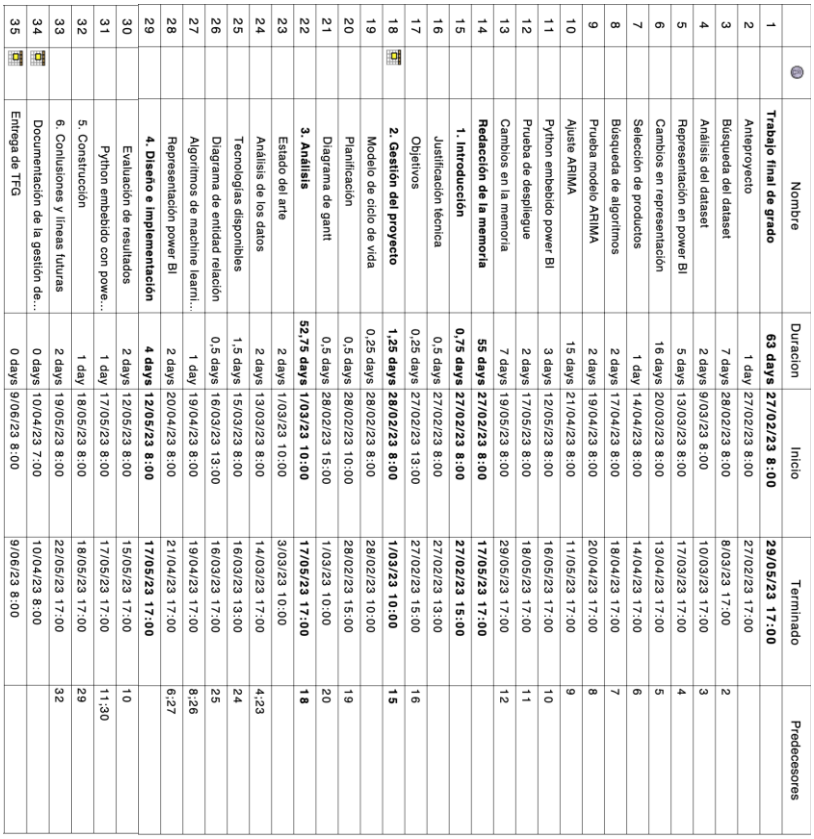
Las tareas que mayores cambios han sufrido respecto a la planificación inicial del proyecto han sido las siguientes:

- Búsqueda del dataset: Esta tarea resultó ser más complicada de lo esperado, lo cual llevó a extender el plazo de trabajo inicialmente previsto de cuatro días a una semana completa. Esto se debió principalmente a las dificultades para encontrar conjuntos de datos de tiendas disponibles, y, además, los hallados no contaban con el histórico de datos necesario de al menos dos años para cumplir con los requisitos establecidos.
- Representación y cambios en Power BI: Este ha sido otro trabajo que ha llevado más tiempo del programado en un principio. Finalmente, la correcta representación y cambios realizados a la misma se han alargado dos semanas más, tras compartir impresiones con el profesor en diferentes reuniones sobre modificaciones para potenciar al máximo la información extraíble.
- Ajuste ARIMA: Por último, la calibración de los parámetros del modelo ARIMA ha supuesto un gran desafío ya que se obtenían predicciones realmente poco precisas al comparar los datos de pronóstico con los reales. A pesar de que los datos cumplían con las características para ser evaluados por el algoritmo ARIMA, como son la estacionalidad, estacionariedad y continuidad, la predicción obtenida resultaba ser una media aritmética del conjunto de datos histórico. Se logró arreglarlo al

añadir otro parámetro de estacionalidad que el algoritmo por defecto no estaba teniendo en consideración.

Además, se han añadido dos tareas, una denominada prueba de despliegue, ya que se realizó la puesta en marcha según se indica en el apartado de construcción y otra de cambios en la memoria que se realizó tras leerla detenidamente, consensuando con el tutor las modificaciones oportunas.

Por todo ello, el proyecto se ha alargado finalmente hasta el día 29 de mayo.



		Nombre	Duración	Inicio	Terminado	Predecesores
1		**Trabajo final de grado**	**63 days**	**27/02/23 8:00**	**29/05/23 17:00**	
2		Anteproyecto	1 day	27/02/23 8:00	27/02/23 17:00	
3		Búsqueda del dataset	7 days	28/02/23 8:00	8/03/23 17:00	2
4		Análisis del dataset	2 days	9/03/23 8:00	10/03/23 17:00	3
5		Representación en power BI	5 days	13/03/23 8:00	17/03/23 17:00	4
6		Cambios en representación	16 days	20/03/23 8:00	13/04/23 17:00	5
7		Selección de productos	1 day	14/04/23 8:00	14/04/23 17:00	6
8		Búsqueda de algoritmos	2 days	17/04/23 8:00	18/04/23 17:00	7
9		Prueba modelo ARIMA	2 days	19/04/23 8:00	20/04/23 17:00	8
10		Ajuste ARIMA	15 days	21/04/23 8:00	11/05/23 17:00	9
11		Python embebido power BI	3 days	12/05/23 8:00	16/05/23 17:00	10
12		Prueba de despliegue	2 days	17/05/23 8:00	18/05/23 17:00	11
13		Cambios en la memoria	7 days	19/05/23 8:00	29/05/23 17:00	12
14		**Redacción de la memoria**	**55 days**	**27/02/23 8:00**	**17/05/23 17:00**	
15		**1. Introducción**	**0,75 days**	**27/02/23 8:00**	**27/02/23 15:00**	
16		Justificación técnica	0,5 days	27/02/23 8:00	27/02/23 13:00	
17		Objetivos	0,25 days	27/02/23 13:00	27/02/23 15:00	16
18		**2. Gestión del proyecto**	**1,25 days**	**28/02/23 8:00**	**1/03/23 10:00**	**15**
19		Modelo de ciclo de vida	0,25 days	28/02/23 8:00	28/02/23 10:00	
20		Planificación	0,5 days	28/02/23 10:00	28/02/23 15:00	19
21		Diagrama de gantt	0,5 days	28/02/23 15:00	1/03/23 10:00	20
22		**3. Análisis**	**52,75 days**	**1/03/23 10:00**	**17/05/23 17:00**	**18**
23		Estado del arte	2 days	1/03/23 10:00	3/03/23 10:00	
24		Análisis de los datos	2 days	13/03/23 8:00	14/03/23 17:00	4;23
25		Tecnologías disponibles	1,5 days	15/03/23 8:00	16/03/23 13:00	24
26		Diagrama de entidad relación	0,5 days	16/03/23 13:00	16/03/23 17:00	25
27		Algoritmos de machine learni...	1 day	19/04/23 8:00	19/04/23 17:00	8;26
28		Representación power BI	2 days	20/04/23 8:00	21/04/23 17:00	6;27
29		**4. Diseño e implementación**	**4 days**	**12/05/23 8:00**	**17/05/23 17:00**	
30		Evaluación de resultados	2 days	12/05/23 8:00	15/05/23 17:00	10
31		Python embebido con powe...	1 day	17/05/23 8:00	17/05/23 17:00	11;30
32		5. Construcción	1 day	18/05/23 8:00	18/05/23 17:00	29
33		6. Conlusiones y líneas futuras	2 days	19/05/23 8:00	22/05/23 17:00	32
34		Documentación de la gestión de...	0 days	10/04/23 7:00	10/04/23 8:00	
35		Entrega de TFG	0 days	9/06/23 8:00	9/06/23 8:00	

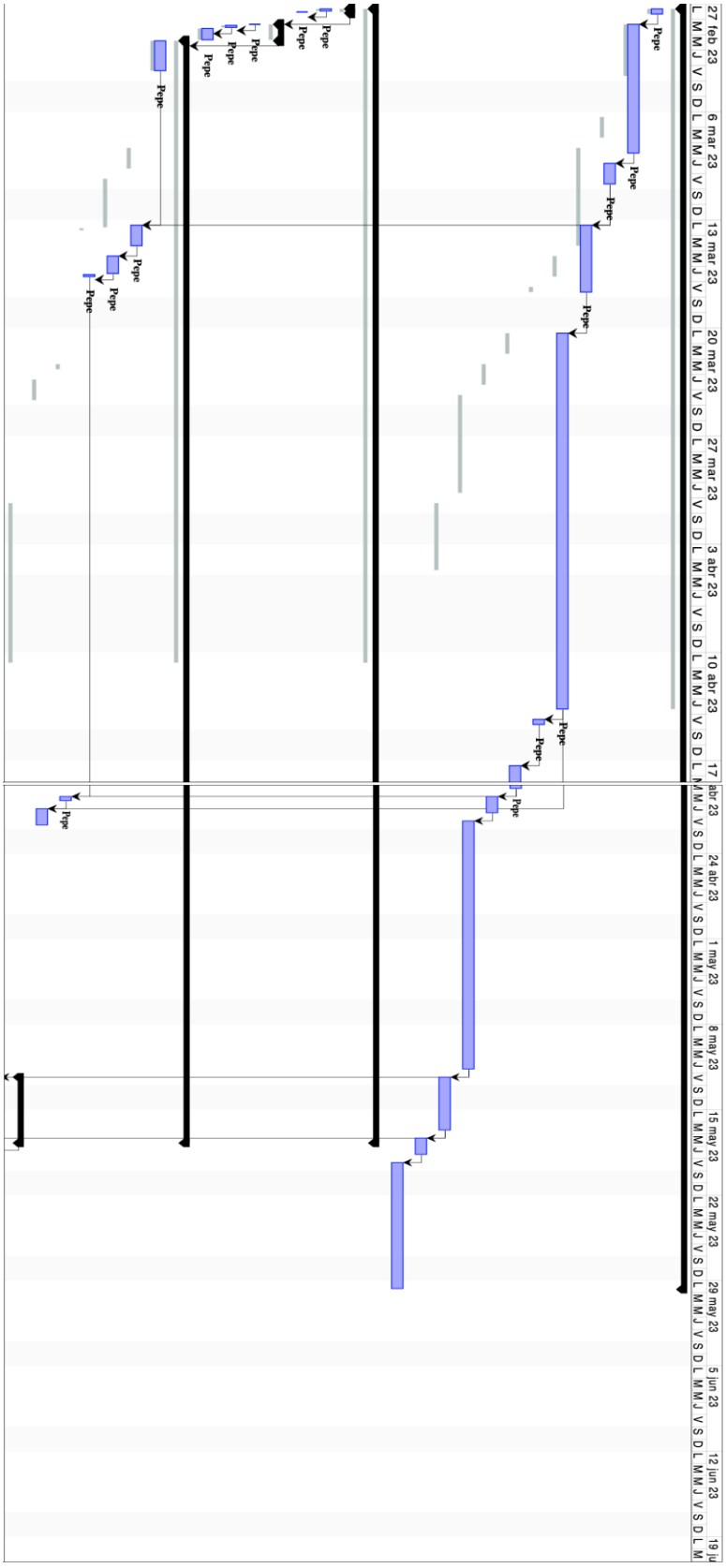


Ilustración 2: Planificación final.

Capítulo 3

Análisis

3.1 Estado del Arte

Generalmente, cuando se habla del mundo empresarial, el activo que más se valora y en el que se pone el foco de atención es el componente humano. No solo los empleados de la propia organización, sino también los clientes, que, en este mercado globalizado en constante evolución, cada vez demandan servicios más rápidos y eficientes de las empresas. En consecuencia, las empresas deben prestar especial atención a la información, ya que para mantenerse competitivas deben estar uno o dos pasos por delante de las posibles necesidades y exigencias del consumidor. Desde el año 2010, los datos han crecido a un ritmo exponencial. No obstante, gran cantidad de los mismos están desestructurados por lo que su análisis supone un reto.

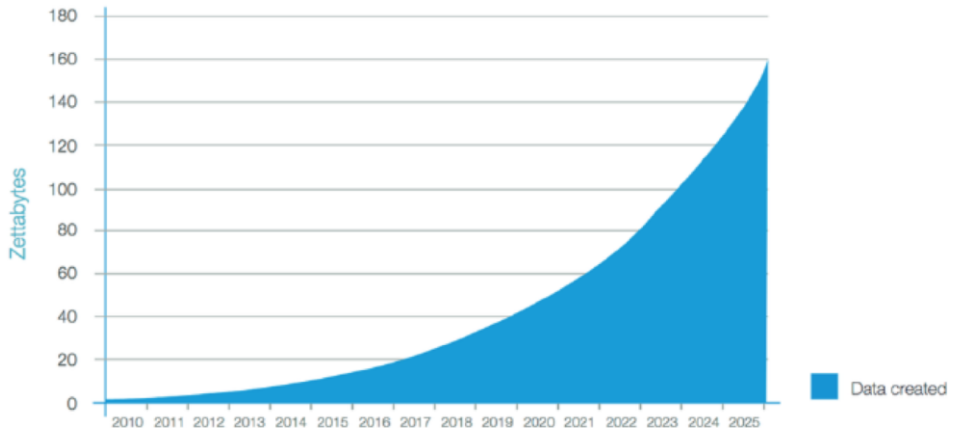


Ilustración 3: Crecimiento de la información.

(zettabyte = unidad de información igual a 10^{21}bytes).

Por tanto, una empresa para mejorar su rendimiento y ser competitiva debe tomar decisiones basadas en información precisa y pertinente, por lo que están

utilizando y confiando en sistemas de Inteligencia Empresarial para mantenerse a la vanguardia de tendencias y eventos futuros. Además, el BI acelera la toma de decisiones, permitiendo a la compañía actuar rápidamente, adelantándose a sus competidores y obteniendo por ello una ventaja competitiva.

El BI se refiere al conjunto de tecnologías, aplicaciones y estrategias para la recolección, integración, análisis y presentación de información de negocios [4]. Permite transformar los datos en conocimiento para mejorar la toma de decisiones. Estos sistemas tienen la capacidad de extraer y analizar grandes cantidades de datos aportando una visión más detallada de éstos, lo cual ayuda a detectar patrones y tendencias comerciales futuras [5].

Loshin habla de que el BI se utiliza para mejorar el rendimiento, así como reducir costes e identificar oportunidades de negocio, incluyendo marketing personalizado, análisis de riegos, identificación de tendencias de compra para predecir el stock necesario o medición, seguimiento y predicción de ventas entre otras muchas [6].

Según el estudio realizado por McKinsey, consultora estratégica global, las empresas pueden aumentar su retorno de inversión (Return On Investment, ROI) entre un 10 y 20% a través de una mayor inversión en recopilación y análisis de datos. Además, estas compañías son un 5% más productivas y un 6% más rentables que sus competidores [7].

Ya se ha visto que la correcta implementación de la inteligencia empresarial puede aportar numerosos beneficios para la firma. Un ejemplo de fructífera implementación de este conjunto de tecnologías es la empresa de seguros EMC Insurances, con más de 2100 empleados, y activos con valor de unos tres mil millones de dólares. La aseguradora contaba con una gran cantidad de datos de seguros, pero una capacidad limitada para analizar la información, por lo que no podía encontrar el equilibrio de reservas para hacer frente a los pagos. El sistema de BI le permitió descubrir correlaciones y patrones ocultos en las bases de datos

de reclamaciones e incluía un modelo predictivo para pronosticar las reservas necesarias. [8]

Una vez encontrado un dataset con una sucesión de datos medidos en determinados momentos y ordenados cronológicamente, es decir, una serie temporal, se deben tomar una serie de decisiones de diseño que permitan cumplir los objetivos fijados anteriormente. En este apartado se exponen y desarrollan tales decisiones.

El Machine Learning consiste en extraer información a partir de los datos. Es un campo de investigación en la intersección de la estadística, inteligencia artificial y la informática, también conocido como análisis predictivo o aprendizaje estadístico. Permite resolver una gran cantidad de problemas como la clasificación de imágenes, recomendación de productos, predicción de demanda o traducción automática entre otros. Sin embargo, se deberá escoger el algoritmo correcto, ya que existen variaciones y diferencias en la forma en la que se enseñan y entrenan los modelos y la condición de los datos que se necesitan.

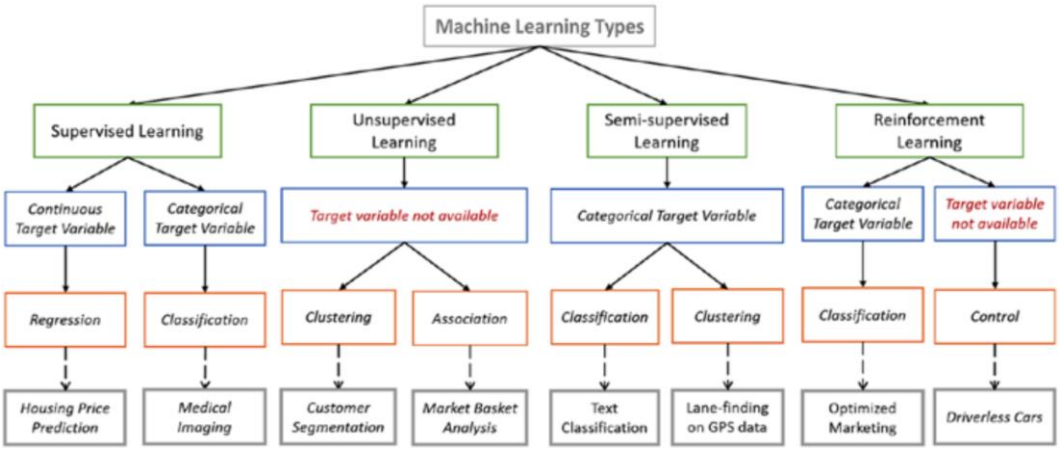


Ilustración 4: Tipos de algoritmos Machine Learning.

Como se puede ver, existen cuatro tipos de clasificación dentro del aprendizaje automático, pero se centrará el estudio en el aprendizaje supervisado y no supervisado [9]:

- Aprendizaje supervisado: El usuario proporciona al algoritmo parejas de datos de entrada y de salida para que dada una entrada este encuentre la salida esperada. El modelo será construido a partir de esas parejas de datos para que comprendan los datos de entrenamiento. Es utilizado siempre que se quiera predecir un determinado "output" a partir de un "input". El objetivo es realizar predicciones certeras para datos nunca vistos.

  Hay dos principales problemas de aprendizaje supervisado:

  - Clasificación: El resultado es una clase entre un número limitado de clases dado el problema en cuestión, por tanto, es categórico. Por ejemplo, si se quiere detectar si un correo es spam o no, sólo hay dos posibles clases.
  - Regresión: El objetivo es predecir un número, dentro de un conjunto infinito de resultados. Por ejemplo, predecir el rendimiento de una granja si es conocido su rendimiento histórico, clima y número de empleados.

  Una forma de distinguir entre clasificación y regresión es la temporalidad de los datos. Si hay continuidad a través del tiempo, entonces es un problema de regresión.

  Los algoritmos más populares para este tipo de problemas son: Naive Bayes, árboles de decisión, redes neuronales y ARIMA.

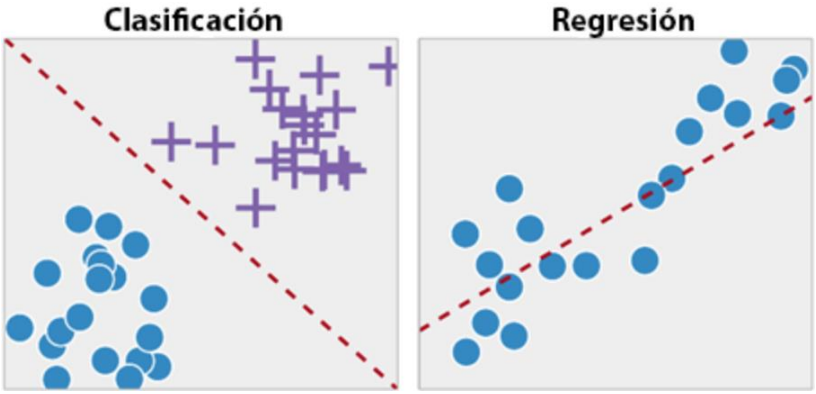


Ilustración 5: Clasificación vs Regresión.

- Aprendizaje no supervisado: En este otro tipo, sólo se proporcionan datos de entrada, no hay ningún dato de salida. Estos algoritmos tratan de detectar patrones para examinar y agrupar los conjuntos de datos, pero no tienen ningún tipo de entrenamiento.

  Se puede distinguir entre agrupamiento o "clustering" y asociación:

  - Clustering: Divide los datos en diferentes grupos con características similares. A diferencia del algoritmo de clasificación del aprendizaje supervisado, los grupos no son conocidos previamente. Por ejemplo, diferenciar segmentos de cliente en base a su comportamiento de compra.
  - Asociación: Se utiliza para extraer reglas y patrones de los datasets. Trata de descubrir relaciones interesantes entre variables y atributos de los datos. Por ejemplo, los clientes que compran una casa tienden a comprar muebles.

  El desafío llega a la hora de supervisar si el modelo ha funcionado correctamente, ya que se aplica a datos sin valores de salida por lo que no se conoce el resultado correcto. Por ello, suelen ser utilizados de forma experimental, cuando se quiere tener una visión más profunda de los datos.

Algunos de los algoritmos de clustering más relevantes son: K-means, K-medoids y FP Growth.

Una vez se conocen los diferentes tipos de Machine Learning existentes, se debe escoger el adecuado para las tareas de predicción que se desean realizar en el conjunto de datos. Claramente, para pronosticar el stock necesario para un producto concreto de la empresa en un futuro se debe utilizar el aprendizaje supervisado (regresión más concretamente) ya que hay continuidad a través del tiempo.

De todos los algoritmos de regresión mencionados anteriormente se va a hacer uso de ARIMA, ya que ha sido recomendado por el tutor del proyecto al ser un modelo que tiene en cuenta parámetros de estacionalidad y autoregresión, además de que es ampliamente utilizado para el pronóstico de series de tiempo y dispone de librerías en Python que facilitan el cálculo e interpretación de los resultados obtenidos.

ARIMA que es la abreviatura de media móvil integrada autorregresiva, es una forma de análisis de regresión con el objetivo de predecir los valores futuros de una serie temporal examinando las diferencias entre los valores de la serie en lugar de utilizar los valores reales.

Si se desglosa el acrónimo puede ser más fácil de comprender:

- Autoregresión (AR): Implica realizar una regresión lineal entre el valor actual de la serie y sus valores previos. El orden del modelo autorregresivo, conocido como “p”, indica cuántos valores anteriores se toman en cuenta.
- Integrado (I): Implica realizar una diferenciación de las observaciones originales para lograr que el conjunto de datos sea estacionario, es decir, se reemplazan los valores de los datos por las diferencias entre los valores actuales y los valores anteriores. Este proceso ayuda a eliminar tendencias y los patrones no estacionarios, por ello si los datos ya son de por sí estacionarios no es necesario aplicar ninguna diferenciación. El parámetro

de integración es conocido como “d”, haciendo referencia al número de veces que las originales son diferenciadas.

- Media móvil (MA): Se tiene en cuenta cómo una observación en una serie de datos puede depender de los errores residuales generados por un modelo de promedio móvil aplicado a mediciones previas, en otras palabras, menos técnicas, lo que ocurre es que los valores pasados pueden estar influenciados por los errores de predicción cometidos en el pasado. El parámetro de media móvil “q” denota el tamaño de la ventana de promedio móvil.

Los parámetros de “*seasonal(P, D, Q, S)”* hacen referencia a los mismos términos que sus homólogos en minúscula, y el componente “S” se refiere al periodo estacional, es decir, la longitud del ciclo estacional de la serie temporal. Por ejemplo, si es un patrón estacional anual se establecería la “S” a 12 ya que son los periodos transcurridos hasta repetirse nuevamente.

Al aplicar este tipo de modelo, es conveniente que la serie de datos tenga como mínimo 50 observaciones, pero se recomienda que disponga de 100 observaciones, es decir, al menos dos años ya que corresponden a 104 semanas para lograr resultados más adecuados.

Es importante enfatizar que este algoritmo realiza una serie de suposiciones. En primer lugar, los datos deben ser estacionarios, es decir, las propiedades estadísticas como la media o la varianza no cambian con el tiempo y se mantienen estables. En caso de que la serie de datos no lo sea, se puede aplicar diferenciación a los datos para convertirlos en estacionarios. También hay que añadir que ARIMA emplea una única variable, sin tener en cuenta la que recoge el tiempo.

3.2 Tecnologías disponibles

3.2.1 Herramientas de Business Intelligence

Tras el estado del arte, se pasarán a exponer las diferentes tecnologías que se encuentran disponibles en el mercado para poder ofrecer una solución al problema propuesto. Para ello se realizará una comparativa entre las aplicaciones disponibles y se elegirá la que mejores prestaciones ofrezca, todo ello de manera argumentada.

Para el proyecto propuesto, se precisa de una herramienta que permita recolectar, procesar y analizar grandes cantidades de datos para que, una vez procesada, se pueda obtener información útil y representarla de forma clara y concisa y con un soporte visual.

Este tipo de tareas las realiza la Inteligencia Empresarial, por lo que se enfocará el proyecto en las aplicaciones que dan soporte a esta área de trabajo. No obstante, existen numerosos programas que se adaptan a las necesidades que se están buscando. Por tanto, se hará uso de la herramienta que mejor encaje con los objetivos, teniendo en cuenta también otros factores como coste o aprendizaje.

Gartner es una empresa consultora y de investigación de las tecnologías de la información que en enero de 2023 publicó el Cuadrante Mágico (MQ) para plataformas de Business Intelligence y Analytics, cuyos resultados se pueden observar en la siguiente imagen:

Ilustración 6: Cuadrante mágico de Gartner para plataforma de BI 2023.

Como se puede apreciar en el cuadrante, se cuentan con numerosas tecnologías, pero destacan tres, ya sea por seguridad o facilidad de uso, que son líderes del sector: Microsoft con Power BI, Salesforce con Tableau y Qlick con QlickView. Por tanto, el análisis comparativo será únicamente entre estas opciones.

Para su clasificación se tienen en cuenta las siguientes capacidades:

- Seguridad de datos
- Data governance
- Análisis en la nube
- Conectividad de fuentes de datos
- Preparación de datos

- Catálogo de datos
- Información automatizada
- Visualización de datos
- Consultas en lenguaje natural
- Storytelling
- Generación de lenguaje natural
- Reporting

En agosto de 2022, la empresa B-eye, especializada en Data Analytics, analizó las amenazas y fortalezas de las tres herramientas en las que se va a centrar nuestra comparativa.

Los resultados obtenidos por este análisis son los siguientes:

Ilustración 7: Comparación de herramientas por B- eye.

Como se puede ver en la imagen, la herramienta que obtiene una puntuación mayor por las funcionalidades ofrecidas es Qlick Sense, seguida de Power BI y por último Tableau.

Si nuestro estudio tuviese en cuenta únicamente estos parámetros y todos ellos estuviesen valorados con la misma importancia, se debería elegir Qlick Sense. Sin

embargo, hay algunos factores que son más relevantes que otros a la hora de realizar este proyecto.

Se debe tener en cuenta que este proyecto es un encargo para un cliente, por ello habrá que buscar la solución que sea más económica y tenga una mayor facilidad de uso al realizar el trabajo. Además, una vez se ha centrado la atención en las tres mejores herramientas disponibles en el mercado, las diferencias que puede haber entre ellas pueden ser mínimas.

A continuación, se profundizará en el coste de utilización y la facilidad de aprendizaje de cada herramienta.

3.2.1.1 Coste de utilización

Tabla 1: Comparativa de precios plataformas de BI

	POWERBI	TABLEAU	QLICK SENSE
PRUEBA GRATIS	30 días	14 días	30 días
LICENCIA	8,40 €/mes	35 $/mes	30 $/mes
VERSIÓN PRO	Precio variable	Licencia de creador: 70 $/mes	No disponible

Las tres aplicaciones disponen de periodos de prueba gratuitos para poder conocer la herramienta y su alcance. Sin embargo, Tableau lo ofrece únicamente durante catorce días mientras que Power BI y Qlick Sense lo hacen durante un mes.

Power BI es con diferencia la más económica de las tres. Además, al ser estudiante de la Universidad San Pablo CEU se cuenta con licencia para poder utilizar el paquete de Microsoft 365 y Power BI de forma completamente gratuita. Cabe destacar que, en 2022, la universidad y la compañía tecnológica llegaron a un

acuerdo para impulsar la aplicación de inteligencia artificial en las empresas con el programa Microsoft AI Business School y la iniciativa Microsoft Garage. [10]

3.2.1.2 Facilidad de aprendizaje

Si se tuviese en cuenta únicamente el precio la elección sería Power BI, pero otro factor para tener en cuenta es la facilidad de uso y el tiempo de aprendizaje de cada herramienta.

Las tres aplicaciones ofrecen tutoriales, ya sea por escrito o mediante vídeos, para poder introducirnos a las características de cada una de ellas.

Por ejemplo, Tableau cuenta en su página web con un centro de ayuda con numerosas secciones que facilitan el aprendizaje. Son una serie de video tutoriales que van desde la conexión a los datos hasta la distribución de los informes.

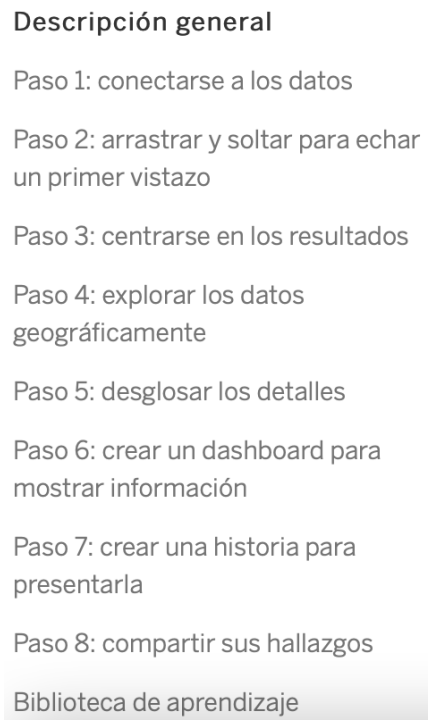


Ilustración 8:Vídeo tutoriales de Tableau.

El problema es que no hay ningún tipo de documentación por escrito, sólo existe el sistema de aprendizaje vía vídeos. QlickSense, al contrario, cuenta únicamente con documentación escrita junto a capturas de pantalla para conocer los detalles de uso de su aplicación.

Finalmente, Power BI dispone tanto de documentación escrita completada con capturas de pantalla como de vídeo tutoriales, así como de rutas de aprendizaje en su página web.

Ilustración 9:Documentación de Power BI.

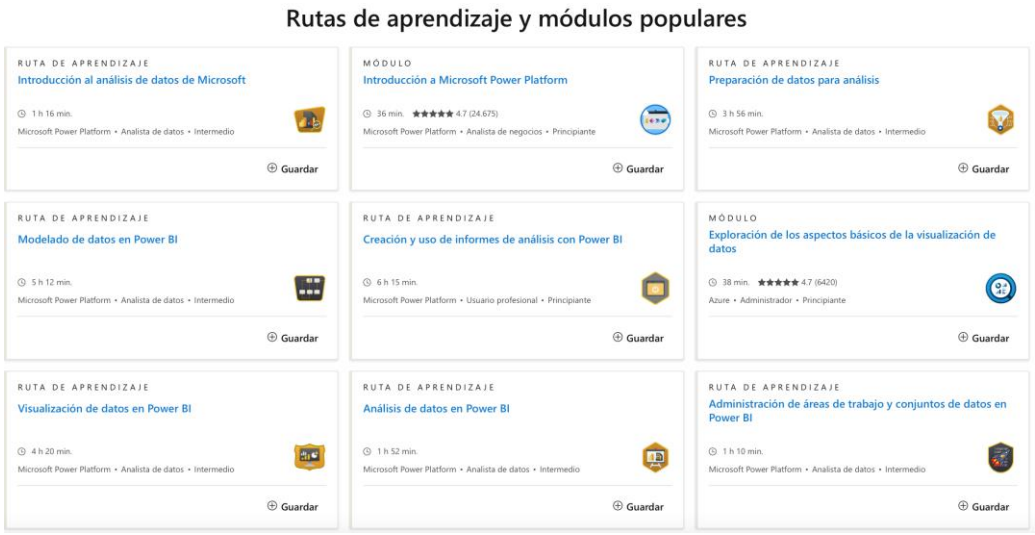


Ilustración 10:Rutas de aprendizaje Power BI.

En consecuencia, Power BI es la herramienta que ofrece una documentación y formación más completa y variada. Adicionalmente, el análisis realizado por B-eye otorgaba una valoración positiva, solo por detrás de Qlick Sense.

Además, cabe destacar que Power BI ya se ha utilizado previamente durante la etapa en la universidad por lo que se cuenta con una base sólida sobre la que se puede trabajar y profundizar.

3.3 Especificación de requisitos

3.3.1 Power BI

- El informe tiene como objetivo proporcionar una visión general sobre los ingresos de la cadena "1C Company" en base a productos, categorías, localización y tipo de tienda.
- Los datos utilizados se extraerán de diferentes archivos en formato ".csv" descargados de Kaggle y se integrarán con otra fuente externa de datos demográficos.
- Se requerirán gráficos de barras, de líneas, Key Performance Indicators (KPIs) y mapas para representar las tendencias de ventas.
- El informe debe utilizar el logotipo de la empresa.
- Los usuarios deben poder filtrar los datos por periodo de tiempo, región y categoría de producto.
- Se debe garantizar un tiempo de carga rápido del informe, aun teniendo grandes volúmenes de datos.
- El informe debe ser compatible con dispositivos móviles para que pueda ser accesible desde smartphones.

3.3.2 ARIMA

- Predecir la demanda futura de un producto en concreto, basado en sus datos históricos en el lenguaje Python.

- Los datos serán preparados adecuadamente para su uso en el modelo ARIMA.
- Se realizará un análisis exploratorio de los datos para identificar patrones de estacionalidad, estacionariedad o correlación.
- Se seleccionará el modelo que mejores resultados ofrezca, tras evaluar la precisión del modelo en base a los parámetros de autoregresión(p), integración(d) y media móvil (q) que se introducen al modelo.
- Se ensayará con los datos de los dos primeros años y se comparará el resultado de la predicción del tercer año con los datos reales.
- Una vez validado el modelo, se realizará el pronóstico para el cuarto año de ventas.
- Se utilizarán métricas de evaluación como el Akaike's Information Criterion(AIC), Bayesian Information Criterion (BIC) o error cuadrático medio (Mean Squared Error, MSE) para medir la precisión del modelo en la predicción de las ventas futuras.
- El pronóstico será representado en una sola gráfica que diferencie los datos reales de los datos pronosticados.
- Se documentarán los pasos seguidos, los parámetros seleccionados y las decisiones tomadas del modelo ARIMA.
- Se presentarán las conclusiones obtenidas del análisis y predicción de la demanda futura.
- Se integrará el script de Python realizado con Power BI para visualizar la previsión de las ventas del producto.

Capítulo 4

Diseño e implementación

4.1 Diseño físico de los datos

El dataset seleccionado contiene los datos de ventas de la empresa 1C Company, empresa rusa que se dedica a la venta de programas informáticos, consolas y videojuegos, desde enero de 2013 hasta octubre de 2015. La información está dividida en cuatro tablas diferentes almacenadas en formato ".csv".

La tabla "sales" cuenta con la siguiente información:

- date: Fecha en formato dd.mm.yy de la venta del producto.
- Date_block_num: Se refiere al número de mes en el que se encuentra. Siendo el 0 para el mes de enero de 2013 y el número 33 para el mes de octubre de 2015.
- Shop_id: Identificador de tienda, para relacionarlo posteriormente con la tabla "shops".
- Item_id: Identificador de producto, para relacionarlo con la tabla "ítems".
- Item_price: Precio del producto en cuestión.
- Item_cnt_day: Número de productos vendidos.

date	date_block_num	shop_id	item_id	item_price	item_cnt_day
02.01.2013	0	59	22154	999	1
03.01.2013	0	25	2552	899	1
05.01.2013	0	25	2552	899	-1
06.01.2013	0	25	2554	1709.05	1
15.01.2013	0	25	2555	1099	1

Ilustración 11: Ejemplo tabla "sales".

La tabla “calendar” contiene los siguientes campos:

- Date: Fecha en formato yyyy-mm-dd.
- Holiday: Indica si es un día festivo, con el valor “1” o si es un día laboral con el valor “0”.
- Weekend: Indica si es fin de semana con el valor “1” o si es entre semana con valor “0”.

date	holiday	weekend
2013-01-01	1	0
2013-01-02	1	0
2013-01-03	1	0
2013-01-04	1	0
2013-01-05	0	1

Ilustración 12: Ejemplo tabla "calendar".

La tabla “item_category” cuenta con la siguiente información:

- item_id: Identificador de producto. Hay 22.149 productos diferentes.
- item_name_translated: Nombre del producto.
- Item_cat1: Nombre de la categoría a la que pertenece el producto. Existen 22 categorías diferentes.
- Item_cat2: Nombre de la subcategoría. Existen 61 subcategorías.

item_id	item_name_translated	item_cat1	item_cat2
5441	PC: Headset HyperX C…	PC	Headsets / Headphones
16255	Headphones PHILIPS SBC HC8680	PC	Headsets / Headphones
16256	Headphones RITMIX RH-120	PC	Headsets / Headphones
16257	Headphones RITMIX RH-124 Black	PC	Headsets / Headphones
5606	PS2: Memory Card 8 M…	Accessories	PS2

Ilustración 13: Ejemplo tabla "item_category".

La tabla “shops-translated” cuenta con la siguiente información:

- Shop_id: Identificador de la tienda. Hay sesenta tiendas numeradas desde el cero hasta el cincuenta y nueve.
- City: Ciudad en la que se encuentra la tienda física.
- Type: El tipo de tienda. Puede ser una tienda (shop), Centro Comercial (SC), Centro de Comercio (TC), Centro Comercial y de Entretenimiento (SEC), Centro de Comercio y Recreación (TRC) y online (web).
- Name: Nombre de la tienda.
- Población: Número de habitantes de cada ciudad.

shop_id ▲	City	Type	Name
0	Yakutsk	Shop	Ordzhonikidze, 56 francs
1	Yakutsk	TC	Central franc
2	Adygea	TC	Mega
3	Balashikha	TRC	October-Kinomir
4	Volzhsky	TC	Volga Mall

Ilustración 14: Ejemplo tabla "shops-translated".

Para complementar la información de la que se dispone con esta serie de tablas se ha decidido extraer la información de la población en las diferentes ciudades que cuentan con tiendas y así unificarla a la tabla “shops-translated”. La tabla en cuestión es la siguiente:

Index	City	Population
0	Adygea	440327
1	Balashikha	215494
2	Chekhov	60720
3	Kaluga	324698
4	Kazan	1143535

Ilustración 15: Ejemplo Tabla "CensoRusia".

Para conocer con mayor detalle el tipo de datos que se está tratando, es importante realizar una exploración descriptiva de los mismos. Esto permite conocer las variables y detectar posibles errores.

En primer lugar, se debe obtener el tipo de datos de cada columna de las tablas.

```
RangeIndex: 2935849 entries, 0 to 2935848
Data columns (total 6 columns):
 #   Column          Dtype
---  ------          -----
 0   date            object
 1   date_block_num  int64
 2   shop_id         int64
 3   item_id         int64
 4   item_price      float64
 5   item_cnt_day    float64
dtypes: float64(2), int64(3), object(1)
```

Ilustración 16: Información tabla "sales".

Esta tabla cuenta con 2.935.849 filas y 6 columnas. Todos los datos cuentan con el formato correcto excepto la variable "date" que tiene un tipo de datos "object" en lugar de "date". Será necesario cambiarlo para que ese campo pueda ser reconocido como una fecha en lugar de un objeto.

```
RangeIndex: 1095 entries, 0 to 1094
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   date     1095 non-null   object
 1   holiday  1095 non-null   int64
 2   weekend  1095 non-null   int64
dtypes: int64(2), object(1)
```

Ilustración 17: Información tabla "calendar".

La tabla “calendar” tiene 1095 filas y 3 columnas, pero tiene el mismo problema que la tabla anterior. La columna “date” no tiene el tipo de datos adecuado por lo que precisará la transformación de los datos, sin embargo, el resto de las columnas son correctas.

```
RangeIndex: 60 entries, 0 to 59
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   shop_id  60 non-null     int64
 1   City     60 non-null     object
 2   Type     60 non-null     object
 3   Name     59 non-null     object
dtypes: int64(1), object(3)
```

Ilustración 18:Información tabla "shops".

Esta tabla tiene 60 filas y 4 columnas. El tipo de datos es correcto, ya que en la librería “pandas” que está siendo utilizada para el tratamiento de datos, el tipo de datos “object” es el equivalente al tipo de datos de Python nativo “String”.

```
RangeIndex: 22170 entries, 0 to 22169
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   item_id              22170 non-null  int64
 1   item_name_translated 22170 non-null  object
 2   item_cat1            22170 non-null  object
 3   item_cat2            22112 non-null  object
dtypes: int64(1), object(3)
```

Ilustración 19:Información tabla "items".

La tabla “ítems” cuenta con 22.169 productos diferentes y 4 columnas. El tipo de datos de las 4 variables es el adecuado por lo que no precisarán de ningún tipo de cambio.

```
RangeIndex: 29 entries, 0 to 28
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   City        29 non-null     object
 1   Population  29 non-null     int64
dtypes: int64(1), object(1)
```

Ilustración 20: Información tabla "censoRusia".

La tabla “censoRusia” cuenta con 29 ciudades diferentes y con dos columnas que se refieren al nombre de la ciudad y la cantidad de población. Los datos son correctos por lo que no requieren ser manipulados.

Además de conocer el tipo de datos con el que se va a trabajar, es necesario conocer si existen valores nulos o ausentes, ya que los algoritmos de machine learning no aceptan variables incompletas o en caso de que las acepten, se ven altamente influenciados por ellas.

Las dos únicas tablas que cuentan con campos vacíos son la tabla “shops” e “ítems”.

```
item_id                0
item_name_translated     0
item_cat1                0
item_cat2               58
dtype: int64
```

Ilustración 21:Valores nulos tabla "items".

Como se puede apreciar, existen 58 valores “item_cat2” que no tienen ningún valor, es decir, es un campo vacío. En un principio, no representa ningún inconveniente dado que la columna "item_cat1" proporciona la información necesaria para los cálculos que se llevarán a cabo.

9	OutboundTrade	Other	nan

Ilustración 22: Valor nulo tabla "shops".

En la tabla “shops” hay un único campo vacío que corresponde al nombre de la tienda. Al analizar la entrada correspondiente, se observa que es una venta que no se ha realizado en una tienda física, sino que ha sido una venta al extranjero, por lo que será un campo para tener en cuenta a la hora de valorar los resultados.

Dado que el modelo de datos utilizado es relacional, serán necesarias relaciones 1 a N para almacenar las ventas realizadas por cada tienda, pues una misma tienda puede realizar numerosas ventas. Además, se pueden hacer muchas ventas en un único día. Finalmente, dado que un producto puede ser vendido en múltiples ocasiones, también se crea una relación 1 a N entre la tabla “sales” e “item_category”.

El diagrama de entidad relación resultante se muestra a continuación:

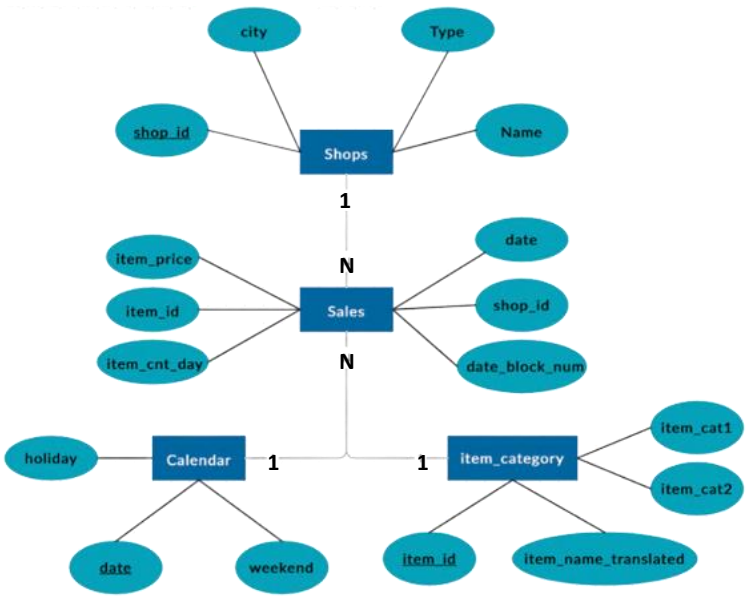


Ilustración 23: Diagrama de Entidad-Relación.

Tras conocer adecuadamente los datos con los que se trabajarán, tanto sus tablas como el formato de cada una de sus columnas, se pueden realizar las modificaciones pertinentes para utilizar el algoritmo ARIMA.

Como se mencionó anteriormente, se ha estandarizado el formato de los campos que contienen fechas para asegurar una gestión adecuada y consistente de dicha información. Además, se han unificado las diferentes tablas para así poder trabajar con una única que contenga todos los datos.

Por último, se han agrupado los datos por semanas, se han filtrado los productos por su identificador de producto para realizar el análisis de un único producto y se ha prescindido de todas las columnas excepto de las columnas que almacenan los datos de fechas y las ventas de producto por semanas. El conjunto de datos resultante, utilizado en el algoritmo ARIMA, se muestra a continuación:

date	item_cnt_day
2013-01-06 00:00:00	235
2013-01-13 00:00:00	158
2013-01-20 00:00:00	141
2013-01-27 00:00:00	129
2013-02-03 00:00:00	113
2013-02-10 00:00:00	96
2013-02-17 00:00:00	87
2013-02-24 00:00:00	118
2013-03-03 00:00:00	81

Ilustración 24: Información tabla df_weeks_bo2

4.2 Diseño de ARIMA

Una vez estudiados y transformados los datos para facilitar su entendimiento y manejabilidad se puede empezar a comprobar si son válidos para realizar el modelo predictivo con el algoritmo ARIMA para series temporales.

En primer lugar, se ha de escoger un producto entre los más vendidos, que cuente con ventas durante todas las semanas desde el mes de enero de 2013 hasta octubre de 2015 incluido, que es la fecha hasta la que se dispone de información. Es importante ya que los resultados del modelo pueden verse afectados si hay valores que sean iguales a cero en la serie temporal, generando por tanto predicciones menos precisas.

El producto seleccionado, fue el *"Call Of Duty: Black Ops II [PC, Jewel, Russian Version]",* ya que contaba con ventas durante todas las semanas de los años 2013,2014 y 2015, aunque otros productos también cumplían esta característica como *"Playstation Store replenishment of wallet: Payment card 2500 rub."* o *"X360: Gamepad Wireless Black - Wireless Controller BLACK (NSF-00002: Microsoft)."*.

Al representarlos el que mejor aspecto presentaba en cuanto a estacionalidad fue el elegido. Como se puede apreciar en el gráfico, en el mes de enero se produce un repunte de las ventas en los tres años de datos de los que se disponen.**¡Error! No se encuentra el origen de la referencia.**

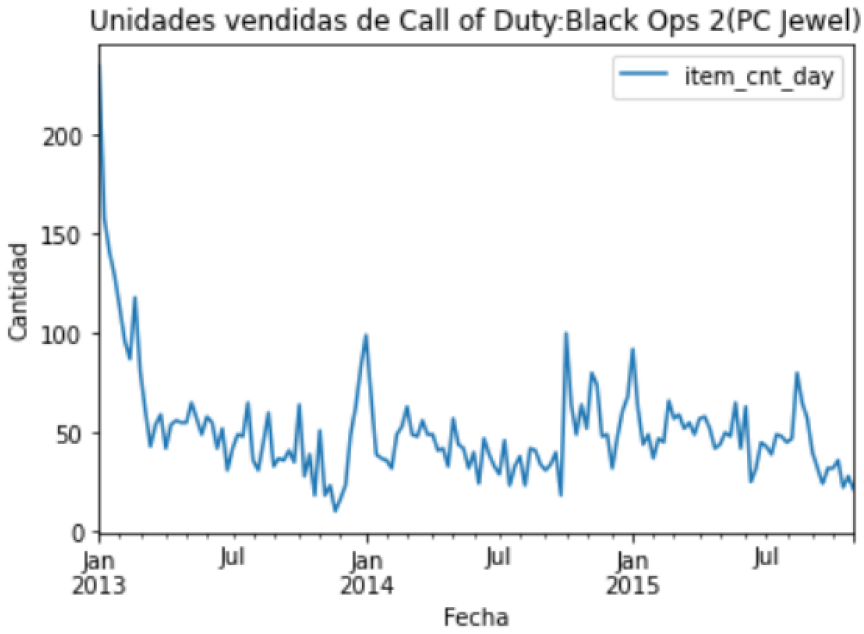


Ilustración 25: Ventas de Call Of Duty: Black Ops II [PC, Jewel, Russian Version].

A continuación, se separan los datos en dos conjuntos de entrenamiento y test. El conjunto de entrenamiento va comprendido desde la primera semana de enero de 2013 hasta la última semana de diciembre de 2014, es decir, 104 semanas para preparar el algoritmo. El conjunto de prueba va desde la primera semana de enero de 2015 hasta la primera semana de noviembre de 2015.

Pese a los indicios, es importante asegurarse que los datos son realmente estacionales y estacionarios. Para ello y para determinar el orden de los parámetros “p” y “q”, se hace uso de la autocorrelación y autocorrelación parcial. La autocorrelación describe la presencia o ausencia de correlación en el conjunto de datos, determinando si las observaciones pasadas influyen en las actuales. La función de autocorrelación (Autocorrelation Function, ACF), mide la correlación entre dos variables separadas por “k” periodos y la función de autocorrelación parcial (Parcial Autocorrelation Function, PACF) realiza la misma medición, pero sin tener en cuenta la dependencia creada por los retardos intermedios existentes entre ambas.

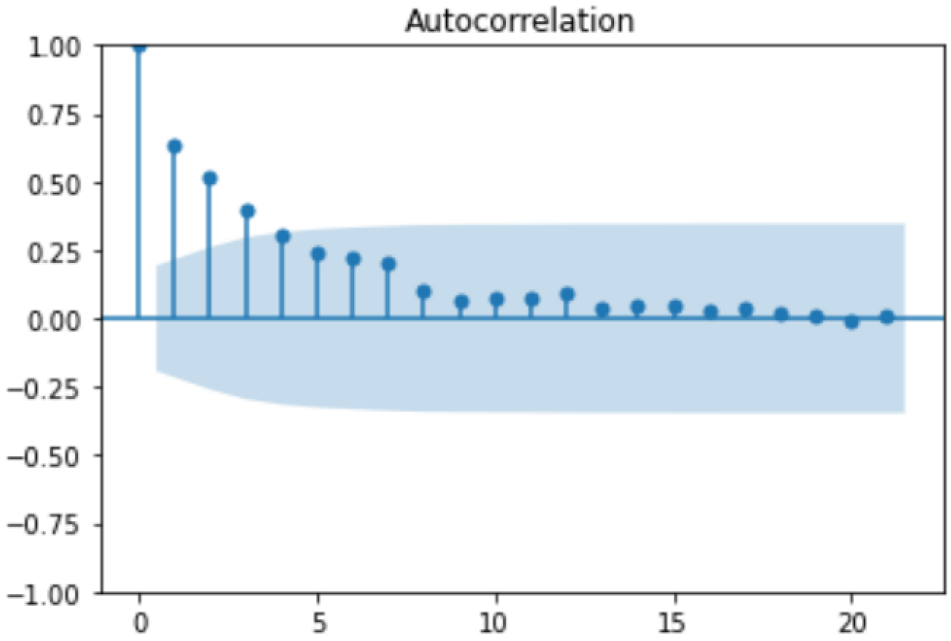


Ilustración 26: Gráfico de autocorrelación.

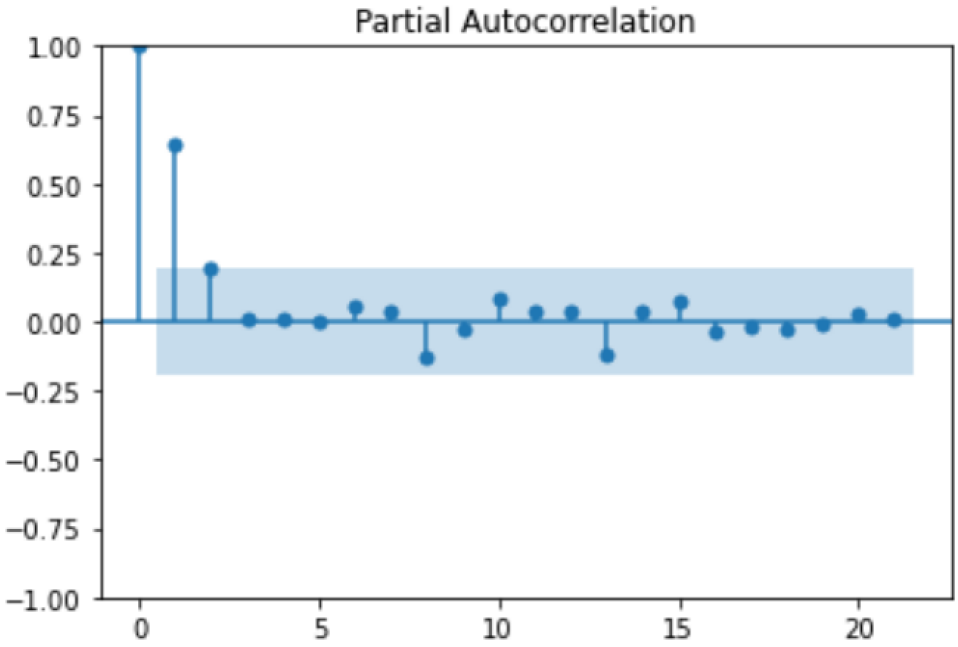


Ilustración 27: Gráfico de autocorrelación parcial.

Se puede observar que hay un gran nivel de correlación entre el retardo igual a 1 y el anterior, lo que haría indicar que si se establecen los parámetros “p” y “q” para ARIMA que se refieren a la autocorrelación (AR) y media móvil (MA) con valor igual a 1, debería dar buenos resultados.

Además, para confirmar que la serie es estacionaria, lo cual es una asunción que realizan los modelos ARIMA, se puede ejecutar el Dickey Fuller test que se basa en la siguiente hipótesis nula y alternativa:

- **H_0:** La serie de datos no es estacionaria, es decir, tiene una raíz unitaria y no tiene una varianza constante durante el tiempo.
- **H_A:** La serie de datos es estacionaria, es decir, no tiene raíz unitaria.

Tras realizar el test el valor obtenido es $p - valor = 7.149652647 \times 10^{-3}$, lo cual indica que es menor que el nivel de significancia ($\alpha = 0.05$) y se puede rechazar la hipótesis nula concluyendo que los datos son estacionarios por lo tanto no es necesario aplicar diferenciación.

$$p - valor < \alpha$$

Haciendo uso de la librería *"statsmodels.tsa.seasonal.seasonal_decompose"* **¡Error! No se encuentra el origen de la referencia.** se pueden obtener gráficas sobre la estacionalidad y los residuos del conjunto de datos que resultan de gran ayuda para visualizar esas características.

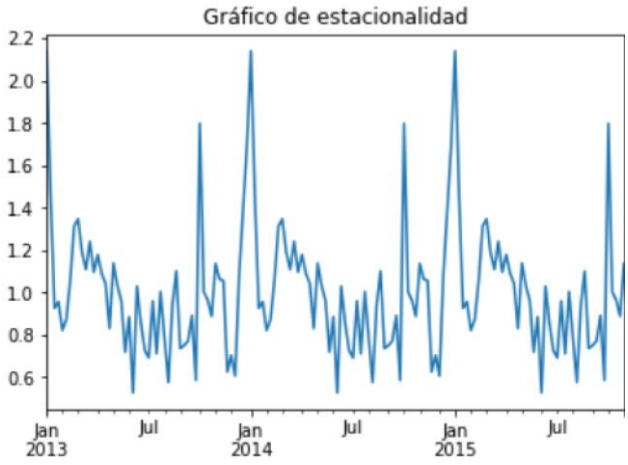


Ilustración 28: Gráfico de estacionalidad.

Resulta evidente que los datos recogidos son exactamente iguales para los años 2013, 2014 y 2015, lo cual confirma la estacionalidad de estos.

4.3 Validación de modelo

Una vez entrenados los datos de los años 2013 y 2014 y realizadas las pruebas con los diferentes parámetros que ofrece el modelo de ARIMA en Python para el año 2015, se han recogido en una tabla los valores obtenidos de diferentes medidas que se utilizan para la evaluación de los resultados de modelos estadísticos.

Cabe destacar que ARIMA ofrece numerosas medidas para verificar si los resultados obtenidos son precisos como HQIC, Ljung-Box o Jarque-Bera entre otras, pero se ha decidido centrarse en el AIC [17], BIC [18]y Mean Squared Error [19] al ser las más utilizadas e identificables en el campo de la estadística:

- AIC: El Criterio de Información de Akaike ofrece un resultado para comprobar cuál de los modelos es mejor en situaciones en las que no se puede comprobar la predicción con valores reales basándose en el ajuste y complejidad del modelo. Cuanto menor es el resultado, mejor es el modelo. La fórmula es la siguiente:

$$AIC = 2k - 2\ln(L)$$

  Donde *k* es el número de parámetros y *L* es el máximo valor para la función de máxima verosimilitud, que indica la calidad del ajuste de los datos.

- BIC: El Criterio de Información Bayesiano es similar al AIC, ya que también selecciona el mejor de los modelos de un mismo conjunto de datos, pero con una penalización para aquellos modelos más complejos, es decir, favorece los modelos simples. Se calcula de la siguiente manera:

$$BIC = -2\ln(L) + kln(n)$$

  Donde *k* es el número de parámetros, *L* el máximo valor para la función de máxima verosimilitud y *n* el tamaño de la muestra.

- MSE: El Error cuadrático medio mide la proximidad de un conjunto de puntos de datos a una línea de regresión como se muestra en la ilustración. Un mayor MSE indica que los puntos están dispersos alrededor de su

media o línea de regresión, prefiriendo por tanto el menor valor posible al tener menos errores.

$$MSE = \frac{1}{n}\, x \sum (y_i - \hat{y}_i)^2$$

Donde *n* es el número total de observaciones, y_i son los valores reales y $\hat{y}_i$ son los valores estimados.

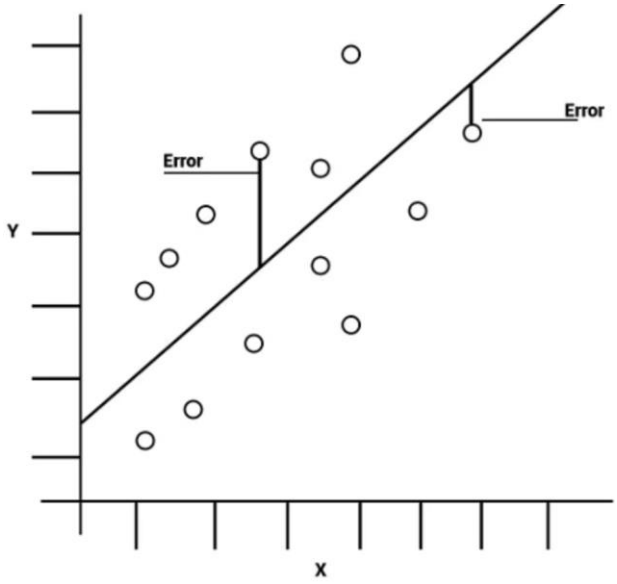


Ilustración 29: Gráfica de línea de regresión.

La tabla obtenida es la siguiente:

Tabla 2: Evaluación de resultados

PARÁMETROS	AIC	BIC	MSE	TYPE DATA
order(1,1,0) seasonal(1,0,0,52)	889.562	897.466	248.197117621843	Normal
order(2,1,1) seasonal(1,0,0,52)	893.136	906.310	248.440281304563	Normal
order(1,1,1) seasonal(1,0,0,52)	891.548	902.087	248.540270626758	Normal
order(2,1,0) seasonal(1,0,0,52)	891.545	902.084	248,642411242033	Normal

order(1,0,0) seasonal(1,0,0,52)	913.090	923.667	392,573726879984	Normal
order(1,0,1) seasonal(1,0,0,52)	906.310	919.532	461,872415073689	Normal
order(1,0,0) seasonal(0,1,0,52)	475.570	479.472	466,557118095753	Normal
order(1,0,0) seasonal(1,0,1,52)	915.067	928.289	495,332377821255	Normal
order(1,0,1) seasonal(1,1,0,52)	471.129	478.934	501,923372587996	Normal
order(1,0,1) seasonal(1,1,1,52)	473.129	482.886	504,692336278177	Normal
order(1,0,0) seasonal(1,1,0,52)	477.490	483.344	516,202963827844	Normal
order(1,0,0) seasonal(1,1,1,52)	479.490	487.295	534,745489548561	Normal

Los parámetros de order y seasonal especifican: ARIMA (order(p,d,q), seasonal(P,D,Q,S)) como se había explicado anteriormente en el [Estado del Arte]. Como se puede apreciar en la tabla 2, que recoge los diferentes valores de medición que se han utilizado para cada prueba realizada, era bastante evidente que el parámetro "p" debía ser igual a 1 en base a los gráficos de autocorrelación y autocorrelación parcial obtenidos y analizados en el punto anterior. Sin embargo, "d" y "q" no eran tan evidentes por lo que se realizaron diferentes pruebas con los valores 0 y 1 para utilizar el que mejores resultados de validación proporcionase. Todas las pruebas tienen el parámetro "S" establecido a 52 porque cada año cuenta con 52 semanas y los datos se han agrupado semanalmente.

La tabla se ha ordenado de mejores a peores resultados tomando como referencia para ello el MSE, ya que cuanto menor sea el valor del error cuadrático medio, mejor ajuste del modelo supondrá, teniendo en cuenta con una menor

ponderación los parámetros de AIC y BIC. Por ello se seleccionó el modelo de ARIMA (order (1,1,0), seasonal (1,0,0,52)) obteniendo la siguiente gráfica:

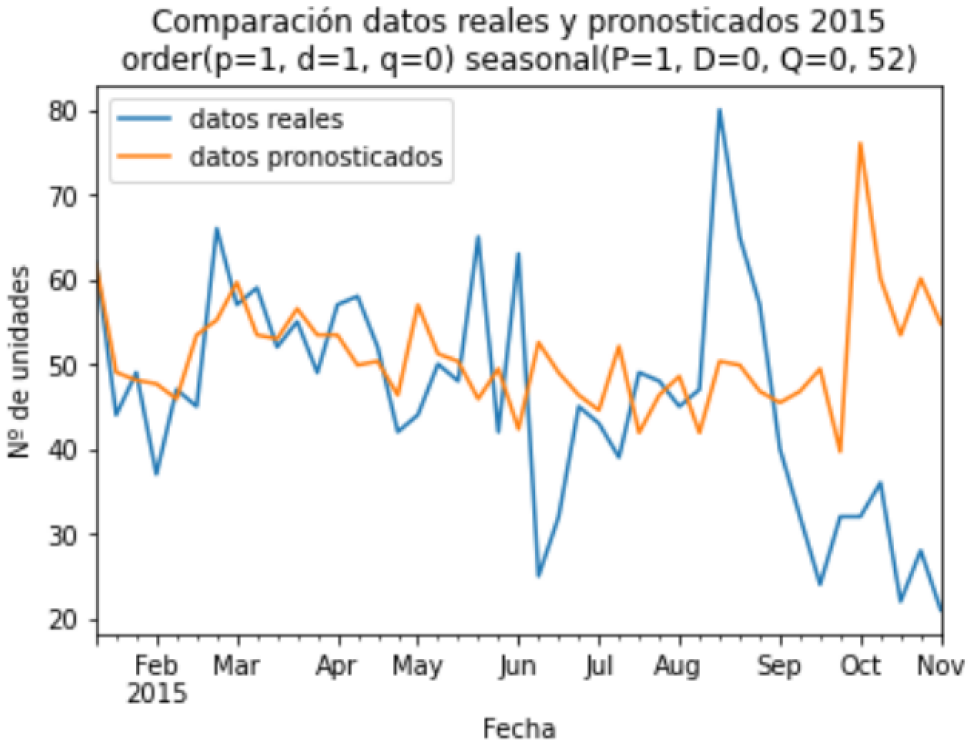


Ilustración 30: Comparación datos reales y pronosticados en 2015.

Como se puede observar en el gráfico, los datos pronosticados entre los meses de enero y agosto resultan bastante fiables mientras que, de agosto hasta noviembre, que es hasta el mes del que se disponen datos no resultan tan precisos, pero es la mejor aproximación que se ha podido obtener.

El siguiente objetivo es utilizar una serie temporal completa una vez conocidos los parámetros que mejores resultados ofrecen, es decir, desde el año 2013 hasta el año 2015 incluido, para pronosticar el número de unidades vendidas de lo que queda de 2015 y de todo 2016, del que no se disponen ningún tipo de dato.

Si se aplican los mismos parámetros de ARIMA (order (1,1,0) seasonal (1,0,0,52), pero utilizando como datos de entrenamiento todos los datos disponibles, el gráfico resultante es el siguiente:

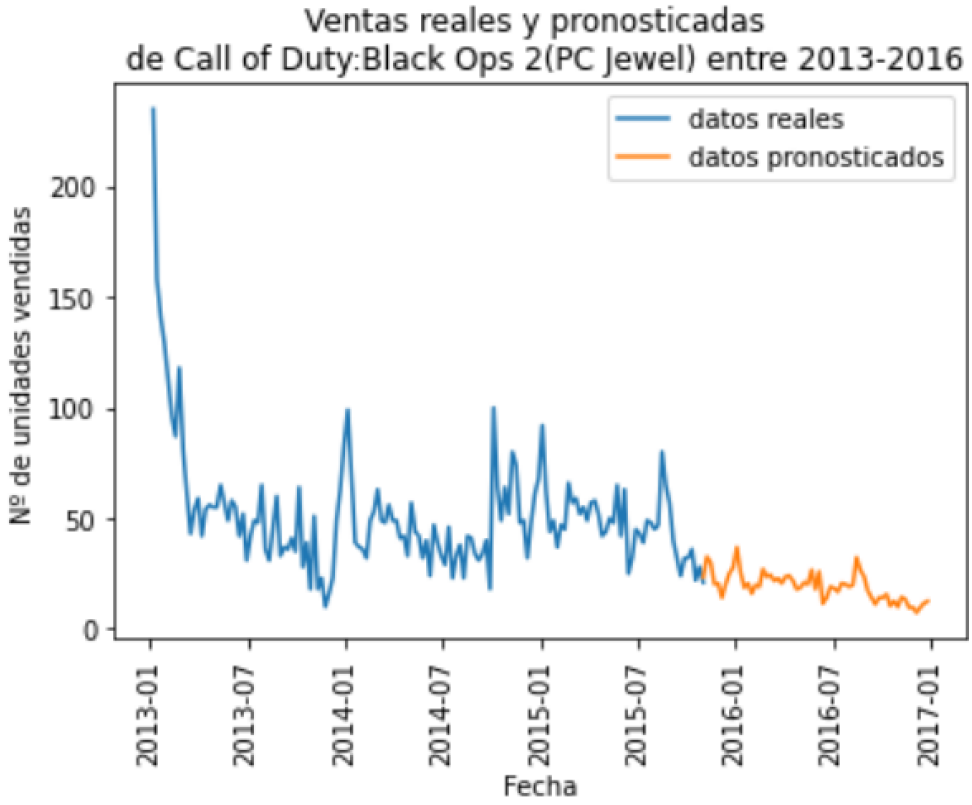


Ilustración 31: Datos reales y pronosticados.

Analizando el gráfico, se puede observar que la tendencia decreciente de las ventas se mantiene para el año pronosticado representado en color naranja. Además, se pueden apreciar los mismos picos de ventas en la primera semana del mes de enero y la segunda semana del mes de agosto. También se debe remarcar que entre febrero y julio se mantienen estables, como en los tres años anteriores al igual que entre agosto y diciembre, por lo que se puede concluir que esta predicción es confiable.

Este estudio permite ayudar a la empresa a planificar su inventario, solicitar financiamiento o realizar inversiones. También influye en la planificación de recursos humanos o el mantenimiento de los equipos. En la gestión de empresas es básico conocer el inventario del que se debe disponer, ya que si se sobreestima la demanda y se cuenta con demasiados productos habrá un exceso de inventario y, por el contrario, si se subestima no se cubrirán las necesidades de los clientes y se beneficiará a la competencia.

Asimismo, influye directamente en sectores como el de marketing, pudiendo orientar sus acciones para aquellos productos que muestren un mejor desempeño y por tanto trabajando los diferentes departamentos de la empresa en un mismo objetivo. [20]

Por último, al anticipar el comportamiento del consumidor y gestionar los recursos de manera inteligente, puede ayudar a fijar precios de productos. Por ejemplo, si se espera una alta demanda de un producto, se pueden establecer precios más altos para maximizar los ingresos. [21]

4.4 Diseño de la interfaz de usuario

En este apartado se mostrarán las diferentes pestañas de las que dispone el informe realizado en Power BI, así como una explicación del tipo de gráfico utilizado para representar la información de la manera más conveniente. [22]

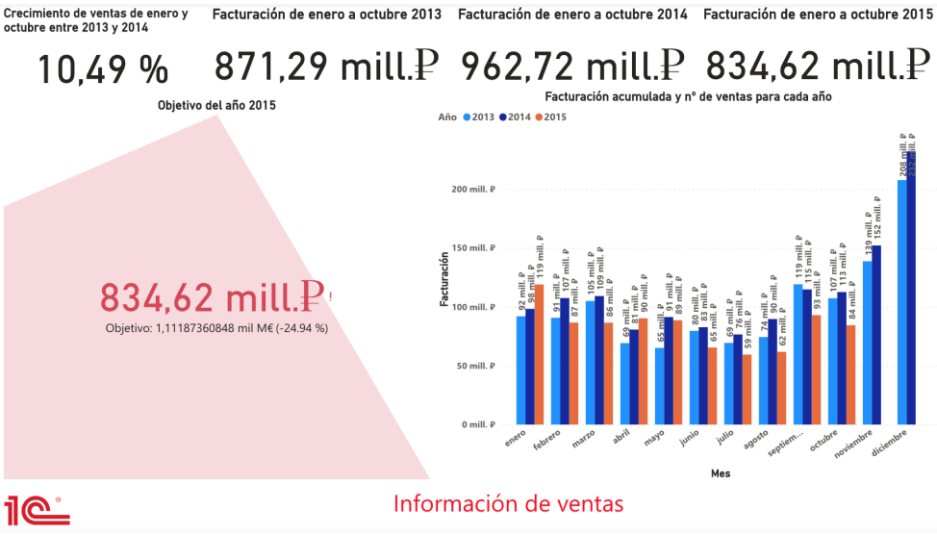


Ilustración 32: Pestaña 1: "Ventas".

En la primera pestaña, se ha realizado una descripción general sobre las ventas de la compañía sin entrar en categorías de productos, países o tipo de tienda lo cual se individualizará en las siguientes pestañas. Se ha decidido mostrar las ventas desde enero a octubre de los años 2013, 2014 y 2015, ya que son los meses de los

que se disponen datos en los tres años y si se realiza una comparativa de todos los meses, los datos de ventas se pueden malinterpretar, ya que en el último año no se conocen las ventas durante noviembre y diciembre provocando que haya menor facturación que los años anteriores.

Además, se han empleado una tarjeta para representar el incremento de ventas en porcentaje del año 2014 frente al año 2013 en los meses de enero hasta octubre y un KPI que representa el progreso realizado para lograr un objetivo cuantificable. Este objetivo se ha calculado como:

$$objetivo_{2015} = facturacion_{2014}\ x\ ((\frac{facturacion_{2014}}{facturacion_{2013}}) + 0.05)$$

Esto implica que el objetivo para 2015 se basa en la proporción de crecimiento de los dos años anteriores y se le suma un 5% para reflejar un crecimiento adicional esperado.

Finalmente, se ha optado por un gráfico de barras de la facturación mensual categorizado por años, ya que es la mejor opción para mostrar la evolución en el tiempo y a la vez realizar una comparativa de los datos.

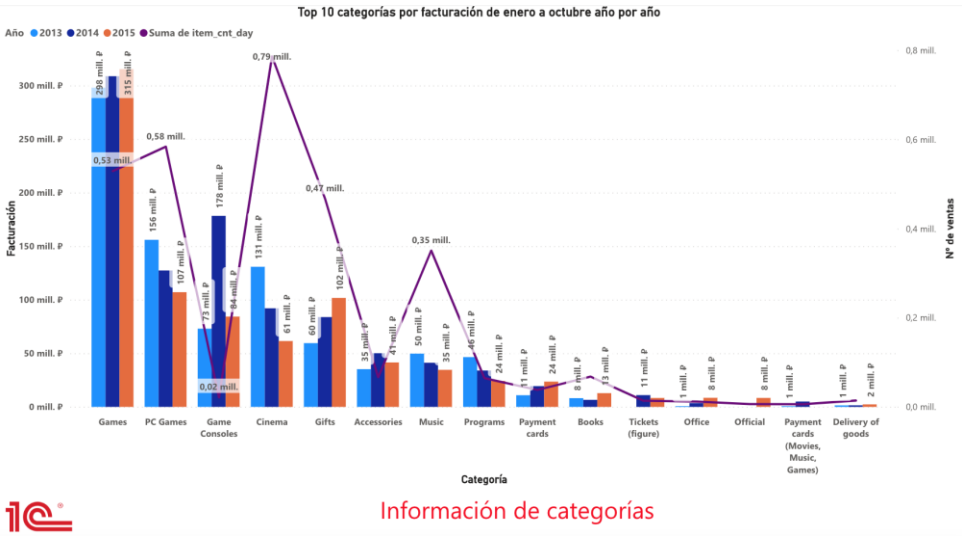


Ilustración 33: Pestaña 2: "Análisis por Categorías".

En esta pestaña, se ha centrado la atención en las categorías de productos existentes en la compañía. El gráfico representa mediante barras la facturación lograda por años junto con las líneas que indican el número de productos que se han vendido, lo que permite visualizar cuales son los precios de cada categoría. Por ejemplo, la categoría cines vende un gran número de artículos, pero no reportan tantos beneficios como las videoconsolas a pesar de vender muchos menos productos.

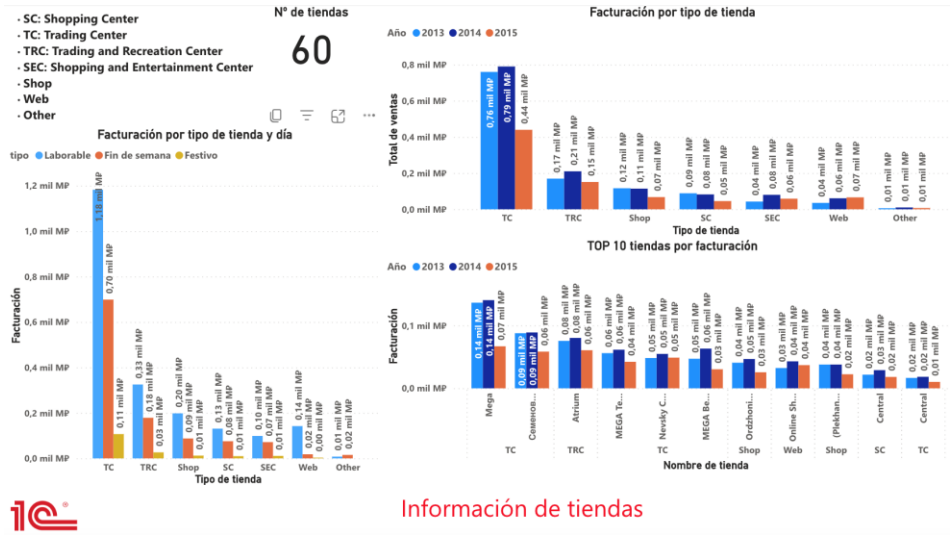


Ilustración 34: Pestaña 3: "Análisis por tiendas".

Dada la imagen, se puede ver que se han empleado una tarjeta y un cuadro de texto para mostrar el número de tiendas y el significado de las siglas de los diferentes establecimientos existentes, respectivamente. Se ha vuelto a optar por los gráficos de barras para representar: La facturación agrupada por años y tipo de tienda ordenadas de mayor a menor facturación, las 10 tiendas con mayor facturación indicando su nombre y tipo correspondiente y la facturación por tipo de tienda en función de si se trata de un día laborable, fin de semana o festivo.

Mediante este análisis, la compañía puede comprender qué tipo de tiendas son las más rentables y si sus clientes prefieren comprar entre semana o durante el fin de semana, ayudando así a calcular la cantidad de personal necesaria para cubrir la afluencia a las tiendas.

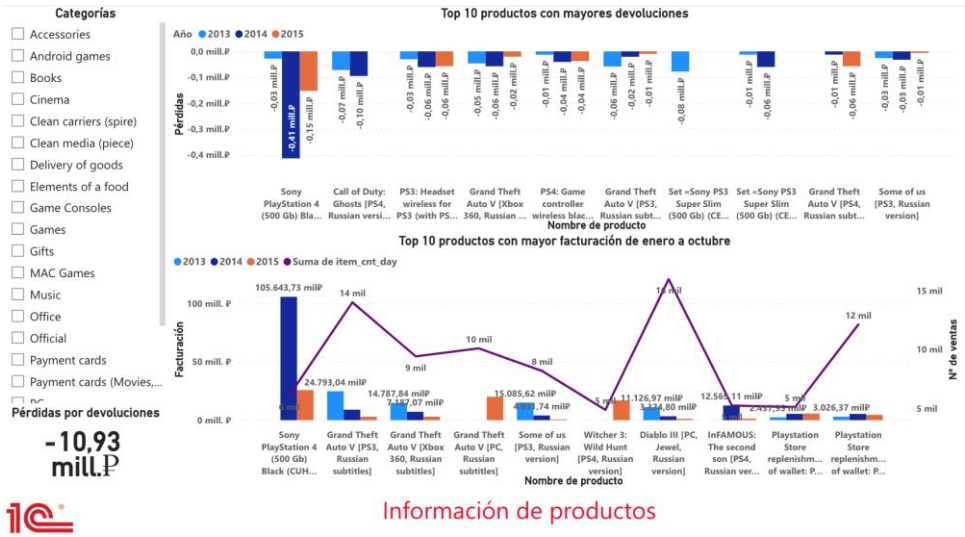


Ilustración 35: Pestaña 4: "Análisis por productos".

Este panel se centra en el comportamiento de los productos. Es importante comprender qué productos reportan los mayores ingresos, así como el número de devoluciones que se pueden producir ya que repercuten en los ingresos de la compañía.

La pérdida de facturación ha sido de 10,93 millones de rublos, lo cual supone una cantidad insignificante para el total de la facturación de la empresa. Mediante la primera gráfica de barras se representa el top 10 de productos con mayores devoluciones agrupado por años. Y mediante la segunda gráfica, se muestran los 10 productos con mayor facturación de enero a octubre utilizando barras, junto con el número de unidades vendidas que se representan con líneas. Además, se ha añadido un filtro en el lado izquierdo del cuadro de mando para poder seleccionar los productos de la categoría de la cual se quiere obtener la información.

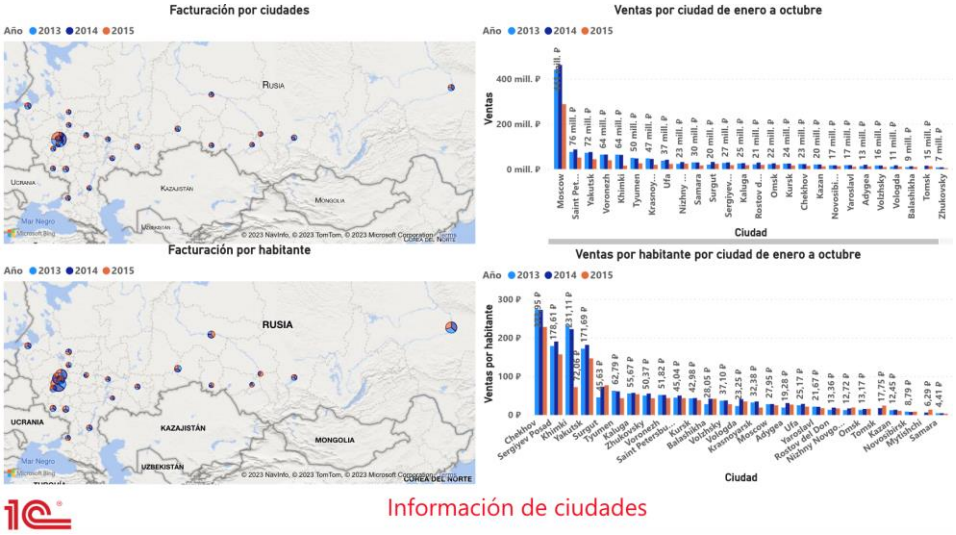


Ilustración 36: Pestaña 5: "Análisis por ciudades".

Esta parte del informe tiene el objetivo de comprender las ventas para cada ciudad en la que existen tiendas de la compañía. Como se ha comentado anteriormente, la empresa cuenta con 60 tiendas repartidas en 29 ciudades de Rusia. Se ha decidido utilizar dos mapas y dos gráficas de barras para mostrar la diferencia existente entre la facturación que hay por cada ciudad y la facturación promedio por habitante en cada ciudad. Gracias al tamaño de las burbujas en el mapa, y a la respectiva gráfica, se puede entender que Moscú tiene un gran volumen de ventas, pero una gran cantidad de habitantes, por lo que al realizar la división la facturación media por habitante es más baja.

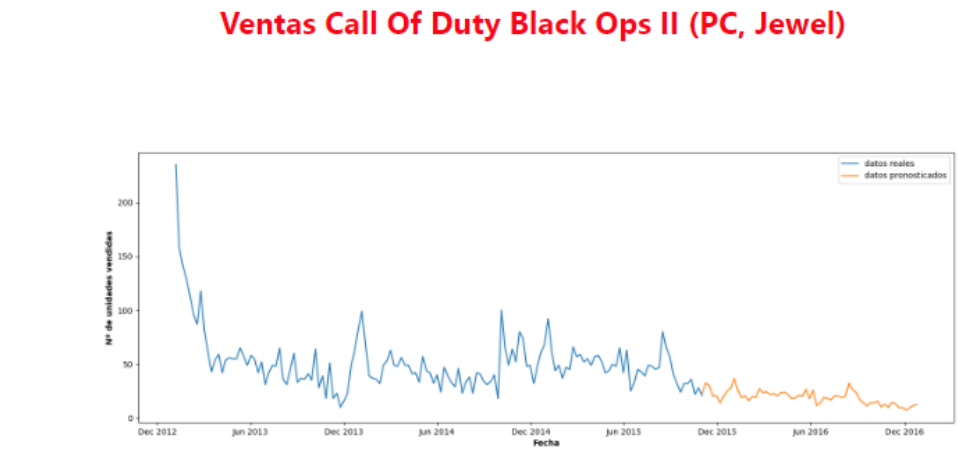


Ilustración 37: Pestaña 6: "Estimación de ventas".

Por último, se ha añadido el gráfico de las ventas entre 2013 y 2016 tanto reales como pronosticadas, del script de Python realizado para pronosticar las ventas del producto al informe de Power BI para así ofrecer una solución con mayor nivel de integración. Para poder hacerlo, se ha seleccionado el objeto visual de Python y se ha introducido el código correspondiente para obtener la gráfica. El mayor cambio ha sido el introducir directamente los datos tras realizar las uniones como aparecen en la Ilustración 24, ya que la tabla "*sales_train*" cuenta con más de dos millones de filas y este tipo de objeto en Power BI sólo permite operaciones con tablas de máximo 150.000 filas.

4.5 Entorno de construcción

Se han valorado dos posibles opciones a la hora de la elección del entorno integrado de desarrollo (Integrated Development Environment, IDE) de acuerdo con las características y capacidades de ciencia de datos que se van a emplear. Las posibilidades que se barajaron fueron; Jupyter Notebook, el cual es gratuito, de código abierto y no consume recursos locales, y Spyder, que goza de una gran integración con numerosas librerías del ámbito de las ciencias como "Pandas", "Numpy" o "matplotlib" las cuales serán explicadas a continuación [23]. Ambas

son muy recomendables para aplicaciones de Machine Learning y data science pero se ha decidido utilizar Spyder ya que se ha utilizado previamente y se conoce su funcionamiento.

4.5.1 Librerías

4.5.1.1 Matplotlib

Es la librería científica principal para representación de datos en Python. Incorpora funciones para la realización de gráficos de dos dimensiones de líneas, de barras o histogramas entre otros. La visualización de los datos y otros aspectos del análisis puede aportar información relevante. Por ejemplo, gracias a esta librería, se puede comprobar si los datos utilizados son continuos, estacionarios y estacionales que son cualidades esenciales para poder realizar la predicción correctamente.

4.5.1.2 Pandas

Es una librería para la gestión y análisis de datos. Está construida a partir de una estructura de datos tabular bidimensional y orientada a columnas, que tiene etiquetas tanto para las filas como para las columnas llamada “Dataframe”. Sería el equivalente de una tabla, como si fuera una hoja de cálculo de Excel. Además, proporciona una gran cantidad de métodos para modificar y operar entre tablas, permitiendo operaciones de tipo SQL y “joins” entre tablas. Pandas combina las capacidades de computación de arrays de alto rendimiento de Numpy con las capacidades de manipulación flexible de datos de hojas de cálculo y bases de datos relacionales (como SQL). Es muy útil porque puede leer y procesar datos de muchos tipos de archivos y bases de datos diferentes como archivos Excel, bases de datos SQL o archivos de valores separados por comas (CSV), lo que permite cargas los datos desde diferentes fuentes, trabajar con ellos en un formato uniforme y remodelar, cortar y seleccionar subconjuntos de datos gracias a la funcionalidad de indexación sofisticada.

4.5.1.3 Numpy

Numpy, abreviatura de Numerical Python, es el paquete fundamental para la computación científica en Python. Entre todas las posibilidades que ofrece, se va a emplear para operaciones de álgebra lineal, así como para realizar cálculos de elementos en matrices u operaciones matemáticas entre matrices.

4.5.1.4 Scikit-learn

Proporciona funciones para evaluar la calidad de los modelos de aprendizaje automático, permitiendo comparar las predicciones de un modelo, en este caso ARIMA, con los datos verdaderos. Incluye variedad de métricas como la precisión, el error cuadrático medio o el coeficiente de determinación para evaluar la precisión del modelo de regresión. También permite realizar evaluación cruzada para garantizar que el modelo no está sobre ajustado.

4.5.1.5 Statsmodels

Proporciona una amplia variedad de métodos estadísticos. Está diseñada para realizar análisis estadísticos avanzados como regresiones, series de tiempo y modelos de datos de panel, así como numerosas herramientas útiles para el modelado predictivo como los algoritmos ARIMA o SARIMAX y medidas de asociación entre valores de series actuales y pasadas como son las funciones de correlación y autocorrelación.

4.5.1.6 sklearn

Es otra biblioteca de aprendizaje automático, útil para la realización de análisis predictivos y modelado de datos que ayuda en tareas como el preprocesamiento de datos, evaluación de modelos y selección de modelos. Estas dos últimas características van a resultar de gran ayuda para valorar la precisión de la predicción respecto con los datos reales.

Capítulo 5 Construcción

5.1 Referencia al repositorio de software

El script de Python en el que se ha realizado el análisis de los datos y la predicción utilizando el modelo ARIMA se encuentra en el siguiente repositorio de github: https://github.com/Butusen/TFG_Jose_Moran

5.2 Manuales

5.2.1 Power BI

Para poder acceder al informe de Power BI se podrá elegir una de las siguientes tres opciones:

- La más sencilla y por tanto más recomendable es abrir el informe en la web de Power BI, ya que el informe puede ser visualizado por todos los componentes de la organización, es decir el CEU, tras haber solicitado permiso para verlo y ser aceptado por parte del alumno, a través del siguiente enlace: Microsoft Power BI

  El problema es que el cuadro de "Estimación de ventas" no puede visualizarse en algunos navegadores por incompatibilidad.
- Otra forma es descargando la aplicación de Power BI Desktop, la cual es gratuita, con el único inconveniente de que solo está disponible para sistemas operativos Windows. Una vez se ha descargado la aplicación, sólo habrá que cargar el informe que estará disponible en el repositorio.
- Por último, se puede acceder al informe desde la aplicación de móvil, disponible tanto en IOS como Google store, registrándose con la cuenta de Microsoft asociada al CEU y en la parte inferior derecha, en el icono "Más",

seleccionar “Compartido conmigo”, tras haber solicitado el acceso desde la web.

5.2.2 Python

Para poder ejecutar el script de Python de manera correcta y visualizar las gráficas creadas, será imprescindible que la máquina en la que se vaya a probar tenga instalado el lenguaje de programación python3, el cual se puede descargar de forma gratuita tanto para sistemas operativos Linux como Windows. La versión utilizada de Python para el proyecto es la 3.9.13 (Python Release Python 3.9.13 | Python.org) y de “pip” ,que es lo que se utiliza para descargar librerías, la 23.1.2.

Una vez descargado e instalado, se deberá instalar en el entorno virtual las librerías que se han utilizado para el desarrollo con el siguiente comando: *“pip install numpy==1.24.3”*. Las librerías son, numpy (1.24.3), pandas (1.5.3), matplotlib (3.7.1), statsmodels (0.13.5), pmdarima (2.0.3) y scikit-learn (1.2.2). Los números entre paréntesis representan la versión que se ha utilizado, al descargar las librerías se deberá utilizar la misma versión para que no haya ningún problema por incompatibilidad de versiones.

Lo más recomendable es utilizar un IDE para facilitar la ejecución y visualización de los datos, el elegido ha sido Spyder, que se puede desplegar desde Anaconda (Files :: Anaconda.org), pero Pycharm, entre otros, también puede ser una buena opción.

Será importante especificar el “path” o ubicación de la carpeta en la que se encuentran los archivos que han sido descargados del repositorio para que funcione correctamente. Por ejemplo, si uno de los archivos que se desea leer tiene la siguiente ruta “~/Desktop/CEU/sales-train.csv”, la ubicación que se deberá detallar será “~/Desktop/CEU”.

En cuanto a Power BI, una vez descargado, habrá que elegir un directorio raíz para Python que deberá ser la ruta en la que se haya descargado e instalado Python.

Para ello, una ver abierta la aplicación de Power BI Desktop, seleccionar: archivo > Opciones y configuración > Opciones > Creación de script de Python y en la ventana emergente “habilitar objetos visuales de script” que aparece al abrir la aplicación, seleccionar habilitar. Tras haber realizado estas configuraciones, se podrá abrir el fichero denominado “tfg.pbix” ubicado en el repositorio para ver todos los informes.

Capítulo 6

Conclusiones y líneas futuras

Durante la realización del proyecto se han extraído una serie de conclusiones, las cuales se van a exponer y detallar con el objetivo de proporcionar una visión completa y precisa de los resultados obtenidos.

En primer lugar, se concluye que el uso de herramientas de inteligencia empresarial y de análisis de datos como Power BI facilitan el procesamiento, visualización y comprensión de la información relacionada con las ventas de la empresa. Gracias a los gráficos interactivos generados, se han podido identificar patrones y relaciones clave que, sin estas herramientas, habrían resultado más complicados de extraer.

Asimismo, el pronóstico de ventas mediante el modelo ARIMA ha demostrado ser un algoritmo efectivo y preciso para predecir el comportamiento futuro de las ventas. La capacidad de modelar las series temporales de ventas brinda a la empresa información valiosa para la toma de decisiones estratégicas, como la planificación de inventario o la fijación de precios de productos para cubrir los costes de producción, permitiendo tomar decisiones más rápidas, precisas y efectivas para los clientes.

Además, queda comprobado que al disponer de una base de datos con los suficientes datos históricos se han podido identificar patrones estacionales y eventos inusuales que han influenciado en el comportamiento de las ventas.

El uso de conocimientos técnicos de programación y negocios, suponen una combinación muy buena ya que el estudio de los datos no se acota a lo conocido, sino que se puede llegar a predecir las tendencias futuras de la demanda.

En resumen, las conclusiones recogidas, confirman la importancia de utilizar herramientas de análisis de datos como Power BI y modelos estadísticos para la predicción de ventas, respaldando la idea de que una gestión basada en los datos puede suponer una ventaja competitiva frente al resto de empresas y contribuir al crecimiento de la empresa en el mercado.

En el futuro, convendría:

- Ampliar el conjunto de productos seleccionados para el pronóstico, pudiendo proporcionar así una visión más completa y precisa de las ventas en general. Al incluir más artículos que cumplan los requisitos del algoritmo ARIMA, se podría mejorar la capacidad del modelo para realizar pronósticos más precisos y capturar patrones de demanda más diversos.
- Relacionado con lo anterior, en lugar de realizar una única predicción global de ventas, desglosar los datos en diferentes segmentos, como categoría o ubicación, y aplicar el modelo ARIMA específico a cada uno.
- Realizar un análisis comparativo entre diferentes modelos de pronóstico, considerando las mismas métricas de evaluación utilizadas para validar el modelo escogido. Esto permitirá seleccionar el modelo más adecuado para las necesidades específicas de pronóstico y maximizar la precisión en las estimaciones futuras.
- Explorar la posibilidad de emplear otros modelos que consideren múltiples factores que puedan repercutir en las ventas, como la existencia de descuentos, la presencia de fines de semana o de días especiales como el Black Friday. Al ser más complejos, se debería valorar si sus resultados son mucho mejores y compensa utilizarlo frente al modelo ARIMA.

Glosario de términos

Las siglas y acrónimos utilizados en esta memoria se hallan listados a continuación, para facilitar al lector su entendimiento en caso de duda.

- **CSV:** Comma-separated values.
- **BI:** Business Intelligence
- **ARIMA:** Autoregressive Integrated Moving Average
- **SC:** Shopping Center
- **TC:** Trade Center
- **SEC:** Shopping and Entertainment Center
- **TRC:** Trade and Recreation Center
- **KPI:** Key Performance Indicator
- **IDE:** Integrated Development Environment
- **AIC:** Akaile's Information Criterion
- **BIC:** Bayesian Information Criterion
- **MSE:** Mean Squared Error
- **ACF:** Autocorrelation Function
- **PACF:** Partial Autocorrelation Function
- **HQIC:** Hannan-Quinn Information Criterion
- **MQ:** Magic Quadrant
- **ROI:** Return On Investment

Bibliografía

[1] 2023). Gartner.com. https://www.gartner.com/doc/reprints?id=1-2CF2LJQ8&ct=230130&st=sb

[2] Why Data Science is the most in-demand skill now and how can you prepare for it? (n.d.). Www.worlddatascience.org. https://www.worlddatascience.org/blogs/why-data-science-is-the-most-indemand-skill-now-and-how-can-you-prepare-for-it

[3] Crece un 92% la demanda de expertos en Big Data y Machine Learning en España.(2023).Equipos&Talento. https://www.equiposytalento.com/noticias/2019/01/18/crece-un-92-la-demanda-de-expertos-en-big-data-y-machine-learning-en-espana

[4] Hedgebeth, D. (2007). Data-driven decision making for the enterprise: an overview of business intelligence applications. Vine, 37(4), 414-420.

[5] Ghasemghaei, M. (2019). Does data analytics use improve firm decision making quality? The role of knowledge sharing and data analytics competency. Decision Support Systems, 120, 14-24.

[6] Loshin, D. (2012). Business intelligence: the savvy manager's guide. Massachusetts: Morgan Kaufmann.

[7] (2023).Insider.com. https://i.insider.com/528f73636da811083d546a67?width=800

[8] Djerdjouri, M. (2020). Data and Business Intelligence Systems for Competitive Advantage: prospects, challenges, and real-world applications. Mercados Y Negocios, 41, 5–18. https://www.redalyc.org/articulo.oa?id=571861494009

[9] Tipos de algoritmos de Machine Learning Types of machine learning algorithms | en.proft.me. (2015, December 24). Proft.me. https://en.proft.me/2015/12/24/types-machine-learning-algorithms/

[10] La Universidad CEU San Pablo confía en Microsoft para mejorar la empleabilidad de sus estudiantes y favorecer la investigación universitaria – Centro de noticias. (n.d.). News.microsoft.com. Retrieved Feb 30, 2023, from https://news.microsoft.com/es-es/2022/01/18/la-universidad-ceu-san-pablo-confia-en-microsoft-para-mejorar-la-empleabilidad-de-sus-estudiantes-y-favorecer-la-investigacion-universitaria/

[11] statsmodels.tsa.seasonal.seasonal_decompose - statsmodels 0.15.0 (+8). (n.d.). Www.statsmodels.org. Retrieved May 7, 2023, from

https://www.statsmodels.org/devel/generated/statsmodels.tsa.seasonal.seasonal_decompose.html

[12] Criterios ARIMA de series temporales. (n.d.). Www.ibm.com. Retrieved May 5, 2023, from https://www.ibm.com/docs/es/spss-modeler/saas?topic=SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/timeseries_arima_criteria.htm

[13] Chen, J. (n.d.). Autoregressive Integrated Moving Average (ARIMA). Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical

[14] Metodología de Box-Jenkins. (n.d.). Rstudio-Pubs-Static.s3.Amazonaws.com. Retrieved May 5, 2023, from https://rstudio-pubs-static.s3.amazonaws.com/574656_3daaa786709c47e29f8d63949bff4e3b.html

[15] Monigatti, L. (2022, August 2). Interpreting ACF and PACF Plots for Time Series Forecasting. Medium. https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c

[16] Zach. (2021, May 25). Augmented Dickey-Fuller Test in Python (With Example). Statology. https://www.statology.org/dickey-fuller-test-python/

[17] What Is Akaike Information Criterion (AIC)? | Built In. (n.d.). Builtin.com. https://builtin.com/data-science/what-is-aic

[18] Stephanie. (2018, March 10). Bayesian Information Criterion (BIC) / Schwarz Criterion. Statistics How To. https://www.statisticshowto.com/bayesian-information-criterion/

[19] Mean Squared Error : Overview, Examples, Concepts and More | Simplilearn. (n.d.). Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error#:~:text=The%20Mean%20Squared%20Error%20measures

[20] Oduka, E. (2022, April 6). Pronóstico de ventas: ¿Por qué es importante para las empresas? Oduka. https://oduka.co/pronostico-de-ventas/

[21] Pronóstico de Ventas: ¿Qué es y cómo hacerlo? (n.d.). Salesforce. Retrieved May 30, 2023, from https://www.salesforce.com/mx/blog/2021/07/pronostico-de-ventas.html#answer2

[22] mihart. (2023, March 18). Tipos de visualización en Power BI - Power BI. Learn.microsoft.com. https://learn.microsoft.com/es-es/power-bi/visuals/power-bi-visualization-types-for-reports-and-q-and-a

[23] McKinney, W. (2022). Python for Data Analysis (3a ed.). O'Reilly Media. 4-6