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

ANÁLISIS DE IMPACTO DE LA COVID-19 EN EL MERCADO LABORAL DE ESPAÑA

JAVIER ARGOS GONZÁLEZ

Dirigido por

CARLOS RAMÍREZ LIZÁN

CURSO 2021-2022

TÍTULO: ANÁLISIS DE IMPACTO DE LA COVID-19 EN EL MERCADO LABORAL DE ESPAÑA

AUTOR: JAVIER ARGOS GONZÁLEZ

TITULACIÓN: GRADO EN INGENIERÍA MATEMÁTICA APLICADA AL ANÁLISIS DE DATOS

DIRECTOR/ES DEL PROYECTO: CARLOS RAMIREZ LIZÁN

FECHA: JUNIO DE 2022

RESUMEN

La pandemia de coronavirus en España y en el mundo ha provocado una crisis sanitaria sin precedentes en los últimos cien años en el mundo que, a su vez, ha derivado en la paralización de ciertos sectores económicos, como el comercio, la hostelería y el turismo, provocando la caída de varios indicadores económicos que alertan de una recesión.

El Banco de España estimó que las provincias de Baleares (-27%), Las Palmas y Santa Cruz de Tenerife, junto a Málaga, Gerona y Alicante registraron las mayores caídas del PIB en 2020, que estuvo marcado por la "heterogeneidad" del impacto del COVID, siendo mayor en las zonas con mayor peso del turismo e incidencia en la movilidad. En España se volvieron a registrar altas cifras de parados. El número total de parados se situó en 3.802.814 personas en el mes de agosto de 2020, sin tener en cuenta a las personas que se encuentran en situación de ERTE (Expediente de Regulación Temporal de Empleo), ya que estos se consideran como ocupados.

Este proyecto plantea un análisis de los datos del PIB y de la población activa, comparándolos con los años previos y el año posterior a la llegada del COVID. Para ello se ha utilizado la herramienta Power BI, capaz de procesar grandes volúmenes de datos para crear visualizaciones y realizar analítica de estos. A partir del conocimiento que se obtiene de los datos, se propondrá el desarrollo de un modelo que trate de predecir los resultados de la población activa en España para finales de 2021, comparándolos con los datos reales.

La finalidad de este proyecto es comprender cuáles han sido los sectores más afectados por la pandemia, en qué CCAA se ha notado más y determinar en qué sectores la población activa crece o decrece.

Palabras clave: COVID-19, PIB, análisis de datos, conocimiento estadístico, población activa en España, sectores, CCAA, modelo, regresión lineal, Power BI, Python.

ABSTRACT

The coronavirus pandemic in Spain and in the world has caused an unprecedented health crisis in the last hundred years in the world which, in turn, has led to the paralysis of certain economic sectors, such as commerce, hospitality and tourism, causing the fall of several economic indicators that warn of a recession.

The Bank of Spain estimated that the provinces of the Balearic Islands (-27%), Las Palmas and Santa Cruz de Tenerife, together with Malaga, Gerona and Alicante registered the greatest falls in GDP in 2020, which was marked by the "heterogeneity" of the impact of COVID, being higher in areas with greater weight of tourism and impact on mobility. In Spain, high numbers of unemployed were registered again. The total number of unemployed stood at 3,802,814 people in the month of August 2020, without taking into account the people who are in a situation of ERTE (Temporary Employment Regulation File), since these are considered as employed .

This project proposes an analysis of GDP and active population data, comparing them with the years before and the year after the arrival of COVID. For this, the Power BI tool has been used, a tool capable of processing large volumes of data. Based on the knowledge obtained from the data, the development of a model that tries to predict the results of the active population in Spain by the end of 2021 will be proposed and compared with the real data.

The purpose of this project is to understand which sectors have been most affected by the pandemic, in which Autonomous Communities have been most noticeable, and determine in which sectors the active population is growing or decreasing.

Keywords: COVID-19, GDP, data analysis, statistical knowledge, active population in Spain, sectors, Autonomous Communities, model, linear regression, Power BI, Python.

AGRADECIMIENTOS

En primer lugar, quiero agradecer a mi novia Andrea, gracias a ella ha sido posible llegar hasta aquí y con su apoyo y su paciencia me ha ayudado a conseguir todos los objetivos que nos hemos fijado. También me gustaría agradecer a mi familia, en especial a mis padres por animarme a seguir en los momentos más difíciles.

Finalmente quiero agradecer a todos aquellos que me han animado a que nunca es demasiado tarde para estudiar, a todos los profesores que han ayudado durante estos 3 años a cimentar las bases de mi futura carrera, especialmente a mi tutor Carlos Ramírez por guiarme en este proyecto.

Gracias a todos.

Cita - frase célebre / Dedicatoria

“No es el conocimiento, sino el acto de aprendizaje; y no la posesión, sino el acto de llegar a ella, lo que concede el mayor disfrute.”

.- Carl Friedrich Gauss

TABLA RESUMEN

	DATOS
Nombre y apellidos:	JAVIER ARGOS GONZÁLEZ
Título del proyecto:	ANÁLISIS DE IMPACTO DE LA COVID-19 EN EL MERCADO LABORAL DE ESPAÑA
Directores del proyecto:	CARLOS RAMÍREZ LIZÁN
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Realizar un estudio del impacto laboral de la pandemia en los principales sectores de España.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

España ha sufrido una transformación en el mercado laboral a raíz de la pandemia causada por el COVID-19. Muchas personas estuvieron en situación de ERTE y muchas otras directamente perdieron su trabajo ya que una gran cantidad de empresas no pudieron hacer frente a esta situación y declararon un ERE. El número total de parados se situó en 3.802.814 personas en el mes de agosto de 2020, sin tener en cuenta a las personas que se encontraban en situación de ERTE.

1.2 Planteamiento del problema

Se pretende realizar una investigación y análisis de cómo ha afectado esta situación a los números de población activa en España, realizando una diferenciación tanto a nivel autonómico como por los principales sectores (Servicios, Agricultura, Industria y Construcción). Utilizando herramientas de visualización de datos, técnicas de extracción, limpieza y carga de datos, se llevará a cabo un análisis exploratorio de los mismos, para finalmente realizar un modelo de predicción que sea capaz de estimar los valores de población activa que tendremos en España a finales del año 2021 comparándolos con los datos reales y adicionalmente realizar una predicción con el modelo de cuáles serán los valores absolutos de población activa, tanto a nivel nacional como para cada uno de los principales sectores económicos, a finales del año en curso, 2022 .

1.3 Objetivos del proyecto

El objetivo general es el del desarrollo de una investigación para determinar qué sectores y en qué CCAA ha tenido mayor impacto la pandemia producida por el COVID-19. Este análisis se desarrollará en una herramienta de visualización la cual está diseñada para poder observar estos datos según el rango de edad de la población activa o por la condición de sexo. Finalmente mediante el uso de técnicas de programación y aprendizaje automático se procederá a la creación de un modelo de predicción.

1.4 Resultados obtenidos

La investigación ha alcanzado los objetivos iniciales por lo que se puede llegar a la conclusión de que la información recogida nos ha sido útil para poder realizar el análisis principal del proyecto. El algoritmo de predicción utilizado en nuestro modelo final nos genera unos resultados realistas teniendo en cuenta los datos de años anteriores.

1.5 Estructura de la memoria

La memoria se compone de la siguiente manera:

- Capítulo 1: Resumen del proyecto. Se incluye un breve resumen del proyecto incluyendo el contexto y justificación del problema, así como el planteamiento del mismo. También se explica resumidamente los objetivos, resultados y la estructura de la memoria.
- Capítulo 2: Antecedentes / Estado del Arte. Capítulo dedicado al planteamiento al análisis del estado del arte, explicación del contexto, justificación y planteamiento del problema de manera más extensa que en el capítulo 1.
- Capítulo 3: Objetivos. Se explican los objetivos, tanto generales como específicos, así como los beneficios del problema.
- Capítulo 4: Desarrollo del proyecto. Explicación del análisis y modelo realizados. También se menciona la planificación del proyecto, los recursos que se han requerido para su desarrollo, un presupuesto del mismo y resultados del proyecto.
- Capítulo 5: Discusión. En esta sección se hace una reflexión sobre los resultados principales del proyecto y otros aspectos que han influido en el desarrollo del mismo.
- Capítulo 6: Conclusiones. Se redactan las conclusiones a las que se ha llegado tras la finalización del proyecto.
- Capítulo 7: Futuras líneas de trabajo. Se exponen posibles trabajos a futuro que se podrían implementar a raíz de esta investigación y de los resultados obtenidos.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

En este apartado, se hablará sobre las bases en las que se centra este proyecto de análisis e investigación, así como las metodologías utilizadas y técnicas que se han puesto en marcha para poder desarrollar este trabajo. hablaremos de cómo ha ido cambiando el mercado laboral en España en los últimos años. De la misma manera se realizará una explicación de la herramienta Power BI, sus aplicaciones así como el concepto del Machine Learning en el que se basará el modelo de predicción.

2.1.1 Mercado Laboral en España

Desde la crisis que golpearía a nuestro país en 2008, el mercado laboral ha sufrido una recuperación lenta hasta la llegada de la pandemia de la COVID-19, obligando a detener muchos servicios debido al confinamiento que se llevó a cabo en nuestro país que tuvo comienzo el 16 de Marzo de 2020.

Este cierre tuvo un impacto directo en las empresas que componen el tejido empresarial, afectando directamente a aquellas dedicadas a actividades no esenciales, produciendo una detención de su actividad durante semanas, dejando de percibir ingresos. Ante esta situación, muchas empresas no han podido hacer frente a los gastos y se han visto obligadas a cerrar definitivamente. Otras tantas, como consecuencia de la pandemia, se han visto obligadas a reducir las plantillas y las rentas de los trabajadores para poder hacer frente a los gastos, y en los casos de cierre temporal ha derivado en un expediente de regulación temporal de empleo (ERTE). Según el Ministerio de Trabajo y Economía Social[1] el número de personas afectadas por EREs es de 974.489 frente a los 967.724 trabajadores afectados por los ERTEs. El ministerio recoge los datos de los ERE a nivel nacional por lo que contando con las autoridades provinciales y autonómicas, se estima que la cifra de personas afectadas puede aumentar hasta los 3,4 millones de trabajadores.

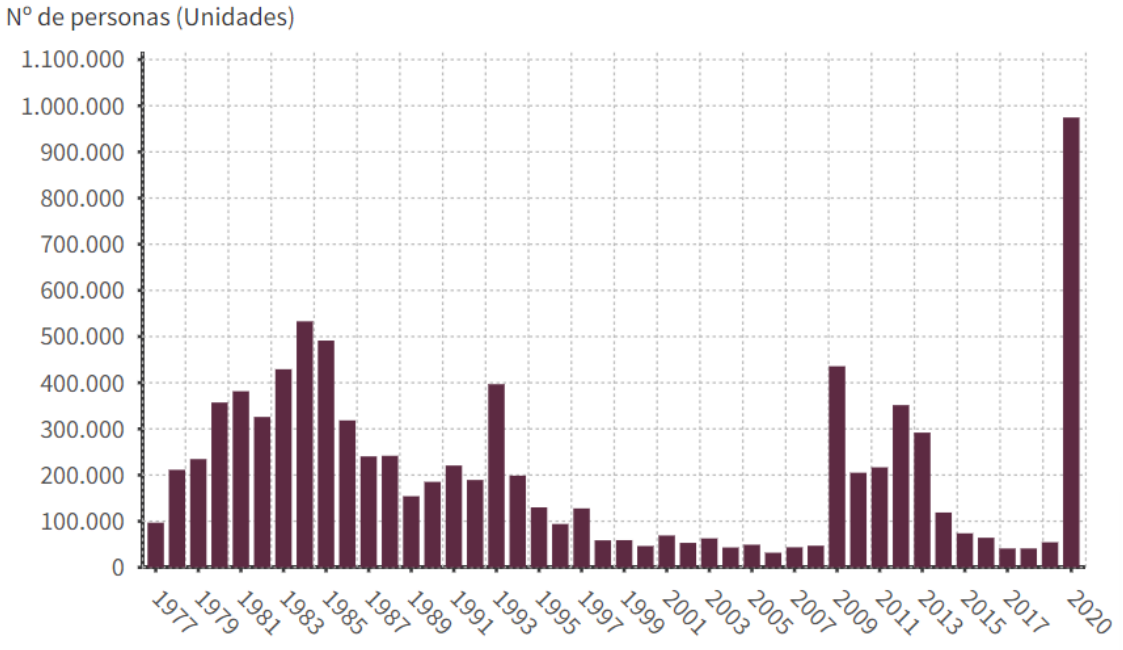

Figura 1. Personas afectadas por un ERE

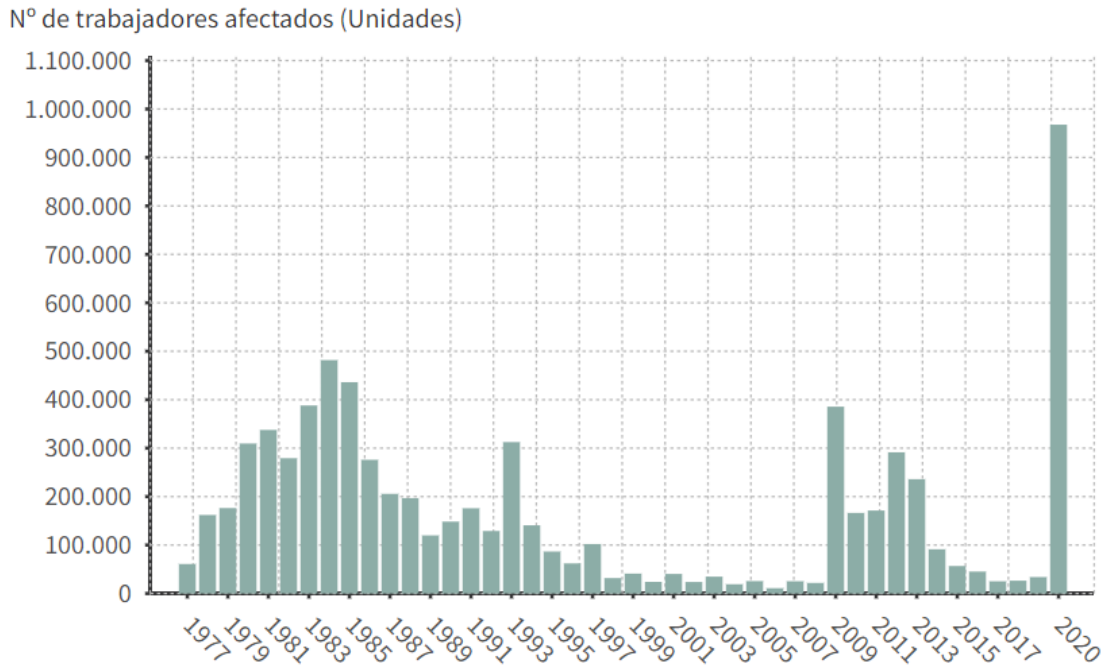

Figura 2. Trabajadores afectados por un ERTE

Como hemos comentado, la crisis producida por la pandemia de la COVID-19 no ha tenido el mismo impacto en todos los países y dentro del nuestro este impacto tampoco ha sido igual según los sectores de actividad. Según informes realizados por el grupo de trabajo mixto, los efectos en los distintos sectores dependen de las características de cada rama productiva, entre otros factores. A pesar de esta situación, en la otra cara de la moneda, se encuentran distintos sectores que han tenido que contratar a más gente para satisfacer la demanda producida, principalmente en el sector sanitario.

En resumen y por las publicaciones realizadas por el Instituto Nacional de Estadística (INE), el año 2020 cerró con unos números muy pesimistas.

- Se destruyeron 622.600 puestos de trabajo.
- Se alcanza la cifra de 3.71 millones de personas en paro.
- El desempleo marca una tasa del 16.13% y sigue recuperándose de la crisis del 2008.
- Los trabajadores en ERTE disminuyen para final de año.

2.1.2 Power BI

Power BI es un servicio gratuito de análisis de datos basado en la nube y visualización. Esta herramienta de Business Intelligence (BI) permite, mediante la creación de dashboards, crear informes y el acceso a datos en tiempo real. Es una herramienta muy útil a la hora de realizar análisis de datos, ya que cuenta con un gran poder de computación y soporta enormes cantidades de datos, permitiendo hacer transformaciones de los datos, consultas y crear datos derivados, siendo muy útil para más tarde poder plasmar esta información en futuros informes. La herramienta permite a los usuarios que la utilizan poder realizar procesos de ETL y de análisis exploratorio de manera más rápida y eficaz que con otras herramientas. Es ampliamente utilizado en agencias de analítica web y empresas especializadas en Business Intelligence.

La gente usa Power BI para tener mayor poder a la hora de generar informes con mayor potencia analítica que la que ofrecen otras herramientas. Los mayores beneficios que ofrece Power BI son:

- Gestión de grandes volúmenes de datos sin dificultad de carga y procesamiento.
- Funciones de aprendizaje automático destinadas al análisis de tendencias y predicciones.
- Numerosas plantillas de gráficos para realizar visualizaciones que faciliten el conocimiento de la información obtenida de los datos.
- Power BI forma parte de Microsoft por lo que tiene la capacidad de integrarse con múltiples herramientas: Azure, Analytics, SQL Server.
- Se garantiza la seguridad de los datos a través de controles de acceso.

Es utilizado principalmente por profesionales de inteligencia empresarial, que lo utilizan para crear modelos de datos e informes.

Power BI está compuesto por aplicaciones, cada una de ellas con sus características propias.

Power Query permite la transformación y optimización de datos por la conexión de distintas fuentes de datos. Power View genera elementos de visualización de datos interactivos como gráficas, mapas u otros elementos visuales para la mejor comprensión de los datos. Power Pivot es una herramienta de modelado de datos. Power Map se usa para crear imágenes en 3D de los datos. Finalmente tenemos Power Q&A, el cual consiste en un motor de preguntas y respuestas para obtener información de los datos utilizando un lenguaje sencillo.

Por todo esto, Power BI es una herramienta idónea para realizar trabajos de investigación y análisis como se viene utilizando en diferentes empresas y sectores, capaz de otorgar conocimiento a través del análisis de los datos e incluso poder realizar predicciones a futuro.

2.2 Contexto y justificación

El mercado laboral español ha sufrido un gran golpe debido a la pandemia producida por el COVID-19 y el confinamiento posterior y cese de la actividad laboral en nuestro país.

Gracias a los datos proporcionados por el INE (recopilados trimestralmente) tenemos registros de cómo ha evolucionado la actividad laboral desde 2008, año en el que estalló la crisis económica. El INE clasifica los principales sectores económicos en los que diferenciar a la población activa/ocupada en cuatro grandes grupos: Agricultura, Industria, Construcción y Servicios.

La crisis sanitaria y económica que estamos atravesando ha tenido impacto directo en el mercado laboral como hemos visto en el punto anterior, por lo que en este proyecto se va a realizar una investigación a gran escala y con profundidad en el detalle de cómo se ha visto afectado el mercado laboral por culpa de esta crisis.

2.3 Planteamiento del problema

El proyecto que se va a desarrollar consistirá en un análisis específico de cómo se ha visto afectado el tejido del mercado laboral de España tanto a nivel nacional como a nivel autonómico, diferenciando a las personas por el sector en el que realicen su profesión, el grupo de edad al que se pertenezca o su género.

Para todo ello va a ser necesario la obtención de los datos por parte del INE y un tratamiento de los mismos realizado íntegramente con la herramienta Power BI, por la cual se va a desarrollar un dashboard que muestre la información solicitada a través de filtros. Como complemento, también se implementa un dashboard adicional en el que ver en qué porcentaje de los sectores diferenciados por el INE contribuyen los perfiles filtrados por periodo al total de personas activas, tanto de la comunidad autónoma o al nacional.

Finalmente se procederá a la creación, con los datos obtenidos, de un modelo que trate de predecir cómo se hubiese recuperado el mercado laboral y sea capaz de realizar una predicción de cuáles serán los valores de personas activas a finales del año 2022 a nivel nacional.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo principal del proyecto es el de realizar un análisis en profundidad del impacto de la crisis producida por la pandemia de la COVID-19 en el mercado laboral español, diferenciando en CCAA, principales sectores y sexo, para determinar qué perfiles de trabajo y donde se hasufrido un mayor impacto. Ligado a esto se creará un modelo que, con los datos analizados, realice una predicción de los valores de población activa para finales del año 2021, evaluando los resultados obtenidos.

3.2 Objetivos específicos

Los objetivos secundarios van a ir ligados a los pasos que se van a seguir para llevar a cabo el proyecto. Estos van a ser los siguientes:

➔ Investigación y comprensión del campo en el cual se va a desarrollar el proyecto.

➔ Búsqueda y obtención de datos. Para un buen desarrollo de nuestra actividad estos datos deben ser íntegros y consistentes con la realidad, así como verificar que estos datos son ciertos.

➔ Realizar un correcto proceso de limpieza y unificación de los datos para la conjunción de un dataset que tenga toda la información.

➔ Llevar a cabo un correcto EDA para extraer conocimiento de los datos que hemos obtenido.

➔ Creación de un modelo capaz de predecir el valor de ocupados en los sectores económicos a finales de año.

➔ Optimización del modelo para que este sea capaz de predecir con mayor precisión.

Finalmente y una vez realizado el proyecto, se podrán definir futuros objetivos dirigidos a la mejora del modelo y a nuevas implementaciones que pueda tener.

3.3 Beneficios del proyecto

El problema del mercado laboral en España, como he mencionado con anterioridad, es un tema que viene de lejos, concretamente desde la crisis económica del 2008. La pandemia de la COVID-19 ha irrumpido drásticamente en el mercado laboral, frenando esta recuperación que venía dándose desde el 2013. Gracias a esta investigación se podrá determinar en qué sectores y comunidades el impacto ha sido mayor y que tipo de perfil de trabajador se ha visto más perjudicado.

Este análisis se ha realizado en PowerBI y cuenta con unos filtros muy intuitivos a través de los cuales cualquier usuario puede consultar y ver como ha sido el impacto, utilizando los datos oficiales del instituto Nacional de Estadística. De este modo y gracias a las visualizaciones llevadas a cabo se muestra de manera más visual cómo ha evolucionado el mercado laboral en el año 2020.

Finalmente, por las previsiones del modelo llevado a cabo, se puede comparar como ha sido la recuperación en el año 2021 frente a la recuperación predicha por el modelo y cuál sería la predicción de cara al final del año 2022.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

La duración del proyecto elaborado ha estado alrededor de los 4 meses, desde su inicio en el mes de enero. Las fases del proyecto se desglosan en 4 sprints.

Sprint 1 Planificación del proyecto

En este sprint se definen los objetivos del proyecto a través de una investigación del campo de estudio en cuestión. Se trata de estudiar el objetivo principal del proyecto y cuál será la finalidad del mismo. De igual manera se debe realizar un análisis de los requisitos necesarios que va a demandar nuestro proyecto, qué herramientas se van a utilizar y en qué lenguaje se va a desarrollar el modelo.

Finalmente, para terminar el sprint de la planificación, vamos a investigar en el INE las fuentes de datos que nos ofrece y con los objetivos definidos se decide cuales son los datos más relevantes a la hora de desarrollar el análisis del proyecto.

Sprint 2 ETL y EDA

En este sprint se descargan los datos definidos en el primer sprint sobre los que se va a cimentar el estudio del proyecto, así como la creación del modelo. Una vez realizada la extracción de los datos, se ha de determinar qué tipos de datos tenemos y realizar las transformaciones necesarias que se llevarán a cabo para subir finalmente estos datos a nuestra herramienta, cerrando así la etapa de ETL.

Finalmente, para cerrar el sprint, se implementarán una serie de dashboards capaces de mostrar mediante gráficas la información extraída de los datos. Estos paneles de control contarán con filtros útiles, capaces de informar de distintos perfiles de trabajadores, así como de distintas identidades en distintos periodos.

Sprint 3 Creación del modelo

Una vez extraído el conocimiento de los datos se propone el desarrollo de un modelo capaz de realizar una predicción de los datos en distintos periodos de tiempo. En el sprint se realizará un análisis de cuál es el mejor algoritmo que se ajuste a nuestros datos, con la finalidad de realizar predicciones de datos de actividad laboral a finales de año.

Se determinan los datos sobre los que se realizará la predicción y las librerías necesarias, el lenguaje y entorno en el cual se va a trabajar con el modelo. Para cerrar el sprint se han de llevar a cabo una serie de pruebas, evaluaciones y discusiones acerca de cuál es el algoritmo que mejores resultados ha proporcionado en el modelo final.

Sprint 4 Mejoras y documentación

Este es un sprint que abarca toda la duración del proyecto, en el que se trata de que a cada avance de la realización de distintas fases del trabajo, se realice una tarea de propuestas de mejora, la viabilidad a la hora de implementarlas y la decisión final.

De la misma manera se debe ir documentando todas las acciones realizadas a lo largo del proyecto para dejar constancia de los pasos llevados a cabo.

Planificación en MS Project

EDT	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
1	**Proyecto TFG**	80 días	lun 31/01/22	vie 20/05/22	
1.1	**Sprint 1**	20 días	lun 31/01/22	vie 25/02/22	
1.1.1	Investigación del campo de estudio	10 días	lun 31/01/22	vie 11/02/22	
1.1.2	Obtención de datos	10 días	lun 14/02/22	vie 25/02/22	3
1.2	**Sprint 2**	25 días	lun 28/02/22	vie 01/04/22	
1.2.1	ETL	15 días	lun 28/02/22	vie 18/03/22	
1.2.1.1	Extracción de datos	5 días	lun 28/02/22	vie 04/03/22	4
1.2.1.2	Transformación de datos	5 días	lun 07/03/22	vie 11/03/22	7
1.2.1.3	Limpieza de datos	5 días	lun 14/03/22	vie 18/03/22	8
1.2.2	Análisis exploratorio	10 días	lun 21/03/22	vie 01/04/22	9
1.3	**Sprint 3**	20 días	lun 04/04/22	vie 29/04/22	
1.3.1	Extracción de conocimiento	5 días	lun 04/04/22	vie 08/04/22	10
1.3.2	Primera versión del modelo	15 días	lun 11/04/22	vie 29/04/22	12
1.4	**Sprint 4**	15 días	lun 02/05/22	vie 20/05/22	

1.4.1	Optimización	15 días	lun 02/05/22	vie 20/05/22	13
1.4.2	Entrega TFG	0 días	vie 20/05/22	vie 20/05/22	15;17
1.5	**Documentación Memoria Final**	20 días	lun 25/04/22	vie 20/05/22	

Tabla 1. Actividades del proyecto

4.2 Descripción de la solución, metodologías y herramientas empleadas

Una vez presentada la planificación del proyecto y las etapas que lo componen, se explicará los datos que se han obtenido, las tecnologías y herramientas utilizadas y seguidamente la metodología aplicada durante el proceso.

4.2.1 Descripción de los datos

Las variables utilizadas para el análisis de los datos provienen del Instituto Nacional de Estadística. Siguiendo la siguiente ruta se llega a los datos que se desean obtener:

INEbase ⟶ Mercado Laboral ⟶ Actividad, ocupación y paro ⟶ Encuesta de población activa ⟶ Series desde el primer trimestre de 2002 ⟶ Resultados por comunidades autónomas.

Seleccionamos todas las variables disponibles para realizar luego el tratamiento de los datos en otras herramientas. El Instituto Nacional de Estadística permite descargarse los datos en el formato deseado. De cara al proyecto se ha descargado un archivo .csv, por lo que no ha sido necesario realizar un scrapping de los datos al ser datos públicos y al alcance de cualquier usuario.

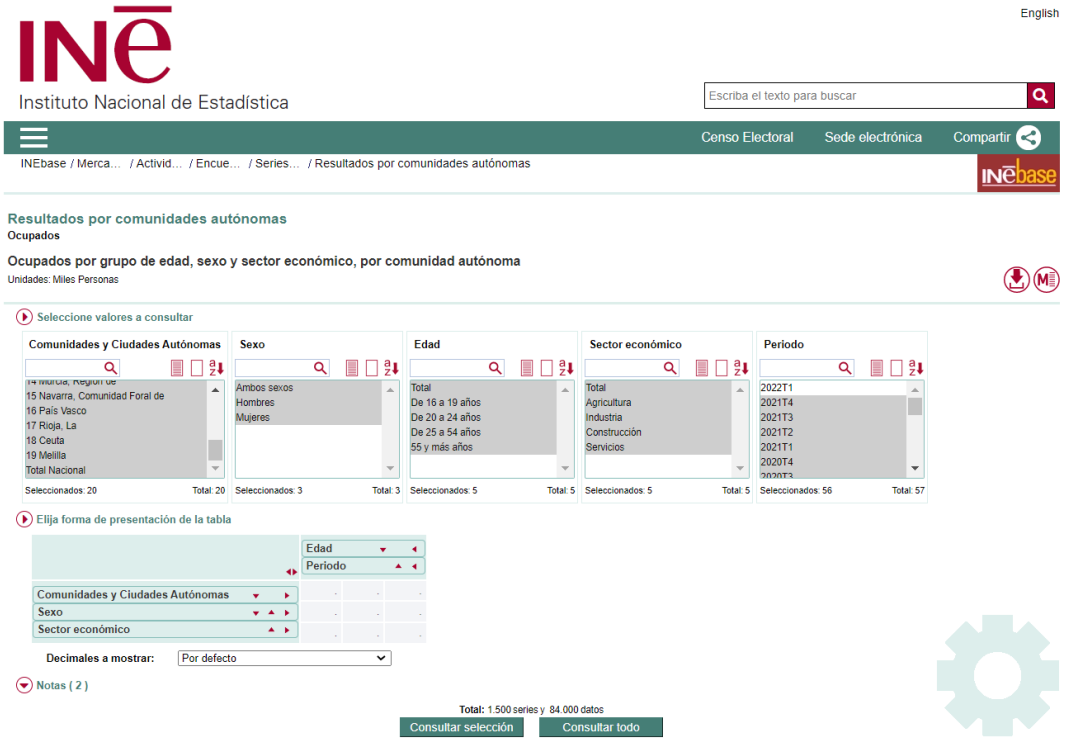

Figura 3. Descarga de los datos INE

De cara a la investigación llevada a cabo en el proyecto se descargan los datos en valores absolutos de resultados de personas ocupadas por grupo de edad, sexo, sector económico y comunidad autonoma, contando con un total de 84.000 registros.

Los datos constan de las siguientes columnas, con su información correspondiente:

- Comunidad Autónoma: código de la comunidad y nombre. Tipo texto.
- Sexo: diferenciación por género. Tipo texto.
- Sector: diferenciación por los sectores que distingue el INE. Tipo texto.
- Edad: diferenciación por rangos de edad. Tipo texto.
- Periodo: diferenciación en trimestres formato (AAAAT0). Tipo texto.
- Total: valor absoluto en miles de personas. Tipo texto.

Para poder realizar un correcto análisis de los datos, se ha procedido a la transformación de los mismos en la herramienta Power BI, la cual nos permite, en una pestaña suplementaria, realizar todas las transformaciones necesarias. En el proyecto se han realizado los siguientes cambios:

- Separación de la columna de 'CCAA' en código identificador de tipo numérico y en el nombre de la CCAA en tipo texto. No se procesa como tipo geográfico ya que no se va a implementar ningún mapa en la solución final.
- Eliminación de los valores acumulativos en todas las columnas de datos.

- Transformación de la columna ‘Periodo' en datos de formato fecha haciendo referencia los últimos días de cada trimestre (ej. T1 se tiene como resultado el 31/03 de cada año).
- Modificación del formato de tipo texto a tipo numérico de la columna ‘Total’.

Gracias a estos cambios realizados se podrá acometer el análisis en las herramientas que se describen a continuación.

4.2.2 Tecnologías y herramientas utilizadas

De cara a trabajar el tratamiento, la visualización y el análisis de los datos, vamos a utilizar la herramienta PowerBI que para ciertas consultas se basa en el lenguaje DAX, muy similar al SQL (para trabajar con bases de datos). De esta herramienta se ha hablado anteriormente en el estado del arte.

Para la creación y evaluación del modelo se ha trabajado en el entorno de trabajo Jupyter Notebook, con el que se permite desarrollar en el lenguaje de programación Python. Finalmente, para la realización del modelo, ha sido necesario implementar una serie de librerías, paquetes y algoritmos de libre acceso para el correcto funcionamiento del código.

- Numpy: una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos. Se puede utilizar como un contenedor multidimensional de datos genéricos. Es capaz de definir tipos de datos arbitrarios. Capaz de realizar integraciones de datos sin problemas y con rapidez.

- Pandas: es un paquete de python que depende de Numpy y permite trabajar con estructuras de datos en formato de dataframes. Los principales datos que se utilizan son datos tabulares y series temporales. Gracias a Pandas se permiten realizar diversas acciones sobre los datos.

- Matplotlib: librería para generar gráficas a partir de datos contenidos en listas, vectores, etc. en el lenguaje de programación Python y en su extensión matemática Numpy.

- Statsmodels: es un módulo de Python que proporciona clases y funciones para la estimación de diferentes modelos, capaz de llevar a cabo pruebas estadísticas y exploración de los datos.

- Scikit-learn: librería que unifica los principales algoritmos y funciones de machine learning, facilitando el procesamiento, entrenamiento, optimización y validación de los modelos predictivos.

- Sk Forecast: es una librería de Python que facilita el uso de regresores de scikit-learn como pronosticadores de varios pasos. Funciona con cualquier regresor compatible con la API scikit-learn.

- XG Boost: Extreme Gradient Boosting. Algoritmo predictivo supervisado que utiliza el principio de boosting. Genera modelos de predicción débiles que van aprendiendo de los modelos anteriores con la finalidad de obtener un modelo final más robusto.

- Cat Boost: Es un algoritmo de aprendizaje automático que utiliza el gradient boosting en los árboles de decisión. Algoritmo de mayor calidad a la hora de manejar parámetros del modelo, más rápido y escalable, capaz de realizar predicciones mejores y más eficaces.

- Random Forest Regressor: es un algoritmo de aprendizaje supervisado que ajusta una serie de árboles de decisión, cada uno entrenado con una muestra distinta de los datos de entrenamiento generada mediante bootstrapping. Los métodos basados en árboles se han convertido en uno de los referentes dentro del ámbito predictivo.

4.2.3 Metodología

Debido a que las tareas desarrolladas a lo largo del proyecto no precisan seguir una secuencia lineal y dado que después de cada sprint se prevén mejoras y nuevos requisitos, la metodología que mejor se ajusta al proyecto es la Agile. Es cierto que algunas metodologías tradicionales como la cascada se podrían adaptar a nuestro proyecto, pero estas resultan ser más lentas y costosas. Dentro de las metodologías Agile se ha determinado que una metodología Scrum es la que mayores ventajas ofrece y está orientada a equipos de trabajo pequeños, donde en cada sprint se evalúan las tareas realizadas y se proponen nuevas para los siguientes sprints.

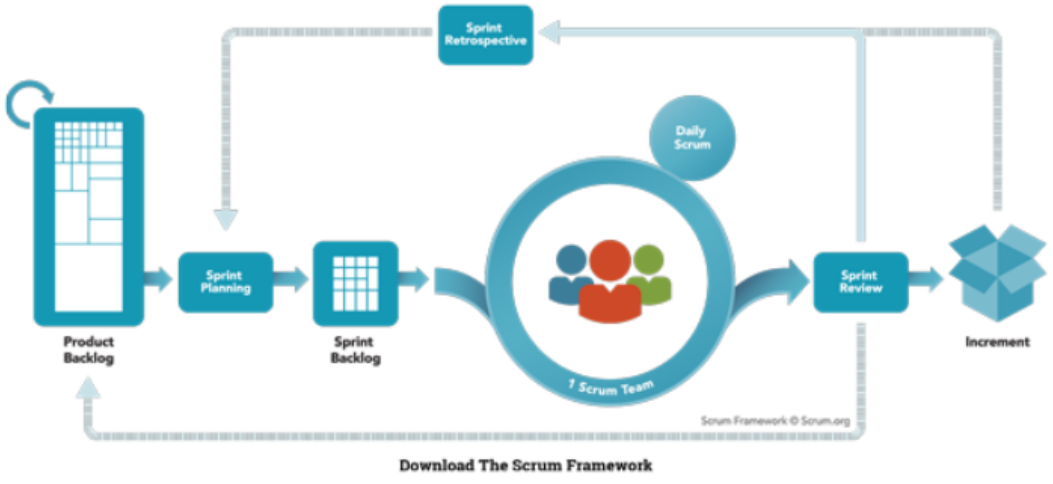

Figura 4. Proceso SCRUM

4.3 Recursos requeridos

A continuación, se enumeran los recursos, tanto humanos como materiales, que han sido utilizados para poder realizar el proyecto de investigación.

Recursos humanos

- Alumno del grado de Ingeniería Matemática (1)

Recursos materiales

- Ordenador portátil Lenovo (1).
- Artículos de papelería y escritorio.
- Paquetes Microsoft Office (última actualización).
- Power BI aplicación de escritorio y en la nube.
- Google Drive para el control de versiones
- Software de distribución Anaconda Navigator

Otros recursos

- Conexión a Internet.
- Gastos de luz e Internet.

4.4 Presupuesto

El Presupuesto supone la evaluación económica total del proyecto. En la siguiente tabla se calcula el presupuesto total del proyecto, teniendo en cuenta todos los recursos mencionados previamente.

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	3600€	Un total de 300 horas para completar los 12 ECTS a 12€/h valor de mercado de un analista de datos junior
Equipo técnico utilizado	1000€	Incluye el valor de mercado del ordenador portátil así como los gastos en el material de papelería y escritorio.
Software utilizado	69€	Valor de mercado del software personal para el uso de Microsoft Office 365. El resto de aplicaciones son Open Source y los

Tipo de coste	Valor	Comentarios
		extras que precisan de pago no se aplican al proyecto.
Otros	119.60€	El coste de internet es aproximado para las necesidades del proyecto de 29.9€/mes por 100 Mbps a precio de mercado, durante la duración del proyecto .

Tabla 2. Presupuesto del proyecto

Por todo lo mencionado anteriormente el presupuesto del proyecto será de un total de 4788.60 €.

4.5 Viabilidad

En este apartado se determinará, mediante el plan de viabilidad, si el proyecto desarrollado será rentable y creará valor a futuro. Dado que el proyecto consiste en el estudio de un suceso y la creación de un informe que va a generar conocimiento, la manera por la que se puede convertir en un estudio viable y rentable es mediante la venta del informe a entidades que precisen de la información recogida en este proyecto.

4.6 Resultados del proyecto

El análisis llevado a cabo en el informe tiene como resultados dos puntos principales: el de la visualización de datos por diferente sectores, comunidades autónomas y sexo de la población, y los resultados que se han obtenido en las predicciones generadas por nuestro modelo.

4.6.1 Resultados del Análisis

En este apartado se destacará la información más notable extraída a partir de la visualización de los datos obtenidos a través del portal del Instituto Nacional de Estadística, tanto a nivel nacional de España como a nivel autonómico, descartando el estudio de las ciudades autónomas de Ceuta y Melilla. Este análisis se realiza principalmente poniendo el foco en el periodo en el cual la pandemia de la COVID-19 aterriza en España, provocando los confinamientos y cierres temporales o totales de algunas empresas, es decir, en marzo de 2020.

- *España*

Se ve reflejada que la mayor caída de empleo ha sido en el sector servicios, con una pérdida de 400 mil puestos de trabajo de los hombres frente a las 600 mil mujeres. Mientras, en el resto de los sectores los hombres han sufrido mayor pérdida de empleo que las mujeres.

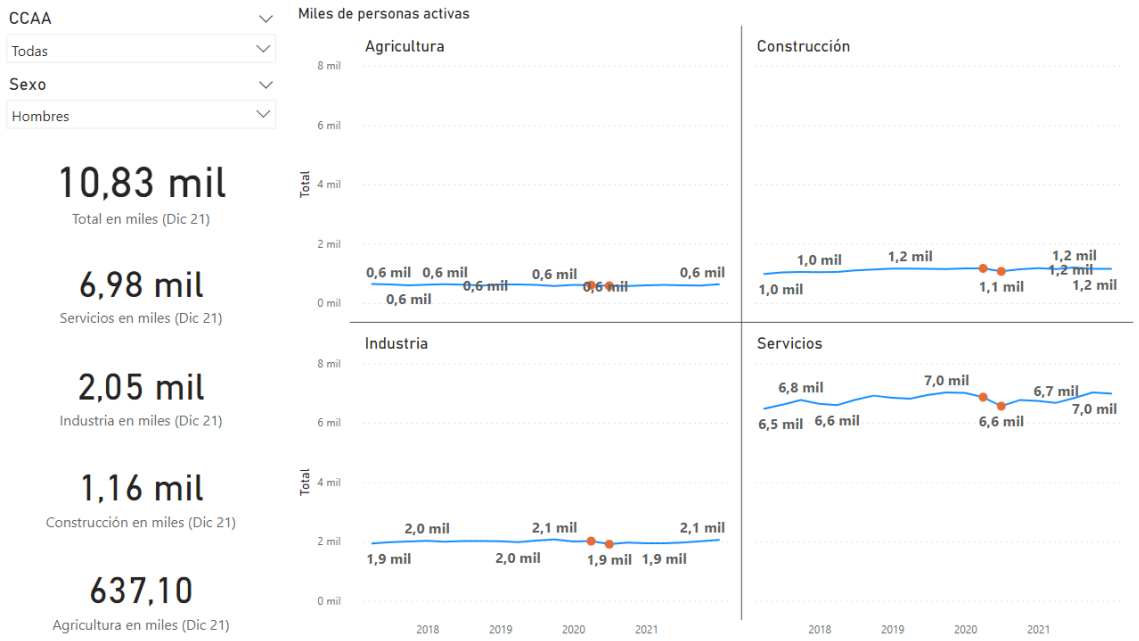

Figura 5. Hombres Activos España

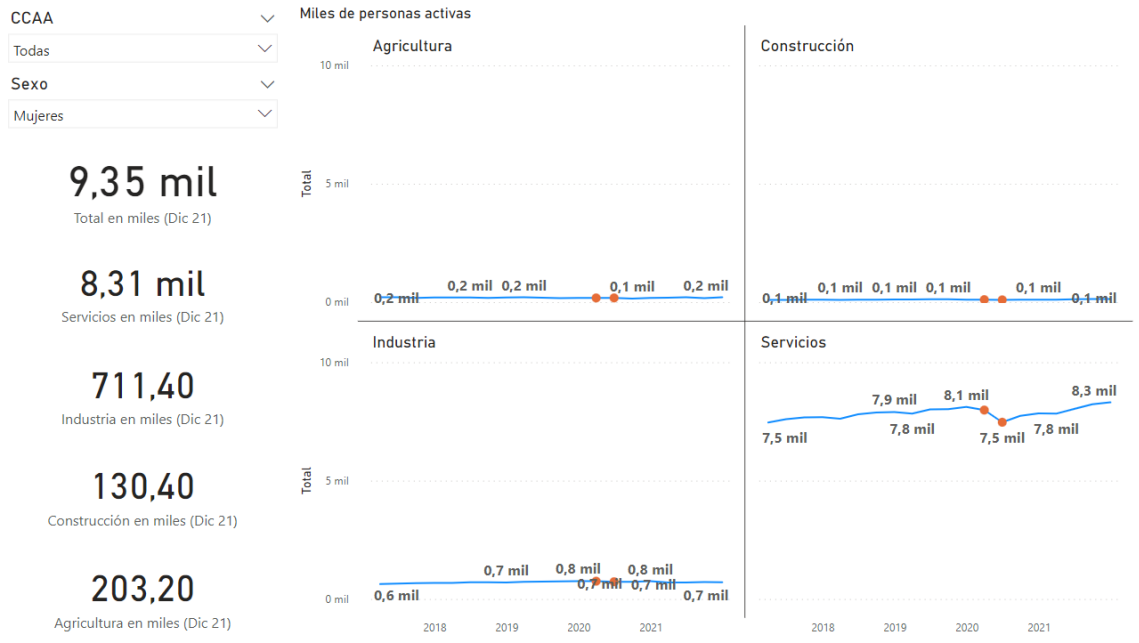

Figura 6. Mujeres Activas España

- *Andalucía*

Los hombres han sufrido mayor caída proporcional en los sectores de la agricultura y la industria. Por el contrario, las mujeres se han visto más golpeadas en el sector servicios, con casi 150 mil empleos perdidos.

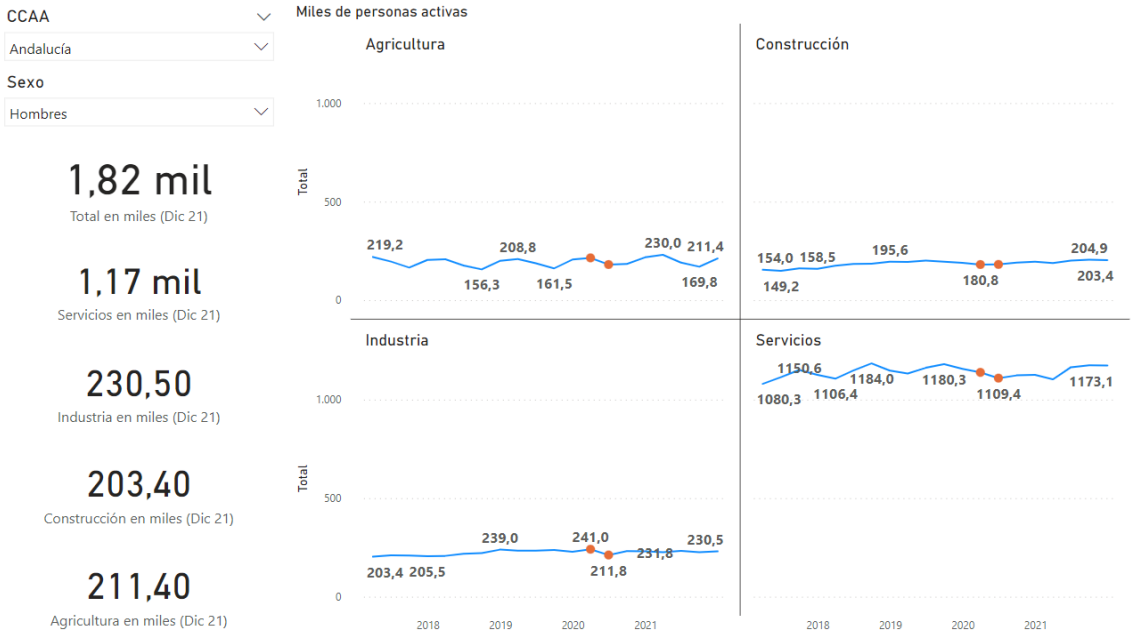

Figura 7. Hombres Activos Andalucía

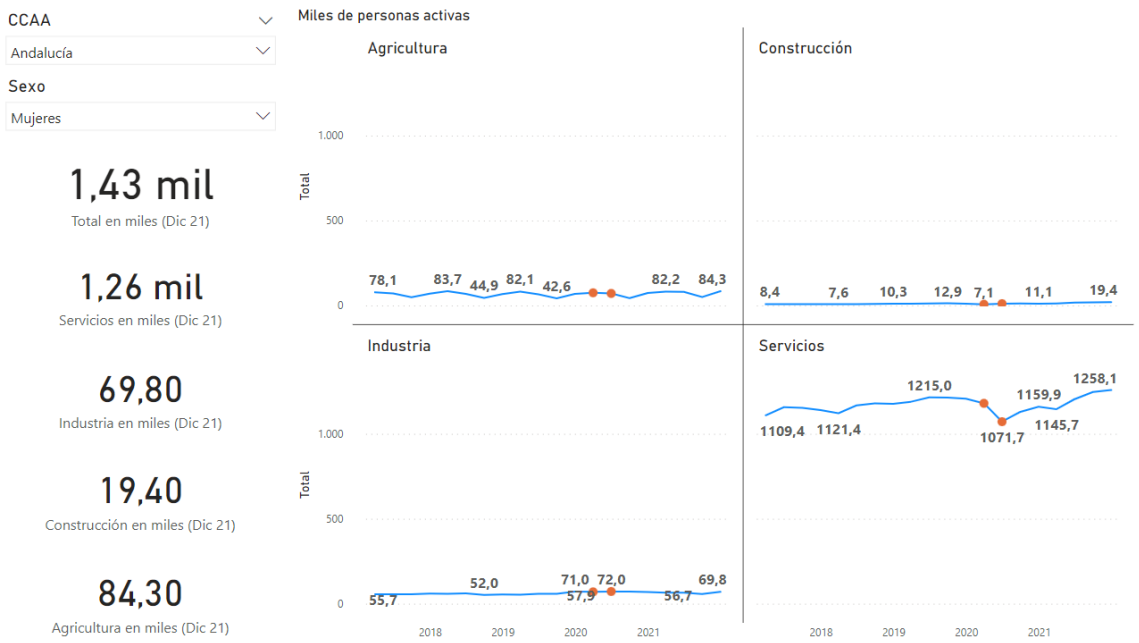

Figura 8. Mujeres Activas Andalucía

- *Aragón*

Las mujeres se han visto más afectadas que los hombres proporcionalmente en cuanto a pérdida de trabajo en el sector servicios. Los hombres se han visto mayormente afectados en el sector de la construcción.

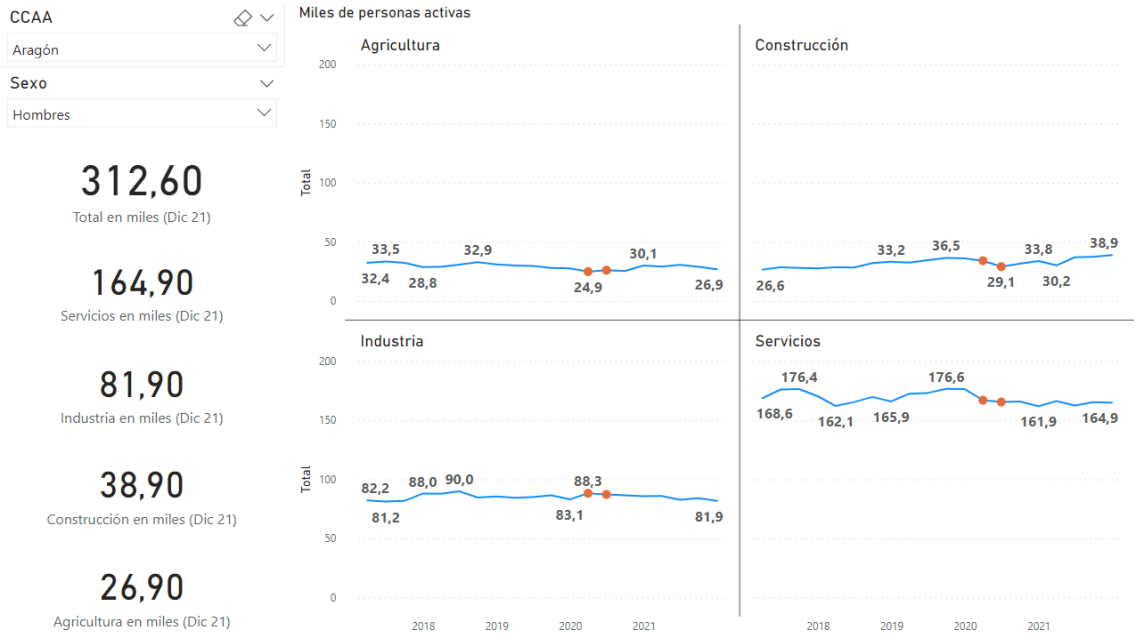

Figura 9. Hombres Activos Aragón

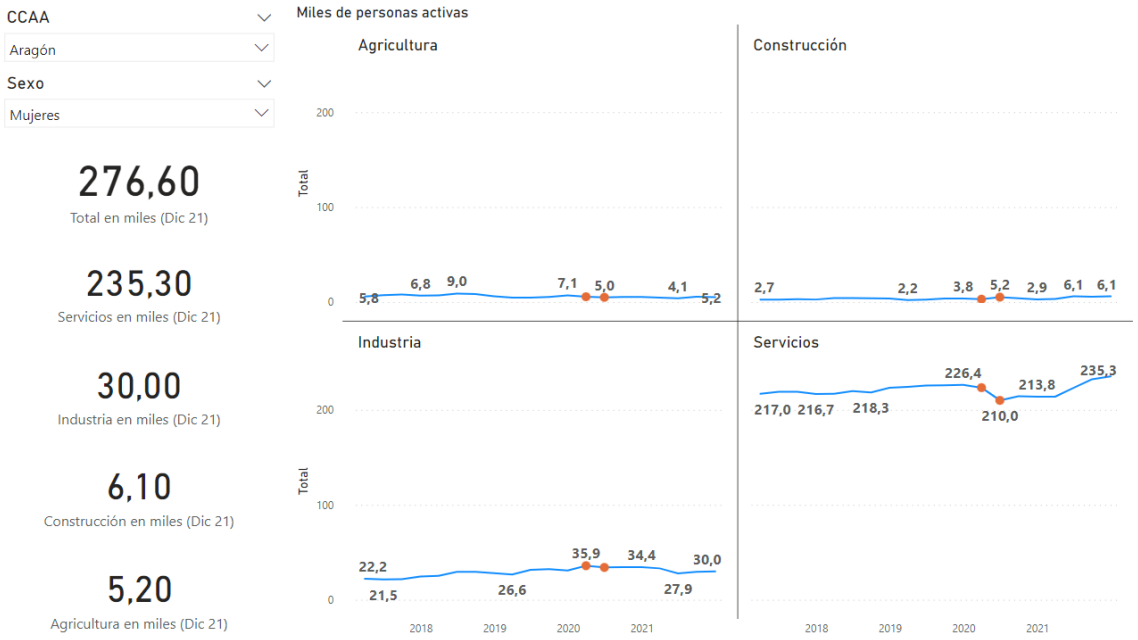

Figura 10. Mujeres Activas Aragón

- *Asturias*

En esta ocasión, los hombres se han visto mayormente perjudicados en todos los sectores salvo en la industria, en la cual han sufrido un aumento de los trabajadores, mientras que las mujeres apenas han sufrido el impacto de la pandemia.

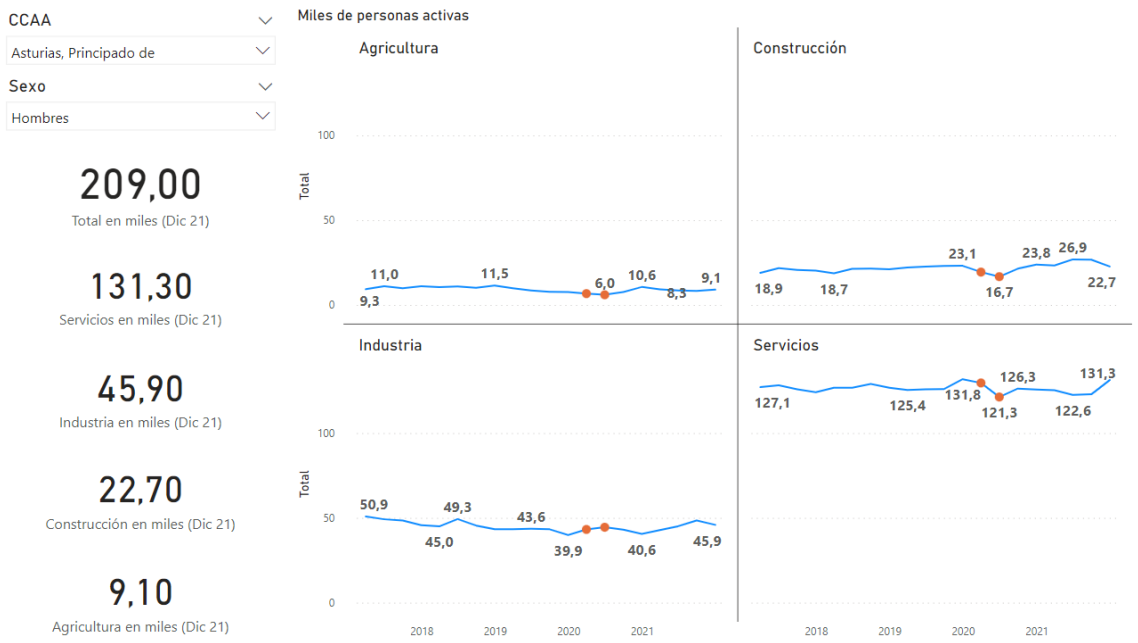

Figura 11. Hombres Activos Asturias

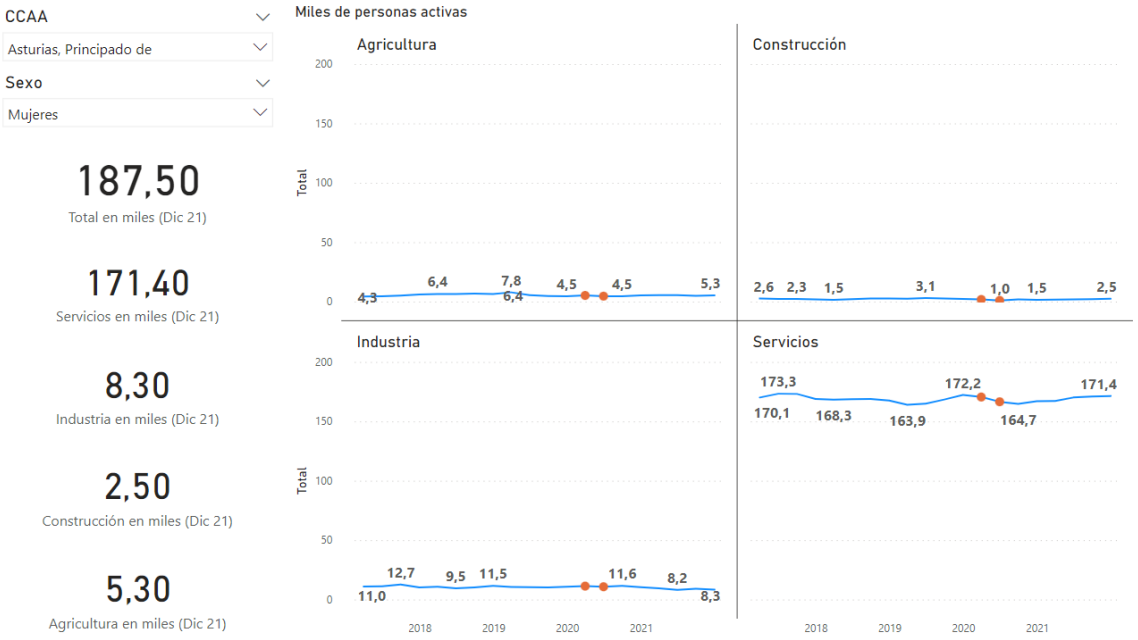

Figura 12. Mujeres Activas Asturias

- *Baleares*

Se observa cómo el sector servicios, en las islas tiene una tendencia estacionaria y que debido al impacto de la pandemia, en el segundo trimestre del año 2020 tanto las mujeres como los hombres no alcanzaron los registros de otros años.

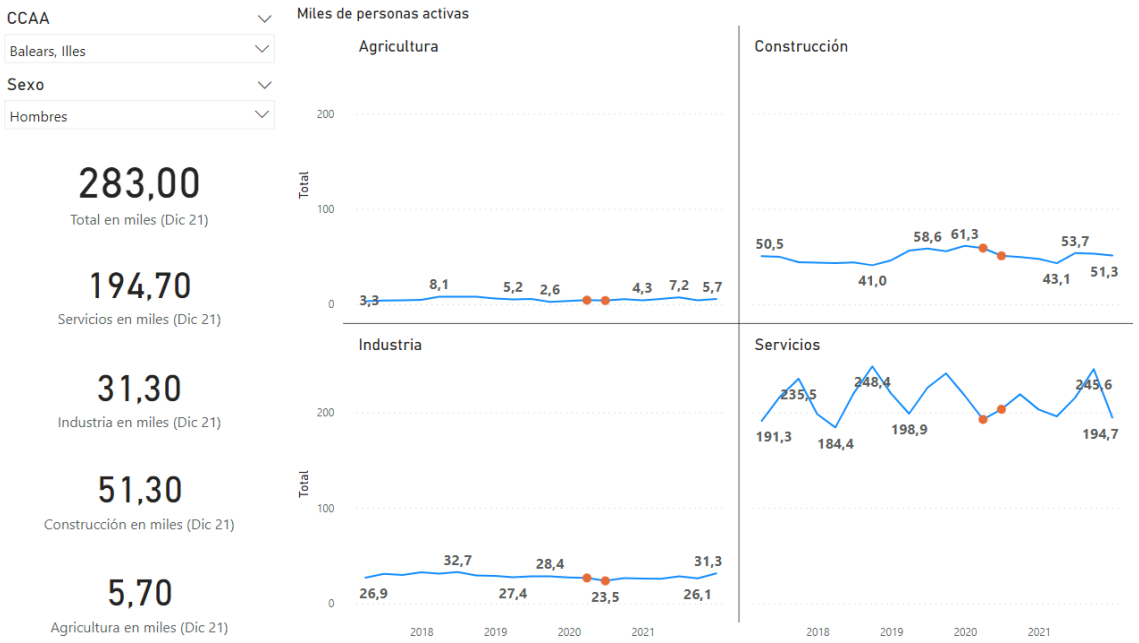

Figura 13. Hombres Activos Baleares

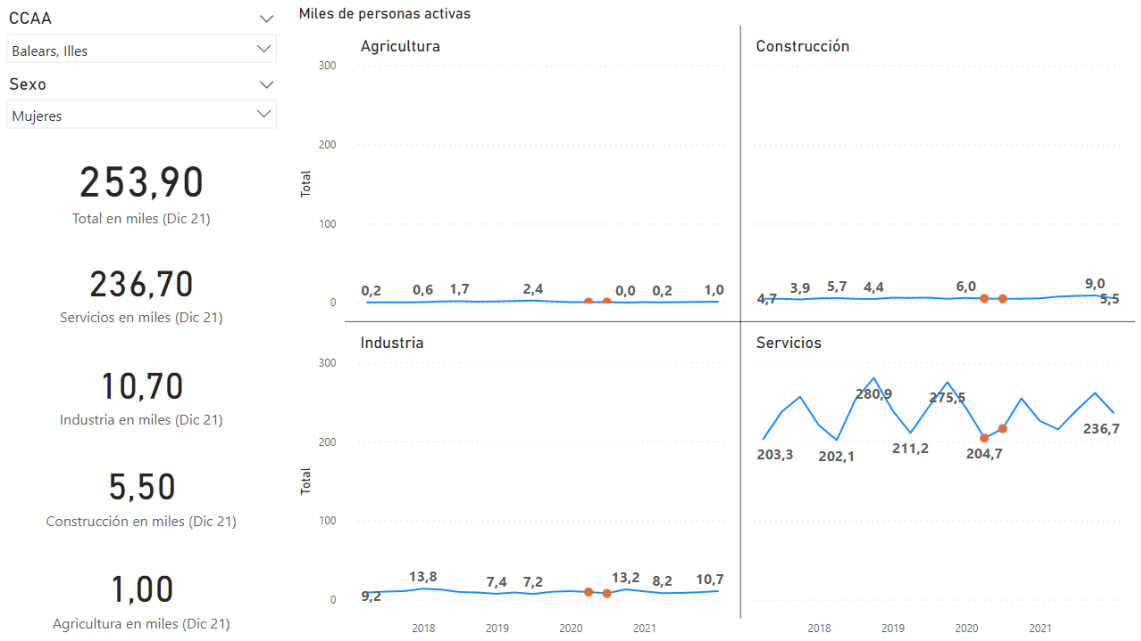

Figura 14. Mujeres Activas Baleares

- *Canarias*

En el sector servicios trabajan casi 40 mil mujeres más que hombres, pero ambos han sufrido caídas similares de número de personas ocupadas en el sector.

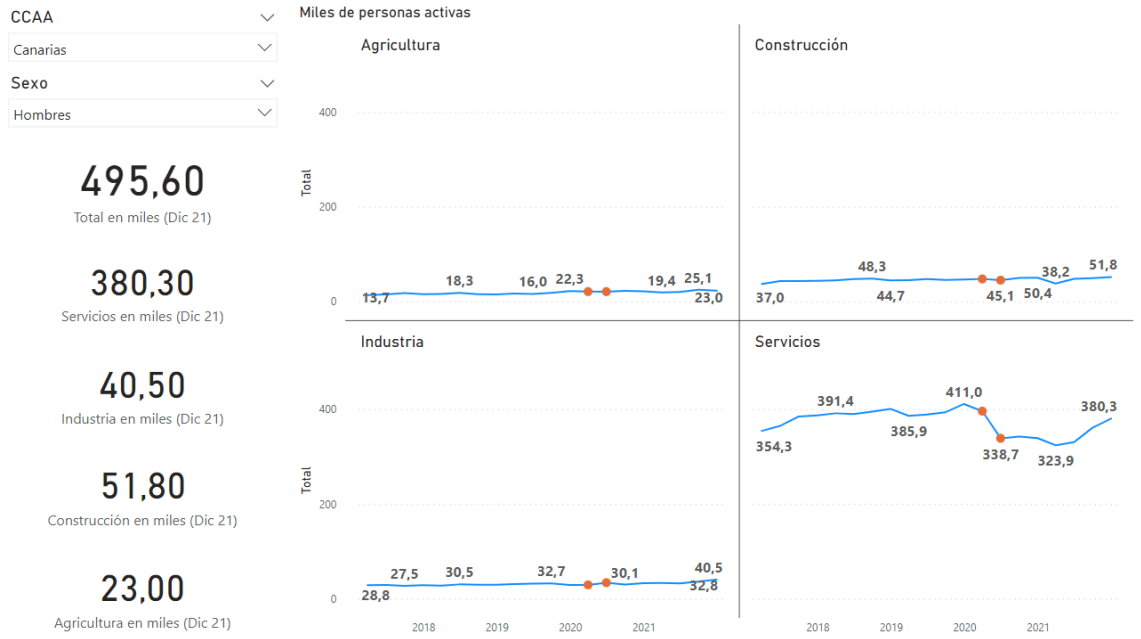

Figura 15. Hombres Activos Canarias

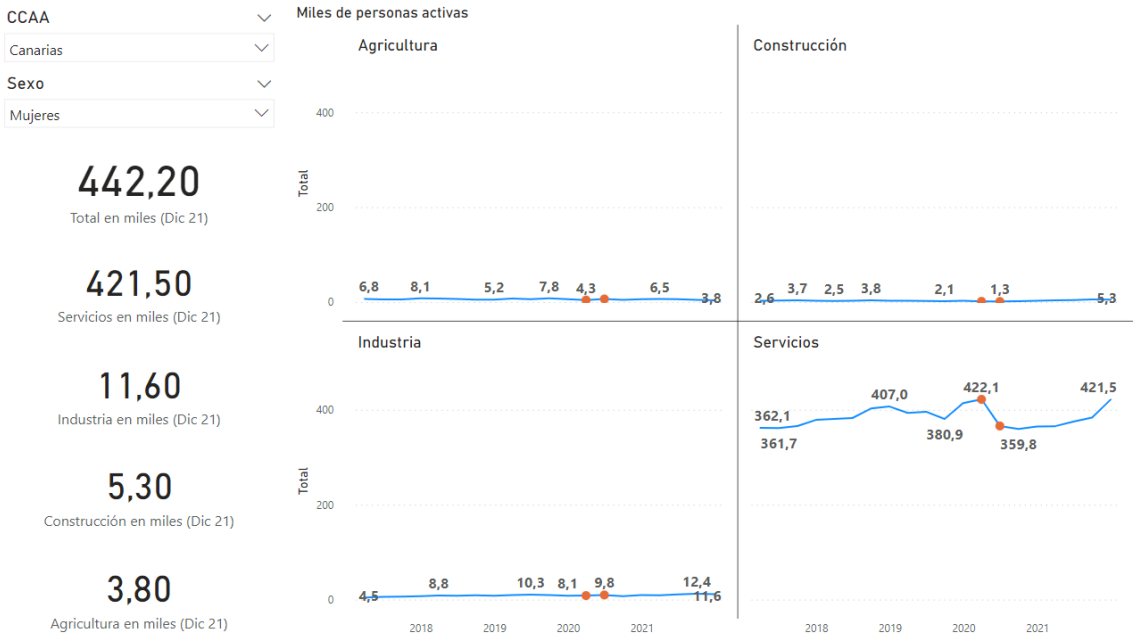

Figura 16. Mujeres Activas Canarias

- *Cantabria*

Los hombres están más repartidos en los distintos sectores que las mujeres y han sufrido el golpe tanto en la industria como en el sector servicios. Para las mujeres se perdieron alrededor de 15 mil puestos de trabajo.

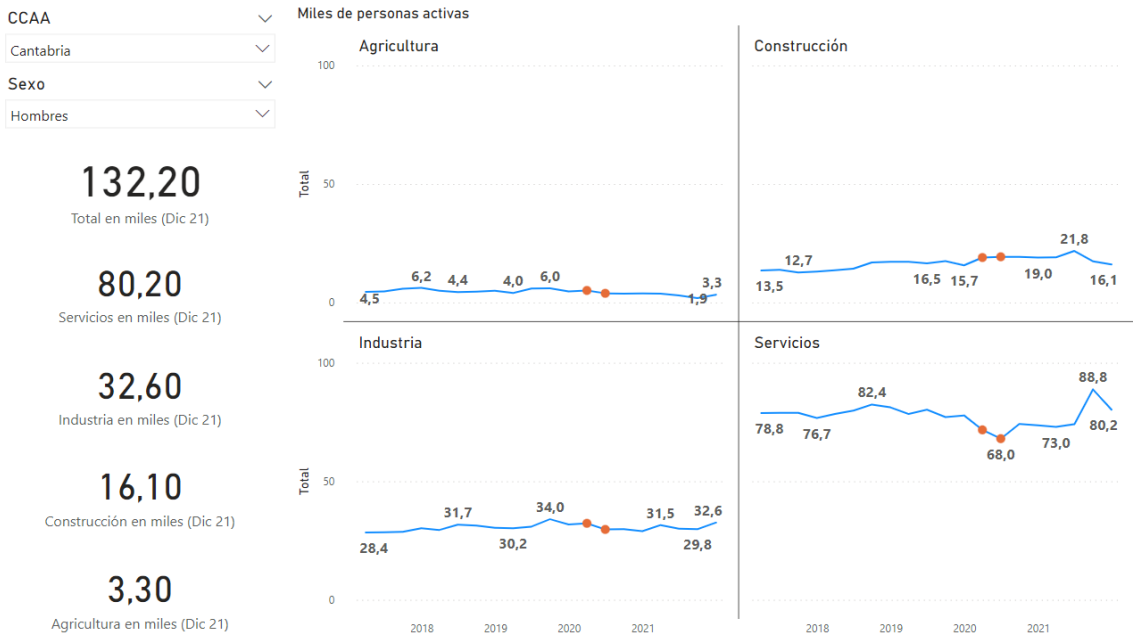

Figura 17. Hombres Activos Cantabria

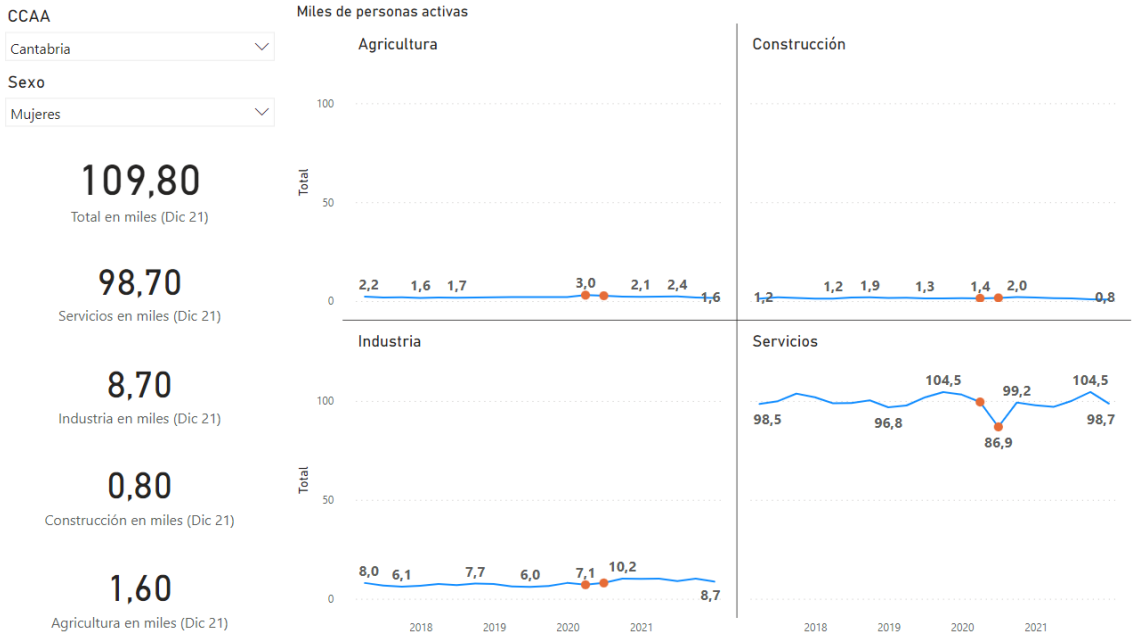

Figura 18. Mujeres Activas Cantabria

- *Castilla La Mancha*

El impacto en los hombres ha sido mayor en la construcción, mientras que las mujeres se han visto perjudicadas en el sector servicios, con aproximadamente una pérdida de 10 mil puestos de trabajo en cada uno, siendo mayor el impacto en los hombres debido a la proporción.

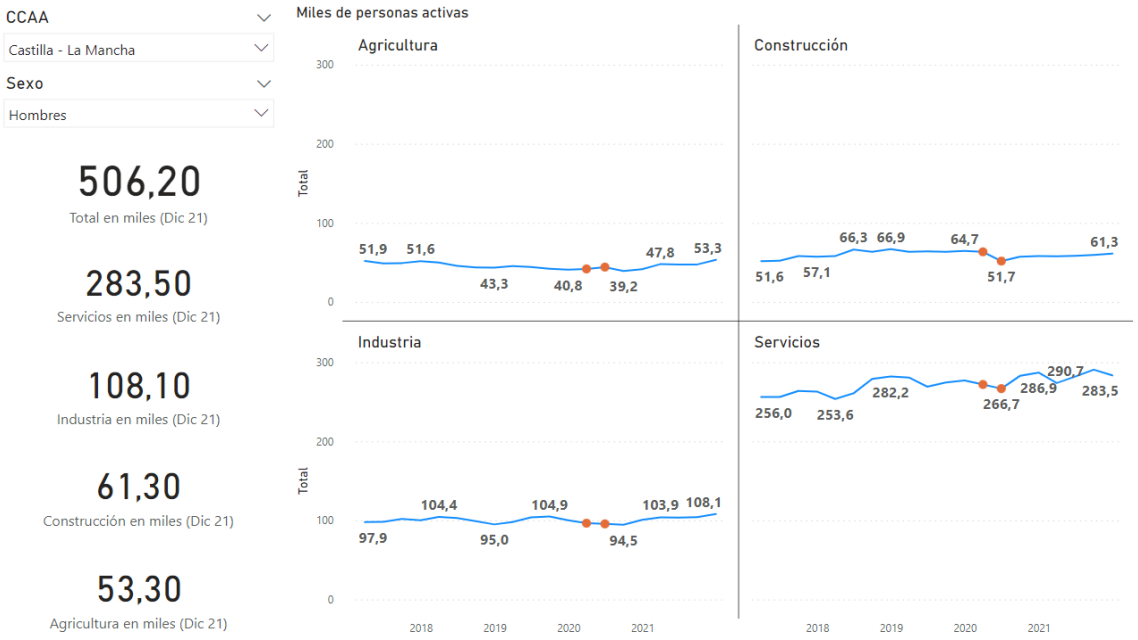

Figura 19. Hombres Activos Castilla La-Mancha

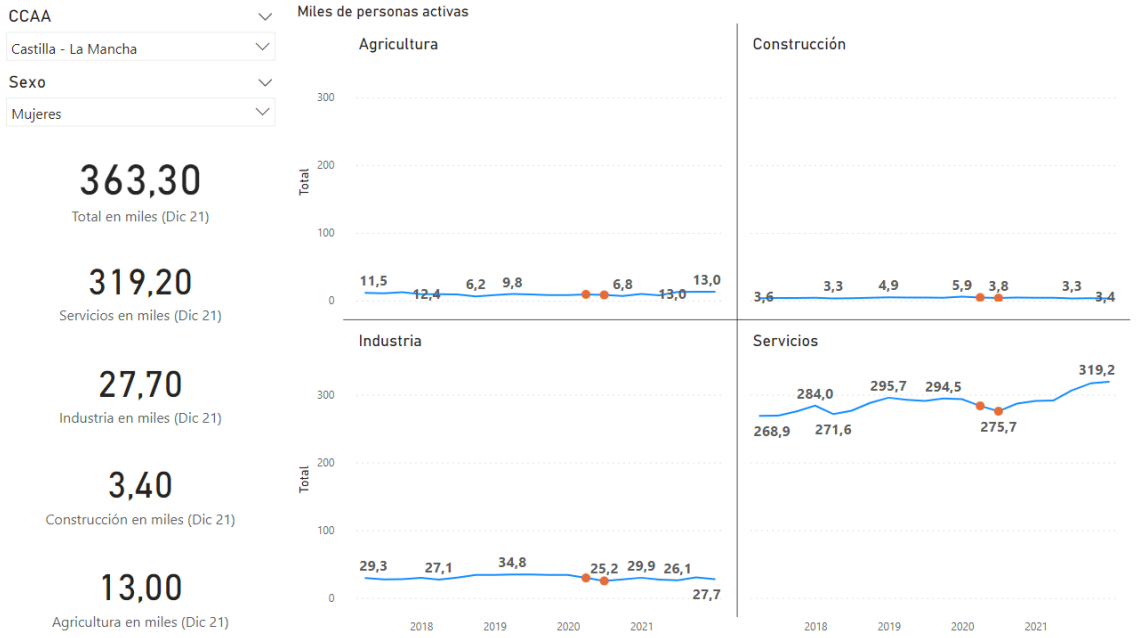

Figura 20. Mujeres Activas Castilla La-Mancha

- *Castilla y León*

Al igual que hemos visto en otras comunidades, los hombres están más repartidos en los sectores. Se han perdido mayormente puestos de trabajo en el sector servicios, seguido de la industria.

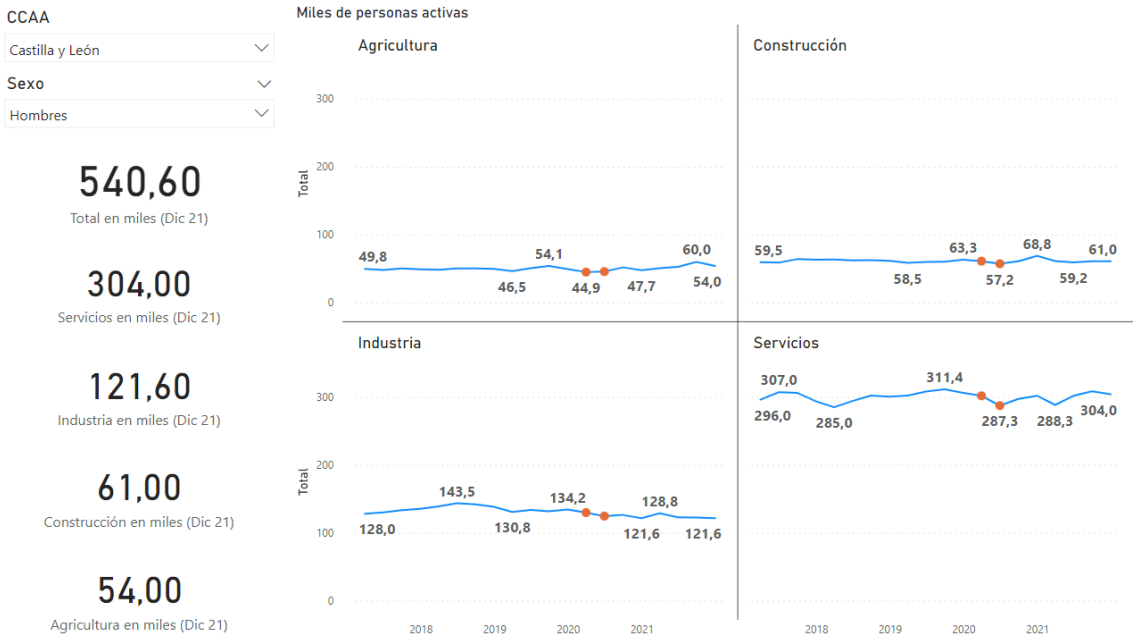

Figura 21. Hombres Activos Castilla y León

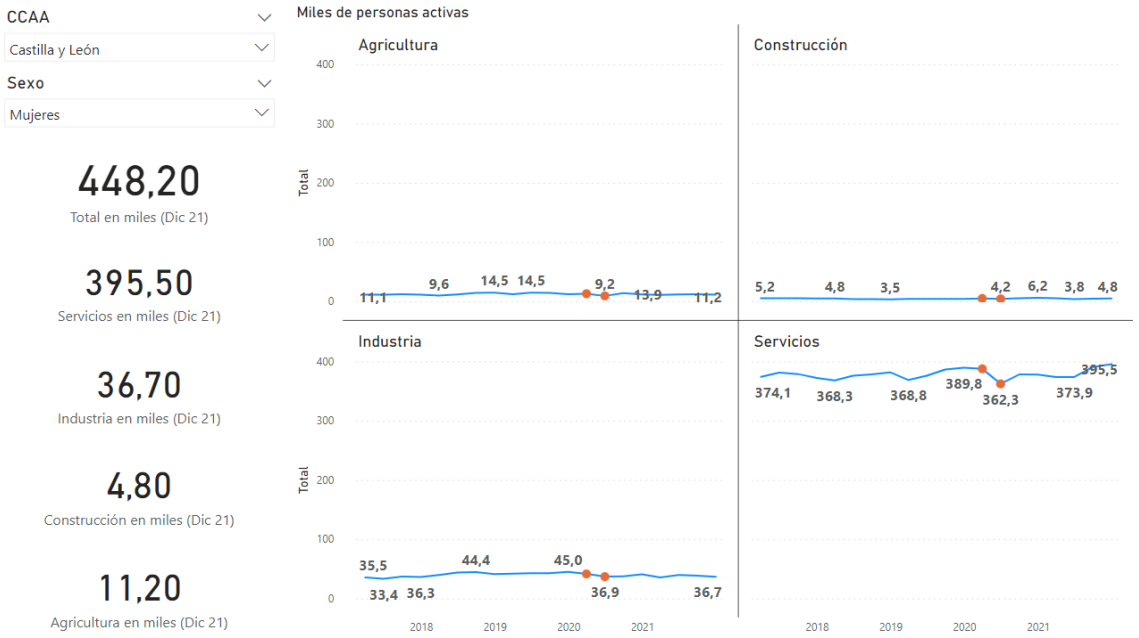

Figura 22. Mujeres Activas Castilla y León

- *Cataluña*

La mayor caída de puestos de trabajo se ha visto en el sector servicios, con mayor pérdida de empleos para las mujeres que para los hombres, con casi 40 mil puestos de trabajos menos.

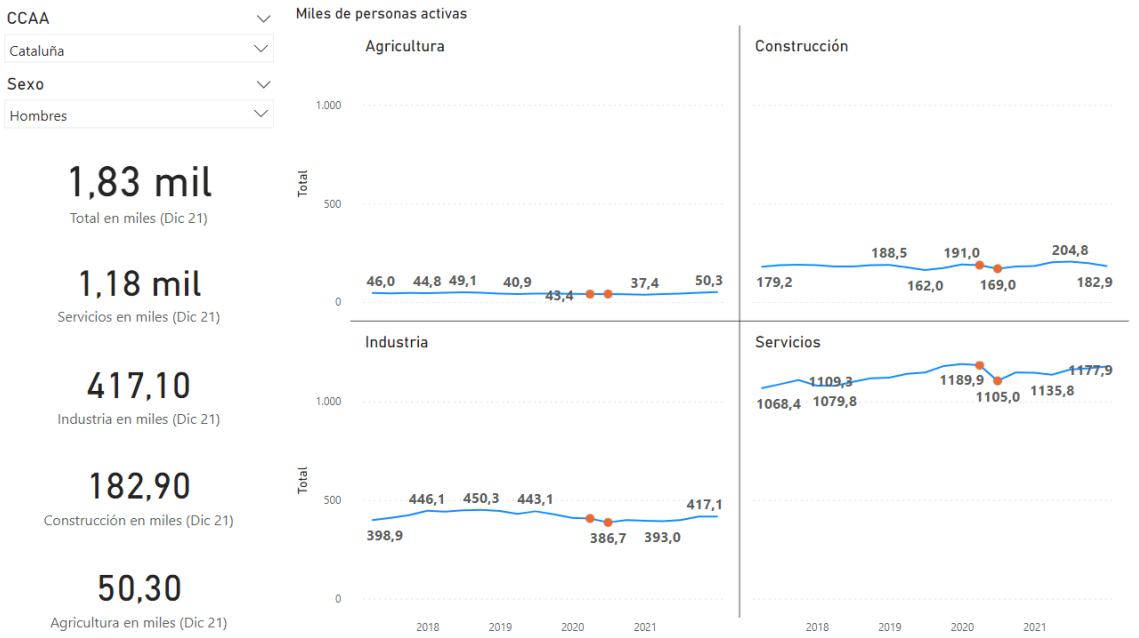

Figura 23. Hombres Activos Cataluña

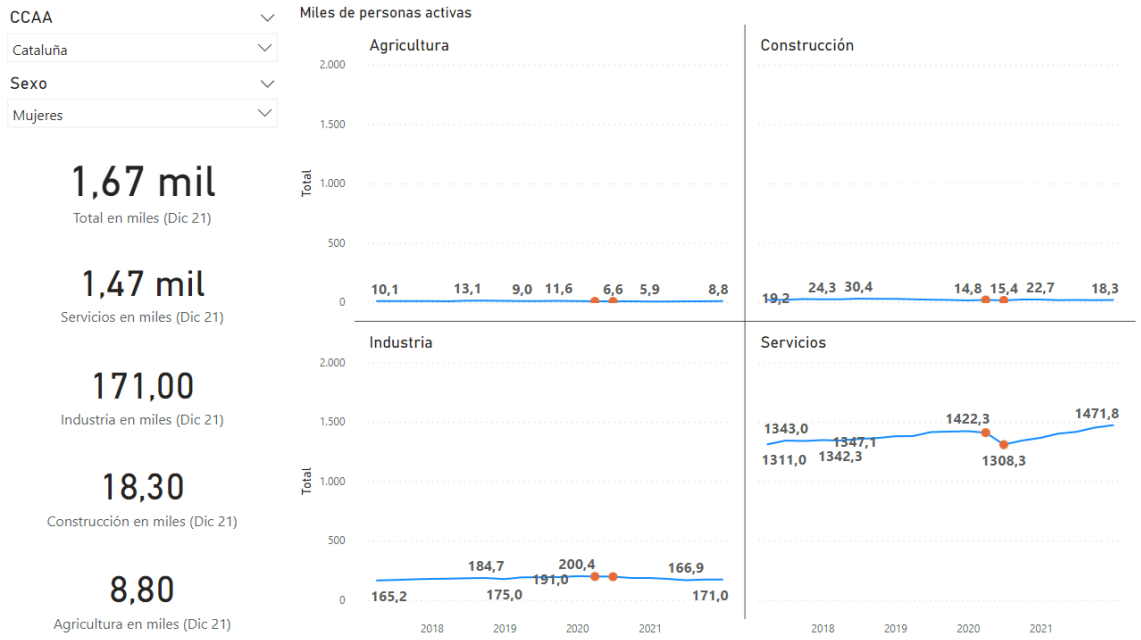

Figura 24. Mujeres Activas Cataluña

- *Comunidad Valenciana*

Se han perdido casi el doble de empleos para mujeres que para hombres en el sector servicios, pero por otro lado, los hombres se han visto más afectados en proporción similar en los sectores de la industria y de la construcción.

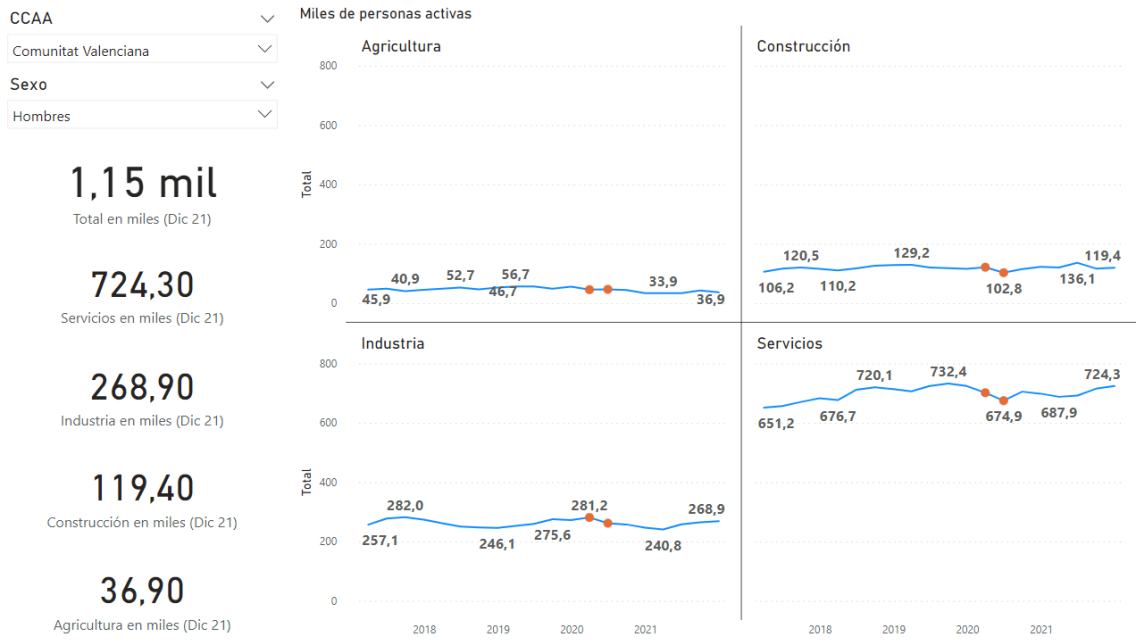

Figura 25. Hombres Activos Comunidad Valenciana

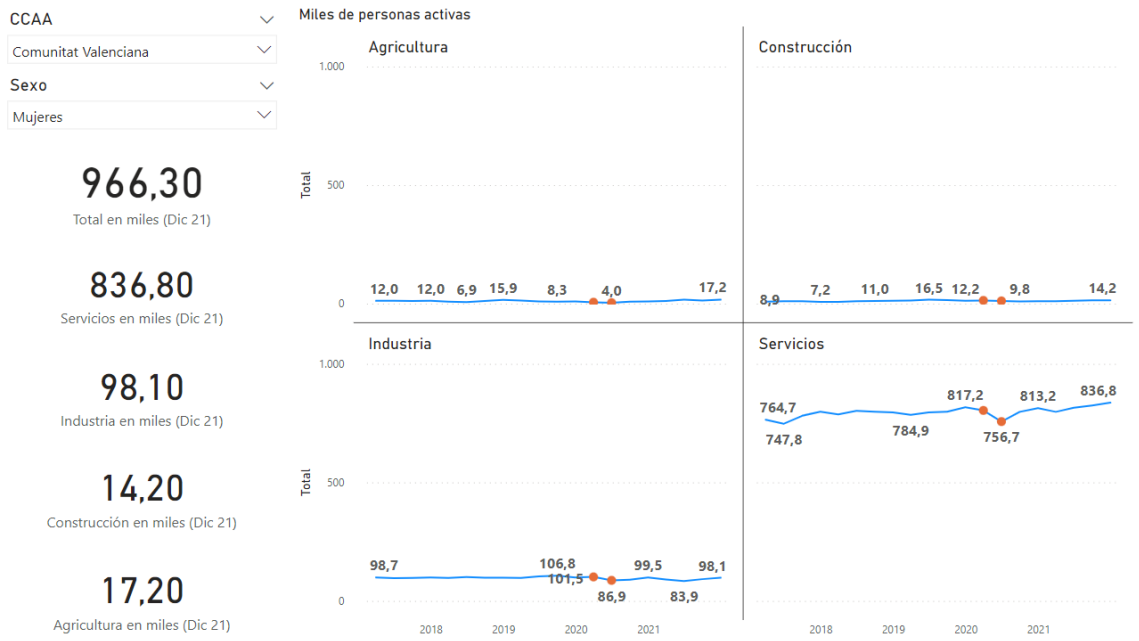

Figura 26. Mujeres Activas Comunidad Valenciana

- *Extremadura*

Las caídas más pronunciadas han sido en el sector de la construcción y la industria para los hombres, mientras que las mujeres apenas han notado la pérdida de empleo (mínimamente en el sector servicios).

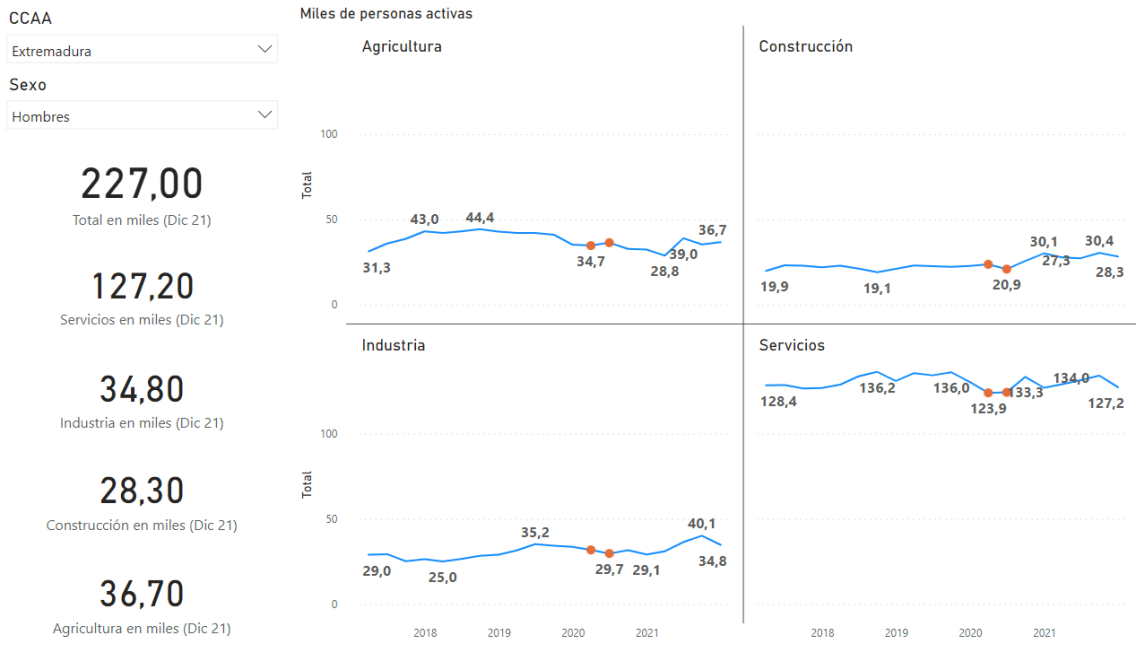

Figura 27. Hombres Activos Extremadura

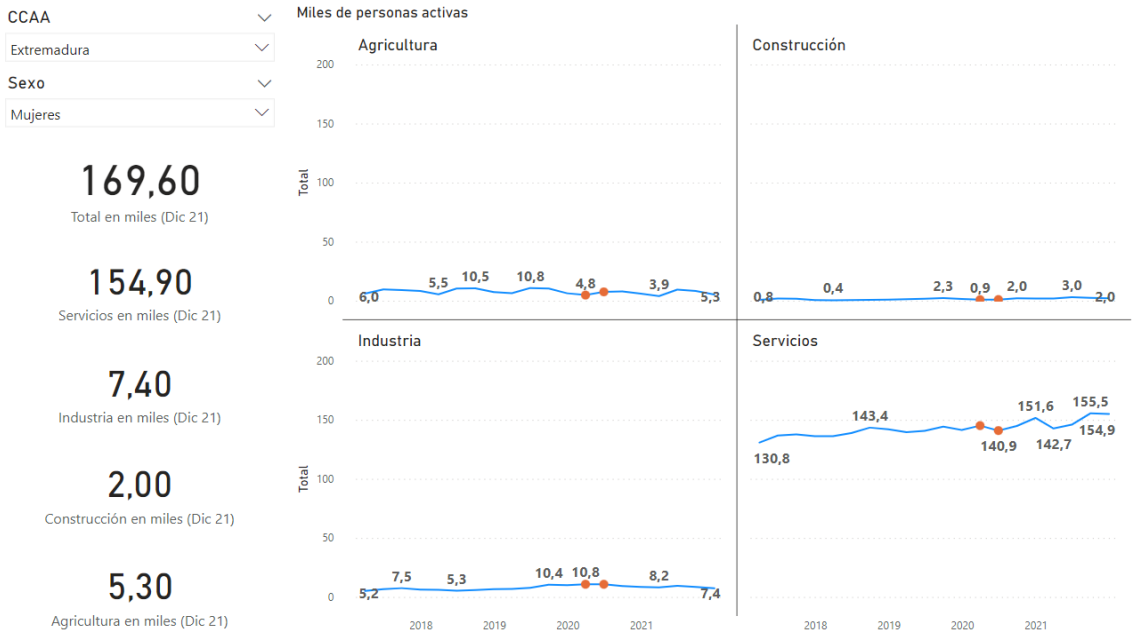

Figura 28. Mujeres Activas Extremadura

- *Galicia*

Se producen pérdidas proporcionales de empleo en los sectores de construcción, industria y servicios en los hombres. Por otro lado, las mujeres pierden 20 mil empleos en el sector servicios, muy similar al total de empleos perdidos por hombres en todos los sectores.

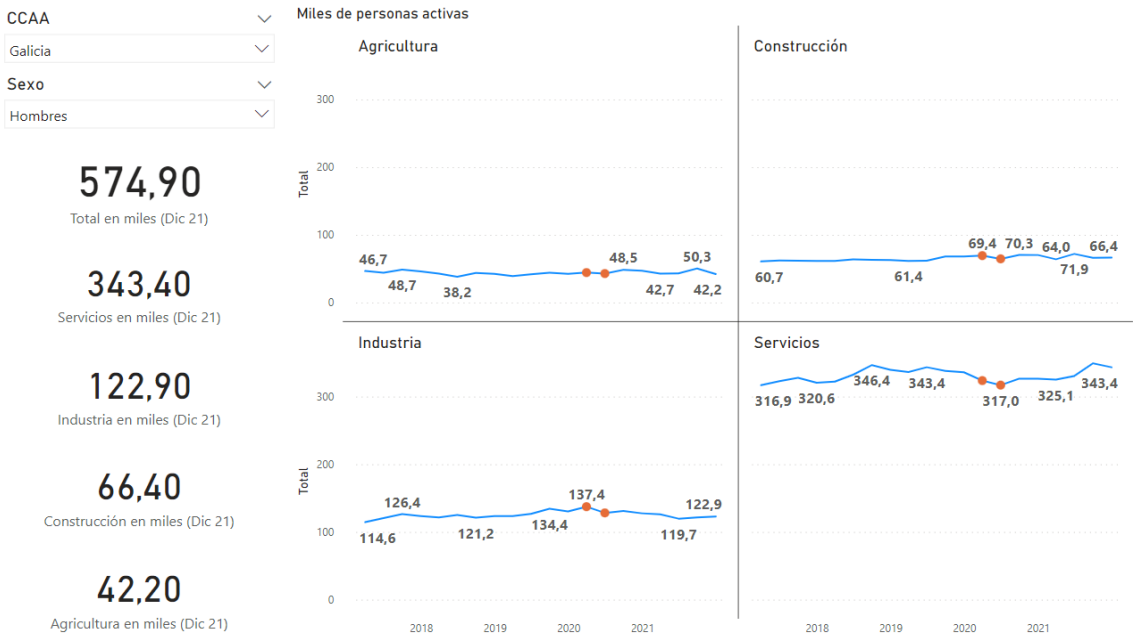

Figura 29. Hombres Activos Galicia

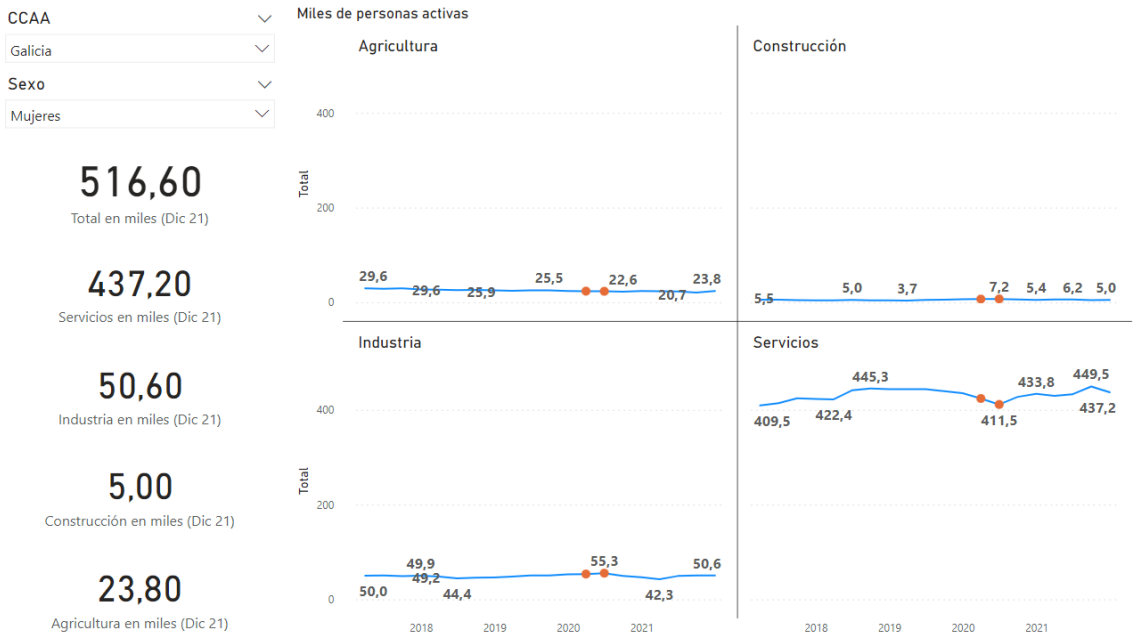

Figura 30. Mujeres Activas Galicia

- *Madrid*

El sector de la industria y de la construcción han sufrido caídas de empleo para los hombres, no tanto así para las mujeres. Por otro lado, el impacto en el sector servicios ha sido de unas pérdidas de 100 mil y 70 mil empleos de mujeres y hombres respectivamente, siendo un sector donde trabajan más mujeres que hombres.

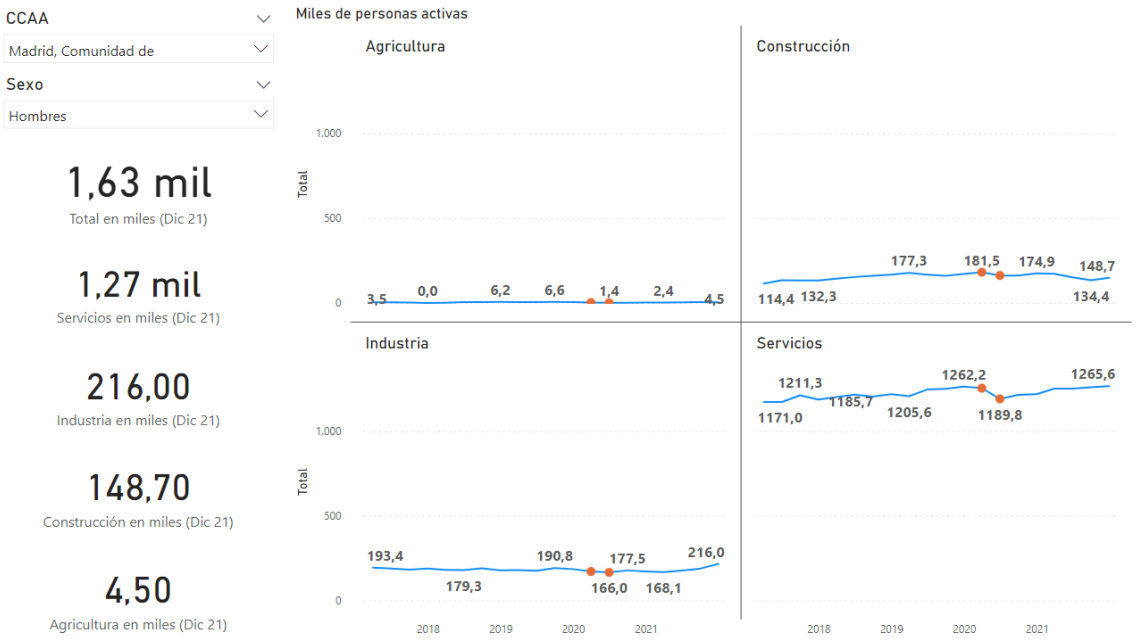

Figura 31. Hombres Activos Madrid

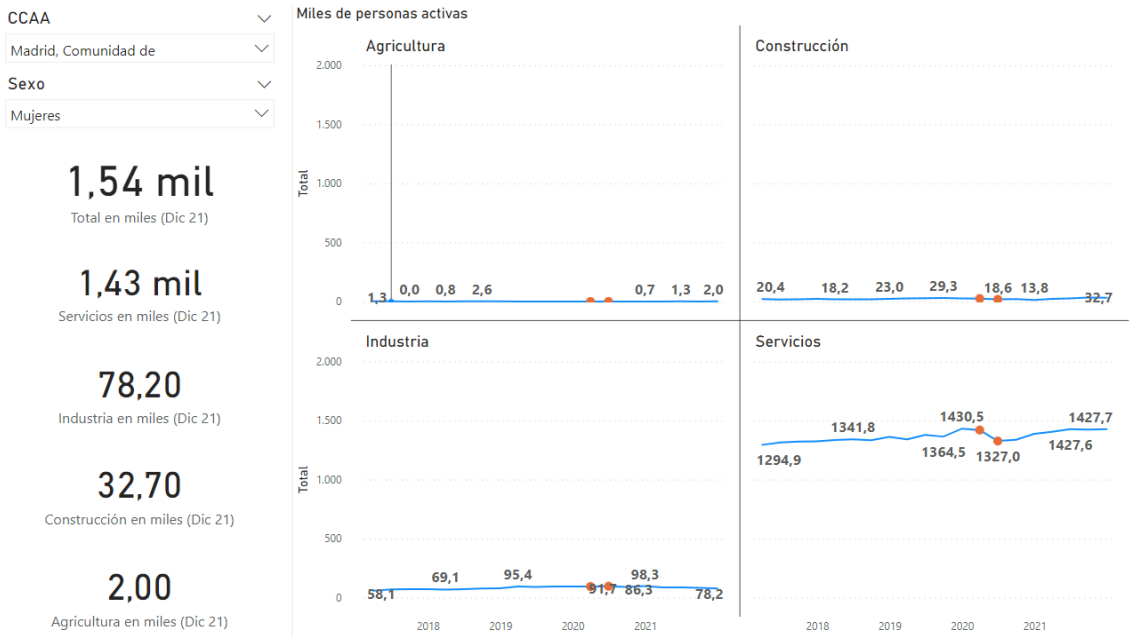

Figura 32. Mujeres Activas Madrid

- *Murcia*

Se aprecia como, en el periodo de confinamiento, la agricultura incrementó los puestos de trabajo, ya que sigue una tendencia estacionaria. El sector servicios y la industria han sido los más damnificados por esta situación, siendo las mujeres las que más empleos pierden.

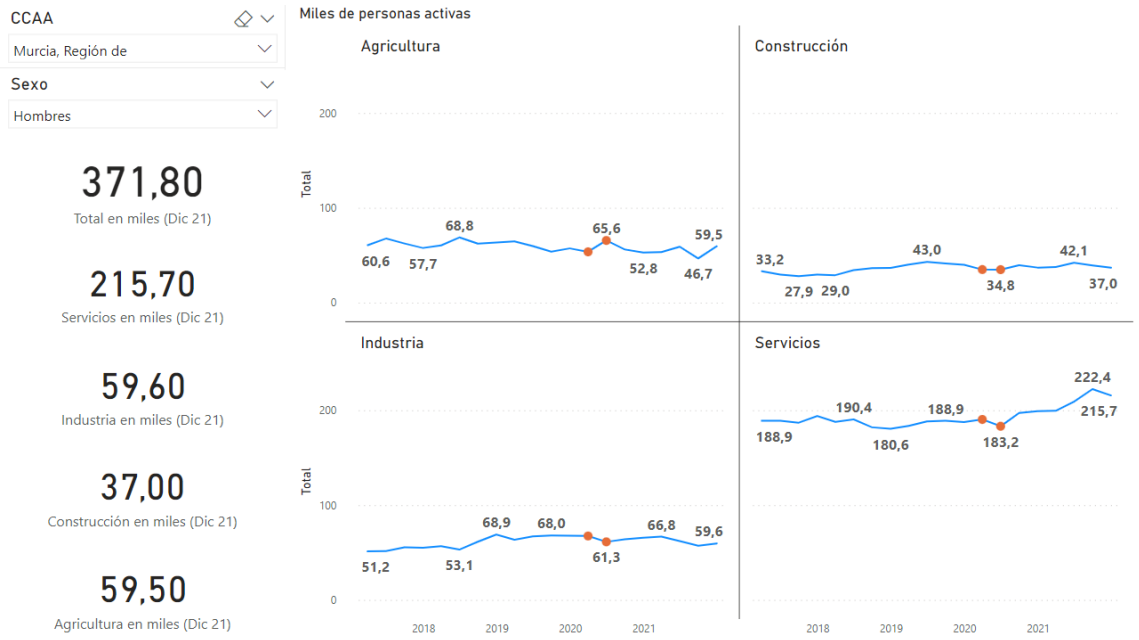

Figura 33. Hombres Activos Murcia

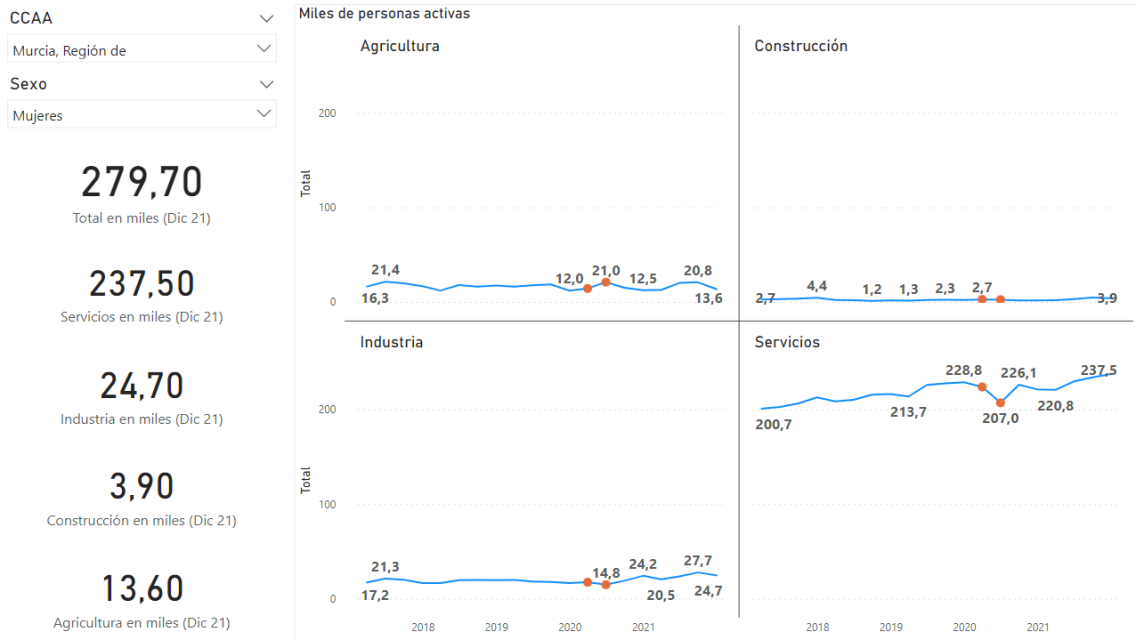

Figura 34. Mujeres Activas Murcia

- *Navarra*

Los hombres se distribuyen de manera similar en los sectores de industria y servicios, con caídas en el número de ocupados de 10 mil puestos de trabajo, lo mismo entre ambos sectores que las mujeres sólo en el sector servicios.

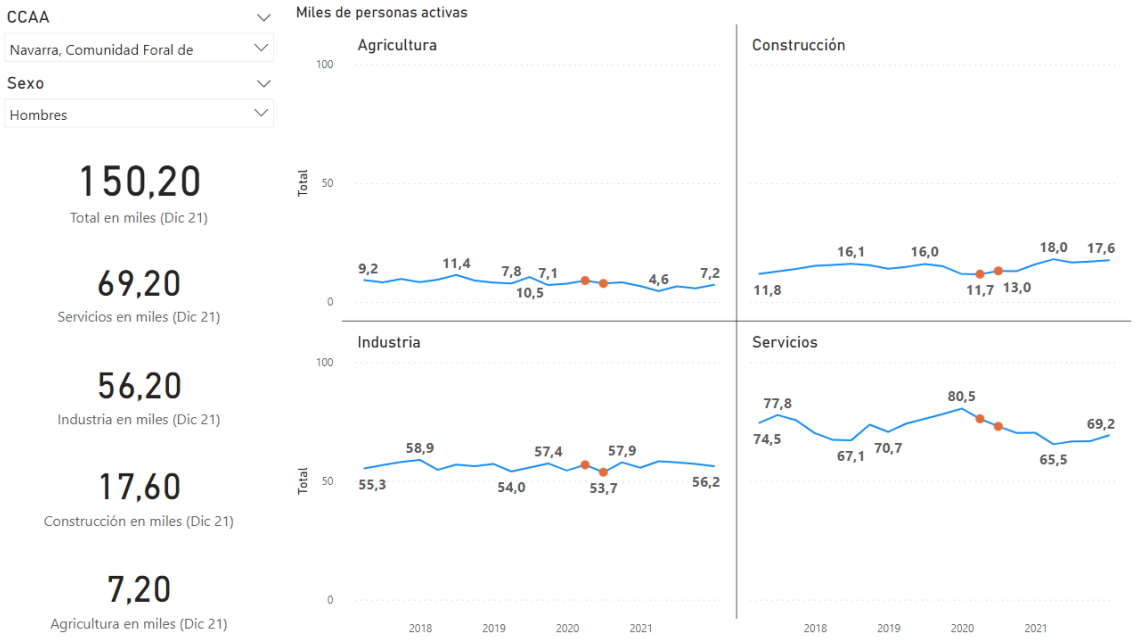

Figura 35. Hombres Activos Navarra

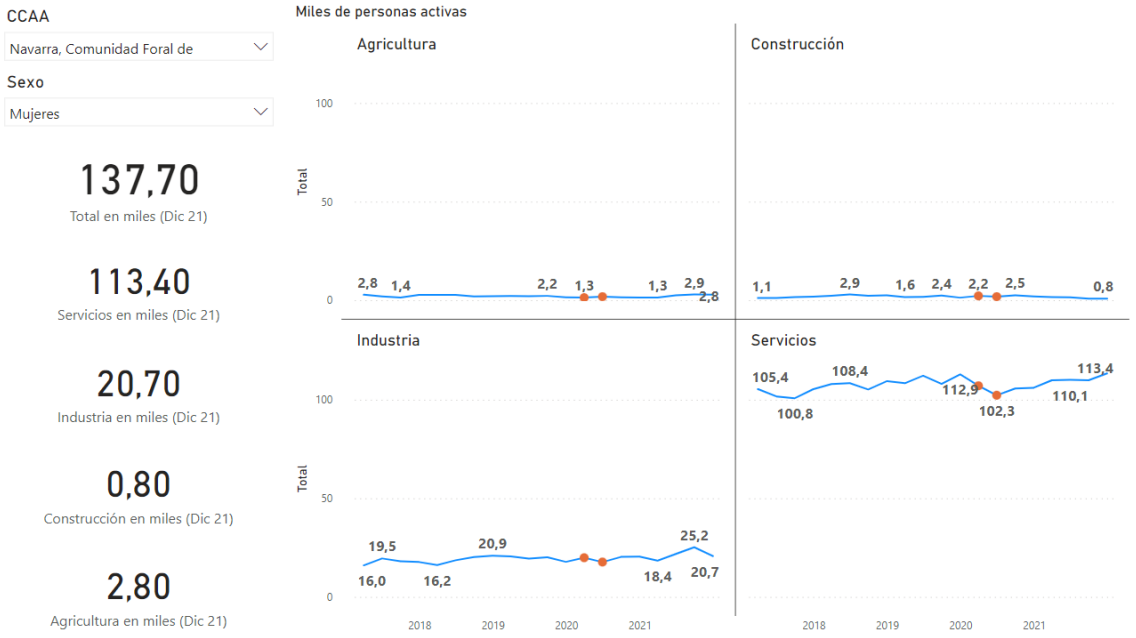

Figura 36. Mujeres Activas Navarra

- *País Vasco*

No ha sufrido mucho impacto la comunidad, en comparación con otras. El mayor golpe se lo ha llevado el sector servicios de manera proporcional entre hombre y mujeres, siendo estas últimas casi un 50% más que trabajadores masculinos.

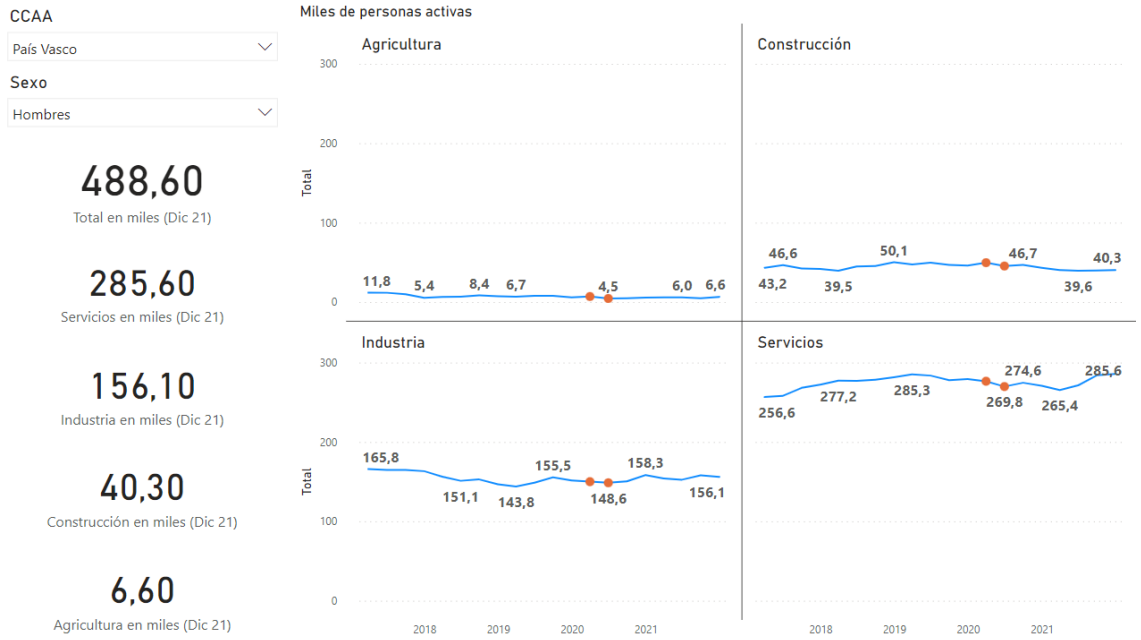

Figura 37. Hombres Activos País Vasco

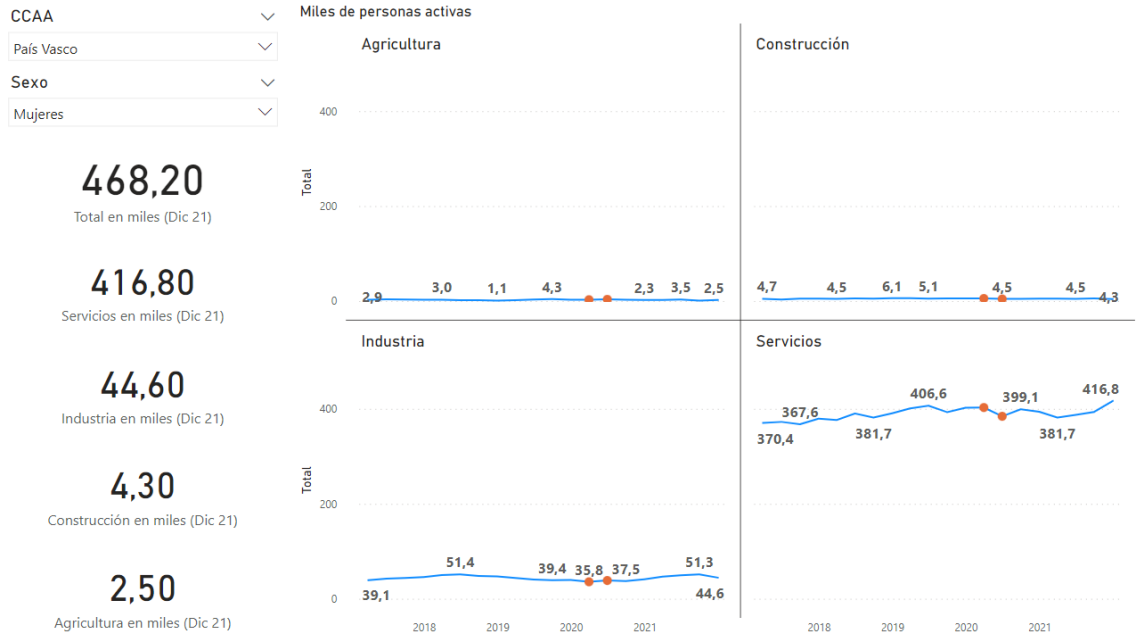

Figura 38. Mujeres Activas País Vasco

- *La Rioja*

Es la comunidad con menor número de trabajadores que hemos analizado en el proyecto. La industria pueda parecer que siga un comportamiento estacionario pero no se puede afirmar. Los hombres han sufrido disminución proporcional tanto en la industria como en el sector servicios, mientras que las mujeres se han visto más afectadas en el sector servicios.

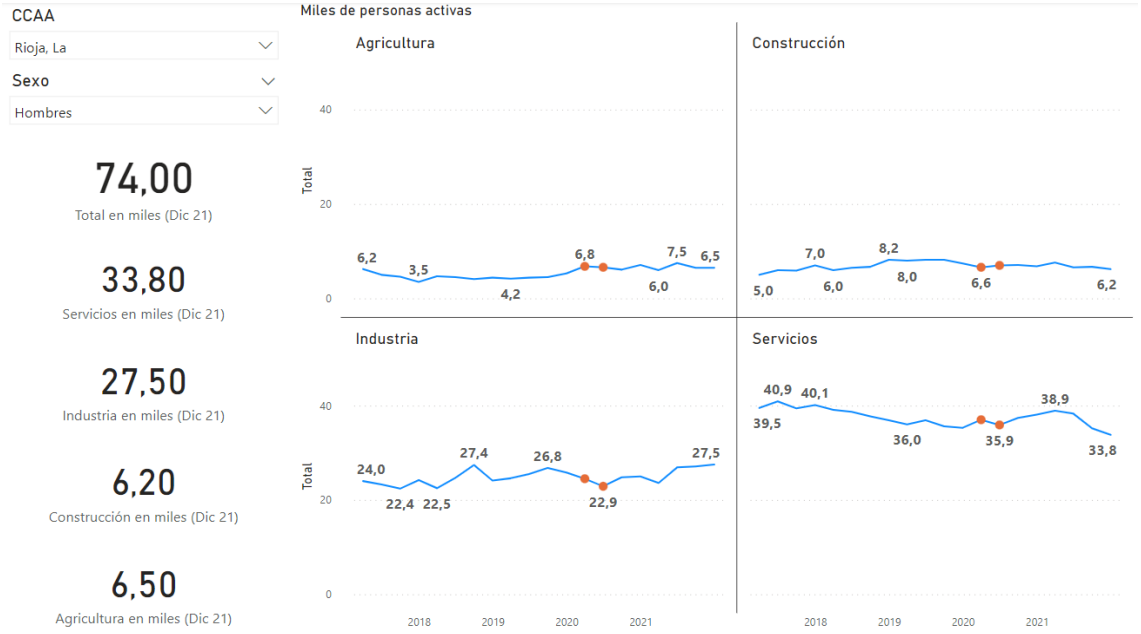

Figura 39. Hombres Activos La Rioja

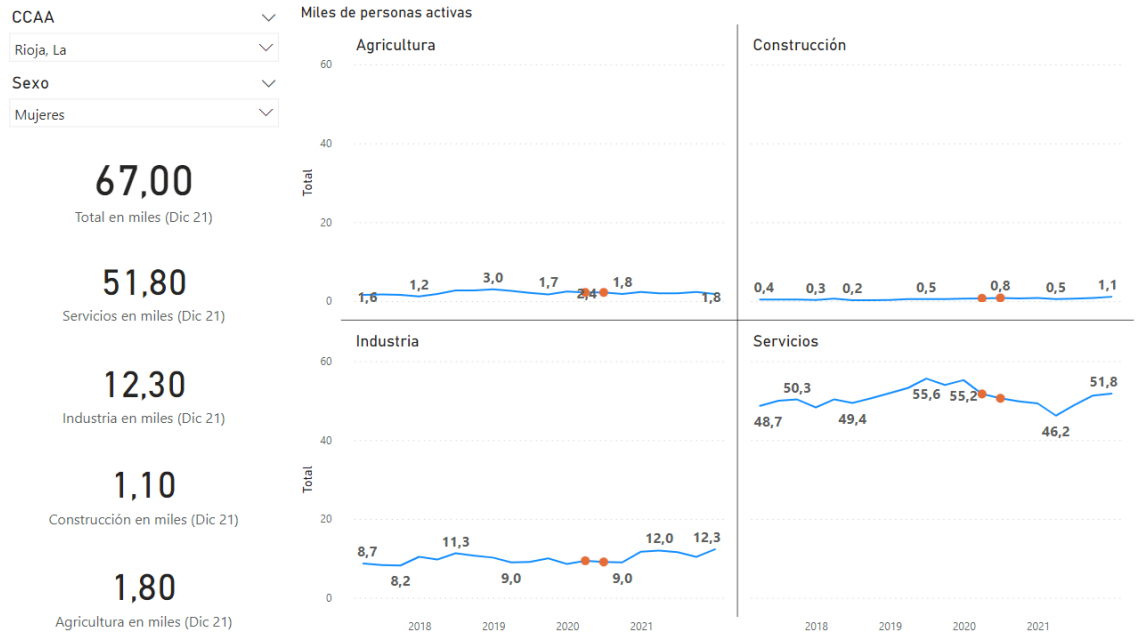

Figura 40. Mujeres Activas La Rioja

4.6.2 Resultados del Modelo

Se ha creado un modelo de machine learning en Python, utilizando las librerías y paquetes mencionados anteriormente, que tenga como objetivo realizar predicciones utilizando los algoritmos RandomForest, XG Boost y Cat Boost.

Previo a la creación del modelo, se han tenido que cambiar los tipos de columna haciendo uso de los paquetes y librerías para manejar y transformar datos, así como definir la variable 'periodo' como la variable índice, especificando que tenemos un modelo que va a predecir en función de una serie temporal. Las predicciones del modelo irán determinadas por la variable 'label' la cual hará referencia a los resultados de salida, en valores absolutos de número de personas ocupadas que vamos a obtener, diferenciando entre: Total, Servicios, Industria, Construcción y Agricultura. En el anexo final del proyecto se adjuntará el código utilizado.

- *Total*

El mejor resultado es obtenido por el algoritmo Random Forest Regressor, con una profundidad de 10 y una cantidad de 100 estimadores. El Algoritmo de Cat Boosting tiene una predicción bastante mala en comparación con los otros algoritmos.

Medida	Random Forest Regressor	Extreme Gradient Boosting	Category Boosting
Profundidad	10	3	3
Estimadores	100	500	100
RMSE	**245350.45**	297579.10	431132.39

Tabla 3. Evaluación Modelos Total

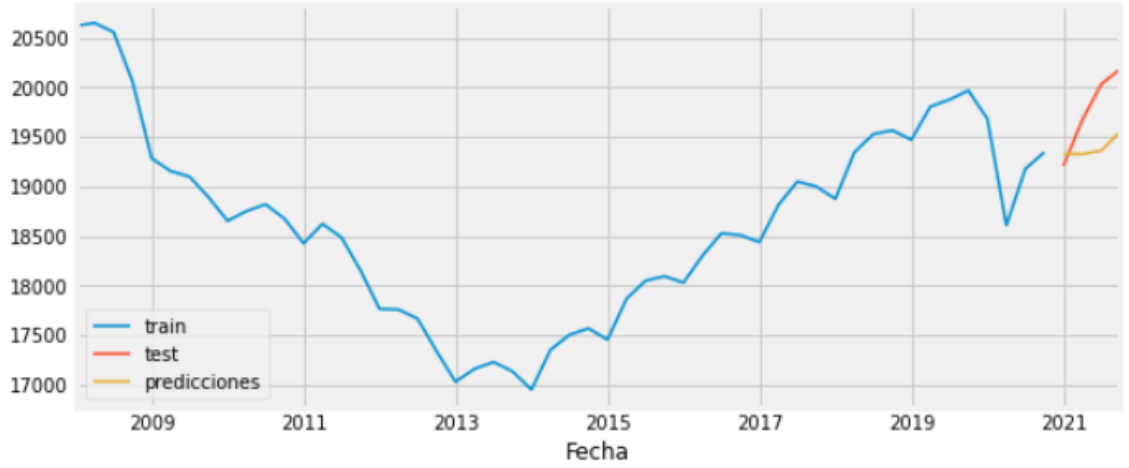

Figura 41. Evaluación Modelo Total Random forest Regressor

- *Servicios*

El mejor resultado es obtenido por el algoritmo XG Boosting con una profundidad de 10 y una cantidad de 500 estimadores. El Algoritmo de Random Forest tiene la peor predicción, pero no se aleja mucho de los otros algoritmos.

Medida	Random Forest Regressor	Extreme Gradient Boosting	Category Boosting
Profundidad	10	10	3
Estimadores	100	500	500
RMSE	201636.64	**142695.40**	158820.56

Tabla 4. Evaluación Modelos Servicios

Figura 42. Evaluación Modelos Servicios XG Boost

- *Industria*

El mejor resultado es obtenido por el algoritmo XG Boosting con una profundidad de 10 y una cantidad de 500 estimadores. El Algoritmo de Cat Boosting tiene la peor predicción muy parecida con el algoritmo XG Boosting.

Medida	Random Forest Regressor	Extreme Gradient Boosting	Category Boosting
Profundidad	3	5	3
Estimadores	500	100	500
RMSE	2178.06	**1588.09**	2192.91

Tabla 5. Evaluación Modelos Industria

Figura 43. Evaluación Modelos Industria XG Boost

- *Construcción*

El mejor resultado es obtenido por el algoritmo Random Forest Regressor con una profundidad de 10 y una cantidad de 500 estimadores. El Algoritmo de Cat Boosting tiene la peor predicción con más del doble de error.

Medida	Random Forest Regressor	Extreme Gradient Boosting	Category Boosting
Profundidad	10	3	3
Estimadores	500	500	100
RMSE	**1440.589**	2609.47	3129.46

Tabla 6. Evaluación Modelos Construcción

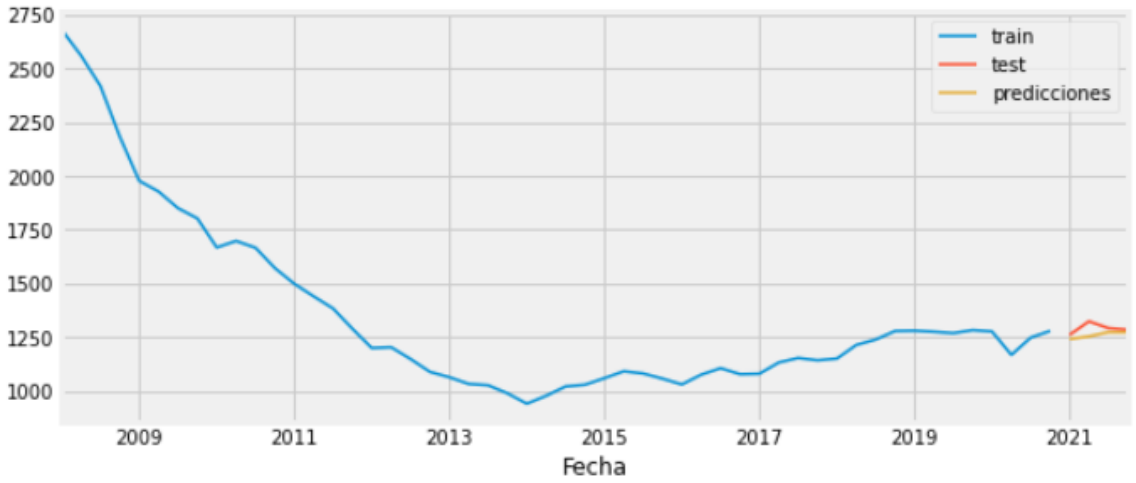

Figura 44. Evaluación Modelos Construcción Random Forest Regressor

- *Agricultura*

El mejor resultado es obtenido por el algoritmo Random Forest Regressor con una profundidad de 10 y una cantidad de 500 estimadores. El Algoritmo de Cat Boosting tiene la peor predicción muy lejana a los demás algoritmos.

Medida	Random Forest Regressor	Extreme Gradient Boosting	Category Boosting
Profundidad	5	10	3
Estimadores	500	500	500
RMSE	**1667.69**	1743.30	3427.66

Tabla 7. Evaluación Modelos Agricultura

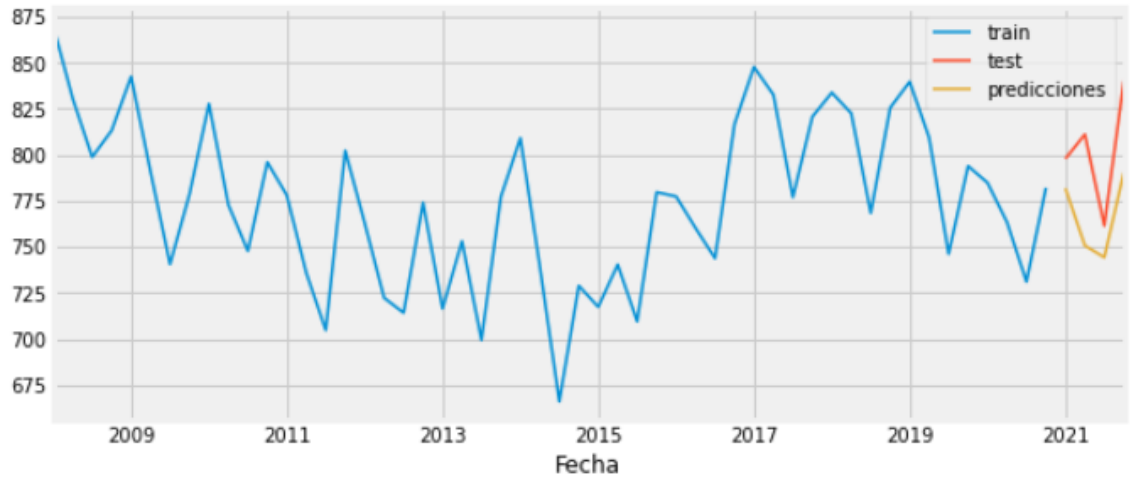

Figura 45. Evaluación Modelos Agricultura Random Forest Regressor

Como se puede comprobar, los algoritmos que mejores resultados obtienen son el Random Forest Regressor y el Extreme Gradient Boosting. Por todo esto se ha decidido realizar una predicción a futuro de cuáles serán los valores de población activa a finales del año 2022, utilizando el algoritmo de Random Forest Regressor, debido a que con dicho algoritmo se han obtenido mejores evaluaciones en tres de los cinco resultados obtenidos. Obteniendo como resultados en valores absolutos de trabajadores los siguientes datos de población activa:

Predicción	TOTAL	Servicios	Industria	Construcción	Agricultura
31/12/2022	19.652.400	14.866.725	2.712.280	1.268.718	790.527

Tabla 8. Predicciones Random Forest Regressor

Capítulo 5. DISCUSIÓN

En esta sección se discuten los resultados principales obtenidos en el capítulo anterior, así como los datos más relevantes del proyecto, tras haber descrito y analizado los resultados que se han obtenido con el desarrollo del análisis del número de personas con un estado de actividad ocupado registrados por el INE y de los resultados que se han obtenido de los modelos de predicción y los algoritmos que se han utilizado.

En primer lugar y como se ha podido observar en todos los análisis realizados, a raíz de las visualizaciones obtenidas, el sector económico que más ha sufrido el impacto ha sido el sector servicios. La crisis que ha golpeado el mercado laboral, producida por la pandemia de la COVID-19, ha impactado en gran parte en el sector de los servicios ya que debido a las restricciones estipuladas en el estado de alarma, se restringió la movilidad de todas las personas que no tuviesen que realizar actividades esenciales. Por esto como se ha expuesto anteriormente a lo largo del proyecto muchas compañías no esenciales se vieron obligadas a cerrar incapaces de implementar la metodología del teletrabajo, gracias a la cual se han salvado grandes cantidades de puestos de trabajo. En total en el trimestre del confinamiento domiciliario en España se perdieron más de un millón de puestos de trabajo, solamente en el sector servicios, de los que de cada 10 personas que perdían su trabajo 4 eran hombres y 6 mujeres aproximadamente, siendo esta tendencia en el sector algo común en la mayoría de las comunidades autónomas.

Aunque pueda parecer que las mujeres han experimentado un mayor golpe debido a la pandemia no se puede afirmar que la causa de estos resultados sea una diferenciación de género, sino una razón proporcional debido a que la mayoría de las mujeres que trabajan en España están en el sector servicios ya que de las 9,35 millones de mujeres con trabajo a finales del 2021, un total de 8,31 millones se encuentran trabajando en el sector servicios, un 88,88% de las mujeres. En el sector servicios cuenta con un total de 15,29 millones de empleados a finales del 2021 mayoritariamente mujeres con una representación del 54,35%. Por todo ello no se puede decir que las mujeres han salido más perjudicadas de la crisis por una razón de género, sino más bien por el sector principal en el que se concentra la actividad laboral de las mujeres españolas, dado que en los otros sectores, aunque no ha habido tanto impacto por culpa de la pandemia, los hombres han sufrido mayores números de pérdida de puestos de trabajo que las mujeres por el mismo motivo que sucede en el sector servicios, principalmente en el sector de la construcción.

Por contraposición se puede observar que el sector que menos ha sufrido la crisis por culpa de la pandemia de la COVID-19 ha sido el sector con la actividad más esencial de todas, la agricultura. Este sector ha sufrido pérdidas de trabajo mínimas en comparación con los otros sectores, llegando incluso a aumentar los puestos de trabajo en algunas comunidades debido a algunos productos cuya producción se lleva a cabo en una temporada concreta del año, para lo que a veces es necesaria la contratación de mayor número de empleados para poder hacer frente a la demanda.

Por último, comentar algunos casos que llaman la atención del análisis realizado en el proyecto. Estos casos afectan a las comunidades de las islas, tanto a Baleares como a Canarias.

En las Islas Baleares llama la atención como es la tendencia estacional del sector servicios dado que la gran mayoría de los contratos de trabajo se realizan de cara a la época de temporada alta en la contratación de personal para hacer frente a las grandes cantidades de turistas que van a pasar sus vacaciones a la comunidad. Es la única comunidad que registra aumentos en el sector servicios en el trimestre en el cual se decretó el estado de alarma, aunque como se puede observar esta subida es mucho menor que los otros picos de aumento laboral del resto de años en las mismas fechas.

En las otras islas, las Islas Canarias, a diferencia de otras comunidades en el sector de la agricultura, muchas de las comunidades se han recuperado muy rápido, pero en Canarias se ve una tendencia más lenta de recuperación, con una caída de empleos en el año 2021, lo que nos lleva a pensar que hay una variable externa a la pandemia por la cual esta recuperación a niveles pre pandemia no se está llevando a cabo, esto es debido a la erupción del volcán de La Palma.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

Tras haber analizado la situación en el mercado laboral en España se ha podido observar cómo la crisis que se preveía en un principio en el mercado no ha sido tanta como se esperaba ya que la recuperación ha avanzado a pasos más rápidos de los previstos inicialmente.

Se ha podido comprobar cómo el sector de los servicios ha sido el sector que más se ha visto perjudicado por la pandemia que inició a finales del año 2019 en china y que supuso el cierre temporal e incluso definitivo de algunas compañías en nuestro país. Dentro de este sector las comunidades más afectadas han sido las grandes comunidades con densidades de población como son Cataluña o Madrid con pérdidas de empleos proporcionales en materia de género.

Por otra parte cabe destacar que las previsiones de recuperación iniciales eran más negativas y a un ritmo más lento del que se ha llevado a cabo finalmente, posicionándose España al cierre del año 2021 con unas cifras similares al año pre pandemia. Cabe destacar que esto ha sido posible gracias al avance social gracias a las nuevas tecnologías, por las cuales se ha explotado en mayor medida una metodología de trabajo que no era muy popular hasta ahora y que cada vez más ocupados optan por esta modalidad, el teletrabajo. También se han evitado mayores pérdidas de empleo gracias a las empresas que declararon ERTEs, ya que pudieron retomar la actividad laboral pasados los estados de alarma, en la época conocida como la “nueva normalidad”.

Para terminar con las conclusiones del modelo se comprueba que debido a la falta de datos, ya que las muestras se toman trimestralmente, esto hace que los modelos no realicen predicciones más acertadas lo que ha supuesto una complicación extra a la hora de realizar el proyecto.

Concluyendo, a pesar de la complejidad que ha podido suponer el proyecto en alguna de sus fases, se ha podido cumplir con los objetivos marcados al principio del mismo a pesar de que los resultados del modelo no sean los más acertados por falta de datos e incluso de otras posibles variables exógenas. Adicionalmente, el trabajo realizado ha servido para comprender mejor el mercado laboral y las distribuciones de personas activas en distintos sectores económicos y comunidades autónomas. Se puede concluir que el proyecto ha tenido una valoración positiva.

6.2 Conclusiones personales

Finalizado el proyecto se exponen a continuación las conclusiones a nivel personal del mismo.

Desde un principio sabía que el proyecto iba a ir ligado al impacto producido por la pandemia del COVID-19 en los principales sectores económicos de España y cuando se lo expuse a mi tutor lo vio con buenos ojos a la hora de enfrentarnos al análisis del mercado laboral ya que este había tenido mucha problemática. Personalmente la mayor complicación que destacaría del proyecto a nivel personal ha sido la necesidad de una buena organización para poder realizar el proyecto final, compaginando las actividades universitarias, las prácticas laborales y

las asignaturas finales de cara a finalizar los estudios. La idea principal ha ido desarrollándose y evolucionando hasta el resultado que se obtiene al final de este análisis. Por todo ello la superación de momentos y fases difíciles a lo largo del proceso suponen una gran satisfacción.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

Tras la conclusión del proyecto y de cara al desarrollo futuro, se proponen una serie de ideas que puedan mejorar el proyecto, actualizarlo y darle una mayor viabilidad. Estas propuestas son las siguientes:

- Nuevo análisis evaluando esta vez los perfiles que se han visto afectados en mayor medida en relación al rango de edad de las personas activas para ver en qué perfil generacional ha sido el impacto de la crisis del mercado laboral.
- Creación y desarrollo de una página web interactiva para los usuarios en la cual se pueda navegar y realizar distintas visualizaciones con conexión a la base de datos generada por las descargas de información trimestral del Instituto Nacional de Estadística.
- Investigación y evaluación de nuevos algoritmos, como por ejemplo de Redes Neuronales, para la predicción de valores más cercanos a la realidad de los datos obtenidos e incluso la valoración de algunas variables exógenas que puedan tener peso en nuestro modelo.
- Expansión del modelo para predecir valores por comunidades autónomas en vez de a nivel nacional.
- Obtención de datos de distintas fuentes por la que los valores recopilados en nuestra base de datos pasen de ser trimestrales a mensuales y preferentemente semanales. Esto tendrá impacto directo en las predicciones de modelos dado que a mayor cantidad de datos se va a poder realizar mejores predicciones, más cercanas a la realidad.
- Ampliación de nuestra base de datos con nuevas variables como datos de entrada, las cuales puedan tener mayor peso a la hora de aplicar los modelos de aprendizaje automático, capaces de explicar en mejor medida la serie temporal de los datos ya que el mercado laboral no sigue estrictamente una línea regresiva sino que puede ser afectada por diversos factores.

Capítulo 8. REFERENCIAS

- Alf, A. c. (4 de Octubre de 2020). Obtenido de La librería Matplotlib: https://aprendeconalf.es/docencia/python/manual/matplotlib/
- Cardellino, F. (20 de Marzo de 2021). Free CodeCamp. Obtenido de La guía definitiva del paquete NumPy para computación científica en Python: https://www.freecodecamp.org/espanol/news/la-guia-definitiva-del-paquete-numpy-para-computacion-cientifica-en-python/
- CatBoost. (2022). Docs Cat Boost . Obtenido de https://catboost.ai/en/docs/
- COMAV, B. a. (s.f.). Obtenido de Pandas: https://bioinf.comav.upv.es/courses/linux/python/pandas.html
- Data, E. (9 de Diciembre de 2021). Ep Data. Obtenido de El número de trabajadores afectados por un ERE: https://www.epdata.es/datos/trabajadores-afectados-ere-graficos/450
- España, B. d. (Julio de 2020). Banco de España. Obtenido de Impacto en el empleo: https://www.bde.es/f/webbde/GAP/Secciones/SalaPrensa/IntervencionesPublicas/DirectoresGenerales/economia/Arc/arce080720.pdf
- España, B. d. (Marzo de 2021). Banco de España. Obtenido de Impacto diferencial por sexos: https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEconomico/Informe%20trimestral/21/Recuadros/Fich/be2103-it-Rec5.pdf
- Estadística, I. N. (2021). INE. Obtenido de Información estadística para el análisis del impacto de la crisis COVID-19: https://www.ine.es/covid/covid_economia.htm
- Huambachano, J. F. (25 de Septiembre de 2017). Scrum.org. Obtenido de ¿Qué es SCRUM?: https://www.scrum.org/resources/blog/que-es-scrum
- Learn, S. (2022). Scikit Learn. Obtenido de sklearn.ensemble.RandomForestRegressor: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
- Microsoft. (2022). Power BI. Obtenido de ¿Qué es Power BI?: https://powerbi.microsoft.com/es-es/what-is-power-bi/
- Rodrigo, J. A. (2021). git hub. Obtenido de skforecast: https://joaquinamatrodrigo.github.io/skforecast/0.4.3/index.html
- Rodrigo, J. A. (Marzo de 2022). Ciencia de datos. Obtenido de Skforecast: forecasting series temporales con Python y Scikit-learn:

https://www.cienciadedatos.net/documentos/py27-forecasting-series-temporales-python-scikitlearn.html

- Rodrigo, J. A. (Marzo de 2022). Ciencia de datos. Obtenido de Forecasting series temporales con gradient boosting: Skforecast, XGBoost, LightGBM y CatBoost: https://www.cienciadedatos.net/documentos/py39-forecasting-series-temporales-con-skforecast-xgboost-lightgbm-catboost.html

- Seabold, S. a. (2010). Statsmodels. Obtenido de statsmodels: Econometric and statistical modeling with python: https://www.statsmodels.org/stable/index.html

- Social, M. d. (Diciembre de 2020). Secretaría de Estado de Empleo y Economía Social. Obtenido de Nota Impacto COVID: https://www.mites.gob.es/ficheros/ministerio/estadisticas/documentos/Nota_impacto_COVID_Diciembre-2020.pdf

- Vega, J. B. (16 de Agosto de 2020). Medium. Obtenido de Tutorial: XGBoost en Python: https://medium.com/@jboscomendoza/tutorial-xgboost-en-python-53e48fc58f73

Capítulo 9. ANEXOS

Código Modelo Random Forest Regresor

```
#!/usr/bin/env python
# coding: utf-8

# Tratamiento de datos
# ==============================================================================
import numpy as np
import pandas as pd

# Gráficos
# ==============================================================================
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
plt.rcParams['lines.linewidth'] = 1.5
get_ipython().run_line_magic('matplotlib', 'inline')

# Modelado y Forecasting
# ==============================================================================
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline

from skforecast.ForecasterAutoreg import ForecasterAutoreg
from skforecast.ForecasterAutoregCustom import ForecasterAutoregCustom
from skforecast.ForecasterAutoregMultiOutput import ForecasterAutoregMultiOutput
from skforecast.model_selection import grid_search_forecaster
from skforecast.model_selection import backtesting_forecaster

from joblib import dump, load

# Descarga de datos
# ==============================================================================
datos = pd.read_csv('activa.csv', sep=';')
datos.info()

# Preparación del dato
# ==============================================================================
datos['Fecha'] = pd.to_datetime(datos['Fecha'], format='%d/%m/%Y')
datos['Agricultura'] = datos['Agricultura'].str.replace(',', '.')
datos['Agricultura'] = datos['Agricultura'].astype(float)
```

```
datos['Industria'] = datos['Industria'].str.replace(',', '.')
datos['Industria'] = datos['Industria'].astype(float)
datos['Construccion'] = datos['Construccion'].str.replace(',', '.')
datos['Construccion'] = datos['Construccion'].astype(float)
datos['Servicios'] = datos['Servicios'].str.replace(',', '.')
datos['Servicios'] = datos['Servicios'].astype(float)
datos['Total'] = datos['Total'].str.replace(',', '.')
datos['Total'] = datos['Total'].astype(float)
datos = datos.set_index('Fecha')
datos = datos.asfreq('Q')
datos = datos.sort_index()
datos.info()

# Verificar si han aparecido missing values tras esta transformación.
# ======================================================================
print(f'Número de filas con missing values: {datos.isnull().any(axis=1).mean()}')

# Verificar que un índice temporal está completo
# ======================================================================
(datos.index == pd.date_range(
                start = datos.index.min(),
                end   = datos.index.max(),
                freq  = datos.index.freq)
).all()

# Etiqueta para predecir introducela en label
label = ''

# Separación datos train-test
# ======================================================================
steps = 4
datos_train = datos[:-steps]
datos_test  = datos[-steps:]

print(f"Fechas train : {datos_train.index.min()} --- {datos_train.index.max()}
(n={len(datos_train)})")
print(f"Fechas test : {datos_test.index.min()} --- {datos_test.index.max()}
(n={len(datos_test)})")

fig, ax = plt.subplots(figsize=(9, 4))
datos_train[label].plot(ax=ax, label=label)
datos_test[label].plot(ax=ax, label=label)
ax.legend();

# Crear y entrenar forecaster
# ======================================================================
```

```
forecaster = ForecasterAutoreg(
                regressor = RandomForestRegressor(random_state=123),
                lags = 6
             )

forecaster.fit(y=datos_train[label])
forecaster
# Predicciones
# ======================================================================
steps = 4
predicciones = forecaster.predict(steps=steps)
predicciones.head(5)

# Gráfico
# ======================================================================
fig, ax = plt.subplots(figsize=(9, 4))
datos_train[label].plot(ax=ax, label='train')
datos_test[label].plot(ax=ax, label='test')
predicciones.plot(ax=ax, label='predicciones')
ax.legend();

# Error test
# ======================================================================
error_mse = mean_squared_error(
                y_true = datos_test[label],
                y_pred = predicciones
            )

print(f"Error de test (mse): {error_mse}")

# Grid search de hiperparámetros
# ======================================================================
steps = 4
forecaster = ForecasterAutoreg(
                regressor = RandomForestRegressor(random_state=123),
                lags      = 6
             )

# Lags utilizados como predictores
lags_grid = [5, 10]

# Hiperparámetros del regresor
param_grid = {'n_estimators': [100, 500],
              'max_depth': [3, 5, 10]}

resultados_grid = grid_search_forecaster(
```

```
                forecaster        = forecaster,
                y                 = datos_train[label],
                param_grid        = param_grid,
                lags_grid         = lags_grid,
                steps             = steps,
                refit             = True,
                metric            = 'mean_squared_error',
                initial_train_size = int(len(datos_train)*0.5),
                fixed_train_size   = False,
                return_best       = True,
                verbose           = False
            )

# Resultados Grid Search
# ==============================================================================
resultados_grid

# Crear y entrenar forecaster con mejores hiperparámetros
# Rellenar max_depth y n_estimators con los valores resultados_grid
# ==============================================================================
regressor = RandomForestRegressor(max_depth= , n_estimators= , random_state=123)
forecaster = ForecasterAutoreg(
             regressor = regressor,
             lags      = 6
         )

forecaster.fit(y=datos_train[label])

# Predicciones
# ==============================================================================
predicciones = forecaster.predict(steps=steps)

# Gráfico
# ==============================================================================
fig, ax = plt.subplots(figsize=(9, 4))
datos_train[label].plot(ax=ax, label='train')
datos_test[label].plot(ax=ax, label='test')
predicciones.plot(ax=ax, label='predicciones')
ax.legend();

# Error de test
# ==============================================================================
error_mse = mean_squared_error(
             y_true = datos_test[label],
             y_pred = predicciones
         )
```

```
print(f"Error de test (mse) {error_mse}")
```

Código Modelo XG Boosting y CatBoosting

```
# Tratamiento de datos
# ==============================================================================
import numpy as np
import pandas as pd

# Gráficos
# ==============================================================================
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.graphics.tsaplots import plot_pacf
import plotly.express as px
plt.style.use('fivethirtyeight')
plt.rcParams['lines.linewidth'] = 1.5
get_ipython().run_line_magic('matplotlib', 'inline')

# Modelado y Forecasting
# ==============================================================================
from xgboost import XGBRegressor
from catboost import CatBoostRegressor

from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import make_pipeline
from sklearn.metrics import mean_squared_error

from skforecast.ForecasterAutoreg import ForecasterAutoreg
from skforecast.ForecasterAutoregMultiOutput import ForecasterAutoregMultiOutput
from skforecast.model_selection import grid_search_forecaster
from skforecast.model_selection import backtesting_forecaster

from joblib import dump, load

# Descarga de datos
# ==============================================================================
datos = pd.read_csv('activa.csv', sep=';')
datos.info()

# Preparación del dato
# ==============================================================================
datos['Fecha'] = pd.to_datetime(datos['Fecha'], format='%d/%m/%Y')
```

```
datos['Agricultura'] = datos['Agricultura'].str.replace(',', '.')
datos['Agricultura'] = datos['Agricultura'].astype(float)
datos['Industria'] = datos['Industria'].str.replace(',', '.')
datos['Industria'] = datos['Industria'].astype(float)
datos['Construccion'] = datos['Construccion'].str.replace(',', '.')
datos['Construccion'] = datos['Construccion'].astype(float)
datos['Servicios'] = datos['Servicios'].str.replace(',', '.')
datos['Servicios'] = datos['Servicios'].astype(float)
datos['Total'] = datos['Total'].str.replace(',', '.')
datos['Total'] = datos['Total'].astype(float)
datos = datos.set_index('Fecha')
datos = datos.asfreq('Q')
datos = datos.sort_index()
datos.info()

# Verificar si han aparecido missing values tras esta transformación.
# ==============================================================================
print(f'Número de filas con missing values: {datos.isnull().any(axis=1).mean()}')

# Verificar que un índice temporal está completo
# ==============================================================================
(datos.index == pd.date_range(
                start = datos.index.min(),
                end   = datos.index.max(),
                freq  = datos.index.freq)
).all()

# Etiqueta para predecir intuducela en label
label = ''

# Separación datos train-val-test
# ==============================================================================
fin_train = '2017-12-31'
fin_validacion = '2020-12-31'
datos_train = datos.loc[: fin_train, :]
datos_val   = datos.loc[fin_train:fin_validacion, :]
datos_test  = datos.loc[fin_validacion:, :]

print(f"Fechas train : {datos_train.index.min()} --- {datos_train.index.max()}
(n={len(datos_train)})")
print(f"Fechas validacion : {datos_val.index.min()} --- {datos_val.index.max()}
(n={len(datos_val)})")
print(f"Fechas test : {datos_test.index.min()} --- {datos_test.index.max()}
(n={len(datos_test)})")

fig, ax = plt.subplots(figsize=(9, 4))
```

```
datos_train[label].plot(ax=ax, label=label)
datos_val[label].plot(ax=ax, label=label)
datos_test[label].plot(ax=ax, label=label)
ax.legend();

# Crear forecaster XGBoost
# ==============================================================================
forecaster = ForecasterAutoreg(
             regressor = XGBRegressor(random_state=123),
             lags = 6
           )

forecaster

# Grid search de hiperparámetros
# ==============================================================================
# Hiperparámetros del regresor
steps = 4
param_grid = {
   'n_estimators': [100, 500],
   'max_depth': [3, 5, 10],
   'learning_rate': [0.01, 0.1]
}

# Lags utilizados como predictores
lags_grid = [5, 10]

resultados_grid = grid_search_forecaster(
      forecaster        = forecaster,
      y                 = datos.loc[:fin_validacion, label], # conjunto de train y validación
      param_grid        = param_grid,
      lags_grid         = lags_grid,
      steps             = steps,
      refit             = False,
      metric            = 'mean_squared_error',
            initial_train_size  = int(len(datos_train)), # El modelo se entrena con los datos de
entrenamiento
      return_best       = True,
      verbose           = False
)

# Resultados Grid Search
# ==============================================================================
resultados_grid.head(10)

# Backtesting
```

```
# ==============================================================================
metric, predicciones = backtesting_forecaster(
    forecaster = forecaster,
    y         = datos[label],
    initial_train_size = len(datos.loc[:fin_validacion]),
    steps     = steps,
    refit     = False,
    metric    = 'mean_squared_error',
    verbose   = False # Change to True to see detailed information
)

print(f"Error de backtest: {metric}")

# Predicciones test
# ==============================================================================
predicciones

# Gráfico estatico predicciones test
# ==============================================================================
fig, ax = plt.subplots(figsize=(11, 4))
datos_test[label].plot(ax=ax, label='test')
predicciones['pred'].plot(ax=ax, label='predicciones')
ax.legend();

# Gráfico completo
# ==============================================================================
fig, ax = plt.subplots(figsize=(9, 4))
datos_train[label].plot(ax=ax, label='train')
datos_val[label].plot(ax=ax, label='validacion')
datos_test[label].plot(ax=ax, label='test')
predicciones.plot(ax=ax, label='predicciones')
ax.legend();

# Crear forecaster CatBoost
# ==============================================================================
forecaster = ForecasterAutoreg(
             regressor = CatBoostRegressor(random_state=123, silent=True),
             lags = 6
         )

forecaster

# Grid search de hiperparámetros
# ==============================================================================
# Hiperparámetros del regresor
param_grid = {
```

```
    'n_estimators': [100, 500],
    'max_depth': [3, 5, 10],
    'learning_rate': [0.01, 0.1]
}

# Lags utilizados como predictores
lags_grid = [5, 10]

resultados_grid = grid_search_forecaster(
                      forecaster         = forecaster,
                      y                  = datos.loc[:fin_validacion, label],
                      param_grid         = param_grid,
                      lags_grid          = lags_grid,
                      steps              = steps,
                      refit              = False,
                      metric             = 'mean_squared_error',
                      initial_train_size = int(len(datos_train)),
                      return_best        = True,
                      verbose            = False
                  )

# Resultados Grid Search
# ==============================================================================
resultados_grid.head(10)

# Backtesting
# ==============================================================================
metric, predicciones = backtesting_forecaster(
    forecaster         = forecaster,
    y                  = datos[label],
    initial_train_size = len(datos.loc[:fin_validacion]),
    steps              = steps,
    refit              = False,
    metric             = 'mean_squared_error',
    verbose            = False
)

print(f"Error de backtest: {metric}")

# Gráfico estatico predicciones test
# ==============================================================================
fig, ax = plt.subplots(figsize=(11, 4))
datos_test[label].plot(ax=ax, label='test')
predicciones['pred'].plot(ax=ax, label='predicciones')
ax.legend();
```

```
# Gráfico completo
# ==============================================================================
fig, ax = plt.subplots(figsize=(9, 4))
datos_train[label].plot(ax=ax, label='train')
datos_val[label].plot(ax=ax, label='validacion')
datos_test[label].plot(ax=ax, label='test')
predicciones.plot(ax=ax, label='predicciones')
ax.legend();
```