UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Desarrollo de un modelo del precio de la vivienda en Valencia por barrios mediante técnicas de análisis multivariante y minería de datos

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Legido Casanoves, José

Tutor/a: Conchado Peiró, Andrea

CURSO ACADÉMICO: 2022/2023

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Desarrollo de un modelo del precio de la vivienda en Valencia por barrios mediante técnicas de análisis multivariante y minería de datos

Dpto. de Estadística e Investigación Operativa

Aplicadas y Calidad

Trabajo fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

Autor: José Legido Casanoves

Directora: Andrea Conchado Peiró

Curso académico 2022-2023

Resumen

En los últimos años la ciudad de Valencia se ha convertido en un rentable objetivo de los inversores de vivienda. En el mercado inmobiliario existen numerosos factores que determinan el precio de las viviendas, pero sin duda la localización en el área urbana resulta determinante. Cada distrito de la ciudad ofrece distintos niveles de accesibilidad a servicios y dispone de determinadas redes de transporte y comunicaciones que, junto con las características particulares de cada inmueble, influyen en la decisión final de los compradores. Muchos trabajos precedentes en esta línea de investigación presentan modelos econométricos basados en características particulares de cada vivienda. Por otro lado, la influencia de la localización en el precio de la vivienda es un resultado ampliamente contrastado. A pesar de ello, existe una carencia de estudios específicos sobre la influencia de las características de cada barrio en el precio agregado de las viviendas construidas en el mismo. El objetivo de este trabajo es desarrollar un modelo del precio de la vivienda en la ciudad de Valencia por barrios. Para ello, se analizarán datos publicados por el Ayuntamiento de Valencia que contienen estadísticas oficiales de los barrios mediante técnicas de reducción de la dimensionalidad y modelos supervisados de aprendizaje automático. Los resultados esperados aportarán un nuevo enfoque a los clásicos modelos econométricos empleados con este fin durante las últimas décadas y ampliarán el conocimiento existente sobre el modo en que la localización influye en el precio de la vivienda.

Palabras clave: Vivienda; Localización; Precio; Análisis multivariante de datos; Minería de datos.

Resum

En els últims anys la ciutat de València s'ha convertit en un rendible objectiu dels inversors d'habitatge. En el mercat immobiliari existeixen nombrosos factors que determinen el preu dels habitatges, però sens dubte la localització en l'àrea urbana resulta determinant. Cada districte de la ciutat ofereix diferents nivells d'accessibilitat a serveis i disposa de determinades xarxes de transport i comunicacions que, juntament amb les característiques particulars de cada immoble, influeixen en la decisió final dels compradors. Molts treballs precedents en aquesta línia d'investigació presenten models economètrics basats en característiques particulars de cada habitatge. D'altra banda, la influència de la localització en el preu de l'habitatge és un resultat àmpliament contrastat. Malgrat això, existeix una manca d'estudis específics sobre la influència de les característiques de cada barri en el preu agregat dels habitatges construïts en aquest. L'objectiu d'aquest treball és desenvolupar un model del preu de l'habitatge a la ciutat de València per barris. Per a això, s'analitzaran dades publicades per l'Ajuntament de València que contenen estadístiques oficials dels barris mitjançant tècniques de reducció de la dimensionalitat i models supervisats d'aprenentatge automàtic. Els resultats esperats aportaran un nou enfocament als clàssics models economètrics emprats a aquest efecte durant les últimes dècades i ampliaran el coneixement existent sobre la manera en què la localització influeix en el preu de l'habitatge.

Paraules clau: Habitatge; Localització; Preu; Anàlisi multivariant de dades; Minería de dades.

Abstract

In recent years the city of Valencia has become a profitable target for property investors. In the real estate market there are numerous factors that determine the price of housing, but the location in the urban area is undoubtedly a determining factor. Each district of the city offers different levels of accessibility to services and has certain transport and communications networks which, together with the particular characteristics of each property, influence the final decision of buyers. Many previous works in this line of research present econometric models based on the particular characteristics of each property. On the other hand, the influence of location on house prices is a widely contrasted result. Despite this, there is a lack of specific studies on the influence of the characteristics of each neighbourhood on the aggregate price of the dwellings built in it. The aim of this paper is to develop a model of house prices in the city of Valencia by neighbourhood. To this end, data published by the Valencia City Council containing official neighbourhood statistics will be analysed using dimensionality reduction techniques and supervised machine learning models. The expected results will provide a new approach to the classical econometric models used for this purpose during the last decades and will extend the existing knowledge on how location influences housing prices.

Keywords: Housing; Location; Price; Multivariate data analysis; Data mining.

ÍNDICE

Índice de figuras

Índice de tablas

1. Introducción

1.1 Aproximación al problema

El mercado inmobiliario es uno de los sectores que más dinero mueve en la economía española, este último año se estima que en España se realizaron más de 600.000 operaciones de compraventa, siendo uno de los países europeos con mayor inversión inmobiliaria.

La inversión inmobiliaria en España en 2022 ha crecido un 27% respecto a 2021, consiguiendo un récord histórico de inversión directa al superar los 15.200 millones. Los mayores inversores han sido los fondos de inversión con una cuota de mercado de 67%, seguido por las compañías aseguradoras con una cuota del 10% y las empresas inmobiliarias con un 9% de cuota de mercado. Cabe destacar que la mayoría de la inversión en activos inmobiliarios en España procede del extranjero (mayoritariamente Europa y EEUU) con un 57%, frente al 43% de los inversores españoles (Forbes, 2022).

Los factores que hacen que España sea uno de los países europeos más atractivos para invertir en el mercado inmobiliario son: su posición geográfica, el turismo, los precios asequibles y la rentabilidad.

Dentro del mercado inmobiliario, el segmento residencial es el que más ha crecido, con un incremento del 10% respecto a 2021 (Idealista, 2022). Este crecimiento se debe sobre todo al primer semestre del año, donde la decisión del Banco central europeo de subir los tipos de interés ocasiono una aceleración en la toma de decisión de compra de viviendas.

Las ciudades que han recibido mayor inversión en inmuebles residenciales han sido Barcelona, Madrid, Málaga, Valencia, Sevilla, Murcia, Girona y Baleares.

A pesar de que la situación actual no ha afectado en gran medida al mercado inmobiliario, ya que el número de operaciones de compraventa ha crecido bastante, llegando casi a un nivel similar que, en 2007, sí que se ha apreciado un aumento en el precio de la vivienda.

Esto se debe a la subida de los materiales de construcción y el coste de financiación, lo que ha ocasionado que el precio de la vivienda de obra nueva suba un 7,1% respecto a 2021. Por lo tanto, existe una escasez de oferta de viviendas (obra nueva y segunda mano) que no puede abastecer a la demanda de compra. (Sociedad de tasación,2023).

En resumen, el precio de la vivienda se ha visto encarecido debido al aumento de los costes de los materiales en una etapa con alta demanda de vivienda, junto a una escalada de los tipos de interés, que ha acelerado muchas decisiones de compra ante la posibilidad de futuros incrementos. Aunque se espera, que a medida que los créditos hipotecarios se vayan encareciendo la demanda de viviendas se vaya moderando, lo que ocasionara que los precios se estabilicen.

Las comunidades autónomas que han sufrido un mayor incremento del precio de la vivienda de segunda mano respecto a 2021 han sido: Navarra (16,9%), Baleares (13,8%), Canarias (10,9%) y Comunidad valenciana (8,8%) (Fotocasa, 2022).

Si se relaciona la situación actual del mercado inmobiliario con la situación que se vivió en 2008, se aprecia que son completamente opuestas, debido a que en 2008 existía una gran facilidad de financiación debido a los tipos de interés muy bajos, y una gran oferta de viviendas.

Por último, si nos centramos en la ciudad de Valencia, se puede comprobar que este 2022 ha sido un gran año para el mercado inmobiliario de la ciudad, consiguiendo más de 38000 operaciones de compraventa, un número de operaciones similar al de 2007.

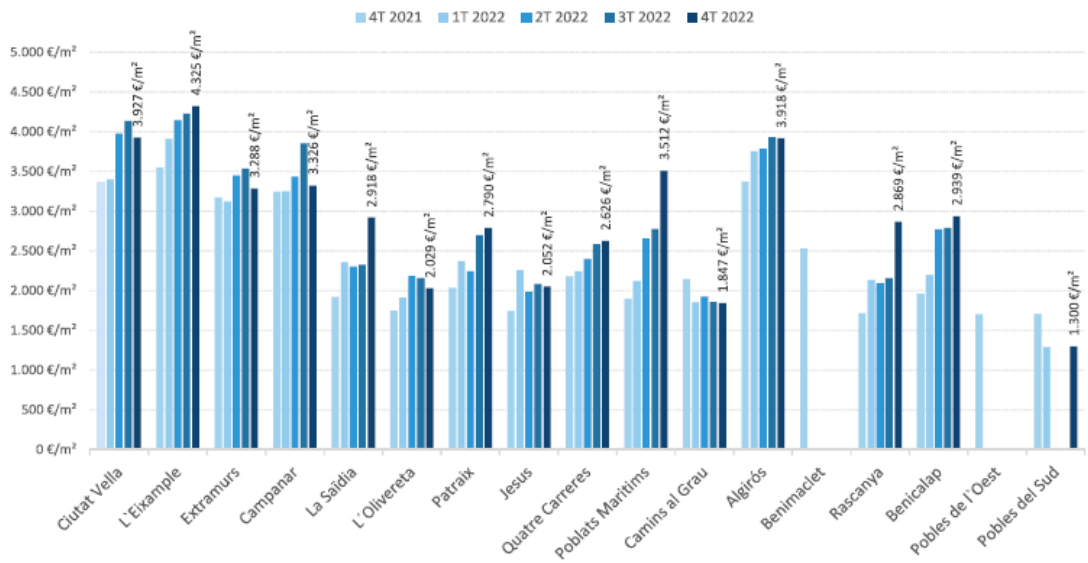
Además, cabe destacar que Valencia ha sido posicionada como la segunda ciudad más rentable para invertir en el mercado del alquiler de Europa con una rentabilidad de 5,9%, siendo los distritos que han recibido mayor inversión: Poblats maritims, Ciutat vella y Benimaclet (elEconomista, 2022).

Por este motivo, el colegio de agentes de la propiedad inmobiliaria (API) de Valencia, ha asegurado que la inversión extranjera se ha incrementado en gran medida en la ciudad de Valencia debido al incremento de las rentas de alquiler, que han aumentado de un precio medio de 875€ a 1201€ solo en dos años año, lo que supone un 38% de incremento en el alquiler (Catedra Observatorio vivienda UPV, 2022). Se estima que el 40% del total de compra de viviendas en Valencia se realizan por personas extranjeras, lo que también ha afectado al incremento de los precios (Llorca, 2022).

Este aumento del alquiler ha ocasionado que algunos ciudadanos no puedan asumir dicho incremento, lo que ha provocado que estas viviendas se destinen al alquiler vacacional, como se ha podido comprobar en los distritos de Ciutat Vella, L'eixample o Extramurs.

Cabe remarcar, que el precio de la vivienda usada en Valencia se ha incrementado un 10,6% respecto a 2021 y el precio de la vivienda de obra nueva en la ciudad se ha incrementado un 26%, debido a la desigualdad entre la oferta y la demanda de vivienda, seguido por un incremento de los costes de los materiales y una escasez de suelo disponible (Catedra Observatorio vivienda UPV, 2022).

Figura 1. Precio unitario medio de la vivienda de obra nueva por distritos de Valencia



Fuente: Catedra Observatorio vivienda UPV

Además, aparte de los factores que se han visto que afectan al precio de la vivienda, la localización es el factor que más influye en el precio de las viviendas, ya que una vivienda con unas características físicas determinadas cambia en gran medida su precio según donde este ubicada, por lo tanto, es importante preguntarse qué características de los barrios son las que aportan mayor valor al precio medio de las viviendas

1.2 Motivación

Actualmente, existen muchos estudios sobre el precio de la vivienda individual, pero la mayoría solo se centran en las características físicas de la vivienda (superficie, altura, habitaciones, baños, terraza, ascensor o garaje), como los principales factores que influyen en el precio, sin embargo, la localización es clave.

La localización de la vivienda es el factor que más repercute en el precio de una vivienda. Debido a que los clientes cuando buscan una vivienda, lo que priorizan es la disponibilidad de servicios, lo que conlleva cercanía al puesto de trabajo, buenos colegios cercanos, bien comunicado, parques, servicios básicos cercanos como bancos, supermercados, comercios, etc. Por lo tanto, dan prioridad al entorno por delante de las características físicas.

Este trabajo va más allá de solo estudiar como las características de la vivienda influyen en el precio, sino que se centra en como la localización afecta al precio de las viviendas, para ello se estudian las características de los barrios para observar cómo afectan al precio medio de las viviendas en el barrio.

La mayoría de los estudios sobre el precio de la vivienda que se han llevado a cabo, se basan en modelos econométricos, los cuales buscan predecir el resultado basándose en el histórico de los datos conocidos. Sin embargo, en este trabajo se utilizarán técnicas de *machine learning*, que a diferencia de las técnicas econométricas analizan cada variable, evalúan las relaciones que existen entre sí y generan una estructura matemática interna compleja que ofrece una predicción, además se puede comprobar la tasa de error de los modelos y algunas técnicas permiten observar las variables que más han influido en la precisión del modelo.

Por lo tanto, este trabajo aporta conocimiento al estudio del precio de la vivienda en función de la localización, debido a que hay muy pocos estudios que se centren en como la localización afecta al precio de las viviendas y que características de los barrios son las que más influyen en el precio medio.

Además, no existe ningún estudio sobre el precio de la vivienda en los barrios de Valencia, lo que nos permitirá descubrir cuáles son las características de los barrios, que más influyen en el precio medio de las viviendas en los barrios de Valencia.

1.3 Objetivos

El presente trabajo tiene como objetivo general, estudiar la relación entre el valor de los barrios y sus características sociodemográficas y económicas, para alcanzar este objetivo global se han establecido dos objetivos específicos:

- Objetivo 1: Describir las relaciones entre las características sociodemográficas y económicas.

- Objetivo 2: Predecir el valor catastral medio y el precio de venta del m^2 medio de los barrios de Valencia a partir de sus características.

2. Marco teórico

2.1 Evolución del mercado inmobiliario español

A lo largo del siglo XXI el mercado inmobiliario español ha sufrido grandes cambios, desde la crisis inmobiliaria de 2008 hasta la crisis sanitaria del COVID, o mismamente en la actualidad donde una elevada inflación está repercutiendo en los precios del mercado inmobiliario.

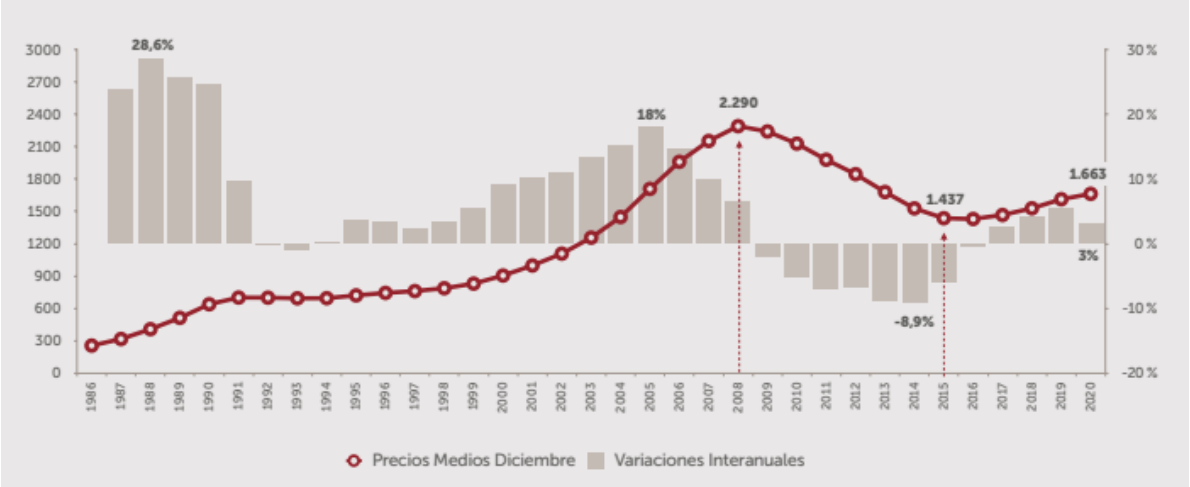
2.1.1 Crisis inmobiliaria 2008

En las dos últimas décadas España ha vivido el mayor auge y caída económica de su historia, desde la mitad del siglo XX la economía española se vio sumergida en un continuo crecimiento, debido principalmente a dos sectores: el sector financiero y sector inmobiliario. A lo que se añadió la entrada de España en la Unión Europea y la posterior introducción del euro, que propició un marco de estabilidad monetaria y bajos tipos de intereses.

Este desenfrenado crecimiento provocó un auge en el mercado inmobiliario, que se vio reflejado en el número de construcciones iniciadas, que en 1996 eran de 282.400 y en 2003 se duplicaron a 636.300 y en 2006 se alcanzaron las 863.600 construcciones iniciadas. El gran incremento de la oferta en el mercado inmobiliario también afectó a los precios, donde el 1996 el precio de venta del m^2 medio se establecía en 694,4 € y en 2003 el precio del m^2 ascendió hasta 1380,3 y se triplicó en 2006 con un precio promedio de 2085,5€/m^2. A pesar del gran incremento de la oferta y teniendo en cuenta el incremento de la población, se estima que solo el 30% de las viviendas construidas fueron destinadas a nuevos hogares familiares.

(Méndez Gutiérrez del Valle, 2016)

Figura 2. Evolución de precios medios y variaciones anuales



Fuente: Sociedad de tasación

El elevado crecimiento en la economía española llegó a tal punto que las viviendas ya no eran consideradas como un bien destinado al consumo, sino como un bien de inversión, ya que ofrecía una alta rentabilidad a corto plazo con un escaso riesgo. Esta situación atrajo a un gran número de inversores, los cuales empezaron a invertir en todas las partes del proceso, desde la compra del suelo hasta la edificación, llegando a tal punto donde existía un mercado con un excedente de viviendas no vendidas que iba creciendo cada año.

Al mismo tiempo, todas las entidades bancarias empezaron a aplicar políticas para ampliar el mercado hipotecario, recurriendo incluso al endeudamiento con otros bancos y fondos de inversión internacionales, mientras competían entre ellos para atraer demanda solvente a sectores sociales con condiciones de precariedad laboral y niveles salariales que aumentaban el riesgo de impago

Además, este aumento del número de hipotecas junto a la especulación existente en los precios del mercado inmobiliario ocasionó un aumento en el importe de las hipotecas demandadas. Lo que llevó a los clientes a un sobreendeudamiento, pasando de destinar el 46% de la renta anual hasta alcanzar en 2007 el 134% de la renta, haciendo prácticamente inasumible el pago de las hipotecas, lo que incremento la morosidad (Méndez Gutiérrez del Valle, 2016).

Debido al aumento de riesgo hipotecario, las entidades empezaron a aplicar fuertes restricciones crediticias que endurecieron las condiciones de acceso a los préstamos, esta medida junto al creciente desempleo, ocasionaron una brusca disminución de la demanda inmobiliaria, así como en la oferta.

Los bancos centraron todo su negocio en los préstamos, por lo tanto, cuando los préstamos hipotecarios se paralizaron junto a una imposibilidad de hacer frente a los pagos de una gran cantidad de clientes, donde se estima que el monto prestado alcanzaba los 319.294 millones de euros en 2008, llevaron a los bancos a la quiebra (Méndez Gutiérrez del Valle, 2016) .

La imposibilidad de pagar las deudas no solo llevo a los bancos a la quiebra, sino que muchas empresas dedicadas al mercado inmobiliario (constructoras, empresas inmobiliarias, industria relacionada) también se declararon en bancarrota, lo que implicó un brusco incremento del desempleo.

Figura 3. Evolución del desempleo en España 2007-2016

Fuente: INE

El desempleo siguió creciendo hasta llegar a su punto máximo a principios de 2013, con una tasa de desempleo del 26,94%.

Tras el “boom” de la burbuja inmobiliaria, la economía española quedo gravemente afectada, con grandes tasas de paro y un mercado inmobiliario en declive, con caídas bruscas en la construcción, venta y precio de las viviendas.

Figura 4. Operaciones de compraventa 2007-2014

Fuente: INE

El número de operaciones de compraventa descendió de las 83.713 operaciones de compraventa a principios de 2007 a apenas 22.099 operaciones a finales de 2013, lo que supone un descenso del 73,6% en el número de operaciones de compraventa en 7 años.

Tabla 1. Evolución del mercado inmobiliario 2006-2013

Años	Viviendas iniciadas (miles)	Viviendas acabadas (miles)	Compraventa de viviendas (miles)	Precio medio vivienda libre (€/m2)	Hipotecas constituidas sobre vivienda	Importe de hipotecas concedidas (millones €)	Saldo crédito a hogares para vivienda (millones €)
2006	863,8	585,6	955,2	1.990,5	1.342.171	188.339,1	519.225
2007	651,4	641,4	836,9	2.085,5	1.238.890	184.427,2	590.580
2008	264,8	615,1	564,5	2.018,5	836.419	116.809,9	621.305
2009	111,1	387,1	463,7	1.892,3	650.889	76.677,1	624.734
2010	91,7	257,4	491,3	1.825,5	607.535	71.041,2	632.437
2011	78,3	167,9	349,1	1.701,8	408.461	45.715,9	626.450
2012	44,2	120,2	363,6	1.531,2	273.873	28.328,9	605.064
2013	33,9	64,6	300,3	1.466,9	197.641	19.732,0	580.764
Evolución 2006-13 (%)	-96,08	-88,97	-68,56	-26,30	-85,27	-89,52	11,85

Fuente: Ministerio de fomento, INE y asociación hipotecaria española

El número de viviendas iniciadas descendió desde las 863.800 viviendas en 2006 a las 33.900 viviendas en 2013, lo que supone un descenso del 96,08% en el número de viviendas iniciadas. También se aprecia un gran descenso en el número de hipotecas concedidas para viviendas, donde en 2006 se concedieron 1.342.171 de hipotecas y en 2013, apenas se concedieron 197.641, por lo tanto, el número de hipotecas concedidas descendió un 89,52% en 8 años.

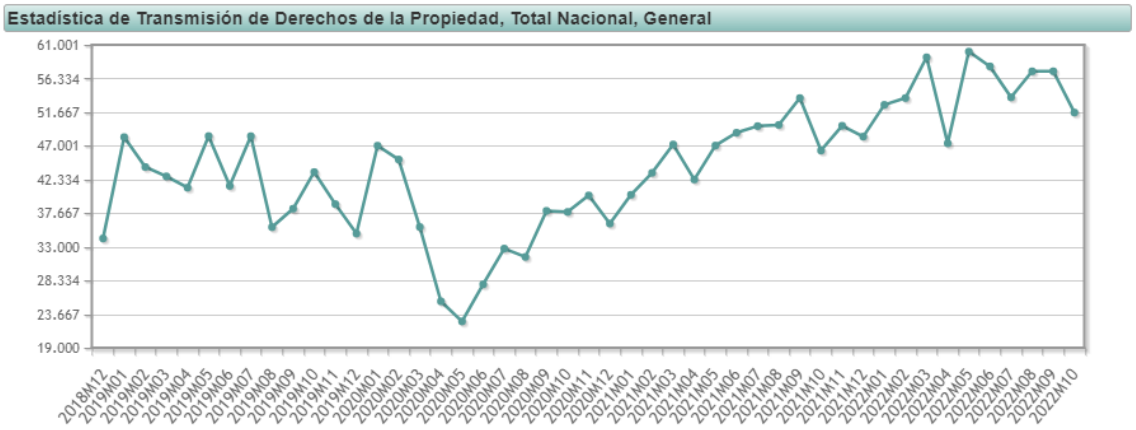
En resumen, el elevado incremento del número de hipotecas concedidas por las entidades sin tener en cuenta la solvencia de los clientes, junto al sobreendeudamiento de los bancos, una sobrevaloración de la inversión en el mercado inmobiliario y un incremento de la morosidad debido la incapacidad de hacer frente a los pagos, provocaron una de las mayores recesiones en la economía española.

2.1.2 Crisis sanitaria COVID

A comienzos de 2020, empezaron a crecer los contagios de COVID de un modo excepcional llegando a declararse una pandemia. Crecieron tanto los contagios que en España se declaró el estado de alarma, lo que supuso un confinamiento de millones de ciudadanos en sus hogares.

El mercado inmobiliario se vio afectado por las medidas adoptadas en la pandemia del COVID, que provocaron un brusco decrecimiento en las operaciones de compraventa durante los primeros meses, llegando solo a 22.652 operaciones a mitad de 2020.

Figura 5. Operaciones de compraventa 2018-2022



Fuente: INE

La pandemia también afectó al tipo de viviendas solicitadas, debido a que empezó un auge en el teletrabajo, lo que permitía a los empleados trabajar desde cualquier sitio, por lo que no era necesario que estuvieran en una zona cercana a las oficinas.

A pesar de la intensa contracción de la actividad económica, los precios de la vivienda no se vieron en gran medida afectados, debido a que no hubo grandes desequilibrios entre la oferta y la demanda.

Figura 6. Índice de precios de la vivienda 2019-2022

Fuente: INE

La crisis económica generada por la pandemia no supuso un aumento de los costes de financiación de la vivienda, sino que estos descendieron hasta alcanzar niveles mínimos

históricos. Sin embargo, sí que se apreció un endurecimiento en los criterios de concesión y en algunas condiciones aplicadas a los préstamos (Alves, 2021)

2.1.3 Previsión del mercado inmobiliario

Tras un 2022 con un alza en los precios, el mercado inmobiliario debería entrar en 2023 en una fase de ajuste. Se prevé que el precio de la vivienda disminuya en España un 3% en 2023 y disminuya un 2% en 2024 (Moreno, 2022).

Los principales factores que influirán en este descenso de los precios de la vivienda son: mayor coste de financiación hipotecaria (debido a la tendencia en aumento de los tipos de interés), menor tasa de ahorro en los hogares y pérdida de atractivo de la inversión en vivienda para el alquiler.

Además, la actividad inmobiliaria también se prevé que se enfríe, por lo que se estima que en 2023 el número de transacciones disminuya un 13% y en 2024 disminuya un 5%. Se trata de estimaciones, que se encuentran por debajo de las estimadas por el Banco Central Europeo para el conjunto de la Eurozona, en la que se estima que el precio de la vivienda disminuya un 9% (5% en España) y el número de transacciones disminuya un 15% en los próximos 2 años (18% en España) (Battistini, 2022).

2.2 El valor de las viviendas

2.2.1 Valoración de viviendas

A la hora de valorar un activo, existe la posibilidad de error, ya que al determinar el precio adecuado y no conocer el precio exacto, se puede cometer dicho error. Esto se debe, a que a la hora de determinar el precio se tiene en cuenta referencias próximas, por lo tanto, no son del todo acertadas, ya que se realizan pocas operaciones y las características de cada inmueble varían. Aunque a la hora vender un inmueble no sea necesario, con el objetivo de minimizar ese error, sí que es recomendable realizar una tasación de la vivienda

Sin embargo, cuando se pretende adquirir una vivienda a través de un préstamo hipotecario, es necesario realizará una tasación para valorar el inmueble, con el objetivo de fijar el límite del importe del crédito a conceder.

Los principales factores que se tienen en cuenta a la hora de realizar la tasación son: Superficie de la vivienda, antigüedad, estado, ubicación, transporte cercano, servicios de la zona, calidad de la construcción, eficiencia energética y situación urbanística.

Según la Orden Ministerial ECO 805/2003 publicada en el BOE sobre las normas de valoración de bienes de inmuebles, existen cuatro métodos para valorar una vivienda: Coste, Comparación, Actualización de rentas y Residual.

- Método del coste.

El método del coste es aplicable en la valoración de toda clase de edificios y elementos de edificios, en proyecto, construcción o rehabilitación.

Para llevar a cabo este método, se calcula el coste que conllevaría construir de nuevo de la misma vivienda, lo que significa:

-El valor de terreno donde se encuentra el edificio (Para calcularlo se utilizarán referencias próximas de suelo que sean similares o por el método residual).

-El coste de edificación o de las obras de rehabilitación (Coste de construcción por contrata).

-Los gastos necesarios para realizar el reemplazamiento (Impuestos, licencias y deducir el deterioro físico del inmueble).

- Método de comparación

El método de comparación es aplicable a la valoración de todo tipo de inmuebles. Además, mediante este método se pueden determinar dos valores técnicos: Valor por comparación (determina el valor de mercado de un bien) y Valor por comparación ajustado (permite establecer su valor hipotecario).

Este método consiste en determinar el valor de mercado de la vivienda a partir del precio de viviendas próximas similares. Para poder aplicar este método es necesarios disponer de al menos 6 datos de transacciones u ofertas en la zona de la vivienda, y que además tengan características comparables.

- Método de actualización de rentas

Se trata de un método aplicable a la valoración de todo tipo de inmueble susceptible de producir rentas. Mediante este método se calcula un valor técnico conocido como valor por actualización, que es utilizado tanto para determinar el valor de mercado como el valor hipotecario.

Para obtener el valor, se calcula una previsión de flujos de ingresos y gastos a futuro, por lo que se tiene en cuenta la duración del contrato, la renta pactada, impuestos y prima de riesgo.

Para obtener el valor es necesario disponer:

- Un mercado representativo de alquileres.
- Que se trate de un inmueble arrendado y por lo tanto disponga de un contrato de arrendamiento.
- Que el activo produzca ingresos como inmueble ligado a actividad económica.

- Método residual

El método residual solo puede utilizarse para valorar terrenos urbanos o urbanizables, así como edificios en proyecto u obras paralizadas. Sin embargo, existen dos procedimientos para calcularlo:

-Estático (se utilizan valores actuales): Solo se puede aplicar este procedimiento a solares e inmuebles en rehabilitación en los que se pueda comenzar la edificación en menos de 1 año.

-Dinámico (se utilizan valores esperados): Se puede aplicar para valorar terrenos urbanos o urbanizables, así como edificios en proyecto u obras paralizadas.

Este método consiste en calcular el valor del inmueble con la construcción finalizada y restar los gastos en los que hay que incurrir para que el inmueble llegue a ese estado.

2.2.2 Valor catastral

El valor catastral es la valoración monetaria que la Administración del Catastro asigna a todos los bienes inmuebles, ya sean rústicos, urbanos o de características especiales. Este valor es utilizado por la Administración de Hacienda como base de cálculo para los impuestos como: IBI (impuesto sobre bienes inmuebles), ITP (impuesto de transmisiones patrimoniales), IP (Impuesto sobre el Patrimonio) e ISD (Impuesto de Sucesiones y Donaciones).

Para determinar el valor catastral de un inmueble se tienen en cuenta los siguientes factores (Catastro, 2023):

- La localización del inmueble, las circunstancias urbanísticas que afectan al suelo y aptitud para la producción.
- El coste de ejecución material de las construcciones, los beneficios de la contrata, honorarios profesionales y tributos que gravan la construcción, el uso, la calidad y la antigüedad edificatoria.
- Las circunstancias y valores del mercado, valor del suelo, valor de la construcción y gastos de producción y beneficios de la actividad empresarial de promoción.

Además, el valor catastral de los inmuebles no podrá superar el valor de mercado. A tal efecto, mediante orden ministerial se ha fijado un coeficiente de referencia al mercado del 0,5 en el momento de aprobación y entrada en vigor de la ponencia.

El valor catastral puede actualizarse anualmente mediante la aplicación de coeficientes aprobados en las Leyes de Presupuestos Generales del Estado.

2.2.3 Índice de precios de la vivienda y evolución

El índice de precios de la vivienda es un índice creado en 2008 por el instituto nacional de estadística (INE) que tiene como objetivo medir la evolución de los precios de compraventa de las viviendas libres, tanto nuevas como de segunda mano, a lo largo del tiempo.

La fuente de información utilizada procede de las bases de datos sobre viviendas escrituradas que proporciona el Consejo General del Notariado, de donde se obtienen los precios de transacción de las viviendas, así como las ponderaciones que se asignan a cada conjunto de viviendas con características comunes (INE, 2023).

El sistema de cálculo del Índice de precios de la vivienda se basa en la combinación de dos factores que reflejan las características del mercado inmobiliario: los precios de la vivienda (oferta y demanda) y las ponderaciones (permiten establecer la importancia de cada tipología de vivienda tiene frente a todas las demás). Esta combinación se realiza mediante la fórmula del índice de Laspeyres encadenado, al igual que en el IPC (INE, 2009).

Evolución del índice de precios de la vivienda (Comunidad Valenciana)

Figura 7. Evolución del índice de precios de la vivienda (Comunidad Valenciana)

Fuente: INE

A raíz de la crisis de 2008 los precios de la vivienda bajaron drásticamente, debido sobre todo al exceso de oferta de vivienda y un difícil acceso a la financiación, lo que provocó que el precio de las viviendas en la comunidad valenciana llegara a situarse en su punto más bajo a finales de 2013. A partir de principios de 2014, los precios fueron recuperándose debido a la decisión del Banco Central Europeo de bajar los tipos de interés que facilito el acceso a los préstamos y provoco un aumento de las operaciones de compraventa.

2.3 Los barrios de Valencia

La ciudad de Valencia es una de las metrópolis más pobladas de todo el continente europeo y una de las ciudades más antiguas del planeta, con más de 2000 años de existencia. A nivel histórico ha sido una de las ciudades más influyentes y ha dejado profundas huellas que muchas veces desconocemos. En el presente trabajo nos adentraremos por los orígenes y los principales núcleos históricos que la conforman: desde el histórico barrio del Carmen, al centro neurálgico de la ciudad como es Ciutat Vella o incluso a los barrios costeros como la Malva-rosa.

La Ciudad de Valencia está dividida en 19 distritos, que a su vez están comprendidos por un total de 87 barrios y pedanías.

DISTRITO	**BARRIOS**
Ciutat Vella	La Seu
	La Xerea
	El Carme
	El Pilar
	El Mercat
	Sant Francesc
Eixample	Russafa
	El Pla del Remei
	Gran via
Extramurs	El Botànic
	La Roqueta
	La Petxina
	Arrancapins
Campanar	Campanar
	Les Tendetes
	El Calvari
	Sant Pau
La Saïdia	Marxalenes
	Morvedre
	Trinitat
	Tormos
	Sant Antoni
El Pla del Real	Exposició
	Mestalla
	Jaume Roig
	Ciutat Universitaria
L' Olivereta	Nou Moles
	Soternes
	Tres Forques
	La Fontsanta
	La Llum
Patraix	Patraix
	Sant Isidre
	Vara de Quart
	Safranar
	Favara
Jesús	La Raïosa

	L'Hort de Senabre
	La Creu Coberta
	Sant Marcel·lí
	Camí Real
Quatre carreres	Mont-Olivet
	En Corts
	Malilla
	La Fonteta de Sant Lluís
	Na Rovella
	La Punta
	C. Arts i les Ciències
Poblats Marítims	El Grau
	Cabanyal - Canyamelar
	La Malva-rosa
	Beteró
	Natzaret
Camins al Grau	Aiora
	Albors
	La Creu del Grau
	Camí Fondo
	Penya-Roja
Algirós	L`Illa Perduda
	Ciutat Jardí
	L`Amistat
	La Bega Baixa
	La Carrasca
Benimaclet	Benimaclet
	Camí de Vera
Rascanya	Els Orriols
	Torrefiel
	Sant Llorenç
Benicalap	Benicalap
	Ciutat Fallera

Figura 8 Distritos de la ciudad de Valencia

Fuente: Ayuntamiento de Valencia

3. Metodología

3.1 Datos

3.1.1 Descripción de la base de datos

La base de datos recoge información de 16 distritos de la ciudad de Valencia, y el desglose en sus respectivos barrios. Esta información corresponde al año 2021, y ha sido obtenida a través de la página web de estadística del ayuntamiento de Valencia (Ayuntamiento de Valencia, 2022), excepto el precio de venta del m^2 que se ha recogido a través del portal inmobiliario Fotocasa (Fotocasa, 2022), que es el portal que utiliza el ayuntamiento de valencia para reflejar los precios del m^2 de los distritos.

En resumen, la base de datos está formada por 70 barrios y 17 variables, que recogen las siguientes características de los barrios:

Densidad de población (Dpoblación): Medida de distribución de la población, que refleja el número de habitantes por hectárea.

Tasa de crecimiento natural (TCrecNat): Tasa que determina al aumento o disminución de una población en un año determinado, donde valores positivos indican un superávit de nacimientos en comparación con las defunciones. La tasa de crecimiento natural se calcula como la diferencia entre la tasa de natalidad y la tasa de mortalidad.

Tasa bruta de emigración intraurbana (TemigIntra): Migración en el interior de un mismo conjunto urbano. Supone cambiar el lugar de residencia (barrio), pero sigue permaneciendo en la misma ciudad.

Tasa bruta de inmigración intraurbana (TinmigIntra): Inmigración en el interior de un mismo conjunto urbano. Hace referencia a las personas que se establecen en el barrio, y provienen de otros barrios de la ciudad.

Tasa bruta de emigración interurbana (TemigInter): Migración entre ciudades. Índica las personas cambian su lugar de residencia a un lugar ajeno a la ciudad.

Tasa bruta de inmigración interurbana (TinmigInter): Inmigración entre ciudades. Hace referencia a las personas que se establecen en el barrio, y provienen de otros lugares ajenos a la ciudad.

% vehículo agrario/industrial (VehIndustoAgrario): Porcentaje de vehículos agrarios o industriales registrados en los barrios. Para calcular este porcentaje se han tenido en cuenta todos los vehículos motorizados registrados en la zona, como son autobuses, camiones, turismos, ciclomotores, motocicletas y tractores, así como remolques.

Valor catastral medio (Valor total): Valor catastral medio de los inmuebles situados en los barrios, teniendo en cuenta tanto el valor del suelo como el valor de la construcción.

Superficie total de aparcamiento por turismo (SuperfAparcTurismo): Superficie de aparcamiento (m^2) por turismo registrado en dicho barrio. Este indicador hace referencia a garajes y zonas privadas de aparcamiento, como parkings.

Abandono escolar (AbEscolar): Jóvenes que no han obtenido el título de educación secundaria obligatoria.

Número de colegios (NColegios): Número de colegios o institutos situados en dicho barrio.

Comercio, restaurante, hospedaje y reparaciones (ComerRestHospyRep): Número total de comercios, restaurantes, hospedajes y reparaciones establecidos en el barrio.

Transporte y comunicación (TranspyComunic): Número total de actividades económicas dedicadas al transporte y comunicación.

Instituciones Financieras, Seguros, Servicios prestados a las empresas y alquileres (InsFinSegSerAlq): Número total de instituciones financieras, seguros, servicios prestados a las empresas y alquileres en dicho barrio.

Otros servicios (Otrosservicios): Número total de actividades económicas no recogidas en las anteriores variables.

%parados (parados): Porcentaje de personas en busca activa de empleo registrados en el barrio. Para calcular este porcentaje se realiza el cociente entre el número total de parados registrados en la zona y el total de población activa que recoge a todos los habitantes en edad laboral que se encuentra trabajando (población ocupada) o bien se encuentran en búsqueda de empleo (población parada).

$$\frac{\text{Total parados}}{\text{Población activa}} \text{ x } 100$$

Precio de venta m^2 (Precioventam^2): Precio de mercado medio del m^2 en dicho barrio.

3.2 Análisis de datos

3.2.1 Aprendizaje no supervisado

El aprendizaje no supervisado es un conjunto de técnicas de aprendizaje automático, que se basan en que no se conoce a priori el objetivo buscado, es decir, cuando los individuos no los tienes previamente clasificados o no conoces ninguna variable respuesta asociada a ellos. Estas técnicas se utilizan mayormente para explorar los datos y encontrar patrones.

En el presente trabajo se utilizarán diferentes técnicas de aprendizaje no supervisado con el objetivo de explorar los datos de los barrios y conocer las relaciones entre las variables. Las técnicas que se utilizarán serán: Análisis de componentes principales, Clustering y Análisis de correspondencias múltiples.

Análisis de componentes principales

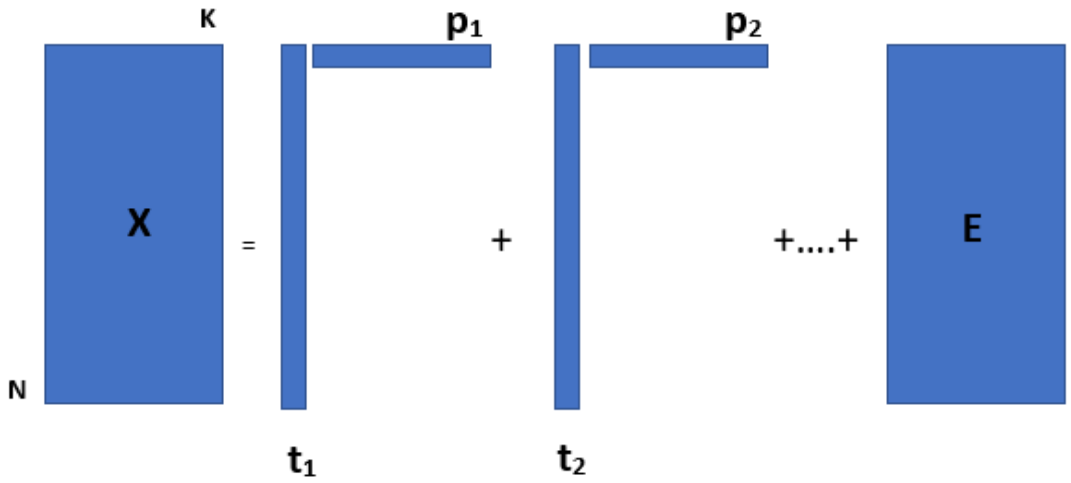
El análisis de componentes principales (PCA, por sus siglas en inglés) es una técnica multivariante de aprendizaje no supervisado que tiene como objetivo condensar la información contenida en k variables primitivas (k es elevado y las variables están correlacionadas entre sí), en un número reducido de nuevas variables (componentes principales) no observables directamente e incorrelacionadas entre sí, definidas como combinaciones lineales de las variables primitivas.

Las principales funciones del PCA son:

- Análisis exploratorio: Descubrir las relaciones existentes entre las diferentes variables y las relaciones entre las diferentes observaciones, así como detectar valores anómalos o atípicos.
- Reducir la dimensionalidad: Representar adecuadamente la información de la matriz de datos (observaciones y variables) con un número menor de variables incorrelacionadas construidas como combinaciones lineales de las originales (componentes principales). Las componentes principales obtenidas explican la mayoría de la variabilidad del modelo, siendo en orden las primeras componentes las que recogen la mayor información. Por lo tanto, la varianza explicada por cada componente principal explicará la variabilidad no explicada por las componentes anteriores.

El funcionamiento del análisis de componentes principales parte como un análisis multivariante, donde se dispone de una matriz de datos X de dimensión NxK con N observaciones o individuos y K variables explicativas.

Figura 9. Descomposición gráfica de la matriz de datos X



En la figura se muestra la descomposición de la matriz X en la matriz de componentes principales, formada por los *scores* (t_n) y los *loadings* (p_n), además de la matriz de residuos(E) que recoge la variabilidad que no es explicada por las componentes seleccionadas.

Para la matriz X se construye un espacio de k dimensiones, donde cada fila de matriz X corresponde a un punto en ese espacio creado definido por las variables, las cuales han sido reescaladas a varianza unitaria, para ello se calcula la media de cada variable y se resta esta media a la matriz de datos, por lo que el origen del sistema de coordenadas pasa a ser el centro de la nube de datos.

Para determinar la primera componente principal se traza una recta en el espacio creado que mejor aproxime los datos (mínimos errores cuadrados), pero pasando por el origen de coordenadas. Siendo el vector de *loadings* (p_n) el que determine la dirección de esta recta.

La segunda componente principal es una recta en el espacio, ortogonal a la primera componente y que pasa por el punto medio de la nube. Ambas componentes forman un plano en el espacio y al proyectar las observaciones sobre estos planos se obtienen los vectores de scores para la primera y segunda componente (Wold, 1987).

En el programa Aspen, se muestra el *R2* y *Q2* de todas las componentes principales extraídas, donde el *R2* hace referencia a la variabilidad total de la modelo explicada por las componentes principales y el *Q2* es la bondad de predicción, que mide la potencia de predicción del modelo *PCA* con A componentes.

Tras realizar la técnica de análisis de componentes principales con A componentes se pueden llevar a cabo diferentes estudios como a través del gráfico de scores donde se observan las relaciones entre las diferentes observaciones y se detectan valores anómalos o atípicos. También, mediante el gráfico de *loadings* se pueden observan las relaciones entre las variables explicativas del modelo.

Cluster

El análisis cluster es una técnica estadística de análisis multivariante que tiene como objetivo agrupar las observaciones en diferentes grupos, en función de sus similitudes o diferencias. En

el presente trabajo, se aplicarán dos métodos de clustering: clustering jerárquico y clustering de partición.

- Clustering jerárquico

El clustering jerárquico se caracteriza en agrupar los datos basándose en la distancia entre cada uno y buscando que los datos que están dentro de un cluster sean los más similares entre sí. Este método se puede dividir en dos grupos:

- Cluster aglomerativo:
  Los métodos aglomerativos, conocidos como ascendentes, se basa en que el agrupamiento se inicia en la base del árbol, donde cada observación forma un cluster individual. Los clusters se van combinado a medida que la estructura crece hasta converger en una única "rama" central. Los métodos del cluster aglomerativo más utilizados son: *Ward, Average, Centroid, Complete or maximum y Single or mínimum*.

  En el presente trabajo se utilizará el método Ward, el cual se base en que se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster (Gutiérrez ,1994).

- Cluster disociativo
  Los métodos disociativos, conocidos como descendentes, es el método opuesto al cluster aglomerativo, se inicia con todas las observaciones contenidas en un mismo cluster y se suceden divisiones hasta que cada observación forma un cluster individual. El método más utilizado del cluster disociativo es Diana (*DIvisive ANAlysis Clustering* (Amat-Rodrigo, 2017)).

  En el método Diana el proceso es inverso al anterior método. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños.

Los métodos jerárquicos permiten la construcción de un árbol de clasificación, conocido como dendograma, que permite observar de forma gráfica el procedimiento de división. Donde en la parte superior se encuentra un solo grupo que se subdivide a medida que se desciende en el dendograma. Por lo tanto, se deberá determinar el número de clusters a formar.

- Clustering de partición

El clustering de partición se basa en métodos que están diseñados para clasificar individuos en una clasificación de K clusters, donde K se especifica a priori o bien se determina como una parte del proceso.

Para determinar el número óptimo de cluster se utilizará el coeficiente de *Silhoutte*, el cual es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de clustering. Esta métrica va desde -1 a 1, donde un valor alto indica que el individuo está bien emparejado con su propio cluster y mal emparejado con los clusters vecinos

En el presente trabajo se aplicarán dos métodos: *K-means* y *k-medoides*.

- *K-means*: El método *K-mean*s se basa encontrar los K mejores clusters, entendiendo como mejor cluster aquel cuya varianza interna (intra-cluster variation) sea lo más pequeña posible.

- *k-medoides*: Se trata de un método bastante similar al *K-means*. La diferencia es que, en *K-medoides*, cada cluster está representado por una observación presente en el cluster (*medoid*), mientras que en *K-means* cada cluster está representado por su centroide, que se corresponde con el promedio de todas las observaciones del cluster (Amat-Rodrigo, 2017).

En todos los métodos se utiliza la distancia euclídea que viene dada por la siguiente expresión:

$$d(i,j) = (W_i - W_j)'(W_i - W_j)$$

Siendo i y j dos individuos, la distancia euclídea mide el grado de desemejanza entre ambas observaciones.

Análisis de correspondencias múltiples

El análisis de correspondencias (le Roux, 2010) es una técnica estadística que analiza la relación entre dos variables categóricas en una tabla de contingencia. Se trata de una técnica que se basa en el análisis de componentes principales donde el objetivo general es aproximar la relación entre variables en un espacio de dimensión reducido. El objetivo principal del análisis de correspondencias en representar gráficamente las similitudes entre las filas y entre las columnas de la tabla, en un espacio dimensional reducido. Por lo que, las filas que están cerca en ese espacio dimensional tienen una distribución condicional similar en las columnas de la tabla.

En el análisis de correspondencia múltiple (*MCA*) se extiende esta idea, pero con un mayor número de variables, siendo el objetivo mostrar geométricamente las filas y columnas de la tabla de datos (donde las filas representan individuos y las columnas las categorías de las variables en un espacio de baja dimensión) de modo que la proximidad en el espacio indica similitud de categorías y de individuos.

El principal objetivo del *MCA* es resumir una gran cantidad de datos en un número reducido de dimensión, con la menor perdida de información posible. Se trata de una técnica que se utiliza principalmente para explorar los datos y analizar las relaciones entre las variables dependientes categóricas.

Esta técnica trata de analizar un conjunto de variables nominales, las cuales comprenden varios niveles y cada uno de estos niveles se codifica como una variable binaria. En análisis de correspondencias múltiples también se pueden utilizar variables cuantitativas, pero es necesario que se recodifiquen, a nivel categórico.

3.2.2 Aprendizaje supervisado

El aprendizaje supervisado es un conjunto de técnicas de aprendizaje automático, donde se conoce a priori el objetivo buscado, es decir, conoces alguna variable respuesta asociada a los individuos.

En el presente trabajo se utilizarán diferentes técnicas de aprendizaje supervisado de regresión como: árbol de regresión, *random forest*, Vecino más próximo, máquinas de soporte vectorial, *Grandient boosting* y *PLS*. El objetivo de aplicar estas técnicas es obtener un modelo predictivo de la variable respuesta con el menor error posible.

Árbol de regresión

El árbol de regresión (Breiman, 1984) es una técnica que tiene como objetivo predecir una variable respuesta Y a partir de diferentes variables predictoras. Para ello, establece sucesivas particiones del conjunto de datos de manera que los subconjuntos resultantes sean los más homogéneos posibles.

La construcción del árbol se realiza a partir de los datos de entrenamiento, y consiste en la partición del espacio predictor en J nodos, para cada uno de los cuales se va a calcular una constante, que es la media de la variable respuesta Y_{Nj} para las observaciones de entrenamiento que caen en dicho nodo. Estas constantes son los valores que se utilizaran para la predicción de nuevas observaciones, por lo que para determinar el valor respuesta Y de una nueva observación será necesario comprobar a que nodo terminal pertenece. (Fernández,2022)

Para realizar la partición del espacio predictor en J nodos, como criterio de error se utiliza la suma de los residuos al cuadrado (RSS). Como no es factible realizar una búsqueda con todas las particiones posibles se sigue un proceso iterativo, en el que se van realizando cortes binarios (Yes/No).

En la primera iteración, se trabaja con todos los datos donde una variable explicativa X_j y un punto de corte *s* definen dos hiperplanos:

$$N1 = \{\mathrm{X} \mid \mathrm{Xj} \leq \mathrm{s}\}$$

$$N2 = \{\mathrm{X} \mid \mathrm{Xj} > \mathrm{s}\}$$

Posteriormente, se seleccionan los valores de j y s que minimicen:

$$\sum_{i \in N_1} (\mathrm{yi} - \hat{y}_{N_1})^2 + \sum_{i \in N_2} (\mathrm{yi} - \hat{y}_{N_2})^2$$

$\hat{y}_{Nj}$: Media de la variable respuesta Y en el nodo j.

A continuación, se repite el proceso en cada uno de los nodos obtenidos N1 y N2, y así sucesivamente hasta alcanzar el criterio de parada, que suele ser un número mínimo de observaciones en un nodo (Fernández, 2022).

El árbol de regresión es una técnica que ofrece una gran interpretabilidad, es robusta frente a valores anómalos y no necesita que las variables presenten normalidad. Además, es una técnica que funciona bien con grandes cantidades de datos, debido a que el coste computacional es inferior al de otras técnicas.

Uno de los problemas del árbol de regresión es el sobreajuste, debido a que se forman árboles demasiado grandes y complejos que predicen muy bien los datos de entrenamiento, pero cuando se utiliza el árbol obtenido para predecir nuevos datos, ofrece un bajo desempeño. Por lo tanto, es importante podar el árbol, ya que evitara el sobreajuste y facilitara la interpretación.

El método utilizado en la poda del árbol es la regla *X-SE*, donde se selecciona el árbol más pequeño con un error estimado menor que $\beta + X * SE$, donde B es la estimación de error más baja y SE es el error estándar de esta estimación B.

Random forest

La técnica Random forest es un algoritmo de aprendizaje automático desarrollado por Breiman (2001)que se basa en la combinación de múltiples árboles de decisión para llegar a un único resultado.

El algoritmo Random forest se compone de un conjunto de árboles de decisión, y cada árbol del conjunto se compone de una muestra de datos extraídos del conjunto de entrenamiento. Posteriormente, se agrega aleatoriedad a través de la función *mtry* (número de variables muestreadas aleatoriamente como candidatas en cada división), lo que añade más diversidad al conjunto de datos y reduce la correlación entre los árboles de decisión. En este caso, al ser un problema de regresión, para determinar la predicción se promediarán los árboles de decisión individuales, sin embargo, si se tratara de un problema de clasificación, se realizará por voto mayoritario, es decir, la variable categórica más frecuente producirá la clase pronosticada (IBM,2021).

La principal ventaja de la técnica Random forest frente a los árboles de regresión, es que reduce el riesgo de sobreajuste, debido a que, al haber una gran cantidad de árboles de decisión, el clasificador no sobreajustara el modelo, ya que el promedio de los árboles no correlacionados reduce la varianza general y el error de predicción.

Otra ventaja de la técnica Random forest es que permite observar cuáles han sido las variables más influyentes en la predicción, a través de la librería *Randomforest* se obtienen las siguientes dos medidas:

- *%incMSE*: Indica la disminución media de la precisión de las predicciones cuando la variable dada se excluye del modelo. Se trata de una medida más robusta e informativa que *IncNodePurity.*
- *IncNodePurity:* Medida de la disminución total de impureza de los nodos que resulta de la división de la variable dada.

Vecino más próximo

La técnica de vecino más próximo pertenece a la clase denominada aprendizaje vago debido a que no se obtiene un modelo, sino que se basa en predecir el valor de una observación en función del valor de las observaciones más cercanas.

El funcionamiento de la técnica de vecino más próximo, parte de un conjunto de datos X con n puntos, donde la técnica permite encontrar los k puntos más cercanos en X a un determinado punto nuevo, teniendo en cuenta la función de distancia utilizada (métrica). Este tipo de técnica es fuertemente dependiente de la noción de similitud entre las observaciones y por tanto de la métrica.

En el caso de problemas de regresión, la predicción se determina a partir de la media de los valores de la variable dependiente de los k vecinos más próximos. Sin embargo, en los problemas de clasificación, se suele coger un número k de vecinos impar, y la predicción se decide por la clase más frecuente.

En el presente trabajo se utilizará la función *train.knn* del paquete *Knnp*, y la métrica utilizada es la distancia de Minkowski:

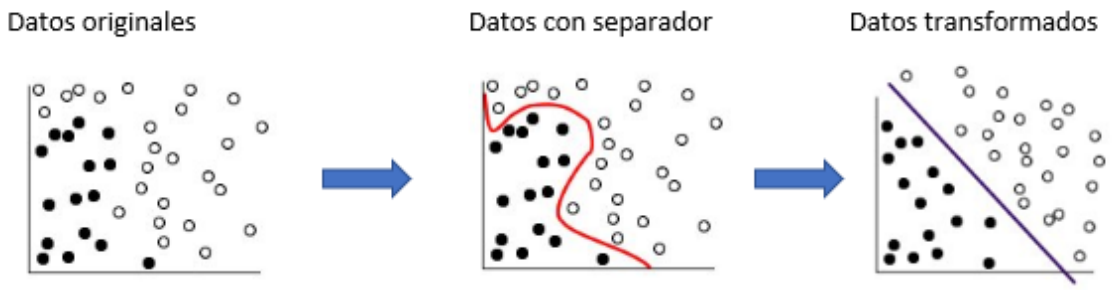
$$[\sum_{i=1}^{k}(|x_i - y_i|)^p]^{\frac{1}{p}}$$

La principal ventaja de la técnica de vecino más próximo es la simplicidad de su algoritmo, por lo que es fácil de implementar. Además, se trata de una técnica robusta frente a valores atípicos (Sun, 2010).

Máquinas de soporte vectorial

Las máquinas de soporte vectorial son un conjunto de algoritmos de aprendizaje supervisado usados para clasificación de datos. Su funcionamiento se basa en correlacionar datos a un espacio de características de grandes dimensiones de forma que los puntos se puedan categorizar. Posteriormente se detecta un separador entre las categorías y los datos se transforman de modo que el separador se puede extraer como un hiperplano. Por lo tanto, las características de una nueva observación se pueden utilizar para predecir el grupo al que puede pertenecer este nuevo individuo (IBM,2021).

Figura 10 Funcionamiento de las máquinas de soporte vectorial



Para realizar la transformación se pueden utilizar varias funciones matemáticas, a través del parámetro *kernel*.

En el presente trabajo se han utilizado dos funciones distintas de máquinas de soporte vectorial y se han configurado con los siguientes *hiperparametros*:

- Función *ksvm* de la librería *kernlab*: en esta función es necesario indicar una serie de parámetros que son:
    1. *Type:* La función *ksvm* puede ser utilizada tanto para clasificación como para regresión, por lo que en este parámetro se indicara que el problema es de regresión, para ello en *type* se indicara "*nu-svr*".

2. *Kernel: Ksvm* te proporciona las funciones *kernel* más populares, tras probar con varias, la que mejores resultados presenta es "*rbdfot*"(*Radial Basis kernel "Gaussian"*)
3. *Kpar:* Contiene los *hiperparametros* del núcleo. Al utilizar en *kernel "rbdfot*", se recomienda asignar "*automatic*" al parámetro *kpar*, el cual usa la heurística para calcular la mejor sigma.

- Función *svm* de la librería e0171: en esta función también es necesario indicar una serie de parámetros, que son:
    1. *Method*: Se indica en el parámetro que se trata de un problema de regresión asignando al parámetro "*eps-regression*".
    2. *Kernel*: Al igual que en la anterior función se asigna la función *kernel* que presenta mejores resultados, que es "radial".
    3. *Cost*: es la constante 'C' del término de regularización en la formulación de Lagrange. Tras probar con varios valores de "*cost*", el que presenta mejores resultados es 10.
    4. Gamma: Para obtener el valor de gama se realiza la división de uno entre el número de variables (14 variables), por lo que se asigna un valor de gamma de 0,07.

Gradient Boosting

Gradient boosting es una técnica desarrollada por (Friedman, 2001), se trata de una técnica que pertenece a los denominados *esamble model*, que significa que se construye por la unión de otros muchos modelos. Esta técnica construye un conjunto de árboles sucesivos poco profundos y débiles, y cada árbol va aprendiendo y mejorando del anterior.

Su funcionamiento se basa en crear varios predictores en secuencia, donde se parte de un primer predictor que usa la media de la variable respuesta(Y) para predecir, posteriormente el segundo predictor explica los errores del primer predictor, el tercer predictor explica los errores del segundo, y así sucesivamente (Hernández ,2021).

Las principales ventajas de los modelos *Gradient Boosting*, es que son muy flexibles debido a que proporciona diferentes opciones de ajuste de *hiperparametros* que hacen que la función se ajuste de manera muy flexible. Además, se trata de una técnica que no requiere un preprocesamiento previo de los datos y es capaz de manejar los datos faltantes.

Para aplicar la técnica de *gradient boosting* se utilizará la función *gbm,* en la cual es necesario indicar una serie de parámetros para mejorar los resultados obtenidos. Los parámetros son:

1. *n.trees*: Número total de árboles para ajustar.
2. *interaction.depht*: Número de divisiones en cada árbol.
3. *shrinkage*: Controla como de rápido el algoritmo desciende por el gradiente. Valores pequeños reducen el peligro de sobreajuste, pero incrementa el tiempo para encontrar el ajuste óptimo.
4. *n.minobsinnode*: Número mínimo de observaciones en el nodo terminal de los árboles.

5. *bag.fraction*: Controla la fracción de las observaciones del conjunto de entrenamiento seleccionadas aleatoriamente para proponer el siguiente árbol en la expansión. Esto introduce aleatoriedad en el ajuste del modelo.

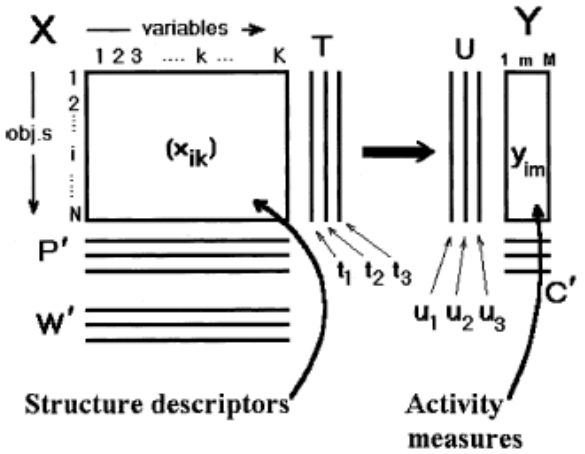
Regresión de mínimos cuadrados parciales

La regresión de mínimos cuadrados parciales (*Partial Least Squares-regression*) es una técnica multivariante, que reduce los predictores a un conjunto más pequeño de componentes no correlacionados y realiza una regresión de mínimos cuadrados sobre estos componentes, en vez de sobre los datos originales.

A diferencia de la regresión lineal múltiple (MRL), el PLSR puede analizar datos con numerosas variables fuertemente correlacionadas y ruidosas, además de modelar simultáneamente varias variables respuesta (Y).

Se trata de un método para relacionar dos matrices de datos X e Y, mediante un modelo multivariante lineal. El modelo PLS es el siguiente:

Figura 11. Gráfico de los elementos del PLS



Las ecuaciones del modelo:

$$X = TP' + E$$

$$Y = TC' + F$$

Donde X es una matriz de dimensiones NxK, siendo N el número de observaciones y K el número de variables explicativas. La matriz Y tiene una dimensión NxM, donde N es número de observaciones y M el número de variables respuesta (Wold, 2001).

La matriz T son las puntuaciones factoriales o scores de la matriz X. Se estiman como combinaciones lineales de las variables originales X con los coeficientes de cargas w^*.

$$T = XW^*$$

La matriz P son las cargas factoriales en el espacio X y la matriz C son las cargas factoriales en el espacio Y. Por último, E y F son las matrices de errores.

Para realizar las predicciones de una nueva observación, se proyectan las coordenadas del nuevo individuo x_i en el plano de los dos primeras componentes en el espacio X, se obtienen los scores (coordenadas) en ambas componentes t_{i1} y t_{i2}.

$$t_1 = XW_1$$

$$t_2 = (X - t_1 P_1^T)W_2$$

A partir de estos pueden predecirse los scores u_{i1} y u_{i2}, y a partir de estos scores se pueden predecir los valores de las variables respuesta para dicho individuo.

Métricas de bondad de ajuste

Para evaluar el desempeño de los modelos predictivos se utilizarán las medidas de bondad de ajuste, que miden la discrepancia entre los valores reales y los predichos.

Existen varias medidas de bondad de ajuste tanto relativas como absolutas, en el presente trabajo para evaluar la calidad de los modelos se utilizarán dos medidas de error:

- Raíz del error cuadrático medio (*RMSE*), representa a la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor predicho.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- Media del error porcentual absoluto (*MAPE*), es el promedio del error absoluto o diferencia entre el valor real y el predicho, expresado como un porcentaje de los valores reales.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{y_i}$$

La principal diferencia entre estas medidas es que el *MAPE* es más robusto que el *RMSE*, ya que no les da tanta importancia a los valores anómalos, debido a que el RMSE al elevar el valor absoluto de la diferencia al cuadrado da más importancia a estos valores anómalos.

3.3 Software

R studio

El R es un software gratuito y de código abierto creado en 1993 por Robert Gentleman y Ross Ihaka, se trata de un ambiente de programación diseñado para el análisis estadístico. El software

está formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones.

Se trata de una herramienta muy poderosa para todo tipo de procesamiento y manipulación de datos, que permite trabajar con grandes volúmenes de datos y consume pocos recursos informáticos.

Además, R ofrece una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, …) y técnicas gráficas con alta calidad (R project, 2022).

RStudio es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, un editor que resalta la sintaxis y admite la ejecución directa del código, así como herramientas para el trazado, el historial, la depuración y la gestión del espacio de trabajo (Rstudio, 2018).

Aspen Pro MV

Aspen Pro MV es un software de análisis multivariante creado por la compañía *Aspentech*, fue desarrollado para tratar una gran cantidad de datos y analizar que está afectando a la variabilidad entre los miles de variables correlacionadas en los procesos de fabricación de las empresas.

El principal objetivo del software es detectar variaciones en el proceso de producción para ello cuenta con advertencias tempranas y precisas sobre el estado del proceso, brindando información para evitar la producción fuera de las especificaciones, mantener la eficiencia y optimizar rápidamente las operaciones (Aspentech, 2018).

Las aplicaciones que ofrece son:

- Análisis de desviación de la calidad
- Análisis de rendimiento unitario
- Análisis de degradación de la capacidad de producción
- Análisis multivariado fuera de línea (descubrimiento y optimización de variables clave)
- Análisis multivariante en línea (supervisión y solución de problemas)
- Análisis de variabilidad del proceso por lotes

4. Resultados

En este apartado se procederá a analizar la base de datos con el objetivo de estudiar los barrios de valencia y realizar los modelos predictivos para la valoración del suelo en barrios de Valencia según el valor catastral y el precio de mercado.

4.1 Análisis descriptivo de las variables

A continuación, se realizará un análisis descriptivo de las variables de las características sociodemográficas y económicas de los barrios:

Tabla 2. Medidas descriptivas de las variables

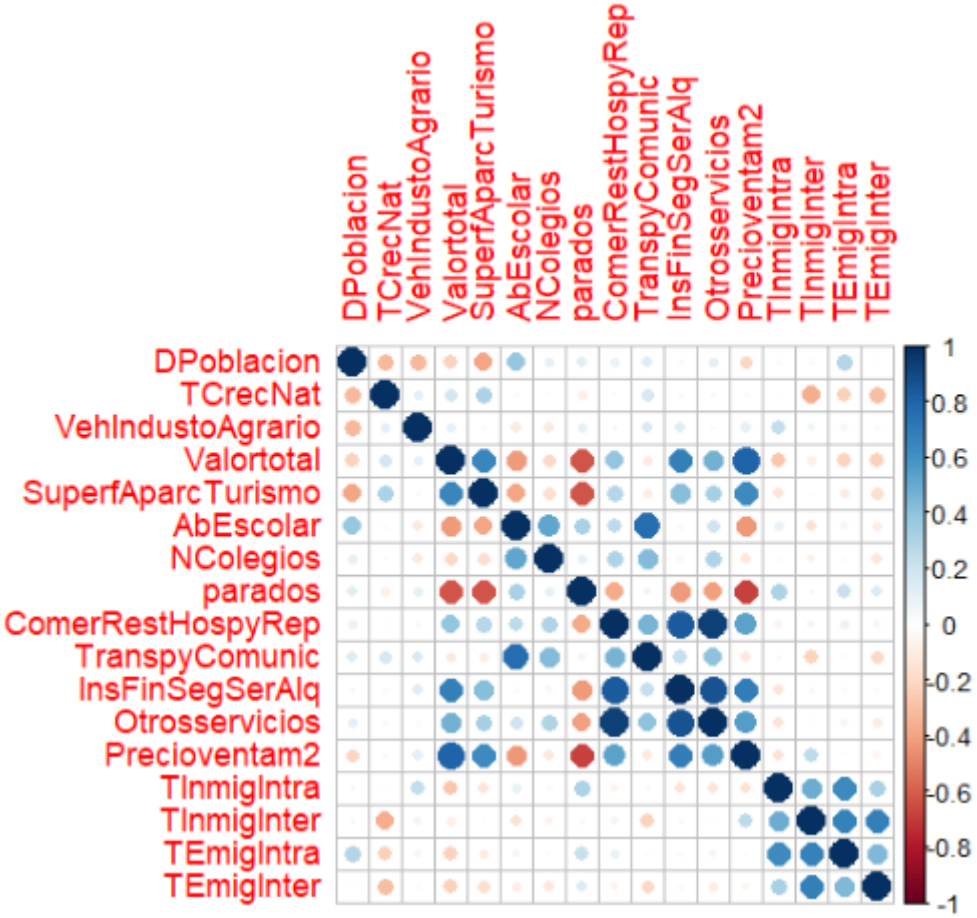
Variable	Media	Mediana	Máximo	Mínimo	Coeficiente de curtosis	Coeficiente de asimetría
Densidad población	224,51	207,84	529,15	4,44	2,229	0,213
T. crecimiento natural	-3,01	-3,1	4,4	-10,7	4,346	-0,114
T. emigración interurbana	32,1	30,7	55,2	20,7	3,9	0,8
T. inmigración interurbana	38,3	35,9	72,7	17,3	4,4	1,1
T. emigración intraurbana	33	33,05	50	19,6	2,8	0,4
T. inmigración intraurbana	32,6	32,6	50,7	15,7	3	0,04
% vehículo agrario industrial	7,4	6	38,8	3,6	20,1	3,8
Valor catastral medio	49067	42206	141292	23993	7,1	1,9
Superficie aparcamiento por turismo	19,4	15,7	70,8	2,6	7,9	2
Abandono escolar	1572,5	1136,5	7971	99	7,9	1,9
Nº de colegios	2,4	2	7	0	3,1	0,7
Comercio, restaurante, hospedaje y reparaciones	441,2	350	1762	62	5,8	1,6
Transporte y comunicación	59,2	46	285	3	10,5	2,3
Financieras, Seguros, Servicios prestados a empresas y alquileres	415,5	264,5	2931	25	14,6	3,1
Otros servicios	199,4	149	769	22	5,6	1,6
% parados	28,7	28,1	51,2	12,45	3,6	0,6
Precio venta $M^{2.}$	2146	2007	3990	1157	2,9	0,7

Para asumir que una variable presenta normalidad, el coeficiente de curtosis y el coeficiente de asimetría, deben estar ambos entre el intervalo [-2,2], como se observa en la tabla ninguna variable cumple con esta condición, por lo que se deduce que ninguna variable presenta normalidad.

Por lo tanto, esto puede condicionar la elección de técnicas de análisis, ya que algunas requieren la normalidad en las variables. Sin embargo, no es un requisito para abordar los objetivos planteados, aunque se tendrá en cuenta para la selección de técnicas.

A continuación, se ha realizado un gráfico de correlaciones de las variables de la base de datos:

Figura 12. Gráfico de correlaciones



Como se observa en el gráfico de correlaciones, existen variables que poseen una elevada correlación entre sí, como puede ser Comercio, restaurante, hostelería y reparación con Otros servicios, el Valor total con el Precio de venta del m^2 o el Transporte y comunicación con el Abandono escolar.

También se observan fuertes correlaciones negativas entre algunas variables, como el Porcentaje de parados y el Valor total o Porcentaje de parados y Superficie de aparcamiento por turismo.

Por último, cabe destacar una significante correlación entre las cuatro variables que representan el movimiento migratorio, así como entre las 4 variables que recogen los servicios de los barrios (Comercio, restaurante, hostelería y reparación, Transporte y comunicación, Instituciones financieras, servicios a empresas, seguros y alquileres y Otros servicios)

4.2 Estudio de los barrios de Valencia

En esta sección se va a realizar un estudio de los barrios de Valencia, con el objetivo de analizar qué diferencias existen entre ellos, lo que nos permitirá conocer mejor los barrios y facilitará posteriormente la interpretación de los resultados obtenidos. La ciudad de Valencia está constituida por 16 distritos, que se dividen en 70 barrios, de los cuales se dispone de 17 variables heterogéneas, que representan distintos aspectos y características de los barrios.

Estas variables tienen distintos niveles de correlación entre ellas, por lo que para aplicar un análisis de componentes principales es necesario un conjunto de variables homogéneas y una base conceptual para analizarlas como conjunto. Por lo que, para analizar la estructura subyacente a un conjunto de datos, es necesario valorar la naturaleza de las dimensiones obtenidas en base a las variables originales.

Tal y como se ha podido apreciar en el gráfico de correlaciones, las relaciones entre las variables pertenecientes a bloques presentan los niveles de correlación más elevados. Por lo tanto, es interesante profundizar en cada uno de estos bloques por separado, ya que se podrá obtener un conocimiento más detallado sobre los barrios.

Por lo tanto, se abordará el análisis de las relaciones entre las variables, considerando los bloques de movimiento migratorio y servicios por separados.

4.2.1 Análisis de los movimientos migratorios en los distintos barrios de Valencia

En primer lugar, se ha decidido estudiar los movimientos migratorios en los barrios de valencia, ya que es interesante conocer el tipo emigración e inmigración que hay en cada barrio. Para ello se dispone de la tasa bruta de emigración e inmigración, tanto intraurbana como interurbana. Para llevar a cabo el estudio, se ha realizado un análisis de componentes principales, con el objetivo de observar las relaciones entre estas variables.

Antes de ajustar el modelo *PCA*, se ha procedido a detectar si existen valores atípicos o anómalos, para ello, se han utilizado los gráficos *SPE* y *T^2 de hotelling*.

Figura 13. Gráfico T^2 de Hotelling

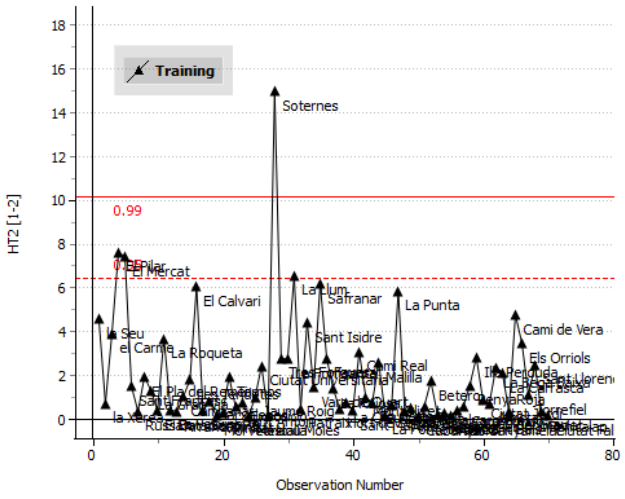
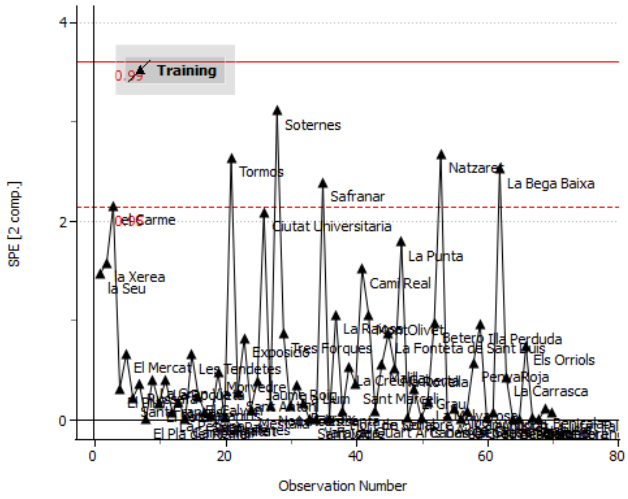


Figura 14. Gráfico Squared Prediction Error



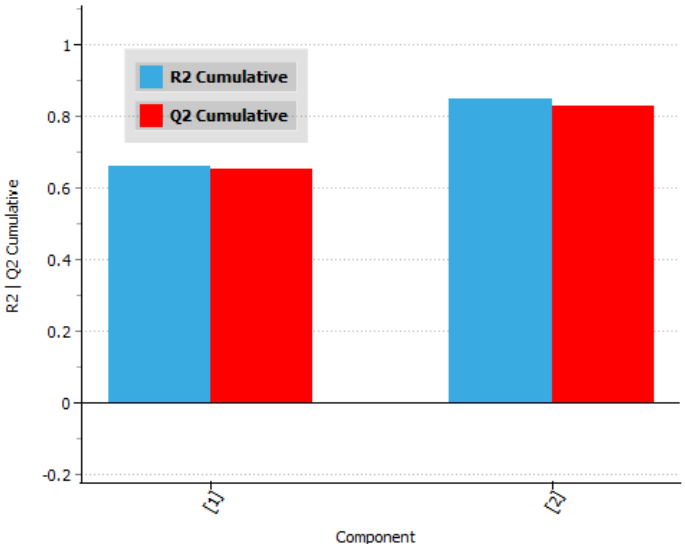
En primer lugar, se observa en el gráfico de T^2 *de Hotelling* que el barrio de Soternes con un valor residual superior al límite de control del 99%. Tras observar el gráfico de contribución se observa un valor elevado en la tasa de emigración interurbana, pero no se trata de un valor demasiado elevado como para considerar eliminarlo.

En segundo lugar, al observar el gráfico *SPE,* no se aprecia ningún barrio que tenga un residuo superior al límite de control del 99%.

Por lo tanto, tras observar ambos gráficos se aprecia que las observaciones respetan la estructura de correlación y tienen unos residuos razonables, por lo que no existen valores atípicos ni anómalos.

A continuación, se ajusta el modelo *PCA* siguiendo el criterio de no sacar más componentes que la mitad del número de variables. El modelo *PCA* ajustado está compuesto por 2 componentes principales que explican el 85% de la variabilidad total.

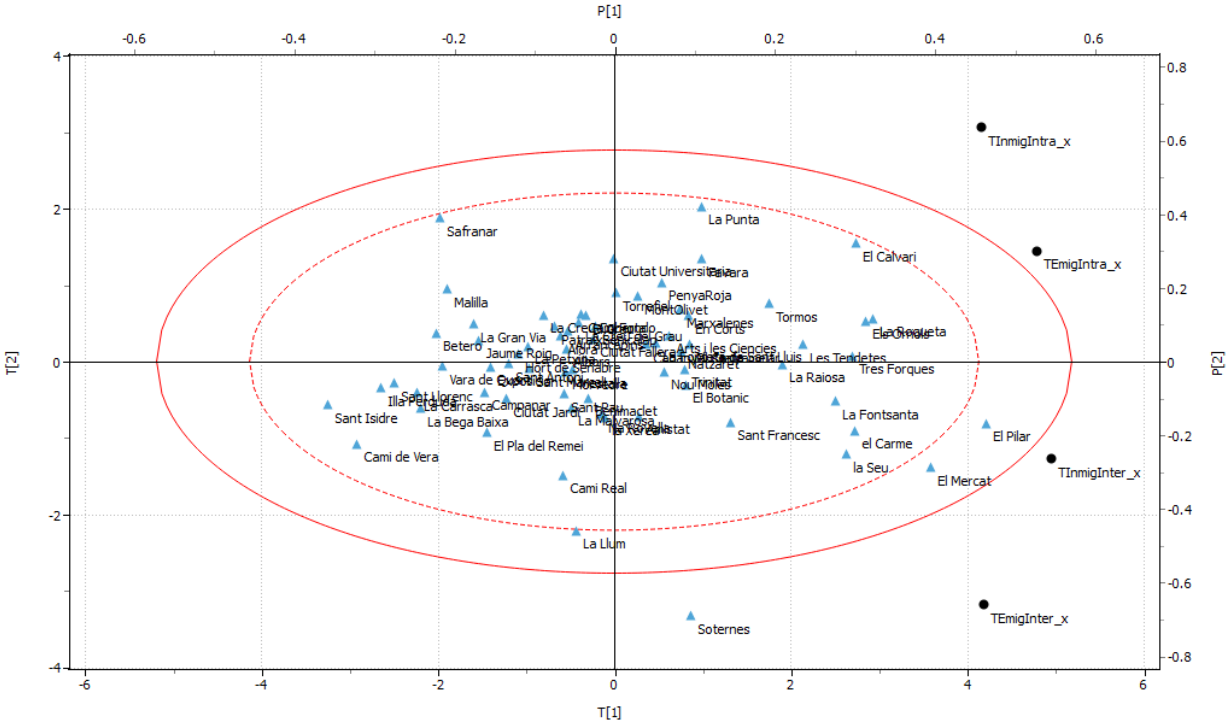
Figura 15. Gráfico de componentes principales



Como se observa en el gráfico, la primera componente principal explica el 66,17% de la variabilidad total, en cambio, la segunda componente explica el 18,86%.

Tras ajustar el modelo se procede a su interpretación, para ello se obtiene el gráfico *biplot*, que muestra las puntuaciones factoriales (*scores*) y las cargas factoriales (l*oadings*) de las 4 variables relativas a movimientos migratorios.

Figura 16. Gráfico Biplot



Como se aprecia en el gráfico, la primera componente diferencia entre barrios con alto y bajo movimiento migratorio, en cambio la segunda componente diferencia entre si dicho movimiento migratorio es de tipo intraurbano o interurbano. Por lo tanto, en los cuadrantes derechos se agrupan los barrios que poseen un alto movimiento migratorio, ya sea de tipo intraurbano o interurbano. Además, se observa que existe una correlación positiva tanto entre los movimientos migratorios interurbanos como entre los intraurbanos.

Por último, los barrios con un alto movimiento migratorio interurbano son: el Pilar, el Mercat y el Carme. En cambio, los barrios con mayor movimiento intraurbano son: el Calvari, la Roqueta y els Orriols.

A fin de resumir las variables sociodemográficas, se ha decidido utilizar las componentes principales obtenidas, en los modelos predictivos que se realizaran en el siguiente objetivo.

Una vez analizadas las relaciones entre los movimientos migratorios a través del *PCA*, con el objetivo de agrupar los barrios en diferentes grupos, en función de sus similitudes sociodemográficas se ha decidido realizar un cluster.

Este análisis pretende Identificar perfiles de barrios en función de los movimientos migratorios. Para ello, se aplicarán dos métodos de clustering: Clustering jerárquico y clustering de partición.

Cluster jerárquico

En primer lugar, se aplicarán los métodos de cluster jerárquico, donde se diferencian dos tipos: los clusters aglomerativos y los clusters disociativos. En el presente trabajo se aplicarán como cluster aglomerativo el método Ward y como cluster disociativo el método Diana.

Método Ward

A continuación, se aplicará un clustering aglomerativo con el método *Ward*, aunque no es necesario en los métodos jerárquicos determinar a priori el número de cluster, se ha decidido obtener el número óptimo de cluster a partir del coeficiente de *Silhoutte*.

Figura 17. Coeficiente de Silhoutte Ward

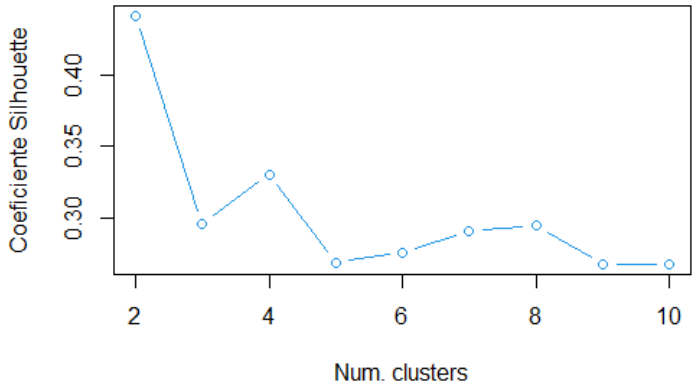
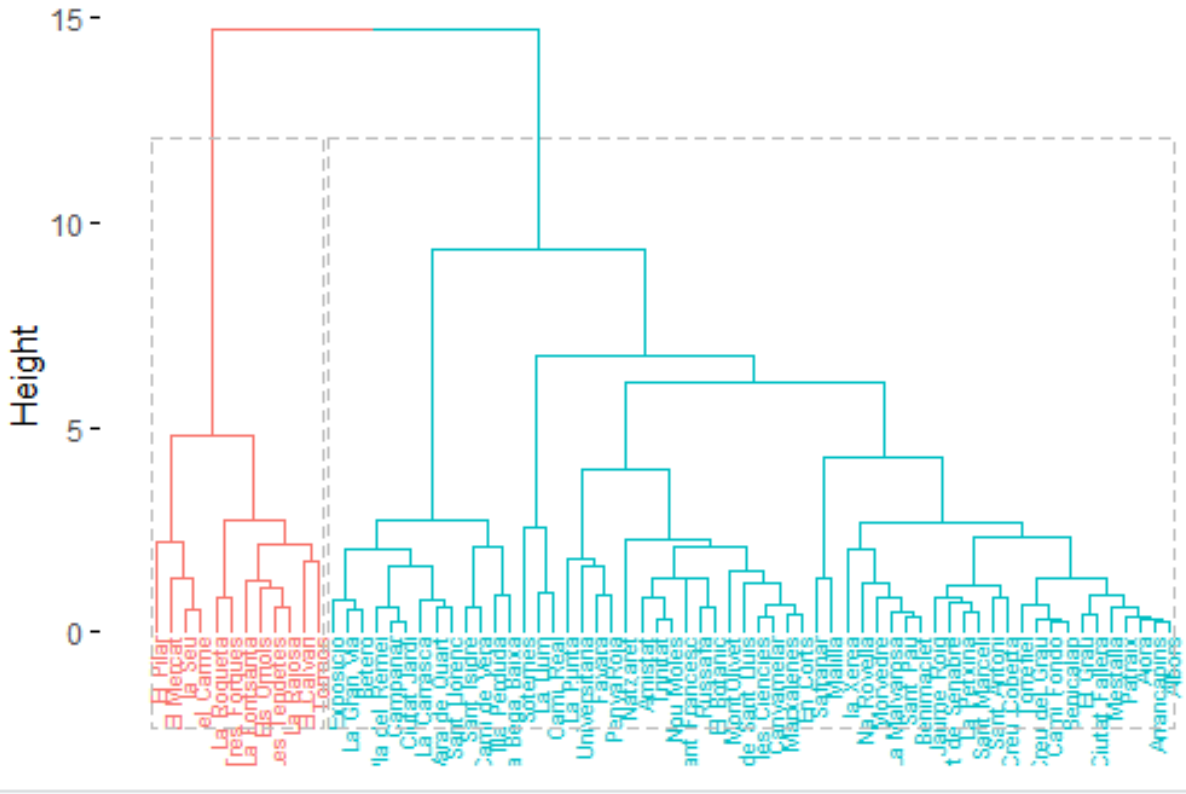


Figura 18. Dendograma método Ward

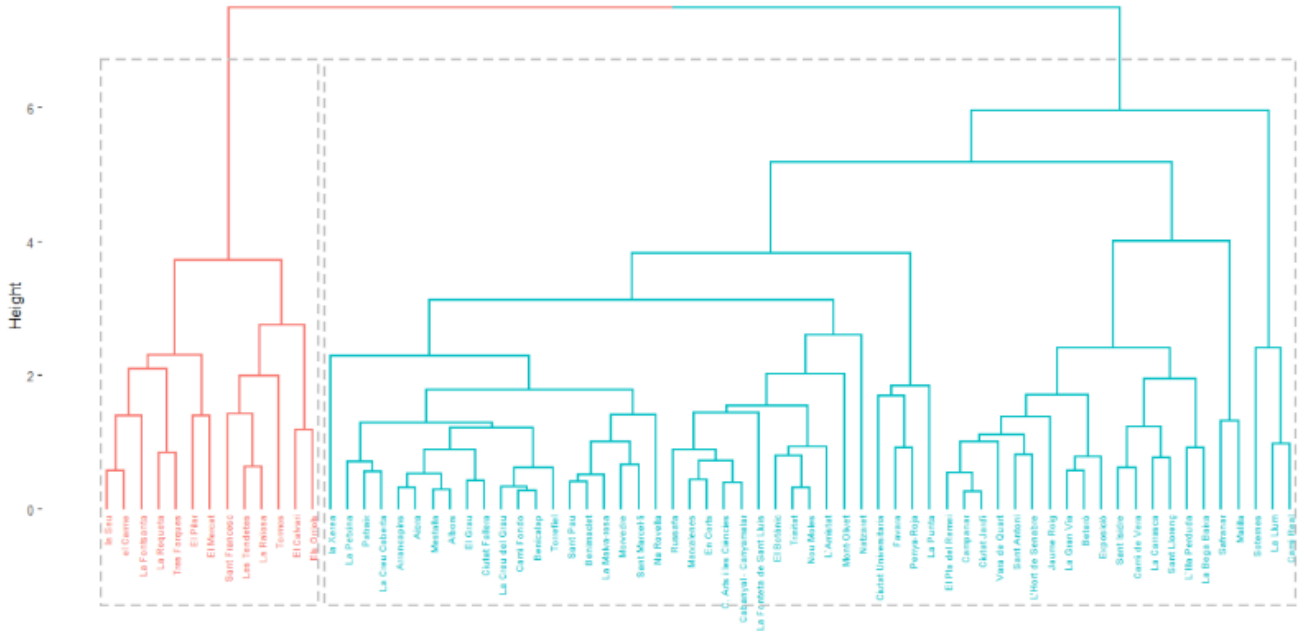


Como se observa en el gráfico, el número de clusters óptimo para el método *Ward* son 2 clusters, donde el primer cluster está formado por 12 barrios y el segundo por 58 barrios.

Método Diana

Como último método jerárquico, se aplicará un cluster disociativo con el método Diana:

Figura 19. Dendograma método DIANA



Utilizando el método diana, los barrios se agrupan en 2 clusters, uno formado por 13 barrios, y otro por 57 barrios

Clustering de partición

En segundo lugar, se utilizará el cluster de partición, donde es necesario determinar a priori el número óptimo de clusters. En el presente trabajo se aplicarán dos métodos*: K-means y k-medoides*.

Método K-means

Para realizar el metodo *k-means*, es necesario determinar el número óptimo de clusters, para ello se utiliza el coeficiente de Silhoutte.

Figura 20. Coeficiente Silhoutte K-means

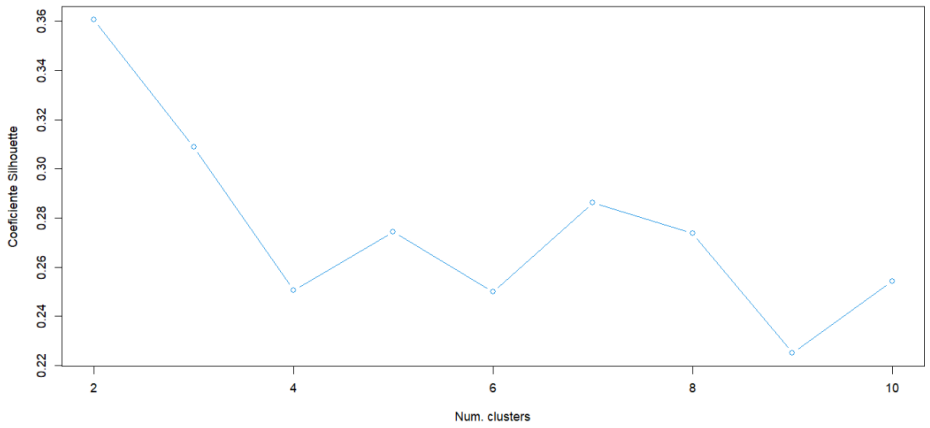
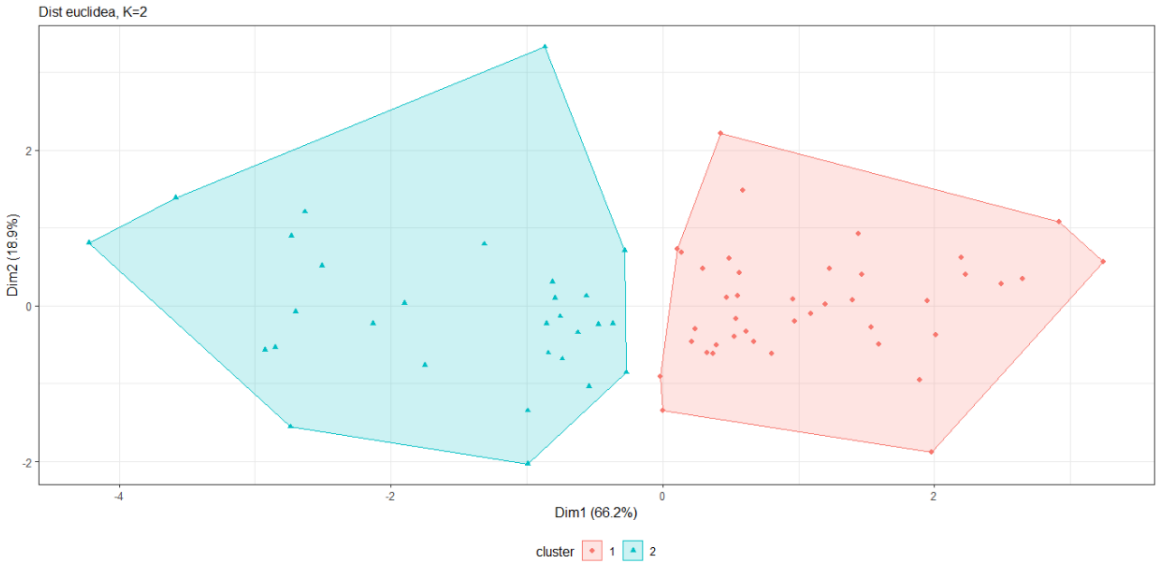


Figura 21. Proyección K-medias



Como se observa en el primer gráfico, el número óptimo de clusters para el método *k-means* es dos. Tras realizar la partición, se obtiene un cluster formado por 41 barrios y otro cluster de 29 barrios.

Método k-medoides

Por último, se utiliza el método *k-medoides*, el cual es bastante similar al *K-means*, y como en el método anterior es necesario determinar a priori el número de clusters.

Figura 22. Coeficiente Silhoutte k-medoides

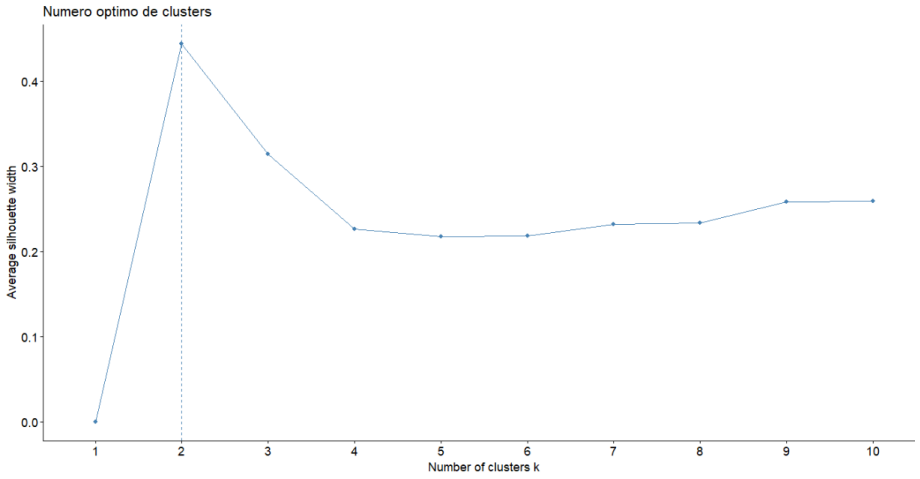
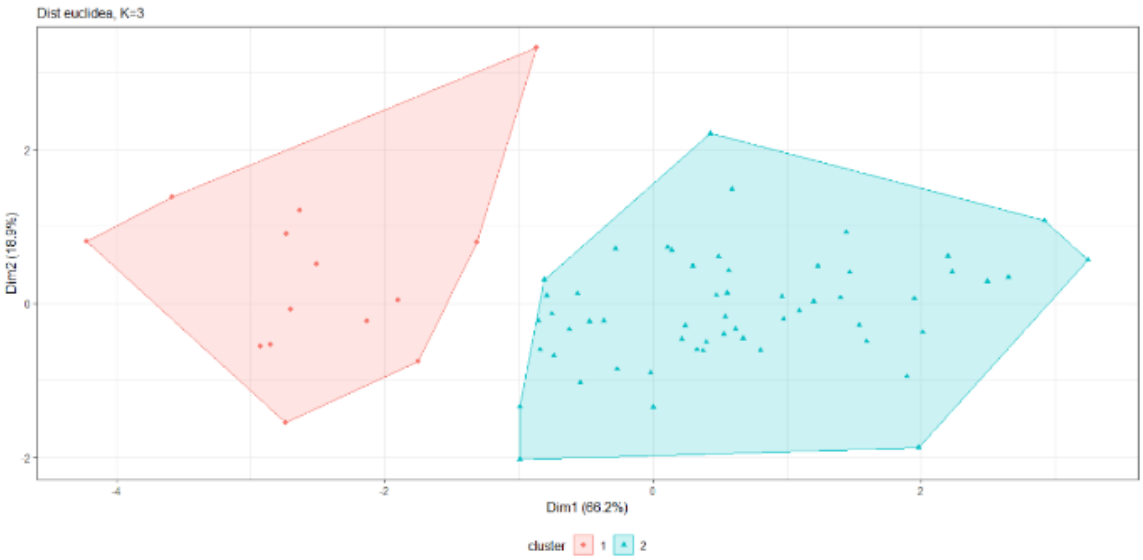


Figura 23. Proyección k-medoides

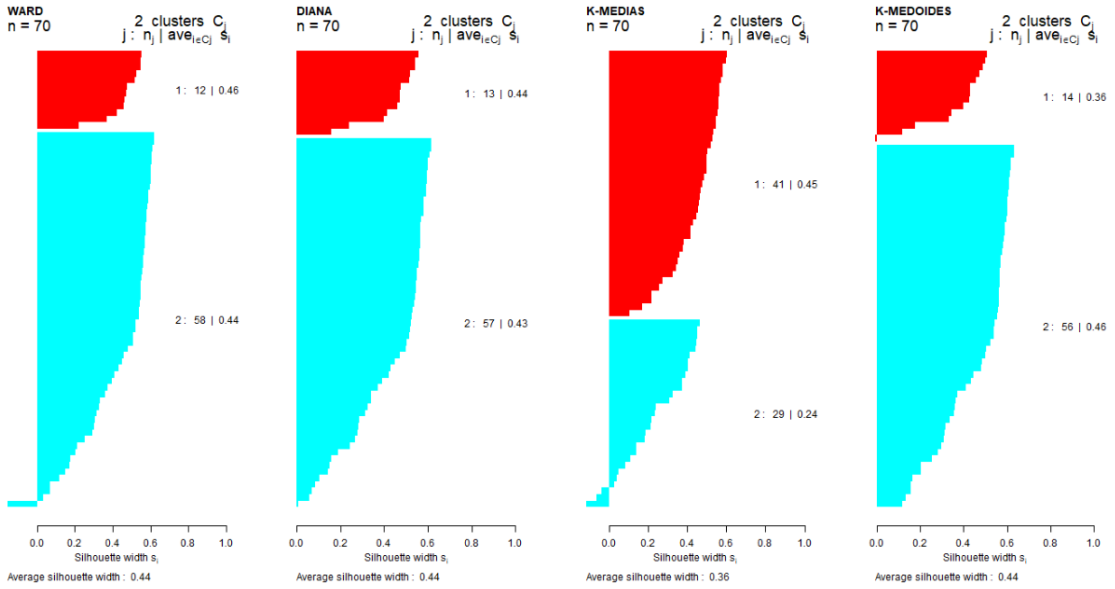


En el primer gráfico, se observa que el número óptimo de cluster para el metodo k-medoides es de dos clusters. El primer cluster esta formado por 14 barrios y el segundo por 56 barrios.

Comparación de resultados

A continuación, se compararán los resultados obtenidos en los diferentes métodos de clustering, y se seleccionará el que presente mejores resultados.

Figura 24. Comparación de los métodos de clustering

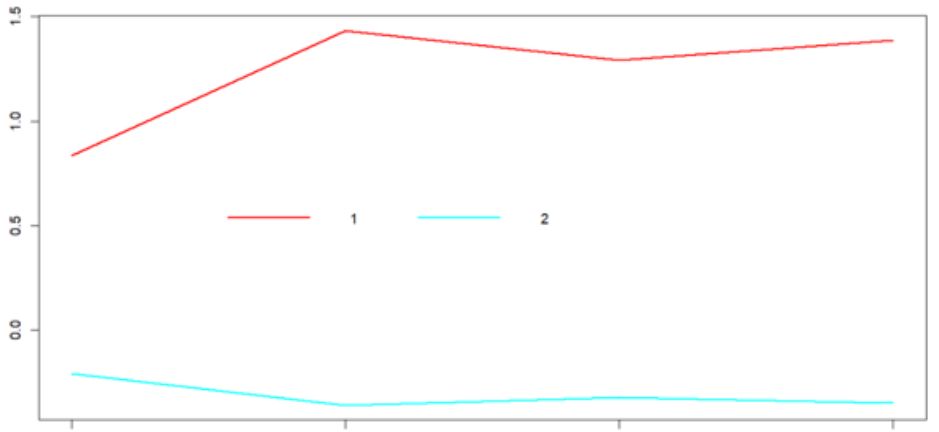


Tras obtener el coeficiente de Silhoutte de todos los métodos, se observa que los métodos *Ward, Diana y k-medoides* tienen un coeficiente de Silhoutte de 0.44. Sin embargo, se aprecia que el método Ward tiene observaciones mal clasificadas (hacia la izquierda), por lo tanto, los mejores métodos de clustering son *Diana y K-medoides*. Por lo tanto, se ha decidido seleccionar el método *K-medoides*, ya que se trata de un método bastante robusto frente a valores anómalos.

Interpretación de los resultados

A continuación, se va a realizar un gráfico descriptivo del perfil de ambos clusteres con el objetivo de ver las diferencias entre ellos. Para ello, se calculará la media de cada variable para cada cluster.

Figura 25. Perfil medio de los clusters



Tasa bruta de inmigración intraurbana	Tasa bruta de inmigración interurbana	Tasa bruta de emigración intraurbana	Tasa bruta de emigración interurbana

Tabla 3. Perfil medio de los clusters

	Cluster 1	Cluster 2
Tasa bruta de inmigración intraurbana	0,84	-0,21
Tasa bruta de inmigración interurbana	1,43	-0,36
Tasa bruta de emigración intraurbana	1,29	-0,32
Tasa bruta de emigración interurbana	1,39	-0,35

Como se ha podido comprobar, el cluster 1 (14 barrios) está formado por barrios con alto movimiento migratorio respecto a la media, en cambio, el cluster 2 (56) son barrios con un menor movimiento migratorio.

Por último, como complemento a lo visto en los resultados,se ha realizado un boxplot para relacionar los clusters con el valor catastral y el precio de venta del m^2.

Figura 26. Boxplot Valor catastral

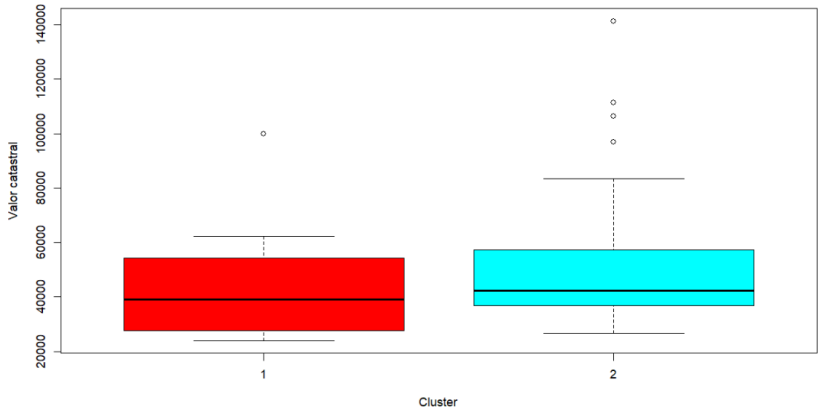
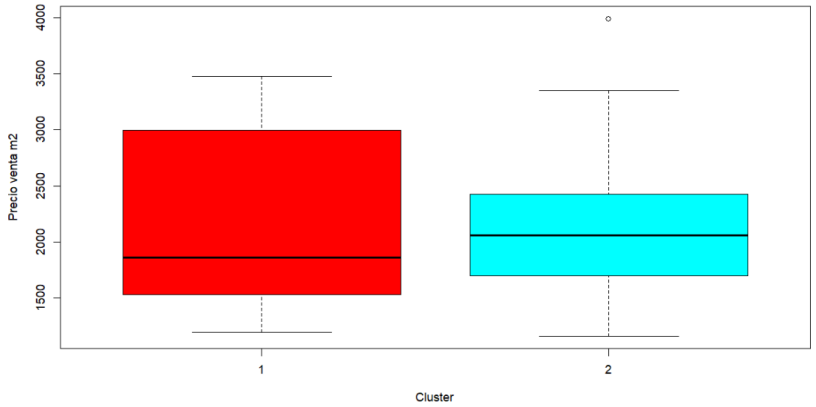


Figura 27. Boxplot precio de venta del m^2



Observando ambos gráficos, en principio no se aprecian diferencias significativas como para afirmar que el movimiento migratorio afecta en el valor catastral y el precio de venta del m^2, aunque esto se comprobara posteriormente.

4.2.2 Análisis de las relaciones entre los servicios de los barrios de Valencia

En segundo lugar, se ha decidido estudiar los servicios de los barrios de Valencia, ya que, a la hora de tomar una decisión de compra, se trata de un factor clave. Para ello se dispone del número de actividades económicas en los barrios, y su tipo: Comercio, restaurante, hostelería y reparación, Transporte y comunicación, Instituciones financieras, servicios a empresas, seguros y alquileres y Otros servicios. Para llevar a cabo el estudio, se ha realizado un análisis de componentes principales, con el objetivo de observar las relaciones entre estas variables.

Antes de ajustar el modelo PCA se comprueba a través de los gráficos T^2 *de Hotelling* y *SPE* si existen valores anómalos o atípicos.

Figura 28. Gráfico T^2 de Hotelling

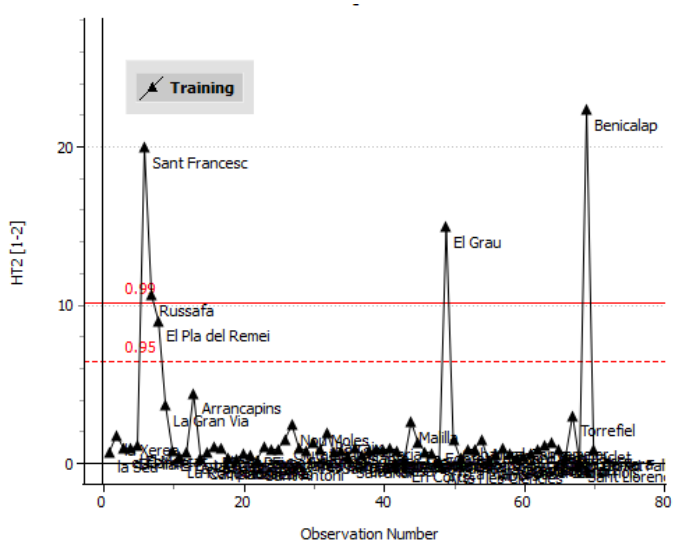
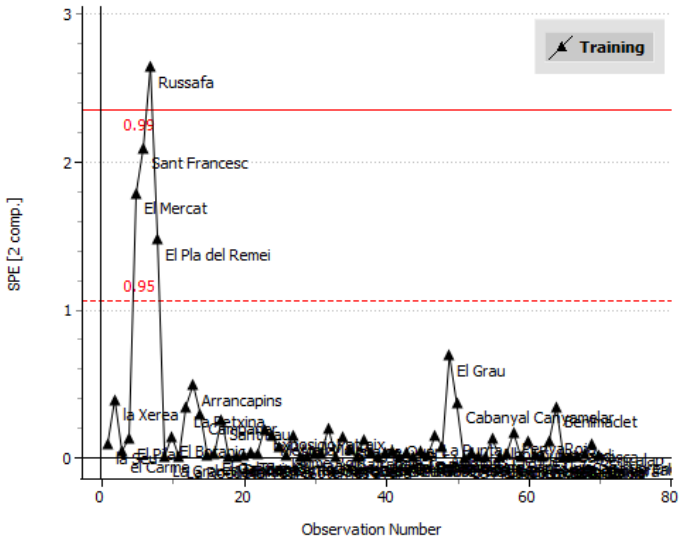


Figura 29. Gráfico SPE



Por un lado, en el gráfico T^2 *de Hotelling* se aprecia un barrio con un alto valor residual, ya que el límite de control este situado en torno a 11 y este barrio supera dos veces ese límite. Tras observar el gráfico de contribución del barrio de Benicalap, se aprecia un valor alto en transporte y comunicación, por lo tanto, se ha considerado no eliminar dicha observación, ya que esta información es correcta y su inclusión en análisis se considera interesante.

Por otro lado, en el gráfico *SPE* se aprecia que el barrio de Russafa tiene valor residual superior al límite de control del 99% (riesgo de 1ºespecie), observando el gráfico de contribución se aprecia un número elevado de Comercios, restaurantes, hostelería y reparaciones, aunque se trata de un valor elevado no difiere en gran medida con otros barrios, por lo que se considera un valor aceptable. Por lo tanto, las observaciones respetan la estructura de correlación, por lo que no existen valores atípicos.

A continuación, se ajusta el modelo *PCA*, que este compuesto por 2 componentes principales que explican el 94,9% de la variabilidad total.

Figura 30. Gráfico de componentes principales

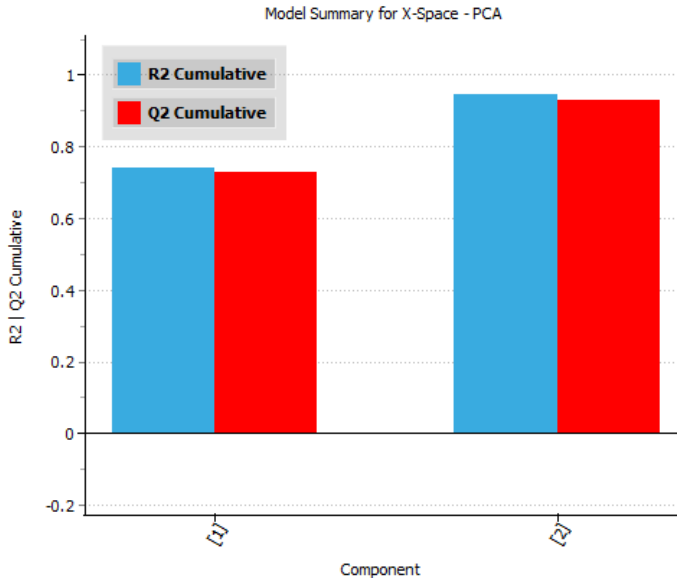
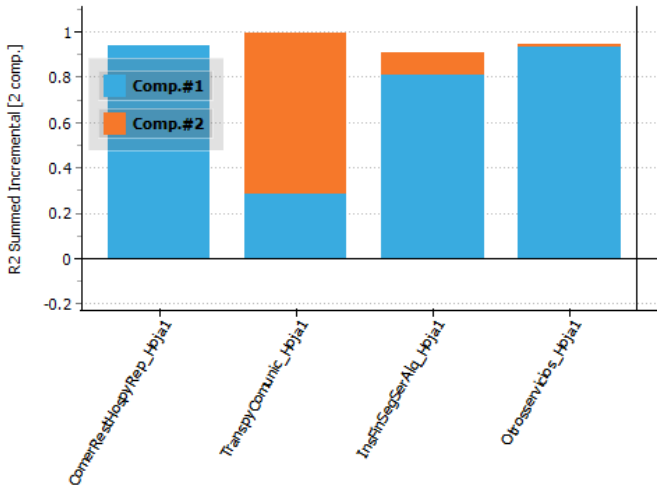


Figura 31. Gráfico R2 de las variables

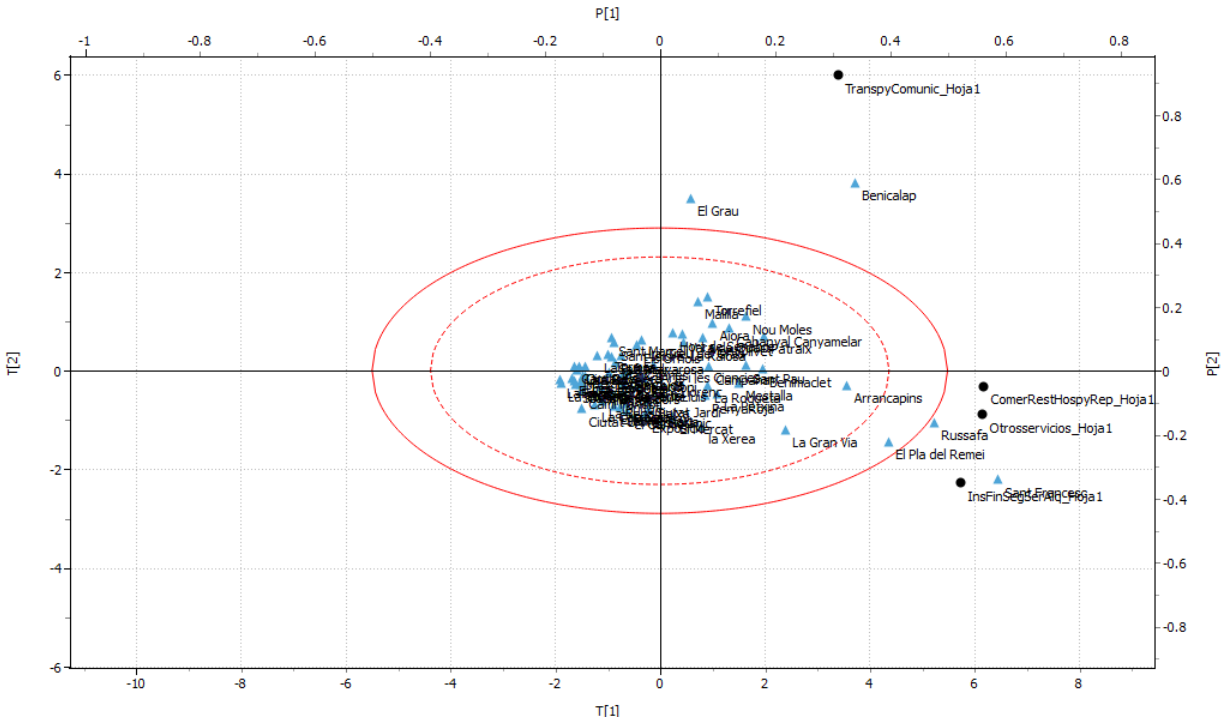


Como se observa en la figura, la primera componente explica el 74,26% de la variabilidad total y la segunda componente el 20,63%. Además, en el gráfico de *R2* de variables se aprecia la variabilidad explicada de cada variable por las componentes principales obtenidas.

Donde la primera componente principal está asociada a: comercios, restaurantes, hostelería, reparaciones, instituciones financieras, seguros y otros servicios, en cambio la segunda componente explica fundamentalmente el transporte y la comunicación.

Tras ajustar el modelo se procede a su interpretación, para ello se obtiene el gráfico *biplot*, que muestra los scores y loadings.

Figura 32. Gráfico Biplot



Observando el *biplot* de la primera y segunda componente, se aprecia una dirección de la variabilidad de forma horizontal debido a 3 variables (Comercio, restaurante, hostelería y reparación, Transporte y comunicación, Instituciones financieras, servicios a empresas, seguros y alquileres y Otros servicios) que están en su mayoría explicada por la primera componente y correlacionadas entre sí. Se aprecia que los barrios que tienen un mayor número de servicios relacionados con estas 3 variables correlacionadas son: Sant Francesc, Russafa y el Pla del Remei.

Figura 33. Gráfico de scores según comercio y hostelería

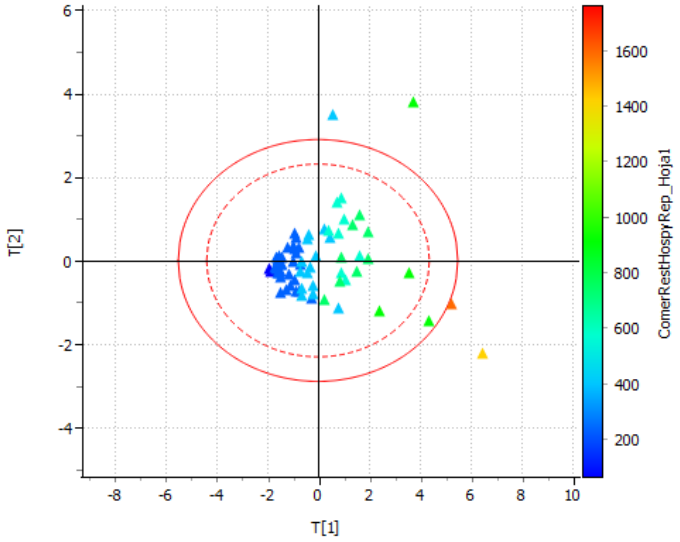
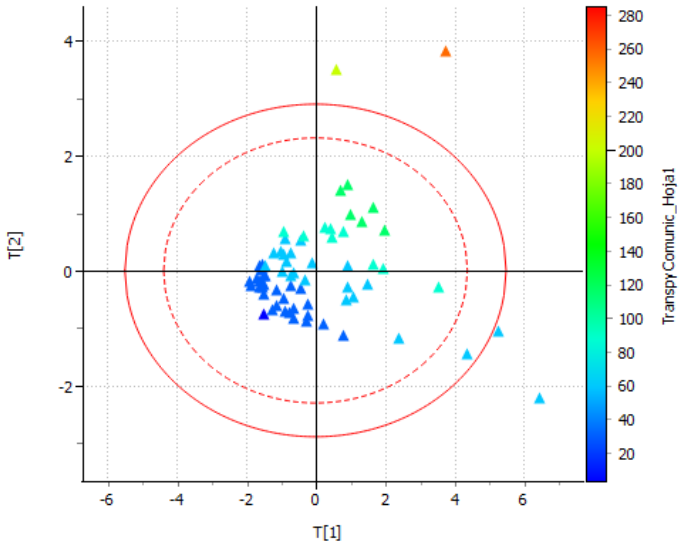


Figura 34. Scores según transporte y comunicación



Además, también se aprecia una dirección de la variabilidad de forma vertical debido a la variable transportes y comunicación que esta explicada por la segunda componente. En el gráfico se observa que los barrios de Benicalap y el Grau tienen un gran número de actividades relacionadas con el transporte y comunicación, lo que era esperable en el barrio del Grau, debido a que se trata de la zona del puerto, por lo que habrá un mayor número de empresas dedicadas al transporte tanto terrestre como marítimo.

Tras analizar las relaciones entre las características económicas de los barrios a través del *PCA,* con el objetivo de agrupar los barrios en diferentes grupos en función de sus similitudes se realizará un análisis cluster. Este análisis pretende identificar perfiles de barrios en función de sus servicios disponibles. Para llevar a cabo este análisis, se aplicarán dos métodos de clustering: Clustering jerárquico y clustering de partición.

Clustering jerárquico

Método Ward

En primer lugar, se aplicará un clustering aglomerativo con el método *Ward*. Aunque no es necesario, se ha utilizado el coeficiente de Silhoutte para determinar el número óptimo de cluster.

Figura 35. Coeficiente de Silhoutte Ward

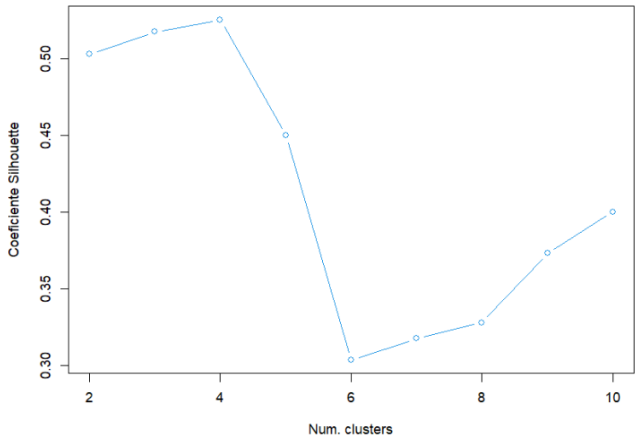
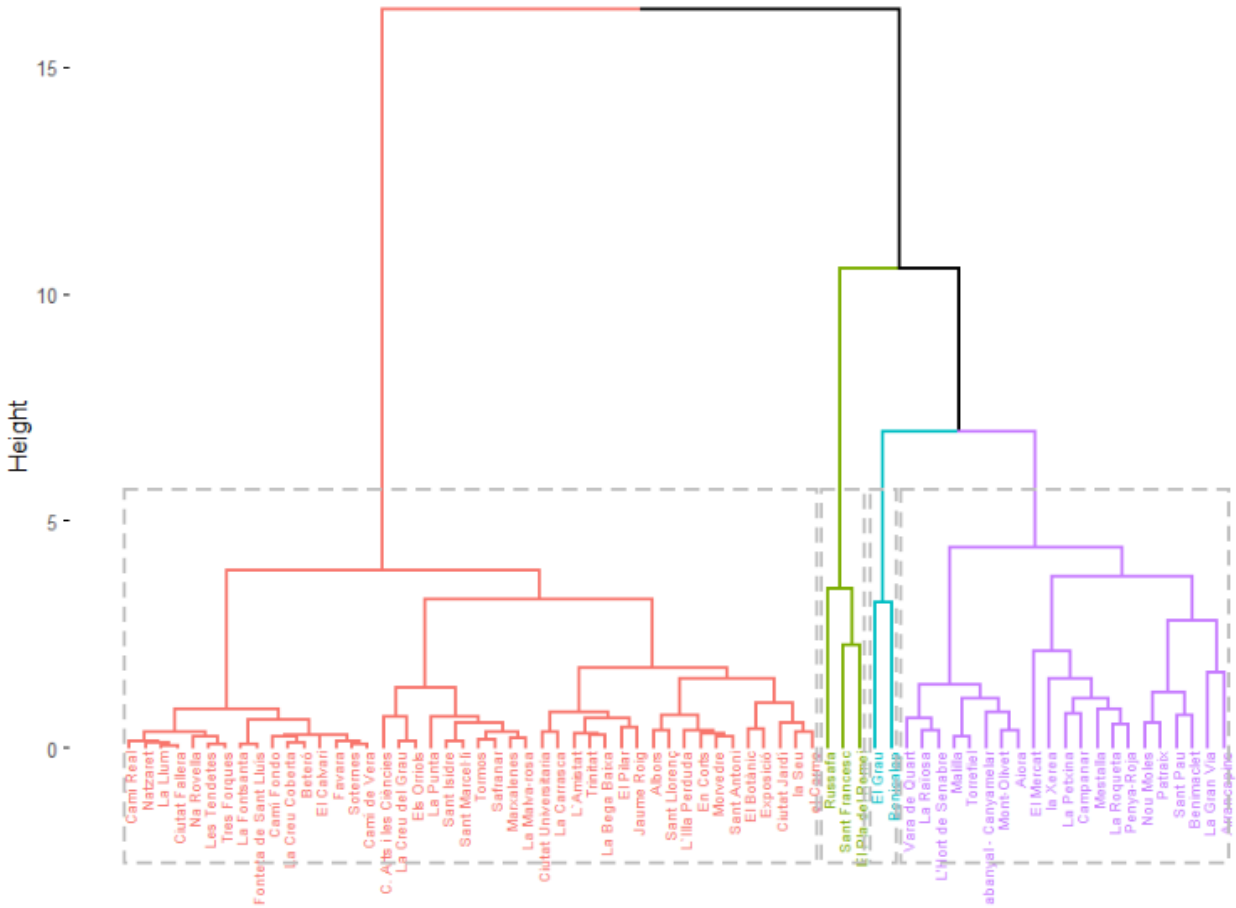


Figura 36. Dendograma Ward



Como se observa en el gráfico, el número de clusters óptimo para el método *Ward* son 4 clusters, donde el primer cluster está formado por 44 barrios, el segundo por 21 barrios, el tercero por 3 barrios y el cuarto por 2 barrios.

Método Diana

En segundo lugar, se aplicará un cluster disociativo con el método Diana. Con el objetivo de determinar el número de clusters se ha realizado el mapa de color, el cual indica que existen 2 cluster.

Figura 37. Mapa de calor

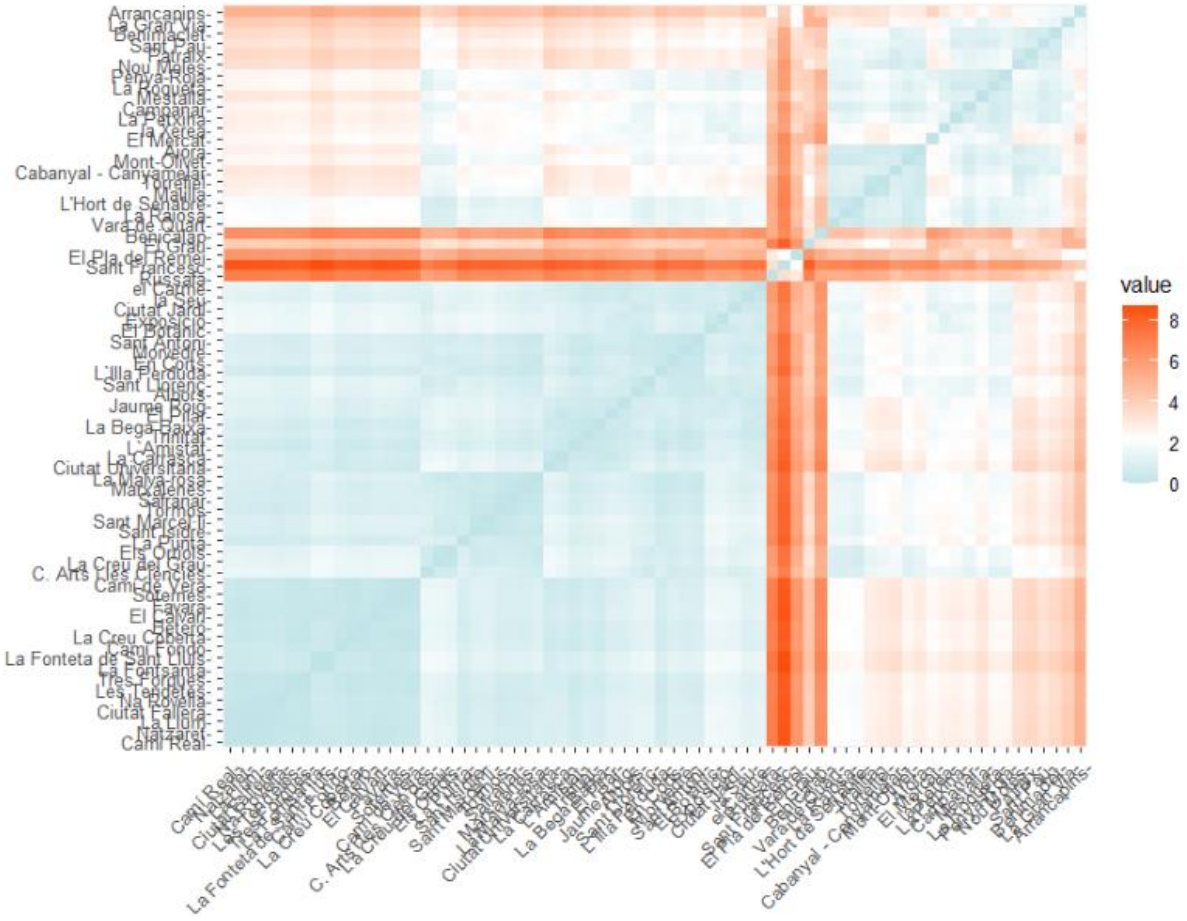
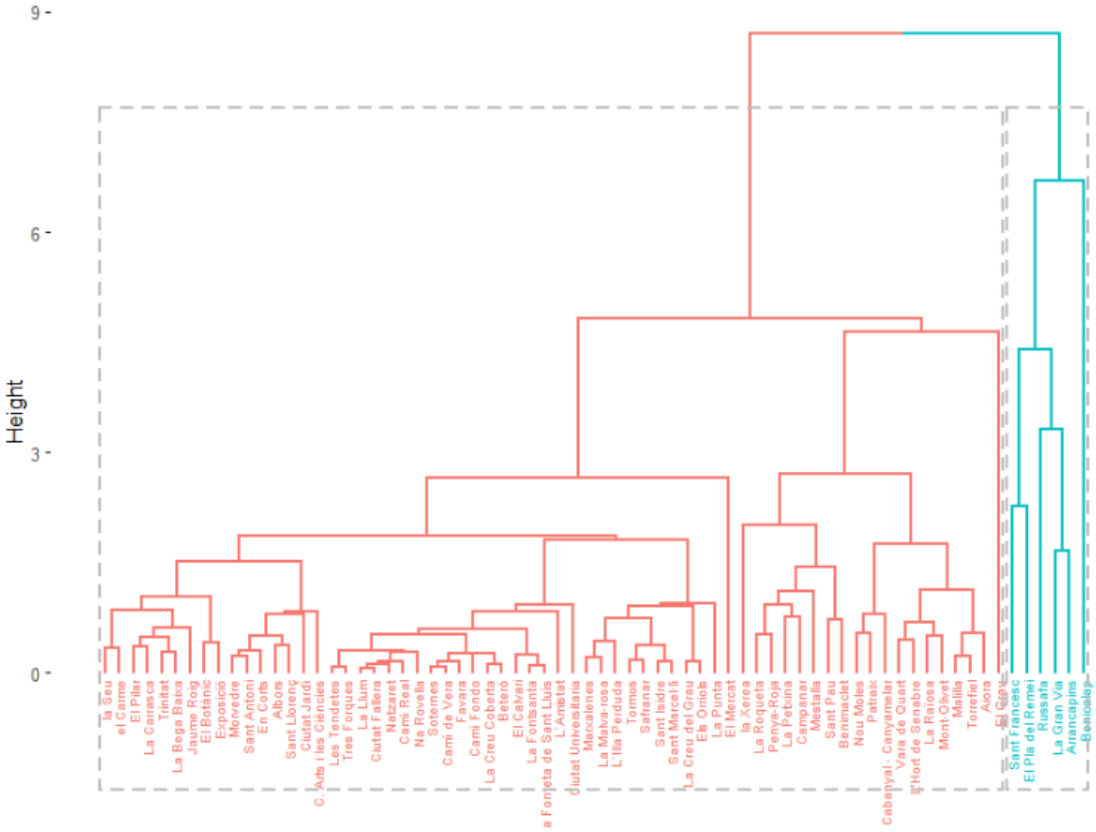


Figura 38. Dendograma DIANA



Aplicando el método Diana, los barrios se agrupan en 2 clusters, uno formado por 64 barrios, y otro por 6 barrios.

Clustering de partición

Método k-means

A continuación, se aplicará el método de clustering *k-means*, para ello es necesario determinar a priori el número de clusters, por lo que se utilizara el coeficiente de Silhoutte para obtener el número óptimo de clusters.

Figura 39. Coeficiente de Silhoutte k-means

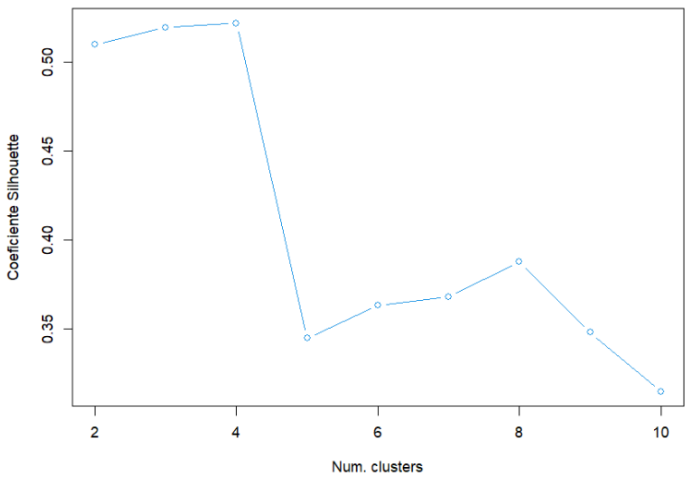
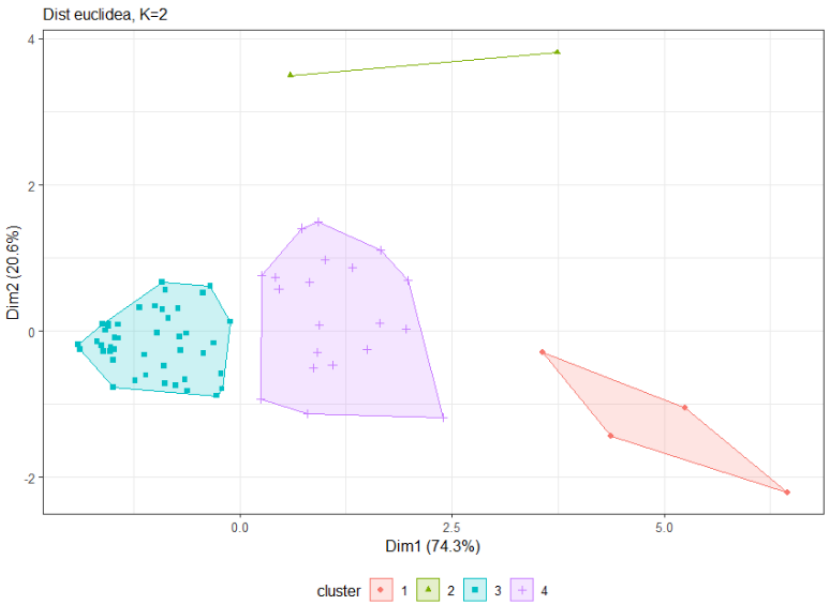


Figura 40. Proyección k-means



Como se observa en el primer gráfico, el número óptimo de clusters para el método *k-means* es cuatro. Tras realizar la partición, se obtiene un primer cluster formado por 4 barrios, un segundo cluster formado por 2 barrios, un tercer cluster formado por 44 barrios y un último cluster formado 20 barrios.

Método k-medoides

Por último, se aplicará el método *k-medoides*, que al igual que el *k-means* es necesario determinar a priori el número de clusters.

Figura 41. Coeficiente de Silhoutte k-medoides

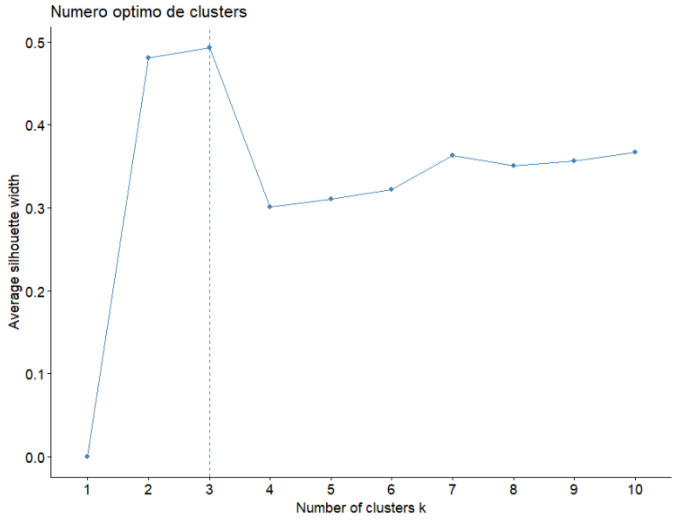
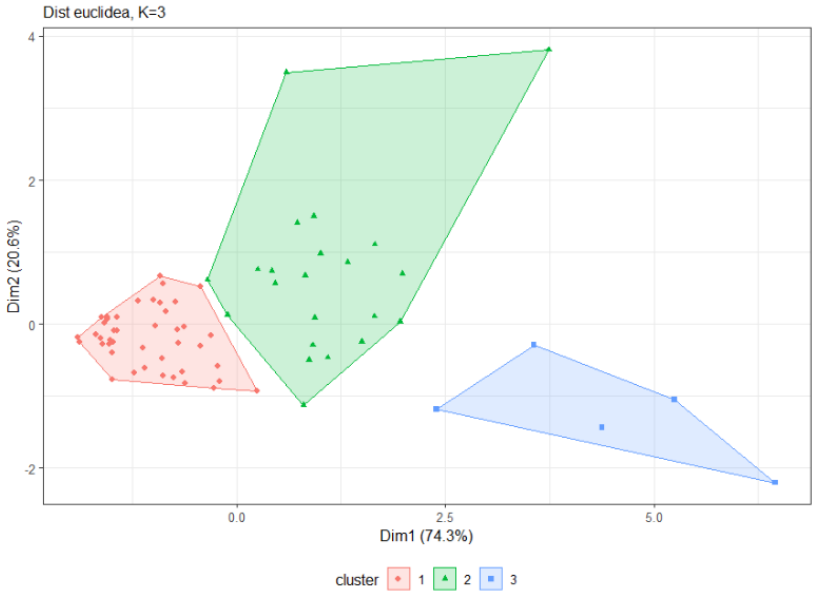


Figura 42. Proyección K-medoides

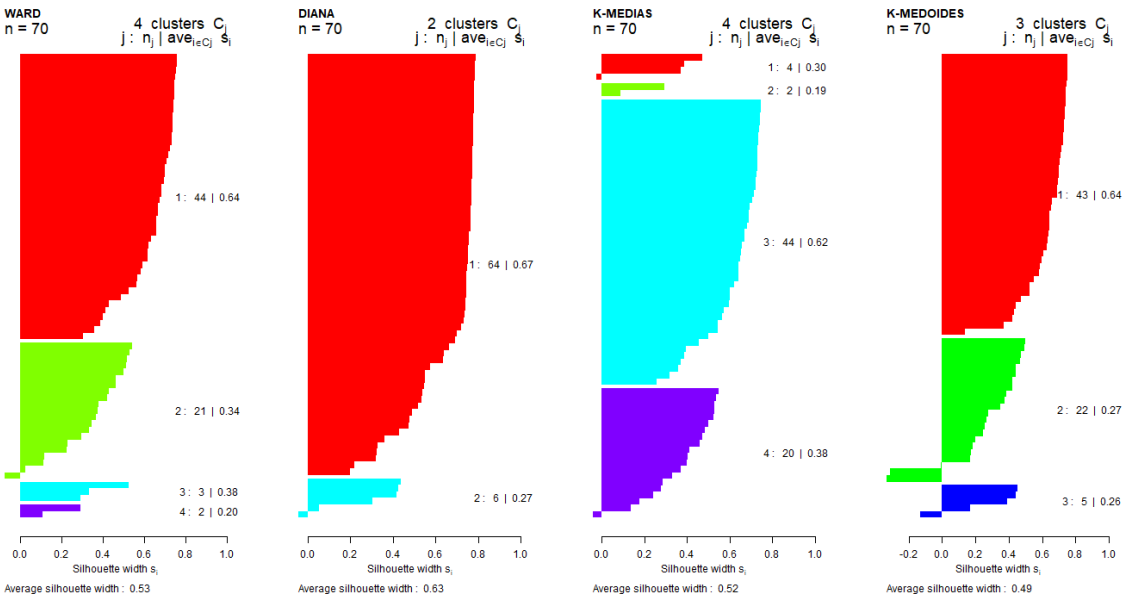


Como se observa en el gráfico, para el método *k-medoides* el número óptimo de clusters es de tres. Tras aplicar el clustering *k-medoides* con tres clusters, se obtiene un primer cluster formado por 43 barrios, un segundo cluster formado por 22 barrios y un último cluster formado por 5 barrios.

Comparación de resultados

En este apartado, se compararán los resultados obtenidos en los diferentes métodos de clustering, y se seleccionará el que presente mejores resultados.

Figura 43. Comparación de métodos a través del coeficiente de Silhoutte



Como se puede observar en el gráfico, el método de clustering Diana es el que presenta mejores resultados, ya que es el que tiene un mayor valor de coeficiente Silhoutte (0,63), por lo tanto, es el método que mejor ha agrupado los barrios. Además, es el que presenta un menor número de barrios mal clasificados (observaciones a la izquierda).

Interpretación de los resultados

A continuación, se va a realizar un gráfico descriptivo del perfil de ambos cluster con el objetivo de ver las diferencias entre ellos. Para ello, se calculará la media de cada variable para cada cluster.

Figura 44. Gráfico del perfil medio de los clusters

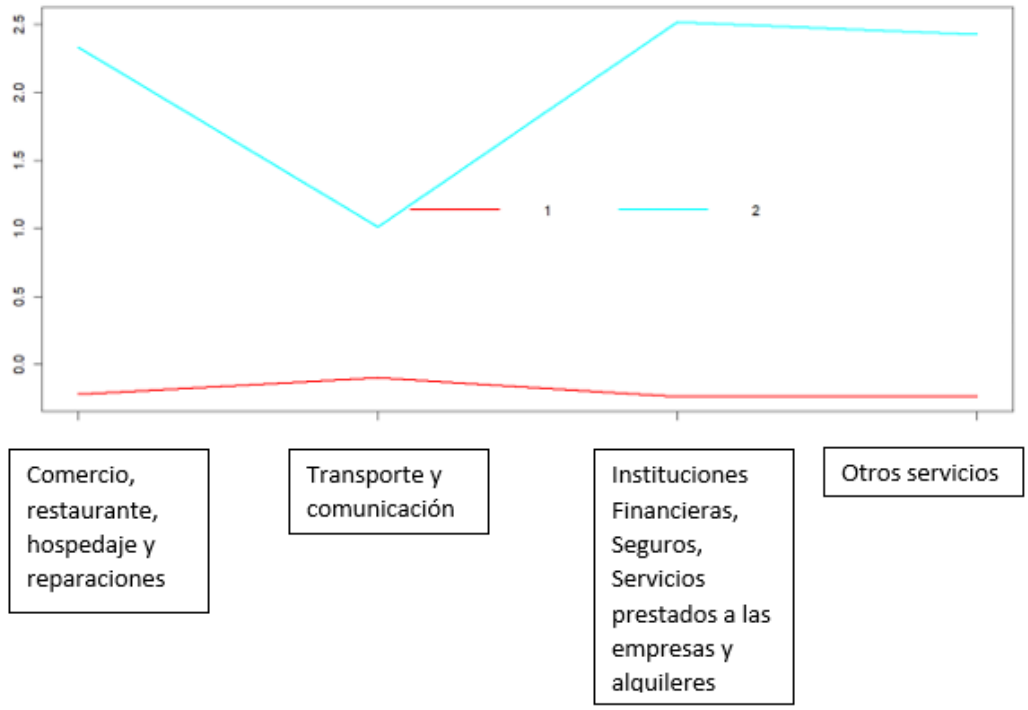


Tabla 4. Perfil medio de los clusters

	Cluster 1	Cluster 2
Comercio, restaurante, hospedaje y reparaciones	-0,22	2,33
Transporte y comunicación	-0,09	1,01
I.Financieras, Seguros, Servicios prestados a las empresas y alquileres	-0,24	2,51
Otros servicios	-0,23	2,42

Como se ha podido comprobar, el cluster 2 (6 barrios) está formado por barrios con alta actividad económica respecto a la media, en cambio, el cluster 1 (64) son barrios con un menor número de servicios. Los barrios que conforman el cluster 2 son: Sant Francesc, Russafa, El Pla del Remei, Gran Via, Arrancapins y Benicalap.

Con el objetivo de enlazar los resultados obtenidos en el cluster con el siguiente objetivo, se ha realizado un *boxplot* para relacionar estos resultados con las variables de valor catastral y precio de venta del m^2.

Figura 45. Boxplot Valor catastral

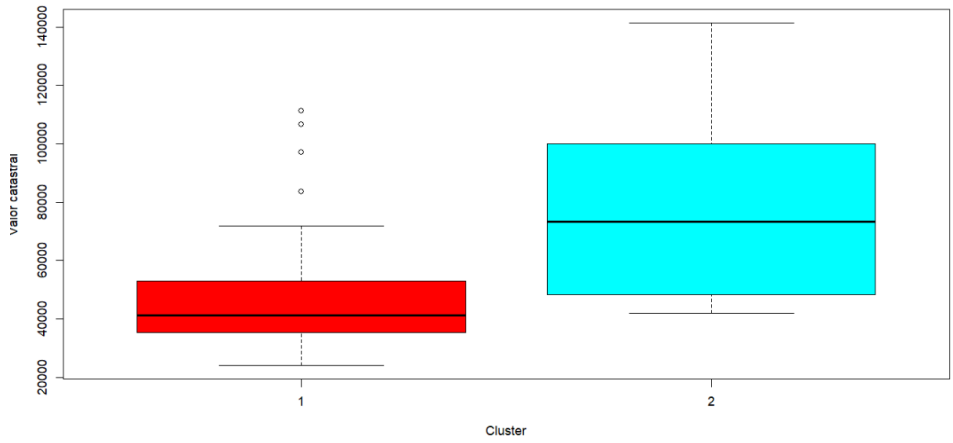
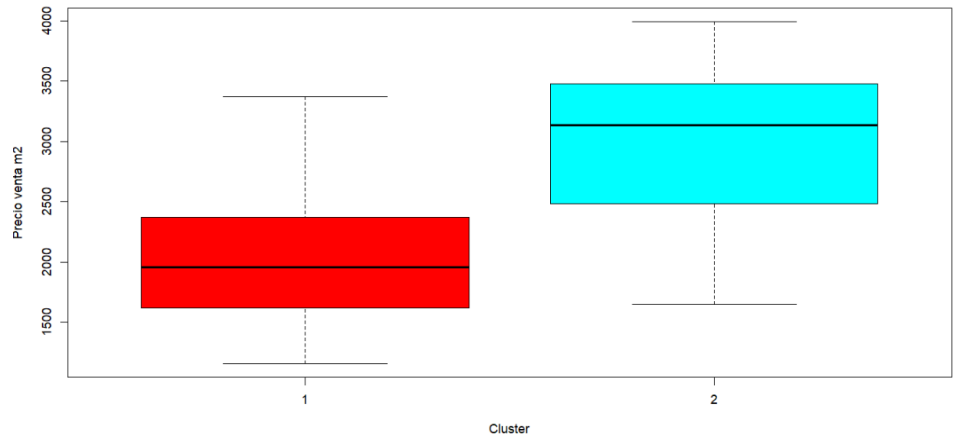


Figura 46. Boxplot Precio de venta del m²



En ambos gráficos, no se observan diferencias significativas, como para afirmar que el número de servicios afecta al Valor catastral o al precio de venta del m^2, aunque esto se confirmara posteriormente.

4.2.3 Análisis de correspondencias múltiples

Tras realizar un análisis especifico de estos bloques, se realizará un análisis de correspondencias múltiples al conjunto de las 17 variables, con el objetivo de resumir toda la cantidad de datos en un número reducido de dimensiones, con la menor pérdida de información posible. Para poder realizar este análisis, es necesario categorizar las variables, para ello se ha decidido, que los barrios que tengan un valor superior al tercer cuartil en dichas variables se clasificaran como valor “alto” de esas variables y el resto como valor “bajo”.

Figura 47. Pesos de las variables en la 1º y 2º dimensión MCA

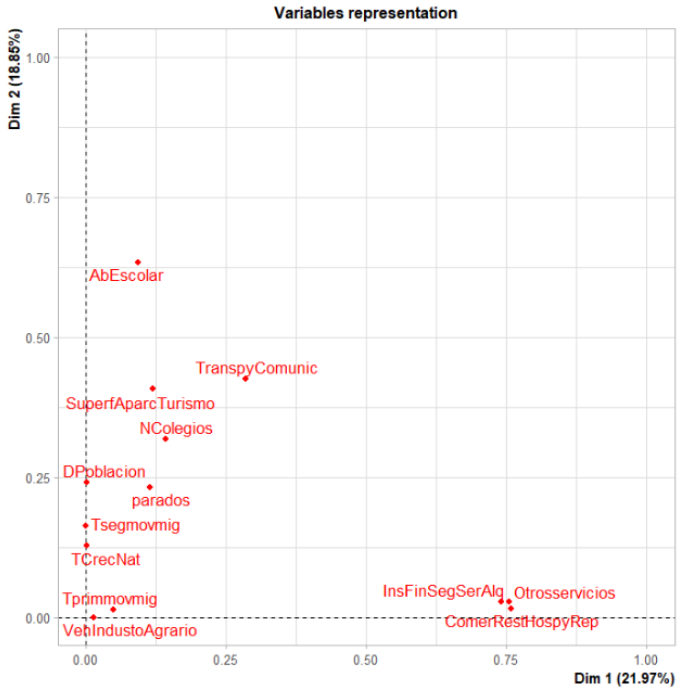


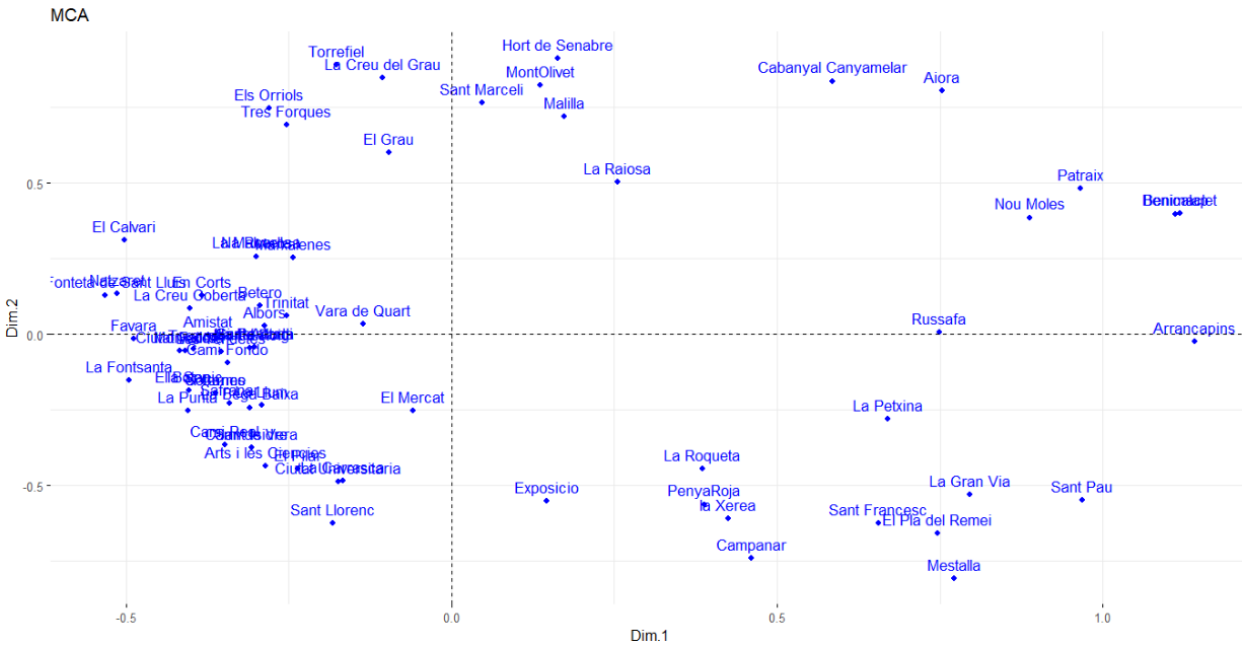
Figura 48. Gráfico biplot MCA

En el primer gráfico se ha representado la primera y segunda dimension, donde la primer componente explica el 22% de la variabilidad total y la segunda componente el 19%.

Se puede observar que las variables que tienen más peso en la primera componente son las variables relacionadas con los servicios que son: Instituciones financieras,Seguros, Comercio, Hosteleria y Otros servicios. En cambio, en la segunda componente las variables que tienen más peso son: Abandono escolar, Transporte y comunicación,Superficie de aparcamiento por turismo y Número de colegios.

En el segundo gráfico, se puede observar las relaciones entre las variables, donde destaca una correlación positiva entre Alto número de colegios, alto abandono escolar y alto transporte y comunicación. Además, se puede observar la correlación positiva vista anteriormente en el *PCA* entre las variables relacionadas con los servicios.

Figura 49. Gráfico de scores MCA



En el gráfico, se puede observar que se forman grupos de barrios, como en la parte inferior derecha que son barrios con una alta actividad económica, bajo porcentaje de paro y mucha superficie de aparcamiento por turismo. Por otro lado, también se forma otro grupo en la parte superior que son barrios con alto abandono escolar, alto número de colegios, alto número de transporte y comunicación,alta densidad de población y alto porcentaje de parados.

Por último, se ha querido relacionar los resultados vistos en el análisis de correspondencias múltiples con las variables de valor catastral total y precio de venta del m^2, con el objetivo de anticiparse a lo que se comprobara en el siguiente objetivo.

Figura 50. Gráfico de scores según el valor catastral

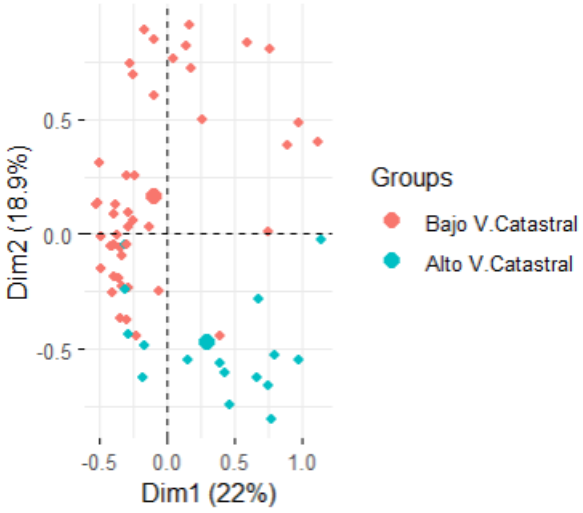
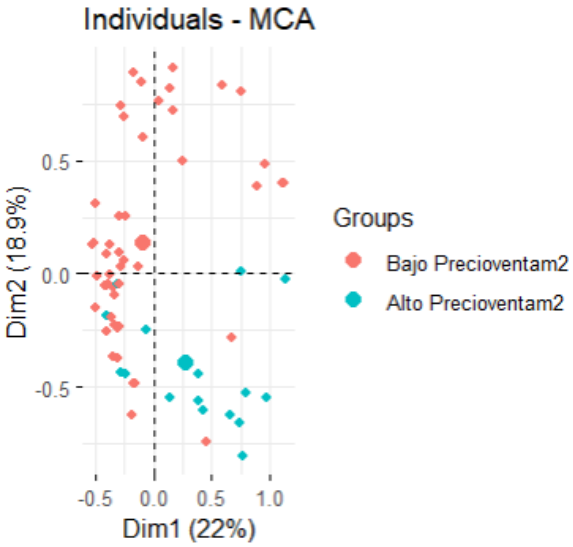


Figura 51. Gráfico de scores según Precio de venta del m^2



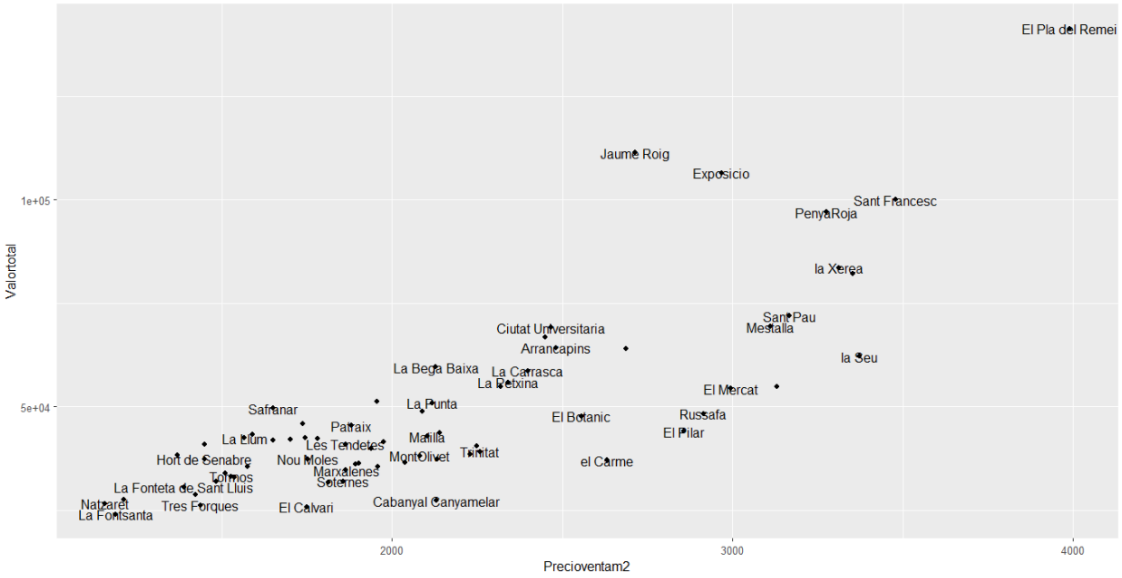
Como se puede observar en ambos gráficos, en el cuadrante inferior derecho se encuentran los barrios con mayor valor catastral y precio de venta del m^2. Estos barrios se caracterizaban por tener una alta actividad económica, bajo porcentaje de paro y mucha superficie de aparcamiento por turismo. Si estas variables realmente influyen en el valor de los barrios se comprobara en el siguiente objetivo.

4.3 Modelos predictivos para la valoración del suelo en barrios de Valencia según el valor catastral y el precio de mercado por m^2

En este apartado, se pretende predecir el valor catastral y el precio de mercado del m^2, a partir de la información disponible de cada barrio. Para lograr este objetivo se utilizarán 6 modelos supervisados de regresión: Árbol de regresión, *Random Forest*, Vecino más próximo, Máquinas de soporte vectorial, *Gradient Boosting* y *PLS*. Se analizarán los resultados obtenidos en las diferentes técnicas y se seleccionará el modelo que presente mejores resultados

Antes de empezar a aplicar las técnicas, se ha realizado un gráfico entre ambas variables a predecir, para observar su relación.

Figura 52. Gráfico del Valor catastral frente el precio de venta del m^2



Como se puede observar en el gráfico, existe una correlación positiva (0,8) entre ambas variables hasta un precio del m^2 de 2500€. A partir de dicho valor se aprecia una dispersión en los datos, donde existen barrios con un valor catastral bajo, pero un precio del m^2 bastante elevado (el Carme y el Pilar), y al contrario un valor catastral alto pero un precio del m^2 bajo (Jaume Roig y Exposicio).

A pesar de ser dos variables, que están bastante relacionadas entre sí, ya que ambas miden el valor de las viviendas, se ha decidido predecir ambas variables. Debido a que no es lo mismo el valor catastral que lo establece el ministerio de Hacienda (que utiliza el valor como base para aplicar impuesto), que el precio del m^2 que lo establece el mercado a través de las operaciones de compraventa que se ejecutan y al realizar tasaciones.

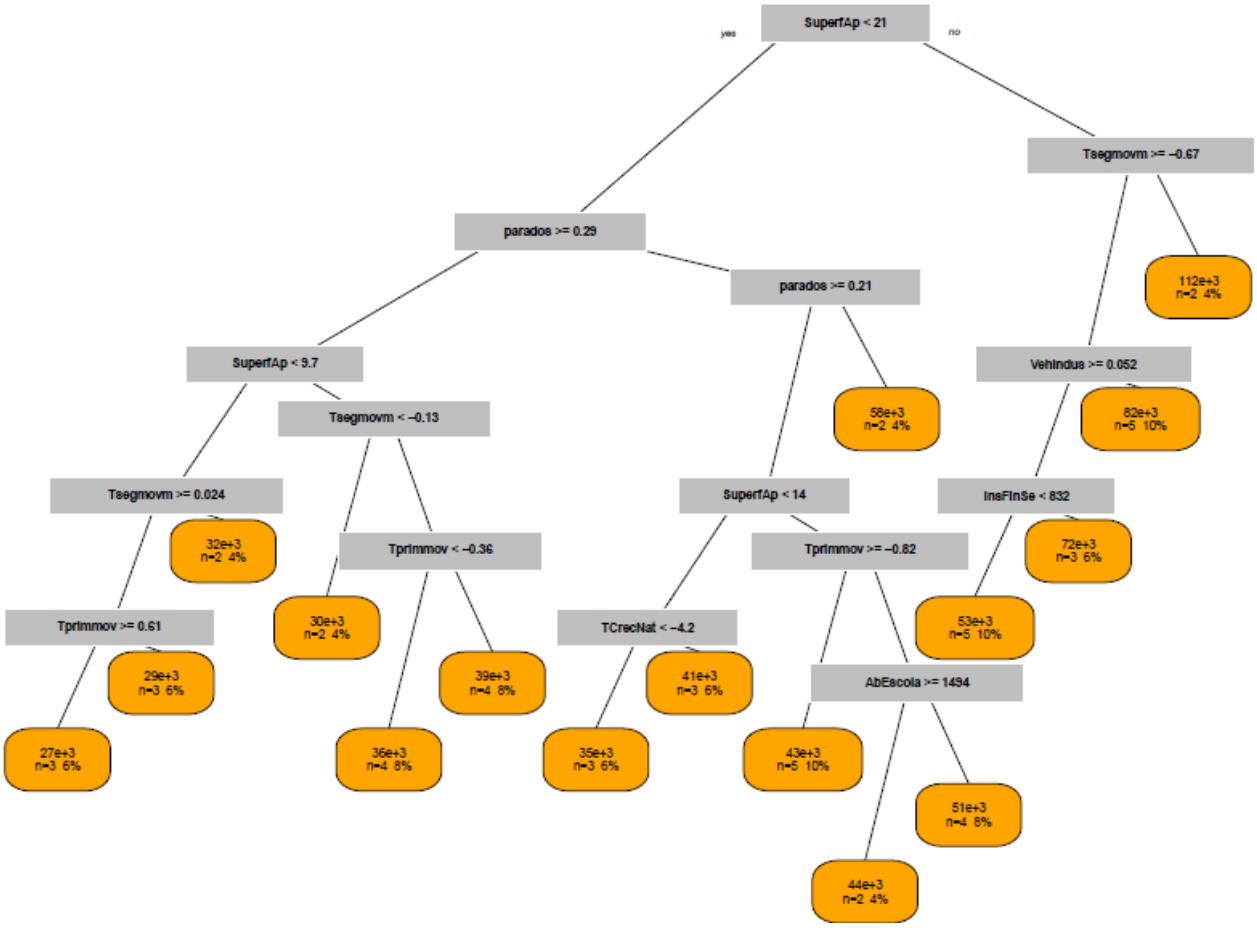
Antes de empezar a aplicar las técnicas, se ha realizado *hold out* donde se han dividido los datos en dos subconjuntos: entrenamiento y validación. Los datos de entrenamiento se utilizarán para construir el modelo y los de validación se usarán para comprobar estimar el error del modelo cuando se aplica a nuevos datos. Para la partición se dividirá aleatoriamente los datos, correspondiendo el 75% del total a entrenamiento y el 25% restante a validación.

Árbol de regresión

La primera técnica que se utilizará será el árbol de regresión. Para llevar a cabo esta técnica, primero se obtendrá el árbol completo, posteriormente se podará y por último se realizarán las predicciones.

Para realizar el árbol completo se utilizará la función *rpart* indicando un valor muy bajo en el estadístico CP (0) y sin indicar el número mínimo de observaciones en el nodo terminal para no limitar el árbol.

Figura 53. Árbol de regresión completo para valor catastral



Tras obtener el árbol completo es necesario podarlo, con el objetivo de evitar sobreajuste y mejorar la interpretabilidad. Para realizar la poda del árbol se utilizará la regla X-SE, a través de la cual se seleccionará el árbol con un menor error estimado.

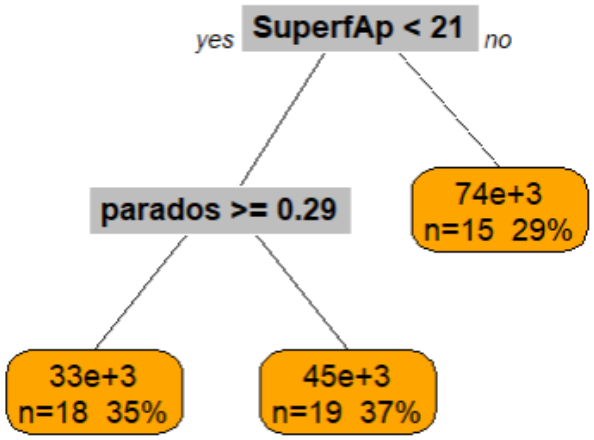
Tabla 5. Poda del árbol valor catastral

	CP	nsplit	Rel.error	xerror	xstd
1	0,53	0	1,00	1,04	0,38
2	0,05	1	0,47	0,66	0,27
3	0,01	2	0,42	0,56	0,23

Si se observa la tabla, el árbol que presenta un menor error estimado es el tercer árbol, que tiene un error de 0,56 y el estadístico CP asociado a este error tiene un valor de 0,01.

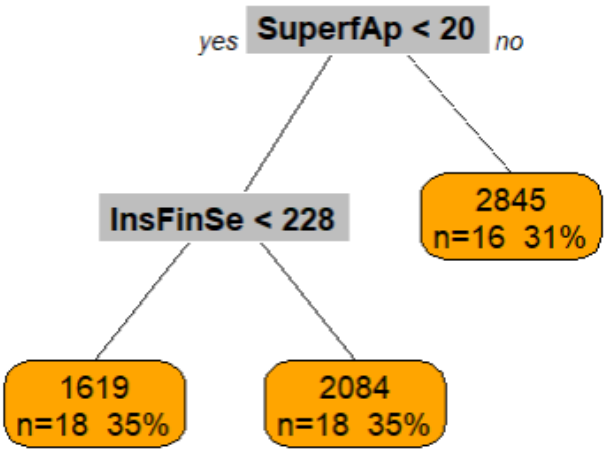
Tras obtener el valor del estadístico CP óptimo, tanto para predecir el valor catastral como para predecir el precio de venta del m^2, se procede a podar los dos árboles a través de la función *prune*, y se obtienen los siguientes árboles.

Figura 54. Árbol de regresión para valor catastral



Se realizan los mismos pasos, pero estableciendo ahora el precio de venta del m^2 como variable respuesta y se obtiene el siguiente árbol podado:

Figura 55. Árbol de regresión para precio de venta del m²



Por un lado, como se puede observar en el árbol de regresión para el valor catastral, las variables más importantes para determinar el valor catastral son la superficie de aparcamiento por turismo y el porcentaje de parados. Si interpretamos el árbol obtenido, se aprecia que se asigna un mayor valor catastral a los barrios que disponen de más de 21 m^2 de superficie de aparcamiento por turismo. En cambio, se establece el menor valor catastral a los barrios que no disponen de más 21 m^2 de aparcamiento por turismo y además tienen un porcentaje de paro superior al 29%.

Por otro lado, en el árbol de regresión para el precio de venta del m^2, las variables más importantes son la superficie de aparcamiento por turismo y el número de actividades económicas dedicadas a instituciones financieras, seguros, servicios prestados a empresas y alquileres. Si interpretamos el árbol obtenido, se aprecia que se asigna un mayor precio del m a los barrios que disponen de más de $20m^2$ de superficie de aparcamiento por turismo. En cambio, se establece el menor precio del m^2 a los barrios que no disponen de más 20 m^2 de aparcamiento por turismo y además tienen menos de 228 actividades económicas dedicadas a finanzas, seguros, servicios a empresas y alquileres.

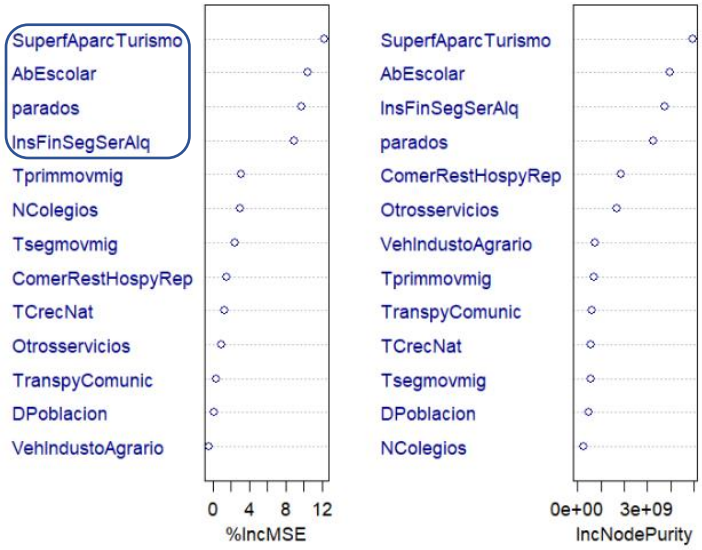
Por último, cabe resaltar que, en ambas variables a predecir, la variable más importante ha sido superficie de aparcamiento por turismo. Sin embargo, a la hora de determinar el valor catastral, la otra variable que ha tenido gran importancia ha sido el porcentaje de parados, en cambio para predecir el precio de venta del m^2 ha sido la variable instituciones financieras, seguros, servicios a empresas y alquileres.

Random Forest

A continuación, se aplicará la técnica Random forest para predecir ambas variables, para ello se utilizará la función *randomforest,* en la cual es necesario indicar el parámetro *mtry* (número de variables muestreadas aleatoriamente como candidatas en cada división), para ello se realiza la división del número de variables (14) entre tres, por lo que se asignara el valor aproximado de 4 al parámetro *mtry*.

En primer lugar, tomando el valor catastral como variable respuesta:

Figura 56. Gráficos %incMSE y IncNodePurity del Valor catastral

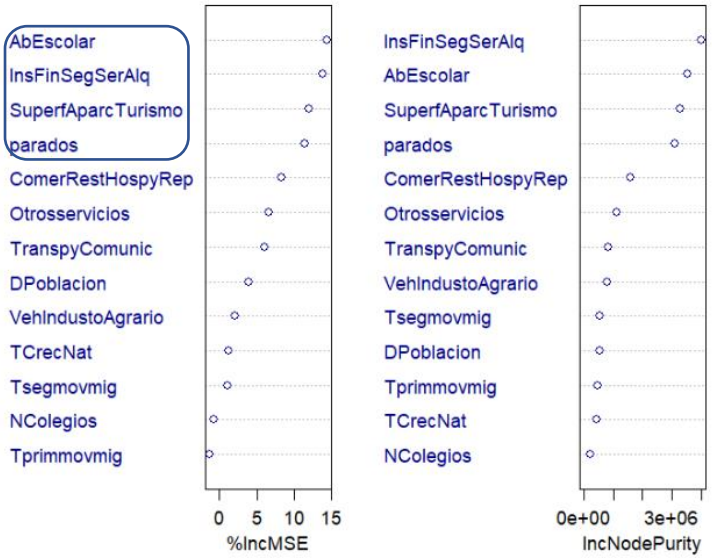


Tras aplicar la técnica Random forest para predecir el valor catastral, se observa la importancia de las variables. En la gráfica de la izquierda *(%incMSE*), se aprecia que las variables que más ayudan a mejorar la precisión del modelo son: Superficie de aparcamiento por turismo, abandono escolar, porcentaje de parados e instituciones financieras, seguros, servicios a empresas y alquileres.

Además, si se observa el gráfico de la derecha, las variables más útiles que logran incrementos en la pureza de los nodos coinciden con las obtenidas en el gráfico *%incMSE.*

En segundo lugar, tras aplicar la técnica Random forest para predecir el precio de venta del m^2, se puede observar en la gráfica de la izquierda (*%incMSE*), que las variables que más ayudan a mejorar la precisión del modelo son: Abandono escolar, instituciones financieras, seguros, servicios a empresas y alquileres, superficie de aparcamiento por turismo y porcentaje de parados.

Figura 57. Gráficos %incMSE y IncNodePurity del Precio de venta del m^2



Vecino más próximo

En este apartado, se utilizará la técnica de vecino más próximo, se trata de un método de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer predicciones sobre la agrupación de un punto de datos individual. A diferencia de las técnicas utilizadas anteriormente, la técnica de vecino más próximo no ofrece la misma interpretabilidad, debido a que no se pueden observar las variables que más influyen a la hora de construir el modelo.

A continuación se aplicará la técnica de vecino más próximo a través de la función *train.kknn* , para predecir el valor catastral y el precio de venta del m^2.

Figura 58. Gráfico de las predicciones del vecino más próximo para Valor catastral

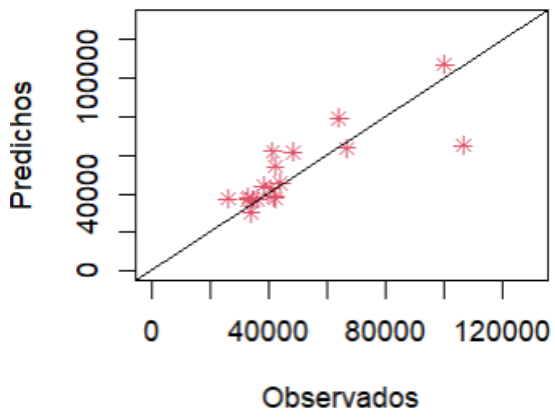
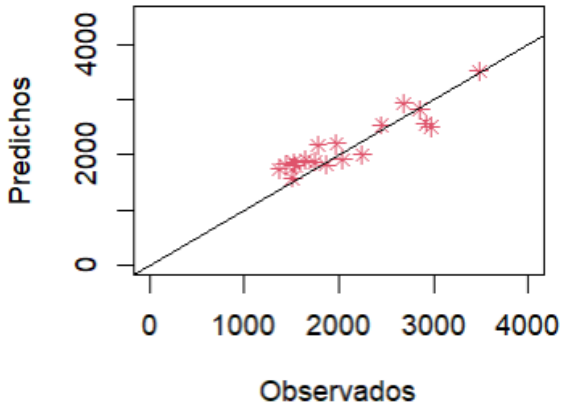


Figura 59. Gráfico de las predicciones del vecino más próximo para Precio de venta el m²



Como se puede observar en ambos gráficos, las predicciones para ambas variables se encuentran bastante cerca de la diagonal, lo que indica que la técnica de vecino más próximo parece predecir bastante bien ambas variables.

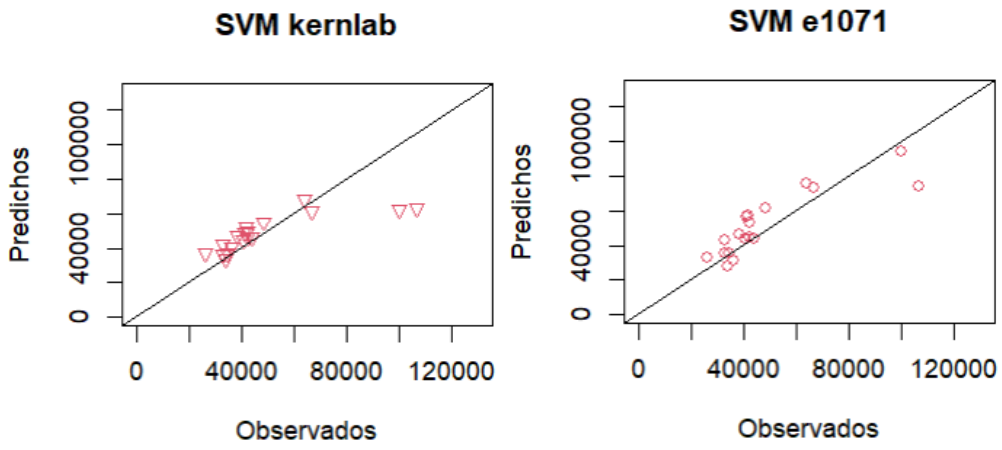
Máquinas de soporte vectorial

A continuación, se aplicará la técnica de máquinas de soporte vectorial a través de dos funciones: Función *ksvm* de la librería *kernlab* y Función *svm* de la librería *e0171*.

Al igual que la técnica de vecino más próximo, la técnica de máquinas de soporte vectorial no ofrece una gran interpretabilidad, debido a que no se pueden observar las variables que más influyen a la hora de construir el modelo

Los resultados de la predicción obtenidos para predecir el valor catastral con ambas funciones son:

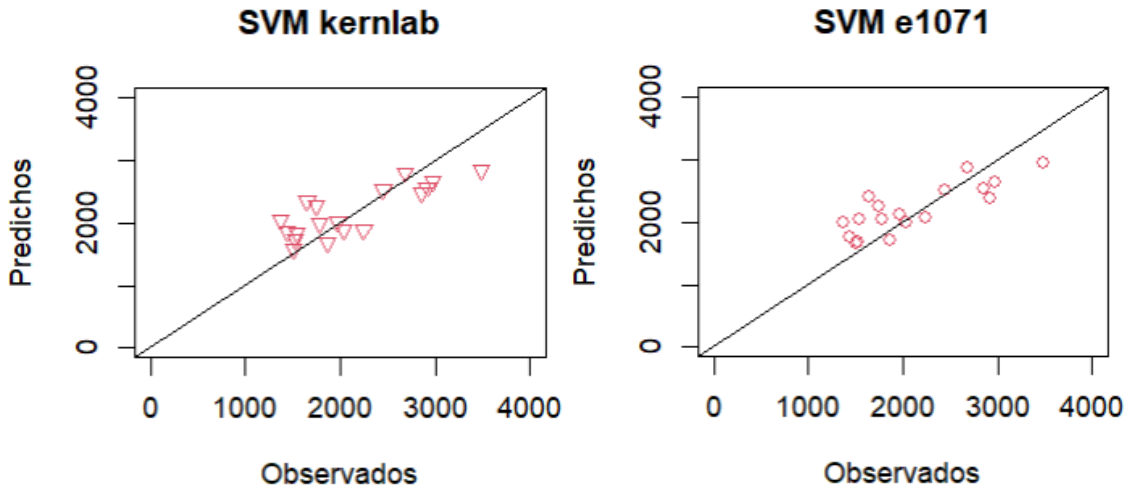
Figura 60. Gráfico de las predicciones con máquinas de soporte vectorial Kernlab y e1071 para Valor catastral



Tras observar los resultados obtenidos con ambas librerias, no se diferencia correctamente con que librería se predice mejor, por lo que se seleccionaran ambos modelo de máquinas de soporte vectorial.

Los resultados de la predicción obtenidos para predecir el precio de venta del m^2 con ambas funciones son:

Figura 61. Gráfico de las predicciones con máquinas de soporte vectorial Kernlab y e1071 para Precio de venta del m^2



Al igual que en los resultados anteriores, no se aprecia correctamente que modelo predice mejor el precio de venta del m^2, por lo que se seleccionan ambos modelos.

Gradient Boosting

A continuación, se utilizará la técnica *gradient boosting*, que al igual que el Random forest es un *esemble model*, es decir, se construye por la unión de otros muchos modelos.

A diferencia del Random forest que construye árboles independientes, la técnica de *gradient boosting* construye un conjunto de árboles sucesivos que son poco profundos y débiles, y con cada árbol va aprendiendo y mejorando del anterior.

Para determinar los parámetros, se utilizará una función, que nos permite obtener los parámetros óptimos para conseguir el mejor árbol posible. Tras aplicar la función se obtiene que los parámetros óptimos para determinar el Valor catastral son:

Tabla 6. Valores óptimos para Gradient boosting

	shrinkage	*interaction.depth*	*n.minobsinnode*	*bag.fraction*	*optimal_trees*	*min_RMSE*	*min_cor*
1	0,10	3	5	1,00	82	7275	0,97
2	0,30	3	5	1,00	44	7283	0,98
3	0,10	3	10	0,80	179	7534	0,97
4	0,10	5	10	0,80	179	7534	0,97
5	0,30	1	5	0,65	29	7618	0,94

En la tabla se muestran los valores óptimos de los parámetros ordenados de menor a mayor RMSE, por lo tanto, en la primera fila se encuentran los parámetros con los que se obtiene el árbol con menor error.

Al igual que el Random forest, la técnica *gradient boosting* también permite observar cuáles han sido las variables más importantes del modelo.

Figura 62. Importancia de las variables para predecir el valor catastral

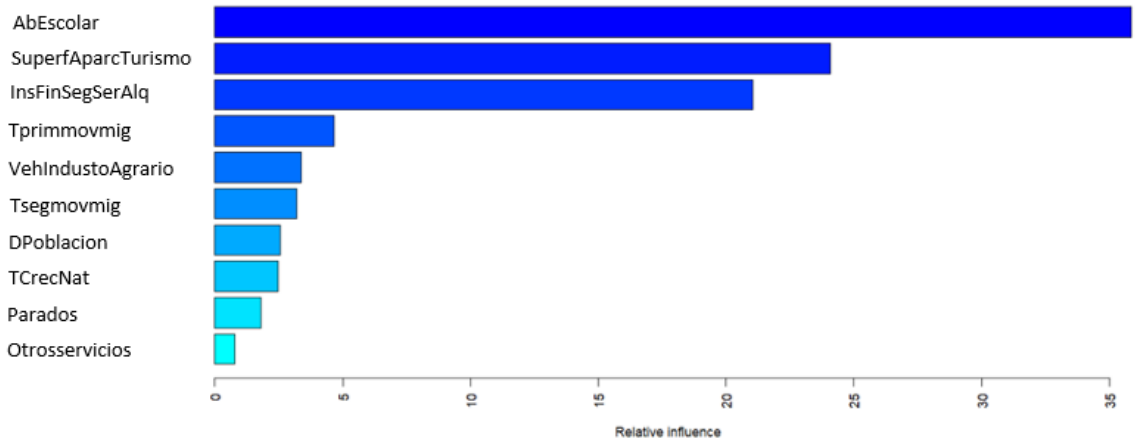


Tabla 7. Importancia de las variables para predecir el valor catastral

	Variable	Relative influence
1	AbEscolar	35,84
2	SuperfAparcTurismo	24,07
3	InsFinSegSerAlq	21,03
4	Tprimmovmig	4,65
5	VehIndustoAgrario	3,36
6	Tsegmovmig	3,19
7	DPoblacion	2,55
8	TCrecNat	2,47
9	Parados	1,80
10	Otrosservicios	0,79

Como se puede observar en el gráfico, las variables que más influido en el modelo han sido: Abandono escolar, Superficie de aparcamiento por turismo, Instituciones financieras, seguros y alquileres, y Movimiento migratorio.

A continuación, se realizará para predecir el precio de venta del m^2:

Tabla 8. Valores óptimos para Gradient boosting

	shrinkage	*interaction.depth*	*n.minobsinnode*	*bag.fraction*	*optimal_trees*	*min_RMSE*	*min_cor*
1	0,30	5	5	0,8	9	292	0,95
2	0,30	3	5	1,0	9	316	0,94
3	0,30	1	5	1,0	28	321	0,93
4	0,30	5	10	1,0	9	327	0,92
5	0,30	3	10	1,0	7	331	0,90

Como se ha comentado anteriormente, la tabla muestra los valores óptimos de los parámetros, para obtener el árbol con menor error.

Tras realizar la técnica de *gradient boosting* con los parámetros óptimos, se observan las variables que más han influido en el modelo.

Figura 63. Importancia de las variables para predecir el precio de venta del m²

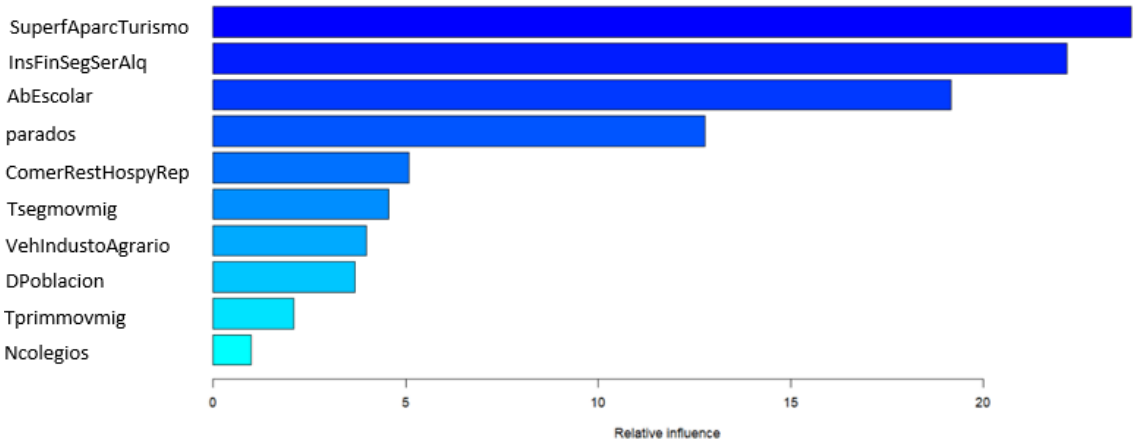


Tabla 9. Importancia de las variables para predecir el precio de venta del m²

	Variable	Relative influence
1	SuperfAparcTurismo	23,83
2	InsFinSegSerAlq	22,17
3	AbEscolar	19,16
4	Parados	12,76
5	ComerRestHospyRep	5,07
6	Tsegmovmig	4,55
7	VehIndustoAgrario	3,98
8	DPoblacion	3,68
9	Tprimmovmig	2,08
10	Ncolegios	0,98

Como se observa en el gráfico, las variables más importantes del modelo son: Superficie de aparcamiento por turismo, Instituciones financieras, seguros, servicios a empresa y alquileres, Abandono escolar y Porcentaje de parados.

4.3.1 Comparación de los resultados obtenidos

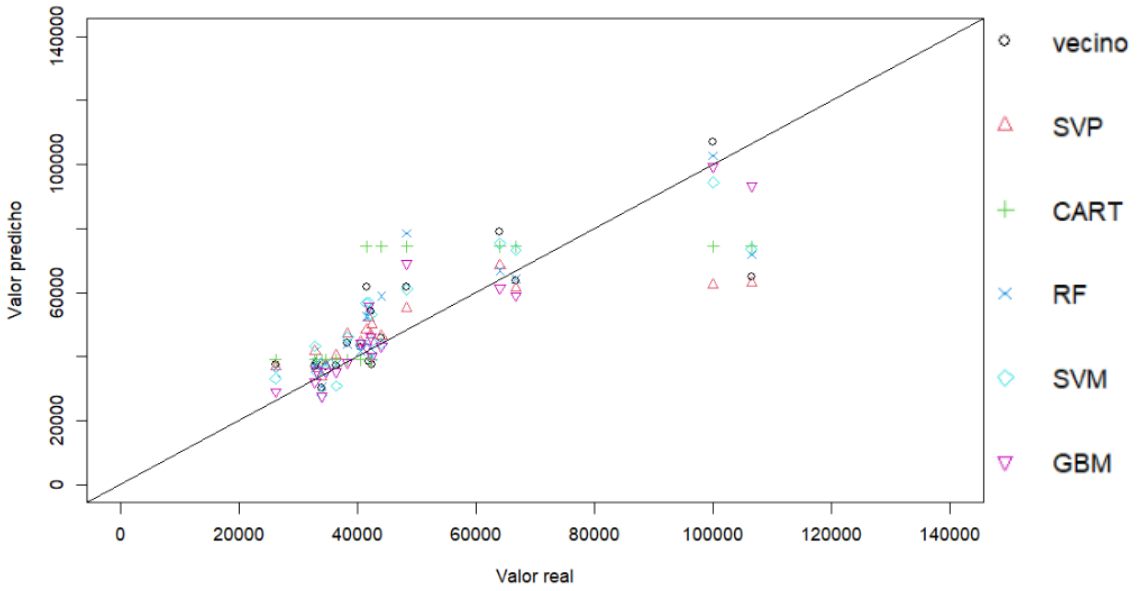
En este apartado se procederá a comparar los resultados obtenidos a través de las diferentes técnicas y seleccionar el modelo que presente mejores resultados.

Comparación gráfica

En primer lugar, se compararán gráficamente las predicciones obtenidas con las diferentes técnicas.

Las predicciones obtenidas para el Valor catastral:

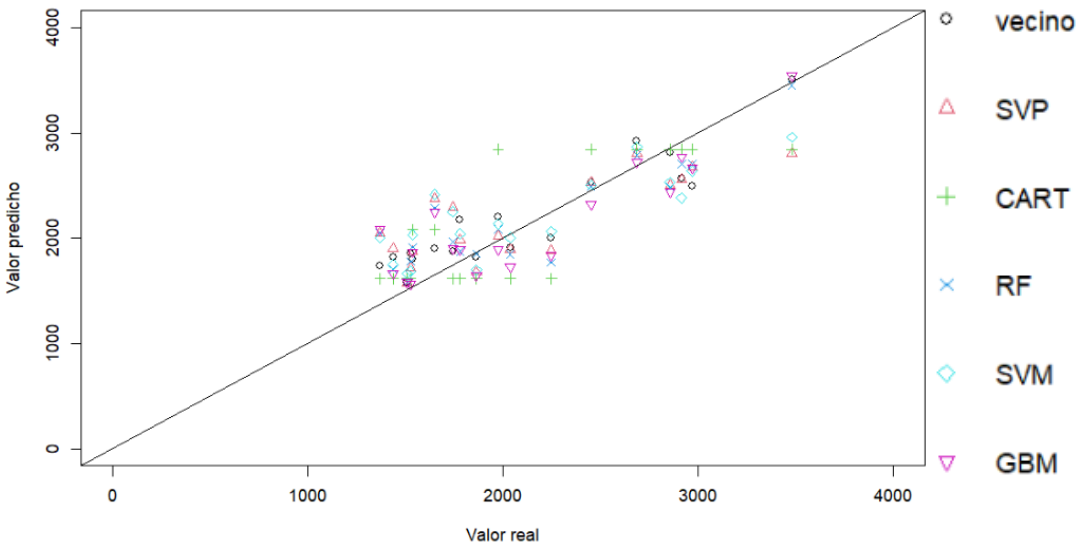
Figura 64. Comparación gráfica de las predicciones de valor catastral



Tras respresentar las predicciones obtenidas, no se aprecia correctamente que técnica ha realizado las mejores predicciones, aunque si que se puede observar que las técnicas que han obtenido las predicciones más cercanas a la diagonal han sido Random Forest, Máquinas de soporte vectorial y Grandient boosting.

Las predicciones obtenidas para el precio de venta del m^2 con las diferentes técnicas:

Figura 65. Comparación gráfica de las predicciones del precio de venta del m^2



En cuanto a las predicciones del precio de venta del m^2, no se identifica con facilidad cual ha sido el metodo que ha realizado mejores predicciones, pero si que se puede apreciar que los metodos que han menor error en sus predicciones han sido: Vecino más próximo, Random Forest y Gradient Boosting.

Comparación numérica

Por último, se procederá a calcular las medidas de bondad RMSE y MAPE, y se seleccionará el modelo que presente mejores resultados.

En cuanto al valor catastral:

Tabla 10. Medidas de bondad de ajuste para valor catastral

	RMSE	*MAPE*
Árbol de regresión	16426,65	24,07%
Vecino más próximo	12913,35	17,45%
SVP "kernlab"	14840,97	17,33%
Random forest	12413,31	16,60%
SVM "e1071"	11380,67	17,97%
GBM	7275,07	9,91%

Tras obtener el *RMSE* y el *MAPE* de las distintas técnicas aplicadas, se aprecia que el modelo que mejor predice el valor catastral es el *Gradient Boosting*, ya que es el que presenta un menor *RMSE* y *MAPE*. Por lo tanto, centrándonos en el resultado del *MAPE*, se podría concluir que la predicción realizada a través del *Gradient Boosting* esta errada en un 9,91%, lo cual es un valor bastante bajo, por lo que esta técnica ha conseguido predecir correctamente el valor catastral medio de los barrios.

Por lo que respecta al precio de venta del m^2:

Tabla 11. Medidas de bondad de ajuste para el precio de venta del m^2

	RMSE	*MAPE*
Árbol de regresión	382,40	15,27%
Vecino más próximo	263,35	11,97%
SVM "kernlab"	387,92	16,91%
Random forest	311,56	13,49%
SVM "e1071"	380,35	16,82%
GBM	306,64	13,06%

También se ha obtenido el *RMSE* y el *MAPE* de todos los modelos aplicados. Como se puede observar en la tabla, la técnica que presenta mejores resultados es el vecino más próximo, con un *RMSE* de 263,35 y un *MAPE* de 11,97%, lo que indica que la predicción realizada esta errada en un 11,97%, lo cual es un error bastante pequeño.

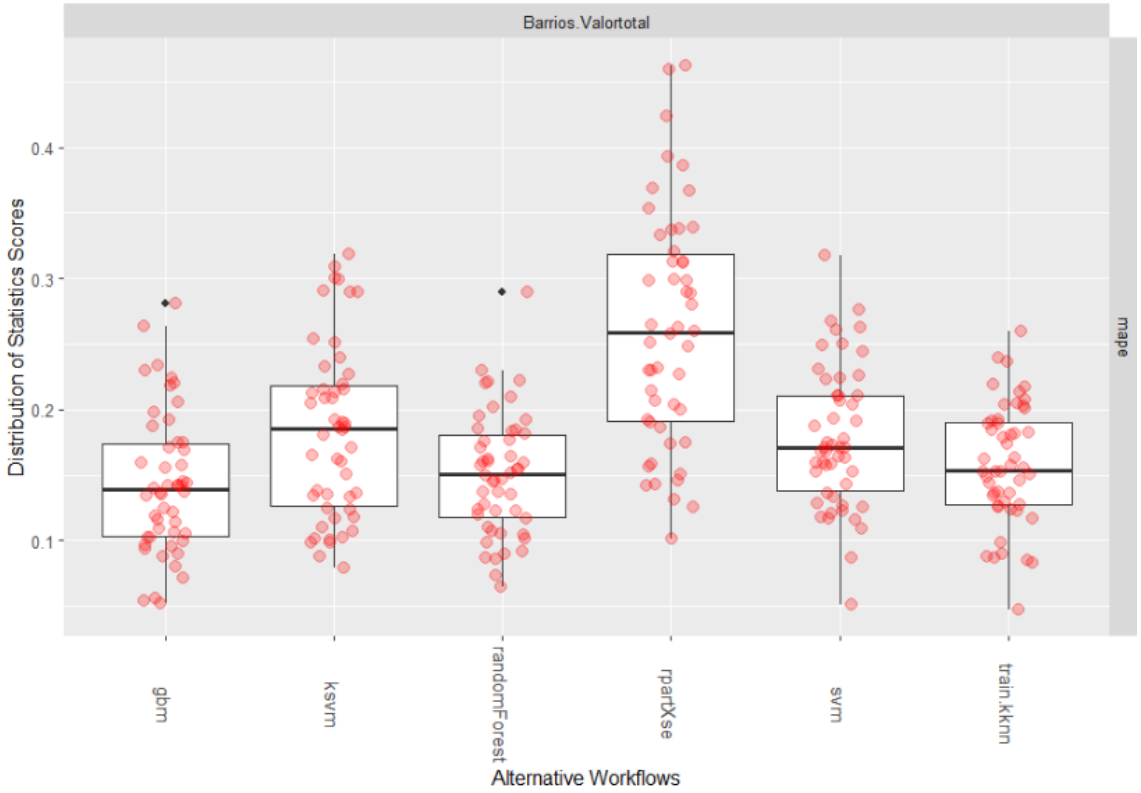
4.3.2 Validación cruzada

Como método de evaluación, se ha realizado una validación cruzada de los modelos predictivos, con el objetivo de que los resultados obtenidos sean independientes de la partición realizada. Por lo tanto, la validación cruzada nos permitirá comprobar que método es el que mejor predice sin importar la partición que se realiza, lo que evita el sobreajuste de los modelos.

A continuación, se aplicará la validación cruzada con 10 subconjuntos y 5 repeticiones, como medida de bondad de ajuste se utilizará el *MAPE*.

Los resultados obtenidos con validación cruzada para predecir el valor catastral son:

Figura 66. Gráfico de validación cruzada para valor catastral



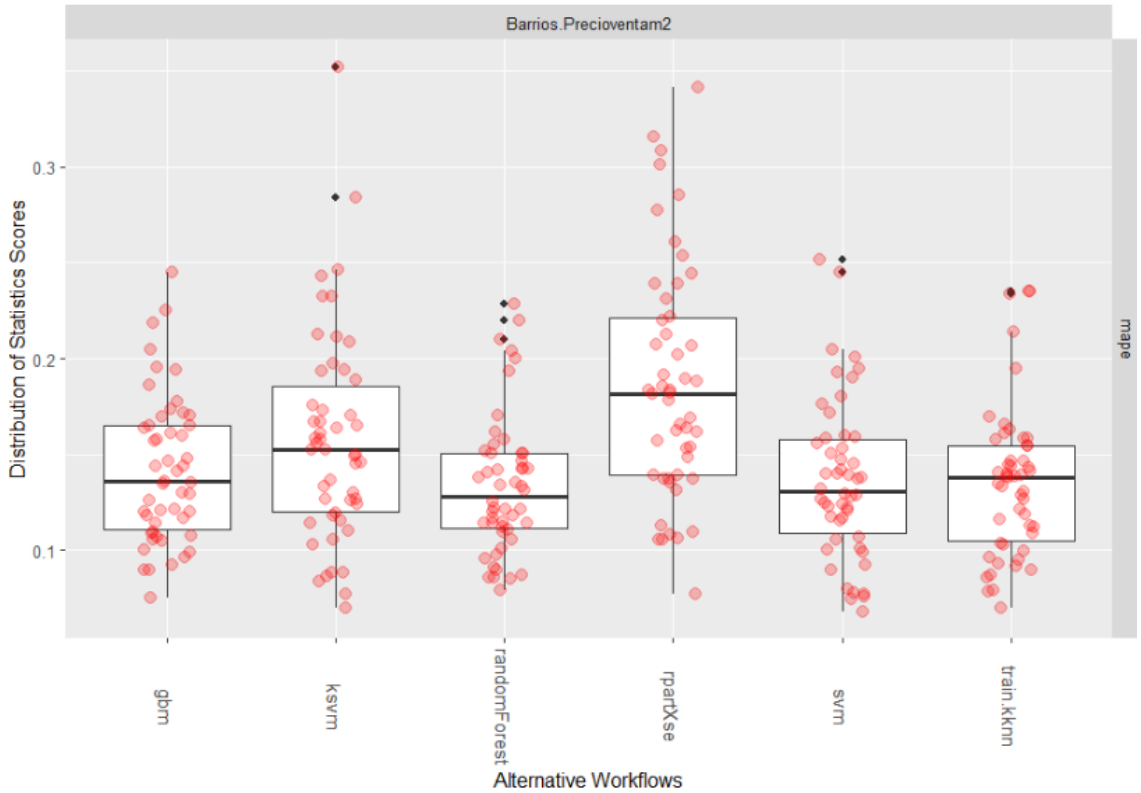
Gbm=Grandient Boosting Method; Ksvm=Máquinas de soporte vectorial "Kernlab"; Random Forest=Bosque aleatorio; rpartXse=árbol de regresión; svm= Máquinas de soporte vectorial "e1071"; train.kknn= vecino más próximo

Tras aplicar validación cruzada con 10 subconjuntos y 5 repeticiones en todas las técnicas utilizadas, se muestra en el gráfico, el *MAPE* obtenido en todos los subconjuntos. Se aprecia que las técnicas que presentan mejores resultados son *Grandient Boosting* con un *MAPE* promedio de 13,86% y Random Forest con un *MAPE* promedio de 15,1%, ambas con una variabilidad similar (rango), por lo que el modelo que mejor predice el valor catastral medio es el *Gradient Boosting*.

Como se ha comentado anteriormente, las variables que más han influido en la técnica de *Gradient Boosting* para predecir el valor catastral medio, han sido: Abandono escolar, Superficie de aparcamiento por turismo, Instituciones financieras, seguros y alquileres, y Movimiento migratorio.

Los resultados obtenidos con validación cruzada para predecir el precio de venta del m^2 son:

Figura 67. Gráfico de validación cruzada para precio de venta del m²



Gbm=Grandient Boosting Method; Ksvm=Máquinas de soporte vectorial "Kernlab"; Random Forest=Bosque aleatorio; rpartXse=árbol de regresión; svm= Máquinas de soporte vectorial "e1071"; train.kknn= vecino más próximo

Para evaluar las técnicas aplicadas para predecir el precio de venta del m^2, también se ha utilizada la validación cruzada con 10 subconjuntos y 5 repeticiones. Como se observa en el gráfico, las técnicas que presentan mejores resultados son *Random Forest* con un *MAPE* promedio de 13,3% y máquinas de soporte vectorial (*e1071*) con un *MAPE* promedio de 13,64%. Por lo tanto, el método que mejor predice el precio de venta del m^2 es el *Random Forest*.

Como se ha visto anteriormente, las variables que más han mejorado la precisión del modelo de *Random Forest* para predecir el precio de venta del m^2, han sido: Abandono escolar, instituciones financieras, seguros, servicios a empresas y alquileres, superficie de aparcamiento por turismo y porcentaje de parados.

4.3.3 Regresión de mínimos cuadrados parciales con ambas variables respuesta

Por último, debido a la correlación entre las variables respuesta, se ha realizado un modelo *PLS* con ambas: Valor catastral y Precio de venta del m^2. Para ello, se utilizará la partición realizada anteriormente, donde los datos de entrenamiento servirán para crear el modelo *PLS* y los datos de validación para comprobar que tan bien predice el modelo ajustado. Para llevar a cabo la técnica PLS se utilizará el software *Aspen Pro MV.*

En primer lugar, tras ajustar el modelo PLS se procederá a comprobar a través de los gráficos T^2 *de Hotelling* y *SPE*, si existen valores anómalos o atípicos.

Figura 68. Gráfico T^2 de Hotelling

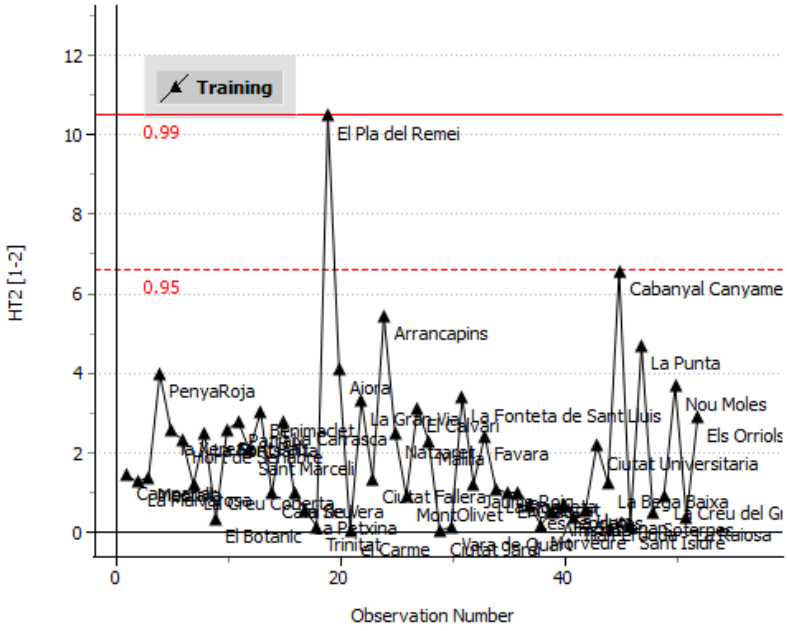
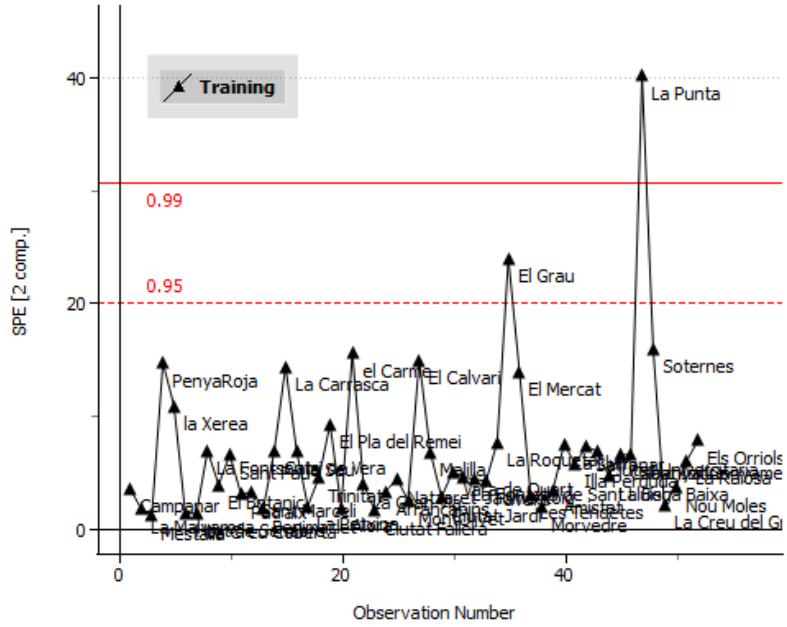


Figura 69. Gráfico SPE

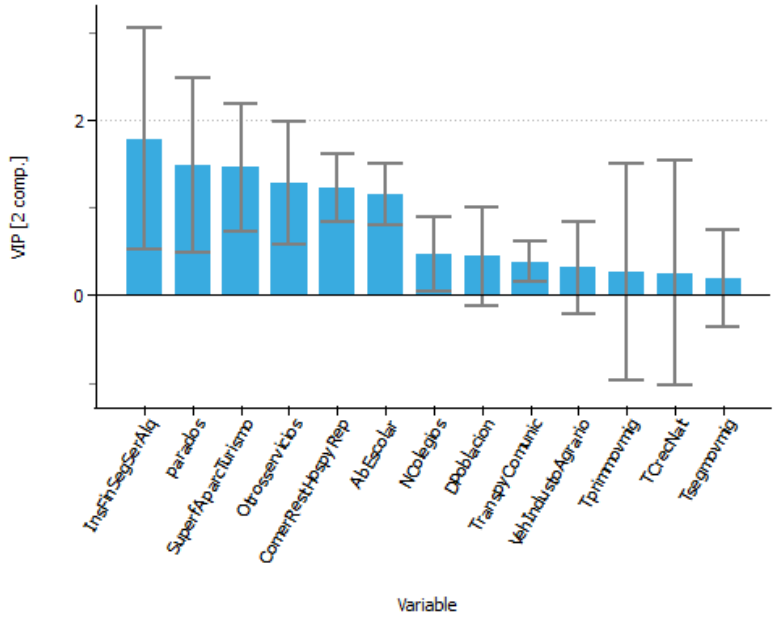


Por un lado, observando el gráfico de *T^2 de Hotelling* no se aprecian barrios con un alto valor residual. Debido a que el límite de control del 99% este situado en torno a 10 y ninguna observación supera dicho limite. Por lo tanto, no se han detectado observaciones extremas.

Por otro lado, observando el gráfico *SPE* no se aprecia ningún barrio con una suma de cuadrados residual alta. Como se puede observar el límite de control del 99% este situado en torno a 31 y ningún barrio supera dos veces este límite de control. Por lo tanto, la observación que supera dicho limite (La Punta) es esperables, ya que se está trabajando con una base de datos de 52 barrios y teniendo en cuenta la tasa de falsas alarmas (1%), se espera que 1 barrio tenga un residuo superior. Por lo tanto, no se detectan valores atípicos.

En segundo lugar, se realizará el gráfico *VIP* para comprobar que regresores tienen mayor importancia en el modelo PLS ajustado. Se considerarán regresores importantes los que tengan un valor superior a 1. Este paso nos permitirá depurar el modelo.

Figura 70. Gráfico VIP

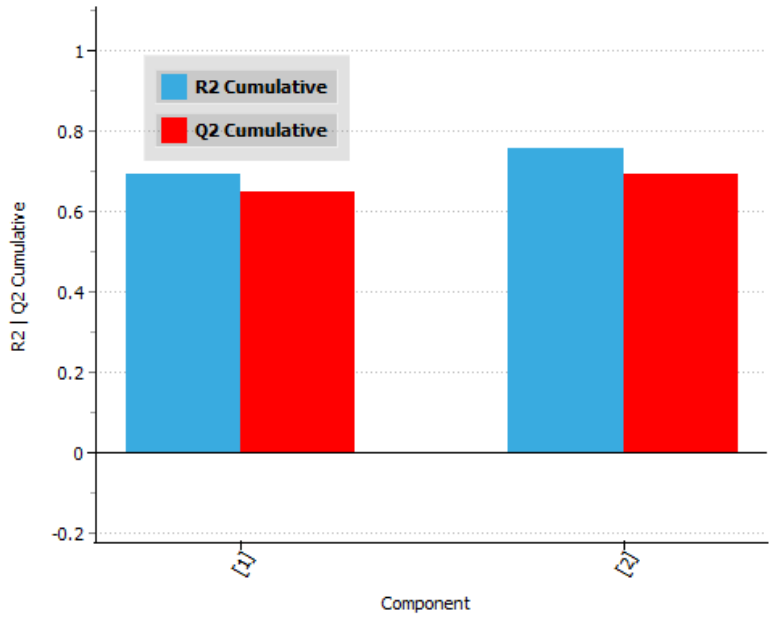


Como se observa en el gráfico *VIP*, los regresores más importantes son los relacionados con las actividades económicas (Instituciones financieras, seguros, servicios a empresas y alquileres, Comercio, hostelería y reparaciones y Otros servicios), Porcentaje de parados, Superficie de aparcamiento por turismo y Abandono escolar.

También, se puede apreciar que las variables como Número de colegios, Densidad de población, Transporte y comunicaciones, Porcentaje de vehículo industrial o agrario, Movimiento migratorio, Tasa de crecimiento natural y Tipo de movimiento migratoria no son estadísticamente significativos. Por lo tanto, se ha decidido eliminarlas.

El modelo *PLS* depurado está formado por 6 regresores, por lo tanto, el número máximo de componente *PLS* que se pueden sacar son 6. El modelo PLS ajustado, está formado por 2 componentes *PLS* que explican el 76,09% de la variabilidad total.

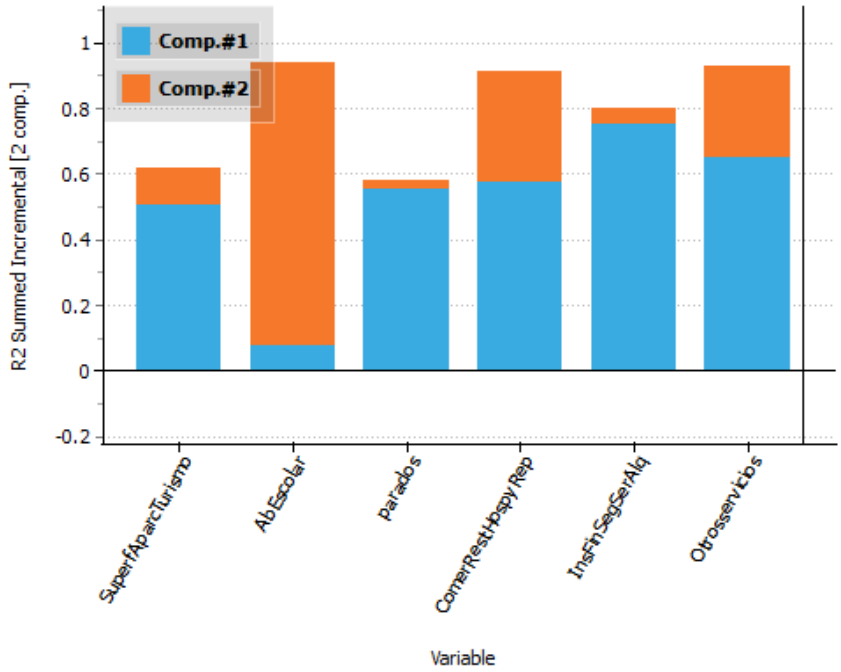
Figura 71. Gráfico de las componentes PLS



Como se observa en el gráfico y en la tabla, la primera componente *PLS* explica la mayoría de la variabilidad, el 69,65%, en cambio la segunda componente *PLS* apenas explica un 6,43%.

A continuación, para comprobar la variabilidad explicada de cada variable por cada componente *PLS*, se obtiene el gráfico *R2* de las variables.

Figura 72. Gráfico de la variabilidad explicada por las componentes PLS



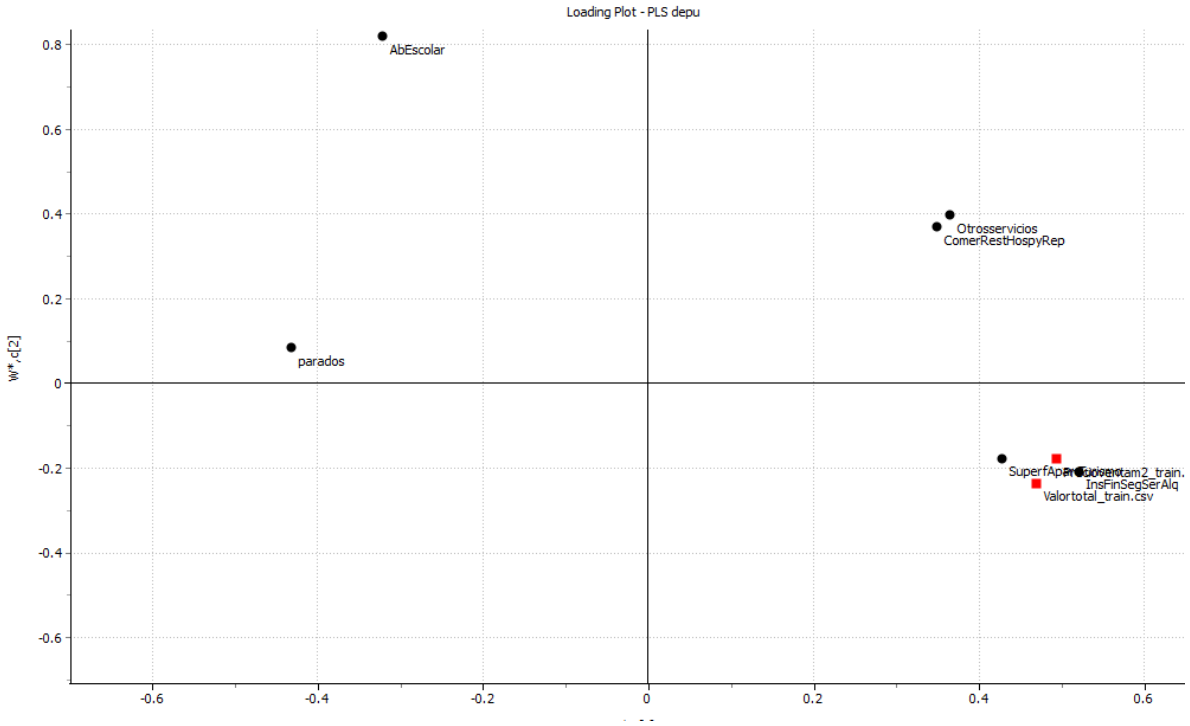
Como se observa en el gráfico *R2*, la primera componente *PLS* está asociada a Superficie de aparcamiento por turismo, Instituciones financieras, seguros, servicios a empresas y alquileres

y Porcentaje de parados, además explica la mayoría de la variabilidad de Comercio, restaurante, hostelería y reparaciones y Otros servicios.

Sin embargo, la segunda componente *PLS* también explica una parte de la variabilidad de Comercio, restaurante, hostelería y reparaciones y Otros servicios, pero esta componente esta fundamentalmente asociada a la variable Abandono escolar.

Para interpretar el modelo *PLS* obtenido, se representará el gráfico de cargas, donde se podrá observar las relaciones entre los regresores y las variables respuesta.

Figura 73. Gráfico de cargas



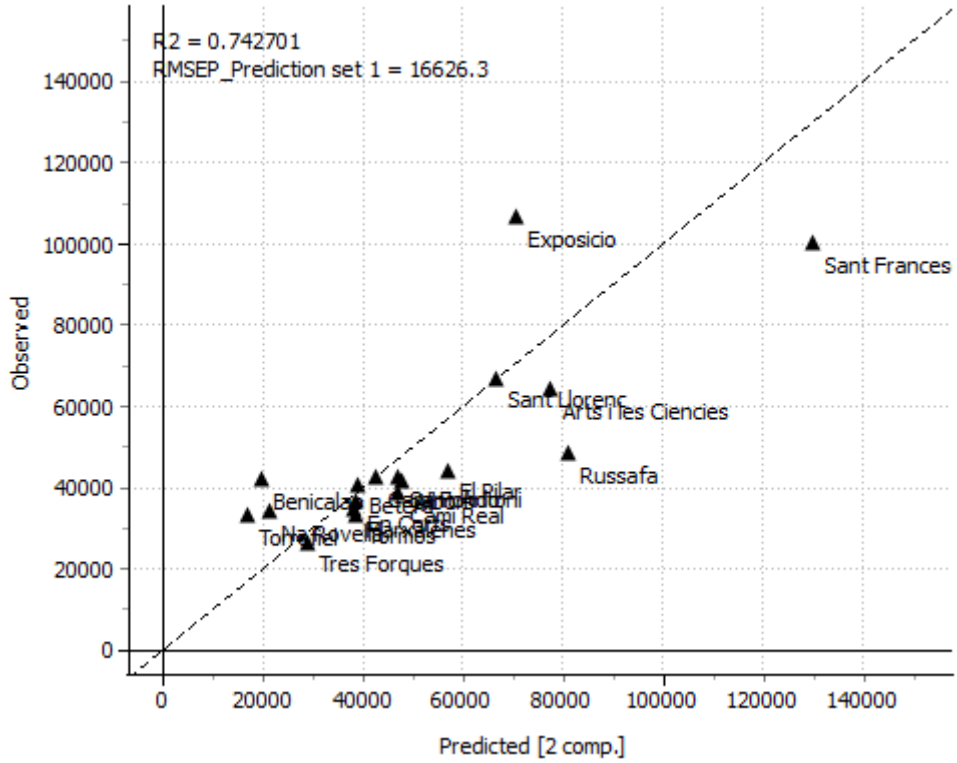
Por un lado, se observa en el gráfico de cargas, que ambas variables respuesta están bastante correlacionadas positivamente con los regresores Instituciones financieras, seguros, servicios a empresas y alquileres y Superficie de aparcamiento por turismo. Por lo tanto, los barrios con mayor valor catastral y precio de venta del m^2 tienden a tener mayor Superficie de aparcamiento por turismo y un mayor número de servicios dedicadas a finanzas, seguros, servicios a empresas y alquileres.

Por otro lado, se observa una correlación negativa entre las variables respuesta y el regresor porcentaje de parados. Por lo tanto, los barrios con un mayor porcentaje de paro, son barrios con un menor valor catastral y precio de venta del m^2.

Tras analizar el modelo *PLS* obtenido a traves de los datos de entrenamiento, se procede a realizar las predicciones. Para ello se utilizaran los datos de validacion, lo que nos permitira evaluar el desempeño del modelo ajustado en nuevos barrios.

Las predicciones del valor catastral que se han obtenido para el conjunto de datos de validación son las siguientes:

Figura 74. Gráfico de las predicciones del valor catastral con PLS



Se observa en el gráfico, que la mayoría de las predicciones obtenidas a través del modelo *PLS* ajustado se encuentran bastante cercanas a la diagonal, a excepción de 3 barrios (Sant Francesc, Russafa y Exposició). Por lo tanto, el modelo *PLS* ha sido capaz de predecir correctamente la mayoría de los barrios, por lo que se considera un buen modelo para predecir el valor catastral.

A continuación, se han obtenido las medidas de bondad de ajuste *RMSE* y *MAPE* de las predicciones con *PLS* y se han comparado con las técnicas anteriores.

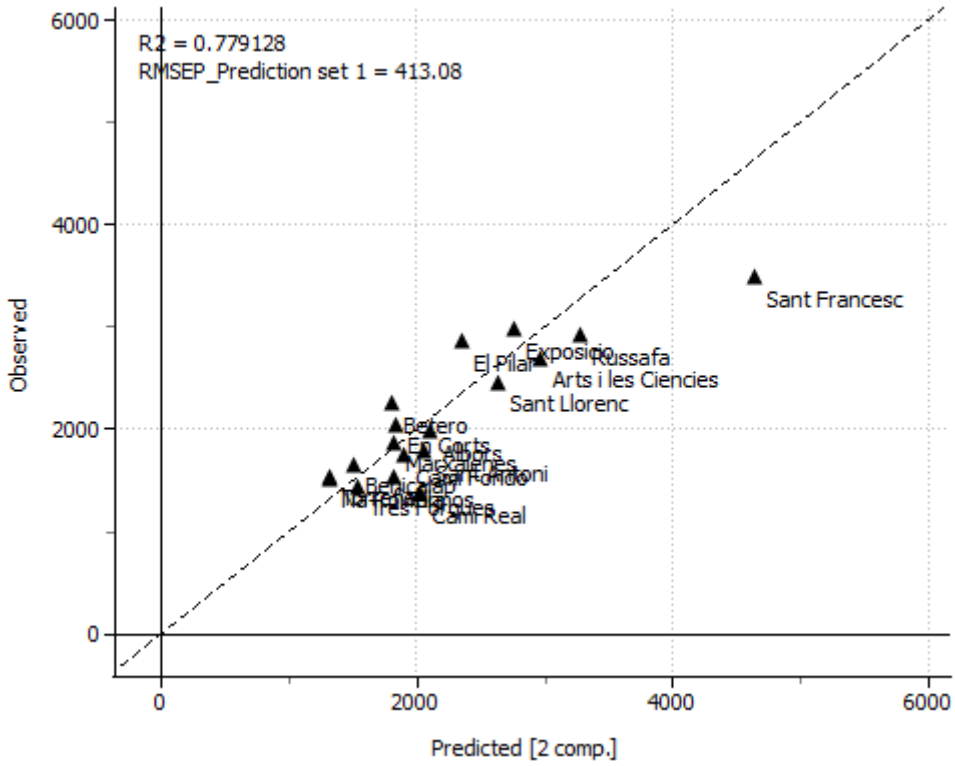
Tabla 12. Medidas de bondad de ajuste del valor catastral con PLS

	RMSE	MAPE
PLS	15915,87	22,07%
Árbol de regresión	16426,65	24,07%
Vecino más próximo	12913,35	17,45%
SVM “kernlab”	14840,97	17,33%
Random forest	12413,31	16,6%
SVM “e1071”	11380,67	17,97%
GBM	7275,07	9,91%

Como se observa en la Tabla 12, las predicciones del *PLS* tienen un *MAPE* del 22,07%, se trata de un error superior a la mayoría de los modelos, por lo que el PLS no ha sido capaz de mejorar las predicciones obtenidas con el resto de las técnicas.

Las predicciones del precio del m^2 que se han obtenido para el conjunto de datos de validación son las siguientes:

Figura 75. Gráfico de las predicciones del precio de venta del m² con PLS



Como se observa en el gráfico, las predicciones del modelo *PLS* ajustado se encuentran bastante cercanas a la diagonal, a excepción del barrio de Sant Francesc, en el cual la predicción determina un precio del m^2 bastante superior al precio real. En general se trata de un buen modelo para predecir.

Por último, se han obtenido las medidas de bondad de ajuste *RMSE* y *MAPE* de las predicciones con la técnica *PLS* y se han comparado los resultados con las técnicas anteriores.

Tabla 13. Medidas de bondad de ajuste del precio de venta del m² con PLS

	RMSE	MAPE
PLS	455,77	15,00%
Árbol de regresión	382,40	15,27%
Vecino más próximo	263,35	11,97%
SVM “kernlab”	387,92	16,91%
Random forest	311,56	13,49%
SVM “e1071”	380,35	16,82%
GBM	306,64	13,06%

Como se observa en la tabla 13, las predicciones obtenidas con el modelo *PLS* tienen un *MAPE* del 15%, se trata de un erro bastante similar al obtenido en el resto de las técnicas, sin embargo, no consigue mejorar las predicciones obtenidas con *Random Forest*.

5. Conclusión

En resumen, este trabajo ha cumplido con los dos objetivos de investigación definidos: estudiar los barrios de valencia en función de sus características sociodemográficas y sus servicios, así como predecir el precio medio de la vivienda a partir de las características propias de cada barrio.

Como se ha visto en la primera parte del trabajo, se disponían de diferentes características heterogéneas de los barrios, entre las cuales se han identificado dos bloques de variables con una elevada correlación. Estos bloques son los movimientos migratorios, que recogen las variables que hacen referencia a la emigración e inmigración, tanto interurbana como intraurbana. Otro bloque lo conforman las variables relacionadas con los servicios disponibles en cada barrio, como son el transporte, la comunicación, el comercio, la hostelería, las instituciones financieras, entre otros.

En primer lugar, al estudiar los movimientos migratorios de los barrios, se han analizado las relaciones entre estas variables a través de un análisis de componentes principales, donde se ha podido observar que la primera componente diferencia entre los barrios con alto y bajo movimiento migratorio. Por otro lado, la segunda componente distingue entre barrios con movimiento migratorio intraurbano o interurbano. Además, se ha visto una elevada correlación entre la emigración e inmigración intraurbana, como entre los movimientos interurbanos. Tras observar estos resultados, se han identificado distintos perfiles de barrios según sus movimientos migratorios. Se ha identificado un primer conglomerado formado por 14 barrios con un perfil con alto movimiento migratorio intraurbano e interurbano, frente a otro conglomerado de 56 barrios con un bajo nivel de movimientos migratorios. Este resultado nos ha dado información de las diferencias migratorias que existen entre los barrios de Valencia. Posteriormente, se ha relacionado estos resultados con el precio de la vivienda, y no se han apreciado diferencias significativas como para afirmar que el nivel de movimiento migratorio afecte al precio medio de las viviendas.

En segundo lugar, se ha estudiado el bloque de los servicios de los barrios, a través de un análisis de componentes principales, donde se ha podido observar una elevada correlación entre las variables: Instituciones financieras, seguros, servicios a empresas y alquileres, Comercio, restaurante, hostelería, reparación y Otros servicios, que estaban explicadas por la primera componente. Por otro lado, la segunda componente estaba asociada al transporte y comunicación. Por lo tanto, se han podido identificar una serie de barrios con una gran cantidad y tipo de servicios muy similar, que son: Sant Francesc, Russafa y el Pla del Remei. Igualmente, se han apreciado dos barrios con un alto número de servicios de transporte y comunicación, que eran Benicalap y el Grau, donde era esperable, debido a que se trata de la zona del puerto. Tras observar los estos resultados, se ha pretendido identificar el perfil de los barrios según los servicios que disponen, donde a través la técnica cluster Diana, se han identificado 2 perfiles de barrios. Un grupo de 6 barrios con un alto número de servicios y un grupo de 64 barrios con un número mucho menor de servicios, lo que nos ha permitido conocer un grupo de barrios con una gran diferencia de número de servicios, que como se ha comentado anteriormente se trata de un factor clave a la hora de tomar la decisión de compra. Por lo tanto, se ha relacionado estos resultados con el precio de la vivienda, y en principio no se han apreciado grandes diferencias como para afirmar que contra mayor número de servicios tenga el barrio, mayor serán los precios de las viviendas.

Posteriormente, se ha realizado un análisis de correspondencias múltiples con todas las características de los barrios, con el objetivo de analizar las relaciones entre las variables. Como resultado se ha identificado un grupo de barrios con mayor precio de la vivienda que se caracterizaban por tener un alto número de servicios, un bajo porcentaje de paro y mucha superficie de aparcamiento por turismo.

A continuación, se ha procedido a cumplir el segundo objetivo planteado, donde en primer lugar se han desarrollado diferentes técnicas de minería de datos para predecir el valor catastral y el precio medio del m^2 . Por un lado, a la hora de predecir el valor catastral se ha comprobado que la técnica que mejor predice este valor era el Gradient Boosting, con un error porcentual absoluto del 9,91% que es aceptable, siendo las variables que más han influido en la predicción: Abandono escolar, Superficie de aparcamiento por turismo e Instituciones financieras, seguros, servicios a empresas y alquileres. Por otro lado, la técnica que ha obtenido mejores resultados para predecir el precio de venta del m^2 es el vecino más próximo con un error porcentual absoluto del 11,97%.

En segundo lugar, con el objetivo de que los resultados obtenidos sean independientes de la partición realizada, se ha aplicado validación cruzada con 10 subconjuntos y 5 repeticiones. Por un lado, tras observar los resultados con validación cruzada para predecir el valor catastral, se ha observado que la técnica *Gradient Boosting* es la que presenta mejores resultados, con un error porcentual absoluto promedio del 13,86%. En cambio, para predecir el precio de venta del m^2 la técnica que presenta mejores resultados es el *Random Forest* con un error porcentual absoluto promedio de 13,3%.

Por lo tanto, se ha comprobado que solo con las características de los barrios (localización), se han obtenido buenos resultados de predicción, lo que resalta la gran influencia que tiene la localización en el precio de la vivienda.

Por último, se ha intentado mejorar los resultados obtenidos a través de una nueva técnica, la regresión por mínimos cuadrados parciales utilizando ambas medidas de precio como variables respuestas. Tras ajustar el modelo y realizar las predicciones no se ha conseguido mejorar los resultados de predicción obtenidos anteriormente.

En conclusión, se ha comprobado a través del trabajo la importancia que tiene la localización en el valor de la vivienda, así como que características de los barrios son las que más afectan a los precios de la vivienda en la ciudad de Valencia.

La mayoría de los trabajos publicados sobre la predicción del precio de las viviendas se basan en técnicas econométricas, por lo que este trabajo es innovador debido a que utiliza técnicas de aprendizaje automático, a partir de las cuales se ha podido descubrir información valiosa tanto a nivel de investigación como para los inversores. Este trabajo aporta valor a nivel práctico, ya que los inversores pueden conocer que factores son los que afectan en mayor media al precio, con el objetivo de realizar una buena inversión y obtener una buena rentabilidad.

Para finalizar, el trabajo ha presentado buenos resultados, pero deben considerarse ciertas limitaciones en el alcance del trabajo. En primer lugar, faltan variables importantes sobre los barrios, como son: el número de viviendas, la superficie media de las viviendas o antigüedad media de las viviendas. Asimismo, hubiera aportado una información valiosa el análisis de las características individuales de cada vivienda, como el tamaño o las características.

Como futura línea de trabajo, sería interesante no centrarse solo en la ciudad de Valencia, sino añadir información de otros municipios, debido a que en cada ciudad las características de la localización que más influyen en el precio pueden variar, por lo que disponer de esta información podría mejorar el conocimiento sobre el mercado inmobiliario. Además, también sería interesante no estudiar el precio de forma estática, sino que estudiarlo de forma dinámica, con el objetivo de observar la evolución temporal y analizar como los efectos macroeconómicos afectan al precio de la vivienda.

6. Referencias

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/https://doi.org/10.1016/S0169-7439(01)00155-1

Alves, P., & San Juan, L. (2021). *El impacto de la crisis sanitaria del COVID-19 sobre el mercado de la vivienda en España. Banco de España. Boletín Económico 2/2021.*

Amat-Rodrigo, J. (2017). *Clustering y heatmaps: aprendizaje no supervisado con R*. https://rpubs.com/Joaquin_AR/310338

Aspentech (2018). *Aspen ProMV ™ Brochure*.

Ayuntamiento de Valencia. (2022). *Dades estadístiques de la ciutat de València - València*. https://www.valencia.es/es/cas/estadistica/inicio

Battistini, N., Gareis, J., & Roma, M. (2022). *The impact of rising mortgage rates on the euro area housing market*. https://www.ecb.europa.eu/pub/economic-bulletin/focus/2022/html/ecb.ebbox202206_04~786da4a23a.es.html

Breiman, L. (1984). *Classification and Regression Trees*. Routledge.

Breiman, L. (2001). *Random Forests* (Vol. 45).

Catastro (2022). *Catastro*. http://www.catastro.minhap.gob.es/esp/faqs.asp

Catedra Observatorio vivienda UPV (2022). *Análisis oferta de obra nueva edificios plurifamiliares VALENCIA ESPAÑA*.

elEconomista (2022). *Comprar una vivienda para alquilar: Valencia, Madrid y Barcelona son más rentables que Londres o París - elEconomista.es*. https://www.eleconomista.es/vivienda-inmobiliario/noticias/11816778/06/22/Comprar-una-vivienda-para-alquilar-Valencia-Madrid-y-Barcelona-son-mas-rentables-que-Londres-o-Paris.html

Fernández Casal, R., Costa Bouzas, J., & Oviedo de la Fuente, M. (2022). *Árboles de regresión CART | Aprendizaje Estadístico*. https://rubenfcasal.github.io/aprendizaje_estadistico/%C3%A1rboles-de-regresi%C3%B3n-cart.html

Forbes (2022). *El inmobiliario bate récord en inversión directa al superar los 15.200 millones de euros en 2022 en España - Forbes España*. https://forbes.es/ultima-hora/213890/el-inmobiliario-bate-record-en-inversion-directa-al-superar-los-15-200-millones-de-euros-en-2022-en-espana/

Fotocasa (2022). *El precio de la vivienda de segunda mano cierra 2022 con una subida del 7,5%, la tercera más abultada de los últimos 17 años - fotocasa*. https://prensa.fotocasa.es/el-precio-de-la-vivienda-de-segunda-mano-cierra-2022-con-una-subida-del-75-la-tercera-mas-abultada-de-los-ultimos-17-anos/

Fotocasa (2022). *Precio viviendas Valencia Capital m^2 - enero de 2023 | Fotocasa*. https://www.fotocasa.es/indice-precio-vivienda/valencia-capital/todas-las-zonas

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Gutiérrez, R., González, A., Torres, F., & Gallardo, J. A. (1994). *Técnicas de Análisis de datos multivariable. Tratamiento Computacional*.

Hernandez Barajas, F. (2021). *7 Gradient Boost | Modelos Predictivos*. https://fhernanb.github.io/libro_mod_pred/gradboost.html#ejemplo-7

IBM (2021). *Funcionamiento de SVM - Documentación de IBM*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works

IBM (2021). *What is Random Forest? | IBM*. https://www.ibm.com/topics/random-forest#anchor-2083981824

Idealista (2022). *La inversión inmobiliaria crece en España — idealista/news*. https://www.idealista.com/news/inmobiliario/vivienda/2022/12/12/800511-la-inversion-inmobiliaria-en-espana-crece-un-10-en-2022-segun-inviertis

INE (2009). *Índice de Precios de Vivienda*.

INE (2022). *INEbase / Nivel y condiciones de vida (IPC) /Índices de precios de consumo y vivienda /Índice de precios de vivienda / Últimos datos*. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736152838&menu=ultiDatos&idp=1254735976607

le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis* (Vol. 163). Sage.

Llorca, J. (2022). *Valencia: la ciudad española más rentable para invertir en vivienda*. https://www.levante-emv.com/economia/2022/12/28/valencia-ciudad-espanola-rentable-invertir-vivienda-76874501.html

Méndez Gutiérrez del Valle, R., & Plaza Tabasco, J. (2016). Housing crisis and mortgage foreclosures in Spain: A geographical perspective. *Boletin de La Asociacion de Geografos Espanoles*, *2016*(71), 99–127. https://doi.org/10.21138/bage.2276

Moreno, J. (2022). *Sector Inmobiliario*. https://www.bankinter.com/broker/analisis/informes

R Project (2022). *R: What is R?* https://www.r-project.org/about.html

Rstudio (2018). *RStudio | The Popular Open-Source IDE* . https://www.rstudio.com/products/rstudio/

Sociedad de tasación (2023). *El precio de la vivienda nueva aumenta un 7,1% durante el último año*. Sociedad de Tasación. https://www.st-tasacion.es/es/mas-alla-del-valor/el-precio-de-la-vivienda-nueva-aumenta-un-7-1-durante-el-ultimo-ano.html

Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, *1*, 91–94. https://doi.org/10.1109/FSKD.2010.5569740

Anexo 1. Alineamiento con los ODS

Los objetivos de desarrollo sostenible (ODS) fueron instaurados en 2015 por todos los estados miembros de la ONU, a través de la aprobación de la Agenda 2030 sobre el desarrollo sostenible.

El objetivo de la ONU es poner fin a la pobreza, proteger el planeta y garantizar que todas las personas gocen de paz y prosperidad, entre otros, para el año 2030. Para conseguir estos objetivos, la agenda cuenta con 17 Objetivos de desarrollo sostenible.

El presente trabajo se encuentra muy relacionado con los **ODS 1: Fin de la pobreza** y el **ODS 11: Ciudades y comunidades sostenibles**, en el cual se trata el problema de la rápida urbanización del planeta, lo que ha ocasionado una sobrepoblación en las ciudades.

Este crecimiento desmesurado ha obligado a un gran número de personas a trasladarse a los barrios más pobres, ha afectado a las infraestructuras y servicios debido a la sobrecarga, ha incrementado el número de emisiones contaminantes y ha fomentado un crecimiento urbanístico descontrolado.

El presente trabajo da un punto de vista sobre como la sobrepoblación ha impactado en la ciudad de Valencia, llegando a afectar en gran medida a los precios de la vivienda debido al escaso suelo disponible, lo que está provocando un crecimiento del precio de las viviendas en la ciudad y un incremento de los alquileres ante la escasa oferta de vivienda disponible.

También se ha podido observar como estos incrementos en los alquileres, ha obligado a muchas familias a tener que desplazarse a las afueras de la ciudad o a municipios cercanos, por la imposibilidad de hacer frente a esta subida del alquiler.