GRADO DE INGENIERÍA INFORMÁTICA
TRABAJO DE FIN DE GRADO
CURSO 2022-20223 Business Intelligence

Análisis del uso de las principales fuentes de energía en España: Análisis causal y predicción mediante técnicas de Business Intelligence

TRABAJO FIN DE GRADO

PEDRO JUEZ MARTEL

CURSO 2023/24

INDICE

Abstract en inglés

Introduction

The current public energy policies, as well as the Agenda 2030, highlight the importance of promoting the use of clean energy at home. However, the choice of energy source for households depends on variables that are often unknown. It is important for public policies and energy companies to know this variables and predict the consumer behaviour to take correct decisions.

Objectives

The objective of the work is twofold:

- Apply Artificial Intelligence and multivariate analysis techniques to model and predict the energy source a household will use.
- Apply these same techniques to identify the explanatory variables of the use of each energy source

This can be use later for the public policies and the business strategy of energy companies.

Methodology

To study the causality and the prediction of the different energy sources we will use:

1. Multivariate analysis techniques such as contingency tables or the study of standardized residuals to detect relationships between variables.

2. Use of classification models like the logit or probit model.

3 Use of Artificial Intelligence techniques belonging to the field of machine learning, such as decision trees and random forest and Artificial Intelligence techniques belonging to the field of deep learning, such as neural networks.

Conclusions

- At an explanatory level, we find that the most polluting energy sources are more strongly correlated with rural households, cold temperatures in the Autonomous Community, and elementary education.
- Gas is more closely associated with Autonomous Communities with cold temperatures, new and urban housing, higher education, and women.
- Among the three models, the multilayer perceptron achieves the highest accuracy, reaching up to 70%. Both the random forest and the

perceptron models achieve higher accuracy than the multinomial probit model.

- Thanks to these AI techniques, we have achieved high prediction accuracy and also conducted a causal analysis of the main energy sources, performing an analysis for each of them. Of all, electricity achieves the lowest level of precision, while gas is predicted more accurately and its causal relationships are easier to explain.
- Electricity is the least predictably modeled energy source. It is particularly linked to Autonomous Communities with high temperatures and single-member households. It is the most challenging energy source to predict.

Introducción

Las políticas públicas actuales de energía, así como la Agenda 2030 resaltan la importancia del fomento en el uso de energías limpias en el hogar. Sin embargo, la elección por parte de los hogares de la fuente de energía para su vivienda, depende de variables que en muchas ocasiones se desconocen.

El poder asimismo predecir el comportamiento de los consumidores en la elección de la fuente de energía para su hogar de acuerdo a su nivel socioeconómico, género, nivel educativo, temperatura de la Comunidad Autónoma en la que reside o las características de la vivienda es un aspecto fundamental tanto para las políticas públicas como de las empresas del sector energético.

Objetivos del trabajo

El objetivo del trabajo es doble:

- Aplicar técnicas de Inteligencia Artificial y de análisis multivariante que permitan modelizar y predecir la fuente de energía que usará un hogar.
- Aplicar estas mismas técnicas para encontrar las variables explicativas del uso de cada fuente de energía.

En concreto, en el trabajo emplearemos de este modo las siguientes herramientas:

- Técnicas de análisis multivariante como tablas de contingencia o estudio de residuos estandarizados para detectar relaciones entre variables
- Empleo de modelos de clasificación como el modelo probit

- Técnicas de Inteligencia Artificial pertenencientes al campo del machine learning como árboles de decisión y random forest
- Técnicas de Inteligencia Artificial pertenecientes al campo del deep learning como son las redes neuronales.

Situación actual del tema: Análisis de la literatura científica

Con carácter previo a realizar los modelos y herramientas de IA que predizcan la elección de la fuente de energía es muy importante analizar el estado del tema.

El desafío sin precedentes del cambio climático ha llevado a la urgencia de implementar una transición energética descarbonizada. Esto debería basarse en la adopción de fuentes de energía más limpias y la electrificación de los usos finales de la energía, incluido el calentamiento. El uso de energía térmica en edificios, principalmente para calefacción y agua caliente, representa una cuarta parte del uso final de la energía en todo el mundo. A nivel mundial, alrededor del 85% de la calefacción y refrigeración en edificios se satisface con combustibles fósiles (AIE, 2020).

Varias fuentes de energía pueden utilizarse para la calefacción, incluidos combustibles fósiles sólidos (carbón), gas, petróleo, energías renovables y electricidad. El calor puede producirse de manera centralizada, en plantas de energía especializadas desde donde se transporta mediante tuberías a los consumidores o, como es el caso en España, la producción de calor puede ser descentralizada, como cuando los ciudadanos utilizan energía para calentar edificios (AEE 2023). El foco de este documento se centra en las elecciones de calefacción en el sector residencial y, en particular, en la electricidad, que, según el IPCC (2022) y Rosenow et al (2023), es una opción clave de "calefacción limpia". Fomentar la adopción de esas opciones de calefacción más limpias es un desafío, ya que a menudo se han encontrado dependencias de trayectorias en el suministro de calor residencial, por ejemplo, en la Unión Europea (UE) (Bertelsen et al., 2020).

El acceso a servicios energéticos modernos, incluida la calefacción, es un objetivo que los gobiernos buscan, y se han adoptado medidas para mitigar la pobreza energética y facilitar dicho acceso. Por ejemplo, el gobierno español afirmó que, en 2019, entre el 6.6% y el 16.7% de la población estaba en situación de pobreza energética (Barrella et al., 2021). Según Eurostat (2023b), la pobreza energética (medida como porcentaje de hogares que no pueden mantener adecuadamente caliente su hogar) ha aumentado sustancialmente entre 2019 y 2022 en la UE (del 6.9% en 2019 al 9.3% en 2022). Sin embargo,

este aumento fue mayor en España (del 7.5% en 2019 al 17.1% en 2022), lo que sugiere una mayor vulnerabilidad de los hogares españoles a la pobreza energética.

Ambos objetivos pueden estar interrelacionados, ya que una mayor pobreza energética dificultaría que los hogares utilicen combustibles de calefacción más limpios (pero más caros). Estos trade-offs son un problema bien conocido para lograr los objetivos del llamado trilema energético (acceso a la energía, energía limpia y energía asequible). Además, los dos eventos de la COVID y la guerra en Ucrania probablemente hayan influido en esos objetivos y, por lo tanto, en su relación, ya sea dificultando el acceso a los servicios energéticos o aumentando sus costos. Esto es especialmente cierto en la calefacción residencial, ya que los precios de los combustibles han aumentado (Trading Economics 2023).

Se hace necesario para seleccionar variables y lograr modelos de inferencia adecuados conocer la situación existente de la literatura a nivel mundial. Es por ello que nos vamos a analizar la literatura anterior identificando factores clave que influyen en las elecciones de calefacción, explorando los efectos de las características de los hogares, las características de la vivienda y la ubicación.

La decisión sobre las elecciones de calefacción es un problema con implicaciones en la transición energética descarbonizada y la pobreza energética, ya que grandes partes de los presupuestos familiares se destinan a la calefacción. Dado que la calefacción de espacios es la principal fuente de emisiones de CO2 en el sector de edificios residenciales en todo el mundo, la elección de fuentes de calefacción hacia fuentes más baratas, pero no necesariamente más limpias, tendría efectos perjudiciales en esas emisiones.

Un análisis empírico a nivel micro de los factores que influyen en la elección de una fuente de calefacción específica (incluida la pobreza energética) revelaría implicaciones políticas relevantes. Podría permitir la identificación de una combinación de políticas, es decir, políticas en los tres ámbitos (apoyo a la transición energética descarbonizada, mitigación de la pobreza energética y mitigación de los efectos de la COVID y la guerra en Ucrania) que aborden eficazmente esos desafíos. También podría permitir la identificación de conflictos, sinergias y complementariedades entre esos ámbitos, lo que a su vez podría llevar a intervenciones políticas beneficiosas para todos, es decir, instrumentos que aborden un objetivo y también tengan efectos secundarios en otro objetivo o intenten equilibrar los trade-offs entre ellos. Por último también

seria una herramienta para empresas energéticas y la necesaria interrelación que éstas deben tener con las políticas públicas de los gobiernos en este tema.

Si bien hay varias contribuciones sobre los determinantes de la elección de fuentes de calefacción y también hay muchos documentos sobre los determinantes de la pobreza energética (ver, por ejemplo, Belaid y Flambard (2023) y Charlier et al 2021), el impacto de la pobreza energética en la decisión de elegir diferentes opciones de calefacción no se ha estudiado mucho (con la excepción de Burguillo et al., 2022) y ninguno ha investigado la influencia de los eventos recientes de la COVID y la guerra en Ucrania en dicha decisión. De hecho, se ha prestado poca atención al análisis cuantitativo de la elección de las fuentes de energía.

El calentamiento tiene una gran importancia en el consumo de energía y las emisiones de GEI de los países. En la Unión Europea (UE), la calefacción de espacios representa el 64.4% del consumo de energía de los hogares y el 17% del consumo final de energía en general (Eurostat 2023a), siendo la actividad que más energía utiliza en los hogares y, por lo tanto, el elemento más destacado del gasto energético relacionado con los hogares. El sector residencial representa el 35% de las emisiones de GEI relacionadas con la energía en la UE (AEE 2023).

Debido al papel importante del petróleo y el gas en la calefacción residencial, esta está fuertemente vinculada a consideraciones políticas sobre el calentamiento global, la seguridad del suministro de energía y el aumento de los precios de la energía (Michelsen y Madlener 2013). Se espera que los sistemas de energía baja en carbono dependan en gran medida de la electrificación de los usos finales, con electricidad producida con bajas emisiones de GEI utilizada para calefacción de edificios e industrial (IPCC 2022). La electrificación generalizada de los usos finales, incluida la calefacción de espacios, debe ir acompañada de un aumento de las energías renovables en los sistemas eléctricos para lograr una reducción neta de CO2.

En España, el consumo de energía para calefacción en el sector residencial ha disminuido en los últimos 10 años en un 27%, como resultado de mejoras en la eficiencia energética y un mejor rendimiento de los equipos (ADL 2023). La energía para todas las fuentes de calefacción ha disminuido, excepto para la solar térmica, y la electrificación de la calefacción no ha experimentado cambios significativos en la última década. Las participaciones de gas natural, petróleo y biomasa oscilan entre el 33% y el 25%, mientras que la electricidad representa solo el 8% del consumo de calefacción.

Estos datos sugieren que hay un considerable potencial para aumentar el uso de la calefacción eléctrica en España. Sin embargo, los sistemas de calefacción tienen características especiales que dificultan su adopción por parte de los propietarios: son productos duraderos (utilizados durante aproximadamente 20 años), costosos de adquirir (con costos de inversión entre 9000 y 17 000 euros) y con costos variables futuros desconocidos (Braun 2010, Decker y Menrad 2015). Diferentes tipos de calefacción tienen características distintas, siendo la calefacción eléctrica la alternativa más costosa (Martinopoulos et al. 2018, Rosenow et al. 2023, Wang et al. 2019, Sunderland y Gibb 2022).

La elección de sistemas de calefacción a lo largo del tiempo se ha explicado tradicionalmente con el llamado modelo de escalera energética, que sugiere que los hogares abandonan tecnologías ineficientes, baratas y contaminantes y eligen energías más sofisticadas a medida que aumenta el ingreso (Belaid y Massié 2022). Pasan de combustibles más primitivos (coque de estiércol, residuos agrícolas y leña) a combustibles de transición (carbón, carbón vegetal y queroseno) a combustibles modernos (GLP, gas natural y electricidad). Con el aumento de los ingresos, los hogares pueden soportar los mayores costos de calefacción para obtener un ambiente interior más saludable (Wang et al. 2019). Este papel importante del ingreso también se asume en este documento, que se centra en la influencia de la pobreza energética en las elecciones de calefacción. Pero se consideran factores adicionales que influyen en la adopción de fuentes más limpias, como las dependencias de trayectorias causadas por hábitos e inversiones previas en infraestructura (viviendas) y varias características de la vivienda y el hogar (por ejemplo, ubicación rural, edad y educación).

Metodología empleada

Se han analizado las tres últimas Encuestas de Presupuestos Familiares (EPF) la de 2019, 2021 y 2022 con una muestra de 66.033 hogares. No se ha considerado 2020 por el efecto que podría existir por la pandemia y también para poder analizar los cambios anteriores y posteriores.

Situación actual y evolución en los años previos

Las muestras de la ENCUESTA DE PRESUPUESTOS FAMILIARES (EPF) de los tres años estudiados muestran porcentajes de uso de las fuentes de energía muy similares con muy pocas variaciones interanuales.

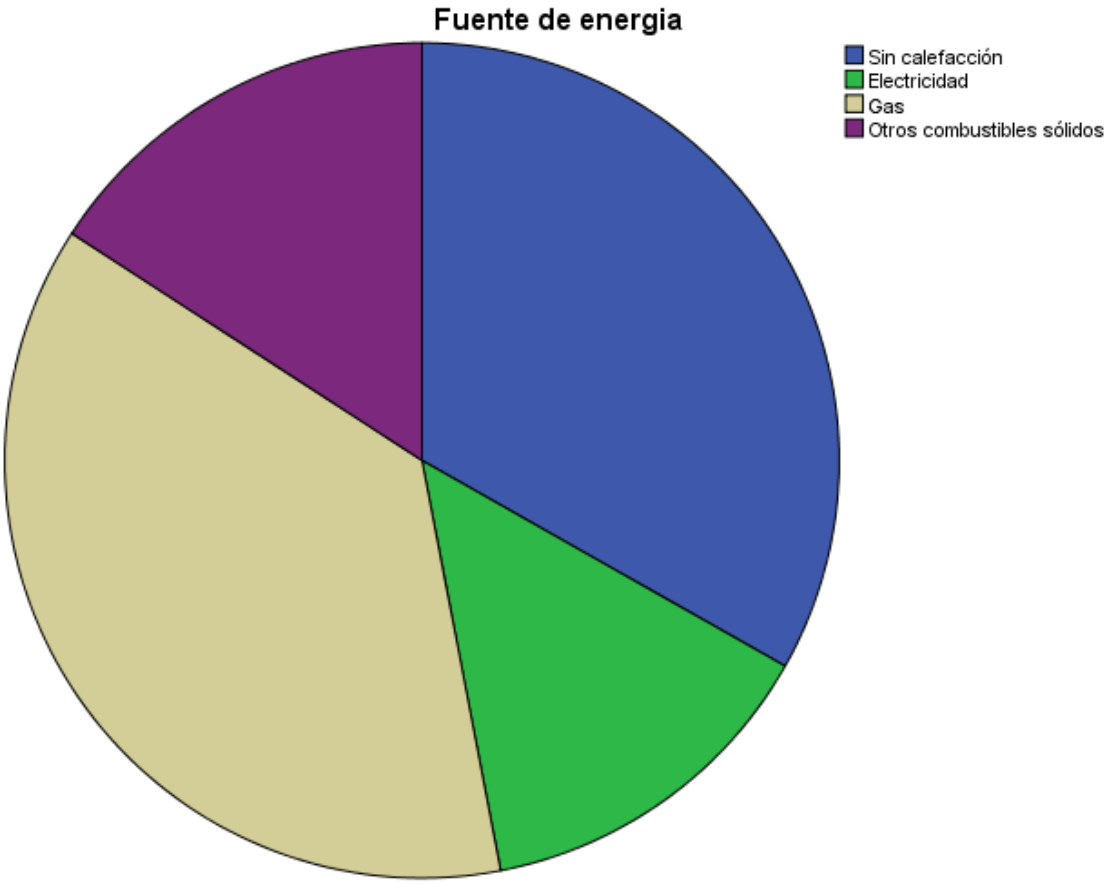
En primer lugar hemos realizado tal y como se informó en el proyecto inicial Aproximadamente la tercera parte de los hogares no tiene calefacción,

con una ligera disminución interanual, alrededor de un 39% usan gas como fuente de energía habiendo variado ligeramente al alza desde 2019, un 14% usa electricidad sin apenas variación interanual y un 16% otros combustibles solidos sin apenas variación interanual.

AÑO 2019

Fuente de energía

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Sin calefacción	6910	33,2	33,2	33,2
	Electricidad	2869	13,8	13,8	47,0
	Gas	7738	37,2	37,2	84,1
	Otros combustibles sólidos	3300	15,9	15,9	100,0
	Total	20817	100,0	100,0	

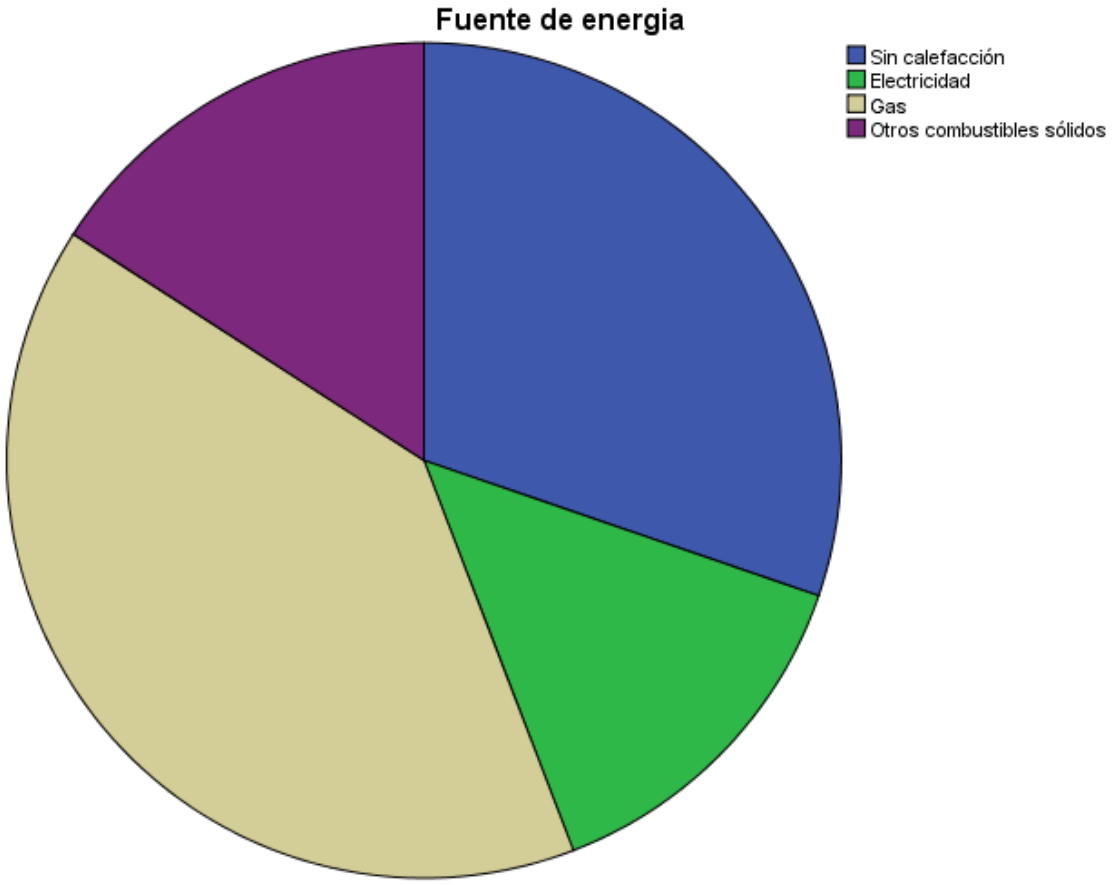


GRADO DE INGENIERÍA INFORMÁTICA
TRABAJO DE FIN DE GRADO
CURSO 2022-20223 Business Intelligence

AÑO 2021

Fuente de energía

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Sin calefacción	7451	30,3	30,3	30,3
	Electricidad	3429	13,9	13,9	44,2
	Gas	9839	39,9	39,9	84,1
	Otros combustibles sólidos	3912	15,9	15,9	100,0
	Total	24631	100,0	100,0	

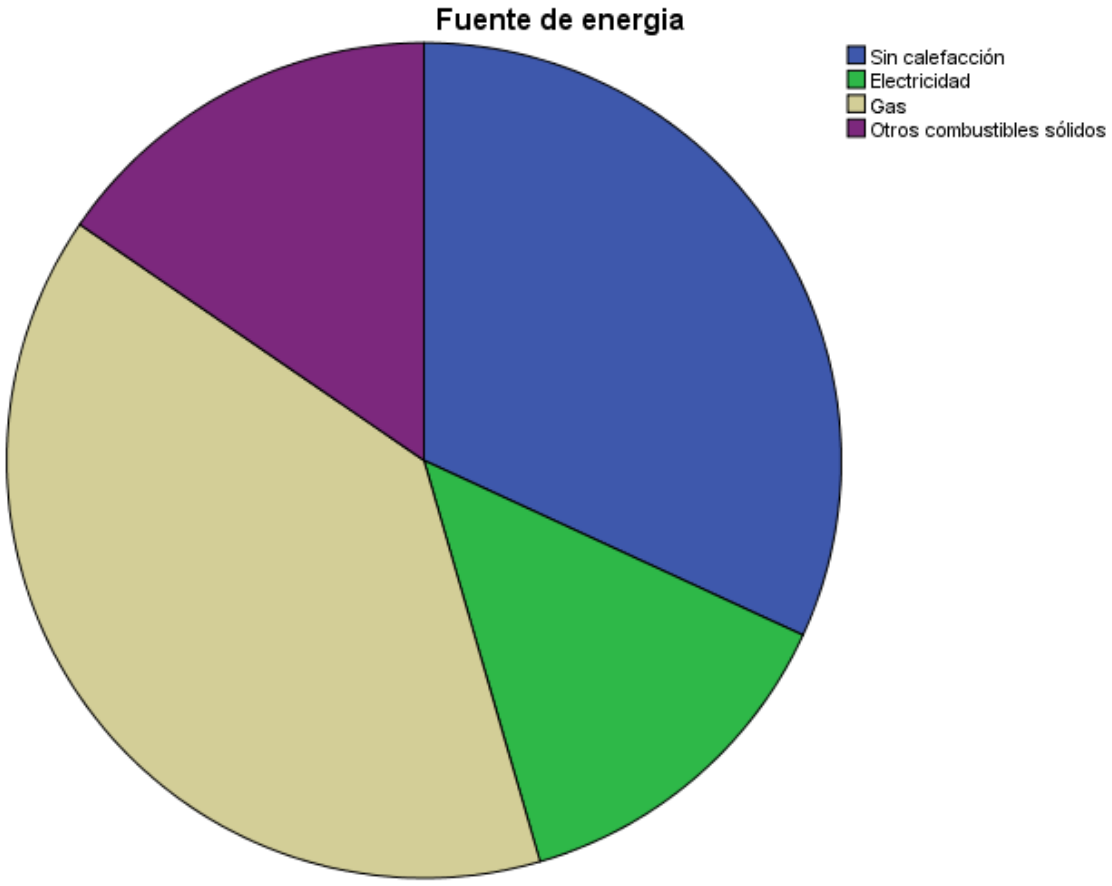


AÑO 2022

Fuente de energía

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado

Válido	Sin calefacción	6560	31,9	31,9	31,9
	Electricidad	2809	13,6	13,6	45,5
	Gas	8037	39,0	39,0	84,6
	Otros combustibles sólidos	3179	15,4	15,4	100,0
	Total	20585	100,0	100,0	



Hacia la construcción de un modelo: Análisis de las variables que pueden incidir en la elección de fuentes de energía mediante técnicas de análisis multivariante: estudio de los residuos estandarizados y tablas de contingencia. Aplicación en la EPF

A continuación voy a analizar la relación entre el uso de las fuentes de energía y las variables que pueden incidir en ella de manera específica.

En las tablas la codificación de las fuentes de energía es la siguiente: 1= Electricidad; 2 = Gas; y 3 = Carbón y otros combustibles sólidos.

En concreto las siguientes.

Relación con la pobreza energética

Se ha usado el índice Linares (1= Hogar pobre energético; 0 =Hogar no pobre energético).

El análisis mediante tablas de contingencia muestra que existe una clara relación entre esta variable y las fuentes de energía de acuerdo al coeficiente chi-cuadrado.

Al realizar un análisis direccional por niveles vemos -a través de los residuos corregidos- que los hogares que son pobre energéticos muestran una correlación positiva significativa directa (+9,4) en primer lugar con el uso de combustibles sólidos y siendo el uso del gas que tiene la correlación inversa más elevada (-4,5) la que menos usan. Por su parte, los hogares no pobres usan más el gas como fuente de energía.

RELACIÓN FUENTES DE ENERGIA CON NIVEL DE POBREZA ENERGÉTICA MEDIDO CON EL INDICE LINARES

Tabla cruzada FUENTES DE ENERGÍA*LINARES

			LINARES		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	3175	254	3429
		% dentro de FUENTES DE ENERGÍA	92,6%	7,4%	100,0%
		% dentro de LINARES	20,4%	15,7%	20,0%
		Residuo corregido	4,5	-4,5	
	2	Recuento	8997	842	9839
		% dentro de FUENTES DE ENERGÍA	91,4%	8,6%	100,0%
		% dentro de LINARES	57,8%	52,1%	57,3%
		Residuo corregido	4,4	-4,4	
	3	Recuento	3393	519	3912
		% dentro de FUENTES DE ENERGÍA	86,7%	13,3%	100,0%
		% dentro de LINARES	21,8%	32,1%	22,8%
		Residuo corregido	-9,4	9,4	
Total		Recuento	15565	1615	17180
		% dentro de FUENTES DE ENERGÍA	90,6%	9,4%	100,0%
		% dentro de LINARES	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	92,862[a]	2	,000
Razón de verosimilitud	87,208	2	,000
Asociación lineal por lineal	77,267	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 322,34.

Relación con la temperatura en invierno de la Comunidad Autónoma

En relación a esta variable se establecen dos niveles diferenciando aquellas CCAA con mayor temperatura e inviernos menos fríos de aquellas donde el invierno tiene un clima más gélido.

Se establece así la variable Temperature (1=CCAA con inviernos fríos con mediana menor a 10 grados; 0=CCAA con mediana de temperatura superior a 10 grados)

Mostramos la salida obtenida mediante SPSS y con posterioridad realizamos el análisis

RELACIÓN FUENTES DE ENERGÍA CON TEMPERATURA

Tabla cruzada

			TEMPERATURE 0	TEMPERATURE 1	Total
FUENTES DE ENERGÍA	1	Recuento	1291	2138	3429
		% dentro de FUENTES DE ENERGÍA	37,6%	62,4%	100,0%
		% dentro de TEMPERATURE	60,1%	14,2%	20,0%
		Residuo corregido	49,7	-49,7	
	2	Recuento	530	9309	9839
		% dentro de FUENTES DE ENERGÍA	5,4%	94,6%	100,0%

		% dentro de TEMPERATURE	24,7%	61,9%	57,3%
		Residuo corregido	-32,7	32,7	
	3	Recuento	328	3584	3912
		% dentro de FUENTES DE ENERGÍA	8,4%	91,6%	100,0%
		% dentro de TEMPERATURE	15,3%	23,8%	22,8%
		Residuo corregido	-8,9	8,9	
Total		Recuento	2149	15031	17180
		% dentro de FUENTES DE ENERGÍA	12,5%	87,5%	100,0%
		% dentro de TEMPERATURE	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	2497,180[a]	2	,000
Razón de verosimilitud	2028,243	2	,000
Asociación lineal por lineal	1306,024	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 428,92.

La interpretación de los residuos estandarizados muestra como los coeficientes positivos y significativos. Esto significa que, cuanto más fría es una CCAA, más probable es que elija gas y carbón que electricidad. No hay diferencias sustanciales en el tiempo.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado.

Al realizar un análisis direccional por niveles vemos -a través de los residuos corregidos- que las comunidades más frías usan en primer lugar el gas como fuente de energía mostrando una alta correlación positiva en su residuo (+32,7)

y con los combustibles sólidos en menor proporción (+8,9). Por su parte las comunidades menos frías lo hacen con el empleo de la electricidad (+49,7).

Relación con el tipo de Municipio

Hemos diferenciado los tipos de municipios en si son o no rurales, De este modo, en la codificación seguida para el análisis 1 = rural; 0 = no rural.

Mostramos a continuación la salida obtenida

Tabla cruzada

			municipio		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	2747	682	3429
		% dentro de FUENTES DE ENERGÍA	80,1%	19,9%	100,0%
		% dentro de municipio	21,0%	16,7%	20,0%
		Residuo corregido	6,0	-6,0	
	2	Recuento	8413	1426	9839
		% dentro de FUENTES DE ENERGÍA	85,5%	14,5%	100,0%
		% dentro de municipio	64,2%	34,9%	57,3%
		Residuo corregido	33,1	-33,1	
	3	Recuento	1936	1976	3912
		% dentro de FUENTES DE ENERGÍA	49,5%	50,5%	100,0%
		% dentro de municipio	14,8%	48,4%	22,8%
		Residuo corregido	-44,7	44,7	
Total		Recuento	13096	4084	17180
		% dentro de FUENTES DE ENERGÍA	76,2%	23,8%	100,0%
		% dentro de municipio	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	2039,532[a]	2	,000
Razón de verosimilitud	1857,143	2	,000

Asociación lineal por lineal	1047,146	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 815,14.

La interpretación de los residuos estandarizados muestra que los coeficientes para gas son negativos, mientras que los de carbón son positivos. Son todos significativos. Esto indica que si el municipio es rural, la probabilidad de utilizar

electricidad es mayor con respecto al gas que si el municipio es no rural, y que la probabilidad de utilizar carbón es mayor con respecto a utilizar electricidad que si el municipio es no rural.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional por niveles observamos una alta correlación positiva de los municipios rurales con el uso de los combustibles sólidos. En el caso de los municipios urbanos la fuente preferida es en primer lugar el gas, y la electricidad y la ausencia de calefacción a continuación.

Relación con el tamaño del hogar en función del número de miembros

La variable se muestra por número de miembros por hogar y se muestra a continuación la salida obtenida

RELACION FUENTES DE ENERGÍA CON HOGAR

Tabla cruzada

			tipohogar		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	2516	913	3429
		% dentro de FUENTES DE ENERGÍA	73,4%	26,6%	100,0%
		% dentro de tipohogar	19,3%	21,9%	20,0%
		Residuo corregido	-3,6	3,6	
	2	Recuento	7401	2438	9839

		% dentro de FUENTES DE ENERGÍA	75,2%	24,8%	100,0%
		% dentro de tipohogar	56,9%	58,4%	57,3%
		Residuo corregido	-1,7	1,7	
	3	Recuento	3089	823	3912
		% dentro de FUENTES DE ENERGÍA	79,0%	21,0%	100,0%
		% dentro de tipohogar	23,8%	19,7%	22,8%
		Residuo corregido	5,4	-5,4	
Total		Recuento	13006	4174	17180
		% dentro de FUENTES DE ENERGÍA	75,7%	24,3%	100,0%
		% dentro de tipohogar	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	33,946[a]	2	,000
Razón de verosimilitud	34,475	2	,000
Asociación lineal por lineal	31,899	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 833,10.

Podemos observar como en lo que se refiere a los residuos estandarizados el signo es negativo para gas y carbón (y significativo). El signo negativo indica que hogares unipersonales es más probable que elijan electricidad frente a gas o carbón (comparado con los no unipersonales).

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional por niveles observamos una alta correlación -a través de los residuos corregidos- de los hogares que tienen sólo un miembro o más de 5 miembros con la ausencia de calefacción. La electricidad es más usada en los hogares con un solo miembro, no existiendo significatividad en la relación del tamaño con el resto de fuentes de energía. Por su parte el gas es la

fuente preferida al mostrar mayor correlación significativa en los hogares de cuatro miembros. Por último, en el caso de los combustibles sólidos observamos una relación directa en el uso en el caso de los hogares de dos miembros e inversa en el caso de que el tamaño sea una sola persona no siendo significativo el resto de los tamaños.

Relación con el nivel de educación

Cruzamos ahora la variable relativa a fuentes de energía con el nivel educativo. Hemos separado aquellos que tienen educación superior de aquellos que no la tienen.

La codificación es en concreto la siguiente. Consideramos la educación recibida por el sustentador principal.:

La variable se muestra así; Educación (1=educación superior; 0=no educación superior)

Mostramos a continuación la salida obtenida

RELACIÓN FUENTES DE ENERGÍA CON EDUCACIÓN

Tabla cruzada

			Educ		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	1970	1459	3429
		% dentro de FUENTES DE ENERGÍA	57,5%	42,5%	100,0%
		% dentro de educ	20,3%	19,5%	20,0%
		Residuo corregido	1,2	-1,2	
	2	Recuento	5132	4707	9839
		% dentro de FUENTES DE ENERGÍA	52,2%	47,8%	100,0%
		% dentro de educ	52,8%	63,1%	57,3%
		Residuo corregido	-13,5	13,5	
	3	Recuento	2614	1298	3912
		% dentro de FUENTES DE ENERGÍA	66,8%	33,2%	100,0%
		% dentro de educ	26,9%	17,4%	22,8%
		Residuo corregido	14,7	-14,7	
Total		Recuento	9716	7464	17180

	% dentro de FUENTES DE ENERGÍA	56,6%	43,4%	100,0%
	% dentro de educ	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	246,245[a]	2	,000
Razón de verosimilitud	250,205	2	,000
Asociación lineal por lineal	76,382	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 1489,76.

Los coeficientes de los residuos estandarizados son siempre positivos y significativos con la excepción: no significativa del carbón en 2021. Significa que a mayor educación, mayor probabilidad de elegir gas o combustibles solidos que electricidad para calefacción (curioso resultado, tal vez contraintuitivo).

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional por niveles vemos -a través de los residuos corregidos- que los hogares sin educación superior muestran mayor correlación con la ausencia de calefacción en primer lugar y con el empleo de combustibles sólidos en segundo lugar, y los hogares con estudios universitarios lo hacen con el gas en primer lugar y la electricidad en segundo lugar.

Relación con los años de construcción del edificio

La codificación realizada es la siguiente. En ella, 1= construcción <25 años; y 0= >25 años.

A continuación mostramos la relación de las fuentes de energía con la antigüedad del edificio

Presentamos a continuación la salida

RELACIÓN FUENTES DE ENERGÍA CON ANTIGÜEDAD DEL EDIFICIO

Tabla cruzada

			buildingage		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	2240	1189	3429
		% dentro de FUENTES DE ENERGÍA	65,3%	34,7%	100,0%
		% dentro de buildingage	19,3%	21,4%	20,0%
		Residuo corregido	-3,2	3,2	
	2	Recuento	6402	3437	9839
		% dentro de FUENTES DE ENERGÍA	65,1%	34,9%	100,0%
		% dentro de buildingage	55,1%	61,8%	57,3%
		Residuo corregido	-8,2	8,2	
	3	Recuento	2973	939	3912
		% dentro de FUENTES DE ENERGÍA	76,0%	24,0%	100,0%
		% dentro de buildingage	25,6%	16,9%	22,8%
		Residuo corregido	12,8	-12,8	
Total		Recuento	11615	5565	17180
		% dentro de FUENTES DE ENERGÍA	67,6%	32,4%	100,0%
		% dentro de buildingage	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	162,866[a]	2	,000
Razón de verosimilitud	169,278	2	,000
Asociación lineal por lineal	102,946	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 1110,73.

El signo para los residuos estandarizados es positivo para gas (y significativo): indica que edificios de más reciente construcción es más probable que tengan gas para calefacción que electricidad.

Asimismo observamos que existe signo negativo para carbón (y significativo): indica que edificios de más reciente construcción es más probable que tengan electricidad para calefacción que carbón.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional por niveles observamos -a través de los residuos corregidos- que los hogares con antigüedad superior a 25 años tienen una correlación directa y mayor tanto con la ausencia de calefacción como con el empleo de combustibles sólidos. Por su parte los hogares con antigüedad superior a 25 años tienen correlación positiva con el gas en primer lugar y con la electricidad en segundo lugar.

Relación con el tipo de vivienda

Se considera aquí si la vivienda es en régimen de comunidad o de tipo unifamiliar. La codificación asociada es la siguiente:

Tipo de vivienda (1=unifamiliar; 0=no unifamiliar)

Mostramos a continuación la salida y análisis de los residuos estandarizados para esta variable.

RELACIÓN FUENTES DE ENERGÍA CON TIPOLOGIA DE LA VIVIENDA

Tabla cruzada

			dwellingtype		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	2471	958	3429
		% dentro de FUENTES DE ENERGÍA	72,1%	27,9%	100,0%
		% dentro de dwellingtype	20,2%	19,4%	20,0%
		Residuo corregido	1,1	-1,1	
	2	Recuento	8361	1478	9839

		% dentro de FUENTES DE ENERGÍA	85,0%	15,0%	100,0%
		% dentro de dwellingtype	68,3%	30,0%	57,3%
		Residuo corregido	45,9	-45,9	
	3	Recuento	1415	2497	3912
		% dentro de FUENTES DE ENERGÍA	36,2%	63,8%	100,0%
		% dentro de dwellingtype	11,6%	50,6%	22,8%
		Residuo corregido	-55,2	55,2	
Total		Recuento	12247	4933	17180
		% dentro de FUENTES DE ENERGÍA	71,3%	28,7%	100,0%
		% dentro de dwellingtype	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	3258,817[a]	2	,000
Razón de verosimilitud	3093,172	2	,000
Asociación lineal por lineal	1307,313	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 984,59.

En este caso, observamos que en el caso de los residuos estandarizados el signo negativo para gas y positivo para carbón (significativo para todos los años). Esto implica que las viviendas unifamiliares es más probable que elijan combustibles sólidos y gas para calefacción que viviendas colectivas (detached, flat). No hay patrón claramente definido.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional por niveles observamos que los hogares unifamiliares tienen una correlación directa y mayor tanto con el empleo de combustibles sólidos, los hogares no unifamiliares por su parte tienen

correlación positiva con el gas en primer lugar y con la electricidad en segundo lugar

Relación con la edad del sustentador principal

Se presenta a continuación la relación con la edad del sustentador principal queriendo comprobar si la edad puede influir sobre la elección del tipo de fuente de energía.

La codificación es la siguiente

Edad (1= >67 años; 0= <67 años).

La salida obtenida es la siguiente:

RELACIÓN FUENTES DE ENERGÍA CON LA EDAD DEL SUSTENTADOR PRINCIPAL

Tabla cruzada

			Age		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	2649	780	3429
		% dentro de FUENTES DE ENERGÍA	77,3%	22,7%	100,0%
		% dentro de Age	20,5%	18,2%	20,0%
		Residuo corregido	3,3	-3,3	
	2	Recuento	7580	2259	9839
		% dentro de FUENTES DE ENERGÍA	77,0%	23,0%	100,0%
		% dentro de Age	58,8%	52,7%	57,3%
		Residuo corregido	7,0	-7,0	
	3	Recuento	2664	1248	3912
		% dentro de FUENTES DE ENERGÍA	68,1%	31,9%	100,0%
		% dentro de Age	20,7%	29,1%	22,8%
		Residuo corregido	-11,4	11,4	
Total		Recuento	12893	4287	17180
		% dentro de FUENTES DE ENERGÍA	75,0%	25,0%	100,0%
		% dentro de Age	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	130,656[a]	2	,000
Razón de verosimilitud	125,860	2	,000
Asociación lineal por lineal	87,985	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 855,65.

El análisis de los residuos estandarizados muestra como los coeficientes son negativos para el gas en todos los años, pero no significativos. Para el carbón, son positivos y significativos. Esto significa que los hogares con personas más mayores tienen una mayor probabilidad de utilizar carbón para calefacción que electricidad, con respecto a hogares con moradores más jóvenes.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional por niveles observamos -a través de los residuos corregidos- que los mayores de 67 años tienen una correlación directa y mayor con el uso de combustibles sólidos, no existiendo significatividad respecto a la ausencia de calefacción. Por su parte los menores de 67 años tienen una correlación directa mayor por el gas en primer lugar y la electricidad en segundo lugar, siendo negativa la correlación con el uso de combustibles sólidos.

Relación con la nacionalidad del sustentador principal

A continuación estudiamos la posible relación entre la nacionalidad del sustentador principal y la elección de la fuente de energía.

La codificación que se muestra es la siguiente:

Nationalidad (1=español; 0=no español).

RELACIÓN FUENTES DE ENERGÍA CON NACIONALIDAD

Tabla cruzada

			NATIONALITY		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	173	3256	3429
		% dentro de FUENTES DE ENERGÍA	5,0%	95,0%	100,0%
		% dentro de NATIONALITY	32,0%	19,6%	20,0%
		Residuo corregido	7,1	-7,1	
	2	Recuento	307	9532	9839
		% dentro de FUENTES DE ENERGÍA	3,1%	96,9%	100,0%
		% dentro de NATIONALITY	56,7%	57,3%	57,3%
		Residuo corregido	-,3	,3	
	3	Recuento	61	3851	3912
		% dentro de FUENTES DE ENERGÍA	1,6%	98,4%	100,0%
		% dentro de NATIONALITY	11,3%	23,1%	22,8%
		Residuo corregido	-6,5	6,5	
Total		Recuento	541	16639	17180
		% dentro de FUENTES DE ENERGÍA	3,1%	96,9%	100,0%
		% dentro de NATIONALITY	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	72,868[a]	2	,000
Razón de verosimilitud	74,001	2	,000
Asociación lineal por lineal	72,408	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 107,98.

El análisis de los residuos estandarizados muestran signo positivo para todos los combustibles y años (excepción: gas en 2019, que aparece como negativo y no significativo). Indica que la probabilidad de elegir gas o carbón como fuente de calefacción (con respecto a electricidad) es mayor si el sustentador principal es español que si no lo es.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional -a través de los residuos corregidos- observamos que los hogares españoles tienen mayor correlación con los combustibles sólidos frente a los no españoles que lo tienen con el uso de electricidad. Respecto al gas no hay diferencias significativas al ser el valor del residuo inferior a 1,96.

Relación con el tipo de propiedad

Analizamos a continuación la relación que puede existir con la propiedad o no de la vivienda del sustentador principal. La codificación es la siguiente:

Vivienda en propiedad (1=vivienda en propiedad; 0=vivienda en alquiler)

La salida obtenida es la siguiente:

RELACIÓN FUENTES DE ENERGÍA CON VIVIENDA EN PROPIEDAD

Tabla cruzada FUENTES DE ENERGÍA*VIVIENDAPROP

			VIVIENDAPROP		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	2971	458	3429
		% dentro de FUENTES DE ENERGÍA	86,6%	13,4%	100,0%
		% dentro de VIVIENDAPROP	19,3%	25,7%	20,0%
		Residuo corregido	-6,4	6,4	
	2	Recuento	8707	1132	9839
		% dentro de FUENTES DE ENERGÍA	88,5%	11,5%	100,0%
		% dentro de VIVIENDAPROP	56,6%	63,4%	57,3%

		Residuo corregido	-5,5	5,5	
	3	Recuento	3717	195	3912
		% dentro de FUENTES DE ENERGÍA	95,0%	5,0%	100,0%
		% dentro de VIVIENDAPROP	24,1%	10,9%	22,8%
		Residuo corregido	12,6	-12,6	
Total		Recuento	15395	1785	17180
		% dentro de FUENTES DE ENERGÍA	89,6%	10,4%	100,0%
		% dentro de VIVIENDAPROP	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	Df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	168,322[a]	2	,000
Razón de verosimilitud	191,747	2	,000
Asociación lineal por lineal	143,764	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 356,27.

En este caso el estudio de los residuos estandarizados muestra como el signo de los coeficientes es negativo para todos los años y combustibles (y significativo). Sugiere que tener vivienda en propiedad hace más probable utilizar electricidad y gas para calefacción que con respecto a tener vivienda en alquiler que usan más combustibles sólidos como el carbón.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional -a través de los residuos corregidos- observamos que los hogares que no son en propiedad tienen mayor correlación con la electricidad y los combustibles sólidos frente a los que son en propiedad donde hay más correlación con el gas. El resto de fuentes de energía no se muestra significativo por el hecho de ser en propiedad o no serlo.

Relación con el género

Por último estudiamos la relación entre la elección de la fuente de energía y el género del sustentador principal. Esta es la codificación que se ha seguido.

Género (1=Hombre; 0=Mujer)

Se presenta a continuación la salida obtenida

RELACIÓN FUENTES DE ENERGÍA CON GENDER

Tabla cruzada

			GENDER		
			0	1	Total
FUENTES DE ENERGÍA	1	Recuento	1222	2207	3429
		% dentro de FUENTES DE ENERGÍA	35,6%	64,4%	100,0%
		% dentro de GENDER	20,7%	19,6%	20,0%
		Residuo corregido	1,7	-1,7	
	2	Recuento	3481	6358	9839
		% dentro de FUENTES DE ENERGÍA	35,4%	64,6%	100,0%
		% dentro de GENDER	58,9%	56,4%	57,3%
		Residuo corregido	3,2	-3,2	
	3	Recuento	1205	2707	3912
		% dentro de FUENTES DE ENERGÍA	30,8%	69,2%	100,0%
		% dentro de GENDER	20,4%	24,0%	22,8%
		Residuo corregido	-5,4	5,4	
Total		Recuento	5908	11272	17180
		% dentro de FUENTES DE ENERGÍA	34,4%	65,6%	100,0%
		% dentro de GENDER	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	28,947[a]	2	,000
Razón de verosimilitud	29,316	2	,000
Asociación lineal por lineal	20,277	1	,000
N de casos válidos	17180		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 1179,19.

Respecto a la interpretación de los residuos estandarizados solo es significativo (con signo negativo) en 2022 para gas. Significa que no hay relación. El de 2022 significa que los hombres tienen más probabilidad de elegir combustibles sólidos que gas para calentarse que las mujeres.

El análisis mediante tablas de contingencia muestra también relación entre esta variable y fuentes de energía de acuerdo al coeficiente chi-cuadrado que es significativo.

Al realizar un análisis direccional -a través de los residuos corregidos- observamos que los hogares donde el sustentador principal es hombre tienden a usar más combustibles sólidos, no siendo significativas el resto de fuentes de energía. En el caso de la mujer es el gas la fuente de energía preferida no siendo significativo el resto de fuentes de energía por el hecho de ser hombre o mujer.

Elección de las variables del primer modelo

La herramienta metodológica adecuada en este primer modelo que estimaremos es el modelo multinomial (logit o probit). Por lo tanto, nos hemos centrado en las contribuciones en la literatura que utilizan esta metodología.

De hecho, con este apartado pretendemos adoptar un marco teórico pragmático que se basa en las variables incluidas en la abundante literatura sobre los determinantes de diferentes fuentes de calefacción en el sector residencial. Es especialmente relevante la literatura que analiza específicamente los factores que llevan a los hogares a elegir un tipo específico de calefacción en lugar de otros.

Las variables dependientes en esos documentos suelen incluir electricidad, gas, carbón o petróleo, al igual que en este artículo, utilizando una de ellas como categoría de referencia. Las variables explicativas se refieren ya sea a las características del hogar (por ejemplo, tamaño del hogar, edad,

ingresos, nivel educativo, nacionalidad y género del sostén de familia), la vivienda (por ejemplo, tamaño y edad) o la ubicación de la vivienda (por ejemplo, en áreas rurales o urbanas, en regiones con un clima más frío o más cálido). Algunas variables explicativas (ingresos, edad del jefe del hogar, tamaño del hogar, educación y edad del edificio) se incluyen con mayor frecuencia en los modelos, mientras que otras rara vez se utiliza

Table 2. Variables incluidas en los modelos para determinar la elección de fuentes de energia en la literatura

	Propiedad de la vivienda	Edad del sustentador principal	Género	Tamaño del hogar	Area para calentar	Ingreso	Número de miembros del hogar	Existencia de niños	Educacion	Antiguedad del edificio	Localizacion)	Rural
1.Decker and Menrad (2015)		*				* (not included in MNL)	*		*			
2.Belaïd and Massié (2022)	*	*		*		*	*					
3.Bai et al (2023)		*	*	*	*	*	*		*			
4.Michelsen and Madlener (2012)	*	*	*	*		*			*	*	*	*
5.Wang et al (2019)					*	*	*		*			
6. Laureti and Secondi (2012)				*		*	*	*	*	*	*	
7.Braun (2010)						*	*	*	*	*	*	*
8.Couture (2012)	*	*		*		*	*			*		
9.Chen (2021)		*	*			*	*	*	*			
10. Liao and Chang (2002)	*	*	*	*	*	*	*			*	* (if dwelling is in city, town or suburbs)	

Fuente: Elaboración propia

Análisis de las principales variables en la literatura y su relación con la elección de fuentes de energia

Pobreza Energética:

Relacionada con los ingresos, siendo más probable que los hogares con mayores ingresos adopten fuentes de calefacción más limpias como el gas y la electricidad.

Hallazgos contradictorios sobre la relación entre los ingresos y la elección de la electricidad.

Temperatura:

Se prefiere el gas sobre la electricidad y el petróleo en regiones más frías debido a la eficiencia del gas y a los costos más bajos.

Hallazgos mixtos sobre la relación entre vivir en regiones frías y la elección de las fuentes de calefacción.

Ruralidad de la Ubicación de la Vivienda:

Menor probabilidad de usar gas para la calefacción en áreas rurales debido a la menor desarrollo de las redes de gas y al mayor espacio para el almacenamiento externo de combustibles.

Tamaño del Hogar:

Impacto ambiguo en la elección del sistema de calefacción; los hogares más grandes pueden usar menos electricidad, pero también se ha observado una correlación positiva con la calefacción eléctrica.

La calefacción eléctrica es favorable para hogares pequeños con flexibilidad en el uso.

Nivel Educativo:

Influencia ambigua en las elecciones de calefacción.

Se asocia la educación superior con la elección de gas en algunos estudios, mientras que otros muestran una correlación positiva con la calefacción eléctrica.

Edad del Edificio:

Los edificios más recientes tienen más probabilidades de tener calefacción eléctrica y gas.

Las casas más antiguas tienen más probabilidades de usar petróleo o carbón debido a tecnologías históricas.

Tipo de Vivienda:

Evidencia limitada sobre la influencia del tipo de vivienda, con algunos estudios indicando preferencias por el gas en casas adosadas y calefacción eléctrica en edificios de apartamentos.

Edad del Sustentador del Hogar:

Impacto ambiguo en las elecciones de calefacción; la edad puede estar asociada con preferencias por fuentes de calefacción específicas.

Los propietarios mayores pueden preferir el petróleo, mientras que los más jóvenes pueden ser más abiertos a fuentes alternativas.

Nacionalidad:

Evidencia limitada, solo Belaid y Massié (2022) incluyeron esta variable en su modelo, encontrando que las personas de nacionalidad extranjera al nacer (no franceses) eran un 10% más propensas a usar electricidad y un 16% más propensas a usar otras energías en lugar de gas.

Propiedad (propietarios vs. inquilinos):

La influencia del estado de propiedad en las elecciones de calefacción es ambigua en la literatura.

Propietarios pueden preferir la opción de petróleo, mientras que inquilinos pueden ser más propensos a elegir gas, según diferentes estudios.

Género:

No se espera una influencia unidireccional del género en la elección de tipos de calefacción.

Michelsen y Madlener (2012) encontraron que la variable "mujer" solo fue significativa para la calefacción a gas, mientras que Chen (2021) mostró que los hogares con más mujeres tenían una mayor preferencia por la electricidad sobre la leña.

Estimación del primer modelo de predicción: Modelo probit multinomial

La primera manera en el que vamos a estudiar el impacto de la pobreza energética en la decisión de utilizar diferentes fuentes de calefacción e identificar si los eventos mencionados anteriormente han afectado esta decisión es empleando un modelo probit. Se estima un modelo probit multinomial utilizando información de una amplia base de datos de hogares españoles (la Encuesta de Presupuestos Familiares) en 2019, 2021 y 2022 del que ya hemos hablado en la segunda PEC. El modelo nos permite desentrañar la influencia de diferentes factores en la probabilidad de aplicar un modo de calefacción residencial (gas, electricidad o petróleo).

Modelo probit multinomial.

Con el fin de aislar el efecto de los diferentes determinantes en la probabilidad de elegir una de las tres alternativas de calefacción (electricidad, gas y petróleo/carbón), se especifica un modelo econométrico. La probabilidad incondicional p_jt de que un hogar español i utilice una fuente de energía específica j para la calefacción en lugar de otra, sujeta a un conjunto de características X_it en el período t, se estima mediante un modelo de elección discreta multinomial no ordenado, en el que la variable dependiente toma tres valores (1, 2 y 3 si el hogar elige electricidad, gas o petróleo/carbón para la calefacción, respectivamente).

Como se muestra en la Tabla de estimación del modelo, los modelos multinomiales no ordenados generalmente se han estimado utilizando la alternativa logit multinomial (MNL). Sin embargo, el MNL realiza la fuerte suposición de la independencia de las alternativas irrelevantes (IIA, por sus siglas en inglés). Cuando la IIA se viola, el MNL está incorrectamente especificado, y los coeficientes estimados son sesgados e inconsistentes (Jumbe

y Angelsen 2011, Kropko 2008). Para superar este problema, estimamos un modelo probit multinomial no ordenado (MNP), ya que el modelo MNP no asume la IIA y produce estimaciones robustas en comparación con las estimaciones del MNL (Álvarez y Nagler 1998, Jumbe y Angelsen 2011, Mensah y Adu 2015). Sin embargo, el MNP no está exento de problemas. Evita imponer la IIA a través del término de perturbación en lugar del componente sistemático, lo cual puede no ser un enfoque totalmente satisfactorio para resolver el problema de la IIA (Álvarez y Nagler 1998). Las estimaciones se realizaron para 2019, 2021 y 2022 utilizando el método de máxima verosimilitud.

Un problema importante en la evaluación de las elecciones de calefacción en los hogares es que el análisis a menudo se realiza con aquellos hogares que informan el uso de una alternativa de calefacción dada, ignorando el hecho de que algunos hogares no utilizan ningún sistema de calefacción en absoluto. Esto es problemático desde un punto de vista econométrico, ya que conduce al conocido sesgo de selección de muestra. Sin embargo, en nuestro caso, este problema se mitiga, ya que se analiza toda la muestra (incluidos los hogares con y sin calefacción).

Resultado del primer modelo

Las Tablas resumen los resultados del modelo. Todos los coeficientes tienen el signo esperado y la mayoría de ellos son estadísticamente significativos. El modelo también es estadísticamente significativo de manera conjunta según la prueba de chi-cuadrado de razón de verosimilitud (valores p de 0,000 en los tres años, es decir, al menos uno de los coeficientes en el modelo no es cero). Las tablas de clasificación muestran que el modelo clasifica correctamente el 62%, 63% y 64% de los casos en 2019, 2021 y 2022, respectivamente.

En el modelo probit multinomial, los parámetros estimados proporcionan el impacto de la variable explicativa en la probabilidad de elegir la categoría de uso en comparación con la categoría de referencia (electricidad) (Tabla 6). Dado que los valores de los coeficientes no se pueden interpretar directamente, sino solo su signo, también se proporcionan los efectos marginales (EM) (en medias muestrales) para ayudar a comprender mejor las relaciones de sustitución entre diferentes tipos de calefacción (Tabla 7). Dado que todas las variables explicativas son variables ficticias, los EM se interpretan en términos de diferencias de probabilidad (1 versus 0), considerando la probabilidad predicha de que cada tipo de calefacción sea elegido por un hogar (Chen 2021).

Pobreza energética

Los coeficientes para la pobreza energética son todos positivos y significativos (excepción: gas en 2022). Por lo tanto, ser más energéticamente pobre aumenta la probabilidad de usar gas y petróleo para la calefacción en comparación con la electricidad. Los efectos marginales son negativos para gas y electricidad y positivos para el petróleo, es decir, ser energéticamente pobre aumenta la probabilidad de usar petróleo y reduce la probabilidad de usar gas y electricidad.

Los efectos marginales son menores en 2022 que en 2019 y 2021, lo que sugiere que la probabilidad de elegir petróleo ha disminuido después de la COVID y la guerra en Ucrania. Sin embargo, la COVID tuvo un impacto mayor (los EM disminuyeron fuertemente entre 2019 y 2021 y aumentaron ligeramente entre 2021 y 2022). En contraste con el petróleo, sin embargo, los EM son mayores en 2021 que en 2019, lo que sugiere que la COVID aumentó la probabilidad de que los pobres energéticos eligieran gas. La falta de significación estadística en 2022 indica que, de manera similar al petróleo, la guerra en Ucrania no ha afectado tanto la elección de los modos de calefacción como lo hizo la COVID. En contraste con el petróleo/gas, ser energéticamente pobre reduce la probabilidad de elegir electricidad (el signo negativo del EM, aunque solo es significativo para 2021).

Temperatura

Los coeficientes son todos positivos y significativos, mientras que los efectos marginales son todos significativos: negativos para electricidad y positivos para petróleo y gas. Por lo tanto, cuanto más fría sea la región donde se encuentra la vivienda, mayor será la probabilidad de que se elija gas y petróleo (en comparación con la electricidad). Belaid y Massié (2022) también encontraron que las viviendas ubicadas en regiones más frías eran más propensas a usar gas que electricidad y petróleo. Pero nuestros resultados contrastan con Vaage (2000), quien encontró que las condiciones climáticas y la elección de la calefacción estaban relacionadas y, especialmente, Secondi y Laureti (2012), quienes mostraron que la ubicación de la vivienda en regiones más frías influyó positivamente en la probabilidad de elegir petróleo, influyó negativamente en la probabilidad de elegir gas y no tuvo relación con la elección de la electricidad. Los EM no son sustancialmente diferentes a lo largo del tiempo, es decir, ni la COVID ni la guerra en Ucrania han tenido un fuerte efecto en tal elección.

Ruralidad

Los coeficientes y los efectos marginales son todos significativos. Los coeficientes para gas son negativos y positivos para el petróleo. Los efectos marginales son positivos para electricidad y petróleo y negativos para gas. Por lo tanto, si la vivienda está ubicada en una zona rural, es más probable que se utilice electricidad y petróleo en comparación con las ubicaciones en áreas no rurales. En las viviendas rurales, es más probable que se utilice electricidad que gas (en comparación con áreas no rurales). La menor probabilidad de usar gas en áreas rurales está en línea con la literatura (Braun 2010, Liao y Chang 2002, Michelsen y Madlener 2012, Laureti y Secondi 2012, Belaid y Massié 2022) y se explica por una red de gas subdesarrollada y el mayor espacio disponible para el almacenamiento de combustible en áreas rurales.

No se pueden observar diferencias sustanciales entre los efectos marginales (EM) para gas y petróleo a lo largo del tiempo. En contraste, los EM para electricidad disminuyeron sustancialmente entre 2019 y 2021 y entre 2021 y 2022. Por lo tanto, tanto la COVID como la guerra en Ucrania influyeron en los hogares rurales para que utilicen menos electricidad.

Tamaño del hogar:

El signo positivo (y significativo) de los coeficientes para gas y petróleo en todos los años sugiere que, cuanto mayor sea el tamaño del hogar, mayor será la probabilidad de usar petróleo o gas (en comparación con la electricidad). Los efectos marginales son todos significativos, con un signo positivo para gas y petróleo y uno negativo para electricidad. Por lo tanto, un tamaño mayor aumenta la probabilidad de elegir gas y petróleo y reduce la probabilidad de usar electricidad. Estos hallazgos coinciden en gran medida con Belaid y Massié (2022) y Bai et al (2023), quienes mostraron que los hogares más grandes usan menos electricidad (por las razones mencionadas de flexibilidad en el uso del combustible y costos), pero en contraste con Wang et al (2019), quienes encontraron que los hogares más grandes eran más propensos a elegir electricidad. Los EM para electricidad y petróleo se han mantenido constantes a lo largo del tiempo, mientras que los EM para gas aumentaron sustancialmente entre 2019 y 2021 y disminuyeron entre 2021 y 2022. Por lo tanto, la COVID ha tenido el efecto más fuerte, llevando a un mayor incentivo para que los hogares más grandes usen gas.

Educación:

Los coeficientes son siempre positivos y significativos para gas y petróleo (excepción: no significativo para petróleo en 2021). Por lo tanto, un nivel educativo alto está relacionado con una mayor probabilidad de elegir petróleo o gas en lugar de electricidad. Esto coincide con Laureti y Secondi (2012), Braun (2010), Belaid y Massié (2022) y Michelsen y Madlener (2012), quienes encontraron que un nivel educativo más alto está relacionado con la elección de gas para la calefacción, pero en contraste con Braun (2010) (en cuanto al petróleo), Decker y Menrad (2015), Bai et al (2023) y Chen (2021). Los EM son todos significativos y positivos, pero, como los EM para electricidad son más bajos, un nivel educativo más alto lleva a un aumento mayor en el uso de gas y petróleo que en la electricidad. Aunque positivos, los EM siguen una tendencia a la baja, especialmente entre 2019 y 2021 y para el petróleo, lo que sugiere un efecto mayor de la COVID que de la guerra en Ucrania en este sentido.

Edad del Edificio

El signo para el gas es positivo y significativo, mientras que es negativo y significativo para el petróleo. Por lo tanto, es más probable que los edificios más recientes se calienten con gas en lugar de electricidad y más probable que utilicen electricidad en lugar de petróleo para la calefacción. Esto concuerda con Braun (2010), Laureti y Secondi (2012), Vaage (2000), Belaid y Massié (2022) y Liao y Chang (2002). Los efectos marginales (ME) son positivos para la electricidad y el gas y negativos para el petróleo. Por lo tanto, es más probable que los edificios más recientes utilicen electricidad y gas. El petróleo para la calefacción se aplicaba principalmente en casas antiguas (Laureti y Secondi 2012), mientras que los edificios nuevos probablemente tienen gas (Michelsen y Madlener 2012) o calefacción eléctrica (Vaage 2000, Liao y Chang 2002). La reducción en los efectos marginales entre 2019 y 2021 para la electricidad y el gas sugiere que el impacto de la COVID ha sido mayor que el de la guerra en Ucrania.

Tipo de Vivienda

El signo negativo para el gas y el positivo para el petróleo (y significativo) indica que es más probable que se utilice electricidad para la

calefacción (en comparación con el gas) y petróleo (en comparación con la electricidad) en viviendas unifamiliares que en viviendas no unifamiliares (pisos). Los efectos marginales son todos significativos y negativos para la electricidad y el gas y positivos para el petróleo. Por lo tanto, es más probable que las viviendas unifamiliares utilicen petróleo y menos probable que utilicen electricidad y gas que las viviendas no unifamiliares. Braun (2010) también encontró que el gas era una opción más atractiva para las casas adosadas en comparación con las casas unifamiliares y lo contrario se encontró para el petróleo. En cambio, se encontró una alta preferencia por la calefacción eléctrica en bloques de apartamentos en Vaage (2000). Belaid y Massié (2022) encontraron que las casas individuales (en comparación con las casas gemelas) eran más propensas a utilizar electricidad en lugar de gas debido a las dificultades para conectarse a las redes de gas. Los efectos marginales son menos positivos para el petróleo y menos negativos para la electricidad y el gas en 2021 que en 2019. Los cambios entre 2021 y 2022 son inexistentes (gas) o menores (electricidad y petróleo).

Edad

Los coeficientes para la edad no son significativos. Son positivos y significativos para el petróleo. Por lo tanto, los hogares con miembros mayores tienen una mayor probabilidad de usar petróleo que electricidad para la calefacción que los hogares con miembros más jóvenes, pero la edad no está relacionada con la elección entre gas y electricidad. Los ME son positivos y significativos para el petróleo, pero no son significativos para la electricidad y el gas. Por lo tanto, es más probable que las personas mayores elijan petróleo, pero la edad no está relacionada con la elección de electricidad y gas. Nuestros resultados contrastan con Chen (2021), quien encontró que el envejecimiento aumenta la probabilidad de elegir electricidad. Sin embargo, están en línea con Belaid y Massié (2022), quienes mostraron que la elección de petróleo con respecto al gas no estaba relacionada con la edad, pero que el envejecimiento reducía la probabilidad de elegir electricidad, tal vez debido a la conveniencia (Liao y Chang 2002). Michelsen y Madlener (2012) encontraron que los propietarios mayores preferían el petróleo, mientras que los más jóvenes estaban más abiertos a las bombas de calor, reflejando diferentes aversiones al riesgo. Las familias mayores usaban más calefacción de petróleo en Liao y Chang (2002). Los ME disminuyen sustancialmente entre 2019 y 2021 para el petróleo. Por lo tanto, aunque la probabilidad de elegir petróleo aumenta con la edad tanto en 2019 como en 2021, la COVID ha reducido dicha probabilidad.

Nacionalidad

El signo positivo para la mayoría de los combustibles y años sugiere que la probabilidad de elegir gas o petróleo (en comparación con la electricidad) es mayor si el sustentador de la familia nació en España. Los ME son negativos para la electricidad y positivos para el petróleo y el gas. Por lo tanto, ser español aumenta la probabilidad de usar petróleo y gas y reduce la probabilidad de usar electricidad. Esta variable no ha recibido mucha atención en la literatura. Belaid y Massié (2022) encontraron que las personas de nacionalidad extranjera al nacer en lugar de francesa eran más propensas a usar electricidad y otras energías en lugar de gas. Los efectos marginales no siguen un patrón claro a lo largo del tiempo.

Propiedad

El signo de los coeficientes es negativo y significativo para todos los años y combustibles, lo que sugiere que ser propietario de la vivienda aumenta la probabilidad de usar electricidad en lugar de gas y petróleo (en comparación con ser inquilino). Los Efectos marginales no son significativos para la electricidad y son negativos y significativos para el gas y el petróleo. Por lo tanto, ser propietario de la casa no influye en la elección de electricidad y afecta negativamente la elección de gas y petróleo. Nuestros resultados contrastan con Braun (2010), quien encontró que los propietarios preferían el petróleo, y Laureti y Secondi (2012), quienes mostraron que los propietarios eran más propensos a elegir gas que las familias que vivían en viviendas alquiladas, mientras que los propietarios eran menos propensos a elegir electricidad. Belaid y Massié (2022) encontraron que, en comparación con los propietarios, los inquilinos eran más propensos a usar petróleo. Sin embargo, la elección entre gas, electricidad y otras fuentes no se vio afectada significativamente por el estado de ocupación. Los efectos marginales para el gas se vuelven menos negativos y más negativos para el petróleo a lo largo de los años. Por lo tanto, la COVID y la guerra en Ucrania han disminuido la probabilidad de que los propietarios elijan gas y petróleo.

Género

El género no está relacionado con la elección de modos de calefacción, ya que el coeficiente solo es significativo (y negativo) en 2022 para el gas. En

comparación con las mujeres, los hombres tienen más probabilidades de elegir electricidad en lugar de gas en 2022. Los efectos marginales solo son significativos (y positivos) para el petróleo. Por lo tanto, ser hombre aumenta la probabilidad de usar petróleo. La falta de significancia estadística del género coincide con Michelsen y Madlener (2012).

Coeficientes del probit multinomial probit (Resultado base: electricidad).

	GAS			PETROLEO/CARBÓN		
	2019	2021	2022	2019	2021	2022
Pobreza	0.105*	0.317***	0.209***	0.221***	0.022	0.183***
Temperatura	1.689***	1.186***	1.788 ***	1.208 ***	1.768 ***	1.173 ***
Rural	-0.381 ***	0.357 ***	-0.419 ***	0.388***	-0.350***	0.435***
Tamaño hogar	0.047***	0.050***	0. 075***	0.051***	0.075***	0.051***
Educacion	0.244***	0.148 ***	0.226 ***	0.010	0.238 ***	0.090 **
Antigüedad edificio	0.339 ***	-0.372 ***	0.266 ***	-0.383***	0.333 ***	-0.349 ***
Tipo de vivienda	-0.497 ***	1.058 ***	-0.487 ***	0.944 ***	-0.486 ***	0.976 ***
Edad	0.039	0.194 ***	0.017	0.119 ***	0.058	0.168 ***
Nationalidad	-0.036	0.218 **	0.235***	0.391 ***	0.218***	0.245**
Propiedad	-0.281***	-0.268***	-0.294 ***	-0.402***	-0.250 ***	-0.433 ***
Genero	-0.014	0.021	0.017	0.043	-0.073**	0.033

Nota: *, **, y *** indican significatividad a niveles de un 10%, 5%, y 1% respectivamente (*p < 0.05, **p < 0.01, ***p < 0.001).

Efectos marginales de las estimaciones del modelo probit multinomial

	ELECTRICIDAD			GAS			PETROLEO/CARBON		
	2019	2021	2022	2019	2021	2022	2019	2021	2022
Pobreza	-0.044***	-0.045***	-0.015	-0.009	0.034**	-0.022	0.054***	0.011	0.037***
Temperature	-0.105***	-0.131***	-0.112***	0.569***	0.615***	0.622***	0.174***	0.165***	0.165***
Rural	0.013***	0.007***	0.003***	-0.144***	-0.181***	-0.152***	0.088***	0.091***	0.096***
Tamaño hogar	-0.007***	-0.009***	0.008***	0.007**	0.196***	0.020***	0.004*	0.003*	0.004*
Educacion	0.022***	0.017***	0.017***	0.143***	0.136***	0.137***	0.050***	0.017***	0.032***
Antigüedad	0.283***	0.032***	0.018***	0.193***	0.179***	0.183***	-0.584***	-0.060***	-0.060***
Tipo de vivienda	-0.045***	-0.030***	-0.035***	-0.289***	-0.279***	-0.279***	0.194***	0.188***	0.182***
Edad	-0.012	-0.003	-0.008	-0.002	0.002	0.012	0.032***	0.024***	0.029***
Nationalidad	0.004	-0.041***	-0.029**	-0.008	0.039**	0.051***	0.051***	0.054***	0.030*
Propiedad	-0.007	0.008	0.003	-0.135***	-0.124***	-0.108***	-0.064***	-0.084***	-0.098***
Genero	0.013**	0.005	0.010	0.015*	0.009	-0.016*	0.015**	0.010*	0.015**

Nota: *, **, y *** indican significatividad a niveles de un 10%, 5%, y 1% respectivamente (*p < 0.05, **p < 0.01, ***p < 0.001).

Aplicación de modelos de IA a la resolución del problema

Las técnicas de Inteligencia Artificial (IA) se pueden clasificar en varias categorías. Una de las más usadas es clasificarlas en herramientas de Machine Learning (Aprendizaje Automático) y de Deep Learning (Aprendizaje Profundo).

La Inteligencia Artificial (IA) puede clasificarse también en:

IA débil (o estrecha): Se refiere a sistemas de IA diseñados para realizar tareas específicas sin capacidad para generalizar más allá de esas tareas. Ejemplos incluyen asistentes virtuales, sistemas de recomendación y chatbots.

IA fuerte (o general): Se refiere a sistemas que poseen la capacidad de comprender, aprender y aplicar conocimientos de manera similar a los humanos. La IA fuerte es un objetivo a largo plazo y aún está en desarrollo.

A su vez las técnicas de IA pueden clasificarse en

Aprendizaje Automático (Machine Learning - ML):

Aprendizaje Supervisado: Algoritmos entrenados con un conjunto de datos que incluye ejemplos con entradas y salidas deseadas. El modelo generaliza a partir de estos ejemplos para hacer predicciones sobre nuevos datos.

Aprendizaje No Supervisado: Algoritmos que se entrenan en conjuntos de datos sin etiquetas. El modelo busca patrones y estructuras inherentes en los datos.

Aprendizaje por Reforzamiento: El modelo toma decisiones secuenciales para maximizar una recompensa acumulativa. Aprende a través de la retroalimentación de las consecuencias de sus acciones.

Aprendizaje Profundo (Deep Learning - DL)

En esta área los principales modelos son los siguientes:

Redes Neuronales Artificiales: Modelos inspirados en la estructura del cerebro humano. Pueden tener varias capas (redes profundas) y son especialmente efectivas para tareas de procesamiento de imágenes, sonido y texto.

Redes Neuronales Convolucionales (CNN): Especializadas en el procesamiento de datos de rejilla -en grid-. Son muy buenas en el análisis de imágenes. Utilizan capas convolucionales para detectar patrones locales. Son muy usadas por ello en campos como la Radiología.

Redes Neuronales Recurrentes (RNN): Diseñadas para trabajar con datos secuenciales, como texto o series temporales. Mantienen una "memoria" de estados anteriores.

Redes Neuronales Generativas (GAN): Consisten en un generador que crea datos y un discriminador que evalúa su autenticidad. Utilizado en la generación de contenido nuevo, como imágenes realistas.

Técnicas Híbridas:

Entre las que encontramos

Transferencia de Aprendizaje: Utiliza conocimientos aprendidos en una tarea para mejorar el rendimiento en otra tarea relacionada.

Aprendizaje Semi-Supervisado: Combina datos etiquetados y no etiquetados en el entrenamiento de modelos.

Aprendizaje por Reforzamiento Profundo (Deep Reinforcement Learning): Combina técnicas de aprendizaje profundo con el aprendizaje por reforzamiento.

Es importante tener en cuenta que estas categorías no son mutuamente excluyentes, y a menudo se utilizan de manera combinada para abordar problemas complejos en la práctica.

Técnicas de IA que aplicaremos a la resolución del problema: Random Forest y Perceptrón Multicapa

Random Forest

Es una técnica de machine learning que se ubica dentro de los métodos de aprendizaje supervisado.

Es un algoritmo que construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones para obtener un resultado más robusto y preciso. Cada árbol en el bosque -forest- se construye de manera independiente, y la predicción final se realiza mediante la votación o promedio de las predicciones individuales.

A diferencia de los modelos de una única máquina de aprendizaje, como el perceptrón multicapa o las redes neuronales, que pertenecen al ámbito del aprendizaje profundo, Random Forest es un método de aprendizaje automático tradicional basado en árboles de decisión. No es parte del aprendizaje profundo, ya que no utiliza arquitecturas de redes neuronales profundas. En cambio, se destaca por su simplicidad, eficacia y capacidad para manejar conjuntos de datos grandes con muchas características.

El algoritmo Random Forest es una técnica de ensamble que combina múltiples árboles de decisión para mejorar la precisión y la generalización del modelo. Cada árbol se entrena en una submuestra aleatoria del conjunto de datos, y las predicciones se combinan a través de votación (en clasificación) o promediado (en regresión). A continuación, se presenta una formulación matemática simplificada para la construcción de un árbol de decisión, que luego se extiende al caso de Random Forest.

Árbol de Decisión: La formulación matemática es la siguiente:

1. **Definiciones:**
   - X: Conjunto de características.
   - Y: Variable objetivo (etiqueta en clasificación o valor en regresión).
   - D: Conjunto de datos de entrenamiento $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$.
2. **Función Objetivo para la División de un Nodo:**
   - Supongamos que estamos en un nodo que representa un conjunto de datos D.
   - Queremos encontrar la mejor característica j y el mejor umbral t para dividir D en dos conjuntos $D_{\text{izquierda}}$ y D_{derecha}.

   La función objetivo para la división puede ser formulada como:

$$\text{Objetivo}(D, j, t) = \text{Impureza}(D) - \left(\frac{|D_{\text{izquierda}}|}{|D|} \cdot \text{Impureza}(D_{\text{izquierda}}) + \frac{|D_{\text{derecha}}|}{|D|} \cdot \text{Impureza}(D_{\text{derecha}})\right)$$

   - $\text{Impureza}(\cdot)$: Una medida de impureza en el conjunto de datos, como la entropía o el índice Gini.
3. **Criterio de Parada:**
   - Establecer un criterio de parada para decidir cuándo detener la subdivisión (profundidad máxima, número mínimo de muestras en un nodo, etc.).

1. **Construcción de Muestras Bootstrap:**
   - Generar B conjuntos de datos de entrenamiento $\{D_1, D_2, ..., D_B\}$ mediante muestreo con reemplazo (bootstrap) del conjunto de datos original D.
2. **Entrenamiento de Árboles:**
   - Para cada conjunto de datos D_i, construir un árbol de decisión utilizando la formulación anterior.
3. **Votación (Clasificación) o Promediado (Regresión):**
   - Para una nueva instancia de entrada x, la predicción de Random Forest se obtiene por votación en el caso de clasificación o promediado en el caso de regresión.

$$\text{Predicción}_{\text{Random Forest}}(x) = \frac{1}{B}\sum_{i=1}^{B} \text{Predicción}_{\text{Árbol}_i}(x)$$

   - $\text{Predicción}_{\text{Árbol}_i}(x)$: Predicción del i-ésimo árbol.

Este algoritmo lo programaremos en Python con el objetivo de comprobar si podemos realizar un mejor ajuste con una de las técnicas más usadas de Machine Learning.

Perceptrón Multicapa

Es una técnica que pertenece a las Redes Neuronales Artificiales dentro del Deep Learning. El Perceptrón Multicapa (MLP) es un precursor y componente fundamental de las arquitecturas

de aprendizaje profundo más avanzadas. A medida que aumenta el número de capas ocultas en un MLP, la red se vuelve más capaz de aprender representaciones jerárquicas y abstracciones complejas de los datos, lo cual es una característica esencial del aprendizaje profundo.

Es una forma de red neuronal artificial que se encuentra bajo el paraguas más amplio del aprendizaje automático y, más específicamente, del aprendizaje profundo. Es una arquitectura clave que ha sido fundamental en el desarrollo de técnicas más avanzadas en el campo de la inteligencia artificial.

La estructura de un MLP consta de tres tipos de capas:

Capa de entrada: Neuronas que representan las características de entrada.

Capas ocultas: Neuronas adicionales entre la capa de entrada y la capa de salida. Cada neurona en una capa oculta realiza una transformación no lineal de las entradas. La presencia de capas ocultas permite al MLP aprender representaciones más abstractas y complejas.

Capa de salida: Neuronas que producen la salida final de la red.

Matemáticamente, el cálculo realizado por cada neurona en una capa oculta o de salida puede describirse de la siguiente manera:

1. **Entrada de la neurona:**
   $z_j^l = \sum_{i=1}^{n^{l-1}} w_{ij}^l a_i^{l-1} + b_j^l$
   Donde:
   - z_j^l es la entrada ponderada a la neurona j en la capa l.
   - w_{ij}^l es el peso de la conexión entre la neurona i en la capa $l-1$ y la neurona j en la capa l.
   - a_i^{l-1} es la salida de la neurona i en la capa $l-1$.
   - b_j^l es el sesgo de la neurona j en la capa l.
   - n^{l-1} es el número de neuronas en la capa $l-1$.
2. **Activación de la neurona:**
   $a_j^l = f(z_j^l)$
   Donde $f(\cdot)$ es una función de activación. Comúnmente se usa la función sigmoide, tangente hiperbólica o ReLU.
3. **Función de pérdida:**
   $L(y, \hat{y})$
   Donde L es la función de pérdida, y es la etiqueta verdadera y $\hat{y}$ es la salida predicha por la red.

Algoritmo de entrenamiento:

El entrenamiento del MLP implica ajustar los pesos y sesgos para minimizar la función de pérdida. Esto se realiza mediante algoritmos de optimización como el descenso del gradiente.

La retropropagación (backpropagation) es un algoritmo comúnmente utilizado para entrenar MLP. Consiste en propagar el error desde la capa de salida hacia atrás a través de la red, ajustando los pesos y sesgos en función de la contribución de cada conexión al error total.

Es importante destacar que la elección de la arquitectura de la red (número de capas y número de neuronas en cada capa), así como las funciones de activación y la función de pérdida, dependen del problema específico que se esté abordando. La selección de hiperparámetros y la arquitectura de la red pueden requerir experimentación y ajuste para obtener un rendimiento óptimo.

Comparación de los tres modelos

Comparar el Perceptrón Multicapa (MLP), Probit y Random Forest implica tener en cuenta diversas características, fortalezas y debilidades de cada técnica. Aquí hay una comparación general entre estas tres técnicas de clasificación:

Perceptrón Multicapa (MLP)

Modelo Neuronal

Utiliza una arquitectura de red neuronal con capas ocultas.

Capacidad de Aprendizaje

Capaz de aprender funciones no lineales y patrones complejos en los datos.

Flexibilidad

Puede modelar relaciones no lineales de manera efectiva debido a su arquitectura profunda.

Interpretación

Puede ser más difícil de interpretar en comparación con modelos lineales como Probit debido a su complejidad.

Requiere Ajuste de Hiperparámetros

La elección de la arquitectura de la red y los hiperparámetros son cruciales y requieren ajustes.

Probit

Respecto al modelo probit que ya hemos estimado éste tiene las siguientes características:

Modelo Probabilístico

El probit asume una distribución normal acumulativa (función de distribución acumulativa probit) para modelar probabilidades.

Interpretabilidad:

Proporciona parámetros interpretables directamente relacionados con la contribución de cada característica.

Suposiciones que necesita:

Requiere que las observaciones sean independientes y que la relación sea lineal. Es por lo tanto un modelo adecuado para Datos Lineales. Al no poder suponer que existe la citada relación lineal estimamos estos dos modelos.

Puede ser más adecuado cuando la relación entre las características y la variable dependiente es aproximadamente lineal.

Random Forest

Como ya expusimos es un ensamble de Árboles de Decisión. Combina múltiples árboles de decisión para mejorar la precisión y generalización.

Manejo de No Linealidades

A diferencia de los modelos probit. Puede manejar eficazmente relaciones no lineales y complejas en los datos.

Robustez Frente a Overfitting (Sobreajuste)

Este modelo es menos propenso al sobreajuste en comparación con un solo árbol de decisión.

Flexibilidad

No asume una relación lineal y es capaz de modelar patrones no lineales.

Consideraciones Generales sobre los tres modelos

Complejidad del Problema

- El MLP puede ser más adecuado para problemas altamente no lineales y complejos.
- Probit puede ser apropiado para problemas más simples con relaciones lineales.
- Random Forest es versátil y puede ser eficaz en una variedad de problemas, especialmente cuando se enfrenta a no linealidades.

Interpretación vs. Rendimiento

La elección entre modelos puede depender de la importancia de la interpretación frente al rendimiento puro en la tarea de clasificación.

Requerimientos Computacionales

- MLP puede ser más intensivo en recursos de cómputo que modelos más simples como Probit o Random Forest.
- La elección entre estas técnicas dependerá del problema específico, la naturaleza de los datos y los objetivos del modelado. Puede ser útil probar múltiples modelos y evaluar su rendimiento en conjuntos de datos de prueba antes de tomar una decisión final.

Estimación del modelo de Deep learning: perceptrón multicapa: Código en Python y análisis del mismo

Realizamos la estimación del perceptrón usando el siguiente código de Python

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.metrics import confusion_matrix, accuracy_score,
precision_score, recall_score

# Lee la base de datos desde un archivo Excel
ruta_archivo_excel = 'c:/Users/Uned/proyecto/Basedatos2019.xls'
df = pd.read_excel(ruta_archivo_excel)

# Selecciona las columnas deseadas (excluyendo la variable objetivo)
variables_explicativas = df[['NATIONALITY', 'GENDER', 'TEMPERATURE',
'buildingage', 'dwellingtype', 'municipio', 'Age', 'educ', 'LINARES']]
variable_objetivo = df['ENERCAL2']

# Codificación one-hot para variables categóricas
variables_explicativas = pd.get_dummies(variables_explicativas,
columns=['NATIONALITY', 'GENDER', 'TEMPERATURE', 'buildingage',
'dwellingtype', 'municipio', 'Age', 'educ', 'LINARES'])

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test =
train_test_split(variables_explicativas, variable_objetivo,
test_size=0.2, random_state=42)

# Codificar la variable objetivo si es necesario
le = LabelEncoder()
y_train_encoded = le.fit_transform(y_train)
y_test_encoded = le.transform(y_test)

# Define el modelo
modelo = Sequential()
modelo.add(Dense(10, input_dim=variables_explicativas.shape[1],
activation='relu'))
modelo.add(Dense(4, activation='softmax'))

# Compila y entrena el modelo
modelo.compile(loss='sparse_categorical_crossentropy', optimizer='adam',
metrics=['accuracy'])
modelo.fit(X_train, y_train_encoded, epochs=50, batch_size=32,
validation_data=(X_test, y_test_encoded))

# Predicciones en el conjunto de prueba
y_pred = modelo.predict(X_test)
y_pred_labels = np.argmax(y_pred, axis=1)
```

```
# Calcula la matriz de confusión
conf_mat = confusion_matrix(y_test_encoded, y_pred_labels)

# Calcula la sensibilidad y especificidad
sensibilidad = recall_score(y_test_encoded, y_pred_labels,
average='macro')
especificidad = recall_score(y_test_encoded, y_pred_labels, pos_label=0,
average='macro')

# Calcula la precisión
precision = precision_score(y_test_encoded, y_pred_labels,
average='macro')

# Imprime los resultados
print("Matriz de Confusión:")
print(conf_mat)
print("Sensibilidad:", sensibilidad)
print("Especificidad:", especificidad)
print("Precisión:", precision)
```

Tras ejecutar el código en Python obtenemos la salida del modelo.

Epoch 1/50

521/521 [==============================] - 2s 2ms/step - loss: 1.1329 - accuracy: 0.5588 - val_loss: 1.0302 - val_accuracy: 0.6150

Epoch 2/50

521/521 [==============================] - 1s 2ms/step - loss: 1.0060 - accuracy: 0.6245 - val_loss: 1.0093 - val_accuracy: 0.6206

Epoch 3/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9969 - accuracy: 0.6252 - val_loss: 1.0092 - val_accuracy: 0.6208

Epoch 4/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9955 - accuracy: 0.6267 - val_loss: 1.0058 - val_accuracy: 0.6222

Epoch 5/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9939 - accuracy: 0.6252 - val_loss: 1.0056 - val_accuracy: 0.6222

Epoch 6/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9936 - accuracy: 0.6260 - val_loss: 1.0083 - val_accuracy: 0.6203

Epoch 7/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9929 - accuracy: 0.6252 - val_loss: 1.0059 - val_accuracy: 0.6222

Epoch 8/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9925 - accuracy: 0.6266 - val_loss: 1.0060 - val_accuracy: 0.6201

Epoch 9/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9921 - accuracy: 0.6273 - val_loss: 1.0051 - val_accuracy: 0.6225

Epoch 10/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9920 - accuracy: 0.6267 - val_loss: 1.0043 - val_accuracy: 0.6213

Epoch 11/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9916 - accuracy: 0.6264 - val_loss: 1.0047 - val_accuracy: 0.6201

Epoch 12/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9915 - accuracy: 0.6260 - val_loss: 1.0043 - val_accuracy: 0.6215

Epoch 13/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9909 - accuracy: 0.6271 - val_loss: 1.0040 - val_accuracy: 0.6215

Epoch 14/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9907 - accuracy: 0.6270 - val_loss: 1.0045 - val_accuracy: 0.6227

Epoch 15/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9908 - accuracy: 0.6263 - val_loss: 1.0038 - val_accuracy: 0.6213

Epoch 16/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9903 - accuracy: 0.6266 - val_loss: 1.0067 - val_accuracy: 0.6208

Epoch 17/50

521/521 [==========] - 1s 2ms/step - loss: 0.9904 - accuracy: 0.6259 - val_loss: 1.0036 - val_accuracy: 0.6225

Epoch 18/50

521/521 [==========] - 1s 2ms/step - loss: 0.9899 - accuracy: 0.6269 - val_loss: 1.0033 - val_accuracy: 0.6225

Epoch 19/50

521/521 [==========] - 1s 2ms/step - loss: 0.9899 - accuracy: 0.6271 - val_loss: 1.0041 - val_accuracy: 0.6213

Epoch 20/50

521/521 [==========] - 1s 2ms/step - loss: 0.9897 - accuracy: 0.6269 - val_loss: 1.0041 - val_accuracy: 0.6208

Epoch 21/50

521/521 [==========] - 1s 2ms/step - loss: 0.9897 - accuracy: 0.6278 - val_loss: 1.0028 - val_accuracy: 0.6196

Epoch 22/50

521/521 [==========] - 1s 2ms/step - loss: 0.9900 - accuracy: 0.6263 - val_loss: 1.0042 - val_accuracy: 0.6196

Epoch 23/50

521/521 [==========] - 1s 2ms/step - loss: 0.9894 - accuracy: 0.6260 - val_loss: 1.0046 - val_accuracy: 0.6225

Epoch 24/50

521/521 [==========] - 1s 2ms/step - loss: 0.9892 - accuracy: 0.6272 - val_loss: 1.0029 - val_accuracy: 0.6215

Epoch 25/50

521/521 [==========] - 1s 2ms/step - loss: 0.9890 - accuracy: 0.6268 - val_loss: 1.0040 - val_accuracy: 0.6227

Epoch 26/50

521/521 [==========] - 1s 2ms/step - loss: 0.9891 - accuracy: 0.6266 - val_loss: 1.0039 - val_accuracy: 0.6225

Epoch 27/50

521/521 [==========] - 1s 2ms/step - loss: 0.9887 - accuracy: 0.6258 - val_loss: 1.0024 - val_accuracy: 0.6215

Epoch 28/50

521/521 [==========] - 1s 2ms/step - loss: 0.9888 - accuracy: 0.6270 - val_loss: 1.0033 - val_accuracy: 0.6225

Epoch 29/50

521/521 [==========] - 1s 2ms/step - loss: 0.9888 - accuracy: 0.6265 - val_loss: 1.0035 - val_accuracy: 0.6227

Epoch 30/50

521/521 [==========] - 1s 2ms/step - loss: 0.9888 - accuracy: 0.6259 - val_loss: 1.0021 - val_accuracy: 0.6227

Epoch 31/50

521/521 [==========] - 1s 2ms/step - loss: 0.9886 - accuracy: 0.6255 - val_loss: 1.0046 - val_accuracy: 0.6227

Epoch 32/50

521/521 [==========] - 1s 2ms/step - loss: 0.9881 - accuracy: 0.6266 - val_loss: 1.0039 - val_accuracy: 0.6203

Epoch 33/50

521/521 [==========] - 1s 2ms/step - loss: 0.9878 - accuracy: 0.6268 - val_loss: 1.0029 - val_accuracy: 0.6210

Epoch 34/50

521/521 [==========] - 1s 2ms/step - loss: 0.9875 - accuracy: 0.6267 - val_loss: 1.0023 - val_accuracy: 0.6227

Epoch 35/50

521/521 [==========] - 1s 2ms/step - loss: 0.9874 - accuracy: 0.6265 - val_loss: 1.0015 - val_accuracy: 0.6227

Epoch 36/50

521/521 [==========] - 1s 2ms/step - loss: 0.9875 - accuracy: 0.6273 - val_loss: 1.0010 - val_accuracy: 0.6222

Epoch 37/50

521/521 [==========] - 1s 2ms/step - loss: 0.9869 - accuracy: 0.6263 - val_loss: 1.0035 - val_accuracy: 0.6222

Epoch 38/50

521/521 [==========] - 1s 2ms/step - loss: 0.9876 - accuracy: 0.6255 - val_loss: 1.0020 - val_accuracy: 0.6222

Epoch 39/50

521/521 [==========] - 1s 2ms/step - loss: 0.9863 - accuracy: 0.6268 - val_loss: 1.0010 - val_accuracy: 0.6227

Epoch 40/50

521/521 [==========] - 1s 2ms/step - loss: 0.9865 - accuracy: 0.6257 - val_loss: 1.0025 - val_accuracy: 0.6227

Epoch 41/50

521/521 [==========] - 1s 2ms/step - loss: 0.9860 - accuracy: 0.6262 - val_loss: 1.0000 - val_accuracy: 0.6225

Epoch 42/50

521/521 [==========] - 1s 2ms/step - loss: 0.9863 - accuracy: 0.6265 - val_loss: 0.9997 - val_accuracy: 0.6225

Epoch 43/50

521/521 [==========] - 1s 2ms/step - loss: 0.9859 - accuracy: 0.6272 - val_loss: 1.0010 - val_accuracy: 0.6225

Epoch 44/50

521/521 [==========] - 1s 2ms/step - loss: 0.9862 - accuracy: 0.6266 - val_loss: 1.0018 - val_accuracy: 0.6215

Epoch 45/50

521/521 [==========] - 1s 2ms/step - loss: 0.9859 - accuracy: 0.6273 - val_loss: 1.0006 - val_accuracy: 0.6225

Epoch 46/50

521/521 [==========] - 1s 2ms/step - loss: 0.9855 - accuracy: 0.6260 - val_loss: 1.0022 - val_accuracy: 0.6215

Epoch 47/50

521/521 [==========] - 1s 2ms/step - loss: 0.9853 - accuracy: 0.6260 - val_loss: 0.9997 - val_accuracy: 0.6225

Epoch 48/50

521/521 [==========] - 1s 2ms/step - loss: 0.9854 - accuracy: 0.6266 - val_loss: 1.0006 - val_accuracy: 0.6218

Epoch 49/50

521/521 [==========] - 1s 2ms/step - loss: 0.9850 - accuracy: 0.6268 - val_loss: 1.0001 - val_accuracy: 0.6215

Epoch 50/50

521/521 [==============================] - 1s 2ms/step - loss: 0.9850 - accuracy: 0.6273 - val_loss: 1.0008 - val_accuracy: 0.6225

131/131 [==============================] - 0s 1ms/step

Matriz de Confusión:

[[913 0 278 145]

[230 0 305 78]

[105 0 1353 120]

[75 0 236 326]]

Sensibilidad: 0.5131429056368761

Especificidad: 0.5131429056368761

Precisión: 0.45008022691991567

Matriz de Confusión (Porcentaje):

[[**68.33832335** 0. 20.80838323 10.85329341]

[37.52039152 **0.** 49.75530179 12.72430669]

[6.6539924 0. **85.74144487** 7.60456274]

[11.77394035 0. 37.04866562 **51.17739**03]]

La precisión es solo de un 45% de manera global. La matriz de confusión en porcentaje para cada categoría de fuente de energía, en concreto la primera que es uso de ninguna fuente de energía para calentar se predice de manera correcta en un 68.33%. el uso de electricidad en un 0%, en un 86.74% el uso del gas y en un 51,17% el uso de carbón y energías más contaminantes.

Para intentar mejorar el modelo cambiamos el ajuste del número de epochs que cambiamos a 30 y comprobamos de nuevo el resultado. Mostramos el código en Python con el ajuste.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.metrics import confusion_matrix, accuracy_score,
precision_score, recall_score
```

```
# Lee la base de datos desde un archivo Excel
ruta_archivo_excel = 'c:/Users/Uned/proyecto/Basedatos2019.xls'
df = pd.read_excel(ruta_archivo_excel)

# Selecciona las columnas deseadas (excluyendo la variable objetivo)
variables_explicativas = df[['NATIONALITY', 'GENDER', 'TEMPERATURE',
'buildingage', 'dwellingtype', 'municipio', 'Age', 'educ', 'LINARES']]
variable_objetivo = df['ENERCAL2']

# Codificación one-hot para variables categóricas
variables_explicativas = pd.get_dummies(variables_explicativas,
columns=['NATIONALITY', 'GENDER', 'TEMPERATURE', 'buildingage',
'dwellingtype', 'municipio', 'Age', 'educ', 'LINARES'])

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test =
train_test_split(variables_explicativas, variable_objetivo,
test_size=0.2, random_state=42)

# Codificar la variable objetivo si es necesario
le = LabelEncoder()
y_train_encoded = le.fit_transform(y_train)
y_test_encoded = le.transform(y_test)

# Define el modelo
modelo = Sequential()
modelo.add(Dense(10, input_dim=variables_explicativas.shape[1],
activation='relu'))
modelo.add(Dense(4, activation='softmax'))

# Compila y entrena el modelo
modelo.compile(loss='sparse_categorical_crossentropy', optimizer='adam',
metrics=['accuracy'])
modelo.fit(X_train, y_train_encoded, epochs=30, batch_size=32,
validation_data=(X_test, y_test_encoded))

# Predicciones en el conjunto de prueba
y_pred = modelo.predict(X_test)
y_pred_labels = np.argmax(y_pred, axis=1)

# Calcula la matriz de confusión
conf_mat = confusion_matrix(y_test_encoded, y_pred_labels)

# Calcula la sensibilidad y especificidad
sensibilidad = recall_score(y_test_encoded, y_pred_labels,
average='macro')
```

```
especificidad = recall_score(y_test_encoded, y_pred_labels, pos_label=0,
average='macro')

# Calcula la precisión
precision = precision_score(y_test_encoded, y_pred_labels,
average='macro')

# Imprime los resultados
print("Matriz de Confusión:")
print(conf_mat)
print("Sensibilidad:", sensibilidad)
print("Especificidad:", especificidad)
print("Precisión:", precision)

# Calcula la matriz de confusión en porcentaje
conf_mat_porcentaje = conf_mat.astype('float') / conf_mat.sum(axis=1)[:,
np.newaxis] * 100

# Imprime la matriz de confusión en porcentaje
print("Matriz de Confusión (Porcentaje):")
print(conf_mat_porcentaje)
```

Con estos nuevos ajustes, observamos como el modelo mejora notablemente. Mostramos la salida de los mismos.

521/521 [==============================] - 2s 2ms/step - loss: 1.1248 - accuracy: 0.5516 - val_loss: 1.0276 - val_accuracy: 0.6110

Epoch 2/30

521/521 [==============================] - 1s 2ms/step - loss: 1.0075 - accuracy: 0.6237 - val_loss: 1.0139 - val_accuracy: 0.6191

Epoch 3/30

521/521 [==============================] - 1s 2ms/step - loss: 1.0012 - accuracy: 0.6257 - val_loss: 1.0083 - val_accuracy: 0.6206

Epoch 4/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9984 - accuracy: 0.6254 - val_loss: 1.0067 - val_accuracy: 0.6225

Epoch 5/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9980 - accuracy: 0.6260 - val_loss: 1.0052 - val_accuracy: 0.6230

Epoch 6/30

521/521 [==========] - 1s 2ms/step - loss: 0.9972 - accuracy: 0.6264 - val_loss: 1.0059 - val_accuracy: 0.6230

Epoch 7/30

521/521 [==========] - 1s 2ms/step - loss: 0.9965 - accuracy: 0.6261 - val_loss: 1.0078 - val_accuracy: 0.6227

Epoch 8/30

521/521 [==========] - 1s 2ms/step - loss: 0.9962 - accuracy: 0.6266 - val_loss: 1.0034 - val_accuracy: 0.6225

Epoch 9/30

521/521 [==========] - 1s 2ms/step - loss: 0.9962 - accuracy: 0.6264 - val_loss: 1.0041 - val_accuracy: 0.6230

Epoch 10/30

521/521 [==========] - 1s 2ms/step - loss: 0.9960 - accuracy: 0.6265 - val_loss: 1.0040 - val_accuracy: 0.6230

Epoch 11/30

521/521 [==========] - 1s 2ms/step - loss: 0.9954 - accuracy: 0.6267 - val_loss: 1.0051 - val_accuracy: 0.6222

Epoch 12/30

521/521 [==========] - 1s 2ms/step - loss: 0.9964 - accuracy: 0.6263 - val_loss: 1.0040 - val_accuracy: 0.6222

Epoch 13/30

521/521 [==========] - 1s 2ms/step - loss: 0.9954 - accuracy: 0.6260 - val_loss: 1.0040 - val_accuracy: 0.6230

Epoch 14/30

521/521 [==========] - 1s 2ms/step - loss: 0.9952 - accuracy: 0.6266 - val_loss: 1.0050 - val_accuracy: 0.6218

Epoch 15/30

521/521 [==========] - 1s 2ms/step - loss: 0.9952 - accuracy: 0.6264 - val_loss: 1.0040 - val_accuracy: 0.6230

Epoch 16/30

521/521 [==========] - 1s 2ms/step - loss: 0.9949 - accuracy: 0.6260 - val_loss: 1.0057 - val_accuracy: 0.6215

Epoch 17/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9949 - accuracy: 0.6263 - val_loss: 1.0077 - val_accuracy: 0.6218

Epoch 18/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9951 - accuracy: 0.6267 - val_loss: 1.0048 - val_accuracy: 0.6234

Epoch 19/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9945 - accuracy: 0.6264 - val_loss: 1.0045 - val_accuracy: 0.6220

Epoch 20/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9945 - accuracy: 0.6261 - val_loss: 1.0099 - val_accuracy: 0.6215

Epoch 21/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9945 - accuracy: 0.6260 - val_loss: 1.0063 - val_accuracy: 0.6222

Epoch 22/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9945 - accuracy: 0.6264 - val_loss: 1.0046 - val_accuracy: 0.6218

Epoch 23/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9939 - accuracy: 0.6260 - val_loss: 1.0056 - val_accuracy: 0.6222

Epoch 24/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9943 - accuracy: 0.6268 - val_loss: 1.0090 - val_accuracy: 0.6196

Epoch 25/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9937 - accuracy: 0.6260 - val_loss: 1.0047 - val_accuracy: 0.6222

Epoch 26/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9942 - accuracy: 0.6262 - val_loss: 1.0038 - val_accuracy: 0.6225

Epoch 27/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9936 - accuracy: 0.6266 - val_loss: 1.0071 - val_accuracy: 0.6237

Epoch 28/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9934 - accuracy: 0.6245 - val_loss: 1.0070 - val_accuracy: 0.6198

Epoch 29/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9930 - accuracy: 0.6268 - val_loss: 1.0028 - val_accuracy: 0.6246

Epoch 30/30

521/521 [==============================] - 1s 2ms/step - loss: 0.9923 - accuracy: 0.6261 - val_loss: 1.0034 - val_accuracy: 0.6237

131/131 [==============================] - 0s 1ms/step

Matriz de Confusión:

[[894 0 279 163]

[227 1 311 74]

[98 0 1377 103]

[60 0 252 325]]

Sensibilidad: 0.513405163448539

Especificidad: 0.513405163448539

Precisión: 0.7020637956599556

Matriz de Confusión (Porcentaje):

[[**66.91616766** 0. 20.88323353 12.2005988]

[37.03099511 **0.16313214** 50.73409462 12.07177814]

[6.2103929 0. **87.26235741** 6.52724968]

[9.41915228 0. 39.56043956 **51.02040816**]]

La precisión sube al 70,01% y analizamos el porcentaje de precisión por categoría. Ahora el porcentaje correctamente de la categoría no uso de ninguna fuente de energía asciende al 66,91%, el de electricidad mejora hasta el 16,31%, al 87,26% el de gas y el 51,02% el de carbón y fuentes más contaminantes.

Con el siguiente código obtendremos la predicción de la fuente de energía conforme a las variables explicativas.

```
# Supongamos que tienes una nueva observación en el formato adecuado
nueva_observacion = pd.DataFrame({
    'NATIONALITY': ['valor_nacionalidad'],
    'GENDER': ['valor_genero'],
    'TEMPERATURE': ['valor_temperatura'],
    'buildingage': ['valor_edad_edificio'],
    'dwellingtype': ['valor_tipo_vivienda'],
```

```
    'municipio': ['valor_municipio'],
    'Age': ['valor_edad'],
    'educ': ['valor_educacion'],
    'LINARES': ['valor_linares']
})

# Aplica la codificación one-hot
nueva_observacion_encoded = pd.get_dummies(nueva_observacion,
columns=['NATIONALITY', 'GENDER', 'TEMPERATURE', 'buildingage',
'dwellingtype', 'municipio', 'Age', 'educ', 'LINARES'])

# Realiza la predicción
prediccion = modelo.predict(nueva_observacion_encoded)

# Convierte las probabilidades predichas a etiquetas
etiqueta_predicha = np.argmax(prediccion, axis=1)

# Imprime la etiqueta predicha
print("Etiqueta Predicha:", etiqueta_predicha)
```

Análisis de la predicción del modelo de deep learning de perceptrón multicapa

El modelo ha mejorado la precisión global del modelo respecto al modelo probit. En ambos casos el principal problema surge al predecir la electricidad. Para intentar ver si logramos mejorar el modelo usaremos otra técnica: el random forest.

Estimación del modelo de machine learning de random forest (bosques aleatorios): Código en Python y análisis del mismo

Para estimarlo usaremos en Python el siguiente código

```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, recall_score,
precision_score

# Lee la base de datos desde un archivo Excel
ruta_archivo_excel = 'c:/Users/Uned/proyecto/Bd2019.xls'
df = pd.read_excel(ruta_archivo_excel)

# División de datos en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(df[df.columns[:-1]],
df["ENERCAL2"], test_size=0.2, random_state=42)

# Crear un modelo de bosque aleatorio
bosque = RandomForestClassifier(n_estimators=100, criterion="gini",
max_features="sqrt", bootstrap=True, max_samples=2/3, oob_score=True)
```

```
# Entrenar el modelo
bosque.fit(X_train, y_train)

# Evaluar el modelo en el conjunto de entrenamiento
train_accuracy = bosque.score(X_train, y_train)
print("Precisión en el conjunto de entrenamiento:", train_accuracy)

# Evaluar el modelo en el conjunto de prueba
test_accuracy = bosque.score(X_test, y_test)
print("Precisión en el conjunto de prueba:", test_accuracy)

# Hacer predicciones en el conjunto de prueba
y_pred = bosque.predict(X_test)

# Crear una matriz de confusión
conf_mat = confusion_matrix(y_test, y_pred)
print("Matriz de Confusión:")
print(conf_mat)

# Calcular sensibilidad, especificidad y precisión para cada clase
sensibilidades = []
especificidades = []
precisiones = []

for i in range(bosque.n_classes_):
    tp = conf_mat[i, i]
    fn = conf_mat[i, :].sum() - tp
    fp = conf_mat[:, i].sum() - tp
    tn = conf_mat.sum() - tp - fn - fp

    sensibilidad = recall_score(y_test == bosque.classes_[i], y_pred ==
bosque.classes_[i])
    especificidad = tn / (tn + fp) if (tn + fp) > 0 else 0
    precision = precision_score(y_test == bosque.classes_[i], y_pred ==
bosque.classes_[i])

    sensibilidades.append(sensibilidad)
    especificidades.append(especificidad)
    precisiones.append(precision)

# Imprimir sensibilidad, especificidad y precisión para cada clase
for clase, sens, espec, prec in zip(bosque.classes_, sensibilidades,
especificidades, precisiones):
    print(f"Clase {clase}: Sensibilidad = {sens:.4f}, Especificidad =
{espec:.4f}, Precisión = {prec:.4f}")
```

Vamos a obtener con este código tanto la precisión por categoría de la variable dependiente como la sensibilidad y especificidad de cada categoría. La sensibilidad es la capacidad que tiene el modelo para predecir y discriminar los casos de esa categoría y especificidad es la capacidad que tiene el modelo para predecir y discriminar los casos que no pertenecen a esa categoría.

De acuerdo a este código, esta sería la salida obtenida:

Precisión en el conjunto de entrenamiento: 0.6488320422746652

Precisión en el conjunto de prueba: 0.6143131604226705

Matriz de Confusión:

[[930 29 256 121]

 [224 19 296 74]

 [133 21 1331 93]

 [106 8 245 278]]

Clase 1: Sensibilidad = 0.6961, Especificidad = 0.8363, Precisión = 0.6676

Clase 2: Sensibilidad = 0.0310, Especificidad = 0.9837, Precisión = 0.2468

Clase 3: Sensibilidad = 0.8435, Especificidad = 0.6918, Precisión = 0.6255

Clase 4: Sensibilidad = 0.4364, Especificidad = 0.9183, Precisión = 0.4912

El modelo mejora la predicción del probit que es solo del 56% aunque es peor que la del perceptrón multicapa que llegaba a una precisión superior al 70%.

De nuevo la categoría que se predice peor es la relativa a la electricidad. La razón la observamos al estudiar los residuos estandarizados y las tablas de contingencia. No existe un patrón claro de predicción para esta variable. Es más usada en las CCAA de mayor temperatura en invierno y a la vez en hogares pequeños pero es muy difícil encontrar modelos que predigan correctamente cuando el usuario empleará electricidad a diferencia de las otras tres fuentes de energía en las que sí existe un patrón claro de predicción. A partir de este modelo es sencillo poder usarlo y obtener predicciones de un modo operativo con el siguiente comando de Python.

```
# Supongamos que tienes un nuevo caso representado por un DataFrame llamado
'nuevo_caso'

nuevo_caso = pd.DataFrame({

   'variable1': valor_variable1,

   'variable2': valor_variable2,

   # ... (proporciona valores para todas las características)

})

# Realiza la predicción para el nuevo caso
```

```
prediccion_nuevo_caso = bosque.predict(nuevo_caso)

# Imprime la predicción

print("Predicción para el nuevo caso:", prediccion_nuevo_caso)
```

Esto se podría automatizar en cualquier web usando herramientas como Flask.

Conclusiones

- El objetivo era el lograr modelos que permitieran explicar y predecir las fuentes de energía a emplear por el usuario
- Esto tiene mucha relevancia tanto para las políticas públicas energéticas con el objetivo de poder dirigirse tanto a los colectivos como determinar las subvenciones que más éxito tendrán.
- Asimismo determinar modelos y averiguar las variables que permiten predecir las fuentes de energía es muy importante para todas las empresas vinculadas al sector energético.
- Se han determinado cuatro niveles de fuentes de energía para calentar el hogar: el no uso de ninguna, la electricidad, el gas y el carbón junto con otras fuentes contaminantes.
- Se han analizado todas las variables que de acuerdo a la literatura pueden tener vinculación para el caso de España por tenerlo a nivel mundial: edad del sustentador principal, el género y la nacionalidad del mismo, la tipología de la vivienda, la tipología del hogar, la temperatura en la CCAA donde se ubica la vivienda, si es rural o urbana, el número de miembros del hogar, y la educación del sustentador principal.
- Se han realizado cuatro tipos diferentes de análisis: tablas de contingencia y estudio de residuos estandarizados con el objetivo de estudiar la relación de las variables explicada y las explicativas.
- A un nivel de explicación encontramos que las fuentes de energía más contaminantes se correlacionan en mayor medida con hogares rurales, de temperatura fría en la Comunidad y de educación elemental.
- El gas está más unido a CCAA de temperaturas frías, vivienda nueva y urbana, educación superior y mujeres.
- La electricidad es la fuente de peor predicción. Está especialmente vinculada con las CCAA de temperatura elevada y hogares con un solo miembro. Es la fuente de energía de más difícil predicción.
- Para la predicción mediante modelos hemos usado un modelo probit multinomial, un modelo de IA del campo del Deep learning: el perceptrón multicapa y otro de la esfera del machine learning: el modelo de bosques aleatorios o random forest -en inglés-.
- Respecto a las semejanzas de las tres técnicas encontramos
  - ✓ Probit multinomial y Random Forest se utilizan para problemas de clasificación, mientras que el perceptrón multicapa también se puede utilizar para clasificación.
  - ✓ Random Forest y perceptrón multicapa son modelos más complejos y flexibles en comparación con el probit multinomial, ya que este último asume una relación lineal entre las características y la variable dependiente.
  - ✓ No Linealidad: Tanto Random Forest como el perceptrón multicapa pueden aprender relaciones no lineales en los datos debido a sus capacidades para modelar interacciones complejas.

- En relación a las diferencias de las tres técnicas
  - ✓ Probit multinomial es un modelo paramétrico basado en regresión, mientras que Random Forest y el perceptrón multicapa son modelos no lineales más flexibles y no paramétricos.
  - ✓ Probit multinomial proporciona coeficientes que pueden interpretarse directamente en términos de la relación entre las características y la variable dependiente. Random Forest y perceptrón multicapa son modelos más difíciles de interpretar debido a su complejidad.
  - ✓ Random Forest utiliza un enfoque de conjunto combinando múltiples árboles, mientras que el perceptrón multicapa utiliza múltiples neuronas en capas ocultas para aprender representaciones más complejas de los datos.
- De los tres modelos es el del perceptrón multicapa el que logra una precisión más elevada que llega a un 70%. Tanto el random forest como el modelo de perceptrón logran una mayor precisión que el del probit multinomial
- Gracias a estas técnicas de IA hemos logrado una alta predicción y también un análisis causal de las principales fuentes de energía habiendo realizado un análisis por cada una de ellas. De todas es la de la electricidad la que logra menor nivel de precisión y la del gas la que se predice mejor y en la que es más fácil explicar sus relaciones causales.

Bibliografía

ADL (2023). Demanda de calor en los hogares: Una transición energética eficiente. ADL and Fundación Naturgy.

Alvarez, R., Naagler, J. (1998). When Politics and Models Collide: Estimating Models of Multiparty Elections. American Journal of Political Science, 42(1), 55-96.

Bai, C., Zhan, J., Wang, H., Yang, Z., Liu, H., Liu, W., Wang, C., Chu, X., Teng, Y. (2023). Energy Policy, 178, 113617

Barrella, R., Linares, J., Romero, J., Arenas, E., Centeno, E. (2021). Does cash money solve energy poverty? Assessing the impact of household heating allowances in Spain. Energy Research & Social Science, 80, 102216

Belaïd F., Flambard, V. (2023). Impacts of income poverty and high housing costs on fuel poverty in Egypt: An empirical modeling approach. Energy Policy, 175, 113450

Belaïd F. (2022). Mapping and understanding the drivers of fuel poverty in emerging economies: The case of Egypt and Jordan. Energy Policy, 162, 112775

Bertelsen, N., Vad Mathiesen, B. (2020). EU-28 Residential Heat Supply and Consumption: Historical Development and Status. Energies, 13(8), 1894.

Braun, F. (2010). Determinants of households' space heating type: a discrete choice analysis for German households. Energy Policy, 38(10), 5493–5503.

Burguillo, M., Barisone, M., Juez-Martel, P. 2022. Which cooking and heating fuels are more likely to be used energy-poor households? Exploring energy and fuel poverty in Argentina. Energy Research & Social Sciences, 87, 102481.

Cameron, C., Trivedi, P. (2005). Microeconomics: Methods and Applications. Cambridge University Press. New York.

Charlier, D., Legendre, B. Ricci, O. (2021). Measuring fuel poverty in tropical territories: A latent class model. World Development, 140, pp.105278.

Chen, Q. (2021). District or distributed space heating in rural residential sector? Empirical evidence from a discrete choice experiment in South China. Energy Policy, 148, 111937

Costa-Campi, T., Jové-Llopis, E., Trujillo-Baute, E. (2019) Energy poverty in Spain: an income approach analysis, Energy Sources, Part B: Economics, Planning, and Policy, 14(7-9), 327-340

Couture, S., Garcia, S., Reynaud, A. (2012). Household energy choices and fuelwood consumption: an econometric approach using French data. Energy Econ., 34, 1972–1981.

Decker, T., Menrad, K. (2015). House owners' perceptions and factors influencing their choice of specific heating systems in Germany. Energy Policy, 85, 150–161

Duscha, V., Wachsmuth, J., Eckstein, J., Pfluger, B. 2019. GHG-neutral EU2050 – a 18 scenario of an EU with net-zero greenhouse gas emissions and its implications. German Environmental Agency.

Eurostat (2023a). Energy consumption in households https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households#Energy_consumption_in_households_by_type_of_end-use

Eurostat (2023b). Inability to keep home adequately warm - EU-SILC survey. https://ec.europa.eu/eurostat/databrowser/view/ILC_MDES01/default/table?lang=en

EEA (2023). Decarbonising heating and cooling — a climate imperative. https://www.eea.europa.eu/publications/decarbonisation-heating-and-cooling

Feng, T., Du, H., Coffman, D., Qu, A., Dong, Z. (2021). Clean heating and heating poverty: A perspective based on cost-benefit analysis. Energy Policy, 152, 112205

Greene, W. (2000). Econometric Analysis. Fourth edition ed. Prentice Hall. Upper Saddle River, NJ.

Hills, J. (2012). Getting the Measure of Fuel Poverty, Centre for Analysis of Social Exclusion, The London School of Economics and Political Science, London.

IDAE (2023). Consumo por usos del sector residencial. https://informesweb.idae.es/consumo-usos-residencial/descargas.php

International Energy Agency (IEA) 2020. World Energy Balances 2020. Paris

IPCC (2022). Climate Change 2022. Mitigation of Climate Change. Working Group III contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. https://report.ipcc.ch/ar6/wg3/IPCC_AR6_WGIII_Full_Report.pdf

Jumbe, C., Angelsen, A. (2011). Modeling choice of fuelwood source among rural households in Malawi: A multinomial probit analysis. Energy Economics, 33, 732–738

Kato, E., Kurosawa, A. 2019. Evaluation of Japanese energy system toward 2050 with TIMES-Japan - Deep decarbonization pathways. Energy Procedia, 158, 4141–4146.

Kropko, J. (2008). Choosing between multinomial logit and multinomial probit models for analysis of unordered choice data. PhD thesis. University of North Carolina at Chapel Hill.

Laureti, T., Secondi, L. (2012). Determinants of Households' Space Heating type and Expenditures in Italy. Int. J. Environ. Res., 6(4), 1025-1038,

Liao, H.-C., Chang, T.-F. (2002). Space-heating and water-heating energy demands of the aged in the US. Energy Economics 24 (3), 267–284.

Lopez-Bernabé, E., Linares, P., Galarraga, I. (2022). Energy-efficiency policies for decarbonising residential heating in Spain: A fuzzy cognitive mapping approach. Energy Policy, 171, 113211

Martinopoulos, G., Papakostas, K., Papadopoulos, A. (2018). A comparative review of heating systems in EU countries, based on efficiency and fuel cost. Renewable and Sustainable Energy Reviews, 90, 687–699.

Mensah, J., Adu, G. 2015. An empirical analysis of household energy choice in Ghana. Renewable and Sustainable Energy Reviews, 51, 1402–1411

Michelsen, C., Madlener, R. (2012). Homeowners' preferences for adopting innovative residential heating systems: A discrete choice analysis for Germany. Energy Economics, 34, 1271–1283

Michelsen, C., Madlener, R. (2013). Motivational factors influencing the homeowners' decisions between residential heating systems: An empirical analysis for Germany. Energy Policy, 57, 221–233

Miyanaga, S. (2021). What is 'the energy trilemma' and what do we need to do about it? Spectra 2021-02-02 https://spectra.mhi.com/what-is-the-energy-trilemma-and-what-do-we-need-to-do-about-it

Nesbakken, R. (1999). Price sensitivity of residential energy consumption in Norway. Energy Economics, 21(6), 493–515.

Newell, R.G., Pizer, W.A. (2008). Carbon mitigation costs for the commercial building sector: discrete-continous choice analysis of multifuel energy demand. Resource and Energy Economics, 30(4), 527–539.

Ortega-Izquierdo, M., Paredes-Salvador, A., Montoya-Rasero, C. (2019). Analysis of the decision making factors for heating and cooling systems in Spanish households. Renewable and Sustainable Energy Reviews, 100, 175–185

Romero, J.C., Linares, P., López, X. (2018). The policy implications of energy poverty indicators. Energy Policy, 115, Pages 98-108.

Rosenbloom, D., Markard, J., Geels, F., Fuenfschilling, L. (2020). Why carbon pricing is not sufficient to mitigate climate change—and how "sustainability transition policy" can help. Proceedings of the National Academy of Sciences of the United States of America, 117, 8664–8668

Rosenow, J., Thomas, S., Gibb, D., Baetens, R., De Brouwer, A., Cornillie, J. (2023). Clean heating: Reforming taxes and levies on heating fuels in Europe. Energy Policy, 173, 113367

Selectra (2013). Les inégalités dans la consommation d'énergie. https://selectra.info/energie/actualites/expert/inegalites-consommation-energie#_ftn1.

Sovacool, B., Cabeza, L., Pisello, A., Fronzetti, A., Madani, M., Dawoud, B., Martiskainen, M. (2021). Decarbonizing household heating: Reviewing demographics, geography and low-carbon practices and preferences in five European countries. Renewable and Sustainable Energy Reviews 139, 110703.

Sunderland, L., Gibb, D. (2022). Taking the burn out of heating for low-income households. Regulatory assistance project. Brussels (Belgium).

Siksnelyte-Butkiene, I., Streimikiene, D., Lekavicius, V., Balezentis, T. (2021). Energy poverty indicators: A systematic literature review and comprehensive analysis of integrity. Sustainable Cities and Society, 67, 102756

Tinbergen, J. (1952). On the Theory of Economic Policy. Amsterdam: North-Holland.

Trading Economics (2023). Markets. http://tradingeconomics.com

Vaage, K. (2000). Heating technology and energy use: a discrete continuous choice approach to Norwegian household energy demand. Energy Economics, 22 (6), 649–666.

Walled, K., Mirza, F. (2023). Examining fuel choice patterns through household energy transition index: an alternative to traditional energy ladder and stacking models. Environment, Development and Sustainability, 25, 6449–6501

Wang, Z., Li, C., Cui, C., Liu, H., Cai, B. (2019). Cleaner heating choices in northern rural China: household factors and the dual substitution policy. J. Environ. Manag., 249, 109433.