UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

ESCUELA POLITÉCNICA SUPERIOR DE ELCHE

GRADO EN INGENIERÍA INFORMÁTICA EN
TECNOLOGÍAS DE LA INFORMACIÓN

"El voto español: análisis y predicción del comportamiento electoral en la sociedad actual"

TRABAJO FIN DE GRADO

Junio - 23

AUTOR: Vicente Candela Pérez
DIRECTOR/ES: Lidia Ortiz Henarejos
Jesús Javier Rodríguez Sala

AGRADECIMIENTOS

Me gustaría agradecer, en primer lugar, a mi familia, tanto consanguínea como política, por haberme apoyado durante toda la carrera y animado en los momentos más duros de ella.

En segundo lugar, quisiera agradecer a los profesores de la carrera la formación ofrecida a lo largo de estos años, consiguiendo superar las adversidades presentadas por el COVID-19, lo cual me ha permitido adquirir muchos conocimientos, tanto académicos como personales, que son igualmente importantes para la vida, tanto personal como profesional.

En tercer y último lugar, me gustaría hacer una mención especial a Mari Carmen Pérez y Vicente Candela, mis padres, a quienes les debo todo lo que he conseguido y conseguiré, gracias a su cariño y amor incondicional.

RESUMEN

El objetivo de este TFG es describir y predecir, mediante algoritmos de minería de datos, a los distintos votantes de la sociedad española en función del partido político al que votan.

Para lograr dicho objetivo se ha realizado un análisis descriptivo para analizar las distintas diferencias y similitudes que existen entre los distintos electores de los partidos políticos, para, posteriormente, realizar la elaboración de un árbol de clasificación para predecir al partido que pertenecen estos.

Los datos empleados en este proyecto fueron recogidos por el Centro de Investigación Sociológicas (CIS) donde obtuvieron casi cuatro mil individuos encuestados, lo que ha permitido realizar un análisis con mayor registro y fiabilidad.

En el primer capítulo de este TFG, la *Introducción*, se expone qué es y qué significa un estudio de demoscopia, para poder poner en contexto al lector de los pasos que se van a realizar durante el proyecto. Asimismo, en este punto, se exponen los objetivos del TFG, siendo el fin último la creación e interpretación de un árbol de clasificación capaz de predecir la intención de voto del electorado.

En el siguiente bloque, la *Metodología*, se muestran de manera teórica las técnicas de minería de datos y Machine Learning necesarios para la elaboración del proyecto. El fin de realizar un Análisis de Correspondencias Simple radica en el interés de visualizar la relación existente entre dos variables previamente seleccionadas, con el fin de ver posibles dependencias entre estas. Además, en esta sección, también se exponen los métodos empleados para llevar a cabo diferentes árboles de clasificación, con sus correspondientes matrices de confusión, donde se explica qué significado recibe cada posición de dicha matriz.

En el tercer bloque, *Hipótesis de trabajo*, se realiza una explicación del material utilizado para realizar el proyecto, donde se muestra una breve descripción del tipo de datasets obtenido, el tipo de entorno utilizado para realizar el TFG, el lenguaje empleado para realizarlo y las distintas librerías que han sido necesarias para poder realizar la totalidad del trabajo.

En el siguiente bloque, en los *Resultados*, se muestran todos los análisis que se han ido obteniendo tras la ejecución de los algoritmos expuestos en el bloque 2. En primer lugar, se realiza una descripcion de la muestra y poblacion objetivo a la que estaba enfocada las encuestas del barómetro del CIS, después se explica el tipo de preguntas que componen el cuestionario que se ha utilizado para el estudio. Seguidamente, se expone el tratamiento previo que se ha realizado al dataset para poder trabajar con él. Posteriormente, se realizan representaciones gráficas de los distintos porcentajes obtenidos de los votantes de la muestra. Mencionar que todos los partidos que obtuvieron un porcentaje menor al 3% de

representación fueron agrupados en un solo partido bajo el nombre de “Otro partido”. A continuación, se realiza una representación del voto por sexo y edad del encuestado. Después, se muestran diferentes gráficos que arrojan los votantes del PP y PSOE acerca de diversos aspectos como la valoración de la situación social actual de España, los principales problemas que encuentran los votantes en España, las tendencias religiosas, las franjas salariales que perciben en sus hogares y la fidelidad de voto que tiene cada partido sobre sus votantes. Seguidamente, se realizan diversos Análisis de Correspondencias de distintas variables, para mostrar si estas poseen algún tipo de relación. Y para finalizar el bloque, se realizan distintas selecciones de atributos para distintos árboles de clasificación con el objetivo de predecir la intención de voto en función de las respuestas del encuestado.

Y por último, el gran bloque de este trabajo, las *Conclusiones*. En él se recogen de forma resumida las conclusiones más relevantes obtenidas tras haber realizado el estudio.

La combinación de los partidos políticos PP y PSOE representa casi el 50% de los votos, mientras que el otro 50% se reparte entre el resto de opciones, partidos minoritarios e indecisos obtienen más de un 25% y el resto, se reparte entre los que no votarían, PODEMOS, VOX y los que votan en blanco.

La comparativa del voto en funcion del sexo, indica que la mujer concentra en mayor medida en comparación con el hombre, el voto en el PP y PSOE, mientras que los hombres tienen una mayor frecuencia que estas por el partido de derechas VOX. Además, ellas tienen más clara la decisión del voto que los varones, aunque esta decisión sea no votar. En siguiente lugar, los resultados en función de la edad muestran cómo los votantes más jóvenes se dejan seducir por los partidos más nuevos y emergentes, mientras que la población más veterana tiene un voto más tradicional y recurre a los partidos con más recorrido. En cuanto a la valoración social en España, los votantes de la oposición son bastante críticos con la situación que está afrontando el país en la actualidad, mientras que los votantes del partido que está en el gobierno son más laxos con la situación y valoran en mejor medida la situación social. Los encuestados indican que los principales problemas de España, a pesar de las distintas ideologías, todos coinciden en que son la crisis económica y los problemas políticos en general. El análisis de las tendencias religiosas en función del partido político al que pertenecen, queda reflejado que los votantes de los partidos de derechas profesan una mayor creencia religiosa que los votantes de los partidos de izquierdas. En cuanto a los resultados de las franjas salariales en función del partido político al que votan, estos indican que los votantes de VOX son los que, por lo general, poseen menor capacidad adquisitiva, seguidos del PSOE y luego el PP. La representación de la frecuencia del voto en función del partido al que se vota, los resultados indican que los partidos de derechas tienen un alto porcentaje de votantes momentáneos y muy pocos votantes fieles, mientras que los partidos de izquierdas poseen un alto porcentaje de voto fiel.

A la hora de establecer asociación entre las variables del estudio cabe destacar la relación existente entre el partido político y la preocupación existente por el medio ambiente, donde a

los votantes de derechas les importa poco o nada el medio ambiente, y a los votantes de izquierdas sí que les preocupa más este tema.

Por último, en los árboles de clasificación se muestran cómo a partir de 4 preguntas son capaces de predecir, en un alto porcentaje de acierto, la tendencia ideológica que tiene cada votante en función de las respuestas proporcionadas por las preguntas del cuestionario.

Índice General

Índice Figuras

Índice de tablas

Capítulo 1 - Introducción

En este primer capítulo se describe brevemente el concepto de demoscopia, técnica que se ha utilizado para realizar el presente Trabajo Fin de Grado. Se detalla la motivación del tema elegido, los objetivos que se tratan de conseguir así como las limitaciones que existen a la hora de realizarlo.

1.1.- Demoscopia

La demoscopia [1] es una disciplina que se encarga de estudiar y analizar las opiniones, actitudes y comportamientos de la población a través de encuestas y sondeos. El objetivo principal es obtener información veraz y precisa sobre la opinión pública en diversos campos como el cultural, político, social y económico, entre otros.

La demoscopia se realiza a través de diferentes estrategias y técnicas, entre las que podemos destacar las encuestas telefónicas, las encuestas en línea, las encuestas presenciales, los grupos de difusión y las entrevistas en profundidad. Estas técnicas se aplican en función de los objetivos de la investigación y de los grupos de población que se analicen.

En el ámbito político, la demoscopia se utiliza para conocer la opinión pública sobre los candidatos, partidos políticos, temas de interés público y demás. Este estudio es fundamental para la toma de decisiones de los gobiernos y partidos políticos, ya que permite conocer las preferencias de la población y adaptar sus programas y políticas a las necesidades y demandas de la sociedad.

Elecciones municipales de 1995: comparación entre los datos brutos(intención explícita y extrapolación), los resultados registrados y la estimación anunciada

	Intención explícita	Extrapolación	Resultados	Estimación
PP	21,8	34,9	35,1	36,9
PSOE	19,0	30,4	30,8	28,0
IU	8,4	13,4	11,7	13,3
CiU	3,1	5,0	4,6	5,3
Otros	10,2	16,3	17,8	16,5
Blanco	2,0			
Abstención	4,4			
Ns/Nc	31,1			

Tabla 1.1: Elecciones municipales de 1995. Fuente: Gonzalez[2]

La demoscopia lleva realizándose desde hace bastante tiempo midiendo la opinión pública sobre los diferentes partidos políticos, entre ellos, se encuentra el estudio realizado en 1996 por Gonzalez [2], en el que se muestra una comparación entre los resultados de las elecciones y la estimación que se hizo.

Otro ejemplo de demoscopia, es la realizada en el año 2011, por el Ministerio de medio ambiente y medio rural y marino, acerca de la opinión de los españoles sobre el cambio climático [3]. En el informe se muestra cual es el perfil más común para los negacionistas y los preocupados por el tema.

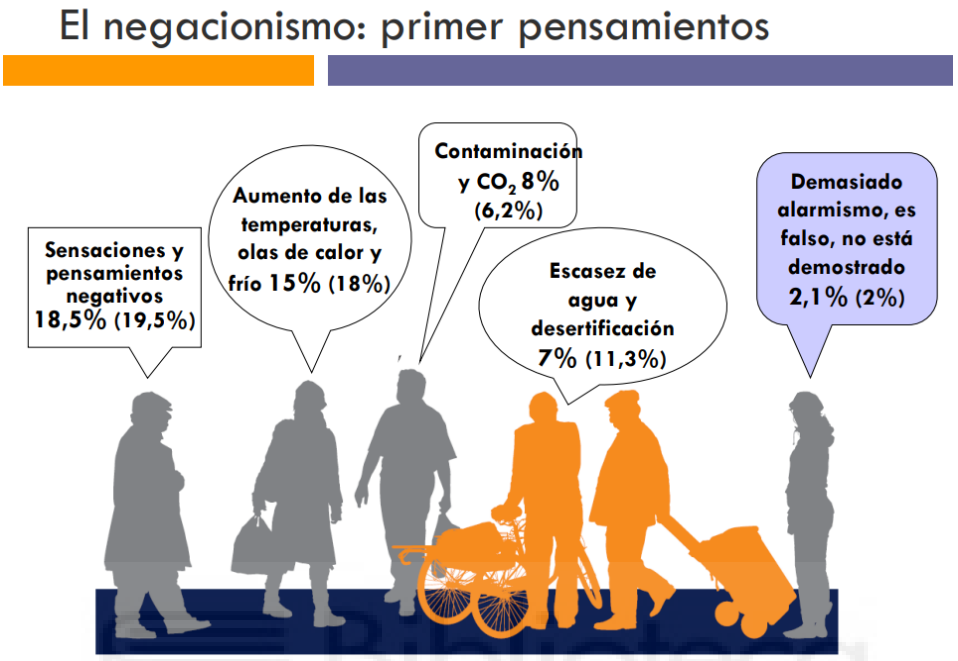

Figura 1.1: El primer pensamiento de los negacionistas.Fuente: El Ministerio de medio ambiente y medio rural y marino

En una sociedad democrática, comprender las preferencias y comportamientos de los votantes es fundamental para los partidos políticos y las estrategias electorales. La capacidad de predecir el voto con precisión permite a los políticos adaptar sus mensajes, enfoques y estrategias de campaña para maximizar su apoyo y aumentar sus posibilidades de éxito en las elecciones.

En este escenario, el uso de técnicas de aprendizaje automático se ha vuelto cada vez más relevante. Permitiendo así comprender mejor el comportamiento de los votantes y realizar predicciones precisas.

En particular, los árboles de clasificación han demostrado ser herramientas efectivas para la predicción del voto. Su capacidad para modelar relaciones complejas entre características y clases objetivo lo convierte en una elección natural para este tipo de análisis. Al construir un árbol de clasificación, se pueden identificar las características más relevantes que influyen en las decisiones de voto y utilizarlas para hacer predicciones precisas sobre las preferencias electorales.

En este proyecto se ha realizado un estudio de demoscopia, haciendo uso de técnicas y algoritmos de aprendizaje automático, sobre los partidos con mayor representación, PP y PSOE, partidos con un bagaje amplio y partidos nuevos, VOX y PODEMOS, los cuales se

sitúan a los extremos de los partidos anteriormente mencionados, uno a la derecha y otro a la izquierda, respectivamente.

El organismo oficial que se encarga a día de hoy de realizar las encuestas de demoscopia es el Centro de Investigaciones Sociológicas, CIS [4], cuya la finalidad es el estudio cientifico de la sociedad española.

1.2.- Justificación del proyecto

La finalidad de este estudio es conocer las preferencias y opiniones de la población sobre temas públicos y políticos poder representarlos de forma visual. La información que proporcionan los ciudadanos puede resultar muy valiosa a la hora de realizar una toma de decisión en el ámbito político, empresarial y económico.

Este TFG puede resultar de interés para un partido político en los siguientes aspectos:

- Conocer las preferencias y opiniones de la población, permitiendo a un partido político conocer las prioridades de la población en relación a temas de interés público. Esta información es fundamental para adaptar las propuestas y políticas del partido a las necesidades y la demanda de la sociedad.

- Evaluar la imagen del partido y sus líderes, ya que esta información es importante para identificar las fortalezas y amenazas, y diseñar estrategias para mejorar la imagen del partido.

- Planificar campañas electorales, de modo que la información obtenida permita conocer la intención del voto de la población y diseñar estrategias que se adapten a sus votantes.

El presente Trabajo Fin de Grado se estructura de la siguiente forma, los objetivos y los límites del proyecto se describen en el presente Capítulo, la metodología del estudio se detalla en el Capítulo 2, la hipótesis del trabajo junto con los materiales empleados se exponen en el Capítulo 3, a continuación se procede a mostrar los resultados arrojados por el estudio realizado en el Capítulo 4, y por último, se realizan las conclusiones obtenidas de los capítulos anteriores en el Capítulo 5.

1.3.- Objetivos

Los objetivos a conseguir mediante la realización del TFG se pueden dividir en objetivos principales y secundarios.

Objetivos principales:

- Describir los rasgos de los votantes en función del partido político al que votan.
- Predecir la intención de voto de la población a través de un breve cuestionario mediante técnicas estadísticas de minería de datos y aprendizaje automático.

Objetivos secundarios:

- Usar el lenguaje de programación Python y Google Colab para el tratamiento y análisis de la base de datos.
- Realizar depuración de la base de datos de un cuestionario mediante técnicas de minería de datos relacionadas con el tratamiento y el preprocesamiento de datasets.
- Describir la muestra, mediante técnicas estadísticas descriptivas y de visualización.
- Describir la representación del voto español, así como la representación por sexo y edad.
- Describir la relación existente entre variables del dataset, relacionadas con la intención de voto, haciendo uso de técnicas de Machine Learning no supervisado.
- Seleccionar los atributos más interesantes para aplicar algoritmos de clasificación.
- Predecir la intención de voto haciendo uso de algoritmos de árboles de clasificación.

1.4.- Límites del proyecto

Queda fuera del proyecto predecir el número de escaños que va a obtener cada partido en las próximas elecciones generales del 23 de julio de 2023.

Las variables tratadas en este trabajo son, en su mayoría, variables categóricas y variables categóricas ordinales, por tanto, una gran limitación de este trabajo es el uso de técnicas y algoritmos que requieran variables continuas.

Capítulo 2 - Métodos y materiales

Al gobierno y a los partidos políticos siempre les ha resultado de interés conocer cuál es la opinión pública de la sociedad. Por eso, gracias a las técnicas de minería de datos y a técnicas estadísticas se puede predecir con mayor facilidad y precisión cuál es esa opinión.

En este apartado se describen los métodos empleados para la realización del estudio, tanto los usados en el preprocesamiento de los datos, como las técnicas estadísticas y los algoritmos empleados en minería de datos para el análisis exhaustivo de dichos datos.

2.1.- Análisis preliminar

El primer paso a la hora de realizar un modelo predictivo es entender el problema que se quiere resolver y conocer cuáles son los datos que hay disponibles. El análisis preliminar [5] es un proceso muy importante para desarrollar un modelo de aprendizaje automático que permita predecir el comportamiento de un evento futuro.

Para elaborar este TFG, se ha hecho uso de los datos del barómetro del CIS realizado durante el mes de febrero de 2023 [6], pero si hubiese que recoger los datos de forma manual se debería plantear cuál es la muestra adecuada a obtener y cuáles son las características que se desean recoger. En este proyecto, al igual que cuando una empresa privada proporciona datos, se deberá comprobar que estos son coherentes.

2.2.- Depuración de los datos

La depuración de datos consiste en realizar la identificación y corrección de errores e inconsistencias, seleccionar los atributos más relevantes y detectar los valores atípicos en los datos, ya que estos pueden afectar en la precisión y eficacia del modelo predictivo. Antes de trabajar sobre el modelo predictivo, se deben revisar los datos disponibles para no incurrir en este tipo de errores. A continuación se describen los procesos de depuración realizados sobre los datos empleados en este TFG.

2.2.1.- Selección de atributos

La selección de atributos en los modelos predictivos es un proceso de identificación y selección de las variables más importantes y relevantes para predecir una respuesta o resultado en un conjunto de datos. El objetivo es reducir el número de variables de entrada en el modelo y mejorar su precisión, eficiencia y generalización.

Este método ayuda a evitar la sobrecarga de información y reducir el ruido y la redundancia en los datos, lo que permite mejorar la calidad de las predicciones y reducir el tiempo de procesamiento. Además, ayuda a la mejora de la interpretación del modelo al identificar las variables más importantes de la predicción.

Existen varios métodos para la selección de atributos [6], que se pueden clasificar en dos categorías: filtros y wrappers. Los filtros son métodos de selección de atributos que se basan en medidas estadísticas o de información para evaluar la relevancia de cada atributo. Los atributos se ordenan según su relevancia y se seleccionan los más importantes.

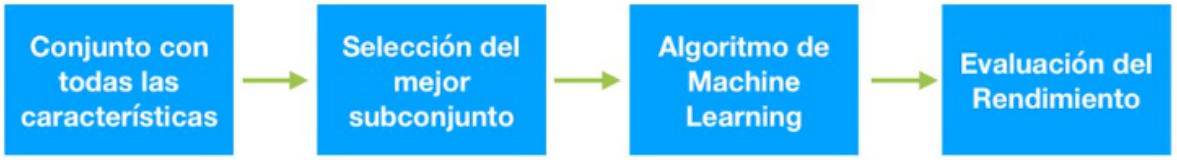

Figura 2.1: Etapas de los métodos de filtro. Fuente: AprendeIA con Ligdi Gonzalez [7]

Los wrappers son métodos que incorporan la selección de atributos directamente en el proceso de aprendizaje del modelo. El modelo se ajusta iterativamente con diferentes combinaciones de atributos de entre todos los disponibles, hasta que se encuentra la mejor de dichas combinaciones.

Figura 2.2: Etapas de los métodos de wrappers. Fuente: AprendeIA con Ligdi Gonzalez [7]

2.2.2.- Tratamiento de valores nulos

El tratamiento de valores nulos se refiere al proceso de manejar y procesar los datos faltantes o nulos en un conjunto de datos estructurados. Un valor nulo es un valor que no está presente en una celda de la tabla de datos. El tratamiento adecuado de los valores nulos es esencial para garantizar la calidad y precisión de los análisis de datos. Existen varias técnicas para tratar los valores nulos de un conjunto de datos [8], y se pueden clasificar en tres categorías principales: eliminación, imputación y predicción.

La eliminación implica eliminar filas o columnas que contienen valores nulos. Esta técnica es útil cuando el número de valores nulos es pequeño y no afecta significativamente a la calidad de los datos restantes. Sin embargo, si la cantidad de los valores nulos es alta, puede resultar una pérdida de información importante.

La imputación implica reemplazar los valores nulos por valores estimados o calculados. Existen varias técnicas de imputación, como la imputación por la media, por la mediana y por la moda. Estas técnicas son útiles cuando los valores nulos son una pequeña fracción del conjunto de datos y se pueden calcular con precisión.

Adicionalmente, se pueden emplear modelos predictivos para estimar los valores nulos en un conjunto de datos. Esta técnica es útil cuando los valores nulos son una gran fracción del conjunto de datos y las técnicas de imputación descritas anteriormente no son adecuadas. La predicción puede ser realizada mediante técnicas como la regresión, el análisis de series de tiempo y el aprendizaje automático.

2.2.3.- Tratamiento de valores anómalos

El tratamiento de valores anómalos [9] es el proceso por el cual se identifican y corrigen valores que son inusualmente altos o bajos en comparación con el resto de los valores en el conjunto de datos. Los valores anómalos también se conocen como valores atípicos (outliers).

A la hora de trabajar con este tipo de datos, se pueden realizar dos tratamientos distintos: corrección o eliminación. La corrección implica corregir los valores anómalos en el conjunto de datos. Las técnicas de corrección incluyen la imputación, la sustitución y la transformación de los valores.

La imputación implica la sustitución con valores estimados o calculados. La sustitución consiste en el reemplazo con valores específicos, como la media o la mediana de los valores no anómalos. La transformación implica la transformación de los valores anómalos a través de técnicas matemáticas, como la transformación logarítmica. La eliminación implica suprimir los valores del conjunto de datos, esta técnica es útil cuando los valores anómalos son una pequeña fracción del conjunto de datos y no afecta significativamente a la calidad de los datos restantes. Sin embargo, si la cantidad es alta, puede implicar una pérdida de información importante.

2.3.- Técnicas de Machine Learning

Las técnicas para la construcción de modelos supervisados y no supervisados [10] se utilizan en el aprendizaje automático para analizar y predecir datos. La elección entre un tipo u otro de técnica dependerá de la naturaleza de los datos disponibles y de cuál es el objetivo del estudio que se desea realizar.

2.3.1.- Modelos no supervisados

Estas técnicas se utilizan para analizar y descubrir patrones de datos no etiquetados, es decir, datos que no tienen una variable objetivo previamente definida. Existen varios tipos de algoritmos de aprendizaje no supervisado, cada uno con diferentes enfoques y técnicas, entre ellas se encuentra el clustering, el análisis de componentes principales, el análisis de correspondencias. A continuación, se describen algunas de las técnicas más empleadas en Machine Learning.

2.3.1.1- Clustering

El clustering [11] es una técnica que se utiliza para agrupar un conjunto de objetos o datos similares en subconjuntos llamados clústers. El objetivo es identificar patrones intrínsecos en los datos y agrupar objetos que comparten características similares, sin tener conocimiento previo de las etiquetas o clases a las que pertenecen.

Existen diversos métodos para realizar agrupación mediante la técnica del clustering. En particular, el algoritmo k-medias realiza una agrupación de los datos en función de la similitud entre ellos. El objetivo es dividir los datos en subconjuntos de datos, de tal forma que los datos dentro de cada grupo sean entre sí similares, pero diferentes entre el resto de datos de los otros grupos.

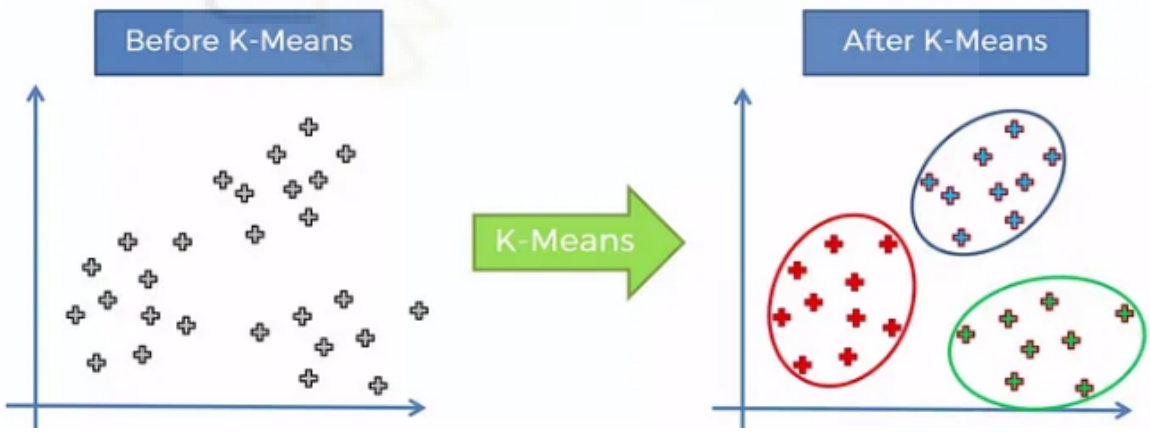

Figura 2.3:Algoritmo de K-means. Fuente: towardsdatascience [12]

El algoritmo utiliza medidas de distancia para determinar qué datos son similares entre sí y los agrupa en clústeres. Existen varios tipos de clustering disponibles, en los que se utilizan diferentes tipos de distancias y cada uno de ellos posee sus fortalezas y sus debilidades.

Figura 2.4: Diferentes distancias en el clustering. Fuente: Keepcoding [13]

El clustering de distancia euclidiana utiliza la distancia lineal que hay entre dos puntos P y Q en un plano n-dimensional. Una de sus ventajas es su intuitividad para calcular la distancia que existe entre dos puntos, pero este método es bastante sensible a valores atípicos.

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (1)$$

Por su parte, el clustering de distancia Manhattan utiliza la distancia calculada de la suma de la diferencia absoluta entre las medidas de todas las dimensiones de dos puntos. Una de sus ventajas es que este método resulta útil para datos con estructuras no lineales y para la detección de clusters de diferentes formas y tamaños, pero este método utiliza muchos más recursos computacionales que el método anterior.

$$d_{Man}(P,Q) = \sum_{i=1}^{n} |p_i - q_i| \quad (2)$$

Finalmente, la distancia Minkowsky es la unión de la distancia euclídea y la distancia Manhattan. Una de sus ventajas es que utiliza una técnica flexible y adaptable que permite manejar diferentes tipos de datos y patrones de distribución, pero este método es el más difícil de interpretar debido a su flexibilidad y la necesidad de ajustar varios parámetros.

$$d_{Min}(P,Q) = \left(\sum_{i=1}^{n} |p_i - q_i|^X \right)^{\frac{1}{X}} \quad (3)$$

2.3.1.2- Reglas de asociación

Este tipo de algoritmos [14] se utilizan para descubrir patrones frecuentes y asociaciones entre variables en un conjunto de datos transaccionales. Estas reglas se utilizan para descubrir relaciones interesantes entre las variables.

El algoritmo de reglas de asociación más conocido es el Apriori que fue propuesto por Rakesh Agrawal en 1994 [15], y busca reglas con una frecuencia mínima de ocurrencias, llamada soporte. La medida de soporte se utiliza para identificar los elementos que aparecen con frecuencia suficiente en las transacciones para ser considerados significativos. Después, se utilizan otras medidas como la confianza y la elevación, para identificar las reglas más relevantes y útiles.

La confianza sirve para medir la probabilidad de que un suceso ocurra dado que se ha producido su antecedente. Se calcula dividiendo el número de veces que se observa la consecuencia y el antecedente juntos por el número de veces que se observa el antecedente.

La elevación mide la importancia relativa de la relación entre el antecedente y la consecuencia. Se calcula dividiendo la confianza de la regla por la frecuencia de la consecuencia.

2.3.1.3- Análisis de correspondencias simples

El análisis de correspondencias simples (ACS) [16] es una técnica estadística utilizada para analizar y visualizar las relaciones entre variables categóricas en un conjunto de datos. Se utiliza principalmente en campos como la sociología, la psicología, la biología y la mercadotecnia para analizar y visualizar patrones en datos de encuestas, cuestionarios y otras formas de datos categóricos.

El ACS fue desarrollado por el estadístico francés Jean-Paul Benzécri en la década de 1970 como una extensión del análisis de componentes principales (ACP), técnica utilizada para analizar relaciones entre variables continuas, pero en el caso del ACS, se enfoca en las variables categóricas, como pueden ser las preguntas de opción múltiple en un cuestionario.

El objetivo es encontrar relaciones entre las variables categóricas y resumirlas en un conjunto de componentes principales. Estos componentes principales se denominan "dimensiones" y se pueden visualizar en un gráfico de dispersión bidimensional llamado "mapa de correspondencias". El mapa de correspondencias muestra la posición relativa de las diferentes categorías de las variables en cada dimensión y ayuda a identificar patrones y relaciones entre ellas.

La técnica del análisis de correspondencias se utiliza para responder a preguntas como "¿hay patrones en las respuestas a las preguntas de una encuesta que puedan ayudar a identificar

grupos de personas con características similares?" o "¿qué factores influyen en la elección de un partido político?".

Para realizar el ACS, primero se crea una tabla de contingencia [18] que muestra la frecuencia observada de las categorías de las variables en el conjunto de datos. Seguidamente, se utiliza esta tabla para calcular las proporciones esperadas de las categorías de las variables como si no hubiera ninguna relación entre ellas. Estas proporciones esperadas se comparan con las proporciones observadas en los datos para calcular las desviaciones entre las proporciones observadas y las esperadas.

A continuación, se utiliza un algoritmo iterativo para encontrar las dimensiones principales que explican la mayor parte de la variación en las desviaciones calculadas. Estas dimensiones se visualizan en el mapa de correspondencias y se interpretan en función de las categorías de las variables que están más fuertemente asociadas con cada dimensión.

La Figura 2.5 es un ejemplo de un mapa de correspondencias realizado entre dos variables categóricas, “País” y “Carrera más elegida por los universitarios”. En concreto, el gráfico muestra que existe cierta asociación entre el grado de Literatura e Italia, es decir, una de las carreras más elegidas en Italia es este grado, aspecto que también se percibe cierta asociación entre el grado de Medicina con Canadá y USA.

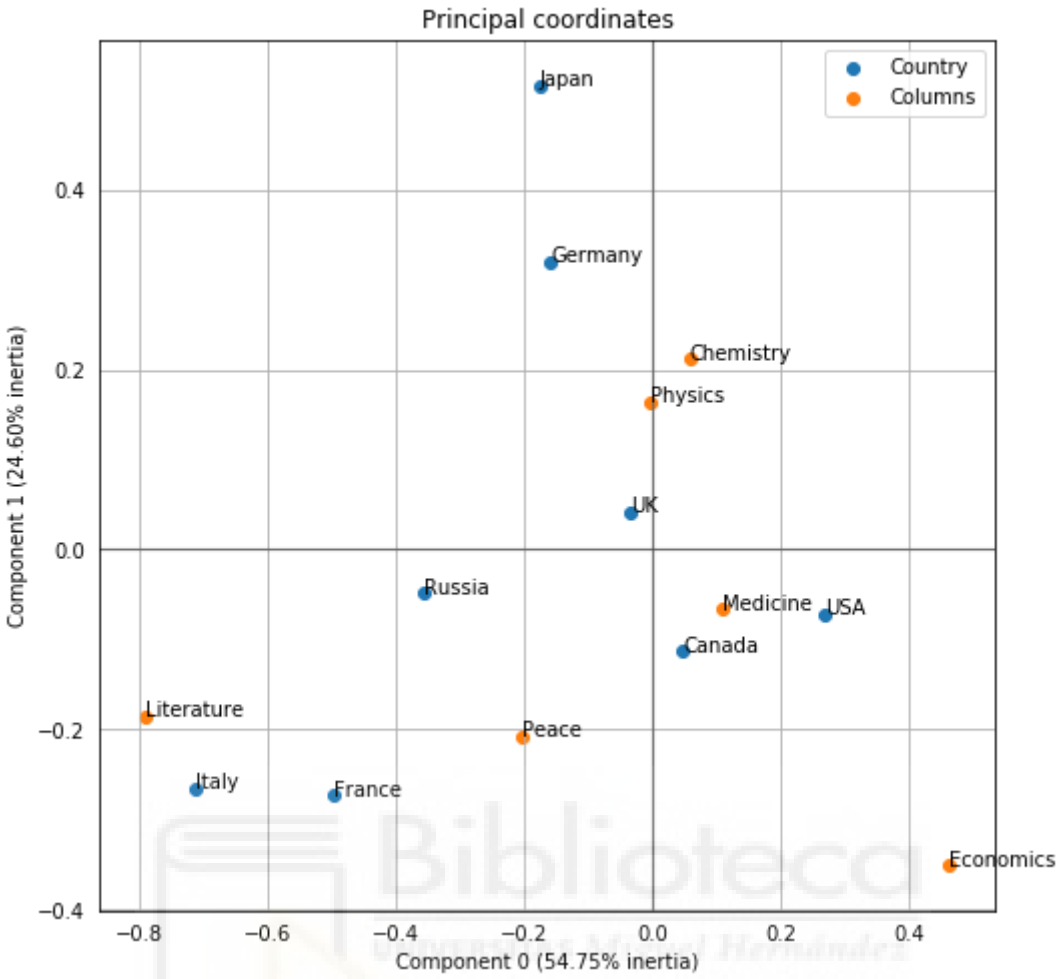

Figura 2.5: Tabla de correspondencias. Fuente: napsterinblue [17]

2.3.1.4- Prueba de Chi-cuadrado

La prueba de chi-cuadrado [19] fue desarrollada por el estadístico Karl Pearson a principios del siglo XX. El objetivo de la prueba es determinar si hay una relación significativa entre dos variables categóricas, como la opinión sobre un tema y el género, la edad, el nivel educativo o la región geográfica.

La prueba de chi-cuadrado se utiliza comúnmente para analizar datos de encuestas y estudios de opinión pública y en diferentes campos como la biología, la psicología y la medicina con el fin de analizar datos experimentales y observacionales. La prueba se utiliza para determinar si las frecuencias observadas son significativamente diferentes de las esperadas.
Para realizar la prueba de chi-cuadrado, al igual que en ACS, primero se crea una tabla de contingencia que contengan las frecuencias observadas de cada categoría. A continuación, se calculan las frecuencias esperadas suponiendo que las dos variables son independientes. Estas frecuencias esperadas se comparan con las frecuencias observadas para determinar si hay una relación significativa entre las dos variables.

El resultado de la prueba de chi-cuadrado es un valor estadístico de chi-cuadrado (χ^2), que se calcula como la suma de las diferencias al cuadrado entre las frecuencias observadas y las esperadas, divididas por las frecuencias esperadas. Si el valor de χ^2 es grande, indica que hay una relación significativa entre las dos variables categóricas.

Para interpretar el resultado de la prueba de chi-cuadrado, se puede realizar de dos formas, en la primera se compara el valor de χ^2 con un valor crítico de la distribución de chi-cuadrado. Este valor crítico se calcula en función del número de grados de libertad de la prueba. Los grados de libertad se calculan como el número de filas menos uno multiplicado por el número de columnas menos uno en la tabla de contingencia.

$$\chi^2 > \chi^2_{(n-1)(p-1),\alpha} \Leftrightarrow pvalor = P\left(X > \chi^2\right) \quad (4)$$

Si el valor de χ^2 es mayor que el valor crítico de la distribución de chi-cuadrado, se rechaza la hipótesis nula de no existencia de relación significativa entre las dos variables categóricas analizadas. En otras palabras, se concluye que entre las dos variables estudiadas sí hay una relación significativa.

La segunda forma de interpretar el resultado es mediante el pvalor [20], el cual se utiliza comúnmente en el contexto de las pruebas de hipótesis estadísticas para evaluar si los resultados obtenidos son significativos o no. Si el pvalor es menor que el nivel de significación predefinido (α) normalmente 0.05, se considera que los resultados son significativos y se rechaza la hipótesis nula.

2.3.2.- Modelos supervisados

Estas técnicas se utilizan para predecir la salida de un suceso introduciendo los datos de entrada, previamente etiquetados. Los datos de entrada se denominan características o atributos, y la salida que se quiere predecir se llama variable objetivo o también variable de clase cuando se trata de una variable categórica.

Para generar un modelo supervisado se proporciona al algoritmo un conjunto de datos de entrenamiento con valores o etiquetas conocidas. El algoritmo utiliza estos datos para aprender a relacionar las características de entrada con la variable objetivo y crear un modelo que pueda hacer predicciones precisas sobre nuevos datos que no se han utilizado en la fase de entrenamiento. Existen varios tipos de técnicas y algoritmos de aprendizaje supervisados, como pueden ser los Modelos de Regresión Lineal Múltiple y de algoritmos de Clasificación, cada uno con diferentes enfoques y técnicas. A continuación se describen brevemente ambas técnicas.

2.3.2.1- Modelos de Regresión lineal Múltiple

La regresión lineal [22] es una técnica estadística utilizada para predecir una variable continua. El objetivo es encontrar un modelo lineal que se ajuste a los datos y se pueda usar para hacer predicciones precisas.

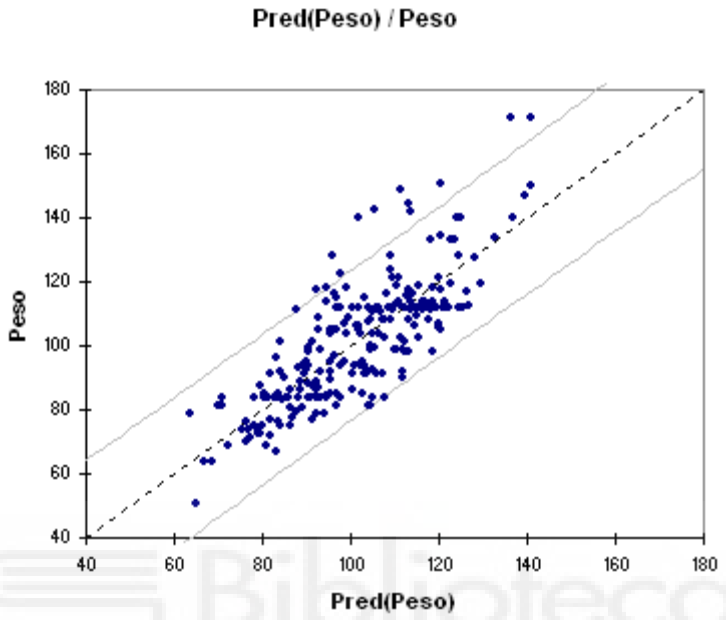

Figura 2.6: Ejemplo de regresión lineal. Fuente: xlstat [21]

Este modelo utiliza una fórmula matemática para describir la relación entre las variables. La ecuación es de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \quad (5)$$

Donde 'Y' es la variable dependiente que se desea predecir, X_1, X_2, …, X_p son las variables independientes, β_0 es el coeficiente de interceptación de la línea de regresión, β_1, β_2, ..., β_p son los coeficientes de regresión que indican cómo afecta cada variable independiente a la variable dependiente, y ε es el término de error.

El proceso de ajustar el modelo lineal a los datos implica minimizar la distancia entre los valores reales de la variable dependiente y los valores predichos por el modelo. Este proceso se puede realizar utilizando diversas técnicas estadísticas como, por ejemplo, el método de los mínimos cuadrados.

2.3.2.2- Clasificación

Un árbol de clasificación [23] es un método que utiliza una estructura de árbol que modela las decisiones y los resultados de un proceso de toma de decisiones. Cada nodo en el árbol representa una pregunta o una característica del conjunto de datos y las ramas representan las posibles respuestas o valores que se pueden tomar.

Los árboles de clasificación se utilizan para predecir valores de variables objetivo categóricas a partir de variables de entrada, es decir, para clasificar los datos en diferentes categorías. La construcción de estos árboles implica dividir repetidamente los datos en grupos diferentes, con la mejor característica que maximiza la separación entre los grupos y minimiza la impureza. Esta medida de impureza indica cómo de mezcladas están las clases en un grupo determinado. Cuanto menor sea la impureza, más homogéneas serán las clases y más informativa será la división. Por lo tanto, el objetivo es encontrar la característica que maximice la separación entre los grupos y minimice la impureza.

El árbol se construye a partir de una raíz, que representa la variable de entrada más importante y se divide en nodos que representan las variables de entrada secundaria. Cada nodo se divide en ramas, que representan los valores posibles para esa variable. Los nodos hoja representan las salidas o resultados finales del proceso de toma de decisiones.

Una de las ventajas de los árboles de decisión es que son fáciles de interpretar y visualizar.

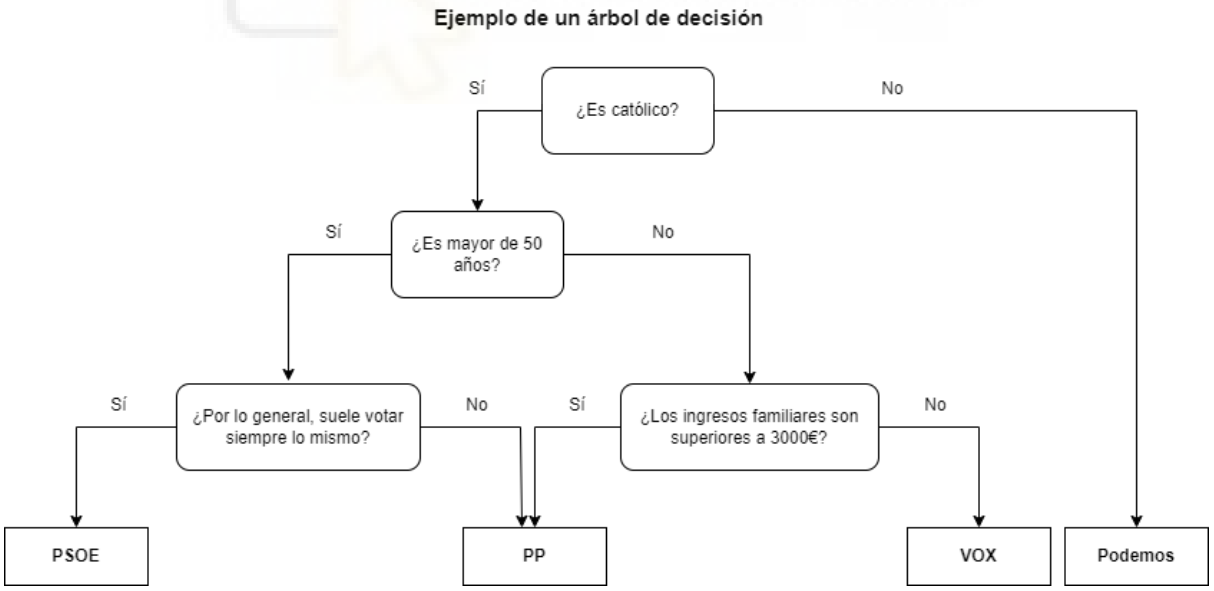

Figura 2.7: Ejemplo de árbol de decisión. Fuente: Elaboración propia

En el ejemplo anterior (Figura 2.7), se puede observar un árbol de clasificación para categorizar a un grupo de personas encuestadas según su adscripción a un determinado partido político. A través de las respuestas obtenidas en cuatro preguntas generales se

determina cuál es el partido al que mayoritariamente pertenece cada grupo de individuos. En la figura presentada, se pregunta al entrevistado si se considera católico, donde se le clasifica directamente en el grupo de Podemos en el caso de recibir una respuesta negativa. A continuación, se realiza una pregunta sobre la edad, y si esta es menor de 50 años, el árbol predice que la inclinación política del entrevistado es hacia la derecha. Siguiendo esa línea de razonamiento, se plantea la cuestión del salario, y en función del rango salarial, se sitúa al entrevistado en el partido político PP si el salario es superior a 3000€, o en VOX si el salario es inferior. Por otro lado, si se toma la otra rama del árbol, se indaga sobre la frecuencia con la que el entrevistado vota por los partidos políticos. Si la frecuencia es alta, el árbol sitúa al entrevistado en el partido PSOE, mientras que si la frecuencia es menor, se le ubica en el PP.

Estos árboles de clasificación poseen un porcentaje de precisión que puede variar según el contexto, permitiendo un cierto margen de error.

A la hora de calcular la precisión del árbol se puede recurrir a dos métodos distintos: F1-score [24] y 'accuracy'. Para poder calcular el F1-score, se necesitan otras dos variables, el 'recall' y la precisión. La precisión, lo que mide, es la proporción de votantes clasificados correctamente como positivos con respecto al total de votantes clasificados como positivos (verdaderos positivos y falsos positivos), esta medida es importante cuando lo que se desea es minimizar los falsos positivos, es decir, evitar clasificar incorrectamente a los votantes. El "recall" lo que mide es la proporción de votantes positivos clasificados correctamente como positivos con respecto al total de votantes que realmente son positivos (verdaderos positivos y falsos negativos), ya que con este lo que se busca es minimizar los falsos negativos.

Para calcular el F1-score se ha utilizado la siguiente fórmula:

$$F1 - score = \frac{2*precisión*recall}{precisión + recall} \quad (5)$$

En cuanto a la variable 'accuracy', que se usa para medir la precisión cuando los datos sí están balanceados, se utiliza para medir la capacidad general del modelo cuando éste está equilibrado. La manera de calcular el 'accuracy' en la siguiente:

$$Accuracy = \frac{TN + TP}{TN+TP+FP+FN} \quad (6)$$

donde:

- TP: significa verdaderos positivos, casos donde el clasificador predice TRUE, y la clase verdadera es TRUE.
- TN: significa verdaderos negativos, casos donde el clasificador predice FALSE, y la clase verdadera es FALSE.
- FP: significa falsos positivos, casos donde el clasificador predice TRUE, pero la clase correcta es FALSE

- FN: significa falsos negativos, casos donde el clasificador predice FALSE, pero es TRUE.

Y para finalizar, se procede a mostrar su matriz de confusión asociada al árbol de clasificación, donde los números correspondientes a la diagonal pertenecen a los verdaderos positivos, es decir, las personas que el algoritmo ha sido capaz de clasificar de manera correcta dentro del partido político al que le pertenece.

El triángulo inferior son los falsos negativos, es decir, gente que sí es votante de un partido político en concreto pero el modelo predijo que era de otro partido político. Y el triángulo superior, que son los votantes de otros partidos y que el modelo predijo que eran de un partido político en concreto.

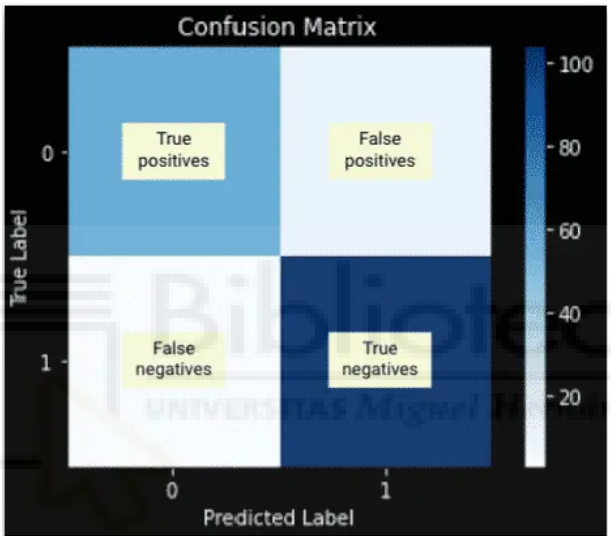

Figura 2.8: Ejemplo de matriz de confusión. Fuente: jcchouinard [25]

Capítulo 3 - Hipótesis de trabajo

En el presente capítulo se realiza una explicación detallada tanto del conjunto de datos utilizado para realizar el Capítulo 4, como del software, librerías, algoritmos y entornos empleados para la realización de este TFG.

3.1.- Dataset

El dataset [26] o conjunto de datos, que se ha empleado para realizar el proyecto se ha obtenido a través de la página oficial del Centro de Investigación Sociológicas [4]. Dicho archivo contiene los microdatos de las encuestas realizadas durante la primera quincena de febrero de 2023, correspondiente al Barómetro del CIS del mes de febrero, en el que se recogen, a través de un encadenado de preguntas, los datos sociodemográficos y de opinión de los ciudadanos encuestados.

El cuestionario utilizado por el CIS para recopilar la información consta de 57 preguntas que tratan, principalmente, temas sociales, económicos e ideológicos (véase cuestionario del Barómetro en el Anexo I). La base de datos obtenida a través de este cuestionario contiene 3.935 filas y 57 columnas correspondientes a las variables del cuestionario, más otras columnas añadidas para numerar o categorizar algunas variables.

CCAA	PROV	MUN	CAPITAL	TAMAÑO MUNIC.	SEXO	…	SINCERIDAD
Andalucía	Almería	<=100.00	Otros municipios	<=2.000 habitantes	Mujer	…	Bastante
Andalucía	Sevilla	<=100.00	Otros municipios	2.001 a 10.000 habitantes	Hombre	…	Mucha
…	…	…	…	…	…	…	…
Melilla	Melilla	Melilla	Capital de CC.AA.	50.001 a 100.000 habitantes	Mujer	…	Poca

Tabla 3.1: Dataset Intención de voto Febrero 2023 (extracto). Fuente: CIS[4]

3.2.- Google Colab

Para llevar a cabo este proyecto, se ha empleado Python, el lenguaje de programación más utilizado actualmente en el análisis de datos. Con el fin de utilizar este lenguaje sin requerir la instalación de ningún compilador adicional en el equipo, se ha decidido realizar el Trabajo de Fin de Grado (TFG) en las máquinas virtuales proporcionadas por Google.

Google Colab [27] es una plataforma gratuita basada en la nube para escribir y ejecutar código en Python. Esta herramienta es muy útil para desarrolladores y científicos de datos que necesitan trabajar con grandes conjuntos de datos y modelos de aprendizaje automático.

Una de las principales ventajas de Google Colab es que proporciona acceso gratuito a los recursos informáticos de Google, incluidas GPU y CPU, lo que permite ejecutar códigos más rápido que en una computadora local. Además, la plataforma está diseñada para facilitar la

colaboración en tiempo real, lo que significa que los usuarios pueden compartir su trabajo y colaborar en proyectos en línea.

Otra característica clave de Google Colab es su integración con otras herramientas populares, como Google Drive y GitHub. Los usuarios pueden importar y exportar fácilmente archivos desde y hacia estas plataformas, lo que facilita la gestión de proyectos y la colaboración con otros usuarios.

Google Colab también incluye una variedad de bibliotecas y herramientas populares preinstaladas, como NumPy, Pandas y Matplotlib, las cuales se explicarán posteriormente. Esto hace que sea fácil comenzar a trabajar en proyectos de aprendizaje automático y análisis de datos sin tener que preocuparse por instalar todas las herramientas y bibliotecas necesarias.

3.2.1.- Python

Python [28] es un lenguaje de programación de alto nivel, interpretado y multiplataforma, que se ha vuelto muy popular en los últimos años. Es conocido por su sintaxis clara y concisa, lo que hace que sea fácil de aprender y leer para los programadores de todos los niveles de experiencia. Python también es muy versátil, lo que lo convierte en una excelente opción para una amplia variedad de aplicaciones, como el análisis de datos, la inteligencia artificial, la automatización de tareas, el desarrollo web o la creación de juegos.

Una de las principales ventajas de Python es su gran comunidad de usuarios y desarrolladores [29] que han creado una amplia variedad de bibliotecas y herramientas para extender la funcionalidad del lenguaje.

Además, Python es un lenguaje de código abierto, lo que significa que cualquiera puede usarlo, modificarlo y distribuirlo libremente. Esto ha llevado a la creación de una amplia variedad de recursos educativos y de capacitación en línea, lo que hace que sea fácil para los nuevos programadores aprender el lenguaje.

3.2.2.- Máquina virtual de ejecución

Para realizar el proyecto se han realizado las pruebas en una máquina virtual de Google, la cual utiliza Python 3. La máquina sobre la que se ha ejecutado el código tiene las siguientes características:

Memoría	12,68 GB
Disco	107.72 GB

Tabla 3.2: Características de la máquina virtual. Fuente: elaboración propia

Al tratarse de una máquina virtual, los datos como el procesador que utiliza, la tarjeta gráfica que posee o el tipo de memoria están ocultos para el usuario.

3.3.- Librerías

Las librerías [30] en Python son un conjunto de módulos y funciones creadas por otros programadores que se pueden importar y utilizar en el código para realizar tareas específicas. Estas pueden ayudar a acelerar el proceso de programación al proporcionar soluciones para tareas comunes, como manipular archivos, entre otras cosas. Las librerías empleadas para realizar el TFG se explican a continuación.

3.3.1.- Matplotlib

Matplotlib [31] es una biblioteca de visualización de datos en 2D para el lenguaje de programación Python. Fue desarrollado por John D. Hunter en 2003 y es una de las herramientas más utilizadas en el campo del análisis de datos y la ciencia de datos.

La biblioteca Matplotlib se utiliza para crear gráficos y visualizaciones a partir de datos numéricos en Python. Puede crear varios tipos de gráficos, como gráficos de líneas, gráficos de barras, gráficos de dispersión, gráficos de contorno, gráficos de histograma, entre otros. También es posible personalizar y manipular estos gráficos para adaptarlos a las necesidades específicas de un proyecto.

Matplotlib se utiliza normalmente en campos como la investigación científica, la ingeniería, la estadística, las finanzas, la informática y muchos otros. También es una herramienta valiosa para la visualización de datos en el ámbito empresarial y para la presentación de informes y análisis.

3.3.2.- Numpy

NumPy [32] es una biblioteca de Python que se utiliza para realizar cálculos numéricos y científicos en Python. Es una de las bibliotecas más populares en el campo del análisis de datos y la ciencia de datos.

Se centra en la manipulación de arrays y matrices numéricas en Python. Proporciona un conjunto de funciones matemáticas y científicas que permiten a los usuarios realizar operaciones aritméticas, lógicas, estadísticas y de álgebra lineal en grandes conjuntos de datos numéricos de manera eficiente. NumPy también tiene una gran cantidad de funciones

para la manipulación de arrays, lo que facilita la creación de algoritmos personalizados y el procesamiento de datos.

Es conocida por su capacidad para realizar operaciones vectorizadas, lo que significa que puede realizar operaciones en grandes conjuntos de datos numéricos en paralelo, lo que lo hace mucho más rápido que los bucles de Python regulares. Esto hace que NumPy sea una excelente opción para cualquier proyecto que involucre el procesamiento de grandes cantidades de datos.

3.3.3.- Pandas

Pandas [33] es una biblioteca de Python utilizada para el análisis de datos. Fue desarrollado por Wes McKinney en 2008 y se basa en el lenguaje de programación Python. Pandas permite manipular y analizar grandes conjuntos de datos de manera rápida y eficiente. La biblioteca se centra en dos estructuras de datos principales: Series y DataFrame. Una Serie es un array unidimensional de datos etiquetados, mientras que un DataFrame es una estructura de datos bidimensional, similar a una tabla, que consiste en filas y columnas etiquetadas.

Las principales características de Pandas incluyen la limpieza y preparación de datos, la selección y filtrado de datos, el manejo de datos faltantes, la agregación y agrupación de datos, el análisis de series de tiempo y la integración con otras bibliotecas de Python.

3.3.4.- Seaborn

Seaborn [34] es una biblioteca de visualización de datos de Python basada en Matplotlib. Fue desarrollado por Michael Waskom en 2012 con el objetivo de crear gráficos de alta calidad con menos código y una sintaxis más simple que Matplotlib.

Seaborn proporciona una amplia variedad de gráficos estadísticos. Además, ofrece una amplia gama de paletas de colores personalizables para resaltar y distinguir datos importantes.

Una de las principales ventajas es su capacidad para trabajar con datos estadísticos complejos y visualizarlos de manera intuitiva. También proporciona herramientas para la exploración y el análisis de datos. Ofrece funciones para la manipulación de datos faltantes y la agrupación de datos. Además, permite crear gráficos utilizando subconjuntos de datos.

3.3.5.- Prince

La librería de Python Prince [35] es una herramienta de análisis de componentes principales y análisis de correspondencias que permite la exploración y la visualización de datos multivariables. Prince es una librería open-source que se utiliza habitualmente en el análisis de datos y en el aprendizaje automático.

En particular, esta librería permite llevar a cabo un análisis de componentes principales de datos cuantitativos, lo que a su vez permite la selección de las k mejores variables. Adicionalmente, se ofrece la posibilidad de visualizar los datos en diferentes formatos, tales como gráficos de dispersión, gráficos de barras y gráficos de calor. Finalmente, esta herramienta permite también la reducción de la dimensionalidad y la imputación de datos faltantes presentes en el conjunto de datos analizado.

3.3.6.- Stats

Stats [36] es una librería de Python que se utiliza para el análisis y manipulación de datos, y para la realización de cálculos estadísticos, lo que la convierte en una herramienta esencial en el análisis de datos y la investigación estadística.

La librería Stats ofrece una amplia gama de funciones y herramientas, tales como la generación de números aleatorios, el cálculo de distribuciones de probabilidad, el ajuste de modelos de regresión y la realización de pruebas de hipótesis. También contiene herramientas para la realización de análisis exploratorio de datos, como la visualización de distribuciones de datos y la realización de pruebas de normalidad.

Además, proporciona funciones para la estimación de parámetros estadísticos, como la media, la desviación estándar y la coincidencia. Asimismo, ofrece herramientas para la manipulación de datos estadísticos, como el agrupamiento de datos, la agregación de datos y la filtración de datos.

3.3.7.- Sklearn

La librería Scikit-learn [37], también conocida como sklearn, es una biblioteca de aprendizaje automático de código abierto en Python que proporciona herramientas para el análisis y modelado de datos. Sklearn es ampliamente utilizada en diversas áreas, como la investigación científica, la ingeniería, la ciencia de datos y la industria.

Ofrece una amplia variedad de herramientas para el preprocesamiento de datos, selección de características, clasificación, regresión, agrupamiento, reducción de la dimensionalidad,

selección de modelos y evaluación de modelos. La biblioteca proporciona una serie de algoritmos de aprendizaje automático predefinidos y funciones útiles para crear y evaluar modelos. Además, ofrece una interfaz sencilla y coherente para que los usuarios puedan trabajar con datos y modelos de manera intuitiva.

También dispone de funciones de selección de características, que permite a los usuarios seleccionar las variables más importantes para el modelado. La biblioteca ofrece una variedad de métodos de selección de características, incluyendo el análisis de componentes principales, la selección de características basada en árboles y la selección de características basada en estadísticas.

En cuanto a la clasificación y la regresión, Scikit-learn dispone de una amplia gama de algoritmos para modelos de clasificación binaria, modelos de clasificación multiclase y modelos de regresión. Entre los algoritmos de clasificación se incluyen el análisis discriminante lineal (LDA), el clasificador de máxima verosimilitud de Naïve Bayes y la máquina de vectores de soporte (SVM). Entre los algoritmos de regresión se incluyen la regresión lineal, la regresión logística y los árboles de decisión.

En cuanto a la evaluación de modelos, Scikit-learn contiene una amplia gama de herramientas para evaluar el rendimiento de los modelos. Esto incluye funciones para la validación cruzada, la matriz de confusión, la curva ROC y la curva de precisión-recuperación. Además, Scikit-learn ofrece una variedad de métricas para evaluar el rendimiento de los modelos, como la precisión, la sensibilidad, la especificidad y el F1-score.

3.3.8.- Google

La librería Google en Python consiste en un conjunto de paquetes y módulos desarrollados por Google para proporcionar una amplia gama de funcionalidades y servicios de Google. Estos paquetes y módulos permiten a los desarrolladores interactuar con diversos servicios y productos de Google, como Google Drive, Google Cloud, Gmail, Google Maps, Google Calendar, YouTube, entre otros.

La librería está diseñada para facilitar la integración de aplicaciones y servicios con los productos y servicios de Google, utilizando las API y las herramientas proporcionadas por Google.

Capítulo 4 - Resultados

En el siguiente capítulo se realiza una descripción objetiva de las características más significativas del dataset. También se realiza una selección de los atributos más importantes de esta, para posteriormente llevar a cabo un trabajo de predicción del voto en función de los atributos anteriormente seleccionados.

4.1.- Introducción

Como se ha explicado anteriormente, la población actual muestra un interés importante en cuanto a los temas de política se refiere y los políticos también tienen un especial interés en saber cuál es la opinión pública. Uno de los objetivos de este trabajo es poder satisfacer las necesidades del político y proporcionarle de manera segmentada las características más comunes del votante en función del partido al que pertenecen y poder predecir, por medio de un árbol de clasificación, el partido al que pertenecen en función de unas pocas preguntas correspondientes al barómetro realizado por el CIS [4].

4.2.- Descripción de la muestra y población objetivo

Para realizar este trabajo se ha hecho uso de los datos del barómetro de febrero de 2023 del CIS [4]. Tanto el tamaño de la muestra como las preguntas del cuestionario fueron diseñadadas por el CIS. En este estudio, el CIS planteó un número inicial de entrevistas de 4000, de las cuales, finalmente, se realizaron 3935. Las entrevistas se realizaron entre el 1 y el 11 de febrero de 2023. El público objetivo de la muestra fue toda la población española de ambos sexos de 18 años o más. La tabla 4.1 muestra las encuestas realizadas por comunidades autónomas.

Código	Comunidad Autónoma	Muestra
01	Andalucía	677
02	Aragón	103
03	Asturia (Principado de)	96
04	Baleares (Islas)	99
05	Canarias	160
06	Cantabria	97
07	Castilla La Mancha	168
08	Castilla y León	195
09	Cataluña	562
10	Comunidad Valenciana	375
11	Extremadura	85
12	Galicia	250
13	Madrid (Comunidad de)	565
14	Murcia (Región de)	120
15	Navarra (Comunidad Foral de)	102
16	País Vasco	174
17	La Rioja	91
18	Ceuta (Ciudad autónoma de)	8
19	Melilla (Ciudad autónoma de)	8
	Total:	3995

Tabla 4.1: Número de encuestas realizadas por CCAA. Fuente: CIS [4]

Según la ficha técnica proporcionada por el CIS, la distribución de la muestra fue proporcional por comunidades autónomas. Las muestras fueron recogidas en las 50 provincias de España, y se obtuvieron en 1233 municipios distintos.

El procedimiento de selección de la muestra fue aleatorio por el teléfono fijo y móvil con un porcentaje del 27,2% y del 72,8%, respectivamente. La selección de individuos se realizó aplicando cuotas de sexo y edad.

En este estudio, se utilizó un error muestral a un nivel de confianza del 95,5% (dos sigmas) y P = Q = 0.5, con un error de muestreo del $\pm\ 1,6\%$ para el conjunto de la muestra, en el supuesto de muestreo aleatorio simple.

4.3.- Cuestionario utilizado

El cuestionario empleado para recoger los casi 4000 registros que elaborado por el CIS [4] se recogen las respuestas de las variables más relevantes para predecir la intención de voto. En casi su totalidad se basan en preguntas con una respuesta categórica, con múltiples opciones de respuesta. El cuestionario se compone de 57 preguntas agrupadas en los siguientes bloques:

- Preguntas que tratan sobre la situación social y económica que está atravesando el entrevistado y España.

- Preguntas ideológicas y partidistas: "Suponiendo que mañana se celebrasen elecciones generales, ¿a qué partido votaría usted?" o "Cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Situándonos en una escala que va del 1 al 10, en la que 1 significa lo más a la izquierda y 10 lo más a la derecha, ¿en qué casilla se colocaría usted?".

- Pregunta sobre el nivel de estudios que posee el entrevistado.

- Preguntas sobre materia religiosa, la posición del entrevistado en este ámbito, y con qué frecuencia celebra algún acto de su religión.

- Preguntas sobre cuál es la situación laboral y también económica en la que se encuentra el entrevistado actualmente.

En el Anexo I del presente trabajo se muestra íntegra y detalladamente el cuestionario utilizado para hacer las entrevistas a las personas que participaron en la encuesta.

4.4.- Tratamiento previo del dataset

Antes de realizar cualquier análisis descriptivo o de aplicar cualquier algoritmo se debe realizar un preprocesado a los datos para poder trabajar sobre él de una manera más correcta y rápida.

En primer lugar, no todas las variables que están en el dataset resultan de interés para el estudio, ya sea porque el tipo de información que proporciona es innecesaria, o porque los registros de dicha columna son univaluados y no tiene sentido dejar todas las preguntas en el dataset.

A continuación, mediante la Figura 4.2 se puede observar cuáles fueron las variables que fueron eliminadas del dataset y que pueden consultarse, para más detalle, en el Anexo I.

Variable	Significado
ENTREV	Identificador del entrevistador
TIPO_TEL	Tipo de teléfono al que se llama
ESTUDIO	(Univaluada) Código del estudio
REGISTRO	Número del registro
CUES	Número del cuestionario
LIDERESCONOCE	Pregunta si conoce o no a diferentes líderes políticos
VALORALIDERES	Pide que valore a diferentes líderes políticos

Tabla 4.2: Variables eliminadas. Fuente: Elaboración propia

Adicionalmente, se descartaron las encuestas en las que se identificó falta de sinceridad o una baja calidad de las respuestas proporcionadas por parte de los encuestados.

En segundo lugar, en algunos casos, nuestra variable objetivo, la P14 ("*¿A qué partido votaría Ud.?*"), no fue contestada, lo cual limita su utilidad para el estudio. Por lo tanto, se optó por eliminar los registros que cumplían la condición de "N.C" (No Contesta). En este caso particular, la respuesta "No sabe todavía" fue tratada como si se tratara de otro partido político.

Para el resto de variables, el tratamiento del N.S y N.C fueron catalogadas como si de la misma respuesta se tratase, y en función del número que hubiese de estas, se ha optado por un procesamiento u otro.

Cuando una columna del dataset posee menos de 50 registros con la categoría N.S./N.C., se ha optado por reemplazar esos valores por la moda de la columna, por otra parte, en el caso de que esa columna posea más de 800 registros con dicha respuesta, se ha optado por

suprimir la columna, y por último, en el caso intermedio, se ha tratado esa respuesta como una opción más.

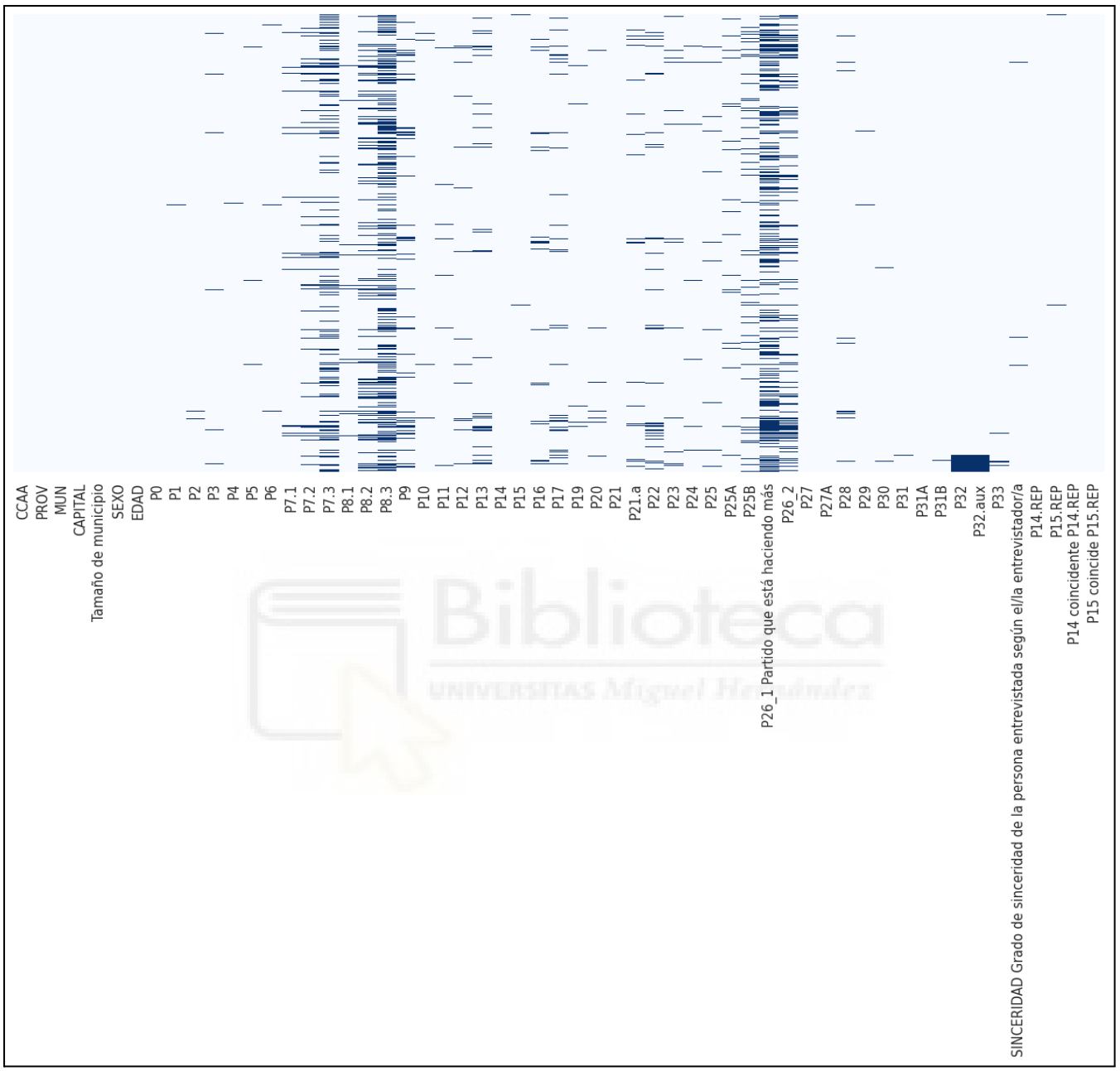

Figura 4.1: Gráfico de valores nulos en el dataset. Fuente: Elaboración propia

Como se observa en la Figura 4.1, las franjas azules muestran los valores nulos de cada una de las variables y que han sido tratadas con el procedimiento anterior para dejar un dataset sin ningún valor por responder.

En tercer lugar, en la pregunta P14 ("*Suponiendo que mañana se celebrasen las elecciones generales, es decir, al Parlamento español, ¿a qué partido votaría Ud.?*".), que es la variable objetivo del estudio y cuyas categorías de respuestas son: PSOE, PP, VOX, PODEMOS, IU (Izquierda Unida), Unidas Podemos, En Comú Podem, En Común - Unidas Podemos y.

Ciudadanos, se ha realizado un tratamiento de las categorías para poder tratar mejor los datos. Por una parte, se han agrupado las diferentes variantes de PODEMOS (Unidas Podemos, En Comú Podem y Podemos) en un único partido llamado PODEMOS, ya que realmente es el mismo partido, lo único que cambia es el nombre según la comunidad autónoma. Y por otra parte, se han agrupado todos los partidos minoritarios con o sin representación parlamentaria en un único grupo político (Otro partido), para poder representar con mayor facilidad los partidos con una mayor representación en el congreso.

En concreto, dentro de la muestra de "Otro partido" se han agrupado a las siguientes formaciones políticas:

IU (Izquierda Unida), Ciudadanos, Más País, ERC (Esquerra Republicana de Catalunya), JxCat (Junts per Catalunya), CUP, EAJ-PNV (Partido Nacionalista Vasco), EH Bildu (Euskal Herria Bildu), CC-PNC (Coalición Canaria – Partido Nacionalista Canario), Nueva Canarias, UPN (Unión del Pueblo Navarro), Compromís, BNG (Bloque Nacionalista Galego), PRC (Partido Regionalista de Cantabria), Teruel Existe, PACMA (Partido Animalista), FAC (Foro Asturias) y Voto nulo.

4.5.- Análisis descriptivo

El análisis descriptivo es una técnica estadística que se utiliza para resumir y visualizar los datos de una muestra. En este proyecto, se ha utilizado esta técnica para describir las características de la muestra, tanto de las variables sociodemográficas como de las variables relacionadas con la política. Para ello, se han empleado gráficos y tablas para visualizar los resultados obtenidos.

4.5.1.- Representación de intención del voto

En la siguiente sección, proporcionamos una descripción detallada de los procedimientos utilizados y los resultados obtenidos.

El primer análisis a realizar es en relación a la pregunta 14, donde se le realizó al entrevistado la cuestión: "*Suponiendo que mañana se celebrasen las elecciones generales, es decir, al Parlamento español, ¿a qué partido votaría Ud.?*". Ya que es interesante saber las proporciones de los diferentes partidos políticos que existen dentro la muestra de la población. Pero, para este caso, sólo resulta de interés saber cuáles son los partidos políticos con más representación entre nuestros entrevistados, por lo que los partidos políticos que reciben un porcentaje inferior al 3% de representación se han englobado en un partido único llamado "Otro partido".

Por lo que, los grupos que compondrán el gráfico (Figura 4.2) son: PP, PSOE, Otro partido, No sabe todavía, No votaría, PODEMOS, VOX, En blanco.

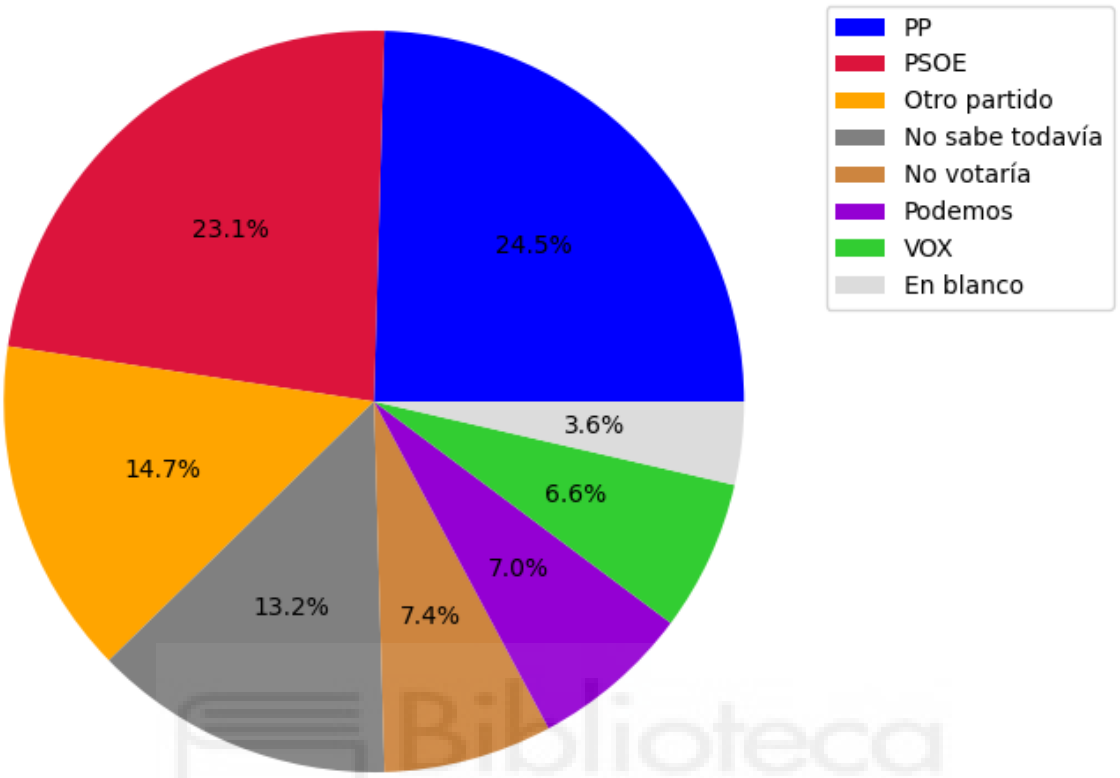

Figura 4.2: Porcentaje de los partidos según el voto. Fuente: Elaboración propia

Como se muestra en la Figura 4.2, el Partido Popular y el PSOE ocupan casi la mitad del porcentaje del voto. Después le sigue la formación de Otro partido con un 14,4%, seguidamente, se encuentra el grupo de los indecisos, es decir, gente que a fecha de febrero no sabe todavía en qué formación política va a depositar su voto. Posteriormente, se encuentra el grupo abstencionista, gente que ya sabe que no va a ir a votar el día de las elecciones y finalmente, le suceden PODEMOS, seguida de VOX y por último el voto en blanco.

Para posteriores análisis, las respuestas de ciertas formaciones políticas como "Otros partidos" o el "Voto en blanco", se han eliminado de los análisis para no añadir ruido innecesario al estudio, ya que se trata de una representación mínima en la muestra.

4.5.2.- Representación por sexo

A la hora de querer hacer una representación por sexos (SEXO) y analizar las tendencias tendencias politicas de estos (P14), se ha de tener cuidado de que no se tengan unos datos desbalanceados. La muestra de nuestro dataset (una vez eliminados los registros asociados a

“Otros partidos”) se encuentra balanceada ya que no domina ningun sexo sobre otro, aspecto que se puede observar en la Figura 4.3.

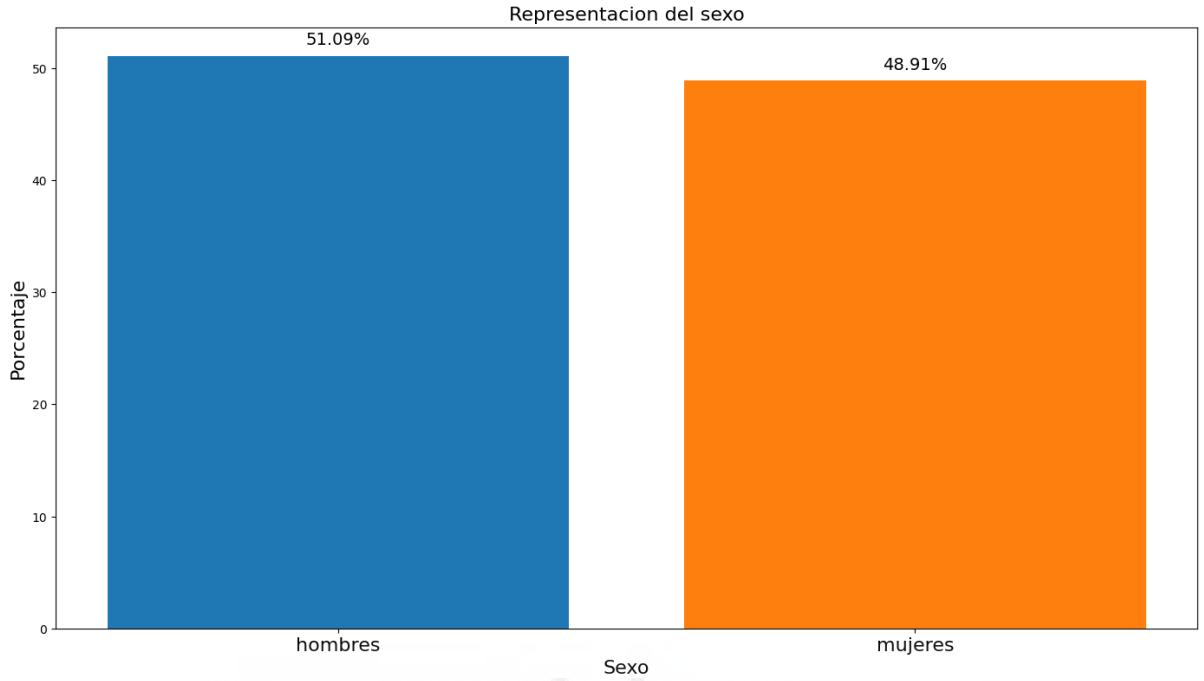

Figura 4.3: Representación por sexo en el dataset. Fuente: Elaboración propia

En concreto, tal y como indica la figura anterior, nuestra muestra se compone de 1681 hombres y 1609 mujeres, es decir, el 51.09% está compuesto de hombres y el 48.91% de mujeres.

En cuanto a la comparativa sobre la decision del voto en funcion del sexo, es decir, lo qué votan los hombres y lo qué votan las mujeres, se ha realizado el análisis de la intención de voto pero diferenciando las respuestas de los encuestados. Como se puede observar en la siguiente figura, los hombres prefieren votar más por los partidos más tradicionales de España, como es el caso del PP y del PSOE, con el 28,53% de los votos y el 26,43% respectivamente. Seguidamente, la tercera opción más representativa dentro de los hombres es la incertidumbre, ya que el 17,18% de ellos todavía no sabe a qué partido votaría en las elecciones generales. En cuarto lugar, se sitúa VOX, el partido más conservador, con un 9,74%, seguido del abstencionismo, es decir, de los hombres que no irían a votar, con un 9,33% y por último, se encuentra PODEMOS, con un 8,17% de representación del voto masculino.

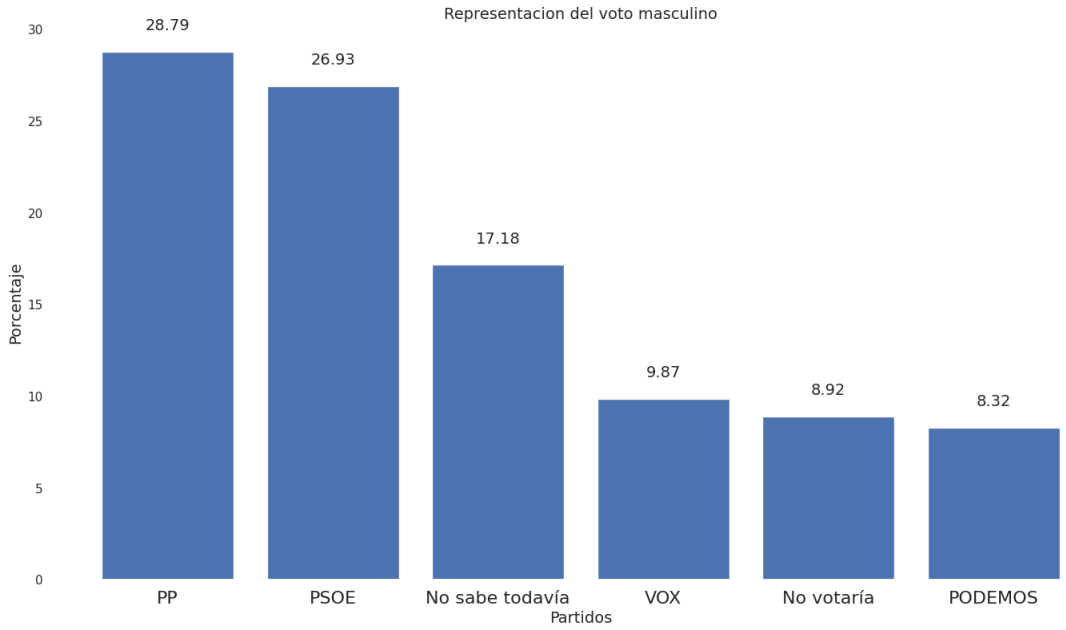

Figura 4.4: Representación de voto masculino. Fuente: Elaboración propia

En el caso de las mujeres, se puede ver en la siguiente figura que, los dos primeros resultados se parecen a los porcentajes de respuesta masculina, donde el PP seguiría ocupando la primera posición, seguido del PSOE, con un 31,29% y 29,74%, respectivamente. En tercer lugar, se sitúan las abstencionistas con una representación del 15,13% de las mujeres encuestadas, acompañado por PODEMOS, que en este caso, obtendría una representación del 9,52%, después se encuentran las mujeres indecisas con un 8,63% y por último se sitúa VOX con un 5,68% del voto femenino para este partido.

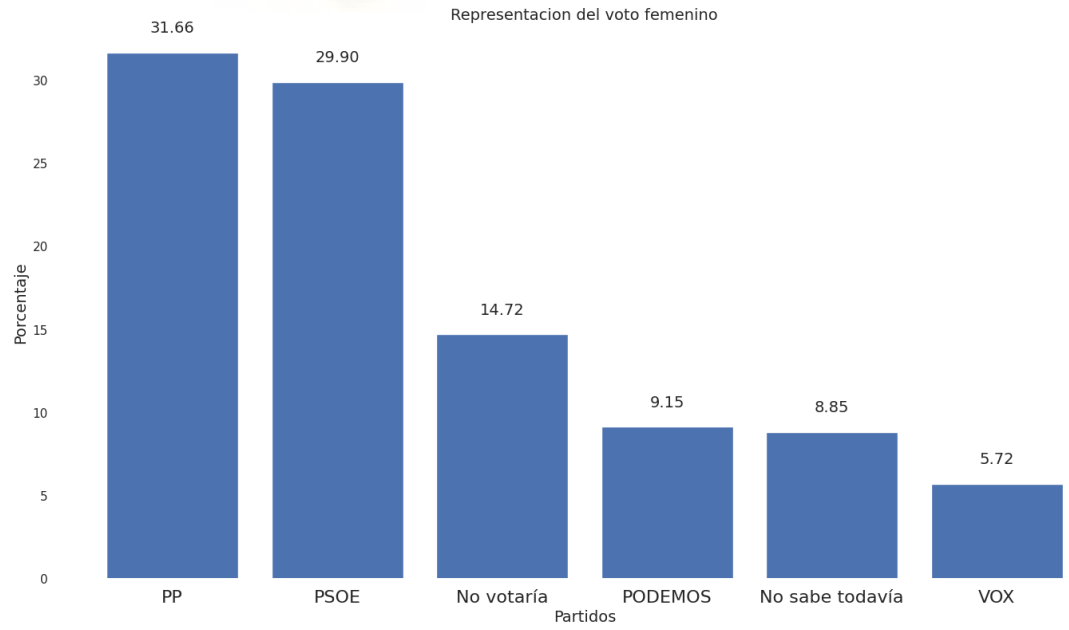

Figura 4.5: Representación de voto femenino. Fuente: Elaboración propia

En la Figura 4.6 se muestra una comparativa de la intención de voto de los hombres y la intención de voto de las mujeres.

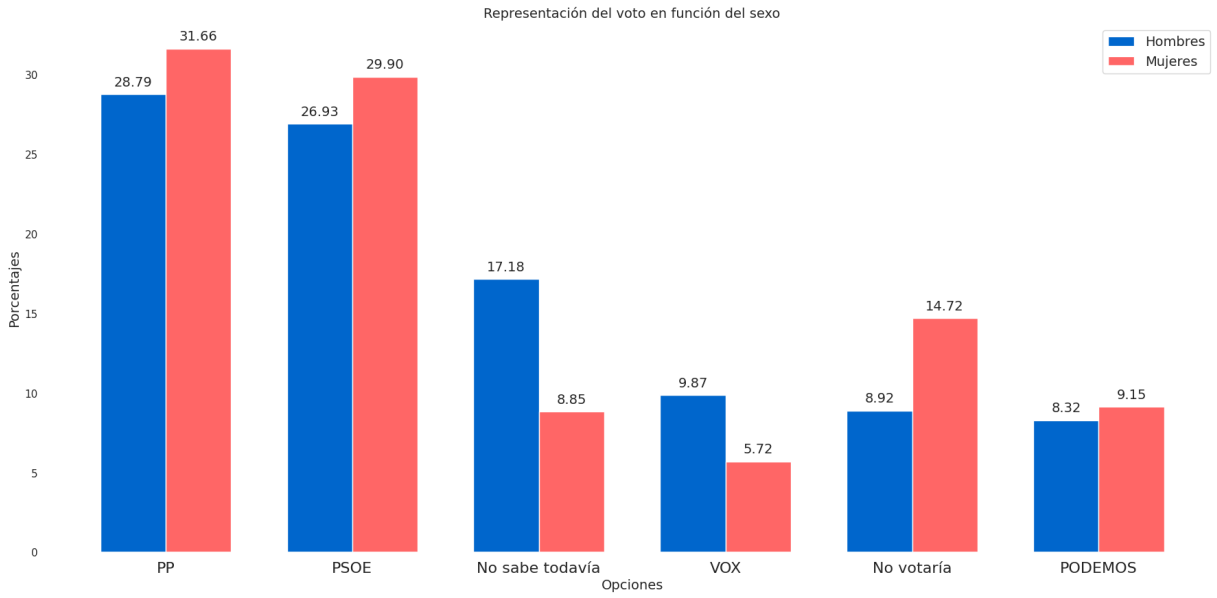

Figura 4.6: Representación de las diferencias del voto por sexo. Fuente: Elaboración propia

Las conclusiones que se pueden obtener observando el gráfico anterior es que las mujeres optan por depositar su voto, más frecuentemente, en partidos tradicionales, como sería el caso de PP y PSOE, con una representación ligeramente mayor que la de los hombres, aunque la tendencia de los hombres es similar a la de las mujeres. Por otra parte, las mujeres son la mitad de indecisas que los hombres, ya que las mujeres tienen más claro donde van a depositar su voto, aunque este sea no votar. Y finalmente, se observa como los hombres tienen una mayor tendencia hacia la formación de VOX que las mujeres.

4.5.3.- Representación por edad

En el ámbito político, comprender cómo se distribuye el voto entre diferentes grupos demográficos es fundamental para entender las preferencias y tendencias electorales de una sociedad. Uno de los factores demográficos que influye significativamente en las elecciones es la edad de los votantes. El análisis del voto por representación de la edad proporciona información valiosa sobre cómo las preferencias políticas varían a lo largo de las diferentes generaciones y cómo estas diferencias pueden afectar los resultados electorales.

A lo largo de la historia, hemos presenciado cambios significativos en las preferencias políticas de distintos grupos de edad. Diferentes generaciones pueden tener perspectivas y prioridades políticas únicas, lo que se refleja en la forma en que ejercen su derecho al voto. Comprender estas diferencias puede ayudar a los partidos políticos y candidatos a diseñar estrategias efectivas de campaña, así como a los investigadores a profundizar en las dinámicas políticas y sociales de una sociedad.

En este contexto, realizar un análisis de la intención del voto (P14) por representación de la edad (EDAD) implica examinar cómo se distribuye el apoyo electoral entre los diferentes grupos de edad y qué patrones o tendencias pueden surgir. Esto implica identificar cómo las preferencias políticas pueden variar entre los jóvenes, los adultos y las personas mayores, y cómo estas diferencias pueden influir en la toma de decisiones electorales.

Por eso, en este estudio se ha segmentado la edad del encuestado en función de los cuatro partidos más representativos.

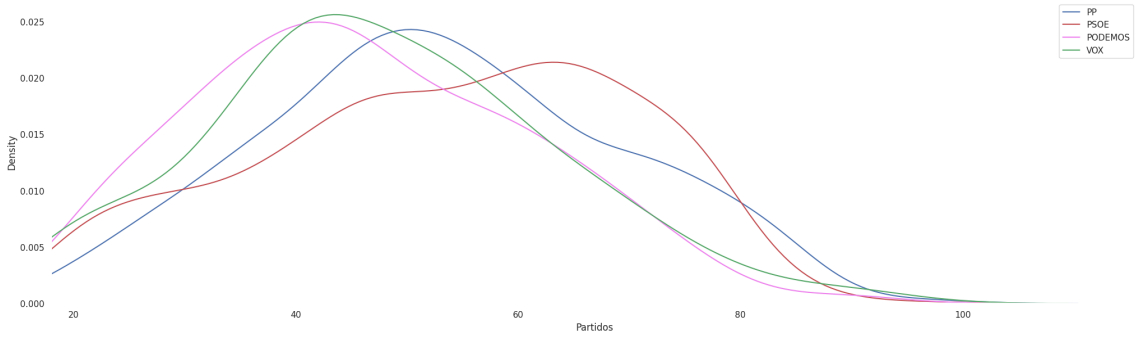

Figura 4.7: Representación de la edad por partidos. Fuente: Elaboración propia

En la Figura 4.7 se ha representado a la población española en función del partido al que votan y la edad que tienen, con el fin de realizar un análisis de la distribución de la edad de cada partido. Como se puede observar, el partido que más jóvenes posee es PODEMOS, donde el pico de edad (la moda) se alcanza a los 40 años y después va decreciendo el número de votantes. Los resultados indican que la edad media de un votante de PODEMOS es de 46 años con una desviación típica de 14,89 años.

El segundo partido con más jóvenes es VOX, el cual se sitúa cerca de PODEMOS al tener el pico de sus votantes en los 43 años, pero a diferencia de PODEMOS, donde se observa que tiene una frecuencia ascendente normal, VOX empieza a partir de los 25 a tener un crecimiento muy acusado hasta alcanzar su pico máximo, y posteriormente, va teniendo una pérdida ligera de manera progresiva conforme la edad va aumentando. La edad promedio de un votante de este grupo político es de 47 años con una desviación típica de 15,44 años.

Seguidamente, se sitúa el PP con su pico máximo de votantes en torno a los 53 años, donde se puede apreciar que, en la primera etapa, es decir, la gente más joven, no decanta su voto por este partido, y es más hacia la zona media hasta el final donde este partido acumula más representación. En este caso, la edad media de un votante del PP es de 53 años con una desviación típica de 16,05 años.

Finalmente nos encontramos con el PSOE, el cual se observa que realiza tres pequeñas subidas hasta llegar a su pico de votos en la franja de 64 años. Hecho que coincide con la edad aproximada de jubilación. Este partido posee una minoritaria representación en las

primeras franjas de edad con respecto al resto de partidos, pero se observa cómo mantiene bien el voto en las franjas más avanzadas. La edad media de un votante del PSOE es de 53 años, y su desviación típica es de 16,74 años.

Por lo que se puede concluir que la gente mayor prefiere votar a los partidos "de toda la vida" y depositar su voto en partidos como PP y PSOE, mientras que la gente de edad joven y edad media, prefieren probar alternativas diferentes y novedosas, apostando con su voto por partidos más nuevos, como es el caso de PODEMOS y VOX, cuyas propuestas suelen ser más populistas.

4.5.4.- Valoración de la situación social en España

La situación social de un país, incluyendo aspectos como el empleo, la desigualdad, la pobreza, la movilidad social y otros indicadores socioeconómicos, desempeña un papel crucial en la forma en que los ciudadanos perciben y evalúan las opciones políticas. Las experiencias y percepciones individuales de la situación social pueden influir en la confianza en el sistema político, las preferencias ideológicas y la elección de los candidatos.

El análisis del voto por representación de la situación social implica examinar cómo se distribuye el apoyo electoral en función de indicadores socioeconómicos relevantes. Esto implica identificar cómo las preferencias políticas pueden variar en función del empleo, el nivel de ingresos, la educación, la desigualdad y otros factores socioeconómicos pertinentes.

En las siguientes figuras se puede visualizar las valoraciones sobre la situación social de España (P3) que realizan los votantes (P14) del PSOE, el partido que gobierna actualmente España y posteriormente, las valoraciones que realizan los votantes del PP, partido de la oposición.

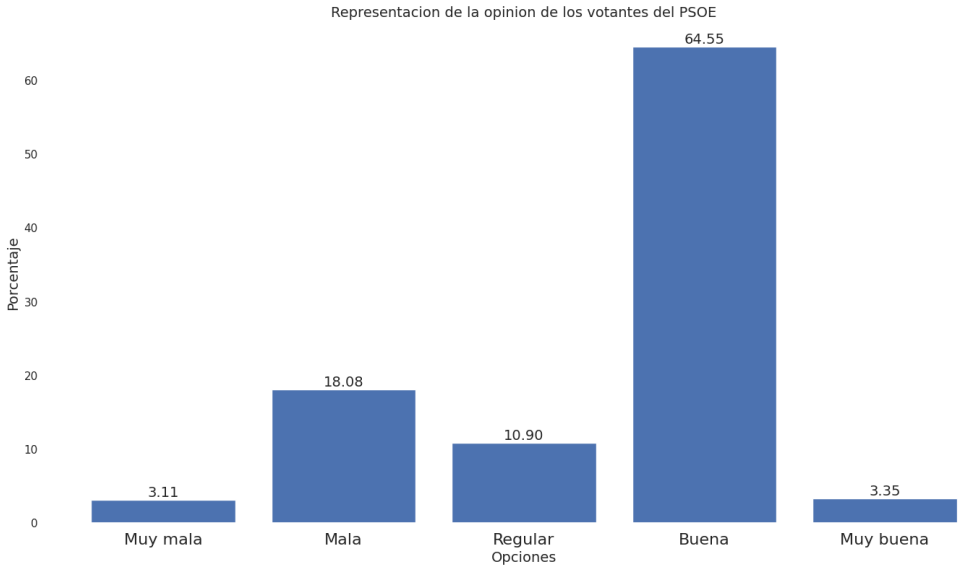

Figura 4.8: Representación de la valoración social por votantes del PSOE. Fuente: Elaboración propia

Como podemos ver, la mayoría de los votantes del PSOE considera que la situación social actual en España es buena.

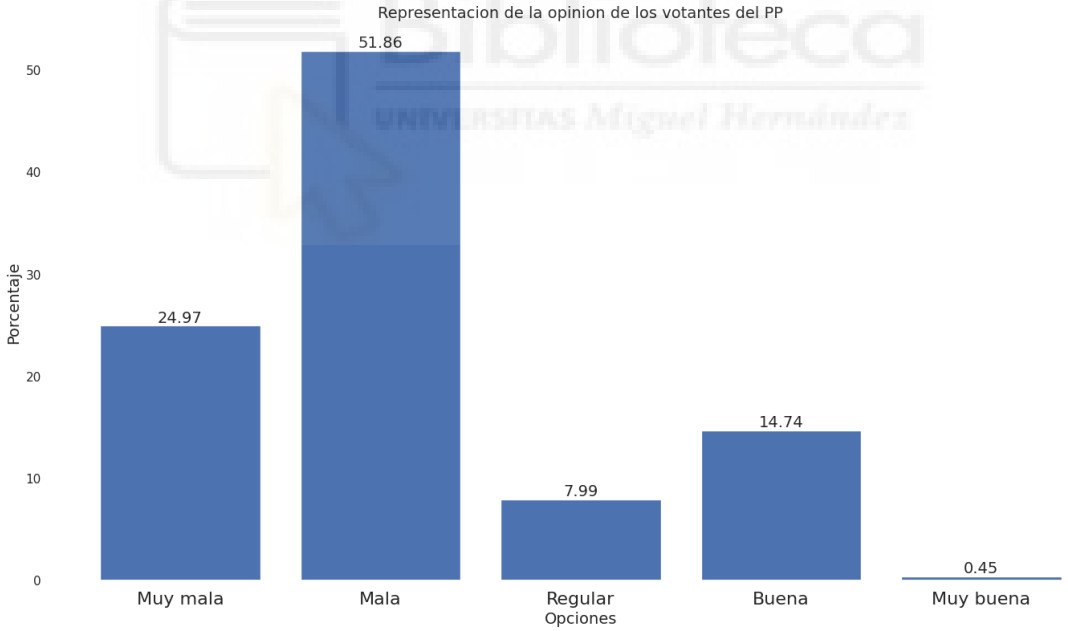

Figura 4.9: Representación de la valoración social por votantes del PP. Fuente: Elaboración propia

Mientras que los votantes del PP (Figura 4.9), en contraposición, consideran que, actualmente, se está pasando por una situación mala o muy mala. Casualmente, estas opiniones coinciden con que el gobierno del país recae en el PSOE por lo que sus votantes consideran que la situación es favorable y viceversa, los votantes pertenecientes a la oposición, como su partido no es el que representa a España opinan que la situación es bochornosa.

4.5.5.- Los principales problemas de España según los votantes

Los principales problemas de un país abarcan una amplia gama de temas que son de interés y preocupación para la sociedad. Estos pueden incluir cuestiones como el desempleo, la economía, la corrupción, la educación, la sanidad, la vivienda, el cambio climático, la migración y otros desafíos cruciales que enfrenta una nación. El análisis del voto en relación con estos problemas permite comprender cómo las posiciones políticas y las propuestas de los partidos y candidatos se alinean con las preocupaciones y expectativas de los votantes.

A continuación, se exponen los tres principales problemas de España (P7) indicados por los encuestados y analizados en función de la formación política a la que pertenecen (P14). Para analizar este apartado se ha considerado exclusivamente a los partidos más representativos del congreso de los diputados.
En concreto, para los votantes de la formación azul, los principales problemas que destacan los encuestados son: la crisis económica en primer lugar, los problemas políticos en general, y por último, el gobierno y partidos políticos concretos, porcentajes que se observan en la Figura 4.10 (el resto de problemas al ser menos representativos se decidió omitirlos de la gráfica).

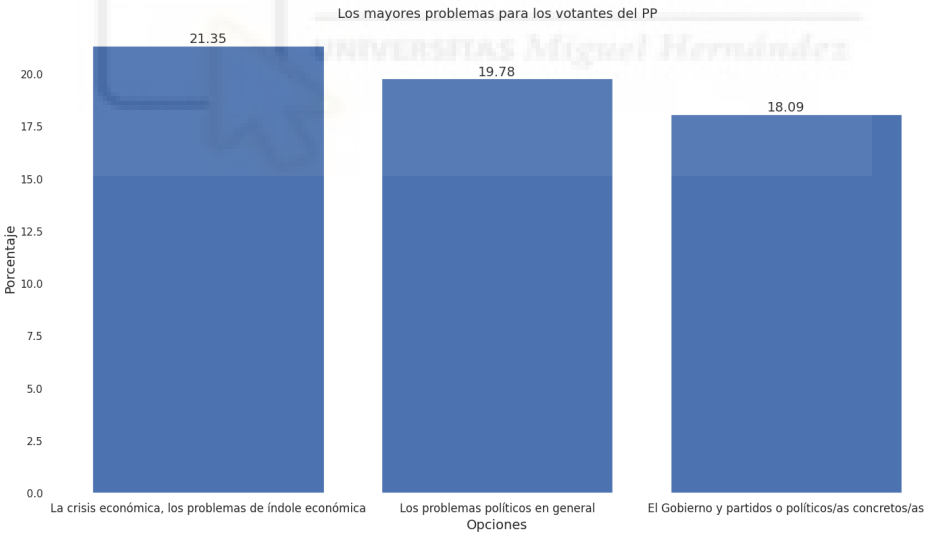

Figura 4.10: Representación de los principales problemas de España según los votantes del PP. Fuente: Elaboración propia

Seguidamente, se agruparon los principales problemas de España, pero esta vez para los votantes del PSOE, y como se observa en la siguiente gráfica, los tres primeros problemas que consideran los socialistas son: la crisis económica, el paro y los problemas políticos en general. A diferencia de los populares, y como es lógico, los socialistas no consideran su principal problema, el gobierno y partidos en concreto.

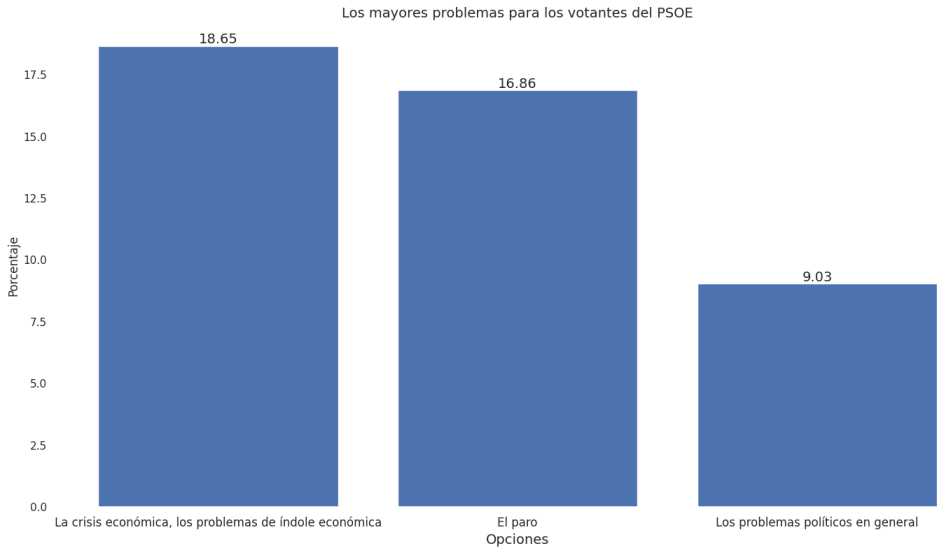

Figura 4.11: Representación de los principales problemas de España según los votantes del PSOE. Fuente: Elaboración propia

En cuanto a los votantes del otro grupo de izquierdas, como se puede observar, coincide con los problemas indicados por los socialistas, en que la crisis económica y los problemas políticos en general son de sus principales preocupaciones, pero en este caso, el tercer problema más recurrente entre este colectivo, es la desigualdad de género (véase Figura 4.12).

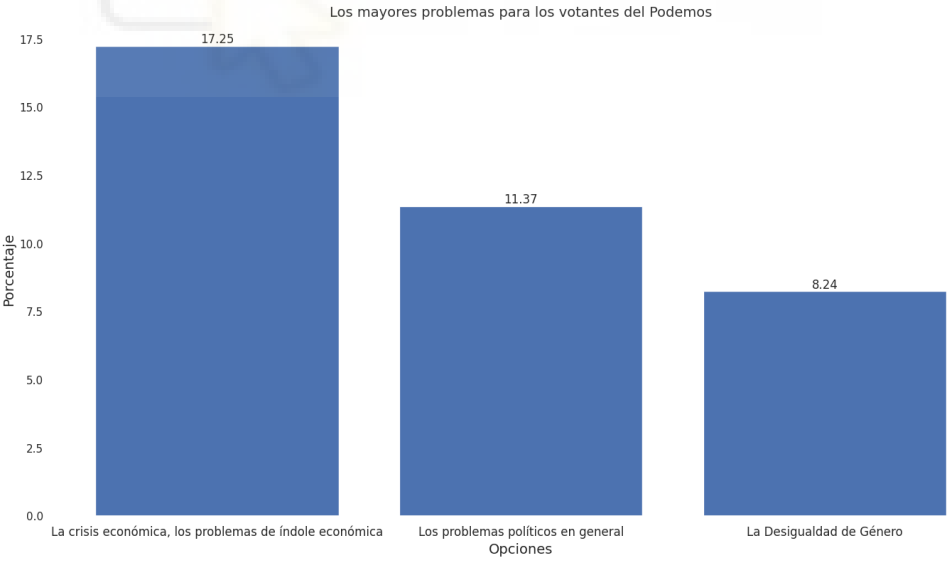

Figura 4.12: Representación de los principales problemas de España según los votantes de PODEMOS. Fuente: Elaboración propia

En cuarto lugar, corresponde a los votantes de VOX, los cuales priorizan la preocupación en temas políticos, dando lugar, en las primeras posiciones, a los problemas relacionados con el gobierno y partidos concretos, seguido de los problemas políticos en general y finalmente, la preocupación por la crisis económica, aspectos que se muestran en el siguiente gráfico.

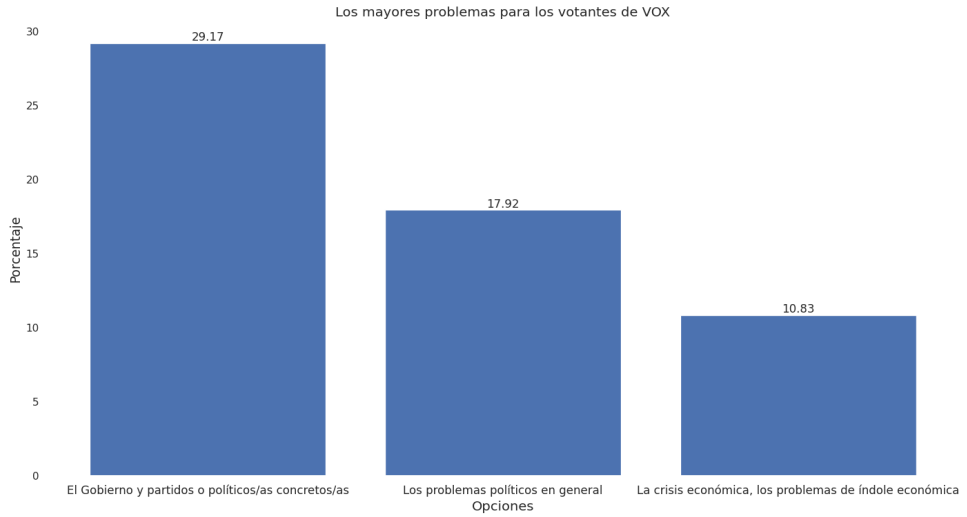

Figura 4.13: Representación de los principales problemas de España según los votantes de VOX. Fuente: Elaboración propia

Y por último, cabe mencionar cómo se comportan los encuestados pertenecientes al grupo de los indecisos, los cuales, como indica el siguiente gráfico (Figura 4.14), se asemejan bastante a las principales preocupaciones de los socialistas, aunque el orden de los problemas políticos en general y el paro están invertidos.

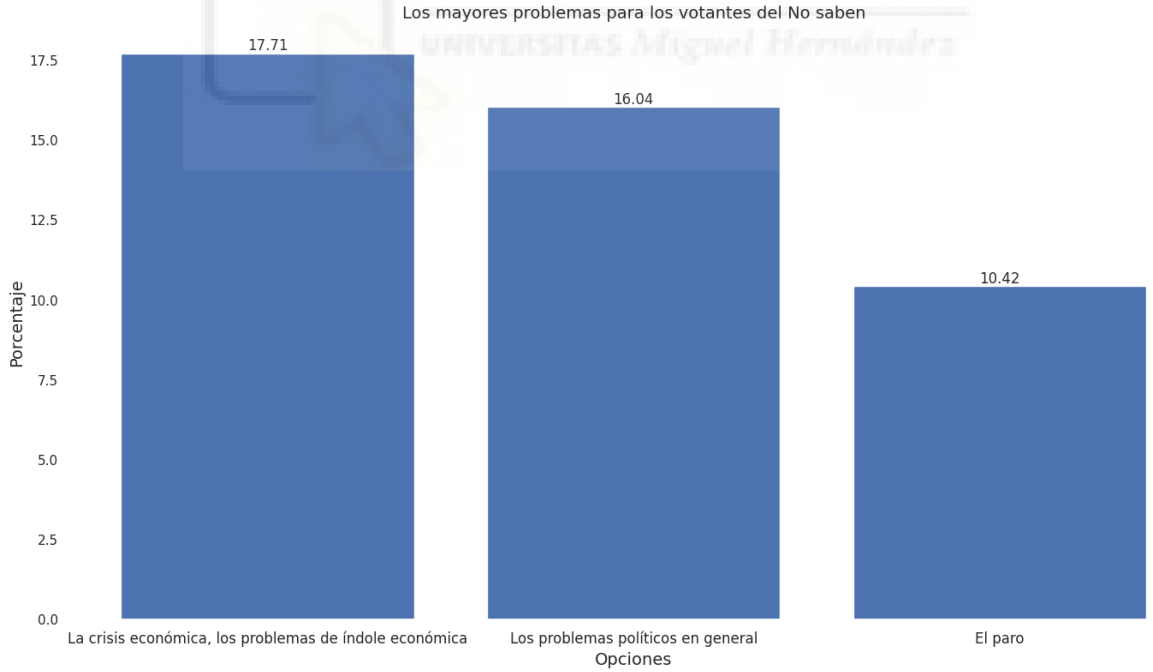

Figura 4.14: Representación de los principales problemas de España según los votantes indecisos. Fuente: Elaboración propia

En conclusión y a tenor de los resultados obtenidos, desde el partido más de izquierdas al partido más de derechas de España, pasando incluso por los indecisos, todos coinciden en dos de los tres problemas (la crisis económica y los problemas políticos en general), por lo que se

puede llegar a la conclusión de que la tendencia ideológica no define tanto la percepción de cuáles son los principales problemas para el país.

4.5.6.- Tendencias religiosas según el partido al que votan

A lo largo de la historia, diferentes gobiernos de España y la Iglesia Católica han trabajado para poder ejercer el culto de esta religión dentro de sus fronteras, pero en la actualidad, según el artículo publicado el 23 de abril de 2022 en el périodico La Razón [38], la religión católica parece estar perdiendo influencia entre los ciudadanos españoles. Hace 50 años, casi todos los españoles eran cristianos practicantes, pero en la actualidad, la realidad es otra y el número de católicos practicantes ha disminuido considerablemente. Por lo que surge la siguiente pregunta: ¿Está la religión alineada con una ideología política en concreto? En los siguientes gráficos, se representa de manera porcentual, un desglose de las creencias de los votantes de cada partido.

Por lo que en este punto se pretende analizar las variables del partido a que votan (P14) y la tendencia religiosa que precesión (P28).

En primer lugar, se analizan los votantes de VOX, donde, como se aprecia en la gráfica, más del 75% de ellos son católicos frente al otro 25% que no lo son. De ese 75% de católicos, se agrupan los no practicantes, con un 45,8% y los practicantes, con un 30%.

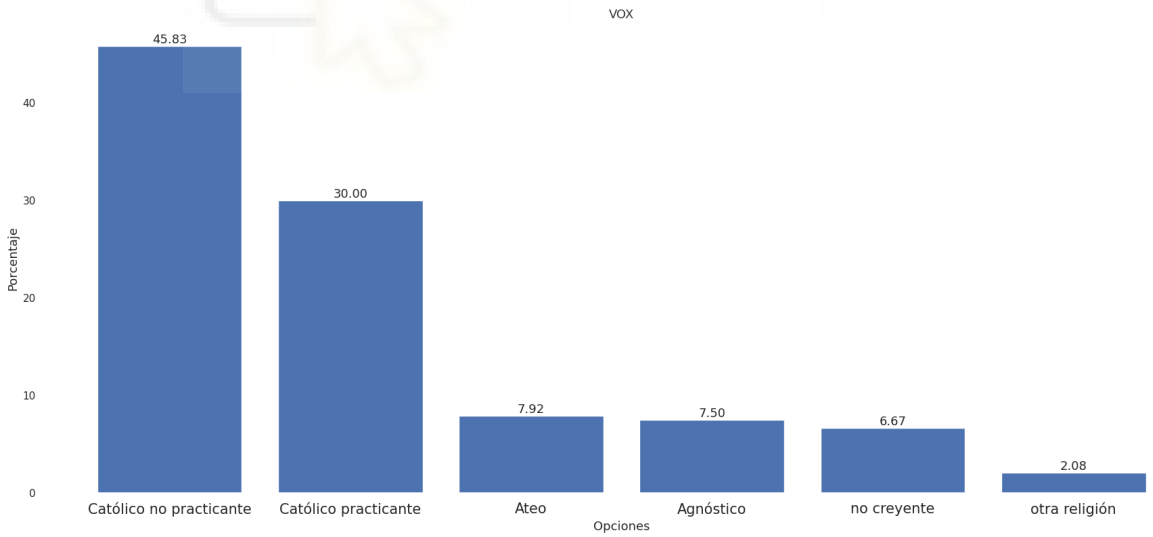

Figura 4.15: Representación religiosa según los votantes de VOX. Fuente: Elaboración propia

Los votantes del Partido Popular también son de creencias cristianas, incluso superando a los votantes de VOX en un 5% y declarándose como católicos el 80% de sus electores, aspecto que se aprecia en en la Figura 4.16.

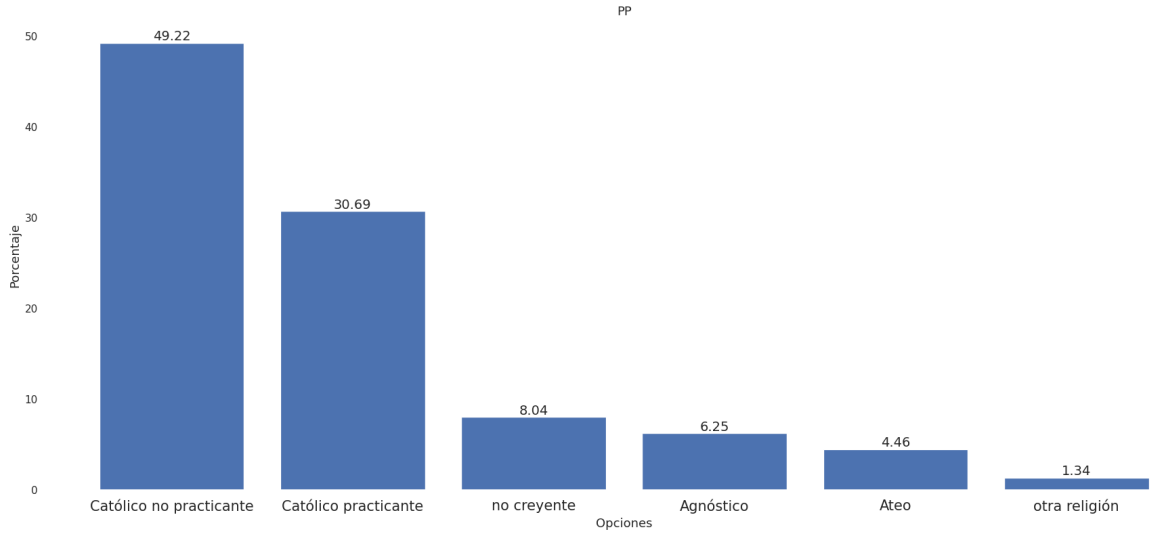

Figura 4.16: Representación religiosa según los votantes del PP. Fuente: Elaboración propia

El análisis de los votantes del Partido Socialista Obrero Español indica que, no obtiene ni siquiera el 50% de católicos entre sus votantes. Un 38,48% de ellos se declara católico no practicante y un 11,4% católico practicante, es decir que, aproximadamente, 1 de cada 2 socialistas se considera católico, resultados que se reflejan en la Figura 4.17.

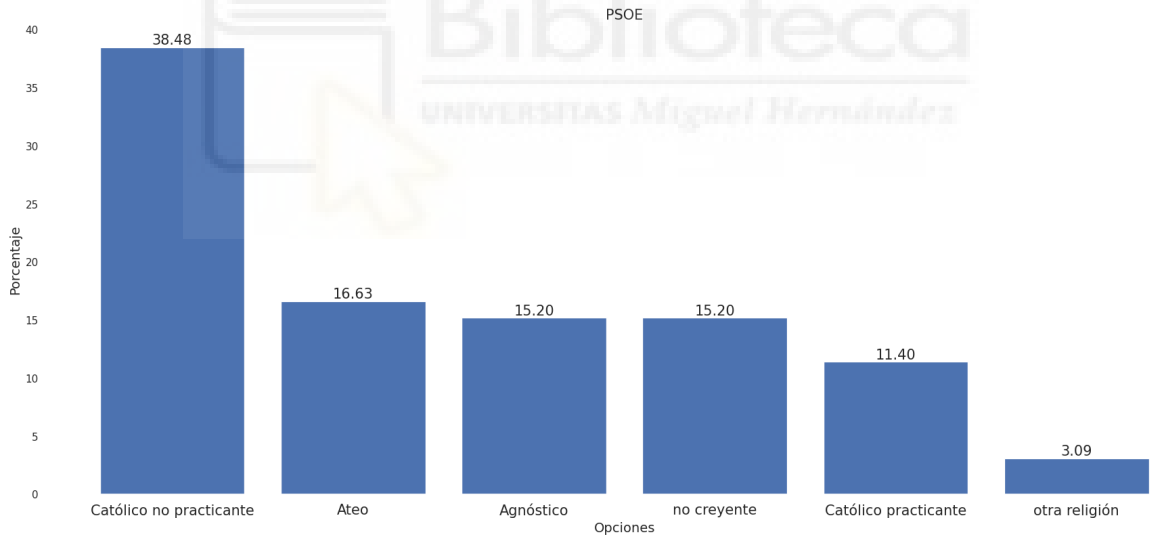

Figura 4.17: Representación religiosa según los votantes del PSOE. Fuente: Elaboración propia

Seguidamente, entre los votantes de PODEMOS se obtiene algo menos del 15% que declara ser católico no practicante, siendo sus votantes mayoritariamente ateos, con un 44,71%, y donde hay un porcentaje algo más elevado en otras confesiones, comparado con el porcentaje de católicos practicantes (véase Figura 4.18).

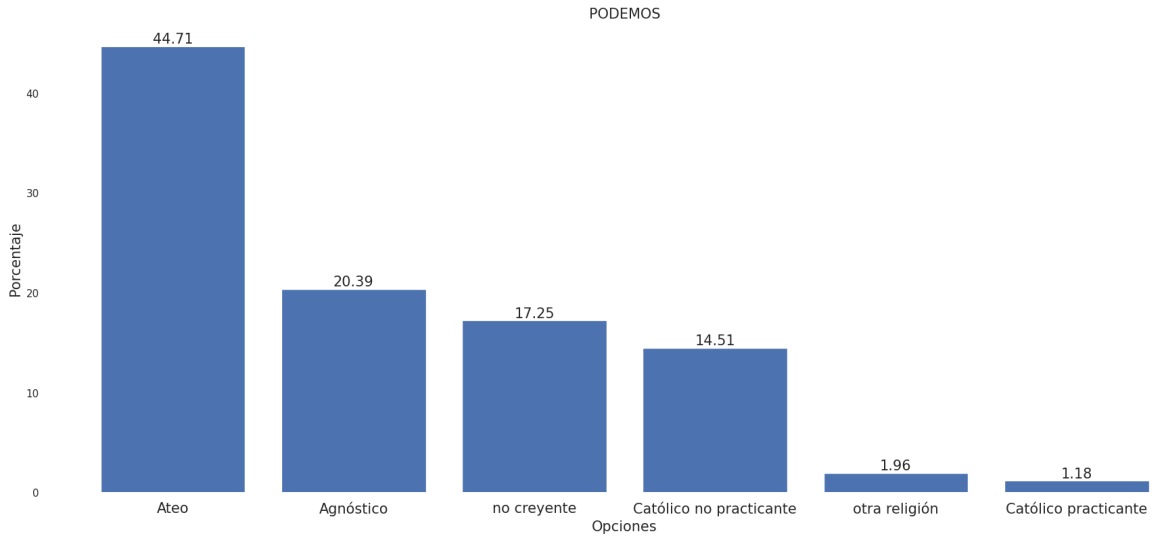

Figura 4.18: Representación religiosa según los votantes de PODEMOS. Fuente: Elaboración propia

Y por último, el grupo de los indecisos, el cual posee un 58% de católicos, (practicantes y no practicantes), porcentajes intermedios entre los votantes de VOX (75%) y los votantes del PSOE (49%), enfocándose más hacia el PSOE que hacia VOX.

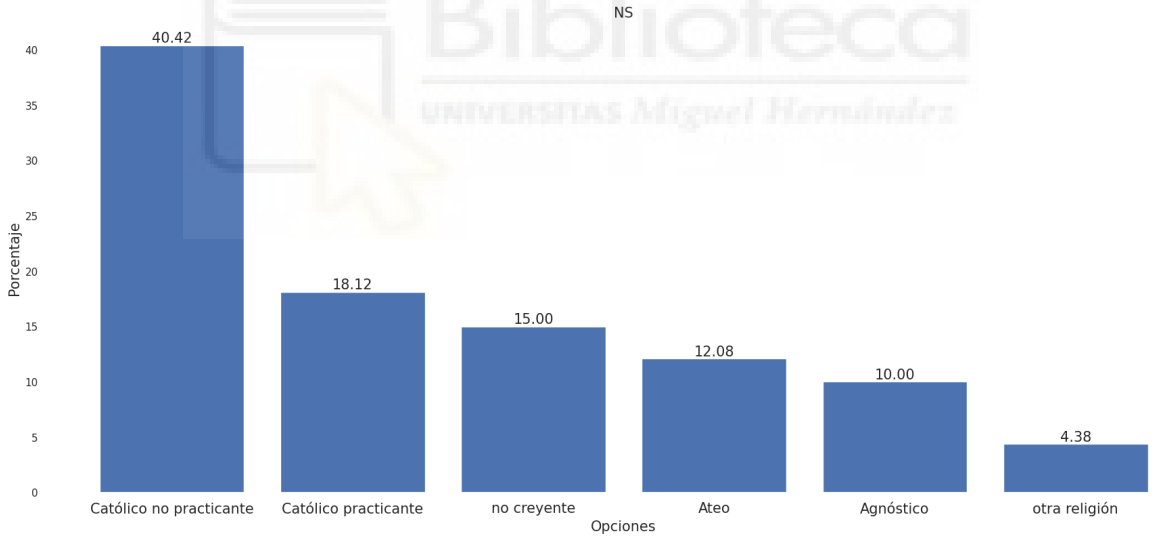

Figura 4.19: Representación religiosa según los votantes indecisos. Fuente: Elaboración propia

En conclusión y a la vista de los resultados, parece que sí puede haber cierto vínculo entre la tendencia religiosa y la ideología de una persona, donde ser creyente parece estar vinculado a ser de derechas, mientras que no serlo está más vinculado a ser de izquierdas. Aunque serlo o no serlo no es concluyente con la ideología, ya que como se ha visto, existen todos las tendencias dentro de todos los grupos políticos.

4.5.7.- Franjas salariales del hogar en función del voto

Las franjas salariales siempre han ido asociadas a cierto estatus social y económico que separa y divide a la sociedad en clases sociales. Según la publicación de la revista de cultura y pensamiento La U [39], las clases sociales más humildes, las que menos ingresos generan, han pertenecido a ciertos grupos o formaciones políticas, mientras que las clases sociales más adineradas han pertenecido a otros partidos políticos.

En la Figura 4.20 se encuentran representadas las densidades de las franjas salariales del hogar (P32) de los votantes de cada grupo político (P14), la representación se ha realizado mediante diferentes colores que recorren el gráfico por el eje X, resultados que corresponden a la pregunta P32 (véase Anexo 1 para mayor información).

La línea verde pertenece a los salarios de los hogares de los votantes de VOX, la línea roja pertenece a los salarios de los votantes de PSOE, la línea azul a los del PP y la línea amarilla a los de PODEMOS.

Apreciando la información del gráfico (Figura 4.20), se observa que, nuestros resultados arrojan que la mayoría de los votantes de VOX se sitúa al principio de la gráfica, en torno a los 1.800 - 1.700€, después decrece bastante rápido y aguanta con fuerza en las partes últimas de las gráficas, donde las franjas salariales son mayores (6.000-7.000€). El sueldo medio para los hogares de los votantes de VOX es de 2.743,16€, con una desviación típica de 1.687,21€.

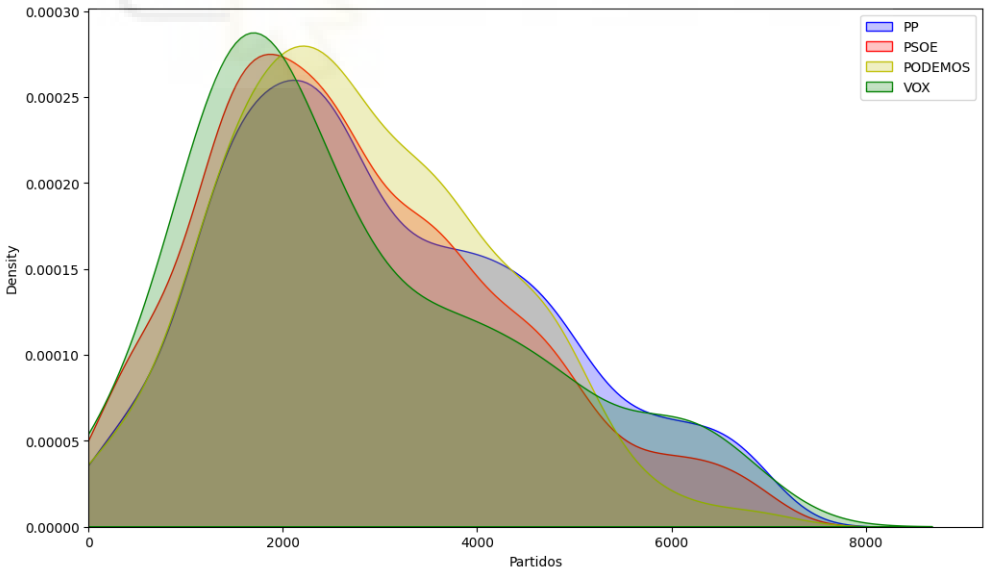

Figura 4.20: Representación salarial de los votantes según su partido. Fuente: Elaboración propia

El segundo partido con el pico de salarios más bajos lo tiene el PSOE, la moda de estos votantes se sitúa sobre los 1.900€ y se va reduciendo poco a poco a lo largo de toda la gráfica, teniendo una densidad significativa en los rango intermedios de 2.000 a 4.500€, y una

representación un poco menor comparada con otros partidos de derechas en las franjas finales del gráfico. Según los resultados de la encuesta, el sueldo medio para los hogares de un votante del PSOE es de 2.741,03€, con una desviación típica de 1.537,48€.

En cuanto al PP, su pico se sitúa en la franja salarial de los 2.000€, sufre un fuerte descenso hasta llegar a la franja de los 3.700€ donde se puede apreciar que reside bastante votante entre estas dos franjas hasta llegar a los 5.500€ donde PP y VOX acumulan mayoritariamente todo el voto de esta zona. El salario medio para los hogares de un votante del Partido Popular es de 3.028€, con una desviación típica de 1.624,75€.

Y por último se encuentra PODEMOS, donde su moda es la más atrasada con respecto a los tres partidos anteriores, por lo que, el grueso de los votantes de este partido, suele ganar más dinero que los votantes de los otros partidos, y lidera, en densidad, las franjas intermedias hasta llegar a los 4.000€, donde empieza a descender más bruscamente. El sueldo medio para los hogares de un votante de PODEMOS es de 2.786,09€ y su desviación típica es de 1.360,36€.

4.5.8.- Frecuencia de voto sobre un partido en concreto

Con la salida de nuevos partidos emergentes en esta última década, y la posible desilusión por los partidos de siempre, resulta de interés saber qué partido ha conseguido fidelizar mejor su marca y conseguir que los votantes confíen en los partidos tradicionales para depositar cada cuatro años su voto en ellos. La recurrencia del voto no suele ser lo normal en un partido político, ya que como se observa en la España actual, cada 4 años, como mucho 8 años, el gobierno de España sufre un cambio con unos pensamientos totalmente opuestos al partido que estaba anteriormente.

En este apartado, se analizan cuáles son los partidos con más votantes fanáticos por la formación (P14) y cuáles son los que sus votantes tienen el voto más volátil, es decir, su voto está relacionado con los intereses del momento (P11).

En primer lugar, se describe a los votantes de VOX (Figura 4.21), que como se aprecia, tres de cada cuatro votantes suyos, les votan porque es lo que más les convence en este momento, mientras que el otro 25%, sí que suele recurrir a ese partido a la hora de depositar el voto.

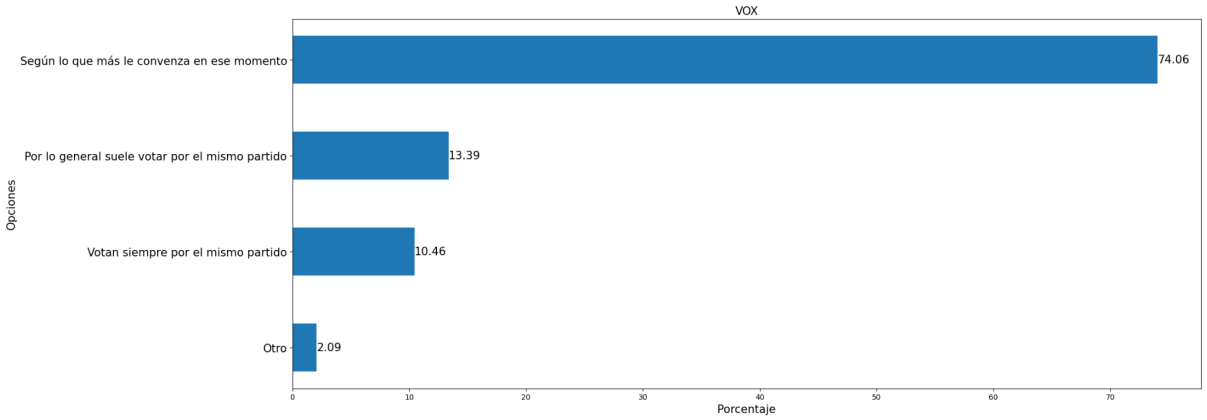

Figura 4.21: Representación de la frecuencia del voto de los votantes de VOX. Fuente: Elaboración propia

Por lo que, como se puede ver, VOX tiene un votante bastante "volátil" y por tanto, si las propuestas que exponen no resultan de interés puede sufrir una penalización de los votantes como le ha ocurrido a la formación de "Ciudadanos".

En segundo lugar, en el análisis de PODEMOS, se aprecia que baja a menos de la mitad su voto eventual, haciendo así que, aproximadamente, 1 de cada 2 personas que votan a PODEMOS indican en la encuesta que su voto siempre es al mismo partido, mientras que sólo un 45% de las personas, votan a PODEMOS porque es lo que más les interesa en ese momento.

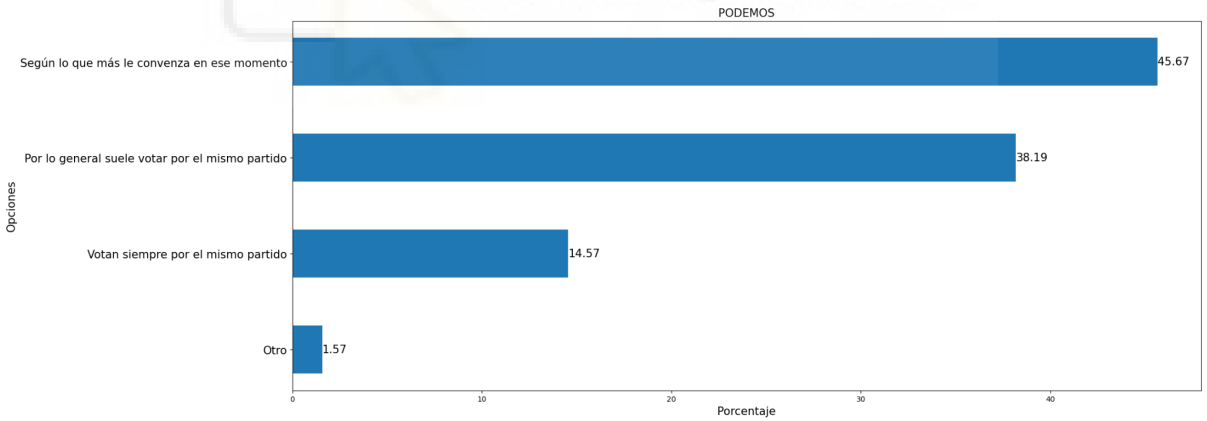

Figura 4.22: Representación de la frecuencia del voto de los votantes de PODEMOS. Fuente: Elaboración propia

En tercer lugar, se tiene al Partido Popular, el cual tiene similitudes en cuanto a la frecuencia del voto con VOX ya que como indica la siguiente figura (Figura 4.23), más del 60% de sus votantes, en función del momento, podrían cambiar su voto hacia otro partido. Este resultado podría explicar que cada cierto tiempo, el PP sufre una fuerte bajada en el número de votantes en las elecciones, mientras que otros partidos no descienden tanto. También es cierto que, casi el otro 40% de los votantes del PP, suele votar por el mismo partido, haciendo así una fidelidad en la marca de la formación, pero no tanto como la del siguiente grupo político.

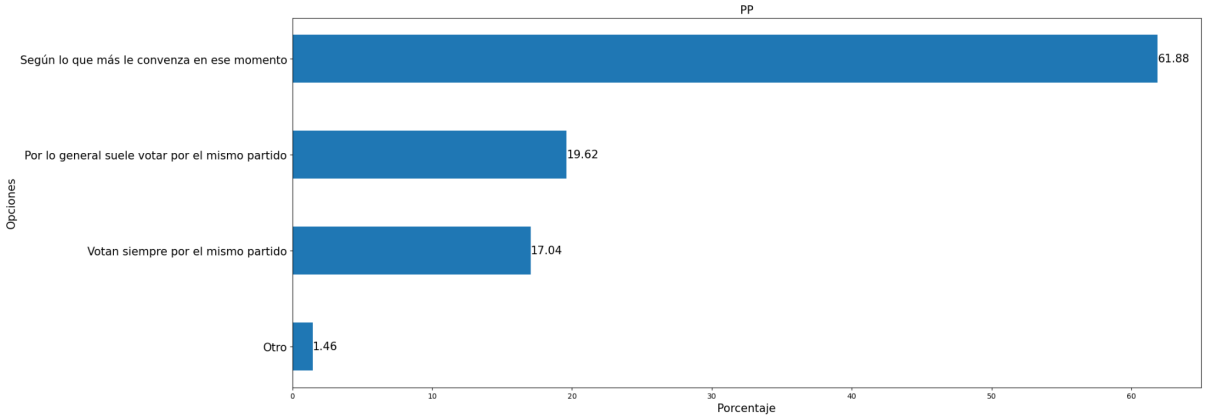

Figura 4.23: Representación de la frecuencia del voto de los votantes del PP. Fuente: Elaboración propia

Por último, en cuanto al PSOE, se observa que es el partido que mejor ha conseguido fidelizar a sus votantes. Como se indica en la Figura 4.24, 2 de cada 3 votantes de esta formación, volverán a votar al PSOE pase lo que pase en la política española, y sólo 1 de cada 3 votarán al PSOE en función de lo que más les interese en ese momento.

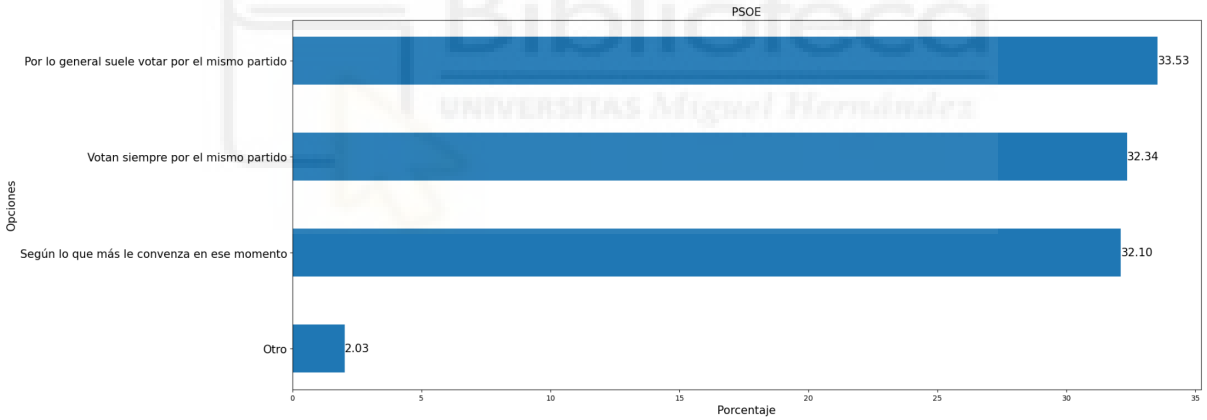

Figura 4.24: Representación de la frecuencia del voto de los votantes del PSOE. Fuente: Elaboración propia

A modo de conclusión, se puede decir que, el Partido Socialista es el que mejor se comporta a la hora de conseguir mantener a sus votantes entre sus filas, de igual modo, PODEMOS también consigue retener, con cierto éxito, a dichos votantes, mientras que, en los partidos de derechas abunda más la indecisión y el interés del momento a la hora de confiar el voto en estos grupos políticos.

4.6.- Relación entre las variables del dataset

La relación entre las variables del dataset es un aspecto muy importante en el análisis de datos. Un conjunto de datos está compuesto por diversas variables que representan diferentes medidas o características de interés en el conjunto de observaciones. Estas variables pueden estar relacionadas o no de diversas formas, por lo que proporcionan información clave a la hora de entender los patrones, las interacciones presentes en los datos y las tendencias.

La exploración de la relación entre variables tiene como objetivo principal identificar posibles dependencias, asociaciones o correlaciones que puedan existir entre ellas. Estas relaciones pueden ser lineales o no lineales, positivas o negativas, directas o indirectas, y pueden revelar conexiones significativas o patrones subyacentes en los datos.

En este proyecto se ha aplicado un análisis de correspondencias simple (ACS), donde se analiza la posible asociación entre aquellas variables que se han considerado más interesantes para el estudio. Estas variables han sido analizadas una frente a otra para comprobar si estaban relacionadas, es decir, se ha realizado un contraste de hipótesis donde, si se da el caso de que no tienen relación, se dice que no se rechaza la hipótesis nula de independencia entre variables, y si se rechaza la hipótesis nula, indica que existe una relación estadísticamente significativa entre las variables.

El primer análisis que se ha realizado es valorar si las variables, "sueldo hogar" y "partido al que vota el encuestado" están relacionadas.

¿De cuántos ingresos al mes disponen en su hogar, después de la deducción de impuestos (o sea, ingresos netos)?

P14 → P32 ↓	PSOE	PP	PODEMOS	VOX
Menos de 1.100€	10,88%	1,97%	7,94%	10,78%
De 1.100 a 1.800€	20,90%	20,10%	17,86%	26,72%
De 1.801 a 2.700€	24,69%	26,39%	26,98%	22,41%
De 2.701 a 3.900€	22,00%	20,10%	26,59%	15,52%
De 3.901 a 5.000€	13,20%	18,00%	15,87%	13,36%
Más de 5.000€	8,31%	13,44%	4,76%	11,21%
Total	100,00%	100,00%	100,00%	100,00%
$Chi^2 = 44.22$			P = 0,0001	

Tabla 4.3: Tabla de contingencia de las variables P14 y P32. Fuente: Elaboración propia

Tras analizar la tabla de contingencia de las variables "sueldo hogar" y "partido al que vota" (Tabla 4.3), es decir, P14 y P32 (véase anexo), se obtiene un valor de Chi^2 de 44.22 y un pvalor de 0,0001. De forma que, se puede afirmar que existe una asociación estadísticamente significativa entre las dos variables.

Como se aprecia tanto en la tabla (Tabla 4.3) como en el mapa de correspondencias (Figura 4.23), se corrobora la afirmación de que los votantes del PP son votantes con dinero, sin embargo VOX, siendo un grupo político de derechas, recoge también a un sector significativo que cobra menos de 1100€.

Mientras que el grupo socialista es una mezcla de todas las franjas salariales, el grupo de PODEMOS tiene una fuerte relación con las franjas intermedias/altas de las bandas salariales, que suelen corresponder a salarios habituales de funcionarios que ocupan cierto tiempo en el puesto de trabajo y tienen complementos salariales como trienios y sexenios.

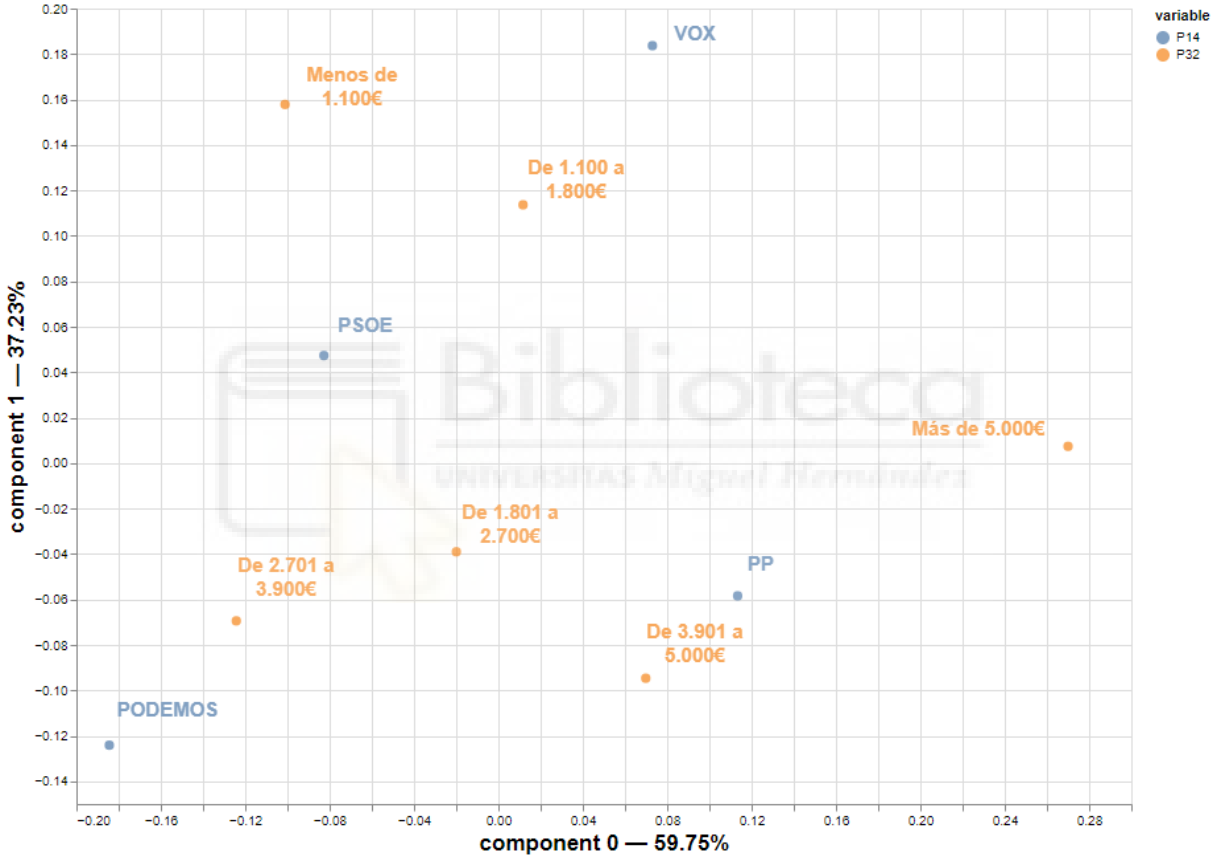

Figura 4.25: Representación de la asociación entre las variables P14 y P32. Fuente: Elaboración propia

El segundo análisis que se ha realizado se valora si las variables, "Cuándo decide a qué partido votar" y "A qué le da más importancia a la hora de votar" están relacionadas.

Y a la hora de votar en unas elecciones, ¿a qué le da Ud. más importancia?

P12 → P13 ↓	Aún no ha votado en ningunas elecciones	Lo decide mucho antes del inicio de la campaña electoral	Lo decide al comienzo de la campaña electoral	Lo decide durante la última semana de la campaña electoral	Lo decide durante la jornada de reflexión, la víspera de las elecciones	Lo decide el mismo día de las elecciones
A todo por igual	0,00%	8,11%	3,65%	4,14%	3,17%	0,00%
Al programa	0,00%	5,74%	5,94%	6,21%	4,76%	4,17%
Al partido político	83,33%	52,97%	55,71%	47,93%	50,79%	52,08%
Al/a la candidato/a	16,67%	33,18%	34,70%	41,72%	41,27%	43,75%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
$Chi^2 = 28,53$					P = 0,018	

Tabla 4.4: Tabla de contingencia de las variables P12 y P13. Fuente: Elaboración propia

Tras analizar la tabla de contingencia de las variables, "Cuándo decide qué votar" y "A qué le da más importancia" (Tabla 4.4.), es decir, P12 y P13 (véase Anexo I), se obtiene un valor de Chi^2 de 28,53 y un pvalor de 0,018. Por lo que existe una asociación estadísticamente significativa entre las variables analizadas.

A tenor de los resultados mostrados tanto en la tabla (Tabla 4.4) como en el gráfico (Figura 4.26), las personas que aún no han votado en ningún comicio, se dejan influenciar de manera considerable por las siglas del partido, se puede observar que, el 83,33% de estos encuestados indican que para ellos lo más importante es el partido político, mientras que en el resto de opciones, este porcentaje disminuye hasta estar cercano al 50%, es decir, a parte del partido político, también les dan una importancia considerable al candidato y al programa. Por otra parte, la opción asociada al candidato toma un mayor peso cuanto más cerca está la fecha de las elecciones.

Finalmente, se observa que, la importancia que los votantes le dan al programa electoral y a todo por igual, prácticamente no toma casi ninguna relevancia en ninguno de los casos.

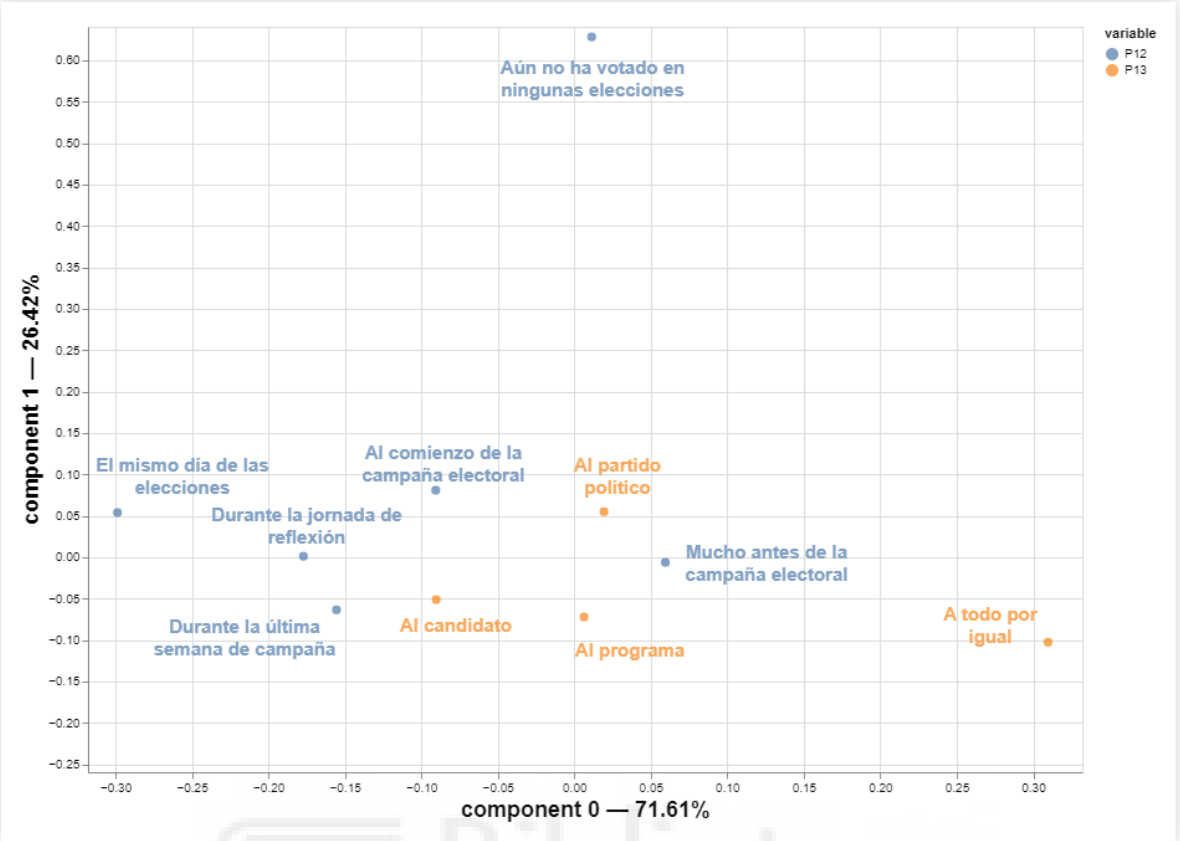

Figura 4.26: Representación de la asociación entre las variables P12 y P13. Fuente: Elaboración propia

En el tercer análisis que se ha realizado se desea valorar si las variables, "Tamaño del municipio" y "A que le dan más importancia a la hora de votar" están relacionadas.

¿Podría decirme el tamaño del municipio en el que reside?

TAMMUN → P13↓	Menos o igual a 2000 habitantes	2001 a 10000 habitantes	10001 a 50000 habitantes	50001 a 100000 habitantes	100001 a 400000 habitantes	400001 a 1000000 habitantes	Más de 1000000 habitantes
A todo por igual	5,31%	8,93%	7,84%	5,67%	5,71%	5,70%	6,10%
Al programa	8,85%	4,47%	5,40%	4,33%	5,92%	8,29%	6,57%
Al partido político	45,13%	48,80%	49,13%	54,33%	56,53%	58,55%	51,64%
Al candidato	40,71%	37,80%	37,63%	35,67%	31,84%	27,46%	35,68%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	$Chi^2 = 25.11$					P = 0,12	

Tabla 4.5: Tabla de contingencia de las variables TAMMUN y P13. Fuente: Elaboración propia

Tras analizar la tabla de contingencia de las variables "tamaño del municipio" y "A qué le da más importancia", es decir, TAMMUN y P13, se obtiene un valor de Chi^2 de 25.11 y un valor de 0,12. Por lo que, al ser superior al 0,05, se descarta que haya asociación entre ambas.

Como se observa, al igual que en la relación anterior, lo que más predomina entre los diversos municipios a la hora de votar, es el partido y le sigue de cerca el candidato, quedando así reflejado que los municipios, tanto pequeños como grandes, al menos a la hora de votar a nivel nacional, no eligen, mayormente, en función del candidato sino del partido. Pero también señalar que, a medida que el número de habitantes aumenta, parece que el candidato pierde cierta relevancia y gana algo más el partido político, aunque estas diferencias no son estadísticamente significativas.

El cuarto análisis que se ha realizado se valora si las variables, "Preocupación por el medio ambiente" y "Partido al que vota el encuestado" están relacionadas.

¿Diría Ud. que en estos momentos el cambio climático le preocupa mucho, bastante, poco o nada?

P14 → P6 ↓	VOX	PP	PSOE	PODEMOS
Bastante	23,85%	41,47%	48,63%	43,53%
Mucho	7,53%	21,13%	39,90%	50,59%
Regular	0,42%	1,02%	0,72%	0,78%
Poco	30,54%	28,25%	9,32%	3,92%
Nada	37,66%	8,14%	1,43%	1,18%
Total	100,00%	100,00%	100,00%	100,00%
	$Chi^2 = 608,98$		$P < 0,001$	

Tabla 4.6: Tabla de contingencia de las variables P14 y P6. Fuente: Elaboración propia

En cuanto a la tabla de contingencia de las variables "Partido al que votan" e "Importancia por el medio ambiente", es decir, P14 y P6 (véase Anexo), se obtiene un valor de Chi^2 de 608,98 y una $p < 0,001$. Por lo que existe una asociación estadísticamente significativa entre estas dos variables.

Como se observa en la tabla (Tabla 4.6) y en el gráfico (Figura 4.27), existe una clara tendencia, cuanto más a la derecha se sitúa la persona en el voto, menos le importa el medio ambiente, como es el ejemplo de los votantes de VOX, donde al 37,66% de ellos no les importa nada el medio ambiente, y cuanto más a la izquierda se sitúa esta persona, mayor preocupación le genera este problema, como es el ejemplo de los votantes de PODEMOS, donde al 43,53% les importa bastante el medio ambiente.

Esta asociación también queda reflejada en el mapa de correspondencias, indicando que los votantes de VOX no le dan nada de importancia a los temas de medio ambiente, a los votantes del PP muy poca o regular importancia, y así en aumento, conforme más a la izquierda se sitúa el partido político.

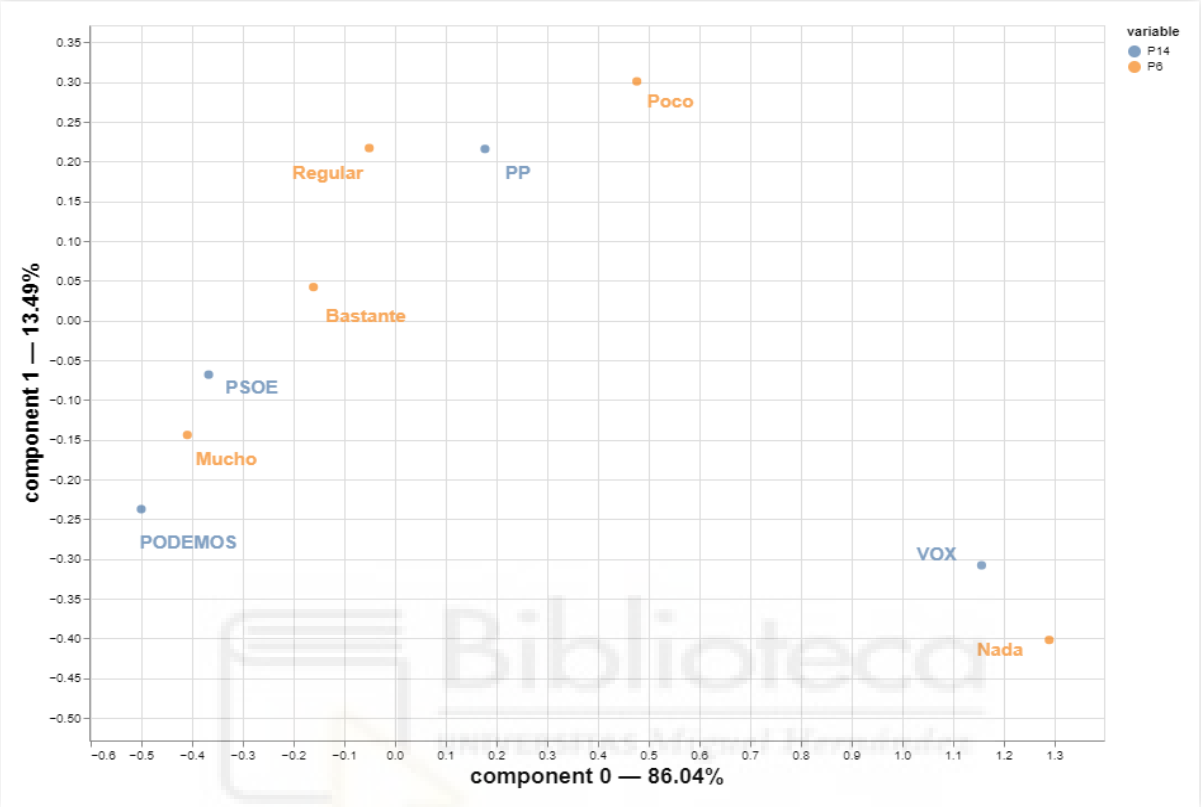

Figura 4.27: Representación de la asociación entre las variables P14 y P6. Fuente: Elaboración propia

Y por último, en el quinto análisis que se ha realizado se valora si las variables, "A que le da más importancia a la hora de votar" y "partido al que vota el encuestado" están relacionadas.

Y a la hora de votar en unas elecciones, ¿a qué le da Ud. más importancia?

P14 → P13 ↓	VOX	PP	PSOE	PODEMOS
A todo por igual	5,56%	7,01%	7,33%	4,76%
Al programa	9,40%	4,48%	5,38%	8,33%
Al partido político	43,16%	41,72%	62,47%	64,68%
Al/a la candidato/a	41,88%	46,78%	24,82%	22,22%
Total	100,00%	100,00%	100,00%	100,00%
$Chi^2 = 132,57$			$P < 0,001$	

Tabla 4.7: Tabla de contingencia de las variables P14 y P13. Fuente: Elaboración propia

Finalmente, el análisis de la tabla de contingencia de las variables "Partido al que votan" y "A que le da más importancia", es decir, P14 y P13 (véase Anexo), se obtiene un valor de Chi^2 de 132,57 y una $p < 0{,}001$. Por lo que existe una asociación estadísticamente significativa entre las variables estudiadas.

Como se observa en la tabla (Tabla 4.7) y en el mapa de correspondencias (Figura 4.28), existe, por parte de las izquierdas, un interés mayor por las siglas del partido (64,68% PODEMOS y 62,47% PSOE), mientras que las derechas y en especial el Partido Popular, se aprecia más el peso en la figura del candidato (46,78% PP y 41,88% VOX), aspecto que las izquierdas dejan como algo más residual.

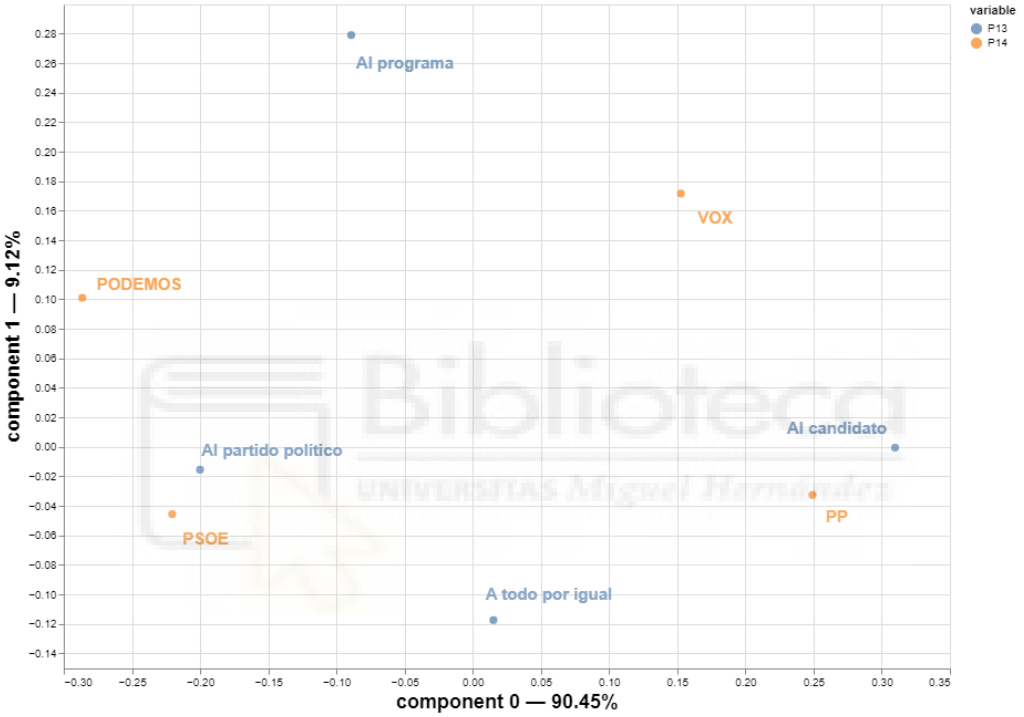

Figura 4.28: Representación de la correlación entre las variables P14 y P13. Fuente: Elaboración propia

4.7.- Selección de los atributos

A la hora de trabajar con un dataset se debe tener claro qué es lo que se pretende conseguir y cuales son las mejores variables para llegar a ese resultado.

En este TFG, uno de los objetivos principales es predecir mediante técnicas de Machine Learning, cuál es la intención del voto de los ciudadanos españoles. En este estudio se cuenta con un dataset con 57 variables distintas y como es lógico, no todas las variables tienen el mismo peso a la hora de aplicar el modelo, por lo que de manera previa a la aplicación de las técnicas y algoritmos de clasificación, se realiza una selección de los k mejores atributos para así ayudar al algoritmo del árbol de clasificación y reducir el ruido de este.

En esta ocasión y con el fin de hacer algo sencillo y visual se procede a obtener los 10 mejores atributos de todo el dataset para predecir la variable objetivo, es decir, a qué partido votarían a día de hoy si se diese el caso de unas elecciones generales (P14, véase Anexo I).

En primer lugar, se debe elegir el dataset con el que se realiza la selección de atributos, para el caso del presente estudio, se han realizado cuatro análisis de selección de atributos distintos, y posteriormente, se han realizado cuatro árboles de clasificación haciendo uso de la selección de atributos, cada uno con un nivel de acierto diferente.

Cada árbol tiene características diferentes al resto de árboles, donde dos de esos árboles clasifican a los votantes de los cuatro partidos mencionados anteriormente, mientras que los otros dos árboles sólo emplean los datos de los votantes del PP y PSOE. Además, de los cuatro árboles, dos están compuestos por preguntas elegidas por el algoritmo de selección de atributos y dos por preguntas que el autor del TFG ha considerado interesantes para realizar la clasificación de los votantes.

Para el primer análisis de clasificación, se ha elegido un dataset donde solamente existen votantes de VOX, PP, PSOE y PODEMOS y se han eliminado algunas de las preguntas más conflictivas, es decir, preguntas con sesgo ideológico. Por tanto, no se consideran para esta selección de atributos las siguientes preguntas: ¿Quién preferiría que fuese el presidente del gobierno?, ¿Qué confianza le inspira Pedro Sanchez?, ¿Qué partido considera más cercano a sus ideas? y ¿Qué partido votó en las pasadas elecciones? Esta última es la variable objetivo del análisis de clasificación que se explicará en la siguiente sección y por tanto, se guarda en una variable aparte para poder predecirla en dicho análisis.

La función SelectKBest [40] se utiliza en la selección de características para el aprendizaje automático. Esta función permite seleccionar un número específico de las mejores características del conjunto de datos basado en una función de puntuación. Esto ayuda a reducir la dimensionalidad del conjunto de datos y mejorar la precisión del modelo.

La segunda función utilizada es la función chi2 que se emplea para realizar una prueba chi-cuadrado de independencia en los datos, lo que permite evaluar la relación entre dos variables categóricas. Es comúnmente utilizada en la selección de características en el aprendizaje automático y la minería de datos.

Lo primero que se debe hacer es crear una instancia del objeto SelectKBest, donde el score_func se establece en chi2, lo que indica que se utiliza el método de la prueba chi-cuadrado para evaluar las relevancias de las características y el parámetro k = 10 ya que es el número de atributos a seleccionar. Seguidamente, se debe ajustar el modelo de SelectKBest a los datos de entrada X (compuesto por preguntas económicas, sociales, salariales, estudios…) y la variable objetivo (voto en las próximas elecciones) utilizando el método fit. Lo que hace calcular la relevancia de cada característica con respecto a la variable objetivo.

Finalmente, sabiendo cuáles son las variables con más relevancia del dataframe se trasladan a un dataframe nuevo que únicamente llevará esas columnas.

El hecho de eliminar preguntas que dirigen bastante la ideología hace que nuevas variables, cuestiones que no son tan directas, pasen a liderar el ranking de las variables principales, el cual hará que baje el porcentaje de acierto.

Y por último, las preguntas que conforman los datos de entreno en los análisis 3 y 4, que corresponden a un dataframe de 4 partidos y de 2 partidos,son elegidas por el autor del TFG con la intención de comprobar si, eligiendo preguntas interesantes para el estudio, se puede conseguir un árbol más interesante sin que el nivel de precisión de acierto se vea gravemente reducido.

Para estos dos últimos análisis se ha optado por agregar al dataframe preguntas tales como: ¿Cuál es su edad?, ¿Cree que las manifestaciones del 8M sirven de algo?, ¿Cómo se define usted en materia religiosa? y ¿De cuántos ingresos dispone al mes en su hogar? Entre otras preguntas. Los resultados del análisis de selección de atributos y los análisis de clasificación se especificarán en el siguiente apartado.

4.8.- Árboles de clasificación

Los árboles de clasificación se han utilizado en este proyecto como una técnica de aprendizaje automático para abordar el desafío de predecir la intención de voto en las próximas elecciones.

Mediante la aplicación de algoritmos de árboles de clasificación se aborda el objetivo principal del presente Trabajo Fin de Grado, es decir, lo que se busca es tratar de conseguir un árbol que, mediante una serie de preguntas, sea capaz de clasificar al individuo en un grupo político y predecir, con una cierta capacidad de acierto, al partido que votaría.

El árbol consiste en un nodo principal donde se tiene agrupada a toda la muestra y este se va segmentando en diferentes ramas, hasta llegar al nodo hoja, donde clasifica a los individuos en función de la variable objetivo, es decir, el partido político al que votarían, y muestra las agrupaciones de los distintos votantes en cada hoja.

La explicación que se va a proceder a dar es aplicable a los cuatro análisis anteriormente mencionados, la única diferencia es en el cambio de los datos de entrenamiento, pero el código es el mismo para todos.

En primer lugar, se procede a explicar cómo se ha obtenido el árbol de clasificación y que técnica de medición de precisión se ha empleado, para posteriormente centrar el estudio en los resultados que arrojan los árboles de clasificación e interpretarlos.

Para obtener el árbol de clasificación se debe dividir el dataset en los datos de entrenamiento y los de prueba, para ello se utiliza la función train_test_split, donde se requieren unas características de entrada (X) y la variable objetivo (Y) representa. Asimismo, la función requiere de un parámetro llamado Random_state que se establece a 1 para poder garantizar la reproducibilidad de la división de los datos.

En segundo lugar, se crea una instancia del objeto DecisionTreeClassifier, donde el criterio se establece en 'entropy', lo que significa que el algoritmo buscará las divisiones que maximicen la ganancia de información, es decir, que reduzcan la impureza y aumenten la homogeneidad en los nodos del árbol. En este caso, también se requiere el parámetro 'max_depth' que limita la profundidad máxima del árbol, donde para estos casos, se establece en 4 niveles.

En siguiente lugar, se debe ajustar el modelo del árbol a los datos de entrenamiento utilizando el método fit. De este modo, el modelo irá aprendiendo a partir de los datos de entrenamiento y empezará a construir el árbol de decisión en función de las características y de la variable objetivo. Por último, el algoritmo almacena los resultados de sus predicciones basándose en el árbol anteriormente construido.

Para medir la precisión del algoritmo, se ha utilizado la variable f1-score que representa la media armónica entre precisión y 'recall' [23] (véase Capítulo 2), y proporciona un equilibrio entre ambas medidas. Es importante utilizar esta medida frente a otras medidas de precisión como el 'accuracy' ya que nuestra muestra está desbalanceada, habiendo así más respuestas por parte del PP y PSOE que de los votantes de VOX y PODEMOS, por lo que utilizar dicha variable garantiza una representación adecuada para todos los partidos, sean grandes o pequeños.

También es importante interpretar la matriz de confusión para entender cuales son los resultados y porcentajes que arroja el algoritmo.

4.8.1.- Primer árbol de clasificación

El siguiente árbol, está compuesto por los cuatro partidos principales (PSOE, PP, VOX y PODEMO). En este caso, se han eliminado todas las preguntas que resultan más conflictivas que pueden dejar entrever la ideología política, se han eliminado preguntas tales como P17 (escala del 1 al 10 según tu ideología) o P21.a (Partido al que votó las pasadas elecciones) ya que marcan demasiado la ideología.

Cabe destacar que para este caso, el dataframe para entrenar el árbol está compuesto por las columnas P23, P28, P24, P6, P3, P29, P5, EDAD, P11 y P2 (véase Anexo I) que han sido extraídas conforme al orden que las ha obtenido el algoritmo de selección de atributos.

En consecuencia, al elegir estas preguntas tan poco determinantes, la precisión del algoritmo es del 73% para los votantes del PP y PSOE, del 21% para los votantes de VOX y del 0% para los votantes de PODEMOS.

		Votantes pronosticados			
	Matriz de Confusión	PSOE	PP	VOX	PODEMOS
Votantes Reales	PSOE	148	30	1	0
	PP	41	154	6	0
	VOX	6	39	7	0
	PODEMOS	29	4	0	0

Tabla 4.8: Matriz de confusión del primer árbol de clasificación. Fuente: Elaboración propia

Como se puede observar en la Tabla 4.8 y en el árbol generado (Figuras 4.29 y 4.30), no resulta posible predecir cuales son los votantes de PODEMOS, ya que el algoritmo los clasifica de manera incorrecta dentro del partido socialista. También mencionar que bastantes votantes del PP los clasifica dentro del PSOE y a muchos votantes de VOX los clasifica dentro del Partido Popular, mientras que, la clasificación de votantes del PSOE dentro de otros partidos no es tan alta.

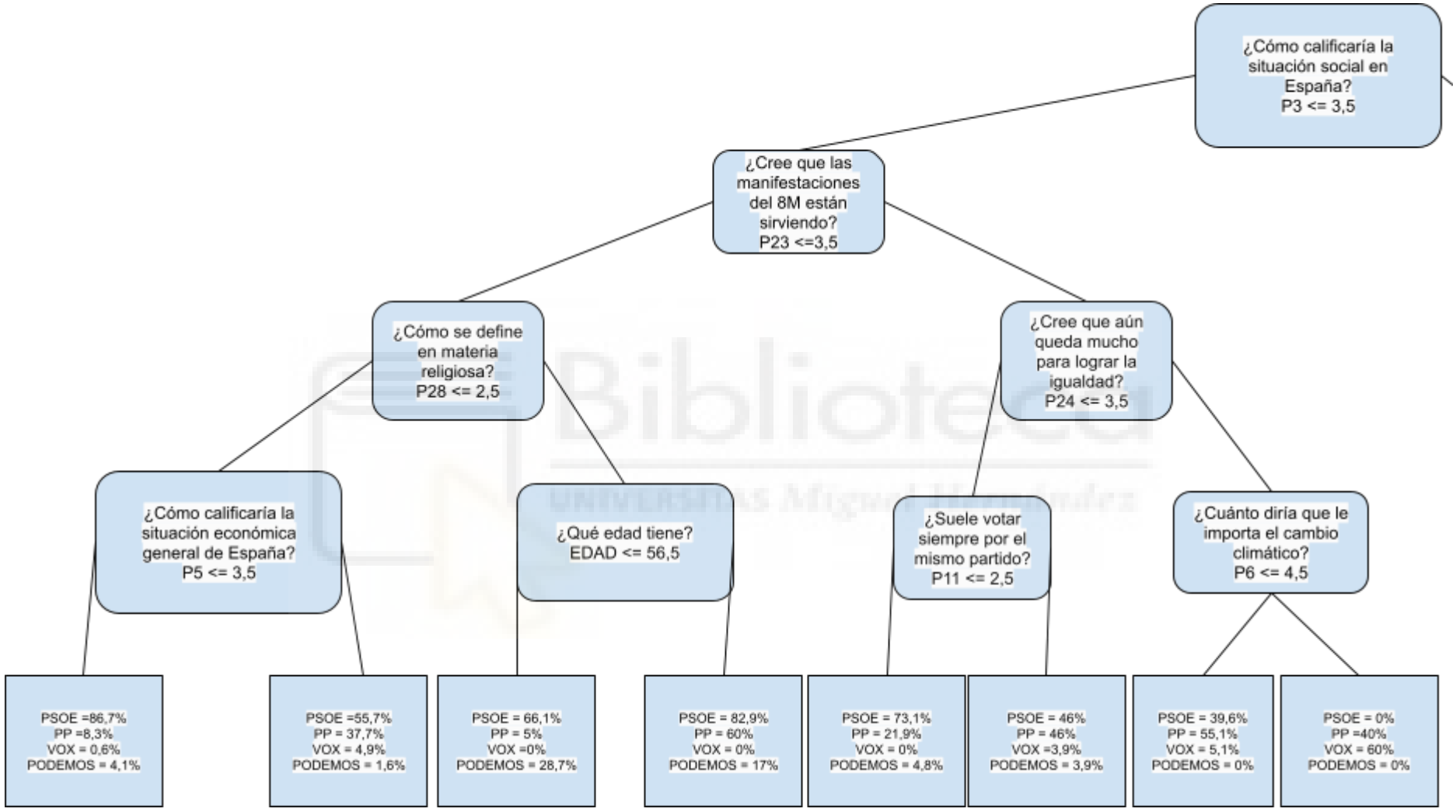

Figura 4.29: Árbol de clasificación 1 Parte I. Fuente: Elaboración propia

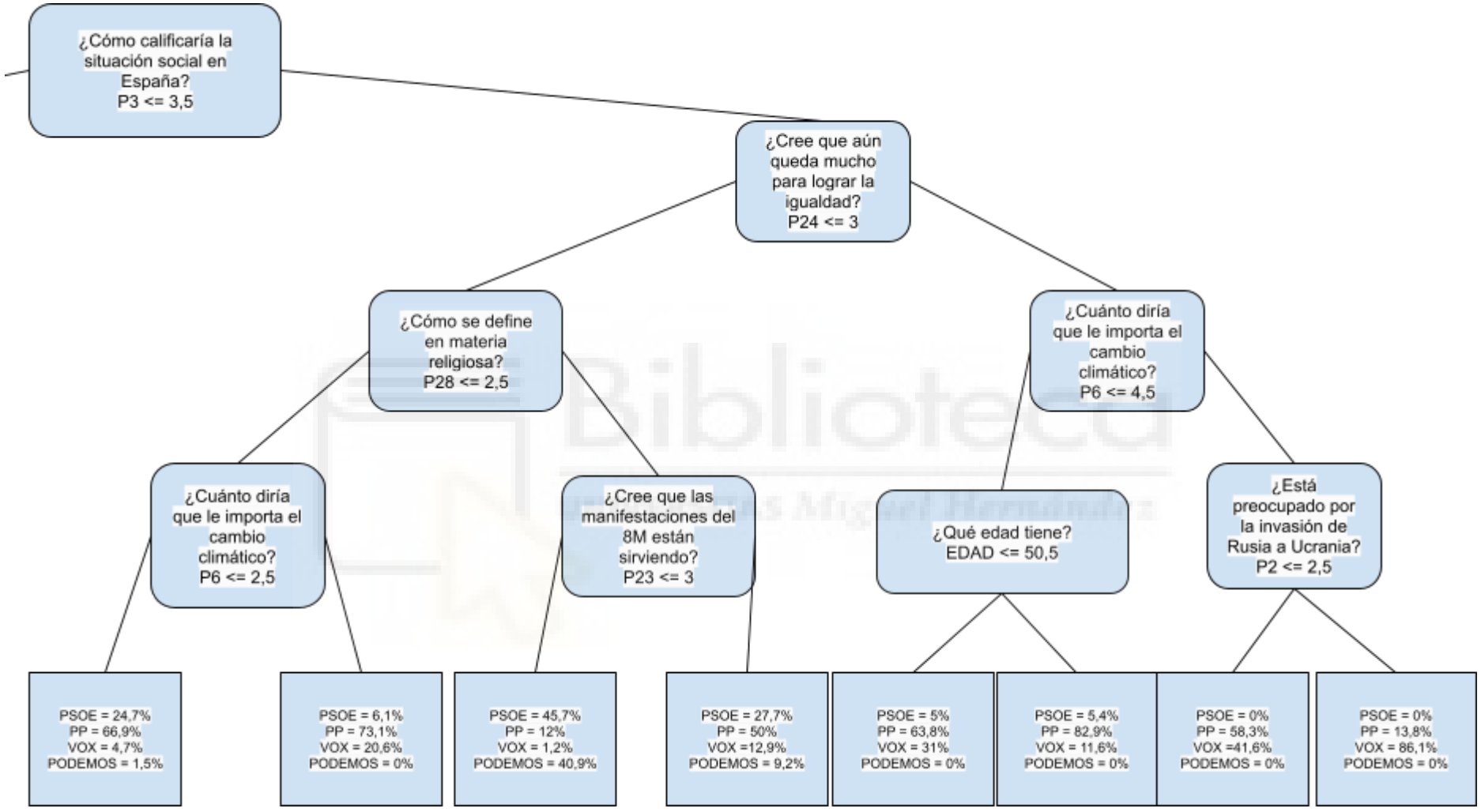

Figura 4.30: Árbol de clasificación 1 Parte II. Fuente: Elaboración propia

En las figuras anteriores (Figura 4.29 y Figura 4.30), se puede observar un árbol de clasificación para categorizar a un grupo de personas encuestadas a un determinado partido político. En la figura presentada, en el primer nodo se pregunta al entrevistado que realice una valoración de la situación social de España (P3), donde en el caso de ser la respuesta 3 o menos (muy buena, buena o regular), se realiza un desvío hacia la izquierda al cumplirse la condición del nodo, y derecha en caso de que diga 4 o más (mala o muy mala). A continuación, para este caso, se va a seguir el trazado por la rama de la derecha, es decir, como si hubiese respondido 4 o más. La siguiente cuestión que se le realiza al encuestado es sobre si cree que aún queda mucho para lograr la igualdad (P24), donde si responde con un 3 o menos (queda bastante, mucho o regular) se realizaría el desvío hacia la izquierda al cumplir la condición del nodo y hacia la derecha si dijese 4 o más, al no cumplirla (queda poco o ya existe la igualdad). Para este caso, el encuestado dice que le importa 5 (ya existe dicha igualdad), por lo que al no cumplirse la condición se realiza el desvío hacia la derecha. Después, se le pregunta al entrevistado por su preocupación por el cambio climático (de mucha preocupación a nada), donde responde que 5 (nada), por lo que al no cumplirse la condición, se realiza el desvío hacia la derecha, y seguidamente, se le realiza la pregunta acerca de la preocupación por la invasión de Rusia a Ucrania (de mucha preocupación a nada), donde el encuestado dice que 3 (algo preocupado), lo que daría como resultado otro desvío más hacia la derecha hasta llegar al nodo hoja. Entonces los resultados que arroja nuestro árbol de clasificación es que, esta persona, tiene una probabilidad del 13,8% de ser del PP y un 86,1% de VOX.

Entonces, se puede llegar a la conclusión de que esta persona seguramente será de derechas y con una alta probabilidad, pertenecerá al grupo político VOX.

4.8.2.- Segundo árbol de clasificación

En el segundo análisis de clasificación, se repite la fórmula realizada en el primer árbol, pero la variable objetivo a predecir solo es entre PP y PSOE.

El algoritmo de selección de atributos nos muestra con el siguiente orden que los atributos más relevantes del dataframe son: P23, P3, P28, P24, P5, P29, P6, P11, P33 y EDAD (véase Anexo I).

Al dataframe creado se le administra el algoritmo de clasificación mostrando una precisión que se sitúa en un 80% (al tener muestras balanceadas, se aplica la precisión 'accuracy').

		Votantes pronosticados	
	Matriz de Confusión	PSOE	PP
Votantes Reales	PSOE	159	24
	PP	46	151

Tabla 4.9: Matriz de confusión del segundo árbol de clasificación. Fuente: Elaboración propia

Como se puede observar en la matriz de confusión (Tabla 4.9) y en el árbol de clasificación obtenido (Figuras 4.31 y 4.32), en este caso, a diferencia del árbol anterior, ha mejorado la capacidad de clasificar correctamente a los votantes del PSOE ya que, esta vez, solo ha clasificado a 24 votantes que eran del PSOE como votantes del PP, mientras que las métricas para el Partido Popular han empeorado a la hora de clasificar correctamente.

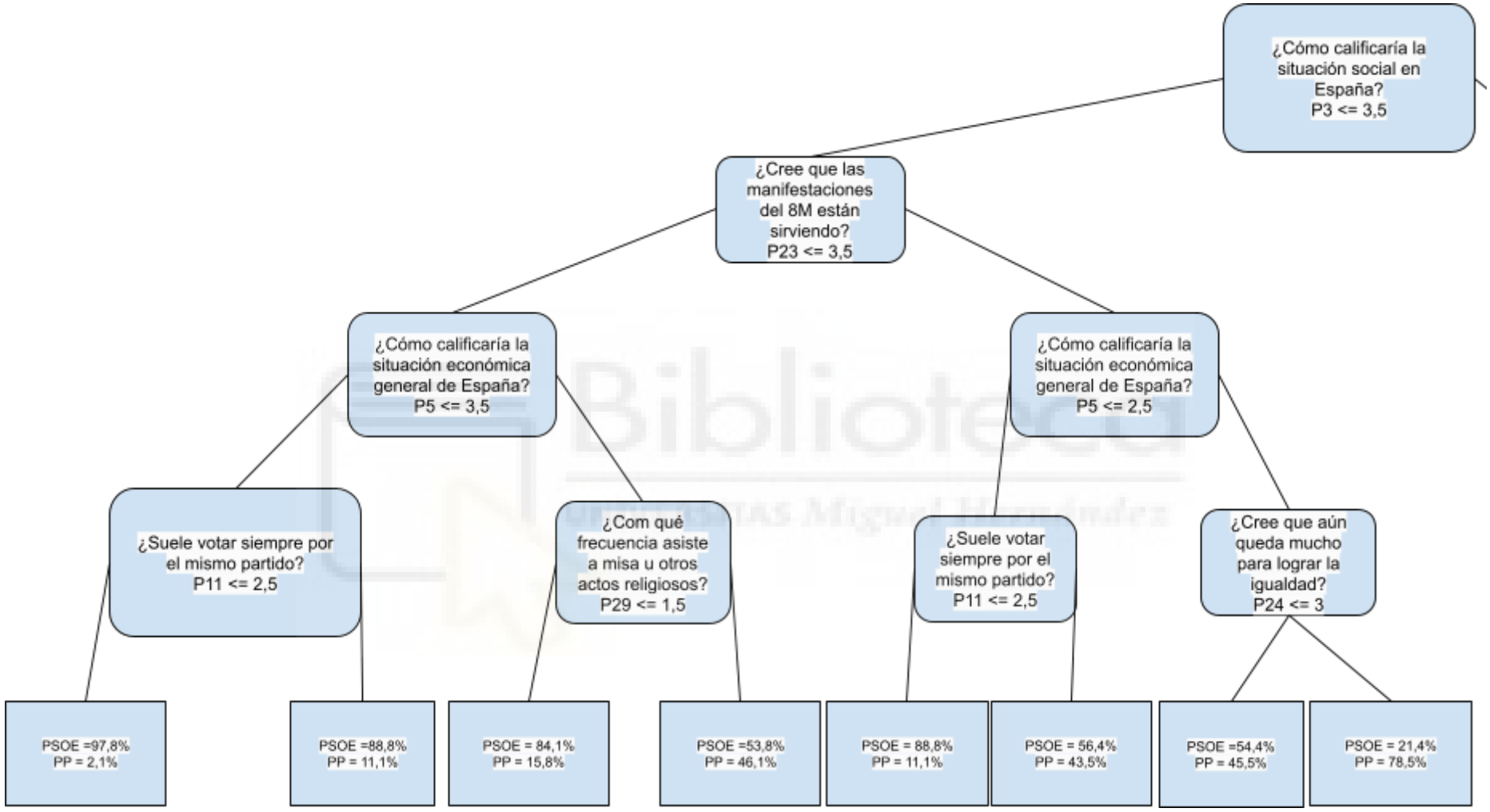

Figura 4.31: Árbol de clasificación 2 Parte I. Fuente: Elaboración propia

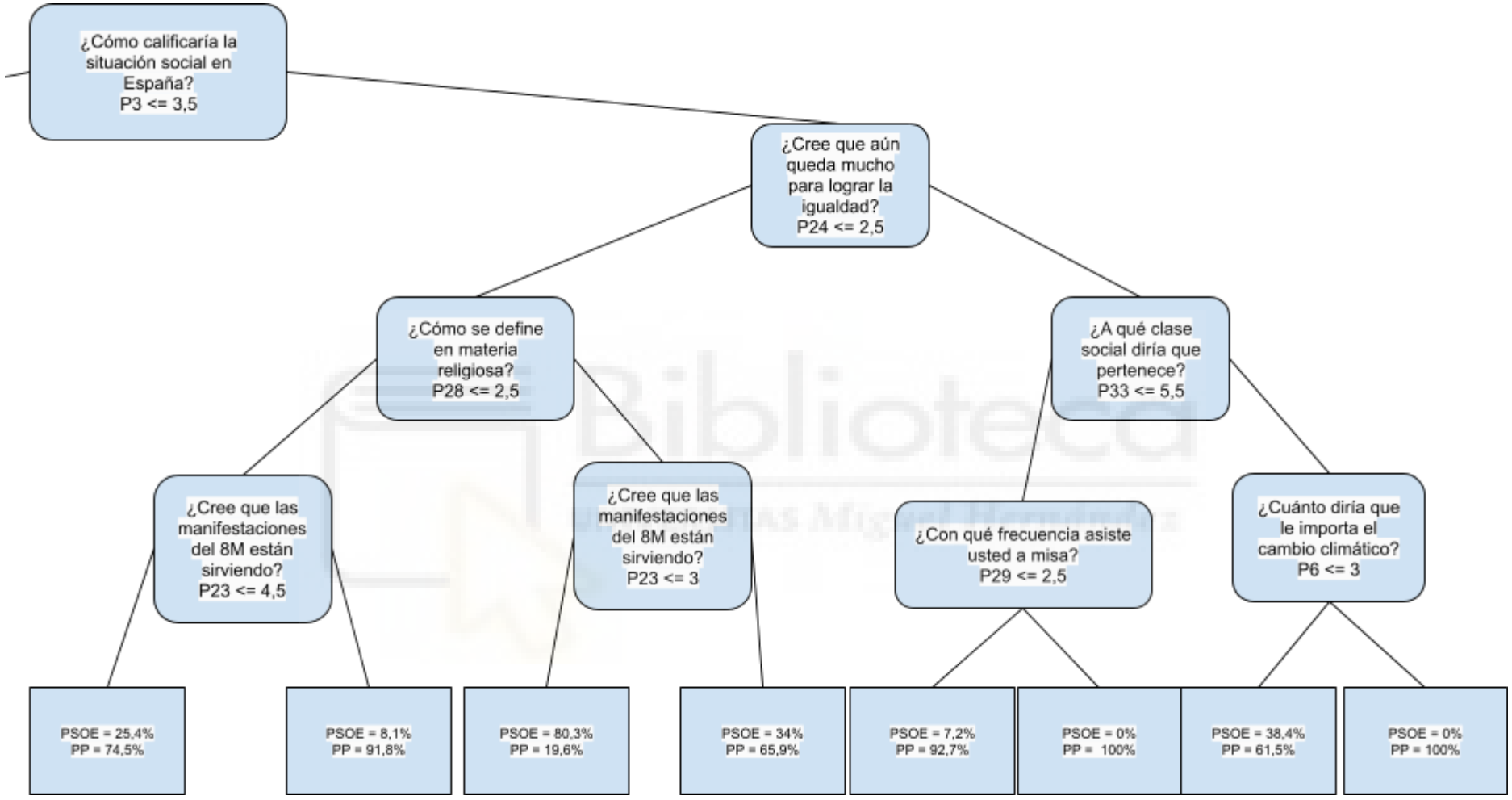

Figura 4.32: Árbol de clasificación 2 Parte II. Fuente: Elaboración propia

En las figuras anteriores (Figura 4.31 y Figura 4.32), se puede observar un árbol de clasificación para categorizar a un grupo de personas encuestadas a un determinado partido político, pero para este caso, solo existen en el árbol los partidos políticos de PP y PSOE. En la figura presentada, se pregunta al entrevistado que realice una valoración de la situación social de España (P3), donde en el caso de ser la respuesta 3 o menos (muy buena, buena o regular), se realiza un desvío hacia la izquierda al cumplirse la condición del nodo, y derecha en caso de que diga 4 o más (mala o muy mala). A continuación, y para este caso, si se pregunta al entrevistado y nos da una respuesta con el valor 4 (mala), al no cumplirse la condición se realiza el desvío hacia la derecha. Seguidamente, se le pregunta al entrevistado por si cree que aún queda mucho para lograr la igualdad (P24, con categorías de respuesta que van desde 1, aún queda mucho, hasta 5, ya existe la igualdad), donde responde que 1 (aún queda mucho), por lo que al cumplirse la condición, se realiza el desvío hacia la izquierda. La siguiente pregunta es acerca de cómo se define en materia religiosa (P28, con categorías de respuesta que van desde 1, católico/a practicante, hasta 6, ateo), donde el encuestado responde que es 4 (Agnóstico), lo que hace que el desvío sea hacia la derecha al no cumplirse la condición. Por último, se le pregunta al entrevistado si cree que las manifestaciones del 8M están sirviendo (P23, con categorías de respuesta que van desde 1, están sirviendo mucho, hasta 5, no están sirviendo de nada) y en este caso el entrevistado contesta 2, es decir, que están sirviendo bastante, por lo que daría como resultado un desvío hacia la izquierda llegando al nodo hoja. En este caso, los resultados que arroja nuestro árbol de clasificación es que esta persona tiene una probabilidad del 19,6% de ser del PP y un 80,3% de ser del PSOE.

Como conclusión, se puede decir que en un alto porcentaje esta persona pertenece al partido socialista.

4.8.3.- Tercer árbol de clasificación

Para este caso, se vuelve a intentar clasificar dentro de los 4 partidos principales que hemos mencionado durante todo el TFG, pero a diferencia de las ocasiones anteriores, se ha procedido a hacer una selección de atributos de manera manual, es decir, lo que se pretende es elegir de manera intencional cuales son las preguntas que se consideran de mayor interés y realizar el entrenamiento con esos atributos.

El dataframe elegido para este caso tiene atributos tales como: EDAD, SEXO, P11, P23, P24, P25, P28, P32 y P33 (véase Anexo I).

El porcentaje de acierto con las variables elegidas por el autor del TFG, a diferencia de cuando las elige el algoritmo, sufren cierta variación a la hora de predecir el resultado, siendo tan solo de un 6% para el peor de los casos, correspondiente a la clasificación de los votantes del PP.

Para este caso, al tratarse otra vez de los cuatro partidos y ser una muestra desbalanceada, se mide la precisión con el f1-score, el cual arroja un porcentaje de precisión del 70% para el PSOE, del 67% del PP, pero del 17% para VOX y del 0% para PODEMOS.

		Votantes pronosticados			
	Matriz de Confusión	PSOE	PP	VOX	PODEMOS
Votantes Reales	PSOE	139	54	2	0
	PP	34	140	9	0
	VOX	0	38	5	0
	PODEMOS	39	4	0	0

Tabla 4.10: Matriz de confusión del tercer árbol de clasificación. Fuente: Elaboración propia

Como se observa en la tabla (Tabla 4.10), el algoritmo no tiene la precisión suficiente para clasificar a los votantes de PODEMOS, y en casi su totalidad, los clasifica como votantes del PSOE. También cabe mencionar que pasa algo parecido a los votantes de VOX aunque en este caso, sí que consigue clasificar correctamente a algún votante de VOX, pero la gran mayoría los engloba dentro del Partido Popular.

También se puede ver como el algoritmo tiene cierta imprecisión con los votantes del Partido Socialista y el Partido Popular, intercambiando 54 y 34 votantes, respectivamente.

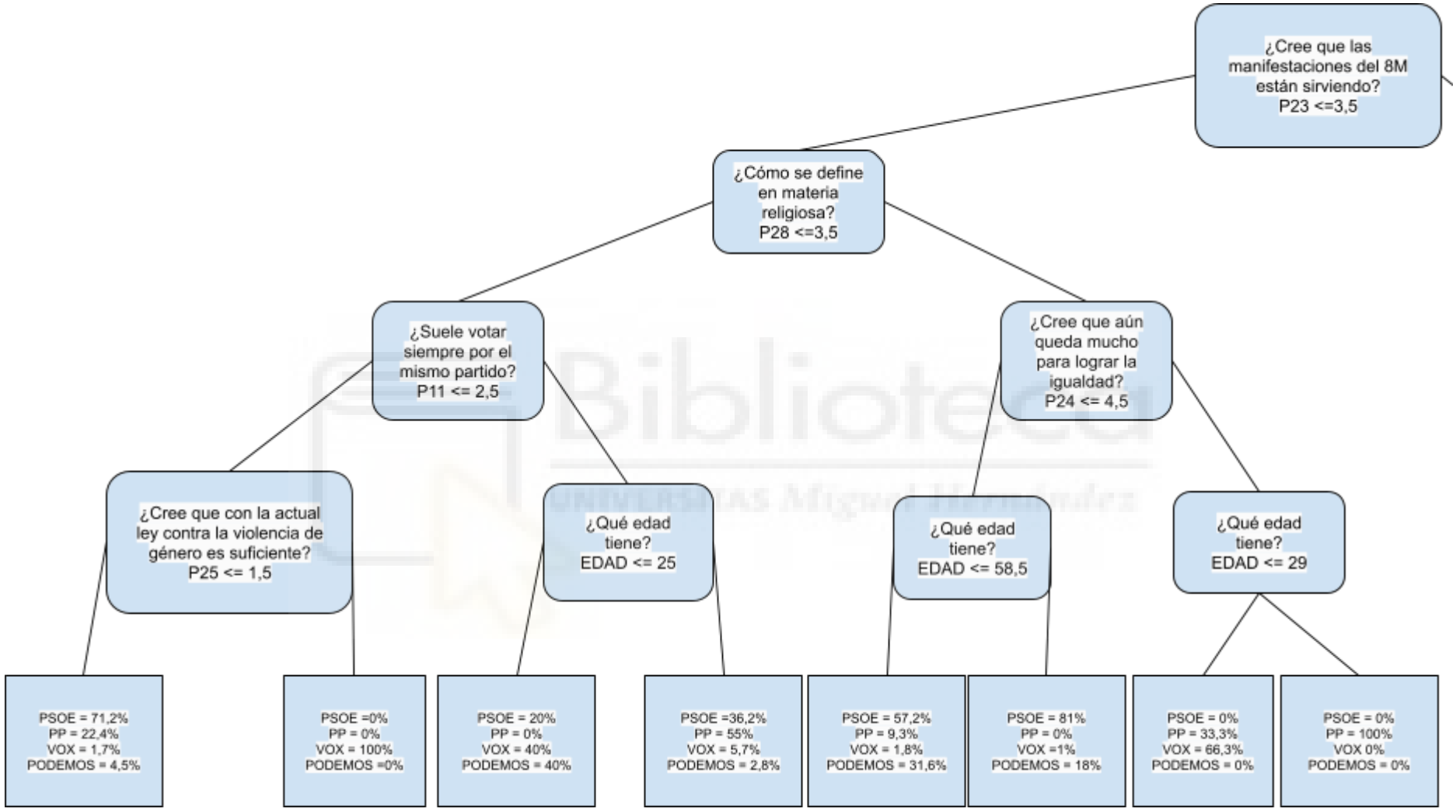

Figura 4.33: Árbol de clasificación 3 Parte I. Fuente: Elaboración propia

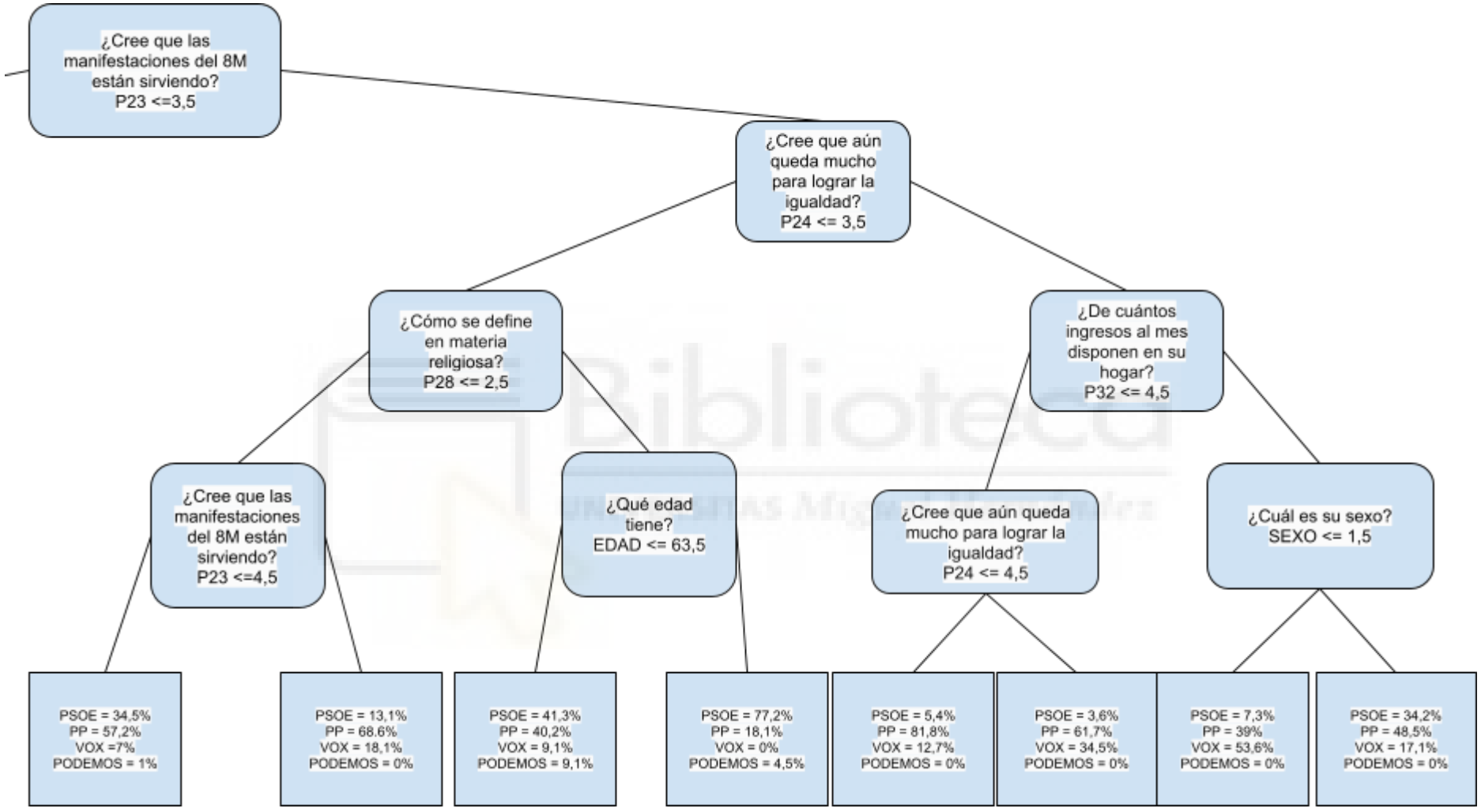

Figura 4.34: Árbol de clasificación 3 Parte II. Fuente: Elaboración propia

En las figuras anteriores (Figura 4.33 y Figura 4.34), se puede observar un árbol de clasificación para categorizar a un grupo de personas encuestadas a un determinado partido político. En la figura presentada, se pregunta al entrevistado sobre su opinión de las manifestaciones que se realizan el 8M (P23, con categorías de respuesta que van desde 1, están sirviendo mucho, hasta 5, no están sirviendo de nada), donde en el caso de ser la respuesta 3 o menos, se realiza un desvío hacia la izquierda al cumplirse la condición del nodo, y derecha en caso de que diga 4 o más. A continuación, y para este caso, se va a seguir el caso por la rama de la derecha, es decir, como si hubiese respondido 4 o más. La siguiente cuestión que se le realiza al encuestado es sobre si cree que queda mucho para lograr la igualdad (P24, con categorías de respuesta que van desde 1, aún queda mucho, hasta 5, ya existe la igualdad), donde si responde con un 2 o menos se realizaría el desvío hacia la izquierda al cumplir la condición del nodo y hacia la derecha si dijese 3 o más al no cumplirla. Para este caso, el encuestado dice que 4 (ya existe dicha igualdad), por lo que al no cumplirse la condición, se realiza el desvío hacia la derecha. Después se le pregunta al entrevistado por la cantidad de ingresos que entran en su hogar a final de mes (P32, con categorías que van desde 1, más de 5000 €, hasta 6, menos de 1.100€) donde responde que 3 (de 2.701 a 3.900), por lo que al cumplirse la condición se realiza el desvío hacia la izquierda, y ahora se le realiza la pregunta acerca de si cree que aún quede mucho para lograr la igualdad (P24, con categorías de respuesta que van desde 1, aún queda mucho, hasta 5, ya existe la igualdad), donde el encuestado dice que 4 (cree que queda poco para lograr la igualdad), lo que daría como resultado otro desvío más hacia la izquierda llegando al nodo hoja. En este caso, los resultados que arroja nuestro árbol de clasificación es que, esta persona, tiene una probabilidad del 81,8% de ser del PP, en un 12,7% de VOX y de 3,6% del PSOE.

Como conclusión, se puede decir que, en un alto porcentaje, esta persona pertenece al PP.

4.8.4.- Cuarto árbol de clasificación

Por último, este caso es similar al anterior, se han elegido manualmente las variables que se consideraban más importantes, pero a diferencia del caso anterior, ahora sólo hay 2 partidos políticos diferentes para la clasificación, PP y PSOE.

Para este caso, el algoritmo ha elegido las preguntas: P23, P28, P24, P11, P33, P32, EDAD, P12, P25 y SEXO (véase Anexo I).

Al dataframe creado se le administra el algoritmo de clasificación, mostrando una precisión que se sitúa en un 76% según el 'accuracy' (igual que en el segundo árbol, al tener unas muestras balanceadas, tiene sentido usar la precisión 'accuracy').

		Votantes pronosticados	
	Matriz de Confusión	PSOE	PP
Votantes Reales	PSOE	120	57
	PP	30	162

Tabla 4.11: Matriz de confusión del cuarto árbol de clasificación. Fuente: Elaboración propia

Como se puede observar en la matriz de confusión (Tabla 4.11), a diferencia del árbol anterior, donde únicamente estaba el PP y el PSOE, ha empeorado la clasificación de los votantes del PSOE ya que, en este caso, ha clasificado a 57 votantes que eran del PSOE como votantes del PP, y 30 votantes del PP como votantes del PSOE, por lo que en este caso y a diferencia del árbol anterior, ha mejorado clasificando a los votantes del PP.

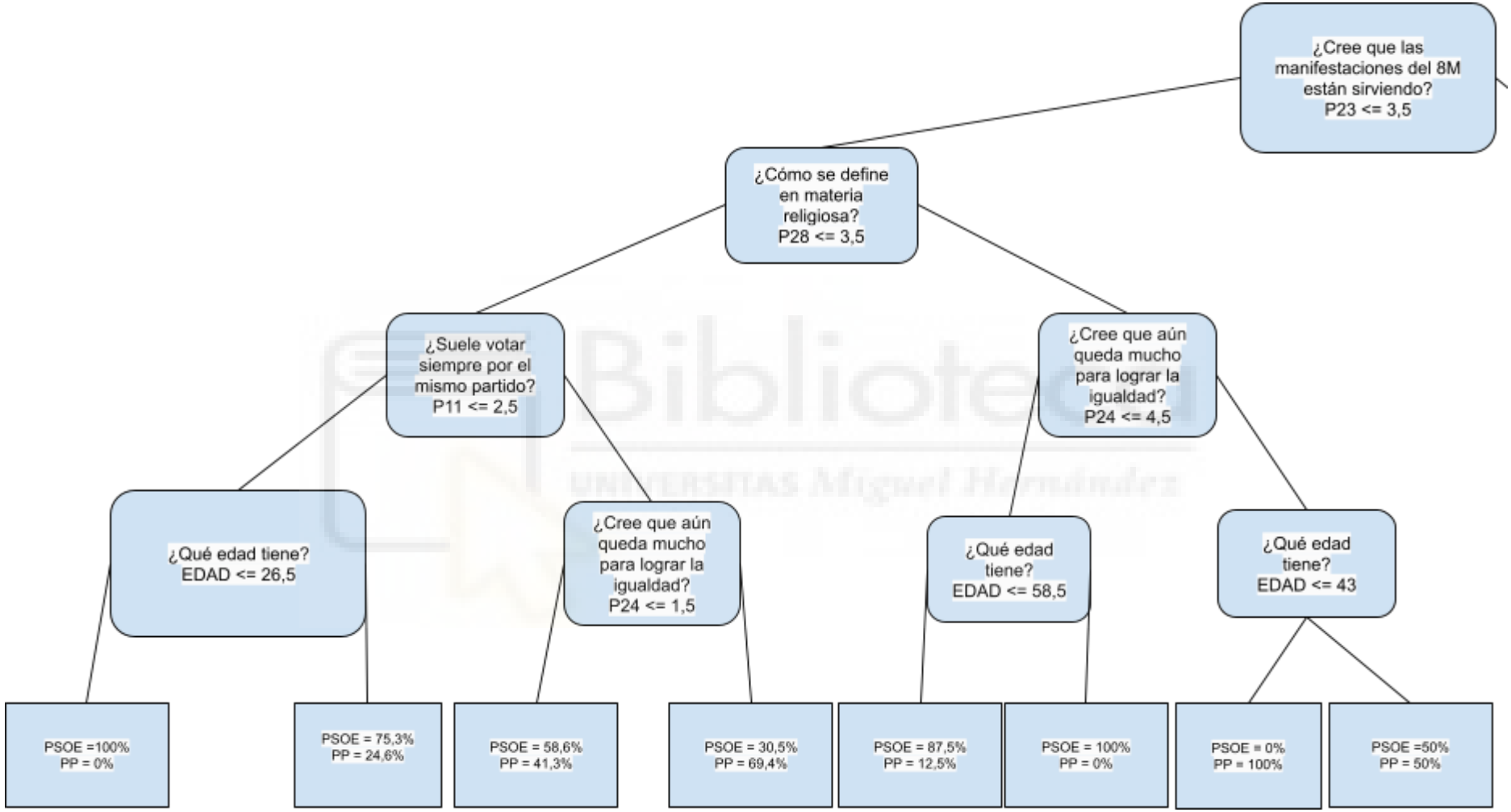

Figura 4.35: Árbol de clasificación 4 Parte I. Fuente: Elaboración propia

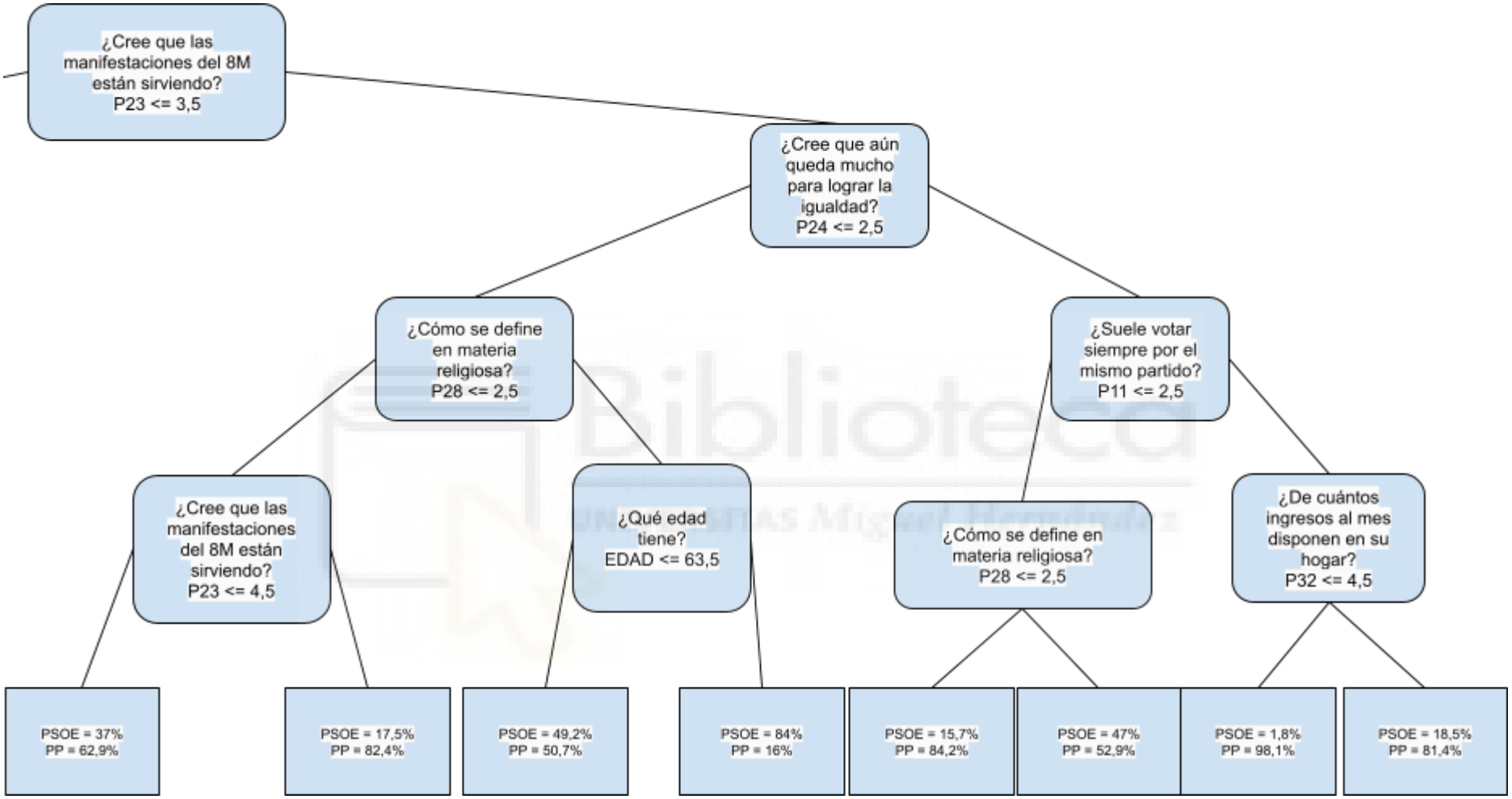

Figura 4.36: Árbol de clasificación 4 Parte II. Fuente: Elaboración propia

En las figuras anteriores (Figura 4.35 y Figura 4.36), se puede observar un árbol de clasificación para categorizar a un grupo de personas encuestadas a un determinado partido político, pero para este caso, solo existen en el árbol los partidos de PP y PSOE. En la figura presentada, se pregunta al entrevistado sobre su opinión de las manifestaciones que se realizan el 8M P23, con categorías de respuesta que van desde 1, están sirviendo mucho, hasta 5, no están sirviendo de nada), donde en el caso de ser la respuesta 3 o menos (están sirviendo mucho, bastante o regular), se realiza un desvío hacia la izquierda al cumplirse la condición del nodo, y derecha en caso de que diga 4 o más (están sirviendo poco o nada). Para este caso, el entrevistado nos da una respuesta con el valor 4 (están sirviendo de poco), por lo que al no cumplirse la condición se realiza el desvío hacia la derecha. Seguidamente, se le pregunta al entrevistado si cree que queda mucho para lograr la igualdad (P24, con categorías de respuesta que van desde 1, aún queda mucho, hasta 5, ya existe la igualdad), donde responde 2 (queda bastante), por lo que al cumplirse la condición, se realiza el desvío hacia la izquierda. Después, la siguiente pregunta es acerca de cómo se define en materia religiosa (P28, con categorías de respuesta que van desde 1, católico/a practicante, hasta 6, ateo), donde el encuestado responde 6 (Ateo), lo que hace que el desvío sea hacia la derecha al no cumplirse la condición. Por último, se le pregunta al entrevistado por su edad (EDAD) y en este caso el entrevistado indica 68 años, por lo que daría como resultado un desvío hacia la derecha llegando al nodo hoja.

A tenor de los resultados que arroja nuestro árbol de clasificación, esta persona tiene una probabilidad del 16% de ser del PP y 84% de ser del PSOE.

Capítulo 5 - Conclusiones y trabajo futuro

En este capítulo se finaliza la memoria elaborada durante el TFG y se exponen las conclusiones obtenidas después de analizar todas las métricas obtenidas en el proyecto y el grado de consecución de los objetivos planteados en el primer capítulo.

5.1.- Conclusiones

Tras haber finalizado el proyecto sobre el análisis de la intención de voto mediante técnicas estadísticas y algoritmos de aprendizaje automático, se va a realizar una valoración sobre el TFG desarrollado y los objetivos establecidos y finalmente alcanzados.

Una vez contextualizado el tema a abordar en el TFG mediante el ejemplo de demoscopia, se procede a realizar un estudio de características similares con los datos del barómetro realizado por el Centro de Investigación Sociológico en febrero de 2023.

Durante todo el Capítulo 2 se exponen de manera teórica los algoritmos de minería de datos que van a ser empleados para realizar el proyecto, donde se hace especial hincapié en el análisis de correspondencias y en los árboles de clasificación, ya que son las técnicas que se emplean en el Capítulo 4.

En el siguiente capítulo, se exponen cuáles han sido las herramientas empleadas para llevar a cabo el TFG, desde el entorno hasta las librerías empleadas, pasando por el lenguaje de programación utilizado en el desarrollo del trabajo.

A lo largo del Capítulo 4 se detalla de manera precisa las diferentes características que tienen en común y en qué difieren los votantes según el partido al que depositan su voto en las elecciones.

Tras realizar estudio, se puede concluir que las técnicas estadísticas y algoritmos de aprendizaje han permitido alcanzar el objetivo de describir los rasgos del votante y predecir la intención de voto. En concreto, las conclusiones del trabajo se pueden resumir en las siguientes:

- A la hora de decidir el voto, el hombre es el doble de indeciso que la mujer a la hora de destinar su voto hacia algún partido o ninguno.

- Los partidos emergentes (PODEMOS y VOX), han sabido recoger mejor el voto de la gente más joven del país, mientras que los partidos tradicionales (PSOE y PP), no tienen apenas gente joven entre sus votantes pero sí un gran número de gente veterana, hecho que no han sabido realizar los partidos más nuevos.

- Los principales problemas en España suelen ser los mismos para todos los votantes, indistintamente del partido político al que vote, y estos son: la crisis económica y los problemas políticos en general.

- Según los resultados obtenidos, se confirma la creencia que la religión, ser católico (practicante y no practicante) está más asociado con ideologías de derechas.

- Los resultados obtenidos mediante este TFG hacen no corroborar que la gente con dinero es de derechas y la gente sin recursos es de izquierdas.

- Nuestros análisis han indicado que la gran mayoría de votantes del PSOE son muy fieles a su partido, en comparación con el resto de partidos políticos.

- Las personas que no han votado en ningún comicio, se dejan influenciar, algo más que el resto de los votantes, por las siglas del partido político.

- El análisis de las respuestas de los encuestados en función del tamaño del municipio indica que, tanto municipios pequeños como grandes, a la hora del voto a nivel nacional, lo eligen, con mayor frecuencia, en función del partido político.

- Cuanto más a la derecha se sitúa la persona en el voto, menos importancia le da a aspectos relacionados con el medio ambiente.

- Existe, por parte de los votantes de partidos de izquierdas, un mayor interés por las siglas del partido, mientras que los votantes de partidos de derechas y en especial, el Partido Popular, se decantan más por la figura del candidato del partido político.

- Los cuatro árboles de clasificación analizados en este TFG han cumplido el objetivo de predecir a un votante en función de las respuestas dadas. Cuatro preguntas que están relacionadas con la política, con la edad, con la situación económica y con la materia religiosa. Estos árboles han dado como resultado un porcentaje de precisión distinto variando en función del tipo de preguntas y el número de partidos a predecir. Como es lógico, cuantos menos partidos tenga que predecir el algoritmo de clasificación y más preguntas partidistas, mejor es el porcentaje de predicción de este árbol.

Asimismo, añadir una mención especial al partido SUMAR de Yolanda Díaz, que aunque este análisis, como ya se ha comentado anteriormente, se ha realizado con los datos obtenidos en el mes de febrero de 2023, parece que este nuevo partido ocupará la posición de PODEMOS en las próximas elecciones, ya que se presentarán en coalición junto con 13 formaciones políticas. Aun así, sigue existiendo mucha incertidumbre acerca de esta nueva coalición para las próximas elecciones.

5.2.- Posibles desarrollos futuros

Existen varios desarrollos que pueden plantearse una vez concluido este trabajo, entre ellos, un posible desarrollo que podría realizarse a futuro es generar un árbol de clasificación con mejores porcentajes de predicción, ya que para este estudio, algunos árboles no terminan de tener buena precisión en ciertos partidos políticos.

Otra línea futura sería desarrollar un árbol de clasificación capaz de predecir el voto indeciso, para poder desarrollar políticas adaptadas a las necesidades de estos votantes con la intención de atraer su voto.

Otro proyecto que se podría desarrollar a futuro, es un estudio similar a este, pero con una encuesta y una muestra a nivel municipal, donde en el tamaño muestral se realizaría, en vez de por comunidades autónomas, se realizaría proporcional a los barrios de Elche, y el tipo de preguntas estarían enfocadas a aspectos interesantes del municipio.

Bibliografía

[1] Demoscopia
https://economipedia.com/definiciones/demoscopia.html
Fecha de consulta: 03/04/2023

[2] Política y demoscopia, los sondeos y las elecciones generales de 1996
Juan Jesus Gonzalez / https://dialnet.unirioja.es/descarga/articulo/199624.pdf
EMPIRIA(1998)/ Fecha de consulta: 03/04/2023

[3] La sociedad española frente al cambio climático
https://acortar.link/ha47qo
Gobierno de España / Fecha de consulta: 04/04/2023

[4] Centro de Investigación Sociológica
https://www.cis.es/cis/opencms/ES/index.html
Fecha de consulta: 03/04/2023

[5] Cómo aplicar un análisis preliminar de datos para Recursos Humanos
https://youtu.be/47bsOwCxAlo
Escuela De Bayes / Fecha de consulta: 10/04/2023

[6] Avance de resultados del estudio 3395 barómetro de febrero 2023
https://acortar.link/M6k2bh
Fecha de consulta: 18/04/2023

[7] Métodos de selección de características machine learning
https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/
Fecha de consulta: 12/04/2023

[8] Imputar y amputar valores nulos y faltantes en Dataframes Pandas
Raúl Valerio / https://youtu.be/O2q7tSPLCJg
Raúl Valerio - Statistics / Fecha de consulta: 11/04/2023

[9] Tratamientos de valores atípicos en los datos con RStudio
Naren Castellon / https://youtu.be/OC7hLL0DtSM
Naren Castellón / Fecha de consulta: 12/04/2023

[10] Aprendizaje supervisado y no supervisado
https://universidadeuropea.com/blog/aprendizaje-supervisado-no-supervisado/
Fecha de consulta: 12/04/2023

[11] Algoritmos de agrupación de clases
http://www.dia.fi.upm.es/~dmaravall/raf2015/raf2013-clustering-libro.pdf
Fecha de consulta: 22/05/2023

[12] Algoritmo de K-medias
https://acortar.link/U31t1R
Fecha de consulta: 13/04/2023

[13] Distancias en el clustering o agrupamiento
https://keepcoding.io/blog/que-es-clustering-o-agrupamiento/
Fecha de consulta: 13/04/2023

[14] Reglas de asociación y algoritmo Apriori con R
Joaquín Amat / https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion
Fecha de consulta: 13/04/2023

[15] Reglas de asociación en una Base de datos del área médica
Agustín Sáenz López / https://www.redalyc.org/journal/1939/193954081005/html/
Revista de Arquitectura e Ingeniería (2017) / Fecha de consulta: 10/06/2023

[16] Análisis de correspondencias simples y múltiples
Santiago de la Fuente Fernández / https://acortar.link/UUf1nz
UAM(2011) / Fecha de consulta: 10/05/2023

[17] ¿Qué es una tabla de contingencia?
https://www.questionpro.com/blog/es/que-es-una-tabla-de-contingencia/
Fecha de consulta: 22/05/2023

[18] Análisis de correspondencia
https://napsterinblue.github.io/notes/stats/techniques/correspondence/
Fecha de consulta: 31/05/2023

[19] Prueba de chi-cuadrado: ¿Qué es y cómo se realiza?
https://www.questionpro.com/blog/es/prueba-de-chi-cuadrado-de-pearson/
Fecha de consulta: 10/05/2023

[20] ¿Qué es el p-valor y cómo interpretarlo con ejemplos sencillos?
https://conceptosclaros.com/que-es-el-p-valor/
Fecha de consulta: 10/05/2023

[21] ¿Qué es la regresión lineal?
https://aws.amazon.com/es/what-is/linear-regression/
Fecha de consulta: 13/04/2023

[22] Regresión Lineal Múltiple: tutorial en Excel
https://help.xlstat.com/es/6685-regresion-lineal-multiple-tutorial-en-excel
Fecha de consulta: 19/04/2023

[23] Modelos de árboles de decisión
https://www.ibm.com/docs/es/spss-modeler/saas?topic=trees-decision-tree-models
Fecha de consulta: 13/04/2023

[24] Precisión frente a F1-Score
https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2
Fecha de consulta: 09/06/2023

[25] Cómo usar Confusion Matrix en Scikit-Learn
JC Chouinard \ https://www.jcchouinard.com/confusion-matrix-in-scikit-learn/
Fecha de consulta: 30/05/2023

[26] ¿Qué son los datasets y dataframes?
https://acortar.link/jkpTyY
Fecha de consulta: 01/05/2023

[27] Introducción a Google Colab
https://youtu.be/CcdE3xYVrb4
Escuela De Bayes / Fecha de consulta: 01/05/2023

[28] ¿Qué es Python?
https://aws.amazon.com/es/what-is/python/
Fecha de consulta: 01/05/2023

[29] Curso de PYTHON desde CERO para PRINCIPIANTES
Brais Moure \ https://youtu.be/Kp4Mvapo5kc
Fecha de consulta: 01/05/2023

[30] Librerías de Python, ¿qué son y cuales son las mejores?
https://immune.institute/blog/librerias-python-que-son/
Fecha de consulta: 01/05/2023

[31] Matplotlib: visualización con Python
https://matplotlib.org/
Fecha de consulta: 01/05/2023

[32] NumPy documentación
https://numpy.org/devdocs/
Fecha de consulta: 01/05/2023

[33] Documentación de pandas
https://pandas.pydata.org/docs/
AprendeIA con Ligdi Gonzalez / Fecha de consulta: 01/05/2023

[34] Seaborn: visualización de datos estadísticos
https://seaborn.pydata.org/
Fecha de consulta: 01/05/2023

[35] Prince
https://github.com/MaxHalford/prince
Fecha de consulta: 22/05/2023

[36] Funciones estadísticas
https://docs.scipy.org/doc/scipy/reference/stats.html
Fecha de consulta: 12/05/2023

[37] scikit-aprender: Aprendizaje automático en Python
https://scikit-learn.org/stable/
Fecha de consulta: 12/05/2023

[38] Los españoles se confiesan mayoritariamente católicos
https://www.larazon.es/opinion/20220423/5gcwlh5tcjgoxmu4tp26rvu4yi.html
Fecha de consulta: 10/06/2023

[39] ¿A quién vota la clase trabajadora en España?
https://la-u.org/a-quien-vota-la-clase-trabajadora-en-espana/
Fecha de consulta: 10/06/2023

[40] Algoritmo SelectKBest
https://acortar.link/LVSTLc
Fecha de consulta: 05/06/2023

Anexo I: Informe CIS

Este anexo contiene la encuesta realizada por el Centro de Investigaciones Sociológicas [4] con todas las preguntas del cuestionario que conforman el dataset, además están desglosadas numéricamente y por nombre de variable. También se puede observar la puntuación que recibe cada respuesta, que es la que se ha empleado en el proyecto para realizar los análisis.

Se puede consultar el estudio del barómetro realizado por el CIS en:
https://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=14692

y el cuestionario completo realizado por el CIS en:
https://www.cis.es/cis/export/sites/default/-Archivos/Marginales/3380_3399/3395/cues3395.pdf

A modo resumen, se incluyen en la Tabla A.1 las preguntas del cuestionario que se realizaron al entrevistado:

Pregunta	Variable	Título y categorías
-	ESTUDIO	Código del estudio 3395.- 3395
-	REGISTRO	Número de registro
-	CUES	Cuestionario
-	CCAA	Comunidad autónoma 1.- Andalucía 2.- Aragón 3.- Asturias (Principado de) 4.- Balears (Illes) 5.- Canarias 6.- Cantabria 7.- Castilla-La Mancha 8.- Castilla y León 9.- Cataluña 10.- Comunitat Valenciana 11.- Extremadura 12.- Galicia 13.- Madrid (Comunidad de) 14.- Murcia (Región de) 15.- Navarra (Comunidad Foral de) 16.- País Vasco 17.- Rioja (La) 18.- Ceuta (Ciudad Autónoma de) 19.- Melilla (Ciudad Autónoma de)
-	PROV	Provincia (Código aparte)
-	MUN	Municipio (Código aparte)
-	CAPITAL	Capital 1.- Capital de CC.AA. 2.- Capital de provincia 3.- Otros municipios
-	ENTREV	Entrevistador/a
-	TAMUNI	Tamaño de municipio 1.- Menos o igual a 2.000 habitantes 2.- 2.001 a 10.000 habitantes 3.- 10.001 a 50.000 habitantes 4.- 50.001 a 100.000 habitantes 5.- 100.001 a 400.000 habitantes 6.- 400.001 a 1.000.000 habitantes 7.- Más de 1.000.000 habitantes
-	TIPO_TEL	Tipo de teléfono 1.- Fijo 2.- Móvil
P0a	SEXO	Sexo de la persona entrevistada 1.- Hombre 2.- Mujer
P0b	EDAD	Edad de la persona entrevistada 99.- N.C.
P0c	P0	Nacionalidad de la persona entrevistada

		1.- La nacionalidad española 2.- La nacionalidad española y otra 3.- Otra nacionalidad
P1	P1	Grado de preocupación ante la situación del coronavirus COVID-19 1.- Mucho 2.- Bastante 3.- (NO LEER) Regular 4.- Poco 5.- Nada 8.- N.S. 9.- N.C.
P2	P2	Grado de preocupación por la invasión de Rusia a Ucrania 1.- Muy preocupado/a 2.- Bastante preocupado/a 3.- Algo preocupado/a 4.- Poco preocupado/a 5.- Nada preocupado/a 6.- (NO LEER) No tiene criterio 7.- (NO LEER) Le es indiferente 8.- N.S. 9.- N.C.
P3	SITSOC	Valoración de la situación social de España 1.- Muy buena 2.- Buena 3.- (NO LEER) Regular 4.- Mala 5.- Muy mala 8.- N.S. 9.- N.C.
P4	ECOPER	Valoración de la situación económica personal actual 1.- Muy buena 2.- Buena 3.- (NO LEER) Regular 4.- Mala 5.- Muy mala 8.- N.S. 9.- N.C.
P5	ECOESP	Valoración de la situación económica general de España 1.- Muy buena 2.- Buena 3.- (NO LEER) Regular 4.- Mala 5.- Muy mala 8.- N.S. 9.- N.C.
P6	P6	Grado de preocupación por el cambio climático 1.- Mucho 2.- Bastante 3.- (NO LEER) Regular 4.- Poco 5.- Nada 8.- N.S. 9.- N.C.
P7	PESPANNA1,	Problemas principales que existen actualmente en España

	PESPANNA2, PESPANNA3	(MR) * Primer problema * Segundo problema * Tercer problema (Código aparte)
P8	PPERSONAL1, PPERSONAL2, PPERSONAL3	Problemas sociales que personalmente afectan más (MR) * Primer problema * Segundo problema * Tercer problema (Código aparte)
P9	PREFPTE	Preferencia personal como presidente del Gobierno central 1.- Pedro Sánchez 2.- Alberto Núñez Feijóo 3.- Santiago Abascal 4.- Yolanda Díaz 5.- Alberto Garzón 6.- Inés Arrimadas 7.- Íñigo Errejón 8.- Isabel Díaz Ayuso 9.- Irene Montero 96.- (NO LEER) Otro/a 97.- (NO LEER) Ninguno/a de ellos/as 98.- N.S. 99.- N.C.
P10	PROBVOTO	Escala de probabilidad (0-10) de votar en las próximas elecciones generales 0.- 0 Con toda seguridad no iría a votar 1.- 1 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- 10 Con toda seguridad, iría a votar 98.- N.S. 99.- N.C
P11	FIDEVOTO	Fidelidad de voto en elecciones 1.- Votan siempre por el mismo partido 2.- Por lo general suele votar por el mismo partido 3.- Según lo que más le convenza en ese momento votan por un partido u otro o no votan 4.- (NO LEER) Votan en blanco o nulo 5.- (NO LEER) No suelen votar 6.- (NO LEER) Es la primera vez que votan 8.- N.S. 9.- N.C
P12	MVOTO	Momento de la decisión del voto por un partido 1.- Lo decide mucho antes del inicio de la campaña electoral 2.- Lo decide al comienzo de la campaña electoral 3.- Lo decide durante la última semana de la campaña electoral 4.- Lo decide durante la jornada de reflexión, la víspera de las elecciones 5.- Lo decide el mismo día de las elecciones

		7.- (NO LEER) Aún no ha votado en ningunas elecciones 8.- N.S. 9.- N.C.
P13	IMPORTVOTO	Prioridad entre partido político y candidato/a en la decisión de voto 1.- Al partido político 2.- Al/a la candidato/a 3.- (NO LEER) Al programa 4.- (NO LEER) A todo por igual 8.- N.S. 9.- N.C.
P14	INTENCIONG	Intención de voto en supuestas elecciones generales (Código aparte)
P15	INTENCIONGALTER	Intención de voto alternativo en supuestas elecciones generales (Código aparte)
P16	SIMPATIA	Partido político por el que se siente más simpatía en las elecciones generales (con filtro) (Código aparte)
P17	ESCIDEOL	Escala de autoubicación ideológica (1-10) 1.- 1 Izquierda 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- 10 Derecha 98.- N.S. 99.- N.C
P18	LIDERESCONOCE_1, LIDERESCONOCE_2, LIDERESCONOCE_3, LIDERESCONOCE_4, LIDERESCONOCE_5, LIDERESCONOCE_6	Conocimiento de líderes políticos nacionales * Pedro Sánchez * Alberto Núñez Feijóo * Yolanda Díaz *Santiago Abascal * Inés Arrimadas * Íñigo Errejón 1.- Conoce 7.- No conoce 9.- N.C.
P18a	VALORALIDERES_1, VALORALIDERES_2, VALORALIDERES_3, VALORALIDERES_4, VALORALIDERES_5, VALORALIDERES_6	Escala de valoración (1-10) de líderes nacionales * Pedro Sánchez * Alberto Núñez Feijóo * Yolanda Díaz *Santiago Abascal * Inés Arrimadas * Íñigo Errejón 1.- 1 Muy mal 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- 10 Muy bien 98.- N.S.

		99.- N.C. 0.- N.P.
P19	CONFIANZAPTE	Grado de confianza en el presidente del Gobierno central: Pedro Sánchez 1.- Mucha confianza 2.- Bastante confianza 3.- Poca confianza 4.- Ninguna confianza 8.- N.S. 9.- N.C. 0.- N.P.
P20	CONFIANZAOPOSIC	Grado de confianza en el líder del principal partido de la oposición (PP): Alberto Núñez Feijóo 1.- Mucha confianza 2.- Bastante confianza 3.- Poca confianza 4.- Ninguna confianza 8.- N.S. 9.- N.C. 0.- N.P.
P21	PARTICIPACIONG	Participación electoral en las elecciones generales de noviembre de 2019 1.- Votó 2.- No votó 3.- No tenía edad para votar 4.- No tenía derecho a voto 8.- No recuerda 9.- N.C.
P21a	RECUVOTOG	Recuerdo de voto en las elecciones generales de noviembre de 2019 de los votantes (Código aparte)
P22	CERCANIA	Partido político que considera más cercano a sus ideas (Código aparte)
P23	P23	Grado utilidad de las manifestaciones del Día Internacional de la Mujer en las reivindicaciones sobre las libertades y derechos de igualdad de las mujeres 1.- Están sirviendo mucho 2.- Están sirviendo bastante 3.- (NO LEER) Están sirviendo regular 4.- Están sirviendo poco 5.- No están sirviendo de nada 6.- (NO LEER) No le interesa 8.- N.S. 9.- N.C.
P24	P24	Valoración sobre lo que falta por avanzar para lograr la igualdad de derechos y oportunidades entre hombres y mujeres 1.- Queda mucho 2.- Queda bastante 3.- (NO LEER) Queda regular 4.- Queda poco 5.- Ya existe dicha igualdad 8.- N.S. 9.- N.C.
P25	P25	Opinión sobre la violencia de género como problema

		preocupante para la sociedad española 1.- Cree que es un problema preocupante 2.- Cree que no supone un problema 8.- N.S. 9.- N.C.
P25a	P25A	Opinión sobre si la ley contra la violencia de género española es suficiente para enfrentarse a la violencia de género 1.- Cree que con la actual ley es suficiente 2.- Cree que habría que hacer más cosas 8.- N.S. 9.- N.C. 0.- P25
P25b	P25B	Medidas adicionales contra la violencia de género 1.- Endurecer las leyes. Penas más severas 2.- Modificar y mejorar la actual ley del 'sí es sí' 3.- Mayor protección, seguimiento y apoyo a las víctimas 4.- Mayor vigilancia y control de los agresores 5.- Mayor control de posibles reincidentes 6.- Investigación y seguimiento de cada caso denunciado 7.- Educación desde la infancia en igualdad y valores afectivosexuales en la escuela y en la familia 8.- Educación y concienciación social en igualdad, principios y valores. Concienciación en la necesidad de denunciar 9.- Fomentar la igualdad real entre hombres y mujeres 10.- Presunción de inocencia. Vigilancia de denuncias falsas 11.- Retirada de leyes que discriminan positivamente a la mujer. Igualdad de penas para hombres y mujeres 12.- Mayor dotación presupuestaria. Ayudas económicas y sociales. Apoyo psicológico 13.- Mejora en la Administración de Justicia. Aplicación de la ley 14.- Consenso político. Escuchar a los expertos en la materia 15.- Derogar la ley actual 16.- Implicarse, esforzarse más, seguir trabajando 96.- Otras 98.- N.S. 99.- N.C. 0.- P25 y P25a
P26	P26_1, P26_2	Partido que está haciendo más en España para apoyar la igualdad de derechos y oportunidades de las mujeres * Partido que está haciendo más * Partido que está haciendo menos (Código aparte)
P27	ESCUELA	Escolarización de la persona entrevistada 1.- No, es analfabeto/a 2.- No, pero sabe leer y escribir 3.- Sí, ha ido a la escuela 9.- N.C.
P27a	NIVELESTENTREV	Nivel de estudios alcanzado por la persona entrevistada 1.- Menos de 5 años de escolarización 2.- Educación Primaria (Educación Primaria de LOGSE, 5º curso de EGB, enseñanza primaria antigua) 3.- Cualificación profesional grado inicial (FP grado inicial). PCPI (Programas de Cualificación Profesional Inicial, que no precisan de titulación académica de la primera etapa de

		secundaria para su realización). Programas de garantía social 4.- Educación secundaria (ESO, EGB. Graduado Escolar. Certificado de Escolaridad, Bachillerato Elemental) 5.- FP de Grado Medio (ciclo/módulo formativo de FP (grado medio), de Artes Plásticas y Diseño, Música y Danza, enseñanzas deportivas, FP I, Bachiller Laboral Elemental. Oficialía Industrial; Bachillerato Comercial) 6.- Bachillerato (Bachillerato LOGSE, BUP, Bachillerato Superior (6º), Bachillerato Universitario (7º), incluidos COU y PREU) 7.- FP de Grado Superior (ciclo/módulo formativo de FP (grado superior) de Artes Plásticas, Diseño, Música y Danza, Deporte, FP II, Bachillerato Laboral Superior, Maestría industrial, perito/a mercantil; Secretariado de 2º grado; Grado Medio conservatorio) 8.- Arquitectura/Ingeniería Técnica (aparejador/a; peritos/as) 9.- Diplomatura (ATENCIÓN: sólo diplomaturas oficiales, no codificar aquí los tres primeros años de una licenciatura o grado con mayor duración) 10.- Grado (estudios de grado, enseñanzas artísticas equivalentes (desde 2006)) 11.- Licenciatura (titulaciones con equivalencia oficial: 2º ciclo INEF; Danza y Arte Dramático (desde 1992); Grado Superior de Música) 12.- Arquitectura/Ingeniería 13.- Máster oficial universitario (especialidades médicas o equivalente) 14.- Doctorado 15.- Títulos propios de posgrado (máster no oficial, etc.) 16.- Otros estudios 98.- N.S./No recuerda 99.- N.C. 0.- N.P.
P28	RELIGION	Religiosidad de la persona entrevistada 1.- Católico/a practicante 2.- Católico/a no practicante 3.- Creyente de otra religión 4.- Agnóstico/a (no niegan la existencia de Dios pero tampoco la descartan) 5.- Indiferente, no creyente 6.- Ateo/a (niegan la existencia de Dios) 9.- N.C
P29	PRACTICARELIG6	Frecuencia de asistencia a oficios religiosos 1.- Nunca 2.- Casi nunca 3.- Varias veces al año 4.- Dos o tres veces al mes 5.- Todos los domingos y festivos 6.- Varias veces a la semana 9.- N.C. 0.- N.P.
P30	ECIVIL	Estado civil de la persona entrevistada 1.- Casado/a 2.- Soltero/a 3.- Viudo/a

		4.- Separado/a 5.- Divorciado/a 9.- N.C.
P31	SITLAB	Situación laboral de la persona entrevistada 1.- Trabaja 2.- Jubilado/a o pensionista (anteriormente ha trabajado) 3.- Pensionista (anteriormente no ha trabajado) 4.- En paro y ha trabajado antes 5.- En paro y busca su primer empleo 6.- Estudiante 7.- Trabajo doméstico no remunerado 8.- Otra situación 9.- N.C
P31a	RELALAB	Situación profesional de la persona entrevistada 1.- Asalariado/a por cuenta ajena 2.- Empresario/a o profesional con asalariados/as 3.- Profesional o trabajador/a autónomo/a (sin asalariados/as) 4.- Ayuda familiar (sin remuneración reglamentada en la empresa o negocio de un/a familiar) 5.- Miembro de una cooperativa 6.- Otra situación 9.- N.C. 0.- N.P.
P31b	CNO11	Ocupación de la persona entrevistada 1.- Directores/as y gerentes 2.- Profesionales, científicos e intelectuales 3.- Técnicos/as y profesionales de nivel medio 4.- Personal de apoyo administrativo 5.- Trabajadores/as de los servicios y vendedores/as de comercios y mercados 6.- Agricultores/as y trabajadores/as cualificados/as agropecuarios, forestales y pesqueros/as 7.- Oficiales/as, operarios/as, artesanos/as y trabajadores/as de artes mecánicas y de otros oficios 8.- Operadores/as de instalaciones y máquinas y ensambladores/as 9.- Ocupaciones elementales 10.- Ocupaciones militares y cuerpos policiales 11.- Otra/o 99.- N.C. 0.- N.P.
P32	INGRESHOG	Nivel de ingresos netos del hogar 1.- Más de 5.000 € 2.- De 3.901 a 5.000 € 3.- De 2.701 a 3.900 € 4.- De 1.801 a 2.700 € 5.- De 1.100 a 1.800 € 6.- Menos de 1.100 € 8.- N.S. 9.- N.C.
P33	CLASESOCIAL	Clase social subjetiva de la persona entrevistada 1.- Clase alta 2.- Clase media-alta 3.- Clase media-media 4.- Clase media-baja 5.- Clase trabajadora/obrera

		12.- Clase baja 6.- Clase pobre 7.- Infraclase 8.- Proletariado 9.- A los/as de abajo 10.- Excluidos/as 11.- A la gente común 96.- Otras 97.- No cree en las clases 98.- No sabe, duda 99.- N.C
-	SINCERIDAD	Grado de sinceridad de la persona entrevistada según el/la entrevistador/a 1.- Mucha 2.- Bastante 3.- Poca 4.- Ninguna 8.- N.S.

Tabla A.1: Cuestionario intención del voto. Fuente: CIS[4]

Anexo II: Ficheros empleados en el Capítulo 4

Se realiza la entrega de un archivo comprimido zip con el siguiente índice de ficheros del Capítulo 4

Dataset en bruto
- 3395_etiq.csv
- 3395_num.csv

Apartado 4.4 - Tratamiento previo del dataset
- CSV FINAL.csv
- RESULTADOS_NUMERICOS.csv

Apartado 4.5 en adelante
- CodigoTFG.ipynb