

Grado en Estadística y Empresa
2020-2021

Trabajo Fin de Grado

Procedimiento de imputación múltiple basado en Random Forest

Beatriz Román Carrascosa

Tutor

Andrés Modesto Alonso Fernández

Getafe, Julio 2021

RESUMEN

Dentro del clima político actual, las encuestas electorales han tomado cada vez más importancia en lo que concierne a la comprensión de la situación política actual. A pesar del alto número de respuestas, dichas encuestas electorales cuentan con numerosos valores faltantes en preguntas clave tales como la intención de voto ante unas futuras elecciones generales. En este estudio se trata de abordar este problema por medio de la imputación múltiple utilizando el procedimiento de *Random Forest*.

Haciendo uso de la base de datos del barómetro electoral realizado por el Centro de Investigaciones Sociológicas (CIS) en diciembre de 2020, se ilustrará la utilización del algoritmo *Random Forest* en variables de menor a mayor número de valores faltantes de forma reiterada para realizar así la imputación múltiple. De esta forma, se obtendrán las estimaciones finales de la intención de voto de los encuestados teniendo en cuenta el peso del muestreo.

Una vez realizadas las estimaciones, se tiene que tanto en el conjunto de entrenamiento como en el conjunto de prueba los resultados son notablemente precisos y se ajustan a las predicciones proporcionados por el CIS. Es por ello que, el algoritmo *Random Forest* dada su gran flexibilidad, resulta una alternativa interesante para tratar de abordar el problema de los valores faltantes en las encuestas electorales.

Palabras Clave: imputación múltiple, CIS, intención de voto, *Random Forest*.

AGRADECIMIENTOS

A mis padres, Julián y Rosa María, por su esfuerzo y sacrificio durante estos cuatro años, apoyándome en cada paso que doy. A mi hermana Celia, gracias por acompañarme en esta etapa tan importante para mí. A Ángel Manuel, por ser mi gran apoyo estos últimos cuatro años. A todos mis familiares y en especial a mi querido abuelo Jesús Carrascosa Espada, víctima del virus SARS-CoV-2, por su fuerza y valentía en los momentos difíciles, siempre estarás con nosotros.

A todos mis amigos, por ser una pieza fundamental en esta etapa, en especial a Belén Rodríguez. Gracias por tu apoyo durante estos últimos meses, por escucharme y por creer en mí desde el primer día, siempre juntas. A Enrique Segura por su inestimable ayuda en esta recta final y a toda la network de Emerald, gracias por hacerme formar parte de algo tan bonito y especial.

Por último, agradecer a mi profesor y tutor Andrés Modesto Alonso, por confiar en mí para realizar este trabajo y brindarme su ayuda desde el primer día. Gracias por tu paciencia, comprensión y ánimos durante estos últimos meses.

ÍNDICE DE CONTENIDO

1. INTRODUCCIÓN.....	1
2. PREPROCESO	3
2.1 Selección de variables.....	3
2.1.1 Preguntas electorales.....	3
2.1.2 Preguntas ideológicas.....	4
2.1.3 Preguntas valorativas	4
2.1.4 Preguntas valorativas	4
2.2 Recodificación	5
3. ANÁLISIS EXPLORATORIO DE DATOS (EDA).....	7
3.1 Análisis univariante	9
3.2 Análisis bivariante	13
4. METODOLOGÍA	18
4.1 Árboles de clasificación.....	19
4.2 Random Forest	20
4.3 Partición de la muestra y procedimiento de imputación.....	21
5. RESULTADOS.....	23
5.1 Imputación de la situación laboral del encuestado	23
5.2 Imputación de la simpatía partidista	25
5.3 Imputación de la auto-ubicación ideológica	26
5.4 Imputación de la intención de voto.....	28
5.5 Precisión de los modelos.....	29
6. INTERPRETACIÓN DEL ALGORITMO	31
6.1 Árbol de decisión para el grado de confianza en el presidente del gobierno.....	31
6.2 Árbol de decisión para la preferencia como presidente del gobierno	32
6.3 Árbol de decisión para la intención de voto.....	33
7. PREDICCIÓN FINAL DE LA INTENCIÓN DE VOTO	35
7.1 Comparación con los resultados del CIS	36
8. CONCLUSIÓN.....	37
BIBLIOGRAFÍA.....	38

ÍNDICE DE TABLAS

Tabla 3.1 Resumen de variables 1-12	7
Tabla 3.2 Resumen de variables (13-30)	8
Tabla 3.4 Tau de Goodman and Kruskal variables (1-12).....	15
Tabla 3.5 Tau de Goodman and Kruskal variables (13-30).....	16
Tabla 5.1 Matrices de confusión de las imputaciones en la situación laboral	23
Tabla 5.2 Matrices de confusión de las imputaciones en la simpatía partidista	25
Tabla 5.3 Matrices de confusión de las imputaciones en la simpatía partidista	26
Tabla 5.4 Matrices de confusión de las imputaciones en la intención de voto	28
Tabla 5.5 Precisión y error estándar de los modelos generados con Random Forest.....	30
Tabla 7.1 Estimación de la intención de voto en % sobre el voto válido	35
Tabla 7.2 Comparación de las predicciones con los resultados del CIS.....	36

ÍNDICE DE FIGURAS

Figura 1.1 Valores nulos (NA) por atributo	5
Figura 3.1 Respuestas NS/NC por atributo	9
Figura 3.2 Clasificación ideológica de los líderes políticos.....	10
Figura 3.3 Intención de voto	11
Figura 3.4 Valoración de los líderes políticos en función de la intención de voto	12
Figura 4.1 Esquema general de Random Forest	20
Figura 5.1 Importancia de variables en el modelo de la situación laboral del encuestado	24
Figura 5.2 Importancia de variables en el modelo la simpatía partidista.....	26
Figura 5.3 Importancia de variables en el modelo de la auto-ubicación ideológica.....	27
Figura 5.4 Importancia de variables en el modelo de la intención de voto.....	29
Figura 6.1 Árbol de clasificación para el grado de confianza en el presidente del gobierno.....	32
Figura 6.2 Árbol de clasificación para la preferencia como presidente del gobierno.....	33
Figura 6.3 Árbol de clasificación para la intención de voto	34

1. INTRODUCCIÓN

El año 2020 ha supuesto un período de convulsión política para España. Por primera vez en la democracia, se formó a principios de 2020 un gobierno de coalición presidido por Pedro Sánchez, líder del Partido Socialista Obrero Español (PSOE), y en alianza con Unidas Podemos, partido político que nació de la crisis de 2008 (Antena 3 Noticias, 2020). Todo ello, bajo un contexto político-sanitario que no ha sido fácil tras la llegada del SARS-CoV-2 y que ha ocasionado crispaciones políticas fruto de la falta de acuerdos entre los diferentes líderes parlamentarios (Bocanegra, 2020).

Ante esta inestabilidad política, las encuestas electorales han despertado en los ciudadanos un especial interés por conocer la situación política actual entre los diferentes partidos más allá de unas próximas elecciones (Orriols, 2021). En este ámbito, el Centro de Investigaciones Sociológicas (CIS) pone a disposición de los españoles barómetros electorales de periodicidad mensual con el objetivo de medir la opinión pública (CIS, 2021). Para ello, se realiza un muestreo polietápico, estratificado por conglomerados, donde se seleccionan las unidades primarias y secundarias, municipios y secciones respectivamente, de forma aleatoria proporcional, y las unidades últimas, es decir, los individuos, por rutas aleatorias y cuotas de sexo y edad (CIS, 2021).

Los barómetros proporcionados por el CIS constan de gran importancia y presencia en diversos medios de comunicación, ya que no sólo generan conocimientos políticos, inclusive, aportan notoriedad a aquellos partidos cuyos candidatos son desconocidos y consiguen influir en la intención de voto de los ciudadanos, así como en su movilización y disposición para acudir a las urnas, a la vez que refuerzan su fidelidad (Orriols, 2021).

No obstante, la fiabilidad del barómetro ha sido cuestionada en numerosas ocasiones, incluso José Félix Tezanos, presidente del CIS, reconoce que los márgenes de error son amplios (Calleja, 2020). Esto puede asociarse a diversos factores, entre ellos, el error irreducible del muestreo utilizado en las encuestas, la dificultad en la predicción de los votantes indecisos o la claridad y transparencia interna del pronóstico electoral, aspecto en el cual el CIS destaca positivamente (Escobar et al., 2014).

Al margen de dichas controversias, tanto el barómetro del CIS como el resto de las encuestas electorales, cuentan con un inconveniente añadido, los llamados *missing values* o valores faltantes. Concretamente, los datos suministrados por el CIS en el barómetro de diciembre de 2020 presentan un alto porcentaje de valores faltantes, especialmente en la variable de interés, intención de voto, donde se alcanza el 21.59% de respuestas “no sabe/no contesta” (NS/NC), porcentaje que no pasa desapercibido.

Es por ello que, en el presente trabajo, se tratará de abordar dicha cuestión. Se realizará un procedimiento de imputación múltiple en una selección de las variables del barómetro electoral que están relacionadas con la intención de voto y que presentan valores faltantes mediante un algoritmo de aprendizaje automático. Concretamente se utilizarán los llamados “bosques aleatorios”, también conocidos como Random Forest. Para implementar el algoritmo y calcular las posteriores predicciones se utilizará el software libre R, pudiendo consultar el código empleado en la página web: <https://beatrizromancarrascosa.github.io/codigoR.TFG.BRC/>.

El resto de la memoria consta de varias secciones dedicadas al análisis exploratorio y descriptivo de las variables seleccionadas en el estudio, seguido del procedimiento del algoritmo a utilizar con los resultados correspondientes. Con todo, se empleará dicho algoritmo de clasificación con el fin de obtener estimaciones precisas y que muestren un reflejo de la situación política que atravesaba España en diciembre de 2020. Posteriormente, se presentarán las conclusiones del estudio y una comparación de con los resultados proporcionadas por el CIS, dejando a relucir así, la exactitud del algoritmo a utilizar.

2. PREPROCESO

La encuesta realizada por el CIS en el barómetro de diciembre de 2020 cuenta con un abanico de preguntas que tratan de recopilar una amplia información social y demográfica de los encuestados para su posterior análisis (CIS, 2021). Dada la complejidad de las encuestas, tanto en el ámbito técnico como en el organizativo, la precisión de los resultados se ve comprometida (Escobar et. al, 2014). Por tanto, es necesario realizar una apropiada selección de aquellas variables más relevantes en el estudio, para simplificar así la metodología y asegurar la calidad de las predicciones. Tras ello, se recodificarán las variables seleccionadas con el objetivo de mejorar la comprensión y posterior interpretación de los resultados.

2.1 Selección de variables

En primer lugar, cabe resaltar que los cuestionarios del CIS son pequeños y monotématicos, aunque cuentan con una extensión variable (Escobar et al., 2014). No obstante, la ficha técnica del barómetro emitida por el CIS en diciembre de 2020 informa que se ha ampliado el tamaño muestral en un rango de 3000 a 4000 entrevistas telefónicas obteniendo finalmente 3817 respuestas. Dicho esto, el cuestionario cuenta con 31 preguntas y adicionalmente una sección para los ciudadanos catalanes, la cual no se analizará.

Una vez importado el conjunto de datos se tienen 251 atributos o variables a analizar, lo cual es una elevada cifra y a pesar de que *Random Forest* es un algoritmo dotado de gran flexibilidad, surge la necesidad de realizar una selección de atributos (Breiman, 2001).

De acuerdo con Escobar et al. (2014) las preguntas incluidas en los cuestionarios electorales pueden clasificarse en cuatro temáticas troncales: preguntas electorales, ideológicas, evaluadoras y sociodemográficas. A continuación, se describirán las categorías y los atributos seleccionados en cada una de ellas.

2.1.1 Preguntas electorales

Son preguntas referidas al comportamiento del votante potencial en las futuras o pasadas elecciones. En este ámbito se han seleccionado como atributos relevantes:

- P21: Intención de voto en supuestas elecciones generales
- P21a: Intención de voto alternativo en supuestas elecciones generales
- P26: Participación electoral en las elecciones generales de noviembre de 2019
- P26a: Recuerdo de voto en las elecciones generales de noviembre de 2019 de los votantes

2.1.2 Preguntas ideológicas

Este tipo de preguntas hacen relación a las creencias o ideas estables de los encuestados sobre el ámbito político. Incluyen autodefiniciones personales en cuanto a ideología, así como ideología hacia los partidos, simpatía o cercanía hacia los mismos (Escobar et al., 2014). Es por ello que, los atributos seleccionados son:

- P20: Preferencia personal como presidente de Gobierno Central
- P22: Partido político por el que siempre más simpatía
- P23: Escala de auto-ubicación ideológica
- P24: Escala de ubicación ideológica de líderes políticos nacionales
- P27: Partido que considera más cercano a sus ideas

2.1.3 Preguntas valorativas

Esta categoría abarca las preguntas relacionadas con la valoración de los líderes parlamentarios, de la situación económica y a pesar de que pueden jugar un papel secundario en el pronóstico electoral, son de vital importancia en la imputación de valores faltantes (Escobar et al., 2014). Como atributos valorativos se tienen:

- P13: Valoración de la situación económica general de España
- P18: Grado de confianza en el presidente del Gobierno central: Pedro Sánchez
- P19: Grado de confianza en el líder del principal partido de la oposición (PP): Pablo Casado
- P25: Escala de valoración de líderes políticos nacionales

2.1.4 Preguntas sociodemográficas

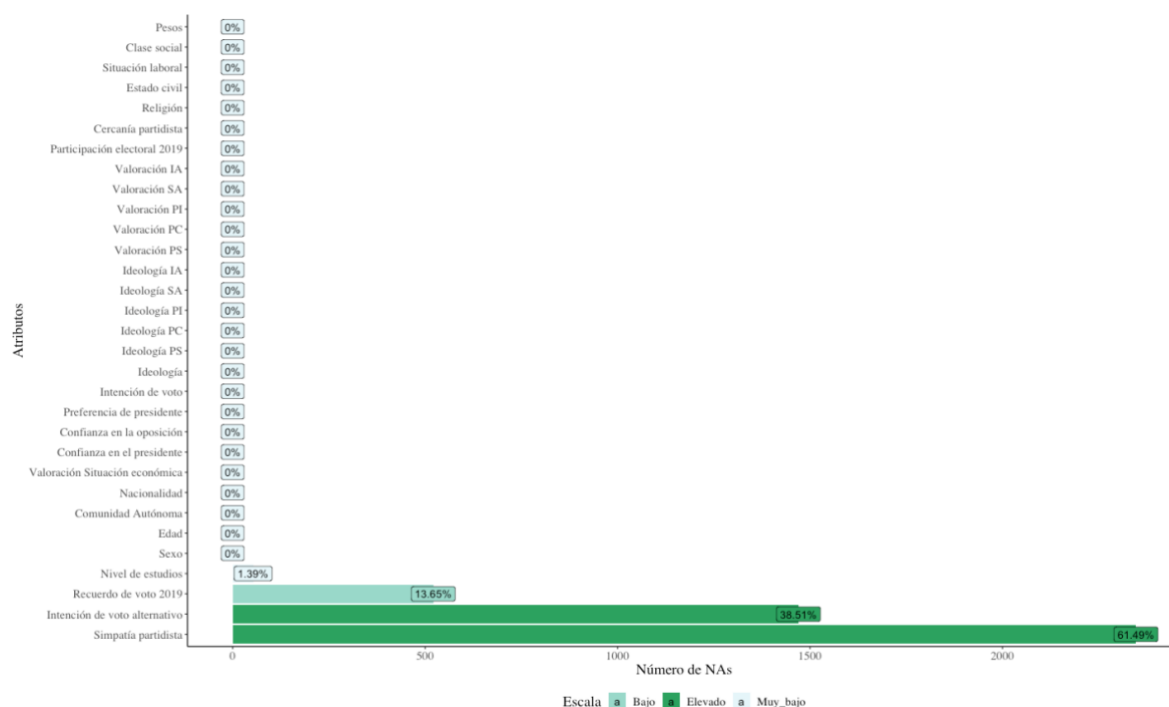
Las preguntas pertenecientes a esta categoría hacen referencia a atributos relacionados con aspectos sociodemográficos tales como la edad, el sexo, estudios del encuestado y su situación laboral. Todas ellas tienen como objetivo estudiar conductas electorales y calibrar la muestra (Escobar et al., 2014).

- P-: Comunidad autónoma
- P0a: Sexo de la persona entrevistada
- P0b: Edad de la persona entrevistada
- P0c: Nacionalidad de la persona entrevistada
- P28a: Nivel de estudios alcanzado por la persona entrevistada
- P29: Religiosidad de la persona entrevistada
- P30: Estado civil de la persona entrevistada
- P31: Situación laboral de la persona entrevistada
- P32: Clase social subjetiva de la persona entrevistada

2.2 Recodificación

Un aspecto a tener en cuenta es que la base de datos original contaba con valores faltantes distintos a NS/NC. Éstos son valores nulos que R califica como NA y son ocasionados por las variables tipo filtro que incluye el cuestionario, es decir, dependiendo de la respuesta del entrevistado, se le excluye de contestar ciertas preguntas. La **figura 1.1** muestra el porcentaje de NAs iniciales mediante una escala de aceptación y como se puede ver únicamente cuatro atributos contienen valores nulos: *nivel de estudios del entrevistado*, *recuerdo de voto en las elecciones generales de 2019*, *intención de voto alternativo* y *simpatía partidista*, siendo este último, el atributo con mayor número de NA, incluso supera el 50% de valores nulos.

Figura 1.1 Valores nulos (NA) por atributo



Fuente: Elaboración propia con RStudio.

Ante este escenario, se han imputado dichos atributos en función del motivo de su exclusión. En el caso de *nivel de estudios*, los valores NA se corresponden con aquellos encuestados que no recibieron ningún tipo de educación, es por ello por lo que se les excluyó de contestar a su nivel de estudios. Por tanto, los valores nulos se han imputado como una nueva categoría que hace referencia a la ausencia de estudios.

Por otro lado, el atributo *recuerdo de voto* contiene valores NA debido a que, si el encuestado no asistió a votar en las pasadas elecciones, carece de recuerdo de voto. Se ha creado así una nueva categoría: “No votó” que envuelve a los individuos que no asistieron a las urnas.

En cuanto a la *intención de voto alternativo*, se ha identificado que los valores NA se corresponden con las categorías de “Voto nulo”, “En blanco” o “NS/NC” en la variable *intención de voto*, es entonces que, carece de sentido preguntar cuál sería el voto alternativo de los encuestados en estos casos. Se imputan así los nulos en *intención de voto alternativo* con su respuesta a la pregunta *intención de voto*.

Por último, en el caso del atributo referente a la *simpatía partidista*, los valores nulos se relacionan con aquellas personas que no responden a la *intención de voto alternativo*, es decir, aquellas que responden a *intención de voto*. En este caso, se han empleado las respuestas a *intención de voto* para imputar la *simpatía partidista*.

Una vez imputados los valores NA, se han recodificado todas las variables seleccionadas como relevantes en el estudio para tratar de mejorar la interoperabilidad de los resultados. El código utilizado para la realización de este trabajo está disponible en <https://beatrizromancarrascosa.github.io/codigoR.TFG.BRC/>.

Los atributos pertenecientes a cuestiones electorales han sido recodificados en función de once clases, donde se han considerado los partidos políticos más influyentes incluyendo también partidos nacionalistas dado la gran relevancia en las respuestas, así como una categoría “Otros partidos” donde se han incluido partidos pequeños y de poca popularidad. Inclusive se ha creado una categoría de “Abstención” y se ha codificado como 99 a la categoría NS/NC.

En las variables ideológicas, se ha respetado la escala suministrada por el CIS, ya que, dado el nivel de abstracción y subjetividad, una posible agrupación puede hacer perder información al estudio.

Para las variables valorativas, se han considerado cinco clases en cuanto a la valoración de la situación económica de España: “muy bien”, “bien”, “regular”, “mal” y “muy mal”. La valoración de los líderes políticos cuenta con una escala similar. Por último, los atributos sociodemográficos han sido codificados con el fin de reducir el número de clases y facilitar la interpretación de los mismos. La recodificación completa de las variables seleccionadas puede consultarse en el **Anexo A**.

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

De acuerdo con Salvador et al. (2003), el análisis exploratorio de Datos o *Exploratory Data Analysis* (EDA) es esencial para realizar cualquier análisis estadístico. Se trata de un conjunto de técnicas estadísticas utilizadas para proporcionar una comprensión básica de los datos y realizar evaluaciones previas al estudio.

Tras la recodificación, el EDA, trata de aportar una idea inicial de la información contenida en el conjunto de datos a utilizar. En primer lugar, cabe destacar que estamos tratando con variables categóricas a pesar de que inicialmente R las consideró como numéricas, por lo que se ha realizado una transformación en la tipología de las variables.

La **tabla 3.1** muestra un análisis básico con las observaciones contenidas por clase los atributos sociodemográficos. Atributos básicos como *Sexo*, *CCAA*, *Edad* y *Nacionalidad* no cuentan con observaciones NS/NC (99), mientras que los encuestados ven la clase social como el atributo más difícil en el que autoclasificarse. Por otro lado, hay algunas clases desbalanceadas, es el caso de *Situación Económica* tan sólo 9 encuestados declaran que la situación económica en España es “muy buena”, vemos cómo las respuestas se concentran en el extremo opuesto y lo clasifican como “muy mala”.

Tabla 3.1 Resumen de variables 1-12

Sexo	Comunidad Autónoma	Edad	Nacionalidad	Valoración Situación económica	Nivel de estudios
Hombre:1831	Cataluña :1384	De 18 a 24 años: 264	Nacionalidad española :3735	Muy buena: 9	Sin estudios / Estudios Primarios : 438
Mujer :1986	Andalucía : 517	De 25 a 34 años: 417	Nacionalidad española y otra: 82	Buena : 165	Estudios Secundarios / Formación Profesional:1883
	Madrid (Comunidad de): 401	De 35 a 44 años: 703		Regular : 263	Estudios Superiores :1483
	Comunitat Valenciana : 290	De 45 a 54 años: 773		Mala :1573	Otros estudios : 3
	Galicia : 189	De 55 a 64 años: 656		Muy mala :1698	NS/NC : 10
	Castilla y León : 166	Más de 65 años :1004		NS/NC : 109	
	(Other) : 870				
Religión	Estado civil	Situación laboral	Clase social	Confianza en el presidente	Confianza en la oposición
Católico :2269	Casado/a :2123	Trabajadores :2038	Clase alta : 13	Mucha confianza : 181	Mucha confianza : 49
Creyente de otra religión: 103	Soltero/a:1413	Jubilados o pensionistas :1078	Clase media :2790	Bastante confianza: 934	Bastante confianza: 333
No creyente : 823	Viudo/a : 270	Parados : 386	Clase baja : 782	Poca confianza :1207	Poca confianza :1438
Ateo/a : 574	NC : 11	Estudiantes : 158	Otra clase : 36	Ninguna confianza :1380	Ninguna confianza :1824
NC : 48		Trabajo doméstico no remunerado: 141	No cree en las clases: 24	NS/NC : 115	NS/NC : 173
		Otra situación : 7	NS/NC : 172		
		NC : 9			

Fuente: Elaboración propia con RStudio

La **tabla 3.2** contienen las variables relacionadas con las preguntas electorales y valorativas. Como vemos, los encuestados muestran preferencia por Pedro Sánchez como *presidente del Gobierno*, no obstante, la mayoría no muestra confianza hacia él, ni tampoco hacia la oposición. En cuanto a la simpatía partidista, los encuestados parecen decantarse por el PSOE, mientras que, en las valoraciones de los líderes políticos, coinciden mayoritariamente en la categoría 1, “muy mal” lo cual podría explicarse al contexto político-sanitario que atravesaba el país en diciembre de 2020.

Tabla 3.2 Resumen de variables (13-30)

Preferencia de presidente	Intención de voto	Intención de voto alternativo	Simpatía partidista	Ideología	Clasificación ideológica PS
Pedro Sánchez :1047	PP :384	PP : 243	PP :460	1 - Izquierda:345	1 - Izquierda:571
Pablo Casado : 310	PSOE :731	PSOE : 238	PSOE :948	2 :194	2 :293
Santiago Abascal: 183	Podemos :284	Podemos : 278	Podemos :338	3 :522	3 :593
Pablo Iglesias : 187	Ciudadanos:247	Ciudadanos : 288	Ciudadanos:310	4 :513	4 :568
Alberto Garzón : 70	VOX :229	VOX : 113	VOX :247	5 :988	5 :680
Inés Arrimadas : 281	ERC :156	ERC : 94	ERC :198	6 :343	6 :293
Iñigo Errejón : 108	JxCat : 82	JxCat : 62	JxCat : 99	7 :229	7 :145
Otro/a : 218	EAJ-PNV : 25	EAJ-PNV : 15	EAJ-PNV : 29	8 :132	8 : 69
Ninguno/a :1029	Otros :377	Otros : 495	Otros :259	9 : 37	9 : 12
NS/NC : 384	Abstención:478	No votaría a ningún otro partido: 85	Ninguno :683	10 - Derecha :101	10 - Derecha : 32
	NS/NC :824	Abstención : 802	NS/NC :246	NS/NC :413	No conoce : 2
		NS/NC :1104			NS/NC :559

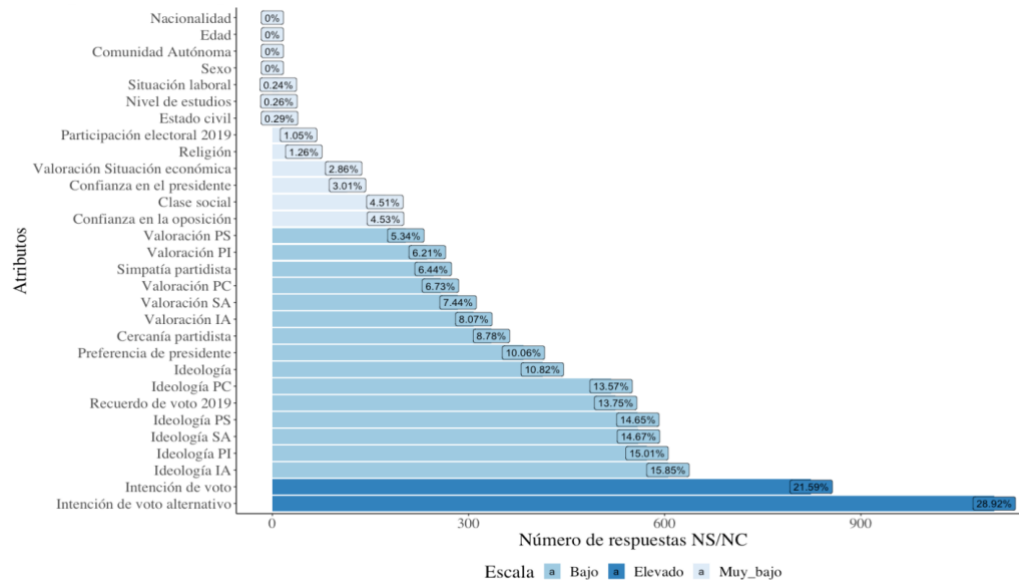
Clasificación ideológica PC	Clasificación ideológica PI	Clasificación ideológica SA	Clasificación ideológica IA	Valoración PS	Valoración PC
1 - Izquierda: 33	1 - Izquierda:1495	1 - Izquierda: 69	1 - Izquierda: 46	Muy mal :1123	Muy mal :1428
2 : 22	2 : 584	2 : 14	2 : 33	Mal : 604	Mal :1039
3 : 40	3 : 427	3 : 22	3 : 64	Regular : 904	Regular : 742
4 : 68	4 : 271	4 : 20	4 :143	Bien : 761	Bien : 244
5 :265	5 : 276	5 : 77	5 :762	Muy bien : 219	Muy bien : 74
6 :381	6 : 80	6 : 74	6 :681	No conoce: 2	No conoce: 33
7 :719	7 : 40	7 : 124	7 :456	NS/NC : 204	NS/NC : 257
8 :759	8 : 30	8 : 239	8 :354		
9 :423	9 : 6	9 : 368	9 :206		
10 - Derecha :556	10 - Derecha : 27	10 - Derecha :2173	10 - Derecha :305		
No conoce : 33	No conoce : 8	No conoce : 77	No conoce :162		
NS/NC :518	NS/NC : 573	NS/NC : 560	NS/NC :605		

Valoración PI	Valoración SA	Valoración IA	Participación electoral 2019	Recuerdo de voto 2019	Cercanía partidista
Muy mal :1615	Muy mal :2440	Muy mal :1132	Votó :3296	PP :407	PP :401
Mal : 677	Mal : 414	Mal : 954	No votó : 432	PSOE :905	PSOE :889
Regular : 785	Regular : 338	Regular : 932	No tenía derecho a voto: 49	Podemos :360	Podemos :356
Bien : 413	Bien : 163	Bien : 267	NS/NC : 40	Ciudadanos :267	Ciudadanos:354
Muy bien : 82	Muy bien : 101	Muy bien : 62		VOX :172	VOX :200
No conoce: 8	No conoce: 77	No conoce: 162		ERC :219	ERC :211
NS/NC : 237	NS/NC : 284	NS/NC : 308		JxCat :120	JxCat :113
				EAJ-PNV : 29	EAJ-PNV : 33
				Otros :267	Otros :238
				Abstención : 25	Ninguno :687
				No fué a votar:521	NS/NC :335
				NS/NC :525	

Fuente: Elaboración propia con RStudio

Son precisamente las variables valorativas e ideológicas en las que los encuestados declaran un mayor número de respuestas NS/NC. La **figura 3.1** muestra el porcentaje de valores faltantes en las variables del estudio y en ella se puede ver que *intención de voto* y *intención de voto alternativo* son las variables con mayor tasa de no respuesta.

Figura 3.1 Respuestas NS/NC por atributo



Fuente: Elaboración propia RStudio

3.1 Análisis univariante

A continuación, se realizará un análisis descriptivo de las variables más influyentes en el cuestionario del CIS.

La principal variable de estudio es la intención de voto, obtenida mediante la conducta electoral. La ideología es un poderoso predictor de la conducta de voto y es considerada una de las variables claves para producir una buena predicción de los resultados (Escobar et. al, 2013).

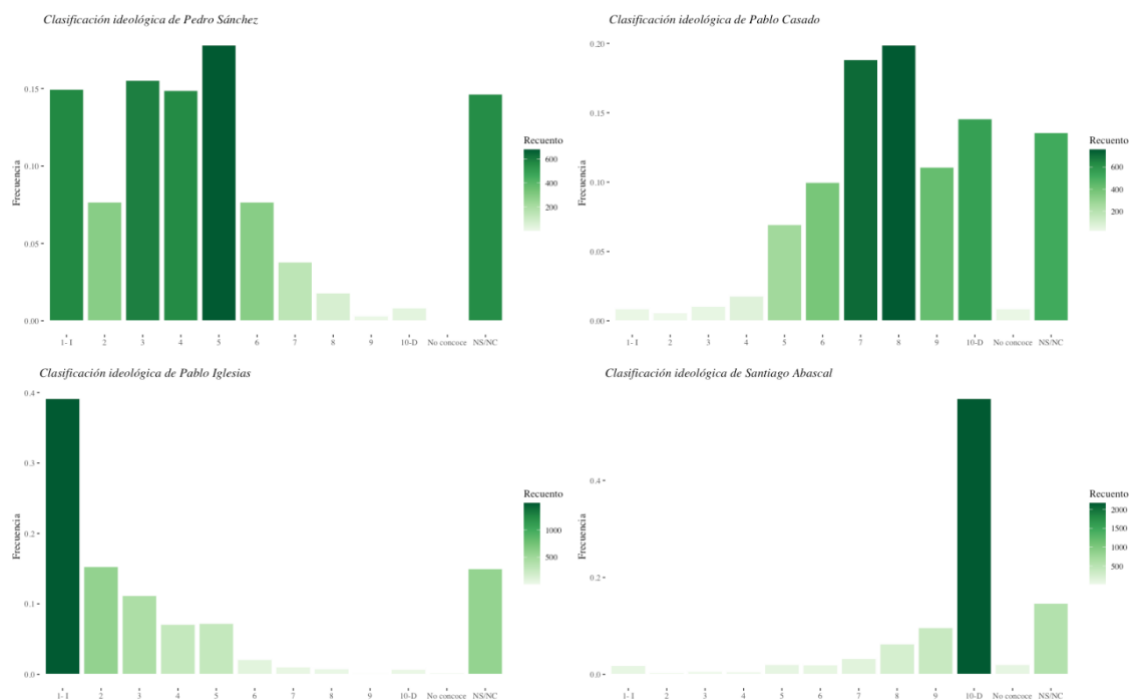
La **figura 3.2** muestra la clasificación ideológica de los diferentes líderes políticos en una escala del 1 al 10 siendo 1 completamente de izquierdas y 10 completamente de derechas, y como se puede ver, los encuestados clasifican mayoritariamente a Pedro Sánchez, líder del PSOE, entre un rango del 1 a 5, incluso mayoritariamente, se decantan por una escala ideológica de 5, que correspondería a partidos de centro. El caso contrario ocurre con Pablo Casado, donde los individuos lo clasifican en un rango entre 5 y 10, mayoritariamente en la clase 8 muy próxima a la extrema derecha.

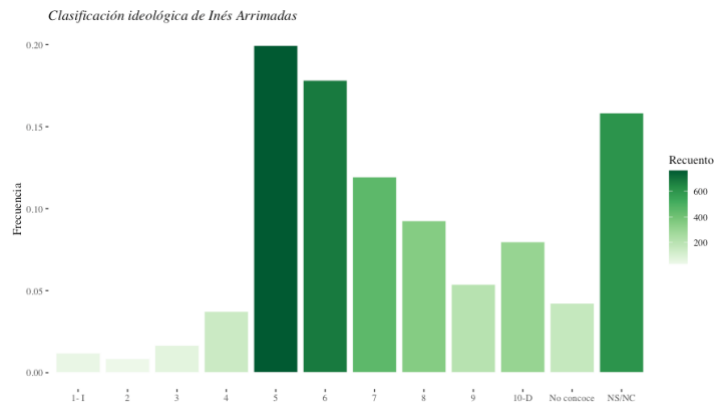
Por otro lado, la extrema izquierda es ocupada por Pablo Iglesias, líder de Podemos, donde la mayoría de encuestados coinciden clasificándolo en la clase 1, y con apenas observaciones de la clase 5 en adelante. No es el caso de Santiago Abascal, líder de VOX, donde más del 50% de encuestados lo clasifican dentro de la extrema derecha. No obstante, Inés Arrimadas, líder de Ciudadanos, es considerada por los encuestados, líder de un partido de centroderecha con valores entre las clases 5 y 10, mayoritariamente en la clase 5 y 6. Coincide con Pedro Sánchez en la clase 5, y ambos alcanzan un porcentaje superior al 15% en dicha clase.

Un aspecto a destacar es la categoría “No conoce”, donde Inés Arrimadas obtiene la mayor proporción de respuestas. Este resultado puede relacionarse con una menor popularidad de Ciudadanos en comparación con el resto de los partidos políticos. Notar que Pedro Sánchez, al ser el actual presidente del gobierno está muy presente en el marco político actual y, por tanto, carece de observaciones en dicha clase.

En cuanto a las observaciones NS/NC, vemos que la clasificación ideológica de los líderes políticos es una de las cuestiones con mayor número de valores faltantes, incluso entran en el top 10 de variables con mayor número de datos faltantes del cuestionario del CIS. En esta línea, Inés Arrimadas, es la líder que obtiene un mayor porcentaje, lo cual puede entenderse como indecisión en la clasificación de los partidos de centro. Sorprendentemente, Pedro Sánchez y Pablo Casado, líderes de partidos tradicionales de izquierda y derecha respectivamente también cuentan con un alto porcentaje de valores faltantes.

Figura 3.2 Clasificación ideológica de los líderes políticos



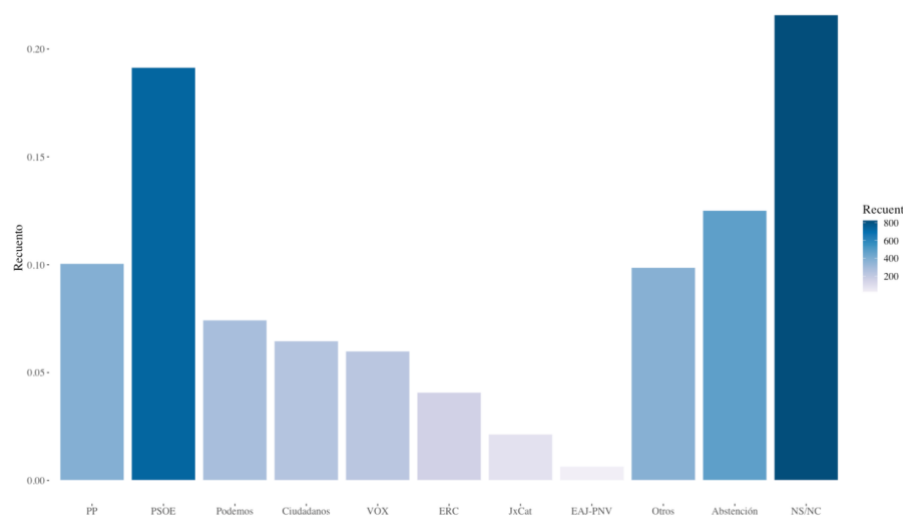


Fuente: Elaboración propia con RStudio

Por otro lado, la valoración de los líderes políticos generalmente no es informativa ya que incluye valoraciones de encuestados que nunca votarían a ciertos partidos e incluso los partidos pequeños pueden tener valoraciones más bajas que los partidos grandes. (Orriols, 2021) Lo correcto es conocer la valoración de sus propios votantes. En este aspecto, la **figura 3.3** muestra un primer acercamiento a la intención de voto de los encuestados, donde se ve claramente que parecen decantarse por el PSOE, incluso la abstención está muy presente. No obstante, PP sería la segunda fuerza política, seguida de Podemos, por lo que, junto con los partidos nacionalistas catalanes, la ideología izquierda predomina en los resultados del CIS.

Por encima de todo, las observaciones NS/NC tienen un gran peso en la variable intención de voto, incluso es la categoría más recurrida con un 21,59% de respuestas.

Figura 3.3 Intención de voto



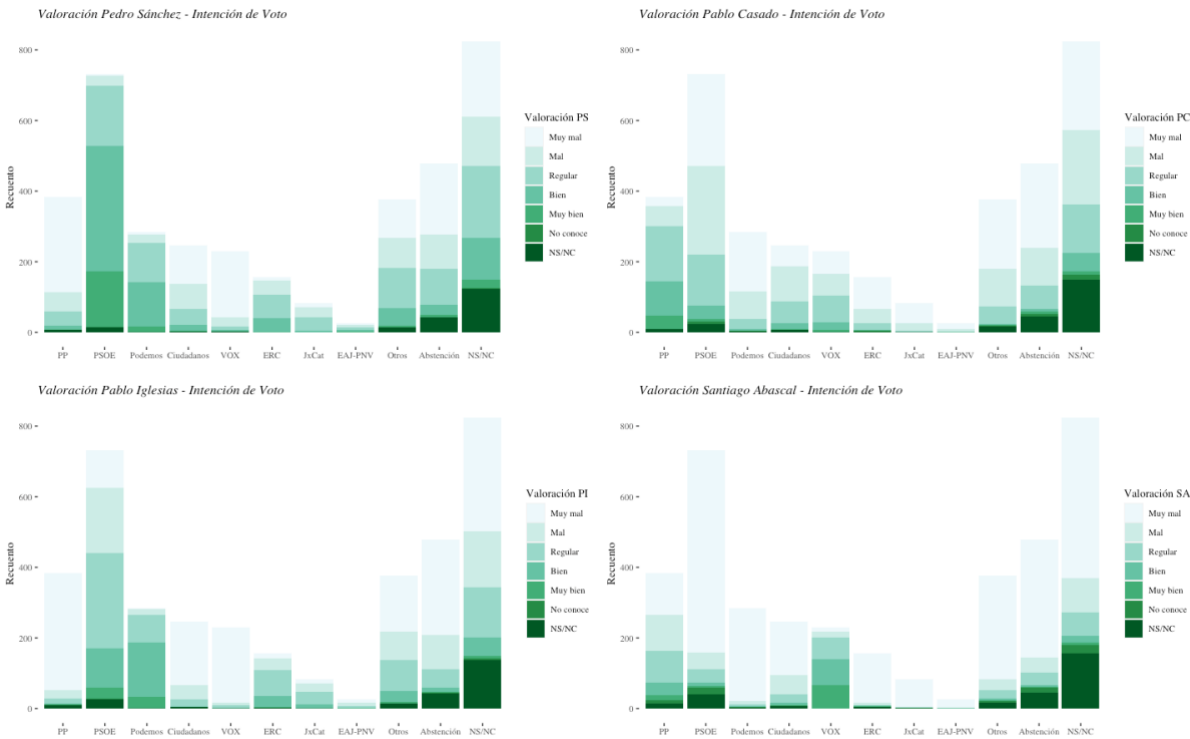
Fuente: Elaboración propia con RStudio

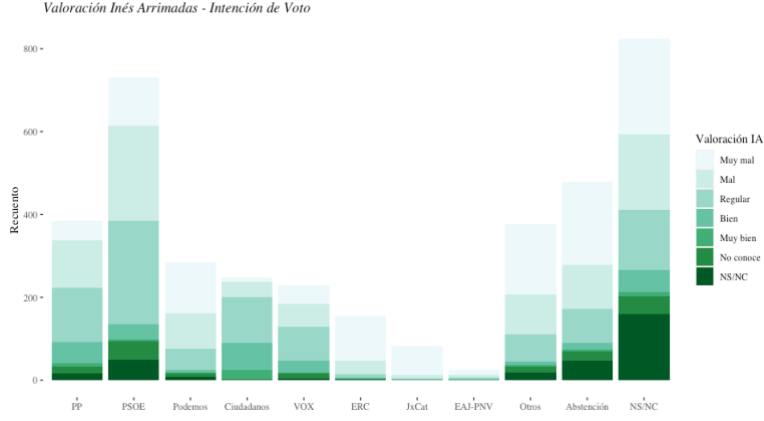
En relación con lo anterior, la **figura 3.4** muestra la valoración de los líderes políticos en función de la intención de voto de los encuestados. Como aspectos relevantes, notar que generalmente los propios votantes de un partido suelen valorar positivamente a los líderes del partido político, pero más allá de esto, son valorados negativamente por el resto de los partidos excepto aquellos en los que comparten ideología, donde las valoraciones son similares.

Es el caso de líderes como Pedro Sánchez, que recibe valoraciones positivas tanto de los miembros de su partido como de los de Podemos y viceversa. Lo mismo ocurre con Pablo Casado y votantes de derecha. En general, se observa una tendencia negativa en las valoraciones, sobre todo, en el caso de Santiago Abascal, líder de VOX, donde a excepción de votantes del PP, cuenta con muy malas valoraciones.

Por último, los encuestados que declaran NS/NC en intención de voto, parecen mostrar un patrón similar en cuanto a las valoraciones de los diferentes líderes políticos, sobre todo en valoraciones hacia Pedro Sánchez, Pablo Casado y Inés Arrimadas, a excepción de Pablo Iglesias y Santiago Abascal líderes que partidos de extrema izquierda y extrema derecha respectivamente donde vemos que cuentan con más valoraciones negativas de este grupo.

Figura 3.4 Valoración de los líderes políticos en función de la intención de voto





Fuente: Elaboración propia con RStudio

3.2 Análisis bivalente

Con el fin de identificar posibles asociaciones entre las variables del cuestionario del CIS, se realiza un análisis bivalente por medio de la medida *Tau de Goodman and Kruskal*.

De acuerdo con Pearson, 2020, la medida *Tau* (τ) *de Goodman and Kruskal* es una medida de asociación asimétrica entre variables categóricas, es decir, la asociación entre x y y generalmente no es la misma que entre y y x . Dicha asimetría es inherente a la forma en la que se define la medida *Tau* y es una característica muy útil de esta medida de asociación.

Ante dos variables categóricas x y y , con K y L valores distintos, se trata de cuantificar la medida de asociación entre dichas variables. Supongamos que tenemos N observaciones, donde x puede tomar el posible valor i y del mismo modo, y puede tomar el valor j . Se define así la tabla de contingencia N_{ij} que contabiliza el número de veces que $x = i$ e $y = j$:

$$N_{ij} = |\{k \mid x_k \rightarrow i, y_k \rightarrow j\}| \quad (1)$$

Se define así π_{ij} como la probabilidad de la clase $x = i$ e $y = j$.

$$\pi_{ij} = \frac{N_{ij}}{N} \quad (2)$$

La medida *Tau* de Goodman and Kruskal viene dada por:

$$\alpha(x, y) = \frac{V(y) - E[V(y|x)]}{V(y)} \quad (3)$$

$$V(y) = 1 - \sum_{j=1}^L \pi_{+j}^2 \quad (4)$$

$$E[V(y|x)] = 1 - \sum_{i=1}^K \sum_{j=1}^L \frac{\pi_{ij}^2}{\pi_{i+}} \quad (5)$$

donde π_{i+} y π_{+j} son las marginales definidas como:

$$\pi_{i+} = \sum_{j=1}^L \pi_{ij} \quad (6)$$

$$\pi_{+j} = \sum_{i=1}^K \pi_{ij} \quad (7)$$

Por tanto, sustituyendo estas expresiones en la Ecuación (1), se tiene finalmente la expresión explícita para la τ de Goodman and Kruskal:

$$\tau(x, y) = \frac{\sum_{i=1}^K \sum_{j=1}^L \left(\frac{\pi_{ij}^2 - \pi_{i+}^2 \pi_{+j}^2}{\pi_{+j}} \right)}{1 - \sum_{j=1}^L \pi_{+j}^2} \quad (8)$$

Notar que en la expresión (8) tenemos π_{+j} en el denominador a lo que provoca la asimetría al calcular $\tau(y, x)$. La τ de Goodman and Kruskal es una medida de asociación basada en la reducción proporcional del error, es decir, pertenece a una familia de medidas de asociación que tratan de cuantificar la variabilidad en una variable objetivo y que puede explicarse por variaciones en la variable x . Esta medida comprende valores entre 0 y 1 donde 0 indica no asociación y 1 una asociación perfecta. (Pearson, 2020).

La **tabla 3.4** muestra los resultados obtenidos tras aplicar la τ de Goodman and Kruskal. Se han considerados tres rangos: valores comprendidos entre 0.1 y 0.2, sombreados en rojo; valores entre 0.2 y 0.5, sombreados en amarillo; y por último valores superiores a 0.5, sombreados en verde. Esto es dado por la posibilidad de que para una variable haya varias variables que estén asociadas a ella con un valor de tau (τ) relativamente pequeño, pero al utilizar la información de todas las variables la predicción resulte razonable.

En cuanto a asociaciones fuertes tendríamos *simpatía partidista* e *intención de voto* (0.579), cuya relación es bidireccional a pesar de tener valores diferentes en la Tau, no obstante, es más fuerte en sentido opuesto: *intención de voto* – *simpatía partidista* (0.65). Por otro lado, *recuerdo de voto* y *participación en las elecciones de 2019* están fuertemente asociadas (0.832), mientras que inversamente dicha relación es débil (0.156).

Si miramos asociaciones comprendidas en el segundo rango, se tienen relaciones entre variables valorativas e ideológicas, es el caso de *preferencia de presidente del gobierno* con *confianza en el presidente* (0.204), *intención de voto* (0.255) y *simpatía partidista* (0.27), incluso asociaciones más débiles con la *valoración* de los diferentes líderes

políticos, el *recuerdo de voto* (0.12) y la *cercanía partidista* (0.179). Por otro lado, *intención de voto* guarda relación con las variables citadas anteriormente, aunque con diferentes valores de Tau. En el caso de la valoración de los líderes político, cabe destacar que en el caso de *valoración de Pablo Casado* existe asociación con *confianza en la oposición*, liderada por Pablo Casado (0.215), por lo que a la vista de los resultados las asociaciones parecen razonables.

En cuanto a variables sociodemográficas, existe una ligera asociación entre *edad* y *situación laboral* (0.218), al igual que entre *edad* y *estado civil* (0.437). No obstante, esta asociación no se cumple en sentido contrario, ya que *estado civil* no está relacionada con ninguna variable, y en el caso de *situación laboral* también presenta asociación con la *edad* del encuestado (0.258)

Tabla 3.4 Tau de Goodman and Kruskal variables (1-12)

	Sexo	Comunidad Autónoma	Edad	Nacionalidad	Valoración Situación económica	Nivel de estudios
Sexo	2	0	0.001	0	0	0.003
Comunidad Autónoma	0.006	19	0.004	0.006	0.009	0.012
Edad	0.003	0.002	6	0.003	0.008	0.076
Nacionalidad	0	0	0.001	2	0.002	0.001
Valoración Situación económica	0.004	0.002	0.009	0.008	6	0.01
Nivel de estudios	0.011	0.003	0.062	0.002	0.005	5
Religión	0.023	0.007	0.024	0.014	0.008	0.015
Estado civil	0.027	0.001	0.078	0.001	0.004	0.028
Situación laboral	0.045	0.003	0.258	0.007	0.004	0.081
Clase social	0.004	0.002	0.009	0.001	0.004	0.048
Confianza en el presidente	0.004	0.01	0.006	0.008	0.083	0.006
Confianza en la oposición	0.006	0.01	0.007	0.003	0.007	0.012
Preferencia de presidente	0.017	0.021	0.015	0.005	0.051	0.018
Intención de voto	0.02	0.059	0.018	0.007	0.047	0.016
Intención de voto alternativo	0.011	0.034	0.013	0.005	0.023	0.012
Simpatía partidista	0.018	0.068	0.016	0.006	0.045	0.018
Ideología	0.016	0.016	0.011	0.006	0.026	0.036
Clasificación ideológica PS	0.015	0.024	0.014	0.005	0.034	0.037
Clasificación ideológica PC	0.025	0.016	0.015	0.011	0.012	0.049
Clasificación ideológica PI	0.024	0.032	0.012	0.013	0.03	0.039
Clasificación ideológica SA	0.037	0.007	0.014	0.012	0.013	0.054
Clasificación ideológica IA	0.031	0.043	0.014	0.008	0.016	0.049
Valoración PS	0.017	0.015	0.013	0.005	0.079	0.016
Valoración PC	0.017	0.008	0.01	0.006	0.011	0.019
Valoración PI	0.009	0.022	0.01	0.001	0.05	0.01
Valoración SA	0.016	0.01	0.009	0.007	0.012	0.025
Valoración IA	0.019	0.021	0.009	0.004	0.013	0.022
Participación electoral 2019	0.002	0.001	0.012	0.014	0.004	0.015
Recuerdo de voto 2019	0.017	0.074	0.024	0.013	0.035	0.024
Cercanía partidista	0.016	0.071	0.018	0.004	0.042	0.02

	Religión	Estado civil	Situación laboral	Clase social	Confianza en el presidente	Confianza en la oposición
Sexo	0.01	0.005	0.008	0.001	0	0.001
Comunidad Autónoma	0.019	0.008	0.005	0.006	0.021	0.023
Edad	0.061	0.218	0.437	0.011	0.008	0.008
Nacionalidad	0.001	0.001	0.002	0.001	0.001	0.001
Valoración Situación económica	0.007	0.005	0.009	0.009	0.067	0.009
Nivel de estudios	0.024	0.026	0.113	0.072	0.004	0.008
Religión	5	0.044	0.032	0.008	0.016	0.022
Estado civil	0.044	4	0.095	0.011	0.003	0.002
Situación laboral	0.046	0.123	7	0.022	0.007	0.007
Clase social	0.005	0.007	0.021	6	0.005	0.01
Confianza en el presidente	0.02	0.004	0.008	0.011	5	0.038
Confianza en la oposición	0.027	0.003	0.012	0.017	0.029	5
Preferencia de presidente	0.054	0.014	0.024	0.021	0.204	0.09
Intención de voto	0.072	0.023	0.024	0.021	0.191	0.09
Intención de voto alternativo	0.045	0.01	0.017	0.016	0.085	0.071
Simpatía partidista	0.087	0.027	0.021	0.021	0.191	0.095
Ideología	0.068	0.013	0.018	0.033	0.097	0.081
Clasificación ideológica PS	0.043	0.011	0.027	0.035	0.112	0.041
Clasificación ideológica PC	0.03	0.011	0.028	0.042	0.042	0.107
Clasificación ideológica PI	0.041	0.013	0.023	0.045	0.083	0.044
Clasificación ideológica SA	0.029	0.009	0.034	0.041	0.038	0.051
Clasificación ideológica IA	0.038	0.012	0.032	0.044	0.047	0.068
Valoración PS	0.022	0.007	0.022	0.022	0.365	0.027
Valoración PC	0.036	0.006	0.017	0.023	0.016	0.215
Valoración PI	0.045	0.009	0.015	0.025	0.201	0.035
Valoración SA	0.033	0.006	0.018	0.024	0.048	0.09
Valoración IA	0.017	0.008	0.02	0.026	0.012	0.099
Participación electoral 2019	0.005	0.019	0.016	0.006	0.004	0.006
Recuerdo de voto 2019	0.085	0.04	0.031	0.02	0.145	0.074
Cercanía partidista	0.085	0.025	0.028	0.023	0.161	0.091

Fuente: Elaboración propia con RStudio

Tabla 3.5 Tau de Goodman and Kruskal variables (13-30)

	Preferencia de presidente	Intención de voto	Intención de voto alternativo	Simpatía partidista	Ideología	Clasificación ideológica PS
Sexo	0.002	0.003	0.003	0.002	0.002	0.002
Comunidad Autónoma	0.012	0.016	0.01	0.021	0.01	0.015
Edad	0.008	0.01	0.01	0.008	0.006	0.008
Nacionalidad	0.001	0.001	0	0.001	0.001	0.001
Valoración Situación económica	0.027	0.016	0.009	0.018	0.008	0.014
Nivel de estudios	0.006	0.007	0.009	0.006	0.011	0.015
Religión	0.008	0.015	0.01	0.017	0.018	0.013
Estado civil	0.003	0.005	0.003	0.005	0.005	0.006
Situación laboral	0.008	0.008	0.008	0.007	0.007	0.01
Clase social	0.006	0.005	0.007	0.005	0.011	0.011
Confianza en el presidente	0.136	0.083	0.024	0.091	0.029	0.045
Confianza en la oposición	0.045	0.036	0.023	0.038	0.028	0.024
Preferencia de presidente	10	0.255	0.069	0.27	0.052	0.051
Intención de voto	0.287	11	0.393	0.65	0.058	0.055
Intención de voto alternativo	0.085	0.336	12	0.143	0.034	0.033
Simpatía partidista	0.259	0.579	0.121	11	0.07	0.062
Ideología	0.062	0.063	0.035	0.084	11	0.071
Clasificación ideológica PS	0.043	0.042	0.025	0.048	0.05	12
Clasificación ideológica PC	0.035	0.034	0.026	0.038	0.063	0.114
Clasificación ideológica PI	0.039	0.034	0.026	0.041	0.052	0.155
Clasificación ideológica SA	0.029	0.029	0.026	0.032	0.051	0.086
Clasificación ideológica IA	0.034	0.036	0.027	0.043	0.055	0.09
Valoración PS	0.136	0.093	0.03	0.104	0.036	0.077
Valoración PC	0.043	0.039	0.029	0.043	0.034	0.04
Valoración PI	0.093	0.067	0.037	0.075	0.049	0.059
Valoración SA	0.055	0.053	0.032	0.059	0.042	0.042
Valoración IA	0.041	0.033	0.027	0.036	0.024	0.031
Participación electoral 2019	0.006	0.014	0.012	0.01	0.008	0.007
Recuerdo de voto 2019	0.129	0.277	0.078	0.318	0.059	0.053
Cercanía partidista	0.182	0.305	0.077	0.446	0.073	0.064

	Clasificación ideológica PC	Clasificación ideológica PI	Clasificación ideológica SA	Clasificación ideológica IA	Valoración PS	Valoración PC
Sexo	0.004	0.004	0.005	0.005	0.002	0.003
Comunidad Autónoma	0.012	0.029	0.012	0.019	0.021	0.015
Edad	0.01	0.008	0.011	0.008	0.008	0.007
Nacionalidad	0.001	0.001	0.003	0.001	0.001	0.002
Valoración Situación económica	0.007	0.018	0.016	0.008	0.05	0.011
Nivel de estudios	0.019	0.018	0.032	0.018	0.006	0.009
Religión	0.01	0.02	0.031	0.011	0.013	0.02
Estado civil	0.006	0.008	0.009	0.005	0.004	0.004
Situación laboral	0.011	0.009	0.017	0.011	0.007	0.006
Clase social	0.015	0.021	0.023	0.014	0.008	0.011
Confianza en el presidente	0.02	0.06	0.041	0.022	0.266	0.018
Confianza en la oposición	0.041	0.042	0.054	0.028	0.024	0.151
Preferencia de presidente	0.04	0.08	0.064	0.043	0.153	0.064
Intención de voto	0.038	0.08	0.065	0.047	0.148	0.062
Intención de voto alternativo	0.027	0.052	0.053	0.033	0.069	0.05
Simpatía partidista	0.046	0.096	0.077	0.056	0.15	0.073
Ideología	0.078	0.099	0.131	0.062	0.076	0.066
Clasificación ideológica PS	0.113	0.234	0.147	0.096	0.117	0.042
Clasificación ideológica PC	12	0.154	0.282	0.155	0.043	0.113
Clasificación ideológica PI	0.118	12	0.192	0.114	0.078	0.039
Clasificación ideológica SA	0.146	0.143	12	0.126	0.039	0.053
Clasificación ideológica IA	0.163	0.146	0.219	12	0.044	0.069
Valoración PS	0.04	0.089	0.069	0.039	7	0.093
Valoración PC	0.075	0.054	0.081	0.05	0.073	7
Valoración PI	0.047	0.118	0.082	0.05	0.265	0.095
Valoración SA	0.051	0.063	0.178	0.058	0.073	0.174
Valoración IA	0.048	0.057	0.078	0.147	0.052	0.208
Participación electoral 2019	0.01	0.012	0.023	0.01	0.004	0.006
Recuerdo de voto 2019	0.038	0.088	0.076	0.048	0.117	0.053
Cercanía partidista	0.046	0.095	0.08	0.055	0.126	0.071

	Valoración PC	Valoración PI	Valoración SA	Valoración IA	Participación electoral 2019	Recuerdo de voto 2019	Cercanía partidista
Sexo	0.003	0.001	0.002	0.003	0	0.002	0.002
Comunidad Autónoma	0.015	0.033	0.022	0.025	0.007	0.022	0.022
Edad	0.007	0.01	0.007	0.005	0.023	0.015	0.009
Nacionalidad	0.002	0	0.002	0.001	0.008	0.002	0.001
Valoración Situación económica	0.011	0.04	0.013	0.011	0.011	0.013	0.016
Nivel de estudios	0.009	0.004	0.015	0.009	0.012	0.007	0.008
Religión	0.02	0.028	0.032	0.01	0.015	0.02	0.017
Estado civil	0.004	0.004	0.005	0.003	0.013	0.008	0.006
Situación laboral	0.006	0.007	0.008	0.006	0.025	0.011	0.009
Clase social	0.011	0.009	0.018	0.009	0.008	0.006	0.006
Confianza en el presidente	0.018	0.192	0.054	0.015	0.013	0.055	0.068
Confianza en la oposición	0.151	0.035	0.105	0.066	0.017	0.026	0.031
Preferencia de presidente	0.064	0.173	0.136	0.059	0.02	0.12	0.179
Intención de voto	0.062	0.168	0.133	0.068	0.073	0.29	0.338
Intención de voto alternativo	0.05	0.094	0.092	0.048	0.042	0.074	0.084
Simpatía partidista	0.073	0.177	0.152	0.08	0.041	0.303	0.437
Ideología	0.066	0.108	0.13	0.041	0.039	0.067	0.082
Clasificación ideológica PS	0.042	0.103	0.068	0.041	0.025	0.04	0.048
Clasificación ideológica PC	0.113	0.059	0.106	0.063	0.044	0.031	0.037
Clasificación ideológica PI	0.039	0.131	0.068	0.048	0.032	0.037	0.04
Clasificación ideológica SA	0.053	0.05	0.206	0.052	0.052	0.03	0.032
Clasificación ideológica IA	0.069	0.068	0.095	0.213	0.045	0.038	0.043
Valoración PS	0.093	0.323	0.132	0.072	0.016	0.065	0.079
Valoración PC	7	0.09	0.271	0.194	0.021	0.031	0.04
Valoración PI	0.095	7	0.151	0.077	0.014	0.052	0.063
Valoración SA	0.174	0.1	7	0.149	0.022	0.042	0.054
Valoración IA	0.208	0.067	0.219	7	0.023	0.028	0.038
Participación electoral 2019	0.006	0.004	0.011	0.007	4	0.156	0.011
Recuerdo de voto 2019	0.053	0.138	0.125	0.065	0.832	12	0.325
Cercanía partidista	0.071	0.152	0.143	0.077	0.046	0.322	11

Fuente: elaboración propia con RStudio

A la vista de los resultados, la existencia de asociaciones leves hace que se plantee si los valores de Tau son verdaderamente significativos, es decir, si realmente son distintos de cero. Para ello se ha utilizado un procedimiento bootstrap para comprobar que esas dependencias son no nulas. Se ha utilizado un intervalo de confianza t-bootstrap (Efron y Tibshirani, 1993).

Como resultado de este procedimiento, se tiene que los intervalos calculados para los tres rangos de la medida Tau, no contienen el valor cero, por tanto, se concluye que los coeficientes son no nulos. Los intervalos están disponibles en el **Anexo B**.

4. METODOLOGÍA

Los datos faltantes de la encuesta del CIS se imputarán mediante un procedimiento de imputación múltiple (Rubin, 1987). La imputación múltiple consiste en crear predicciones de cada observación faltante m veces, creando así m bases de datos completas. En cada base de datos completa los valores observados son los mismos, pero los valores faltantes presentan valores faltantes diferentes (Rivas et al., 2010).

La imputación múltiple es necesaria para inferencias conjuntas sobre múltiples parámetros, dado que incluso con pequeñas proporciones de valores faltantes un procedimiento de imputación simple puede verse afectado (Schafer, 1999). En este sentido, la imputación simple estima el parámetro, pero se ignora la estimación de la variabilidad, produciendo una subestimación de los errores (Barceló, 2008). La eliminación de las observaciones no es una opción, ya que cuando las observaciones a eliminar difieren sistemáticamente del resto, las estimaciones pueden estar sesgadas

Una de las principales cuestiones sobre la imputación múltiple es fijar el valor del parámetro m . Según Rubin, 1987 la eficiencia relativa de una estimación es aproximadamente $(1+\lambda/m)^{-1}$, por lo que con un 50% de valores faltantes una estimación en $m = 5$ imputaciones tiene una desviación estándar que es tan sólo un 5% más elevada que una estimación basada en $m = \infty$, por lo que a menos que el porcentaje de datos faltantes sean muy elevado resulta de poca utilidad fijar una m superior a cinco imputaciones. Generalmente, valores de $m = 3$ funcionan bien en la práctica (Rubin, 1996).

La imputación múltiple puede dividirse en tres pasos: imputación, análisis y combinación (Horton y Lipstiz, 2001). Para realizar la imputación múltiple es necesario especificar un modelo, en este las estimaciones se realizan por medio del algoritmo *Random Forest*. Una vez estimados los parámetros con dicho algoritmo, se tomará la media aritmética de las estimaciones de cada conjunto de datos generado, mientras que los errores estándar vendrán dados por una los errores estándar dentro de los conjuntos de datos imputados (varianza dentro de la imputación) y la varianza entre las imputaciones (Marín y Restrepo, 2012).

La varianza dentro de la imputación viene dada por la media aritmética de los errores estándar al cuadrado:

$$W = \frac{\sum_{t=1}^m SE_t^2}{m}, \quad (9)$$

donde t es el conjunto de datos imputado y m el número total de imputaciones ($m = 3$).

La varianza entre imputaciones se calcula como:

$$B = \frac{\sum_{t=1}^m (\widehat{\theta}_t - \bar{\theta})^2}{m - 1}, \quad (10)$$

donde $\widehat{\theta}_t$ es la variable estimada del conjunto imputado t y $\bar{\theta}$ la media de la variable estimada en las m imputaciones.

Por último, el error estándar combinado se calcula por medio de la varianza dentro de la imputación y la varianza entre imputaciones

$$SE = \sqrt{W + B + (B / m)} \quad (11)$$

Por tanto, la imputación múltiple corrige el problema de la subestimación de los errores estándar de la imputación simple ya que incorpora la varianza entre imputaciones en el error estándar (Marín y Restrepo, 2012)

4.1 Árboles de clasificación

Previamente a la definición del algoritmo *Random Forest*, es conveniente introducir un breve contexto del algoritmo de los árboles de clasificación o *Classification and Regression Trees* (CART) (Breiman et al., 1984).

Lo que sigue se basa en la descripción de los árboles de decisión que hacen Alfaro et al. (2009). Los árboles de clasificación constituyen un método de clasificación no paramétrico dotado de gran flexibilidad. Gráficamente pueden representarse mediante nodos y ramas donde cada nodo es una decisión sobre uno de los atributos o variables. De cada nodo pueden salir dos o más ramas, dependiendo de si la respuesta es binaria y al alcanzar los nodos u hojas finales, el árbol toma una decisión sobre la clase asignada.

La división de los nodos se realiza en función de una prueba que se plantea sobre el valor de alguno de los atributos o variables. Si se trata de nodos binarios, las observaciones que cumplen una de las condiciones se asignan a un nodo hijo y los restantes a otro. De todas las posibles particiones se elige aquella que obtiene un mayor incremento en la homogeneidad de los nodos, para lo cual se utiliza la medida de Entropía o el Índice de Gini.

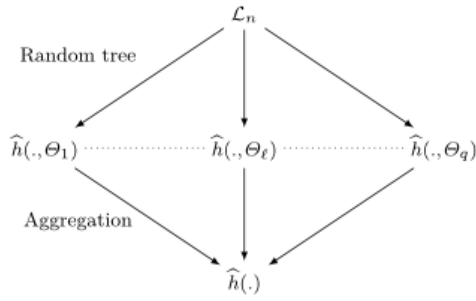
Finalmente, la subdivisión finaliza cuando el nodo es puro, es decir, todas las observaciones del nodo son de la misma clase, o cuando se supere un determinado nivel de pureza en términos de la proporción de la clase mayoritaria (Alfaro et al., 2009)

4.2 Random Forest

Una vez introducidos los árboles de clasificación, se puede entender la definición general del algoritmo de *Random Forest* dada por Leo Breiman (2001):

“El clasificador Random Forest o bosque aleatorio es un clasificador que consta de una colección de árboles estructurados en $\{h(x, \Theta_q), q = 1, \dots, \}$ donde Θ_q son vectores aleatorios independientes e idénticamente distribuidos y cada árbol emite un voto unitario para la clase más popular en la entrada x ” (Breiman, 2001) El predictor $\hat{h}_{RF}$ se obtiene mediante la agregación de árboles aleatorios (Genuer y Poggi, 2020). Dicha agregación se realiza acorde a la **figura 4.1**.

Figura 4.1 Esquema general de Random Forest



Fuente: Genuer y Poggi (2020)

Por tanto, el clasificador $\hat{h}_{RF}$ viene dado por:

$$\hat{h}_{RF} = \arg \max \sum_{\ell=1}^q 1_{\hat{h}(x, \Theta_{\ell}) = c} \quad (12)$$

De acuerdo con Livingston (2005), el procedimiento de *Random Forest* comienza formando un conjunto de datos *inbag* mediante un muestreo con reemplazamiento del conjunto de entrenamiento. Esta técnica se conoce como “*bootstrapping*”. En dicho conjunto *inbag* el número de observaciones es igual a la del conjunto de datos de entrenamiento. Al realizarse un muestreo con reemplazamiento, puede haber observaciones duplicadas del conjunto de entrenamiento. En general, un tercio de los datos de entrenamiento suele no estar en el conjunto *inbag*, dichos datos se conocen como observaciones *out-of-bag*.

A continuación, se elige un número aleatorio de atributos para cada árbol y dichos atributos forman los nodos y hojas mediante el algoritmo estándar de los árboles de clasificación que describimos en la sección anterior.

Tras ello, se repite el mismo procedimiento desarrollando múltiples árboles aleatorios sin podar. Las observaciones *out-of-bag* se utilizan para validar tanto los árboles individuales como el bosque completo. Se obtiene así el error *out-of-bag*, el cual es el error medio de cada observación calculado mediante los árboles que no contienen dicha observación en la muestra que se ha utilizado para estimarlos (Hastie et al., 2009). Este error es muy útil para predecir el comportamiento de Random Forest y evitar el sesgo optimista de la estimación del error cuando se utilizan todas las observaciones del conjunto de entrenamiento. Además, aporta tasas de error muy similares al error en el conjunto de prueba y se puede utilizar para optimizar los parámetros de Random Forest (Genuer y Poggi, 2020)

El algoritmo Random Forest cuenta con varios parámetros ajustables, entre los que destacan el número de árboles o *ntree*. Este parámetro no es crucial, ya que los bosques aleatorios no se sobreajustan a medida que se agregan más árboles, sino que producen un valor límite en el error, es decir, el error se estabiliza cuando se alcanza un número de árboles (Breiman, 2001). El parámetro más importante de Random Forest es el número de variables elegidas en cada nodo o *mtry*, diferentes valores de *mtry* pueden producir predicciones muy diferentes. El valor óptimo de *mtry* depende en gran medida del número total de variables y del número de variables relacionadas con el atributo a predecir (Genuer y Poggi, 2020). Para elegir el valor óptimo de *mtry* se ha utilizado un procedimiento de validación cruzada con cinco submuestras variando este hiperparámetro con valores entre dos y diez. Dicho procedimiento consiste en dividir aleatoriamente el conjunto de observaciones en *k* grupos aproximadamente del mismo tamaño. En este caso, se ha fijado *k* igual a cinco. El primer grupo se utiliza como conjunto de validación y como entrenamiento se utilizan los *k-1* grupos restantes. Se repite *k* veces este procedimiento tomando cada vez un grupo diferente de validación (James et al., 2013).

Generalmente, con un *mtry* demasiado bajo el error *out-of-bag* es relativamente alto, mientras que si aumentamos el *mtry* el error disminuye rápidamente hasta llegar a un mínimo, luego, el error aumenta gradualmente (Genuer y Poggi, 2020)

4.3 Partición de la muestra y procedimiento de imputación

La variable de interés en el estudio es la *intención de voto* por lo que la partición se realiza con el objetivo de predecir dicha variable. Para realizar la partición de la muestra vamos a subdividir la muestra dos veces:

- Por un lado, tenemos las observaciones en las que las personas encuestadas han respondido a intención de voto, concretamente son 2993, por tanto, las observaciones con “NS/NC” como respuesta son 824. Partiendo de esta base,

dividiremos aleatoriamente en una muestra de entrenamiento y validación las observaciones **con respuesta** en *intención de voto*. Esto se debe a que carece de sentido evaluar el modelo con observaciones sin respuesta en *intención de voto*. Esta partición se realiza utilizando dos tercios de las observaciones como conjunto de entrenamiento y un tercio como conjunto de validación.

- Por otro lado, los valores **sin respuesta** en *intención de voto* se agruparán junto con el conjunto de entrenamiento creado formando un nuevo conjunto de datos que utilizaremos para implementar el algoritmo *Random Forest*. No obstante, no utilizaremos este conjunto completo como entrenamiento del modelo, sino las observaciones **sin valores faltantes** en la variable que vamos a imputar.

En este sentido, diferenciaremos entre los datos que tienen respuesta y aquellos con valores faltantes, que llamaremos datos a imputar. Notar que tendremos valores faltantes tanto en esta segunda subdivisión, como posibles valores faltantes de las variables distintas a intención de voto que hayan entrado en el conjunto de validación inicial.

En definitiva, una vez entrenado el modelo, realizaremos predicciones tanto para los datos a imputar, como para el conjunto de validación inicial donde utilizaremos las observaciones con respuesta en la variable imputar como validación, dado que para realizar dicha validación es necesario conocer el valor de la variable.

Una vez realizada la validación, vamos a imputar ambas predicciones que correspondan a valores “NS/NC” en la variable que estamos imputando y formaremos así el primer conjunto de datos imputado. Al tratarse de una imputación múltiple se repetirá la estimación del modelo tres veces más y finalmente contaremos con tres conjuntos de datos imputados completos.

5. RESULTADOS

El siguiente paso tras realizar la partición de la muestra es implementar el algoritmo *Random Forest* para cada una de las variables que vamos a imputar. El procedimiento de imputación se realizará de menor a mayor número de respuestas “NS/NC”. Si recordamos, las variables sociodemográficas eran las que tenían menor número de valores faltantes por lo que serán las primeras en imputarse, finalizando con las preguntas clave *intención de voto* e *intención de voto alternativo*. Para cada una de las variables se estimará el modelo tres veces por lo que, por consiguiente, se obtienen tres resultados, que se mostrarán en una matriz de confusión conjunta calculada como el promedio de la matriz de confusión resultante en las tres ejecuciones.

No obstante, dado el alto número de atributos a imputar, únicamente se mostrarán algunos ejemplos de las variables más relevantes en la imputación de *intención de voto*, el resto de los atributos con sus matrices de confusión y gráficos correspondientes, podrán consultarse en el **Anexo C**. En este sentido, se desglosarán variables con un número alto y bajo de valores faltantes para comparar el funcionamiento del modelo ambos casos. Dentro del bloque de preguntas sociodemográficas y, por ende, preguntas con un menor número de valores faltantes, elegiremos el caso con menos valores faltantes, *situación laboral del encuestado*; dentro de las preguntas ideológicas elegiremos la simpatía partidista y la auto-ubicación ideológica ambas con un número medio de valores faltantes; y por último intención de voto y voto alternativo, variables con mayor número de respuestas “NS/NC”.

Al final del capítulo, se mostrará la exactitud media de los modelos de todos los atributos de forma conjunta para una mejor comprensión e interpretación de los resultados.

5.1 Imputación de la situación laboral del encuestado

Dado que la imputación se realizará de menor a mayor número de valores faltantes, la primera variable que imputaremos será la situación laboral del encuestado la cual tenía nueve respuestas con “NS/NC”, un número altamente bajo. La tabla 5.1 muestra los resultados en las tres implementaciones sucesivas del modelo.

Tabla 5.1 Matrices de confusión de las imputaciones en la situación laboral

Matriz de confusión Situación laboral 1ª imputación							Matriz de confusión Situación laboral 2ª imputación						
Predicción	Referencia						Predicción	Referencia					
	1	2	3	4	5	6		1	2	3	4	5	6
1	529	43	103	39	19	1	1	531	43	103	39	19	1
2	18	225	1	0	12	0	2	17	225	1	0	12	0
3	1	0	0	0	0	0	3	0	0	0	0	0	0
4	0	0	0	5	0	0	4	0	0	0	5	0	0
5	0	0	0	0	0	0	5	0	0	0	0	0	0
6	0	0	0	0	0	0	6	0	0	0	0	0	0

Matriz de confusión Situación laboral 3ª imputación

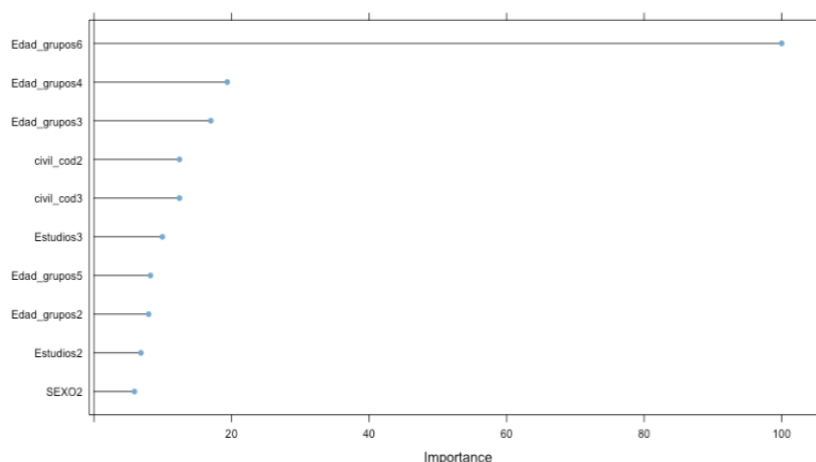
Predicción	Referencia					
	1	2	3	4	5	6
1	531	43	103	38	19	1
2	17	225	1	0	12	0
3	0	0	0	0	0	0
4	0	0	0	6	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Fuente: Elaboración propia con RStudio

Como se puede ver, el modelo tiene valores altos en la diagonal, signo de una alta precisión. Las matrices de confusión son muy similares en las tres imputaciones, como vemos difieren en clasificar alguna categoría de la 3 a la 6 como clase 1, es decir, fallar ubicando las observaciones en la categoría 1 correspondiente a los trabajadores. Este problema puede estar relacionado con que *situación laboral* tiene clases minoritarias, en este caso, las categorías 1 y 2 cuentan con 1490 y 810 observaciones respectivamente, pero el resto de las categorías no son mayoritarias y por ello, el modelo tiende a fallar. Una posible extensión de este trabajo es incorporar alguna medida para la corrección de este problema de clases poco representadas.

Aún así el modelo aporta una exactitud media en las tres imputaciones superior al 70%, concretamente un 74.4%, y el error *out-of-bag*, lo cual interpretarse como el error del conjunto de validación es de un 22.9%. A la vista de los resultados, cabe preguntarse qué variables han sido decisivas en el modelo y cuáles constan de mayor importancia. La **figura 5.1** muestra la importancia de las variables que entran dentro del modelo. Cabe destacar que las variables importantes resultan ser las mismas en las tres imputaciones.

Figura 5.1 Importancia de variables en el modelo de la situación laboral del encuestado



Fuente: Elaboración propia con RStudio

Las variables que tienen mayor importancia en el modelo corresponden a variables sociodemográficas. La *edad* consta de gran importancia, siendo el grupo 6 correspondiente a personas mayores de 65 la más importante. Vemos que el grupo 1 correspondiente a encuestados con una edad entre 18 y 24 años no parece constar de importancia en la predicción de la situación laboral, esto puede deberse a que normalmente el salto al mundo laboral suele darse sobre los 24 años. Por otro lado, también vemos variables como el *estado civil* y el *sexo*, pero su importancia es inferior a un 20%, por lo que el mayor peso recae en la *edad*.

En cuanto al parámetro *mtry* estimado mediante validación cruzada de 5-*folds* en la primera y tercera imputación el valor elegido fue 10, mientras que en la segunda eligió 9, no obstante, la medida de *accuracy* fue similar en los tres casos con un 0.79.

5.2 Imputación de la simpatía partidista

La simpatía partidista consta de 246 valores faltantes, por lo que en este caso se evaluará el funcionamiento del modelo con un número de valores faltantes medio-bajo. La **tabla 5.2** muestra la matriz de confusión de las 3 imputaciones realizadas.

Tabla 5.2 Matrices de confusión de las imputaciones en la simpatía partidista

Matriz de confusión Simpatía partidista 1ª imputación											Matriz de confusión Simpatía partidista 2ª imputación										
Prediction	Reference										Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	122	0	0	2	0	0	0	0	0	2	1	121	0	0	2	0	0	0	0	0	2
2	1	283	0	1	0	0	0	0	2	10	2	1	283	0	1	0	0	0	0	2	10
3	0	1	93	1	0	0	0	1	0	0	3	0	1	93	0	0	0	0	1	0	1
4	1	0	0	77	0	0	0	0	0	3	4	1	1	0	78	0	0	0	0	0	3
5	0	0	0	0	85	0	0	0	0	0	5	1	0	0	0	85	0	0	0	0	0
6	0	0	0	0	0	55	0	0	0	0	6	0	0	0	0	0	55	0	0	0	0
7	1	0	0	0	0	0	28	0	1	0	7	1	0	0	0	0	0	29	0	1	1
8	0	0	0	0	0	0	0	8	0	0	8	0	0	0	0	0	0	0	8	0	0
9	0	0	1	1	0	2	0	0	63	7	9	0	0	1	0	0	2	0	0	62	5
10	5	14	4	0	0	5	2	2	6	99	10	5	13	4	1	0	5	1	2	7	99

Matriz de confusión Simpatía partidista 3ª imputación										
Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10
1	123	0	0	2	0	0	0	0	0	3
2	1	283	0	1	0	0	0	0	2	10
3	0	1	92	0	0	0	0	1	0	0
4	1	0	0	78	0	0	0	0	0	2
5	0	0	0	0	85	0	0	0	0	0
6	0	0	0	0	0	57	0	0	0	0
7	0	0	0	0	0	0	28	0	1	0
8	0	0	0	0	0	0	0	8	0	0
9	0	0	1	0	0	0	0	0	61	7
10	5	14	5	1	0	5	2	2	8	99

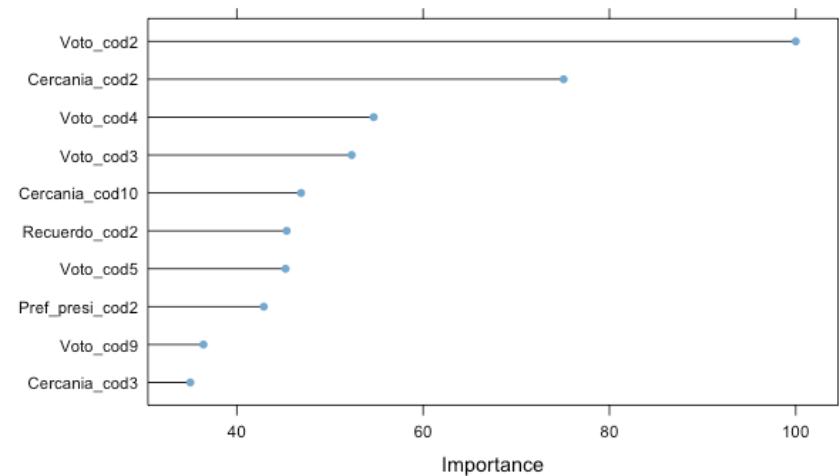
Fuente: Elaboración propia con RStudio

Al igual que con la variable anterior, se sigue teniendo valores altos en la diagonal, pero en esta ocasión vemos que el modelo falla en apenas muy pocas observaciones, en la clase que más difieren los resultados en el caso de la categoría 10 que indica que el individuo no siente simpatía por ningún partido. El modelo consigue una precisión del 92.3%, lo

cual es un resultado muy bueno y por otro lado el error *out-of-bag* es aproximadamente un 15%, por lo que los resultados han mejorado con respecto al modelo obtenido para la situación laboral.

Si miramos las variables más importantes para construir el modelo ilustradas en la **figura 5.2** vemos que en este caso cobran mayor importancia las variables relacionadas con la *intención de voto* y la *cercanía partidista*, muy similar a la idea de la simpatía. En ambos casos es la clase 2 la que tiene un mayor peso, en este caso pertenece al PSOE. También influyen el *recuerdo de voto* e incluso la *preferencia de presidente del gobierno*.

Figura 5.2 Importancia de variables en el modelo la simpatía partidista



Fuente: Elaboración propia con RStudio

En esta ocasión las tres imputaciones coinciden en un *mtry* de 10 alcanzando una *accuracy* del 0.84 en los tres casos, una vez más, superior al anterior modelo.

5.3 Imputación de la auto-ubicación ideológica

Una de las variables que guarda mayor relación con la intención de voto es la propia auto-ubicación ideológica de los encuestados, ya que esto les hará decantarse por un partido u otro. Dicha variable cuenta con un número medio-alto de valores faltantes, concretamente 413 respuestas “NS/NC”.

Tabla 5.3 Matrices de confusión de las imputaciones en la simpatía partidista

Matriz de confusión Ideología 1ª imputación											Matriz de confusión Ideología 2ª imputación										
Prediction	Reference										Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	40	4	14	8	4	1	0	1	2	1	1	35	5	10	8	6	1	0	1	2	1
2	2	1	1	0	0	0	0	0	0	0	2	2	1	1	0	0	0	0	0	0	0
3	38	30	74	34	25	6	2	1	0	0	3	35	28	77	38	23	7	2	1	0	0
4	6	3	27	32	15	3	3	0	0	0	4	10	5	29	27	15	3	2	0	0	0
5	28	12	34	57	193	59	46	18	4	10	5	32	11	32	58	191	60	48	18	4	8
6	0	0	0	0	1	7	6	2	0	1	6	0	0	0	0	1	7	5	2	0	0
7	0	0	0	0	2	9	9	8	4	1	7	0	0	0	0	3	8	10	8	2	1
8	0	0	0	0	1	0	1	0	0	0	8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	1	2	2	5	2	13	10	0	0	1	0	2	1	2	5	4	16

Matriz de confusión Ideología 3ª imputación

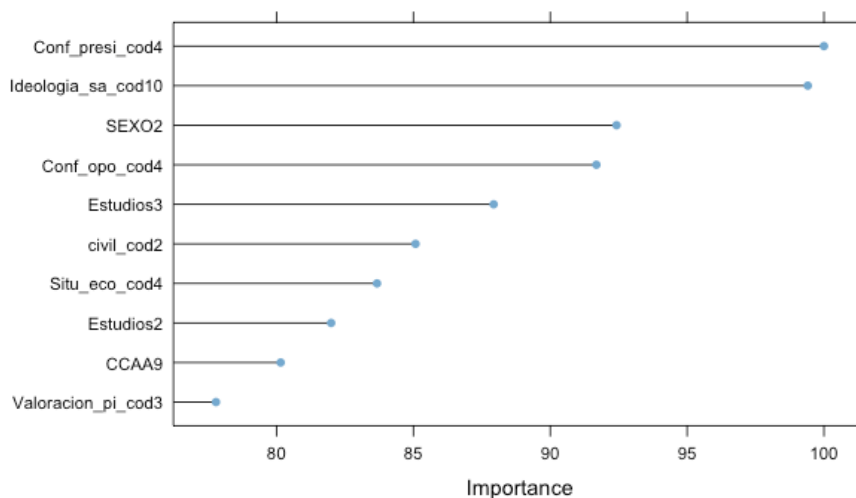
Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10
1	32	3	11	6	4	1	0	1	0	1
2	1	0	1	0	0	0	0	0	0	0
3	41	33	75	30	23	8	2	1	0	0
4	7	4	29	32	16	4	3	0	0	0
5	33	10	34	63	197	64	59	26	7	12
6	0	0	0	0	0	4	2	1	0	0
7	0	0	0	0	2	5	3	5	2	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0	1	3	13

Fuente: Elaboración propia con RStudio

Como se ve en la **tabla 5.3** el modelo no presenta valores altos en la diagonal, es más difiere en sobre todo en las clases de la 3 a la 5 que corresponderían a una ideología de centro o centro-izquierda. La precisión media de las imputaciones del modelo es de 0.39, corrobora con la matriz de confusión la baja precisión en la predicción de esta variable. De igual forma, el error OOB es superior al 50% en las tres imputaciones. Por otra parte, aunque la precisión entendida como predicción de la categoría exacta es baja, puede ocurrir que se dé una predicción con categorías contiguas. También se observa que hay categorías que están menos representadas, en particular, las correspondientes a la derecha más extrema, 9 a 10.

En cuanto a la importancia de las variables en la predicción de la *auto-ubicación ideológica*, vemos que la *confianza en el presidente* y en la *confianza en la oposición* cobren gran importancia. Por otro lado, también la ubicación ideológica del líder de VOX en la clase 10, que se correspondía con la extrema derecha, atributos que pueden ayudar a auto-clasificar la ideología del encuestado con una importancia superior al 90%.

Figura 5.3 Importancia de variables en el modelo de la auto-ubicación ideológica



Fuente: Elaboración propia con RStudio

El *mtry* elegido por el modelo predictivo de la auto-ubicación ideológica difiere en las tres imputaciones, es de 7, 8, y 5 respectivamente para cada una con la *accuracy* más baja hasta el momento, un 0.38.

5.4 Imputación de la intención de voto

La intención de voto es la variable más relevante en el estudio y cuenta con un elevado número de valores faltantes, se han registrado 824 respuestas “NS/NC”.

Como se ve en la **tabla 5.4** los tres modelos predicen bastante bien. La diagonal contiene valores altos, señal de una alta precisión e incluso en los partidos nacionalistas que contaban con un menor número de observaciones parecen estar correctamente clasificadas en los tres casos. El modelo parece diferir muy levemente en las clases 9 y 10 que correspondían a la abstención y a otros partidos. Señal de esta buena precisión es la precisión media de las tres imputaciones que alcanza un 93.6%.

Tabla 5.4 Matrices de confusión de las imputaciones en la intención de voto

Matriz de confusión Intención de Voto 1ª imputación											Matriz de confusión Intención de Voto 2ª imputación										
Prediction	Reference										Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	118	0	0	0	0	0	0	0	0	0	1	119	0	0	0	0	0	0	0	0	1
2	1	259	0	0	0	0	0	0	10	10	2	1	259	0	0	0	0	0	10	12	
3	0	0	87	0	0	0	0	1	0	9	3	0	0	87	0	0	0	0	1	0	9
4	0	0	0	73	0	0	0	0	4	3	4	0	0	0	73	0	0	0	0	4	3
5	0	0	0	0	84	0	0	0	1	0	5	0	0	0	0	84	0	0	0	1	0
6	0	0	0	0	0	54	0	0	2	2	6	0	0	0	0	0	54	0	0	2	4
7	0	0	0	0	0	0	27	0	1	2	7	0	0	0	0	0	0	27	0	1	2
8	0	0	0	0	0	0	0	7	0	0	8	0	0	0	0	0	0	0	7	0	0
9	1	1	0	0	0	0	0	0	108	2	9	0	0	0	0	0	0	0	0	108	2
10	6	1	0	1	0	0	0	1	2	120	10	6	2	0	1	0	0	0	1	2	115

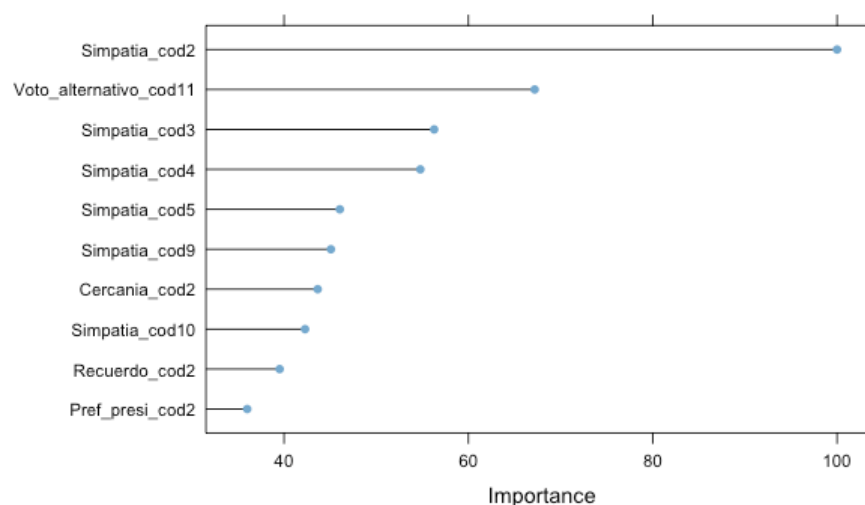
Matriz de confusión Intención de Voto 3ª imputación										
Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10
1	119	0	0	0	0	0	0	0	0	0
2	1	259	0	0	0	0	0	0	10	12
3	0	0	87	0	0	0	0	1	0	9
4	0	0	0	73	0	0	0	0	4	3
5	0	0	0	0	84	0	0	0	1	0
6	0	0	0	0	0	54	0	0	2	5
7	0	0	0	0	0	0	27	0	1	2
8	0	0	0	0	0	0	0	7	0	0
9	0	0	0	0	0	0	0	0	108	2
10	6	2	0	1	0	0	0	1	2	115

Fuente: Elaboración propia con RStudio

La **figura 5.4** muestra la importancia de variables en el modelo, y como vemos *Simpatía partidista* ocupa el primer puesto en la clase 2, que corresponde al PSOE, pero también encontramos el resto de las clases con una menor importancia. *Simpatía partidista* era una de las variables que de acuerdo con la medida Tau de Goodman and Kruskal guardaba gran relación con la *intención de voto*. También vemos el *voto alternativo* en la clase 11, que corresponde a no votar a ningún otro partido, es decir, no cambiar su intención de

voto tiene alta importancia, superior al 50% y así como la *cercanía partidista* y la *preferencia de presidente del gobierno*.

Figura 5.4 Importancia de variables en el modelo de la intención de voto



Fuente: Elaboración propia con RStudio

El *mtry* elegido por validación cruzada en las tres imputaciones es 10 con una *accuracy* de aproximadamente 0.93, hasta ahora la más alta de los modelos estimados. El OOB es relativamente bajo, tan solo un 7% aproximadamente, lo cual es muy bueno.

5.5 Precisión de los modelos

Una vez implementado el algoritmo en todas las variables, podemos analizar la precisión de los bosques aleatorios en cada una de ellas mediante la *accuracy* medio de las tres imputaciones. La **tabla 5.5** muestra una medida la *accuracy* para todas las variables que contenían valores faltantes, así como su error estándar.

Como se puede ver los modelos cuentan en su mayoría con una precisión relativamente alta superior al 50% de precisión. No obstante, hay algunos modelos que cuentan con una precisión inferior, como es el caso de la *ideología de los líderes políticos* y la *auto-ubicación ideológica*, esto puede deberse a varios motivos: (i) son variables con muchas categorías y, en algunos casos, con desequilibrios en las categorías, y (ii) que la selección inicial de las variables se hizo enfocada a la intención de voto y puede ser que no haya variables relevantes o suficientemente explicativas de esas variables. En cuanto a los errores estándar, son inferiores a un 1% lo que apunta a una gran estabilidad en los modelos estimados.

Tabla 5.5 Precisión y error estándar de los modelos generados con Random Forest

Situación laboral	0.764 (0.001)	Valoración Santiago Abascal	0.77 (0.002)
Nivel de Estudios	0.614 (0.003)	Valoración Inés Arrimadas	0.591 (0.003)
Estado civil	0.685 (0.001)	Cercanía partidista	0.762 (0.002)
Participación anteriores elecciones	0.988 (0.001)	Preferencia presidente del Gobierno	0.69 (0.001)
Religión	0.656 (0.002)	Ideología	0.396 (0.004)
Valoración Situación económica	0.582 (0.005)	Ideología Pablo Casado	0.401 (0.004)
Confianza en el actual presidente	0.725 (0.002)	Recuerdo de voto	0.802 (0.002)
Clase social	0.771 (0.001)	Ideología Pedro Sánchez	0.379 (0.004)
Confianza en el presidente de la oposición	0.702 (0.002)	Ideología Santiago Abascal	0.733 (0.001)
Valoración Pedro Sánchez	0.644 (0.003)	Ideología Pablo Iglesias	0.525 (0.005)
Valoración Pablo Iglesias	0.626 (0.001)	Ideología Inés Arrimadas	0.421 (0.003)
Simpatía partidista	0.923 (0)	Intención de voto	0.936 (0.001)
Valoración Pablo Casado	0.617 (0.004)	Voto alternativo	0.563 (0.002)

Fuente: Elaboración propia con RStudio

6. INTERPRETACIÓN DEL ALGORITMO

Los bosques aleatorios se caracterizan por incluir una gran cantidad de árboles de decisión individuales y esto hace que la interpretación del modelo de forma individual se complique. La importancia de variables que entran en el modelo nos hace hacernos una idea de cómo se están construyendo los árboles y qué variables se están teniendo en cuenta. No obstante, no podemos graficar el propio algoritmo en sí porque no está formado por un único árbol.

El que *Random Forest* sea un algoritmo formado por múltiples árboles de clasificación, nos hace pensar que mirar a un árbol individual nos puede ayudar a tener una idea conjunta del algoritmo. Por ello, a continuación, graficaremos un árbol de clasificación de tres variables con diferente número de valores faltantes para tratar de comprender el funcionamiento de *Random Forest*.

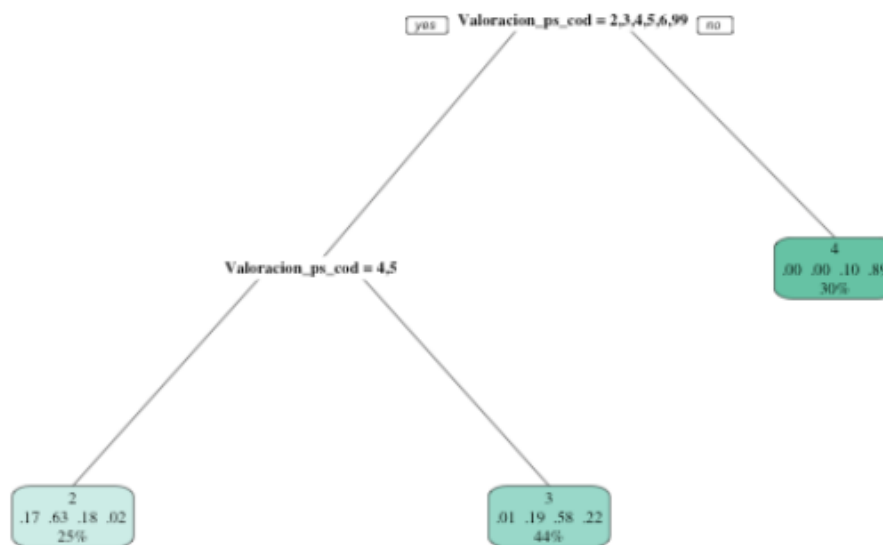
Para construir los árboles hemos utilizado la misma muestra de entrenamiento utilizada para *Random Forest*, y en este caso elegiremos tres variables con un número bajo, medio y alto de valores faltantes. En cuanto a los parámetros de los árboles de clasificación, fijaremos su parámetro principal, *minsplit*, entendido como el número más pequeño de observaciones para poder dividir un nodo, fijaremos este mínimo como el parámetro *mtry* que estimamos en *Random Forest*.

6.1 Árbol de decisión para el grado de confianza en el presidente del gobierno

El grado de confianza en el actual presidente del gobierno, Pedro Sánchez, cuenta con cinco clases que van desde mucha confianza hasta ninguna, además la variable cuenta con 173 valores faltantes. Se ha fijado el valor de *minsplit* en nueve.

La **figura 6.1** muestra el árbol de clasificación para el grado de confianza en el presidente del gobierno actual. En este caso se ha tomado la variable valoración de Pedro Sánchez como nodo raíz donde nacen dos ramas en función de si la valoración comprende entre las clases 2 y 6, es decir, si la valoración se encuentra entre mal y muy bien o por defecto el encuestado no conoce quién es Pedro Sánchez o no tiene clara su valoración. Si no se cumple ninguno de estos casos, es decir, si la valoración hacia Pedro Sánchez es un 1, “Muy mal”, directamente el árbol clasifica la observación en la clase 4, “Ninguna confianza”. Por el contrario, encontramos un nuevo nodo, aún con la misma variable, en este caso, si la valoración de Pedro Sánchez pertenece a la clase 4 o 5, es decir, “Bien” o “Muy bien”, el algoritmo clasifica la observación en la clase 2, “Bastante confianza”, si por el contrario, la valoración es 2,3,6 o 99 es decir, “mal”, “regular”, “no conoce a Pedro Sánchez” o “NS/NC”, el algoritmo lo clasifica en la clase 3, “Poca confianza”.

Figura 6.1 Árbol de clasificación para el grado de confianza en el presidente del gobierno



Fuente: Elaboración propia con RStudio

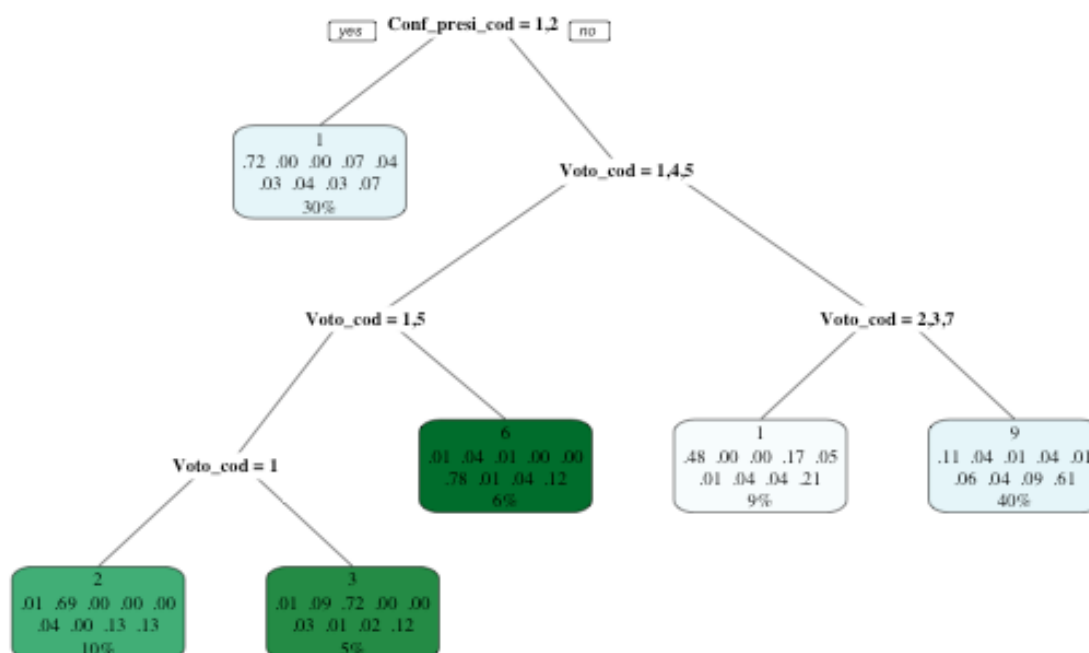
6.2 Árbol de decisión para la preferencia como presidente del gobierno

La siguiente variable por analizar es la *preferencia de presidente del gobierno*, la cual cuenta con 384 valores faltantes. Al igual que el árbol de clasificación anterior se ha fijado nueve como valor de *minsplit*.

La **figura 6.2** muestra el árbol de clasificación para predecir la *preferencia como presidente de gobierno*. En este caso se ha tomado la variable *grado de confianza en el actual presidente*, Pedro Sánchez, como nodo raíz donde nacen dos ramas en función de si la valoración es 1 o 2, es decir, si la valoración es “mucha confianza” o “bastante confianza”. En el caso de pertenecer a alguna de estas clases, el árbol clasifica la observación en la clase 1, “Pedro Sánchez” como preferencia de presidente de gobierno. En caso contrario, se ha elegido la variable *intención de voto* en el siguiente nodo. En este caso se dividen las ramas en función de si votaría a PP, Ciudadanos o VOX. Si no se vota a ninguno de estos partidos, se vuelve a dividir el árbol en función del voto nuevamente, donde si se vota a PSOE, Podemos o JxCat, el árbol clasificaría la observación como la clase 1 nuevamente, “Pedro Sánchez”, en el caso contrario lo clasificaría en la clase 9, “Ninguno”.

Por otro lado, si la intención de voto inicial es PP, Ciudadanos o VOX, vuelve a dividirse nuevamente en intención de voto, diferenciando entre si votaría a PP y VOX. En caso de no votar a ninguno de estos partidos, votaría a Ciudadanos y el árbol lo clasifica en la clase 6, “Inés Arrimadas” como preferencia de presidente del gobierno. Si votaría a PP o VOX, se vuelve a dividir en función de si votaría al PP, de forma que en caso afirmativo la observación se clasificaría en la clase 2, “Pablo Casado” y en caso contrario, es decir, votante de VOX, en la clase 3 “Santiago Abascal”.

Figura 6.2 Árbol de clasificación para la preferencia como presidente del gobierno



Fuente: Elaboración propia con RStudio

6.3 Árbol de decisión para la intención de voto

La intención de voto es la variable decisiva y para predecirla se ha utilizado un *minsplit* con valor igual a nueve. Una vez fijado se ha estimado un árbol de clasificación ilustrado en la **figura 6.3**.

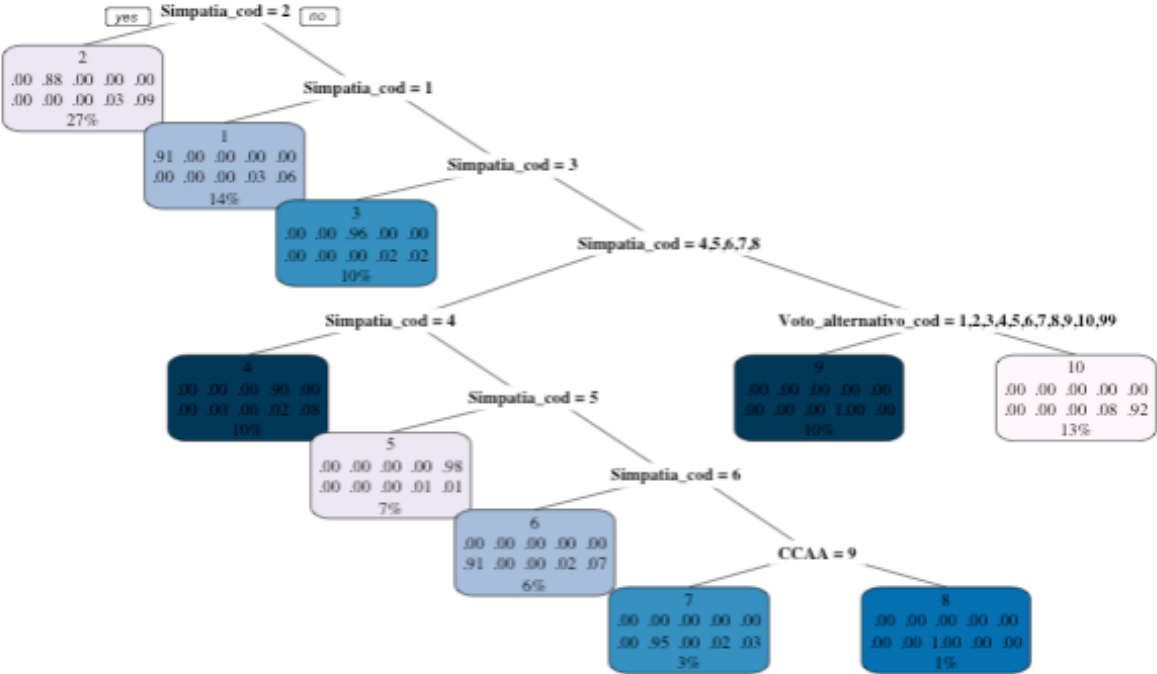
Como nodo raíz se ha tomado la variable *simpatía partidista*, en el caso de que la simpatía sea 2, es decir, el encuestado sienta simpatía por el PSOE, entonces, su intención de voto será PSOE. En el caso opuesto, encontramos un nuevo nodo cuya división se realiza en función de si se siente simpatía por el PP. Una vez más, en caso afirmativo, la intención de voto sería PP. Lo mismo ocurre con el siguiente nodo, pero esta vez con Podemos, si el encuestado siente simpatía por Podemos, entonces su intención de voto será Podemos. El siguiente nodo presenta una subdivisión en función de si la simpatía es hacia Ciudadanos, Vox, JxCat, PNV y ERC.

En caso negativo, se dividiría según el voto alternativo, en el sentido de que, si su voto alternativo pertenece a la clase 11, es decir, no cambiaría su voto, su intención de voto sería “Abstención”, pero si su voto alternativo es cualquier partido, entonces el algoritmo lo clasifica como “Otros”.

Volviendo al nodo anterior, en caso afirmativo, se subdivide en función de simpatía por Ciudadanos, donde la intención de voto sería Ciudadanos, simpatía por VOX, donde

entonces votaría a VOX y simpatía por ERC, donde votaría a ERC. En caso negativo de todos los partidos anteriores, se subdivide en función de si el encuestado es catalán, donde en caso positivo, votaría a JxCat, y por el contrario votaría a EAJ-PNV.

Figura 6.3 Árbol de clasificación para la intención de voto



Fuente: Elaboración propia con RStudio

7. PREDICCIÓN FINAL DE LA INTENCIÓN DE VOTO

Tras realizar la imputación en cada una de las variables con valores faltantes, tenemos tres conjuntos de datos completos y por consiguiente tres estimaciones diferentes para la variable *intención de voto*. Ante este escenario, se calculan las estimaciones de voto para cada uno de los partidos en los tres conjuntos de datos, así como el error estándar combinando de acuerdo con las fórmulas presentadas por Marín y Restrepo, 2012 expuestas en el capítulo cinco.

Los resultados se presentan en la **tabla 7.1** y cabe destacar que las estimaciones se han realizado sobre el voto válido, es decir, en el cálculo se ha excluido la abstención y hemos tenido en cuenta los coeficientes de ponderación que permiten el tratamiento conjunto de la muestra y que tienen en cuenta que el diseño maestral usa diferentes tasas de muestreo. Además, se ha incluido el error estándar combinado entre paréntesis.

A la vista de los resultados, el PSOE es el partido con un mayor porcentaje de votos, un 28.4%, seguido del PP con un 20.9% y Ciudadanos con un 10.5%. Por debajo del 10% se encuentran Unidas Podemos con un 9.8% y VOX con un 9.2%, mientras que dentro de los partidos nacionalistas ERC tiene un mayor porcentaje de votantes, concretamente un 2.4%, seguido de JxCat con un 1.2% y en último lugar PNV con un 1.2% de los votos. Por otro lado, el porcentaje de encuestados que votarían a otros partidos es alto, con un 16.5% que se colocaría en tercera posición entre PP y Ciudadanos.

Tabla 7.1 Estimación de la intención de voto en % sobre el voto válido

*Estimación de voto en % sobre voto
válido*

PSOE	28.4 (0.009)
PP	20.9 (0.007)
Unidas Podemos	9.8 (0.005)
Ciudadanos	10.5 (0.006)
VOX	9.2 (0.005)
ERC	2.4 (0.003)
JxCat	1.2 (0.002)
EAJ-PNV	1.2 (0.002)
Otros	16.5 (0.01)

Fuente: Elaboración propia con RStudio

En cuanto al error estándar combinado calculado por medio de la varianza dentro de la imputación y la varianza entre imputaciones utilizando la expresión (11), vemos que en todos los partidos políticos es inferior al 1% lo cual dota a las estimaciones de buena precisión.

7.1 Comparación con los resultados del CIS

Con objeto validar la exactitud de los resultados, se presentan los resultados del barómetro del CIS en diciembre de 2020 en la **tabla 7.2**. Vemos que los resultados difieren levemente: en ambos casos, el partido ganador es el PSOE, seguido del PP, pero en el caso del tercer puesto, en los resultados del CIS lo ocuparía VOX con un 14% de votantes, mientras que, en la estimación realizada, VOX ocupa el quinto puesto con un 9.2% de votantes. Las estimaciones dadas para Ciudadanos y Unidas Podemos son similares, salvo que, en el caso de las estimaciones del CIS, Unidas Podemos está por delante de Ciudadanos, mientras que en nuestra estimación es al contrario. En cuanto a los partidos nacionalistas, ERC tiene una intención de voto superior a JxCat y PNV en ambos casos y las estimaciones son similares no distan mucho de la estimación del CIS.

Por otro lado, “Otros” alcanza un 8% en el caso del CIS sin contar el voto en blanco. No obstante, para que los resultados sean comparables, se añade esta categoría a “Otros” y por tanto el porcentaje de votantes de otros partidos sería un 10% frente a la estimación proporcionada por el modelo, un 16.5%.

Tabla 7.2 Comparación de las predicciones con los resultados del CIS

	Voto directo en la encuesta (en % sobre censo)	Estimación de voto CIS (en % sobre el voto válido)	Agrupación estimación de voto CIS	Estimación de voto Imputación múltiple
PSOE	19.4	29.5	29.5	28.4
PP	11.8	19.2	19.2	20.9
VOX	7.0	14.0	14.0	9.2
Unidas Podemos	7.1	10.8	10.8	9.8
Ciudadanos	7.5	10.5	10.5	10.5
ERC	1.6	2.9	2.9	2.4
JxCat	0.9	1.7	1.7	1.2
EAJ-PNV	0.9	1.4	1.4	1.2
Otros	0.9	1.2	10.0	16.5
EH Bildu	1.1	1.2		
Más	1.6	2.4		
País/Compromís				
CUP	0.4	0.8		
CCa-NC	0.2	0.5		
NA+	0.2	0.4		
BNG	0.5	0.5		
PRC	0.1	0.1		
PACMA	0.7	0.9		
En blanco	4.5	2.0		
Voto nulo	1.3			
Abstención	11.5			
NS	16.8			
NC	4.0			

Fuente: Centro de Investigaciones Sociológicas (2020) y elaboración propia con RStudio

8. CONCLUSIÓN

Las encuestas electorales son determinantes para analizar el clima político actual y aportar notoriedad a los diferentes partidos políticos. No obstante, presentan un problema difícil de abordar: el tratamiento de los valores faltantes. Estimar la intención de voto de los encuestados que no responden a su intención de voto es de vital importancia y para ello, el CIS ha tratado de mejorar las estimaciones proporcionadas mediante la imputación de dichos valores faltantes durante años.

En el presente trabajo se ha tratado de abordar esta cuestión imputando los valores faltantes por medio de una imputación múltiple basada en el algoritmo *Random Forest*. A través de dicho algoritmo se han realizado predicciones de aquellas variables con presencia de valores faltantes en tres estimaciones sucesivas formando así tres conjuntos de datos completos y obteniendo de esta forma la estimación de la intención de voto.

Los modelos estimados están dotados de gran exactitud, salvo en el caso de aquellas variables donde las clases están desbalanceadas donde, por consiguiente, el modelo tiende a fallar. A pesar de la gran flexibilidad de *Random Forest*, el elevado número de árboles utilizado por el algoritmo hace que se dificulte la interpretación de los modelos, por lo que en este aspecto un único árbol de decisión aporta una mayor claridad. Implementado el árbol de clasificación, en el caso de la *intención de voto*: la *simpatía partidista*, el *voto alternativo* y si el encuestado procede de Cataluña, han resultado variables determinantes en la predicción de la variable. Inclusive, el modelo proporcionado por Random Forest para la intención de voto es uno de los modelos con una mayor precisión de entre los calculados.

Es precisamente esta exactitud, lo que hace que el método de la imputación múltiple mediante este tipo de algoritmos sea útil en la imputación de las encuestas electorales. La comparación de los resultados con el barómetro real del CIS lo cercioran dado que las estimaciones proporcionadas han respetado el mismo ranking de resultados del CIS, salvo en el caso de VOX, Unidas Podemos y Ciudadanos donde el modelo ha presentado estimaciones de voto ligeramente diferentes. No obstante, a pesar de ello, las predicciones proporcionadas por el modelo son notablemente precisas y sería interesante abordar como futuro estudio aumentar la exactitud de los resultados mediante procedimientos para variables que presenten clases desbalanceadas.

BIBLIOGRAFÍA

Antena 3 Noticias (31 de diciembre, 2020). El año político 2020: un Gobierno de coalición en estado de alarma. *Antena 3 Noticias*. https://www.antena3.com/noticias/espana/el-ano-politico-2020-un-gobierno-de-coalicion-en-estado-de-alarma_202012315fedf6ae3005ce0001fc8a8a.html.

Bocanegra, J. (22 de diciembre, 2020). Los ciudadanos se cansan de la crispación y reclaman más acuerdos a sus políticos. *EL CONFIDENCIAL*. https://www.elconfidencial.com/espana/2020-12-22/crispacion-politica-problemas-poblacion-cis_2881704/.

Breiman, Leo. (2001) Random Forest. *Machine Learning*, 45, 5–32.

Calleja, M (19 de noviembre, 2020). No confíen en el CIS, según Tezanos. *ABC*. https://www.abc.es/espana/abci-no-confien-segun-tezanos-202011182031_noticia.html

CIS (2021). Barómetros. *Centro de Investigaciones Sociológicas*. http://www.cis.es/cis/opencm/ES/11_barometros/index.jsp

CIS (2020). *BARÓMETRO DE DICIEMBRE FICHA TÉCNICA*. (Estudio CIS Nº 3303). http://www.cis.es/cis/export/sites/default/-Archivos/Marginales/3300_3319/3303/Ft3303.pdf

CIS (2021). Metodología. *Centro de Investigaciones Sociológicas*. http://www.cis.es/cis/export/sites/default/-Archivos/NotasdeInvestigacion/NI004_MetodologiaBarometros_Informe.pdf.

Escobar, M., Rivière, J. y Cilleros, R (2014). *Los pronósticos electorales con encuestas*. Centro de Investigaciones Sociológicas.

Escobar, M. y Jaime, A.M. (2013). Métodos de imputación múltiple para predecir resultados electorales. En: Mendoza, A. (Ed.), *Aplicaciones en Economía y Ciencias Sociales con Stata*. College Station: Stata.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, London.

Gómez, M., Alfaro, E., Alfaro, J., y García, N. (2009). Árboles de clasificación para el análisis de gráficos de control multivariantes. *Revista De Matemática Teoría Y Aplicaciones*, 16(1): 30-42.

Hastie, T., Tibshirani, R. y Friedman, J (2009). *The Elements of Statistical Learning*. Springer.

Horton, N. J. y Lipsitz, S. R. (2001). “Multiple imputation in practice: Comparison of software packages for regression models with missing variables”, *The American Statistician*, 55(3):244-254.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Marín, J.M. y Restrepo, M.I (2012). Imputación de ingresos en la Gran Encuesta Integrada de Hogares (GEIH) de 2010. *Revista Desarrollo y Sociedad*. 70(70):219-243.

Orriols, L. (2021). *Webinar Alumni UC3M 28-04-2021– CAMPAÑAS, ENCUESTAS Y PROCESOS ELECTORALES – Lluís Orriols*. [Video]. YouTube. <https://www.youtube.com/watch?v=1teaLZPe650>.

Pearson, R. (2020). The GoodmanKruskal package: Measuring association between categorical variables. <https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html>.

Rivas, C., Martínez, M.M. y Galindo, P (2010). “La Imputación Múltiple como Alternativa al Análisis de la No Respuesta en la Variable Intención de Voto”. *Revista Española de la Ciencia Política*, 22(22):99-118.

Rubin, D. B. (1996). “Multiple imputation after 18 + years”, *Journal of the American Statistical Association*, 91(434):473-489.

Salvador Figueras, M. y Gargallo, P. (2003). Análisis Exploratorio de Datos. *5campus, Estadística*. <http://www.5campus.com/leccion/aed>.

Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3-15.

ANEXO A

Tabla A.1

P0a: Sexo

Valores iniciales	
1	Hombre
2	Mujer

Fuente: Elaboración propia con RStudio

Tabla A.2

P0b: Edad

Recodificación	
1	De 18 a 24 años
2	De 25 a 34 años
3	De 35 a 44 años
4	De 45 a 54 años
5	De 55 a 64 años
6	Más de 65 años

Fuente: Elaboración propia con RStudio

Tabla A.3

P0c: Nacionalidad del entrevistado

Valores iniciales	
1	Nacionalidad española
2	Nacionalidad española y otra

Fuente: Elaboración propia con RStudio

Tabla A.4

P-: Comunidad autónoma del entrevistado

Valores iniciales	
1	Andalucía
2	Aragón
3	Asturias (Principado de)
4	Baleares (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 Floral de)
16	País Vasco
17	Rioja (La)
18	Ceuta (Ciudad Autónoma de)
19	Melilla (Ciudad Autónoma de)

Fuente: Elaboración propia con RStudio

Tabla A.5*P28a: Nivel de estudios alcanzado por la persona entrevistada*

Valores iniciales		Recodificación	
1	Menos de 5 años de escolarización	1	Sin estudios / Estudios Primarios ¹
2	Primaria (enseñanza primaria o 5º curso de EGB)	2	Estudios Secundarios / Formación Profesional ²
3	Formación Profesional inicial	3	Estudios Superiores ³
4	Secundaria (ESO completa, EGB completa, Bachillerato Elemental)	4	Otros estudios
5	Formación Profesional de Grado Medio (FP I)	99	NS/NC
6	Bachillerato/COU/PREU (Bachillerato superior, BUP)		
7	Formación Profesional de grado superior (FP II)		
8	Arquitectura / Ingeniería Técnica (aparejador/a; peritos/as)		
9	Diplomatura oficial		
10	Estudios de Grado		
11	Licenciatura		
12	Arquitectura/Ingeniería		
13	Máster oficial universitario		
14	Doctorado		
15	Títulos propios de posgrado (máster no oficial, etc.)		
16	Otros estudios		
98	N.S/No recuerda		
99	N.C		

¹ NA, Menos de 5 años de escolarización y Primaria (enseñanza primaria o hasta 5º de EGB)² Formación Profesional inicial, Secundaria (Bachillerato Elemental, EGB completa, ESO completa), Formación Profesional de grado medio (FP I), Bachillerato/ COU/ PREU (Bachillerato superior, BUP, bachillerato) y Formación Profesional de grado superior (FP II)³ Universitarios medios (diplomatura, arquitectura o ingeniería técnica) y Universitarios superiores (licenciatura, grado, máster oficial, doctorado, títulos propios de postgrado no oficiales)

Fuente: Elaboración propia con RStudio

Tabla A.6*P31: Situación laboral de la persona entrevistada*

Valores iniciales		Recodificación	
1	Trabaja	1	Trabajadores
2	Jubilado/a o pensionista (trabajó con anterioridad)	2	Jubilados o pensionistas ¹
3	Pensionista (no trabajó con anterioridad)	3	Parados ²
4	En paro y ha trabajado con anterioridad	4	Estudiantes
5	En paro y en busca de su primer empleo	5	Trabajo doméstico no remunerado
6	Estudiante	6	Otra situación
7	Trabajo doméstico no remunerado	99	NC
8	Otra situación		
9	NC		

¹ Jubilado/a o pensionista (trabajó con anterioridad), Pensionista (no trabajó con anterioridad)² En paro y ha trabajado con anterioridad, En paro y en busca de su primer empleo

Fuente: Elaboración propia con RStudio

Tabla A.7*P29: Religiosidad de la persona entrevistada*

Valores iniciales		Recodificación	
1	Católico/a practicante	1	Católico ¹
2	Católico/a no practicante	2	Creyente de otra religión
3	Creyente de otra religión	3	No creyente ²
4	Agnóstico/a	4	Ateo/a
5	Indiferente, no creyente	99	NC
6	Ateo/a		
9	NC		

¹ Católico/a practicante, Católico/a no practicante

² Agnóstico/a, Indiferente, no creyente

Fuente: Elaboración propia con RStudio

Tabla A.8*P30: Estado civil de la persona entrevistada*

Valores iniciales		Recodificación	
1	Casado/a	1	Casado/a
2	Soltero/a	2	Soltero/a ¹
3	Viudo/a	3	Viudo/a
4	Separado/a	99	NC
5	Divorciado/a		
9	NC		

¹ Soltero/a, Divorciado/a, Separado/a

Fuente: Elaboración propia con RStudio

Tabla A.9*P32: Clase social subjetiva de la persona entrevistada*

Valores iniciales		Recodificación	
1	Clase alta	1	Clase alta
2	Clase media-alta	2	Clase media ¹
3	Clase media-media	3	Clase baja ²
4	Clase media-baja	4	Otra clase
5	Clase trabajadora/obrero	5	No cree en las clases
6	Clase pobre	99	NS/NC
8	Proletariado		
9	A los/as de abajo		
10	Excluidos/as		
11	A la gente común		
12	Clase baja		
96	Otras		
97	No cree en las clases		
98	No sabe, duda		
99	NC		

¹ Clase media-alta, Clase media-media, Clase media-baja

² Clase trabajadora/obrero, Clase pobre, Proletariado, A los/as de abajo, Excluidos/as, A la gente común, Clase baja

Fuente: Elaboración propia con RStudio

Tabla A.10*P21: Intención de voto en nuevas elecciones generales*

Valores iniciales		Recodificación	
1	PP	1	PP
2	PSOE	2	PSOE
3	Podemos	3	Podemos ¹
4	Ciudadanos	4	Ciudadanos
5	IU	5	VOX
6	En Comú Podem	6	ERC
7	Compromís	7	JxCat
8	ERC	8	EAJ-PNV
9	JxCat	9	Otros
11	EAJ-PNV	10	Abstención
12	EH Bildu	99	NS/NC
13	CC-PNC		
16	Nueva Canarias		
17	PACMA		
18	VOX		
19	CUP		
20	Los Verdes		
21	Unidas Podemos		
22	PAR		
24	BNG		
29	Partido Libertario		
30	CHA		
34	Geroa Bai		
39	PCPE		
43	PRC		
44	UPL		
50	Más País		
67	En Común-Unidas Podemos		
68	Teruel Existe		
71	PDeCAT		
77	Voto nulo		
95	Otro partido		
96	En blanco		
97	No votaría		
98	No sabe todavía		
99	N.C		

¹ Formado por la coalición de En Comú Podem, IU, EQUO, Unidas Podemos, En Común - Unidas Podemos

Fuente: Elaboración propia con RStudio

Tabla A.11*P21a: Voto alternativo en futuras elecciones generales*

Valores iniciales		Recodificación	
1	PP	1	PP
2	PSOE	2	PSOE
3	Podemos	3	Podemos ¹
4	Ciudadanos	4	Ciudadanos
5	IU	5	VOX
6	En Comú Podem	6	ERC
7	Compromís	7	JxCat
8	ERC	8	EAJ-PNV
9	JxCat	9	Otros
11	EAJ-PNV	10	No votaría a ningún otro partido
12	EH Bildu	11	Abstención
13	CC-PNC	99	NS/NC
16	Nueva Canarias		
17	PACMA		
18	VOX		
19	CUP		
20	Los Verdes		
21	Unidas Podemos		
22	EQUO		
24	BNG		
25	MÉS		
27	Escaños en Blanco		
29	Partido Libertario		
30	CHA		
38	UPyD		
39	PCPE		
43	PRC		
44	UPL		
50	Más País		
61	Democaracia Nacional		
67	En Común-Unidas Podemos		
68	Teruel Existe		
71	PDeCAT		
72	PNC (Partido Nacionalista Catalán)		
77	Voto nulo		
93	No votaría a ningún otro partido		
95	Otro partido		
96	En blanco		
97	No votaría		
98	N.S		
99	N.C		

¹ Formado por la coalición de En Comú Podem, IU, EQUO, Unidas Podemos, En Común - Unidas Podemos

Fuente: Elaboración propia con RStudio

Tabla A.12*P22: Partido político por el que siente más simpatía**P27: Partido que considera más cercano a sus ideas*

Valores iniciales		Recodificación	
1	PP	1	PP
2	PSOE	2	PSOE
3	Podemos	3	Podemos ¹
4	Ciudadanos	4	Ciudadanos
5	IU	5	VOX
6	En Comú Podem	6	ERC
7	Compromís	7	JxCat
8	ERC	8	EAJ-PNV
9	JxCat	9	Otros
11	EAJ-PNV	10	Ninguno
12	EH Bildu	99	NS/NC
13	CC-PNC		
14	UPN		
16	Nueva Canarias		
17	PACMA		
18	VOX		
19	CUP		
20	Los Verdes		
21	Unidas Podemos		
22	EQUO		
24	BNG		
26	Falange Española de las JONS		
29	Partido Libertario		
30	CHA		
34	Geroa Bai		
38	UPyD		
40	PI		
42	IGRE		
43	PRC		
44	UPL		
50	Más País		
67	En Común-Unidas Podemos		
68	Teruel Existe		
71	PDeCAT		
72	PNC (Partido Nacionalista Catalán)		
95	Otro partido		
97	Ninguno		
98	N.S		
99	N.C		

¹ Formado por la coalición de En Comú Podem, IU, EQUO, Unidas Podemos, En Común - Unidas Podemos

Fuente: Elaboración propia con RStudio

Tabla A.13

P26a: Recuerdo de voto en las elecciones generales (nov. 2019)

Valores iniciales		Recodificación	
1	PP	1	PP
2	PSOE	2	PSOE
3	Podemos	3	Podemos ¹
4	Ciudadanos	4	Ciudadanos
5	IU	5	VOX
6	En Comú Podem	6	ERC
7	Compromís	7	JxCat
8	ERC	8	EAJ-PNV
9	JxCat	9	Otros
11	EAJ-PNV	10	Abstención
12	EH Bildu	11	No fué a votar ²
13	CC-PNC	99	NS/NC
16	Nueva Canarias		
17	PACMA		
18	VOX		
19	CUP		
20	Los Verdes		
21	Unidas Podemos		
22	PAR		
24	BNG		
29	Partido Libertario		
30	CHA		
34	Geroa Bai		
39	PCPE		
43	PRC		
44	UPL		
50	Más País		
67	En Común-Unidas Podemos		
68	Teruel Existe		
71	PDeCAT		
77	Voto nulo		
95	Otro partido		
96	En blanco		
97	No votaría		
98	No sabe todavía		
99	N.C		

¹ Formado por la coalición de En Comú Podem, IU, EQUO, Unidas Podemos, En Común - Unidas Podemos

² Formado por NAs de la variable inicial

Fuente: Elaboración propia con RStudio

Tabla A.14

P13: Valoración de la situación económica general de España

Valores iniciales		Recodificación	
1	Muy buena	1	Muy buena
2	Buena	2	Buena
3	Regular	3	Regular
4	Mala	4	Mala
5	Muy mala	5	Muy mala
8	NS	99	NS/NC
9	NC		

Fuente: Elaboración propia con RStudio

Tabla A.15

P18: Grado de confianza en el actual presidente: Pedro Sánchez

P19: Grado de confianza en el principal líder de la oposición: Pablo Casado (PP)

Valores iniciales		Recodificación	
1	Mucha confianza	1	Mucha confianza
2	Bastante confianza	2	Bastante confianza
3	Poca confianza	3	Poca confianza
4	Ninguna confianza	4	Ninguna confianza
8	NS	99	NS/NC
9	NC		

Fuente: Elaboración propia con RStudio

Tabla A.16

P20: Preferencia como presidente del gobierno central

Valores iniciales		Recodificación	
1	Pedro Sánchez	1	Pedro Sánchez
2	Pablo Casado	2	Pablo Casado
3	Santiago Abascal	3	Santiago Abascal
4	Pablo Iglesias	4	Pablo Iglesias
5	Alberto Garzón	5	Alberto Garzón
6	Inés Arrimadas	6	Inés Arrimadas
7	Iñigo Errejón	7	Iñigo Errejón
96	Otro/a	8	Otro/a
97	Ninguno/a	9	Ninguno/a
98	NS	99	NS/NC
99	NC		

Fuente: Elaboración propia con RStudio

Tabla A.17*P23: Escala de autoubicación ideológica (1-10)*

Valores iniciales		Recodificación	
1	1 - Izquierda	1	1 - Izquierda
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10 - Derecha	10	10 - Derecha
98	NS	99	NS/NC
99	NC		

Fuente: Elaboración propia con RStudio

Tabla A.18

*P24: Escala de ubicación ideológica (1-10) de líderes políticos nacionales: *Pedro Sánchez * Pablo Casado * Pablo Iglesias * Santiago Abascal * Inés Arrimadas*

Valores iniciales		Recodificación	
1	1 - Izquierda	1	1 - Izquierda
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10 - Derecha	10	10 - Derecha
97	No conoce	11	No conoce
98	NS	99	NS/NC
99	NC		

Fuente: Elaboración propia con RStudio

Tabla A.19

*P25: Escala de valoración (1-10) de líderes políticos nacionales: *Pedro Sánchez * Pablo Casado * Pablo Iglesias * Santiago Abascal * Inés Arrimadas*

Valores iniciales		Recodificación	
1	1 - Muy mal	1	Muy mal ¹
2	2	2	Mal ²
3	3	3	Regular ³
4	4	4	Bien ²
5	5	5	Muy bien ³
6	6	6	No conoce
7	7	99	NS/NC
8	8		
9	9		
10	10 - Muy bien		
97	No conoce		
98	NS		
99	NC		

¹ Valoración entre 1 y 2
² Valoración entre 3 y 4
³ Valoración entre 5 y 6
⁴ Valoración entre 7 y 8
⁵ Valoración entre 9 y 10

Fuente: Elaboración propia con RStudio

Tabla A.20

P26: Participación electoral en las anteriores elecciones generales (nov. 2019)

Valores iniciales		Recodificación	
1	Fue a votar y votó	1	Votó ¹
2	No tenía edad para votar	2	No votó ²
3	Fue a votar pero no pudo hacerlo	3	No tenía derecho a voto ³
4	No fue a votar porque no pudo	99	NS/NC
5	Prefirió no votar		
6	No tenía derecho a voto		
7	Votó por correo		
8	No recuerda/NS		
9	NC		

¹ Fue a votar y votó; Votó por correo
² Fue a votar pero no pudo hacerlo; No fue a votar porque no pudo
³ No tenía edad para votar; No tenía derecho a voto

Fuente: Elaboración propia con RStudio

ANEXO B

Tabla B.1

Intervalos Bootstrap para la Tau entre 0.1 y 0.2 (Primeras 33 variables)

Variable 1	Variable 2	K-tau	Intervalo Bootstrap
Situ	civil_cod	0.123	(0.11 ; 0.135)
Estudios	Situ	0.113	(0.101 ; 0.127)
Voto_cod	Conf_presi_cod	0.191	(0.175 ; 0.205)
Simpatia_cod	Conf_presi_cod	0.191	(0.178 ; 0.201)
Ideologia_ps_cod	Conf_presi_cod	0.112	(0.09 ; 0.125)
Recuerdo_cod	Conf_presi_cod	0.145	(0.129 ; 0.158)
Cercania_cod	Conf_presi_cod	0.161	(0.144 ; 0.175)
Ideologia_pc_cod	Conf_opo_cod	0.107	(0.096 ; 0.114)
Conf_presi_cod	Pref_presi_cod	0.136	(0.123 ; 0.149)
Valoracion_ps_cod	Pref_presi_cod	0.136	(0.124 ; 0.149)
Recuerdo_cod	Pref_presi_cod	0.129	(0.109 ; 0.144)
Cercania_cod	Pref_presi_cod	0.182	(0.166 ; 0.194)
Simpatia_cod	Voto_alternativo_cod	0.121	(0.108 ; 0.13)
Voto_alternativo_cod	Simpatia_cod	0.143	(0.128 ; 0.153)
Valoracion_ps_cod	Simpatia_cod	0.104	(0.092 ; 0.113)
Ideologia_pc_cod	Ideologia_ps_cod	0.114	(0.099 ; 0.128)
Ideologia_pi_cod	Ideologia_ps_cod	0.155	(0.139 ; 0.166)
Ideologia_ps_cod	Ideologia_pc_cod	0.113	(0.098 ; 0.123)
Ideologia_pi_cod	Ideologia_pc_cod	0.118	(0.105 ; 0.129)
Ideologia_sa_cod	Ideologia_pc_cod	0.146	(0.129 ; 0.157)
Ideologia_ia_cod	Ideologia_pc_cod	0.163	(0.147 ; 0.176)
Ideologia_pc_cod	Ideologia_pi_cod	0.154	(0.133 ; 0.165)
Ideologia_sa_cod	Ideologia_pi_cod	0.143	(0.125 ; 0.157)
Ideologia_ia_cod	Ideologia_pi_cod	0.146	(0.13 ; 0.161)
Valoracion_pi_cod	Ideologia_pi_cod	0.118	(0.108 ; 0.129)
Ideologia_cod	Ideologia_sa_cod	0.131	(0.117 ; 0.146)
Ideologia_ps_cod	Ideologia_sa_cod	0.147	(0.123 ; 0.162)
Ideologia_pi_cod	Ideologia_sa_cod	0.192	(0.173 ; 0.212)
Valoracion_sa_cod	Ideologia_sa_cod	0.178	(0.162 ; 0.192)
Ideologia_pc_cod	Ideologia_ia_cod	0.155	(0.142 ; 0.166)
Ideologia_pi_cod	Ideologia_ia_cod	0.114	(0.101 ; 0.123)
Ideologia_sa_cod	Ideologia_ia_cod	0.126	(0.111 ; 0.138)
Valoracion_ia_cod	Ideologia_ia_cod	0.147	(0.132 ; 0.159)

Fuente: Elaboración propia con RStudio

Tabla B.2*Intervalos Bootstrap para la Tau entre 0.1 y 0.2 (Variable 34 en adelante)*

Variable 1	Variable 2	K-tau	Intervalo Bootstrap
Pref_presi_cod	Valoracion_ps_cod	0.153	(0.141 ; 0.163)
Voto_cod	Valoracion_ps_cod	0.148	(0.135 ; 0.156)
Simpatia_cod	Valoracion_ps_cod	0.15	(0.135 ; 0.161)
Ideologia_ps_cod	Valoracion_ps_cod	0.117	(0.104 ; 0.127)
Recuerdo_cod	Valoracion_ps_cod	0.117	(0.104 ; 0.125)
Cercania_cod	Valoracion_ps_cod	0.126	(0.115 ; 0.134)
Conf_opo_cod	Valoracion_pc_cod	0.151	(0.137 ; 0.164)
Ideologia_pc_cod	Valoracion_pc_cod	0.113	(0.094 ; 0.127)
Valoracion_sa_cod	Valoracion_pc_cod	0.174	(0.158 ; 0.186)
Conf_presi_cod	Valoracion_pi_cod	0.192	(0.181 ; 0.202)
Pref_presi_cod	Valoracion_pi_cod	0.173	(0.158 ; 0.184)
Voto_cod	Valoracion_pi_cod	0.168	(0.153 ; 0.178)
Simpatia_cod	Valoracion_pi_cod	0.177	(0.164 ; 0.187)
Ideologia_cod	Valoracion_pi_cod	0.108	(0.09 ; 0.12)
Ideologia_ps_cod	Valoracion_pi_cod	0.103	(0.09 ; 0.111)
Ideologia_pi_cod	Valoracion_pi_cod	0.131	(0.117 ; 0.141)
Valoracion_sa_cod	Valoracion_pi_cod	0.1	(0.087 ; 0.115)
Recuerdo_cod	Valoracion_pi_cod	0.138	(0.124 ; 0.146)
Cercania_cod	Valoracion_pi_cod	0.152	(0.135 ; 0.163)
Conf_opo_cod	Valoracion_sa_cod	0.105	(0.09 ; 0.117)
Pref_presi_cod	Valoracion_sa_cod	0.136	(0.118 ; 0.151)
Voto_cod	Valoracion_sa_cod	0.133	(0.116 ; 0.145)
Simpatia_cod	Valoracion_sa_cod	0.152	(0.133 ; 0.166)
Ideologia_cod	Valoracion_sa_cod	0.13	(0.113 ; 0.142)
Ideologia_pc_cod	Valoracion_sa_cod	0.106	(0.084 ; 0.119)
Valoracion_ps_cod	Valoracion_sa_cod	0.132	(0.111 ; 0.151)
Valoracion_pi_cod	Valoracion_sa_cod	0.151	(0.135 ; 0.165)
Recuerdo_cod	Valoracion_sa_cod	0.125	(0.11 ; 0.137)
Cercania_cod	Valoracion_sa_cod	0.143	(0.133 ; 0.154)
Valoracion_pc_cod	Valoracion_ia_cod	0.194	(0.18 ; 0.207)
Valoracion_sa_cod	Valoracion_ia_cod	0.149	(0.137 ; 0.162)
Pref_presi_cod	Recuerdo_cod	0.12	(0.106 ; 0.13)
Participacion_cod	Recuerdo_cod	0.156	(0.147 ; 0.166)
Pref_presi_cod	Cercania_cod	0.179	(0.163 ; 0.194)

Fuente: Elaboración propia con RStudio

Tabla B.3*Intervalos Bootstrap para la Tau entre 0.2 y 0.5*

Variable 1	Variable 2	K-tau	Intervalo Bootstrap
Situ	Edad_grupos	0.258	(0.242 ; 0.271)
Edad_grupos	civil_cod	0.218	(0.194 ; 0.24)
Edad_grupos	Situ	0.437	(0.416 ; 0.457)
Pref_presi_cod	Conf_presi_cod	0.204	(0.189 ; 0.219)
Valoracion_ps_cod	Conf_presi_cod	0.365	(0.344 ; 0.379)
Valoracion_pi_cod	Conf_presi_cod	0.201	(0.188 ; 0.214)
Valoracion_pc_cod	Conf_opo_cod	0.215	(0.192 ; 0.24)
Voto_cod	Pref_presi_cod	0.287	(0.266 ; 0.308)
Simpatia_cod	Pref_presi_cod	0.259	(0.229 ; 0.283)
Pref_presi_cod	Voto_cod	0.255	(0.233 ; 0.272)
Voto_alternativo_cod	Voto_cod	0.336	(0.319 ; 0.349)
Recuerdo_cod	Voto_cod	0.277	(0.256 ; 0.296)
Cercania_cod	Voto_cod	0.305	(0.289 ; 0.323)
Voto_cod	Voto_alternativo_cod	0.393	(0.378 ; 0.406)
Pref_presi_cod	Simpatia_cod	0.27	(0.252 ; 0.289)
Recuerdo_cod	Simpatia_cod	0.318	(0.297 ; 0.338)
Cercania_cod	Simpatia_cod	0.446	(0.424 ; 0.462)
Ideologia_ps_cod	Ideologia_pi_cod	0.234	(0.214 ; 0.246)
Ideologia_pc_cod	Ideologia_sa_cod	0.282	(0.257 ; 0.299)
Ideologia_ia_cod	Ideologia_sa_cod	0.219	(0.196 ; 0.233)
Conf_presi_cod	Valoracion_ps_cod	0.266	(0.253 ; 0.28)
Valoracion_pi_cod	Valoracion_ps_cod	0.265	(0.244 ; 0.281)
Valoracion_ia_cod	Valoracion_pc_cod	0.208	(0.185 ; 0.225)
Valoracion_ps_cod	Valoracion_pi_cod	0.323	(0.304 ; 0.336)
Ideologia_sa_cod	Valoracion_sa_cod	0.206	(0.183 ; 0.227)
Valoracion_pc_cod	Valoracion_sa_cod	0.271	(0.246 ; 0.299)
Valoracion_ia_cod	Valoracion_sa_cod	0.219	(0.199 ; 0.234)
Ideologia_ia_cod	Valoracion_ia_cod	0.213	(0.194 ; 0.228)
Voto_cod	Recuerdo_cod	0.29	(0.275 ; 0.301)
Simpatia_cod	Recuerdo_cod	0.303	(0.286 ; 0.315)
Cercania_cod	Recuerdo_cod	0.322	(0.301 ; 0.342)
Voto_cod	Cercania_cod	0.338	(0.315 ; 0.354)
Simpatia_cod	Cercania_cod	0.437	(0.417 ; 0.457)
Recuerdo_cod	Cercania_cod	0.325	(0.305 ; 0.341)

Fuente: Elaboración propia con RStudio

Tabla B.4*Intervalos Bootstrap para la Tau mayor que 0.5*

Variable 1	Variable 2	K-tau	Intervalo Bootstrap
Simpatia_cod	Voto_cod	0.579	(0.559 ; 0.601)
Voto_cod	Simpatia_cod	0.65	(0.63 ; 0.668)
Recuerdo_cod	Participacion_cod	0.832	(0.799 ; 0.859)

Fuente: Elaboración propia con RStudio

ANEXO C

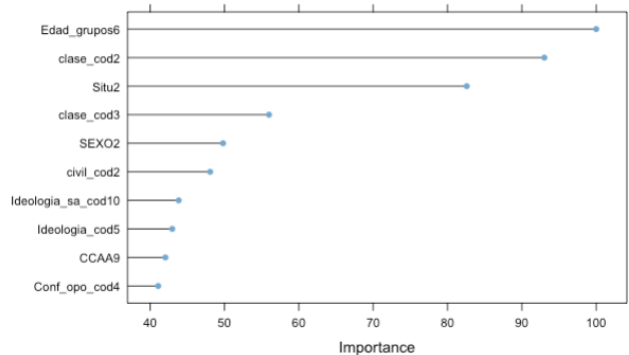
Figura C.1

Matriz de confusión e importancia de variables para el nivel de Estudios

Matriz de confusión Nivel de Estudios

Prediction	Reference			
	1	2	3	4
1	30.67	10	0.67	1
2	68	350.67	161.67	0
3	3.33	140.33	230.67	0
4	0	0	0	0

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

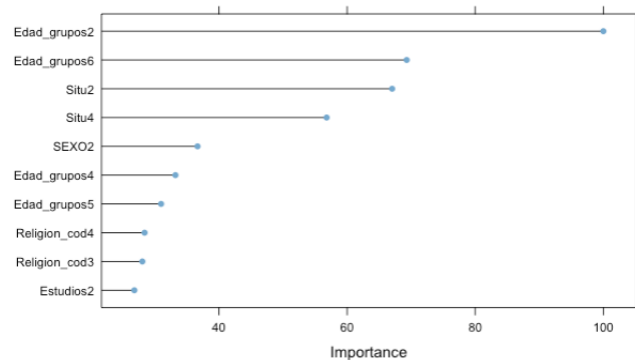
Figura C.2

Matriz de confusión e importancia de variables para el estado civil

Matriz de confusión Estado Civil

Prediction	Reference		
	1	2	3
1	498.33	188.67	62
2	59.67	182.33	1
3	1	1	1

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

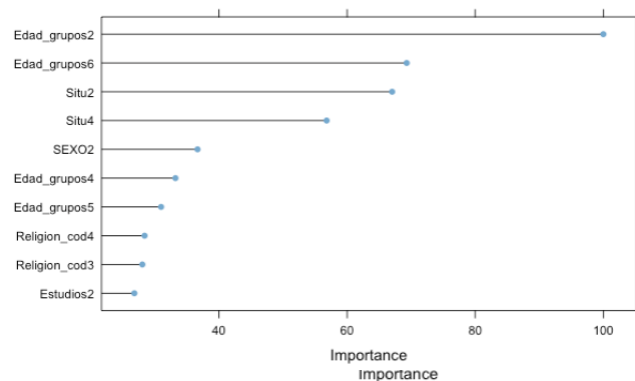
Figura C.3

Matriz de confusión e importancia de variables para la participación electoral

Matriz de confusión Participación electoral

Prediction	Reference		
	1	2	3
1	862.67	3	2.67
2	0.33	115	5.33
3	0	1	1

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

Figura C.4

Matriz de confusión e importancia de variables para la religión

Matriz de confusión Religión

Prediction	Reference			
	1	2	3	4
1	573	25.33	139.67	84.67
2	0	0	0	0
3	14.33	2.67	32.67	27.67
4	8.67	0	34.67	39.67

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero

Fuente: Elaboración propia con RStudio

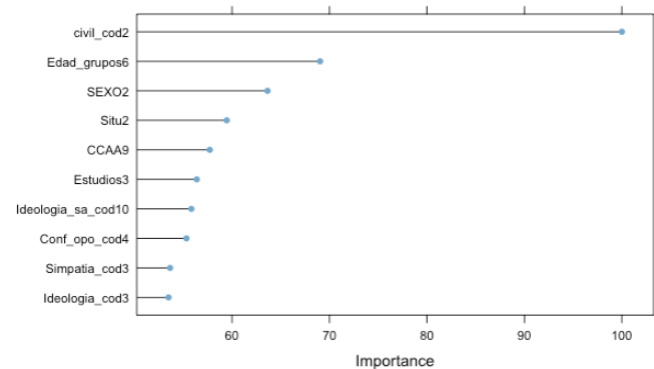


Figura C.5

Matriz de confusión e importancia de variables para la situación económica

Matriz de confusión Situación económica

Prediction	Reference				
	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	1	33.67	57	274.67	166
5	1	9.33	10	130.33	293

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero

Fuente: Elaboración propia con RStudio

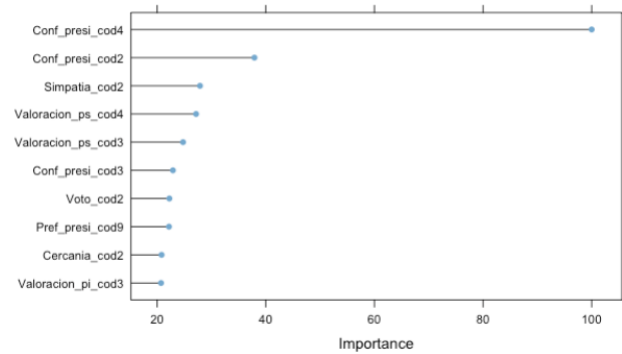


Figura C.6

Matriz de confusión e importancia de variables para la confianza en el presidente del gobierno

Matriz de confusión Confianza en el Presidente del gobierno

Prediction	Reference			
	1	2	3	4
1	1	4	0	0
2	49.67	208.33	65	7
3	2.33	44	182.33	41.33
4	1	1.67	53.67	318.67

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero

Fuente: Elaboración propia con RStudio

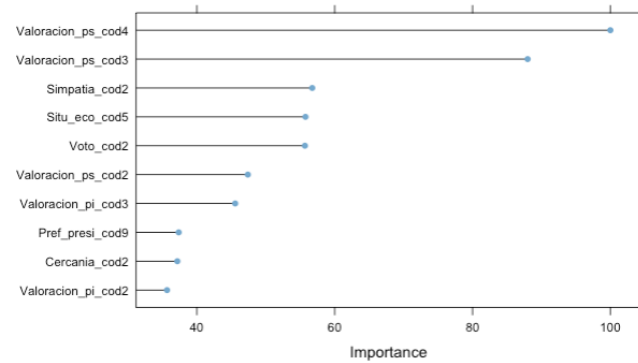


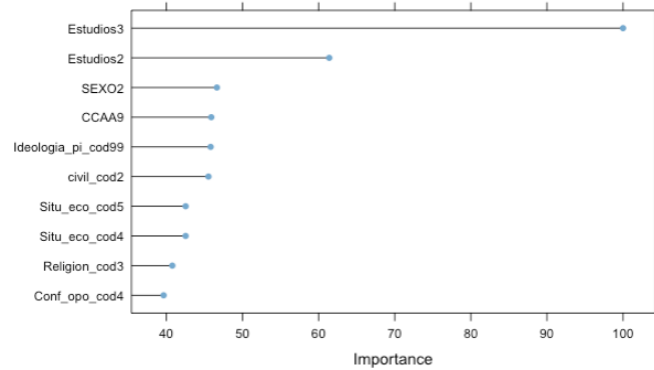
Figura C.7

Matriz de confusión e importancia de variables para la clase social del encuestado

Matriz de confusión Clase social

Prediction	Reference				
	1	2	3	4	5
1	0	0	0	0	0
2	2	737	200	10	4
3	0	4	3	0	0
4	0	0	0	0	0
5	0	0	0	0	0

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

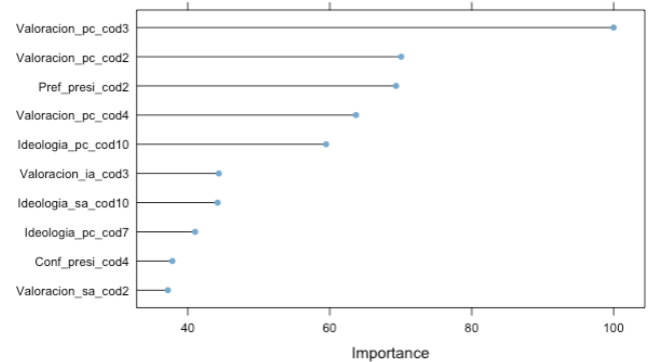
Figura C.8

Matriz de confusión e importancia de variables para la confianza en el presidente de la oposición

Matriz de confusión Confianza en el presidente de la oposición (Pablo Casado)

Prediction	Reference			
	1	2	3	4
1	0	0	0	0
2	7.33	22.33	8.67	1
3	8.67	61	244.67	79
4	0	9.67	114.67	417

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

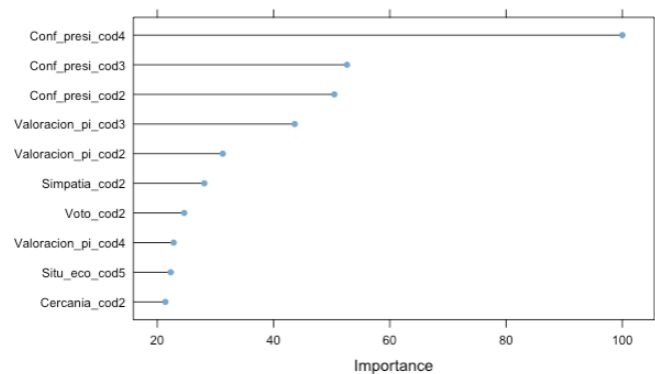
Figura C.9

Matriz de confusión e importancia de variables para la valoración de Pedro Sánchez

Matriz de confusión Valoración Pedro Sánchez

Prediction	Reference					
	1	2	3	4	5	6
1	275.33	45.33	14	1	2	1
2	15.33	62	24.67	1.67	0	0
3	5.33	52.67	124.33	53	7.33	0
4	2	5	69	150.67	41.33	0
5	0	0	0	4.67	13.33	0
6	0	0	0	0	0	0

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

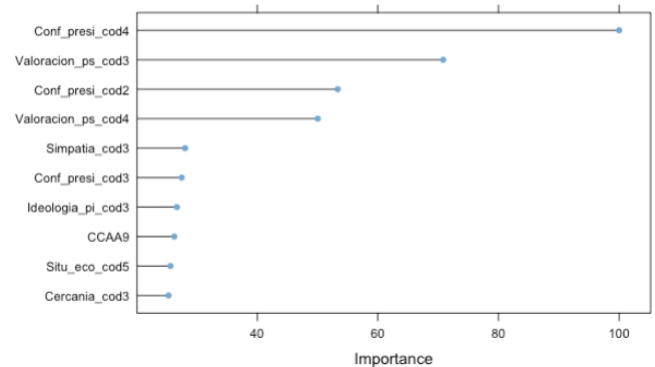
Figura C.10

Matriz de confusión e importancia de variables para la valoración de Pablo Iglesias

Matriz de confusión Valoración Pablo Iglesias

Prediction	Reference					
	1	2	3	4	5	6
1	384.33	53.67	25.67	7.67	4	2
2	20.67	50.33	29.33	3.33	1.33	0
3	16	80	126.33	65.67	14	0
4	1	5	16.67	40.33	12.33	1
5	0	0	0	0	0.33	0
6	0	0	0	0	0	0

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

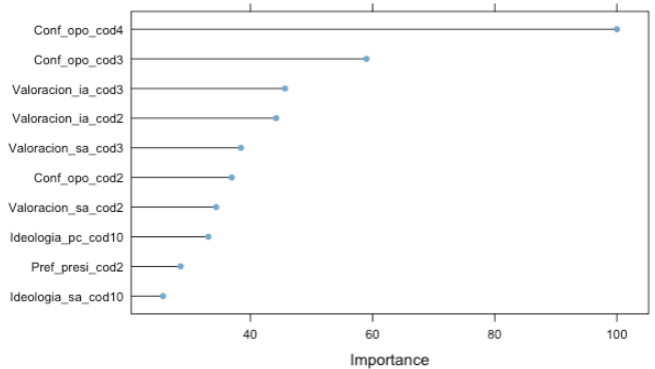
Figura C.11

Matriz de confusión e importancia de variables para la valoración de Pablo Casado

Matriz de confusión Valoración Pablo Casado

Prediction	Reference					
	1	2	3	4	5	6
1	337.33	83.33	24.67	3.67	5	2.33
2	66	142.33	78.33	11	1	0.67
3	10	29	93	36	7.33	0
4	0.67	0.33	1	14.33	6.67	0
5	0	0	0	0	1	0
6	0	0	0	0	0	4

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

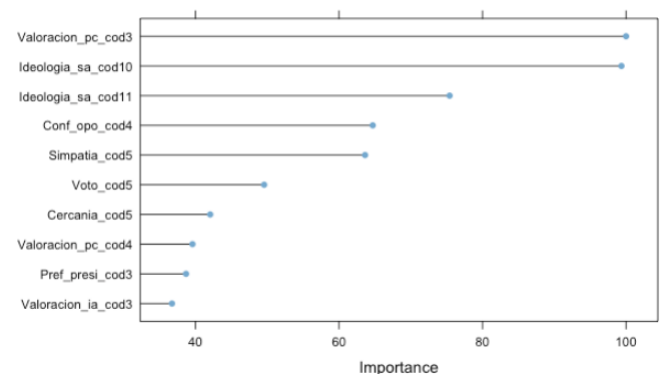
Figura C.12

Matriz de confusión e importancia de variables para la valoración de Santiago Abascal

Matriz de confusión Valoración Santiago Abascal

Prediction	Reference					
	1	2	3	4	5	6
1	653	85.33	47.67	13.67	5.67	2
2	4.33	8.67	10.67	0	0	0
3	6.67	4.67	20.33	13.33	6.33	0
4	0	2.33	4	17.33	5.33	0
5	0	0	4.33	3.67	19.67	0
6	0	0	0	0	0	16

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

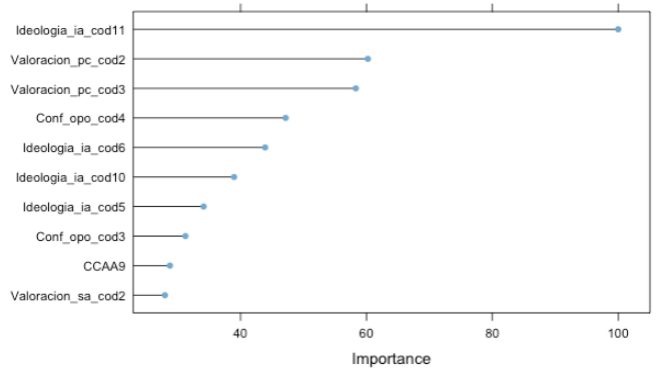
Figura C.13

Matriz de confusión e importancia de variables para la valoración de Inés Arrimadas

Matriz de confusión Valoración Inés Arrimadas

Prediction	Reference					
	1	2	3	4	5	6
1	251.33	58.33	13.67	4.33	5.67	0
2	57.33	114.67	58.67	6	2.67	0
3	24.33	85	149.67	54.67	13.67	0
4	0	0	2	6	1	0
5	0	0	0	0	0	0
6	0	0	0	0	0	38

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

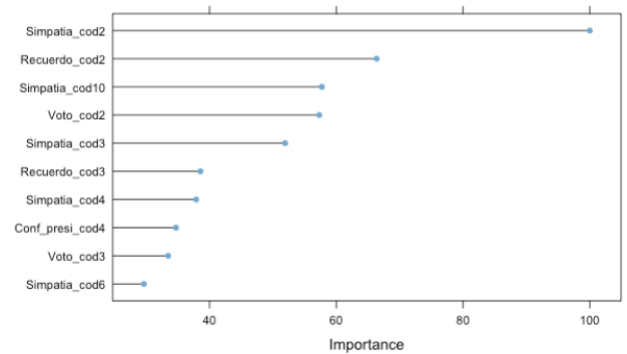
Figura C.14

Matriz de confusión e importancia de variables para la cercanía partidista

Matriz de confusión Cercanía partidista

Prediction	Reference									
	1	2	3	4	5	6	7	8	9	10
1	91	2	1	9	4.67	0	0	1	0	6
2	3.33	222.67	18	8	1	6	3	2	10.33	18.33
3	0	2	77	0	0	5.67	0	0	8	4
4	3	4	1	65	0	0	0	0	1	5
5	9	3	0	7	58.33	0	0	0	1	5
6	0	1	1	0	0	51.67	1.33	0	2.33	3
7	0	0	0	0	0	0.33	24.67	0	0	1
8	0	0	0	0	0	0	0	6	0	0
9	0	2.67	5	0	0	1	3	2	40	4.33
10	7.67	10.67	3	9	2	2.33	3	3	7.33	93.33

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

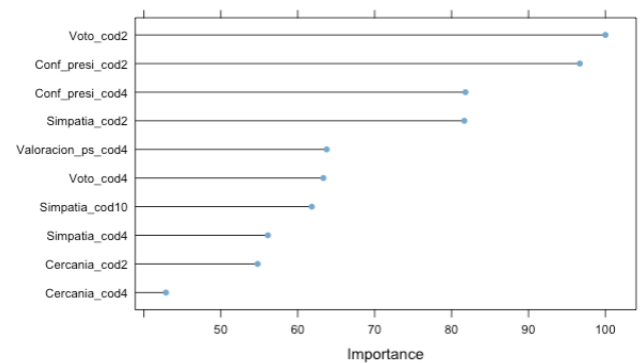
Figura C.15

Matriz de confusión e importancia de variables para la preferencia de presidente

Matriz de confusión Preferencia como presidente del gobierno

Prediction	Reference								
	1	2	3	4	5	6	7	8	9
1	276	2	1	26.67	4.67	1	17.67	19.67	36.33
2	0	69.67	2.67	0	0	8	0	7	13
3	1	3	62.33	0	0	0	0	4	9
4	2	0	0	18.33	4.67	0	3.33	1.67	1.67
5	0	0	0	0	0	0	0	0	0
6	2	1	1	0	0	51	1	5	17.33
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	18	15.33	2	18	4.67	11	6	21.67	175.67

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

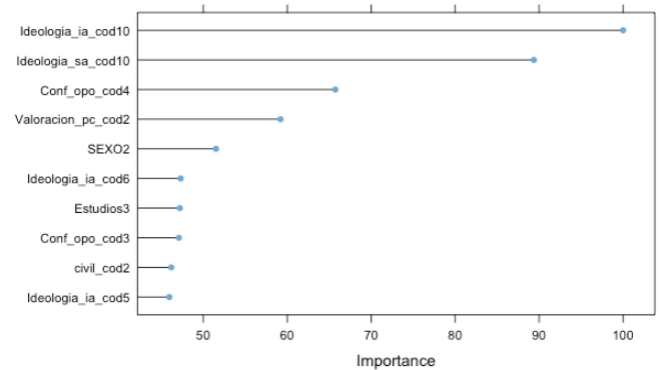
Figura C.16

Matriz de confusión e importancia de variables para la clasificación ideológica de Pablo Casado

Matriz de confusión Clasificación ideológica Pablo Casado

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0.33	0	0	0	0	0
5	3	0	2.67	6.67	23.67	12	8.33	4.67	2	4	0
6	0	0	1.33	2.67	12.67	15.67	14.67	5	1	1.67	0
7	2	0	1.67	4.67	24	41.33	100.67	54.67	16.67	11.33	0
8	2.33	0	4.33	2	8.33	15.33	64	104.67	53.33	34.33	0.33
9	0	0	0	0	0	0	2.67	9.67	12.67	5.33	0
10	1.67	0	0	4	3.33	4.33	15.67	27.33	40.33	98.33	1
11	0	0	0	0	0	0	0	0	0	0	5.67

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

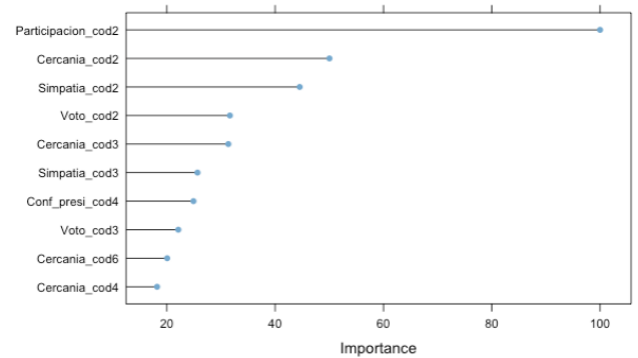
Figura C.17

Matriz de confusión e importancia de variables para el recuerdo de voto en las anteriores elecciones

Matriz de confusión Recuerdo de Voto

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	90	2.33	1	11	5	0	1	0.67	4.33	0	1
2	4	234.67	15.33	8.33	1	6	3	2	15	3.33	4.67
3	0	2	83	0	0	1.33	0	1	5	0	2
4	3	6	1	46.67	3.67	0	0	0	0	0	0
5	13	1	0	3	46	0	0	0	0	0	6
6	0	1	3	0	0	52.67	3	0	1.33	0	0.33
7	0	1	0	0	0	4	21.67	0	3	0	0
8	0	0	0	0	0	0	0	6.33	0	0	0
9	2	3	7	1	0	1	0.33	1	46	2	0.67
10	0	0	0	0	0	0	0	0	0	0	0
11	1	1	1.67	1	1.33	0	0	0	3.33	3.67	120.33

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

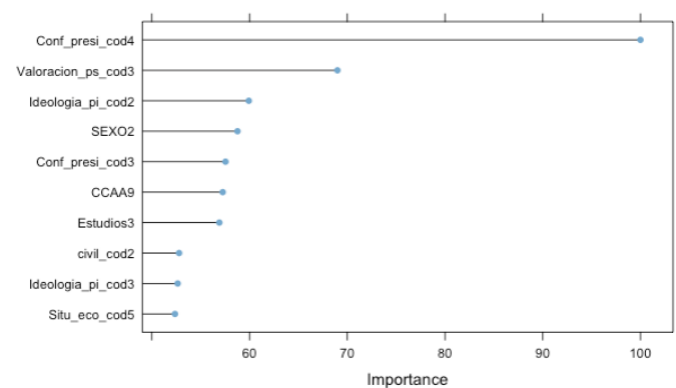
Figura C.18

Matriz de confusión e importancia de variables para la clasificación ideológica de Pedro Sánchez

Matriz de confusión Clasificación ideológica Pedro Sánchez

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	118.33	37.33	42	17.67	15.33	2	0.67	4.33	0	6.33	1
2	0.67	0	0	0	0	0	0	0	0	0	0
3	17	28	54	44.67	22.67	7.67	3.67	0	0	1	0
4	3	4	26.33	40	26.67	6.67	1.33	0.33	0	0	0
5	17	10.33	31.33	56.67	118.33	61	22.33	10.33	1	0.67	0
6	0	0.33	1.33	0	3.67	4	8.33	1	0	0	0
7	0	0	0	0	0.33	1.67	1.67	1.67	0	1	0
8	0	0	0	0	0	0	0	1	1.33	0	1
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

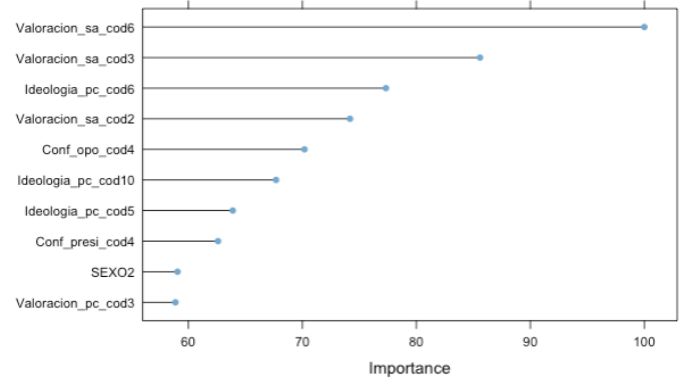
Figura C.19

Matriz de confusión e importancia de variables para la clasificación ideológica de Santiago Abascal

Matriz de confusión Clasificación ideológica Santiago Abascal

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	2.33	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	1	0	0	0	0.67	0
8	0	0	0	0	0	0.67	3.67	4.33	4	2	0
9	0	0	0	0	0	0	1	2.67	5	4.33	0
10	18	4	4	4	11.67	16.33	25.33	40	91	630	4.67
11	0	0	0	0	0	0	0	0	0	0	13.33

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

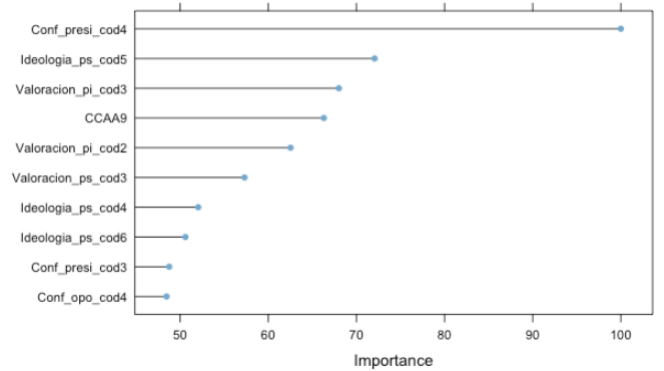
Figura C.20

Matriz de confusión e importancia de variables para la clasificación ideológica de Pablo Iglesias

Matriz de confusión Clasificación ideológica Pablo Iglesias

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	369.33	81.33	59.67	41	33.67	18.33	3.67	5.33	1	11	3
2	20.67	43.33	18.67	8	7.67	1.67	0.33	0	0	0	0
3	11	19.33	26.33	15.33	11.67	2.33	1.33	0	0	0	0
4	1	3	4	6	8.67	2.67	0.67	0	0	0	0
5	1	3	2.33	6.67	16	4	2.33	0.67	0	0	0
6	0	0	0	0	0.33	1	1.67	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

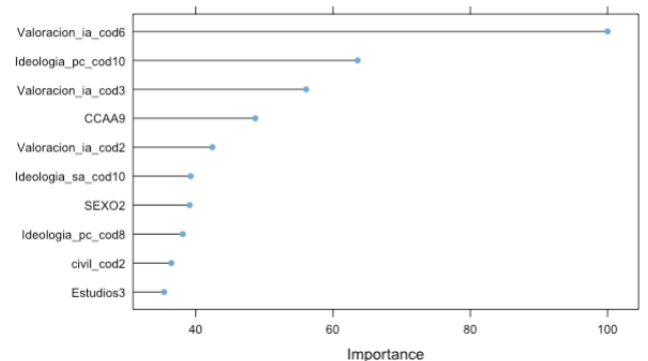
Figura C.21

Matriz de confusión e importancia de variables para la clasificación ideológica de Inés Arrimadas

Matriz de confusión Clasificación ideológica Inés Arrimadas

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0.33	0	0	0	0	0	0	0	0
5	10	12.67	10	26.33	138.33	93.33	40	11.67	5.67	4.67	0
6	0	0.67	2.67	4.67	41	69.67	39.33	20.33	5.67	3	0
7	0	0.67	0	2.33	8.67	13.67	19.67	23.67	9	6.33	0
8	2.67	0	0	1.67	5	9.67	15	22	15.67	9.67	0
9	0	0	0	1	2	0	2.33	4	12.33	1.67	0
10	0.33	0	0	0	1	2.33	7.67	13.33	15.67	69.67	0
11	0	0	0	0	0	0.33	0	0	0	1	38

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio

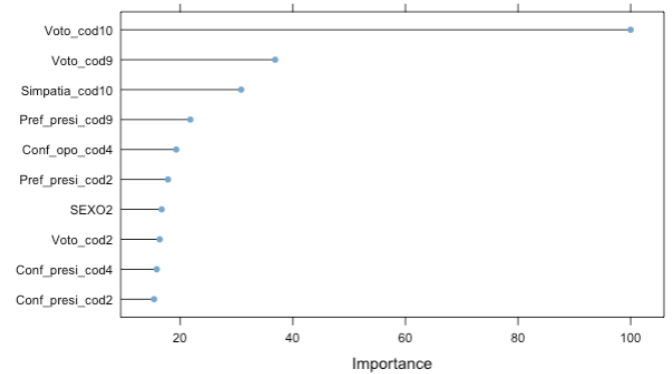
Figura C.22

Matriz de confusión e importancia de variables para la intención de voto alternativo

Matriz de confusión Intención de Voto alternativo

Prediction	Reference										
	1	2	3	4	5	6	7	8	9	10	11
1	63.67	6.67	1	6	3.67	0	0	1	7	2.33	15
2	1.33	51.67	8.33	0	0	3	0	1	10	1.33	6
3	2.33	2.67	44.67	10	1	4.33	3.33	1	15	6	19.67
4	4	0.33	2.67	52.33	14.67	0	0	0	10.33	5.67	17
5	0	0	0	2	7	0	0	0	3	2.67	1.67
6	1	1	1	0	0	10	0.67	0	3	0	2.33
7	0	3.33	4.33	0	0	0	3.33	0	1.33	0.67	1
8	0	0	0	0	0	0	0	0	0	0	0
9	2.67	14	17	4	0.33	6	14.33	0	93.33	3.33	22.33
10	0	0	0	0	0	0	0	0	0	0	0
11	17	4.33	12	12.67	5.33	5.67	0.33	1	18	9	179

* Al tomar la media de los elementos de las matrices de confusión, es posible que no se obtenga un número entero



Fuente: Elaboración propia con RStudio