Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: Informática, Multimedia y Telecomunicación

# Predicción de panel attrition con Machine Learning: El caso de la Encuesta Financiera de las Familias

---

Autor: Carlos Luis Gento de Celis

Tutor: Jordi Escayola Mansilla

Profesor: Antonio Lozano Bagén

---

Madrid, 27 de diciembre de 2023

# Créditos/Copyright

# FICHA DEL TRABAJO FINAL

| | |
|---|---|
| Título del trabajo: | Predecir panel attrition con Machine Learning: Un análisis con la Encuesta Financiera de las Familias |
| Nombre del autor: | Carlos Luis Gento de Celis |
| Nombre del colaborador/a docente: | Jordi Escayola Mansilla |
| Nombre del PRA: | Antonio Lozano Bagén |
| Fecha de entrega (mm/aaaa): | 10/2023 |
| Titulación o programa: | Máster Universitario en Ciencia de Datos |
| Área del Trabajo Final: | Informática, Multimedia y Telecomunicación |
| Idioma del trabajo: | Español |
| Palabras clave | predictive models, machine learning, panel attrition |

# Resumen

La Encuesta Financiera de las Familias (EFF) es una encuesta bienal cuyo objetivo es recoger información sobre la situación económico-financiera de los hogares que residen en España, y su evolución a lo largo del tiempo. Para ello, los hogares seleccionados pueden participar en hasta cuatro olas consecutivas de la encuesta. Sin embargo, hay hogares que abandonan el estudio antes de tiempo. Aunque estos abandonos no interrumpan el estudio, sí podrían afectar a sus resultados si el número de abandonos es demasiado grande, o si se concentra en colectivos específicos de la población. Es importante analizar las causas de estos abandonos y desarrollar herramientas que puedan evitarlos.

Este documento, en primer lugar, analiza las características de los hogares que participaron en la Encuesta Financiera de las Familias (EFF) en sus ediciones de 2017 y 2020, y si participaron en las olas de 2020 y de 2022. A continuación, plantea una serie modelos de predicción basados en métodos de Machine Learning y evalúa su capacidad para predecir si un hogar abandonará la EFF en 2020 o en 2022. Finalmente, interpreta los resultados el modelo que mejor ha funcionado.

**Palabras clave**: predictive models, panel attrition, machine learning, household surveys

# Abstract

The Spanish Survey of Household Finances (EFF) is a biennial survey whose goal is to collect information about the economical and financial situation of households in Spain, and its evolution through time. To do so, selected households may take part in up to four consecutive waves of this survey. However, some households cease their participation prematurely. Although withdrawals do not interrupt the research study, they might affect its results if the number of withdrawals is too high, or if it is more likely to happen for certain groups of people. It is important to analyse the causes of this phenomenon and develop tools to prevent it.

Firstly, this paper analyses the characteristics of households that responded to the EFF in its editions of 2017 and 2020, and their participation during the 2020 or 2022 waves. Next, a series of predictive models based on machine learning methods are considered and evaluated for the exercise of predicting panel attrition in the EFF. Finally, it interprets the results of the most successful model.

**Key words**: predictive models, panel attrition, machine learning, longitudinal surveys, household surveys

# Índice general

# Capítulo 1

# Introducción

Las encuestas son una forma de recolección de datos que se fundamenta en plantear preguntas relevantes a una muestra de unidades muestrales, y utilizar sus respuestas para entender a una población en su conjunto. Desde una perpectiva temporal, hay dos tipos de encuestas: las de sección cruzada, y las longitudinales o panel. Las encuestas de sección cruzada realizan sus preguntas a las unidades muestrales en un momento concreto del tiempo, mientras que las longitudinales plantean sus preguntas de manera repetida a las mismas unidades muestrales durante un período de tiempo (semanas, meses, años...). En este documento nos vamos a centrar en las encuestas longitudinales.

La dimensión temporal de las encuestas longitudinales las convierten en una herramienta muy útil para poder analizar relaciones causales, ya que recogen cambios de opiniones, comportamientos o estados de los mismos encuestados panel a lo largo del tiempo. Sin embargo, la calidad de esos análisis depende de la cooperación exitosa y continuada de dichos encuestados durante las sucesivas ediciones u olas de la encuesta. El abandono prematuro y acumulado en el tiempo de participantes en un panel se conoce como **Panel Attrition** (Watson and Wooden (2009).

Conceptualmente, las causas del Panel Attrition pueden clasificarse en tres categorías secuenciales: no-localización, no-contacto y no-cooperación (Lepkowski et al. (2002)). La no-localización se refiere a no conseguir localizar a los panelistas durante el período de campo. Esto suele ocurrir por cambios o errores en la información de contacto recogida durante la ola anterior (direcciones de residencia, teléfonos o email). Si un participante panel cambia de residencia o cambia de número de teléfono, es más costoso o incluso inviable localizarle.

La segunda causa es el no-contacto, que es el hecho de, después de localizar exitosamente al panelista, no conseguir establecer un contacto. Para que esto suceda, es necesario que el intento de contacto por parte de los encuestadores coincida temporalmente con la disponibilidad del panelista. Esto depende completamente del método de recolección de los datos. Por ejemplo,

en una entrevista personal, el panelista debe estar en su residencia justo en el mismo momento en el que el entrevistador hace la visita al hogar.

Finalmente, después de establecer el contacto, es necesario convencer al panelista para que vuelva a participar otra vez en la encuesta. La falta de cooperación en encuestas en general se ha analizado bastante en la literatura, y suelen destacar factores como las catacterísticas demográficas de los encuestados o la temática de la encuesta (para un desarrollo más detallado, puede consultarse Groves et al. (1992)). Sin embargo, en el caso de las encuestas longitudinales, los encuestados tienen experiencia previa por haber participado anteriormente, y este hecho puede potenciar el efecto de factores como la duración de la entrevista, la carga cognitiva por pensar las respuestas o la fatiga por haber participado en varias ediciones anteriores (Laurie et al. (1999), Watson and Wooden (2009), Lynn (2018)).

Con respecto a las consecuencias del Panel Attriton, Lynn (2018) destaca dos principales problemas. Por un lado, si la tasa de attrition es alta, el tamaño de la muestra se reducirá drásticamente con el paso de las olas, lo que provocará que la precisión de los estimadores de la encuesta sea muy baja, y además limitará o incluso imposibilitará el análisis de subgrupos dentro de la muestra. Por otro lado, si el Panel Attrition no es aleatorio, y por tanto los panelistas que abandonan la encuesta son sistemáticamente diferentes a los que se mantienen, existe el riesgo de introducir un sesgo de no-respuesta en los estimadores.

Tradicionalmente, los efectos del Panel Attrition se han mitigado con el uso de métodos de imputación múltiple (Rubin (1987)), reponderando de pesos muestrales (Groves et al. (2009)) e introduciendo muestras de refresco para sustituir a las unidades muestrales perdidas (Hirano et al. (1998)). Pero en las últimas décadas se ha extendido el uso de los diseños adaptativos y reactivos (adaptative and responsive designs, Groves and Heeringa (2006), Wagner (2008), Schouten et al. (2017), Tourangeau et al. (2017)). La idea detrás de estos diseños se fundamenta en utilizar toda la información que se genera durante la elaboración de encuestas (respuestas al cuestionario, paradata u observaciones de los entrevistadores) para diseñar implementaciones informadas cuyo objetivo sea mejorar la calidad de los datos, reducir los costes, o ambos. Por ejemplo, se han utilizado para revisar incentivos a participar para grupos concretos de encuestados Mcgonagle et al. (2022) o revisar el orden de las preguntas (Early et al. (2017). En este sentido, las encuestas longitudinales ofrecen una gran oportunidad para estos diseños porque contienen mucha información tanto de los trabajos de campo que se estén desarrollando, como los que se realizaron en olas anteriores. Por ejemplo, en Kreuter and Müller (2015) utilizan los registros de llamadas a panelistas en ediciones anteriores para optimizar las estrategias de contacto en las ediciones siguientes.

Dentro de este contexto de los diseños adaptativos y reactivos, en las últimas décadas se ha desarrollado el uso algoritmos de Machine Learning en la metodología de encuestas (Buskirk

et al. (2018), Kern et al. (2019)). Y de manera particular, han presentado resultados prometedores a la hora de predecir resultados de participación en los trabajos de campo, especialmente los modelos basados en árboles de decisión (Kern et al. (2019), Kern et al. (2021), Liu (2020)). En el contexto de Panel Attrition, en Beste et al. (2023) utilizaron información de ediciones pasadas de una encuesta a hogares en alemania para entrenar un Random Forest e identificar hogares panelistas con una baja propensión a participar. Luego, utilizaron esa información para crear un diseño experimental en la siguiente ola de la encuesta, en el cual se asignaba un incentivo monetario adicional a la mitad de esos hogares. Sus resultados mostraron incrementos en las tasas de respuesta de los hogares tratados, y animaron a sus responsables a seguir utilizando este diseño adaptativo en futuras ediciones.

El objetivo de este Trabajo de Fin de Máster es adaptar la implementación de machine mearning vista en Beste et al. (2023) para predecir la participación de los hogares panel en el contexto de la Encuesta Financiera de las Familias (EFF). En cada uno de estos artículos se utiliza información sobre hogares que participaron en olas pasadas de dos encuestas, y se utiliza esa información para predecir una variable binaria de participación o no participación en una ola posterior. El rendimiento de todos los modelos se compara con un modelo de referencia, que es una regresión logística. En el caso de este proyecto, utilizamos datos de las olas 4, 5, 6 y 7 de la EFF (que se corresponden con los años 2011, 2014, 2017 y 2020) para entrenar varios modelos de machine learning. A continuación, utilizamos dichos modelos para predecir la participación de los hogares panel en la ola 8 (que se corresponde con el año 2022), y comparamos su rendimiento con respecto a un modelo de regresión logistica.

El resto de este documento se organiza de la siguiente manera. En el siguiente capítulo se exponen los objetivos, la planificación y las motivaciones personales de este proyecto. En el tercer capítulo se presenta la EFF, los datos que se han utilizado y la metodología aplicada. En el cuarto capítulo se presentan los resultados de la implementación descrita en el tercer capítulo, junto con los desafíos encontrados durante el proceso. Finalmente, el último capítulo contiene las conclusiones obtenidas de los resultados del proyecto y reflexiones sobre cuáles podrían ser los próximos pasos para continuar con este proyecto en el futuro.

# Capítulo 2

# Objetivos del proyecto, planificación y motivación personal

## 2.1. Objetivos del proyecto

Este proyecto se contextualiza dentro del marco de la metodología de encuestas. En concreto, se centra en las encuestas longitudinales a hogares, y dentro de ese campo pone el foco en la predicción de la participación de los hogares en olas posteriores a su primera colaboración. La encuesta elegida para el proyecto es la Encuesta Financiera de las Familias (EFF), una encuesta de referencia para investigaciones sobre finanzas de los hogares.

El **objetivo principal** de este proyecto es desarrollar un modelo de predicción basado en métodos de machine learning que ayude a predecir si un hogar que ha participado en al menos una ola de la EFF volverá a hacerlo en olas posteriores.

Para poder desarrollar ese modelo, es necesario completar una serie de **objetivos secundarios**. Estos objetivos se agruparán en 5 fases:

1. Hacer una revisión del estado del arte sobre el Panel Attrition y las metodologías de Machine Learning aplicadas a su predicción.

2. Recolección de datos. Preferentemente, obtener un conjunto de datos que contenga:

   *a*) Las respuestas de los hogares al cuestionario de la EFF.

   *b*) Paradata sobre el proceso de creación de la encuesta.

   *c*) Características de los entrevistadores.

3. Análisis exploratorio de los datos. Identificar patrones dentro de los datos que puedan estar relacionados con el panel attrition.

4. Preprocesar los datos para entrenar modelos de machine learning.
5. Modelado y evaluación de modelos de predicción basados en métodos de machine learning.
6. Interpretación del mejor modelo y redacción de conclusiones.

## 2.2. Planificación del proyecto

La planificación de este proyecto se va a dividir en 5 fases. La asignación temporal a cada una de ellas se ha realizado de acuerdo a la estructura de contenidos del Plan Docente y al calendario de la asginatura.

1. Fase 1: Contiene la definición del proyecto y la planificación del TFM. Abarca las dos primeras semanas del proyecto, del 27 de septiembre al 10 de octubre.
2. Fase 2: Contiene las tareas de la revisión de la literatura y la caracterización del panel attrition de la EFF. Abarcará desde el 11 de octubre al 23 de octubre.
3. Fase 3: Contiene las tareas de recolección de datos, análisis exploratorio de dichos datos, preprocesamiento para entrenar los modelos de machine learning, modelado y evaluación de los modelos de predicción, y la interpretación de los modelos y la redacción de conclusiones. Esta fase es la más duradera y abarcará desde el 17 de octubre al 19 de diciembre.
4. Fase 4: En esta fase se procederá a la redacción de la memoria del TFM y la preparación de una presentación audiovisual sobre el proyecto. Abarcará desde el 20 de diciembre al 18 de enero.
5. Fase 5: En esta fase final se procederá a la defensa del TFM. Esa etapa abarcará desde el 22 de enero hasta el 4 de febrero.

En la figura 2.1 puede observarse la distribución temporal y el tiempo dedicado a cada tarea en un diagrama de Gantt.

## 2.3. Motivación personal

Trabajo para el Banco de España, y formo parte del equipo que elabora la Encuesta Financiera de las Familias (EFF) desde principios de 2015. He participado en la elaboración de sus cuatro últimas ediciones (EFF2014, EFF2017, EFF2020 y EFF2022) y con los años he adquirido cada vez más interés por la metodología de encuestas y el potencial que tienen para recoger

| | Nombre actividad | Fecha inicio | Duración en días | Fecha fin |
|---|---|---|---|---|
| Fase 1 | 1. Definición y planificación del TFM | 27-Sep | 13 | 10-Oct |
| Fase 2 | 2. Revisión literatura | 11-Oct | 12 | 23-Oct |
| | 3. Caracterizar panel attrition | 11-Oct | 5 | 16-Oct |
| Fase 3 | 4. Recolección de datos | 17-Oct | 5 | 22-Oct |
| | 5. Análisis exploratorio | 23-Oct | 18 | 10-Nov |
| | 6. Preprocesamiento de datos | 10-Nov | 33 | 13-Dec |
| | 7. Modelización y evaluazión | 13-Nov | 30 | 13-Dec |
| | 8. Interpretación y conclusiones | 14-Dec | 5 | 19-Dec |
| Fase 4 | 9. Memoria | 20-Dec | 20 | 09-Jan |
| | 10. Presentación audiovisual | 20-Dec | 29 | 18-Jan |
| Fase 5 | 11. Defensa | 22-Jan | 13 | 04-Feb |

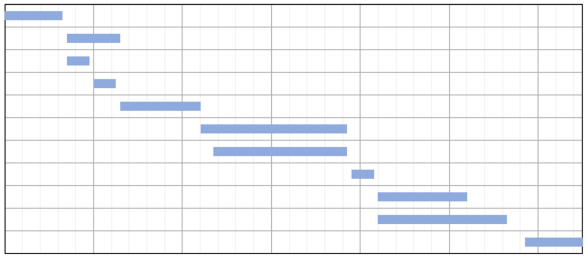


Figura 2.1: Planificación de las actividades del TFM

información sobre fenónemos que de otra manera serían difíciles de captar. La EFF es una fuente de información de referencia en el campo de las finanzas de los hogares, y el aumento del interés que se ha generado en los últimos años hace más importante realizar trabajos y esfuerzos para garantizar e incluso mejorar la calidad de sus datos. En ese sentido, la gran cantidad de paradatos que se generan durante cada ola ofrecen muchas oportunidades para aprender sobre todo el proceso, y también mejorarlo.

Por otro lado, en los últimos años ha aumentado la variedad de formatos en los que recibimos los datos, tomando una gran relevancia los audios de las entrevistas y los comentarios de texto escritos por los entrevistadores. Los métodos y las herramientas desarrolladas en el campo de la ciencia de datos abren un mundo de posibilidades para poder analizar toda esa información, y que hace apenas unos años no eramos capaces ni de imaginar.Es un buen momento y una gran oportunidad para trabajar en este campo.

Finalmente, sobre el Panel Attrition, considero que la etapa más importante de la elaboración de la EFF es el trabajo de campo. En ella se contacta a los hogares, se les convence para participar en la encuesta y se realizan las entrevistas. Marca el devenir las siguientes etapas, y también el nivel de calidad de los datos. Por esa razón, es importante conseguir la colaboración de los hogares. Especialmente la de los hogares panel, ya que representan una proporción importante de la muestra, y no convercerles puede llegar a ser muy costoso. Y es un área que en la EFF no se ha podido explorar hasta. Es una gran oportunidad para aprender.

# Capítulo 3

# Panel Attrition: Causas, predicción y machine learning

## 3.1. Causas del panel attrition, tratamiento y métodos adaptativos y reactivos

Conceptualmente, las causas del panel attrition pueden clasificarse en tres categorías secuenciales (Lepkowski et al. (2002)): no-localización, no-contacto y no-cooperación.

La no-localización se refiere a no conseguir localizar a los hogares panelistas durante el período de campo. Esto suele ocurrir por cambios o errores en la información de contacto recogida durante la ola anterior (direcciones de residencia, teléfonos o email) (Couper and Ofstedal (2009)). Algunos factores que pueden contribuir al éxito o fracaso en la localización son el método de recolección de datos, la propensión a los cambios de localización de los encuestados entre diferentes olas, el tiempo transcurrido entre olas o el presupuesto (Lynn (2009)).

La segunda causa es el no-contacto. Se trata de, tras localizar exitosamente al panelista, no conseguir establecer un contacto. Para que haya contacto exitoso, es necesario que el intento de contacto por parte de los encuestadores coincida temporalmente con la disponibilidad del panelista. Esto depende completamente del método de recolección de los datos. Por ejemplo, en una entrevista personal, el panelista debe estar en su residencia justo en el mismo momento en el que el entrevistador hace la visita al hogar.

Finalmente, después de establecer el contacto, es necesario convencer al panelista para que vuelva a participar otra vez en la encuesta. La falta de cooperación es una preocupación común en todo tipo de encuestas, y suelen destacar factores como las características socio-demográficas de los encuestados o la temática de la encuesta (Groves et al. (1992)). En el caso particular de las encuestas longitudinales, los encuestados además tienen experiencia previa por haber

participado en ediciones pasadas, y esto puede potenciar el efecto de factores como la duración de la entrevista, la carga cognitiva de pensar las respuestas o la fatiga por haber participado en ediciones anteriores (Laurie et al. (1999), Watson and Wooden (2009), Lynn (2018)).

Con respecto a las consecuencias del Panel Attriton, Lynn (2018) destaca dos principales problemas. Por un lado, si la tasa de attrition es alta, el tamaño de la muestra se reducirá drásticamente con el paso de las olas, lo que provocará que la precisión de los estimadores de la encuesta sea muy baja, y además limitará o incluso imposibilitará el análisis de subgrupos dentro de la muestra. Por otro lado, si el Panel Attrition no es aleatorio, y por tanto los panelistas que abandonan la encuesta son sistemáticamente diferentes a los que se mantienen, existe el riesgo de introducir un sesgo de no-respuesta en los estimadores.

Tradicionalmente, los efectos del Panel Attrition se han mitigado con la implementación de métodos de imputación múltiple (Rubin (1987)), reponderando los pesos muestrales (Groves et al. (2009)) e introduciendo muestras de refresco para sustituir a las unidades muestrales perdidas (Hirano et al. (1998)). Pero en las últimas décadas se ha extendido el uso de los llamados diseños adaptativos y reactivos (adaptative and responsive designs, Groves and Heeringa (2006), Wagner (2008), Schouten et al. (2017), Tourangeau et al. (2017)). Estos diseños se fundamentan en utilizar toda la información que se genera durante la elaboración de encuestas (respuestas al cuestionario, paradata, observaciones de los entrevistadores, información de olas pasadas...) para diseñar implementaciones informadas cuyo objetivo sea mejorar la calidad de los datos, reducir los costes, o ambos. Por ejemplo, se han utilizado para revisar incentivos a participar para grupos concretos de encuestados Mcgonagle et al. (2022), revisar el orden de las preguntas (Early et al. (2017) o utilizar los registros de llamadas a panelistas en ediciones anteriores para optimizar las estrategias de contacto en las ediciones siguientes (Kreuter and Müller (2015)).

Dentro de este contexto de diseños adaptativos y reactivos, en las últimas décadas destaca el desarrollo del uso de algoritmos de machine learning en la metodología de encuestas, y en particular para predecir no-respuesta y panel attrition (Buskirk et al. (2018), Kern et al. (2019)). Se comenta con más detalle en el siguiente apartado.

## 3.2. Predicción de Panel Attrition con machine learning

Los modelos utilizados en metodología de encuestas pueden clasificarse en dos categorías según sus objetivos: modelos para explicar, y modelos para predecir. Los modelos explicativos utilizan toda la información disponible para explorar las relaciones entre diferentes variables observadas, identificar causalidad entre ellas y realizar ejercicios de inferencia y contrastes de hipótesis. Los modelos predictivos, en cambio, buscan predecir o clasificar con precisión

el valor de ciertas variables para escenarios que todavía no han ocurrido. Por construcción, sólo utilizan la información disponible antes de que ocurra el suceso a predecir. Aunque los modelos basados en métodos de machine learning pueden ser utilizados para ambas tareas, son particularmente interesantes para realizar tareas de predicción. En Buskirk et al. (2018) destacan particularmente la flexibilidad de los modelos de machine learning con respecto a otros modelos tradicionales utilizados en la metodología de encuestas, como la regresión logística o los mínimos cuadrados ordinarios (OLS). Para muchos algoritmos de machine learning no es necesario hacer supuestos sobre las distribuciones de las variables, hacer una especificación explícita de las relaciones entre variables antes de estimar los modelos, y además soportan el uso de un gran número de variables. Esto les permite detectar patrones y relaciones complejas entre variables y los convierte en una herramienta muy útil para realizar tareas de predicción.

En el contexto de la predicción del panel attrition, muchos estudios comparan el rendimiento de modelos de machine learning con el de modelos que se han utilizado tradicionalmente para analizar panel attrition, generalmente una regresión logística o Logit de efectos principales, es decir, sin considerar interacciones entre variables ni relaciones no linales. En Kern et al. (2019) y Kern et al. (2021) utilizan datos de dos paneles de hogares en alemania para comparar el rendimiento de un Logit con varios modelos basados en árboles de decisión. Los resultados son prometedores ya que todos los modelos basados en árboles mostraron mejores rendimientos que el Logit, y destacan por su facilidad para ser interpretados. Sin embargo, estos resultados no parecen ser extrapolables a todo tipo de encuestas. Otro ejemplo prometedor puede verse es en Beste et al. (2023). Este estudio se divide en dos partes. En primer lugar, se utiliza información de ediciones pasadas de una encuesta a hogares en alemania para comparar, de nuevo, el rendimiento de un Logit con un algoritmo k-nearest neighbours (kNN), un árbol de clasificación (CART), un Random Forest (RF) y un Gradient Boosting Machine (GBM). El modelo que ofreció mejores resultados fue el Random Forest. Y en la segunda parte, utilizaron ese modelo de Random Forest para identificar hogares panelistas con una baja propensión a participar en la edición siguiente de la encuesta. Esa información les sirvió para crear un diseño experimental en el cual se asignaba un incentivo monetario adicional a la mitad de ésos hogares. Sus resultados del experimento mostraron incrementos en las tasas de respuesta de los hogares tratados, y animaron a sus responsables a seguir utilizando este diseño adaptativo en futuras ediciones.

Aunque sin duda estos resultados son prometedores, es importante recalcar que deben ser contextualizados y valorados para cada caso de análisis. Por ejemplo, en Liu (2020) se utiliza un panel de individuos en Estados Unidos para predecir la participación de panelistas en la segunda edición de dicha encuesta, y se compara de nuevo el rendimiento de un Logit con los de un Random Forest (RF), un modelo de Máquinas de Soporte Vectorial (SVM) y un LASSO.

Sólo el LASSO mostró una mejora con respecto al modelo de regresión logística. Otro ejemplo puede verse en Jankowsky and Schroeders (2022). En este estudio se compara el rendimiento de un modelo Logit con el de un modelo GBM (Gradient Boosting Machine) en dos encuestas con diseños bastante diferentes (una encuesta es en EEUU y la otra en alemania, una se ha realizado cada nueve años y la otra anualmente, una es telefónica y la otra es una entrevista presencial...). Para ambas encuestas apenas se observa que el GBM mejore los resultados de la regresión logística.

Como resumen, podemos decir que los algoritmos de machine learning poseen características que los hacen atractivos como herramientas de apoyo para el desarrollo de diseños adaptativos y reactivos en el área de la metodología de encuestas, y en particular para predecir. De manera particular, algunos modelos, especialmente los basados en árboles de decisión, han mostrado buenos resultados a la hora de predecir la participación de panelistas en futuras ediciones de encuestas longitudinales. Sin embargo, es importante recalcar que estos resultados no son necesariamente generalizables a cualquier tipo de encuestas, y su uso debe ser valorado para cada caso en particular.

# Capítulo 4

# Datos y metodología

## 4.1. Datos: La Encuesta Financiera de las Familias

Para este proyexto se van a utilizar los datos de la Encuesta Financiera de las Familias (EFF). La EFF es una encuesta oficial a hogares elaborada por el Banco de España y está incluida en el Plan Estadístico Nacional. Su primera edicion se realizó en el año 2002, y se ha producido de manera trienal hasta el año 2020. Desde entonces, se produce de manera bienal.

El objetivo de la EFF es recabar información sobre las condiciones financieras de los hogares residentes en España. Es la única fuente estadística que permite relacionar información sobre activos, deudas, ingresos y gastos de los hogares españoles. Esto permite analizar las decisiones de inversión y financiación de las familias y conocer su situación patrimonial, y gracias a ello tener un mayor conocimiento de la economía española y poder utilizarlo para hacer un diseño adecuado de politicas públicas. Por nombrar algunos ejemplos de estudios realizados con la EFF, se ha utilizado para cuantificar el ahorro adicional generado para los partícipes en planes de pensiones de empresa (Gómez and Villanueva (2022)), para caracterizar cómo afectó la pandemia del Covid-19 a la situación patrimonial de los trabajadores más afectados por dicha crisis (Alvargonzález et al. (2020) o para analizar las diferencias en aceptación y uso de tarjetas de crédito y banca online entre diferentes grupos de hogares desde el año 2002 (Crespo et al. (2023)).

El diseño de la muestra de la EFF tiene dos características importantes: un sobremuestreo de hogares ricos y un componente longitudinal o panel. El sobremuestreo de ricos[1] garantiza poder analizar con suficiente precisión el comportamiento de los hogares de la parte alta de la distribución de riqueza. Este detalle es importante porque la distribución de la riqueza entre

[1]La muestra de la EFF es seleccionada por el Instituto Nacional de Estadística, en colaboración con la Agencia Tributaria, a partir de las declaraciones individuales más recientes en el Impuesto sobre el Patrimonio. Para una descripción más detallada del proceso, puede consultarse Barceló et al. (2021).

los hogares es asimétrica, por lo que sólo unos pocos hogares, especialmente los más ricos, son los que invierten en ciertos activos. Por otro lado, el compomente panel indica que se vuelve a entrevistar a hogares que participaron en ediciones anteriores. Esto permite monitorearles durante períodos de hasta diez años, y observar los cambios en las variables de interés de la encuesta. El número máximo de ediciones en las que un hogar puede participar en la EFF es de cuatro olas consecutivas. Si cesa su participación antes de completar sus cuatro ediciones, se descarta de la muestra y no vuelve a ser contactado en olas posteriores. Finalmente, para sustituir a los hogares descartados del panel rotatorio y mantener la representatividad de la muestra, en cada nueva ola de la EFF se añade una nueva muestra de refresco a la muestra panel.

Volviendo a los datos, en este proyecto se utiliza información que proviene tanto de las respuestas al cuestionario de la EFF, como del paradata recopilado durante la producción de los datos. En la figura 4.1. puede observarse un resumen de este proceso. La producción de la EFF se divide en dos grandes fases: Campo y Post-Campo. Durante el Campo se contacta con los hogares, se realizan las entrevistas personales, se procesan los datos y se realiza parte de la revisión de los mismos. En el Post-Campo se termina el proceso de revisión, se evalúa el grado de no-respuesta de los datos de cada hogar para eliminar las entrevistas con poco contenido informativo, y se procede a la imputación de la no-respuesta. Tras este último proceso, se obtienen los datos finales.

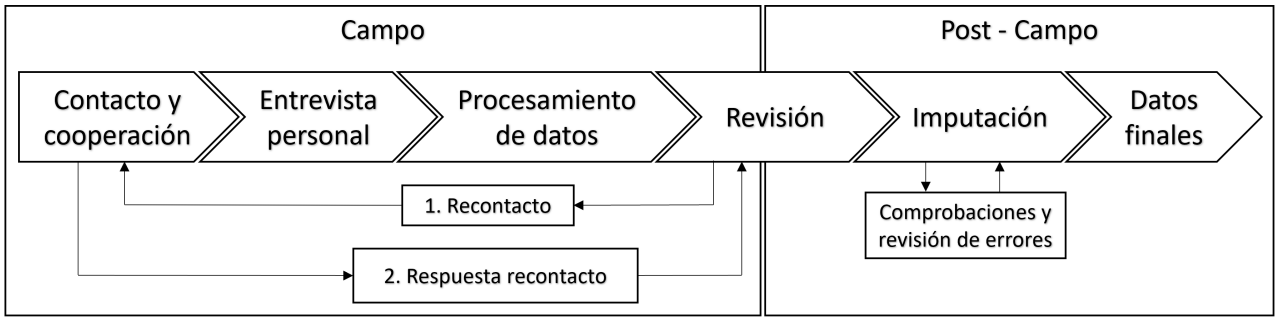


Figura 4.1: Fases de la creación de datos de la EFF

A continuación se hace una breve descripción de los aspectos más relevantes de las subfases y que son importantes para la selección de variables para este estudio.

- **Contacto y cooperación:** Los entrevistadores realizan visitas personales a los hogares en sus domicilios. Esto puede requerir varios intentos, ya que es necesario que algún miembro del hogar esté físicamente en su hogar cuando tenga lugar la visita presencial. La información sobre cada intento (fecha y hora, resultado...) se recoge en un ordenador.

Antes de establecer el contacto, los entrevistadores rellenan un cuestionario que recoge información sobre las características del barrio y del edificio en el que vive el hogar.

- **Entrevista personal:** Se realiza una entrevista personal a la persona con más conocimiento de las finanzas del hogar, que se denomina Persona de Referencia o PR. La PR puede ser miembro del hogar, o un representante del mismo, demoninado proxy, siempre y cuando sea la persona con mayor conocimiento de las finanzas del hogar. También pueden participar otros miembros del hogar. Antes de empezar la entrevista, se pide a la PR su consentimiento para que algunas partes de la misma puedan ser grabadas en audio por motivos de calidad de los datos. La entrevista puede realizarse aunque no haya grabación. Los entrevistadores recogen las respuestas en un ordenador o una tablet (CAPI), y también pueden anotar en comentarios de texto todos los detalles que consideren importantes para la revisión. La PR puede decidir no contestar a ciertas preguntas. Su valor se asignará como missing, y se imputará más adelante. Las cantidades monetarias pueden responderse en valor puntual, o dentro de un rango de valores. Tras la entrevista, y sin la presencia de los hogares, los entrevistadores rellenan un cuestionario con información sobre el desarrollo de la entrevista, en el que por ejemplo se recoge el nivel de comprensión de la PR a las preguntas, el nivel de interés mostrado por la PR o cuántas personas han participado en la entrevista.

- **Procesamiento de datos:** Los ordenadores y servidores de la empresa de campo procesan los datos recogidos durante el contacto con los hogares y durante las entrevistas. Se crean tres ficheros, uno con las respuestas al cuestionario, uno con la información de los contactos con cada hogar, y finalmente uno con la información del paradata recogido por el ordenador durante la entrevista[2].

- **Revisión:** Un equipo de personas revisa individualmente todas las entrevistas y corrige los errores que puedan detectarse. Si hay información relevante que se ha recogido erróneamente o se ha omitido, se contacta de nuevo con el hogar para corregirlo o recuperar esa información. Este recontacto se hace por teléfono. Toda la información sobre la revisión (cambios sobre variables, recontactos...) se recoge en una aplicación informática y puede ser exportada en ficheros de diversos formatos (csv, excel...).

- **Imputación:** Se analiza la proporción de preguntas sin responder dentro de cada entrevista (no-respuesta), y se eliminan las que no superen ciertos umbrales de calidad. Todas

[2]También se crea un fichero con los comentarios de texto recogidos por los entrevistadores durante la entrevista, pero no se ha podido incluir en este estudio por falta de tiempo.

las variables que contienen no-respuesta se imputan mediante técnicas de imputación múltiple[3]. Se crean 5 ficheros con datos imputados.

Con respecto a los procedimientos de Contacto y Entrevista personal, es necesario mencionar que algunos elementos tuvieron que modificarse durante la EFF2020, ya que el campo tuvo lugar entre noviembre de 2020 y junio de 2021, y se vió afectado por la pandemia del Covid-19. Durante esa ola, los entrevistadores siguieron visitando personalmente a los hogares para conseguir su colaboración[4], pero siempre respetando las medidas de distanciamiento social. Las entrevistas se realizaron de manera telefónica asistida por una tablet (CATI). El resto de procedimientos se mantuvieron como en otras ediciones.

Para la elaboración de este proyecto se utiliza la información recopilada durante las olas de la EFF2017, EFF2020 y EFF2022 [5]. Estas ediciones son las únicas que cuentan con la información más detallada de paradata y de las características de los entrevistadores. En ediciones anteriores esta información o no está disponible, o contiene errores de medida que no son fácilmente corregibles. A pesar de esto, es posible identificar en cuántas ediciones ha participado cada hogar, por lo que es posible utilizar información que se remonta hasta la edición de la EFF2011. Los ficheros de datos que se han utilizado durante este proyecto son:

- **Fichero de trabajo:** Registro de hogares con entrevistas válidas. Contiene la información de las respuestas de los hogares, incluyendo las correcciones y ediciones de la revisión. Se indican los datos que deben ser imputados. También incluye variables auxiliares generadas para el proceso de imputación (características del hogar, de los miembros, del municipio...) y contadores de no respuesta de cada entrevista. Es el fichero que se utiliza para imputar.

- **Fichero de datos imputados:** Registro de hogares que contiene las respuestas al cuestionario después de haber imputado los datos con no-respuesta de los hogares. Por simplicidad, sólo utilizaremos uno de los conjuntos de datos de la imputación multiple. Sólo contiene variables imputadas e indicadores de las características del hogar.

- **Fichero de contactos:** Registo de hogares contactados durante una ola de la EFF. Contiene información sobre el número de intentos de contacto con cada hogar, la fecha en que se produjo, el resultado de cada uno de ellos (aplazamientos, rechazos...) y las respuestas al cuestionario de vecindario rellenado por los entrevistadores.

[3]Los métodos de imputación múltiple utilizados en la EFF pueden consultarse en Barceló (2006).

[4]Durante el principio del campo, algunos hogares panel fueron contactados sólo por teléfono, pero a las pocas semanas se decidió establecer la visita personal como el procedimiento estándar.

[5]En el momento de escribir este documento, la EFF2022 se encontraba en pleno proceso de imputación, por lo que el equipo del Banco de España ya conocía los resultados de participación de esa ola y era posible utilizarlos para este proyecto.

- **Fichero de revisión:** Registro de hogares e incidencias que contiene información sobre el proceso de revisión y los recontactos. Para cada hogar hay varios registros. Uno contiene información general sobre el proceso de revisión (por ejemplo, si se ha realizado un recontacto), y el resto de registros son incidencias en los datos que se detectado (por ejemplo, si se ha omitido una propiedad inmobiliaria, se abrirá un registro indicando esa incidencia y cómo se ha solucionado).

- **Censo de entrevistadores:** Registro de entrevistadores que contiene la información disponible sobre los entrevistadores que han participado desde la EFF2014 a la EFF2022.

- **Fichero de paradata:** Registro con información sobre las pantallas del software CAPI que se utilizó durante la entrevista. Para cada hogar, contiene información detallada de la interacción del entrevistador con el ordenador durante la entrevista. En concreto, contiene el flujo de pantallas que se visualizaron en el ordenador en cada entrevista, y para cada pantalla se registra la fecha y hora en la que entró, si volvió a la pantalla anterior, o el tiempo que estuvo en cada pantalla.

## Recopilación de datos

Como acabamos de comprobar, en la EFF hay, por un lado, información sobre las entrevistas a los hogares repartida en varios ficheros para cada ola en la que han participado; y por otro lado, información sobre los entrevistadores que han colaborado en la EFF recogidas en un censo. La unidad de análisis de este estudio son hogares que ya han participado al menos una vez en alguna edición de la EFF, y que son elegibles para ser entrevistados en la siguiente edición. Por tanto, para cada ola se selecciona sólo a los hogares que han participado como máximo en tres ediciones, ya que los que han participado en la cuarta son eliminados de la muestra de cara a la siguiente edición. Los datos de una ola en concreto se utilizan como variables explicativas en los modelos, y como variable objetivo se extraerá el estatus de participación de ese mismo hogar del fichero de contactos de la ola siguiente. La vinculación de la información entre ambas ediciones se realiza a través de un identificador común que tienen los hogares en los ficheros de datos de ambas olas.

Por otro lado, la vinculación con la información del censo de entrevistadores se realiza a través de un identificador único que tiene cada entrevistador y que permanece inalterable a lo largo de las olas. Para cada hogar se conoce el identificador de la persona que realizó la entrevista en cada una de las olas en las que participó.

## 4.2. Metodología

Tomando como referencia las estrategias de investigación propuestas en Oates et al. (2022), en este proyecto se presenta un 'Caso de estudio'. Se busca tener un conocimiento profundo sobre la participación de los hogares panel en el caso específico de la EFF, y buscar maneras de predecirla. Para el ejercicio de predicción, se adapta la implementación de un modelo de predicción basado en algoritmos de machine learning hecha por Beste et al. (2023) al caso de la EFF, y comprobar si se obtienen resultados similares.

La metodología de este proyecto se fundamenta en cuatro pilares. En primer lugar, en la recopilación de los datos de las entrevistas y los entrevistadores que participaron en las olas EFF2017, EFF2020 y EFF2022. El segundo pilar es la realización de un análisis descriptivo de un conjunto de variables con potencial para explicar la Attrition. En tercer lugar, en el entrenamiento de varios modelos basados en métodos de machine learning con datos de la EFF2017 y la EFF2020, y la evaluación de su rendimiento para predecir la participación de los hogares panel en la EFF2022. Finalmente, se realiza una valoración de los resultados obtenidos y qué pasos podrían darse en el futuro para seguir desarrollando este proyecto.

El resto de este apartado se divide en tres partes. Es primer lugar se definen las variables dependientes y explicativas de los modelos de predicción. A continuación se da una breve descripción del proceso de recopilación de datos. Finalmente, se cierra el apartado y el capítulo con la descripción de la estrategia de entrenamiento y validación de los modelos.

### 4.2.1. Variable Attrition y variables explicativas

Se busca predecir si un hogar que ha participado en la ola $t$ de la EFF no volverá a participar en la siguiente ola $t+1$. La variable a predecir, o dependiente, es una variable dicotómica *Attrition* que toma valor 0 si el hogar vuelve a participar, y valor 1 si no vuelve a participar:

$$Attrition = \begin{cases} 0 & \text{el hogar participa en la ola } t+1 \\ 1 & \text{el hogar no participa en la ola } t+1 \end{cases} \tag{4.1}$$

Para predecir *Attrition* en la ola $t+1$, se utiliza como predictores o variables independientes un conjunto de variables recopiladas durante la ola anterior, la ola $t$. La selección de estas variables se ha inspirado en la realizada por Beste et al. (2023) y Kern et al. (2021), y adicionalmente se han incluido otras variables específicas que se consideran de interés para la EFF, como por ejemplo si un hogar se ha recontactado durante la revisión. La selección final de 57 variables explicativas utilizadas en este estudio puede consultarse en el cuadro 4.1.

### 4.2.2. Algoritmos de Machine Learning: Entrenamiento, validación y test

La estrategia de entrenamiento y metodología aplicada en este proyecto sigue la aplicada en Beste et al. (2023). Para la evaluación de los modelos de panel attrition, el procedimiento habitual consiste en la elección de una regresión logística (Logit) como modelo base, y se realiza una comparación del rendimiento de cada modelo de machine learning con este modelo base (Lepkowski et al. (2002), Kern et al. (2021), Beste et al. (2023)). En la implementación de este documento se toma como modelo base una regresión logística de efectos primarios (es decir, sin interacciones entre variables).

Los modelos de machine learning que se evaluarán son un árbol de decisión (CART), un Random Forest (RF), un eXtreme gradient boosting (XGBOOST) y un Naive Bayes (NB). La elección de estos modelos se debe a los buenos resultados que han mostrado para predecir panel attrition en otros estudios (Kern et al. (2019), Kern et al. (2021), Beste et al. (2023)), y a que son fácilmente interpretables en comparación con otros algoritmos de machine learning, como pueden ser las máquinas de soporte venctorial(SVM) o las redes neuronales. Esto facilita que sean utilizados en diseños adaptativos.

Con respecto al proceso de entrenamiento y evaluación, a la hora de elegir los conjuntos de entrenamiento, validación y test se ha tenido en cuenta el uso que se querría dar al algoritmo en la práctica. Siguiendo la filosofía de los diseños adaptativos, se quiere predecir con antelación qué hogares tienen más probabilidad de abandonar el estudio en la ola siguiente, y utilizar esa información para diseñar algún mecanismo dirigido a esos hogares. Eso implica establecer un criterio de separación temporal entre los conjuntos de entrenamiento y test, y asegurar que en los predictores de los modelos no se incluye información desconocida de antemano. En ese sentido, siguiendo la idea propuesta en Beste et al. (2023), utilizaremos exclusivamente información disponible en una ola concreta para predecir el resultado de la participación en la siguiente. Por esa razón, para el entrenamiento y la validación de los modelos, se utiliza la información recopilada en la EFF2017 para predecir la participación en EFF2020, y para el ejercicio del test se utiliza la información recopilada en la EFF2020 para predecir la participación en la EFF2022.

Para el tuning de los hiperparámetros de los modelos, se sigue una estrategia de k-fold cross-validation para los conjuntos de entrenamiento y validación, con k=5. Con respecto a los valores específicos de los hiperparámetros, como no existe certeza sobre cuáles son los valores más adecuados, se considera un rango bastante amplio para cada hiperparámetro de todos los algoritmos. Para los modelos Logit y CART se utiliza un algoritmo de búsqueda de red (grid search) para probar cada combinación de hiperparámetros. Con respecto a los algoritmos RF y XGBOOST, dados los elevados tiempos y costes de ejecución de cada uno, se utiliza un algoritmo

de búsqueda aleatoria (random search) con 2500 iteraciones. Aunque no se prueben todos los valores de los hiperparámetros, se ha comprobado que la búsqueda aleatoria es una alternativa válida en contextos de alta dimensionalidad de hiperparámetros (Bergstra and Bengio (2012)). Finalmente, a la hora de seleccionar las mejores combinaciones de modelos, la métrica que se busca maximizar es la ROC AUC.

| Fuente de información | Variables |
|---|---|
| Características del hogar | Número de adultos con empleo, Número de adultos jubilados, Propietario vivienda principal, Tamaño del hogar, Tiene otras propiedades, Tiene joyas, Posee negocios, Posee cuentas para pagos, Posee acciones que cotizan, Posee acciones que no cotizan, Posee renta fija, Posee fondos de inversión, Posee cuentas para no pagos, Posee planes de pensiones, Les deben dinero, Poseen vehículos, Percentil de renta, Percentil de riqueza Bruta, Pareja vive en el hogar, Poseen otros activos financieros, Tienen deuda, Tienen ingresos de activos, Hijos viven en el hogar, Nivel de satisfacción con la vida |
| Características PR | Nivel de satisfacción con la vida, Es panel, Nivel educativo, Estado de salud, Edad, Estado Civil - Casada, Estado Civil - Viuda, Sexo, Situación laboral - Asalariado, Situación laboral - Jubilado, Situación laboral - Inactivo |
| Valoraciones FI | Recelo tras la entrevista, Edificio Unifamiliar, Barreras - portero automático, Barreras - Sin barreras, Entendimiento de las preguntas por parte de PR, Interés de PR, Razones colaborar - Interesado en estos estudios, Razones para colaborar - Lo lleva el BdE, Razones para colaborar - Relevancia de la encuesta, Razones - Favor al entrevistador, Razones - Dar su opinión, Razones - Otras |
| Paradata | PR consiente grabar la entrevista, Hogar recontactado, Número de olas en las que se ha participado, Tamaño del municipio, Entrevista con proxy, Número de miembros del hogar que participan en la entrevista, Porcentaje de preguntas monetarias respondidas en valor puntual, Porcentaje de preguntas monetarias respondidas (incluye intervalos), Número de variables con valores missing, Duración de la entrevista en segundos |

Cuadro 4.1: Selección de variables para entrenar los modelos de predicción

# Capítulo 5

# Resultados

## 5.1. Análisis exploratorio de los datos

El objetivo del análisis exploratorio de datos es investigar las características de los datos que se van a utilizar. Por un lado se observan las características particulares de cada variable, y por otro las relaciones que existan entre ellas. Esta información es importante porque ayuda a identificar y tratar rasgos de las variables que pueden afectar a los modelos de machine learning que se quieren entrenar.

Como veremos a continuación, en este proyecto se ha manejado una gran cantidad de datos de gran diversidad de origen y formato. El análisis exploratorio de toda esa información es muy amplio y no es posible incluir todo en esta sección. Por esa razón, sólo se muestran resultados que son de interés para el análisis del panel attrition, o resultados que han ayudado a la toma de decisiones para la selección o transformaciones de variables.

### Número de registros y variables

El Cuadro 5.1 contiene el número de registros y variables disponibles en cada uno de los ficheros utilizados en este proyecto. Se utilizan ficheros de las olas de la EFF2017, EFF2020 y EFF2022[1]. Estos números contienen todos registros y variables que se han recogido durante la producción de datos de la EFF, y por tanto incluyen a hogares que no son objeto de este estudio, como por ejemplo hogares que nunca han llegado a participar en la EFF u hogares que han participado por primera vez en la EFF2022. Tras realizar el filtrado de hogares elegibles para el estudio (panelistas de 2017 y 2020 que pueden volver a participar en la respectiva ola

[1]De la EFF2022 sólo se utiliza el fichero de contactos ya que es el que contiene la información sobre la participación de los hogares panel.

siguiente), obtenemos que los hogares elegibles para la EFF2020 son 5,937[2] y los hogares elegibles para la EFF2022 son 5,505. Con respecto a los entrevistadores, en la EFF2017 participaron 69 entrevistadores. En la EFF2020 participaron 65 entrevistadores, de los cuales 25 también participaron en la EFF2017.

| | **EFF2017** | | **EFF2020** | |
|---|---|---|---|---|
| **Nombre del fichero** | **Registros** | **Variables** | **Registros** | **Variables** |
| Fichero de trabajo | 6,413 | 6,103 | 6,313 | 6,497 |
| Fichero de datos imputados | 6,413 | 659 | 6,313 | 787 |
| Fichero de contactos | 14,456 | 640 | 15,457 | 636 |
| Fichero de revisión | 44,760 | 22 | 35,217 | 51 |
| Fichero paradata | 2,807,091 | 13 | 3,121,437 | 12 |

| | **EFF2022** | |
|---|---|---|
| | **Registros** | **Variables** |
| Fichero de contactos | 15,182 | 636 |

| **Censo de entrevistadores** | |
|---|---|
| **Registros** | **Variables** |
| 260 | 56 |

Cuadro 5.1: *Número de registros y variables de los ficheros de datos*

Con respecto al número de variables, el cuadro 5.1 puede observarse que hay ficheros que tienen más de 6,000 variables. Sin embargo, la inmensa mayoría de estas variables almacenan las respuestas a las preguntas del cuestionario. La gran mayoría de las preguntas sólo se plantean si el hogar posee ciertos activos o si está formado por varios miembros. Esto hace que la mayoría de preguntas no tengan datos para todos los hogares, por lo que es posible seleccionar sólo aquellas que sí tengan datos. Esto reduce considerablemente el número de variables para manejar. Por otro lado, muchas de estas variables en su estado original no son informativas y necesitan ser combinadas con otras para poder obtener información interpretable, lo cual también reduce el número de variables a utilizar. También hay variables que están almacenadas en varios ficheros de datos. Por ejemplo, todas las variables que aparecen en el fichero de datos imputados también aparecen en el fichero de trabajo, por lo que también es posible eliminar variables duplicadas. Del fichero de datos imputados se usan las variables con los valores missing imputados, y el fichero de seleccionan los indicadores de calidad de los datos, indicadores de no-respuesta y otras variables de interés que no aparecen en el fichero de datos imputados. Finalmente, para que los modelos de predicción pudieran aplicarse tanto para datos de la EFF2017 como

[2]Originalmente se identificaron 5938 hogares de la EFF2017 elegibles para la EFF2020. Pero se detectó que uno de esos hogares no tenía registros en el fichero de paradata, y se decidió eliminarlo del conjunto final de datos.

para la EFF2020, sólo se seleccionaron las variables que estaban disponibles en ambas olas. Esto implicó dedicar bastante tiempo a revisar la codificación de cada variable seleccionada, y homogeneizarlas.

## Variable target: *Attrition*

En el cuadro 3.1 puede observarse que, en las olas de 2020 y 2022, la mayoría de los hogares panel volvieron a participar. Es importante destacar que, aunque en la EFF2020 parezca que existe sólo un ligero desbalance entre las observaciones de cada clase de Attrition, durante los primeros entrenamientos y evaluaciones de los modelos se comprobó que las predicciones se asignaban masivamente a la clase mayoritaria ($Attrition = 0$). Para evitar esto, y dado el tamaño limitado de la muestra, previo al entrenamiento se procede al balanceo de la clase Attrition realizando un sobre-muestreo de dicha clase usando el algoritmo SMOTE (Syntetic Monirity Over-sampling Technique, Chawla et al. (2002)).

| Attrition | EFF2020 | EFF2022 |
|---|---|---|
| Participa | 3830 | 3974 |
| No participa | 2107 | 1531 |

Cuadro 5.2: *Distribución de Attrition en EFF2020 y EFF2022*

El análisis exploratorio de las variables es fundamental para poder ver cómo son las distribuciones de las variables que se van a utilizar, las relaciones que hay entre ellas, y especialmente la que hay con la variable de interés. También ayuda a detectar posibles errores o particularidades que puedan afectar a los modelos de predicción. Dada la enorme cantidad de variables que se han analizado en este proyecto, en este apartado sólo se presentan algunos resultados que han llamado bastante la atención y se considera que merecen la pena ser nombrados porque ofrecen conocimiento sobre la participación de los panelistas en futuras olas y que podría ser útil en la práctica.

En la figura 4.1 presenta cuatro gráficos de mekko que presentan las propociones de hogares panel que vuelven a participar (región azul) frente a los que no vuelven a hacerlo (región roja) para las variables de número de olas en las que se ha participado (arriba a la izquierda), el nivel de recelo después de la entrevista (arriba a la derecha), el estado de salud reportado por la PR (abajo a la izquerda) y la edad de la PR (abajo a la derecha).

En la figura puede observarse que la proporción de abandono de los hogares que han participado sólo una ola es mayor que la de los que han participado en más de una. Este resultado sugiere que los hogares que llevan menos tiempo en la encuesta podrían ser más complicados de retener, y que podría ser interesante hacer un análisis enfocado en este tipo de hogares.

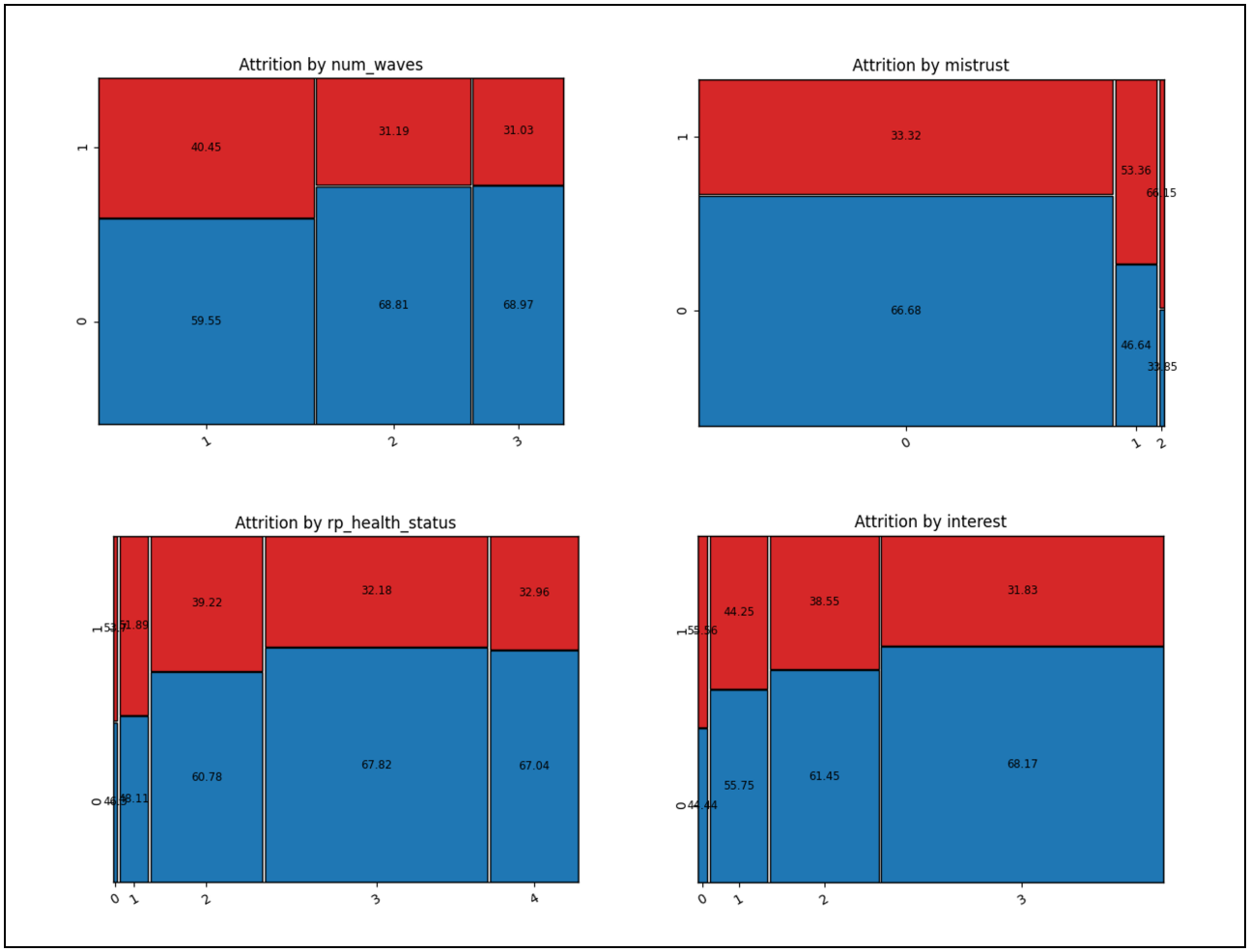


Figura 5.1: Categorical

El segundo resultado a destacar es el del recelo mostrado por el hogar tras la entrevista. Recordemos que en la EFF se hacen preguntan que pueden considerarse delicadas, como por ejemplo si se poseen joyas, o cuál es el saldo que el hogar tiene en su cuenta corriente. Puede ser comprensible que se genere recelo. Este se mide con tres niveles: nada receloso, algo receloso y muy receloso. En el gráfico se observa que la proporción de hogares que abandonan la EFF aumenta con el nivel de recelo. Este resultado es bastante obvio, pero es muy importante resaltarlo porque a un entrevistador podría resultarle muy útil saber si el hogar con el que está a punto de contactar se mostró muy receloso después de hacer la entrevista en la ola anterior, y en su estrategia de contacto y de ganar la cooperación podría hacer más énfasis en los aspectos de la encuesta que causan más recelo.

Finalmente, los dos gráficos de la parte de inferior hacen referencia a características de la PR. En el de la izquierda se observa que la proporción de hogares que abandona la encuesta es mayor cuando menor es el nivel de salud que reporta la PR. De nuevo, esto puede parecer algo obvio porque alguien que tiene peor salud seguramenten o quiera gastar su tiempo en realizar una encuesta. Pero, de nuevo, la estrategia de contacto y de ganar la cooperación de un entrevistador puede adaptarse si ya saben que van a ver a alguien que seguramente tenga mala salud, ya sea ofreciendo hacer la entrevista en varias sesiones, o en un ambiente en el que la PR se muestre más cómoda.

Finalmente, el último gráfico muestra el nivel de interés que mostró la PR durante la entrevista durante la edición anterior. Se observa que la proporción de hogares que vuelven a participar es mayor cuanto mayor es el interés que mostraron durante la ola anterior. De nuevo, se trata de un resultado obvio. Pero, al igual que con el nivel de recelo, este podría ser un dato que podría ser muy útil para el entrevistador cuando esté preparando su estategia de contacto con el hogar.

## 5.2. Evaluación de modelos

Para la selección de las variables que han entrenado los modelos se han seguido X criterios. En primer lugar, se han seleccionado variables que aparecen en las implementaciones de Kern et al. (2021) y Beste et al. (2023) en sus modelos de machine learning. En segundo lugar, se han seleccionado variables sobre factores mencionados en la revisión de Lynn (2018) y que han mostrado potencial para predecir la no participación de hogares panel, como por ejemplo la duración o información sobre los contactos con los hogares. En tercer lugar, se han seleccionado variables específicas que se recogen en la EFF, como la tenencia de activos, deudas o el recelo de los hogares percibido por los entrevistadores. Finalmente, tras seleccionar todas las variables anteriores, se han realizado diversos tests gráficos y estadísticos para detectar variables con

correlaciones o dependencias muy altas, y se han descartado. La selección final de variables es la que aparece en el cuadro 4.2.

El cuadro 4.3 contiene los resultados del rendimiento de los modelos entrenados sobre el conjunto de test, es decir, los resultados para predecir la no participación de los panelistas en la EFF2022. Las métricas de evaluación que se utilizan son Accuracy, Precision, Recall, F1 y ROC AUC. La métrica de referencia que utilizamos para la evaluación es la ROC AUC, que se encuentra en la parte derecha del cuadro. Esta métrica mide el rendimiento entre la tasa de falsos positivos y falsos negativos. Toma valores de 0.5 a 1, con 1 siendo un predictor perfecto, y 0.5 el que se obtendría con una estimación realizada de manera aleatoria.

El modelo de referencia de este estudio, el Logit, presenta una ROC AUC de 0,5959. Se considera que un valor inferior a 0,6 es un resultado malo, por lo que el modelo Logit no es un buen predictor. Con respecto a los otros modelos, observamos que el CART y el Naïve Bayes presentan valores más bajos que el Logit, de 0,5521 y 0,5798 respectivamente. El Random Forest y el XGBooster, en cambio, presentan valores de 0,5821 y 0,5911 respectivalente, que son ligeramente superiores al del modelo Logit.

De estos resultados podemos ver que los modelos de Random Forest y XGBooster mejoran al Logit de referencia. Y entre estos dos, el Random Forest presenta valores un poco más altos en Accuracy y en Precision, y el XGBooster es mejor en Recall y F1. Aun así, es necesario recalcar que el valor de ROC AUC sigue estando por debajo de 0,6, por lo que, aunque se mejore el rendimiento con respecto al modelo Logit, los resultados de los test son malos.

Estos resiltados indican que hay modelos de machine learning que mejoran en rendimiento de la predicción con respecto a un Logit. Sin embargo, las métricas señalan que el rendimiento del mejor modelo es regular. La pregunta que hay que preguntarse entonces es, ¿por qué no está funcionando bien la predicción?

Una de las causas puede ser el overfitting, es decir, que los modelos no son capaces de generalizar bien porque se han adaptado demasiado bien a los datos con los que han sido entrenados, y por tanto su rendimiento no es bueno cuando se les presenta con datos nuevos. una manera de comprobar esto es haciendo el ejercicio predicción sobre los datos de entrenamiento con estos mismos modelos, y ver cómo son esos resultados. Estos resultados pueden verse en el cuadro 4.4.

Como era de esperar, los resultados de las predicciones de los modelos sobre los datos de entrenamiento son mejores que los que se observan con respecto a los datos test. En este caso, el modelo que presenta mejor rendimiento es el Random Forest, con una métrica de ROC AUC de 0,6726, ligeramente superior a la del XGBooster. El resto de métricas del Ranfom Forest también son ligeramente mejores que las del XGBooster. Sin embargo, esta métrica de ROC AUC está entre 0,6 y 0,75, que lo clasifica como un test regular.

| Fuente de información | Variables |
|---|---|
| Características del hogar | Número de adultos que trabajan, Número de adultos jubilados, Propietario vivienda principal, Tamaño del hogar, Tiene otras propiedades, Tiene joyas, Posee negocios, Posee cuentas para pagos, Posee acciones que cotizan, Posee acciones que no cotizan, Posee renta fija, Posee fondos de inversión, Posee cuentas para no pagos, Posee planes de pensiones, Les deben dinero, Poseen vehículos, Percentil de renta, Percentil de riqueza Bruta, Pareja vive en el hogar, Poseen otros activos financieros, Tienen deuda, Tienen ingresos de activos, Hijos viven en el hogar, Nivel de satisfacción con la vida |
| Características PR | Nivel de satisfacción con la vida, PR es panel, PR - nivel educativo, PR- Edad, PR - Casada, PR - Viuda, PR - Sexo, Situación laboral - Asalariado, Situación laboral - Jubilado, Situación laboral - Inactivo, PR - Estado de salud |
| Valoraciones FI | Recelo tras la entrevista, Tipo de edificio - Unifamiliar, Barreras - portero automático, Barreras - No hay barreras, Entendimiento de las preguntas PR, Interés PR, Razones colaborar - Interesado en estos estudios, Razones para colaborar - Lo lleva el BdE, Razones para colaborar - Relevancia de la encuesta, Razones - Favor al entrevistador, Razones - Dar su opinión, Razones - Otras |
| Paradata | PR consiente grabar la entrevista, Hogar con recontacto exitoso, Número de olas en las que se ha participado, Tamaño del municipio, Entrevista con proxy, Número de miembros del hogar que participan, Ratio estricto, Ratio amplio, Número de variables con valores missing, Duración de la entrevista |

Cuadro 5.3: Selección de variables para entrenar los modelos de predicción

| Modelo | Accuracy | Precision | Recall | F1 | ROC AUC |
|---|---|---|---|---|---|
| Logit | 0,6556 | 0,3749 | 0,3573 | 0,3659 | 0,5959 |
| CART | 0,6434 | 0,3338 | 0,2835 | 0,3066 | 0,5521 |
| Random Forest | 0,6489 | 0,3529 | 0,3148 | 0,3328 | 0,5821 |
| XGBooster | 0,6718 | 0,3772 | 0,2769 | 0,3194 | 0,5911 |
| Naive Bayes | 0,6254 | 0,3504 | 0,4063 | 0,3763 | 0,5798 |

Cuadro 5.4: Métricas de evaluación de los modelos de predicción en el conjunto de test

| Modelo | Accuracy | Precision | Recall | F1 | ROC AUC |
|---|---|---|---|---|---|
| Logit | 0,6389 | 0,4905 | 0,4518 | 0,4704 | 0,6517 |
| CART | 0,6126 | 0,4594 | 0,5178 | 0,4868 | 0,6188 |
| Random Forest | 0,6596 | 0,5223 | 0,4770 | 0,4986 | 0,6726 |
| XGBooster | 0,6544 | 0,5144 | 0,4675 | 0,4898 | 0,6722 |
| Naive Bayes | 0,6251 | 0,4741 | 0,5178 | 0,4950 | 0,6435 |

Cuadro 5.5: Métricas de evaluación de los modelos de predicción en el conjunto de entrenamiento

Continuando con el Random Forest, una ventaja de los modelos basados en árboles con respecto a otros modelos es que pueden ser interpretados. En el caso del Radom Forest, es posible consultar qué variables han tenido más peso a la hora de clasificar la participación de los hogares. Aunque los resultados de la predicción del training no sean buenos, merece la pena echarle un ojo por los patrones que haya podido detectar. El cuadro 4.5 presenta las 10 variables con más importancia en el modelo de Random Forest.

| Variable | Importancia |
|---|---|
| PR es panel | 0,0822 |
| Interés PR | 0,0790 |
| Razones para colaborar - Relevancia de la encuesta | 0,0605 |
| Ratio amplio | 0,0594 |
| Poseen vehículos | 0,0539 |
| Tienen deuda | 0,0496 |
| Situación laboral - Asalariado | 0,0475 |
| Razones colaborar - Interesado en estos estudios | 0,0380 |
| Posee planes de pensiones | 0,0353 |
| PR consiente grabar la entrevista | 0,0348 |

Cuadro 5.6: Selección de variables para entrenar los modelos de predicción

La importancia de una variable en el Random Forest mide el peso relativo que ha tenido una variable concreta a la hora de crear las ramificaciones de los diferentes árboles de decisiones que va generando el Random Forest durante su entrenamiento. En este caso de la participación en la EFF2020, las tres variables que más importancia tuvieron fueron que la PR fuera panel, que la PR mostrase interés durante la entrevista, y que el entrevitador indicase que el hogar colaboró

por la relevancia de la encuesta. También es interesante destacar la variable ratio amplio, que es un indicador de la no-respuesta del hogar a la hora de facilitar cantidades monerarias[3]. Finalmente, otra variable que es importante para hacer la clasificación es si la PR consintió que se grabase la entrevista.

[3]El ratio amplio se define como el cociente entre el número de preguntas en euros respondidas por el hogar, ya fuera como valor puntual o como intervalos, sobre el número total de preguntas planteadas. Cuando mayor es su valor, más información ha facilitado el hogar.

# Capítulo 6

# Conclusiones y futuras líneas de trabajo

En este Trabajo de Final de Master se ha intentado aplicar la implementación de Beste et al. (2023) para predecir la no participación de hogares panel en encuestas longitudinales en olas futuras al caso de la Encuesta Financiera de las Familias. Para ello, se ha usado un modelo de Regresión Logística como referencia, y se han entrenado modelos CART, Random Forest, XGBooster y Naïve Bayes. El conjunto de predictores está formado por variables que potencialmente pueden explicar la participación de un hogar en la edición siguiente de la encuesta, como por ejemplo el interés mostrado por la persona que respondió a la encuesta, el nivel de recelo que mostró durante la entrevista, su nivel de salud o si consintió que la entrevista fuese grabada.

Los modelos de Random Forest y XGBooster presentaron mejores rendimientos que el modelo de referencia Logit en todas las métricas de evaluación consideradas, pero el valor de ROC AUC del mejor de los modelos no superó el valor de 0.6, lo cual lo clasifica como un predictor malo. Para intentar aprender sobre cómo se ha hecho la predicción, se hizo el ejercicio de usar los mismos modelos para hacer la predicción con el conjunto de entrenamiento. El modelo que funcionó mejor fue el Random Forest, pero su ROC AUC no superó el valor de 0,7, que lo clasifica como un predictor regular. Al observar las variables que más importancia tuvieron durante el entrenamiento del Random Forest destabacan que la persona que contestó a la estrevista a ya formase parte del hogar desde al menos dos ediciones antes, el valor del interés que mostró también era importante, y que el entrevistador considerase que el hogar participó por la relevancia de la encuesta.

A partir de los resultados que se han visto en este proyecto, se plantean las siguientes reflexiones y los posibles pasos que se podrían dar en el futuro para este proyecto:

1. La exploración de los datos sugiere que las variables seleccionadas para entrenar los mo-

delos de predicción guardan cierta relación con la variable de attrition. Pero no son suficientes para predecir bien si un hogar panel dejará de participar en la ola siguiente. Tal y como hemos comentado al principio de este capítulo, la EFF genera una gran cantidad de variables, y se hizo un filtrado inicial basado en las implementaciones de Beste et al. (2023) y Kern et al. (2021). Es posible que haya **variables que no se han incluido, y que tengan poder para predecir el attrition**. Una vía de trabajo para el futuro es **volver a revisar toda la información disponible y plantear una nueva selección de variables**. El uso de métodos de machine learning para selección de variables es una opción que se podría implementar.

2. Las olas de la EFF seleccionadas para el estudio son las de los años 2017, 2020 y 2022 porque son las que tienen mayor cantidad de datos y también los de mayor calidad. Todos los modelos entrenados buscan utilizar datos de lo que había en 2017 para predecir algo que iba a ocurrir en 2020, que es el año en el que tuvo lugar la crisis del **Covid-19**. Y posteriormente, en el test, se utiliza información recogida durante 2020 para predecir lo ocurrido en el año 2022, cuando el mundo estaba ya superando la crisis del Covid-19. El año 2020 fue un año especial en la EFF porque seguramente mucha gente fuera más reticente a participar por el covid, lo cual afectaría a la variable target del entrenamiento). También se tuvo que cambiar la metodología de la entrevista, pasando de entrevista personal a entrevista telefónica, que afecta a la calidad de los datos (Lynn (2018)), y por tanto a los datos de los predictores usados en el test. Ante esta problemática, se plantean dos posibles alternativas:

   *a*) Las olas de **EFF2002 a EFF2017 son homogéneas** en lo que a metodología se refiere. Todas las entrevistas fueron personales, y no hubo una crisis como la del Covid-19. Una alternativa interesante sería crear **nuevos modelos de predicción**, pero utilizando sólo la información disponible en esas olas. Serían **menos variables** (las respuestas de los hogares y la información recogida por los entrevistadores), pero podría ser que el rendimiento fuese mejor.

   *b*) La EFF va a continuar realizándose en los próximos años, y con una frecuencia bienal. Se puede volver a **plantear esta misma metodología dentro unos años, utilizando sólo ediciones completadas después de 2020**, y con el beneficio de recoger toda la información que se ha estado recogiendo en las últimas ediciones y que no está disponible para antes de 2017.

3. Los resultados de las predicciones sobre los datos de entrenamiento también podría explicarse por la **existencia de información no observable en el momento de hacer la predicción, y que además tenga más peso para determinar el resultado de la**

**participación que la información recabada durante la ola anterior**. El Covid-19 es un buen ejemplo de algo que no se puede prever, pero otra cosa que también se desconoce es qué entrevistadores habrá durante la siguiente edición. En todas las ediciones hay entrevistadores nuevos, y algunos funcionan muy bien, y otros no. Tal y como comentan Lynn (2018) y Groves (2006), el papel del entrevistador es importante para la colaboración de los hogares y la calidad de los datos. Esto puede investigarse identificando a los hogares para los que no se han hecho buenas predicciones, y analizar por un lado cómo son las características que tienen en la ola anterior, y por otro la información que se tenga sobre la ola que se intenta predecir, y comprobar si las variables que tienen más peso para explicar la participación son las de la ola corriente o las de la ola anterior.

4. Una opción que siempre hay que considerar es un **cambio de enfoque**. Hay dos alternativas que son interesantes:

   *a*) En la exploración de los datos vimos que los hogares que han participado sólo en una edición muestran más proporción de abandonos que los que han participado más de dos años. Una posible explicación de esto es que los hogares que han participado más de una vez están más comprometidos con el estudio, y seguramente merezca la pena **enfocar el análisis en la predicción de la participación de los paneles en su segunda ola** en vez de hacer la predicción para todos los hogares.

   *b*) En vez de predecir un resultado binario, de participar o no participar, se puede plantear hacer un ejercicio de **análisis de supervivencia** (survival analysis), e intentar predecir el número de ediciones en las que participará un hogar de la EFF antes de abandonar el estudio. Esta información está disponible y se podría utilizar información de todas las olas de la EFF.

# Bibliografía


Alvargonzález, P., Pidkuyko, M., and Villanueva, E. (2020). La situación financiera de los trabajadores más afectados por la pandemia: un análisis a partir de la encuesta financiera de las familias. *Boletín Económico/Banco de España, 3/2020*.

Barceló, C. (2006). Imputation of the 2002 wave of the spanish survey of household finances (eff). *Banco de Espana Research Paper No. OP-0603*.

Barceló, C., Crespo, L., Garcia, S., Gento, C., Gómez, M., and de Quinto, A. (2021). The spanish survey of household finances (eff): Description and methods of the 2017 wave. *Banco de Espana Occasional Paper*, 2033.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13, No. 2.

Beste, J., Frodermann, C., Trappmann, M., and Unger, S. (2023). Case prioritization in a panel survey based on predicting hard to survey households by machine learning algorithms: An experimental study. In *Survey Research Methods*, volume 17, No. 3, pages 243–268.

Buskirk, T. D., Kirchner, A., Eck, A., and Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1).

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Couper, M. P. and Ofstedal, M. B. (2009). Keeping in contact with mobile sample members. *Methodology of longitudinal surveys*, pages 183–203.

Crespo, L., El Amrani, N., Gento, C. L., and Villanueva, E. (2023). Heterogeneidad en el uso de los medios de pago y la banca online: un análisis a partir de la encuesta financiera de las familias (2002-2020). *Documentos Ocasionales/Banco de España, 2308*.

Early, K., Mankoff, J., and Fienberg, S. E. (2017). Dynamic question ordering in online surveys. *Journal of Official Statistics*, 33(3):625–657.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *International Journal of Public Opinion Quarterly*, 70(5):646–675.

Groves, R. M., Cialdini, R. B., and Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public opinion quarterly*, 56(4):475–495.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(3):439–457.

Gómez, M. and Villanueva, E. (2022). El efecto de los planes de pensiones de empresa sobre el ahorro privado de los hogares. *Boletín Económico/Banco de España, 2/2022*.

Hirano, K., Imbens, G., Ridder, G., and Rebin, D. B. (1998). Combining panel data sets with attrition and refreshment samples.

Jankowsky, K. and Schroeders, U. (2022). Validation and generalizability of machine learning prediction models on attrition in longitudinal studies. *International Journal of Behavioral Development*, 46(2):169–176.

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. In *Survey research methods*, volume 13, No. 1, page 73. NIH Public Access.

Kern, C., Weiß, B., and Kolb, J.-P. (2021). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. *Journal of Survey Statistics and Methodology*, page smab009.

Kreuter, F. and Müller, G. (2015). A note on improving process efficiency in panel surveys with paradata. *Field Methods*, 27(1):55–65.

Laurie, H., Smith, R. A., and Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15(2):269–282.

Lepkowski, J. M., Couper, M. P., et al. (2002). Nonresponse in the second wave of longitudinal household surveys. *Survey nonresponse*, pages 259–272.

Liu, M. (2020). Using machine learning models to predict attrition in a survey panel. *Big data meets survey science: A collection of innovative methods*, pages 415–433.

Lynn, P. (2009). Methods for longitudinal surveys. *Methodology of longitudinal surveys*, pages 1–19.

Lynn, P. (2018). Tackling panel attrition. *The Palgrave handbook of survey research*, pages 143–153.

Mcgonagle, K. A., Sastry, N., and Freedman, V. A. (2022). The effects of a targeted "early bird" incentive strategy on response rates, fieldwork effort, and costs in a national panel study. *Journal of Survey Statistics and Methodology*, page smab042.

Oates, B. J., Griffiths, M., and McLean, R. (2022). *Researching information systems and computing*. Sage.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schouten, B., Peytchev, A., and Wagner, J. (2017). *Adaptive survey design*. CRC Press.

Tourangeau, R., Michael Brick, J., Lohr, S., and Li, J. (2017). Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(1):203–223.

Wagner, J. R. (2008). *Adaptive survey design to reduce nonresponse bias*. PhD thesis, University of Michigan.

Watson, N. and Wooden, M. (2009). Identifying factors affecting longitudinal survey response. *Methodology of longitudinal surveys*, pages 157–181.