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: Ismael Benito Altamirano

Madrid, 7 de enero de 2024

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

El abandono de panelistas en encuestas longitudinales, también conocido como Panel Attrition, es un asunto preocupante porque puede sesgar y afectar la eficiencia de los resultados de estas encuestas. En este contexto, en las últimas décadas ha aumentado el uso de diseños adaptativos y reactivos. Estos diseños utilizan información de panelistas recogida en olas anteriores para desarrollar intervenciones en los procesos de creación de datos con el objetivo de disminuir el Panel Attrition en olas posteriores. Los modelos de predicción basados en algoritmos de machine learning son herramientas interesantes para el diseño de estas implementaciones porque han mostrado buenos resultados para predecir Panel Attrition en encuestas.

Tomando como referencia la implementación hecha en Beste et al. (2023), este Trabajo de Fin de Máster busca desarrollar un modelo basado en algoritmos de machine learning para predecir Panel Attrition en el caso de la la Encuesta Financiera de las Familias (EFF) utilizando información de hogares panel de olas anteriores. Este proyecto se divide en tres partes. Primero, se realiza un análisis exploratorio de los datos generados durante la producción de la EFF, con especial interés en la relación entre las variables y el Panel Attrition. En la segunda parte, se entrenan cuatro modelos basados en algoritmos de machine learning con información de los hogares de la ola anterior para predecir Panel Attrition en la ola siguiente, se evalúan utilizando datos nuevos, y se compara su rendimiento con el de un modelo de Regresión Logística o Logit. Finalmente, se analizan todos los resultados y se consideran posibles líneas de trabajo para el futuro.

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

Abstract

The drop-out of panel members in longitudinal surveys, also known as Panel Attrition, is a worrying issue because it potentially bias survey estimates and affects their efficiency. In this context, during the last decades the use of adaptative and responsive designs has increased. These designs use panelist's information from previous waves to develop informed interventions in data production processes aimed at decreasing Panel Attrition in subsequent waves. Prediction models based on machine learning algorithms are interesting tools for designing these implementations because they have shown good performance at predicting Panel Attrition.

Inspired by the implementation at Beste et al. (2023), this Master Project aims to develop a model based on machine learning algorithms for predicting Panel Attrition in the case of the Survey os Spanish Household Finances (EFF) using panel household information from previous waves. This project is divided in three parts. Firstly, an exploration analysis is performed on all the data generated during the production of the EFF, with special focus on the relationship between each variable and Panel Attrition. Secondly, four machine-learning-based models are trained to predict Panel Attrition using household's information from the previous wave, then they are tested using new data, and their performance is compared with the one shown by a Logistic Regression or Logit. Finally, all results are analized and future lines of work are discussed.

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

Índice general

Capítulo 1

Introducción

Los estudios longitudinales son proyectos de investigación en los que se hace un seguimiento a un grupo de unidades muestrales (personas, hogares...) a lo largo de un período de tiempo. En el ámbito de las ciencias sociales, la recolección de datos de muchos de estos estudios se realiza mediante el uso de encuestas, de tal manera que las mismas personas responden a las preguntas del mismo cuestionario de manera repetida durante un tiempo, que pueden ser semanas, meses o incluso años. Esto da lugar a las llamadas encuestas longitudinales o encuestas panel. La dimensión temporal de estas encuestas las convierten en una herramienta muy útil para poder analizar relaciones causales, ya que permiten observar cambios en opiniones, comportamientos o estados de los mismos panelistas a lo largo del tiempo. Sin embargo, la calidad de esos análisis depende de la cooperación exitosa y continuada de dichos panelistas 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)).

Lynn (2018) destaca que el Panel Attrition presenta dos principales problemas. Por un lado, si la tasa de abandonos es alta, el tamaño de la muestra se reducirá drásticamente en pocas 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 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 de la encuesta.

Tradicionalmente, los métodos utilizados para mitigar los efectos del Panel Attrition se han centrado en el impacto estadístico que provoca, principalmente con el uso de métodos de imputación múltiple (Rubin (1987)), la reponderación 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 ha aumentado el interés por tratarlo durante

los procesos de creación y recolección de datos, lo que ha llevado a la extensión del uso de los llamados diseños adaptativos y reactivos (adaptative and responsive designs, 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. 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, se ha utilizado información sobre intentos de contacto en olas pasadas para revisar la estrategia de incentivos para hogares a los que cuesta volver a entrevistar Mcgonagle et al. (2022) o para optimizar las estrategias de contacto en las ediciones siguientes (Kreuter and Müller (2015)).

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, especialmente para predecir Panel Attrition (Buskirk et al. (2018)). Por ejemplo, en Beste et al. (2023) utilizan información de ediciones pasadas de una encuesta a hogares en alemania para entrenar varios modelos basados en algoritmos de Machine Learning para identificar hogares panelistas con una baja propensión a participar en la edición siguiente. Posteriormente, utilizaron el mejor de esos modelos para predecir qué hogares panelistas que iban a participar en la nueva edición tenían menor probabilidad de colaborar de nuevo, y utilizaron esas predicciones para crear un diseño experimental enfocado en esos hogares.

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 al caso de estudio de la Encuesta Financiera de las Familias (EFF). La EFF es una encuesta a hogares representativa de los hogares que residen en España, y es creada por el Banco de España. La idea central del ejercicio de predicción consiste en utilizar información de hogares en olas pasadas de la EFF para predecir una variable binaria de participación o no participación en la ola siguiente. Para ello, se entrenan cuatro modelos basados en algoritmos de machine learning y se compara su rendimiento con un modelo de referencia utilizado tradicionalmente en análisis de Panel Attrition, que en este caso es una Regresión Logística o Logit. En este proyecto, para el entrenamiento de los modelos se utiliza información sobre hogares que han participado en las olas 4, 5, 6 y 7 de la EFF (que se corresponden con los años 2011, 2014, 2017 y 2020), y a continuación se utilizan dichos modelos para predecir la participación de los hogares panel elegibles para la ola 8 (que se corresponde con el año 2022).

Para poder realizar estos ejercicios de predicción ha sido necesario realizar un extenso y muy laborioso trabajo de exploración de los datos y transformación de variables. Durante el

desarrollo de estas tareas se han encontrado resultados muy interesantes y que ofrecen ideas para poder mejorar los procesos de creación de datos. Por esa razón, dentro del capítulo de resultados se incluye un apartado adicional en el que se comentan los resultados más destacables de este proceso de exploración de datos.

Los cinco capítulos restantes de este documento se organizan de la siguiente manera. En el siguiente capítulo se exponen los objetivos, la planificación original de este proyecto y las motivaciones personales para realizarlo. En el tercer capítulo se describe cuáles son las causas del Panel Attrition, y cómo se han utilizado algoritmos de machine learning para predecirlo. En el cuarto capítulo se describen por un lado los datos de la EFF y el proceso de producción que los genera, y por otro la metodología utilizada para el análisis exploratorio de datos y para el ejercicio de entrenamiento, validación y test de los modelos de predicción. En el quinto capítulo se comentan los resultados del análisis exploratorio de los datos y los resultados de la evaluación de los modelos de predicción. Finalmente, en el sexto capítulo se exponen las conclusiones y reflexiones sobre el proyecto y cuáles pueden ser las líneas de trabajo para 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.

	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

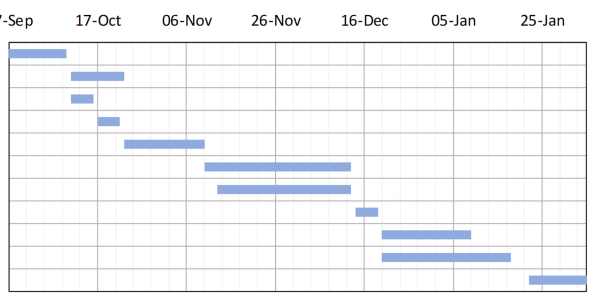

Figura 2.1: Planificación de las actividades del TFM

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 ha crecido mi interés por la metodología de encuestas y el potencial que tienen para recoger 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 eso 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 encontrar maneras de mejorarlo.

Por otro lado, en los últimos años ha aumentado cantidad de datos que se generan durante la encuesta y también su variedad, tomando espacecial relevancia los audios de las entrevistas y los comentarios de texto escritos por los entrevistadores, que se han convertido en herramientas fundamentales de los procesos de revisión de calidad de los datos. Los métodos y las herramientas desarrolladas en el campo de la ciencia de datos abren un mundo de posibilidades para poder analizar y explotar toda esa información.

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

Estado del Arte

3.1. Causas del Panel Attrition

Tal y como hemos mencionado en la introducción, los métodos adaptativos y reactivos buscan implementar intervenciones en los procesos de creación de datos de encuestas. Para poder implementarlas para reducir el abandono de panelistas, es importante conocer qué provoca el abandono de los hogares panelistas. 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 localizar exitosamente a un encuestado durante una ola posterior. Generalmente, esto se debe a cambios en la información de contacto (dirección de residencia, número teléfono, correo electrónico...) obtenida del participante durante la ola anterior (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 e incluso el presupuesto (Lynn (2009)). Por ejemplo, las encuestas con entrevistas cara a cara utilizan métodos de rastreo y búsqueda que suelen ofrecer altos índices de localización y cooperación (De Leeuw et al. (2005), Couper and Ofstedal (2009)), pero también requieren esfuerzos adicionales para localizar a los participantes panel que se hayan mudado, como por ejemplo reembolsar a los entrevistadores los gastos derivados del proceso de búsqueda.

La segunda causa es el no-contacto. Tras localizar exitosamente al encuestado, es necesario establecer un contacto. Para que sea 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 la persona que quiere entrevistarle hace la visita al hogar. Algo parecido sucede con las entrevistas telefónicas,

ya que el panelista debe tener su teléfono accesible cuando se realiza la llamada para intentar el contacto. Tradicionalmente, dos estrategias de contacto que han presentado buenos resultados han sido utilizar un número de intentos de contacto alto y diversificado en horarios (mañana, tarde, fines de semana...), y establecer períodos para realizar entrevistas lo suficientemente largos (Nicoletti and Peracchi (2005), Watson and Wooden (2009)).

Finalmente, tras 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 poseen una experiencia previa por haber participado en ediciones pasadas. Esto hace que sea relativamente probable que vuelvan a hacerlo, en parte por ser consistentes con su comportamiento de conformidad mostrado anteriormente (Groves et al. (1992)), pero también puede potenciar el efecto negativo de factores como la duración de la entrevista, la carga cognitiva que supone pensar algunas respuestas o la fatiga por haber participado en varias ediciones anteriores (Laurie et al. (1999), Watson and Wooden (2009), Lynn (2018)).

En el siguiente apartado se comenta cómo la disponibilidad de información sobre todas estas causas puede ser aprovechada por los algoritmos de machine learning para crear modelos que ayuden a predecir Panel Attrition, y posibilitar la implementación de diseños adaptativos y reactivos con el objetivo de reducir el abandono de hogares panelistas en encuestas longitudinales.

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 además destacan por su facilidad para ser interpretados.

Otro ejemplo prometedor, y que además sirve para implementar un diseño adaptativo, puede verse 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. Los 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 particular 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 Panel Attrition. 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 proyecto 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. Su cuestionario principal está compuesto por nueve secciones. Las secciones 1 y 6 recogen información sociodemográfica individualizada de todos los miembros del hogar y también información individualizada sobre la situación laboral y los ingresos de cada miembro del hogar mayor de 16 años. El resto de secciones recogen información detallada sobre los activos, deudas, gastos y uso de medios de pago del hogar en su conjunto, y contienen módulos específicos que recogen información individualizada de ciertos elementos patrimoniales si el hogar los posee, como propiedades inmobiliarias, negocios gestionados por cuenta propia o planes de pensiones. Los hogares formados por más miembros mayores de edad y que posean más elementos patrimoniales responderán a más preguntas. Por poner un ejemplo con números, en la EFF2017 a cada hogar se le plantearon entre 137 y 594 preguntas, siendo la mediana de 259 preguntas (Barceló et al. (2021)).

La EFF 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 los hogares es asimétrica, por lo que sólo unos pocos hogares, especialmente los más ricos, son los que invierten en ciertos activos. Posteriormente, para poder hacer análisis válidos para la población española, a cada hogar se le asigna un peso muestral que indica a cuántos hogares de la población española representa ese hogar[2]. 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. Es importante conocer las etapas de este proceso de producción para poder entender 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.

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:** Se envía una carta a los hogares para informar sobre su

[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).

[2] Para más detalle sobre el cálculo de los pesos muestrales, consultar Bover (2004).

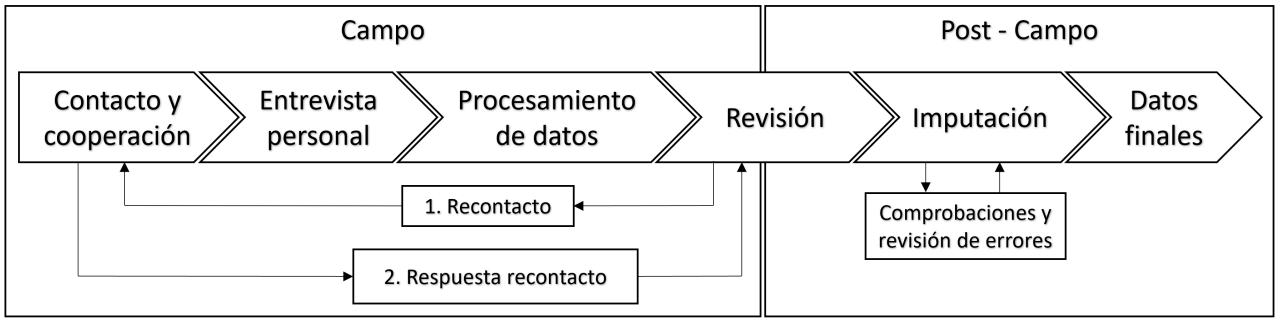

Figura 4.1: Fases de la creación de datos de la EFF

selección para la encuesta y que una persona les visitará personalmente en su domicilio para hacerles una entrevista personal. El contacto puede requerir varias visitas, 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. Los entrevistadores también llevan documentación sobre la encuesta que pueden utilizar para intentar convencer al hogar, como por ejemplo un folleto con noticias de prensa en las que se ha hablado de la EFF.

- **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. Las entrevistas suelen durar entre una hora y hora y media (Barceló et al. (2021)). 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[3].
- **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 las variables que contienen no-respuesta se imputan mediante técnicas de imputación múltiple[4]. 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[5], 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 [6]. 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 utilizan durante este proyecto son:

- **Fichero de trabajo:** Registro de hogares con entrevistas válidas. Hay uno para cada

[3]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.

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

[5]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.

[6]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.

edición de la encuesta. 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. Se utilizan dos ficheros, el de la EFF2017 y el de la EFF2020.

- **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. Hay cinco ficheros para cada edición, pero por simplicidad sólo se utiliza uno de los conjuntos de datos de la imputación multiple. Sólo contiene variables imputadas e indicadores de las características del hogar. Se utilizan dos ficheros, los de EFF2017 y EFF2020.
- **Fichero de contactos:** Registo de hogares contactados durante una ola de la EFF. Hay uno para cada 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. Se utilizan tres ficheros, los de EFF2017, EFF2020 y EFF2022.
- **Fichero de revisión:** Registro de incidencias durante el proceso de revisión y los recontactos. hay uno para cada edición de la EFF. 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). Se utilizan dos ficheros, los de EFF2017 y EFF2020.
- **Fichero de paradata:** Registro de las pantallas visualizadas durante el uso del software CAPI durante cada entrevista. Hay uno para cada edición de la EFF. Cada registro es una pantalla visualizada durante una entrevistacada entrevista y se recoge cómo fue la interacción del entrevistador con el ordenador en esa pantalla durante la entrevista. En concreto, contiene la fecha y hora en la que se cargó la pantalla, si se pasó a la pantalla siguiente (se dió una respuesta), el tiempo que se estuvo visualizando dicha pantalla, si se volvió a la pantalla anterior o incluso si se seleccionó parar la entrevista en esa pantalla. Si en una entrevista se pasó varias veces por una pantalla, esa pantalla tendrá tantos registros como veces se cargó esa pantalla. Es posible agrupar los registros por entrevista, pregunta o sección, y ver el flujo de pantallas que se siguió durante la entrevista. Se utilizan dos ficheros, los de EFF2017 y EFF2020.
- **Censo de entrevistadores:** Registro de entrevistadores que contiene la información disponible sobre los entrevistadores que han participado desde la EFF2014 a la EFF2022. Sólo hay un fichero.

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 un modelo para predecirla.

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 evaluar su rendimiento para predecir la participación de los hogares panel en la EFF2022. Para este ejercicio se adapta la implementación hecha en Beste et al. (2023) al caso de la EFF. Finalmente, se realiza una valoración de los resultados obtenidos y qué pasos podrían darse en el futuro para seguir desarrollando este proyecto.

El tratamiento de datos y el análisis descriptivo en este proyecto se ha realizado con lenguaje Python (3.8.8), y los paquetes scikit-learn (1.2.2, Pedregosa et al. (2011)), imbalanced-learn (0.10.1, Lemaître et al. (2017)) y xgboost (1.5.2, Chen and Guestrin (2016)) para la construcción de los modelos de machine learning.

El resto de esta sección se divide en tres apartados. Es primer lugar se da una breve descripción del proceso de recopilación de los datos y el análisis descriptivo. En el siguiente apartado se define la variable dependiente *Attrition* y también las variables que se van a utilizar como predictores de los modelos. Finalmente, se cierra este capítulo con la descripción de la estrategia de entrenamiento y validación de los modelos.

Recopilación de datos y análisis descriptivo

En este proyecto hay que vincular información de doce ficheros de datos. La vinculación se realiza a través de dos identificadores, uno a nivel de hogar y otro a nivel de entrevistador. Cada hogar que participa en una ola concreta de la EFF tiene un identificador único para esa edición, y ese identificador es común para todos los ficheros de datos generados durante esa edición. Esto permite vincular los datos del mismo hogar entre los diferentes ficheros de datos de una edición. Además, si ese hogar es panelista, los ficheros de trabajo, datos imputados y contactos también incluyen el identificador único que tiene ese hogar en la ola anterior. Esto permite hacer la vinculación con la información de ese hogar recogida en ediciones anteriores de la EFF.

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. Este indicador aparece como una variable adicional en el fichero de trabajo de cada ola, por lo que es posible identificar a la persona que realizó cada entrevista en cada una de las olas en las que participó. Esto además permite vincular los ficheros de hogares con los de entrevistadores.

Para realizar el análisis descriptivo, se han realizado tres tipos de tareas:

1. Exploración de las estructuras de datos de los ficheros. Esto ha permitido identificar los diferentes tipos de registros y la cantidad, tipología y codificación de las variables.
2. Análisis de distribuciones individuales de todas las variables mediante cálculos de estadísticos descriptivos y representaciones visuales mediante histogramas y gráficos box-plot para las variables numéricas, y gráficos de columnas para las variables categóricas.
3. Análisis de las relaciones entre las variables, y de manera particular con el Panel Attrition. Para analizar la relación entre variables continuas se han utilizado printipalmente gráficos de nubes de puntos y se han calculado los coeficientes de correlación entre cada variable numérica. Las relaciones entre variables categóricas se han realizado mediante el uso de gráficos de columnas, gráficos de mekko y test de independencia chi-cuadrado. Finalmente, la relación entre variables categóricas y numéricas se ha realizado visualmente con histogramas y box-plots de las variables continuas para cada categoría, y un análisis ANOVA en los casos que se cumplían los supuestos para este análisis.

Variable *Attrition* y predictores

El objetivo del estudio es encontrar un modelo para predecir si un hogar que ha participado en la ola t de la EFF no volverá a participar en la siguiente ola $t+1$. Para ello, se define la variable *Attrition* como una variable dicotómica que toma valor 0 si el hogar vuelve a participar, y valor 1 si no vuelve a participar:

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

En el cuadro 4.1 puede observarse que, en las olas de 2020 y 2022, la mayoría de los hogares panel volvieron a participar. Esto provoca un ligero desbalance entre las observaciones de cada clase de Attrition, y durante los primeros entrenamientos y evaluaciones de los modelos se comprobó que las predicciones se asignaban masivamente a la clase mayoritaria, en este caso $Attrition = 0$.

Para evitar esto, y dado el tamaño limitado de la muestra, de manera previa se implementa una técnica de sobre-muestreo sobre la clase $Attrition = 1$ para balancear la muestra y obtener

Attrition	**EFF2020**	**EFF2022**
Participa (attrition = 0)	3830	3974
No participa (attrition = 1)	2107	1531

Cuadro 4.1: *Distribución de Attrition en EFF2020 y EFF2022*

el mismo número de instancias de cada clase de *Attrition*. El sobre-muestre se realiza con el algoritmo SMOTE (Syntetic Monirity Over-sampling Technique, Chawla et al. (2002)), que es un algoritmo que genera nuevas instancias de la clase minoritaria basándose en el algoritmo de vecinos más cercanos.

En el otro lado de un modelo de predicción se encuentran las variables que se eligen como predictores. Por definición, estas variables deben contener información conocida y observable antes de que se producza el fenómeno que se quiere predecir. En ese sentido, para predecir *Attrition* en $t+1$ se utilizan como predictores variables que se recogieron durante la ola anterior, la ola t. La selección concreta 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 de interés para la EFF, como por ejemplo si un hogar se ha recontactado durante la revisión o si la entrevista se realizó con un proxy. La selección final de 57 variables explicativas utilizadas en este estudio puede consultarse en el cuadro 4.2.

Entrenamiento, validación y test de modelos de machine learning

La estrategia de entrenamiento y validación utilizada en este proyecto se inspira en la implementada en Beste et al. (2023). En ese trabajo utilizan la información de las olas 4 a 12 de una encuesta a hogares para entrenar y validar sus modelos de machine learning, y la ola 13 para evaluar el rendimiento de dichos modelos. Como predictores utilizan la información de la encuesta inmediatamente anterior, y como información histórica utilizan el número total de olas en las que se ha participado y la proporción de olas en las que se ha participado desde que el hogar fue elegido para la muestra[7]. Para la implementación de este proyecto se utiliza la información de la EFF2017 y la participación en EFF2020 para el entrenamiento y la validación de los modelos, y para evaluación se utiliza la información de la EFF2020 para predecir la participación en la EFF2022. La información histórica viene dada por en número de ediciones de la EFF en la que ha participado cada hogar.

Se entrenan cuatro modelos de predicción basados en algoritmos de machine learning, y se compara su rendimiento con un modelo de referencia que se utiliza tradicionalmente para anali-

[7]La encuesta usada en Beste et al. (2023) permite que un hogar que no ha participado en una edición pueda volver a hacerlo en la edición siguiente. Pero esta caraterística no está presente en el diseño de la EFF, por lo que no puede usarse.

Fuente de información	Variables
Características del hogar	Número de adultos con trabajo, 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, Posee al menos un vehículo, Percentil de renta, Percentil de riqueza Bruta, Pareja vive en el hogar, Poseen otros activos financieros, Tiene deudas pendientes, Tiene ingresos de activos, Hijos viven en el hogar, Nivel de satisfacción con la vida
Características de 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 entrevistador	Recelo tras la entrevista, Edificio Unifamiliar, Barreras - portero automático, Barreras - Sin barreras, Nivel de comprensión de las preguntas por PR, Interés de PR, Razones colaborar - Interesado en estos estudios, Razones para colaborar - El estudio 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, Proporción de preguntas monetarias respondidas en valor puntual, Proporción de preguntas monetarias respondidas (incluye intervalos), Número de variables con valores missing, Duración de la entrevista (en segundos)

Cuadro 4.2: Selección de variables para entrenar los modelos de predicción

zar Panel Attrition. Este modelo base es una Regresión Logística o Logit de efectos primarios, es decir, sin interacciones entre variables y sin especificaciones no lineales. Los modelos de machine learning que se entrenan y evalúan son un árbol de decisión (CART, Breiman et al. (1984)), un Random Forest (Breiman (2001)), un eXtreme Gradient Boosting (XGBooster, Chen and Guestrin (2016)) y un Naive Bayes (Webb et al. (2010)). La elección de estos modelos se fundamenta, por un lado, en los buenos resultados que han mostrado para predecir panel attrition en otras encuestas para hogares (Kern et al. (2019), Kern et al. (2021), Beste et al. (2023)) y, por otro lado, en que estos modelos 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 para diseños adaptativos y reactivos. El cuadro 4.3 contiene una breve descripción de los modelos de clasificación que se entrenan.

Algoritmo	**Descripción**
Regresión logística	Modelo de regresión utilizado comúnmente para analizar problemas de clasificación binaria. Estima la probabilidad de respuesta binaria basada en un conjunto de regresores.
Árbol de decisión (CART)	Modelo de clasificación o regresión que subdivide recursivamente el espacio de datos en regiones disjuntas de tal manera que todos los elementos de una región pertenezcan a la misma clase.
Random Forest	Modelo de clasificación o regresión basado en la combinación de árboles de decisión que creados de manera independiente con muestreos aleatorios tanto del conjunto original de datos como de las variables.
XGBooster	Método de clasificación o regresión en el que se constuye una secuencia de árboles en el que cada modelo repondera los elementos erróneamente clasificados por del modelo anterior para darles más peso.
Naïve Bayes	Modelo que clasifica asignado la clase que maximiza la probabilidad condicional de dicha clase dados unos atributos.

Cuadro 4.3: Modelos de clasificación

Antes del entrenamiento de los modelos se siguen los siguientes pasos de preparación de los datos:

1. Se revisan los datos en busca de valores faltantes o missing. Este problema se soluciona utilizando el fichero de datos imputados de la EFF. También se elimina un hogar porque no tenía valores en el fichero de paradata.

2. Se revisan los datos en busca de valores atípicos. Sólo se encuentran en las variables de

duración de la entrevista y se procede a su imputación. Los detalles de esta imputación se describen en la sección 5.1.

3. Las variables categóricas ordinales se recodifican para que tengan valores crecientes que empiecen por valor 1. Las variables categóricas no ordinales se codifican creando variables binarias de valores 0 y 1 para cada categoría de dichas variables (One-Hot encoding).

Para el entrenamiento y validación de los modelos se sigue una estrategia de k-fold cross-validation para los conjuntos de entrenamiento y validación, con k=5. Para evitar que haya fuga de datos (data leakage) del conjunto de entrenamiento al de validación, los estimadores se obtienen se implementando un Pipeline de scikit-learn con dos pasos: Primero, se realiza el sobremuestreo de la clase minoritaria utilizando el algoritmo SMOTE, y a continuación estandarizan los valores de las variables monetarias.

Con respecto al tuning de hiperparámetros, en la tabla 4.2 puede consultarse la estrategia de búsqueda de valores de hiperparámetros para cada algoritmo, así como el espacio de valores de hiperparámetros considerados para cada hiperparámetro. Como no existe certeza sobre cuáles son los valores más adecuados, se considera un rango bastante amplio para cada hiperparámetro de cada algoritmo. Con respecto a los algoritmos de búsqueda, se implementa un algoritmo de búsqueda de res (Grid search) para los modelos Logit y CART. Este algoritmo prueba cada combinación de hiperparámetros. Sin embargo, los algoritmos Random Forest y XGBOOST tienen elevados tiempos y costes de ejecución, por lo que se utiliza un algoritmo de búsqueda aleatoria (Random Search) con 2500 iteraciones. Este algoritmo realiza una búsqueda aleatoria de combinaciones de hiperparámetros en lugar de comprobar cada combinación de los mismos. Se ha comprobado que este tipo de 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 de referencia que se busca maximizar en el entrenamiento y validación es la ROC AUC. Adicionalmente, por si es necesario comparar podemos con ROC AUC similares, también se calculan las métricas de Accuracy, Precision, Recall y F1. En el cuadro 4.5 presenta las definiciones de todas las métricas de evaluación de modelos de clasificación.

Algoritmo	**Método de búsqueda**	**Espacio de hiperparámetros**
Regresión Logística	Grid	C: np.logspace(-1.5,3,10) penalty: l1, l2
Árbol de decisión CART	Grid	criterion: gini, entropy max_depth: 2, 4, 8, 16, 32, None min_samples_split: 2, 4, 8, 16, 32 min_samples_leaf: 2, 4, 8, 16, 32, 64 max_features: sqrt, log2
Random Forest	Random	min_samples_leaf: 2, 4, 8, 16, 32, 64 n_estimators: 25, 50, 70, 100 max_features: sqrt, log2 max_samples: 0.6, 0.7, 0.8, 0.9, None min_samples_split: 2, 4, 8, 16, 32 max_depth: 2, 4, 8, 16, 32, None
Extreme Gradient Boosting	Random	max_depth: 2, 3, 4, 5, 6 n_estimators: 100, 300, 500 reg_alpha: np.linspace(1, 11, 20) reg_lambda: np.linspace(1, 11, 25) base_score: np.linspace(0.1, 0.6, 10) subsample: np.arange(0.5, 1, 0.05) learning_rate: 0.1, 0.05
Naive-Bayes	-	-

Cuadro 4.4: Algoritmos, estrategias de búsqueda y espacio de hiperparámetros

Métrica	**Descripción**
Accuracy	Número de predicciones correctas sobre el número total de predicciones.
Precission	Número de predicciones correctas de clase Attrition sobre el total de predicciones de Attrittion hechas (verdadero Attrition y falso Attrition)
Recall	Número de predicciones de Attrition correctas sobre el total de casos de Attrition reales (predicciones correctas de Attrition y predicciones incorrectas de no Attrition)
F1	Media armónica de Precission y Recall.
ROC AUC	Área bajo la curva ROC. La curva ROC mide el rendimiento de predicciones correctas de Attrition y las predicciones incorrectas de Attrition, y AUC es el área bajo dicha curva. Toma valores de 0,5 a 1, siendo 0,5 un clasificador aleatorio y 1 un clasificador perfecto

Cuadro 4.5: Definiciones de métricas de evaluación de modelos de clasificación

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, y de manera especial la que tienen con la variable Attrition. 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 el trabajo 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 transformación de variables para los modelos de predicción.

Este análisis se separa en cinco bloques. El primero describe la gran cantidad de datos que se generan en la EFF y el proceso de filtrado que se ha realizado para este proyecto. Los tres boques siguientes abordan tres temas que preocupan a los productores de encuestas porque suelen afectar a la participación de los panelistas: la experiencia en la encuesta y las entrevistas, la persona que responde a la encuesta, y las características de los hogares. Finalmente, se aborda el problema de valores atípicos detectados en las duraciones de las entrevistas.

Número de registros y variables

Una sola edición de la EFF produce una gran cantidad de datos. El Cuadro 5.1 muestra el número de registros y variables disponibles en cada uno de los ficheros utilizados en este proyecto. Estos números incluyen a todos hogares contactados y a todas variables generadas

durante la producción de datos de la EFF en sus ediciones de 2017, 2020 y 2022[1]. Tras realizar el filtrado de hogares elegibles para el estudio, se obtiene que los hogares de la EFF2017 elegibles para la EFF2020 son 5,937[2] y los hogares elegibles de la EFF2020 para la EFF2022 son 5,505. Con respecto al censo de entrevistadores, sus números incluyen a los 260 entrevistadores que han participado en las ediciones de 2014, 2017, 2020 y 2022. Tras filtrar por las ediciones de 2017 y 2020, se obtiene que en la EFF2017 participaron 69 entrevistadores, mientras que en la EFF2020 participaron 65 entrevistadores, de los cuales 25 personas 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*

En el cuadro 5.1 también puede observarse que hay ficheros que almacenan más de 6,000 variables. Esto supone un problema de dimensionalidad ya que hay más variables que registros en los datos. Sin embargo, hay cuatro maneras para reducir drásticamente el número de variables a manejar sin perder información relevante, y obtener las variables mencionadas en el cuadro 4.2. La primera es que la inmensa mayoría de variables almacenan las respuestas al cuestionario principal de la EFF. En el segundo párrafo de la sección 4.1 se comentó que el número de preguntas que se formulan depende del número de miembros del hogar, sus edades, y los activos y deudas que posea el hogar, y que en la EFF2017 se plantearon entre 137 y 594 preguntas a cada hogar. Como los modelos de predicción requieren de variables que contengan datos para todos los registros, es posible descartar muchas variables por no tener valores para todos los hogares.

[1] De la EFF2022 sólo se utiliza el fichero de contactos ya que sólo se necesita la información sobre la participación de los hogares panel en dicha edición.

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

La segunda razón para descartar variables es que muchas no son informativas en su estado original y necesitan ser combinadas con otras para poder obtener información interpretable, o se utilizan como apoyo para la imputación. Por ejemplo, la información sobre cantidades monetarias se recoge en cuatro variables que permiten declarar valores en intervalos a los hogares que no quieran o no puedan dar un valor puntual (Barceló et al. (2021)). Esto se utiliza en la imputación para estimar valores puntuales dentro del rango declarado por el hogar. Al usar el fichero de datos imputados para entrenar los modelos, todas esas variables auxiliares se descartan.

En tercer lugar, hay variables duplicadas porque están almacenadas en varios ficheros de datos, por lo que sólo es necesario extraerlas de uno de esos ficheros. Por ejemplo, todas las variables que aparecen en el fichero de datos imputados también aparecen en el fichero de trabajo. Del fichero de trabajo se extraen indicadores de no-respuesta y otras variables de interés que no aparecen en el fichero de datos imputados, y de éste último se extraen las variables con los valores missing imputados.

Finalmente, para que los modelos de predicción puedan aplicarse tanto para datos de la EFF2017 como para la EFF2020, sólo se seleccionan las variables que estaban disponibles en ambas olas. Esta tarea ha requerido una gran dedicación de esfuerzo y tiempo, ya que en algunos ficheros se detectaron variables que no mantuvieron su nomeclatura, el tipo de dato almacenado o la codificación de los datos entre diferentes olas. Para asegurar la homogeneidad, se ha revisado de manera individualizada la nomenclatura y la codificación de cada variable para ambas ediciones de la EFF.

La experiencia en la encuesta y las entrevistas

En la sección 3.1 se comentó que los hogares panel poseen experiencia previa sobre la encuesta que puede afectar a su participación en olas posteriores. Esta experiencia puede abarcar varias ediciones, pero también puede ser informativo observar datos sobre la ola más reciente.

En la figura 5.1 hay cuatro gráficos de mekko que muestran cómo fue la participación en la EFF2020 de hogares elegibles de la EFF2017 según su ola de entrada en la EFF, si consintieron grabar la entrevista de la EFF2017, si dicha entrevista se realizó con un proxy, y el nivel de recelo que mostraron después de realizarla. Los gráficos de mekko son gráficos de columnas apiladas 100 % en los que la anchura de cada columna muestra la proporción de hogares que hay de una categoría dentro de la muestra. Las regiones superiores o rojas de cada columna muestran la proporción de hogares que no participaron en la EFF2020, mientras que las regiones inferiores o azules muestran la proporción de hogares que sí participaron.

En la figura superior izquierda de la figura 5.1 se observa que la proporción de abandono de hogares que participaron por primera vez en 2017 es mayor que la de los que lo hicieron por

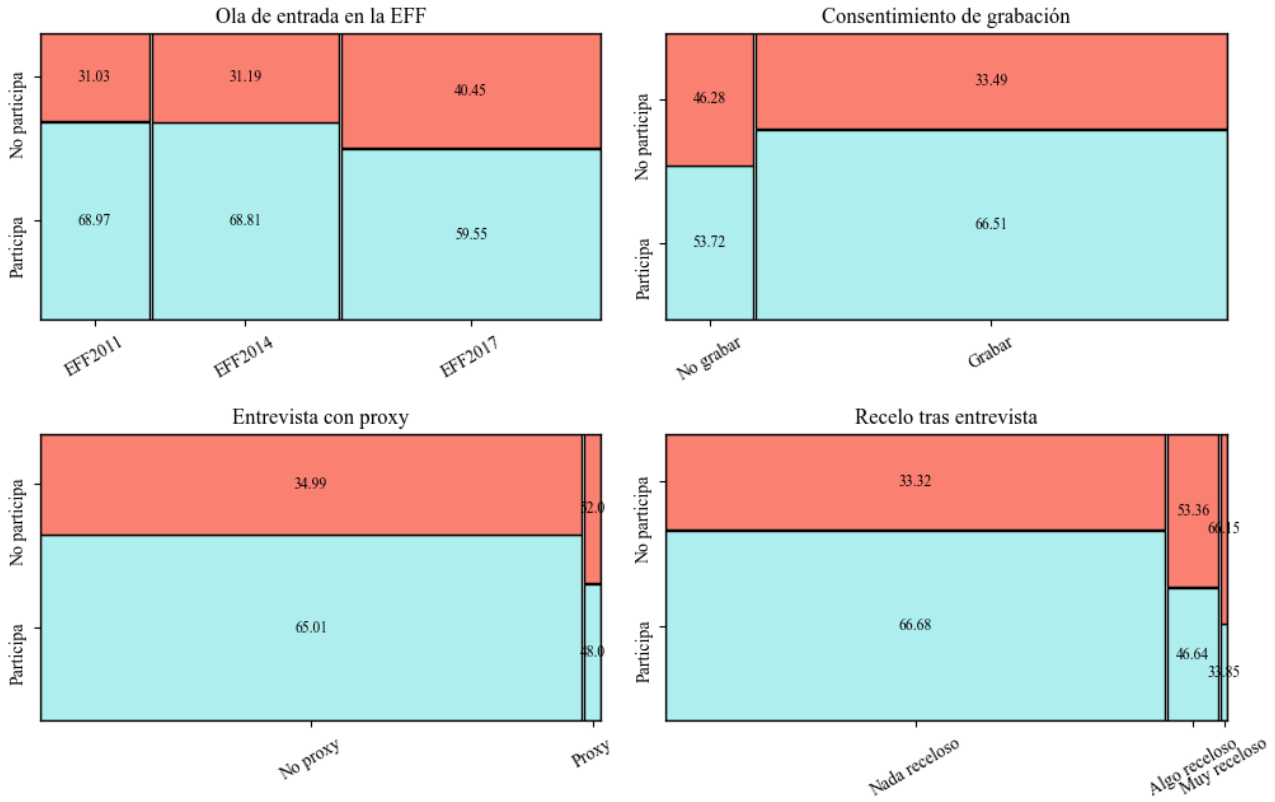

Figura 5.1: Participación en EFF2020 por experiencia en la EFF y en la entrevista de EFF2017

primera vez en 2014 o en 2011, y entre estas dos últimas la proporción es similar. Esto sugiere que los hogares que van a participar en su segunda ola podrían ser más complicados de retener que los que llevan más tiempo.

La figura superior derecha muestra que la proporción de no participación es mayor entre los hogares que no consintieron grabar la entrevista en 2017. Esto puede ser una señal de recelo hacia la encuesta, lo que puede dificultar la participación en las siguientes ediciones. En ese sentido también es interesante ver si los hogares se mostraban recelosos tras la entrevista, que es lo que se observa en la figura inferior derecha. La gran mayoría de hogares no se mostraron recelosos tras la entrevista, pero se observa que la proporción de hogares que no participaron en la EFF2020 es mayor a medida que aumenta el nivel de recelo.

Finalmente, en la figura inferior izquierda, se ve que la proporción de abandonos en 2020 fue mayor entre los hogares que hicieron la entrevista con proxy en 2017. Una situación que ha ocurrido bastantes veces en la EFF y que puede encajar con un abandono es el de un hogar formado por personas muy mayores en el que quien lleva las finanzas y termina respondiendo a la entrevista es un hijo o un familiar. Muchos de estos familiares se muestran muy recelosos y, comprensiblemente, quieren que no se moleste a sus familiares. En este tipo de situaciones

puede ser más importante volver a convencer a estos familiares que a los propios miembros del hogar, ya que al final son ellos quienes conocen la información del hogar.

Algunos de estos resultados pueden parecer poco útiles porque es razonable pensar que un hogar que se mostró receloso durante la entrevista seguramente será más complicado de convencer para volver a participar en la siguiente edición. Sin embargo, para un entrevistador que está a punto de entrevistar a un hogar, puede ser muy útil saber si ese hogar se mostró receloso tras la anterior entrevista. A la hora convercer al hogar puede centrarse más en utilizar argumentos relacionados con la confidencialidad y la seguridad de los datos y no tanto en hablar de la relevancia de la encuesta o del eco que ha tenido en los medios de comunicación. Estos resultados pueden servir para justificar este tipo de análisis y encontrar qué información puede ayudar al trabajo de los entrevistadores.

Características y comportamiento de la PR

En la sección 3.1 se vió que las características de la persona que responde a una encuesta puede ser relevante para el panel attrition. Es este apartado vamos a ver cómo se relacionan algunas características de la PR en 2017 con la participación del hogar en la EFF2020.

La figura 5.2 contiene los gráficos de mekko de la participación en la EFF2020 según el nivel máximo de estudios alcanzados de la PR, su edad y su satisfacción con la vida en 2017 y el interés que mostró durante la entrevista de la EFF2017. El gráfico superior izquierdo muestra que la proporción de hogares que no participó en la EFF2020 es mayor cuanto menor era el nivel de estudios de la PR en 2017. Una posible explicación de este resultado podría deberse a la menor tenencia de productos financieros por parte de hogares de menor nivel educativo (Hospido et al. (2023)). El cuestionario de la EFF contiene muchas preguntas sobre muchos productos financieros diferentes, y es razonable pensar que alguien que no posee este tipo de activos piense que no tiene sentido participar en esta encuesta. Un entrevistador con este conocimiento podría preparar un argumentario orientado a destacar que sin su participación no se podría identificar a hogares con sus características y así para poder diseñar políticas económicas específicas para esos hogares.

En el gráfico superior derecho se ve que la proporción de hogares que participó en 2020 aumenta a medida que aumenta la edad de la PR en 2017, excepto cuando ésta tenía menos de 35 años, que presenta la segunda proporción más alta de los grupos de edad. El resultado para los hogares más jóvenes podría explicarse por el hecho de que cada vez menos de estos hogares son propietarios de su vivienda principal (Banco de España (2017), Banco de España (2019), Banco de España (2022)) y esto puede hacer que sean más propensos a mudarse, y por tanto ser más difíciles de localizar. El caso de las PR de mayor edad podría explicarse por motivos de fallecimiento.

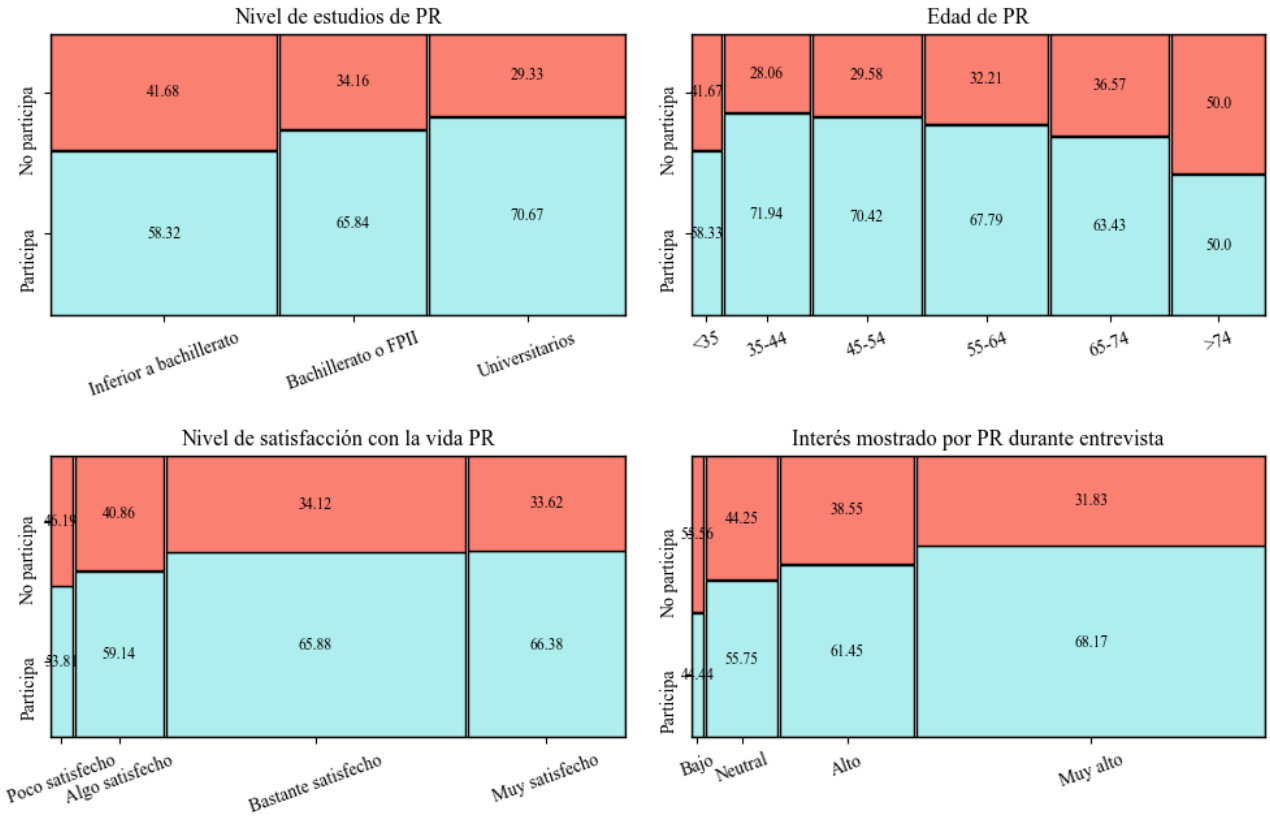

Figura 5.2: Participación en EFF2020 por características y actitudes de PR en EFF2017

Finalmente, en los gráficos de la parte inferior de la figura 5.2 vemos, por un lado, que la proporción de hogar que dejan de participar en 2020 se reduce a medida aumenta el nivel de satisfacción con la vida de la PR en 2017, y por otro, que la proporción de hogares que participaron en 2020 aumenta a medida que aumenta el interés por la encuesta. Este último resultado es razonable y útil de saber para el entrevistador porque puede basar su argumentario para convencer al hogar en ese interés.

Características hogar

En este apartado se analiza a nivel exploratorio la posible relación que puedan tener las características del hogar con el panel attrition. En concreto, se analizan las variables de renta, riqueza y tenencia de deudas. La renta y la riqueza son el eje central de la EFF y es habitual incluirlas de alguna manera en cualquier análisis que se haga con los datos de la encuesta.

En la figura 5.3 hay tres gráficos de mekko en el que se observa la proporción de hogares que participaron en la EFF2020 según la posición relativa de cada hogar en las distribuciones

de renta anual y riqueza bruta[3] de los hogares españoles en 2017, y también si el hogar tenía deudas pendientes en 2017. La posición relativa de cada hogar dentro de la distribución de renta se muestra indicando los percentiles de la distribución total entre los que se sitúa cada hogar. Si el nivel de renta de un hogar lo sitúa entre los percentiles 60 y 80 del total de la renta de hogares en España, entonces ese hogar está en la categoría "P60-P80"de la distribución de renta, y si se sitúa por encima del percentil 90, la categoría es ">P90".

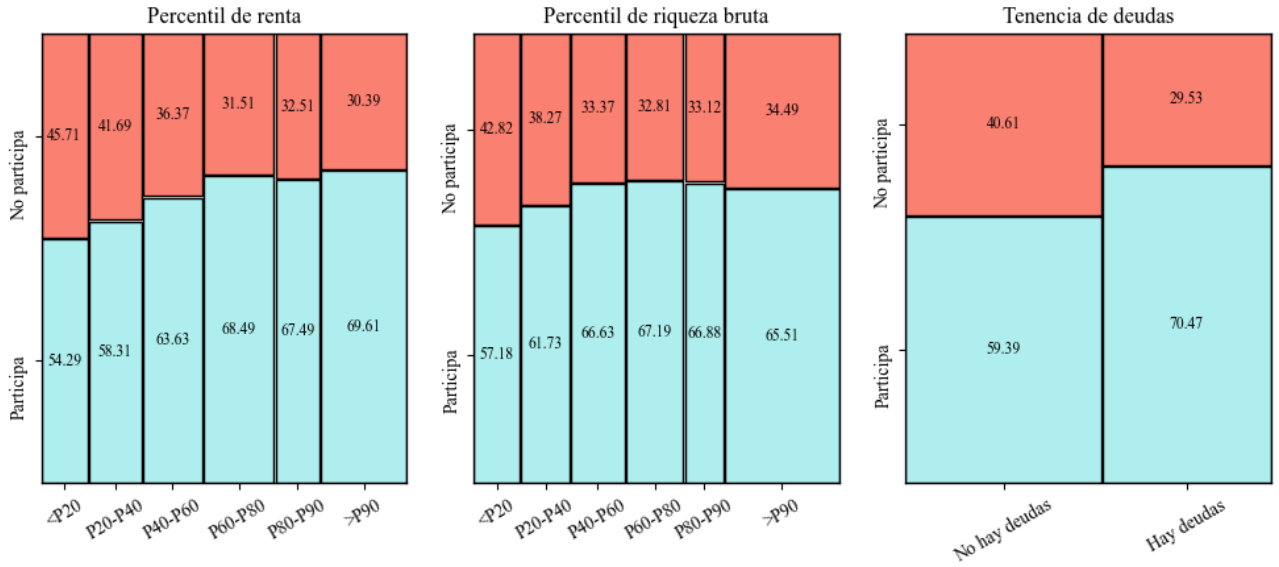

Figura 5.3: Participación en EFF2020 por renta, riqueza y tenencia de deudas en EFF2017

En el gráfico de la izquierda de la figura 5.3 se observan dos tendencias diferentes dependiendo del nivel de renta anual de 2017. Por encima del percentil 60 no se aprecian grandes diferencias en la proporción de panelistas que no participaron en 2020, pero por debajo de ese percentil se ve que la proporción de hogares que no participaron es mayor cuanto menor es el nivel de renta. Una posible explicación puede venir de que los niveles de tenencia de activos son menores cuanto menor es el nivel de renta de los hogares (Banco de España (2019)) y, como se ha comentado para el nivel educativo, es razonable pensar que un hogar que tiene pocos activos considere que no tiene sentido participar en una encuesta como la EFF. Estas dos tendencias también se observan para la distribución de riqueza bruta, pero en este caso la tendencia se observa por debajo del percentil 40 de riqueza bruta. La explicación de por qué la proporción de abandonos es mayor sería similar a la de renta. Si no se tienen activos, se puede pensar que no tiene sentido participar en la EFF.

[3]La riqueza bruta se define como la suma del valor de todos los activos que posee el hogar (activos reales + activos financieros = riqueza bruta).

Finalmente, el gráfico de la derecha muestra la proporción de hogares que participaron en la EFF2020 según si tenían deudas pendientes en 2017. Se observa que la proporción de hogares que no participan en 2020 es mayor entre los hogares que no tenían deudas en 2017.

Duración de las entrevistas

En esta última pieza de análisis exploratorio se comenta el tratamiento de valores atípicos detectados en la duración de las entrevistas. En el entrenamiento de los modelos de la sección 5.2 se ha utilizado la duración total de la entrevista en segundos. Este valor se obtiene del fichero paradata calculando, para cada hogar, la suma de los segundos que pasaron en cada pantalla del CAPI durante la entrevista. Al analizar la distribución de las duraciones se encontraron valores atípicos que indicaban que hubo entrevistas que duraron más de 20 horas, cuando las entrevistas suelen durar entre una hora y hora y media (Barceló et al. (2021)).

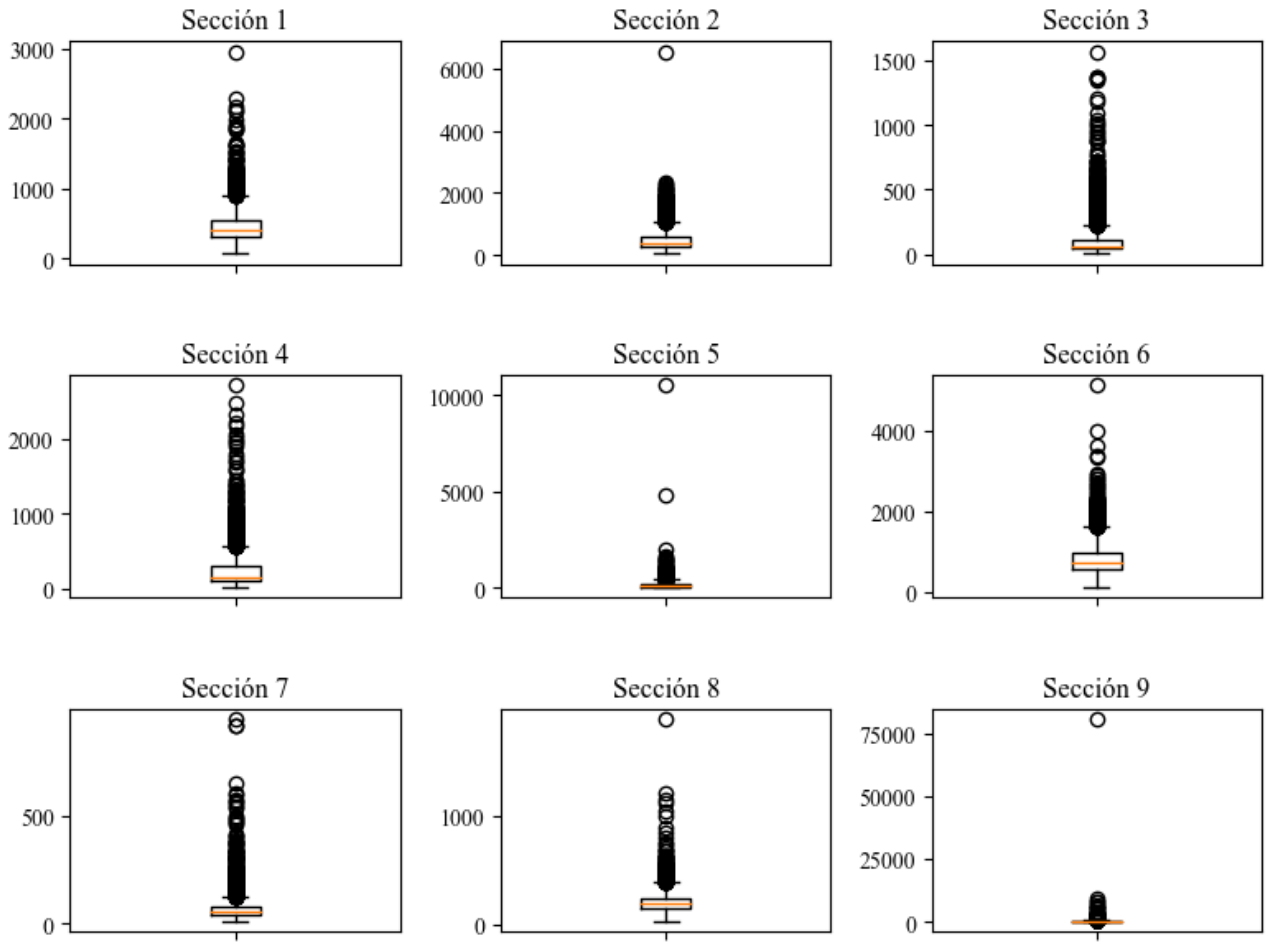

Figura 5.4: Duración por sección del cuestionario en la EFF2017

A partir de la información en el fichero paradata se calcularon las duraciones de cada una de las nueve secciones de la EFF. La figura 5.4 contiene los gráficos box-plot de las duraciones de las 9 secciones del cuestionario de la EFF2017. En estos gráficos se ve que hay valores atípicos en todas las secciones, y que algunos son particularmente altos, especialmente en la sección 9. Se sospecha que los entrevistadores no cerraron bien la aplicación del ordenador o la tablet al tomar un descanso, o al interrumpir una entrevista, o al terminarla.

Estos valores tan altos representan un error de medición importante que puede afectar a los modelos de predicción, por lo que es necesario tratarlos. Como hay secciones que contienen más preguntas que otras, la duración en cada sección dependerá de cuántas preguntas se formulen en cada sección. Como el fichero paradata contiene esa información, se decide imputar la duración a nivel de sección, y en concreto se imputan los valores que estén por encima del percentil 99.9 de la duración de dicha sección. Se utiliza el algoritmo kNN (*k-nearest neighbors*) con k=5 y distancia euclídea. Como rasgos de cada sección se calculan el número de preguntas que se formulan una sola vez, el número total de preguntas formuladas, el número de veces que se vuelve a una pantalla anterior, el número de paradas, el número de preguntas categóricas con 1-4 opciones de respuesta, 5-9 opciones de respuesta o más de diez opciones de respuesta, y el tiempo pasado en preguntas monetarias. Tras la imputación, se calcula la duración total de la entrevista como la suma de las duraciones de todas las secciones. Este procedimiento se implementa por separado para las duraciones de cada ola.

5.2. Evaluación de modelos

Rendimiento de los modelos sobre el conjunto de test

El cuadro 5.2 contiene los resultados de la evaluación de los modelos entrenados con algoritmos de machine learning con respecto al modelo de referencia de Regresión Logística Logit. Recordemos que los modelos se han entrenado con datos de la EFF2017 para predecir la no participación de los hogares panelistas en la EFF2020. El ejercicio del test consistía en predecir la participación de los hogares panel en la ola EFF2022 utilizando información de la EFF2020.

Las métricas de evaluación que se utilizan son Accuracy, Precision, Recall, F1 y ROC AUC. La métrica de referencia que utilizada 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. La del modelo CART es 0,5521, la del Random Forest es 0,5821, la del XGBooster es 0,5911, y la del

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 5.2: Métricas de evaluación de los modelos de predicción en el conjunto de test

Naive Bayes es 0,5798. Ninguno de estos valores mejora a la ROC AUC del modelo Logit. Los que se quedan más cerca son el Random Forest y el XGBooster. De todas maneras, es necesario recalcar que el valor más alto de ROC AUC, el del Logit, sigue estando por debajo de 0,6, por lo que resultados de los test son malos.

¿Por qué el rendimiento de los modelos es malo?

Los resultados de la evaluación de los modelos con el conjunto de test indican que hay modelos de machine learning, aunque se quedan cerca, no mejoran en rendimiento de la predicción con respecto a un Logit. Sin embargo, la métrica ROC AUC por el Logit indica que el rendimiento de este modelo es malo. ¿Por qué los modelos no hacen bien la predicción?

Una posible causa podría ser el overfitting, es decir, que los modelos han aprendido demasiado de los datos con los que han sido entrenados y no son capaces de generalizar bien cuando se enfrentan a datos nuevos. Si esta fuese la principal razón, lo esperable sería que su rendimiento fuese bueno si se hace el ejercicio de predicción sobre los datos de entrenamiento. En el cuadro 5.3 se presenta una comparación entre las métricas de entrenamiento del ejercicio de predecir la Attrition con los datos de test y de entrenamiento (Train).

Modelo	Datos	Accuracy	Precision	Recall	F1	ROC AUC
Logit	Test	0,6556	0,3749	0,3573	0,3659	0,5959
	Train	0,6389	0,4905	0,4518	0,4704	0,6517
CART	Test	0,6434	0,3338	0,2835	0,3066	0,5521
	Train	0,6126	0,4594	0,5178	0,4868	0,6188
Random Forest	Test	0,6489	0,3529	0,3148	0,3328	0,5821
	Train	0,6596	0,5223	0,4770	0,4986	0,6726
XGBooster	Test	0,6718	0,3772	0,2769	0,3194	0,5911
	Train	0,6544	0,5144	0,4675	0,4898	0,6722
Naive Bayes	Test	0,6254	0,3504	0,4063	0,3763	0,5798
	Train	0,6251	0,4741	0,5178	0,4950	0,6435

Cuadro 5.3: Comparación métricas evaluación entre conjuntos Test y Train

En esta tabla hay dos tipos de resultados. En primer lugar, se observa que los resultados del

test decaen puntos con respecto al conjunto de entrenamiento. Esta degradación es esperable, ya que los algoritmos han sido entrenados con el conjunto de entrenamiento y conocen los patrones internos de los datos mejor que los del conjunto de testo. La caída del entrenamiento al test es baja cuando se compara la métrica de accuracy, pero la caída es más grande para el resto de métricas. Por ejemplo, la ROC AUC del Random Forest cae de 0,6726 para el conjunto de entrenamiento a 0,5821 para el conjunto de test, y la del Logit cae de 0,6517 para el entremaniento a 0,5959 para el test.

El segundo tipo de resultados se centra en la comparación del rendimiento entre los modelos para el conjunto de entrenamiento. En este caso la ROC AUC de todos los modelos sí supera el valor de 0,6, y también se observa que la de dos de los modelos, el Random Forest y XGBooster, que son 0,6726 y 0,6722 respectivamente, sí son más altas que la ROC AUC del modelo Logit, que es 0,6517. Del resto de métricas, sólo el recall del XGBooster presenta un valor más bajo que el del modelo Logit. De los dos modelos, el mejor sería el Random Forest porque presenta mejores valores en todas las métricas. Sin embargo, aunque este valor sí esté por encima de 0,6, al estar por debajo de 0,75 a este modelo se le clasifica como un predictor regular. Este resultado descarta el overfitting para explicar el mal rendimiento de los modelos, e invita a considerar otras alternativas e incluso nuevos enfoques. Estas opciones se comentan con más detalle en el capítulo 6.

Importancia de las variables en la predicción

Tal y como se comentó en la sección 4.2, una ventaja de los modelos basados en árboles de decisión 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 5.4 presenta las 20 variables con más importancia en el modelo de Random Forest ordenadas por valor de importancia.

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 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 importancia de proporción de preguntas monetarias respondidas por el hogar (incluyendo intervalos), y algunas variables que hemos mencionado en la sección 5.1, como son si la PR consintió que se grabase la entrevista, el nivel educativo de la PR y el número de olas en las que el hogar ha participado.

Variable	**Importancia**
PR es panel	0,0822
Interés PR	0,0790
Razones para colaborar - Relevancia de la encuesta	0,0605
Proporción de preguntas monetarias respondidas (incluye intervalos)	0,0594
Posee al menos un vehículo	0,0539
Tiene deudas pendientes	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
Hijos viven en el hogar	0,0318
Nivel educativo PR	0,0313
Número de olas en las que ha participado el hogar	0,0304
Sexo de PR	0,0291
Número de adultos con trabajo	0,0277
Posee fondos de inversión	0,0206
Nivel de comprensión de las preguntas por PR	0,0206
Proporción de preguntas monetarias respondidas en valor puntual	0,0198
Razones para colaborar - El estudio lo lleva el BdE	0,0195
Estado Civil - Casada	0,0190

Cuadro 5.4: Importancia de las variables en el modelo de Random Forest

Capítulo 6

Conclusiones y futuras líneas de trabajo

El objetivo principal de este Trabajo de Fin de Máster es desarrollar un modelo basado en algoritmos de machine learning que ayude a predecir si un hogar panel de la Encuesta Financiera de las Familias (EFF) dejará de participar en la siguiente edición. Para ello, se ha seguido la implementación de entrenamiento y evaluación realizada en Beste et al. (2023), pero adaptada al caso de estudio de la EFF.

Este trabajo se ha divido en dos piezas de análisis. La primera es un análisis exploratorio de las variables de la EFF y ver su posible relación con la participación de los hogares en la siguiente edición. En la segunda se entrenan cuatro modelos basados en algoritmos de machine learning (CART, Random Forest, XGBooster y Naïve-Bayes) para predecir el Panel Attrition en la EFF2020 con datos de la EFF2017, y a continuación se compara y evalúa su rendimiento comparado con el ofrecido por un modelo de Regresión Logística de efectos principales, que es un modelo que se ha utilizado tradicionalmente para analizar el Panel Attrition.

El trabajo de análisis exploratorio ha sido muy extenso y complejo porque ha requerido manejar y combinar un total de doce ficheros de datos repartidos entre tres ediciones de la EFF. Siete de ellos contenían registros de hogares, otros dos registros de incidencias, otros dos registros de pantallas de interacción con un software y el último era un censo de entrevistadores. Además, dos de los ficheros contenían más de 6,000 variables, y otros dos más de 600. En una misma edición de la encuesta había algunas variables duplicadas en diferentes ficheros, y también algunas variables comunes entre olas tenían diferentes codificaciones dependiendo de la ola. Tras eliminar duplicados, redundancias y homogeneizar variables, se consiguió un conjunto de datos manejable.

El análisis exploratorio ha mostrado varios resultados interesantes que pueden ser útiles para hacer implementaciones de diseños adaptativos. La proporción de hogares panel que no

participaron en la EFF2020 era mayor para hogares que en 2017 participaron por primera vez, se mostraron más recelosos, no consintieron grabar la entrevista, la realizaron con un proxy. También lo era en aquellos hogares cuya PR en 2017 tenía un menor nivel de estudios, tenía menos de 35 años o más de 74, mostraron un menor nivel de satisfacción con su vida y un menor interés durante la entrevista. Finalmente, la proporción de abandonos en 2020 también fue mayor para los hogares cuya renta de 2017 se situaba en las tres quintilas inferiores de la distribución de renta de los hogares españoles en 2017, su riqueza bruta en 2017 se encontrada en dos las quintilas inferiores de la distribución de riqueza bruta de todos los hogares españoles en 2017, y si no poseían deudas. Toda esta información puede ser útil para un entrevistador que va a visitar a un hogar panel porque puede adaptar los argumentos que utilice para convencerles. Por ejemplo, si el hogar se mostró receloso en la edición anterior, puede ser más útil enfocarse en hablar sobre la confidencialidad de la encuesta que la presencia del estudio en los medios de comunicación.

En la parte del entrenamiento y evaluación de los modelos de machine learning, ninguno de los modelos de mahcine learning entrenados llega a superar el rendimiento del modelo de Regresión Logística para la predicción del conjunto de test. Además, el valor de la ROC AUC de este modelo Logit no supera el valor de 0.6, lo cual lo clasifica como un predictor malo. Para intentar aprender sobre los errores de predicción, se han usado esos mismos modelos para hacer la predicción con el conjunto de entrenamiento y hacer una comparación entre los resultados del conjunto de test y los de entrenamiento. Todos los modelos presentan mejores resultados para el conjunto que entrenamiento que para el de test, y además los modelos de Random Forest y XGBooster sí presentan valores más altos de ROC AUC que el modelo Logit. Sin embargo, la mejor métrica, la del Random Forest, no llega a superar el valor de 0,7, que lo clasifica como un predictor regular. No parece que los modelos hagan overfitting.

Finalmente, para aprender cómo ha sido el entrenamiento del Random Forest, se observa la importancia de los veinte predictores con mayor valor de importancia. Destacan que la PR ya formase parte del hogar desde al menos dos ediciones antes, el valor del interés mostrado por la PR, que el entrevistador considerase que el hogar participó por la relevancia de la encuesta y algunas variables analizadas durante la exploración de los datos, en concreto la tenencia de deudas pendientes, el número de olas en las que ha participado el hogar, que la PR consintiera grabar la entrevista y el nivel educativo de la PR.

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. **Revisar la selección de variables:** Un buen modelo predictivo requiere estar construido sobre variables que tengan poder predictivo sobre la variable de interés. Los resultados del análisis descriptivo muestran que las variables seleccionadas tienen relación con la

participación de los hogares, pero no tienen suficiente poder para predecir la participación de los hogares en las ediciones siguientes. La selección de variables se inspiró en las implementaciones de Kern et al. (2021) y Beste et al. (2023), que son ejemplos de buenas predicciones de Panel Attrition de hogares. Pero aunque sean encuestas a hogares, tienen métodos de recolección de datos, frecuencias y muestreos diferentes al de la EFF, y en Jankowsky and Schroeders (2022) mencionan que esos factores pueden afectar a los resultados de los modelos e impedir su generalización a otras encuestas. Afortunadamente, en la EFF se recogen muchas variables que no se han utilizado en este estudio, y su inclusión podría mejorar el rendimiento de los modelos hasta niveles aceptables. Esto abre vías de trabajo futuro continuistas:

   a) Revisar la selección de variables actual, añadir las que se hayan dejado fuera y puedan tener poder predictivo, y eliminar aquellas que puedan estar generando ruido. Implementar procesos de feature selection basados en algoritmos de machine learning.
   b) Hacer un análisis más profundo del error de predicción. En concreto, analizar de manera específica aquellos hogares que los modelos predicen mal y ver qué características tienen.

2. **Plantear un ejercicio explicativo en vez de predictivo:** Todo ejercicio de predicción requiere utilizar variables que se conozcan ex-ante. Por esa razón, los modelos de predicción de Panel Attrition utilizan los datos de ediciones anteriores como predictores. Pero la predicción sólo saldrá bien si esos datos tienen mayor influencia sobre el resultado de la variable objetivo que otras variables que todavía se desconocen. Por ejemplo, el papel del entrevistador es muy importante para conseguir la colaboración de los hogares y para conseguir datos de calidad (Lynn (2018) y Groves (2006)). Pero es imposible saber qué entrevistadores estarán disponibles para el trabajo de campo hasta que llegue el momento de hacer las entrevistas. En este sentido, antes de continuar con el ejercicio predictivo, puede ser interesante utilizar todos los datos disponibles de todas las olas y hacer un análisis explicativo de los factores que determinan la participación de los hogares panel, diferenciando entre variables conocidas ex-ante y variables ex-post, y comprobando cuál de los dos grupos tienen más influencia.

3. **El posible papel del Covid-19:** Las olas de la EFF seleccionadas para el estudio son EFF2017, EFF2020 y EFF2022 porque son las que tienen mayor cantidad de datos y también los de mayor calidad. Sin embargo, como se comentó en la sección 4.1, la pandemia del Covid-19 obligó a cambiar la metodología del campo de la EFF2020 para poder hacer la encuesta respetando las condiciones de seguridad y distancia social. Todos los modelos

entrenados utilizan datos de 2017 para predecir 2020, que es año de Covid. Y posteriormente, en el test se utiliza información recogida durante 2020, de nuevo Covid, para predecir 2022, cuando ya se estaba superando la pandemia y fue posible las entrevistas presenciales. Es muy razonable pensar que muchos hogares rechazaron participar en 2020 por miedo al Covid, y también pensar que cambiar el modo de entrevista de personal a teléfónica pudo afectar a los datos. Esto puede implicar que las causas de participación en 2020 sean muy diferentes a las de 2022. Ante esta problemática, se plantean dos posibles alternativas:

   a) **Hacer un modelo de predicción con los datos de las olas previas a 2020:** La metodología de todas estas olas es homogénea porque todas las entrevistas fueron personales, y no hubo una crisis como la del Covid-19. El inconveniente es que sólo se podría utilizar la información de las respuestas de los hogares. Pero no está claro si necesariamente el rendimiento sería peor.
   b) **Repetir el actual ejercicio cuando haya más ediciones posteriores a 2020:** La EFF va a continuar realizándose en los próximos años 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 recoge actualmente y que no está disponible para antes de 2017.

4. **Cambiar el enfoque:**: Siempre existe la opción de considerar enfoques alternativos sean interesantes y más adecuados para el problema que se quiere abordar. En ese sentido, hay dos alternativas bastante interesantes:

   a) **Predecir la participación de los paneles en su segunda ola**: En la exploración de los datos se vio que los hogares que han participado sólo en una edición muestran más proporción de abandonos que los que han participado más de dos años. Esto invita a pensar 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 panelistas que sólo han participado una vez en lugar de predecir la participación de todos los hogares.
   b) **Realizar un análisis de supervivencia:** En vez de predecir un resultado binario, de participar o no participar, se puede plantear hacer un ejercicio de análisis de superviviencia (survival analysis), e intentar predecir el número de ediciones en las que participará un hogar de la EFF antes de abandonar el estudio.

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*.

Banco de España (2017). Encuesta financiera de las familias (EFF) 2014: métodos, resultados y cambios desde 2011. *Boletín Económico*, MAR.

Banco de España (2019). Encuesta financiera de las familias (EFF) 2017: métodos, resultados y cambios desde 2014. *Boletín Económico*, 4/2019.

Banco de España (2022). Encuesta financiera de las familias (EFF) 2020: métodos, resultados y cambios desde 2017. *Boletín Económico*, 3/2022.

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.

Bover, O. (2004). Encuesta financiera de las familias españolas (EFF): Descripción y métodos de la encuesta de 2002. *Banco de España, Documentos Ocasionales*, 0409.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Cart. *Classification and regression trees*.

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.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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*.

De Leeuw, E. D. et al. (2005). To mix or not to mix data collection modes in surveys. *Journal of official statistics*, 21(5):233–255.

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.

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.

Hospido, L., Machelett, M., Pidkuyko, M., and Villanueva, E. (2023). Encuesta de competencias financieras (ECF) 2021: principales resultados y cambios desde 2016. *Encuesta de Competencias Financieras/Banco de España*.

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.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17):1–5.

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.

Nicoletti, C. and Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the european community household panel. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168(4):763–781.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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

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.

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

Webb, G. I., Keogh, E., and Miikkulainen, R. (2010). Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714.