FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2017/2018

Trabajo de Fin de Máster

TITULO: Análisis y predicción del canal de contratación en el Sector Bancario

Alumno: Mª Montserrat Pérez Méndez

Tutor: Aida Calviño Martínez

Noviembre de 2018

UNIVERSIDAD COMPLUTENSE
MADRID

AGRADECIMIENTOS

A mi familia, por ser un apoyo incondicional en cada paso de mi vida.

A los profesores y compañeros, de los que tanto he aprendido, y en especial a Aida por su esfuerzo y dedicación en tutelar este trabajo.

ÍNDICE GENERAL

1. Introducción

El mercado está inmerso en un continuo cambio, los negocios y las empresas evolucionan desde hace ya unas décadas hacia una *Era Digital*. Se crean nuevos sistemas y formas de trabajo, que sustituyen a las formas tradicionales, en las que la transformación digital deja de ser una opción y pasa a ser una necesidad.
Se trata de un cambio de carácter generacional, en el que ya no existen barreras y donde no se conciben los negocios sin el impacto de la tecnología.

Esta creciente transformación digital de los negocios influye en gran medida en el sector bancario. Este sector se ha caracterizado, durante décadas, por su tradicionalismo y conservadurismo. Nos encontramos ante una nueva situación en la que las entidades bancarias deben innovar para seguir siendo competitivas en el sector y aprovechar así los nuevos canales de relación con el cliente.

Según esta visión, nos encontramos ante dos grandes cambios en el sector. En primer lugar, aparecen nuevas formas de relación con los clientes (Web y Aplicaciones), dejando en un segundo plano el canal tradicional (las oficinas bancarias). En segundo lugar, cambian las formas de decisión de los consumidores en las que, no solo se buscan las mejores condiciones de financiación, sino mayores facilidades y accesibilidad al control inmediato de los servicios financieros que ofrece la Banca Online.

Este reto digital, en el que se encuentran inmersas las principales entidades bancarias del país, implica, no solo un aumento de recursos para financiar la transformación, sino una disminución de recursos a los canales tradicionales de comunicación con el cliente, llevando a cabo una reestructuración en el modelo de negocio, buscando una mayor eficiencia.

Tanto es así que, según los datos más recientes del Banco de España, el número de oficinas bancarias ha disminuido considerablemente en la última década, pasando de **40.202** oficinas en **2011** a **27.882** en el 3º Trimestre del **2017**, lo que supone una **bajada total de más del 30%** desde 2011.
Esto supone que, desde el año 2011 hasta 2017, la **variación** interanual del número de oficinas ha sido siempre negativa, encontrándose la mayor variación **entre los años 2012 y 2013**, con una **bajada del 13,17%.**

Previo a la definición del objetivo del proyecto, se identifican las ventajas y desventajas de esta transformación digital en el sector bancario, tanto para los clientes como para las entidades bancarias y, en base a esa identificación, se justificará el proyecto escogido.

Las principales ventajas del uso de la banca online en los clientes son, entre otras, una mayor comodidad y sencillez en la contratación y operaciones, facilidad de control de las finanzas personales, accesibilidad permanente a la situación personal con la entidad, mayor personalización de las finanzas, etc.

Para las entidades bancarias esta transformación también supone ventajas, como son las mejoras en el cumplimiento de la normativa (recogida de la firma, datos identificativos, etc.), facilidad de análisis del perfil de sus clientes, imposibilidad de pérdida de documentación o información del cliente, aumento del público objetivo, mayor transparencia, facilidad de acceso a la información de los clientes, asesoramiento personalizado, reducción de costes, etc.

Sin embargo, en la transformación digital, no todo es positivo. Se debe tener en cuenta que para los bancos también pueden existir ciertas desventajas, tales como la fuerte inversión inicial para su adaptación, pérdida del trato "*one to one*" con el cliente, existencia de clientes reticentes al cambio, mayor probabilidad de fraude, etc.

En base a lo desarrollado, y sabiendo que la transformación digital aporta más beneficios que perjuicios al sector, se decide realizar como Trabajo fin de Máster una predicción del comportamiento de los clientes en un Banco, estudiando las operaciones realizadas en banca online frente a las operaciones realizadas en oficina bancaria y en busca de una previsión del comportamiento de sus clientes en el futuro.

2. Objetivos del proyecto

El objetivo principal del trabajo es la propuesta de un nuevo plan estratégico a un Banco, apostando por la transformación digital, basado en el estudio estadístico de predicción del uso de los servicios bancarios.

Para ello, se buscará conocer el comportamiento futuro de la cartera de clientes, para lo que será necesario estudiar los principales atributos y características particulares de los mismos (edad, sexo, hijos, estudios, etc.).

Con el resultado de la predicción podremos aconsejar, o no, una mayor inversión en banca online y destinar menos recursos a las oficinas bancarias.

Para conseguir el objetivo principal, se deberán desarrollar los siguientes objetivos secundarios:

- Conocimiento, preparación y obtención de información útil de los datos.

- Análisis de las variables y estudio de las relaciones existentes entre las mismas.

- Análisis y comparación de los diferentes modelos de predicción, optando por el que mejor prediga el comportamiento de los clientes.

3. Naturaleza de los datos

Para este estudio se han obtenido datos reales de una Entidad Bancaria española.
Por motivos de la Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal, en la que se garantizan los datos personales de las personas físicas, no se hace referencia en el proyecto a datos personales de los clientes de la Entidad que puedan identificarlos, como son el nombre, apellidos, número de identificación fiscal (NIF), etc.

Los datos recogen una muestra de la contratación de diferentes productos realizada por los clientes en el transcurso de los años 2016 y 2017. En cada operación se quedan registrados una serie de datos de los clientes que será lo que se analice en el proyecto.

La información se recibió en formato Access en dos ficheros con una única tabla cada uno. El contenido de los ficheros se reparte de la siguiente manera:

- Fichero Access 1: contratación del año 2016 con un total de 20.250 observaciones.

- Fichero Access 2: contratación del año 2017 con un total de 13.048 observaciones.

Ambos ficheros contenían las mismas variables, y con el objetivo de facilitar el análisis, se unificaron en un único fichero con un volumen total de 33.298 observaciones.

En una primera visualización de los datos se comprueba que se debe realizar una depuración de los mismos ya que se observan algunos errores.

4. Metodología

Para el estudio del proyecto se utilizará SAS Enterprise Miner Workstation 14.1 y SAS Base 9.4.

El objetivo principal del trabajo es conocer el comportamiento fututo de la cartera de clientes mediante la predicción del uso de los servicios bancarios.

Se buscará predecir, según los atributos y características particulares de los clientes, qué canal de contratación utilizarán: el canal objetivo (online) o el canal de contratación alternativo (sucursal).

Para ello, los métodos estadísticos que se desarrollarán a lo largo del trabajo son los siguientes: Regresión Logística, Redes Neuronales, Gradient Boosting, Random Forest, Bagging y Support Vector Machine.

Previo a los métodos estadísticos de predicción, se realizará un análisis de correspondencias simple, como parte del análisis descriptivo, con el objetivo de detectar posibles relaciones entre las variables.

4.1. Descripción de los métodos estadísticos

Análisis de Correspondencias Simple

Mediante el Análisis de Correspondencias Simple se analiza la relación existente entre pares de variables mediante la visualización de las Tablas de frecuencias, o contingencia, de las variables cualitativas.

Las Tablas de frecuencias (o contingencia) se obtienen al cruzar dos variables nominales al repartir la muestra según el número de individuos que presentan cada categoría de las variables. Las columnas representan los niveles de una variable, y las filas de la tabla de frecuencias (o contingencia) representan los niveles de la otra variable.
Para el estudio de la relación entre ambas variables se obtienen los perfiles fila y columna, representando el reparto en porcentaje del interior de las filas y de las columnas. Si los perfiles (fila y columna) son parecidos a los perfiles medios, se considera que las variables son independientes, es decir, el valor que toma una variable no influye en los valores de la otra.

El principal objetivo del ACS es representar gráficamente la relación entre las variables categóricas. Al tener un elevado número de niveles cada variable, no es posible representarlas gráficamente, y por ello mediante el ACS se busca representar la relación mediante un número reducido de dimensiones, buscando la menor pérdida de información posible.
Para ello, se consideran las filas como observaciones y las columnas como variables, y al revés, las columnas como observaciones y las filas como variables.
Existen múltiples métodos para determinar el número de dimensiones a representar, pero en este trabajo nos centraremos en los siguientes:

- Elegir las dimensiones que expliquen una variabilidad aceptable, por encima del 60%.
- Elegir las dimensiones correspondientes a las inercias mayores que $\frac{I}{\min\{r-1,c-1\}}$, lo que se conoce como "average rule"

Siendo "***I***" (Inercia) una medida de dispersión que representa la distancia χ^2de los perfiles al perfil medio ponderados por la masa de los perfiles. De esta manera, permite evaluar la hipótesis de independencia.

Tras la elección de las dimensiones a representar, se examina la calidad de cada factor. En caso de que dicha calidad fuera inferior a 0,5 para alguna de las categorías se incluirían más dimensiones para no perder información.

Por último, se representan las gráficas donde se proyecten las dimensiones escogidas, donde se podrán explicar las relaciones de las variables (y sus categorías) según la ubicación y distancia entre ellas.

Regresión logística

La Regresión Logística es un tipo de análisis de regresión que busca determinar la existencia de relación entre las variables independientes con la variable dependiente, así como predecir la probabilidad de que se cumpla el objetivo sobre la probabilidad de que no se cumpla.

Las variables independientes son de tipo categóricas o continuas (x_1, x_2,…, x_n), mientras que la variable dependiente a predecir es de tipo dicotómica (*Y*).

Al ser la variable que predecir binaria, se puede construir un modelo de regresión lineal cuya variable objetivo sea una variable *dummy*, obtenida a partir de la variable original. La variable dependiente (*Y*) representa la ocurrencia o no de un suceso (objetivo). Si se cumple el objetivo, la variable dependiente toma valor *Y=1* y, si no se cumple el objetivo, toma valor *Y=0*.

Este modelo de regresión logística se representa en la siguiente relación, siendo $p_1$la probabilidad de que ocurra *Y=1*:

$$p_1 = P(Y = 1 \parallel x_1, x_2, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

De lo que se deduce:

$$p_0 = 1 - p_1 = P(Y = 0 \parallel x_1, x_2, \dots, x_m) = \frac{e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

Con lo que obtenemos el logaritmo de la razón de probabilidades u *odds ratio* (logit):

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m;\ p_1 = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 (x=1))}}$$

$$odds\ (x = 1) = \frac{p\ (x = 1)}{1 - p(x = 1)} = \ e^{\beta_0 + \beta_1}$$

$$odds\ (x = 0) = \frac{p\ (x = 0)}{1 - p(x = 0)} = \ e^{\beta_0}$$

$$\boldsymbol{Odds\ Ratio = \frac{odds\ (x = 1)}{odds\ (x = 0)} = \ e^{\beta_1}}$$

En la Figura 4.1.1 se encuentra la representación gráfica de una Regresión Logística:

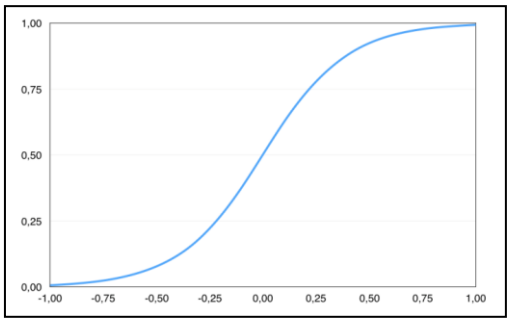


Figura 4.1.1. Representación gráfica de la función de Regresión Logística.

Redes Neuronales

Las Redes Neuronales son un método estadístico basado en un conjunto de unidades neuronales que imitan la estructura y comportamiento de las neuronas en el cerebro. La información que entra en la red neuronal se analiza y transforma, generando valores de salida.

El esquema más sencillo que representa las redes neuronales se encuentra en la Figura 4.1.2.:

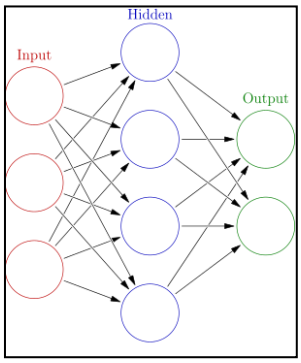


Figura 4.1.2. Representación gráfica de una Red Neuronal

En las conexiones de la red neuronal se distinguen tres capas: la **capa de entrada**, que contiene los nodos con información que proviene del exterior; la **capa oculta** en la que se realizan las transformaciones de la información del exterior para dar lugar a la **capa salida**, que aporta los nodos con información para el exterior.

Random Forest

Método de predicción basado en la combinación de una gran cantidad de árboles de decisión, en el que se construyen árboles no correlacionados para realizar posteriormente un promedio.

Se trata de generar multitud de árboles de decisión, construyéndose a partir de datos de entrada distintos, por lo que se debe alterar el conjunto inicial de partida.

En definitiva, se hace una media final de los resultados de cada árbol de decisión, representada gráficamente en la Figura 4.1.4.:

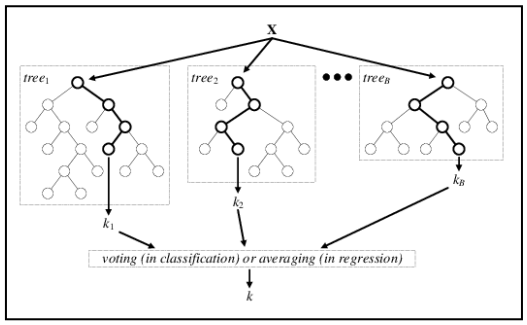


Figura 4.1.4. Representación gráfica de árboles no correlacionados y la obtención promedio

Un método particular de Random Forest es el método de predicción ***Bagging***. Sigue el mismo algoritmo que Random Forest, con la diferencia que en Random Forest se selecciona aleatoriamente el conjunto de predictores.
En caso de que los predictores escogidos sean los totales, entonces el método Bagging y Random Forest son equivalentes.

Gradient Boosting

Se trata de un método estadístico de predicción en el que se realiza un modelo predictivo basado en un conjunto de modelos de predicción débiles, principalmente árboles de decisión.
Se trata de construir un modelo de forma escalonada, repitiendo arboles de decisión, de manera que se van minimizando los errores y ajustando las predicciones en la construcción de dichos árboles.

El aumento de gradiente va construyendo un conjunto de árboles, uno a uno, para posteriormente combinar las predicciones de dichos árboles individuales. En la Figura 4.1.3. se representa gráficamente la combinación de árboles de decisión individuales:

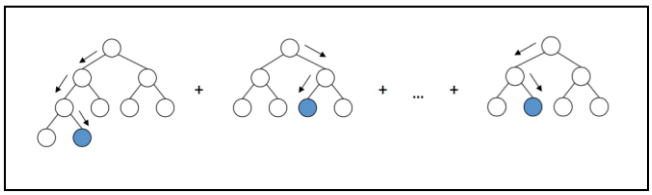


Figura 4.1.3. Representación gráfica de combinación de árboles de decisión individuales

Cada nuevo árbol que va creando aprende de los errores del modelo anterior, tratando de minimizar el error del conjunto y complementar a los árboles ya existentes. De esta manera, cuantos más árboles se construyan, el residuo es cada vez más pequeño, aunque no se debe abusar del número de árboles, porque puede dar lugar a sobreajuste. Además, se crea un árbol residual, con la diferencia entre la función objetivo y las predicciones del modelo.

Ensamblado de Modelos

El Ensamblado de modelos consiste en la construcción de un modelo más exacto a partir de la combinación de varios modelos y su ensamblado combinándolos entre ellos.

Los modelos deben tener una baja correlación entre ellos para que el modelo ensamblado obtenga lo mejor de cada modelo.

En este trabajo se combinarán los mejores modelos de cada uno métodos de predicción descritos.

5. Variables

5.1. Análisis exploratorio de las variables

Comenzamos con un análisis exploratorio de las variables disponibles, previo a la fase de predicción, para poder comprender y conocer mejor la información a predecir y realizar las modificaciones que se consideren necesarias.
Partimos de un total de 8 variables iniciales, que son las siguientes:

* **Tipo Producto**: variable que indica el producto contratado por el cliente en esa operación. Se trata de una variable categórica con 18 niveles y con grandes diferencias en las frecuencias de contratación, como se observa en la Tabla 5.1.1:

TIPO DE PRODUCTO	% DE FRECUENCIA
TARJETA VISA	**22,14%**
FONDO DE INVERSIÓN	**19,93%**
CUENTA DE AHORRO	**17,28%**
CUENTA CORRIENTE	**13,73%**
PRESTAMO	**9,04%**
PLAN PENSIONES	**4,03%**
OTROS (***11 tipos***)	**13,86%**

Tabla 5.1.1. Frecuencias de las categorías de la variable Tipo Producto

La frecuencia de la contratación de los productos está concentrada, principalmente, en 6 tipos de producto, ya que, de los 18 niveles, nos encontramos con 11 para los que la suma de la frecuencia apenas llega al 14% de la frecuencia total.

* **Fecha de Apertura:** variable que aporta la fecha en la que se ha formalizado la operación. Se trata de una variable que nos aporta 3 datos: el día, el mes y el año. Por lo que más adelante, desglosaremos esta variable en tres. En la Tabla 5.1.1. se representa el diagrama de barras con la frecuencia de las categorías de la variable:

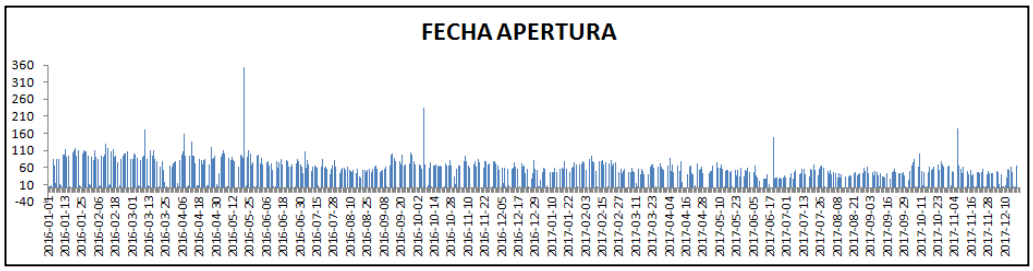


Figura 5.1.1. Diagrama de barras de la frecuencia de las categorías de la variable Fecha Apertura

* **Canal:** se trata de una de las variables más importantes para el proyecto, que nos aporta el medio por el cual el cliente ha realizado la operación de contratación. De esta variable obtendremos la variable objetivo (binaria).
Los niveles de esta variable son 3, y el canal de uso más frecuente es el de "Sucursal" frente a los otros (Banca Electrónica y Teléfono).

Las frecuencias de la variable y sus porcentajes sobre el total se indican en la Tabla 5.1.2:

CANAL	FRECUENCIA	%
BANCA ELECTRONICA	8.298	**24,92%**
SUCURSAL	24.681	**74,12%**
TELEFONICO	319	**0,96%**

Tabla 5.1.2. Frecuencias de la categorías de la variable Canal

* **Nivel de Estudios:** variable con el detalle del nivel de estudios de los clientes registrados.
En esta variable no hay grandes diferencias en las frecuencias de los distintos niveles, lo que nos aportará información de valor sobre el perfil del cliente.

En la Figura 5.1.2. a continuación se representa un diagrama de barras horizontales con la frecuencia de cada una de las categorías de esta variable:

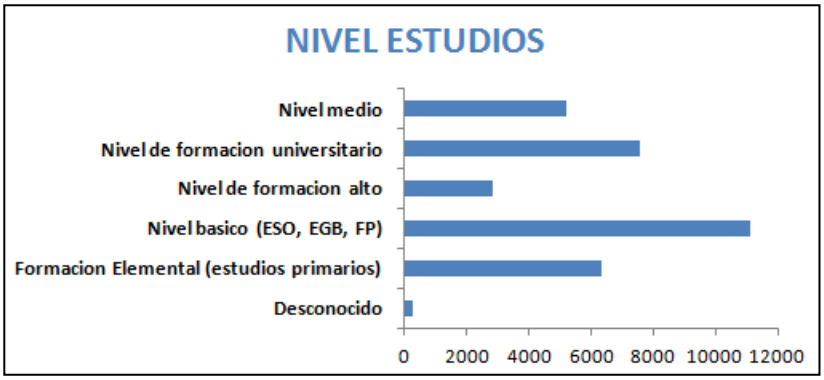


Figura 5.1.2. Diagrama de barras horizontales de la frecuencia de las categorías de la variable Nivel Estudios

* **Sexo:** información sobre el sexo de los clientes que operan.

En la Tabla 5.3.1 se indica la frecuencia de ambas categorías:

SEXO	FRECUENCIA	%
HOMBRE	20.651	**62,02%**
MUJER	12.647	**37,98%**

Tabla 5.1.3. Frecuencias de las categorías de la variable Sexo

* **Estado Civil:** variable que indica el estado civil de los clientes. Existen grandes diferencias en la frecuencia de los estados, teniendo la mitad de las observaciones estado civil “Casado/a”.

Los demás estados se reparten el 50% restante, por lo que, en la depuración de datos, analizaremos las agrupaciones óptimas para los niveles de esta variable.

La Tabla 5.1.4 recoge la frecuencia de cada uno de los estados civiles de los clientes, y el porcentaje de frecuencia sobre el total:

ESTADO CIVIL	FRECUENCIA	%
Casado/a	17.004	**51,07%**
Soltero/a	12.503	**37,55%**
Divorciado/a	1.709	**5,13%**
Viudo/a	1.331	**4,00%**
Separado/a legal	589	**1,77%**
Separado/a de hecho	155	**0,47%**
Desconocido	7	**0,02%**

Tabla 5.1.4. Frecuencias de las categorías de la variable Estado Civil

* **Número de Hijos:** información sobre los hijos que tienen los clientes.

En la Figura 5.1.3. observamos que más del 70% no tienen hijos en el momento de la contratación, y únicamente un 4% tienen más de 3 hijos.

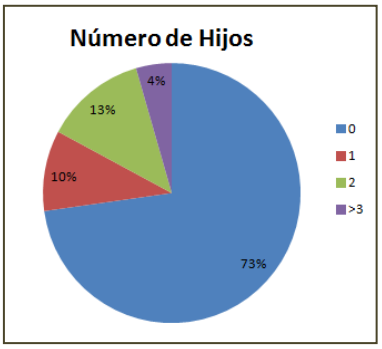


Figura 5.1.3. Gráfico circular de la frecuencia de las categorías de la variable Número de Hijos

* **Edad**: el estudio del comportamiento del uso de los canales bancarios de los clientes y su edad nos permite conocer el comportamiento de la población futura. Esta información contiene datos de gran relevancia para el estudio.

En la Figura 5.1.4 observamos posibles valores fuera de rango (0, 101, 105…) que se limpiarán en la fase de depuración de datos.

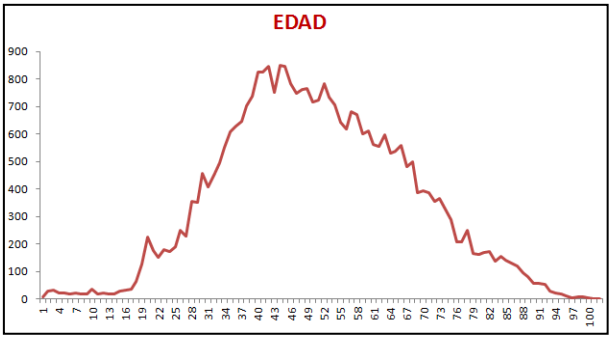


Figura 5.1.4. Gráfico de la frecuencia de las categorías de la variable Edad

5.2. Creación de nuevas variables

A raíz de las 8 variables originales se crean 5 variables adicionales que podrán aportar información útil para el estudio del proyecto.
En primer lugar, a partir de la variable "Fecha de Apertura" del contrato podemos obtener 3 nuevas variables: Día de Apertura, Mes de Apertura y Año de Apertura.

De esta manera, podremos analizar, de un solo dato como es la fecha, tres aspectos que pueden influir en la decisión de elección del canal de contratación.

En la Tabla 5.2.1 observamos las características de las nuevas variables obtenidas de la variable de origen "Fecha de Apertura":

Variable Origen	Nueva Variable	Tipo de Variable	Niveles
Fecha de Apertura	Día de Apertura	Intervalo	1-31
	Mes de Apertura	Clase	Enero-Diciembre
	Año de Apertura	Intervalo	2016-2017

Tabla 5.2.1. Creación de nuevas variables desde la variable Fecha Apertura

A continuación, de la variable de clase "Canal" con 3 niveles, se obtiene la nueva variable que será la variable objetivo del estudio: "Canal Objetivo". Se trata de una variable dicotómica que puede tomar dos valores posibles:

* SI: si el canal de contratación es on-line;
* NO: si el canal de contratación es en sucursal

En la Tabla 5.2.2 observamos las características de la nueva variable obtenida de la variable de origen "Canal":

Variable Origen	Nueva Variable	Tipo de Variable	Niveles
Canal	Canal Objetivo	Dicotómica (Objetivo)	**Si**: Banca Electrónica **No**: Sucursal y Teléfono (Desconocido)

Tabla 5.2.2. Creación de nuevas variables desde la variable Canal

En esta nueva variable, destacan las diferencias en las frecuencias de ambos niveles, representadas en la Tabla 5.2.3:

CANAL OBJETIVO	%
SI	**24,92%**
NO	**75,07%**

Tabla 5.2.3. Frecuencias de las categorías de la nueva variable Canal Objetivo

Por último, de la variable "Número de Hijos" obtenemos una variable dicotómica en la que agrupamos la información en dos valores: "Hijos Sí" e "Hijos No".

En la Tabla 5.2.3 observamos las características de la nueva variable obtenida de la variable de origen "Número de Hijos":

Variable Origen	Nueva Variable	Tipo de Variable	Niveles
Número de Hijos	Hijos	Dicotómica	Si/No

Tabla 5.2.3. Creación de la nueva variable desde la variable Número de Hijos

Tras estas modificaciones se obtiene el fichero final con el que se va a trabajar, formado por un total de 33.298 observaciones y 13 variables.

5.3. Depuración de las variables originales

Tras el análisis exploratorio de las variables y la creación de variables adicionales, se procede a depurar los datos, con el fin de corregir las incorrecciones en los datos, que puedan entorpecer el estudio estadístico del proyecto.
Se realizará con el Software SAS Enterprise Miner Workstation 14.1, analizando la presencia de missings, categorías incorrectas, valores fuera de rango, etc.

En primer lugar, obtenemos mediante el Nodo "DMDB" las principales características tanto de las variables de intervalo como de clase, representadas en las Tabla 5.3.1 y Tabla 5.3.2:

* ***Variables de Intervalo***:

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar
ANO_APERTURA		0	33298	2016	2017	2016.392	0.488172
DIA_APERTURA		0	33298	1	31	15.74013	8.687094
EDAD	EDAD	0	33298	0	105	49.73058	16.38401
NUMERO_DE_HIJOS		0	33298	0	20	0.507298	0.959158

Tabla 5.3.1. Características de las Variables de Intervalo

En los datos importados observamos que, para las variables de intervalo, no hay datos ausentes (missings). Se trata de datos de obligado cumplimiento para el Banco, por lo que el sistema de censado garantiza la recogida de todos los datos sin permitir campos vacíos. Por ello, lo que se debe analizar es la calidad de los datos recogidos.

Esta tabla aporta también información sobre los posibles valores fuera de rango:

- La variable EDAD tiene un rango de 0 a 105, lo que implica valores erróneos.
- La variable NUMERO_DE_HIJOS tiene un valor máximo de 20, lo que muestra que pueden no ser datos correctos.

* ***Variables de clase***:

Variable	Etiqueta	Tipo	Número de niveles	Ausente
CANAL_OBJETIVO		C	2	0
ESTADO_CIVIL		C	7	0
HIJOS	HIJOS	C	2	0
MES_APERTURA		C	12	0
NIVEL_ESTUDIOS		C	6	0
SEXO	SEXO	C	2	0
TIPO_PRODUCTO		C	18	0

Tabla 5.3.2. Características de las Variables de Clase

En los datos importados de las variables de clase observamos que tampoco hay datos ausentes.

Se indican los niveles de cada variable y, según lo analizado en la descripción de las variables, las categorías son correctas.

En primer lugar, observamos gráficamente en las Figuras 5.3.1 y 5.3.2 las variables en las que se han detectado valores fuera de rango:

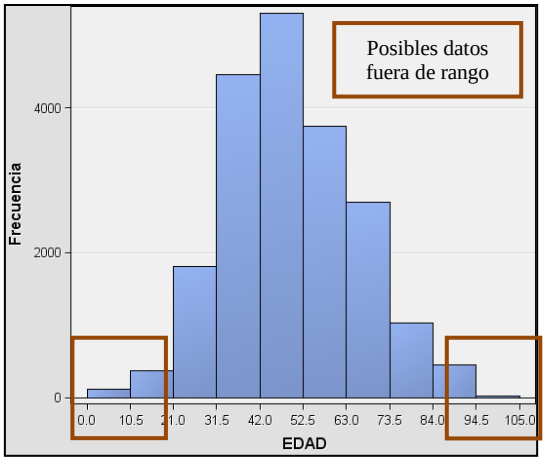


Figura 5.3.1. Frecuencia de la Variable Edad

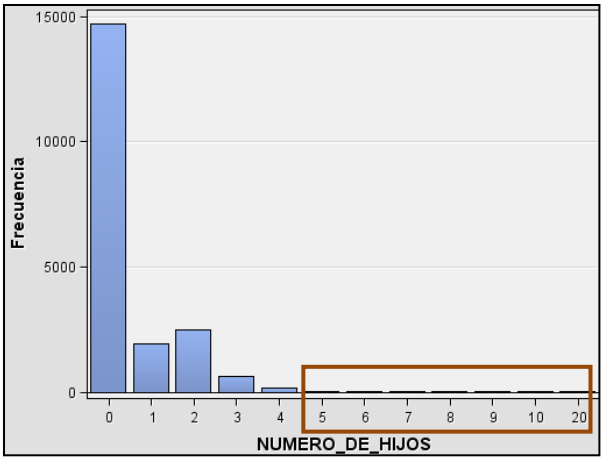


Figura 5.3.2. Frecuencia de la Variable Número Hijos

Además, analizamos los gráficos de relación entre las variables de clase con la variable objetivo.

* *Relación entre la variable CANAL_OBJETIVO y ESTADO_CIVIL,* representada en la Figura 5.3.3:

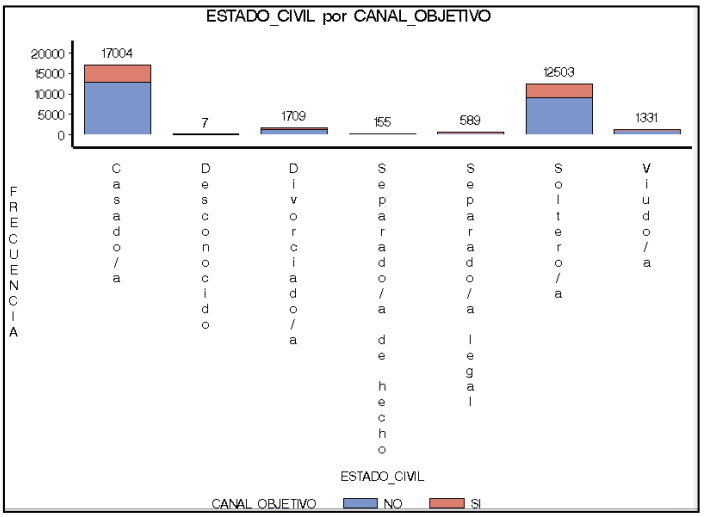


Figura 5.3.3. Frecuencia de las categorías de las variables Estado Civil y Canal Objetivo

Destaca la contratación en ambos canales (sucursal y online) en los clientes con Estado Civil Casado/a y Soltero/a.

* *Relación entre la variable CANAL_OBJETIVO e HIJOS*

En la Tabla 5.3.3. se representa la frecuencia de las categorías de las variables Hijos (Si/No) y Canal Objetivo (Si/No). Observamos que el 25,76% del total de clientes sin hijos (24.234) operan en el canal objetivo (online), y del total de los clientes con hijos (9.064), el 22,67% operan en este canal.

Con ello deducimos que el volumen de contratación en el canal objetivo (Si) es mayor en los clientes sin hijos.

HIJOS/ CANAL OBJETIVO	CANAL OBJETIVO (NO)	CANAL OBJETIVO (SI)	TOTAL	% OBJETIVO SI S/NIVEL HIJOS
HIJOS (NO)	17.991	6.243	24.234	**25,76%**
HIJOS (SI)	7.009	2.055	9.064	**22,67%**
Total	**25.000**	**8.298**	**33.298**	

Tabla 5.3.3. Frecuencia de las categorías de las variables Hijos y Canal Objetivo

* *Relación entre la variable CANAL_OBJETIVO y MES_APERTURA:*

En la Tabla 5.3.4 se representan las frecuencias de las categorías de las variables en la que se observa una mayor contratación en los primeros meses del año (Enero-Mayo). Además, analizamos que, sobre el total de contratación de cada mes, es mayor la contratación por el canal objetivo en los meses de Enero, Abril y Diciembre, y menor en los meses de Julio y Septiembre.

MES/CANAL OBJETIVO	CANAL OBJETIVO (NO)	CANAL OBJETIVO (SI)	TOTAL	% TOTAL OBJETIVO S/NIVEL MES
ENERO	2.287	899	3.186	**28,22%**
FEBRERO	2.690	876	3.566	**24,57%**
MARZO	2.327	863	3.190	**27,05%**
ABRIL	2.275	856	3.131	**27,34%**
MAYO	2.549	854	3.403	**25,10%**
JUNIO	1.850	637	2.487	**25,61%**
JULIO	1.986	423	2.409	**17,56%**
AGOSTO	1.637	444	2.081	**21,34%**
SEPTIEMBRE	2.074	465	2.539	**18,31%**
OCTUBRE	2.172	662	2.834	**23,36%**
NOVIEMBRE	2.012	716	2.728	**26,25%**
DICIEMBRE	1.141	603	1.744	**34,58%**
Total general	25.000	8.298	33.298	

Tabla 5.3.4. Frecuencia de las categorías de las variables Mes y Canal Objetivo

* *Relación entre la variable CANAL_OBJETIVO y NIVEL_ESTUDIOS*:

En la Tabla 5.3.5 de frecuencias de las categorías de las variables observamos que el mayor porcentaje, sobre el total de contratación de cada nivel de estudios, es el del Nivel de formación universitario, además de ser el mayor porcentaje de contratación total (sin tener en cuenta el nivel "Desconocido").

ESTUDIOS/CANAL OBJETIVO	CANAL OBJETIVO (NO)	CANAL OBJETIVO (SI)	TOTAL	% TOTAL OBJETIVO S/NIVEL ESTUDIOS
Desconocido	117	153	270	**56,67%**
Formación Elemental (estudios primarios)	5.666	681	6.347	**10,73%**
Nivel básico (ESO, EGB, FP)	9.520	1.593	11.113	**14,33%**
Nivel de formación alto	1.838	993	2.831	**35,08%**
Nivel de formación universitario	4.139	3.405	7.544	**45,14%**
Nivel medio	3.720	1.473	5.193	**28,37%**
Total general	25.000	8.298	33.298	

Tabla 5.3.5. Frecuencia de las categorías de las variables Estudios y Canal Objetivo

* *Relación entre la variable CANAL_OBJETIVO y SEXO:*

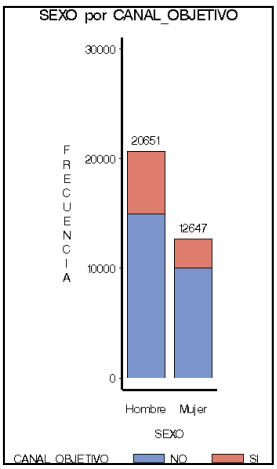


Figura 5.3.4. Frecuencia de las categorías de las variables Sexo y Canal Objetivo

En la figura 5.3.4 se representa la frecuencia de las categorías de las variables Canal_Objetivo y Sexo.
Es mayor la contratación por parte de clientes hombres, y también es más frecuente el uso del canal on-line en estos.

* *Relación entre la variable CANAL_OBJETIVO y TIPO_PRODUCTO:*

Analizamos en la Tabla 5.3.6 los porcentajes de contratación del canal objetivo para cada uno de los niveles de la variable. Observamos que hay ciertos productos que no tienen ninguna contratación mediante el canal objetivo:

TIPO PRODUCTO/CANAL OBJETIVO	CANAL OBJETIVO (NO)	CANAL OBJETIVO (SI)	TOTAL	% TOTAL OBJETIVO S/TIPO PRODUCTO
TARJETAS EXTERNAS	10	19	29	65,52%
FONDO DE INVERSION	258	329	587	56,05%
FONDO DE INVERSIÓN	3.745	2.892	6.637	43,57%
CUENTA DE AHORRO	3.853	1.901	5.754	33,04%
CUENTA CORRIENTE	3.347	1.224	4.571	26,78%
PLAN DE PENSIÓN INDIVIDUAL	651	142	793	17,91%
PRESTAMO	2.517	492	3.009	16,35%
TARJETA PRIVADA	234	41	275	14,91%
TARJETA VISA	6.318	1.054	7.372	14,30%
PLAN DE PENSION INDIVIDUAL	440	60	500	12,00%
PLAN PENSIONES	1.197	144	1.341	10,74%
FINANCIACIÓN DE EXPORTACIONES EN MONEDA EXTRANJERA	1	0	1	0,00%
HIPOTECAS	395	0	395	0,00%
PLAN DE EMPLEO	113	0	113	0,00%
PLAN DE PENSIÓN COLECTIVO	1	0	1	0,00%
PRESTAMO PERSONAL	1.183	0	1.183	0,00%
SEGUROS	736	0	736	0,00%
SEGUROS DE CAMBIO	1	0	1	0,00%
Total general	**25000**	**8298**		**33298**

Tabla 5.3.6. Frecuencia de las categorías de las variables Tipo Producto y Canal Objetivo

Observamos que dos de las categorías deberían unificarse (Fondo de Inversión y Plan de Pensión Individual), ya que al estar escritas de diferente manera se desglosan en 2.

Para comprender y estudiar las variables, observamos también gráficos con otras relaciones entre las variables restantes:

* *Relación entre la variable DIA_APERTURA y TIPO_PRODUCTOS*:

En la Figura 5.3.5 observamos que hay productos que se contratan durante todo el mes (tarjetas visa, cuenta corriente, cuenta de ahorro…) y hay otros productos que se concentran en unos días (plan de empleo, tarjetas externas, etc.).

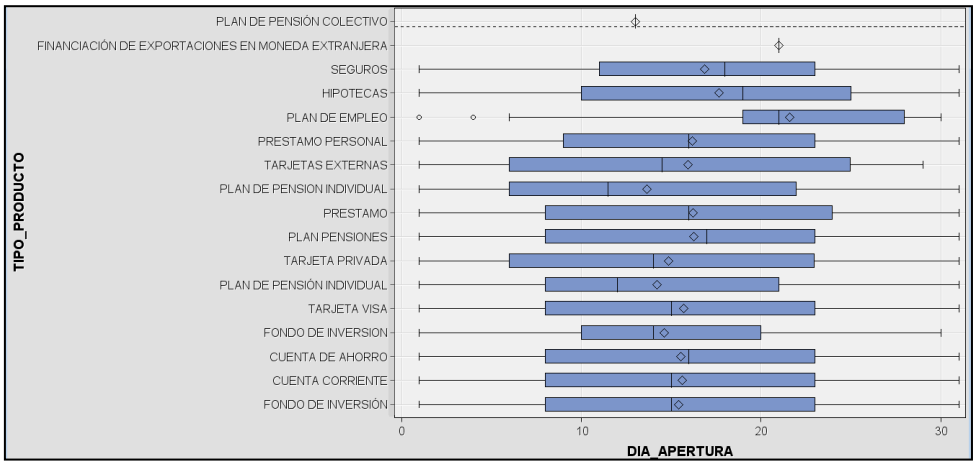


Figura 5.3.5. Frecuencia de las categorías de las variables Tipo Producto y Día Apertura

* Relación de las variables que aportan *más información sobre la variable objetivo* representadas en la Figura 5.3.6:

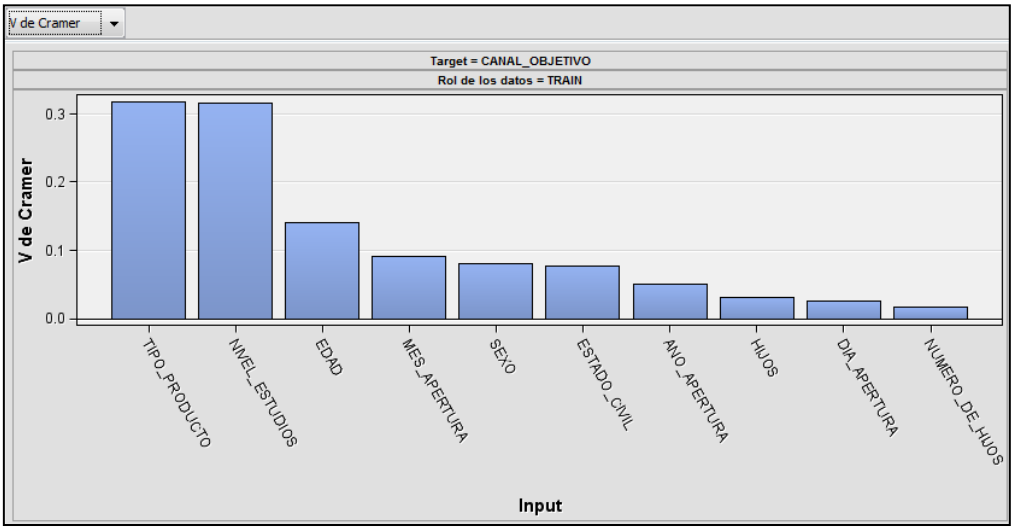


Figura 5.3.6. V de Cramer- Relación entre las variables con la variable Objetivo

La V de Cramer es una medida de asociación entre dos variables, basado en el estadístico Chi-Cuadrado de Pearson, siendo su valor entre 0 (no asociación) y 1 (total asociación). En este caso, las variables con mayor asociación con la variable objetivo son el TIPO_PRODUCTO y NIVEL_ESTUDIOS, con una V de Cramer de 0.3, seguido de la EDAD y MES_APERTURA.

Tras el análisis de las relaciones de las variables con la variable objetivo, y algunas relaciones entre ellas, procedemos a la modificación de los errores que se han detectado hasta el momento.

* *Categorías erróneas*: en la variable TIPO_PRODUCTO hay dos categorías "Fondo de Inversión", una con tilde y otra sin tilde. Mediante el Nodo Reemplazo se unifica la categoría indicando el mismo nombre.
  Lo mismo ocurre para las categorías "Préstamo Personal" y "Plan de Pensión Individual".

* *Datos fuera de rango*:

  - ✓ EDAD: se establece un límite inferior de 5 y superior de 95.
  - ✓ NUMERO_HIJOS: se establece un máximo de 8 hijos.

* *Porcentaje mínimo*: para las variables de clase se exige un porcentaje mínimo de observaciones por categoría, si no se supera, se eliminan las observaciones de dichas categorías.

  - ✓ ESTADO_CIVIL: se eliminan las observaciones de las categorías "Desconocido" y "Separado/a de hecho". Quedando la variable con 5 niveles.
  - ✓ NIVEL_ESTUDIOS: se eliminan las observaciones de la categoría "Desconocido". Quedando la variable con 5 niveles.
  - ✓ TIPO_PRODUCTO: se eliminan las observaciones de las categorías "Financiación de Exportaciones en M.E.", "Plan de Empleo", "Seguros de Cambio", "Tarjeta Privada", "Tarjetas Externas", etc.

Con estos reemplazos y eliminaciones, queda reducido el número de observaciones, quedando en 32.479.

Mediante el Nodo DMDB obtenemos las Tablas 5.3.7 y 5.3.8 con los datos finales tras las modificaciones en las variables:

Variable	Etiqueta	N	Ausente	Mínimo	Máximo	Media	Desviación estándar
ANO_APERTURA		32479	0	2016	2017	2016.395	0.488936
DIA_APERTURA		32479	0	1	31	15.72148	8.681944
REP_EDAD	Replacement: EDAD	32479	0	5	95	49.72475	16.39086
REP_NUMERO_DE_HIJOS	Replacement: NUMERO_DE_HIJOS	32479	0	0	8	0.506758	0.950111

Tabla 5.3.7. Características de las Variables de Intervalo

Variable	Etiqueta	Tipo	Número de niveles	Ausente
CANAL_OBJETIVO		C	2	0
ESTADO_CIVIL		C	5	0
HIJOS	HIJOS	C	2	0
MES_APERTURA		C	12	0
NIVEL_ESTUDIOS		C	5	0
REP_TIPO_PRODUCTO	Replacement: TIPO_PRODUCTO	C	9	0
SEXO	SEXO	C	2	0

Tabla 5.3.8. Características de las Variables de Clase

5.4. Análisis de Correspondencias

Importamos el fichero con las modificaciones en SAS 9.4, para realizar un análisis de correspondencias simple entre las variables de entrada y la variable objetivo.

Relación de las variables NIVEL DE ESTUDIOS, TIPO PRODUCTO y CANAL_OBJETIVO:

Mediante el código SAS (**10.2.1.1. /*IMPORTAR DATOS */** y **10.2.1.2. /*PROCEDIMIENTO ACS*/**) importamos los datos a estudiar, y ejecutamos el procedimiento *PROC CORRESP*, identificando como niveles las categorías de la variable NIVEL DE ESTUDIOS, así como el cruce del Tipo de Producto y el Canal.

En la 5.4.1. se muestra la tabla de contingencias con los datos de entrada para el ACS.

Se puede comprobar, por ejemplo, que el 4,21% de los clientes tienen un nivel de Formación Elemental y contratan Tarjetas Visa en la Sucursal, o que hay un 5,059% de clientes con nivel de Formación Universitario que contratan Fondos de Inversión on-line.

Contingency Table						
Percents	**Formacion Elemental (estudios primarios)**	**Nivel basico (ESO, EGB, FP)**	**Nivel de formacion alto**	**Nivel de formacion universitario**	**Nivel medio**	**Sum**
NOCTACOR	1.940	3.673	0.890	2.124	1.564	**10.191**
NOCTAAHO	3.793	4.310	0.745	1.389	1.521	**11.758**
NOFONDO	2.925	3.916	1.065	2.346	2.017	**12.269**
NOHIPO	0.160	0.511	0.123	0.228	0.188	**1.210**
NOPLANPE	1.244	2.885	0.570	1.179	1.139	**7.017**
NOPRESTA	2.565	4.748	0.631	1.727	1.576	**11.247**
NOSEGURO	0.382	1.025	0.157	0.320	0.376	**2.260**
NOTARVIS	4.212	7.688	1.364	3.153	2.830	**19.246**
SICTACOR	0.477	0.930	0.366	1.179	0.760	**3.713**
SICTAAHO	0.551	0.742	0.991	2.248	0.973	**5.505**

Contingency Table						
Percents	**Formacion Elemental (estudios primarios)**	**Nivel basico (ESO, EGB, FP)**	**Nivel de formacion alto**	**Nivel de formacion universitario**	**Nivel medio**	**Sum**
SIFONDO	0.594	1.623	1.004	5.059	1.586	**9.865**
SIPLANPE	0.040	0.108	0.105	0.533	0.206	**0.991**
SIPRESTA	0.142	0.530	0.197	0.342	0.286	**1.496**
SITARVIS	0.274	0.893	0.329	1.047	0.687	**3.230**
Sum	**19.299**	**33.582**	**8.538**	**22.873**	**15.709**	**100.000**

Tabla 5.4.1. Tabla de Contingencias de las variables Nivel Estudios, Canal y Tipo Producto

Con esta tabla comprobamos la idoneidad de los datos y sus categorías, considerando que estas son idóneas siempre que los porcentajes de las filas y columnas de la tabla de contingencias superen el 5%.
En este caso, las columnas superan el 5%, pero de las 14 filas, 6 no superan el porcentaje, por lo que se tendrán que agrupar.

Se unificarán las dos variables de CANAL_OBJETIVO "No" (No-hipoteca y No-Seguro) a la variable "No- Plan de Pensiones". Por otro lado, agruparemos las 4 variables de CANAL_OBJETIVO "Si" que no superan el 5%.

Tras las agrupaciones de las filas indicadas (**10.2.1.3. /*AGRUPACIÓN DE LAS FILAS */**) y realizar de nuevo el procedimiento *PROC CORRESP* (**10.2.1.4. /*PROCEDIMIENTO ACS CON FILAS AGRUPADAS*/**) indicando el número de componentes a retener, observamos que todos los porcentajes superan el 5%. (Tabla 10.1.1.)

En la Figura 5.4.1. se muestra la Descomposición de la Inercia, así como las distancias Chi-Cuadrado.
En este caso, la inercia total toma un valor de 0,12707 y el estadístico Chi-Cuadrado el valor de 4.127,08. Además, se indica el p-valor del test de Pearson, probabilidad de equivocarse al rechazar la hipótesis de independencia de los datos, y en este caso es menor que 0,0001. De esta manera, se puede rechazar la hipótesis de independencia a cualquier nivel de significación razonable.

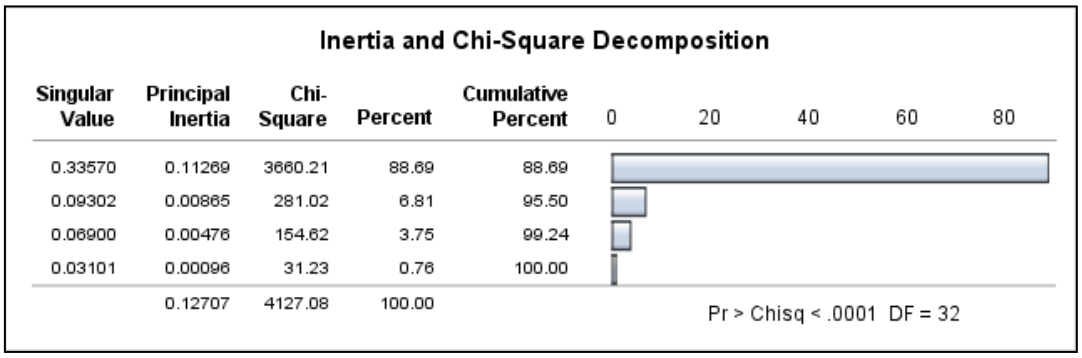


Figura 5.4.1. Inercia y Descomposición del Chi-Cuadrado

Con el fin de definir el número de factores a retener utilizamos a continuación los métodos indicados en la metodología:

- Tomar los primeros ejes significativos, que expliquen la variabilidad aceptable (>60%), con este criterio nos quedaríamos únicamente con el primer factor.

- Elegir las dimensiones correspondientes a las inercias mayores que **0,0317** (0,12707/4), por lo que nos quedaríamos solo con el primer factor.

Al realizar SAS por defecto el modelo con dos dimensiones, comprobamos en la Tabla 54.2. los datos de las calidades para decidir:

Summary Statistics for the Row Points			
	Quality	**Mass**	**Inertia**
NOCTACOR	0.8114	0.1019	0.0030
NOCTAAHO	0.9889	0.1176	0.1425
NOFONDO	0.8808	0.1227	0.0174
NOPRESTA	0.9573	0.1125	0.0576
NOHISPL	0.9343	0.1049	0.0354
NOTARVIS	0.9941	0.1925	0.0561
SICTAAHO	0.9189	0.0551	0.1791
SIFONDO	0.9641	0.0986	0.4151
SICPPT	0.9332	0.0943	0.0937

Tabla 5.4.2. Tabla de estadísticos del perfil fila de las variables Nivel Estudios, Canal y Tipo Producto

En las variables de fila, la calidad de las categorías es superior al 80%, por lo que, de momento, damos por valido utilizar dos dimensiones.

En relación con las columnas, en la Tabla 5.4.3 observamos que tienen porcentajes de calidad más bajos, en especial la categoría “Nivel Medio”, que no supera el 50%:

Summary Statistics for the Column Points			
	Quality	**Mass**	**Inertia**
Formacion Elemental (estudios primarios)	0.9933	0.1930	0.2283
Nivel basico (ESO, EGB, FP)	0.9944	0.3358	0.1933
Nivel de formacion alto	0.6560	0.0854	0.0697
Nivel de formacion universitario	0.9832	0.2287	0.4889
Nivel medio	0.4864	0.1571	0.0199

Tabla 5.4.3. Tabla de estadísticos del perfil columna de las variables Nivel Estudios, Canal y Tipo Producto

Teniendo en cuenta lo anterior, consideramos que nos podemos quedar con dos dimensiones para explicar las relaciones entre las variables.

En la Tabla 5.4.4. se indican las Contribuciones parciales a la Inercia de las filas. La Dimensión 1 contiene principalmente información de los productos contratados on-line y los préstamos en sucursal y Cuenta de Ahorro, y en la Dimensión 2 se explican la mayoría de los productos contratados en sucursal.

En la Tabla 5.4.5. se indican las Contribuciones parciales a la Inercia de las columnas. La Dimensión 1 explica los niveles de formación Elemental y Universitario y la Dimensión 2 los niveles de formación Elemental y Básico.

Partial Contributions to Inertia for the Row Points		
	Dim1	**Dim2**
NOCTACOR	0.0019	0.0108
NOCTAAHO	0.1299	0.3775
NOFONDO	0.0089	0.1097
NOPRESTA	0.0590	0.0421
NOHISPL	0.0184	0.2457
NOTARVIS	0.0604	0.0335
SICTAAHO	0.1759	0.1269
SIFONDO	0.4513	0.0001
SICPPT	0.0945	0.0537

Tabla 5.4.4. Contribución parcial a la inercia del perfil fila

Partial Contributions to Inertia for the Row Points		
	Dim1	**Dim2**
Formacion Elemental (estudios primarios)	0.2135	0.5486
Nivel basico (ESO, EGB, FP)	0.1876	0.3787
Nivel de formacion alto	0.0480	0.0466
Nivel de formacion universitario	0.5417	0.0034
Nivel medio	0.0092	0.0227

Tabla 5.4.5. Contribución parcial a la inercia del perfil columna

En la Figura 5.4.2. se representa el gráfico de ambas dimensiones:

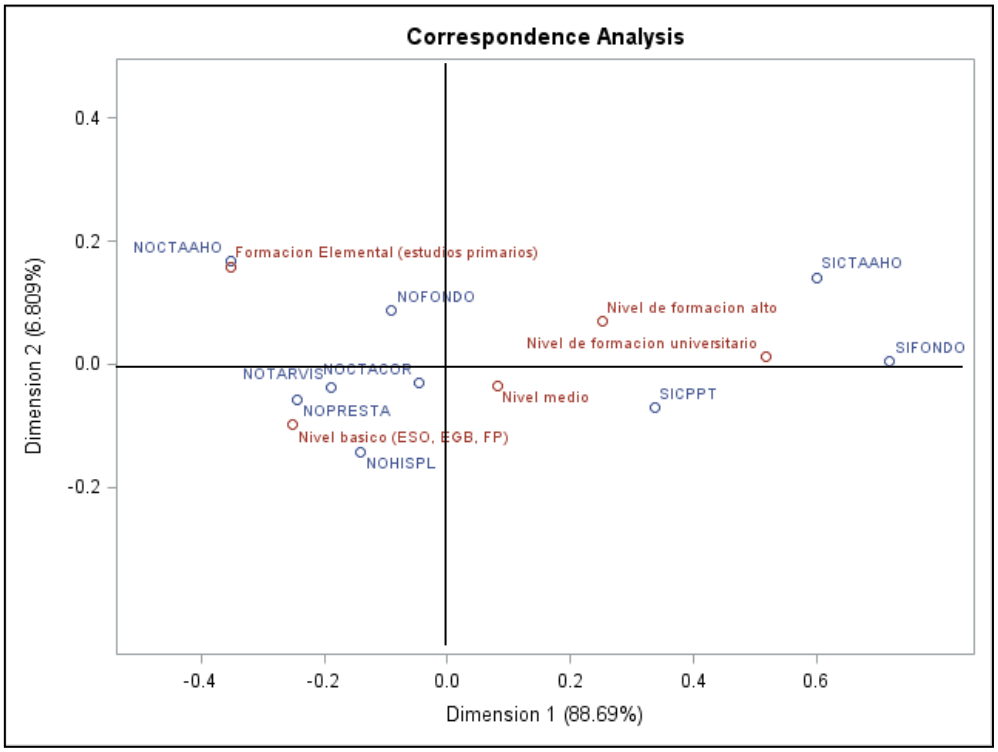


Figura 5.4.2. Gráfico de las dimensiones 1 y 2 del ACS para las variables Nivel de Estudios, Canal y Tipo de Producto

De este gráfico podemos deducir las siguientes conclusiones:

- Para el nivel de ***Formación Elemental*** destaca la contratación en sucursal del producto Cuenta de Ahorro.
Además, este nivel de formación tiene una baja contratación de los productos on-line.
- El nivel de ***Formación Básico*** destaca en la contratación en sucursal del grupo de productos Hipotecas, Seguros y Planes de Pensiones. Tiene un nivel alto de contratación en otros productos de sucursal como los Préstamos, Cuentas Corrientes y Tarjetas Visa. Destaca el elevado número de productos diversos que contrata en sucursal, y al igual que el Nivel de Formación Elemental, tiene una muy baja contratación de los productos on-line.

- El perfil de estudios ***Nivel Medio*** es el grupo de clientes con un nivel muy parecido de contratación online frente a sucursal, no destaca en ninguno de los dos grupos de contratación.

- El nivel de ***Formación Alto*** destaca en la contratación on-line de las Cuentas de Ahorro. Se trata del grupo con menor contratación total de productos.

- El nivel de ***Formación Universitario*** es el grupo con mayor contratación online de todos los productos, especialmente de Fondos de Inversión.

Se trata del grupo de clientes por excelencia que contrata todos los productos del canal objetivo on-line.

Con esto y observando el eje de la Dimensión 1 en el gráfico, vemos que el nivel educativo está ordenado de menor a mayor. Además, los productos contratados en sucursal están a la izquierda y los contratados online están a la derecha. Por lo tanto, concluimos que, a mayor nivel educativo, mayor propensión a la adquisición online.

Relación de las variables ESTADO CIVIL, TIPO PRODUCTO y CANAL_OBJETIVO:

Mediante los códigos SAS (**10.2.1.5. /*IMPORTAR DATOS*/** y **10.2.1.6. /*PROCEDIMIENTO ACS */**) importamos los datos a estudiar, y ejecutamos el procedimiento *PROC CORRESP*, identificando como niveles las categorías de la variable ESTADO CIVIL, así como el cruce del Tipo de Producto y el Canal.

En la Tabla 5.4.6. se muestra la tabla de contingencia con los datos de entrada para el ACS.

Contingency Table						
Percents	**Casado**	**Divorciado**	**Separado**	**Soltero**	**Viudo**	**Sum**
NOCTACOR	4.440	0.671	0.139	4.646	0.296	**10.191**
NOCTAAHO	5.114	0.477	0.157	5.440	0.570	**11.758**
NOFONDO	7.288	0.385	0.200	3.430	0.967	**12.269**
NOHIPO	0.600	0.092	0.034	0.459	0.025	**1.210**
NOPLANPE	4.665	0.360	0.194	1.644	0.154	**7.017**
NOPRESTA	6.013	0.936	0.400	3.347	0.551	**11.247**
NOSEGURO	1.265	0.148	0.077	0.745	0.025	**2.260**
NOTARVIS	9.268	1.105	0.345	7.676	0.853	**19.246**
SICTACOR	1.555	0.194	0.040	1.854	0.071	**3.713**
SICTAAHO	3.039	0.169	0.052	2.106	0.139	**5.505**
SIFONDO	5.296	0.271	0.058	3.990	0.249	**9.865**
SIPLANPE	0.486	0.028	0.009	0.462	0.006	**0.991**
SIPRESTA	0.834	0.132	0.040	0.446	0.043	**1.496**
SITARVIS	1.459	0.203	0.046	1.466	0.055	**3.230**
Sum	**51.322**	**5.173**	**1.792**	**37.711**	**4.003**	**100.000**

Tabla 5.4.6. Tabla de Contingencias de las variables Estado Civil, Canal y Tipo Producto

De nuevo, con esta tabla comprobamos la idoneidad de los datos y sus categorías, considerando que estas son idóneas siempre que los porcentajes de las filas y columnas de la tabla de contingencias superen el 5%.
En este caso, no todas las columnas y filas superan el 5%. Se unificarán las dos variables de CANAL_OBJETIVO "No" (No-hipoteca y No-Seguro) a la variable "No- Plan de Pensiones". Por otro lado, agruparemos las 4 variables de CANAL_OBJETIVO "Si" que no superan el 5%.
En relación con las columnas, agruparemos los estados civiles "Separado" y "Viudo".

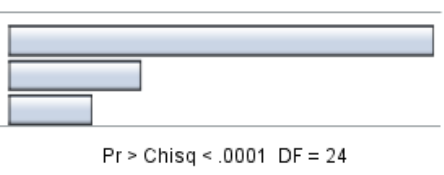
Tras las agrupaciones de las filas indicadas (**10.2.1.7. /*AGRUPACIÓN DE LAS FILAS Y COLUMNAS */**) y realizar de nuevo el procedimiento *PROC CORRESP* (**10.2.1.8. /*PROCEDIMIENTO ACS CON FILAS Y COLUMNAS AGRUPADAS */**) indicando el número de componentes a retener, observamos que todos los porcentajes superan el 5% (Tabla 10.1.2.).

En la Figura 5.4.3. se muestra la Descomposición de la Inercia, así como las distancias Chi-Cuadrado. En este caso, la inercia total toma un valor de 0,03503 y el estadístico Chi-Cuadrado el valor de 1.137,86.

Además, se indica el p-valor del test de Pearson, probabilidad de equivocarse al rechazar la hipótesis de independencia de los datos, y en este caso es menor que 0,0001. De esta manera, se puede rechazar la hipótesis de independencia a cualquier nivel de significación razonable.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.15264	0.02330	756.72	66.50	66.50
0.08495	0.00722	234.36	20.60	87.10
0.06723	0.00452	146.78	12.90	100.00
	0.03503	1137.86	100.00	



Pr > Chisq < .0001 DF = 24

Figura 5.4.3. Inercia y Descomposición del Chi-Cuadrado

A continuación, definimos el número de factores a retener:

- Tomar los primeros ejes significativos, que expliquen la variabilidad aceptable (>60%), con este criterio nos quedaríamos con el primer factor.

- Elegir las dimensiones correspondientes a las inercias mayores que **0,01167** (0,03503/3), por lo que nos quedaríamos solo con el primer factor.

Al realizar SAS por defecto el modelo con dos dimensiones, comprobamos los datos de las calidades para decidir en las Tablas 5.4.7 y 5.4.8:

Summary Statistics for the Row Points			
	Quality	**Mass**	**Inertia**
NOCTACOR	0.9605	0.1019	0.1051
NOCTAAHO	0.7443	0.1176	0.1142
NOFONDO	0.8199	0.1227	0.2444
NOPRESTA	0.9584	0.1125	0.1576
NOHISPL	0.7848	0.1049	0.1648
NOTARVIS	0.9650	0.1925	0.0228
SICTAAHO	0.9870	0.0551	0.0328
SIFONDO	0.9988	0.0986	0.0754
SICPPT	0.8999	0.0943	0.0830

Tabla 5.4.7. Tabla de estadísticos del perfil fila de las variables Nivel Estudios, Canal y Tipo Producto

En las variables de fila, la calidad de las categorías es superior al 70%, por lo que, de momento, damos por valido utilizar dos dimensiones.
En relación con las columnas, también se superan los porcentajes de calidad del 70%:

Summary Statistics for the Column Points			
	Quality	**Mass**	**Inertia**
Casado	0.9770	0.5132	0.2250
Divorciado	0.6597	0.0517	0.1623
Separado-Viudo	0.9866	0.3771	0.3924
Soltero	0.7125	0.0579	0.2203

Tabla 5.4.8. Tabla de estadísticos del perfil columna de las variables Estado Civil, Canal y Tipo Producto

Teniendo en cuenta lo anterior, consideramos que nos podemos quedar con dos dimensiones para explicar las relaciones entre las variables.

En la Tabla 5.4.9. se indican las Contribuciones parciales a la Inercia de las filas. La Dimensión 1 contiene principalmente información de los productos contratados en sucursal (Cta. Corriente, Cta. de Ahorro, Fondo de Inversión, Hipotecas, Seguros y Planes de Pensiones). En la Dimensión 2 los productos online y los préstamos en sucursal.

En la Tabla 5.4.10. se indican las Contribuciones parciales a la Inercia de las columnas. La Dimensión 1 explica los estados civiles de Casados, Separados y Viudos y la Dimensión 2 los estados civiles de Divorciados y Solteros.

Partial Contributions to Inertia for the Row Points		
	Dim1	Dim2
NOCTACOR	0.1396	0.0392
NOCTAAHO	0.1263	0.0049
NOFONDO	0.2984	0.0093
NOPRESTA	0.1180	0.3524
NOHISPL	0.1749	0.0632
NOTARVIS	0.0157	0.0562
SICTAAHO	0.0005	0.1554
SIFONDO	0.0145	0.3191
SICPPT	0.1122	0.0003

Tabla 5.4.9. Contribución parcial a la inercia del perfil fila

Partial Contributions to Inertia for the Column Points		
	Dim1	Dim2
Casado	0.2784	0.1682
Divorciado	0.0000	0.5199
Separado-Viudo	0.5822	0.0001
Soltero	0.1394	0.3118

Tabla 5.4.10. Contribución parcial a la inercia del perfil columna

En la figura 5.4.4 se representa el gráfico de ambas dimensiones:

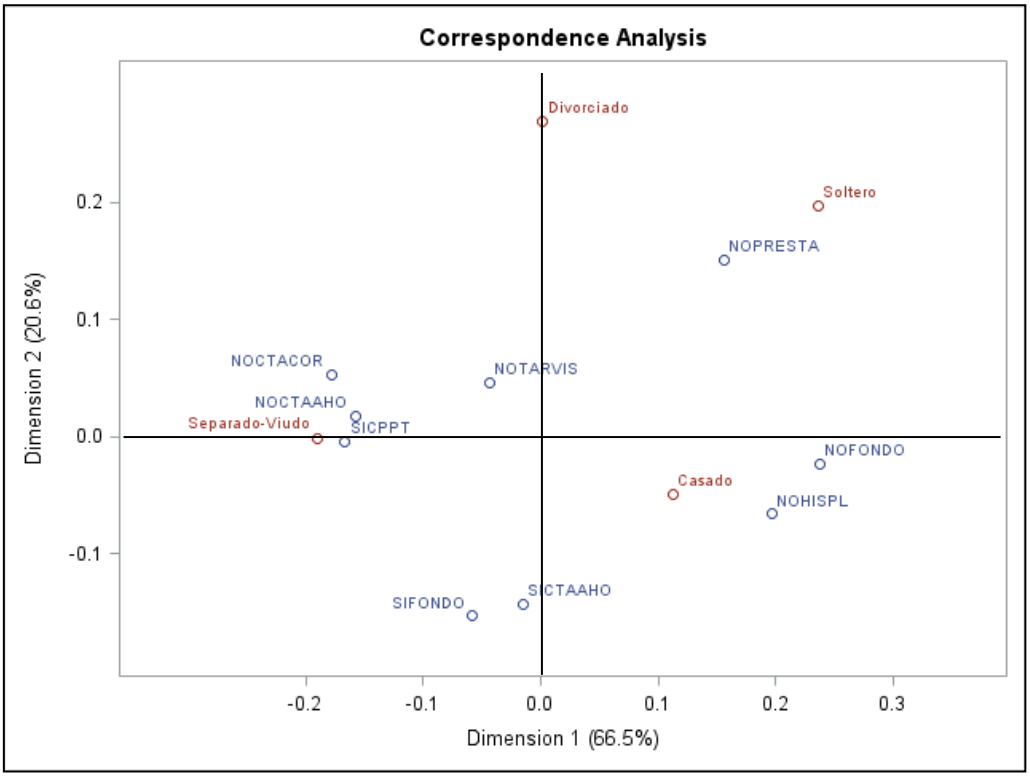


Figura 5.4.4. Gráfico de las dimensiones 1 y 2 del ACS para las variables Estado Civil, Canal y Tipo de Producto

De este gráfico podemos deducir las siguientes conclusiones:

- En el estado civil ***Casado*** destaca la contratación en sucursal, tanto del grupo de productos (Plan de Pensiones, Hipoteca y Seguros) como de los Fondos de Inversión.

- Los ***Divorciados*** tienen un nivel muy bajo de contratación (tanto online como sucursal). Únicamente se podría destacar la contratación de préstamos en sucursal.

- El grupo de los ***Separados-Viudos*** destacan por la contratación en sucursal de Cuenta Corriente y Cuenta de Ahorro.
Es el grupo con mayor contratación on-line (Cuenta Corriente, Préstamos, Planes e Pensiones y Tarjetas Visa)

- Los ***Solteros*** destacan en la contratación de Fondos de Inversión y Préstamos en sucursal.

Relación de las variables EDAD, TIPO PRODUCTO y CANAL_OBJETIVO:

Mediante los códigos SAS (**10.2.1.9. /*IMPORTAR DATOS */** y **10.2.1.10. /*PROCEDIMIENTO ACS */**) importamos los datos a estudiar, y ejecutamos el procedimiento *PROC CORRESP*, identificando como niveles las categorías de la variable EDAD, así como el cruce del Tipo de Producto y el Canal.

En la Tabla 5.4.11. se muestra la tabla de contingencia con los datos de entrada para el ACS.

Contingency Table					
Percents	**low-23**	**24-46**	**47-69**	**70-high**	**Sum**
NOCTACOR	0.951	4.735	3.488	1.016	**10.191**
NOCTAAHO	1.601	4.009	4.187	1.961	**11.758**
NOFONDO	0.228	2.069	5.847	4.126	**12.269**
NOHIPO	0.022	0.779	0.379	0.031	**1.210**
NOPLANPE	0.022	2.322	4.449	0.225	**7.017**
NOPRESTA	0.185	4.899	5.339	0.825	**11.247**
NOSEGURO	0.018	1.204	1.013	0.025	**2.260**
NOTARVIS	1.188	8.504	7.020	2.534	**19.246**
SICTACOR	0.111	2.189	1.216	0.197	**3.713**
SICTAAHO	0.028	2.097	2.780	0.600	**5.505**
SIFONDO	0.065	4.735	4.187	0.877	**9.865**
SIPLANPE	0.000	0.533	0.459	0.000	**0.991**

Contingency Table					
Percents	**low-23**	**24-46**	**47-69**	**70-high**	**Sum**
SIPRESTA	0.000	0.637	0.822	0.037	**1.496**
SITARVIS	0.135	1.854	1.102	0.139	**3.230**
Sum	**4.554**	**40.565**	**42.289**	**12.593**	**100.000**

Tabla 5.4.11. Tabla de Contingencias de las variables Edad, Canal y Tipo Producto

Con esta tabla comprobamos la idoneidad de los datos y sus categorías, considerando que son idóneas siempre que los porcentajes de las filas y columnas de la tabla de contingencias superen el 5%.

En este caso, no todas las columnas y filas superan el 5%. Se unificarán las dos variables de CANAL_OBJETIVO "No" (No-hipoteca y No-Seguro) a la variable "No- Plan de Pensiones". Por otro lado, agruparemos las 4 variables de CANAL_OBJETIVO "Si" que no superan el 5%. En relación con las columnas, uniremos los grupos de edad quedando de la siguiente manera: <25, 26-45, 46-69 y >70.

Tras las agrupaciones de las filas indicadas (**10.2.1.11. /*AGRUPACIÓN DE LAS FILAS Y COLUMNAS */**) y realizar de nuevo el procedimiento *PROC CORRESP* (**10.2.1.12. /*PROCEDIMIENTO ACS CON FILAS Y COLUMNAS AGRUPADAS */)** indicando el número de componentes a retener, observamos que todos los porcentajes superan el 5% (Tabla 10.1.3.)

En la Figura 5.4.3. se muestra la Descomposición de la Inercia y la inercia total, así como las distancias de Chi-Cuadrado.
En este caso, la inercia total toma un valor de 0,13947y el estadístico Chi-Cuadrado el valor de 4.529,70.

Además, se indica el p-valor del test de Pearson, probabilidad de equivocarse al rechazar la hipótesis de independencia de los datos, y en este caso es menor que 0,0001. De esta manera, se puede rechazar la hipótesis de independencia a cualquier nivel de significación razonable.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.29222	0.08540	2773.56	61.23	61.23
0.21730	0.04722	1533.59	33.86	95.09
0.08278	0.00685	222.55	4.91	100.00
	0.13947	4529.70	100.00	

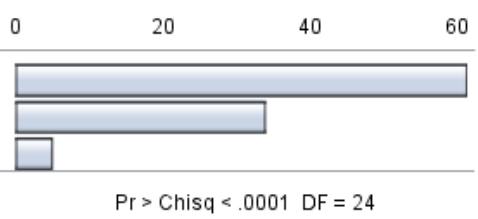


Figura 5.4.5. Inercia y Descomposición del Chi-Cuadrado

Al realizar SAS por defecto el modelo con dos dimensiones, y ser las únicas dimensiones posibles, analizamos el gráfico de representación de ambas dimensiones en la Figura 5.4.6:

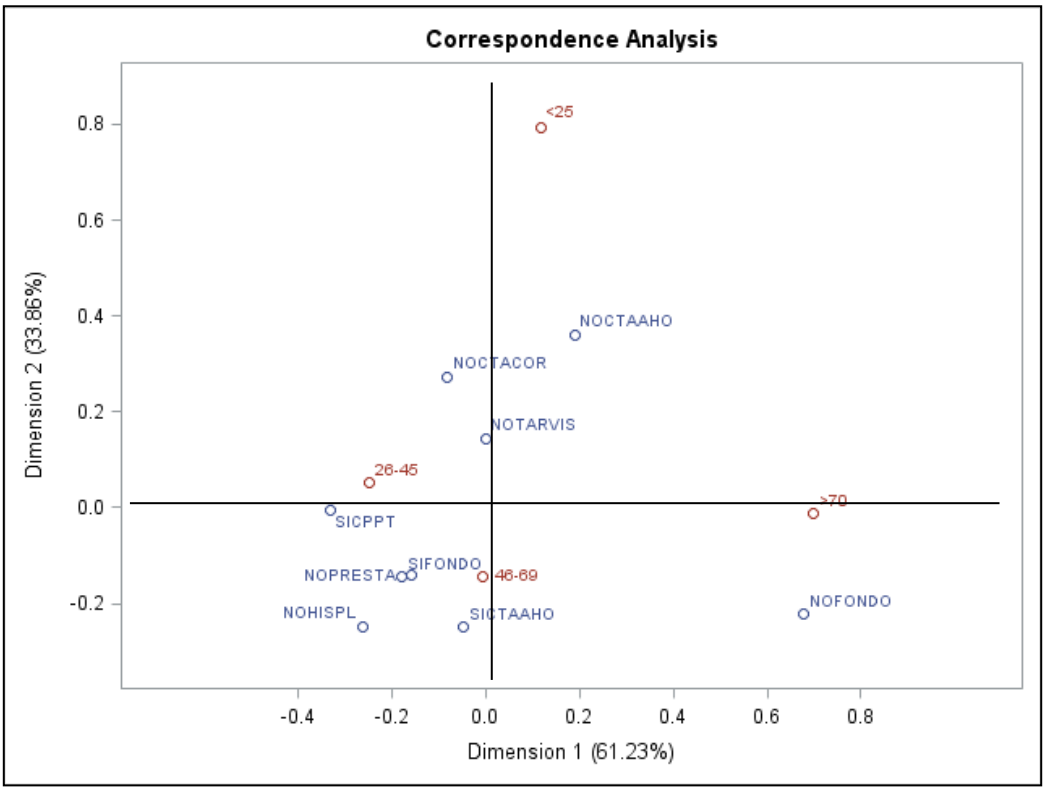


Figura 5.4.6. Gráfico de las dimensiones 1 y 2 del ACS para las variables Edad, Canal y Tipo de Producto

De este gráfico podemos deducir las siguientes conclusiones:

- El rango de edad **<25 años** tiene un nivel muy bajo de contratación (tanto online como sucursal). Únicamente se podría destacar la contratación de Cuentas Corrientes y de Ahorro en sucursal.

- Para el rango de edad de ***26 a 45 años*** destaca la contratación online, tanto del grupo de productos Cuenta Corriente, Préstamos, Planes de Pensiones como de Fondos de Inversión.

- En el rango de los ***46 a 69 años***, destaca la contratación online de los Fondos de Inversión y de las Cuentas de Ahorro.

- Para el grupo de más edad, ***>70 años,*** destaca la contratación en sucursal de los Fondos de Inversión.

Conclusiones finales del Análisis de Correspondencias Simples:

* **Nivel de estudios**:

El único nivel en el que destaca la contratación on-line es el nivel de ***Formación Universitario***. Del análisis ACS concluimos que, a mayor nivel educativo, mayor propensión a la adquisición online.
Para los diferentes niveles destacan los siguientes productos/canal:

- ✓ *Formación Elemental*: Cuenta de Ahorro en sucursal. Destaca por su baja contratación online, y alta en sucursal.
- ✓ *Nivel Básico*: Hipotecas, Seguros y Planes de Pensiones en sucursal
- ✓ *Nivel Medio*: nivel de contratación muy parecido de sucursal frente online
- ✓ *Formación Alto*: Cuenta de Ahorro online.
- ✓ *Formación Universitario*: Fondo de Inversión online. Grupo por excelencia que opera online.

* **Estado civil**:

Ninguno de los estados civiles destaca en una mayor contratación online.
Para los diferentes estados destacan los siguientes productos/canal:

- ✓ *Casado*: Planes de Pensiones, Hipoteca y Seguros en sucursal
- ✓ *Divorciados*: Préstamos en sucursal
- ✓ *Separados y Viudos*: Cuenta Corriente y de Ahorro en sucursal
- ✓ *Solteros*: Fondos de Inversión y Préstamos en sucursal

* **Edad**:

El grupo de edad en el que destaca la contratación on-line (objetivo) es el de los clientes entre ***26 y 45 años***.
Para los diferentes grupos de edad destacan los siguientes productos/canal:

- ✓ *<25 años:* Cuentas Corrientes y de Ahorro en sucursal.
- ✓ *Entre 26 y 45 años*: Cuenta Corriente, Préstamos, Planes de Pensiones y Fondos de Inversión online
- ✓ *Entre 46 y 69 años*: Fondos de Inversión y de las Cuentas de Ahorro online.
- ✓ *>70 años*: Fondos de Inversión en sucursal.

* **Perfiles de clientes de los dos canales**:

- ✓ **Online**: Formación Universitaria, cualquier Estado Civil y con una edad entre los 26 y 45 años.
- ✓ **Sucursal**: Formación Elemental, cualquier Estado Civil y menor de 25 años o mayor de 70.

6. Modelos de Predicción

6.1. Introducción

Tras el análisis, depuración y estudio de las relaciones entre las variables, comenzamos a elaborar los modelos de predicción, que nos permitirán conocer el comportamiento de los clientes.
Para la elaboración de los modelos, realizamos una agrupación de las categorías, analizamos las posibles interacciones para incorporar a los modelos y hacemos una selección de las variables con dichas interacciones.

6.1.1. Agrupación de Categorías

En el Software SAS Base 9.4 ejecutamos una macro específica para la agrupación de categorías mediante árboles (**10.2.2.1. /*MACRO DE AGRUPACIÓN DE CATEGORÍAS*/**), que posteriormente aplicaremos sobre nuestro conjunto de datos (**10.2.2.2./*AGRUPACIÓN CATEGORÍAS EN LOS DATOS*/**).

Obtenemos las siguientes agrupaciones para las variables Mes Apertura, Estado Civil y Tipo de Producto:
En primer lugar, en la Tabla 6.1.1.1. la variable ***Mes Apertura*** se agrupa en cuatro categorías:

Obs	MES_APERTURA	MES_APERTURA_G	COUNT	PERCENT
1	ABRIL	1	3043	9.3691
2	ENERO	1	3112	9.5816
3	FEBRERO	1	3481	10.7177
4	JUNIO	1	2425	7.4664
5	MARZO	1	3076	9.4707
6	MAYO	1	3311	10.1943
7	NOVIEMBRE	1	2688	8.2761
8	DICIEMBRE	2	1690	5.2034
9	AGOSTO	3	2037	6.2717
10	OCTUBRE	3	2777	8.5501
11	JULIO	4	2353	7.2447
12	SEPTIEMBRE	4	2486	7.6542

Tabla 6.1.1.1. Agrupación de Categorías de la variable Mes Apertura

En relación con la variable ***Estado Civil***, en la Tabla 6.1.1.2. se agrupan los niveles en cuatro categorías, unificando en una misma categoría a los Separados y Viudos:

Obs	ESTADO_CIVIL	ESTADO_CIVIL_G	COUNT	PERCENT
1	Casado/a	1	16669	51.3224
2	Soltero/a	2	12248	37.7105
3	Separado/a legal	3	582	1.7919
4	Viudo/a	3	1300	4.0026
5	Divorciado/a	4	1680	5.1726

Tabla 6.1.1.2. Agrupación de Categorías de la variable Estado Civil

Por último, en la Tabla 6.1.1.3. observamos la variable ***Tipo de Producto*** agrupada en 6 categorías:

Obs	TIPO_PRODUCTO	TIPO_PRODUCTO_G	COUNT	PERCENT
1	FONDO DE INVERSION	1	7189	22.1343
2	CUENTA CORRIENTE	2	4516	13.9044
3	CUENTA DE AHORRO	3	5607	17.2635
4	TARJETA VISA	4	7300	22.4761
5	PLAN DE PENSIONES	5	2601	8.0083
6	PRESTAMO	5	4139	12.7436
7	HIPOTECAS	6	393	1.2100
8	SEGUROS	6	734	2.2599

Tabla 6.1.1.3. Agrupación de Categorías de la variable Tipo Apertura

La variable categórica ***Nivel de Estudios*** no se agrupa en nuevas categorías, mantiene los 5 niveles de origen.

Procedemos a guardar el fichero con las variables agrupadas para utilizarlo en los modelos de predicción (**10.2.2.3. /*ARCHIVO FINAL CON AGRUPACIÓN DE CATEGORÍAS*/** y **10.2.2.4. /*GUARDAR EL FICHERO FINAL*/**).

6.1.2. Interacciones de Variables

Tras las agrupaciones de las categorías para las variables Mes de Apertura, Estado Civil y Tipo de Producto, estudiamos las posibles interacciones de las variables que podrán usarse como nuevas variables de los modelos.

Generamos un listado de las principales interacciones entre variables categóricas y entre variables categóricas y continuas, que se ordenan por su relación con la variable dependiente con el método AIC. (**10.2.2.5. /*MACRO DE INTERACCIONES*/ y 10.2.2.6. /*INTERACCIONES EN NUESTROS DATOS*/**)

Obs	variable	AIC	FValue	ProbF
1	DIA_APERTURA	-22070	12.59	0.0004
2	NUMERO_DE_HIJOS	-22085	27.26	<.0001
3	HIJOS*NUMERO_DE_HIJOS	-22085	27.26	<.0001
…	…	…	…	…
47	ESTADO_CIVIL_G*TIPO_PRODUCTO_G	-25669	168.09	<.0001
48	MES_APERTURA_G*TIPO_PRODUCTO_G	-25946	181.61	<.0001
49	NIVEL_ESTUDIOS*TIPO_PRODUCTO_G	-28973	267.90	<.0001

Tabla 6.1.2.1. Interacciones de las variables

Tabla completa en Anexo (Tabla 10.1.4)

6.1.3. Selección de Variables

Por último, y antes de comenzar con los modelos de predicción, realizamos una selección de variables, teniendo en cuenta las interacciones obtenidas.

Para ello, compilamos la macro (**10.2.2.7. /*MACRO DE SELECCIÓN DE VARIABLES*/**), que a partir del remuestreo repetido presenta los mejores modelos seleccionados por el método Stepwise. Para ejecutar posteriormente en nuestros datos (**10.2.2.8. /*SELECCIÓN DE VARIABLES EN LOS DATOS*/**).

Obtenemos los siguientes resultados en la selección de variables, que tendremos en cuenta para elaborar los modelos de predicción de Regresión Logística:

Obs	Efecto	COUNT	PERCENT
1	MES_APERTURA_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS MES_APERTURA_G*TIPO_PRODUCTO_G NIVEL_ESTUDIOS*TIPO_PRODUCTO_G	**12**	**57.1429**
2	ANO_APERTURA MES_APERTURA_G EDAD*TIPO_PRODUCTO_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS MES_APERTURA*TIPO_PRODUCTO_G NIVEL_ESTUDIOS*TIPO_PRODUCTO_G	**6**	**28.5714**
3	SEXO MES_APERTURA_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS MES_APERTURA_G*TIPO_PRODUCTO_G NIVEL_ESTUDIOS*TIPO_PRODUCTO_G	**2**	**9.5238**
4	ANO_APERTURA SEXO MES_APERTURA_G EDAD*TIPO_PRODUCTO_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS MES_APERTURA_G*TIPO_PRODUCTO_G NIVEL_ESTUDIOS*TIPO_PRODUCTO_G	**1**	**4.7619**

Tabla 6.1.3.1. Selección de modelos con interacciones

6.2. Regresión Logística

Tras la agrupación de variables, estudios de las interacciones y selección de las variables, comenzamos con la elaboración de los modelos de Regresión Logística.

Realizaremos validación cruzada, repitiendo los mismos modelos en diferentes particiones, para garantizar que los resultados son independientes.

En primer lugar, mediante la macro compilada (**10.2.3.1. /*MACRO DE REGRESIÓN LOGÍSTICA*/**) ejecutamos la Regresión logística para los cuatro modelos obtenidos en la selección de variables (Tabla 6.1.3.1.) y con la misma semilla cada modelo, Semilla de Inicio: 12345 y Semilla Final: 12356 (**10.2.3.2. /*REGRESIÓN LOGÍSTICA EN LOS DATOS*/**).

Obtenemos el gráfico de comparación de los 4 modelos de Regresión Logística con la misma semilla (Figura 6.2.1) y observamos que el mejor modelo, con una menor tasa de fallos, es el modelo 4. Además de tener una tasa menor, la variabilidad de los errores también es más baja.

De esta manera, para los cuatro modelos de Regresión Logística con la misma semilla, elegimos el **modelo 4**.

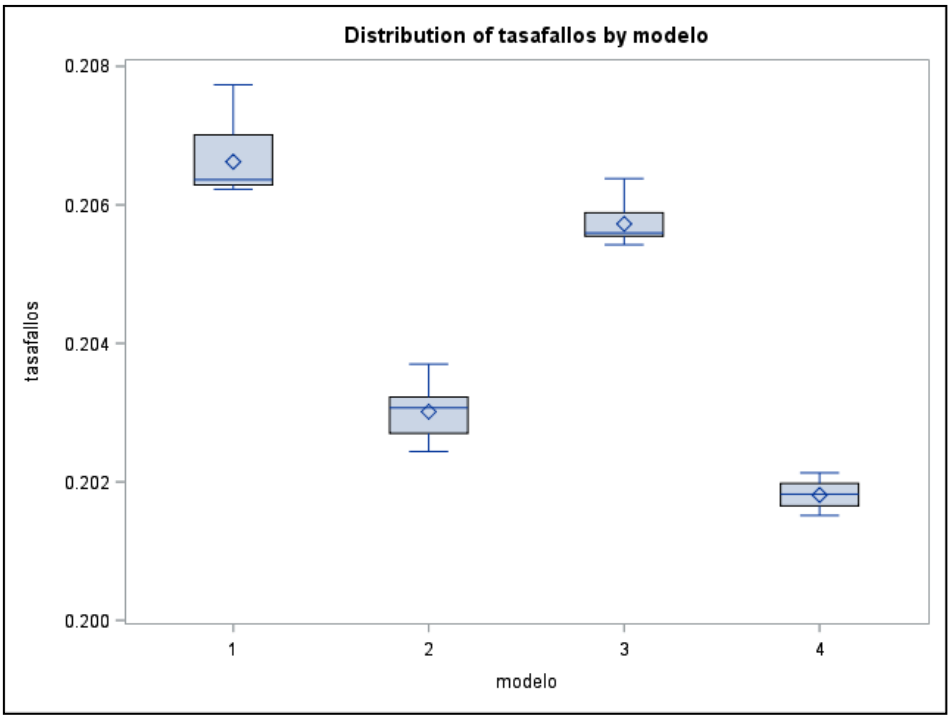


Figura 6.2.1. Modelos de Regresión Logística I

Probamos con otra semilla los mismos modelos de variables con interacciones, para comprobar y asegurar que el modelo 4 es el de menor error.

Mediante el Código en (**10.2.3.2. /*REGRESIÓN LOGÍSTICA EN LOS DATOS*/**) ejecutamos la Regresión logística para los cuatro modelos obtenidos en la selección de variables y con una nueva semilla, lo que modifica la selección de datos sobre los que se realiza el modelo.

Obtenemos en la Figura 6.2.2. el gráfico de comparación de los 4 modelos de Regresión Logística con la misma semilla (***Semilla Inicio***: 54321- ***Semilla Final***: 54340)

El mejor modelo, con una menor tasa de fallos, es el modelo 8. En este caso, la variabilidad de los errores no es la más baja, pero la tasa de fallos es considerablemente menor, por lo que escogeríamos el **Modelo 8**.

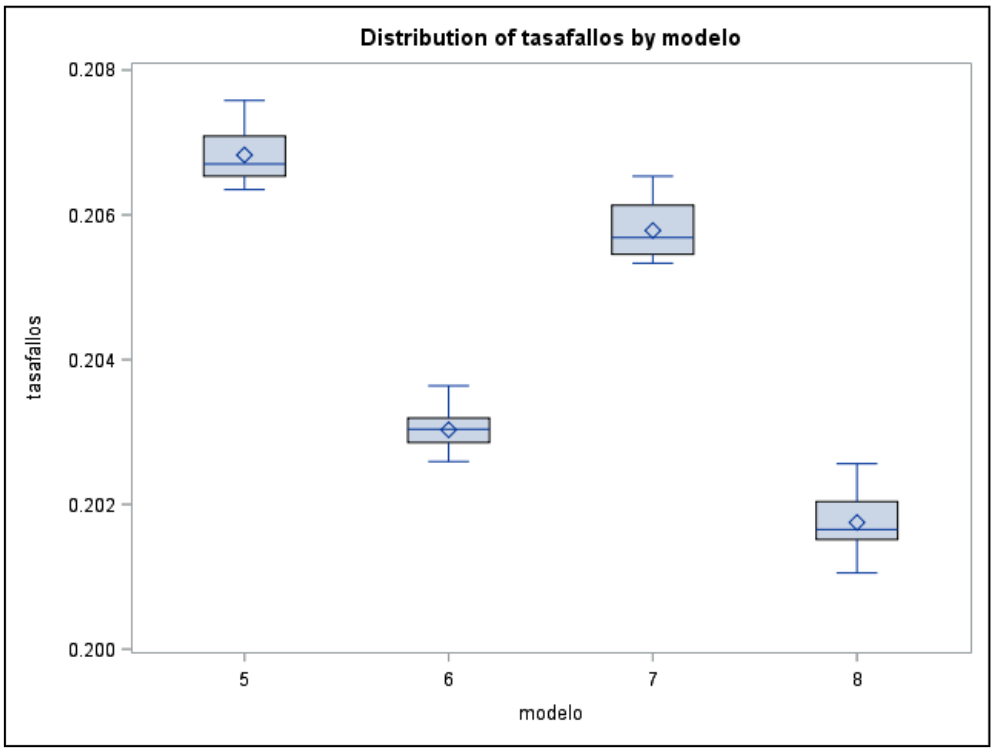


Figura 6.2.2. Modelos de Regresión Logística II

De esta manera, para los ocho modelos de Regresión Logística con diferentes semillas, el mejor modelo de variables con interacciones es el mismo (Modelo 4 y Modelo 8):

ANO_APERTURA SEXO MES_APERTURA_G EDAD*TIPO_PRODUCTO_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS MES_APERTURA_G*TIPO_PRODUCTO_G NIVEL_ESTUDIOS*TIPO_PRODUCTO_G

En la Tabla 6.2.1. se indican algunos de los valores de los parámetros para el mejor modelo de la Regresión Logística. Tabla completa en Anexo (Tabla 10.1.5.).

Según lo explicado en la metodología, la Regresión Logística es un tipo de análisis de regresión que busca determinar la existencia de relación entre las variables independientes con la variable dependiente, así como predecir la probabilidad de que se cumpla el objetivo sobre la probabilidad de que no se cumpla.

Este modelo de regresión logística se representa en la siguiente relación, siendo $p_1$la probabilidad de que ocurra *Y=1*:

$$p_1 = P(Y = 1 \parallel x_1, x_2, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

Con lo que se obtiene el logaritmo de la razón de probabilidades u *odds ratio* (logit):

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

En este caso, el parámetro Intercept es β_0, con un valor estimado de 840.9 y un error estándar de 78.9755.
El parámetro $\beta_1 x_1$ se compone de β_1 con un valor estimado de -0.1740 y x_1 la variable Sexo (Categoría: Hombre). Este parámetro tiene un error estándar de 0.0180.

Así todos los parámetros de la expresión hasta $\beta_m x_m$, siendo *m=55.*

Calculamos la razón de probabilidades u *odds ratio* (logit) con el primer parámetro, de la variable Sexo (Hombre):

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m;\ p_1 = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 (x=1))}}$$

$$odds\ (x = 1) = \frac{p\ (x = 1)}{1 - p(x = 1)} = \ e^{\beta_0 + \beta_1} = \ e^{840.9 + (-0.1740)}$$

$$odds\ (x = 0) = \frac{p\ (x = 0)}{1 - p(x = 0)} = \ e^{\beta_0} = \ e^{840.9}$$

$$Odds\ Ratio = \frac{odds\ (x = 1)}{odds\ (x = 0)} = \ e^{-0.1740} = 0{,}84$$

Teniendo en cuenta únicamente los dos primeros parámetros de la Regresión Logística (β_0 y β_1), la probabilidad de que ocurra el suceso (x=1=canal online) en los clientes de Sexo Hombre, frente a que no ocurra, es de 0,84.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	840.9	78.9755	113.3750	<.0001
SEXO	**Hombre**	**1**	**-0.1740**	**0.0180**	**93.2390**	**<.0001**
ANO_APERTURA		1	-0.4152	0.0358	134.2298	<.0001
MES_APERTURA_G	**1**	1	-0.0316	35.7033	0.0000	0.9993
MES_APERTURA_G	**2**	1	-0.4436	68.8739	0.0000	0.9949
MES_APERTURA_G	**3**	1	0.1621	47.1324	0.0000	0.9973
EDAD*TIPO_PRODUCTO_G	**1**	1	0.0394	0.00211	347.0247	<.0001
…	**…**	…	…	…	…	…
NIVEL_ESTUDIO*TIPO_PRODUCTO	**Nivel de formacion universitario 5**	1	-0.0416	47.9313	0.0000	0.9993

Tabla 6.2.1. Parámetros del mejor modelo de Regresión Logística

6.3. Red Neuronal

Probamos con otros modelos de predicción que nos pueden aportar más información sobre el mejor modelo de predicción del canal de contratación de los clientes.

Procedemos a elaborar diferentes modelos con redes neuronales, modificando los Nodos y Algoritmos, en búsqueda del mejor modelo.

Mediante la macro compilada (**10.2.4.1. /*MACRO DE RED NEURONAL*/**) ejecutamos dos modelos de Redes Neuronales con Nodos de 2 a 10 (variando de 2 en 2), para los algoritmos BROP y QUANEW. Y otros dos modelos de Redes Neuronales, modificando la semilla, con Nodos de 3 a 11 (variando de 3 en 3), para los algoritmos BROP y QUANEW (**10.2.4.2. /*RED NEURONAL EN LOS DATOS (I)*/**)

En la Tabla 6.3.1. observamos las características de los cuatro modelos a construir:

Modelo	Nombre Modelo	Nº Nodos	Inicio Nodos	Final Nodos	Incremento	Algoritmo	Semilla Inicio	Semilla Final
Modelo 9	***Neural1***	5	2	10	2	BROP	12345	12354
Modelo 10	***Neural2***	5	2	10	2	QUANEW	12345	12354
Modelo 11	***Neural3***	4	3	11	3	QUANEW	54321	54340
Modelo 12	***Neural4***	4	3	11	3	BROP	54321	54340

Tabla 6.3.1. Modelos de Red Neuronal I

De los cuatro modelos, observamos en la Figura 6.3.1. que los dos modelos con el algoritmo QUANEW (Neural2 y Neural3) tienen una media más baja de error, aunque una variabilidad mayor a la de los modelos Neural1 y Neural4.

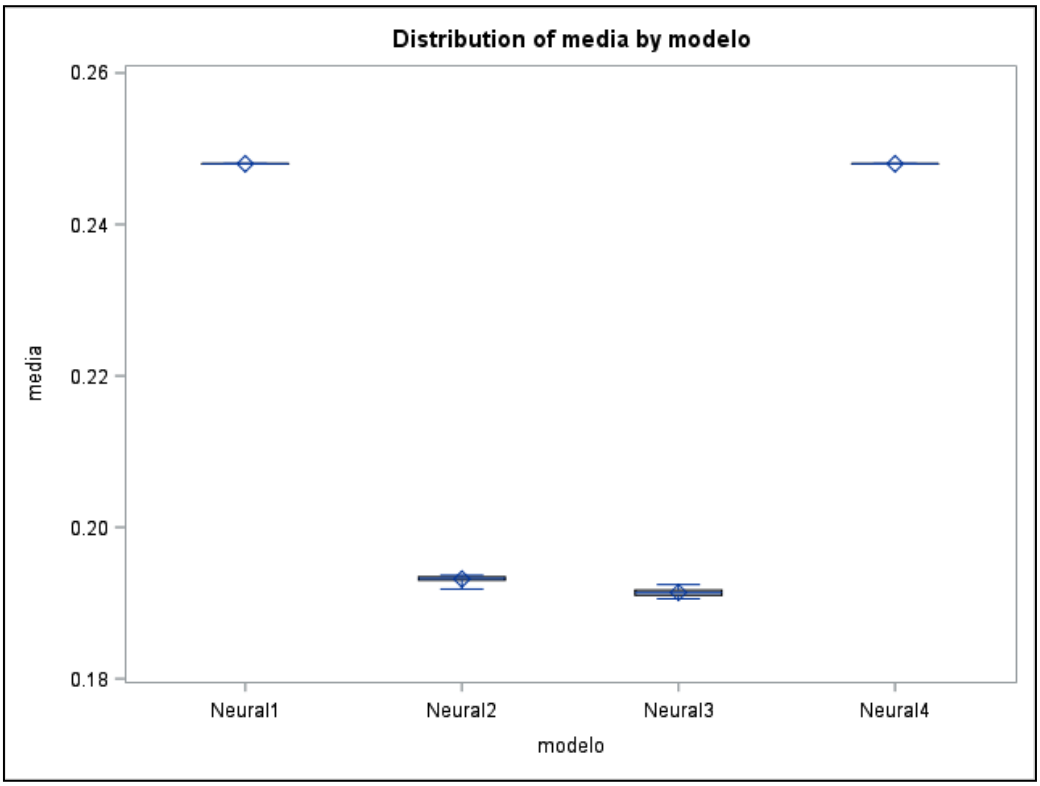


Figura 6.3.1. Modelos de Red Neuronal I

Probamos, para el mismo logaritmo QUANEW varios modelos con diferentes Nodos, para comprobar cuál aporta una menor tasa de fallos (**10.2.4.3. /*RED NEURONAL EN LOS DATOS (II)*/**):

En la Tabla 6.3.2. observamos las características de los nuevos modelos a construir:

Modelo	Nombre Modelo	Nº Nodos	Inicio Nodos	Final Nodos	Incremento	Algoritmo	Semilla Inicio	Semilla Final
Modelo 13	***Neural1***	4	2	8	2	QUANEW	12345	12354
Modelo 14	***Neural2***	5	2	10	2	QUANEW	54321	54340
Modelo 15	***Neural3***	6	2	12	2	QUANEW	54321	54340
Modelo 16	***Neural4***	7	2	14	2	QUANEW	54321	54340

Tabla 6.3.2. Modelos de Red Neuronal II

De los cuatro modelos con el algoritmo QUANEW el mejor modelo es ***Neural-4,*** con una tasa de fallos por debajo de 0,192, representado en la Figura 6.3.2.

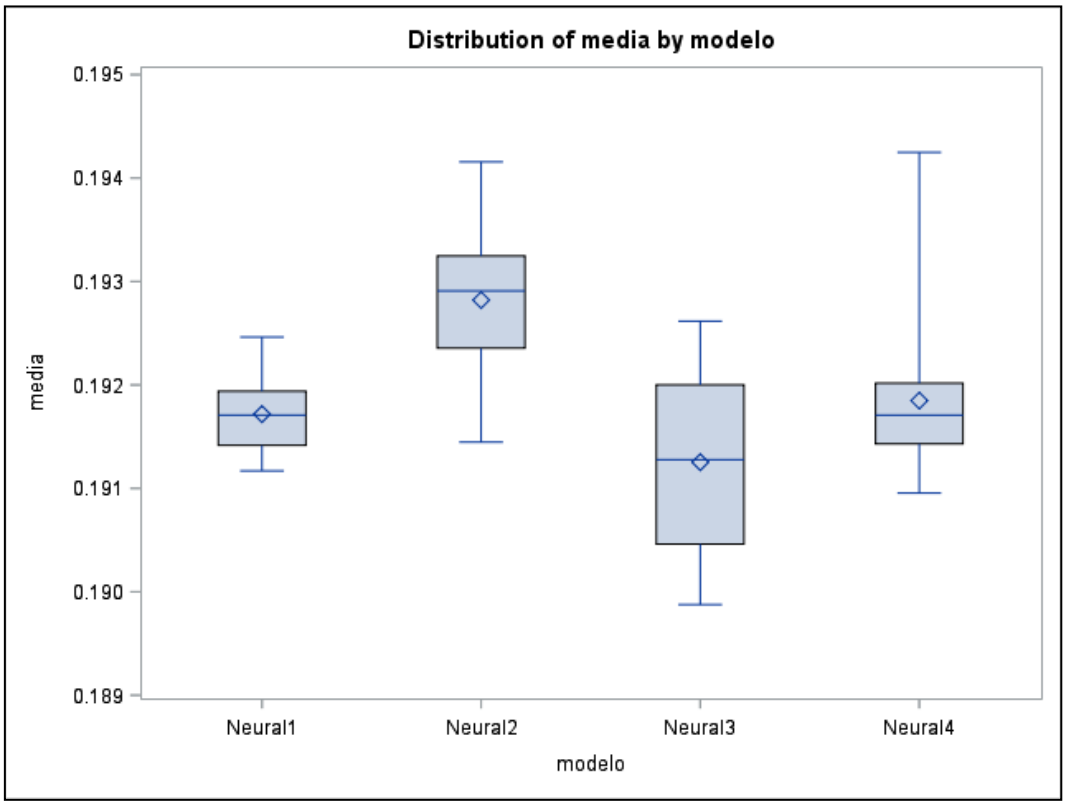


Figura 6.3.2. Modelos de Red Neuronal II

6.4. Gradient Boosting

A continuación, utilizamos el método Gradient Boosting, método que, mediante la construcción de árboles de clasificación va reduciendo el error medio con la modificación de las predicciones.

Compilamos la macro (**10.2.5.1. /*MACRO DE GRADIENT BOOSTING*/**) y la ejecutamos en nuestros datos modificando el número de árboles de cada modelo.
Además, en cada uno de los modelos modificamos las características de los árboles (tamaño y número de hojas) (**10.2.5.2. /*GRADIENT BOOSTING EN LOS DATOS*/**).

En la Tabla 6.4.1. observamos las características de los modelos a construir:

Modelo	Nombre Modelo	Tamaño hoja	Número hoja	Nº Árboles	Criterio	Semilla Inicio	Semilla Final
Modelo 17	***BTG1***	5	10	20	ProbF	12345	12356
Modelo 18	***BTG2***	8	20	30	ProbF	12345	12356
Modelo 19	***BTG3***	10	30	40	ProbF	54321	54340
Modelo 20	***BTG4***	13	40	50	ProbF	54321	54340

Tabla 6.4.1. Modelos de Gradient Boosting

En la Figura 6.4.1. observamos que el modelo de **Gradient Boosting 4** (50 árboles, tamaño de hoja 13 y nº de hojas 40) tiene una media más baja de error, aunque la

variabilidad del error no es la más pequeña. Pese a ello, seleccionamos este modelo por tener una media de error más baja.

Es lógico que el modelo que más se ajusta, y comete menos errores, es aquel con mayor número de árboles, con un tamaño y número de hojas mayor.

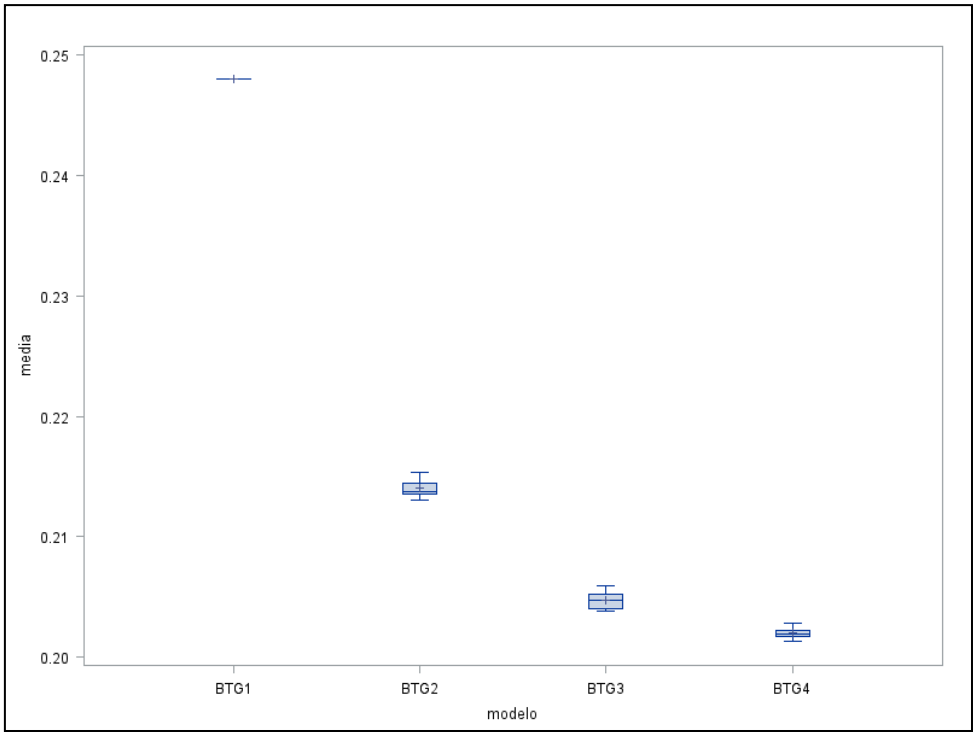


Figura 6.4.1. Modelos de Gradient Boosting

6.5. Random Forest

Este algoritmo construye los modelos mediante el re-muestreo de observaciones y de variables, modificando el tamaño de las hojas, con el objetivo de reducir el sobreajuste. Compilamos la macro (**10.2.6.1. /*MACRO DE RANDOM FOREST*/**) y la ejecutamos en nuestros datos variando el tamaño de las hojas (**10.2.6.2. /*RANDOM FOREST EN LOS DATOS (I)*/**).

En la Tabla 6.5.1. observamos las características de los modelos a construir:

Modelo	Nombre Modelo	Tamaño hoja	Máximo árboles	Semilla Inicio	Semilla Final
Modelo 21	***FOREST1***	8	350	12345	12356
Modelo 22	***FOREST2***	5	250	12345	12356
Modelo 23	***FOREST3***	10	450	54321	54340

Tabla 6.5.1. Modelos de Random Forest I

En la Figura 6.5.1 observamos los tres modelos en los que la tasa de fallos es menor en el ***FOREST-3***, aunque la variabilidad de la misma es menor en el ***FOREST-1.***

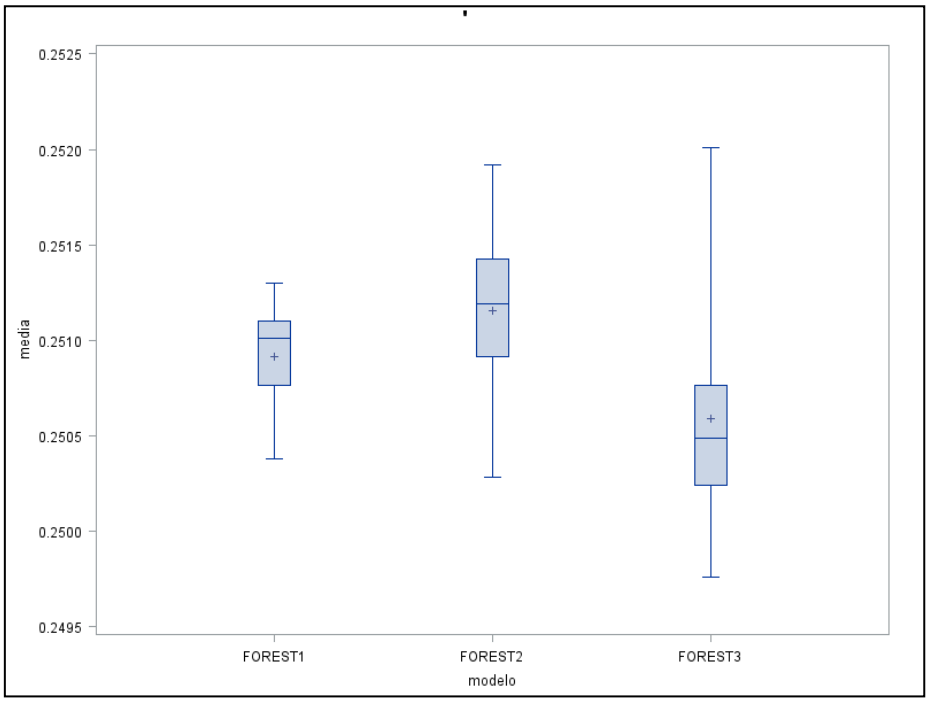


Figura 6.5.1. Modelos de Random Forest I

Comprobamos si, manteniendo el número máximo de árboles, pero variando el tamaño de la hoja, la variabilidad del error varía (**10.2.6.3. /*RANDOM FOREST EN LOS DATOS (II)*/**).

En la Tabla 6.5.2. observamos las características de los nuevos modelos a construir:

Modelo	Nombre Modelo	Tamaño hoja	Máximo árboles	Semilla Inicio	Semilla Final
Modelo 24	***FOREST1***	8	450	12345	12356
Modelo 25	***FOREST2***	5	450	12345	12356
Modelo 26	***FOREST3***	10	450	54321	54340

Tabla 6.5.2. Modelos de Random Forest II

En la Figura 6.5.2. observamos que el mejor modelo, con 450 árboles, es el de mayor tamaño de hoja (**FOREST-3**), al tener una menor tasa de fallos.

Aun sí, comprobamos que la tasa de fallos es muy elevada (0.2505) en comparación con los métodos realizados hasta el momento.

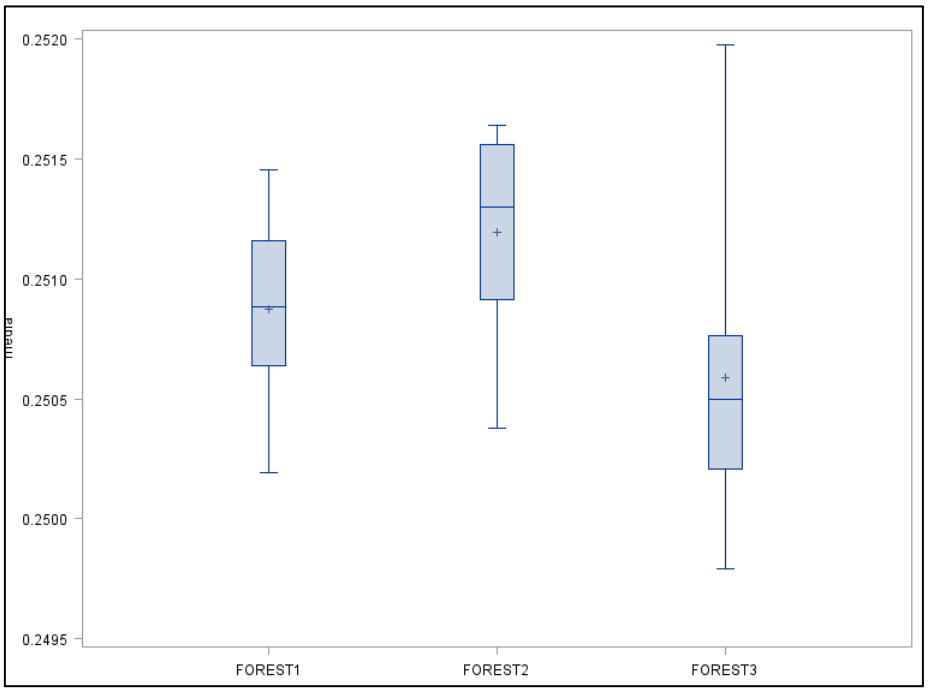


Figura 6.5.2. Modelos de Random Forest II

Como caso particular del método Random Forest está el método Bagging que busca el sobreajuste en la selección de variables, modificando el número de hojas y el tamaño.

Compilamos la macro (**10.2.7.1. /*MACRO DE BAGGING*/**) y la ejecutamos en nuestros datos modificando el número de hojas y tamaño (**10.2.7.2. /*BAGGING EN LOS DATOS*/**)

En la Tabla 6.5.3. observamos las características de los modelos a construir:

Modelo	Nombre Modelo	Tamaño hoja	Número Hojas	Semilla Inicio	Semilla Final
Modelo 27	***BAGGING1***	15	45	12345	12356
Modelo 28	***BAGGING2***	20	40	12345	12356
Modelo 29	***BAGGING3***	15	20	12345	12356
Modelo 30	***BAGGING4***	10	30	12345	12356

Tabla 6.5.3. Modelos de Bagging

En la Figura 6.5.3 observamos los cuatro modelos, siendo la tasa de fallos menor en el ***BAGGING-1***:

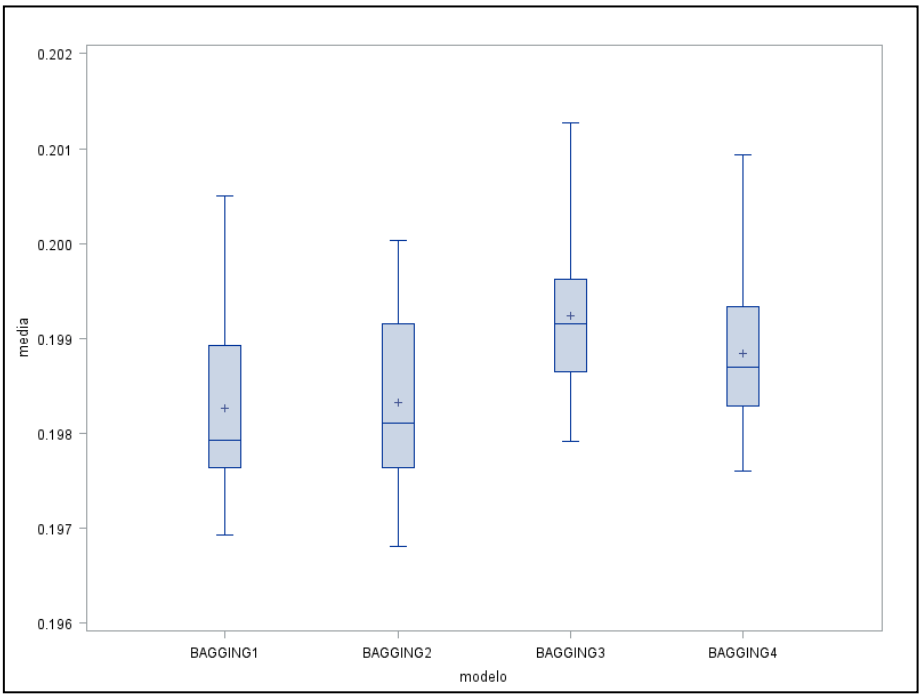


Figura 6.5.3. Modelos de Bagging

6.6. Comparación de Modelos:

Una vez realizados los 30 modelos de prediccion, comparamos los mejores modelos de cada algoritmo.

En la Tabla 6.6.1. observamos los mejores modelos de cada método de predicción elaborados hasta el momento que se van a comparar:

ALGORITMO	MODELO
Regresión Logística	Modelo 4
Redes Neuronales	Neural4
Gradient Boosting	BTG4
Random Forest	FOREST3
Bagging	BAGGING1

Tabla 6.6.1. Comparación de Modelos

A continuación, en las Tablas 6.6.2, 6.6.3, 6.6.4, 6.6.5 y 6.6.6, se detallan las características principales de los mejores modelos de cada método:

* En la Tabla 6.6.2. observamos las características del mejor modelo de la Regresión Logística, Modelo 4, con una menor tasa de fallos.

  Se obtuvo como resultado un porcentaje de error del **20,2%.**

Modelo	Variables	Semilla Inicio	Semilla Final
Modelo 4	ANO_APERTURA SEXO MES_APERTURA_G EDAD*TIPO_PRODUCTO_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS MES_APERTURA_G*TIPO_PRODUCTO_G NIVEL_ESTUDIOS*TIPO_PRODUCTO_G	12345	12356

Tabla 6.6.2. Mejor modelo de Regresión Logística

* En la Tabla 6.6.3. observamos las características del mejor modelo de Redes Neuronales, NEURAL-3, con una menor media de fallos.

  Se obtuvo como resultado un porcentaje de error del **19,2%.**

Modelo	Nombre Modelo	Nº Nodos	Inicio Nodos	Final Nodos	Incremento	Algoritmo	Semilla Inicio	Semilla Final
Modelo 15	***Neural3***	6	2	12	2	QUANEW	54321	54340

Tabla 6.6.3. Mejor modelo de Redes Neuronales

* En la Tabla 6.6.4 observamos los parámetros del mejor modelo de Gradient Boosting, BTG4.

  Se obtuvo como resultado un porcentaje de error del **20%**

Modelo	Nombre Modelo	Tamaño hoja	Número hoja	Interacciones	Criterio	Semilla Inicio	Semilla Final
Modelo 20	***BTG4***	13	40	50	ProbF	54321	54340

Tabla 6.6.4. Mejor modelo de Gradient Boosting

* En las Tablas 6.6.5 y 6.6.6. observamos las características de los mejores modelos de los métodos Random Forest y Bagging:

Modelo	Nombre Modelo	Tamaño hoja	Máximo árboles	Semilla Inicio	Semilla Final
Modelo 26	***FOREST3***	10	450	54321	54340

Tabla 6.6.5. Mejor modelo de Random Forest

Obteniendo como resultado un porcentaje de error del **25,05%**

Modelo	Nombre Modelo	Tamaño hoja	Número Hojas	Semilla Inicio	Semilla Final
Modelo 27	***BAGGING1***	15	45	12345	12356

Tabla 6.6.6. Mejor modelo de Bagging

Obteniendo como resultado un porcentaje de error del **19,8%**

De los cuatro modelos, el modelo de Random Forest se desvía de la media total de la tasa de fallos (25,05%). Por lo que ese modelo no se escogerá para predecir y explicar el comportamiento de los clientes.

Entre los modelos restantes (Regresión Logística, Red Neuronal, Gradient Boosting y Bagging), el modelo con menor media en la tasa de fallos es la **Red Neuronal (19,02%)**.

6.7. Ensamblado de modelos

Una vez realizados todos los modelos de predicción, y analizadas las comparaciones entre los mejores modelos de cada tipo, utilizamos el método de ensablado de modelos.

Este método consiste en la construccion de un modelo a partir de la combinacion de varios modelos y su ensamblado combinándolos entre ellos.

Los modelos que combinaremos serán los mejores modelos realizados de forma independiente a lo largo del trabajo: Regresión Logística, Red Neuronal, Random Forest y Gradient Boosting.
Las combinaciones de ensamblaso entre los modelos serán las siguientes:

Modelo 1: Red Neuronal (RED)
Modelo 2: Regresión Logística (LOG)
Modelo 3: Random Forest (RFOR)
Modelo 4: Gradient Boosting (BOOST)
Modelo 5: Red Neuronal y Regresión Logística (RLOG)
Modelo 6: Red Neuronal y Random Forest (REDFOR)
Modelo 7: Red Neuronal y Gradient Boosting (REDBOO)
Modelo 8: Regresión Logística y Random Forest (LRFOR)
Modelo 9: Regresión Logística y Gradient Boosting (LBOOST)
Modelo 10: Random Forest y Gradient Boosting (RFORBOO)
Modelo 11: Red Neuronal, Regresión Logística, Random Forest y Gradient Boosting (R-L-RF-BOO)

Para ello, compilamos la macro (**10.2.9.1. /*MACRO ENSAMBLADO DE MODELOS*/**) y la ejecutamos en nuestros datos (**10.2.9.2. /*ENSAMBLADO DE MODELOS EN LOS DATOS*/**).

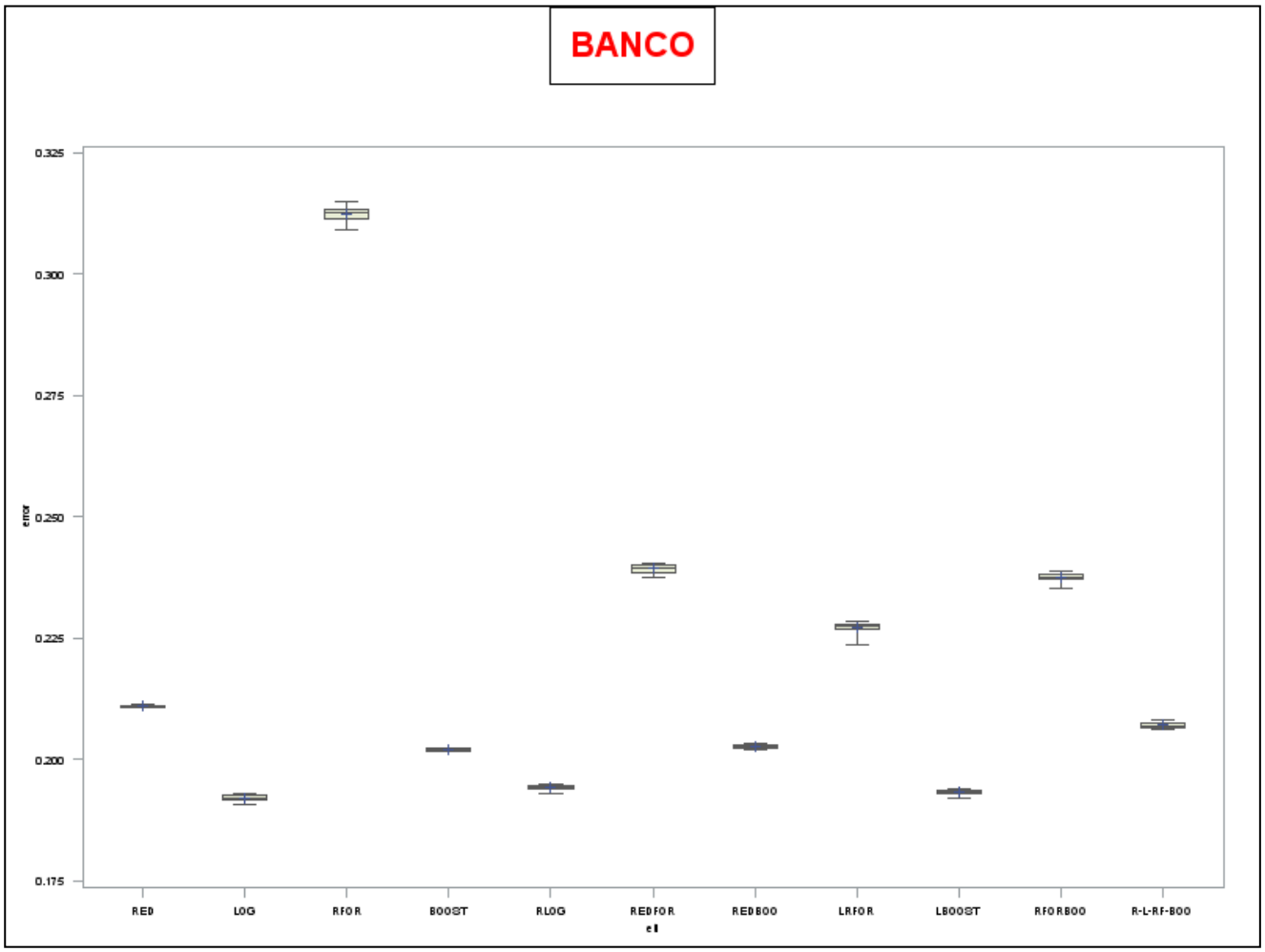


Figura 6.7.1. Comparación de Modelos

En la figura 6.7.1. observamos la representacion de los 11 modelos, siendo el modelo con menor error el ensamblado de la Regresión Logística y Red Neuronal (Modelo 5). Los mejores modelos desarrollados en el trabajo para estos métodos tuvieron unos errores del 20.2% y 19.2% respectivamente.

Otro de los modelos ensamblado con un porcentaje de error bajo es la Regresión Logística y Gradient Boosting (Modelo 9).

7. Selección del modelo

Tras el análisis de todos los modelos planteados, su ensamblado y combinacion, y la comparacion entre ellos, el modelo final seleccionado para explicar el comportamiento de los clientes y predecirlo es la **Red Neuronal** (con 6 Nodos y el Algoritmo QUANEW) y un porcentaje de error por debajo del 20%.

La Red Neuronal es un método de predicción con habilidad de aprendizaje a partir de casos anteriores abstrae características esenciales.

Además, es un modelo con capacidad de autoorganización, creando su propia estructura de la información.

Otra de las ventajas del uso de este método es la tolerancia a los fallos, y la flexibilidad en el manejo de los cambios de información.

Con esto se considera que es un buen método escogido para la predicción del comportamiento del canal de contratación de la cartera de clientes.

Si bien se debe indicar que el modelo es bastante débil al tener una tasa de fallos tan elevada.

8. Conclusiones

El objetivo principal de este proyecto era conocer y predecir el uso del canal de los servicios bancarios, conociendo el comportamiento de la cartera de clientes, mediante el estudio de sus principales atributos y características (edad, sexo, hijos, estudios, etc.), lo cual se ha alcanzado a lo largo del estudio, tanto al analizar las variables y sus relaciones como en la selección el modelo final de predicción.

A lo largo del trabajo se ha definido el perfil del cliente que contrata en el canal objetivo (online), y se ha concluido que el cliente que opera online tiene una edad media entre 26 y 45 años, con un nivel de formación alto, sin hijos y de diversos estados civiles.

Con esto, y teniendo en cuenta principalmente la edad, se puede aconsejar al banco a realizar una mayor inversión en banca online y destinar menos recursos a las oficinas bancarias.
La inversión de recursos destinados al desarrollo de la banca online debe tener en cuenta las características y necesidades de este perfil de clientes. Además, se debe tener en cuenta que el principal producto contratado online por estos clientes son los Fondos de Inversión.

Además, se establecieron unos objetivos secundarios que también se han alcanzado con éxito:

* Conocimiento, preparación y obtención de información útil de los datos: mediante el "Análisis Exploratorio de las Variables" hemos conocido las principales características de las variables, sus deficiencias y errores, así como las principales relaciones, mediante gráficos, entre la variable objetivo y las variables independientes.
  A raíz de este análisis exploratorio se ha propuesto la creación de nuevas variables, lo que ha aportado más información al estudio y a los modelos.

La preparación de los datos se ha realizado a través de la depuración de las variables originales, que ha permitido limpiar los errores de las variables y sus categorías, así como las observaciones que pudiesen desvirtuar los resultados o modelos.

* En segundo lugar, se propuso conocer mediante el análisis de las variables las relaciones existentes entre las mismas.
  Para la consecución de este objetivo, se ha obtenido información útil de los datos mediante el "Análisis de Correspondencias Simple", que ha permitido conocer las dependencias entre las variables independientes más relacionadas con la variable objetivo. Análisis del que se obtiene un perfil de cliente para cada canal de contratación (variable objetivo).

* Por último, se propuso analizar y comparar los diferentes modelos de predicción, optando por el que mejor prediga el comportamiento de los clientes.
  Para ello, se han utilizado los diferentes métodos de predicción: Regresión Logística, Red Neuronal, Gradient Boosting y Ramdom Forest.
  En cada método de predicción se han elaborado diferentes modelos, variando sus características (número de árboles, tamaño hoja, algoritmo, criterio, interacciones, etc.), con el objetivo de obtener el mejor modelo de cada método.
  Tras la obtención del mejor modelo de cada método, se han comparado entre ellos, para elegir aquel modelo que mejor se ajuste a los datos, minimizando la tasa de fallos.

9. Bibliografía

(1) Banco de España, www.bde.es/bde/es/secciones/informes/

(2) Teodoro Luque Martínez (2012). **Técnicas de análisis de datos en investigación de mercados**

(3) S. Sarma, Kattamury (2007). **Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications**.

(4) Lucio A. Pat Fernández, Aarón H. Martínez Menchaca, Juan M. Pat Fernández (2013). **Introducción a los modelos de regresión.**

(5) Andrea. Beltratti, Sergio. Margarita, Pietro. Terna (1996). **Neural networks for economic and financial modelling**

(6) César. Pérez López, Daniel. Santín González (2007). **Minería de datos: técnicas y herramientas**

(7) Javier Portela (2017-2018), **Apuntes REDES NEURONALES Y ALGORITMOS GENÉTICOS del Máster de Minería de Datos e Inteligencia de Negocios (UCM)**

10. Anexos

10.1. Tablas

Análisis de Correspondencias Simple:

Relación de las variables NIVEL DE ESTUDIOS, TIPO PRODUCTO y CANAL_OBJETIVO

Contingency Table						
Percents	**Formacion Elemental (estudios primarios)**	**Nivel basico (ESO, EGB, FP)**	**Nivel de formacion alto**	**Nivel de formacion universitario**	**Nivel medio**	**Sum**
NOCTACOR	1.940	3.673	0.890	2.124	1.564	10.191
NOCTAAHO	3.793	4.310	0.745	1.389	1.521	11.758
NOFONDO	2.925	3.916	1.065	2.346	2.017	12.269
NOPRESTA	2.565	4.748	0.631	1.727	1.576	11.247
NOHISPL	1.786	4.421	0.850	1.727	1.703	10.487
NOTARVIS	4.212	7.688	1.364	3.153	2.830	19.246
SICTAAHO	0.551	0.742	0.991	2.248	0.973	5.505
SIFONDO	0.594	1.623	1.004	5.059	1.586	9.865
SICPPT	0.933	2.460	0.998	3.100	1.940	9.431
Sum	19.299	33.582	8.538	22.873	15.709	100.000

Tabla 10.1.1. Tabla de Contingencias de las variables Nivel Estudios, Canal y Tipo Producto Agrupada

Relación de las variables ESTADO CIVIL, TIPO PRODUCTO y CANAL_OBJETIVO

Contingency Table					
Percents	**Casado**	**Divorciado**	**Separado-Viudo**	**Soltero**	**Sum**
NOCTACOR	4.440	0.671	4.646	0.434	10.191
NOCTAAHO	5.114	0.477	5.440	0.727	11.758
NOFONDO	7.288	0.385	3.430	1.167	12.269
NOPRESTA	6.013	0.936	3.347	0.951	11.247
NOHISPL	6.530	0.600	2.848	0.508	10.487

Contingency Table					
Percents	**Casado**	**Divorciado**	**Separado-Viudo**	**Soltero**	**Sum**
NOTARVIS	9.268	1.105	7.676	1.198	19.246
SICTAAHO	3.039	0.169	2.106	0.191	5.505
SIFONDO	5.296	0.271	3.990	0.308	9.865
SICPPT	4.335	0.557	4.227	0.311	9.431
Sum	51.322	5.173	37.711	5.795	100.000

Tabla 10.1.2. Tabla de Contingencias de las variables Estado Civil, Canal y Tipo Producto Agrupada

Relación de las variables EDAD, TIPO PRODUCTO y CANAL_OBJETIVO

Contingency Table					
	<25	**26-45**	**46-69**	**>70**	**Sum**
NOCTACOR	381	1398	1201	330	3310
NOCTAAHO	574	1169	1439	637	3819
NOFONDO	97	585	1963	1340	3985
NOPRESTA	103	1441	1841	268	3653
NOHISPL	48	1275	1992	91	3406
NOTARVIS	533	2492	2403	823	6251
SICTAAHO	12	633	948	195	1788
SIFONDO	41	1457	1421	285	3204
SICPPT	125	1565	1252	121	3063
Sum	1914	12015	14460	4090	32479

Tabla 10.1.3. Tabla de Contingencias de las variables Edad, Canal y Tipo Producto Agrupada

Modelos de Predicción: Interacciones de Variables

DIA_APERTURA
NUMERO_DE_HIJOS
HIJOS*NUMERO_DE_HIJOS
HIJOS
HIJOS*DIA_APERTURA
MES_APERTURA_G*NUMERO_DE_HIJOS
EDAD
SEXO*NUMERO_DE_HIJOS
HIJOS*EDAD
ANO_APERTURA
ESTADO_CIVIL_G*NUMERO_DE_HIJOS
HIJOS*ANO_APERTURA
ESTADO_CIVIL_G*DIA_APERTURA
SEXO*DIA_APERTURA
ESTADO_CIVIL_G

MES_APERTURA_G*DIA_APERTURA
SEXO
ESTADO_CIVIL_G*EDAD
HIJOS*ESTADO_CIVIL_G
MES_APERTURA_G
SEXO*HIJOS
SEXO*EDAD
ESTADO_CIVIL_G*ANO_APERTURA
HIJOS*MES_APERTURA_G
MES_APERTURA_G*EDAD
SEXO*ANO_APERTURA
MES_APERTURA_G*ANO_APERTURA
SEXO*ESTADO_CIVIL_G
MES_APERTURA_G*ESTADO_CIVIL_G
SEXO*MES_APERTURA_G
NIVEL_ESTUDIOS*NUMERO_DE_HIJOS
TIPO_PRODUCTO_G*NUMERO_DE_HIJOS
TIPO_PRODUCTO_G*DIA_APERTURA
NIVEL_ESTUDIOS*DIA_APERTURA
TIPO_PRODUCTO_G*EDAD
NIVEL_ESTUDIOS*EDAD
TIPO_PRODUCTO_G
HIJOS*TIPO_PRODUCTO_G
TIPO_PRODUCTO_G*ANO_APERTURA
NIVEL_ESTUDIOS
HIJOS*NIVEL_ESTUDIOS
NIVEL_ESTUDIOS*ANO_APERTURA
ESTADO_CIVIL_G*NIVEL_ESTUDIOS
SEXO*TIPO_PRODUCTO_G
MES_APERTURA_G*NIVEL_ESTUDIOS
SEXO*NIVEL_ESTUDIOS
ESTADO_CIVIL_G*TIPO_PRODUCTO_G
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G

Tabla 10.1.4. Todas Interacciones de las variables

Modelos de Predicción

Regresión Logística:

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	840.9	78.9755	113.3750	<.0001
ANO_APERTURA		1	-0.4152	0.0358	134.2298	<.0001
SEXO	**Hombre**	1	-0.1740	0.0180	93.2390	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter			**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
MES_APERTURA_G	**1**		1	-0.0316	35.7033	0.0000	0.9993
MES_APERTURA_G	**2**		1	-0.4436	68.8739	0.0000	0.9949
MES_APERTURA_G	**3**		1	0.1621	47.1324	0.0000	0.9973
EDAD*TIPO_PRODUCTO_G	**1**		1	0.0394	0.00211	347.0247	<.0001
EDAD*TIPO_PRODUCTO_G	**2**		1	0.00730	0.00252	8.3909	0.0038
EDAD*TIPO_PRODUCTO_G	**3**		1	-0.0175	0.00232	56.6757	<.0001
EDAD*TIPO_PRODUCTO_G	**4**		1	0.0127	0.00266	22.7831	<.0001
EDAD*TIPO_PRODUCTO_G	**5**		1	0.00991	0.00380	6.7913	0.0092
TIPO_PRODUCTO_G	**1**		1	-5.5464	31.8743	0.0303	0.8619
TIPO_PRODUCTO_G	**2**		1	-3.0047	31.8743	0.0089	0.9249
TIPO_PRODUCTO_G	**3**		1	-1.6730	31.8743	0.0028	0.9581
TIPO_PRODUCTO_G	**4**		1	-2.6839	31.8743	0.0071	0.9329
TIPO_PRODUCTO_G	**5**		1	-2.4284	31.8746	0.0058	0.9393
NIVEL_ESTUDIOS	**Formacion Elemental (estudios primarios)**		1	0.7928	48.8556	0.0003	0.9871
NIVEL_ESTUDIOS	**Nivel basico (ESO, EGB, FP)**		1	0.4504	36.2180	0.0002	0.9901
NIVEL_ESTUDIOS	**Nivel de formacion alto**		1	-0.4083	65.7267	0.0000	0.9950
NIVEL_ESTUDIOS	**Nivel de formacion universitario**		1	-0.6579	47.9312	0.0002	0.9890
MES_APERT*TIPO_PRODU	**1**	**1**	1	0.1031	35.7033	0.0000	0.9977
MES_APERT*TIPO_PRODU	**1**	**2**	1	-0.0666	35.7033	0.0000	0.9985
MES_APERT*TIPO_PRODU	**1**	**3**	1	-0.7474	35.7034	0.0004	0.9833
MES_APERT*TIPO_PRODU	**1**	**4**	1	0.1756	35.7033	0.0000	0.9961
MES_APERT*TIPO_PRODU	**1**	**5**	1	0.4588	35.7033	0.0002	0.9897
MES_APERT*TIPO_PRODU	**2**	**1**	1	0.0934	68.8739	0.0000	0.9989
MES_APERT*TIPO_PRODU	**2**	**2**	1	0.2126	68.8740	0.0000	0.9975
MES_APERT*TIPO_PRODU	**2**	**3**	1	0.3033	68.8741	0.0000	0.9965
MES_APERT*TIPO_PRODU	**2**	**4**	1	-0.1057	68.8740	0.0000	0.9988
MES_APERT*TIPO_PRODU	**2**	**5**	1	-0.8777	68.8739	0.0002	0.9898
MES_APERT*TIPO_PRODU	**3**	**1**	1	-0.2762	47.1325	0.0000	0.9953
MES_APERT*TIPO_PRODU	**3**	**2**	1	-0.2637	47.1325	0.0000	0.9955

Analysis of Maximum Likelihood Estimates							
Parameter			**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
MES_APERT*TIPO_PRODU	**3**	**3**	1	0.3563	47.1326	0.0001	0.9940
MES_APERT*TIPO_PRODU	**3**	**4**	1	-0.0556	47.1325	0.0000	0.9991
MES_APERT*TIPO_PRODU	**3**	**5**	1	0.3168	47.1325	0.0000	0.9946
NIVEL_EST*TIPO_PRODU	**Formacion Elemental (estudios primarios)**	**1**	1	0.1268	48.8557	0.0000	0.9979
NIVEL_EST*TIPO_PRODU	**Formacion Elemental (estudios primarios)**	**2**	1	-0.4191	48.8557	0.0001	0.9932
NIVEL_EST*TIPO_PRODU	**Formacion Elemental (estudios primarios)**	**3**	1	0.5133	48.8557	0.0001	0.9916
NIVEL_EST*TIPO_PRODU	**Formacion Elemental (estudios primarios)**	**4**	1	0.2742	48.8557	0.0000	0.9955
NIVEL_EST*TIPO_PRODU	**Formacion Elemental (estudios primarios)**	**5**	1	0.2990	48.8558	0.0000	0.9951
NIVEL_EST*TIPO_PRODU	**Nivel basico (ESO, EGB, FP)**	**1**	1	0.0263	36.2180	0.0000	0.9994
NIVEL_EST*TIPO_PRODU	**Nivel basico (ESO, EGB, FP)**	**2**	1	-0.1103	36.2180	0.0000	0.9976
NIVEL_EST*TIPO_PRODU	**Nivel basico (ESO, EGB, FP)**	**3**	1	0.6305	36.2180	0.0003	0.9861
NIVEL_EST*TIPO_PRODU	**Nivel basico (ESO, EGB, FP)**	**4**	1	-0.1081	36.2180	0.0000	0.9976
NIVEL_EST*TIPO_PRODU	**Nivel basico (ESO, EGB, FP)**	**5**	1	0.0211	36.2180	0.0000	0.9995
NIVEL_EST*TIPO_PRODU	**Nivel de formacion alto**	**1**	1	0.0595	65.7268	0.0000	0.9993
NIVEL_EST*TIPO_PRODU	**Nivel de formacion alto**	**2**	1	0.3266	65.7268	0.0000	0.9960
NIVEL_EST*TIPO_PRODU	**Nivel de formacion alto**	**3**	1	-0.5883	65.7268	0.0001	0.9929
NIVEL_EST*TIPO_PRODU	**Nivel de formacion alto**	**4**	1	0.0397	65.7268	0.0000	0.9995
NIVEL_EST*TIPO_PRODU	**Nivel de formacion alto**	**5**	1	-0.2274	65.7268	0.0000	0.9972
NIVEL_EST*TIPO_PRODU	**Nivel de formacion universitario**	**1**	1	-0.2388	47.9313	0.0000	0.9960

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
NIVEL_EST*TIPO_PRODU	Nivel de formacion universitario	2	1	0.3303	47.9313	0.0000	0.9945
NIVEL_EST*TIPO_PRODU	Nivel de formacion universitario	3	1	-0.5782	47.9313	0.0001	0.9904
NIVEL_EST*TIPO_PRODU	Nivel de formacion universitario	4	1	-0.0435	47.9313	0.0000	0.9993
NIVEL_EST*TIPO_PRODU	Nivel de formacion universitario	5	1	-0.0416	47.9313	0.0000	0.9993

Tabla 10.1.5. Parámetros del mejor modelo de Regresión Logística

10.2. Código SAS

10.2.1. Variables: Análisis de Correspondencias Simple

Relación de las variables NIVEL DE ESTUDIOS, TIPO PRODUCTO y CANAL OBJETIVO:

10.2.1.1. /*IMPORTAR DATOS */

```
DATA DATOS.ESTUDIOS;
INPUT CANAL_PRODUCTO $ X1-X5;
LABEL X1='Formacion Elemental (estudios primarios)'
X2='Nivel basico (ESO, EGB, FP)' X3='Nivel de formacion
alto' X4='Nivel de formacion universitario' X5='Nivel
medio';
DATALINES;
NOCTACORR 630 1193  289 690 508
NOCTAAHO  1232  1400  242 451 494
NOFONDO 950 1272  346 762 655
NOHIPO  52  166 40  74  61
NOPLANPEN 404 937 185 383 370
NOPRESTAMO  833 1542  205 561 512
NOSEGUROS 124 333 51  104 122
NOTARVIS  1368  2497  443 1024  919
SICTACORR 155 302 119 383 247
SICTAAHO  179 241 322 730 316
SIFONDO 193 527 326 1643  515
SIHIPO  0 0 0 0 0
SIPLANPEN 13  35  34  173 67
SIPRESTAMO  46  172 64  111 93
SISEGUROS 0 0 0 0 0
SITARVIS  89  290 107 340 223
;
```

10.2.1.2. /*PROCEDIMIENTO ACS */

```
PROC CORRESP DATA=DATOS.ESTUDIOS ALL CHI2P PRINT=BOTH;
VAR X1-X5;
ID CANAL_PRODUCTO;
RUN;
```

10.2.1.3. /*AGRUPACIÓN DE LAS FILAS */

```
DATA DATOS.ESTUDIOS2;
INPUT CANAL_PRODUCTO $ X1-X5;
LABEL X1='Formacion Elemental (estudios primarios)'
X2='Nivel basico (ESO, EGB, FP)' X3='Nivel de formacion
alto' X4='Nivel de formacion universitario' X5='Nivel
medio';
DATALINES;
NOCTACORR 630 1193  289 690 508
NOCTAAHO  1232  1400  242 451 494
NOFONDO 950 1272  346 762 655
NOPRESTAMO  833 1542  205 561 512
NOHISPL 580 1436  276 561 553
NOTARVIS  1368  2497  443 1024  919
SICTAAHO  179 241 322 730 316
SIFONDO 193 527 326 1643  515
SICPPT  303 799 324 1007  630
;
```

10.2.1.4. /*PROCEDIMIENTO ACS CON FILAS AGRUPADAS */

```
PROC CORRESP DATA=DATOS.ESTUDIOS2 ALL CHI2P PRINT=BOTH;
VAR X1-X5;
ID CANAL_PRODUCTO;
ods output CellChiSq = Aportaciones;
ods output RowProfiles = PerfilFila;
ods output ColProfiles = PerfilColumna;
RUN;
```

Relación de las variables ESTADO CIVIL, TIPO PRODUCTO y CANAL_OBJETIVO:

10.2.1.5. /*IMPORTAR DATOS */

```
DATA DATOS.ESTADO;
INPUT CANAL_PRODUCTO $ X1-X5;
LABEL X1='Casado' X2='Divorciado' X3='Separado'
X4='Soltero' X5='Viudo';
DATALINES;
NOCTACORR 1442  218 45  1509  96
NOCTAAHO  1661  155 51  1767  185
NOFONDO 2367  125 65  1114  314
NOHIPO  195 30  11  149 8
NOPLANPEN 1515  117 63  534 50
```

```
NOPRESTAMO  1953  304 130 1087  179
NOSEGUROS 411 48  25  242 8
NOTARVIS  3010  359 112 2493  277
SICTACORR 505 63  13  602 23
SICTAAHO  987 55  17  684 45
SIFONDO 1720  88  19  1296  81
SIHIPO  0 0 0 0 0
SIPLANPEN 158 9 3 150 2
SIPRESTAMO  271 43  13  145 14
SISEGUROS 0 0 0 0 0
SITARVIS  474 66  15  476 18
;
```

10.2.1.6. /*PROCEDIMIENTO ACS */

```
PROC CORRESP DATA=DATOS.ESTADO ALL CHI2P PRINT=BOTH;
VAR X1-X5;
ID CANAL_PRODUCTO;
RUN;
```

10.2.1.7. /*AGRUPACIÓN DE LAS FILAS Y COLUMNAS */

```
DATA DATOS.ESTADO2;
INPUT CANAL_PRODUCTO $ X1-X4;
LABEL X1='Casado' X2='Divorciado' X3='Separado-Viudo'
X4='Soltero';
DATALINES;
NOCTACORR 1442  218 1509  141
NOCTAAHO  1661  155 1767  236
NOFONDO 2367  125 1114  379
NOPRESTAMO  1953  304 1087  309
NOHISPL 2121  195 925 165
NOTARVIS  3010  359 2493  389
SICTAAHO  987 55  684 62
SIFONDO 1720  88  1296  100
SICPPT  1408  181 1373  101
;
```

10.2.1.8. /*PROCEDIMIENTO ACS CON FILAS Y COLUMNAS AGRUPADAS */

```
PROC CORRESP DATA=DATOS.ESTADO2 ALL CHI2P PRINT=BOTH;
VAR X1-X4;
ID CANAL_PRODUCTO;
ods output CellChiSq = Aportaciones;
ods output RowProfiles = PerfilFila;
ods output ColProfiles = PerfilColumna;
RUN;
```

Relación de las variables EDAD, TIPO PRODUCTO y CANAL_OBJETIVO:

10.2.1.9. /*IMPORTAR DATOS */

```
DATA DATOS.EDAD;
INPUT CANAL_PRODUCTO $ X1-X4;
LABEL X1='low-23' X2='24-46' X3='47-69' X4='70-high';
DATALINES;
NOCTACORR 309 1538  1133  330
NOCTAAHO  520 1302  1360  637
NOFONDO 74  672 1899  1340
NOHIPO  7 253 123 10
NOPLANPEN 7 754 1445  73
NOPRESTAMO  60  1591  1734  268
NOSEGUROS 6 391 329 8
NOTARVIS  386 2762  2280  823
SICTACORR 36  711 395 64
SICTAAHO  9 681 903 195
SIFONDO 21  1538  1360  285
SIHIPO  0 0 0 0
SIPLANPEN 0 173 149 0
SIPRESTAMO  0 207 267 12
SISEGUROS 0 0 0 0
SITARVIS  44  602 358 45
;
```

10.2.1.10. /*PROCEDIMIENTO ACS */

```
PROC CORRESP DATA=DATOS.EDAD ALL CHI2P PRINT=BOTH;
VAR X1-X4;
ID CANAL_PRODUCTO;
RUN;
```

10.2.1.11. /*AGRUPACIÓN DE LAS FILAS Y COLUMNAS */

```
DATA DATOS.EDAD2;
INPUT CANAL_PRODUCTO $ X1-X4;
LABEL X1='<25' X2='26-45' X3='46-69' X4='>70';
DATALINES;
NOCTACORR 381 1398  1201  330
NOCTAAHO  574 1169  1439  637
NOFONDO 97  585 1963  1340
NOPRESTAMO  103 1441  1841  268
NOHISPL 48  1275  1992  91
NOTARVIS  533 2492  2403  823
SICTAAHO  12  633 948 195
SIFONDO 41  1457  1421  285
SICPPT  125 1565  1252  121
;
```

10.2.1.12. /*PROCEDIMIENTO ACS CON FILAS Y COLUMNAS AGRUPADAS */

```
PROC CORRESP DATA=DATOS.EDAD2 ALL CHI2P PRINT=BOTH;
VAR X1-X3;
ID CANAL_PRODUCTO;
ods output CellChiSq = Aportaciones;
ods output RowProfiles = PerfilFila;
ods output ColProfiles = PerfilColumna;
RUN;
```

10.2.2. Modelos de Predicción: Introducción

Agrupación de Categorías:

10.2.2.1. /*MACRO DE AGRUPACIÓN DE CATEGORÍAS*/

```
%macro AgruparCategorias(
archivo=,vardep=,vardeptipo=,listclass=,criterio=,directori
o=c:);
%if &criterio eq %then %do;
 %if &vardeptipo=I %then %let criterio=PROBF;
 %if &vardeptipo=N %then %let criterio=PROBCHISQ;
%end;
  data archivosa;
    set &archivo (KEEP = &vardep &listclass);
  run;
  data _null_;
    file
"&directorio\tempAgrupacionClasesVariableNominal.txt";
    put ' ';
  run;
  /* data temporal;
    retain variable ' ';
  run;
  */
  data _null_;
    length clase $ 10000 ;
  /* Cuento el número de variables */
  clase="&listclass";
    ncate= 1;
    do while (scanq(clase, ncate) ^= '');
      ncate+1;
    end;
    ncate+(-1);put;
    put // ncate= /;
    call symput('ncate',left(ncate));
  run;
  /* Bucle arboretum */
  %do i=1 %to &ncate;
    %let vari=%qscan(&listclass,&i);
    %if %upcase(&vardeptipo)=I %then %do;
      proc arboretum data=archivosa criterion=&criterio;
```

```
        input &vari / level=nominal;
        target &vardep / level=interval;
        save model=tree1;
      run;
    %end;
    %else %do;
      proc arboretum data=archivosa criterion=&criterio;
        input &vari / level=nominal;
        target &vardep / level=nominal;
        save model=tree1;
      run;
    %end;
    proc arboretum inmodel=tree1;
      score data=archivosa out=archivosa2 ;
      subtree best;
    run;
    data archivosa;
      set archivosa2;
    run;

    proc freq data=archivosa noprint;
      tables &vari /out=sal1;
    proc freq data=archivosa noprint;
      tables _leaf_ /out=sal2;
    data _null_;
      if _n_=1 then set sal1 nobs=nume1;
      if _n_=1 then set sal2 nobs=nume2;
      if _n_=1 then do;
        if nume1=nume2 then noagrupa=1;
        else noagrupa=0;
        call symput ('noagrupa',left(noagrupa));
      end;
      if noagrupa=1 then do;
        put 'NOAGRUPA  ' "&vari";
        file
"&directorio\tempAgrupacionClasesVariableNominal.txt" mod;
        put "&vari";
      end;
      stop;
    run;

      proc freq data=archivosa noprint;
      tables _leaf_ /out=sal1;
    run;
    data _null_;
      set sal1 nobs=nume;
      call symput ('seunentodas',left(nume));
      if nume=1 then do;
        put 'SE UNEN TODAS   ' "&vari";
        file
"&directorio\tempAgrupacionClasesVariableNominal.txt" mod;
```

```
        put "&vari";
      end;
    run;
    %if &noagrupa eq 0 and &seunentodas ne 1 %then %do;
      data _null_;koko2=cats("&vari",'_G');call
symput('koko',left(koko2));run;
      data archivosa (drop=_node_ );
        set archivosa;
        rename _leaf_=&koko;
      run;
      data _null_;
        file
"&directorio\tempAgrupacionClasesVariableNominal.txt" mod;
        h="&koko";h=left(h);
        put h;
      run;
    %end;
    %else %do;
      data archivosa(drop=_leaf_ _node_);
        set archivosa;
      run;
    %end;
  %end;
  data archivofinal (drop=P_&vardep R_&vardep);
    merge &archivo archivosa;
  run;
  data _null_;
    length c $ 300;
    if _n_=1 then put ' '//'LISTA DE GRUPOS CREADOS Y NO
CREADOS'//'****************************************' ;
      infile
"&directorio\tempAgrupacionClasesVariableNominal.txt" ;
      input c $;
      put c @@;
  run;
  data _null_;put
//'****************************************' ;run;
/* COMPROBAR GRUPOS CREADOS */
%do i=1 %to &ncate;
  %let vari=%qscan(&listclass,&i);
  data _null_;retain control 0;length c $ 300;infile
"&directorio\tempAgrupacionClasesVariableNominal.txt"
;input c $;
    c3=cats("&vari",'_G');
    if c=c3 then control=1;
    call symput('control',left(control));
    call symput('grupo',left(c3));
  run;
  %if &control=1 %then %do;
   proc freq data=archivofinal noprint;tables &vari*&grupo
/out=sal;run;
```

```
  proc sort data=sal;by &grupo;
  proc print data=sal;run;
  %end;
%end;
%mend;
```

10.2.2.2. /*AGRUPACIÓN CATEGORÍAS EN LOS DATOS*/

```
%AgruparCategorias(archivo= DATOS.BANCO,
vardep=CANAL_OBJETIVO, vardeptipo=I,
listclass=SEXO HIJOS MES_APERTURA ESTADO_CIVIL
NIVEL_ESTUDIOS TIPO_PRODUCTO, criterio=PROBF,
directorio=C:\Users\Montserrat\Desktop\tfm1\TXT\Prediccion)
;
```

10.2.2.3. /*ARCHIVO FINAL CON AGRUPACIÓN DE CATEGORÍAS*/

```
proc contents data=archivofinal;run;
```

10.2.2.4. /*GUARDAR EL FICHERO FINAL*/

```
data DATOS.BANCO_G(drop=_WARN_ ESTADO_CIVIL
LG10_ANO_APERTURA MES_APERTURA OPT_DIA_APERTURA OPT_EDAD
OPT_NUMERO_DE_HIJOS TIPO_PRODUCTO);
set archivofinal;
run;
```

Interacciones:

10.2.2.5. /*MACRO DE INTERACCIONES*/

```
%macrointeracttodo(archivo=,vardep=,listclass=,listconti=,i
nterac=1,directorio=c:);
proc printto print="&directorio\kaka.txt";run;
data _null_;file "&directorio\inteconti.txt";put ' ';file
"&directorio\intecategor.txt";put ' ';run;

data _null_;
length clase conti con cruce1 $ 32000 cruce2 $ 32000;
clase="&listclass";
conti="&listconti";
  ncate= 1;
  do while (scan(clase, ncate) ^= '');
    ncate+1;
  end;
  ncate+(-1);
  put ncate=;
  nconti= 1;
  do while (scan(conti, nconti) ^= '');
    nconti+1;
  end;
  nconti+(-1);
```

```
  put nconti=;

call symput('ncate',left(ncate));
call symput('nconti',left(nconti));

%if &interac=1 %then %do;
cruce2=' ';
do i=1 to ncate;
  do j=1 to nconti;
    ca=scan(clase,i);
    con=scan(conti,j);
    cruce1=cats(ca,'*',con);
    file "&directorio\inteconti.txt" mod;
    put cruce1;
  end;
end;

cruce2=' ';
do i=1 to ncate-1;
  do j=i+1 to ncate;
    ca=scan(clase,i);
    con=scan(clase,j);
    if i ne j then cruce1=cats(ca,'*',con);else cruce1=' ';
    file "&directorio\intecategor.txt" mod;
    put cruce1;
  end;
end;
run;
%end;
data union;run;

%if &listclass ne %then %do i=1 %to &ncate;
data _null_;cosa="&listclass";va=scanq(cosa,&i);
call symput ('vari',va);
run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
class &vari;
model &vardep=&vari /selection=none;
run;

proc print data=tanova;run;

data a;set ajuste (where=(Label1='AIC'));AIC=cvalue1;keep
AIC;run;
data b(keep=Fvalue probf);set tanova;if _n_=1 then
output;stop;run;
data c;length variable $ 1000;merge a
b;variable="&vari";run;
data union;set union c;run;
```

```
%end;

%if &interac=1 %then %do;

%if &ncate>1 %then %do;

data pr234;
length vari $ 1000;
infile "&directorio\intecategor.txt";
input vari;
run;
data _null_;set pr234 nobs=nume;ko=nume;
call symput('nintecat',left(ko));stop;
run;

%if &listclass ne %then %do i=1 %to &nintecat;
data _null_;ko=&i;
set pr234 point=ko;
var1=scan(vari,1);
var2=scan(vari,2);
lista1=compbl(var1||'  '||var2);
call symput('lista1',left(lista1));
call symput('vari',left(vari));
stop;
run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
class &lista1;
model &vardep=&vari / selection=none;
run;

data a;set ajuste (where=(Label1='AIC'));
AIC=cvalue1;keep AIC;
data b(keep=Fvalue probf);set tanova;if _n_=1 then
output;stop;run;
data c;length variable $ 1000;merge a
b;variable="&vari";run;
data union;set union c;run;

%end;
data _null_;if _n_=1 then put 'LISTA CLASE E
INTERACCIONES';set union;put variable @@;run;
%end;

%end;

%if &listconti ne %then %do i=1 %to &nconti;
data _null_;cosa="&listconti";va=scanq(cosa,&i);
call symput ('vari',va);
```

```
run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
model &vardep=&vari /selection=none;
run;

data a;set ajuste (where=(Label1='AIC'));AIC=cvalue1;keep
AIC;run;
data b(keep=Fvalue probf);set tanova;if _n_=1 then
output;stop;run;
data c;length variable $ 1000;merge a
b;variable="&vari";run;
data union;set union c;run;
%end;

%if &interac=1 %then %do;
data pr235;
length vari $ 1000;
infile "&directorio\inteconti.txt";
input vari;
run;

data _null_;set pr235 nobs=nume;ko=nume;
call symput('ninteconti',left(ko));stop;
run;

%if (&listclass ne) and (&listconti ne) %then %do i=1 %to
&ninteconti;
data _null_;ko=&i;
set pr235 point=ko;
var1=scan(vari,1);
var2=scan(vari,2);
call symput('lista1con',left(var1));
call symput('varicon',left(vari));
stop;
run;

ods output FitStatistics=ajuste anova=tanova;
proc glmselect data=&archivo ;
class &lista1con;
model &vardep=&varicon / selection=none;
run;

data a;set ajuste (where=(Label1='AIC'));AIC=cvalue1;keep
AIC;
data b(keep=Fvalue probf);set tanova;if _n_=1 then
output;stop;run;
data c;length variable $ 1000;merge a
b;variable="&varicon";run;
data union;set union c;run;
```

```
%end;
%end;
proc printto;run;
data union;set union;if _n_=1 then delete;run;
proc sort data=union;by AIC;
proc print data=union;run;
data _null_;set union;put variable @@;run;
%mend;
```

10.2.2.6. /*INTERACCIONES EN NUESTROS DATOS*/

```
%interacttodo(archivo=DATOS.BANCO_G,
vardep=CANAL_OBJETIVO,
listclass=SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
listconti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
interac=1,
directorio=
C:\Users\Montserrat\Desktop\tfm1\TXT\Prediccion);
```

Selección de Variables:

10.2.2.7. /*MACRO DE SELECCIÓN DE VARIABLES*/

```
%macrorandomselectlog(data=,listclass=,vardepen=,modelo=,si
nicio=,sfinal=,fracciontrain=,directorio=);
options nocenter linesize=256;
proc printto print="&directorio\kk.txt";run;
data;file "&directorio\cosa2.txt" ;run;
%do semilla=&sinicio %to &sfinal;
proc surveyselect data=&data rate=&fracciontrain out=sal1234
seed=&semilla;run;
%if &listclass ne %then %do;
ods output type3=parametros;
proc logistic data=sal1234;
class &listclass;
model &vardepen= &modelo/ selection=stepwise;
run;
data parametros;length effect $20. modelo $ 20000;retain
modelo " ";set parametros end=fin;effect=cat(' ',effect);
if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then
do;variable=modelo;output;end;
run;
%end;
%else %do;
ods output Logistic.ParameterEstimates=parametros;
proc logistic data=sal1234;
model &vardepen= &modelo/ selection=stepwise;
run;
%end;
```

```
ods graphics off;
ods html close;
data;file "&directorio\cosa2.txt" mod;set parametros;
%if &listclass ne %then %do; put variable @@;%end;
%else %do; if _n_ ne 1 then put variable @@;%end;
run;
%end;
proc printto ;run;
data todos;
infile "&directorio\cosa2.txt";
length efecto $ 400;
input efecto @@;
if efecto ne 'Intercept' then output;
run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;
data todos;
infile "&directorio\cosa2.txt";
length efecto $ 200;
input efecto $ &&;
run;
proc freq data=todos;tables efecto /out=sal;run;
proc sort data=sal;by descending count;
proc print data=sal;run;
data;set sal;put efecto;run;
%mend;
```

10.2.2.8. /*SELECCIÓN DE VARIABLES EN LOS DATOS*/

```
%randomselectlog(data= DATOS.BANCO_G,
listclass= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
vardepen=CANAL_OBJETIVO,
modelo=DIA_APERTURA NUMERO_DE_HIJOS HIJOS*NUMERO_DE_HIJOS
HIJOS HIJOS*DIA_APERTURA
MES_APERTURA_G*NUMERO_DE_HIJOS EDAD SEXO*NUMERO_DE_HIJOS
HIJOS*EDAD ANO_APERTURA
ESTADO_CIVIL_G*NUMERO_DE_HIJOS HIJOS*ANO_APERTURA
ESTADO_CIVIL_G*DIA_APERTURA SEXO*DIA_APERTURA
ESTADO_CIVIL_G MES_APERTURA_G*DIA_APERTURA SEXO
ESTADO_CIVIL_G*EDAD HIJOS*ESTADO_CIVIL_G
MES_APERTURA_G SEXO*HIJOS SEXO*EDAD
ESTADO_CIVIL_G*ANO_APERTURA HIJOS*MES_APERTURA_G
MES_APERTURA_G*EDAD SEXO*ANO_APERTURA
MES_APERTURA_G*ANO_APERTURA SEXO*ESTADO_CIVIL_G
MES_APERTURA_G*ESTADO_CIVIL_G SEXO*MES_APERTURA_G
NIVEL_ESTUDIOS*NUMERO_DE_HIJOS
TIPO_PRODUCTO_G*NUMERO_DE_HIJOS
TIPO_PRODUCTO_G*DIA_APERTURA NIVEL_ESTUDIOS*DIA_APERTURA
```

```
TIPO_PRODUCTO_G*EDAD NIVEL_ESTUDIOS*EDAD TIPO_PRODUCTO_G
HIJOS*TIPO_PRODUCTO_G
TIPO_PRODUCTO_G*ANO_APERTURA NIVEL_ESTUDIOS
HIJOS*NIVEL_ESTUDIOS NIVEL_ESTUDIOS*ANO_APERTURA
ESTADO_CIVIL_G*NIVEL_ESTUDIOS SEXO*TIPO_PRODUCTO_G
MES_APERTURA_G*NIVEL_ESTUDIOS
SEXO*NIVEL_ESTUDIOS ESTADO_CIVIL_G*TIPO_PRODUCTO_G
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
sinicio=12345,sfinal=12365,fracciontrain=0.7,
directorio=
C:\Users\Montserrat\Desktop\tfm1\TXT\Prediccion);
```

10.2.3. Modelos de Predicción: Regresión Logística

10.2.3.1. /*MACRO DE REGRESIÓN LOGÍSTICA*/

```
%macro
cruzadalogistica(archivo=,vardepen=,conti=,categor=,ngrupos
=,sinicio=,sfinal=,objetivos=,corte=0.5,porcaptura=0,direct
orio= C:\Users\Montserrat\Desktop\REDES\REDES TRABAJO2);
data final;run;
/* contar objetivos */
data _null_;length clase $ 300;
clase="&objetivos";
nobje= 1;
do while (scanq(clase, nobje) ^= '');
nobje+1;
end;
nobje+(-1);
call symput('nobje',left(nobje));
run;
proc printto print="&directorio\outp.txt"
log="&directorio\log.txt";run;/*SE PUEDE QUITAR EL PROC
PRINTTO, POR SI ACASO HAY PROBLEMAS*/
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;
data dos ;
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;
data fantasma;run;
%do exclu=1 %to &ngrupos;
data tres;set dos;if grupo ne &exclu then vardep=&vardepen;
ods output ROCAssociation=roca;
proc logistic data=tres ROCOPTIONS(NODETAILS)
PLOTS=NONE;/*<<<<<******SE PUEDE QUITAR EL NOPRINT */
%if (&categor ne) %then %do;class &categor;model
```

```
vardep=&conti;%end;
%else %do;model vardep=&conti;%end;
output out=sal p=predi;roc;run;
data sal2;set sal;pro=1-predi;if pro>&corte then pre11=1;
else pre11=0;
if grupo=&exclu then output;run;
proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
data estadisticos (drop=count percent pre11 &vardepen);
retain vp vn fp fn suma 0;
if _n_=1 then set roca;
set sal3 nobs=nume;
suma=suma+count;
if pre11=0 and &vardepen=0 then vn=count;
if pre11=0 and &vardepen=1 then fn=count;
if pre11=1 and &vardepen=0 then fp=count;
if pre11=1 and &vardepen=1 then vp=count;
if _n_=nume then do;
porcenVN=vn/suma;
porcenFN=FN/suma;
porcenVP=VP/suma;
porcenFP=FP/suma;
sensi=vp/(vp+fn);
especif=vn/(vn+fp);
tasafallos=1-(vp+vn)/suma;
tasaciertos=1-tasafallos;
precision=vp/(vp+fp);
F_M=2*Sensi*Precision/(Sensi+Precision);
Mcc=VP*VN-
FP*FN;be=(VP+FP)*(VP+FN)*(VN+FP)*(VN+FN);be=sqrt(be);
MCC=MCC/be;
Youden=especif+sensi-1;
AUC=Area;
output;
end;
run;

%if &porcaptura ne 0 %then %do;
proc sort data=sal2;by descending pre11;
data sal4;retain suma1 0;set sal2 nobs=nume;
if &vardepen=1 then suma1=suma1+1;
if _n_=int(&porcaptura*nume) then
do;ncapturados=suma1;capturados=suma1/_n_;ntot=_n_;output;
stop;end;
run;
data estadisticos;set estadisticos;if _n_=1 then set
sal4;run;
%end;
data estadisticos;set estadisticos;
keep AUC F_M Mcc Youden ncapturados ntot capturados especif
fn fp porcenFN porcenFP porcenVN porcenVP precision
sensi tasaciertos tasafallos vn vp;
```

```
run;
data fantasma;set fantasma estadisticos;run;
%end;
proc means data=fantasma sum noprint;var &objetivos;
output out=sumaresi sum=suma mean=media1-media&nobje;
run;
data sumaresi;set sumaresi;semilla=&semilla;
data final (keep=suma media1-media&nobje semilla);set final
sumaresi;if suma=. then delete;run;
/* renombrar objetivos para entender mejor */
%end;
data _null_;
file "&directorio\kk.txt";
put 'data final;set final;array media{'"&nobje" '};';
%do i=1 %to &nobje;
%let vari=%qscan(&objetivos,&i);
put "&vari" '=media{'"&i"'};';
%end;
put 'drop suma media1-media'"&nobje"';output;run;';
run;
%include "&directorio\kk.txt";
proc printto;run;
proc print data=final;run;
%mend;
```

10.2.3.2. /*REGRESIÓN LOGÍSTICA EN LOS DATOS*/

Semilla Inicio: 12345- Semilla Final: 12356

```
%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= MES_APERTURA_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=12345,sfinal=12356,objetivos=tasafallos);
data final1;set final;modelo=1;

%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= ANO_APERTURA MES_APERTURA_G EDAD*TIPO_PRODUCTO_G
TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=12345,sfinal=12356,objetivos=tasafallos);
data final2;set final;modelo=2;
```

```
%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= SEXO MES_APERTURA_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=12345,sfinal=12356,objetivos=tasafallos);
data final3;set final;modelo=3;

%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= ANO_APERTURA SEXO MES_APERTURA_G
EDAD*TIPO_PRODUCTO_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=12345,sfinal=12356,objetivos=tasafallos);
data final4;set final;modelo=4;run;

data union;set final1 final2 final3 final4;
proc boxplot data=union;plot tasafallos*modelo;run;
```

Semilla Inicio: 54321- Semilla Final: 54340

```
%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= MES_APERTURA_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,objetivos=tasafallos);
data final5;set final;modelo=5;

%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= ANO_APERTURA MES_APERTURA_G EDAD*TIPO_PRODUCTO_G
TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,objetivos=tasafallos);
data final6;set final;modelo=6;
```

```
%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= SEXO MES_APERTURA_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,objetivos=tasafallos);
data final7;set final;modelo=7;

%cruzadalogistica
(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti= ANO_APERTURA SEXO MES_APERTURA_G
EDAD*TIPO_PRODUCTO_G TIPO_PRODUCTO_G NIVEL_ESTUDIOS
MES_APERTURA_G*TIPO_PRODUCTO_G
NIVEL_ESTUDIOS*TIPO_PRODUCTO_G,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,objetivos=tasafallos);
data final8;set final;modelo=8;

data union;set final5  final6  final7  final8;
proc boxplot data=union;plot tasafallos*modelo;run;
```

10.2.4. Modelos de Predicción: Red Neuronal

10.2.4.1. /*MACRO DE RED NEURONAL*/

```
%macrovariar(seminicio=,semifin=,inicionodos=,finalnodos=,i
ncrenodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;

%neuralbinariabasica(archivo=redes.datos_grupos,
listconti=EDAD HIJOS,
listclass= IMP_REP_ESTADO_CIVIL_G IMP_REP_NIVEL_ESTUDIOS_G
TIPO_PRODUCTO,vardep=SEXO2,nodos=&nodos,corte=50,semilla=&s
emilla,porcen=0.70,algo=brop);
data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
```

```
proc boxplot data=union;plot (porcenVN porcenFN porcenVP
porcenFP
sensi especif tasafallos tasaciertos precision
F_M)*nodos;run;
%mend;

%macro
cruzadabinarianeural(archivo=,vardepen=,conti=,categor=,ngr
upos=,sinicio=,sfinal=,nodos=,algo=,objetivo=,early=300,
acti=tanh);
title '';
data final;run;
proc printto print='c:\basura.txt';
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;
data dos (drop=nume);
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;
data fantasma;run;
%do exclu=1 %to &ngrupos;
data trestr tresval;
set dos;if grupo ne &exclu then output trestr;else output
tresval;
PROC DMDB DATA=trestr dmdbcat=catatres;
target &vardepen;
var &conti;
class &vardepen;
%if &categor ne %then %do;class &categor &vardepen;%end;
run;
proc neural data=trestr dmdbcat=catatres random=789 ;
input &conti;
%if &categor ne %then %do;input &categor
/level=nominal;%end;
target &vardepen /level=nominal;
hidden &nodos /acti=&acti
/*nloptions maxiter=500*/;
netoptions randist=normal ranscale=0.15 random=15459;
/*prelim 0 */
prelim 15 preiter=10 pretech=&algo;
train maxiter=&early outest=mlpest technique=&algo;
score data=tresval role=valid out=sal ;
run;
data sal2;set sal;pro=1-%str(p_&vardepen)0;if pro>0.5 then
pre11=1; else pre11=0;run;
proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
data estadisticos (drop=count percent pre11 &vardepen);
retain vp vn fp fn suma 0;
```

```
set sal3 nobs=nume;
suma=suma+count;
if pre11=0 and &vardepen=0 then vn=count;
if pre11=0 and &vardepen=1 then fn=count;
if pre11=1 and &vardepen=0 then fp=count;
if pre11=1 and &vardepen=1 then vp=count;
if _n_=nume then do;
porcenVN=vn/suma;
porcenFN=FN/suma;
porcenVP=VP/suma;
porcenFP=FP/suma;
sensi=vp/(vp+fn);
especif=vn/(vn+fp);
tasafallos=1-(vp+vn)/suma;
tasaciertos=1-tasafallos;
precision=vp/(vp+fp);
F_M=2*Sensi*Precision/(Sensi+Precision);
output;
end;
run;
data fantasma;set fantasma estadisticos;run;
%end;
proc means data=fantasma sum noprint;var &objetivo;
output out=sumaresi sum=suma mean=media;
run;
data sumaresi;set sumaresi;semilla=&semilla;
data final (keep=suma media semilla);set final sumaresi;if
suma=. then delete;run;
%end;
proc printto ;
proc print data=final;run;
%mend;
```

10.2.4.2. /*RED NEURONAL EN LOS DATOS (I)*/

```
%variar(seminicio=12345,semifin=12354,inicionodos=2,finalno
dos=10,increnodos=2);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor=SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=12345,sfinal=12356,nodos=5,algo=bprop
mom=0.2 learn=0.1);
data final9;set final;modelo='Neural1'; run;
proc print data=final9;run;

%variar(seminicio=12345,semifin=12354,inicionodos=2,finalno
dos=10,increnodos=2);
```

```
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=12345,sfinal=12356,nodos=5,algo=quanew
mom=0.2 learn=0.1);
data final10;set final;modelo='Neural2';run;
proc print data=final10;run;

%variar(seminicio=54321,semifin=54340,inicionodos=3,finalno
dos=11,increnodos=3);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,nodos=3,algo=quanew
mom=0.2 learn=0.1);
data final11;set final;modelo='Neural3';run;
proc print data=final11;run;

%variar(seminicio=54321,semifin=54340,inicionodos=3,finalno
dos=11,increnodos=3);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,nodos=3,algo= bprop
mom=0.2 learn=0.1);
data final12;set final;modelo='Neural4';run;
proc print data=final12;run;

data union;set final9 final10 final11 final12;
proc boxplot data=union;plot media*modelo;run;
```

10.2.4.3. /*RED NEURONAL EN LOS DATOS (II)*/

```
%variar(seminicio=54321,semifin=54340,inicionodos=2,finalno
dos=8,increnodos=2);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
```

```
ngrupos=4,sinicio=54321,sfinal=54340,nodos=4,algo=quanew
mom=0.2 learn=0.1);
data final13;set final;modelo='Neural1';run;
proc print data=final13;run;

%variar(seminicio=54321,semifin=54340,inicionodos=2,finalno
dos=10,increnodos=2);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,nodos=5,algo=quanew
mom=0.2 learn=0.1);
data final14;set final;modelo='Neural2';run;
proc print data=final14;run;

%variar(seminicio=54321,semifin=54340,inicionodos=2,finalno
dos=12,increnodos=2);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,nodos=6,algo=quanew
mom=0.2 learn=0.1);
data final15;set final;modelo='Neural3';run;
proc print data=final15;run;

%variar(seminicio=54321,semifin=54340,inicionodos=2,finalno
dos=14,increnodos=2);
%cruzadabinarianeural(archivo=DATOS.BANCO_G,vardepen=CANAL_
OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
objetivo=tasafallos,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,sinicio=54321,sfinal=54340,nodos=7,algo=quanew
mom=0.2 learn=0.1);
data final16;set final;modelo='Neural4';run;
proc print data=final16;run;

data union;set final13 final14 final15 final16;
proc boxplot data=union;plot media*modelo;run;
```

10.2.5. Modelos de Predicción: Gradient Boosting

10.2.5.1. /*MACRO DE GRADIENT BOOSTING*/

```
%macrocruzadatreeboostbin(archivo=,vardepen=,conti=,categor
=,ngrupos=,sinicio=,sfinal=,criterion=ProbF,leafsize=,nleav
es=,
iteraciones=,shrink=0.01,maxbranch=2,maxdepth=4,mincatsize=
15,minobs=20,objetivo=tasafallos,);
data final;run;
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;
data dos ;
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;
data fantasma;run;
%do exclu=1 %to &ngrupos;
data tres;set dos;if grupo ne &exclu then vardep=&vardepen;
proc treeboost data=tres
exhaustive=1000 intervaldecimals=max
leafsize=&leafsize iterations=&iteraciones
maxbranch=&maxbranch
maxdepth=&maxdepth mincatsize=&mincatsize
missing=useinsearch shrinkage=&shrink
splitsize=&minobs;
%if (&categor ne) %then %do;
input &categor/level=nominal;
%end;
input &conti/level=interval;
target vardep /level=binary;
save fit=iteraciones importance=impor model=modelo
rules=reglas;
subseries largest;
score out=sal;
data sal2;set sal;pro=1-p_vardep0;if pro>0.5 then pre11=1;
else pre11=0;
if grupo=&exclu then output;run;
proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
data estadisticos (drop=count percent pre11 &vardepen);
retain vp vn fp fn suma 0;
set sal3 nobs=nume;
suma=suma+count;
if pre11=0 and &vardepen=0 then vn=count;
if pre11=0 and &vardepen=1 then fn=count;
if pre11=1 and &vardepen=0 then fp=count;
if pre11=1 and &vardepen=1 then vp=count;
if _n_=nume then do;
porcenVN=vn/suma;
porcenFN=FN/suma;
porcenVP=VP/suma;
```

```
porcenFP=FP/suma;
sensi=vp/(vp+fn);
especif=vn/(vn+fp);
tasafallos=1-(vp+vn)/suma;
tasaciertos=1-tasafallos;
precision=vp/(vp+fp);
F_M=2*Sensi*Precision/(Sensi+Precision);
output;
end;
run;
data fantasma;set fantasma estadisticos;run;
%end;
proc means data=fantasma sum noprint;var &objetivo;
output out=sumaresi sum=suma mean=media;
run;
data sumaresi;set sumaresi;semilla=&semilla;
data final (keep=suma media semilla);set final sumaresi;if
suma=. then delete;run;
%end;
proc print data=final;run;
%mend;
```

10.2.5.2. /*GRADIENT BOOSTING EN LOS DATOS*/

```
%cruzadatreeboostbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=12345,
sfinal=12356,
leafsize=5,nleaves=10,iteraciones=20,shrink=0.03,maxbranch=
2,maxdepth=4,mincatsize=15,minobs=20,objetivo=tasafallos);
data final17;set final;modelo='BTG1';RUN;

%cruzadatreeboostbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=12345,
sfinal=12356,
leafsize=8,nleaves=20,iteraciones=30,shrink=0.03,maxbranch=
2,maxdepth=4,mincatsize=15,minobs=20,objetivo=tasafallos);
data final18;set final;modelo='BTG2';RUN;

%cruzadatreeboostbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
```

```
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=54321,
sfinal=54340,
leafsize=10,nleaves=30,iteraciones=40,shrink=0.03,maxbranch
=2,maxdepth=4,mincatsize=15,minobs=20,objetivo=tasafallos);
data final19;set final;modelo='BTG3';RUN;

%cruzadatreeboostbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
conti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
categor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=54321,
sfinal=54340,
leafsize=13,nleaves=40,iteraciones=50,shrink=0.03,maxbranch
=2,maxdepth=4,mincatsize=15,minobs=20,objetivo=tasafallos);
data final120;set final;modelo='BTG4';RUN;

data union; set final17 final18 final19 final20;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;
```

10.2.6. Modelos de Predicción: Random Forest

10.2.6.1. /*MACRO DE RANDOM FOREST*/

```
%macrocruzadarandomforestbin(archivo=,vardep=,listconti=,li
stcategor=,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=2,tamhoja
=5,maxdepth=10,pvalor=0.1,
ngrupos=4,sinicio=12345,sfinal=12356,objetivo=tasafallos);
data final;run;
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;
data dos ;
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;
data fantasma;run;
%do exclu=1 %to &ngrupos;
data tres;set dos;if grupo ne &exclu then vardep=&vardep;
ods listing close;
proc hpforest data=tres
maxtrees=&maxtrees
```

```
vars_to_try=&variables
trainfraction=&porcenbag
leafsize=&tamhoja
maxdepth=&maxdepth
alpha=&pvalor
exhaustive=5000
missing=useinsearch ;
target vardep/level=nominal;
input &listconti/level=interval;
%if (&listcategor ne) %then %do;
input &listcategor/level=nominal;
%end;
score out=salo;
run;
ods listing ;
data salo;merge salo tres;
if p_vardep1>0.5 then pre11=1;else pre11=0;
if grupo=&exclu;
run;
proc freq data=salo;tables pre11*&vardep/out=sal3;run;
data estadisticos (drop=count percent pre11 &vardep);
retain vp vn fp fn suma 0;
set sal3 nobs=nume;
suma=suma+count;
if pre11=0 and &vardep=0 then vn=count;
if pre11=0 and &vardep=1 then fn=count;
if pre11=1 and &vardep=0 then fp=count;
if pre11=1 and &vardep=1 then vp=count;
if _n_=nume then do;
porcenVN=vn/suma;
porcenFN=FN/suma;
porcenVP=VP/suma;
porcenFP=FP/suma;
sensi=vp/(vp+fn);
especif=vn/(vn+fp);
tasafallos=1-(vp+vn)/suma;
tasaciertos=1-tasafallos;
precision=vp/(vp+fp);
F_M=2*Sensi*Precision/(Sensi+Precision);
output;
end;
run;
data fantasma;set fantasma estadisticos;run;
%end;/* fin grupos */
proc means data=fantasma sum noprint;var &objetivo;
output out=sumaresi sum=suma mean=media;
run;
data sumaresi;set sumaresi;semilla=&semilla;
data final (keep=suma media semilla);set final sumaresi;if
suma=. then delete;run;
%end;/* fin semillas validación cruzada repetida*/
```

```
proc print data=final;run;
%mend;
options mprint=0;
```

10.2.6.2. /*RANDOM FOREST EN LOS DATOS (I)*/

```
%cruzadarandomforestbin(archivo=DATOS.BANCO_G,
vardep= CANAL_OBJETIVO,
listconti= ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
maxtrees=350,variables=3,porcenbag=0.70,maxbranch=2,
tamhoja=8,maxdepth=10,pvalor=0.1,ngrupos=4,sinicio=12345,sf
inal=12356,objetivo=tasafallos);
data final21;set final;modelo='FOREST1';run;

%cruzadarandomforestbin(archivo=DATOS.BANCO_G,
vardep= CANAL_OBJETIVO,
listconti= ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
maxtrees=250,variables=3,porcenbag=0.70,maxbranch=2,
tamhoja=5,maxdepth=10,pvalor=0.1,ngrupos=4,sinicio=12345,sf
inal=12356,objetivo=tasafallos);
data final22;set final;modelo='FOREST2';run;

%cruzadarandomforestbin(archivo=DATOS.BANCO_G,
vardep= CANAL_OBJETIVO,
listconti= ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
maxtrees=450,variables=3,porcenbag=0.70,maxbranch=2,
tamhoja=10,maxdepth=10,pvalor=0.1,ngrupos=4,sinicio=54321,s
final=54340,objetivo=tasafallos);
data final23;set final;modelo='FOREST3';run;

data union; set final21 final22 final23;run;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;
```

10.2.6.3. /*RANDOM FOREST EN LOS DATOS (II)*/

```
%cruzadarandomforestbin(archivo=DATOS.BANCO_G,
vardep= CANAL_OBJETIVO,
listconti= ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
maxtrees=450,variables=3,porcenbag=0.70,maxbranch=2,
tamhoja=8,maxdepth=10,pvalor=0.1,ngrupos=4,sinicio=12345,sf
inal=12356,objetivo=tasafallos);
data final24;set final;modelo='FOREST1';run;

%cruzadarandomforestbin(archivo=DATOS.BANCO_G,
vardep= CANAL_OBJETIVO,
listconti= ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
maxtrees=450,variables=3,porcenbag=0.70,maxbranch=2,
tamhoja=5,maxdepth=10,pvalor=0.1,ngrupos=4,sinicio=12345,sf
inal=12356,objetivo=tasafallos);
data final25;set final;modelo='FOREST2';run;
```

```
%cruzadarandomforestbin(archivo=DATOS.BANCO_G,
vardep= CANAL_OBJETIVO,
listconti= ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
maxtrees=450,variables=3,porcenbag=0.70,maxbranch=2,
tamhoja=10,maxdepth=10,pvalor=0.1,ngrupos=4,sinicio=54321,s
final=54340,objetivo=tasafallos);
data final26;set final;modelo='FOREST3';run;

data union; set final24 final25 final26;run;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;
```

10.2.7. Modelos de Predicción: Bagging

10.2.7.1. /*MACRO DE BAGGING*/

```
%macrocruzadabaggingbin(archivo=,vardepen=,listconti=,listc
ategor=,
ngrupos=4,
sinicio=12345,sfinal=12356,
siniciobag=12345,sfinalbag=12356,
porcenbag=0.80,maxbranch=2,nleaves=6,tamhoja=5,reemplazo=1,
objetivo=tasafallos);
data final;run;
/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;
data dos ;
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;
data fantasma;run;
%do exclu=1 %to &ngrupos;
data tres;set dos;if grupo ne &exclu then vardep=&vardepen;
data tresbis trespred;set tres;if grupo ne &exclu then
output tresbis;else output trespred;run;
/* bagging: */
%do sem=&siniciobag %to &sfinalbag;
data;
numero=&sem-&siniciobag+1;call
symput('numero',left(numero));
total=&sfinalbag-&siniciobag+1;call
symput('total',left(total));run;
%if &reemplazo=0 %then %do;
proc surveyselect data=tresbis out=muestra2 outall
method=srs seed=&sem samprate=&porcenbag noprint;run;
%end;
```

```
%else %do;
proc surveyselect data=tresbis out=muestra2 outall
method=urs seed=&sem samprate=&porcenbag noprint;run;
%end;
data entreno1 ;set muestra2;if selected=1 then output
entreno1;drop selected;run;
proc arbor data=entreno1 ;
input &listconti/level=interval;
%if (&listcategor ne) %then %do;
input &listcategor/level=nominal;
%end;
target vardep /level=nominal;
interact largest;
train maxbranch=&maxbranch leafsize=&tamhoja;
subtree nleaves=&nleaves;
score data=trespred out=sal;
run;
data sal;set sal;vardepen&numero=p_vardep1;run;
/*
%if &numero=1 %then %do;data uni;set sal;keep vardepen1-
vardepen&numero &vardep;run;%end;
%else %do;data uni; merge uni sal;keep vardepen1-
vardepen&numero &vardep;run;%end;
%end;
data uni;merge uni muestra1;ypredi=mean(of vardepen1-
vardepen&total);run;
*/
%if &numero=1 %then %do;data uni;set
sal;ypredi=vardepen&numero;keep ypredi &vardepen;run;%end;
%else %do;data uni; merge uni
sal;ypredi=vardepen&numero+ypredi;keep ypredi
&vardepen;run;%end;
%end;/* fin bagging */
data sal2 ;set uni ;ypredi=ypredi/&total;
if ypredi>0.5 then pre11=1;else pre11=0;run;
proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;
data estadisticos (drop=count percent pre11 &vardepen);
retain vp vn fp fn suma 0;
set sal3 nobs=nume;
suma=suma+count;
if pre11=0 and &vardepen=0 then vn=count;
if pre11=0 and &vardepen=1 then fn=count;
if pre11=1 and &vardepen=0 then fp=count;
if pre11=1 and &vardepen=1 then vp=count;
if _n_=nume then do;
porcenVN=vn/suma;
porcenFN=FN/suma;
porcenVP=VP/suma;
porcenFP=FP/suma;
sensi=vp/(vp+fn);
especif=vn/(vn+fp);
```

```
tasafallos=1-(vp+vn)/suma;
tasaciertos=1-tasafallos;
precision=vp/(vp+fp);
F_M=2*Sensi*Precision/(Sensi+Precision);
output;
end;
run;
data fantasma;set fantasma estadisticos;run;
%end;/* fin grupos */
proc means data=fantasma sum noprint;var &objetivo;
output out=sumaresi sum=suma mean=media;
run;
data sumaresi;set sumaresi;semilla=&semilla;
data final (keep=suma media semilla);set final sumaresi;if
suma=. then delete;run;
%end;/* fin semillas validación cruzada repetida*/
proc print data=final;run;
%mend;
```

10.2.7.2. /*BAGGING EN LOS DATOS*/

```
%cruzadabaggingbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
listconti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
listcategor=SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=12345,
sfinal=12356,
siniciobag=12345,
sfinalbag=12356,
porcenbag=0.7,maxbranch=2,
nleaves=45,tamhoja=15,
reemplazo=1,objetivo=tasafallos);
data final27;set final;modelo='BAGGING1';RUN;

%cruzadabaggingbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
listconti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
listcategor=SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=12345,
sfinal=12356,
siniciobag=12345,
sfinalbag=12356,
porcenbag=0.7,maxbranch=2,
nleaves=40,tamhoja=20,
reemplazo=1,objetivo=tasafallos);
data final28;set final;modelo='BAGGING2';RUN;
```

```
%cruzadabaggingbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
listconti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
listcategor=SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=12345,
sfinal=12356,
siniciobag=12345,
sfinalbag=12356,
porcenbag=0.7,maxbranch=2,
nleaves=20,tamhoja=15,
reemplazo=1,objetivo=tasafallos);
data final29;set final;modelo='BAGGING3';RUN;

%cruzadabaggingbin(archivo=DATOS.BANCO_G,
vardepen=CANAL_OBJETIVO,
listconti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
listcategor=SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
ngrupos=4,
sinicio=12345,
sfinal=12356,
siniciobag=12345,
sfinalbag=12356,
porcenbag=0.7,maxbranch=2,
nleaves=30,tamhoja=10,
reemplazo=1,objetivo=tasafallos);
data final30;set final;modelo='BAGGING4';RUN;

data union; set final27 final28 final29 final30 ;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;
```

10.2.8. Modelos de Predicción: Comparación de Modelos

10.2.8.1. /*COMPARACIÓN DE MODELOS*/

```
data union; set final4 final16 final120 final26 final27;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;
```

10.2.9. Modelos de Predicción: Ensamblado de Modelos

10.2.9.1. /*MACRO ENSAMBLADO DE MODELOS*/

```
 %macro cruzadastack
(archivo=,vardepen=,listcategor=,listconti=,ngrupos=,semini
cio=,semifinal=,nodos=7);
```

```
data final;run;
*proc printto print='
C:\Users\Montserrat\Desktop\TFM1\ca.txt' log='
C:\Users\Montserrat\Desktop\TFM1\loga.txt';run;
%do semilla=&seminicio %to &semifinal;/*<<<<<******AQUI SE
PUEDEN CAMBIAR LAS SEMILLAS */
data dos;set &archivo;u=ranuni(&semilla);
proc sort data=dos;by u;run;

data dos (drop=nume);
retain grupo 1;
set dos nobs=nume;
if _n_>grupo*nume/&ngrupos then grupo=grupo+1;
run;

data fantasma;run;
data unionsalfin;run;
data unifin;run;

%do exclu=1 %to &ngrupos;

data tres;set dos;if grupo ne &exclu then
vardep=&vardepen*1;run;

/* LOGISTICA */
proc logistic data=tres noprint;/*<<<<<******SE PUEDE
QUITAR EL NOPRINT */
class &listcategor;
model vardep=&listconti &listcategor;
score out=saco;
;run;

data sal1 (drop=p_1);set saco;predi1=p_1;run;

/*RED NEURONAL */
PROC DMDB DATA=tres dmdbcat=catatres;
target vardep ;
var &listconti;
class vardep &listcategor;
;run;
proc neural data=tres dmdbcat=catatres ;
input &listconti/ id=i;
input &listcategor /level=nominal;
target vardep/ id=o level=nominal;
hidden &nodos/ id=h act=tanh;
netoptions randist=normal ranscale=0.15 random=15459;
prelim 15 preiter=10 pretech=quanew mom=0.2 learn=0.1;
train maxiter=100 technique=quanew mom=0.2 learn=0.7;
score data=tres out=salred;
run;
```

```
data sal2 (keep=&vardepen predi2 grupo vardep);set
salred;predi2=p_vardep1;run;

/*RANDOM FOREST*/
proc hpforest data=tres
maxtrees=450
vars_to_try=3
trainfraction=0.7
leafsize=10
maxdepth=10
exhaustive=5000
missing=useinsearch ;
target vardep/level=nominal;
input &listconti/level=interval;
input &listcategor/level=nominal;
score out=salo;
run;

data sal3 (keep=&vardepen predi3 grupo vardep);set
salo;predi3=p_vardep1;run;


/*GRADIENT BOOSTING */
        proc treeboost data=tres
        exhaustive=1000 intervaldecimals=max
        leafsize=13 iterations=50 maxbranch=2
        maxdepth=4 mincatsize=15 missing=useinsearch
shrinkage=0.03
        splitsize=15;
        input &listcategor/level=nominal;
        input &listconti/level=interval;
        target vardep /level=binary;
        subseries largest;
        score out=salboost;
run;
data sal4 (keep=&vardepen predi4 grupo vardep);set
salboost;predi4=p_vardep1;run;


/* PRUEBAS CON STACKING */

data unionsal (drop=ygorro);merge sal1 sal2 sal3 sal4;
predi5=(predi1+predi2)/2; /* RED -LOG */
predi6=(predi1+predi3)/2;/* RED -RFOR */
predi7=(predi1+predi4)/2;/* RED -BOOST*/
predi8=(predi2+predi3)/2;/* LOG-RFOR */
predi9=(predi2+predi4)/2;/* LOG-BOOST */
predi10=(predi3+predi4)/2;/* RFOR-BOOST */
predi11=(predi1+predi2+predi3)/3;/* RED -LOG-RFOR */
predi12=(predi1+predi2+predi4)/3;/* RED -LOG-BOOST*/
predi13=(predi1+predi3+predi4)/3;/* RED -RFOR-BOOST*/
```

```
predi14=(predi2+predi3+predi4)/3;/* LOG-RFOR-BOOST*/
predi15=(predi1+predi2+predi3+predi4)/4;/* RED-LOG-RFOR-
BOOST*/
predi17=(predi1*0.2+predi2*0.1+predi3*0.5+predi4*0.2);/*
RED-LOG-RFOR-BOOST ponderado*/
run;

proc logistic data=unionsal;
class &listcategor;
model vardep=predi1 predi3 predi4 &listconti
&listcategor/stepwise;
score out=saco;
run;

data salfin (keep=&vardepen vardep predi1-predi17
grupo);set saco;predi16=p_1;if grupo=&exclu then
output;run;

data unionsalfin;set unionsalfin salfin;run;

data salbis;
array predi{17};
array pre{17};
set salfin;
do i=1 to 17;
if predi{i}>0.5 then pre{i}=1;
if predi{i}<=0.5 then pre{i}=0;
end;
run;
data salbos;run;
%do j=1 %to 17;
proc freq data=salbis noprint;tables pre&j*&vardepen
/out=salconfu;run;
data confu&j (keep=tasa&j);retain buenos 0 malos 0;set
salconfu nobs=nume;
if &vardepen=pre&j then buenos=buenos+count;
if &vardepen ne pre&j then malos=malos+count;
if _n_=nume then do;tasa&j=malos/(malos+buenos);output;end;
run;
data salbos;merge salbos confu&j;run;
;
%end;

data fantasma;set fantasma salbos;run;

%end;
/* FIN GRUPOS */
proc means data=fantasma noprint;var tasa1-tasa17;
output out=mediaresi mean=ase1-ase17 ;
run;
data mediaresi;set mediaresi;semilla=&semilla;run;
```

```
data final (keep=ase1-ase17 semilla);set final mediaresi;if
ASE1=. then delete;run;

data unifin;set unifin unionsalfin;run;
%end;
proc printto; run;
proc print data=final;run;
%mend;
```

10.2.9.2. /*ENSAMBLADO DE MODELOS EN LOS DATOS*/

```
libname discoc
'C:\Users\Montserrat\Desktop\tfm1\TXT\Prediccion';
data uno;set
DATOS.BANCO_G;CANAL_OBJETIVO2=CANAL_OBJETIVO*1;drop
CANAL_OBJETIVO;run;
data uno;Set uno;CANAL_OBJETIVO=CANAL_OBJETIVO2;drop
CANAL_OBJETIVO2;run;


%cruzadastack(archivo=uno,
vardepen=CANAL_OBJETIVO,
listcategor= SEXO HIJOS MES_APERTURA_G ESTADO_CIVIL_G
NIVEL_ESTUDIOS TIPO_PRODUCTO_G,
listconti=ANO_APERTURA DIA_APERTURA EDAD NUMERO_DE_HIJOS,
ngrupos=4,seminicio=12345,semifinal=12356);

proc corr data=salfin;var predi1-predi4;run;

proc corr data=final;var ase1-ase4;run;

data cajas;
array ase{17};
set final;
do i=1 to 17;
modelo=i;
error=ase{i};
output;
end;
run;

proc sort data=cajas;by modelo;
data eti;length eti $ 13;
input modelo eti $;
cards;
1 RED
2 LOG
3 RFOR
4 BOOST
5 RLOG
```

```
6 REDFOR
7 REDBOO
8 LRFOR
9 LBOOST
10 RFORBOO
15 R-L-RF-BOO
;
data cajas2;merge cajas eti;by modelo;
title1
h=2 box=1 j=c c=red 'BANCO' j=c ;

options font="Courier New" bold 8;
run;goptions htext=5pt;

ods graphics off;

proc boxplot data=cajas2;plot error*ETI /
cboxes        = dagr
cboxfill      = ywh;
/* vaxis=0.20 to 0.35 by 0.01 */
;run;

proc boxplot data=cajas2;plot error*ETI /
cboxes        = dagr
cboxfill      = ywh;
/* vaxis=0.20 to 0.35 by 0.01 */
where (modelo ne 16);
;run;

/*data discoc.final1;set final;run;
data discoc.stack1;set cajas2;run;*/

data unifin;set unifin;if bad=. then delete;
RED=predi1;
LOG=predi2;
RFOR=predi3;
BOOST=predi4;
ENSAMBLADO=predi15;
run;

symbol v=dot;
axis1 order=0 to 1;
proc gplot data=unifin;
plot RED*LOG=bad RED*RFOR=bad RED*BOOST=bad LOG*RFOR=bad
LOG*BOOST=bad RFOR*BOOST=bad
RED*ENSAMBLADO=bad
LOG*ENSAMBLADO=bad
RFOR*ENSAMBLADO=bad
BOOST*ENSAMBLADO=bad
/
```

```
vaxis=axis1 haxis=axis1 href=0.5 vref=0.5;
run;
```