FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2019/2020

Trabajo de Fin de Grado

TITULO:

MODELOS PREDICTIVOS PARA DETECTAR EL FRAUDE EN EL CONSUMO DE AGUA

***Alumno*: José Miguel Ayora López**

***Tutor*: Rosa Espínola Vílchez**

Javier Castro Cantalejo

Julio de 2020

UNIVERSIDAD COMPLUTENSE
MADRID

Contenido

RESUMEN

El presente estudio abordará el análisis de posibles fraudes cometidos en el seno de una compañía de servicios de agua en España. Para ello, se hará uso de diferentes datos obtenidos de los clientes de una comunidad autónoma de nuestro país, a través de los cuales se tratará de obtener el mejor modelo de predicción de fraude con la finalidad de evitar que estos hechos se sigan cometiendo.

La metodología para llevar a cabo este trabajo se basará en diferentes técnicas estadísticas, comenzando por la depuración de nuestra base de datos y avanzando hasta la Regresión Logística, Redes Neuronales o Bosques Aleatorios. Estas tareas ayudarán a obtener un modelo robusto y fiable para la detección de irregularidades en base a las características y comportamientos de estos clientes.

El objetivo final será el de detectar todos los usuarios fraudulentos de la empresa mediante el modelo obtenido para que la misma pueda visitar a todos los clientes que indique el mismo.

Palabras clave: agua, fraude, irregularidad, consumo, anomalía, predicción

ABSTRACT

This study will address the analysis of possible fraud committed within a water services company in Spain. To do this, different customer data will be use from an autonomous community in our country, through which the study will try to obtain the best fraud prediction model with the resolution of preventing these acts from continuing to be committed.

The methodology will be based on different statistical techniques, starting with the debugging of our database and progressing to logistic regression, neural networks or random forests. These tasks help us to obtain a robust and reliable model for detecting irregularities based on the characteristics and behaviour of these clients.

The final objective will be to detect all the fraudulent users of the company through our model so the company could visit all the clients indicated by the model.

Keywords: water, fraud, irregularity, consumption, anomaly, prediction

1. INTRODUCCION

Uno de los bienes fundamentales para nuestra existencia es el agua, hecho que ha motivado numerosas polémicas en los últimos años, bien por las sequias acaecidas en nuestro país, por cultivos de regadíos o por quejas de ciudadanos por trasvases realizados de unos ríos a otros.

En España, el comienzo de la gestión hidrológica data de principios del siglo XX y está considerada como la herramienta más eficaz para la distribución eficiente y solidaria del recurso disponible. La planificación hidrológica combina la gestión de oferta y demanda de este bien a la par que promueve un uso eficiente, sostenible y que pueda satisfacer las demandas de la sociedad.

Dicha política hidrológica nos ha permitido pasar de 900.000 Ha de regadío a 3.400.000 Ha, de 200 MW de potencia hidroeléctrica instalada a 17.000 MW, de 296 Km de canalización a decenas de miles de kilómetros de canales, de 57 grandes presas a más de 1.200, de unos consumos urbanos de 10 l/hab y día a otros de 300 l/hab y día. (MINISTERIO PARA LA TRANCISIÓN ECOLÓGICA Y EL RETO DEMOGRAFICO, 2020)

Sin embargo, tal y como indica el gobierno español: *"España es, y sigue siendo, paradigma de la lucha del hombre para poder utilizar responsablemente sus recursos de agua. Su acusada irregularidad en el espacio y en el tiempo ha hecho necesario desarrollar una potente y continuada actuación para poner el agua al servicio del hombre y del desarrollo sostenible."*

He aquí el porqué de la inquietud que lleva a plantear este estudio, ya que son numerosas las noticias a las que se puede acceder a diario relacionadas con fraudes, evasiones de impuestos o casos de corrupción. Por ello, se plantean una serie de cuestiones tales como:

- ¿Siguen los ciudadanos españoles cometiendo irregularidades en el uso del agua?
- ¿De verdad pagan por todo lo que consumen?
- Continuas son las noticias que aparecen en televisión acerca de fraudes fiscales en continuo aumento. ¿Son extensibles dichos datos a los del consumo de agua?
- ¿Existe fraude dentro de la lucha por los regadíos?

Para conseguir dar respuesta a dichas cuestiones se ha planteado el siguiente trabajo de investigación en el que se partirá de un conjunto de datos que proporciona información acerca de los clientes que engloban una compañía de suministros de agua. Dichos datos serán adquiridos en bruto por lo que la primera parte del trabajo estará centrada en realizar la depuración de los mismos y prepararlos para llevar a cabo la tarea planteada.

Dentro de la depuración de datos se llevarán a cabo tareas tales como la corrección de posibles errores en la escritura de las observaciones, tratamiento de valores perdidos y datos atípicos o agrupación variables poco representadas y relacionadas entre sí para poder ser estudiadas. El objetivo de la depuración no es otro que la detección y posterior corrección de los datos incorrectos o corruptos de forman parte de la base de datos adquirida.

Las variables que conforman el conjunto de datos informarán acerca de las cantidades de consumo de agua realizadas por los usuarios, incidencias que hayan presentado con anterioridad, anomalías detectadas o problemas encontrados al realizar la lectura de sus consumos.

A continuación, se estudiarán los comportamientos de las distintas variables de que se dispone, analizando los posibles valores que pueden tomar y sus estadísticos básicos y, posteriormente, se analizará la relación entre las mismas con el objetivo de crear nuevas variables que aporten mayor información y simplifiquen el volumen de la base de datos. Asimismo, se seleccionarán las posibles variables respuesta en base a los datos de que se disponen relativos a irregularidades observadas previamente.

Tras realizar dichas tareas se estará en disposición de aplicar técnicas predictivas a los datos con la finalidad de obtener un modelo que permita clasificar que usuarios están cometiendo alguna irregularidad y cuáles no.

Entre las metodologías más importantes que serán usadas destacan la Regresión Logística, las Redes Neuronales Artificiales y los Bosques Aleatorios.

Al final de dicho estudio, se dispondrá de un modelo que permitirá indicar a cualquier empresa que clientes presentan una mayor probabilidad de cometer fraude con la finalidad de que pueda visitar a los mismos y evitar que continúen haciendo uso de estas técnicas irregulares.

2. OBJETIVOS

El principal objetivo de este estudio se centra en **la generación de un modelo estadístico** que permita la localización de cualquier cliente de empresas de agua que este cometiendo alguna irregularidad con la finalidad de subsanarla.

No obstante, este objetivo plantea una serie de objetivos secundarios que también serán de ayuda para cualquier empresa dedicada a la gestión de dicho servicio. Se tratarán como objetivos secundarios los siguientes:

- **Procesado y almacenamiento de datos.** En ocasiones, una tarea tan simple como pueda ser la lectura de datos se puede convertir en algo bastante engorroso para cualquier empresa. Para ello, se tratará de generar un código que con su simple ejecución realice la lectura de los datos con los que llevar a cabo dicha tarea.
- **Depuración de los datos.** En cualquier ámbito de la estadística en que se trabaje, será poco habitual que los datos obtenidos sean correctos. Será necesario establecer unas pautas que permitan detectar aquellas variables para las que no se disponga de información en algunos clientes, así como de la presencia de datos atípicos, bien sea por errores de transcripción al procesar los datos, o bien, por tratarse de datos reales pero que difieren de forma considerable con los demás.
- **Análisis exploratorio de los datos.** Al realizar cualquier estudio, es muy importante el conocimiento sobre los datos. Se sabe cuál es la situación a tratar, pero es necesario conocer la naturaleza del conjunto de datos para poder relacionarlos con la misma.
- **Generar la variable respuesta, así como nuevas variables explicativas.** En un principio solo se dispone de un gran conjunto de datos, pero de que sirve ello si no se sabe a qué concepto atenerse para detectar que clientes están cometiendo fraude y cuáles no. Para ello, es necesario generar una variable respuesta, la cual sea capaz de indicar aquellos clientes sospechosos de estar cometiendo fraude. Asimismo, será necesario realizar un análisis exhaustivo de todas las variables presente en el conjunto de datos, ya que la transformación de muchas de ellas será de gran ayuda.

El cumplimiento de dichos objetivos al finalizar este estudio desembocará en un ahorro de tiempo para la evaluación y detección de un potencial cliente de riesgo, así como una mayor celeridad a la hora de subsanar los posibles problemas que pudieran surgir derivado de ello.

3. DESCRIPCIÓN DEL CONJUNTO DE DATOS

En este capítulo se abordará la descripción del conjunto de datos. Se comenzará indicando que los datos de los que se va a hacer uso serán camuflados por motivos de confidencialidad. Es por ello que a lo largo de dicho estudio es posible que se presenten problemas de interpretación.

Dado que los datos han venido en diferentes ficheros y que estos tienen diferente número de observaciones, se hará la descripción de cada uno de ellos de forma individual para así poder abordar sus variables de un modo más personalizado. Los pasos a seguir serán los siguientes:

1. Para cada uno de los ficheros seleccionados se va a realizar una selección de las variables que pudieran aportar información al modelo a calcular. Es necesaria esta selección puesto que pueden aparecer variables presentes en distintos ficheros y que, por tanto, con tenerla en alguno de ellos será más que suficiente, teniendo siempre presente la finalidad principal de no perder información por realizar el descarte de alguna de ellas.
2. Análisis de valores perdidos, que son las posiciones en las que las variables sometidas a estudio no disponen de ningún valor en la base de datos.
3. Estudio de datos atípicos. Tras realizar un análisis exploratorio previo se han observado valores que podrían desvirtuar el modelo y que no deben ser tenidos en cuenta a la hora de la creación del mismo.

Es importante reseñar que a medida que se avance en el trabajo será necesaria la unión de todos los archivos en un único conjunto para poder operar con todas las variables seleccionadas. Por ello, debemos indicar que la variable *Nº Contrato* se encuentra presente en todos los ficheros. Esta variable es única para cada usuario, ya que corresponde al código alfanumérico que identifica cada casa o negocio.

- **Clientes**: Se trata de un archivo Excel que consta de 270381 observaciones y 37 variables. Indicará los datos básicos de cada uno de los clientes de que se compone nuestra base de datos. Se seleccionan las variables de interés, las cuales serán Tarifa (la tarifa contratada por cada cliente), Descripción Ubicación Contador (que nos dirá el lugar en que se halla el contador), Sociedad (la empresa que gestiona el servicio), Código Postal – Municipio – Provincia (datos en que se encuentra posicionado el cliente), Paridad (0 si la facturación es en mes impar y 1 si corresponde a mes par), Tipo Contador (modelo del contador), Marca Contador (marca del contador), Año Fabricación Contador (año en que se fabricó el contador), Descripción Propiedad Contador (indica si el propietario del contador es la empresa o el cliente), Coordinate X – Coordinate Y (las coordenadas de localización del suministro), Diámetro (diámetro del contador) y Situación Suministro (estado actual en que se encuentra el servicio).
- **Facturación:** Al igual que en el fichero de clientes, se trabajará con un archivo Excel que cuenta con 270381 observaciones. Dicho fichero englobará el consumo que cada cliente ha realizado cada mes. En este caso, a cada mes para el que se realizan observaciones se le asigna una variable con el consumo de dicho mes y que está formada por la concatenación del año en cuatro cifras y el mes en dos al que corresponden las mismas. Por tanto, todas las variables serán válidas y no será necesario realizar ninguna selección.
- **Estado:** En este archivo se encontrarán 3 variables y 270381 observaciones. Dicho fichero será de ayuda, puesto que indica la situación en que se encuentra el suministro

actualmente, el último mes que cada cliente tuvo una suspensión de suministro y motivo por el que dicho servicio fue suspendido. Además de ello, incluye una pestaña llamada histórico en la que se podrá observar mes a mes el estado en que se encuentra el suministro para cada cliente. Las variables de esta pestaña corresponden a la concatenación de años y meses para cada observación.

- **Pasos Contador:** El conjunto de esta base de datos ha sido obtenido en una base de datos en formato Access. Cuenta con 6438020 observaciones, las cuales fueron comprobadas con las del fichero facturación, puesto que ambas deben coincidir, es decir, en un determinado mes de un determinado año, la diferencia entre lecturas correlativas debe coincidir con el consumo. Se hará uso de las variables Fecha de Lectura (la fecha en que se toma nota de la evolución del consumo) y Tipo de Lectura (nos ayudará a saber si la lectura fue facilitada por el cliente o realizada por un empleado de la empresa). Como se puede observar, se descarta el valor de la lectura, ya que el mismo será obtenido del archivo de facturación.
- **Proyectos:** Aquí se describen las tareas llevadas a cabo sobre los usuarios. Corresponde a un archivo en formato Excel con 10 variables y 68513 observaciones. Aportará las variables Código Proyecto (código numérico de la variable Proyecto) y esta misma. Además de ellas, se necesitarán las variables Fecha Real Fin Visita (la fecha en que se cerró la operación), Estado Proyecto (indicará el estado en que ha quedado el proyecto), Incidencia Visita (aportará información acerca de posibles adversidades surgidas al llevar a cabo la operación) y Acciones (que dará una descripción de las tareas realizadas en la operación).
- **Campañas Compañía:** Dicho archivo guarda cierta similitud con el de proyectos. Se debe diferenciar que en el anterior correspondían a tareas específicas realizadas por la empresa, mientras que en este se encuentras campañas organizadas puntualmente. Cuenta con 24 variables y 20544 observaciones. Dentro de las variables seleccionadas aparecerán IdCampaña y Campaña (código alfanumérico de cada campaña con su respectiva descripción), Tipo Campaña (hace relación a la campaña a que se encuentran adscritas las dos variables anteriores), Fecha (fecha en que se realizó la campaña), Unidad Organizativa (la unidad que gestiona la campaña), Departamento (el departamento dentro de la unidad organizativa), Dirección (el grupo que dirige campaña), Jefatura, Expedientes (tomará el valor SI cuando un cliente comete fraude y NO cuando no), Contrata (la empresa que lleva a cabo la campaña), Estado (indicará si la campaña se encuentra anulada, finalizada o pendiente) y Resuelta (que dirá si la campaña se encuentra resuelta, no resuelta o pendiente).
- **Irregularidades:** El último de los archivos que conforman la base de datos también fue obtenido en formato Excel. Presenta 12515 observaciones y 35 variables. Corresponde a un fichero que será muy importante para el desarrollo de esta labor, puesto que en él se encuentran todos los clientes que han cometido algún tipo de irregularidad. Por tanto, en el mismo se encontrarán variables como: Estado (indicará el estado en que se encuentra un determinado expediente), Fecha de Alta (que hará referencia a la fecha en que comenzó dicho expediente) y Observaciones Plataforma (facilita información acerca de las irregularidades detectadas en dicho expediente).

3.1. DIVISION DE ARCHIVOS

Partiendo de cada uno de los archivos anteriores se generarán 61 ficheros alternativos, los cuales recibirán el nombre del archivo matriz seguido del mes a que corresponden. Se especifica que el mes no corresponde al calendario, sino que, dado que se dispone de datos desde enero del año 2015 hasta enero de 2020 inclusive, se comenzará por el 1 y se terminará en el mes 61 que corresponderá a este último.

Se omiten en este apartado los ficheros de **Clientes** y **Estado – histórico**, para los cuales no es necesario realizar esta partición.

El hecho de realizar esta tarea viene motivado porque una vez comience el estudio, una parte de los datos será destinado a conjunto de entrenamiento y la parte restante, a conjunto de validación.

Para llevar a cabo esta tarea se hace uso de las fechas que eran descritas en el análisis exploratorio y que están presentes en cada uno de los archivos. Así pues, para la división del archivo de facturación, cada una de las variables de que constan estos (61 en total) se formarán los 61 archivos, mientras que para el resto de archivos se partirá de la variable fecha que forma parte de cada uno de ellos. Se utilizará el mes de dicha variable para establecer el fichero a que corresponde y así poder procesarlo.

4. DEPURACION DE DATOS

"La depuración es una actividad inevitable y difícil dentro de la producción de datos estadísticos. Si se define, planifica y ejecuta de forma deficiente, puede tener un impacto muy negativo en la calidad de los datos, en los costes de producción y en los plazos de difusión de los datos." (Villán Criado & García Rubio, 1995)

Esta definición tan simple muestra como una tarea tan sencilla como puede ser corregir pequeños errores que existan en cualquier conjunto de datos pueda mandar a la basura un arduo trabajo que en ocasiones puede haber abarcado meses de desarrollo.

Esto lleva a considerar este capítulo como uno de los más importantes para poder obtener un buen modelo, dada la relevancia de las tareas a desarrollas. El capítulo será dividido en 4 pilares fundamentales para el desarrollo del trabajo:

- Agrupación en categorías variables que se encuentren poco representadas.
- Corrección de los posibles errores de escritura que pueda presentar la base de datos, tales como observaciones en mayúsculas o minúsculas o palabras mal escritas que puedan duplicar una observación
- Establecimiento de pautas que determinen qué valores perdidos deberán ser eliminados del conjunto de datos.
- Realización de un análisis exhaustivo de los valores que toman las distintas observaciones para detectar la posible presencia de datos atípicos.

4.1. AGRUPACION DE VARIABLES

El hecho de contar con archivos con numerosas observaciones lleva a encontrar valores que representan categorías idénticas o categorías con poca representación que requieren de la unión con categorías similares para poder ser estudiadas.

Para evitar esta problemática se recorrerán todos los archivos para ver si en el conjunto de sus variables existiese alguna cuyas observaciones tienen significados similares y así poder realizar la agrupación de las mismas.

Se comenzará por el fichero de **Clientes**, en el cual se realizan dos agrupaciones. En primer lugar, la variable Situación Suministro, que inicialmente puede tomar 16 valores diferentes, queda reducida a solo 4. Para ello se crea la variable *gruposit* que tomará el valor 1 para todos los cierres totales por falta de pago, 2 para el resto de cierres totales, 3 para los cierres parciales y 4 para el resto de valores. A continuación, se encontrará la variable Tipo Contador, para la que se dispone de 12 valores distintos. Se crea una nueva variable llamada *grupocont*, que constará de 4 valores, el valor uno para los grupos que comiencen por G-1, 2 para aquellos que comiencen por G-2, 3 para los que tienen un valor que comienza por G-4, quedando el valor 4 para los que comienzan por G-6.

Seguido de ello, se procede a manipular el archivo **Proyectos**. La primera variable a agrupar será Proyecto. Recibirá el nombre de *grupodesc*, tomará los valores de 1 a 8 y englobará los casos de cambios de contador, ceses, cesados, excesos, cierres por comercializadora, inspecciones de

Instalación Receptora Individual (IRI), reaperturas de IRI, reaperturas por pago y revisiones, respectivamente. Otra variable que se puede encontrar en dicho fichero es Incidencia Visita. Para ella se creará la variable *grupoincd*, que tomará el valor 1 para clientes no localizados, 2 para visitas que han sido imposibles de realizar, 3 para las visitas innecesarias, 4 en aquellas que no se encontró incidencia, 5 para trabajos que no se tenían que realizar y 6 para valores perdidos.

También en este fichero, se toma la decisión de modificar la variable Código Proyecto. Anteriormente estaba formada por códigos de 2 cifras y se ha decidido convertir dichos códigos en una única cifra para facilitar su manejo. Así pues, el código 08 pasara a valer 1, el código 90 pasara a 2, el 04 a 3, el 09 a 4, el 63 a 5 y el 89 a 6. Esta variable recibirá el nombre de *grupotipo*. Para finalizar con la manipulación de este archivo, se realiza la transformación de la variable Acciones, para lo que se genera una nueva variable que recibirá el nombre de *grupoacc*. La descripción de los valores que engloba será 1 para las anomalías, 2 para los cierres, 3 para los ceses, 4 para comprobaciones de marca, 5 para levantamientos, 6 cuando no se requieran acciones y 7 para los que no presenten valor alguno.

Otro fichero en el que se encuentran variables con necesidad de manipular es el de **Campañas Compañía**. La primera variable que requiere de actuación es Departamento. Se genera la variable *grupodpto* que toma el valor 1 para el departamento de control, 2 para nuevos departamentos, 3 en el caso de nuevos departamentos y 4 en aquellos de planificación. También es necesario actuar sobre las variables Estado y Resuelta, ya que solo interesarán aquellas que se encuentren en finalizadas y resueltas. La concatenación de ambas variables será guardada en la variable *unión* y se borrarán las 2 mencionadas. La ultima agrupación a realizar en este archivo es para la variable Tipo Campaña. Se genera la variable *grupotipo*, con valores desde el 1 hasta el 14, en función de que corresponda a determinados tipos de análisis (1 - 2), cierres (3), avisos (4-5), consumo nulo (6), proyectos (7), 6.1 (8 – 11), revisiones (12) y diferentes tomas de lectura (13 – 14).

El siguiente archivo que se encuentra corresponde a **Estado (histórico)**. En el mismo será generada la variable *grupo* que agrupará las observaciones de la variable Estado Suministro. Toma el valor 1 en los cierres totales por falta de pago, 2 en el resto de cierres totales, 3 en los parciales y 4 para los datos perdidos.

Finaliza este apartado trabajando con el fichero de **Irregularidades**, el cual incluye dos variables a agrupar. La primera de ellas corresponde a la actividad de la empresa. Recibe el nombre de *grupoact* y alberga los siguientes supuestos: publica (1), industria (2), hostelería (3), local (4), riegos (5), administración (6) y vivienda (7). La otra variable mencionada es Estado. Quedará agrupada en la variable *grupoest*, tomando el valor 1 si esta cancelado, 2 si esta facturado, 3 si se encuentra facturado en ciclo y 4 si es no facturable.

4.2. ERRORES DE ESCRITURA

A lo largo de la revisión de los datos se han detectado diversos errores de escritura. No se puede olvidar detallar que el software utilizado tratará como diferentes estas posibles observaciones erróneas y que están haciendo mención a una misma observación, por lo que es necesario corregirlas. (Cody, 2017)

El primer error localizado es que en el fichero de **Irregularidades** y en el de **Campañas Compañía**, la clave Nº Contrato presenta letras en minúsculas. Es debido subsanar este error, pues en caso contrario, considerará como clientes distintos aquellos que en un fichero tengan su Nº Contrato en minúscula y en otra en mayúscula.

También en el fichero de **Campañas Compañía**, se encontrarán errores de transcripción para la variable contrata. En determinadas observaciones aparece el valor de la variable con distintos valores.

En el archivo **Pasos Contador**, se detecta que la variable Tipo Anomalía presenta el símbolo de comillas (“) para los valores perdidos, por lo que ha sido necesario ajustarlo para que el software lo considerase como un valor perdido.

El ultimo archivo en el que se encuentran errores de este tipo fue el de **Estado**. Tanto en el fichero inicial, como el correspondiente al histórico, la variable situación suministro presenta errores de escritura que generaban distintas observaciones para una misma categoría.

4.3. VALORES PERDIDOS

Es importante destacar que, al obtener el conjunto de datos, estos no han sido filtrados previamente. Posibles problemas que se podrán encontrar serán valores perdidos en las variables de consumo, clientes con consumos negativos o consumos desorbitados. Además, existen observaciones con errores de transcripción que se deben corregir, puesto que esto podría provocar distintos valores en una variable que en teoría deberían ser los mismos. (Cody, 2017)

La principal problemática que se encuentra es que el consumo de cada cliente se factura de forma bimensual, por lo que se dispondrá de valores en su facturación en meses alternos. Es por ello que es debido eliminar todas aquellas observaciones en las que el consumo no presente información.

También al analizar las coordenadas de ubicación de los clientes se observa que presentan valores perdidos. En este caso, el plan de actuación definido será el de imputar las coordenadas de su capital de provincia, ya que en esos casos no se dispone de información como puede ser el nombre de la localidad para poder localizar dicha información por otras vías.

Por último, tras realizar la agrupación de observaciones (véase apartado 4.1) se detectan usuarios que no poseerán valor alguno para determinadas categorías. Cabe reseñar que este supuesto ha sido contemplado en la agrupación de variables, creando una categoría adicional para las mismas, tal y como se indicaba en la descripción de las mismas.

Estas serán las únicas variables en las que se deba ejecutar alguna acción sobre los valores perdidos que se presentan, puesto que la eliminación de valores perdidos en otros ficheros llevaría a cometer el error de eliminar clientes que tienen ese valor perdido pero que podrían aportar información muy valiosa para generar el modelo. Sirva de ejemplo que habrá clientes que no tengan ninguna irregularidad porque nunca han cometido un fraude.

4.4. DATOS ATIPICOS

Una vez realizada una primera exploración de la base de datos se podrán detectar anomalías que pueden interferir en la creación del modelo. Dentro de ellas se engloban los datos atípicos, valores que difieren de la mayoría de las observaciones de que se disponen en el conjunto de datos. (Cody, 2017)

Se localizan valores negativos de consumos cuando todos sabemos que, si un mes no haces uso de agua, el consumo será 0, pero nunca podrá ser negativo. Estos pueden ser debidos a refacturaciones que se hayan realizado a los clientes. Sin embargo, no serán tenidos en cuenta a la hora de calcular el modelo.

También se detectan valores muy elevados en dicha variable. Para evitar obtener resultados que pudieran desvirtuar este estudio, se considerarán también como valores atípicos en el conjunto de datos todos aquellos consumos superiores a 1000 y todos aquellos que excedieran de la media en 3 veces la diferencia entre los percentiles 90 y 10.

Por último, se depurará la variable *Año Fabricación Contador*, ya que esta presenta valores anómalos. A partir de la misma, se creará la variable *Año Fabricación Contador 2*, la cual asignará por defecto el año 1990 como valor de la observación a todos aquellos contadores que presenten un año de fabricación superior al año al que pertenece la observación de referencia.

5. CREACION DE VARIABLES

Este apartado estará a la creación de nuevas variables que serán de ayuda a la hora de construir el modelo. Dichas variables parten de la transformación de las variables de que se dispone en el conjunto de datos inicial y que, con una simple manipulación de las mismas, hará que se obtenga información muy relevante acerca del comportamiento de los usuarios a lo largo del periodo estudiado.

Se destacan dos grandes bloques en este capítulo:

- El primero de ellos se basará en la creación de las posibles variables respuesta, la cuales informarán de aquellos clientes que estén cometiendo irregularidades y cuáles no.
- Continuará con la creación de nuevas variables, cuantitativas y cualitativas, que ampliarán la información de la que ya se disponía en la base de datos mediante el simple hecho de combinar variables de los distintos ficheros que fueron adecuados en el capítulo anterior.

5.1. VARIABLES RESPUESTA

La variable respuesta tiene la finalidad de indicar el número de meses que un determinado usuario lleva cometiendo algún tipo de irregularidad.

Se generarán 12 posibles variables respuesta, desde *Irr1* hasta *Irr12*. El nombre de esta variable se desglosa como Irr (que proviene de irregularidad) seguido de un numero establecido entre 1 y 12 que indicará con cuántos meses de anterioridad ya se estaba cometiendo fraude.

Cualquiera de estas variables Irr tomará el valor 1 cuando el cliente ha sido visitado y se comprueba que está cometiendo fraude, 0 en el caso de que el cliente no ha sido visitado y no está cometiendo fraude y -1 cuando el cliente ha sido visitado, pero se ha detectado que no está cometiendo fraude.

Una vez se tenga el conjunto de variables creadas, se analizará la relación de estas posibles variables respuesta y las variables explicativas para así poder comprobar cuál será la variable respuesta que mejor se relaciona con el resto de variables y, por tanto, conseguir una mejor predicción.

5.2. VARIABLES EXPLICATIVAS

Otro aspecto importante para llevar a cabo en este trabajo será la creación de nuevas variables en función de las variables que conforman el conjunto de datos. Esta parte estará dividida en tres grandes bloques en función del tipo que posea la variable de inicio. Estos apartados serán el de variables numéricas, variables categóricas y variables obtenidas por estudios anteriores.

En primer lugar, se detalla que no existe una metodología predefinida para la creación de estas nuevas variables, sino que ha sido fruto del trabajo realizado al analizar las variables del conjunto

de datos y que, se ha considerado, pueden ser las más adecuadas para obtener un modelo óptimo.

También se indica que dichas variables tratarán de agrupar la información distribuida a lo largo del tiempo con el fin de evitar que el modelo se limite a buscar el fraude en un determinado momento, sino que detecte determinados patrones que localicen estas actividades irregulares agregando información pasada.

Para conseguir todo ello, los bloques de tiempo en que se centrarán este trabajo serán de los 6, 12 y 24 últimos meses, generando las correspondientes observaciones a dichos periodos en cada variable.

5.2.1. VARIABLES NUMÉRICAS

Como su propio nombre indica y, tal y como ha sido indicado anteriormente, el origen de estas variables esta en las variables numéricas que se disponen en el conjunto de datos. Bastará con manipular un poco los datos de que ya se dispone para poder obtener otro punto de vista distinto a un simple número.

Es importante reseñar que todos los cálculos que se realicen tendrán la única finalidad de ver si los comportamientos de las variables que se están creando son los mismos o, por el contrario, al espaciarlos en el tiempo van cambiando.

Se conoce por lógica que el consumo de agua no será el mismo en meses de invierno que en los de verano. Sirva de ejemplo que, en cuestión de regadíos, no será necesaria la misma cantidad de agua en meses lluviosos que en aquellos de sequía. Es por ello que también se crearán nuevas variables teniendo en cuenta estos otros factores que no vienen detallados en el conjunto de datos. Para ello, consideramos 3 escenarios posibles que serán los siguientes:

- Opción 1: Datos naturales.
- Opción 2a: Datos desestacionalizados.
- Opción 2b: Datos sin desestacionalizar.

Por ello, destacamos que en la descripción que realicemos de las variables veremos que en todas aparecerán las siglas *Opc* y estas irán seguidas del número de opción que indique la tipología de los datos.

- Se comenzará creando la variable *IndiceOpc* que aportará el cociente entre el consumo de cada usuario y el consumo medio para la tarifa contratada. Valores superiores a 1 ofrecerán que el consumo se encuentra por encima de la media de los usuarios que poseen la misma tarifa.
- Se calcula el consumo medio para cada cliente en los últimos meses. Se crearán 3 variables: *media6Opc*, *media12Opc* y *media24Opc*. Se crean otras 3 variables similares, pero en este caso, en lugar de con la media, será con la desviación típica. Estas variables recibirán el nombre de *desviacion6Opc*, *desviacion12Opc* y *desviacion24Opc*.
- Variable *Termino_saltoOcp.* Se calcula una variable auxiliar que posteriormente será eliminada del código. Esta variable tomara el valor 0 cuando el índice calculado en el primer punto sea 0, -1 para el caso en que el índice tome valores negativos y 1 en el caso de ser positivo. La variable *Termino_saltoOcp* se obtendrá ponderando el valor de la

variable auxiliar en función de su cercanía al momento temporal estudiado. La suma de estos valores acumulados hasta el fichero en estudio será el valor de la observación para cada cliente.

- Variables *Max_salto* y *Lugar_Max_Salto*. Para la primera de ellas, se hará uso de la variable auxiliar creada en el apartado anterior. En este caso, en lugar de ponderar la variable y, posteriormente, sumarla, directamente se calcula la suma acumulada. Por tanto, esta variable estará acotada y se moverá en el intervalo (-23,23). Es necesario indicar que solo serán de interés aquellos saltos en los que se reduzca el consumo, ya que el hecho de cometer la irregularidad es para consumir menos, por lo que clientes que aumenten su facturación no estarán mostrando indicios de cometer fraude. La segunda de ellas dará número de meses que se ha de retroceder en los ficheros para localizar el máximo salto que se ha producido y que ha sido calculado por la variable antes comentada. El intervalo en que se mueve estará establecido entre (0,23).

Por último, se cierra el bloque de creación de variables numéricas analizando todas aquellas facturas en las que un cliente no presenta consumo, es decir, aquellas en las que la variable consumo toma el valor 0.

- En primer lugar, es necesario saber las facturas sometidas a estudio por parte de cada usuario, para lo que se definen las variables *NumeroFacturas_6*, *NumeroFacturas_12* y *NumeroFacturas_24*, las cuales proporcionarán la cantidad de facturas que se han emitido a un usuario en los últimos 6, 12 y 24 meses.
- El siguiente paso es la creación de las variables *Porcentaje0_6*, *Porcentaje0_12* y *Porcentaje0_24*, que indicarán el porcentaje de facturas con consumo 0 que tiene cada cliente en los últimos meses, número que será indicado por los dos últimos caracteres con que se nombra a la variable.
- Seguido de ello, se generan las variables *Numero0_6*, *Numero0_12* y *Numero0_24* que indicaran el número de facturas que presenta un usuario en los últimos 6, 12 y 24 meses, respectivamente.
- Al igual que en supuestos anteriores, se generará la variable *Termino_salto_0*, que hará una función idéntica a los términos de salto de los que se hacía uso en dichos casos.
- Puesto que también se analizan los saltos en estos bloques, también se crean las variables *Max_salto_0* y *Lugar_Max_Salto_0*, a través de las cuales se podrá saber el valor que toma el máximo salto, así como el número de meses que se habría de retroceder para localizar la correspondiente factura.

Una vez finalizado este último bloque, es necesario detallar también que las variables *consumo* y *Año Fabricación Contador2*, las cuales ya habían sido definidas en la presentación del conjunto de datos, también se tratan de variables numéricas, por lo que también serán usadas para conformar el modelo.

5.2.2. VARIABLES OBTENIDAS POR ESTUDIOS ANTERIORES

En la información adquirida con el conjunto de datos, también se dispone de una serie de variables que han sido de interés en estudios realizados con anterioridad y que, por tanto, también deben ser codificadas en este estudio.

Se abordará la creación de 11 nuevas variables, las cuales se describen a continuación:

- La primera variable que se crea recibe el nombre de *VarOtros1*. Se trata de una variable dicotómica que tomará el valor 1 en aquellos casos en que el consumo correspondiente al mes analizado sea 0 y el valor 0 en caso contrario.
- A continuación, se analizarán los consumos a 0 espaciados en el tiempo. Para ello se generarán las variables *VarOtros2* y *VarAux2*, donde la primera de ellas corresponde también a una variable dicotómica con valor 1 cuando el consumo de las 3 últimas facturas de un usuario sea 0, siendo 0 cuando alguno de los 3 últimos meses presenta un valor para consumo distinto del descrito.
- Otra variable que analiza la ausencia de consumo en el tiempo es *VarOtros7*. Corresponde a una variable de tipo numérico que indicará el número de meses acumulados en los cuales el cliente no ha tenido consumo real.
- La siguiente variable recibe el nombre de *CaidaConsumo0* y proporcionará la diferencia entre consumos entre los últimos 12 meses y los penúltimos 12. Así se podrán comparar caídas de consumos entre años.
- Se continuará con las variables *VarOtros10Avg*, *VarOtros11Avg* y *VarOtros13Avg*, que proporcionarán la diferencia entre consumos entre los últimos 12 meses y los penúltimos 12, por lo que se podrán comparar variaciones de consumo. Las dos primeras serán de tipo dicotómico. *VarOtros10Avg* tomará valor 1 cuando la media del consumo de los 12 penúltimos meses sea positiva y la media de los 12 últimos sea 0, mientras que *VarOtros11* tomará valores 1 cuando el consumo medio de los 12 últimos meses sea positivo y se tengan valores perdidos, negativos o nulos para los 12 penúltimos. Además, se establece el supuesto de que la media de los últimos 12 meses presente un valor perdido, en cuyo caso, dicha variable también tomara el valor 1. Seguido de ellas, la variable *VarOtros13Avg* indicará el porcentaje de variación del consumo. Si este porcentaje es superior al 30%, convertirá automáticamente el valor de la observación para la variable *VarOtros10Avg* en 1.
- Finaliza este bloque con la creación de las variables *VarOtros10Sum*, *VarOtros11Sum* y *VarOtros13Sum*. Estas variables realizarán una función similar a las creadas en el punto anterior. Su diferencia con ellas reside en que, si bien antes, se realizaba la comparación con el consumo medio para el periodo sometido a estudio, en esta ocasión se realizará la toma de decisiones para asignar el valor a las variables dicotómicas en función de la suma de los consumos de los 12 últimos y penúltimos meses. Por último, la variable *VarOtros13Sum* indicará el porcentaje de variación del consumo. Si este porcentaje es superior al 30%, convertirá automáticamente el valor de la observación para la variable *VarOtros10Sum* en 1.

5.2.3. VARIABLES CATEGÓRICAS

Para la creación de estas variables se tendrán en cuenta tanto variables que existían en la base de datos como las agrupaciones realizadas en el capítulo 4.4.

El objetivo que se pretende cumplir con la creación de estas variables no es otro que transformar variables categóricas convirtiéndolas en numéricas y evitar la posibilidad de que se presenten variables con infinidad de categorías.

Las tareas que se desarrollarán serán las siguientes:

- En primer lugar, se generarán una serie de variables que clasificarán las observaciones tomando como parámetros de referencia los números 30 y 100 para determinar el límite de fraudes. El valor que tomará cada observación será el valor medio de fraudes en función de la variable analizada en el conjunto de entrenamiento cuando el número total de fraudes sea superior al indicado en el parámetro de la variable. En caso de que no lo supere, tomará como valor el número medio de fraudes en el mes de observación. Se crearán bajo este supuesto las variables *ProbLocal1, ProbLocal2, Tarifa, Descripcion_Ubi_Cont, Paridad, GrupoCont y Descripcion_Propiedad_Cont.* Todas ellas irán acompañadas del parámetro que delimita el número de fraudes en cada supuesto.
- El siguiente bloque se calculará para periodos que comprenden los 6, 12 y 24 últimos meses de facturación de cada usuario, por lo que se obtendrán 3 ficheros por cada opción descrita dentro de cada variable. Se generarán las variables *NumeroGrupoAcc, NumeroGrupoIncd, NumeroGrupoTipo, Numerofraude, UltimaVisita,* las cuales realizarán el conteo de observaciones que se disponen en la base de datos para el periodo estudiado. Además, para las variables que habían sido agrupadas en el punto 4.4, también se realizará el conteo por categorías, añadiendo el valor de la categoría al nombre de la variable creada.
- De un modo análogo al anterior, se calcularán las variables *PercentTipoLectura* y *PercentEstSuminis* para 6, 12 y 24 meses en función de la categoría de la observación y que calcularán los porcentajes correspondientes a cada categoría y periodo.
- Por último, se generarán las variables *UltimaVisita* para los mismos periodos, las cuales indicarán el resultado de la última visita, tomando el valor 1 en caso de haberse detectado irregularidad, -1 en caso de que no, asignando el 0 en caso de que el cliente no haya sido visitado. Asociadas a las mismas se crearán las variables *TiempoUltimaVisita*, que detallarán el tiempo en meses que ha transcurrido desde la última visita. A ello se añadirá la variable *NumeroAux* que asignará como observación el número total de fraudes observados en el total del conjunto de datos para los usuarios analizados.

Con esto finaliza la generación y descripción de las variables categóricas, por lo que se estará en posición de seleccionar cual de las posibles variables respuesta es la más adecuada, lo cual será analizado en el próximo apartado.

6. METODOLOGIA

Tras las tareas realizadas, se deben introducir las distintas metodologías que serán utilizadas para conseguir el objetivo planteado. Se utilizarán distintos métodos mediante los cuales se pueden generar infinitos modelos con diferentes técnicas que puedan considerarse óptimas. Las técnicas en las que se centrará este trabajo serán:

- Regresión Logística. Tratará la situación en que la variable respuesta solo ofrece dos resultados posibles, frecuentemente considerados como éxito y fracaso. (Montgomery, 2006)
- Redes Neuronales. Se encuentran inspiradas en las redes neuronales biológicas. Su funcionamiento se basa en el aprendizaje continuo, generalizando de ejemplos previos y extrayendo las características principales de los datos. (Basogain Olabe, 2014)
- Random Forest (Bosques Aleatorios), que combinará árboles predictores de tal manera que cada árbol dependa de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de ellos.

Se probarán una gran cantidad de modelos para cada técnica según la parametrización de los mismos. Cuando se elija el mejor o mejores de cada una de las metodologías se realizará una comparativa entre los mejores obtenidos y se elegirá el óptimo.

Sin embargo, se debe comenzar por seleccionar la variable respuesta óptima, para lo cual se usará la Regresión Logística, la cual medirá la relación existente entre las variables explicativas y las distintas variables respuesta.

6.1. SELECCIÓN VARIABLE RESPUESTA

Una vez que se disponga de las posibles variables respuesta es necesario determinar cuál de ellas se relaciona mejor con el resto de variables y, por tanto, proporcionará una mejor predicción.

La calidad del modelo estará vinculada al grado de relación que exista entre las variables del conjunto de datos con las posibles variables respuesta de que se disponga. Para analizar dicha relación, se hará uso de la Regresión Logística.

Será necesario realizar un análisis de regresión por cada una de las candidatas a variable respuesta que contenga cualquier conjunto de datos, tomando siempre como variables explicativas las variables numéricas.

Criterios importantes para la selección de una variable respuesta adecuada serán:

- Porcentaje de individuos clasificados correctamente.
- Estadísticos de ajuste: D-Sommers, Gamma o Tau-c. A mayor calidad de ajuste, el estadístico se situará más próximo a 1.
- Relación de las variables de la base de datos con la variable respuesta. El análisis de dicha relación permitirá observar si los valores reales son similares a los esperados o si, por el contrario, es necesario ampliar la información de la base de datos.

6.2. TECNICAS PREDICTIVAS

Se realizará una breve introducción a cada una de las técnicas que serán utilizadas para la creación de los modelos. Para cada una de ellas, serán destacados sus elementos fundamentales, así como el modo en que procesa los datos para facilitar los modelos calculados.

6.2.1. REGRESION LOGÍSTICA

Se define el análisis de Regresión Logística como una técnica destinada a analizar la relación existente entre una variable dependiente y un conjunto de variables independientes. Su objetivo corresponde a comprobar hipótesis o relaciones causales cuando la variable dependiente es categórica. Se trata de una técnica similar a la regresión lineal múltiple, con la particularidad de que esta última es usada cuando la variable dependiente es de tipo cuantitativo. (López Roldán & Fachelli, 2015)

Las variables explicativas son más flexibles, pues se puede hacer uso tanto de variables continuas como categóricas. Para que estas últimas puedan ser introducidas en el modelo se generan las variables dummy, que solo tienen dos categorías, las cuales tomaran los valores 0 y 1 en función de la pertenencia a la categoría de referencia o no.

Se considera el supuesto en que la variable respuesta solo dispone de dos posibles valores: 0 y 1, los cuales podrán ser asignaciones a una respuesta cualitativa. (Montgomery, 2006)

La Regresión Logística aúna dos partes fundamentales del análisis estadístico:

- El análisis de tablas de contingencia mediante modelos log-lineales
- Análisis de regresión por mínimos cuadrados ordinarios.

En ambas situaciones, se encontrarán limitaciones que la Regresión Logística resolverá, ya que, para el análisis de tablas de contingencia con modelos log-lineales, los modelos de dependencia no aceptaban el uso de variables continuas, mientras que, para el análisis de regresión por mínimos cuadrados ordinarios, las variables de tipo categórico no siempre funcionarían como buenos predictores. (López Roldán & Fachelli, 2015)

Dentro del análisis de Regresión Logística se destacan dos tipos:

- La Regresión Logística Binaria: explica una característica o suceso dicotómico.
- La Regresión Logística Multinomial: explica una variable cualitativa con varias categorías.

El proceso de análisis se puede establecer en varias etapas:

1) Estimación de parámetros del modelo. El método usado para realizar la estimación de los parámetros será el de máxima verosimilitud. Dicho método indica que será elegido como valor estimado para el parámetro aquel que cuenta con mayor probabilidad de suceder en base a los datos observados. Además, proporcionará valores consistentes, eficientes y que trataran de corregir el sesgo. (Montgomery, 2006)

2) Evaluación de los coeficientes individuales del modelo. Se analiza la significatividad individual de cada uno de los parámetros obtenido utilizando el Test de Wald, que establece las siguientes hipótesis:

   - H_0: $\beta_i = 0$
   - H_1: $\beta_i \neq 0$

   El estadístico del contraste es el siguiente:

$$W = \frac{\beta_i}{SE(\beta_i)}$$

   donde:
   - β_i corresponde al parámetro estimado.
   - $SE(\beta_i)$ hace referencia al error estándar del parámetro.

   El valor del estadístico se distribuye según una distribución χ^2, la cual no podrá rechazar la hipótesis nula para valores pequeños de W, indicando así que el parámetro no sería significativo. En caso contrario, valores altos de W, se rechazará la hipótesis nula y, por tanto, se considerará que dicha variable no sería válida para explicar la variable respuesta. (Agresti, 2007)

3) Clasificación de los casos. Una vez obtenida la ecuación del modelo de Regresión Logística, se procede a realizar la clasificación de las observaciones del conjunto de datos en función del de la variable dependiente estudiada. Así pues, se obtendrán dos clasificaciones distintas, la inicial de que se dispone antes de realizar el análisis y la pronosticada en función del modelo obtenido. El porcentaje de casos clasificados de un modo correcto indicará la capacidad predictiva del modelo. (López Roldán & Fachelli, 2015)

La Regresión Logística también ayuda a realizar comparaciones entre la probabilidad de sucesos, lo que se define como Odds Ratio. Estos determinarán la probabilidad de ocurrencia de un determinado suceso frente a que no ocurra. Su cálculo se puede realizar del siguiente modo:

$$Odds\ Ratio = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

Donde p_1 corresponderá a la probabilidad de que un suceso ocurra o tome el valor 1, mientras que p_0 se definirá como la probabilidad de que un suceso no ocurra o tome el valor 0.

Para linealizar la función de respuesta se suelen utilizar distintas funciones de enlace. Su misión será la de relacionar el valor esperado de la variable respuesta con los predictores lineales que forman parte del modelo. Una vez realizada dicha transformación, la relación podrá ser modelada mediante la regresión lineal y permitirá que los valores obtenidos en el modelo se sitúen entre 0 y 1. Entre las mismas destacan: (Gündüz & Fokoué, 2015)

- **Logit:** $\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$

- **Probit:** $\pi(x) = \Phi(\alpha + \beta x)$

- **CLogLog:** $\pi(x) = 1 - \exp(-\exp(\alpha + \beta x))$

- **Cauchit:** $\pi(x) = \frac{1}{\pi}\left(tan^{-1}(\alpha + \beta x) + \frac{\pi}{2}\right)$

Donde α corresponderá a la constante del modelo obtenido, β serán los coeficientes estimados y x corresponderá al conjunto de variables explicativas.

La representación gráfica de las diferentes funciones para observar las diferencias existentes entre ellas quedaría del siguiente modo:

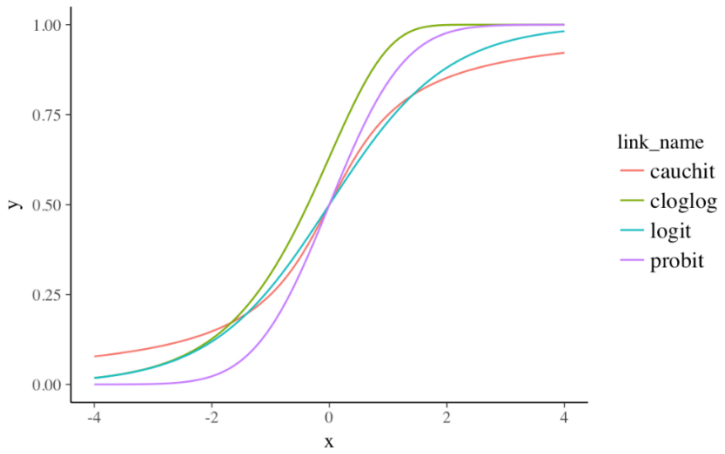

Ilustración 1.- Funciones de enlace

En la ilustración 1 se puede comprobar como las funciones de densidad acumulativas son simétricas para las funciones de enlace logit, probit y Cauchit, a pesar de que estas últimas tienen colas más finas o pesadas que la función logit. Sin embargo, la función cloglog no tiene un comportamiento simétrico alrededor del punto de inflexión a pesar de tener también una forma sigmoidal.

Por último, se establecen los distintos métodos de selección de variables existentes:

- **Método Backward:** Eliminación hacia atrás. Se parte de un modelo en el que se han introducido todas las variables que irán siendo excluidas una tras otra. En cada paso será eliminada la variable menos influyente, finalizando el proceso cuando no exista ninguna variable no significativa. Las variables restantes serán aquellas que conformen el modelo.
- **Método Forward:** Selección hacia delante. Las variables son introducidas en el modelo de forma secuencial, siempre y cuando cumplan el criterio de entrada. La primera variable en ser introducida es la que presenta una mayor correlación con la variable dependiente, mientras que para los sucesivos pasos será aquella que presente una

mayor correlación parcial y aún no se encuentre en el modelo. El procedimiento finaliza cuando ya no se dispone de variables significativas.

- **Método Stepwise:** El método paso a paso combina los dos métodos anteriores. En cada paso, se introducirá la variable explicativa que presente una mayor significatividad. A continuación, se evalúa si alguna variable dentro del modelo no es significativa, en cuyo caso, esta sería eliminada. Se repiten los pasos hasta que no hay más variables candidatas a ser incluidas o eliminadas en el modelo.

6.2.2. REDES NEURONALES ARTIFICIALES

"Es indispensable que las empresas comprendan las potencialidades de las Redes Neuronales Artificiales y puedan aplicarlas para aumentar la competitividad en un ambiente global." (Acuña, 2020)

Los modelos neuronales están caracterizados por su gran capacidad para aprender, generalizar y retener conocimiento de los datos, motivo por el que pueden ser considerados como modelos de regresión o modelos discriminantes no lineales. (Sarle, 1994)

Se definen las Redes Neuronales Artificiales, en adelante RNA, como sistemas de procesamiento de la información que poseen una estructura inspirada en las redes neuronales biológicas. Componen un conjunto de elementos simples de procesamiento también llamados neuronas o nodos, unidos entre sí mediante conexiones a las que se le asignan un valor numérico que actuara como peso.

Ventajas del uso de Redes Neuronales Artificiales serán: (Serrano, Soria, & Martín, 2009)

- Se encuentran distribuidas de forma no lineal, lo que permitirá simular sistemas no lineales que, con sistemas lineales, no se podría llevar a cabo.
- Tolerantes a fallos. Se permite el fallo de algunas neuronas a nivel individual sin modificar significativamente la respuesta.
- Adaptabilidad. Las RNA poseen capacidad para modificar los parámetros de que depende su funcionamiento en función de los cambios que se produzcan en el entorno de trabajo.
- Generan relaciones de tipo no lineal entre los datos, es decir, pueden relacionar dos conjuntos de datos.

Por otro lado, algunos de los problemas que presentan en su aplicación serán la complejidad o tiempo elevado en el aprendizaje cuando se demanden grandes tareas, dificultad en la interpretación del aprendizaje o la falta de reglas que ayuden a definir un determinado tipo de red para un problema dado.

Podemos describir la arquitectura de la red en los siguientes componentes (Serrano, Soria, & Martín, 2009):

- **Capa de entrada:** Constituida por los nodos de entrada, su función será la de recibir información del exterior.
- **Conexiones o pesos:** Determinarán el comportamiento de la neurona. Podrán ser excitadoras (positivas) o inhibidoras (negativas).

- **Función de combinación:** Su misión es sumar las entradas multiplicadas por las distintas conexiones.
- **b_j:** Corresponde el sesgo de la red. Irá acompañado del subíndice que hará referencia a la capa en que se encuentre.

Por último, es necesario detallar que una red puede tener más de una capa oculta, es otro de los parámetros a determinar en la arquitectura de la red, aunque normalmente redes con una capa oculta suelen ser suficientes.

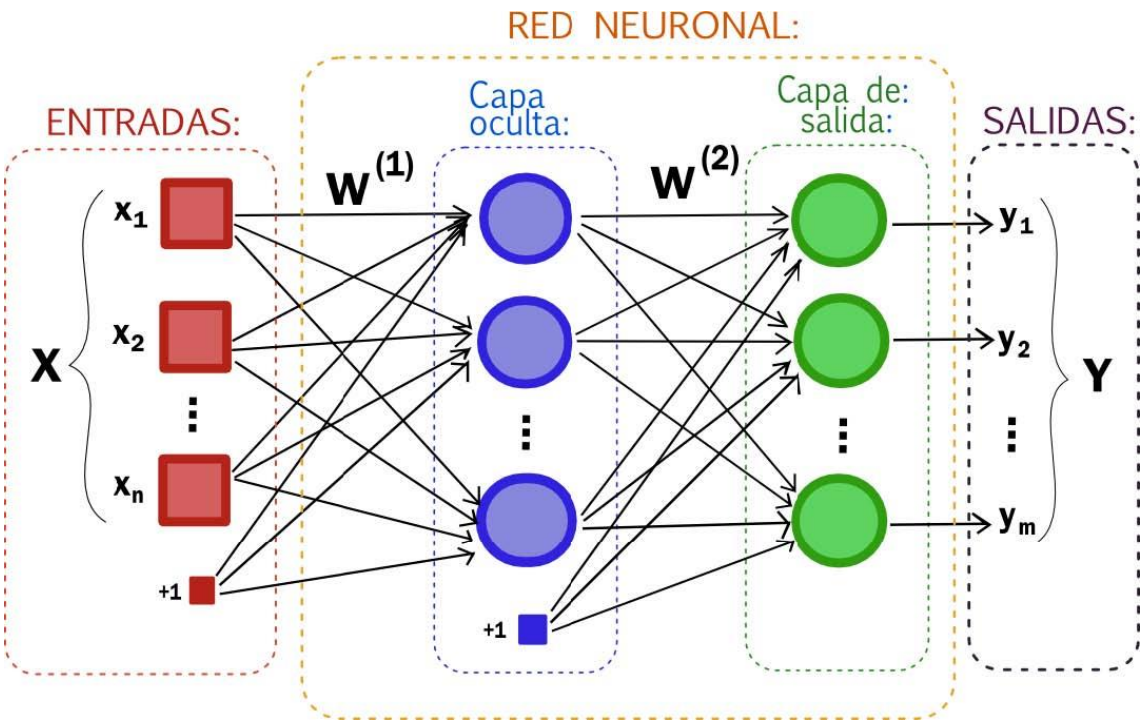

Ilustración 2.- Esquema de Red Neuronal Artificial

La función que engloba todas las entradas que recibe para mandar como entrada para la siguiente capa recibe el nombre de **función de activación**. En ella, una neurona utiliza el valor asociado a su entrada, generando una salida que normalmente es numérica. Las funciones de activación más conocidas son las siguientes:

- Tanh (Tangente Hiperbólica). Transforma los valores introducidos a una escala (-1, 1) de tal manera que los valores altos tiendan de modo asintótico a 1 y los bajos a -1.
- Sigmoide. Similar a la función Tanh, pero con una escala de trabajo de (0, 1).
- Softmax. Convierte las salidas en probabilidades. Por tanto, el sumatorio de todas ellas será 1.
- Relu. Transforma los valores introducidos, anulando los negativos y manteniendo los positivos tan y como entran.

Por último, se definen una serie de medidas para determinar la red óptima, las cuales permitirán decidir que conjuntos de pesos son mejor y cuales peores. Esta parte será usada para determinar los parámetros de la red y, por tanto, el modelo óptimo para realizar una predicción. Destacan entre las mismas (Berzal, 2018):

- Raíz Error Cuadrático Medio (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum(y_j - \widehat{y_j})^2}$$

donde $\widehat{y_j}$ tomará el valor que proporcione la red, y_j corresponderá al valor real y *n* indicará el número de observaciones.

- Error Absoluto Medio (MAE):

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \widehat{y_j}|$$

Define la diferencia en valor absoluto entre predicciones ($\widehat{y_j}$) y observaciones (y_j), donde *n* corresponde nuevamente al número de observaciones.

Dicha función presenta distintos optimizadores, entre los que destacan:

- SGD. Descenso estocástico o gradiente de descenso incremental. Tratará de encontrar mínimos y máximos por iteración.
- Adam. Fijará la ratio de aprendizaje en base a la distribución de los parámetros. En caso de que los parámetros se encuentren muy dispersos esta ratio aumentara.
- Adagrad. Se basa en un gradiente que adapta el radio de aprendizaje a los parámetros.
- Adadelta. Corresponde a una extensión del anterior. En lugar de almacenar ineficientemente los gradientes, estos son definidos de forma recursiva como un promedio decreciente de los anteriores.

Una vez descrita la arquitectura de la red, es necesario saber podrán ser clasificadas las mismas, ya que existen diferentes criterios a tener en cuenta.

Clasificación de Redes Neuronales según la **topología** de red (Serrano, Soria, & Martín, 2009):

- **Red Neuronal Monocapa.** Se trata del modelo más simple. La capa de entrada se proyecta sobre la capa de salida, lugar donde se producen los cálculos.
- **Red Neuronal Multicapa.** Es una generalización de la red monocapa, con la particularidad de que además de las capas de entrada y salida, posee una serie de capas ocultas. Un ejemplo de ella es la que podemos observar en la ilustración 2.
- **Red Neuronal Convulcional.** En este caso, cada neurona no es unida con todas las capas siguientes, sino que "especializamos" un subgrupo de ellas para reducir el número de neuronas necesarias.
- **Red Neuronal Recurrente.** Sustituyen la estructura de capas por conexiones arbitrarias entre neuronas. Tienen la posibilidad de crear ciclos, lo que posibilidad crear la temporalidad para que la red tenga memoria.
- **Red de Base Radial.** Calculan la salida de la función en base a la distancia a un punto denominado centro. La salida es una combinación lineal de las funciones de activación radiales utilizadas por las neuronas.

Clasificación en función del **método de aprendizaje** (Haykin, 1999):

Este tipo de clasificación será dividido en tres bloques. El primero de ellos corresponde a casos en que el aprendizaje es **supervisado**. Se dispone de un entrenamiento controlado por el usuario que determina la respuesta para cada entrada.

- **Aprendizaje por corrección de error.** Ajusta los pesos de las conexiones en función del error cometido.
- **Aprendizaje estocástico.** Se producen cambios aleatorios en los pesos con el objetivo de mejorar la predicción. En caso de que los cambios empeoren los resultados estos son desechados.

Dentro de este apartado se encontrará también el supuesto en que el aprendizaje es **no supervisado** o **autosupervisado**. En este caso, no se requiere de influencia externa para realizar el ajuste de los pesos. (Haykin, 1999)

- **Aprendizaje Hebbiano.** Mide la familiaridad y extrae las características de los datos de entrada.
- **Aprendiza competitivo y comparativo.** Realiza clasificaciones de los datos de entrada. Actúa añadiendo elementos a una clase con el fin de matizar los pesos o construir nuevas clases con ellos.

La última posibilidad que se encontrará es la de aprendizaje **por refuerzo**. Se trata de un tipo más lento que los anteriores, ya que no facilita un conjunto completo para los datos de salida, sino que únicamente indica si el dato es aceptable o no y con ello ajusta los pesos.

Al igual que las redes biológicas, las RNA aprenden por repetición, por lo que un factor importante a tener en cuenta será que cuanto mayor sea nuestro conjunto de entrenamiento mejores serán los resultados que se obtengan.

6.2.3. RANDOM FOREST

Random Forest compone un algoritmo de Machine Learning flexible y de fácil uso que genera grandes resultados. Se trata de uno de los algoritmos de mayor uso dada su simplicidad y también debido al hecho de que se pueden usar tanto para tareas de clasificación como de predicción.

Random Forest corresponde a un algoritmo de aprendizaje supervisado que, parafraseando su propio nombre, crea un bosque y lo convierte en aleatorio, es decir, crea múltiples arboles de decisión y los combina para obtener una predicción más precisa. (Breiman, 2001)

Por tanto, se definirán los arboles de decisión como una forma gráfica y analítica de representar todos los eventos que pueden surgir a partir de una decisión asumida en cierto momento.

Cada árbol es construido mediante la aplicación de un algoritmo en el conjunto de datos de entrenamiento y un vector aleatorio que suele ser usado para muestrear alguna distribución. Se obtiene la predicción en función de las mayorías obtenidas por las predicciones de los árboles de decisión individuales. (Shalev Shwartz & Ben David, 2014)

También se debe distinguir el uso de Random Forest, ya que puede ser usado tanto para clasificación como para regresión. Al ser usado para regresión, las predicciones son promediadas, mientras que, en el caso de la clasificación, como se indicaba, se obtiene un voto

de cada árbol que es clasificado por la mayor predicción obtenida. (Hastie, Tibshirani, & Friedman, 2008)

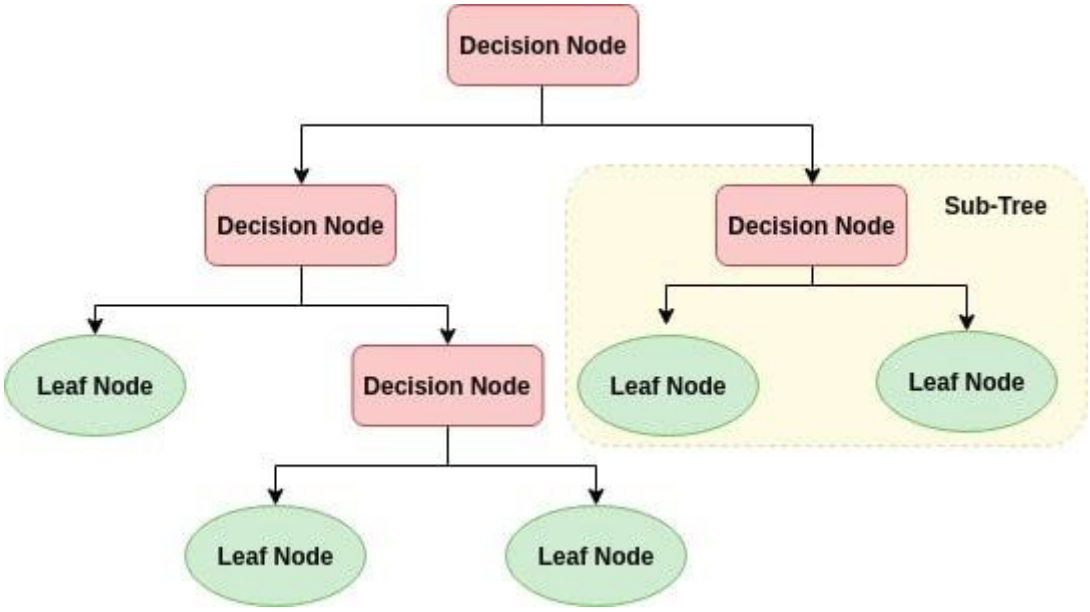

Ilustración 3.- Ejemplo de Árbol de decisión

Dentro de los elementos, también llamados hiperparámetros, que forman parte de esta técnica se definen los más importantes como (Breiman, 2001):

- **Maxtrees.** Número de árboles de decisión de los que estará compuesto el bosque aleatorio.
- **Maxbranch.** Número máximo de características consideradas para dividir un nodo.
- **Leafsize.** Determinará el número mínimo de hojas que se requerirán para dividir un nodo interno.
- **Maxdepth.** Indica el número de procesadores que se puede usar.
- **Random_state.** Hace replicable la salida del modelo.
- **Trainfraction.** Este hiperparámetro nos ayudará a aplicar la validación cruzada adaptada a los bosques aleatorios.

Los pasos a seguir dentro del algoritmo asociado a Random Forest son los siguientes (Hastie, Tibshirani, & Friedman, 2008):

1) Se realiza una representación gráfica para los datos del conjunto de entrenamiento.
2) Se genera el bosque a través de la repetición de los pasos que se mencionan a continuación en cada nodo del árbol hasta alcanzar el nodo de tamaño mínimo.
    a. Se seleccionan m variables al azar entre las que disponemos en el conjunto de datos.
    b. Elección de la mejor variable entre las mismas.
    c. División del nodo en dos sub-nodos.
3) Obtención de la salida del conjunto de árboles.
4) Para realizar la predicción de un nuevo punto:

- Regresión: $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Donde $T_b(x)$ corresponde a la salida del conjunto de árboles y B al número de árboles.

- Clasificación: $\hat{C}_{rf}^{B}(x) = \max(\hat{C}_b(x))$

$\hat{C}_b(x)$ indicará la predicción del b-ésimo árbol.

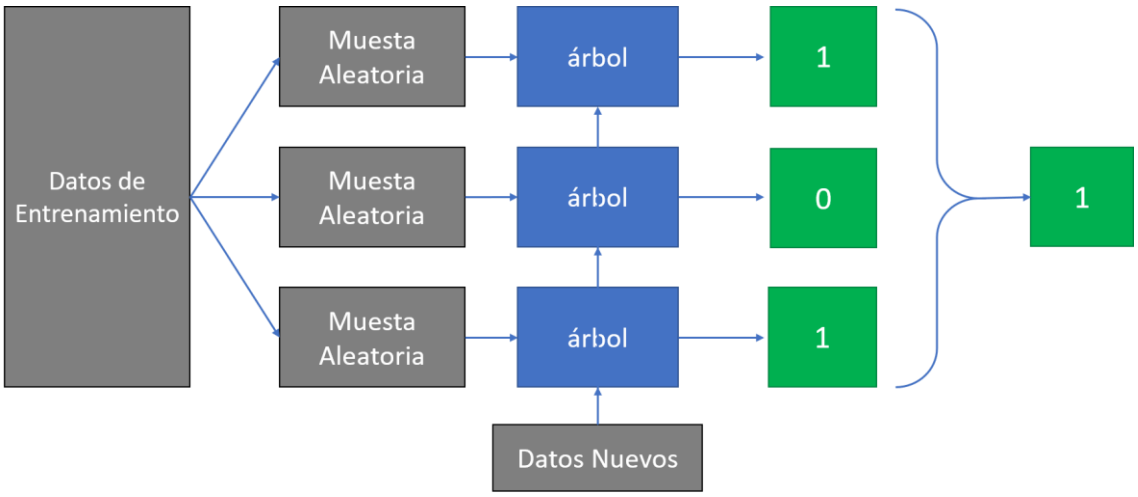

Ilustración 4.- Ejemplo de Random Forest

Se considera como ventaja de Random Forest su facilidad de uso que, a menudo, produce buenos resultados de predicción. Como ultima ventaja, destaca la capacidad del Random Forest para evitar el sobreajuste, uno de los grandes problemas existentes en el aprendizaje automático. (Hastie, Tibshirani, & Friedman, 2008)

Por el contrario, una de las principales limitaciones del Random Forest será que gran cantidad de árboles provocarán que el procedimiento sea lento y poco efectivo para predicciones en tiempo real. Una predicción más precisa requerirá de más árboles provocando dicha lentitud. (Hastie, Tibshirani, & Friedman, 2008)

6.3. SELECCIÓN DEL MODELO ÓPTIMO

Llegados a este punto, se han conocido diferentes técnicas, las cuales serán aplicadas sobre los datos con la finalidad de encontrar un método de predicción de irregularidades óptimo.

Es por ello, que en este apartado se abordarán los criterios que se deben tener en cuenta a la hora de seleccionar el mejor de los modelos generados para cada una de estas técnicas y, finalmente, seleccionar el mejor de todos para su posterior aplicación.

Es necesario comenzar indicando que el conjunto de datos será dividido en dos grandes grupos:

- **Conjunto de datos de entrenamiento:** Englobará los ficheros que serán utilizados para entrenar modelos y llevar a cabo la estimación de parámetros para cada uno de ellos.
- **Conjunto de datos de validación:** Su misión será el cálculo de estadísticos básicos que indicarán la calidad predictiva del modelo obtenido.

Con ello se evitará el sobreajuste en los casos en que no sea posible generalizar un patrón, bien sea por que el conjunto de los datos de entrenamiento sea pequeño o porque el número de parámetros del modelo es grande.

A partir de los datos obtenidos mediante la aplicación de las técnicas descritas hasta aquí, se podrá clasificar cualquier conjunto de datos en cuatro grandes grupos que permitirán continuar realizando cálculos con la finalidad de determinar el modelo óptimo. Dichos grupos se representan a través de la **matriz de confusión** y serán (Benítez, Escudero, & Kanaan, 2013):

- **VP – Verdaderos positivos.** Serán aquellas observaciones que inicialmente poseían la característica sometida a estudio en la variable respuesta y tras la aplicación de un determinado modelo de regresión siguen manteniéndola.
- **FP – Falsos positivos.** Observaciones que en la variable respuesta no poseen la característica en estudio y que explicada por el modelo cambian su valor y, según es, si la poseerían.
- **VN – Verdaderos negativos.** Situación similar a la de VP, con la particularidad de que en este caso se evalúan las observaciones que no poseen la característica de referencia y son correctamente clasificados por nuestro modelo.
- **FN – Falsos negativos.** Al igual que en el punto anterior, se repite la operación, pero con aquellas observaciones que inicialmente poseen la característica de referencia y tras la aplicación del modelo son clasificadas como si no la poseyeran.

Se hará uso de dichos valores para calcular la bondad ajuste del modelo, criterio que será de ayuda una vez llegue el momento de elegir el mejor de ellos entre todos los calculados. Para ello, se calcularán otros cuatro valores que serán los necesarios para llevar a cabo esta evaluación. (Benítez, Escudero, & Kanaan, 2013)

- **Sensibilidad.** Se define este concepto como la capacidad del modelo para identificar una observación que presenta la característica sometida a estudio, en nuestro caso, la capacidad para detectar las irregularidades. Su cálculo viene expresado de la siguiente manera:

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

- **Especificidad.** Este valor tiene como objetivo evaluar la capacidad para identificar correctamente aquellas observaciones que no presentan la característica de referencia, es decir, se centra en evaluar la proporción de negativos clasificados correctamente.

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

- **Valor Predictivo Positivo.** Se establece el VPP como herramienta fundamental en la Regresión Logística, dado que su valor establecerá la bondad de ajuste del modelo. Otra definición del mismo aplicada a este estudio sería la probabilidad de que un usuario este cometiendo fraude una vez que el modelo ha detectado alguna irregularidad.

$$\mathrm{VPP} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

- **Valor Predictivo Negativo.** Corresponde a una evaluación similar a la realizada en el VPP, pero en este caso con aquellas observaciones que no presentan la característica de referencia. Por tanto, se definirá el VPN como la probabilidad de que un cliente no este cometiendo irregularidad alguna si el modelo lo clasifica como no fraudulento.

$$\mathrm{VPN} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FN}}$$

Por último, se detallan los criterios que serán tenidos en cuenta para realizar la selección del modelo óptimo.

- **Bondad de ajuste** del modelo. Para ello, siempre se estudiará aquel que presente un mayor Valor Predictivo Positivo, esto es, aquel que mejor clasifique correctamente a los clientes que comenten irregularidades.
- **Principio de parsimonia** (navaja de Ockam), el cual indica que, *en igualdad de condiciones, la explicación más sencilla suele ser la más probable*. En este caso, será aplicado al número de parámetros dentro del modelo, tomando como óptimo aquel que menos presente. (Lacey, 1976)

A estos criterios se ha de añadir que también se valorará el **número de visitas** pronosticadas por el modelo. Para ello, se simularán las predicciones en 50000 y 100000 visitas, analizando los resultados que ofrecen los modelos, ya que es obvio que, a mayor número de visitas, la bondad de ajuste será menor. Se evaluará, por tanto, la robustez del modelo elegido al aumentar el número de visitas.

7. RESULTADOS

A lo largo de este capítulo se aplicará la metodología anteriormente descrita junto a la depuración previamente realizada a nuestro conjunto de datos.

Cabe recordar que no se incluirán en este apartado los resultados de imputación, análisis de valores perdidos y agrupación y creación de variables, pues ya fueron descritos en capítulos anteriores (véase capítulo 4 y 5).

Se comenzará realizando la selección de la variable respuesta que mejor se relacione con las variables explicativas de que consta la base de datos. Para ello, el procedimiento elegido ha sido la Regresión Logística. Una vez se disponga de la variable respuesta óptima, se procederá a aplicar las distintas técnicas predictivas descritas divididas en distintas secciones.

Así pues, cada sección desarrollará obligatoriamente las siguientes tareas:

- Se indicarán los parámetros que se han definido en cada técnica para calcular diferentes modelos, como pueden ser pvalores, funciones de enlace, métodos asociados a la técnica analizada, etc.
- Evaluación de las salidas ofrecidas por el software en base a todos los criterios dispuestos en la metodología.
- Selección y desarrollo del modelo óptimo para cada una de las técnicas sometidas a estudio.

Como fue indicado previamente, el conjunto de datos fue dividido en dos grandes partes: entrenamiento y validación. Detallar que el conjunto de entrenamiento estará compuesto por los archivos 24 a 45, inclusive, que albergarán el periodo comprendido entre diciembre de 2016 y septiembre de 2018. El conjunto de datos de validación constará del resto de observaciones, es decir, de los archivos 46 a 61, abarcando el periodo de octubre de 2018 a enero de 2020, ambos incluidos. El hecho de que no se haga referencia a los archivos 1 a 23 es porque estos ya fueron usados para la creación de las variables explicativas y, por tanto, su información ya se encuentra recogida en estos datos.

Por último, detallamos que el software utilizado para el desarrollo del presente estudio será el software estadístico SAS 9.4.

7.1. SELECCIÓN VARIABLE RESPUESTA

Una vez que se dispone de las posibles variables respuesta (desde Irr1 a Irr12), las cuales fueron generadas atendiendo a los criterios dispuestos en el apartado 5.1. es momento de determinar cuál de todas ellas se relaciona mejor con el resto de variables y, por tanto, proporcionará una mejor predicción.

Esta tarea corresponde a una de las decisiones más importantes que vamos a tomar antes de generar nuestro modelo, ya que la calidad de este estará supeditada a la relación que exista entre las variables del conjunto de datos con las posibles variables respuesta de que se dispone. Para analizar dicha relación, se hará uso de la Regresión Logística.

Para ello, se repetirá el proceso en 12 ocasiones (una por cada Irr como variable respuesta), en el que por abreviar se han valorado 4 escenarios posibles y que no tendrían por qué ser los más relevantes:

- En el primero de ellos, se analizará únicamente la variable *consumo* con variable regresora.
- El siguiente paso, únicamente tendrá en cuenta como variable explicativa *el Año Fabricación Contador*.
- Para la tercera regresión, se hará uso de las dos variables usadas en los puntos anteriores conjuntamente.
- Finalmente, y al igual que se ha hecho para analizar las correlaciones, se hará uso de todas las variables numéricas que se utilizarán para la creación del modelo.

Se resumen en una tabla los porcentajes de fraude clasificados correctamente para su análisis:

	CONSUMO - AÑO	CONSUMO	AÑO	NUMERICAS
Irr1	83,3	84,0	72,4	-
Irr2	70,7	32,2	69,4	80,2
Irr3	70,8	32,6	69,4	80,1
Irr4	70,3	36,4	68,9	77,1
Irr5	70,3	36,5	68,9	77,1
Irr6	70,9	38,6	69,3	74,1
Irr7	70,9	38,7	69,3	74,1
Irr8	72,0	38,9	70,5	71,9
Irr9	72,0	39,0	70,4	71,8
Irr10	73,7	37,7	72,2	70,4
Irr11	73,6	37,6	72,1	70,3
Irr12	74,9	36,8	73,5	69,7

Tabla 1.- Porcentaje Clasificación Correcta Variables Respuesta

En primer lugar, se comenzará diciendo que se descarta como candidata a variable respuesta la variable Irr1, ya que solo dispone de 9 observaciones válidas y, por tanto, los resultados que se muestran en la tabla omiten los valores para dicha opción.

Se puede comprobar que, para la primera y la tercera opción, la variable respuesta ideal correspondería a la Irr12, mientras que, para la segunda opción, la ideal sería la Irr9. Sin embargo, al realizar la Regresión Logística con todas las variables numéricas, el mejor porcentaje se obtiene para la variable Irr2.

Por tanto, se calcularán algunos estadísticos que determinen el grado de ajuste de este modelo como pueden ser la D-Sommers, el coeficiente Gamma y la Tau-c. Se puede comprobar en la tabla 2 como el mayor valor y, por tanto, el más próximo a 1 es nuevamente el que corresponde a la regresión estudiada estableciendo como variable respuesta Irr2, lo que muestra indicios de que esta debería ser la seleccionada.

No obstante, se analizarán las relaciones entre las variables categóricas del conjunto de datos inicial y las variables respuesta analizadas en cada uno de los cuatro modelos expuestos con la finalidad de determinar cualquier relación inesperada entre estas variables respuesta y dichas variables, lo que podrá confirmar si estamos en lo cierto al tomar alguna de estas tres como nuestra variable respuesta definitiva.

	D-SOMMERS	GAMMA	TAU-C
Irr1	-	-	-
Irr2	0,604	0,604	0,802
Irr3	0,603	0,603	0,801
Irr4	0,541	0,542	0,771
Irr5	0,541	0,541	0,771
Irr6	0,482	0,482	0,741
Irr7	0,482	0,482	0,741
Irr8	0,438	0,438	0,719
Irr9	0,437	0,437	0,718
Irr10	0,408	0,408	0,704
Irr11	0,407	0,407	0,703
Irr12	0,393	0,393	0,697

Tabla 2.- Estadísticos Regresión Logística

Por tanto, se comenzará analizando las visitas para estudiar si estas se han realizado de forma aleatoria y, acto seguido, se creará una nueva tabla en la que se podrá comprobar el porcentaje de fraude cometido en las visitas realizadas. Para un mejor entendimiento de las tablas adjuntas se detalla que si las visitas son excesivas y el fraude también, esa categoría indicará un claro fraude y las marcaremos en rojo. Si las visitas son escasas y el fraude también será evidencia de la ausencia de fraude y, por tanto, serán marcadas en verde. En cualquier otro caso, será necesaria la recopilación de más información.

Se destaca que, si bien el estudio se ha realizado para todo el conjunto de variables, únicamente se adjuntan las variables más relevantes para así poder mostrar las diferentes casuísticas que se han planteado. Dicho esto, las variables que se adjuntan serán las siguientes: *TipoLectura, Tarifa* y *GrupoTipo*.

TIPO LECTURA	% Población	% Visitas Irr2	% Visitas Irr9	% Visitas Irr12
AUSENTE	7,65	10,21	9,58	10,34
BLOQUEADA	0,02	0,13	0,11	0,11
FACILITADA	7,05	4,49	4,61	5,43
REAL	85,27	85,16	85,70	84,12
TOTAL	100,00	100,00	100,00	100,00

Tabla 3.- Porcentaje Visitas Tipo Lectura

La tabla 3 proporciona información acerca de las visitas realizadas. Se puede ver como los usuarios con tipo de lectura ausente y bloqueada reciben un porcentaje de visitas mayor al que se esperaría por el porcentaje en que se distribuyen en nuestro conjunto de datos.

TIPO LECTURA	Fraude Irr2	No Fraude Irr2	Fraude Irr9	No Fraude Irr9	Fraude Irr12	No Fraude Irr12
AUSENTE	24,51	75,49	26,87	73,13	28,19	71,81
BLOQUEADA	42,86	57,14	55,56	44,44	52,00	48,00
FACILITADA	36,99	63,01	38,74	61,26	38,49	61,51
REAL	38,02	61,98	39,48	60,52	41,38	58,62
MEDIA	36,60	63,40	38,26	61,74	39,87	60,13

Tabla 4.- Porcentaje Fraude Tipo Lectura

Por otro lado, en la tabla 4 se comparan los porcentajes de fraude de dichas visitas con respecto a los valores medios. Se observa como el porcentaje de fraude en las visitas de clientes que presentan un tipo de lectura ausente o facilitada es inferior a la media, siendo superiores los de las lecturas bloqueadas o reales.

Uniendo la información que proporcionan ambas tablas, se podrá concluir que las lecturas de tipo bloqueado presentaran un mayor fraude y el de las facilitadas será menor. Sin embargo, se necesitarán más visitas para alcanzar conclusiones más fiables en las categorías ausente y real.

TARIFA	% Población	% Visitas Irr2	% Visitas Irr9	% Visitas Irr12
3.1	38,71	37,82	38,68	39,62
3.2	60,15	51,98	60,19	51,09
3.3	0,24	1,45	0,24	1,48
3.4	0,90	8,75	0,89	7,82
TOTAL	100,00	100,00	100,00	100,00

Tabla 5.- Porcentaje Visitas Tarifa

A excepción de la tarifa 3.2, todas las tarifas están recibiendo más visitas de las esperadas. Estos datos pueden ser contrastados en la tabla 5.

TARIFA	Fraude Irr2	No Fraude Irr2	Fraude Irr9	No Fraude Irr9	Fraude Irr12	No Fraude Irr12
3.1	49,67	50,33	48,48	51,52	44,97	55,03
3.2	57,28	42,72	56,92	43,08	54,38	45,62
3.3	10,48	89,52	7,02	92,98	8,32	91,68
3.4	15,30	84,70	12,56	87,44	12,19	87,81
MEDIA	50,05	49,95	49,27	50,73	39,87	60,13

Tabla 6.- Porcentaje Fraude Tarifa

Sin embargo, según la tabla 6, se observa que es esta tarifa, la 3.2, la única que supera el porcentaje medio de fraude.

El conjunto de esta información indica que la tarifa 3.2 es la que mayor fraude presenta, por lo que se necesitará la realización de más visitas ya que es la que menos visitas recibe.

GRUPOTIPO	% Población	% Visitas Irr2	% Visitas Irr9	% Visitas Irr12
1	4,58	0,42	1,22	1,59
2	18,13	14,02	19,50	18,88
3	10,41	4,96	9,34	10,21
4	1,07	7,08	3,70	4,77
5	56,78	62,75	51,50	53,29
6	1,92	4,82	8,87	6,75
7	7,11	5,95	4,54	4,51
TOTAL	100,00	100,00	100,00	100,00

Tabla 7.- Porcentaje Visitas GrupoTipo

Se puede observar en la tabla 7 como los grupos 2, 4 y 6 reciben más visitas de las esperadas según el porcentaje que representan en la población, mientras que el resto de grupos se encuentran por debajo de estos valores, aunque si bien, hay grupos que se mueven en valores cercanos a los que les correspondería.

GRUPOTIPO	Fraude Irr2	No Fraude Irr2	Fraude Irr9	No Fraude Irr9	Fraude Irr12	No Fraude Irr12
1	0,00	100,00	11,11	88,89	34,48	65,52
2	85,86	14,14	48,61	51,39	48,26	51,74
3	60,00	40,00	45,65	54,35	50,54	49,46
4	42,00	58,00	33,33	66,67	35,63	64,37
5	73,81	26,19	63,40	36,60	56,75	43,25
6	14,71	85,29	12,21	87,79	13,01	86,99
7	45,24	54,76	43,28	56,72	39,02	60,98
MEDIA	32,29	67,71	51,12	48,88	49,40	50,60

Tabla 8.- Porcentaje Fraude GrupoTipo

En la tabla 8 se analizan los porcentajes de fraude, viendo como los grupos 2, 3 y 4 mostrarán valores muy superiores en la realidad.

Concatenando la información aportada por ambas tablas, sería necesaria la recopilación de una mayor información para los grupos 3, 4, 5 y 6 para así poder obtener conclusiones más fiables.

Con este trabajo, también se ha podido comprobar la relación que guardan las posibles variables respuesta con el conjunto de variables cualitativas de que se dispone.

Una vez que se han analizado todas las relaciones entre variables, finalmente, y en base al conjunto del análisis abordado en este apartado, se ha realizado la selección de la **variable Irr2** como la variable respuesta óptima para el modelo.

7.2. REGRESION LOGISTICA

La Regresión Logística será dividida en tres grandes bloques que corresponderán a cada uno de los tres métodos de selección de variables: **Backward**, **Forward** y **Stepwise**.

Se establecerán distintos pvalores de entrada o salida de variables dependiendo del método que serán los siguientes: **0.05, 0.001, 0.0001, 0.00001, 0.000001, 0.000000001 y 0.00000000001**.

Las funciones de enlace utilizadas para este estudio han sido: **logit**, **probit** y **cloglog**.

Por último, se simularán resultados en el conjunto de validación para la realización de 50000 y 100000 visitas.

7.2.1. MÉTODO BACKWARD

Como ya se indicaba en la metodología, el método Backward corresponde a la selección de variables hacia atrás. Se partirá de modelos que incluyan todas las variables y se irá modificado el criterio de salida de variables en base a los pvalores indicados y haciendo uso de las distintas funciones de enlace mencionadas.

Al disponer de 7 pvalores de salida distintos, 3 funciones de enlace distintas y 2 parámetros asociados al número de visitas, se obtendrán un total de 42 modelos.

En base a las características dispuestas, se genera el anexo 1, que presentará la clasificación realizada sobre los datos que conforman el conjunto de validación (Verdaderos Positivos, Verdaderos Negativos, Falsos Positivos, Falsos Negativos), el número de visitas a ejecutar, el número de parámetros de que consta el modelo, la bondad de ajuste del mismo, función de enlace, método de selección de variables y pvalor de salida de variables. Dicho anexo se ordenará en orden descendente según la mayor bondad de ajuste obtenida.

Atendiendo a los distintos criterios definidos en la metodología, y aplicados al conjunto de datos de validación, vemos como por bondad de ajuste, el mejor modelo correspondería al modelo 1, que presenta un valor del 88.13%. Dado que el volumen de visitas pronosticado es de 50000, se observan los datos obtenidos para el mismo modelo con 100000 visitas (modelo 26), donde se comprueba una bondad del ajuste del 80.23%.

Sin embargo, atendiendo al criterio de Ockam, los modelos con un menor número de variables serían los números 21, 22, 24, 28, 32, 37, 38 y 39, que presentarían bondades de ajuste que se situarían entre el 75 y el 81%, todos ellos con una cantidad de visitas pronosticadas entre 50000 y 100000.

Por último, se estudia el mejor modelo con función de enlace logit, ya que esta presentará mayor facilidad a la hora de realizar los cálculos. Dicho modelo presentaría una bondad de ajuste del 85.71% y correspondería al modelo 5. Dicho modelo pronostica la realización de 50000 visitas, 100000 la bondad de ajuste se situaría en el 84.11% (modelo 11).

Por tanto, realizando la combinación de los criterios expuestos para la selección del mejor modelo, se selecciona el modelo 11 como el mejor modelo para la Regresión Logística utilizando el método Backward, ya que la pérdida en bondad de ajuste es mínima, pero se obtiene un aumento considerable en el número de visitas, fundamental para la detección de clientes que cometen irregularidades. Asimismo, se considera que será un modelo robusto, ya que al aumentar el número de visitas el doble, el descenso de la bondad de ajuste se sitúa en el 1.6%.

Las características de dicho modelo son las siguientes:

- Selección de variables: Método Backward
- Función de enlace: Logit
- Numero de parámetros: 14
- Pvalor de salida: 0.00001
- Visitas: 100000

7.2.2. MÉTODO FORWARD

El método Forward corresponde al tipo de selección de variables hacia delante. El modelo inicial no poseerá ninguna variable, las cuales irán siendo introducidas al mismo en función del pvalor definido para que se produzca la entrada.

Al igual que en el método Backward, dado que el número de combinaciones de parámetros que se realizarán es el mismo, también se obtendrán 42 modelos, los cuales serán evaluados atendiendo a los mismos criterios.

Se puede comprobar en el anexo 2 como los mejores modelos obtenidos según la clasificación realizada en el conjunto de validación son los siguientes:

- Atendiendo a la mejor bondad de ajuste, el mejor modelo sería el número 1 y usaría una función de enlace cloglog, constaría de 13 parámetros y situaría la calidad del modelo en un 87.72% de usuarios fraudulentos correctamente clasificados. Dado que las visitas establecidas son 50000, se procede a analizar la bondad de ajuste cuando estas aumentan a 100000, donde se comprueba que el valor para este dato es del 84.15% (modelo 20).
- El menor número de variables queda establecido en 5 para los modelos 15, 18, 37, 39 y 40, con bondades de ajustes situadas en el intervalo del 80-85% y haciendo uso de las distintas funciones de enlace que ya se han descrito y con 50000 – 100000 visitas pronosticadas.
- El mejor modelo con función logit, mayor bondad de ajuste (86.9%) y menor número de parámetros (7) corresponde al modelo número 7. Dicho modelo pronostica 50000 visitas, por lo que se analiza robustez mediante la bondad de ajuste para 100000 visitas (86.05% - Modelo 12), donde se observa que la pérdida es mínima.

Por tanto, se selecciona el modelo 12 como el mejor de los modelos. Se describen sus características:

- Selección de variables: Método Forward
- Función de enlace: Logit
- Numero de parámetros: 7
- Pvalor de salida: 0.000000001
- Visitas: 100000

7.2.3. METODO STEPWISE

Se usará el método Stepwise para mezclar los criterios de selección de variables de los dos métodos anteriores (Backward y Forward). Así, se establecen los pvalores indicados de forma conjunta, definiéndolos simultáneamente como criterio de entrada y de salida. Se partirá de un modelo sin variables en el que se irán introduciendo una a una y, tras cada entrada, se comprobará si alguna variable debe abandonar el modelo y, en ese caso, se eliminará.

Nuevamente se obtienen 42 modelos, que aplicados al conjunto de datos de validación quedan resumidos en el anexo 3.

- Bondad de ajuste: El mejor modelo correspondió a los modelos número 1 al 4, con una clasificación correcta para los clientes que habían cometido alguna irregularidad del 87.3%. Dichos modelos usan función de enlace logit probit y contienen 9 parámetros cada uno. El pronóstico de visitas es de 50000, que al aumentarlo a 100000 reduciría la bondad de ajuste en un 1% para aquellos que usaron función de enlace logit y en un 3% para el caso de la función de enlace probit.
- En base a la razón de Ockam, los modelos con menor número de parámetros fueron los modelos 39, 40, 41 y 42, todos ellos con función cloglog como enlace. La bondad de ajuste rondará el 68-74% en todos los casos.
- El mejor modelo evaluado con función de enlace logit es el modelo 5, que incluye 9 parámetros, una pérdida en bondad de ajuste del 0.3% y 50000 visitas pronosticadas. Si es analizado con 100000 visitas perderemos un 3% (modelo 19).

Evaluando los tres criterios de forma simultánea, se llega a la conclusión del que el mejor modelo obtenido al utilizar el método Stepwise sería el modelo número 19, ya que no posee un número muy elevado de parámetros, la perdida en bondad de ajuste es mínima con respecto al máximo que se obtiene por este método y la función de enlace seria logit. Además, el pvalor asociado a la entrada y salida de variables es más exigente, por lo que se considerará un modelo más robusto.

- Selección de variables: Método Stepwise
- Función de enlace: Logit
- Numero de parámetros: 9
- Pvalor de salida: 0.000001
- Visitas: 100000

7.2.4. MODELO OPTIMO REGRESION LOGISTICA

Una vez analizados los tres métodos de forma individual, se debe analizar el mejor modelo obtenido para cada uno de ellos y seleccionar el modelo óptimo para esta técnica. Se resumen los datos de dichos modelos en la tabla 9:

Obs	VP	FP	FN	VN	visitas	num_parametros	Bondad_Ajuste	Funcion	Metodo	P_Entrada
1	90	17	295	418	100000	14	0.84112	Logit	Backward	.00001000000
2	74	12	311	423	100000	7	0.86047	Logit	FORWARD	.00000000100
3	75	14	310	421	100000	9	0.84270	Logit	STEPWISE	.00000100000

Tabla 9.- Mejores Modelos Regresión Logística

En este caso, todos los modelos usan función de enlace logit, por lo que este no será un factor determinante a la hora de elegir el mejor modelo.

Sí que se puede comprobar que la bondad de ajuste es ligeramente mayor para el método Forward. Dado que en el modelo calculado mediante el método Forward, el número de variables es menor y el número de visitas es idéntico para todos los modelos, nos decantamos por este para seleccionarlo como modelo optimo al aplicar la Regresión Logística y se pasa a detallarlo en profundidad.

Estimación de parámetros:

Se realiza el cálculo de los parámetros del modelo, que quedarán recogidos en la tabla 10 y que indicarán los siguientes datos:

- **DF:** Grados de libertad. Dado que cada contraste se realizará únicamente para la variable a estudiar, siempre tomará el valor 1.
- **Estimador:** Valor que toma el parámetro.
- **Error estándar:** Error estándar del parámetro estimado.
- **Estadístico de Wald** (véase apartado 7.1.1.).
- **P-valor:** Corresponde al pvalor obtenido al realizar el test de Wald.

Análisis del estimador de máxima verosimilitud					
Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	-91.8148	11.7167	61.4064	<.0001
DESCRIPCION_UBIC_CON	1	4.5827	0.6613	48.0242	<.0001
NumeroGrupoAcc12_5	1	-2.4032	0.2612	84.6554	<.0001
PercentEstSuminis12_	1	-1.3844	0.2160	41.0779	<.0001
Tarifa30	1	5.3536	0.6147	75.8522	<.0001
UltimaVisita6	1	1.2943	0.1725	56.3180	<.0001
ProbLocal2_30	1	4.4852	0.2704	275.0724	<.0001
A_O_FABRICACION_CONT	1	0.0428	0.00582	54.1493	<.0001

Tabla 10.- Parámetros Modelo Regresión Logística

La primera conclusión a que se podrá llegar tras estudiar los parámetros seleccionados es la de que todas las variables incluidas en el modelo son significativas, algo trivial si se tiene en cuenta que el método usado para realizar dicha selección, solo las introducía en caso de serlo.

Por tanto, la ecuación del modelo quedaría del siguiente modo:

Z = -91.8148 + (4.5827 * *Descripción Ubicación Contador*) - (2.4032 * *NumeroGrupoAcc12_5*) - (1.3844 * *PercentEstSuminis12_Estado 1*) + (5.3536 * *Tarifa30*) + (1.2943 * *UltimaVisita6*) + (4.4852 * *ProbLocal2_30*) + (0.0428 * *Año Fabricación Contador2*)

$$\text{P}\{\text{Cliente cometa irregularidad}\} = \frac{e^{Z}}{1 + e^{Z}}$$

A continuación, se define la matriz de confusión para el modelo obtenido:

		Predicción	
		Fraude	**No Fraude**
Observación	**Fraude**	74	311
	No Fraude	12	423

La matriz de confusión ayudará a calcular los siguientes valores:

- **Sensibilidad:**

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{74}{74 + 311} = 0.1922$$

El 19.22% de los clientes que están cometiendo alguna irregularidad serán detectados.

- **Especificidad:**

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}} = \frac{423}{423 + 12} = 0.9724$$

El 97.24% de los clientes que no cometen irregularidades será clasificado correctamente.

- **Valor Predictivo Positivo:**

$$\text{VPP} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{74}{74 + 12} = 0.8605$$

El 86.05% de los clientes clasificados como fraudulentos estará cometiendo alguna irregularidad.

- **Valor Predictivo Negativo:**

$$\text{VPN} = \frac{\text{VN}}{\text{VN} + \text{FN}} = \frac{423}{423 + 311} = 0.5763$$

El 57.63% de los clientes que el modelo clasifica como no fraudulentos los estará correctamente.

Para facilitar la interpretación de estos datos, se representa gráficamente la curva ROC, la cual determinará la exactitud diagnóstica de los análisis realizados. Dicha gráfica indica que la clasificación obtenida es correcta, ya que el área bajo la curva toma un valor próximo a uno. El umbral de la curva en que se obtienen los mejores resultados será aquel punto en que la sensibilidad y especificidad se encuentren más próximas, aproximadamente en 0.75.

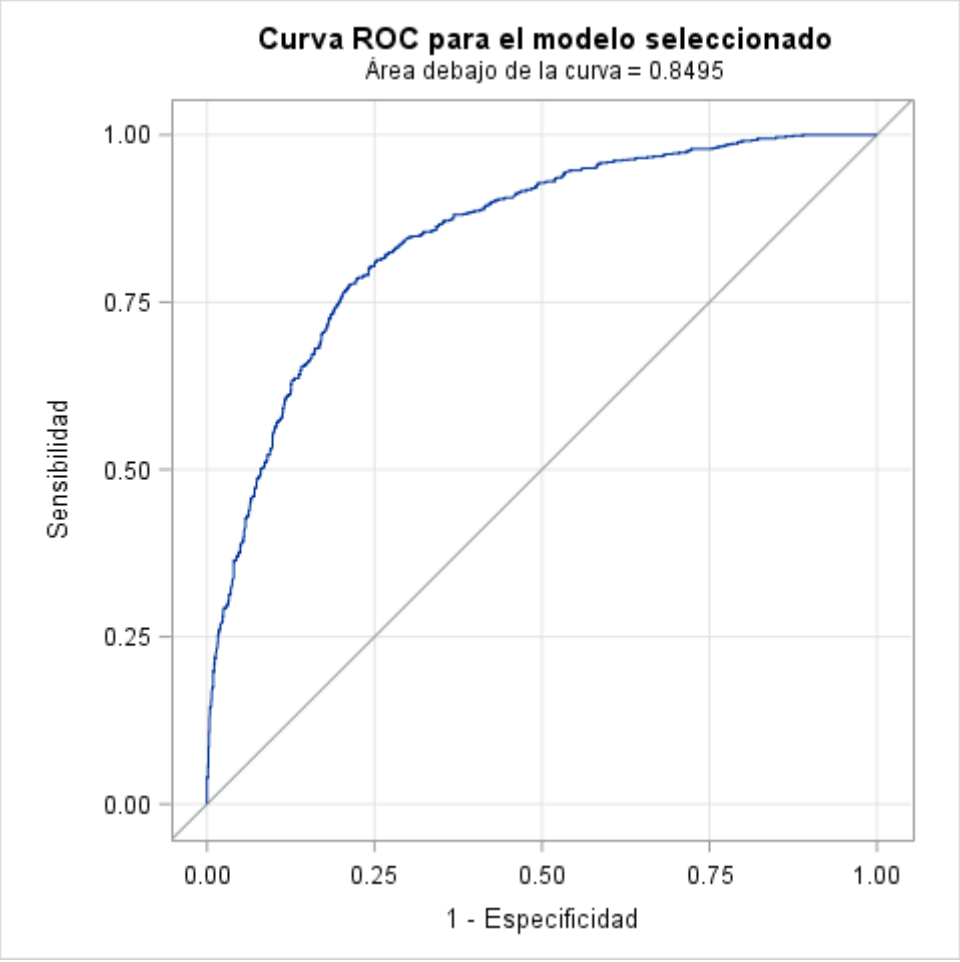

Ilustración 5.- Curva ROC Regresión Logística Óptima

El porcentaje de clientes clasificados correctamente se sitúa en torno al 85% independientemente de que sea un cliente que haya cometido fraude o no, lo cual indica que nos movemos en valores más que aceptables.

Se comprueba también como los estadísticos que analizan la clasificación realizada para las observaciones del conjunto de datos, tales como la D de Somers, la Gamma o la Tau-c de mantienen próximos a 0.7 o superior a este valor en el caso de esta última, lo que prueba, una vez más, la calidad del modelo.

Asociación de probabilidades predichas y respuestas observadas			
Concordancia de porcentaje	84.9	D de Somers	0.699
Discordancia de porcentaje	15.0	Gamma	0.699
Porcentaje ligado	0.0	Tau-a	0.333
Pares	1276050	c	0.850

Tabla 11.- Estadísticos Básicos Regresión Logística

Por último, se analizan los valores obtenidos para los odds ratio junto con sus intervalos de confianza:

Estimadores de cocientes de disparidad;			
Efecto	Estimador del punto	Límites de confianza al 95% de Wald	
DESCRIPCION_UBIC_CON	97.774	26.751	357.361
NumeroGrupoAcc12_5	0.090	0.054	0.151
PercentEstSuminis12_	0.250	0.164	0.382
Tarifa30	211.372	63.360	705.151
UltimaVisita6	3.648	2.602	5.116
ProbLocal2_30	88.691	52.202	150.685
A_O_FABRICACION_CONT	1.044	1.032	1.056

Tabla 12.- Odds Ratio Regresión Logística Óptima

De la tabla adjunta se puede ver como la variable *Año_Fabricación_Contador2* no tiene asociación con la variable respuesta, ya que toma un valor muy próximo a uno.

Además, se podrán extraer datos como que un cliente que ha recibido visita en los 6 últimos meses, tendrá una probabilidad de haber cometido fraude 3.6 veces mayor que uno que no la haya recibido.

También se comprueba como por cada unidad que aumenta la variable Tarifa30, la probabilidad de estar cometiendo fraude aumenta en más del 200%.

Por último, se destaca que para las variables que tomen valores entre 0 y 1 su Odds Ratio no será interpretable.

7.3. REDES NEURONALES

Uno de los principales problemas que se encuentran ante el uso de Redes Neuronales es el de la selección de variables. Tanto para Regresión Logística como para Random Forest, se podrán introducir todas las variables en el modelo y estas técnicas determinan que variables serán significativas y cuáles no, mientras que en el uso de Redes Neuronales, se podrán usar todas las variables como entrada si es lo deseable, pero encontraremos un problema de interpretabilidad en las variables de la red, ya que no se podrá saber con la misma claridad que en el resto de modelos, la influencia ejercida por cada variable sobre la variable respuesta.

Sin embargo, si nos fijamos en los modelos obtenidos mediante los distintos métodos de Regresión Logística, todas las variables seleccionadas se encuentran bastante relacionadas entre sí. Es por ello que se ha decidido dividir el cálculo de modelos de Redes Neuronales en tres bloques. En cada uno de ellos, se introducirán en la red las variables que conforman los mejores modelos de Regresión Logística para cada uno de los distintos métodos de selección de variables.

Se partirá de un modelo con 8 nodos como máximo en el que se aplicarán las funciones de activación tangente hiperbólica, exponencial, arco tangente y Elliott. Destacar que los modelos serán aplicados sobre el conjunto de validación y serán probados con una única capa oculta según la justificación desarrollada en nuestra metodología. Además, se mantendrán los números de visitas a desarrollar que se han pronosticado en el análisis de Regresión Logística y que se establecen en 50000 y 100000 visitas.

7.3.1. VARIABLES REGRESION LOGISTICA BACKWARD

Las variables que serán incluidas para la generación de este primer bloque de modelos corresponden a aquellas que se seleccionaron en el mejor modelo de Regresión Logística usando el método Backward. Dicho grupo está conformado por las siguientes variables: *VarOtros7, VarOtros11Sum, Descripcion_Ubic_Cont30, NumeroGrupoTipo12_4, NumeroGrupoTipo6_2, PercentEstSuminis12_Estado_1, PercentTipoLectura12_Real, PercentTipoLectura24_Facilitada, Tarifa30, TiempoUltimaVisita6, TiempoUltimaVisita24, Ultima_Visita_6, Prob_Local2_30 y Año Fabricación Contador2.*

Tras combinar nodos con funciones de activación y visitas pronosticadas, se obtendrán 64 modelos, los cuales fueron aplicados sobre el conjunto de datos de validación y quedan recogidos en el anexo 4.

La mayor bondad de ajuste queda recogida en el modelo 1, con un 94.74% de clientes fraudulentos clasificados correctamente y un pronóstico de 50000 visitas. Analizado para 100000 visitas (modelo 21), la bondad de ajuste se reduce al 90%, por lo que se considerará este modelo bastante robusto y poco sensible a cambios.

Según el menor número de nodos, se partirá de todos aquellos modelos que presenten un mínimo de tres nodos. Los mejores modelos en base a este criterio corresponden a los modelos 14, 28, 44 y 45. Dado que todos estos modelos pronostican entre 50000 y 100000 visitas, se

analizan los mismos para el número visitas complementario, donde se obtienen diferencias entre el 5-10% en bondad de ajuste.

Revisados todos los datos, se comprueba como el modelo 25 presenta la mayor robustez, pues formado con 4 nodos y pronosticando 100000 visitas a realizar, ofrece un número intermedio de nodos entre los modelos indicados anteriormente y una pérdida insignificante en la bondad de ajuste. Por tanto, se considerará este como el modelo óptimo para este grupo de variables.

- **Función de activación:** Exponencial
- **Nodos:** 4
- **Variables:** 14
- **Visitas:** 100000
- **Bondad de ajuste:** 89.58%

7.3.2. VARIABLES REGRESION LOGISTICA FORWARD

Para este segundo bloque de variables, se hará uso de las variables que fueron seleccionadas para el mejor modelo obtenido en Regresión Logística (Forward). Dispondremos, por tanto, de 9 variables que serán: *Año Fabricación Contador, Prob_Local2_30, Ultima_Visita_6, Tarifa30, PercentEstSuminis12_Estado_1, NumeroGrupoAcc12_5 y Descripcion_Ubic_Cont30.*

De nuevo, y dado que se mantiene el uso de las mismas funciones de activación y de nodos, el número de modelos que se obtienen para evaluar en el conjunto de validación será de 64, los cuales quedan recogidos en el anexo 5.

En base a los criterios definidos en la metodología, los mejores modelos serían los siguientes:

- El modelo 1 presenta una mayor **bondad de ajuste**, situada en un 87.4%. Dicho modelo está compuesto por una función de activación Elliott y 7 nodos y plantearía la realización de 50000 visitas. Para 100000 visitas, la bondad de ajuste quedaría reducida en 5 puntos (modelo 18).
- En base al **criterio de Ockam**, nuevamente los modelos que presentarán menor números de nodos son aquellos que tengan tres nodos en capa oculta. El mejor modelo en base a este criterio corresponde al número 9, con bondad de ajuste del 84,87%, y función de activación Exponencial. El mismo, plantea la realización de 50000 visitas, por lo que, analizada la robustez, vemos como la bondad cae hasta el 79.46% para 100000 visitas (modelo 45).

Por todo lo dispuesto, y al mezclar ambos criterios, se considera como óptimo el modelo 45, ya que al ser aplicado sobre el conjunto de validación ofrece mejores resultados y cuyas características se muestran a continuación:

- **Función de activación:** Exponencial
- **Nodos:** 3
- **Variables:** 7
- **Visitas:** 100000
- **Bondad de ajuste:** 79.46%

7.3.3. VARIABLES REGRESION LOGISTICA STEPWISE

El último bloque incluye el conjunto de variables que fueron obtenidas para el mejor modelo obtenido en Regresión Logística mediante el método Stepwise. Dichas variables son las siguientes: *Año Fabricación Contador2, Prob_Local2_30, Ultima_Visita_6, Tarifa30, PercentEstSuminis12_Estado_1, NumeroGrupoAcc12_5, GrupoCont30, Numero_Facturas_24 y Descripcion_Ubic_Cont30.*

En base a las combinaciones de nodos, funciones de activación y visitas se obtuvieron un total de 64 modelos, que fueron evaluados sobre el conjunto de validación, obteniendo los resultados que se recogen en el anexo 6, que presentará nuevamente la función de activación, la clasificación realizada sobre los individuos (VP, VN, FP, FN), el número de visitas a ejecutar, la bondad de ajuste del modelo y el número de nodos de que consta.

- En base a la mayor **bondad de ajuste**, el mejor modelo corresponde al modelo 1, compuesto por 6 nodos, función de activación arco tangente y 50000 visitas. Establece el porcentaje de clientes con irregularidades clasificado correctamente en el 91.82%. Evaluada la robustez del mismo, la bondad de ajuste para 100000 visitas queda establecida en el 87.42% (modelo 22).
- Atendiendo al **criterio de parsimonia**, los modelos más simples serán aquellos que presenten un menos número de nodos. En este caso, el menor número de nodos contemplado es 3, que se presenta en el modelo 8 (bondad de ajuste del 89.24%). Al aumentar las visitas, vemos como el ajuste se reduce al 85.56% para 100000 visitas (modelo 42).

Se consideran conjuntamente todos los criterios atendiendo también a la robustez de los mismos y se observa como el modelo 42 contemplado para 100000 visitas ofrece una reducción a la mitad del número de nodos con pérdidas mínimas en la bondad de ajuste y, por tanto, queda seleccionado como modelo óptimo para este grupo de variables el modelo 1.

- **Función de activación:** Exponencial
- **Nodos:** 3
- **Variables:** 9
- **Visitas:** 100000
- **Bondad de ajuste:** 85.56%

7.3.4. MODELO OPTIMO REDES NEURONALES

Al igual que se ha realizado en el caso de Regresión Logística, se debe elegir el modelo óptimo entre los mejores modelos obtenidos para los dos grupos de variables que se han usado en Redes Neuronales tras ser aplicados sobre el conjunto de validación. Los resultados han sido los siguientes:

Obs	f_activacion	VP	FP	FN	VN	visitas	Bondad_Ajuste	nodos1	variables	Grupo
1	EXP	172	20	213	415	100000	0.89583	4	14	Backward
2	EXP	147	38	238	397	100000	0.79459	3	7	Forward
3	EXP	160	27	225	408	100000	0.85561	3	9	Stepwise

Tabla 13.- Mejores Modelos Redes Neuronales

En base a los criterios que se usan para analizar modelos, la mayor **bondad de ajuste** se encuentra en el modelo correspondiente a las variables del mejor modelo Backward. Sin embargo, atendiendo al criterio de parsimonia, el modelo más simple según el menor número de nodos y variables corresponde al obtenido con las variables del modelo Forward.

Sin embargo, el modelo obtenido con las variables del modelo Stepwise ofrece valores intermedios para todos los parámetros analizados, y dado que se considera importante la robustez del modelo, la perdida en bondad de ajuste es mínima y se reduce considerablemente el número de variables, se concluye que el modelo óptimo para Redes Neuronales será el obtenido con las variables pertenecientes a las obtenidas mediante el método Stepwise.

Se procede a analizar dicho modelo a través de su matriz de confusión:

		Predicción	
		Fraude	**No Fraude**
Observación	**Fraude**	160	225
	No Fraude	36	399

- **Sensibilidad:**

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{160}{160 + 225} = 0.4156$$

El 41.56% de los clientes que están cometiendo alguna irregularidad serán detectados.

- **Especificidad:**

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}} = \frac{399}{399 + 36} = 0.9172$$

El 91.72% de los clientes que no cometen irregularidades será clasificado correctamente.

- **Valor Predictivo Positivo:**

$$\text{VPP} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{160}{160 + 36} = 0.8163$$

El 81.63% de los clientes clasificados como fraudulentos estará cometiendo alguna irregularidad.

- **Valor Predictivo Negativo:**

$$\text{VPN} = \frac{\text{VN}}{\text{VN} + \text{FN}} = \frac{399}{399 + 225} = 0.6394$$

El 63.94% de los clientes que el modelo clasifica como no fraudulentos los estará correctamente.

Por último, se detalla que se adjuntará la estimación de parámetros realizada para este modelo en el anexo 7. Indicar que los parámetros definidos como BIAS corresponderán a los sesgos, mientras que los parámetros H1 a H6 indicarán la entrada de cada nodo en el que se aglutina las entradas de la capa anterior teniendo en cuenta la función de activación.

7.4. RANDOM FOREST

El último bloque de modelos predictivos que se creará es el correspondiente a la técnica Random Forest. Se comenzará recordando que al igual que en la Regresión Logística, y a diferencia de Redes Neuronales Artificiales, Random Forest si realiza selección de variables, por lo que se partirá de modelos que incluirán todas las variables y esta técnica se encargará de eliminar aquellas que no sean significativas.

Los parámetros establecidos han sido los siguientes:

- **Alpha:** Fijará el criterio para dividir los nodos. Cuando sea inferior al valor establecido, el nodo padre quedará divido en dos subnodos, quedando intacto en caso contrario. Los diferentes valores establecidos para alpha son: **0.05, 0.001, 0.0001, 0.00001, 0.000001, 0.000000001 y 0.00000000001**.
- **Leafsize:** Corresponderá al tamaño mínimo para la hoja final. Se establecen también distintos valores: **1000, 700, 500, 100 y 50.**
- **Variables:** Selección de variables que se realizara sobre el conjunto inicial de variables para introducir en el modelo. Tomará los valores **20, 50 y 80.**
- **Maxtrees:** Número máximo de árboles a usar. Tomará un valor fijo de 500.
- **Visitas:** Al igual que en resto de técnicas planteadas, las visitas pronosticadas a realizar tomarán los valores **50000 y 100000.**

- **Trainfraction:** Porcentaje de datos de entrenamiento utilizados. Establecemos dicho valor en **1**, ya que la base de datos fue previamente dividida en el conjunto de datos de entrenamiento y el de validación y, por tanto, se hará uso del primero al completo.
- **Maxdepth:** Profundidad máxima del árbol. Se establece un valor fijo de **50**.
- **Numvariables:** Numero inicial de variables de que disponemos. Toma el valor **180**.
- **Maxbranch:** Establece el número máximo de ramas en que podremos dividir cada nodo. Toma el valor fijo **2.**

La combinación de dichos parámetros llevará a obtener un total de 315 modelos, entre los que se elegirá el modelo que se considere óptimo en base a los criterios definidos en nuestra metodología al aplicarlo sobre el conjunto de datos de validación.

7.4.1. MODELO OPTIMO RANDOM FOREST

En el caso estudiado solo se dispone de un método de selección de variables, por lo que el modelo óptimo que se seleccione con esta técnica será el definitivo para la misma.

Se resumen los mejores modelos obtenidos en el anexo 8 para realizar la selección del mejor modelo para Random Forest. En dicha tabla aparecerá una nueva variable que toma el nombre **Numrules** y que indicará el número de divisiones a realizar en cada uno de los modelos:

- Ordenados por mayor **bondad de ajuste**, se puede ver como todos los modelos con 100000 visitas son los que obtienen un mayor valor, el cual clasificaría correctamente el 31.13% de los clientes que comenten alguna irregularidad.
- Según el criterio de parsimonia, los modelos más sencillos serán aquellos que presenten un mayor valor para alpha (se realizarán menos divisiones de nodos), un mayor tamaño de hoja y un menor número de variables. Estudiando dichos valores se comprueba como el modelo 1 presenta un alpha de 0.05, 38 divisiones y 20 variables, con una bondad de ajuste del 31.13% y 100000 visitas pronosticadas para realizar.

Se comprueba como el modelo 1 dispone de 100000 visitas y, en base a la combinación de los criterios descritos, será considerado el modelo óptimo para Random Forest.

Se procede a realizar el análisis del mismo:

		Predicción	
		Fraude	**No Fraude**
Observación	**Fraude**	33	352
	No Fraude	73	362

- **Sensibilidad:**

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{33}{33 + 352} = 0.0857$$

El 8.57% de los clientes que están cometiendo alguna irregularidad serán detectados.

- **Especificidad:**

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}} = \frac{362}{362 + 73} = 0.8322$$

El 83.22% de los clientes que no cometen irregularidades será clasificado correctamente.

- **Valor Predictivo Positivo:**

$$\text{VPP} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{33}{33 + 73} = 0.3113$$

El 31.13% de los clientes clasificados como fraudulentos estará cometiendo alguna irregularidad.

- **Valor Predictivo Negativo:**

$$\text{VPN} = \frac{\text{VN}}{\text{VN} + \text{FN}} = \frac{362}{362 + 352} = 0.5070$$

El 50.70% de los clientes que el modelo clasifica como no fraudulentos los estará correctamente.

Dado que esta técnica no presenta posibilidades de interpretación puesto que no se obtienen parámetros, se resumen en la tabla 14 la importancia de las variables que se usan para generar este modelo que se ha considerado como óptimo, donde, por tanto, se podrán ver cuáles son las variables más influyentes en la variable respuesta:

Obs	Variable	NRules	Gini	Margin
1	ProbLocal2_100	70	0.0103202	0.020640
2	ProbLocal2_30	63	0.0092882	0.018576
3	A_O_FABRICACION_CONTADOR2	50	0.0057393	0.011479
4	ProbLocal1_100	33	0.0016651	0.003330
5	ProbLocal1_30	22	0.0012382	0.002476
6	Tarifa30	23	0.0005943	0.001189
7	Tarifa100	22	0.0005684	0.001137
8	Termino_saltoOpc2b	28	0.0005488	0.001098
9	Termino_saltoOcp1	23	0.0002007	0.000401
10	Termino_saltoOpc2a	16	0.0001784	0.000357

11	Max_salto_2a	18	0.0001592	0.000318
12	VarSuya7	20	0.0001403	0.000281
13	Porcentaje0_24	17	0.0001160	0.000232
14	VarSuya1	13	0.0000912	0.000182
15	Lugar_Max_Salto_1	12	0.0000610	0.000122
16	media6Opc1	14	0.0000387	0.000077
17	media12Opc1	8	0.0000160	0.000032
18	Max_salto_1	5	0.0000139	0.000028
19	desviacion12Opc1	7	0.0000116	0.000023
20	desviacion24Opc1	7	0.0000111	0.000022
21	Max_salto_0	6	0.0000092	0.000018
22	Lugar_Max_Salto_2a	2	0.0000041	0.000008
23	Porcentaje0_12	3	0.0000032	0.000006
24	Lugar_Max_Salto_0	2	0.0000029	0.000006
25	media24Opc1	9	0.0000017	0.000003
26	Lugar_Max_Salto_2b	1	0.0000017	0.000003
27	Porcentaje0_6	3	0.0000010	0.000002
28	Termino_salto_0	1	0.0000003	0.000001
29	Numero0_12	1	0.0000003	0.000001
30	Numero0_24	1	0.0000003	0.000001

Tabla 14.- Importancia Mejor Modelo Random Forest

7.5. SELECCIÓN MEJOR MODELO

Como se ha visto en el capítulo anterior, se ha seleccionado el modelo óptimo para cada una de las tres técnicas desarrolladas, Regresión Logística, Redes Neuronales y Random Forest, tras su aplicación sobre el conjunto de datos de validación. El desarrollo de este apartado será breve, pues únicamente se tendrá que realizar la comparación entre los modelos seleccionados y elegir el que será el modelo definitivo.

Comenzamos rememorando los mejores modelos obtenidos para cada una de las tres técnicas:

Obs	VP	FP	FN	VN	visitas	num_parametros	Bondad_Ajuste	Funcion	Metodo	P_Entrada
1	74	12	311	423	100000	7	0.86047	Logit	FORWARD	.000000001

Tabla 15.- Modelo Óptimo Regresión Logística

Obs	f_activacion	VP	FP	FN	VN	visitas	Bondad_Ajuste	nodos1	variables	Grupo
1	EXP	160	27	225	408	100000	0.85561	3	9	Stepwise

Tabla 16.- Modelo Óptimo Redes Neuronales

Obs	VP	FP	FN	VN	Numrules	visitas	Bondad_Ajuste	vars_to_try	leafsize	alpha
1	33	73	352	362	38	100000	0.31132	20	1000	0.05

Tabla 17.- Modelo Óptimo Random Forest

A la vista de los resultados, se puede ver como los valores que ofrece Random Forest son bastante pobres, por lo que, directamente, se podrá descartar dicha técnica para la creación de nuestro modelo.

Como se viene realizando, el primer criterio en que fijarse es el de la **bondad de ajuste**, donde se puede ver como claramente es mayor el valor correspondiente al modelo que hemos obtenido con Regresión Logística, aunque con Redes Neuronales se obtienen valores similares.

Si realizamos la elección mediante la **razón de Ockam** es debido quedarse con aquel modelo que haga uso de un menor número de variables y que en este caso corresponde también a la Regresión Logística.

Un punto a favor de esta técnica también sería la interpretabilidad del modelo, ya que es mucho más sencillo realizar una regresión e interpretarla que cualquiera de los modelos que se hayan obtenido mediante Redes Neuronales o Random Forest.

Sin embargo, se considera muy importante la diferencia en el dato obtenido para la sensibilidad, donde el modelo obtenido por Redes Neuronales detectará un más de un 20% de usuarios con irregularidades.

Por tanto, y dado que el objetivo principal del trabajo es la detección del fraude, se considera el modelo de Redes Neuronales como modelo óptimo.

Características del modelo óptimo:

- **Técnica Predictiva:** Redes Neuronales Artificiales
- **Función de Activación:** Exponencial
- **Número de Nodos:** 3
- **Visitas Pronosticadas:** 100000
- **Numero de Variables:** 9
- **Sensibilidad:** 0.41558
- **Especificidad:** 0.93793
- **Bondad de Ajuste (Valor Predictivo Positivo):** 0.8163
- **Valor Predictivo Negativo:** 0.6394

Por último, recordamos las variables que formarán parte de este modelo: *Año Fabricación Contador, Prob_Local2_30, Ultima_Visita_6, Tarifa30, PercentEstSuminis12_Estado_1, NumeroGrupoTipo12_5, GrupoCont30, Numero_Facturas_24 y Descripcion_Ubic_Cont30.*

8. CONCLUSIONES

A pesar de que las empresas continuamente tratan de mejor las medidas de detección de fraude, el ingenio de los usuarios para cometer irregularidades aumenta, por lo que el riesgo de cualquier empresa para ser estafada posee un valor bastante alto.

La detección del fraude en una empresa de consumo no es algo sencillo, ya que este puede aparecer por diversas vías, por lo que hemos intentado crear un patrón de trabajo que, independientemente de cual sea dicha vía, permita detectar cualquier tipo de irregularidad con la finalidad de solventarla.

El panorama actual, en el que se considera el agua un bien preciado, sumado a las continuas épocas de sequía que sufre nuestro país, hacen que la avaricia de muchas personas vaya en aumento y les lleve a incurrir en dichos comportamientos irregularidades.

Para la prevención de dichas irregularidades nuestro estudio se ha centrado en la consecución de un modelo predictivo que permita adelantarnos al usuario que ha tomado la determinación de cometer cualquier tipo de fraude.

Se han desarrollado 528 modelos divididos en 126 para Regresión Logística, 192 para Redes Neuronales Artificiales y 210 para Random Forest.

De estas técnicas, se han obtenido las mejores clasificaciones para Regresión Logística y Redes Neuronales, alcanzando en ambos casos el 85% de usuarios fraudulentos correctamente clasificados. Sin embargo, a pesar de las múltiples combinaciones realizadas para Random Forest, los resultados obtenidos para dicha técnica fueron bastante pobres, posiblemente motivado por el pequeño tamaño del conjunto de datos de entrenamiento.

Finalmente, se ha optado por el modelo obtenido mediante Redes Neuronales, el cual presenta un porcentaje del 81.63% de clientes fraudulentos correctamente clasificados, sensibilidad de 41.56% y especificidad del 93.79%.

No obstante, para llegar a esta conclusión ha sido necesario el cumplimiento de otros objetivos planteados en este estudio como secundarios y que eran el procesado y almacenamiento de datos, la depuración de los mismos o la generación de nuevas variables.

Se destaca también que, para mejorar la capacidad predictiva de estos modelos, el conjunto de datos fue dividido en dos grandes bloques: el conjunto de entrenamiento y el de validación, siendo este último el conjunto utilizado como referente para la selección de modelos.

Por tanto, dado que el objetivo principal era la construcción de un modelo predictivo capaz de detectar los usuarios que comenten irregularidades en el seno de una compañía de servicios de agua, se puede considerar que el objetivo ha sido cumplido.

Por último, se debe reseñar que el modelo obtenido ha sido calculado en base a los datos de que se disponían, por lo que será necesario ir actualizando los datos mientras que se ha intentado obtener un modelo robusto para que no fuese necesario tener que actualizar los parámetros periódicamente.

9. BIBLIOGRAFÍA

Acuña, G. (2020). ¿Qué son las Redes Neuronales Artificiales que pretenden emular la inteligencia humana? *La Tercera*.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* New Jersey: Editorial John Wiley & Sons.

Basogain Olabe, X. (2014). *Redes Neuronales Artificiales y sus aplicaciones.* Bilbao: Escuela Superior de Ingeniería de Bilbao.

Benítez, R., Escudero, G., & Kanaan, S. (2013). *Inteligencia Artificial Avanzada.* Barcelona: Universitat Oberta de Catalunya.

Berzal, F. (2018). *Redes Neuronales & Deep Learning.* Granada.

Breiman, L. (2001). *Random Forests.* Berkeley: University of California.

Cody, R. (2017). *Cody's Data Cleaning Techniques Using SAS.* Cary: SAS Institute Inc.

Gündüz, N., & Fokoué, E. (2015). *On the Predictive Properties of Binary Link Functions.* Ankara.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning.* Stanford: Ed. Springer.

Haykin, S. (1999). *Neural Networks and Learning Machines.* Ontario: Editorial Pearson.

Lacey, A. (1976). *A Dictionary of Philosophy.* Londres: Editorial Routledge.

López Roldán, P., & Fachelli, S. (2015). *Metodología de la Investigación Social Cuantitativa.* Barcelona: Universidad Autónoma de Barcelona.

MINISTERIO PARA LA TRANCISIÓN ECOLÓGICA Y EL RETO DEMOGRAFICO. (2020). *PLAN HIDROLOGICO NACIONAL.*

Montgomery, D. C. (2006). *Introducción al Análisis de Regresión Lineal.* México: Editorial Continental.

Sarle, W. (1994). Neural Networks and Statistical Models., (págs. 1-13).

Serrano, A., Soria, E., & Martín, J. (2009). *Redes Neuronales Artificiales.* Valencia: Escuela Técnica Superior de Ingenieria - Universidad de Valencia.

Shalev Shwartz, S., & Ben David, S. (2014). *Understanding Machine Learning.* Jerusalén / Waterloo: Cambridge University.

Villán Criado, I., & García Rubio, E. (1995). *La depuracion de datos estadísticos.* Cuadernos aragoneses de economía.

ANEXO 1. MODELOS REGRESIÓN LOGÍSTICA METODO BACKWARD

Obs	VP	FP	FN	VN	visitas	num_parametros	Bondad_Ajuste	Funcion	Metodo	P_Entrada
1	52	7	333	428	50000	19	0.88136	Cloglog	Backward	.00100000000
2	56	8	329	427	50000	12	0.87500	Cloglog	Backward	.00001000000
3	55	8	330	427	50000	17	0.87302	Probit	Backward	.00010000000
4	49	8	336	427	50000	14	0.85965	Cloglog	Backward	.00010000000
5	60	10	325	425	50000	14	0.85714	Logit	Backward	.00001000000
6	53	9	332	426	50000	15	0.85484	Logit	Backward	.00010000000
7	82	14	303	421	100000	19	0.85417	Cloglog	Backward	.00100000000
8	80	14	305	421	100000	12	0.85106	Cloglog	Backward	.00001000000
9	78	14	307	421	100000	14	0.84783	Cloglog	Backward	.00010000000
10	48	9	337	426	50000	11	0.84211	Probit	Backward	.00000100000
11	90	17	295	418	100000	14	0.84112	Logit	Backward	.00001000000
12	51	10	334	425	50000	10	0.83607	Logit	Backward	.00000100000
13	88	18	297	417	100000	15	0.83019	Logit	Backward	.00010000000
14	88	18	297	417	100000	17	0.83019	Probit	Backward	.00010000000
15	71	15	314	420	100000	11	0.82558	Probit	Backward	.00000100000
16	80	17	305	418	100000	11	0.82474	Cloglog	Backward	.00000100000
17	55	12	330	423	50000	12	0.82090	Probit	Backward	.00001000000
18	77	17	308	418	100000	10	0.81915	Logit	Backward	.00000100000
19	54	12	331	423	50000	24	0.81818	Probit	Backward	.00100000000
20	49	11	336	424	50000	11	0.81667	Cloglog	Backward	.00000100000
21	30	7	355	428	50000	5	0.81081	Cloglog	Backward	.00000000100
22	30	7	355	428	50000	5	0.81081	Cloglog	Backward	.00000000001
23	51	12	334	423	50000	56	0.80952	Cloglog	Backward	.05000000000
24	29	7	356	428	50000	5	0.80556	Probit	Backward	.00000000001
25	94	23	291	412	100000	24	0.80342	Probit	Backward	.00100000000
26	69	17	316	418	50000	19	0.80233	Logit	Backward	.00100000000
27	84	21	301	414	100000	12	0.80000	Probit	Backward	.00001000000
28	31	8	354	427	50000	5	0.79487	Logit	Backward	.00000000001
29	72	19	313	416	50000	57	0.79121	Logit	Backward	.05000000000
30	64	17	321	418	50000	40	0.79012	Probit	Backward	.05000000000
31	45	12	340	423	50000	9	0.78947	Probit	Backward	.00000000100
32	54	15	331	420	100000	5	0.78261	Probit	Backward	.00000000001
33	70	20	315	415	100000	9	0.77778	Probit	Backward	.00000000100
34	101	29	284	406	100000	19	0.77692	Logit	Backward	.00100000000
35	79	23	306	412	100000	9	0.77451	Logit	Backward	.00000000100
36	80	24	305	411	100000	56	0.76923	Cloglog	Backward	.05000000000
37	52	16	333	419	100000	5	0.76471	Logit	Backward	.00000000001
38	51	16	334	419	100000	5	0.76119	Cloglog	Backward	.00000000100

39	51	16	334	419	100000	5	0.76119	Cloglog	Backward	.00000000001
40	52	17	333	418	50000	9	0.75362	Logit	Backward	.00000000100
41	103	34	282	401	100000	57	0.75182	Logit	Backward	.05000000000
42	92	32	293	403	100000	40	0.74194	Probit	Backward	.05000000000

ANEXO 2. MODELOS REGRESIÓN LOGÍSTICA METODO FORWARD

Obs	VP	FP	FN	VN	visitas	num_parametros	Bondad_Ajuste	Funcion	Metodo	P_Entrada
1	50	7	335	428	50000	13	0.87719	Cloglog	FORWARD	.00001000000
2	49	7	336	428	50000	21	0.87500	Cloglog	FORWARD	.00100000000
3	55	8	330	427	50000	11	0.87302	Logit	FORWARD	.00001000000
4	55	8	330	427	50000	9	0.87302	Probit	FORWARD	.00000100000
5	54	8	331	427	50000	9	0.87097	Logit	FORWARD	.00000100000
6	54	8	331	427	50000	7	0.87097	Probit	FORWARD	.00000000100
7	53	8	332	427	50000	7	0.86885	Logit	FORWARD	.00000000100
8	46	7	339	428	50000	32	0.86792	Cloglog	FORWARD	.05000000000
9	78	12	307	423	100000	21	0.86667	Cloglog	FORWARD	.00100000000
10	57	9	328	426	50000	16	0.86364	Logit	FORWARD	.00100000000
11	57	9	328	426	50000	16	0.86364	Logit	FORWARD	.00010000000
12	74	12	311	423	100000	7	0.86047	Logit	FORWARD	.00000000100
13	42	7	343	428	50000	12	0.85714	Cloglog	FORWARD	.00000100000
14	76	13	309	422	100000	11	0.85393	Logit	FORWARD	.00001000000
15	56	10	329	425	50000	5	0.84848	Probit	FORWARD	.00000000001
16	50	9	335	426	50000	9	0.84746	Cloglog	FORWARD	.00000000100
17	76	14	309	421	100000	9	0.84444	Probit	FORWARD	.00000100000
18	54	10	331	425	50000	5	0.84375	Cloglog	FORWARD	.00000000001
19	75	14	310	421	100000	9	0.84270	Logit	FORWARD	.00000100000
20	69	13	316	422	100000	13	0.84146	Cloglog	FORWARD	.00001000000
21	53	10	332	425	50000	5	0.84127	Logit	FORWARD	.00000000001
22	68	13	317	422	100000	12	0.83951	Cloglog	FORWARD	.00000100000
23	47	9	338	426	50000	31	0.83929	Logit	FORWARD	.05000000000
24	73	14	312	421	100000	9	0.83908	Cloglog	FORWARD	.00000000100
25	78	15	307	420	100000	18	0.83871	Cloglog	FORWARD	.00010000000
26	55	11	330	424	50000	35	0.83333	Probit	FORWARD	.05000000000
27	45	9	340	426	50000	18	0.83333	Cloglog	FORWARD	.00010000000
28	74	15	311	420	100000	7	0.83146	Probit	FORWARD	.00000000100
29	78	16	307	419	100000	16	0.82979	Logit	FORWARD	.00100000000
30	78	16	307	419	100000	16	0.82979	Logit	FORWARD	.00010000000
31	82	17	303	418	100000	32	0.82828	Cloglog	FORWARD	.05000000000
32	59	13	326	422	50000	11	0.81944	Probit	FORWARD	.00001000000
33	90	20	295	415	100000	11	0.81818	Probit	FORWARD	.00001000000
34	79	18	306	417	100000	21	0.81443	Probit	FORWARD	.00100000000
35	48	11	337	424	50000	21	0.81356	Probit	FORWARD	.00100000000
36	81	19	304	416	100000	14	0.81000	Probit	FORWARD	.00010000000
37	74	18	311	417	100000	5	0.80435	Probit	FORWARD	.00000000001
38	53	13	332	422	50000	14	0.80303	Probit	FORWARD	.00010000000

39	73	18	312	417	100000	5	0.80220	Logit	FORWARD	.00000000001
40	72	19	313	416	100000	5	0.79121	Cloglog	FORWARD	.00000000001
41	79	22	306	413	100000	35	0.78218	Probit	FORWARD	.05000000000
42	74	22	311	413	100000	31	0.77083	Logit	FORWARD	.05000000000

ANEXO 3. MODELOS REGRESIÓN LOGÍSTICA METODO STEPWISE

Obs	VP	FP	FN	VN	visitas	num_parametros	Bondad_Ajuste	Funcion	Metodo	P_Entrada
1	55	8	330	427	50000	10	0.87302	Logit	STEPWISE	.00010000000
2	55	8	330	427	50000	10	0.87302	Logit	STEPWISE	.00001000000
3	55	8	330	427	50000	9	0.87302	Probit	STEPWISE	.00001000000
4	55	8	330	427	50000	9	0.87302	Probit	STEPWISE	.00000100000
5	54	8	331	427	50000	9	0.87097	Logit	STEPWISE	.00000100000
6	54	8	331	427	50000	7	0.87097	Probit	STEPWISE	.00000000100
7	76	12	309	423	100000	10	0.86364	Logit	STEPWISE	.00010000000
8	76	12	309	423	100000	10	0.86364	Logit	STEPWISE	.00001000000
9	57	9	328	426	50000	16	0.86364	Logit	STEPWISE	.00100000000
10	57	10	328	425	50000	31	0.85075	Probit	STEPWISE	.05000000000
11	51	9	334	426	50000	25	0.85000	Cloglog	STEPWISE	.05000000000
12	56	10	329	425	50000	5	0.84848	Probit	STEPWISE	.00000000001
13	50	9	335	426	50000	9	0.84746	Cloglog	STEPWISE	.00100000000
14	50	9	335	426	50000	9	0.84746	Cloglog	STEPWISE	.00010000000
15	50	9	335	426	50000	9	0.84746	Cloglog	STEPWISE	.00001000000
16	50	9	335	426	50000	9	0.84746	Cloglog	STEPWISE	.00000100000
17	76	14	309	421	100000	9	0.84444	Probit	STEPWISE	.00001000000
18	76	14	309	421	100000	9	0.84444	Probit	STEPWISE	.00000100000
19	75	14	310	421	100000	9	0.84270	Logit	STEPWISE	.00000100000
20	48	9	337	426	50000	13	0.84211	Probit	STEPWISE	.00010000000
21	53	10	332	425	50000	5	0.84127	Logit	STEPWISE	.00000000100
22	53	10	332	425	50000	5	0.84127	Logit	STEPWISE	.00000000001
23	73	14	312	421	100000	9	0.83908	Cloglog	STEPWISE	.00100000000
24	73	14	312	421	100000	9	0.83908	Cloglog	STEPWISE	.00010000000
25	73	14	312	421	100000	9	0.83908	Cloglog	STEPWISE	.00001000000
26	73	14	312	421	100000	9	0.83908	Cloglog	STEPWISE	.00000100000
27	72	14	313	421	100000	13	0.83721	Probit	STEPWISE	.00010000000
28	51	10	334	425	50000	33	0.83607	Logit	STEPWISE	.05000000000
29	86	17	299	418	100000	25	0.83495	Cloglog	STEPWISE	.05000000000
30	74	15	311	420	100000	7	0.83146	Probit	STEPWISE	.00000000100
31	78	16	307	419	100000	16	0.82979	Logit	STEPWISE	.00100000000
32	81	19	304	416	100000	14	0.81000	Probit	STEPWISE	.00100000000
33	74	18	311	417	100000	5	0.80435	Probit	STEPWISE	.00000000001
34	53	13	332	422	50000	14	0.80303	Probit	STEPWISE	.00100000000
35	73	18	312	417	100000	5	0.80220	Logit	STEPWISE	.00000000100
36	73	18	312	417	100000	5	0.80220	Logit	STEPWISE	.00000000001
37	83	21	302	414	100000	31	0.79808	Probit	STEPWISE	.05000000000
38	83	22	302	413	100000	33	0.79048	Logit	STEPWISE	.05000000000

39	29	10	356	425	50000	2	0.74359	Cloglog	STEPWISE	.00000000100
40	29	10	356	425	50000	2	0.74359	Cloglog	STEPWISE	.00000000001
41	44	20	341	415	100000	2	0.68750	Cloglog	STEPWISE	.00000000100
42	44	20	341	415	100000	2	0.68750	Cloglog	STEPWISE	.00000000001

ANEXO 4. MODELOS REDES NEURONALES VARIABLES BACKWARD

Obs	f_activacion	VP	FP	FN	VN	visitas	Bondad_Ajuste	nodos1
1	ARC	108	6	277	429	50000	0.94737	6
2	TAN	89	5	296	430	50000	0.94681	6
3	ELL	108	7	277	428	50000	0.93913	6
4	TAN	109	8	276	427	50000	0.93162	7
5	ARC	101	8	284	427	50000	0.92661	8
6	TAN	86	7	299	428	50000	0.92473	8
7	TAN	69	6	316	429	50000	0.92000	2
8	TAN	126	11	259	424	50000	0.91971	4
9	EXP	110	10	275	425	50000	0.91667	4
10	TAN	107	10	278	425	100000	0.91453	2
11	EXP	139	13	246	422	50000	0.91447	7
12	EXP	126	12	259	423	50000	0.91304	5
13	ARC	80	8	305	427	50000	0.90909	5
14	TAN	59	6	326	429	50000	0.90769	3
15	TAN	143	15	242	420	100000	0.90506	6
16	ARC	95	10	290	425	50000	0.90476	7
17	ELL	95	10	290	425	50000	0.90476	8
18	ELL	85	9	300	426	50000	0.90426	5
19	TAN	112	12	273	423	50000	0.90323	5
20	ELL	148	16	237	419	100000	0.90244	5
21	ARC	144	16	241	419	100000	0.90000	6
22	EXP	170	19	215	416	100000	0.89947	7
23	ARC	141	16	244	419	100000	0.89809	5
24	ARC	79	9	306	426	50000	0.89773	2
25	EXP	172	20	213	415	100000	0.89583	4
26	ELL	103	12	282	423	50000	0.89565	7
27	ELL	144	17	241	418	100000	0.89441	6
28	ELL	139	17	246	418	100000	0.89103	3
29	ARC	145	18	240	417	100000	0.88957	7
30	TAN	161	20	224	415	100000	0.88950	4
31	ELL	160	20	225	415	100000	0.88889	8
32	ELL	72	9	313	426	50000	0.88889	2
33	ELL	149	19	236	416	100000	0.88690	7
34	TAN	156	20	229	415	100000	0.88636	5
35	TAN	169	22	216	413	100000	0.88482	7
36	ELL	69	9	316	426	50000	0.88462	4
37	EXP	176	23	209	412	100000	0.88442	5
38	ELL	129	17	256	418	100000	0.88356	2

39	ARC	136	18	249	417	100000	0.88312	2
40	TAN	148	20	237	415	100000	0.88095	8
41	ARC	148	20	237	415	100000	0.88095	8
42	ARC	80	11	305	424	50000	0.87912	4
43	ARC	58	8	327	427	50000	0.87879	1
44	ARC	65	9	320	426	50000	0.87838	3
45	TAN	136	19	249	416	100000	0.87742	3
46	TAN	57	8	328	427	50000	0.87692	1
47	ELL	64	9	321	426	50000	0.87671	1
48	EXP	74	11	311	424	50000	0.87059	2
49	ARC	133	20	252	415	100000	0.86928	3
50	EXP	53	8	332	427	50000	0.86885	1
51	EXP	139	21	246	414	50000	0.86875	3
52	EXP	85	13	300	422	50000	0.86735	8
53	ELL	136	21	249	414	100000	0.86624	4
54	ARC	135	21	250	414	100000	0.86538	4
55	ELL	51	8	334	427	50000	0.86441	3
56	EXP	157	25	228	410	100000	0.86264	8
57	TAN	94	16	291	419	100000	0.85455	1
58	ARC	94	16	291	419	100000	0.85455	1
59	EXP	111	19	274	416	50000	0.85385	6
60	EXP	163	28	222	407	100000	0.85340	6
61	ELL	92	16	293	419	100000	0.85185	1
62	EXP	87	16	298	419	100000	0.84466	1
63	EXP	98	19	287	416	100000	0.83761	2
64	EXP	186	37	199	398	100000	0.83408	3

ANEXO 5. MODELOS REDES NEURONALES VARIABLES FORWARD

Obs	f_activacion	VP	FP	FN	VN	visitas	Bondad_Ajuste	nodos1
1	ELL	118	17	267	418	50000	0.87407	7
2	EXP	125	19	260	416	50000	0.86806	7
3	ELL	110	18	275	417	50000	0.85938	8
4	EXP	122	20	263	415	50000	0.85915	6
5	ARC	105	18	280	417	50000	0.85366	7
6	TAN	128	22	257	413	50000	0.85333	8
7	ARC	127	22	258	413	50000	0.85235	4
8	EXP	103	18	282	417	50000	0.85124	4
9	EXP	101	18	284	417	50000	0.84874	3
10	EXP	128	23	257	412	50000	0.84768	8
11	ARC	126	23	259	412	50000	0.84564	5
12	TAN	109	20	276	415	50000	0.84496	5
13	ARC	141	26	244	409	100000	0.84431	7
14	ARC	106	20	279	415	50000	0.84127	3
15	ARC	117	23	268	412	50000	0.83571	8
16	EXP	120	24	265	411	50000	0.83333	5
17	TAN	114	23	271	412	50000	0.83212	6
18	ELL	145	30	240	405	100000	0.82857	7
19	TAN	101	21	284	414	50000	0.82787	3
20	ARC	109	24	276	411	50000	0.81955	6
21	EXP	149	33	236	402	100000	0.81868	4
22	TAN	112	25	273	410	50000	0.81752	7
23	TAN	156	35	229	400	100000	0.81675	7
24	TAN	160	36	225	399	100000	0.81633	6
25	ELL	91	21	294	414	50000	0.81250	3
26	EXP	99	23	286	412	50000	0.81148	2
27	TAN	150	35	235	400	100000	0.81081	8
28	TAN	55	13	330	422	50000	0.80882	1
29	EXP	55	13	330	422	50000	0.80882	1
30	ARC	55	13	330	422	50000	0.80882	1
31	ELL	55	13	330	422	50000	0.80882	1
32	ELL	147	35	238	400	100000	0.80769	8
33	ELL	96	23	289	412	50000	0.80672	4
34	TAN	158	38	227	397	100000	0.80612	5
35	ARC	158	38	227	397	100000	0.80612	8
36	ARC	106	26	279	409	50000	0.80303	2
37	TAN	158	39	227	396	100000	0.80203	3
38	ARC	162	40	223	395	100000	0.80198	5

39	ARC	163	41	222	394	100000	0.79902	4
40	ELL	107	27	278	408	50000	0.79851	5
41	TAN	162	41	223	394	100000	0.79803	4
42	EXP	161	41	224	394	100000	0.79703	8
43	EXP	149	38	236	397	100000	0.79679	6
44	ARC	151	39	234	396	100000	0.79474	3
45	EXP	147	38	238	397	100000	0.79459	3
46	ELL	145	38	240	397	100000	0.79235	4
47	EXP	163	43	222	392	100000	0.79126	7
48	TAN	87	23	298	412	50000	0.79091	4
49	TAN	79	21	306	414	100000	0.79000	1
50	EXP	79	21	306	414	100000	0.79000	1
51	TAN	94	25	291	410	50000	0.78992	2
52	ELL	100	27	285	408	50000	0.78740	6
53	EXP	147	40	238	395	100000	0.78610	2
54	TAN	143	39	242	396	100000	0.78571	2
55	ELL	138	38	247	397	100000	0.78409	3
56	EXP	148	41	237	394	100000	0.78307	5
57	ARC	79	22	306	413	100000	0.78218	1
58	ELL	79	22	306	413	100000	0.78218	1
59	ELL	142	40	243	395	100000	0.78022	5
60	ARC	138	40	247	395	100000	0.77528	2
61	ARC	149	44	236	391	100000	0.77202	6
62	ELL	144	43	241	392	100000	0.77005	6
63	ELL	95	29	290	406	50000	0.76613	2
64	ELL	121	39	264	396	100000	0.75625	2

ANEXO 6. MODELOS REDES NEURONALES VARIABLES STEPWISE

Obs	f_activacion	VP	FP	FN	VN	visitas	Bondad_Ajuste	nodos1
1	ARC	101	9	284	426	50000	0.91818	6
2	TAN	114	12	271	423	50000	0.90476	6
3	TAN	53	6	332	429	50000	0.89831	2
4	ARC	149	17	236	418	100000	0.89759	7
5	EXP	52	6	333	429	50000	0.89655	2
6	EXP	137	16	248	419	50000	0.89542	4
7	ARC	109	13	276	422	50000	0.89344	7
8	EXP	141	17	244	418	50000	0.89241	3
9	EXP	152	19	233	416	50000	0.88889	5
10	TAN	55	7	330	428	50000	0.88710	1
11	ARC	55	7	330	428	50000	0.88710	1
12	ELL	55	7	330	428	50000	0.88710	1
13	EXP	149	19	236	416	50000	0.88690	7
14	ELL	109	14	276	421	50000	0.88618	5
15	TAN	131	17	254	418	50000	0.88514	8
16	ELL	121	16	264	419	50000	0.88321	7
17	ARC	128	17	257	418	50000	0.88276	8
18	ARC	150	20	235	415	100000	0.88235	3
19	EXP	172	23	213	412	100000	0.88205	7
20	ELL	153	21	232	414	100000	0.87931	7
21	ARC	129	18	256	417	50000	0.87755	5
22	ARC	139	20	246	415	100000	0.87421	6
23	TAN	118	17	267	418	50000	0.87407	5
24	EXP	55	8	330	427	50000	0.87302	1
25	ELL	164	24	221	411	100000	0.87234	3
26	EXP	172	26	213	409	100000	0.86869	5
27	ELL	105	16	280	419	50000	0.86777	3
28	ARC	162	25	223	410	100000	0.86631	8
29	EXP	155	24	230	411	100000	0.86592	4
30	ARC	89	14	296	421	50000	0.86408	3
31	EXP	133	21	252	414	50000	0.86364	6
32	ELL	113	18	272	417	50000	0.86260	8
33	TAN	161	26	224	409	100000	0.86096	8
34	EXP	86	14	299	421	100000	0.86000	2
35	EXP	165	27	220	408	100000	0.85938	6
36	ARC	165	27	220	408	100000	0.85938	5
37	TAN	116	19	269	416	50000	0.85926	7
38	ELL	116	19	269	416	50000	0.85926	4

39	ELL	114	19	271	416	50000	0.85714	6
40	TAN	149	25	236	410	100000	0.85632	6
41	ELL	143	24	242	411	100000	0.85629	5
42	EXP	160	27	225	408	100000	0.85561	3
43	TAN	94	16	291	419	100000	0.85455	2
44	TAN	152	26	233	409	100000	0.85393	5
45	ELL	76	13	309	422	100000	0.85393	1
46	TAN	160	28	225	407	100000	0.85106	4
47	ARC	74	13	311	422	100000	0.85057	1
48	TAN	73	13	312	422	100000	0.84884	1
49	EXP	73	13	312	422	100000	0.84884	1
50	TAN	128	23	257	412	50000	0.84768	4
51	EXP	128	23	257	412	50000	0.84768	8
52	ARC	103	19	282	416	50000	0.84426	2
53	TAN	167	31	218	404	100000	0.84343	7
54	ARC	118	22	267	413	50000	0.84286	4
55	ELL	96	18	289	417	50000	0.84211	2
56	ELL	147	28	238	407	100000	0.84000	4
57	EXP	157	30	228	405	100000	0.83957	8
58	ELL	145	28	240	407	100000	0.83815	8
59	ARC	157	31	228	404	100000	0.83511	4
60	ARC	154	31	231	404	100000	0.83243	2
61	ELL	147	31	238	404	100000	0.82584	6
62	TAN	89	19	296	416	50000	0.82407	3
63	TAN	130	29	255	406	100000	0.81761	3
64	ELL	134	32	251	403	100000	0.80723	2

ANEXO 7. PARÁMETROS MEJOR MODELO REDES NEURONALES

Resultados de optimización			
Parameter Estimates			
N	Parameter	Estimador	Gradient Objective Function
1	A_O_FABRICACION_CONTADOR2_H1	-0.245833	-0.000524
2	DESCRIPCION_UBIC_CONT30_H1	0.087734	0.000156
3	NumeroGrupoAcc12_5_H1	-0.157953	-0.000005297
4	Tarifa30_H1	0.176663	-0.000134
5	UltimaVisita6_H1	-0.140126	-0.000104
6	ProbLocal2_30_H1	0.107844	-0.000115
7	PercentEstSuminis12_Estado_1_H1	-0.442023	-0.000023664
8	Grupocont30_H1	-0.412014	-0.000434
9	NumeroFacturas_24_H1	-0.691291	0.000245
10	A_O_FABRICACION_CONTADOR2_H2	-0.248345	-0.000380
11	DESCRIPCION_UBIC_CONT30_H2	-0.314445	-0.000009532
12	NumeroGrupoAcc12_5_H2	0.111246	-0.000050596
13	Tarifa30_H2	-0.677189	-0.000366
14	UltimaVisita6_H2	0.069992	-0.000088971
15	ProbLocal2_30_H2	0.172584	-0.000007889
16	PercentEstSuminis12_Estado_1_H2	0.360493	0.000008222
17	Grupocont30_H2	0.454904	-0.000028535
18	NumeroFacturas_24_H2	0.381562	0.000180
19	A_O_FABRICACION_CONTADOR2_H3	7.014225	-0.000270
20	DESCRIPCION_UBIC_CONT30_H3	0.028611	0.000224
21	NumeroGrupoAcc12_5_H3	1.147578	-0.000067222
22	Tarifa30_H3	0.086237	-0.000552
23	UltimaVisita6_H3	-0.169959	0.000015341
24	ProbLocal2_30_H3	-0.488424	-0.000023042
25	PercentEstSuminis12_Estado_1_H3	8.820738	-0.000034615
26	Grupocont30_H3	2.380949	-0.000551
27	NumeroFacturas_24_H3	4.037048	0.000529
28	A_O_FABRICACION_CONTADOR2_H4	4.433783	-0.000033541
29	DESCRIPCION_UBIC_CONT30_H4	-2.513449	0.000025082
30	NumeroGrupoAcc12_5_H4	-1.666596	0.000002426
31	Tarifa30_H4	-1.561720	0.000007532
32	UltimaVisita6_H4	-2.569123	0.000012420
33	ProbLocal2_30_H4	1.260481	-0.000034668
34	PercentEstSuminis12_Estado_1_H4	0.963720	-0.000041626
35	Grupocont30_H4	4.611351	-0.000009658

36	NumeroFacturas_24_H4	-0.247704	-0.000029266
37	A_O_FABRICACION_CONTADOR2_H5	0.037456	0.000199
38	DESCRIPCION_UBIC_CONT30_H5	-0.284547	-0.000238
39	NumeroGrupoAcc12_5_H5	-0.282961	0.000014721
40	Tarifa30_H5	-0.566370	0.000278
41	UltimaVisita6_H5	-0.295362	-0.000047754
42	ProbLocal2_30_H5	-0.015888	0.000132
43	PercentEstSuminis12_Estado_1_H5	-0.029226	0.000119
44	Grupocont30_H5	-0.165017	0.000347
45	NumeroFacturas_24_H5	-0.369714	-0.000293
46	A_O_FABRICACION_CONTADOR2_H6	-1.355782	0.000061391
47	DESCRIPCION_UBIC_CONT30_H6	-0.614292	-0.000085552
48	NumeroGrupoAcc12_5_H6	1.621843	-0.000001934
49	Tarifa30_H6	-0.748784	0.000065047
50	UltimaVisita6_H6	-0.520291	0.000015840
51	ProbLocal2_30_H6	0.212592	0.000070006
52	PercentEstSuminis12_Estado_1_H6	0.884467	-0.000005055
53	Grupocont30_H6	0.743391	0.000090975
54	NumeroFacturas_24_H6	-0.218778	-0.000062155
55	BIAS_H1	-0.172233	-0.000347
56	BIAS_H2	0.738564	0.000337
57	BIAS_H3	-1.779078	0.000002545
58	BIAS_H4	-7.534732	0.000085598
59	BIAS_H5	-0.445425	-0.000258
60	BIAS_H6	1.088529	0.000000562
61	H1_Irr21	-5.804623	0.000079332
62	H2_Irr21	-5.684891	0.000065318
63	H3_Irr21	-1.726390	0.000023604
64	H4_Irr21	-1.686354	-0.000006802
65	H5_Irr21	4.645074	-0.000107
66	H6_Irr21	2.721021	0.000017333
67	BIAS_Irr21	2.216930	-0.000100

ANEXO 8. MODELOS RANDOM FOREST

Obs	VP	FP	FN	VN	Numrules	visitas	Bondad_Ajuste	vars_to_try	leafsize	alpha
1	33	73	352	362	38	100000	0.31132	20	1000	.05000000000
2	33	73	352	362	85	100000	0.31132	50	1000	.05000000000
3	33	73	352	362	134	100000	0.31132	80	1000	.05000000000
4	33	73	352	362	500	100000	0.31132	20	700	.05000000000
5	33	73	352	362	500	100000	0.31132	50	700	.05000000000
6	33	73	352	362	500	100000	0.31132	80	700	.05000000000
7	33	73	352	362	875	100000	0.31132	20	500	.05000000000
8	33	73	352	362	933	100000	0.31132	50	500	.05000000000
9	33	73	352	362	937	100000	0.31132	80	500	.05000000000
10	33	73	352	362	6293	100000	0.31132	20	100	.05000000000
11	33	73	352	362	6497	100000	0.31132	50	100	.05000000000
12	33	73	352	362	6591	100000	0.31132	80	100	.05000000000
13	33	73	352	362	12831	100000	0.31132	20	50	.05000000000
14	33	73	352	362	13109	100000	0.31132	50	50	.05000000000
15	33	73	352	362	13205	100000	0.31132	80	50	.05000000000
16	33	73	352	362	38	100000	0.31132	20	1000	.00100000000
17	33	73	352	362	85	100000	0.31132	50	1000	.00100000000
18	33	73	352	362	134	100000	0.31132	80	1000	.00100000000
19	33	73	352	362	500	100000	0.31132	20	700	.00100000000
20	33	73	352	362	500	100000	0.31132	50	700	.00100000000
21	33	73	352	362	500	100000	0.31132	80	700	.00100000000
22	33	73	352	362	875	100000	0.31132	20	500	.00100000000
23	33	73	352	362	933	100000	0.31132	50	500	.00100000000
24	33	73	352	362	937	100000	0.31132	80	500	.00100000000
25	33	73	352	362	6293	100000	0.31132	20	100	.00100000000
26	33	73	352	362	6531	100000	0.31132	50	100	.00100000000
27	33	73	352	362	6591	100000	0.31132	80	100	.00100000000
28	33	73	352	362	12831	100000	0.31132	20	50	.00100000000
29	33	73	352	362	13116	100000	0.31132	50	50	.00100000000
30	33	73	352	362	13180	100000	0.31132	80	50	.00100000000
31	33	73	352	362	38	100000	0.31132	20	1000	.00010000000
32	33	73	352	362	85	100000	0.31132	50	1000	.00010000000
33	33	73	352	362	134	100000	0.31132	80	1000	.00010000000
34	33	73	352	362	500	100000	0.31132	20	700	.00010000000
35	33	73	352	362	500	100000	0.31132	50	700	.00010000000
36	33	73	352	362	500	100000	0.31132	80	700	.00010000000
37	33	73	352	362	875	100000	0.31132	20	500	.00010000000
38	33	73	352	362	933	100000	0.31132	50	500	.00010000000

39	33	73	352	362	937	100000	0.31132	80	500	.00010000000
40	33	73	352	362	6293	100000	0.31132	20	100	.00010000000
41	33	73	352	362	6531	100000	0.31132	50	100	.00010000000
42	33	73	352	362	6580	100000	0.31132	80	100	.00010000000
43	33	73	352	362	12831	100000	0.31132	20	50	.00010000000
44	33	73	352	362	13129	100000	0.31132	50	50	.00010000000
45	33	73	352	362	13166	100000	0.31132	80	50	.00010000000
46	33	73	352	362	38	100000	0.31132	20	1000	.00001000000
47	33	73	352	362	85	100000	0.31132	50	1000	.00001000000
48	33	73	352	362	134	100000	0.31132	80	1000	.00001000000
49	33	73	352	362	500	100000	0.31132	20	700	.00001000000
50	33	73	352	362	500	100000	0.31132	50	700	.00001000000
51	33	73	352	362	500	100000	0.31132	80	700	.00001000000
52	33	73	352	362	875	100000	0.31132	20	500	.00001000000
53	33	73	352	362	933	100000	0.31132	50	500	.00001000000
54	33	73	352	362	937	100000	0.31132	80	500	.00001000000
55	33	73	352	362	6293	100000	0.31132	20	100	.00001000000
56	33	73	352	362	6531	100000	0.31132	50	100	.00001000000
57	33	73	352	362	6580	100000	0.31132	80	100	.00001000000
58	33	73	352	362	12831	100000	0.31132	20	50	.00001000000
59	33	73	352	362	13117	100000	0.31132	50	50	.00001000000
60	33	73	352	362	13146	100000	0.31132	80	50	.00001000000
61	33	73	352	362	38	100000	0.31132	20	1000	.00000100000
62	33	73	352	362	85	100000	0.31132	50	1000	.00000100000
63	33	73	352	362	134	100000	0.31132	80	1000	.00000100000
64	33	73	352	362	500	100000	0.31132	20	700	.00000100000
65	33	73	352	362	500	100000	0.31132	50	700	.00000100000
66	33	73	352	362	500	100000	0.31132	80	700	.00000100000
67	33	73	352	362	875	100000	0.31132	20	500	.00000100000
68	33	73	352	362	933	100000	0.31132	50	500	.00000100000
69	33	73	352	362	937	100000	0.31132	80	500	.00000100000
70	33	73	352	362	6293	100000	0.31132	20	100	.00000100000
71	33	73	352	362	6530	100000	0.31132	50	100	.00000100000
72	33	73	352	362	6591	100000	0.31132	80	100	.00000100000
73	33	73	352	362	12831	100000	0.31132	20	50	.00000100000
74	33	73	352	362	13117	100000	0.31132	50	50	.00000100000
75	33	73	352	362	13134	100000	0.31132	80	50	.00000100000
76	33	73	352	362	38	100000	0.31132	20	1000	.00000000100
77	33	73	352	362	85	100000	0.31132	50	1000	.00000000100
78	33	73	352	362	134	100000	0.31132	80	1000	.00000000100
79	33	73	352	362	500	100000	0.31132	20	700	.00000000100
80	33	73	352	362	500	100000	0.31132	50	700	.00000000100

81	33	73	352	362	500	100000	0.31132	80	700	.00000000100
82	33	73	352	362	875	100000	0.31132	20	500	.00000000100
83	33	73	352	362	933	100000	0.31132	50	500	.00000000100
84	33	73	352	362	937	100000	0.31132	80	500	.00000000100
85	33	73	352	362	6293	100000	0.31132	20	100	.00000000100
86	33	73	352	362	6531	100000	0.31132	50	100	.00000000100
87	33	73	352	362	6576	100000	0.31132	80	100	.00000000100
88	33	73	352	362	12778	100000	0.31132	20	50	.00000000100
89	33	73	352	362	13106	100000	0.31132	50	50	.00000000100
90	33	73	352	362	13198	100000	0.31132	80	50	.00000000100
91	33	73	352	362	38	100000	0.31132	20	1000	.00000000001
92	33	73	352	362	85	100000	0.31132	50	1000	.00000000001
93	33	73	352	362	134	100000	0.31132	80	1000	.00000000001
94	33	73	352	362	500	100000	0.31132	20	700	.00000000001
95	33	73	352	362	500	100000	0.31132	50	700	.00000000001
96	33	73	352	362	500	100000	0.31132	80	700	.00000000001
97	33	73	352	362	875	100000	0.31132	20	500	.00000000001
98	33	73	352	362	933	100000	0.31132	50	500	.00000000001
99	33	73	352	362	937	100000	0.31132	80	500	.00000000001
100	33	73	352	362	6293	100000	0.31132	20	100	.00000000001
101	33	73	352	362	6531	100000	0.31132	50	100	.00000000001
102	33	73	352	362	6591	100000	0.31132	80	100	.00000000001
103	33	73	352	362	12831	100000	0.31132	20	50	.00000000001
104	33	73	352	362	13129	100000	0.31132	50	50	.00000000001
105	33	73	352	362	13208	100000	0.31132	80	50	.00000000001
106	16	56	369	379	38	50000	0.22222	20	1000	.05000000000
107	16	56	369	379	85	50000	0.22222	50	1000	.05000000000
108	16	56	369	379	134	50000	0.22222	80	1000	.05000000000
109	16	56	369	379	500	50000	0.22222	20	700	.05000000000
110	16	56	369	379	500	50000	0.22222	50	700	.05000000000
111	16	56	369	379	500	50000	0.22222	80	700	.05000000000
112	16	56	369	379	875	50000	0.22222	20	500	.05000000000
113	16	56	369	379	933	50000	0.22222	50	500	.05000000000
114	16	56	369	379	937	50000	0.22222	80	500	.05000000000
115	16	56	369	379	6293	50000	0.22222	20	100	.05000000000
116	16	56	369	379	6497	50000	0.22222	50	100	.05000000000
117	16	56	369	379	6591	50000	0.22222	80	100	.05000000000
118	16	56	369	379	12831	50000	0.22222	20	50	.05000000000
119	16	56	369	379	13109	50000	0.22222	50	50	.05000000000
120	16	56	369	379	13205	50000	0.22222	80	50	.05000000000
121	16	56	369	379	38	50000	0.22222	20	1000	.00100000000
122	16	56	369	379	85	50000	0.22222	50	1000	.00100000000

123	16	56	369	379	134	50000	0.22222	80	1000	.00100000000
124	16	56	369	379	500	50000	0.22222	20	700	.00100000000
125	16	56	369	379	500	50000	0.22222	50	700	.00100000000
126	16	56	369	379	500	50000	0.22222	80	700	.00100000000
127	16	56	369	379	875	50000	0.22222	20	500	.00100000000
128	16	56	369	379	933	50000	0.22222	50	500	.00100000000
129	16	56	369	379	937	50000	0.22222	80	500	.00100000000
130	16	56	369	379	6293	50000	0.22222	20	100	.00100000000
131	16	56	369	379	6531	50000	0.22222	50	100	.00100000000
132	16	56	369	379	6591	50000	0.22222	80	100	.00100000000
133	16	56	369	379	12831	50000	0.22222	20	50	.00100000000
134	16	56	369	379	13116	50000	0.22222	50	50	.00100000000
135	16	56	369	379	13180	50000	0.22222	80	50	.00100000000
136	16	56	369	379	38	50000	0.22222	20	1000	.00010000000
137	16	56	369	379	85	50000	0.22222	50	1000	.00010000000
138	16	56	369	379	134	50000	0.22222	80	1000	.00010000000
139	16	56	369	379	500	50000	0.22222	20	700	.00010000000
140	16	56	369	379	500	50000	0.22222	50	700	.00010000000
141	16	56	369	379	500	50000	0.22222	80	700	.00010000000
142	16	56	369	379	875	50000	0.22222	20	500	.00010000000
143	16	56	369	379	933	50000	0.22222	50	500	.00010000000
144	16	56	369	379	937	50000	0.22222	80	500	.00010000000
145	16	56	369	379	6293	50000	0.22222	20	100	.00010000000
146	16	56	369	379	6531	50000	0.22222	50	100	.00010000000
147	16	56	369	379	6580	50000	0.22222	80	100	.00010000000
148	16	56	369	379	12831	50000	0.22222	20	50	.00010000000
149	16	56	369	379	13129	50000	0.22222	50	50	.00010000000
150	16	56	369	379	13166	50000	0.22222	80	50	.00010000000
151	16	56	369	379	38	50000	0.22222	20	1000	.00001000000
152	16	56	369	379	85	50000	0.22222	50	1000	.00001000000
153	16	56	369	379	134	50000	0.22222	80	1000	.00001000000
154	16	56	369	379	500	50000	0.22222	20	700	.00001000000
155	16	56	369	379	500	50000	0.22222	50	700	.00001000000
156	16	56	369	379	500	50000	0.22222	80	700	.00001000000
157	16	56	369	379	875	50000	0.22222	20	500	.00001000000
158	16	56	369	379	933	50000	0.22222	50	500	.00001000000
159	16	56	369	379	937	50000	0.22222	80	500	.00001000000
160	16	56	369	379	6293	50000	0.22222	20	100	.00001000000
161	16	56	369	379	6531	50000	0.22222	50	100	.00001000000
162	16	56	369	379	6580	50000	0.22222	80	100	.00001000000
163	16	56	369	379	12831	50000	0.22222	20	50	.00001000000
164	16	56	369	379	13117	50000	0.22222	50	50	.00001000000

165	16	56	369	379	13146	50000	0.22222	80	50	.00001000000
166	16	56	369	379	38	50000	0.22222	20	1000	.00000100000
167	16	56	369	379	85	50000	0.22222	50	1000	.00000100000
168	16	56	369	379	134	50000	0.22222	80	1000	.00000100000
169	16	56	369	379	500	50000	0.22222	20	700	.00000100000
170	16	56	369	379	500	50000	0.22222	50	700	.00000100000
171	16	56	369	379	500	50000	0.22222	80	700	.00000100000
172	16	56	369	379	875	50000	0.22222	20	500	.00000100000
173	16	56	369	379	933	50000	0.22222	50	500	.00000100000
174	16	56	369	379	937	50000	0.22222	80	500	.00000100000
175	16	56	369	379	6293	50000	0.22222	20	100	.00000100000
176	16	56	369	379	6530	50000	0.22222	50	100	.00000100000
177	16	56	369	379	6591	50000	0.22222	80	100	.00000100000
178	16	56	369	379	12831	50000	0.22222	20	50	.00000100000
179	16	56	369	379	13117	50000	0.22222	50	50	.00000100000
180	16	56	369	379	13134	50000	0.22222	80	50	.00000100000
181	16	56	369	379	38	50000	0.22222	20	1000	.00000000100
182	16	56	369	379	85	50000	0.22222	50	1000	.00000000100
183	16	56	369	379	134	50000	0.22222	80	1000	.00000000100
184	16	56	369	379	500	50000	0.22222	20	700	.00000000100
185	16	56	369	379	500	50000	0.22222	50	700	.00000000100
186	16	56	369	379	500	50000	0.22222	80	700	.00000000100
187	16	56	369	379	875	50000	0.22222	20	500	.00000000100
188	16	56	369	379	933	50000	0.22222	50	500	.00000000100
189	16	56	369	379	937	50000	0.22222	80	500	.00000000100
190	16	56	369	379	6293	50000	0.22222	20	100	.00000000100
191	16	56	369	379	6531	50000	0.22222	50	100	.00000000100
192	16	56	369	379	6576	50000	0.22222	80	100	.00000000100
193	16	56	369	379	12778	50000	0.22222	20	50	.00000000100
194	16	56	369	379	13106	50000	0.22222	50	50	.00000000100
195	16	56	369	379	13198	50000	0.22222	80	50	.00000000100
196	16	56	369	379	38	50000	0.22222	20	1000	.00000000001
197	16	56	369	379	85	50000	0.22222	50	1000	.00000000001
198	16	56	369	379	134	50000	0.22222	80	1000	.00000000001
199	16	56	369	379	500	50000	0.22222	20	700	.00000000001
200	16	56	369	379	500	50000	0.22222	50	700	.00000000001
201	16	56	369	379	500	50000	0.22222	80	700	.00000000001
202	16	56	369	379	875	50000	0.22222	20	500	.00000000001
203	16	56	369	379	933	50000	0.22222	50	500	.00000000001
204	16	56	369	379	937	50000	0.22222	80	500	.00000000001
205	16	56	369	379	6293	50000	0.22222	20	100	.00000000001
206	16	56	369	379	6531	50000	0.22222	50	100	.00000000001

207	16	56	369	379	6591	50000	0.22222	80	100	.00000000001
208	16	56	369	379	12831	50000	0.22222	20	50	.00000000001
209	16	56	369	379	13129	50000	0.22222	50	50	.00000000001
210	16	56	369	379	13208	50000	0.22222	80	50	.00000000001