2016/2017

TÉCNICAS ESTADÍSTICAS PARA LA DETECCIÓN DEL FRAUDE.

ADRIÁN MARTÍN GARCÍA

UNIVERSIDAD COMPLUTENSE
MADRID

TUTOR: JAVIER CASTRO CANTALEJO / ROSARIO ESPINOLA VILCHEZ
FACULTAD DE ESTUDIOS ESTADÍSTICOS
Máster en Minería de Datos e Inteligencia de Negocios

Índice

1. Introducción.

En la actualidad coexisten multitud de técnicas que permiten predecir los valores de una o varias variables dentro de una base de datos, por lo que a lo largo del presente estudio, se pretenderá aplicar estas técnicas bajo la simulación de un problema real que se puede encontrar casi en la totalidad de las empresas del mundo, ya que hablamos de la detección de los clientes que resultan fraudulentos para estas empresas.

Poder lograr aplicar dichas técnicas de manera efectiva para poder conocer de antemano que clientes pueden generar un perjuicio tanto económico como en cualquiera de las distintas casuísticas posibles supone una gran ventaja para las empresas, ya que, entre otras cosas, las empresas serían capaces de:

- Detectar los clientes fraudulentos.
- Conocer los patrones que definen a los clientes fraudulentos para poder actuar en consecuencia, evitando así futuros fraudes.

Por lo tanto, si las empresas son capaces de actuar para reducir al mínimo el efecto del fraude, las pérdidas ocasionadas se reducirían al mínimo, y por lo tanto, este tipo de estudios podrían marcar la diferencia dentro de las empresas, por lo que su importancia podría llegar a ser vital para el futuro de estas.

Como ya se irá viendo a lo largo del estudio, las condiciones de las cuales parte este estudio están situadas en un marco pensado para simular las condiciones reales que se encuentran dentro de las empresas, con el fin de demostrar las grandes ventajas que se obtienen de estos estudios.

2. Naturaleza de los datos.

Para comprender la naturaleza de los datos utilizados en este estudio, debemos indicar que los datos provienen de un estudio real, por lo que para respetar la confidencialidad a la cual se debe someter la información no es posible indicar la procedencia de los datos ni aportar información detallada sobre estos, pero si es posible indicar algunas de las características principales de estos, las cuales trataremos a continuación.

La mayor parte de las variables son variables continuas que serán utilizadas como variables explicativas para modelizar la variable objetivo, en total, la base de datos consta de 129 variables de este tipo, cuya nomenclatura de denominará mediante el prefijo "Var" e ira desde "Var1" hasta "Var129".

El resto de variables de la base de datos se corresponden con una variable para identificar a cada uno de los objetos de estudio de la base de datos, una variable para identificar el marco temporal del estudio y la variable objetivo (denominada "Fraude"), a este conjunto de variables las denominaremos "variables características" ya que cada una de ellas tiene un comportamiento y características generales diferentes entre sí que hace necesario analizarlas de manera independiente.

Todas las variables pertenecientes al estudio serán analizadas formalmente dentro del apartado de análisis descriptivos.

3. Objetivos y metodología.

A continuación desarrollaremos los objetivos del estudio así como la metodología empleada para intentar lograr los objetivos.

3.1. Objetivos.

El objetivo principal de este estudio en términos generales consiste en comprender de la forma más exacta posible el comportamiento de la variable "Fraude" en base al resto de variables, en términos estadísticos este conocimiento se reflejaría en la capacidad de predecir el valor de la variable "Fraude" a partir del conocimiento del resto de variables mediante el uso de técnicas de predicción estadística.

Otro objetivo que se marca en este estudio consistiría en conocer si la muestra de registros investigados (definición que trataremos más adelante) proviene de una muestra aleatoria de la población total, o es un conjunto de registros seleccionados por alguna característica en común o por tener en ellos algún patrón en sus variables, este objetivo se marca por la necesidad de extrapolar los resultados a toda la población, y el hecho de calcular y validar

los modelos únicamente con los registros anteriormente investigados provoca que si estos son distintos a la población total, extrapolar los resultados de manera directa no sería factible.

Dentro de las técnicas de predicción estadística aplicadas para predecir la variable fraude utilizaremos los modelos de Regresión Logística, Redes Neuronales, Random Forest y Gradient Boosting, por lo que dentro de los objetivos quedarían marcadas la obtención del mejor modelo de cada una de estas técnicas, así como el análisis de dicho modelo, para finalmente, obtener el mejor modelo de entre todos los anteriores.

3.2. Metodología.

Como ya hemos visto anteriormente, para modelizar el comportamiento de la variable objetivo, utilizaremos distintos modelos estadísticos, los cuales generarán la probabilidad de que un registro sea fraudulento, dicha probabilidad la denominaremos probabilidad de fraude. Una vez calculados los modelos, debemos conocer cuál es el mejor de ellos, ya que las probabilidades por si solas no permiten conocer que modelos predicen mejor o peor, por lo que a lo largo de este apartado explicaremos el proceso llevado a cabo para ordenar los modelos.

Dentro del estudio dividiremos nuestra base de datos en tres muestras, para determinar a que muestra pertenece cada uno de los registros utilizaremos el valor de la variable "Grupo", ya que esta variable hace referencia al marco temporal del estudio. Las características de cada una de ellas son:

- Muestra de entrenamiento: En esta muestra se incluirán los registros con los que calcularemos los modelos estadísticos con los que se modelizará el comportamiento de la variable fraude.

- Muestra de validación: En esta muestra aplicaremos los modelos generados por la muestra de entrenamiento, por lo que obtendremos una probabilidad de fraude que se podrá comparar con su variable "fraude", con lo que podremos conocer si los modelos aciertan en la predicción o no.

- Muestra de test: A partir de esta muestra obtendremos el resultado final del estudio, estos datos solo se utilizarán una vez determinado el modelo final del estudio, para así poder estimar la capacidad de predicción final del estudio.

Para decidir cómo considerar que un modelo es mejor que otro, tendremos en cuenta que, dentro de este estudio, lo más importante para los modelos obtenidos es que el modelo acierte cuando considere que una observación es fraudulenta.

Para determinar cuando los modelos clasifican un registro como fraudulento o no, generaremos un intervalo mediante un punto de corte inferior y un punto de corte superior, y todos aquellos registros con una probabilidad de fraude dentro de dicho intervalo se considerarían fraudulentos.

A partir de la metodología indicada, obtendremos 4 alternativas distintas una vez estimado el fraude de cada registro, dichas alternativas son:

- Verdadero positivo (VP): Observaciones fraudulentas que el modelo considera fraudulentas.
- Falso positivo (FP): Observaciones no fraudulentas que el modelo considera fraudulentas.
- Verdadero negativo (VN): Observaciones no fraudulentas que el modelo no considera fraudulentas.
- Falso negativo (FN): Observaciones fraudulentas que el modelo no considera fraudulentas.

Una vez definido a que alternativa pertenece cada registro, podemos obtener un criterio que nos permita ordenar de mejor a peor los modelos. Dado que el objetivo es detectar observaciones fraudulentas, podemos ordenar los modelos en función del porcentaje de acierto, siendo mejores los que obtengan un valor más alto, dicho porcentaje de acierto se obtiene mediante la siguiente formula:

- Porcentaje de acierto= VP / (VP+FP).

Por lo que dentro de los modelos que obtengamos, seleccionaremos aquel con mayor porcentaje de acierto.

Una vez decidido el criterio para ordenar los modelos, también debemos añadir una condición que debe cumplir un modelo para ser considerado válido, ya que si únicamente utilizamos el criterio de seleccionar como mejor modelo aquel con mayor porcentaje de acierto, podríamos encontrarnos con la situación de tener como mejor modelo un modelo que

únicamente considera un registro como fraudulento y obtiene un 100% de acierto, pero este modelo no resultaría muy útil.

Para evitar esta situación consideraremos válidos únicamente a aquellos modelos que consideren fraudulentos al menos, al 10% de la población, para obtener el porcentaje de registros fraudulentos utilizaremos el siguiente parámetro:

- Porcentaje de investigados= (VP+FP) / (VP+FP+VN+FN).

Por último, es posible que nos encontremos con dos o más modelos cuyo porcentaje de acierto sea similar, pero difieran significativamente en algún otro criterio que no se ha tenido en cuenta hasta ahora, como por ejemplo, la robustez del modelo, la interpretabilidad o la multicolinealidad, por lo que para todos los modelos que no difieran en más de un 1% en su probabilidad de fraude (lo que denominaremos como diferencia significativa) se tendrán en cuenta estos criterios para decidir cuál es mejor modelo. Dentro de los modelos que difieran en más de un 1% su probabilidad de fraude será siempre seleccionado el que obtenga mayor porcentaje de acierto.

Una vez definido como ordenar los modelos y las características que estos deben tener, solo queda indicar como generar los modelos. Estos modelos provendrán de distintas ramas, siendo estas, la rama de la Regresión Logística, las Redes Neuronales, Random Forest y Gradient Boosting.

Cada una de estas ramas tienen sus propios parámetros, pero dentro de este apartado analizaremos los parámetros comunes a todos ellos.

En primer lugar, se definirán para cada modelo, los valores posibles de los parámetros que se utilizarán para el cálculo de los modelos.

Además, para cada modelo se obtendrán puntos de corte dentro del intervalo [0,1]. Estos puntos de corte los definiremos como punto de corte inferior y punto de corte superior, las características de estos valores son:

- Únicamente tomarán valores múltiplos de 0.05 dentro del intervalo indicado, por lo tanto, tendremos 21 puntos de corte distintos.
- El punto de corte inferior debe ser menor o igual que el punto de corte superior.

- Para cada modelo se calcularán todas las combinaciones posibles teniendo en cuenta los dos aspectos anteriores, por lo que para cada modelo se obtienen 231 combinaciones distintas, a las que denominaremos sub-modelos.

Es evidente que el hecho de poder igualar ambos puntos de corte generará intervalos nulos donde no se encontraría ningún registro, pero en términos de programación simplifica notablemente el proceso.

Además, existe la posibilidad de que dentro de un intervalo no se encuentre la probabilidad de ningún registro, para ayudar en este proceso imaginemos el caso de no tener registros dentro del intervalo (0, 0,1) para un modelo determinado, por lo tanto, los sub-modelos con puntos de corte (0, 0,2), (0,05, 0,2) y (0,1, 0,2) obtendrían los mismos resultados para todos los parámetros a tener en cuenta, por lo que, en estos casos, se seleccionará como mejor opción aquella donde el intervalo sea más pequeño.

Por último, hay que indicar que en ningún modelo se aplicarán las interacciones entre variables como variables explicativas, ya que, dada la complejidad de los modelos y el tamaño de la base de datos el hardware utilizado hace imposible incluirlas.

3.2.1. Modelos predictivos.

A continuación, desarrollaremos las características de las técnicas estadísticas utilizadas a lo largo del estudio.

3.2.1.1. Regresión.

Dado que para el objetivo principal del estudio utilizaremos diversos modelos de predicción basados en la regresión estadística, a continuación, se expondrán las características principales de las diversas técnicas que se utilizarán a lo largo del estudio, así como de algunos términos que ayuden al seguimiento del trabajo.

La regresión estadística se basa en el estudio de la relación entre variables, lo cual permite conocer el efecto que hacen unas variables sobre otras, así como los patrones que reflejan estas relaciones. Obteniendo el conocimiento sobre dichas relaciones se obtienen multitud de ventajas, entre ellas destacaríamos el hecho de poder predecir el valor de una o varias variables a partir del resto y en este punto es donde entran en juego los modelos

predictivos, ya que estos son un conjunto de técnicas usadas para analizar y explorar la relación entre las distintas variables de una base de datos.

3.2.1.2. Regresión Logística.

La Regresión Logística es un modelo de regresión estadística que relaciona linealmente una variable dependiente categórica, con una o varias variables explicativas mediante una función de enlace.

Como ya veremos a lo largo del estudio, para nuestros modelos de Regresión Logística las funciones de enlace que utilizaremos serán las funciones Logic y Probit, donde para cada una de ella tenemos las siguientes formas funcionales:

- Modelo Logit:

$$1.\quad P = F(X, \boldsymbol{\beta}) + \mathbf{u} = \frac{e^{X\boldsymbol{\beta}}}{1+e^{X\boldsymbol{\beta}}}$$

Con lo que podemos expresar la forma funcional del modelo logit como:

$$2.\quad \log\left(\frac{P}{1-P}\right) = X\boldsymbol{\beta} + \mathbf{u}$$

Y, por lo tanto, la función de enlace correspondiente a este modelo quedaría de la forma:

$$3.\quad \log\left(\frac{P}{1-P}\right)$$

La cual se considera la función de enlace Probit, y se encontraría dentro de la familia de los modelos binomiales.

- Modelo Probit:

Para esta función de enlace, la función F de la forma funcional del modelo anterior (Función nº 1) será una función de distribución normal, por lo que la forma funcional resultante será la siguiente:

$$P = F(X, \boldsymbol{\beta}) + \mathbf{u} = (2\pi)^{\frac{-1}{2}} \int_{-\infty}^{X\boldsymbol{\beta}} e^{-\frac{t^2}{2}}\, dt + u$$

Con lo que, como en el caso anterior, podemos expresar el modelo probit de la siguiente manera:

$$\Phi^{-1}(P) = X\boldsymbol{\beta} + \mathbf{u}$$

Y, por lo tanto, la función de enlace del modelo Probit será de la forma $\Phi^{-1}(P)$ y aunque proviene de una distribución normal, también quedaría dentro de la familia de los modelos binomiales.

Dentro de las ventajas de este modelo, podemos resaltar que el hecho de utilizar únicamente relaciones lineales permite poder interpretar la relación entre las variables explicativas y la variable objetivo. Además, estos modelos permiten delimitar el valor a partir del cual una relación se considera estadísticamente significativa, por lo que el propio modelo es capaz de filtrar que variables son útiles y cuáles no.

El aspecto negativo de estos modelos se encuentra como contrapunto al aspecto de la linealidad, ya que el hecho de utilizar únicamente relaciones lineales implica que no recoge bien el resto de relaciones, por lo que su poder predictivo quedaría altamente limitado.

3.2.1.3. Redes Neuronales.

El concepto el cual pretende reflejar los modelos de Redes Neuronales y al cual debe su nombre, se corresponde con el funcionamiento de las Redes Neuronales dentro de un sistema nervioso, por lo que estos modelos son un conjunto de elementos de procesamiento de la información altamente interconectados que tienen la capacidad de alimentarse de la información para simular un proceso de aprendizaje, esta característica se considera una de las principales propiedades, ya que permite introducir información (mediante registros), la cual el modelo puede ser capaz de generalizarla y así ofrecer una salida.

Los modelos de Redes Neuronales requieren de ciertos parámetros para su correcto funcionamiento, dentro de los cuales se encontrarían las capas del modelo, las cuales son:

- capa de entrada.
- capa de salida.
- capas ocultas.

Dentro de estas capas, tenemos un número de nodos los cuales determinan en parte el funcionamiento del modelo.

El último aspecto a tener en cuenta dentro de nuestros modelos de Redes Neuronales serán las funciones de activación, las cuales sirven para relacionar funcionalmente las

ecuaciones que conectan cada uno de los nodos de las distintas capas, en nuestro caso se utilizarán las funciones de enlace arco tangente, tangente hiperbólica, seno y lineal, las cuales tienen la forma funcional que se muestra a continuación:

- Arco tangente:

$$arcotang(t) * \frac{2}{\pi}$$

- tangente hiperbólica:

$$1 - \frac{2}{(1 + e^{(2t)})}$$

- seno:

$$sin(t)$$

- lineal:

$$t$$

Entre las ventajas que presentan estos modelos, podemos encontrar la capacidad de detectar relaciones no lineales entre las variables, además de ser capaces de poder generalizar la información, y por ello no se ven afectadas por el ruido de las variables.

En contra punto, tenemos que los modelos de Redes Neuronales no se pueden interpretar, ya que no es posible determinar cómo se procesa la información dentro del modelo.

3.2.1.4. Árboles de Decisión.

Dentro de los distintos modelos estadísticos de predicción, los Arboles de Decisión reflejan la necesidad de segmentar una población de manera progresiva, ya que a medida que avanza el modelo, los arboles de decisión tratan de encontrar la variable más discriminante para segmentar la población en uno o varios segmentos a los cuales se les repetiría el proceso hasta llegar a tener todas y cada una de las entidades de la población en un grupo independiente al resto.

Hay que tener en cuenta que este modelo no busca relaciones multidimensionales entre las variables explicativas y la variable objetivo, sino que busca una a una cada una de las relaciones entre la variable objetivo y una variable explicativa, para después seleccionar la relación más discriminante.

A partir de la descripción anterior podemos encontrar gran cantidad de elementos que definen un árbol de decisión, entre ellos los más importantes se definen a continuación:

- Numero de hojas: hace referencia al número de subpoblaciones de las cuales no se obtiene una nueva división.
- Profundidad: Número máximo de divisiones entre la población inicial y cada una de las hojas.
- Numero de divisiones por hoja: Es un valor que hace referencia al número máximo de subpoblaciones que se pueden obtener de una población, cuando este número es dos, hablamos de un árbol de división binaria.
- Tamaño de hoja: Hace referencia al número mínimo de registros que debe haber dentro de una hoja, como ya hemos visto anteriormente, el modelo calcula divisiones hasta separar cada uno de los registros, por lo que este parámetro se utiliza para "parar" el cálculo de divisiones y así optimizar el poder predictivo del modelo.

Las ventajas de los árboles de decisión son muy importantes, ya que es de los pocos modelos que permiten ser visualizados gráficamente, lo cual es una cualidad muy útil para entender su funcionamiento, además, el hecho de buscar relaciones entre la variable objetivo y una única variable explicativa hace que su complejidad no sea elevada además de permitir poder realizar un seguimiento del recorrido que realizaría un registro a lo largo del modelo.

En contrapunto a lo anterior, el modelo de manera natural no detectaría las relaciones multidimensionales de las variables (aunque estas relaciones se podrían incluir como nuevas variables explicativas de manera manual), por lo que su capacidad predictiva podría verse limitada.

Por último, indicar que los árboles de decisión no se utilizarán para modelizar la variable objetivo de este estudio, aunque si tendrán su participación para el análisis de las muestras investigadas, además, el resto de modelos que describiremos a continuación tienen su base en los modelos de árboles de decisión.

3.2.1.5. Random Forest.

Los modelos de Random Forest son una evolución de los modelos de Árboles de Decisión, ya que una de las características de estos es que están formados por un número determinado de árboles de decisión, donde en cada uno de ellos, los datos que se utilizan para

el cálculo de cada uno de los árboles es una muestra de la población total, y además, las variables que entran en cada Árbol de Decisión son a su vez una muestra de las variables totales de la base de datos, por lo que a priori, cada modelo de Árbol de Decisión sería distinto al resto, y para concluir el modelo de Random Forest final, se calculará el promedio con cada uno de los Árboles de Decisión.

Las variables que definen estos modelos incluirán todas las variables de los modelos de Árboles de Decisión, y además las variables propias de estos modelos, las cuales son:

- Número de árboles de decisión calculados para el modelo de Random Forest.
- Porcentaje de registros que entran en cada uno de los Árboles de Decisión.
- Porcentaje de variables que entran en cada uno de los Árboles de Decisión.

Como ventaja dentro de estos modelos, debemos destacar su gran capacidad de predicción, por lo que se observa como en la actualidad, son una de las mejores opciones para encontrar modelos predictivos en problemas de predicción estadística.

Como desventajas de estos modelos, es importante indicar que pierden una de las características principales de los árboles de decisión, ya que el hecho de unir multitud de modelos de árboles de decisión provoca que se pierda la posibilidad de interpretar los resultados.

3.2.1.6. Gradient Boosting.

Los modelos de Gradient Boosting se corresponden con una evolución de los modelos de Random Forest, ya que estos modelos obtienen su capacidad predictiva a través de la construcción escalonada de modelos de Random Forest.

Para poder calcular estos modelos es necesario una primera etapa, en la cual se obtiene el modelo inicial, el cual será ajustado en una segunda etapa, ya que la construcción del segundo modelo tendrá en cuenta la información del modelo anterior, donde el proceso de ajuste modificará las predicciones del modelo anterior con el objetivo de minimizar los errores de los modelos, obteniendo así, de manera gradiente modelos que convergen en un modelo final donde los errores son mínimos.

Ya que la base de estos modelos son los modelos de Random Forest, los modelos de Gradient Boosting compartirán las variables que definen los modelos de Random Forest y a su vez ganan nuevos parámetros, los cuales son:

- Iteraciones: reflejan el número de etapas del modelo.
- Parámetro Shrinkage, refleja el grado con el que se ajustará el modelo en cada una de las iteraciones.

Las ventajas e inconvenientes de estos modelos coinciden con las de los modelos de Random Forest, salvo que en la actualidad se observa como a nivel general, las predicciones de los modelos de Gradient Boosting superan ligeramente a los modelos de Random Forest, por lo que los dos formarían el grupo de los modelos con mejores resultados para los problemas de predicción.

3.2.1.7. Metodología Fuzzy.

Esta técnica no se corresponde con un modelo de regresión estadística, sino que quedaría enmarcada dentro de las técnicas de inferencia de rechazados, las cuales tienen como objetivo poder introducir dentro de la construcción de los modelos, registros donde se desconoce el valor de la variable objetivo.

La metodología de esta técnica consiste en predecir la variable objetivo utilizando tanto los registros donde conocemos el valor la variable objetivo como aquellas de las que se desconoce dicho valor

Para ello, se calcula en una primera fase un modelo utilizando únicamente los registros donde si se conoce el valor de la variable objetivo, y una vez obtenido dicho modelo, aplicarlo sobre el resto de registros que se quieren incluir en el cálculo del modelo.

Una vez aplicado el modelo, si la variable objetivo es categórica, obtendremos una probabilidad para cada una de las categorías posibles de la variable objetivo dentro de cada uno de los registros, llegando así a la segunda fase de esta técnica.

En esta fase, para la construcción de los modelos se introducirá una nueva variable, que en nuestro caso denominaremos “peso”, y la cual funcionará como peso de cada registro para el cálculo del modelo, donde el valor de la variable “peso” para cada registro se calcula de la siguiente manera:

- Para las variables de las que conocemos el valor de la variable objetivo tomará el valor 1.
- Para las variables de las que se desconoce el valor de la variable objetivo el proceso es el siguiente:
    1. Replicamos cada registro tantas veces como categorías tenga la variable objetivo.
    2. Para cada caso, el valor de la variable objetivo será cada una de las categorías, y su peso se corresponderá con la probabilidad obtenida para dicho registro y dicho valor de la variable objetivo.

Una vez obtenida la nueva muestra de entrenamiento, ya es posible calcular el modelo correspondiente mediante la metodología Fuzzy.

4. Análisis descriptivos.

En este apartado intentaremos dar una visión global del comportamiento de las variables de nuestra base de datos, para lo cual dividiremos las variables en dos grupos, en el primero incluiremos algunas variables las cuales su comportamiento es único dentro del estudio (las denominadas variables características, ya introducidas anteriormente), y por lo tanto debemos analizarlas individualmente, en el segundo bloque incluiremos a las variables continuas, ya que aunque estas son independientes entre sí (en relación a que cada variable representa aspectos distintos dentro de la población estudiada), todas ellas se utilizarán como variables explicativas dentro del cálculo de los modelos, además de ser tratadas todas ellas como variables continuas.

4.1. Variables características.

Dentro de este apartado incluiremos las variables Identificador, Fraude y Grupo, las cuales analizaremos a continuación:

- Identificador: Es un código identificativo para cada observación, hace referencia al marco poblacional de la muestra, por lo que cada uno de los clientes de la empresa tendría su propio identificador, el cual es único para cada cliente.

- Grupo: Es la variable que indica el marco temporal del estudio, para evitar aportar información al respecto, el primer grupo se define con un 1, y los siguientes grupos a

través de la numeración correspondiente hasta llegar al valor 66, para ayudar a comprender este concepto, podríamos considerar el grupo 1 equivalente al día 1, y así sucesivamente hasta el día 66.

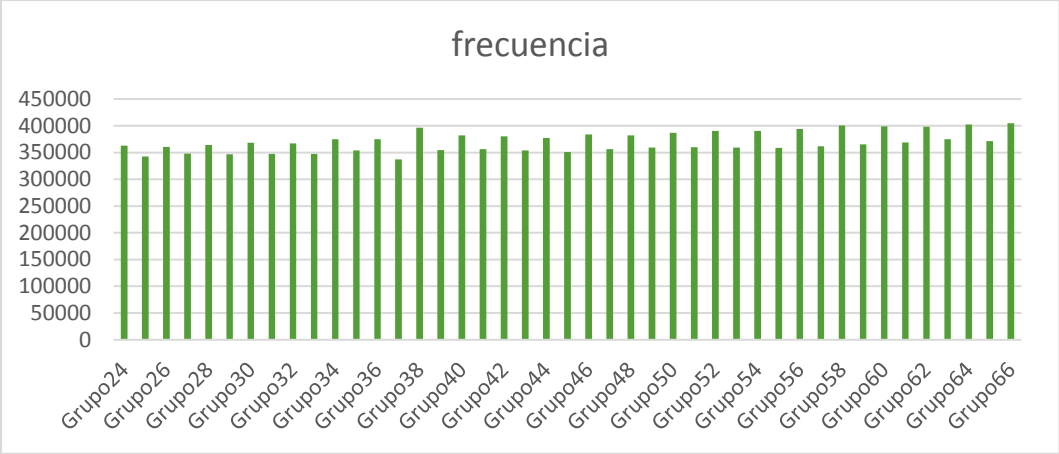

Gráfico 1: Frecuencias variable Grupo

Gráficamente la distribución de la variable "Grupo" se muestra en el Gráfico 1, donde podemos observar como existe una tendencia creciente en el número de registros, esto se puede deber al aumento de clientes en la base de datos, aunque este hecho no se tendrá en cuenta a lo largo del estudio.

- Fraude: Es la variable objetivo del estudio, hace referencia al resultado de una investigación sobre un determinado registro en un grupo en concreto, el valor 1 se corresponde con los casos en los que se encontró un fraude, mientras que el -1 indica que no se encontró fraude, y por último el valor 0, el cual indica que dicha observación en ese grupo concreto no fue investigada y por lo tanto no tenemos información sobre ella.

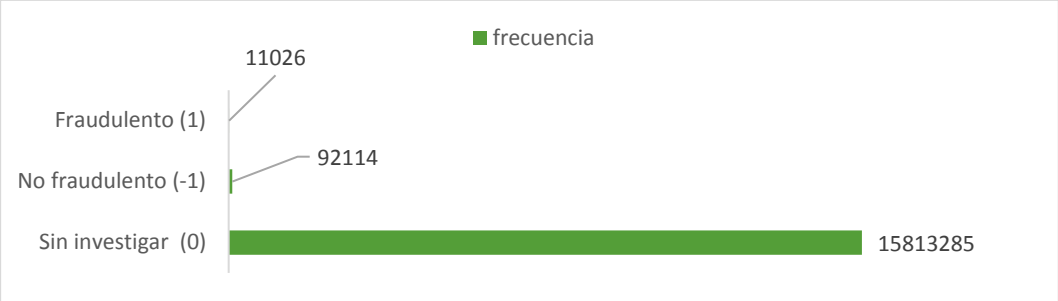

Gráfico 2: Frecuencias variable Fraude.

En el Gráfico 2 observamos cómo el 99,35% de los datos no ha sido investigado, mientras que dentro del 0,64% restante, el 89,3% de las observaciones no son fraudulentas, frente al 10,69% restante que sí lo son, por lo que es evidente que la población que no fue investigada representa gran parte de la población.

4.2. Variables Continuas.

Para las variables continuas del estudio, calcularemos los estadísticos más importantes, siendo estos la media, el mínimo y el máximo, los cuartiles y por último el índice de asimetría y el de kurtosis. A continuación en la Tabla 1 se muestran los resultados para las 6 primeras variables (en la Tabla Anexo 1, se incluye la tabla completa con las 129 variables):

Variable	Media	Mínimo	Cuartil 1	Mediana	Cuartil 3	Máximo	Asimetría	Kurtosis
var1	76,064	0,000	22,000	45,000	82,000	2187725,000	2341,710	6190593,120
var2	126,000	0,000	83,170	97,937	113,320	246015,660	376,891	159256,240
var3	132,146	0,000	96,637	114,135	127,670	100240,200	202,176	54539,260
var4	132,663	0,000	100,989	108,409	137,261	70821,520	146,593	34861,740
var5	132,125	0,000	104,787	120,290	122,206	50100,040	92,742	18047,720
var6	1,000	0,000	0,356	0,758	1,306	1098,630	58,165	33722,120

Tabla 1: Muestra análisis descriptivos.

Nótese como observación que dentro de la base de datos no encontramos ningún valor perdido, además debemos tener en cuenta que no disponemos de datos extremos que pudieran afectar dentro del cálculo de los modelos.

5. Análisis observaciones investigadas.

En este apartado intentaremos conocer si las observaciones que han sido investigadas tienen alguna diferencia sobre el resto o no, en otras palabras, intentaremos descubrir si los usuarios que han sido investigados provienen de un muestreo aleatorio o han sido investigadas aquellas observaciones que tienen un determinado comportamiento.

Para responder a esto, crearemos la variable "Investigados", que tomará el valor 1 para todas las observaciones investigadas, es decir, aquellas que tienen un valor -1 o 1 en la variable "Fraude", y tomará el valor 0 para las observaciones con la variable "Fraude" igual a 0.

La técnica que utilizaremos para conocer si las observaciones investigadas provienen de una muestra aleatoria será intentar modelizar la probabilidad de que una observación sea investigada o no mediante la aplicación de un árbol de decisión, considerando que la muestra no es aleatoria si dicho árbol es capaz de predecir que observaciones han sido investigadas y cuáles no, mientras que, si el árbol no es capaz de predecirlo, sería un primer indicio de que la muestra es aleatoria.

Para obtener un correcto funcionamiento dentro de los modelos de los árboles de decisión, es necesario que la frecuencia entre las distintas categorías de la variable objetivo sea aproximadamente similares (siempre y cuando no se esté utilizando una variable que identifique un peso para cada registro que permita solventar este problema), o al menos con unas diferencias que no sean exageradas, para comprobar si nuestra base de datos reúne dicha condición utilizaremos la Tabla 2 mostrada a continuación:

investigado	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0	15813285	99.35	15813285	99.35
1	103140	0.65	15916425	100.00

Tabla 2: Frecuencias variable "Investigados"

Como observamos en la Tabla 2, las frecuencias no están balanceadas, por lo que es necesario realizar alguna modificación en esta para poder lograr que los resultados del árbol de decisión sean útiles, para ello replicaremos la población de los registros investigados 152 veces

para obtener una población balanceada, con lo que la nueva tabla de frecuencias seria la siguiente:

investigado	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0	**15813285**	50.05	**15813285**	50.05
1	**15780420**	49.94	**31593705**	100.00

Tabla 3: Frecuencias variable "Investigados" balanceada

Para la obtención del árbol se ha utilizado el software SAS Enterprise Miner, en el cual se muestran a continuación los principales valores utilizados para definir las características del árbol de decisión:

- Árbol binario.
- Profundidad máxima 5.
- Tamaño hoja mínimo 1000000 registros.

Gráficamente el árbol obtenido se muestra en la Ilustración 1:

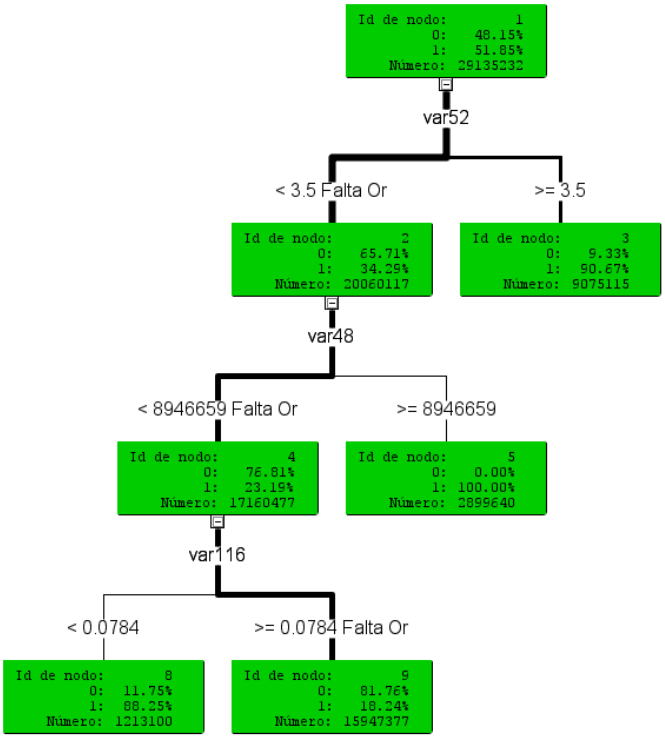

Ilustración 1: Árbol de decisión.

El árbol utiliza únicamente 3 de las 129 variables, pero clasifica correctamente al 86.621% de las observaciones, por lo que parece razonable considerar que las observaciones investigadas se rigen bajo algún patrón, este hecho es muy determinante para el análisis, ya que afectará de manera importante durante el resto del estudio, debido a que las conclusiones pueden verse afectadas y por lo tanto, es posible que debamos realizar ajustes a los modelos obtenidos, aunque todas estas cuestiones serán tenidas en cuenta más adelante.

Parece evidente, dados los resultados obtenidos por el árbol de decisión que existen diferencias entre la población investigada y la que no lo ha sido, por lo tanto, a priori los resultados que se obtengan utilizando la población investigada no se puede extrapolar a la población no investigada, y por lo tanto, intentaremos mejorar dichos resultados mediante las técnicas denominadas como inferencia de rechazados, dentro de las cuales utilizaremos el método Fuzzy, como ya veremos a lo largo del estudio.

5.1. Resultados iniciales.

Aunque a priori no se dispone de información de modelos iniciales (los cuales serían los resultados de modelos de predicción aplicados al problema antes del inicio de este estudio), si podemos obtener información al respecto, ya que se puede considerar que las observaciones de las que se tienen información en la variable fraude es porque éstas han sido investigadas, mientras que no lo han sido aquellas con valor desconocido en la variable fraude, por lo tanto, podemos obtener el porcentaje de investigados del modelo inicial.

La segunda variable que tendremos en nuestros modelos es el porcentaje de acierto, la cual la podemos obtener utilizando la información correspondiente a la variable fraude de las observaciones investigadas, ya que podemos considerar como aciertos las observaciones en las que se detectó fraude (verdadero positivo), y fracasos aquellas investigadas en las que no se detectó fraude (falso positivo) y, por lo tanto, los resultados del modelo inicial son:

- Observaciones totales = 15916425.
- Observaciones investigadas = 103140.
- Observaciones fraudulentas = 11026.
- Observaciones no fraudulentas = 92114.

Consideraremos que el porcentaje de investigados inicial es del 0.648%, con un porcentaje de acierto del 10.690%, por lo que nos marcaremos como objetivo superar este porcentaje de acierto, ya que, si no lo lográsemos, el modelo inicial superaría al nuestro y por lo tanto el obtenido por este estudio no podría ser aplicado (para esta afirmación no estamos teniendo en cuenta el porcentaje de investigados).

6. Modelos de predicción (Inferencia estadística).

Dentro de este apartado procederemos a calcular los distintos modelos de predicción los cuales ya han sido introducidos a lo largo del estudio incluyendo únicamente datos de la muestra de entrenamiento cuyo valor de la variable fraude sea conocido. Dado que este apartado será el núcleo del estudio, formará la mayor parte del trabajo, a continuación, empezaremos con la primera familia de modelos, que en este caso serán los modelos de Regresión Logística.

6.1. Regresión Logística.

Para los modelos de Regresión Logística, las características que podremos modificar serán la función de enlace (Enlace), el método de selección de variables (Selección) y el p-valor (P-Valor), los valores que podrán tomas estas variables son los siguientes:

- Enlace: Logic y Probit.
- Selección: Stepwise, Forward y Backward.
- P-Valor: 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001, 0.00000001, 0.000000001,

Por lo tanto, para estas tres variables tendremos 60 combinaciones distintas, y por lo tanto tendremos 60 modelos distintos para los modelos de Regresión Logística, estos modelos se dividen en 13.860 sub-modelos a partir de los distintos puntos de corte. Una vez obtenidos todos ellos, ordenaremos todos los que cumplen las condiciones necesarias, por lo que a continuación se muestran los 20 mejores modelos:

Modelo	Parámetros	CorteMin	CorteMax	Selección	Enlace	P-Valor	Acierto	Investigados
499	24	0,05	0,8	stepwise	Logic	0,01	0,227855	0,102186
518	**24**	**0,1**	**0,8**	**stepwise**	**Logic**	**0,01**	**0,227855**	**0,102186**
536	24	0,15	0,8	stepwise	Logic	0,01	0,227247	0,101958
5099	28	0	0,8	forward	Logic	0,01	0,227222	0,102471
5119	28	0,05	0,8	forward	Logic	0,01	0,226793	0,102414
5138	28	0,1	0,8	forward	Logic	0,01	0,226793	0,102414
5156	28	0,15	0,8	forward	Logic	0,01	0,226615	0,102243
5330	22	0	0,8	forward	Logic	0,001	0,226436	0,102072
710	21	0	0,8	stepwise	Logic	0,001	0,226164	0,102698
5350	22	0,05	0,8	forward	Logic	0,001	0,226004	0,102015
749	21	0,1	0,8	stepwise	Logic	0,001	0,225735	0,102641
767	21	0,15	0,8	stepwise	Logic	0,001	0,22543	0,102528
5387	22	0,15	0,8	forward	Logic	0,001	0,225266	0,101844
553	24	0,2	0,8	stepwise	Logic	0,01	0,225084	0,101674
7409	26	0	0,8	forward	Probit	0,01	0,224972	0,103495
7429	26	0,05	0,8	forward	Probit	0,01	0,224972	0,103495
569	24	0,25	0,8	stepwise	Logic	0,01	0,22465	0,101617
7448	26	0,1	0,8	forward	Probit	0,01	0,224546	0,103438

5173	28	0,2	0,8	forward	Logic	0,01	0,224456	0,101958
2558	23	0	0,8	stepwise	Probit	0,05	0,224309	0,10304

Tabla 4: Mejores modelos Regresión Logística.

En primer lugar, debemos tener en cuenta la diferencia significativa, podemos comprobar como únicamente los 4 primeros modelos se encuentran dentro de dicho intervalo (utilizando como valor de referencia el porcentaje de acierto más alto). Como criterio para decidir el mejor modelo podemos tener en cuenta la robustez de los modelos, por lo que buscaríamos aquel con menor número de parámetros, aunque esto sigue siendo favorable a los dos primeros modelos (ambos tienen el mismo porcentaje de acierto), por lo que deberemos seleccionar uno de estos dos modelos.

Para decidir entre el modelo 499 y el 518 no podemos utilizar los criterios anteriormente descritos, ya que en lo único que se diferencian ambos modelos es en el punto de corte inferior, ya que dentro del intervalo [0.05, 0.1] el modelo 499 no considera a las observaciones fraudulentas mientras que el modelo 518 sí. Teniendo en cuenta que el resto de parámetros del modelo son exactamente idénticos, llegamos a la conclusión de que dentro de ese intervalo no existe ninguna observación, por lo que finalmente el modelo seleccionado será el 518, ya que no consideramos fraudulentas observaciones dentro de un intervalo para el cual no tenemos información, este criterio se utilizará en todas las ocasiones en las que la única diferencia entre dos modelos sea el intervalo en el que selecciona las observaciones como fraudulentas.

Por lo tanto, las características del mejor modelo de Regresión Logística son las siguientes:

- Método de selección Stepwise
- Función de enlace Logic
- P-valor 0.01
- Punto de corte inferior 0.1
- Punto de corte inferior 0.8

Resultados obtenidos por el modelo:

- Porcentaje de acierto = 22.785%
- Porcentaje de investigados = 10.219%
- Numero de parámetros = 23

Como ya vimos en la Tabla 4, el modelo seleccionado consta de 24 parámetros, es decir, 23 variables explicativas y un parámetro independiente, a continuación, se muestra la tabla que indica la estimación del valor asociado a cada parámetro en la Tabla 5:

Intercept	**-79.099**
var8	-0.0117
var9	-0.00860
var10	-0.00949
var30	0.000014
var32	-0.00002
var36	-0.00006
var49	-0.2214
var50	0.2784
var57	0.1249
var59	-0.0589
var67	-0.0262
var68	75.785
var77	-0.4429
var81	0.0125
var85	0.2782
var86	0.3918
var97	0.000173
var110	45.852
var111	87.795
var113	144.346
var116	71.029
var119	111.129
var127	0.0968

Tabla 5: Estimación parámetros Regresión Logística.

Por lo tanto, si tenemos en cuenta las estimaciones de cada parámetro, y que la función de enlace del modelo seleccionado era la función Logic, podemos obtener la ecuación a través de la cual es posible obtener la estimación en el valor del fraude de cada registro, dicha ecuación es:

$$\frac{e^{X\beta}}{1 + e^{X\beta}}$$

Donde $X\beta$ = -79099*Intercept - 0.0117*var8 - 0.00860*var9 - 0.00949*var10 + 0.000014*var30 - 0.00002*var32 - 0.00006*var36 - 0.2214*var49 + 0.2784*var50 + 0.1249*var57 - 0.0589*var59 - 0.0262*var67 + 75.785*var68 - 0.4429*var77 + 0.0125*var81 + 0.2782*var85 + 0.3918*var86 + 0.000173*var97 + 45852*var110 + 87.795var111 + 144.346*var113 + 71.029*var116 + 111.129*var119 + 0.0968*var127.

Por lo que las observaciones con un valor en la ecuación anterior situado dentro del intervalo (0,1, 0,8) quedarían clasificadas como fraudulentas y por lo tanto, deberían ser investigadas si finalmente este es el modelo final seleccionado, por lo que es evidente que todas las variables que no aparecen en la ecuación no están incluidas en el modelo y por lo tanto, según nuestro modelo de Regresión Logística seleccionado, todas estas variables no aportan información significativa para poder predecir la variable "Fraude".

Por último, para poder estimar la importancia relativa de cada variable dentro del modelo, calcularemos dicha importancia en valor absoluto mediante la multiplicación del parámetro estimado para cada variable por el valor medio dentro de la base de datos de entrenamiento de dicha variable, los resultados se muestran a continuación en la **¡Error! No se encuentra el origen de la referencia.**:

Variable	**Valor medio**	**Parámetro**	**Importancia relativa**
var8	17,52316	-0,0117	0,20502097
var9	12,26192	-0,0086	0,10545251
var10	13,27276	-0,00949	0,12595849
var30	895,7368	0,000014	0,01254032
var32	2666,346	-0,00002	0,05332692
var36	1036,314	-0,00006	0,06217884
var49	0,592638	-0,2214	0,13121005
var50	0,608679	0,2784	0,16945623
var57	0,18632	0,1249	0,02327137
var59	0,47594	-0,0589	0,02803287
var67	3,63119	-0,0262	0,09513718
var68	0,107167	75,785	8,1216511
var77	-0,11129	-0,4429	0,04929034
var81	22,50629	0,0125	0,28132863
var85	-0,10172	0,2782	0,0282985
var86	0,246112	0,3918	0,09642668
var97	39,29202	0,000173	0,00679752
var110	0,109087	45,852	5,00185712
var111	0,107434	87,795	9,43216803

var113	0,106343	144,346	15,3501867
var116	0,107685	71,029	7,64875787
var119	0,107103	111,129	11,9022493
var127	0,854123	0,0968	0,08267911

Tabla 6: Parámetros importancia variables Regresión Logística.

Una vez obtenida la importancia de cada variable, podemos comprobar como claramente las variables var68, var110, var111, var113, var116 y var119 destacan sobre el resto dada su gran importancia, además, observamos cómo estas variables son aquellas que tienen los parámetros estimados más grandes.

6.2. Modelo Redes Neuronales.

Para los modelos de Redes Reuronales, los parámetros que utilizaremos para definir características serán:

- Uso máximo de dos capas ocultas.
- Numero de nodos en la primera capa oculta (valores: de 1 a 10).
- Numero de nodos en la segunda capa oculta (valores: de 0 a 10).
- Función de activación (valores: tangente (TAN), tangente hiperbólica (TANH), lineal (LIN), seno (SEN) y arco tangente (ARC))

Por lo tanto, calcularemos 550 modelos distintos con 127.050 sub-modelos tras aplicar los puntos de corte, una vez obtenidos los modelos, podremos ordenarlos en función del porcentaje de acierto, mostrando en la Tabla 7 los 20 mejores modelos que tienen un porcentaje de investigados superior al 10%.

Activacion	parametros	CorteMin	CorteMax	Nodo1	Nodo2	Acierto	Investigados
TAN	**289**	**0,05**	**0,75**	**2**	**7**	**0,25539**	**0,102983**
TAN	289	0	0,75	2	7	0,2551077	0,1030969
TAN	289	0,1	0,75	2	7	0,2541528	0,1028123
TAN	289	0,15	0,75	2	7	0,2523624	0,1024138
TAN	289	0,2	0,75	2	7	0,250558	0,1020153
TAN	1281	0	0,75	9	10	0,2491506	0,1005351
TAN	289	0,25	0,75	2	7	0,2484577	0,1015029

TAN	1281	0,05	0,75	9	10	0,2478729	0,1003643
TAN	1071	0	0,75	8	3	0,2446689	0,101446
TAN	289	0,3	0,75	2	7	0,2446328	0,1007628
TAN	436	0	0,75	3	9	0,2438753	0,102243
TAN	436	0,05	0,75	3	9	0,2438753	0,102243
ARC	527	0	0,75	4	1	0,243532	0,101218
ARC	527	0,05	0,75	4	1	0,243532	0,101218
ARC	527	0,1	0,75	4	1	0,243532	0,101218
ARC	527	0,15	0,75	4	1	0,243532	0,101218
TAN	1071	0,05	0,75	8	3	0,2433952	0,1012752
TAN	436	0,1	0,75	3	9	0,2430323	0,1021291
TAN	1111	0	0,75	8	7	0,2414519	0,1082204
TAN	551	0	0,75	4	5	0,2408027	0,1021291

Tabla 7: Orden modelos Redes Neuronales.

Al analizar la Tabla 7 observamos como dentro de los modelos que cumplen las condiciones, el primer modelo (y por lo tanto aquel con mayor porcentaje de acierto) obtiene un porcentaje de acierto del 25.54%, igualado por el segundo modelo. Ya que ambos son idénticos salvo en el intervalo de selección.

Como ya vimos en el apartado de metodología del estudio, en el caso de tener modelos que únicamente difieran en los puntos de corte, se seleccionará aquel con el intervalo más pequeño, en este caso sería el primer modelo (0,05, 0,75),

Por último, tenemos que tener en cuenta los modelos que difieran únicamente en un 1% en la probabilidad de acierto, y dado que en este caso no existe ninguno, el modelo seleccionado será finalmente el primero de la Tabla 7.

Una vez determinado el mejor modelo, y al tratarse de un modelo de Redes Neuronales, es necesario realizar un segundo paso, en el cual se repetirán tantas veces el modelo como variables explicativas tengamos (en este caso 129 modelos distintos), donde en cada uno de ellos, será eliminada cada una de las variables explicativas del modelo original, para poder estudiar el impacto que tenga la eliminación cada variable en el modelo, considerando que si el modelo empeora significativamente, es porque la variable eliminada era importante, mientras que si el modelo no empeora, es porque la variable no aportaba información significativa al modelo.

Una vez explicado el procedimiento, obtendremos los 129 modelos correspondientes, en la Tabla 8 podemos analizar los 10 modelos con mayor porcentaje de acierto.

Variable eliminada	Parámetros	Corte Min	Corte Max	Capas nodo 1	Capas nodo 2	% Acierto	% Investigados
var71	**287**	**0,05**	**0,75**	**2**	**7**	**0,257518**	**0,104122**
var119	287	0,05	0,75	2	7	0,237673	0,102755
var123	287	0,05	0,75	2	7	0,265832	0,081806
var16	287	0,05	0,75	2	7	0,265734	0,081407
var73	287	0,05	0,75	2	7	0,265432	0,083001
var10	287	0,05	0,75	2	7	0,265027	0,083343
var14	287	0,05	0,75	2	7	0,264808	0,081692
var61	287	0,05	0,75	2	7	0,264767	0,08192
var29	287	0,05	0,75	2	7	0,264022	0,083229
var52	287	0,05	0,75	2	7	0,263543	0,077764

Tabla 8: Mejores modelos 1º etapa Redes.

Al analizar la Tabla 8 se comprueba como únicamente los dos primeros modelos superan el límite de tener al menos al 10% de la población investigada, y por lo tanto estos son los únicos modelos que podemos considerar válidos, además dentro de ellos únicamente uno supera en porcentaje de acierto al modelo con todas las variables, por lo que este modelo será el elegido como mejor modelo en esta fase, es decir, la variable 71 será eliminada de nuestros modelos.

En la siguiente fase generaremos los 128 modelos posibles tras eliminar la variable var71, es decir, repetiremos el proceso anterior, salvo con la diferencia de que la variable var71 ya ha sido eliminada en todos los modelos. Como en el caso anterior mostraremos los resultados en la Tabla 9.

Variable	Parámetros	CorteMin	CorteMax	Nodo1	Nodo2	Acierto	Investigados
var14	285	0,05	0,75	2	7	0,253941	0,101104
var24	285	0,05	0,75	2	7	0,250418	0,102072
var25	285	0,05	0,75	2	7	0,24797	0,105146
var29	285	0,05	0,75	2	7	0,246006	0,106911
var17	285	0,05	0,75	2	7	0,245557	0,105716
var50	285	0,05	0,75	2	7	0,243984	0,101731
var67	285	0,05	0,75	2	7	0,235608	0,106797

var56	285	0,05	0,75	2	7	0,229155	0,103097
var101	285	0,05	0,75	2	7	0,226404	0,101332
var33	285	0,05	0,75	2	7	0,22597	0,101275

Tabla 9: Mejores modelo 2º etapa Redes.

En esta nueva etapa observamos cómo los 10 primeros modelos si investigan a más del 10% de la población, pero si incluimos en el análisis el porcentaje de acierto, observamos como ninguno de ellos mejora al modelo en el que únicamente se eliminaba la variable var71, y por lo tanto no debemos seguir eliminando variables y ya tenemos nuestro mejor modelo de Redes Neuronales.

Características del mejor modelo de Redes Neuronales.

- Punto de corte inferior 0.05.
- Punto de corte superior 0.75.
- 2 capas en el primer nodo.
- 7 capas en el segundo nodo.
- Función de enlace Tangente (Tan).

Resultados obtenidos para el modelo de Redes Neuronales.

- Tiene un porcentaje de acierto del 25.39%.
- El modelo considera fraudulenta al 10.11% de la población.
- Numero de parámetros = 285.

Para analizar los resultados del modelo de Redes Neuronales debemos utilizar métodos auxiliares, ya que las ecuaciones de los modelos de Redes Neuronales no son interpretables ni permiten un estudio de su funcionamiento.

En primer lugar, utilizaremos los resultados obtenidos anteriormente donde eliminábamos cada una de las variables y repetíamos el modelo, ya que como vimos anteriormente, a través de la eliminación de cada una de las varaibles podemos analizar el impacto de estas sobre el modelo final. Los resultados se muestran en la Tabla 10.

Variable	**Acierto**	**Investigados**
var23	0,278383	0,07321
var1	0,272169	0,083457
var8	0,272059	0,077422

var7	0,270703	0,081806
var9	0,270597	0,080155
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
var67	0,235608	0,106797
var116	0,230363	0,086246
var56	0,229155	0,103097
var101	0,226404	0,101332
var33	0,22597	0,101275

Tabla 10: Resumen porcentaje de acierto variables eliminadas.

Podemos observar como al eliminar la variable var33 el porcentaje de acierto pasa del 25.75% al 22.59%, siendo la mayor diferencia obtenida entre todas las variables, por lo que podríamos considerar esta variable como la más importante, mientras que en el lado inverso nos encontramos la variable var23, la cual no solo no reduce el porcentaje de acierto al eliminarla, sino que dicho porcentaje aumenta, por lo que se puede considerar poco importante.

Para el cálculo de la importancia a de cada variable no se esta teniendo en cuenta como varía el porcentaje de investigados, por lo que la importancia de cada variable es meramente ilustrativa y no un dato exacto ya que, dentro de los modelos de Redes Neuronales, no es posible obtener dicho valor.

En el Gráfico 3 se muestran gráficamente el porcentaje de acierto y el porcentaje de investigados obtenido para cada uno de los modelos correspondientes tras eliminar una de las variables:

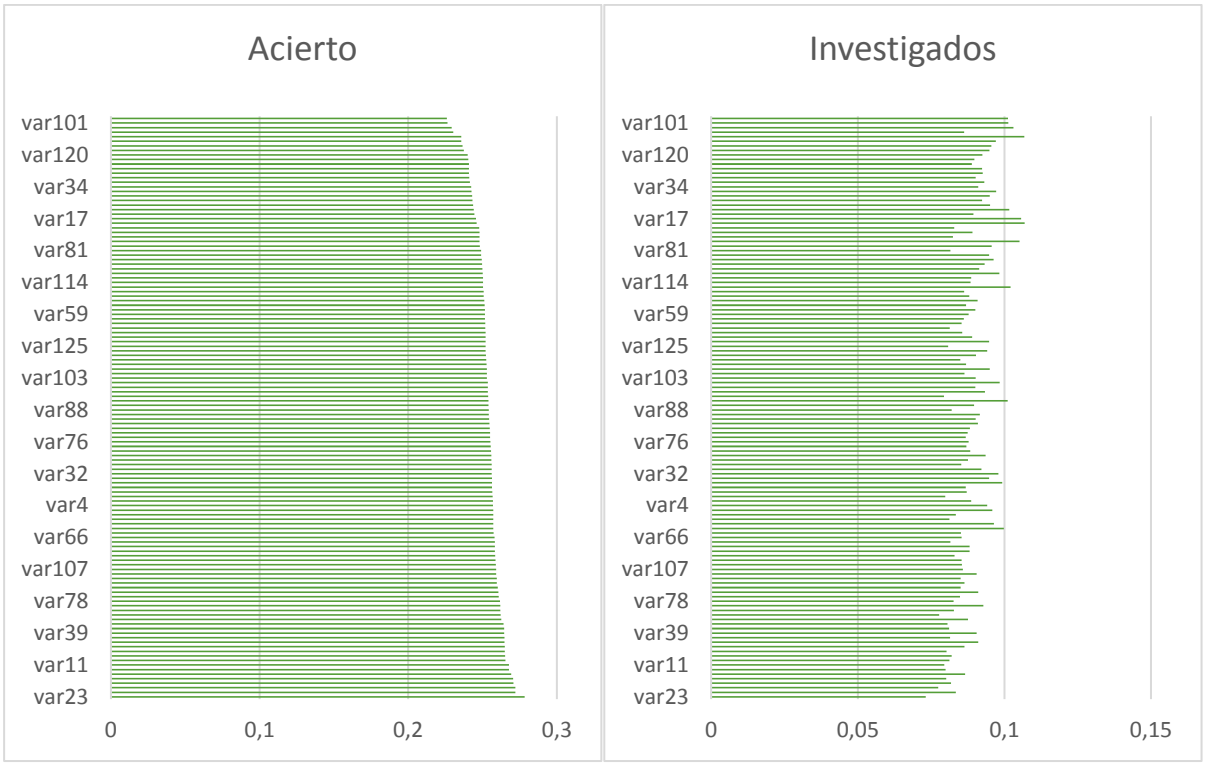

Gráfico 3: Porcentaje de acierto e investigados Redes Neuronales.

Observamos cómo tras ordenar las variables en función del porcentaje de acierto se puede observar cierta relación entre el porcentaje de acierto y el porcentaje de investigados, ya que a medida que uno de estos dos parámetros aumenta, el otro disminuye.

6.3. Random Forest.

En esta sección nos centraremos en los modelos de predicción obtenidos a partir de Random Forest.

Para definir estos modelos, uno de los parámetros a tener en cuenta será el tamaño mínimo que puede tener una hoja dentro de los distintos árboles y tomará los valores 100, 600 y 1100, otro parámetro será la profundidad máxima de los árboles, cuyos valores son 10, 30 y 50, además probaremos con modelos con 30, 60 o 90 árboles, y por último el p-valor a partir del cual se considerarán las divisiones significativas o no (necesario para realizar la poda del árbol), el cual tomará los valores 0.1, 0.05 y 0.01.

Para realizar los arboles utilizaremos como valor fijo el 40% de los datos de la muestra.

Teniendo en cuenta estas 4 variables, con 3 opciones distintas cada una de ella, obtenemos 81 combinaciones diferentes, a partir de las cuales obtendremos 18.711 sub-modelos diferentes tras aplicar los distintos puntos de corte.

Todos estos modelos se compararán entre sí, para ello utilizaremos la Tabla 11 donde se muestran ya ordenados aquellos que cumplen la condición de investigar al 10% de la población por orden del porcentaje de acierto.

Modelos	NumNormas	CorteMin	CorteMax	numArboles	TamañoHoja	profundidad	P-Valor	Acierto	Investigados
701	1440	0	0,35	30	600	10	0,1	0,220708	0,104463
721	1440	0,05	0,35	30	600	10	0,1	0,220708	0,104463
740	1440	0,1	0,35	30	600	10	0,1	0,220708	0,104463
6938	1440	0	0,35	30	600	10	0,05	0,220708	0,104463
6958	1440	0,05	0,35	30	600	10	0,05	0,220708	0,104463
6977	1440	0,1	0,35	30	600	10	0,05	0,220708	0,104463
13175	1440	0	0,35	30	600	10	0,01	0,220708	0,104463
13195	1440	0,05	0,35	30	600	10	0,01	0,220708	0,104463
13214	**1440**	**0,1**	**0,35**	**30**	**600**	**10**	**0,01**	**0,220708**	**0,104463**
16872	29901	0	0,4	90	100	30	0,01	0,219502	0,137197
16892	29901	0,05	0,4	90	100	30	0,01	0,219502	0,137197
758	1440	0,15	0,35	30	600	10	0,1	0,218579	0,104179
6995	1440	0,15	0,35	30	600	10	0,05	0,218579	0,104179
13232	1440	0,15	0,35	30	600	10	0,01	0,218579	0,104179
16911	29901	0,1	0,4	90	100	30	0,01	0,217879	0,136912
2550	19954	0	0,4	60	100	50	0,1	0,215165	0,143402
2570	19954	0,05	0,4	60	100	50	0,1	0,215165	0,143402
8787	19888	0	0,4	60	100	50	0,05	0,214511	0,14437
8807	19888	0,05	0,4	60	100	50	0,05	0,214511	0,14437
2780	2877	0	0,35	60	600	10	0,1	0,213949	0,108562

Tabla 11: Orden modelos Random Forest.

Se ha observado que el p-valor no tiene efecto diferenciador en el modelo con mejor porcentaje de acierto, ya que independientemente del valor (evidentemente dentro de los que se han probado) el resultado no varía si el resto de parámetros es idéntico, por lo que nuestro mejor

modelo obtiene un porcentaje de acierto del 22.07%. Dentro de los 9 modelos que obtienen dicho porcentaje, seleccionaremos como el mejor aquel cuyo intervalo sea menor y además tenga el p-valor más pequeño, ya que así obtenemos el modelo más restrictivo.

Por lo tanto, las características del mejor modelo de Random Forest son las siguientes:

- Número máximo de árboles 30
- Tamaño de hoja mínimo 600
- Profundidad máxima 10
- P-valor 0.01
- Punto de corte inferior 0.1
- Punto de corte inferior 0.35

Resultados obtenidos por el modelo:

- Porcentaje de acierto = 22,07%
- Porcentaje de investigados = 10,44%
- Numero de parámetros = 1440

Las características más importantes dentro del modelo seleccionado serían el número de veces que aparece cada variable dentro de las normas de división de los árboles, y la importancia de estas, por lo que se mostrará dicha información mediante el Gráfico 4 mostrado a continuación:

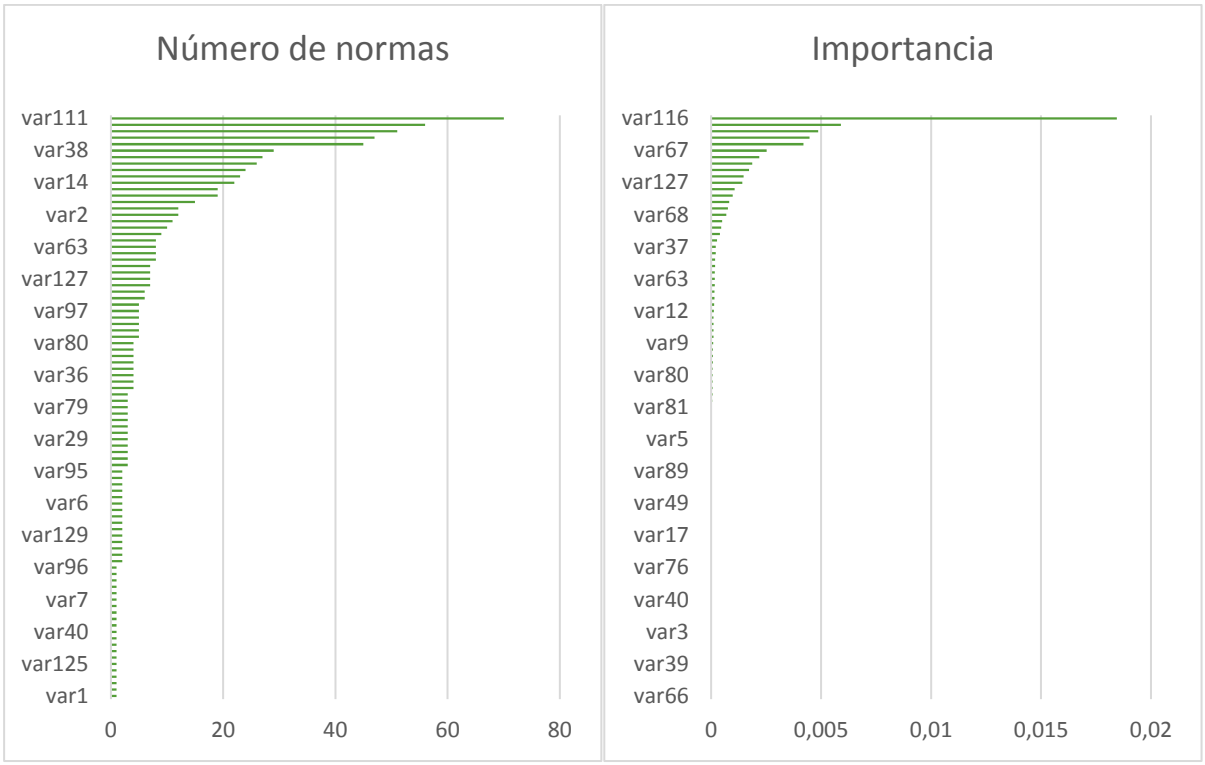

Gráfico 4: Numero de normas e Importancia de las variables en el modelo Random Forest.

En primer lugar, sería destacable el hecho de que la variable var111 es la que aparece en el mayor número de normas, con un total de 70 normas de división, la siguiente variable sería la var115 con 56 apariciones, a partir de este punto el número de apariciones se va reduciendo hasta las 21 variables que únicamente aparecen una vez en cada norma, además para el gráfico se han dejado fuera a 39 variables que no aparecen dentro del modelo.

En segundo lugar, si tenemos en cuenta la importancia de estas variables, destaca claramente sobre el resto la variable 116, la cual acumula más del 30% de la importancia relativa de todas las variables, en contrapunto tenemos la variable 66, la cual tiene una importancia prácticamente nula (hasta el propio software la redondea a 0), aunque esta variable aparece en una norma de división dentro del modelo.

Como final, se muestran en la Tabla 12 las 39 variables que no aportan ninguna información dentro del modelo obtenido y por lo tanto han quedado excluidas de este:

var58	**var22**	**var21**	**var121**	**var57**	**var74**	**var31**
var73	**var119**	**var46**	**var94**	**var11**	**var69**	**var33**
var70	**var84**	**var103**	**var123**	**var53**	**var30**	**var66**

var59	**var25**	**var28**	**var124**	**var101**	**var87**	**var90**
var23	**var102**	**var41**	**var117**	**var42**	**var108**	**var106**
var35	**var43**	**var122**	**var72**			

Tabla 12: Variables que no entran en el modelo de Random Forest.

6.4. Gradient Boosting.

Una vez terminados los modelos anteriores, pasaremos a realizar la búsqueda del mejor modelo de Gradient Boosting, los cuales se definirán mediante los parámetros siguientes:

- Tamaño de la hoja. (2000, 1500, 1000, 500, 100)
- Profundidad máxima. (25, 50, 100)
- Parámetro Shrinkage. (0,01, 0,05, 0,09, 1,3 y 1,7)
- Numero de iteracciones. (10, 50, 100)
- Número máximo de hojas por norma. (2, 5, 10)
- Número máximo de árboles. (100, 200, 500)

Dentro de cada parámetro se han indicado los valores que se han utilizado para la generación de modelos, por lo que se han calculado 2025 modelos distintos, los cuales, una vez obtenidos se les aplicaron los distintos puntos de corte, por lo que al final se obtuvo una tabla con 467.775 sub-modelos distintos tras aplicar los puntos de corte.

A continuación se muestran los 20 primeros modelos con un porcentaje de investigados superior al 10% ordenados en función del porcentaje de acierto en la Tabla 13:

Modelo	Numrules	CorteMin	CorteMax	leafsize	maxdepth	shrinkage	iterations	maxbranch	Maxrees	Acierto	Investigados
15026	46876	0	0,5	100	25	0,05	100	2	100	0,254246	0,11397
15046	46876	0,05	0,5	100	25	0,05	100	2	100	0,2535	0,113856
15065	46876	0,1	0,5	100	25	0,05	100	2	100	0,24798	0,112718
11559	9928	0	0,4	500	25	0,05	100	2	100	0,247065	0,116361
11579	9928	0,05	0,4	500	25	0,05	100	2	100	0,247065	0,116361
4628	3223	0	0,35	1500	25	0,05	100	2	100	0,245653	0,101503
4648	**3223**	**0,05**	**0,35**	**1500**	**25**	**0,05**	**100**	**2**	**100**	**0,245653**	**0,101503**

14794	24873	0	0,45	100	25	0,05	50	2	100	0,245606	0,119834
14814	24873	0,05	0,45	100	25	0,05	50	2	100	0,245606	0,119834
14833	24873	0,1	0,45	100	25	0,05	50	2	100	0,244402	0,119492
12252	9945	0	0,4	500	25	0,09	100	2	100	0,244141	0,116589
12483	9945	0	0,4	500	25	1,3	10	2	100	0,244141	0,116589
12714	9945	0	0,4	500	25	1,3	50	2	100	0,244141	0,116589
12945	9945	0	0,4	500	25	1,3	100	2	100	0,244141	0,116589
13176	9945	0	0,4	500	25	1,7	10	2	100	0,244141	0,116589
13407	9945	0	0,4	500	25	1,7	50	2	100	0,244141	0,116589
13638	9945	0	0,4	500	25	1,7	100	2	100	0,244141	0,116589
12272	9945	0,05	0,4	500	25	0,09	100	2	100	0,243771	0,116532
12503	9945	0,05	0,4	500	25	1,3	10	2	100	0,243771	0,116532
12734	9945	0,05	0,4	500	25	1,3	50	2	100	0,243771	0,116532

Tabla 13: Orden modelos Gradient Boosting

Observamos como el mayor porcentaje de acierto se corresponde con el modelo 15.026, el cual tiene un 25,42% de acierto, por lo que los 17 primeros modelos quedarían dentro del rango del 1% en la diferencia en el porcentaje de acierto, y por lo tanto debemos considerar otros aspectos de los modelos a parte del porcentaje de acierto, y teniendo en cuenta esto, observamos como el modelo 4648 (24.56% de acierto) tiene 3.223 parámetros, frente a los 46.876 del modelo anterior, por lo que dada la alta diferencia entre estos dos modelos, nuestro mejor modelo de Gradient Boosting será el modelo 4.648, el cual analizaremos en el siguiente apartado.

Por lo tanto, las características del mejor modelo de Gradient Boosting son las siguientes:

- Tamaño de hoja mínimo 1500.
- Profundidad máxima 25.
- Parámetro shrinkage 0,05.
- Número de iteracciones 100.
- Número de divisiones 2.
- Número de árboles 100.
- Punto de corte inferior 0,05.
- Punto de corte inferior 0,35.

Resultados obtenidos por el modelo:

- Porcentaje de acierto = 24,5%.
- Porcentaje de investigados = 10,15%.
- Numero de parámetros = 3.223.

Para analizar el modelo generado, tendremos en cuenta dos parámetros, que coinciden con los analizados en el modelo de Random Forest, es decir, el número de normas en las que aparece cada variable y su importancia relativa dentro del modelo, ambos parámetros se muestran a continuación en el Gráfico 5:

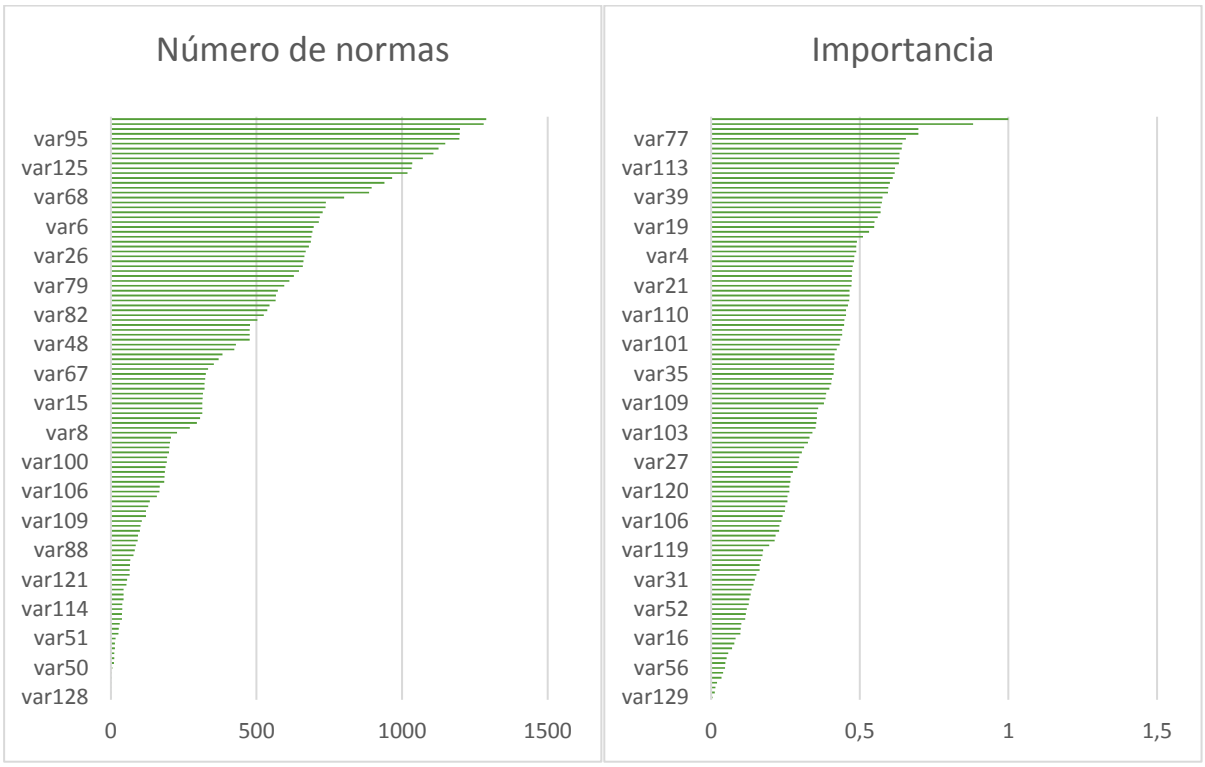

Gráfico 5: Numero de normas e importancia Gradient Boosting.

En lo referente al parámetro que hace referencia al número de normas en las que aparece cada variable observamos cómo no hay una variable que destaque fuertemente sobre el resto, sino que las dos variables con mayor frecuencia (var77 y var78 con 1289 y 1280 normas respectivamente) no difieren mucho de la siguiente variable de la lista (var98 con 1199 normas), esta reducción se mantiene hasta llegar a las variables var128, var86 y var129, las cuales solo aparecen en una norma.

El otro aspecto a tener en cuenta es la importancia de las variables, donde podemos ver como la variable var116 destaca sobre el resto, seguida de la variable var111, mientras que en

el resto de variables la importancia se va reduciendo de manera gradual sin grandes saltos entre cada una de ellas y la siguiente con menor importancia, hasta llegar a las tres últimas variables con menor importancia, las cuales coinciden con el parámetro anterior, es decir, las 3 variables que aparecen una única vez en una norma de división también son las tres con menor importancia relativa.

En último lugar, a continuación, se muestran las variables que no aparecen en ninguna decisión, y por lo tanto su importancia es 0 y no entrarían dentro del modelo seleccionado.

Variable	Normas	Importancia
var99	0	0
var102	0	0
var105	0	0
var89	0	0
var90	0	0
var108	0	0
var69	0	0
var93	0	0
var96	0	0
var30	0	0

Tabla 14: Variables fuera del modelo Gradient Boosting.

6.5. Ensamblado.

Una vez obtenidos los mejores modelos de cada familia, en este apartado obtendremos combinaciones de estos para lograr obtener modelos que mejoren las predicciones. Para generar las combinaciones utilizaremos algunas funciones matemáticas aplicadas a las distintas probabilidades de fraude obtenidas por los distintos modelos. Las combinaciones que utilizaremos serán:

- Media: a partir de las puntuaciones de los distintos modelos para cada observación calcularemos la media y la utilizaremos como probabilidad de fraude.
- Mediana: similar al caso anterior, pero utilizando la Mediana entre los modelos incluidos.
- Máximo: Como el caso anterior, pero se selecciona la probabilidad mayor de los modelos incluidos y esta se utiliza como probabilidad de fraude.

- Mínimo: Idéntico al máximo, pero seleccionando la probabilidad más pequeña.

Evidentemente, para seguir con los patrones de los modelos, una vez obtenida la probabilidad de cada modelo, se aplicarán los distintos puntos de corte para poder seleccionar los intervalos que el modelo considera fraudulentos y cuales no serían investigados.

Modelo	CorteMin	CorteMax	Ensamblado	Acierto	Investigado
4635	0	0,7	mediasinRF	0,265929	0,108107
4655	0,05	0,7	mediasinRF	0,265929	0,108107
4674	0,1	0,7	mediasinRF	0,265543	0,10805
4692	0,15	0,7	mediasinRF	0,263992	0,107822
8789	0	0,5	maximosinRN_RL	0,263596	0,102585
8809	0,05	0,5	maximosinRN_RL	0,263596	0,102585
8828	**0,1**	**0,5**	**maximosinRN_RL**	**0,263596**	**0,102585**
2784	0	0,55	mediasinRL	0,263524	0,106285
2804	0,05	0,55	mediasinRL	0,263524	0,106285
2823	0,1	0,55	mediasinRL	0,263524	0,106285
3938	0	0,5	medianasinRN	0,263274	0,102926
3958	0,05	0,5	medianasinRN	0,263274	0,102926
3014	0	0,5	medianasinRL	0,2631	0,103211
3034	0,05	0,5	medianasinRL	0,2631	0,103211
3053	0,1	0,5	medianasinRL	0,2631	0,103211
4709	0,2	0,7	mediasinRF	0,262963	0,107594
3977	0,1	0,5	medianasinRN	0,262867	0,102869
2841	0,15	0,55	mediasinRL	0,262084	0,106
8846	0,15	0,5	maximosinRN_RL	0,261547	0,1023
3995	0,15	0,5	medianasinRN	0,260821	0,102585

Tabla 15: Orden modelos Ensamblado.

Para entender la Tabla 15 hay que tener en cuenta que el parámetro "Ensamblado" sirve para identificar que cálculo y con qué modelos se ha realizado cada registro, con lo que la primera parte hace referencia al cálculo, y la parte posterior a la palabra "sin" se refiere a los modelos que no se han incluido en el cálculo, donde RL equivale a Regresión Logística, RN se corresponde con las Redes Neuronales, RF con el modelo de Random Forest y por último, aunque no aparece en la tabla, GB que sería la nomenclatura utilizada para los modelos de Gradient Boosting, por lo que explicado esto, ya sería posible analizar la tabla.

En primer lugar, el porcentaje de acierto más elevado se obtiene mediante el modelo "mediasinRL" con puntos de corte 0 y 0,7, es decir, el modelo que utiliza como probabilidad de fraude la media entre la probabilidad de los modelos de Redes Neuronales, Random Forest y Gradient Boosting.

Otro modelo a tener en cuenta sería el modelo 8828, ya que obtiene un porcentaje de acierto casi idéntico (se reduce únicamente en 0,235), y para obtenerlo no es necesario el cálculo del modelo re Redes Neuronales, por lo que este modelo sería preferible frente al anterior.

Además de los modelos que aparecen en la tabla, tenemos que tener en cuenta que existen modelos que difieren menos de un 1% en la probabilidad de fraude comparados con el modelo que obtiene mayor valor que no aparecen en la tabla, uno de ellos sería el modelo de Redes Neuronales, ya que este obtenía un porcentaje de acierto de 0,257518.

Por lo tanto, el modelo de Redes Neuronales estaría dentro del rango y utilizaría únicamente un modelo, por lo que el mejor modelo de ensamblado se correspondería con el modelo de Redes Neuronales el cual ya analizamos anteriormente, dado que, aunque no tiene el porcentaje de acierto más elevado, esta dentro del intervalo del 1% en la probabilidad del fraude y para obtener este modelo únicamente es necesario el cálculo del modelo de Redes Neuronales, no como los otros candidatos que requerían de mas de un modelo para su obtención.

6.6. Comparación de Modelos.

Dentro de este apartado debemos comparar los mejores modelos de cada categoría entre sí, para poder obtener el mejor modelo y poder avanzar dentro del estudio, para ello analizaremos la Tabla 16 que se muestra a continuación:

Modelo	**Parámetros**	**CorteMin**	**CorteMax**	**Acierto**	**Investigados**
Logística	23	0,1	0,8	0,227855	0,102186
Redes Neuronales	287	0,05	0,75	0,257518	0,104122
Random Forest	4516	0,05	0,35	0,2213035	0,11704429
Gradient Boosting	3223	0,05	0,35	0,245653	0,101503
Ensamblado	287	0,05	0,75	0,257518	0,104122

Tabla 16: Resultados mejor modelo de cada familia.

En este apartado, seleccionar el mejor modelo no supone un reto complicado, ya que no existe ningún modelo que difiera menos del 1% en la probabilidad de acierto respecto al modelo de Redes Neuronales, por lo que, hasta este momento, nuestro mejor modelo será el de Redes Neuronales, y por lo tanto este será el utilizado como peso dentro de la inferencia de rechazados la cual trataremos en el siguiente apartado.

7. Modelos de predicción (Inferencia de rechazados).

Dentro de este apartado, repetiremos los denominados como mejores modelos de cada una de las categorías anteriores, con la diferencia de que se calcularán bajo la metodología Fuzzy, y por lo tanto en la fase de entrenamiento se incluirán los registros de los que se conoce el valor de la variable "Fraude" (como en la fase anterior), y además se añadirán los registros de los que no se conoce el valor, lo cual no era posible antes. La muestra de datos para la validación se mantiene igual que en el apartado anterior. Para calcular los pesos de cada registro se utilizará la probabilidad de fraude obtenida mediante el mejor modelo de predicción obtenido en el apartado anterior, es decir, el modelo de Redes Neuronales.

7.1. Método Fuzzy.

Al incluir en la fase de entrenamiento todas las variables de las que se desconoce si eran fraudulentas o no, hemos pasado de tener 61346 registros a tener 9.429.230 registros, por lo que los tiempos necesarios para poder calcular los modelos se han disparado, haciendo imposible repetir los modelos anteriores mediante la metodología Fuzzy, por lo que para solventar este problema, será necesario aplicar dicha metodología únicamente a los denominados mejores modelos de cada familia, ya que el tiempo de ejecución por modelo se ampliaba a aproximadamente una semana.

Otro aspecto importante es que, aunque solo se calcule un modelo para cada familia, si será posible aplicar los distintos puntos de corte y por lo tanto comparar entre sí los 231 sub-modelos de cada rama (ya que una vez obtenido el modelo si es posible aplicar los distintos puntos de corte), los cuales iremos calculando en los siguientes apartados.

7.2. Regresión Logística Fuzzy.

Recordemos que el modelo de Regresión Logística seleccionado anteriormente utilizaba un método de selección de variables Stepwise (paso a paso), la función de enlace era la función Logic y el P-Valor utilizado era 0.01, por lo tanto, el nuevo modelo mantendrá estos parámetros. Una vez calculado, los resultados son los siguientes:

El modelo utiliza 88 variables (68.21%) además del parámetro independiente, por lo que a nivel de robustez este modelo es claramente inferior al modelo obtenido sin incluir la metodología Fuzzy (ya que este solo tenía 23 parámetros), la estimación de cada parámetro se muestra a continuación en la Tabla 17:

Parámetro	Estimación	Parámetro	Estimación	Parámetro	Estimación
Independiente	0.5397	**var38**	0.000025	**var81**	-0.0357
var1	0.000313	**var39**	-1.85E-6	**var82**	-1.31E-7
var2	-0.00312	**var40**	2.21E-6	**var85**	-0.2463
var3	-0.00079	**var42**	-0.00001	**var86**	11.127
var4	0.00213	**var43**	-0.00002	**var89**	-0.4100
var5	0.00171	**var44**	4,16E-03	**var95**	-0.00311
var6	0.1793	**var46**	-0.1112	**var96**	0.00305
var8	-0.00302	**var47**	0.3861	**var97**	0.00102
var9	-0.0141	**var49**	0.2435	**var98**	-0.00090
var10	-0.0116	**var50**	-0.4426	**var106**	-0.00428
var12	0.0172	**var51**	0.5294	**var107**	0.00481
var13	0.0148	**var52**	0.1437	**var109**	-33.160
var15	0.0319	**var53**	0.2463	**var111**	45.486
var16	-0.6110	**var54**	-0.1280	**var112**	167.977
var17	0.00998	**var57**	0.0673	**var113**	31.895
var18	-0.00896	**var58**	-0.0609	**var114**	156.291
var19	-0.0205	**var59**	-0.0658	**var115**	-12.290
var20	0.0186	**var61**	-0.8480	**var116**	16.267
var21	-0.00001	**var62**	19.855	**var117**	-63.507
var22	-0.00003	**var63**	-46.348	**var118**	31.740
var25	-0.00003	**var65**	13.552	**var120**	11.719
var26	-0.00003	**var67**	-0.0160	**var121**	167.536
var27	0.000072	**var68**	-44.034	**var122**	-1.09E-6
var28	0.000378	**var70**	-1.79E-6	**var123**	2,54E-03
var29	0.000154	**var72**	-0.00007	**var124**	0.00475

var30	0.000011	**var75**	-0.00535	**var127**	-0.2083
var31	-0.00005	**var76**	0.6366	**var128**	**-0.5510**
var32	0.000015	**var77**	0.5533		
var33	0.000222	**var78**	-0.6442		
var35	0.000276	**var79**	-25.147		
var37	-0.00002	**var80**	32.800		

Tabla 17: Estimación parámetros Logística Fuzzy.

Una vez analizado el modelo, pasamos a comprobar los resultados con los distintos puntos de corte. Para ello disponemos de la Tabla 18 donde podemos observar como dentro de los modelos que cumplen las condiciones necesarias para ser válidos, el mejor porcentaje de acierto es del 8.57%, por lo que si tenemos en cuenta que en la etapa de validación tenemos un 7.85% de probabilidad de acierto si seleccionamos las observaciones de manera aleatoria, parece evidente que este modelo no está aportando información destacable a la hora de predecir los registros fraudulentos, y es muy inferior a los modelos anteriormente calculados, por lo que parece evidente que la inclusión de la metodología Fuzzy dentro del modelo de Regresión Logística no está mejorando en absoluto los resultados obtenidos anteriormente.

Por último y como dato curioso, podemos observar como el octavo mejor modelo es aquel que considera fraudulentas todas las observaciones (y por lo tanto tiene un acierto del 7.85%), lo cual dejaría a todos los modelos con menor porcentaje de acierto como modelos peores que seleccionar mediante muestreo aleatorio el 10% de la población a investigar.

VP	**FP**	**FN**	**VN**	**CorteMin**	**CorteMax**	**Acierto**	**Investigados**
753	**8028**	**8159**	**626**	**0,95**	**1**	**0,085753**	**0,499886**
1113	12680	3507	266	0,9	1	0,080693	0,78521
1245	14314	1873	134	0,85	1	0,080018	0,885745
1352	15637	550	27	0,75	1	0,079581	0,967152
1311	15199	988	68	0,8	1	0,079406	0,939884
1359	15915	272	20	0,7	1	0,078673	0,983377
1378	16175	12	1	0,05	1	0,078505	0,99926
1379	16187	0	0	0	1	0,078504	1
1376	16172	15	3	0,1	1	0,078413	0,998975
1375	16169	18	4	0,2	1	0,078374	0,998748
1375	16170	17	4	0,15	1	0,07837	0,998805
1374	16159	28	5	0,35	1	0,078367	0,998121
1374	16163	24	5	0,3	1	0,078349	0,998349

1374	16167	20	5	0,25	1	0,078331	0,998577
1362	16030	157	17	0,65	1	0,078312	0,990095
1368	16123	64	11	0,55	1	0,078212	0,99573
1365	16089	98	14	0,6	1	0,078206	0,993624
1369	16141	46	10	0,5	1	0,078184	0,996812
1370	16156	31	9	0,4	1	0,07817	0,997723

Tabla 18: Puntos de corte Regresión Logística Fuzzy.

Dados los resultados de la Tabla 18, observamos como el modelo obtenido no ofrece buenos resultados y por lo tanto estaríamos ante un modelo con una capacidad de predicción muy baja o nula, por lo que a priori, podemos considerar que la metodología Fuzzy no funciona correctamente para el mejor modelo de Regresión Logística obtenido en la primera etapa, donde algunas de las posibles causas que podrían explicar este hecho podrían ser:

- El modelo no es capaz de predecir bien la variable objetivo a través de una relación lineal dado el aumento en el número de registros.
- El modelo seleccionado no es un buen candidato para aplicar la metodología Fuzzy.
- Los modelos de Regresión Logística no son buenos candidatos para aplicar la metodología Fuzzy,

Por último, y repitiendo la metodología del modelo seleccionado anteriormente como el mejor modelo de Regresión Logística, calcularemos la importancia de cada variable multiplicando el valor medio de cada variable por el valor del parámetro asociado a dicha variable, dado el aumento de variables incluidas en el modelo, los resultados se muestran gráficamente en el Gráfico 6, mientras que la tabla correspondiente se encuentra en la Tabla 6 de los anexos:

Gráfico 6: Importancia variables Regresión Logístic Fuzzy.

Observamos como destacan sobre el resto la importancia de las variables var123 y var44, siendo su importancia muy superior al resto de variables incluidas en el modelo.

7.3. Redes Neuronales Fuzzy.

Para analizar la importancia de cada variable dentro de estos modelos recordemos que era necesario repetir el modelo eliminando cada una de las variables, y teniendo en cuenta que el tiempo necesario para el cálculo de dicho modelo ha sido superior a 10 días, es imposible realizar 129 modelos con una duración aproximada similar, por lo que para el cálculo de dicha importancia ha sido necesario recurrir a una muestra de 1.078.800 registros, para así obtener una estimación de los resultados, los resultados se muestran a continuación de manera gráfica:

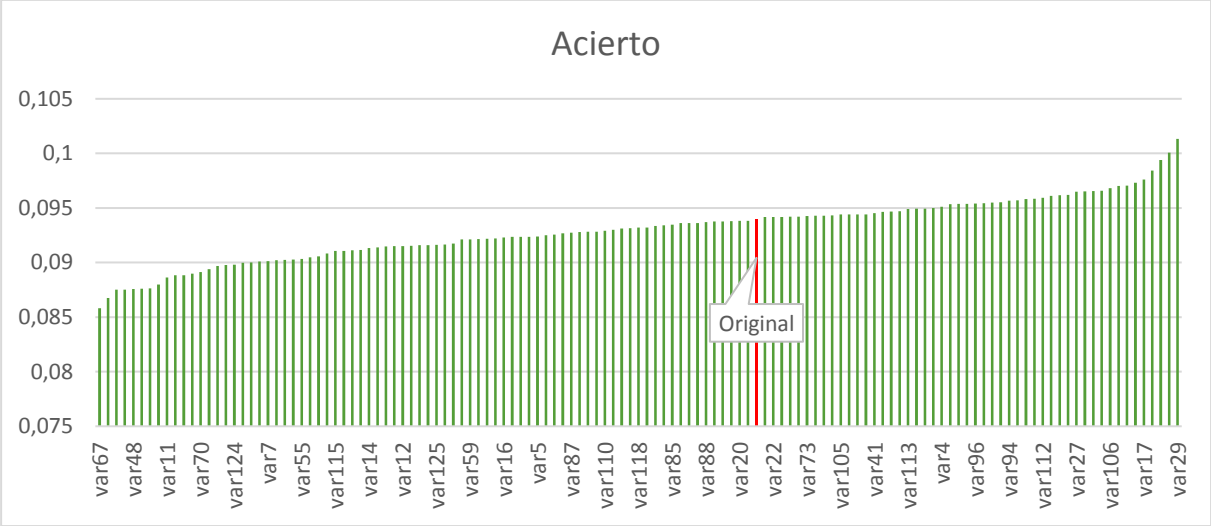

Gráfico 7: Importancia relativa Redes Neuronales Fuzzy.

Observamos en el Gráfico 7, como quedarían los resultados en función del porcentaje de acierto para el mejor modelo de Redes Neuronales tras aplicarle la metodología Fuzzy y eliminar en cada ejemplo la variable indicada. Observamos como la variable que provoca una reducción más notable en el porcentaje de acierto es la variable 67, por lo que dentro de esta metodología podría considerarse como la variable más importante del modelo, mientras que en el lado opuesto tendríamos la variable 29, ya que no solo no empeora los resultados, sino que los mejoraría, aunque no es único de esta variable, ya que ocurre con 50 de ellas. Por lo que si no existieran las actuales limitaciones en el hardware utilizado, sería posible investigar si el modelo obtenido se mejoraría tras eliminar una o varias variables del modelo.

Por último, en el Gráfico 5 mostrado a continuación se muestran conjuntamente los porcentajes de acierto (ordenados por esta variable) e investigados tras eliminar cada una de las variables, y observamos la misma relación que encontrábamos en el modelo original de Redes Neuronales, en la cual en los modelos que aumentaba el porcentaje de investigados tras eliminar una variable, en media también ser reducía el porcentaje de investigados.

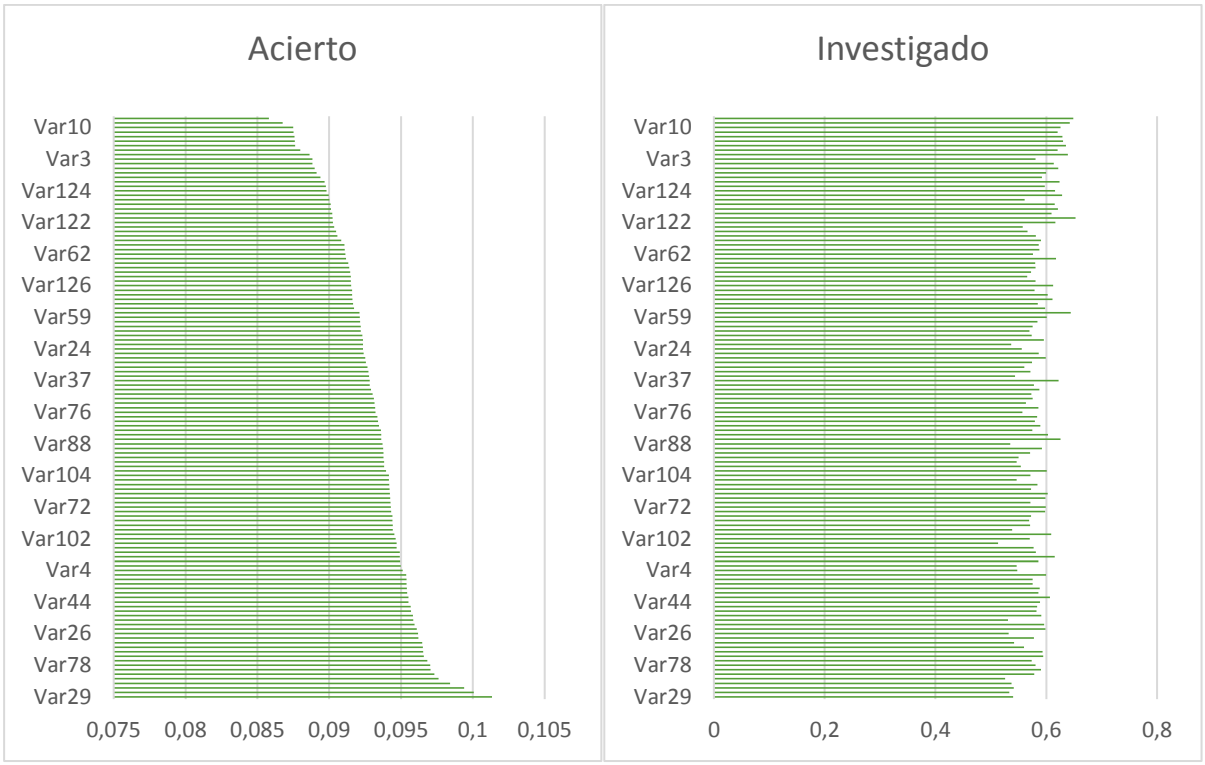

Gráfico 8: Porcentaje de acierto e investigados Redes Neuronales Fuzzy.

Dentro de los modelos de Redes Neuronales, los resultados obtenidos a partir del mejor modelo se muestran a continuación:

PuntoDeCorteMin	PuntoDeCorteMax	acierto	Investigados
0,75	0,95	0,126074	0,139075
0,7	0,95	0,123297	0,171297
0,8	0,95	0,121833	0,105602
0,65	0,95	0,120604	0,207218
0,7	0,9	0,119878	0,130593
0,65	0,85	0,119435	0,132984
0,65	0,8	0,119328	0,101617
0,6	0,95	0,117867	0,243425
0,65	0,9	0,117265	0,166515
0,6	0,85	0,115747	0,16919
0,6	0,8	0,114829	0,137823
0,6	0,9	0,114575	0,202721
0,75	1	0,113821	0,287089
0,7	1	0,113567	0,31931
0,55	0,95	0,113438	0,28003
0,65	1	0,112981	0,355232

0,6	1	0,111984	0,391438
0,5	0,95	0,110895	0,322384
0,8	1	0,110438	0,253615
0,55	0,85	0,110097	0,205795

Tabla 19: Puntos de corte Redes Neuronales Fuzzy.

Como observamos, el mejor modelo mediante la técnica Fuzzy dentro de los que superan el 10% de investigados, obtiene un 12,35% de acierto, siendo esta muy inferior al 25.75% que obteníamos con las Redes Neuronales iniciales, por lo que es evidente que la técnica Fuzzy no mejora el modelo inicial de Redes Neuronales.

7.4. Random Forest Fuzzy.

Una vez calculado el nuevo modelo mediante el método Fuzzy, obtenemos las tablas del número de normas en las que aparece cada variable del modelo y su importancia dentro de dicho modelo:

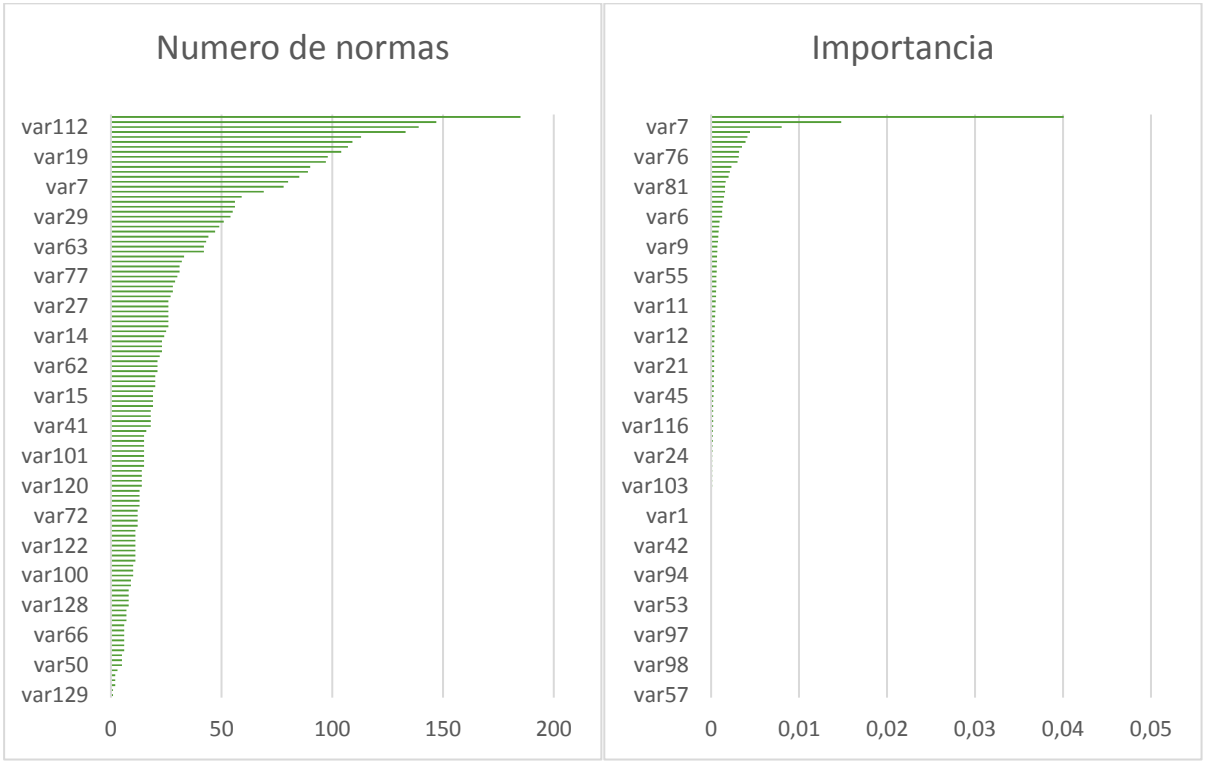

Gráfico 9: Numero de normas e importancia Random Forest Fuzzy.

En primer lugar y a la vista de los resultados, debemos destacar las variables var30, var118, var74, var69, var110, var109, var58, var87, var56, var117, var119 y var89, ya que son las únicas que no aparecen en ninguna norma a lo largo del modelo y por lo tanto su importancia es 0.

Una vez destacadas las variables menos influyentes de nuestro modelo, pasaremos a analizar el gráfico que hace referencia al número de normas en las que aparece cada variable en el modelo, donde destaca la var48 que aparece en un total de 185 normas, cifra que se va reduciendo hasta las variables 129 y 90 que son las únicas que aparecen en una única norma, por último, si analizamos el cómputo global, obtenemos que nuestro modelo dispone de 3674 normas distintas y que cada variable, en media aparece en 28,26 de estas normas.

En el gráfico de las importancias, vuelve a destacar la variable var48, ya que con una importancia de 0,040084 representa más del 30% del total, si además tenemos en cuenta la variable var3 y var7, obtenemos que entre estas 3 variables representan casi el 50% de la importancia total del modelo.

Una vez analizado el modelo y obtenido los resultados, podemos comprobar las distintas opciones en función del punto de corte, para ello utilizaremos la Tabla 20 en el cual se muestran los 20 mejores resultados:

VP	FP	FN	VN	CorteMin	CorteMax	Acierto	Investigados
640	2833	13354	739	0,3	0,6	0,184279	0,197711
667	2953	13234	712	0,3	0,8	0,184254	0,20608
667	2953	13234	712	0,3	0,85	0,184254	0,20608
667	2953	13234	712	0,3	0,9	0,184254	0,20608
667	2953	13234	712	0,3	0,95	0,184254	0,20608
667	2953	13234	712	0,3	1	0,184254	0,20608
666	2952	13235	713	0,3	0,75	0,18408	0,205966
659	2922	13265	720	0,3	0,65	0,184027	0,20386
664	2946	13241	715	0,3	0,7	0,183934	0,205511
618	2760	13427	761	0,3	0,55	0,182948	0,192303
578	2702	13485	801	0,3	0,5	0,17622	0,186724
509	2629	13558	870	0,3	0,45	0,162205	0,178641
405	2414	13773	974	0,3	0,4	0,143668	0,16048

254	1816	14371	1125	0,3	0,35	0,122705	0,117841
1059	7612	8575	320	0,25	0,8	0,122131	0,493624
1059	7612	8575	320	0,25	0,85	0,122131	0,493624
1059	7612	8575	320	0,25	0,9	0,122131	0,493624
1059	7612	8575	320	0,25	0,95	0,122131	0,493624
1059	7612	8575	320	0,25	1	0,122131	0,493624
1058	7611	8576	321	0,25	0,75	0,122044	0,49351

Tabla 20: Puntos de corte Random Forest Fuzzy.

Observamos como todos ellos tienen un porcentaje de registros investigados válido, por lo que seleccionaríamos el modelo con mejor porcentaje de acierto, el cual es aquel con puntos de corte entre 0.3 y 0.6, obteniendo un 18.42% de acierto, muy inferior al 25.4152% conseguido mediante el mismo modelo sin utilizar la metodología Fuzzy.

7.5. Gradient Boosting Fuzzy.

Para el análisis del modelo de Gradient Boosting, analizaremos el número de veces que aparece cada variable en las distintas normas de división, y su importancia dentro del modelo, para ello analizaremos el grafico siguiente:

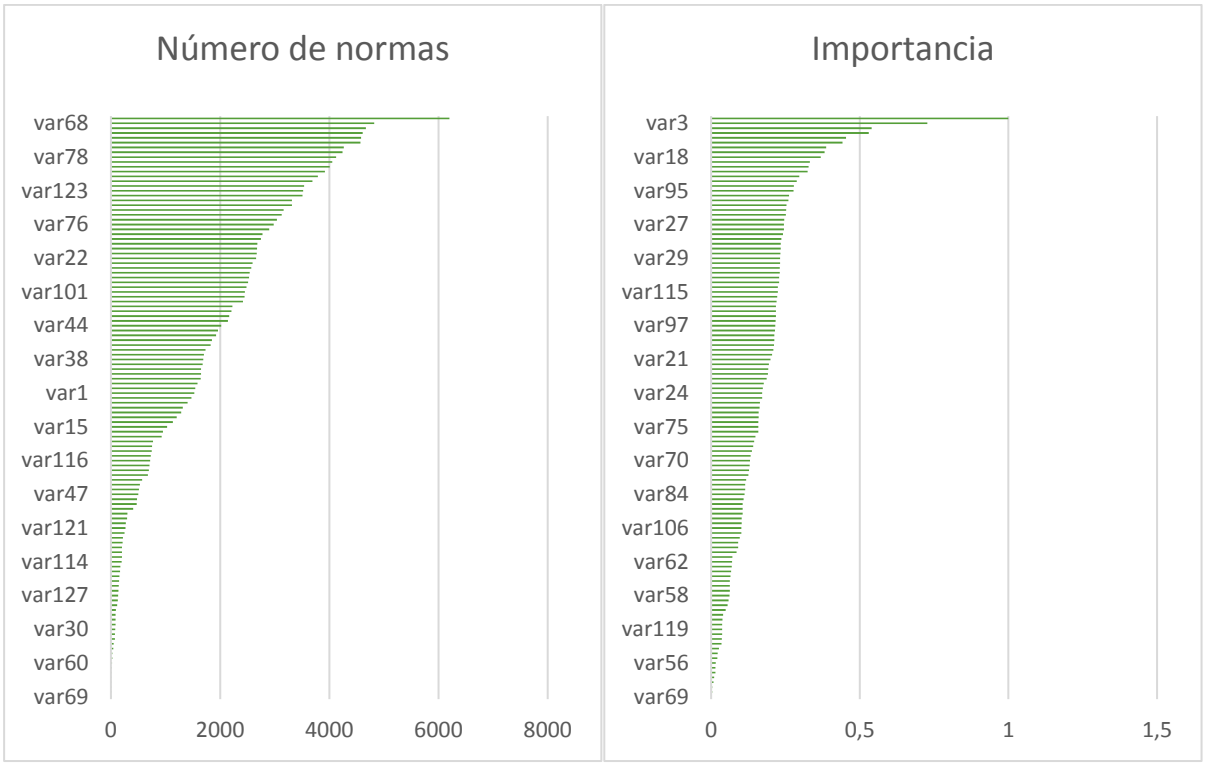

Gráfico 10: Numero de normas e importancia Gradient Boosting Fuzzy.

Observamos cómo respecto al número de normas, destaca la variable var95 con un total de 6202 apariciones, muy por delante del segundo de la lista que sería la variable var68 con 4822 apariciones, a partir de la cual la reducción es gradual hasta llegar a la variable var69, que es la única que aparece únicamente en una norma de las 192154.

El otro aspecto a tener en cuenta es la importancia de cada una de las variables. En el grafico se observa como destaca sobre el resto la variable var19, la cual, si tenemos en cuenta el número de normas, aparece en 2671 normas, por lo que a pesar de ser la variable más importante, se situaría como la variable número 29 en función del número de normas, si este análisis lo realizamos a la inversa, vemos que la variable que aparece en el mayor número de normas sería la decimonovena variable con mayor importancia.

Por último, destacar que las variables var99, var89, var102, var90, var93, var105, var108 y var96 no aparecen en ninguna de las normas (por lo que su importancia también es 0) y por lo tanto son las únicas variables que no aportan información dentro del modelo.

Tras analizar el modelo, analizaremos los porcentajes de acierto para los 20 mejores modelos que cumplen la condición de tener al menos al 10% de observaciones investigadas, los resultados se muestran a continuación:

VP	FP	FN	VN	CorteMin	CorteMax	% Acierto	% Investigados
1238	**1369**	**14818**	**141**	**0,15**	**0,85**	**0,474875**	**0,148412**
1238	1369	14818	141	0,15	0,9	0,474875	0,148412
1238	1369	14818	141	0,15	0,95	0,474875	0,148412
1238	1369	14818	141	0,15	1	0,474875	0,148412
1237	1369	14818	142	0,15	0,8	0,474674	0,148355
1236	1369	14818	143	0,15	0,75	0,474472	0,148298
1227	1369	14818	152	0,15	0,7	0,47265	0,147785
1207	1369	14818	172	0,15	0,65	0,468556	0,146647
1190	1368	14819	189	0,15	0,6	0,465207	0,145622
1142	1364	14823	237	0,15	0,55	0,455706	0,142662
1061	1356	14831	318	0,15	0,5	0,438974	0,137595
992	1340	14847	387	0,15	0,45	0,425386	0,132756
878	1307	14880	501	0,15	0,4	0,401831	0,124388
783	1266	14921	596	0,15	0,35	0,382138	0,116646
628	1191	14996	751	0,15	0,3	0,345245	0,103552
1355	3283	12904	24	0,1	0,85	0,292152	0,264033
1355	3283	12904	24	0,1	0,9	0,292152	0,264033
1355	3283	12904	24	0,1	0,95	0,292152	0,264033
1355	3283	12904	24	0,1	1	0,292152	0,264033

Tabla 21: Puntos de corte Gradient Boosting Fuzzy.

Observamos como el primer modelo (puntos de corte entre 0.15 y 0.85) considera fraudulentos al 14.84% de los registros, y además obtiene un acierto del 47.84% de los casos, muy superior al 25,75% obtenido mediante el modelo de Redes Neuronales y al 24,56% logrado mediante el mejor modelo de Gradient Boosting original, por lo que mediante este modelo hemos logrado superar los resultados anteriores notablemente.

Por último, hay que tener en cuenta el hecho de que el salto en el porcentaje de acierto antes y después de aplicar la metodología Fuzzy es muy alto y aunque se ha tratado de buscar una explicación que justifique este salto en el porcentaje de acierto, no ha sido posible encontrar una explicación lógica más haya de concluir que el modelo de Gradient Boosting utilizado se

ha visto ampliamente beneficiado de la metodología Fuzzy, salvo el hecho de que le número de parámetros se ha disparado hasta 192.441, mientras que hasta este punto, el número de parámetros más alto se encontraba en el modelo de Gradient Boosting original, con 13.956 parámetros, por lo que la diferencia es más que evidente y es posible que el alto poder de predicción del modelo obtenido se deba también al alto número de parámetros.

7.6. Ensamblado modelos Fuzzy.

Para concluir la fase de modelizado, repetiremos el proceso de ensamblado anterior, pero utilizando los resultados obtenidos mediante la metodología Fuzzy, por lo que una vez obtenidos los 4 modelos, se muestran las 20 mejores modelos con un porcentaje de investigados superior al 10% en la Tabla 22 mostrada a continuación:

VP	FP	FN	VN	CorteMin	CorteMax	Variable	acierto	investigados
1234	1352	14835	145	0	0,85	maximosinRF_RN	0,477185	0,147216
1234	1352	14835	145	0,05	0,85	maximosinRF_RN	0,477185	0,147216
1234	1352	14835	145	0,1	0,85	maximosinRF_RN	0,477185	0,147216
1234	1352	14835	145	0,15	0,85	maximosinRF_RN	0,477185	0,147216
1233	1352	14835	146	0,2	0,85	maximosinRF_RN	0,476983	0,147159
1232	1352	14835	147	0,25	0,85	maximosinRF_RN	0,47678	0,147102
1223	1352	14835	156	0,3	0,85	maximosinRF_RN	0,474951	0,14659
1238	1369	14818	141	0	0,85	ProbGradBoos	0,474875	0,148412
1238	1369	14818	141	0,05	0,85	ProbGradBoos	0,474875	0,148412
1238	1369	14818	141	0,1	0,85	ProbGradBoos	0,474875	0,148412
1238	**1369**	**14818**	**141**	**0,15**	**0,85**	**ProbGradBoos**	**0,474875**	**0,148412**
1237	1369	14818	142	0,2	0,85	ProbGradBoos	0,474674	0,148355
1236	1369	14818	143	0,25	0,85	ProbGradBoos	0,474472	0,148298
839	931	15256	540	0	0,75	medianasinRN	0,474011	0,100763
839	931	15256	540	0,05	0,75	medianasinRN	0,474011	0,100763
838	931	15256	541	0,1	0,75	medianasinRN	0,473714	0,100706
836	929	15258	543	0,15	0,75	medianasinRN	0,473654	0,100478
1227	1369	14818	152	0,3	0,85	ProbGradBoos	0,47265	0,147785
842	940	15247	537	0	0,75	minimosinRN_RL	0,472503	0,101446
841	940	15247	538	0,05	0,75	minimosinRN_RL	0,472207	0,101389

Tabla 22: Orden Ensamblado Fuzzy.

El mejor porcentaje de acierto obtenido mediante estos modelos es del 47.71%, el cual se obtiene al calcular la probabilidad de fraude de cada observación como el valor máximo entre la Regresión Logística y los modelos Gradient Boosting, considerando como fraudulentas aquellas observaciones con dicha probabilidad situada entre 0 y 0.85, pero como ya vimos anteriormente (y queda reflejado en esta tabla), si utilizamos únicamente el modelo de Gradient Boosting obtenemos un porcentaje de acierto de 47.48%, por lo que la diferencia en este dato no sería suficiente para utilizar un modelo que requiere del cálculo de dos modelos distintos frente al que requiere únicamente de uno, y por lo tanto, dentro de la metodología Fuzzy, el modelo obtenido mediante Gradient Boosting será considerado como el mejor modelo para este análisis, ya que mediante el ensamblado no ha sido posible mejorarlo sustancialmente.

8. Comparación final modelos.

Una vez ya hemos obtenido tanto los modelos originales como los modelos mediante metodología fuzzy, solo podemos determinar cuál de todos es el que mejores resultados a generado para poder determinar cuál es nuestro mejor modelo y poder pasar a la última etapa del estudio, en la que utilizamos los datos test, que no han sido utilizados a lo largo del estudio para así poder obtener una estimación del porcentaje de acierto e investigados final de este trabajo, para ello en primer lugar compararemos todos los resultados obtenidos anteriormente en la Tabla 23:

Modelo	*Parámetros*	*CorteMin*	*CorteMax*	*Acierto*	*Investigados*
Logística	*23*	*0,1*	*0,8*	*0,227855*	*0,102186*
Redes Neuronales	*287*	*0,05*	*0,75*	*0,257518*	*0,104122*
Random Forest	*4516*	*0,05*	*0,35*	*0,2213035*	*0,11704429*
Gradient Boosting	*13956*	*0,25*	*0,8*	*0,254153*	*0,102812*
Ensamblado	*287*	*0,05*	*0,75*	*0,257518*	*0,104122*
Redes Fuzzy	*376*	*0,95*	*1*	*0,085753*	*0,499886*
Redes Neuronales Fuzzy	*287*	*0,35*	*0,75*	*0,19354839*	*0,01415687*
Random Forest Fuzzy	3961	*0,3*	*0,6*	*0,184279*	*0,197711*
Gradient Boosting Fuzzy	*192441*	*0,15*	*0,85*	*0,474875*	*0,148412*
Ensamblado Fuzzy	*192441*	*0,15*	*0,85*	*0,474875*	*0,148412*

Tabla 23: Tabla resumen modelos finales.

Es evidente, que el modelo con mejor porcentaje de acierto es el modelo obtenido mediante la metodología Fuzzy aplicada al modelo de Gradient Boosting, (empata con el ensamblado Fuzzy pero como ya vimos, ambos son el mismo modelo), el siguiente en la lista sería el modelo de Redes Neuronales original, pero con más de 20 puntos menos en el porcentaje de acierto por lo que no disponemos de ninguna evidencia que permita no concluir que el modelo de Gradient Boosting Fuzzy es el mejor modelo que hemos obtenido a lo largo del estudio, y por lo tanto este debe de ser el modelo final utilizado para predecir la probabilidad de fraude de los clientes en las siguientes etapas del estudio, las cuales veremos a continuación.

Tras aplicar el modelo final, dentro de la muestra de datos Test obtenemos los siguientes resultados:

VP	VN	FP	FN	% Acierto	%Investigados
1491	936	13027	5764	0,61433869	0,11438401

Tabla 24: Resultados fase Test.

Se observa como el porcentaje de acierto ha aumentado respecto a la fase anterior, llegando hasta el 61,43%, por lo que ya tendríamos el porcentaje de acierto estimado final de este estudio, además conocemos el porcentaje de investigados final es del 11,43%.

9. Conclusiones.

Como acabamos de observar, el parámetro más importante de este estudio (es decir el porcentaje de acierto) ha logrado resultados muy superiores al valor del cual partía este estudio, el cual era del 10.690%, mientras que el valor alcanzado por nuestros modelos se sitúa en el 61,43%, por lo que en este aspecto podemos concluir que el objetivo de mejorar el modelo del que se partía ha sido logrado con mucho margen, el otro aspecto que podríamos considerar es que nuestro modelo también mantiene el porcentaje de investigados por encima del 10% por lo que este sería otro aspecto positivo para el estudio.

Otro aspecto que se ha analizado como algo positivo, aunque tiene sus inconvenientes es el alto porcentaje de acierto obtenido por el modelo final en la etapa test, ya que se desconocen las causas que provocan tanta diferencia entre los valores en la etapa de validación y la etapa test, a continuación de muestran los distintos valores en cada una de las etapas:

Etapa	**% Acierto**	**% Investigados**
Entrenamiento	0,663151	0,113694
Validación	*0,474875*	*0,148412*
Test	*0,614339*	0,114384

Tabla 25: Resultados test por etapas.

Es evidente que al existir cierta diferencia en las distintas etapas (principalmente entre la etapa de validación y el resto) y el hecho a que se desconocen las causas de esta diferencia puede provocar cierta incertidumbre, aunque a priori se podría explicar por el hecho de que el porcentaje de investigados en la etapa de validación es superior frente al resto de etapas, por lo que si suponemos que le porcentaje de registros fraudulentos se mantiene constante en cada etapa, la capacidad de acierto se reduce ante un aumento en el porcentaje de investigados, pero no se ha demostrado que esta sea la causa por lo que desconocemos cómo evolucionará el modelo si se aplicase a nuevos datos, y en este tipo de estudios la incertidumbre no es una cualidad deseada.

Otro aspecto que se ha detectado, es el bajo éxito de los modelos de ensamblado, ya que en ningún momento se han mejorado los modelos obtenidos mediante las distintas técnicas utilizadas (aunque en la primera fase si se obtenía mayor porcentaje de acierto, este no era suficiente para justificar la utilización de estos modelos), es importante indicar que dadas las infinitas combinaciones para obtener los ensamblados, es posible que existan modelos que mejorasen significativamente la probabilidad de acierto, pero no se han detectado mediante las pruebas realizadas.

9.1. Futuras líneas de investigación.

A lo largo del estudio se han tratado ciertas líneas de investigación con el fin de encontrar respuesta a los objetivos marcados, pero es evidente que a lo largo del camino no solo

se encontraban respuestas, sino que surgían nuevas preguntas, por lo que en este apartado pretendemos explicar dichas preguntas.

En primer lugar, destacaríamos la necesidad de poder calcular todos los modelos con los datos de entrenamiento al completo, ya que como hemos visto en los modelos de Random Forest Fuzzy y Gradient Boosting Fuzzy, fue necesario obtener los modelos utilizando una muestra ya que el hardware utilizado hacía imposible calcular los modelos con todos los datos, y por lo tanto, se desconoce qué resultados se obtendrían, y dado que esta casuística afecta nada más y nada menos que al resultado final del estudio, sería un aspecto a tener en cuenta.

Otro aspecto que se podría analizar, y que dadas las limitaciones en el hardware utilizado es la implementación de las interacciones dentro de los modelos.

La segunda línea de investigación a tener en cuenta, y la cual si se pudiera aplicar llegaría incluso a ser más importante que la primera, consistiría en poder aplicar la metodología para obtener el mejor modelo de cada familia pero aplicada a los modelos de inferencia de rechazados, ya que como vimos a lo largo del estudio, solo fue posible aplicar dicha metodología al mejor modelo de cada familia, dejando fuera la posibilidad de repetir todos los modelos y encontrar el que mejor funcionase.

Para continuar, otra línea posible de investigación a tener en cuenta se correspondería con ampliar los modelos estadísticos utilizados, ya que al utilizar un número finito de valores dentro de los parámetros que definen los modelos, quedan abiertas infinitas opciones con modelos que no se han tenido en cuenta.

Por último, existe una línea de investigación igualmente interesante, la cual consistiría en probar nuevas familias de modelos predictivos que en este estudio no se han tenido en cuenta, como por ejemplo los Análisis Discriminantes, Super Vector Machine o modelos Knn o futuros modelos estadísticos que se desarrollen en el futuro, así como multiples opciones dentro de la inferencia de rechazados, tales como la metodología Hard Cutoff o la metodología Parceling Augmentatio.

10. Bibliografía.

- Técnicas avanzadas de predicción / Autor Cesar Pérez editorial Garceta grupo editorial.

- EL SISTEMA ESTADISTICO SAS / Autor Cesar Pérez editorial Garceta grupo editorial
- Redes de neuronas artificiales: un enfoque práctico / Autores Inés M. Galván León y Pedro Isasi Viñuela editorial Pearson
- An introduction to statistical learning : with applications in R / Gareth James editorial Springer
- Decision trees for business intelligence and data mining [Recurso electrónico] : using SAS Enterprise Miner / Barry de Ville editorial SAS Institute
- The elements of statistical learning : data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman editorial Springer
- http://support.sas.com/documentation/onlinedoc/miner/em43/neural.pdf
- https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect004.htm
- http://documentation.sas.com/?docsetId=emhpprcref&docsetTarget=emhpprcref_hpforest_overview.htm&docsetVersion=14.2&locale=en
- https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_tree_sect004.htm

11. Anexos.

Variable	Media	Mínimo	Cuartil 1	Mediana	Cuartil 3	Máximo	Asimetría	Kurtosis
var1	76,0642058	0	22	45	82	2187725	2341,71	6190593,12
var2	126,000364	0	83,169758	97,9368705	113,319932	246015,66	376,890724	159256,24
var3	132,146463	0	96,6374776	114,135017	127,669887	100240,2	202,176358	54539,26
var4	132,663318	0	100,989144	108,409194	137,261476	70821,52	146,592517	34861,74
var5	132,125144	0	104,786546	120,2904	122,206218	50100,04	92,742439	18047,72
var6	0,9999026	0	0,3564018	0,7580622	1,3064477	1098,63	58,1646733	33722,12
var7	2,0683209	1	1	1	1	24	4,2729044	17,5358865
var8	13,3617236	1	1	17	23	24	-0,2321215	-1,7806418
var9	10,5744901	1	3	9	19	24	0,2864016	-1,3792396
var10	12,4631195	1	5	13	21	24	-0,0664178	-1,4387478
var11	1,2072225	1	1	1	1	24	10,5382938	134,287721
var12	-0,3680776	-1	-1	-1	-1	24	4,4957621	22,2924981
var13	-5,9485604	-24	-12	-6	1	1	-0,1990659	-1,0784379
var14	-2,5276498	-23	-4	-2	-1	1	-0,9491538	1,4269201
var15	-2,9167868	-23	-4	-3	-1	1	-0,8470987	1,4603057
var16	0,1299186	-1	-1	1	1	24	1,9790446	13,7131136
var17	72,5675791	0	44,6843487	58,4324592	78,8745014	28839,54	70,7571296	14838,57
var18	71,2581453	0	50,0809048	63,0110111	72,6465377	5712,69	13,1126083	232,097392
var19	70,1173569	0	59,1180238	60,3790785	63,0353298	4169,33	12,4035792	167,342577

var20	70,3543535	0	59,3738702	60,7765207	61,9642566	3177,5	12,2424153	159,194233
var21	24511,47	0	13461	18778,67	26396,33	270893954	1842,58	4481461,79
var22	24116,29	0	13570,67	18861,57	26062,83	69967781,5	261,440834	258303,48
var23	23729,79	0	13502,58	18657	25640,77	38864085	91,9277839	35786,57
var24	63331,56	0	27114	42971,33	67926,33	503710524	925,963377	1215155,09
var25	63757,71	0	28831,83	44800,43	69011,83	503710524	949,102621	1400518,22
var26	64293,42	0	30559	46382,38	69387,92	503710524	976,837505	1476821,31
var27	8746,36	0	3746	5932,33	9375	69477314	925,609296	1214532,63
var28	8809,93	0	3985,67	6186,33	9527,5	69477314	948,536697	1399398,81
var29	8889,37	0	4227,17	6409,64	9588	69477314	975,419075	1473889,41
var30	2807,82	0	0	0	0	38775000	664,327677	871480,99
var31	3607,47	0	0	0	0	38775000	620,749647	805189,1
var32	4387,52	0	0	0	0	38775000	446,606982	536007,41
var33	303,769441	-82714	88,5	186	334	4799565	1720,46	4083143,27
var34	297,89103	-82714	94	189,666667	328,714286	1244561	223,507021	199979,71
var35	293,759145	-82714	97,2	190	324,083333	732363	84,3353907	30039,16
var36	1589,3	0	0	0	0	434233210	1371,41	2336025,38
var37	1821,59	0	0	0	0	434233210	1362,62	2312374,85
var38	1992,73	0	0	0	1404,38	434233210	1352,38	2304549,51
var39	63377,53	0	30275	46500	70000	503710500	932,724083	1226887,53
var40	61715,12	0	30883,33	46616,67	68583,33	503710500	960,725714	1423263,13

var41	60345,38	0	31107,14	45925	66891,67	503710500	991,374671	1506120,91
var42	23451,46	0	7449,33	15509,67	27661,33	63851838	217,509799	202624,99
var43	22725,89	0	7793,17	15623,33	26890,83	63851838	147,249789	148219,14
var44	22114,75	0	7922,92	15345	26043,67	63851838	158,825102	170567,05
var45	0,2995494	0	0	0	0	6	2,9216702	8,5252379
var46	0,5541666	0	0	0	0	12	3,1292889	10,5274119
var47	1,0355017	0	0	0	1	24	3,4189818	13,642226
var48	8494304,01	1	8413403	8533018	8771998	8958782	-8,8447649	81,4377372
var49	0,5815378	0	0	1	1	6	1,393553	4,2350022
var50	0,6103145	0	0	1	1	12	2,7855206	15,5714788
var51	0,6375439	0	0	1	1	24	5,1592071	53,1004773
var52	3,1494928	1	3	3	3	6	2,0746624	8,5902587
var53	6,1561175	1	6	6	6	12	0,8892653	8,0251687
var54	11,8808149	1	12	12	12	24	-0,2271539	5,8541171
var55	0,1268994	0	0	0	0	6	2,9101064	9,1754416
var56	0,0477172	0	0	0	0	21	9,1429289	127,765859
var57	0,1258795	0	0	0	0	29	6,2949522	63,7240836
var58	0,233894	0	0	0	0	49	5,0857668	46,2655092
var59	0,4152689	0	0	0	0	65	4,2370664	36,0625858
var60	0,0163926	0	0	0	0	5	8,3544928	76,0102656
var61	0,1011382	0	0	0	0	1	2,6716233	5,9905666

var62	0,1050373	0	0	0	0	1	2,6760278	6,4156625
var63	0,1104061	0	0	0	0,1111111	1	2,7058446	6,9892998
var64	0,1917119	0	0	0,25	0,3333333	1	1,2580832	2,6034638
var65	0,106702	0	0	0,1428571	0,1666667	1	3,0010257	13,8610725
var66	0,0643177	0	0	0,0714286	0,0833333	1	4,8492959	30,7937034
var67	6,6010665	1	1	2	2	66	2,5787984	5,7828813
var68	0,1081577	0,0704062	0,0917832	0,1126352	0,1196189	0,1516366	-0,2934042	-0,5404454
var69	1586,48	0	0	0	0	38775000	827,106541	1160269,83
var70	23847,5	0	6646	15050	28002	106642779	385,555807	563285,9
var71	24601,46	0	13145	18480	26471	539020673	3061,87	11153642,7
var72	304,81764	-374143	77	177	333	9544073	2896,71	10349660,2
var73	8444,2	0	3354	5608	9052	104833202	1204,5	2130219,97
var74	1181,9	0	0	0	0	434233210	1379,17	2352399,71
var75	69,5538766	0	20,0579741	44,6855654	79,3219914	419863,56	333,724586	449851,46
var76	-0,131012	-1	-0,5220099	-0,2940429	0,3333272	1	0,4269363	-1,2837691
var77	-0,0148088	-0,9999631	-0,1631442	0	0,1446445	0,9988576	-0,2081221	-0,1062042
var78	-0,0537714	-0,9999631	-0,1820125	-0,0452175	0,0676236	0,9994472	-0,2005282	0,4438986
var79	-0,4376465	-1	-0,5220099	-0,5220099	-0,4654051	1	1,9563619	5,2829898
var80	-0,4177633	-1	-0,5220099	-0,389932	-0,3333272	1	0,6106978	4,8193883

var81	23,0198114	0	25	25	25	25	-2,8234589	6,6990196
var82	34141,2	0	0	11121	45145	765835336	1396,34	2483041,54
var83	35531,17	0	0	11123	46207	772506430	1367,87	2425019,93
var84	36799,2	0	0	11751	46782	772506430	1315,79	2297352,63
var85	-0,1036237	-1	0	0	0	1	-1,5281794	3,6013259
var86	0,0965937	0	0	0	0	1	2,7312197	5,4595617
var87	0,0441225	0	0	0	0	1	4,4396341	17,7103536
var88	0,3930526	0	0	0	0	24	5,7508348	36,3473584
var89	0,9582707	0	1	1	1	1	-4,583398	19,00754
var90	0,9582707	0	1	1	1	1	-4,583398	19,00754
var91	31,6903284	0	9,2231593	18,4869323	33,5002488	770376,68	2259,76	5881842,74
var92	22,8557791	0	6,7137518	12,0365852	21,88653	144935,9	345,208656	344581,31
var93	31,6903284	0	9,2231593	18,4869323	33,5002488	770376,68	2259,76	5881842,74
var94	33,8869472	0	11,0305361	20,632976	35,5882795	487047,67	1704,52	4007182,64
var95	26,2758615	0	8,54244	14,6325756	25,675791	91281,39	164,099343	77975,11
var96	33,8869472	0	11,0305361	20,632976	35,5882795	487047,67	1704,52	4007182,64
var97	26,0773695	0	5,8594653	13,2035349	26,5769324	1546955,18	2654,66	7307916,06
var98	18,4530184	0	4,2428902	8,705804	17,0752213	296811,75	1142,88	2250082,26
var99	26,0773695	0	5,8594653	13,2035349	26,5769324	1546955,18	2654,66	7307916,06
var100	69,5674616	0	22,5	45,1666667	77,8333333	282396,75	240,701607	221697,68
var101	69,2442433	0	22,6463856	45,4472593	78,0779277	176720,13	82,8716573	30410,84

var102	69,5674616	0	22,5	45,1666667	77,8333333	282396,75	240,701607	221697,68
var103	69,0963286	0	23,2857143	45,3333333	76,9230769	173275	91,6911107	34842,8
var104	69,0169321	0	23,4072068	45,5113461	77,1701607	176720,13	80,6692281	29548,6
var105	69,0963286	0	23,2857143	45,3333333	76,9230769	173275	91,6911107	34842,8
var106	71,479222	0	21,3333333	44,6666667	79,6666667	1098965	1778,04	4267896,95
var107	69,4638165	0	21,7639142	45,1221872	78,6042543	216643,26	159,03215	104557,72
var108	71,479222	0	21,3333333	44,6666667	79,6666667	1098965	1778,04	4267896,95
var109	0,1039641	0,1031025	0,1031025	0,1031025	0,1031025	0,4699903	27,5842341	788,773003
var110	0,1038375	0,0555556	0,1030879	0,1030879	0,1030879	0,4662841	20,4005966	471,810162
var111	0,1083321	0,0367688	0,0862543	0,1074541	0,1180068	0,2179575	0,5611801	-0,29213
var112	0,1082561	0,0091575	0,0994648	0,1126352	0,1196189	0,1304223	-1,2798544	4,272339
var113	0,1065865	0,0576923	0,0942166	0,1081918	0,1119536	0,1294672	0,1394352	-0,7230603
var114	0,107584	0,0688054	0,1096653	0,1096653	0,1096653	0,2258065	-3,7456352	18,4186387
var115	0,1080777	0,0409922	0,0822321	0,1055269	0,1377953	0,2378886	0,7273293	0,2184382
var116	0,100923	0,0275229	0,0942726	0,0942726	0,0942726	0,4085779	3,1286775	11,2611511
var117	0,1072042	0,0876623	0,1075983	0,1075983	0,1075983	0,2396694	13,7533731	512,520513
var118	0,1059154	0,1031072	0,1031072	0,1031072	0,1031072	0,1956122	5,5157487	28,5003915
var119	0,1075618	0,0348837	0,1076319	0,1076319	0,1076319	0,16	-8,1372221	450,562373
var120	0,1122935	0,0636329	0,1073732	0,1073732	0,1152951	0,2142411	2,6860752	9,0947103
var121	0,106626	0,0340136	0,1064911	0,1064911	0,1064911	0,1624365	-2,9089044	30,6696932
var122	61220,18	0	24319	40658	65629	760040727	1204,5	2130219,24

var123	65109,79	0	28800	45700	70400	760040700	1187,93	2090497,52
var124	72,5675791	0	19	43	81	2187725	2997,38	10839016,7
var125	-68,8248215	-3881400	-18,6708861	0	11,407767	100	-779,795432	1160363,1
var126	-47,8984693	-1455462,5	-19,2982456	0	12,529274	100	-345,312956	254323,69
var127	0,2718872	0	0	0	0	3	2,8383095	7,2433758
var128	0,1109103	0	0	0	0	1	2,4781121	4,1410401
var129	0,121675	0	0	0	0	1	2,3145511	3,3571474

Tabla Anexo 1: Descriptivos 129 variables.

NAME	NRULES	IMPORTANCE
var19	2671	1
var3	3315	0,727297
var112	4264	0,539998
var7	705	0,529739
var68	4822	0,453301
var48	2143	0,441433
var81	2210	0,387427

var2	3691	0,382057
var18	2744	0,369233
var4	3040	0,332265
var43	1682	0,32793
var17	3165	0,324494
var5	1924	0,296964
var121	265	0,287783
var125	4568	0,278347
var95	6202	0,277172
var94	4613	0,261957
var67	2164	0,259768
var9	3315	0,253742
var126	4007	0,252757
var10	2900	0,25191
var26	2543	0,245797
var27	745	0,244873
var111	2446	0,244702
var77	4581	0,241879
var23	3125	0,236835
var98	4672	0,234485
var20	1653	0,234385

var44	2019	0,233171
var29	1588	0,231919
var113	2527	0,23152
var91	4054	0,231442
var41	3789	0,230449
var76	2983	0,228992
var78	4124	0,228511
var34	1705	0,224712
var115	2571	0,223944
var40	3511	0,222526
var92	4240	0,220513
var71	2596	0,218325
var39	3538	0,218225
var25	1959	0,21763
var114	196	0,217062
var97	3920	0,215447
var22	2658	0,215284
var79	2513	0,212724
var38	1690	0,212163
var123	3523	0,211216
var80	2484	0,209467

var104	2674	0,205024
var21	2682	0,199678
var28	953	0,194083
var107	2775	0,191839
var8	1314	0,191421
var6	2417	0,187056
var101	2454	0,177296
var42	1647	0,174096
var24	2225	0,172232
var116	717	0,171668
var1	1528	0,164501
var73	1543	0,162589
var63	1202	0,160057
var82	1850	0,159615
var103	1404	0,15921
var75	1824	0,158752
var35	1735	0,158328
var72	1288	0,148829
var13	732	0,144563
var54	696	0,141745
var33	1648	0,137708

var11	85	0,133127
var70	1472	0,130326
var51	93	0,130225
var14	1133	0,127188
var36	408	0,125762
var15	1028	0,117888
var120	474	0,114828
var61	248	0,114195
var84	772	0,112024
var100	932	0,109187
var32	271	0,106414
var66	675	0,106182
var85	200	0,105575
var127	134	0,103311
var55	203	0,102913
var106	756	0,10222
var47	503	0,101801
var37	532	0,096733
var59	573	0,091116
var83	515	0,090988
var122	475	0,085574

var53	216	0,071365
var62	304	0,07073
var65	219	0,069897
var110	127	0,067686
var88	169	0,064893
var64	202	0,063165
var52	148	0,062591
var12	175	0,062573
var58	298	0,061993
var45	117	0,059013
var46	145	0,05527
var57	155	0,048866
var74	71	0,039897
var124	135	0,037651
var16	87	0,037309
var119	41	0,03708
var109	82	0,036664
var60	23	0,035918
var50	47	0,034624
var31	73	0,026608
var118	26	0,021922

var49	26	0,021111
var56	12	0,016066
var30	78	0,01501
var87	7	0,014068
var86	4	0,010511
var117	5	0,008248
var128	4	0,004234
var129	4	0,00379
var69	1	0,000257
var99	0	0
var89	0	0
var102	0	0
var90	0	0
var93	0	0
var105	0	0
var108	0	0
var96	0	0

Tabla Anexo 2: Gradient Boosting Fuzzy.

Variable	Normas	Importancia			Variable	Normas	Importancia			Variable	Normas	Importancia
var116	158	1			var82	525	0,423002			var31	27	0,147393
var111	887	0,881055			var63	334	0,414995			var62	66	0,142474
var115	738	0,697532			var67	326	0,414815			var121	55	0,136436
var98	1199	0,697434			var81	383	0,414033			var57	53	0,132785
var77	1280	0,654935			var25	566	0,41247			var64	65	0,129006
var95	1197	0,643754			var35	477	0,412092			var53	43	0,126631
var91	1125	0,641584			var44	477	0,406211			var52	43	0,119992
var41	1198	0,633511			var112	477	0,404472			var124	43	0,116883
var92	1148	0,633475			var24	503	0,39782			var74	12	0,114256
var78	1289	0,631856			var5	306	0,387545			var46	39	0,101967
var113	737	0,618315			var88	82	0,384514			var45	30	0,099174
var94	1071	0,617069			var109	106	0,379685			var61	38	0,098018
var125	1033	0,611437			var42	353	0,359608			var16	26	0,082838
var40	1108	0,600947			var20	315	0,355921			var118	14	0,077686
var72	295	0,595741			var73	324	0,355533			var11	11	0,070225
var123	1019	0,594834			var33	314	0,3539			var51	16	0,057243
var39	1035	0,576482			var43	370	0,351175			var127	14	0,052546
var75	424	0,574327			var103	314	0,340779			var49	3	0,047865
var97	939	0,570079			var34	322	0,331542			var56	10	0,046884
var126	966	0,569887			var14	316	0,325813			var87	3	0,040187

var23	895	0,560511			var15	314	0,312625			var50	5	0,034601
var68	801	0,550079			var29	271	0,304786			var117	2	0,020124
var19	689	0,548574			var37	183	0,29686			var128	1	0,014268
var6	696	0,53101			var27	201	0,294508			var86	1	0,012613
var38	478	0,511248			var55	92	0,289755			var129	1	0,004372
var76	714	0,490564			var83	168	0,275322			var99	0	0
var22	727	0,488727			var8	227	0,267663			var102	0	0
var71	573	0,488257			var13	200	0,265942			var105	0	0
var4	537	0,481521			var84	185	0,263435			var89	0	0
var2	717	0,481108			var120	193	0,262699			var90	0	0
var104	661	0,476482			var122	120	0,256702			var108	0	0
var10	686	0,474078			var100	192	0,256374			var69	0	0
var26	665	0,47352			var59	203	0,248729			var93	0	0
var32	93	0,472857			var28	206	0,248405			var96	0	0
var21	692	0,472259			var66	184	0,240331			var30	0	0
var48	429	0,465404			var106	166	0,236505					
var9	680	0,465156			var36	102	0,230329					
var17	659	0,464815			var7	187	0,229287					
var79	595	0,459749			var47	134	0,216646					
var18	669	0,453662			var85	64	0,21383					
var110	128	0,453393			var54	120	0,195298					

var3	613	0,448814			var119	64	0,175041					
var70	321	0,447111			var58	99	0,173232					
var80	646	0,440992			var65	78	0,167644					
var107	545	0,440757			var60	38	0,163601					
var1	628	0,434535			var12	85	0,16348					
var101	567	0,432178			var114	39	0,152702					

Tabla Anexo 3: Modelo Gradient Boosting.

Modelo	**Acierto**	**Investigados**
var23	0,278383	0,07321
var1	0,272169	0,083457
var8	0,272059	0,077422
var7	0,270703	0,081806
var9	0,270597	0,080155
var68	0,269382	0,086645
var77	0,267806	0,079927
var11	0,267717	0,079529
var6	0,265406	0,081293
var69	0,265278	0,081977
var3	0,26506	0,080326

var80	0,264997	0,08636
var13	0,264834	0,091142
var72	0,264665	0,081521
var39	0,264613	0,090573
var41	0,264561	0,081123
var75	0,264124	0,08061
var48	0,262508	0,087612
var64	0,262079	0,077764
var40	0,262036	0,082774
var2	0,261963	0,092793
var78	0,261708	0,08266
var44	0,260899	0,08488
var10	0,260625	0,091085
var83	0,260201	0,085108
var113	0,259552	0,086417
var18	0,259532	0,085108
var54	0,259119	0,090516
var107	0,259112	0,085905
var95	0,258988	0,085506
var20	0,258667	0,085392
var52	0,258573	0,083001

var89	0,258231	0,088182
var90	0,258231	0,088182
var99	0,2582	0,081578
var66	0,258	0,085392
var112	0,257353	0,085165
var22	0,257273	0,099795
var21	0,257227	0,096493
var84	0,257183	0,081236
var85	0,257162	0,083457
var12	0,257126	0,095867
var4	0,257108	0,094102
var74	0,256904	0,088637
var126	0,256776	0,079813
var70	0,256527	0,087214
var87	0,256393	0,086815
var30	0,256307	0,099283
var51	0,256303	0,094842
var32	0,256246	0,097973
var111	0,256173	0,092224
var61	0,256171	0,085335
var47	0,25601	0,087612

var63	0,255927	0,093647
var97	0,255799	0,088352
var127	0,255556	0,0871
var76	0,255347	0,08784
var94	0,255082	0,086815
var96	0,255042	0,087499
var108	0,254839	0,088239
var37	0,254693	0,090971
var86	0,254419	0,090174
var62	0,254195	0,091597
var88	0,254167	0,081977
var31	0,253968	0,089662
var14	0,253941	0,101104
var123	0,253763	0,079415
var115	0,253659	0,093362
var104	0,253477	0,09006
var53	0,253472	0,098372
var103	0,252997	0,090231
var82	0,252964	0,086417
var49	0,252846	0,095013
var98	0,252783	0,086929

var122	0,252512	0,084994
var105	0,252207	0,090288
var45	0,252116	0,094159
var125	0,252113	0,080838
var38	0,252101	0,094842
var117	0,252079	0,088979
var65	0,251995	0,08562
var124	0,251924	0,08135
var43	0,251832	0,085449
var128	0,251651	0,086189
var59	0,251621	0,087783
var28	0,25158	0,09006
var121	0,251473	0,086929
var27	0,251096	0,090914
var42	0,250808	0,088068
var129	0,25066	0,086303
var24	0,250418	0,102072
var114	0,250322	0,088466
var106	0,250321	0,088694
var5	0,250145	0,098315
var26	0,249844	0,09137

var109	0,249695	0,093248
var46	0,249409	0,096322
var73	0,2491	0,094842
var81	0,248954	0,081635
var19	0,248214	0,095639
var25	0,24797	0,105146
var92	0,247928	0,082432
var110	0,247923	0,089093
var119	0,247769	0,082944
var29	0,246006	0,106911
var17	0,245557	0,105716
var15	0,24443	0,089434
var50	0,243984	0,101731
var60	0,243567	0,095127
var91	0,243227	0,092451
var118	0,243114	0,09507
var36	0,242531	0,097176
var34	0,242349	0,091142
var58	0,241443	0,093134
var35	0,241162	0,090174
var79	0,240934	0,092622

var57	0,240912	0,092394
var102	0,240871	0,088865
var100	0,24033	0,089776
var120	0,24	0,092508
var16	0,23741	0,094956
var93	0,236591	0,095525
var55	0,235777	0,097063
var67	0,235608	0,106797
var116	0,230363	0,086246
var56	0,229155	0,103097
var101	0,226404	0,101332
var33	0,22597	0,101275

Tabla Anexo 4: Acierto e investigados tras eliminar variables Redes Neuronales.

Variable	**Acierto**	**Investigados**
29	0,101338	0,540419
32	0,100053	0,533132
28	0,0994	0,541216
19	0,098421	0,537345
17	0,097616	0,525447
78	0,097323	0,578504

91	0,09706	0,59063
77	0,097021	0,580895
105	0,096825	0,573836
92	0,096592	0,594672
90	0,096536	0,593248
58	0,096522	0,559718
27	0,096479	0,541671
61	0,096205	0,577536
26	0,096181	0,532107
18	0,096099	0,598315
111	0,095934	0,596379
25	0,095851	0,530969
94	0,09583	0,591085
97	0,095699	0,582375
93	0,095679	0,583684
44	0,095524	0,588808
40	0,095506	0,606797
43	0,095423	0,585848
95	0,095413	0,588296
88	0,095389	0,575316
89	0,095389	0,575316
68	0,09536	0,599966
4	0,095104	0,547706
23	0,09498	0,546624
65	0,094937	0,585848

46	0,094921	0,615336
112	0,094916	0,58118
96	0,094703	0,577081
2	0,094683	0,512866
101	0,09463	0,570306
41	0,094539	0,60879
34	0,094417	0,538427
99	0,094417	0,570989
39	0,094409	0,569225
104	0,0944	0,572299
69	0,094309	0,598201
71	0,094297	0,598884
100	0,094279	0,571217
72	0,094246	0,5986
42	0,094217	0,602414
102	0,094201	0,572299
107	0,09417	0,583969
22	0,094167	0,54651
103	0,094161	0,571331
71	0,093948	0,601104
31	0,093834	0,553911
20	0,093815	0,546738
33	0,093779	0,549983
106	0,093775	0,570648
83	0,093759	0,591996

87	0,093703	0,535239
45	0,093622	0,625697
80	0,09362	0,603211
82	0,093601	0,574747
84	0,093472	0,589548
30	0,093409	0,579586
98	0,093348	0,583627
75	0,093213	0,556985
117	0,093206	0,585734
21	0,09315	0,563475
108	0,093113	0,575316
79	0,093014	0,572868
109	0,092917	0,587556
116	0,092826	0,577707
37	0,092821	0,622509
35	0,092785	0,543607
86	0,092729	0,571559
1	0,092674	0,56023
74	0,092549	0,574519
76	0,092506	0,599397
5	0,092391	0,586588
24	0,092361	0,55596
119	0,092345	0,536946
120	0,092343	0,595525
16	0,092308	0,573551

110	0,092199	0,569908
73	0,092186	0,575544
81	0,092158	0,584368
59	0,092133	0,600592
13	0,092106	0,644028
60	0,09174	0,598201
64	0,091653	0,584481
124	0,09162	0,611408
56	0,091604	0,602812
113	0,091589	0,578675
125	0,091534	0,612604
12	0,091506	0,580439
126	0,091495	0,565581
15	0,091469	0,572583
63	0,091399	0,580496
14	0,09132	0,579756
127	0,09116	0,617614
62	0,091125	0,575999
38	0,091076	0,587556
114	0,091059	0,586417
54	0,09083	0,590402
36	0,09058	0,58135
85	0,090489	0,566207
55	0,090343	0,557668
121	0,090262	0,61619

47	0,090235	0,652966
128	0,090221	0,60953
7	0,090117	0,620972
122	0,09011	0,615336
118	0,090005	0,561027
57	0,08997	0,628316
123	0,089809	0,616133
6	0,089774	0,597347
9	0,089673	0,624046
115	0,089397	0,592224
70	0,089115	0,599852
8	0,088978	0,621883
66	0,088825	0,613344
3	0,088824	0,580667
11	0,088634	0,639075
49	0,087982	0,620517
53	0,087619	0,635432
51	0,087608	0,630309
48	0,087586	0,62917
50	0,087523	0,620517
10	0,087502	0,625868
52	0,08676	0,642377
67	0,085805	0,648867

Tabla Anexo 5: Acierto e investigados Redes Fuzzy al eliminar una variable.

Variable	Media	Parámetro	Importancia
var1	148,301796	0,000313	0,04641846
var2	226,418659	-0,00312	-0,70642622
var3	235,147666	-0,00079	-0,18576666
var4	240,330491	0,00213	0,51190395
var5	249,380379	0,00171	0,42644045
var6	0,69557275	0,1793	0,12471619
var8	17,5231637	-0,00302	-0,05291995
var9	12,2619242	-0,0141	-0,17289313
var10	13,2727643	-0,0116	-0,15396407
var12	1,20465882	0,0172	0,02072013
var13	-8,6495615	0,0148	-0,12801351
var15	-4,11247677	0,0319	-0,13118801
var16	0,15482672	-0,611	-0,09459913
var17	133,041477	0,00998	1,32775394
var18	134,362586	-0,00896	-1,20388877
var19	135,030526	-0,0205	-2,76812579
var20	136,361842	0,0186	2,53633026
var21	39386,796	-0,00001	-0,39386796
var22	39393,0491	-0,00003	-1,18179147
var25	91439,5623	-0,00003	-2,74318687
var26	92701,4556	-0,00003	-2,78104367
var27	12295,7002	0,000072	0,88529041
var28	12640,1968	0,000378	4,7779944
var29	12818,026	0,000154	1,97397601
var30	895,736784	0,000011	0,0098531
var31	1950,58881	-0,00005	-0,09752944

var32	2666,34606	0,000015	0,03999519
var33	571,015266	0,000222	0,12676539
var35	592,40219	0,000276	0,163503
var37	1142,99361	-0,00002	-0,02285987
var38	1143,98476	0,000025	0,02859962
var39	92130,3989	-0,00000185	-0,17044124
var40	92062,4102	0,00000221	0,20345793
var42	34140,3289	-0,00001	-0,34140329
var43	34305,0259	-0,00002	-0,68610052
var44	33756,6247	0,00416	140,427559
var46	1,60267988	-0,1112	-0,178218
var47	2,6512405	0,3861	1,02364396
var49	0,59263848	0,2435	0,14430747
var50	0,60867864	-0,4426	-0,26940117
var51	0,62512633	0,5294	0,33094188
var52	3,25630033	0,1437	0,46793036
var53	6,35956705	0,2463	1,56636136
var54	12,3092622	-0,128	-1,57558556
var57	0,18632022	0,0673	0,01253935
var58	0,30437192	-0,0609	-0,01853625
var59	0,47593975	-0,0658	-0,03131684
var61	0,28577441	-0,848	-0,2423367
var62	0,26287236	19,855	5,21933075
var63	0,23749841	-46,348	-11,0075765
var65	0,10449494	13,552	1,41611542
var67	3,6311903	-0,016	-0,05809904
var68	0,10716677	-44,034	-4,71898154
var70	34864,4586	-0,00000179	-0,06240738
var72	575,18531	-0,00007	-0,04026297
var75	138,777195	-0,00535	-0,74245799
var76	-0,3227499	0,6366	-0,20546259
var77	-0,11129337	0,5533	-0,06157862

var78	-0,14439764	-0,6442	0,09302096
var79	-0,27058154	-25,147	6,80431393
var80	-0,43416169	32,8	-14,2405036
var81	22,5062922	-0,0357	-0,80347463
var82	54793,0741	-1,31E-07	-0,00717789
var85	-0,10171812	-0,2463	0,02505317
var86	0,24611222	11,127	2,73849063
var89	0,89564112	-0,41	-0,36721286
var95	54,6944408	-0,00311	-0,17009971
var96	58,2860011	0,00305	0,1777723
var97	39,2920154	0,00102	0,04007786
var98	36,3523978	-0,0009	-0,03271716
var106	138,147665	-0,00428	-0,59127201
var107	136,054741	0,00481	0,6544233
var109	0,10898419	-33,16	-3,61391585
var111	0,1074336	45,486	4,88672491
var112	0,10720955	167,977	18,0087387
var113	0,1063429	31,895	3,39180693
var114	0,10729998	156,291	16,770021
var115	0,10749807	-12,29	-1,32115128
var116	0,10768506	16,267	1,75171292
var117	0,1073542	-63,507	-6,81774343
var118	0,10793608	31,74	3,42589116
var120	0,11110328	11,719	1,30201931
var121	0,10704877	167,54	17,9345232
var122	88718,8323	-0,00000109	-0,09670353
var123	93172,8116	0,00254	236,658941
var124	138,937388	0,00475	0,65995259
var127	0,85412252	-0,2083	-0,17791372
var128	0,28560949	-0,551	-0,15737083

Tabla Anexo 6: Importancia variables Regresión Logística Fuzzy.